# EDB Postgres Distributed -PGD

リリース *6.4.0*

**EnterpriseDB Corporation**

**2026 年 06 月 04 日**

# 目 次

**第 1 章 foundation** **3**
1.1 基盤 変換された PostgreSQL . . . 3
1.2 継続的な事業継続アクティブ-アクティブ . . . 3
1.3 コミットスコープを介した一貫性の簡素化 . . . 3
1.4 詳細なデータの局所性と主権 . . . 4
1.5 自動化された競合管理 . . . 4
1.6 サブスクライバ専用ノードを使用したグローバル読み取りスケーリング . . . 4
1.7 管理対象データのライフサイクル: AutoPartition と階層化テーブル . . . 5
1.8 シームレスなライフサイクル管理 . . . 5

**第 2 章 EDB Postgres Distributed 6 release notes** **7**
2.1 EDB Postgres Distributed 6.4.0 release notes . . . 7
2.1.1 ハイライト . . . 7
2.1.2 機能 . . . 8
2.1.3 機能強化 . . . 8
2.1.4 バグ修正 . . . 10
2.2 EDB Postgres Distributed 6.3.1 release notes . . . 13
2.2.1 バグ修正 . . . 13
2.3 EDB Postgres Distributed 6.3.0 release notes . . . 14
2.3.1 ハイライト . . . 14
2.3.2 機能 . . . 14
2.3.3 機能強化 . . . 15
2.3.4 の変更 . . . 16
2.3.5 バグ修正 . . . 16
2.4 EDB Postgres Distributed 6.2.0 release notes . . . 19
2.4.1 ハイライト . . . 19
2.4.2 機能 . . . 19
2.4.3 機能強化 . . . 20
2.4.4 バグ修正 . . . 21
2.5 EDB Postgres Distributed 6.1.2 release notes . . . 24
2.5.1 バグ修正 . . . 24
2.6 EDB Postgres Distributed 6.1.1 release notes . . . 24
2.6.1 バグ修正 . . . 24
2.7 EDB Postgres Distributed 6.1.0 release notes . . . 25
2.7.1 ハイライト . . . 25
2.7.2 機能 . . . 26
2.7.3 機能強化 . . . 27
2.7.4 の変更 . . . 28
2.7.5 バグ修正 . . . 28
2.8 EDB Postgres Distributed 6.0.2 release notes . . . 30

2.8.1 バグ修正 . . . 30
2.9 EDB Postgres Distributed 6.0.1 release notes . . . 31
2.9.1 ハイライト . . . 31
2.9.2 機能 . . . 31
2.9.3 機能強化 . . . 33
2.9.4 の変更 . . . 36
2.9.5 バグ修正 . . . 37

**第 3 章 Overview** **39**
3.1 Known issues and limitations . . . 39
3.1.1 既知の問題 . . . 39
3.1.2 制限事項 . . . 41
3.2 PGD compatibility . . . 45
3.2.1 PGD の PostgreSQL バージョンとの互換性 . . . 45
3.2.2 オペレーティングシステムおよびアーキテクチャとの PGD の互換性 . . . 46
3.3 Postgres Distributed terminology . . . 46
3.3.1 非同期レプリケーション . . . 47
3.3.2 コミットスコープ . . . 47
3.3.3 CAMO またはコミット・アット・モスト・ワンス . . . 47
3.3.4 競合 . . . 47
3.3.5 接続マネージャー . . . 47
3.3.6 コンセンサス . . . 47
3.3.7 クラスター . . . 48
3.3.8 DDL データ定義言語 . . . 48
3.3.9 DML データ操作言語 . . . 48
3.3.10 耐久性 . . . 48
3.3.11 熱心な . . . 48
3.3.12 最終的な一貫性 . . . 48
3.3.13 フェイルオーバー . . . 48
3.3.14 グループコミット . . . 48
3.3.15 即時一貫性 . . . 49
3.3.16 遅延制御 . . . 49
3.3.17 論理レプリケーション . . . 49
3.3.18 論理スタンバイノード . . . 49
3.3.19 ノード . . . 49
3.3.20 ノードグループ . . . 49
3.3.21 PGD クラスター . . . 50
3.3.22 PGD ノード . . . 50
3.3.23 物理的レプリケーション . . . 50
3.3.24 Postgres クラスター . . . 50
3.3.25 クォーラム . . . 50
3.3.26 クォーラムコミット . . . 50
3.3.27 レプリケートされた利用可能なフォールトトレランス Raft . . . 51
3.3.28 読み取りスケーラビリティ . . . 51
3.3.29 サブスクリプション . . . 51
3.3.30 スイッチオーバー . . . 51
3.3.31 同期レプリケーション . . . 51

3.3.32 サブスクライバー専用ノード . . . 51
3.3.33 2 フェーズコミット 2PC . . . 52
3.3.34 垂直スケーリングまたはスケールアップ . . . 52
3.3.35 Witness ノード . . . 52
3.3.36 ライトリーダー . . . 52
3.3.37 ライター . . . 52
3.4 Quickstart guide . . . 52
3.4.1 Creating your first cluster . . . 53
3.4.2 Working with SQL and the PGD cluster . . . 55
3.4.3 Loading data into your PGD Cluster . . . 58
3.4.4 Using PGD CLI . . . 60

**第 4 章 Planning your PGD deployment 71**
4.1 Deployment patterns . . . 71
4.1.1 パターンの選択 . . . 71
4.1.2 アーキテクチャ要素 . . . 73
4.2 Choosing a Postgres distribution . . . 93
4.3 Choosing your deployment method . . . 94
4.4 Comparing replication solutions . . . 94
4.4.1 展開 . . . 95
4.4.2 パフォーマンス . . . 95
4.4.3 耐久性 . . . 96
4.4.4 一貫性 . . . 96
4.4.5 データ分散 . . . 96
4.4.6 メンテナンス . . . 97
4.4.7 接続管理 . . . 97
4.4.8 DDL サポート . . . 98
4.4.9 Postgres の互換性 . . . 98
4.4.10 管理 . . . 99
4.4.11 ライセンス . . . 99
4.5 Choosing between Physical Streaming Replication (PSR) and EDB Postgres Distributed (PGD) . . . 99
4.5.1 EFM または Patroni を使用した PSR . . . 100
4.5.2 EDB Postgres Distributed PGD . . . 100
4.6 Other considerations . . . 101
4.6.1 Sizing . . . 101
4.6.2 Time and PGD . . . 102

**第 5 章 Managing lifecycle operations 103**
5.1 Deploying and configuring EDB Postgres Distributed . . . 103
5.1.1 Manually . . . 103
5.1.2 Deploying and configuring PGD using TPA . . . 118
5.1.3 Deploying and configuring PGD on Kubernetes . . . 118
5.1.4 Deploying and configuring PGD on Hybrid Manager . . . 119
5.2 Upgrading . . . 119
5.2.1 PGD のアップグレード . . . 119
5.2.2 Postgres または Postgres と PGD メジャーバージョンのアップグレード . . . 120
5.2.3 その他のアップグレード . . . 120

5.3 Testing and tuning PGD clusters . . . 152
5.3.1 pgd_bench . . . 152
5.3.2 pgd_bench の使用に関する注意事項 . . . 153
5.4 Backup and recovery . . . 154
5.4.1 論理バックアップとリストア . . . 155
5.4.2 物理的なバックアップとリストア . . . 157
5.4.3 最終的な一貫性 . . . 159
5.4.4 モニタリング . . . 161
5.5 Migrating to Postgres Distributed (PGD) . . . 162
5.5.1 Migrating from Postgres Physical Streaming Replication (PSR) to PGD . . . 162

**第 6 章 Configuration and Management 181**
6.1 Node types and capabilities . . . 181
6.1.1 An overview of PGD Node types . . . 181
6.1.2 Witness nodes . . . 182
6.1.3 Logical standby nodes . . . 182
6.1.4 Subscriber-only nodes and groups . . . 184
6.2 Node management . . . 189
6.2.1 Creating PGD nodes . . . 190
6.2.2 Groups and subgroups . . . 192
6.2.3 Creating and joining PGD groups . . . 193
6.2.4 Viewing PGD topology . . . 194
6.2.5 Removing nodes and groups . . . 196
6.2.6 Node restart and down node recovery . . . 197
6.2.7 Automatic synchronization . . . 198
6.2.8 Node UUIDs . . . 201
6.2.9 Replication slots created by PGD . . . 202
6.3 Connection Manager . . . 203
6.3.1 Connection Manager overview . . . 204
6.3.2 Connection Manager Authentication . . . 205
6.3.3 Connection pooling . . . 207
6.3.4 Load Balancing with Connection Manager . . . 209
6.3.5 Configuring Connection Manager . . . 212
6.3.6 Monitoring the Connection Manager . . . 213
6.4 Postgres configuration . . . 215
6.4.1 Postgres 設定 . . . 215
6.4.2 準備されたトランザクションの最大数 . . . 216
6.4.3 グローバル構成に関する考慮事項 . . . 216
6.5 Monitoring . . . 217
6.5.1 Monitoring through SQL . . . 217
6.6 AutoPartition in PGD . . . 232
6.6.1 パーティションの自動作成 . . . 232
6.6.2 AutoPartition の例 . . . 232
6.6.3 RANGE パーティションテーブル . . . 233
6.6.4 パーティションの自動作成の停止 . . . 234
6.6.5 パーティションの作成を待機しています . . . 234
6.6.6 パーティションの検索 . . . 234

6.6.7 自動パーティショニングの有効化または無効化 . . . 235
6.6.8 パーティションの削除またはデタッチ . . . 235
6.7 Commit Scopes . . . 235
6.7.1 のご紹介 . . . 235
6.7.2 コミットスコープの種類 . . . 236
6.7.3 コミットスコープを使用した作業 . . . 236
6.8 Conflict Management . . . 280
6.8.1 Conflicts . . . 281
6.8.2 Column-level conflict detection . . . 298
6.8.3 Conflict-free replicated data types . . . 304

**第 7 章 Advanced functionality** **319**
7.1 DDL replication . . . 319
7.1.1 DDL overview . . . 319
7.1.2 DDL replication options . . . 321
7.1.3 DDL locking details . . . 322
7.1.4 Managing DDL with PGD replication . . . 324
7.1.5 DDL command handling matrix . . . 326
7.1.6 DDL and role manipulation statements . . . 337
7.1.7 Workarounds for DDL restrictions . . . 338
7.1.8 PGD functions that behave like DDL . . . 339
7.2 Decoding worker . . . 340
7.2.1 有効にする . . . 340
7.2.2 無効化 . . . 340
7.3 CDC Failover support . . . 341
7.3.1 背景 . . . 341
7.3.2 CDC フェイルオーバーのサポート . . . 342
7.3.3 CDC フェールオーバーサポートの有効化 . . . 343
7.3.4 Postgres 17+での CDC フェールオーバーのサポート . . . 343
7.3.5 一貫した初期スナップショットの取得 . . . 344
7.3.6 新しく参加したノードへのフェイルオーバー . . . 344
7.3.7 オリジンごとの進行状況の追跡 . . . 344
7.3.8 制限事項 . . . 345
7.4 Parallel Apply . . . 346
7.4.1 パラレル適用とは何ですか? . . . 346
7.5 Replication sets . . . 348
7.5.1 レプリケーションセットを使用する . . . 348
7.5.2 パーティション分割テーブルの動作 . . . 349
7.5.3 外部キーを使用した動作 . . . 349
7.5.4 レプリケーションセットのメンバーシップ . . . 350
7.5.5 レプリケーションセットのリスト . . . 351
7.5.6 DDL レプリケーションフィルタリング . . . 352
7.5.7 選択的レプリケーションの例 . . . 353
7.6 Security and roles . . . 357
7.6.1 Roles . . . 357
7.6.2 Role management . . . 358
7.6.3 PGD predefined roles . . . 359

7.6.4 Roles and replication . . . 364
7.6.5 Access control . . . 365
7.7 Sequences . . . 367
7.7.1 自動シーケンス変換 . . . 368
7.7.2 UUID、KSUUID、およびその他のアプローチ . . . 373
7.7.3 KSUUID v2 ファンクション . . . 374
7.7.4 関連項目 . . . 376
7.8 Replicated large objects . . . 376
7.8.1 レプリケートされたラージオブジェクトを使用する . . . 376
7.8.2 大きなオブジェクトの保存方法 . . . 377
7.8.3 制限事項 . . . 377
7.9 Stream triggers . . . 377
7.9.1 適用中のトリガー実行 . . . 378
7.9.2 行型の用語 . . . 379
7.9.3 競合トリガー . . . 379
7.9.4 Transform トリガー . . . 380
7.9.5 行の内容 . . . 381
7.9.6 実行オーダー . . . 381
7.9.7 ストリームトリガーの例 . . . 381
7.10 Transaction streaming . . . 384
7.10.1 PGD の機能強化 . . . 384
7.10.2 構成 . . . 385
7.10.3 モニタリング . . . 387
7.11 Explicit two-phase commit (2PC) . . . 387
7.11.1 使用する . . . 388

**第 8 章 Guidance for DBAs 389**
8.1 Encrypting columns in PGD . . . 389
8.1.1 暗号化方法の比較 . . . 389
8.1.2 分散クラスターでの暗号化の計画 . . . 390
8.1.3 pgcrypto を使用したカラムの暗号化 . . . 391
8.1.4 DBMS_CRYPTO を使用した列の暗号化 . . . 393
8.1.5 暗号化キーを安全に管理する . . . 394
8.2 Using tiered storage . . . 395
8.2.1 テーブルタイプを理解する . . . 396
8.2.2 階層化ストレージ戦略 . . . 396
8.2.3 メタデータ管理 . . . 397
8.3 Recovering a failed node . . . 408
8.3.1 クラスターの状態の評価 . . . 409
8.3.2 障害が発生したノードのクラスターからの削除 . . . 410
8.3.3 障害が発生したノードの復旧 . . . 411
8.3.4 復旧したノードの検証 . . . 412
8.3.5 クラスターの状態の検証 . . . 413
8.3.6 障害対応 . . . 413
8.3.7 リカバリー後のチェックリスト . . . 413
8.4 Developing applications for geo-replication . . . 414
8.4.1 geo レプリケーションがアプリケーションに影響を与える場合 . . . 414

8.4.2 アプリケーションのジオレプリケーションの準備をする . . . 414
8.5 Setting up access controls . . . 424
8.5.1 Securing inter-node connections . . . 425
8.5.2 Securing client application connections . . . 429
8.6 Handling replication apply errors . . . 432
8.6.1 エラーポリシーの選択 . . . 433
8.6.2 エラーポリシーの構成 . . . 433
8.6.3 無効になったサブスクリプションへの対応 . . . 434
8.6.4 蓄積された障害の監視 . . . 436
8.7 Configuring connection pooling . . . 436
8.7.1 前提条件 . . . 437
8.7.2 接続プーリングの有効化 . . . 437
8.7.3 プールのモニタリング . . . 437
8.7.4 pgBouncer からの移行 . . . 438

**第 9 章 Guidance for Developers** **439**
9.1 Designing applications for PGD . . . 439
9.1.1 Preparing your application for PGD . . . 439
9.1.2 Assessing an existing application for PGD . . . 442
9.2 Managing geo-replication . . . 443
9.2.1 geo レプリケーションを使用する場合 . . . 443
9.2.2 導入ワークフロー . . . 443
9.3 Using connection pooling in your application . . . 467
9.3.1 アクティブプールモードの理解 . . . 467
9.3.2 アプリケーションをトランザクションモードに適応させる . . . 468
9.4 Application use . . . 469
9.4.1 DML and DDL replication and nonreplication . . . 469
9.4.2 Nodes with differences . . . 470
9.4.3 Timing considerations and synchronous replication . . . 472
9.4.4 Using extensions with PGD . . . 473
9.4.5 Use of table access methods (TAMs) in PGD . . . 474
9.4.6 Feature compatibility . . . 475

**第 10 章 Reference** **477**
10.1 PGD Reference . . . 477
10.1.1 Tables, views and functions reference . . . 477
10.1.2 EDB Postgres Distributed Command Line Interface (PGD CLI) . . . 689

EDB Postgres DistributedPGD は、標準の PostgreSQL を分散常時オンデータベースプラットフォームに変換する統合ソフトウェアスタックです。ダウンタイムがもはやオプションではない現代のエンタープライズ向けに設計された PGD は、グローバルデータ分散、高可用性、および簡素化された一貫性管理のための堅牢なフレームワークを提供します。

# 第1章　foundation

## 1.1 基盤 変換されたPostgreSQL

PGD は、標準の PostgreSQL、 EDB Postgres Extended ServerPGE、または EDB Postgres Advanced Server と互換性のある洗練された拡張機能として実装されています。PGD は、その中核として、最新世代の BDR 双方向レプリケーションを搭載しています。市場に出ているいくつかのアクティブ/アクティブシステムは、まだ古い BDR 1.0 アーキテクチャに基づいて構築されていますが、PGD には、最初のリリースから 10 年以上の継続的な機能強化とエンタープライズ強化が組み込まれています。この百戦錬磨のバージョンは、ティア 1 アプリケーションに必要な特定のパフォーマンスと信頼性を提供し、最新の分散ワークロードを処理するために 10 年間の運用で実証された改良を活用しています。

## 1.2 継続的な事業継続アクティブ-アクティブ

複雑で多くの場合数分かかるフェイルオーバープロセスを必要とする従来の物理レプリケーションとは異なり、PGD は論理レプリケーションに基づいて構築されています。

- **ゼロ待機フェイルオーバー** PGD クラスター内のすべてのノードは本質的に書き込み可能です。ノードまたはリージョンに障害が発生すると、アプリケーションは、プライマリ選択または復旧プロセスを待たずに、トラフィックを他のアクティブノードにすぐにルーティングできます。
- **常時アクセス**トラフィックはクラスター全体で流動的なままで、「復旧時間」は分ではなく秒単位で測定されます。

## 1.3 コミットスコープを介した一貫性の簡素化

PGD は、PostgreSQL の真の分散一貫性を提供する唯一のソリューションです。異なるアプリケーションには、CAP 定理一貫性、可用性、およびパーティショントレランスに関する異なるニーズがあることを認識しています。

- **事前定義された保証** 数十の `postgresql.conf` パラメーターを手動で調整する代わりに、PGD はコミットスコープを使用します。これらは事前定義された直感的な名前の構成であり、アプリケーションが必要とするパフォーマンスと耐久性の正確なトレードオフを選択できます。
- **複雑さの削減**単一のグローバル設定を使用して、非同期、同期、またはクォーラムコミットロジックを切り替えることができます。より複雑な要件の場合、簡単なコマンドを使用してこれらのスコープを個々のトランザクションレベルまで変更して、特定のビジネスニーズをその場で満たすことができます。

## 1.4 詳細なデータの局所性と主権

グローバルなユーザーベースと GDPR のような厳しい規制環境の要求を満たすために、PGD は高度なデータルーティングをサポートしています。

- **選択的フィルタリング**どのデータをどのノードに複製するかを正確に制御します。
- **地域主権**統一されたグローバルクラスタービューを維持しながら、機密ユーザーデータを特定の地理的リージョンにピン留めします。
- **レイテンシーの削減**データをアプリケーションインスタンスの物理的に近くに保ちます。エッジでローカルデータの可用性を維持することにより、リージョンを越えたラウンドトリップに関連する「光の速度」の遅延を排除します。

## 1.5 自動化された競合管理

行レベルの競合を最小限に抑えるアプリケーションデザインをお勧めしますが、現実の運用環境は本質的に乱雑であることを認識しています。同時書き込みがオーバーラップする避けられない瞬間に、PGD は堅牢な自動競合解決エンジンを提供します。

重要なセーフティネットとして機能する PGD は、確立されたビジネスルール最後の更新を正とするまたはカスタムトリガーなどに基づいて、不一致を検出および解決します。これにより、手動 DBA の介入または「データクリーンアップ」シフトを必要とせずに、グローバルファブリック全体でデータが同一のままであることが保証されます。

## 1.6 サブスクライバ専用ノードを使用したグローバル読み取りスケーリング

大規模な読み取りキャパシティを必要とするワークロードの場合、PGD はサブスクライバ専用ノードをサポートします。これらのグループを使用すると、標準のクラスターの制限をはるかに超えて水平にスケーリングできます。

- **大規模なスケール** PGD は、最大 200 のサブスクライバノードを含むクラスターを使用した実稼働環境で証明されています。
- **エッジアクセス**これらのノードは世界のどこにでも展開でき、グローバルユーザーに低遅延のローカルアクセスを提供します。
- **柔軟なトポロジ**データの分散方法を選択できます。グループ内のすべてのノードはプライマリに直接接続できます。または、グループごとに 1 つのノードが他のノードにデータを再配布する「ハブ」として機能できます。この再配布モデルには、エッジがオンラインのままであることを保証するための自動高可用性が含まれています。
- **パフォーマンスチューニング**アーキテクトは、エッジへのレプリケーション遅延を最小限に抑えるか、プライマリノードの接続負荷を管理するかを選択できます。

## 1.7 管理対象データのライフサイクル: AutoPartition と階層化テーブル

新しいデータは指数関数的に作成されており、規制要件を満たすには古いデータにアクセスし続ける必要があることがよくあります。PGD は、階層化テーブルを導入することにより、タイムインターバルパーティションの作成を自動化する AutoPartition 機能を基盤としています。この機能は、最新のデータレイクハウスアーキテクチャと直接統合しながら、さまざまなストレージ階層にわたってホットおよびコールドデータを管理するシームレスな方法を提供します。

- **自動管理** PGD は、タイムインターバルの幻影の作成とメンテナンスを自動的に処理します。毎日、毎月の間隔などのポリシーを定義すると、システムはデータが到着する前に必要なパーティションが存在することを確認します。
- **統合クエリインターフェイス**単一のテーブルからホット行ベースとコールド列方向のデータの両方を照会できます。PGD は、さまざまなストレージ形式からデータを取得する複雑さをバックグラウンドで処理します。
- **自動データオフロード** PGD は、古いデータパーティションをオブジェクトストレージの高圧縮のカラムナ形式に自動的に移行します。これにより、アクティブな SSD は高性能、低遅延のトランザクションに集中し続けます。
- **柔軟なストレージとデータレイクターゲット** 階層テーブルは、Amazon S3、MinIO、ローカルファイルシステムなどの費用対効果の高いターゲットをサポートします。さらに、PGD は、Postgres Analytics Accelerator PGAA を介して Apache Iceberg へのレプリケーションをサポートするようになりました。
- **合理化された分析** Apache Iceberg のサポートにより、PGD クラスターからデータレイクにデータを直接ストリーミングし、PGD 内からその Iceberg データを直接照会できます。これにより、複雑な ETL パイプラインを使用せずに、リアルタイムのトランザクションデータを大規模な履歴データセットと結合できます。
- **大幅なコスト効率**履歴データをオブジェクトストレージまたは Iceberg レイクハウスに移動することにより、高性能 SSD と比較してストレージコストを最大 18 倍に削減できます。

## 1.8 シームレスなライフサイクル管理

PGD は、「メンテナンスウィンドウ」を排除します。アーキテクチャが分散され、ノードが独立しているため、一度に 1 ノードずつローリングアップグレードと OS パッチを実行できます。アプリケーションは、データベースのライフサイクル全体を通じてオンラインで、クラスターの残りの部分に接続されたままです。

| Feature | Standard PostgreSQL | EDB Postgres Distributed (PGD) |
|---|---|---|
| Write availability | Single primary only | Multi-master (all nodes) |
| Failover speed | Minutes (detection + promotion) | Near-instant (traffic re-routing) |
| Consistency | Rigid (primary/standby) | Configurable via commit scopes |
| Data locality | All-or-nothing replication | Granular/selective filtering |

次のページに続く

表 1 – 前のページからの続き

| Feature | Standard PostgreSQL | EDB Postgres Distributed (PGD) |
|---|---|---|
| Upgrades | Significant downtime/re-build standbys | Zero-downtime rolling upgrades |

# 第2章　EDB Postgres Distributed 6 release notes

EDB Postgres Distributed ドキュメントでは、マイナーリリースとパッチを含む EDB Postgres Distributed 6 の最新バージョンについて説明しています。リリースノートは、各リリースの新機能に関する情報を提供します。マイナーまたはパッチリリースで導入された新機能の場合、コンテンツには、機能を導入したリリースも示されます。

| Release Date | EDB Postgres Distributed |
|---|---|
| 26 May 2026 | *EDB Postgres Distributed 6.4.0 release notes* |
| 14 Apr 2026 | *EDB Postgres Distributed 6.3.1 release notes* |
| 26 Mar 2026 | *EDB Postgres Distributed 6.3.0 release notes* |
| 05 Dec 2025 | *EDB Postgres Distributed 6.2.0 release notes* |
| 06 Nov 2025 | *EDB Postgres Distributed 6.1.2 release notes* |
| 09 Oct 2025 | *EDB Postgres Distributed 6.1.1 release notes* |
| 19 Aug 2025 | *EDB Postgres Distributed 6.1.0 release notes* |
| 25 Jun 2025 | *EDB Postgres Distributed 6.0.2 release notes* |
| 09 Jun 2025 | *EDB Postgres Distributed 6.0.1 release notes* |

## 2.1 EDB Postgres Distributed 6.4.0 release notes

リリース日 2026 年 5 月 26 日

EDB Postgres DistributedPGD 6.4.0 には、新機能、機能強化、バグ修正が含まれています。

### 2.1.1 ハイライト

- ** クォーラムコミット :** クラスター全体でプリコミットコーディネーションを強制する新しいコミットスコープの種類。ノードがトランザクションをコミットする前に、ノードの大部分が同意する必要があります。合意に達しない場合、トランザクションはあらゆる場所でロールバックし、競合する更新やノード全体での発散がない一貫したグローバルビューを保証します。
- ** *Connection pooling* :** 接続マネージャーは、セッションモードプーリングに加えてトランザクションモード接続プーリングをサポートするようになりました。トランザクションプーリングモードでは、バックエンド接続はトランザクションの間のみ保持され、`max_connections` の圧力が削減され、高い同時実行性の下でのスケーラビリティが向上します。
- ** *Replicated large objects* :** PGD 6.4.0 では、分散展開でのラージオブジェクトのサポートを追加します。PGD は、すべてのノードでラージオブジェクトをレプリケートし、galloc シーケンスを使用してグローバルに一意の OID を生成して、競合を排除します。`lo_import` や `lo_export` を含む標準の

lo_* API は、Connection Manager を介して透過的に動作するため、ラージオブジェクトを使用するアプリケーションは、アーキテクチャを変更せずに PGD に移行できます。

### 2.1.2 機能

| Description | Addresses |
|---|---|
| レプリケーションの失敗を処理するための構成可能な適用エラーポリシーを追加しました。PGD は、レプリケーションの適用の失敗を処理するための構成可能なポリシーを提供するようになりました。トランザクションの適用に失敗すると、PGD は失敗した変更をキャプチャし、新しい SQL ファンクションを介してレビューと解決のために公開できます。 `bdr.alter_node_group_option`を使用して、ノードグループごとにポリシーを構成します。詳細については、<a href="/pgd/current/apply-error-policy">レプリケーション適用エラーの処理</a>を参照してください。</p> | |
| 論理的にレプリケートされたラージオブジェクトのサポートを追加しました。PGD は、ビルトインの Postgres ラージオブジェクトのドロップイン置換として、ラージオブジェクトの独自の実装を提供するようになりました。 PGD ラージオブジェクトは、クラスター内のすべてのノードに論理的にレプリケートされます。この機能は、 `bdr.lo_`ファンクションを介して直接使用するか、 `bdr.override_large_objects`を介して透過的に利用できます。詳細については、<a href="/pgd/current/large-objects">レプリケートされたラージオブジェクト</a>および<a href="/pgd/current/reference/tables-views-functions/large-objects を参照してください。 ">ラージオブジェクトファンクションリファレンス</a>。</p> | |
| Raft の状態を同期するための `pgd raft sync-snapshot`コマンドを追加しました。Raft を無効にすることにより、すべてのノードで Raft 状態を同期する新しい raft `pgd raft sync-snapshot`</a>コマンド、リーダーノードから状態をエクスポートし、すべてのノードにインポートし、Raft を再度有効にします。 `--group`を使用してターゲットグループを指定します。デフォルトはグローバルグループです。</p> | |
| 書き込みと Raft のリーダーシップをローカルノードから転送するために<a href="/pgd/current/reference/tables-views-functions/functions#bdrtransfer_all_leaderships">`bdr.transfer_all_leaderships`</a>を追加しました。書き込みと Raft のリーダーシップをローカルから転送しますすべてのグループのノード。メンテナンス前にノードをドレインするのに役立ちます。</p> | |

### 2.1.3 機能強化

| Description | Addresses |
|---|---|
| パラレル適用がクォーラムコミットと互換性があるようになりました。グループコミットとは異なり、クォーラムコミットはパラレル適用と互換性があります。クォーラムコミットは、`num_writers = 1`でのみサポートされていた古いグループコミットを改善します。</p> | |
| より多くの情報を表示するように`pgd raft show`コマンドを強化しました。raft `pgd raft show`</a>は、デフォルトで 3 つのセクションを表示し、Raft コンセンサスレイヤーのより深い可視性を提供するようになりました。：State コンセンサス状態、Follows リーダーから見たフォロワーステータス、および Journal グローバルジャーナルの詳細。</p> | |
| 接続マネージャーは、接続起動パラメーターで指定された`IntervalStyle`を尊重するようになりました。以前は、`IntervalStyle`は`options`接続パラメーターを介してのみサポートされていました。接続マネージャーは、起動パラメーターとして直接指定された場合に`IntervalStyle`を正しく適用するようになりました。</p> | |
| トポロジと構成の変更に関する接続マネージャーのアイドル接続の処理が向上しました。接続マネージャーは、トポロジまたは構成の変更に応じて古いアイドル接続をプロアクティブに閉じ、サーバー接続エラーをクライアントに直接報告します。</p> | |
| `node_group_dsn`を介してサブグループごとの接続文字列のサポートを追加しました。 `bdr.node_config`の新しい `node_group_dsn`列により、同じサブグループ内のノードは専用の接続文字列を使用できます。これは、Kubernetes およびクラウド環境で下りコストを削減するのに役立ちます。 <a href="/pgd/current/reference/tables-views-functions/nodes-management-interfaces#bdralter_node_option">`bdr.alter_node_option`</a>を介して設定します。</p> | |
| 自動候補選択をサポートするように<a href="/pgd/current/reference/tables-views-functions/functions#bdrrouting_leadership_transfer">`bdr.routing_leadership`を強化しました。`bdr.routing_leadership_transfer`は、`leader_name`の`NULL`を受け入れ、ベースで最適な候補の自動選択をトリガーします書き込みリーダーの進行状況。自動選択は、コミットスコープの種類としてクォーラムコミットを使用する場合に推奨されるアプローチです。 Raft プロトコルバージョン 6004 以降を使用した PGD 6.4 が必要です。</p> | |
| Connection Manager は、`SIGHUP`で Postgres 構成パラメーターを動的にリロードするようになりました。以前は、構成パラメーターは Connection Manager プロセスの開始時にのみロードされていました。接続マネージャーは、再起動を必要とせずに`SIGHUP`にそれらをリロードするようになりました。</p> | |
| `bdr`スキーマのオブジェクトの DDL は、許可されなくなりました。データベースユーザーは、`bdr`スキーマでデータベースオブジェクトを作成または変更したり、オブジェクトの名前空間を `bdr`スキーマに設定したりできません。この制限により、このような操作で発生する可能性のあるカタログの破損と拡張機能のアップグレードの問題が防止されます。詳細については、 <a href="/pgd/current/reference/tables-views-functions/pgd-settings#bdrallow_bdr_schema_changes">`bdr.allow_bdr_schema_changes`</a>を参照してください。</p> | 53224 |

次のページに続く

表 3 – 前のページからの続き

| Description | Addresses |
|---|---|
| CAMO で構成されたグループのノードの参加と分離のためのサニティチェックを追加しました。既存のノードが CAMO パートナーとして構成されているグループにデータノードを参加させようと、またはグループの `default_commit_scope`またはローカルノードの `bdr.commit_scope`が次のように設定されているときにデータノードを分離しようとしている CAMO コミットスコープ、エラーが発生するようになりました。これらの操作を実行するには、最初に CAMO コミットスコープをドロップします。</p> | 55212 |
| 構成パラメーター `bdr.enable_auto_sync_reconcile`が `pg_settings`に表示されるようになりました。`bdr.enable_auto_sync_reconcile`構成パラメーターが `pg_settings`ビューに公開され、モニタリングツールで表示されます。</p> | |
| トランザクション変更トラッカーバッファを超過した場合のログ出力が改善されました。出力されたログメッセージにより、問題が一時的なものであり、ライタの自動再起動後に処理が再開されることが明確になります。</p> | |

### 2.1.4 バグ修正

| Description | Addresses |
|---|---|
| 論理結合の同期中にテーブルスペースに割り当てられるゼロの OID を修正しました。`bdr_tablespace_upsert()`は、新しいテーブルスペースの挿入時に OID を設定せず、論理結合中に複数のユーザーテーブルスペースが同期されたときに重複キー違反を発生させました。</p> | 59715 |
| BDR 拡張機能のない Postgres インスタンスで失敗する `pgd assess`コマンドを修正しました。 `pgd assess`コマンドは、`ERROR: schema "bdr" does not exist`で失敗していました。このコマンドは、最初に BDR スキーマをチェックすることにより、 BDR がインストールされていないインスタンスを正常に処理し、 PGD Essential エディションと Expanded エディションの両方の互換性の提案を使用して完全な評価を実行します。</p> | |
| 参加ノードによって使用される一時レプリケーションスロットの walsender のメモリリークを修正しました。`bdr_copy_to_repl_dest_receiver_receive`のメモリリークを修正しました。参加ノードによって使用される一時レプリケーションスロットの walsender の行ごとの割り当てが正しく解放されません。</p> | 59663 |
| 古い `node_peer_progress_mapping`エントリが原因で発生する無制限の WAL の蓄積を修正しました。キャンセルされた転送サブスクリプションによって残された古い `node_peer_progress_mapping`エントリは、グループレプリケーションスロットを固定し、WAL のクリーンアップを妨げていました。メンテナンスパスは、直接サブスクリプションの進行状況によって置き換えられたエントリを削除するようになりました。</p> | 58999, 59320 |

次のページに続く

表 4 – 前のページからの続き

| Description | Addresses |
|---|---|
| Part および同期ノードステートマシンのデッドロックとストールを修正しました。 Part および sync ステートマシンは、リカバリ終了マーカーが `FORWARD`または`CATCHUP`サブスクリプションを介して配信された場合、デッドロックまたはストールが発生する可能性がありました。複数の修正により、確認待機が`NORMAL`サブスクリプションに制限され、`PART_CATCHUP`中のアトミックオリジンの前進が保証され、`node_peer_progress_mapping`行の二重削除エラーが防止されます。</p> | |
| レプリケーションオリジンが既に使用中のライタのセグフォルトを修正しました。同じオリジンで`sync_node`要求が進行中に`NORMAL`サブスクリプションを有効にすると、2 番目のライタが`SIGSEGV`でクラッシュしました。この場合、PGD は`NORMAL`サブスクリプションの有効化をブロックするようになり、オリジン取得パスはクラッシュする代わりに回復可能なエラーを結果ます。</p> | |
| BDR 5 から PGD 6 のローリングアップグレード中の`pgd manager`クラッシュを修正しました。クラスター内の最後のノードを BDR 5 から PGD 6 にアップグレードする場合、レプリケーションスロットの名前が変更されたときに、サブスクリプションの再構成中に`pgd manager`プロセスが`SIGSEGV`でクラッシュしました。 `pgd manager`は、欠落しているスロットまたは名前が変更されたスロットを正常に処理するようになりました。</p> | |
| アップグレードの失敗の原因となった`pgd node setup`のシステム OID の割り当てを修正しました。`pgd node setup`は、システムレンジの OID で BDR オブジェクトを作成していたため、 `pgd node upgrade`が`schema "bdr" does not exist`で失敗しました。 setup コマンドはユーザー範囲 OID を正しく割り当てるようになり、upgrade コマンドは既存の影響を受けるクラスターの早期に問題を検出します。</p> | |
| `pgaudit`が`shared_preload_libraries`の`bdr`に続く場合、PGD 6.2 から 6.3 または 6.4 へのアップグレードの失敗を修正しました。`shared_preload_libraries`で`pgaudit`が`bdr`をフォローする場合、PGD 6.2 からのアップグレードは`ERROR: pgaudit not loaded before call to pgaudit_ddl_command_end()`で失敗しました。フックの初期化により、フックが`NULL`にリセットされ、マネージャープロセスで pgaudit がクラッシュしていました。この修正では、バイパスマクロを追加して、フックが初期化を解除する代わりに次のハンドラーに遅延するようにします。</p> | 59715 |
| ポストマスターが強制終了された後も停止しない接続マネージャープロセスを修正しました。 | 59631 |
| `bdr_connection.client_addr`更新によって発生した Postgres ワーカーのメモリリークを修正しました。セッションプーリングモードで Connection Manager が使用する`bdr_connection.client_addr`構成パラメーターの更新の処理は、Postgres ワーカーのメモリリークでした。</p> | |
| ノードグループでルーティングが有効または無効になっているときに、古いルーティング構成を使用する接続マネージャーを修正しました。 | |

次のページに続く

表 4 – 前のページからの続き

| Description | Addresses |
|---|---|
| `bdr.stat`共有メモリエントリ処理のセグフォルトを修正しました。コンカレント統計リーダーは、統計エントリが作成またはドロップされたときに部分的に初期化または解放された動的共有領域 DSA ポインタを観察し、`bdr_stat_write_file`が NULL または解放されたメモリを逆参照します。初期化パスは、割り当てが成功するまで無効なエントリをマークし、ドロップパスが排他的ロックを保持し、すべての dshash スキャンパスが無効な DSA ポインタからガードします。</p> | |
| 並列ライターの下の `bdr.alter_sequence_set_kind`の検出できないデッドロックを修正しました。複数のライターがシーケンス種類を同時に変更した場合、2 人のライターが同じシーケンスリレーションで `ShareRowExclusiveLock`を`AccessExclusiveLock`に同時にアップグレードしようとして、デッドロック検出器のビューの外部でデッドロックが発生する可能性がありました。ロックは `AccessExclusiveLock`で事前に取得され、ロックアップグレードウィンドウが排除されるようになりました。</p> | |
| WAL の最後にある `bdr_load_commit_decisions`のアサーションクラッシュを修正しました。コミット決定のロード中に WAL の最後に到達すると、デバッグビルドでアサーション `buf.len > 0`がトリップしました。使用可能なデータがなくなると、ループは通常終了するようになりました。</p> | |
| BDR カタログが利用できる前に、フェールオーバースロットスタンバイ応答フックのクラッシュを修正しました。 BDR カタログが完全に初期化される前にスタンバイ応答フックが起動し、クラッシュが発生する場合がありました。フックはカタログの可用性を確認し、カタログの準備がまだ整っていない場合は早期に終了するようになりました。</p> | |
| レシーバーロールが起動中にリセットされた場合、 Writer-start アサーションを正常な終了に置き換えました。ライタの起動とライタの初期化の間でレシーバーロールがリセットされた場合、ライタはアサーションをトリップしました。ライタは正常に終了するため、コアファイルを生成せずにスーパーバイザによって再起動できます。</p> | |
| 非同期転送を正しく待機し、グループスコープを検証するように `pgd raft set-leader`を修正しました。 `pgd raft set-leader`コマンドは、非同期モードでリーダーシップの変更を適切に待機せず、無効なグループ間転送を許可し、ターゲットグループが存在しない場合に不明瞭なエラーメッセージを生成しました。 3 つの問題はすべて修正されました。</p> | |
| 現在利用可能な Raft リーダーがないときに `routing_leadership_transfer`を再試行させました。`routing_leadership_transfer`は、コンセンサスがリーダーなしを報告すると、他の一時的なコンセンサスエラーのように再試行するのではなく、すぐに返されました。 `BDR_CONSENSUS_NO_LEADER`は再試行可能なエラーセットの一部になったため、リーダーシップの転送は失敗するのではなく、短いリーダー選挙の後に完了します。</p> | |

次のページに続く

表 4 – 前のページからの続き

| Description | Addresses |
|---|---|
| 分析リース更新の障害が `ERROR`から `LOG`にダウングレードされました。analytics Seafowl レプリケーターの一時的なリース更新障害により、ワーカーがクラッシュして再起動しました。これらの障害は次の更新サイクルで自己回復するため、`LOG`レベルでログに記録され、ワーカーは実行を継続します。</p> | |
| AutoPartition の待機信頼性とタスクマネージャーのバックオフの遅延を修正しました。AutoPartition タスクマネージャーの指数バックオフは、作業項目の処理後にリセットされず、不必要な反復遅延の原因となります。 *autopartition_wait_for_partitions_on_all_nodes* `は、シグナリングの向上、リモート接続の` *statement_timeout = 0* `オーバーライド、および` *PGresult* / `PGconn`リークの修正で強化されました。</p> | |
| Postgres の動作と一致するように接続マネージャーのログレベルフィルタリングを修正しました。Connection Manager は `log_min_messages`を誤って解釈し、Postgres と比較してログフィルタリングの一貫性の欠如を引き起こしました。</p> | |
| `bdr.ddl_locking`のデフォルトの表示値を `auto`に変更しました。以前は、値は `true`として表示されていました。これは、同じ意味ですが、下位互換性のエイリアスとしてのみ存在します。この変更には、構成の変更は必要ありません。</p> | |
| PGD ワーカーが `SIGTERM`に応答しなくなったときに、100 の CPU でスピンする pglogical manager プロセスを修正しました。 pglogical manager プロセスがエラーで終了し、他の PGD ワーカーを終了しようとすると、それらのワーカーが不潔に死亡しているか、スタックして受信できない場合にハングしました信号。このような場合、プロセスはすべて正常に終了するようになりました。</p> | 60272 |

## 2.2 EDB Postgres Distributed 6.3.1 release notes

リリース日 2026 年 4 月 14 日

EDB Postgres Distributed 6.3.1 には、安定性と信頼性の向上に焦点を当てたバグ修正のみが含まれています。このリリースには、新機能や機能強化は含まれません。既知の問題のリストについては、*'Known issues and limitations'* を参照してください。

### 2.2.1 バグ修正

| Description | Addresses |
| --- | --- |
| *bdr.enable_auto_sync_reconcile* **がオンになっているときに** *6.3.0* **へのローリングアップグレード中にコンセンサスワーカーのセグフォルトを修正しました。** *`bdr.enable_auto_sync_reconcile = on`* **の場合、古いバージョンのノードが** *`reconcile_2pc_requests`* Raft upsert を 6.3.0 ノードに送信すると、再起動時にクラッシュループが発生する可能性がありました。この設定は、6.x バージョンではデフォルトでオンになり、5.x バージョンではデフォルトでオフになります。</p> | 59715 |
| PGAA 拡張機能が存在する場合の `bdr_init_physical`障害を修正しました。Postgres Analytics AcceleratorPGAA 拡張機能を使用して展開で PGD クラスターを単一のデータグループから複数のデータグループにスケールアップする場合、 `bdr.remove_unwanted_data()`ステップ中に `bdr_init_physical`は `no analytics lease found for database`で失敗しました。この問題は、単一のデータグループで最初にプロビジョニングされ、後でスケールアウトされたクラスターにのみ影響がありました。</p> | |
| ピアノードに複製されないユーザーのログイン属性を修正しました。`CREATE USER`ステートメントの後、ユーザーのログイン属性はピアノードに伝播されず、元のノード以外のノードでのログインを防止しました。この修正により、ロールログイン属性がレプリケーションに正しく含まれるようになります。</p> | |

## 2.3 EDB Postgres Distributed 6.3.0 release notes

リリース日 2026 年 3 月 26 日

EDB Postgres DistributedPGD 6.3.0 には、安定性と信頼性の向上に焦点を当てた新機能、機能強化、バグ修正が含まれています。

### 2.3.1 ハイライト

- **加入者専用ノードの接続マネージャー**加入者専用ノードで接続マネージャーがデフォルトで起動し、高度な接続プーリングおよびセッション管理機能へのアクセスを提供します。また、すべての読み取り専用トラフィックをローカルノードに再ルーティングして、遅延を削減します。
- **PGD CLI を介したコンセンサス管理** PGD CLI には、特定のノードまたはグループで Raft コンセンサスを有効または無効にするコマンドが含まれるようになりました。コマンド *pgd raft enable* および *pgd raft disable* は、ノードの動作をきめ細かく制御します。
- **プラットフォームサポートの更新** PGD は、SUSE Linux Enterprise ServerSLES 15 SP7 6.3.0 以降でサポートされるようになりました。

### 2.3.2 機能

| Description | Addresses |
|---|---|
| 接続マネージャーは、サブスクライバー専用ノードでデフォルトで起動し、新しいヘルスチェックエンドポイントが含まれます。 接続マネージャーは、サブスクライバー専用ノードで自動的に初期化されるようになりました。このリリースでは、 */connection/is-ready-ro* `および/connection/is-ready-rw` HTTP エンドポイントを追加して、読み取り専用および読み取り/書き込みトラフィックの詳細な準備状況チェックを提供します。 | |
| *pgd raft set-leader* コマンドを追加しました。 `pgd raft set-leader`コマンドが導入され、ユーザーは特定のグループの Raft リーダーを手動で変更できます。 | |
| *pgd raft enable* および *pgd raft disable* コマンドを追加しました。 さまざまなノード、グループ、またはクラスターレベルでコンセンサス Raft を有効または無効にするコマンドを追加しました。 | |

### 2.3.3 機能強化

| Description | Addresses |
|---|---|
| グループ削除プロセスを改善して、関連するレプリケーションセットを自動的にドロップします。 同じ名前のレプリケーションセットを自動的にドロップするようにノードグループ削除ロジックを強化しました。 | 57531 |
| `pg_hba.conf`に対する接続マネージャーの一貫性が向上しました。 接続マネージャーがデータベース名とユーザー名を`pg_hba.conf`フィールドと照合する方法を強化しました。主な改善には、`sameuser`マッチングのサポート、洗練された`samerole`ロジック、ロール自分自身と間接メンバーを含むように拡張された`+rolename`マッチングが含まれます。さらに、一貫した認証動作を保証するために、照合する前にユーザー名から引用符が削除されるようになりました。 | |
| 接続マネージャーへのユーザー名マッピング`pg_ident.conf`のサポートを追加しました。 この機能強化により、接続マネージャーはユーザー名マッピングに`pg_ident.conf`を使用できるようになります。現在、この機能は`cert`認証方法に適用され、Connection Manager の動作を PostgreSQL のネイティブ認証 ID マッピングと調整します。 | |
| *bdr.reconcile_2pc_requests* `テーブルを改善し、ビュー`*bdr.reconcile_2pc_requests_summary*`を追加しました。 :ref:`bdr.reconcile_2pc_requests<bdr.reconcile_2pc_requests>` を強化して、要求時間、起点 LSN、コミットタイムスタンプなどの詳細を含めます。さらに、使いやすいビュー *bdr.reconcile_2pc_requests_summary* を追加して、2 フェーズコミット 2PC 調整のモニタリングを簡素化しました。 | |
| グループコミットでのオリジン変更の競合処理が向上しました。 トランザクションのオリジンが変更されたときに熱心な競合検出と解決を実行するためのグループコミットが強化されました。競合処理は、レプリケーションの進行状況の推測ではなく、タイムスタンプベースの検出に依存するようになりました。最適なパフォーマンスを保証し、コミットレイテンシーの増加を回避するには、すべてのノードでシステムクロックを同期する必要があります。 | |

次のページに続く

表 7 – 前のページからの続き

| Description | Addresses |
|---|---|
| モニタリングを強化するために `bdr.stat_reconcile_commit_scope`ビューを追加しました。 新しいシステムビュー *bdr.stat_reconcile_commit_scope* が導入されました。コミットスコープリコンシリエーションのリアルタイム統計を提供します。 | |
| 2 つの新しい `bdr.node_group_config`オプションを介した書き込みリーダー選択ロジックが改善されました。 *bdr.node_group_config* の新しいオプションを導入しました。これにより、クラスターは、リーダーの移行中に最も高度な候補レプリカを選択できます。<ul><li>`leader_progress_tracking_enable`: レプリケーションの進行状況を Raft リーダーに定期的に報告します。</li><li>`leader_progress_tracking_period`: レポートを構成しますインターバル。</li></ul> | |
| *galloc* **`シーケンスの** *`ALTER SEQUENCE ... RENAME`* **および** *`ALTER SEQUENCE ... `*SET SCHEMA`のサポートを追加しました。 `galloc`シーケンスの名前の変更またはスキーマ間での移動が完全にサポートされるようになりました。これは、 `bdr.sequence_kind`カタログのグローバル UUID 識別子 `sequuid`によって有効になり、PGD はこれらのシーケンスを一貫して同時に追跡できます。 PGD は、既存のシーケンスに必要な UUID を自動的に生成して、互換性を確保するため、PGD 6.3 より前のバージョンからアップグレードされたシーケンスでも動作します。 | |
| *bdr.prepared_xacts_summary* **診断ビューを追加しました。 グループコミットスコープによって暗黙的に作成された準備されたトランザクションを監視するためのローカル診断ビュー** *:ref:`bdr.prepared_xacts_summary<bdr.prepared_xacts_summary>* が導入されました。 | |

### 2.3.4 の変更

| Description | Addresses |
|---|---|
| スタンドアロン macOS PGD CLI は ARM64 のみになりました。 macOS 用のスタンドアロン PGD CLI が、Apple Silicon ARM64 アーキテクチャのみで利用できるようになりました。この変更は、スタンドアロンの macOS バイナリにのみ適用されます。 GNU / Linux でのスタンドアロン PGD CLI バイナリの可用性とサポートは変更されません。 | |

### 2.3.5 バグ修正

| Description | Addresses |
|---|---|
| 中断されたパーティションの切断が原因で発生するレプリケーションクラッシュループを修正しました。`DETACH PARTITION CONCURRENTLY`コマンドが中断された場合、論理レプリケーションが無限クラッシュループに入る問題を解決しました。以前は、walsender は`detach-pending`状態のパーティションの WAL をデコードするときに、`relation has no parent because it's being detached`エラーで失敗していました。 | 53628, 59291 |
| パーティションの切断中の`walsender`クラッシュループを修正しました。`ERROR: relation %u has no parent failures`をレポートした`inhdetachpending = true`を使用したパーティションに遭遇すると、`walsender`プロセスがクラッシュループに入る可能性がある問題を修正しました。 | 53628, 59291 |
| `ALTER ROLE`レプリケーション障害を修正しました。`NOSUPERUSER`や`NOREPLICATION`のような属性がレプリケートされたコマンドに自動的に追加されるため、レプリケーションが中断される問題を修正しました。これらのオプションはスーパーユーザーのみが変更できるため、ダウンストリームノードでの適用プロセスは失敗しました。この変更により、レプリケートされたコマンドには、明示的に指定され、ユーザーが十分な特権を持つオプションのみが含まれるようになります。 | 54906 |
| `pgd setup`コマンドのエンコード文字セットの問題を修正しました。 コマンドはリモートノードから直接エンコード値を取得しようとします。使用できない場合はデフォルトの UTF8 に設定されます。 | |
| `pgd group set-option`コマンドのインラインヘルプを修正しました。`pgd group set-option`のヘルプテキストを修正して、グループレベルのオプションを正確に反映しました。 | |
| *pgd replication show* **`コマンドのマトリックスビュー出力を修正しました。 遅延が** *10 MB* **未満の場合、非アクティブ/切断されたスロットに`*`の代わりに!`が表示され、一貫性が復元されます。** | 56295 |
| コンセンサス共有メモリキューのメモリの増加を修正しました。 ノードが到達不能であるときにコンセンサスが使用する共有メモリキューの成長を防ぎ、ノードの停止中にメモリが安定した状態を保証します。 | |
| サブスクライバー専用ノードがダウンしている場合のコンセンサスプロセスのメモリリークを修正しました。 接続文字列の不適切な解放に関連する malloc ベースのメモリリークを解決しました。 | 57094 |
| `use_https`ノードグループ構成オプションが無視される問題を修正しました。接続マネージャーが`use_https`設定を誤って無視するバグを解決しました。PGD 6.3.0 または 5.9.2 にアップグレードする場合、以前の動作を維持するには、アップグレードの前に`use_https`グループオプションを`true`に設定することにより、HTTPS を明示的に有効にする必要があります`postgresql.conf`またはコマンドラインで`ssl_cert_file`および`ssl_key_file`を構成します。 | |
| 構成のリロード中に発生する接続マネージャーのメモリリークを修正しました。 構成のリロード中の Postgres メモリコンテキストの誤った処理が原因で発生した接続マネージャーのメモリリークを解決しました。 | |

次のページに続く

表 9 – 前のページからの続き

| Description | Addresses |
|---|---|
| `pgaudit.role`がデフォルト以外の値に設定されている場合、`COPY`コマンド中に発生するセグメンテーションフォールトを修正しました。`pgaudit.role`構成がデフォルト以外の値に設定されている場合、`COPY`コマンドを実行すると Postgres がセグメンテーションフォールトでクラッシュする問題を解決しました。 | 58060 |
| AutoPartition テーブルで`retention_period`および`analytics_offload_period`を同時に設定すると競合が発生する問題を修正しました。 データ保持と分析オフロード間の調整が向上しました。データ保持ポリシーが適用される前に、Postgres Analytics AcceleratorPGAA を使用してパーティションがオフロードされるようになりました。 | |
| 冗長なレプリケーションプロセスを発生させる可能性がある分析レプリケーターの問題を修正しました。 サブグループと親グループで分析が有効になっているクラスタートポロジでは、バグにより各サブグループの書き込みリーダーが独自の分析レプリケーターインスタンスを開始し、行が重複していました。この修正により、PGD グループごとに 1 つの分析レプリケーターのみがクラスター全体で実行されるようになります。 | |
| サーバー接続が予期せず閉じられたときに発生する接続マネージャーの無限ループを修正しました。 接続プールからの取得プロセス中に予期せずサーバーが切断されると、無限ループがトリガーされる、接続マネージャーの競合状態を解決しました。この問題は、セッションプーリングモードで動作しているクラスターに特に限定されていました。 | |
| 変更適用中のロック処理を改善して、同時バックエンドからのデータの破損を防ぎます。 テーブルとインデックスのロックは、変更を適用するときにトランザクションがコミットするまで保持されるようになりました。これにより、正しいシリアル化が保証され、コンカレント操作中のデータの一貫性の問題が防止され、特にパラレル適用によって導入されるリスクに対処します。 | |
| PGD がコミュニティ Postgres で展開されている場合に、Parallel apply を有効にできるようになりました。コミュニティ PostgreSQL でパラレル適用を有効にすると、頻繁なロックタイムアウトが原因でライタのハングとサブスクリプションのスタックが発生する問題を解決しました。この修正により、回避策として無効にすることなく、パラレル適用を安全に使用できます。 | 54565, 57005, 57327 |
| ファイル記述子が枯渇した際の接続マネージャーの復元力が向上しました。接続マネージャーは、ファイル記述子の枯渇から自動的に回復するようになりました。システム制限に到達すると、プロセスはリッスンソケットを閉じ、30 秒後に自動的に再起動して接続を復元します。 | |
| ノードグループの作成中の誤ったエラーレポートを修正しました。 グループが正常に作成された場合でも、`bdr.create_node_group`がエラーを返す場合がある競合状態を解決しました。ファンクションは、返す前にノードの作成が完全に完了するのを正しく待機するようになりました。 | |
| `bdr_init_physical`後のプロキシ構成の損失を修正しました。 Raft スナップショットの復元がプロキシ設定の永続化に失敗したため、`bdr_init_physical`の実行後にプロキシ構成が保持されなかった問題を修正しました。 | 58319 |

次のページに続く

表 9 – 前のページからの続き

| Description | Addresses |
|---|---|
| `DROP`および `CREATE`後の古い `galloc`シーケンス割り当てエントリを修正しました。 `galloc`シーケンスがドロップされて再作成された後、`sequence_alloc`エントリが持続するバグを修正しました。この修正により、監視ノードを含むクラスター全体に適切なクリーンアップの伝播が保証されます。 | |
| `galloc`シーケンスの開始値を無視する `bdr.alter_sequence_set_kind`を修正した。 シーケンス種類を `galloc`に変更すると、提供された `start_value`がサイレントに破棄される問題を修正しました。さらに、WAL レプリケーション中に `0`の `startval`が失われるバグを解決しました。 | |
| アナリティクスの書き込みリーダーの移行中の重複データのレプリケーションを防止しました。 以前の書き込みリーダーの分析レプリケーターが、新しいリーダーのレプリケーターと同時に実行され続ける可能性がある問題を修正しました。この同期の向上により、新しいプロセスが初期化する前に古いプロセスが完全に停止し、重複データの挿入が防止されます。この修正には、*bdr.group_lease* カタログテーブルと *bdr.group_lease_override* ファンクションの追加が含まれています。 | |

## 2.4 EDB Postgres Distributed 6.2.0 release notes

リリース日 2025 年 12 月 5 日

EDB Postgres Distributed 6.2.0 には、安定性と信頼性の向上に焦点を当てた新機能、機能強化、バグ修正が含まれています。

### 2.4.1 ハイライト

- **Postgres 18 サポートを含む最初のリリース** PGD 6.2.0 は、新しくリリースされた PostgreSQL 18、EDB Postgres ExtendedPGE 18、および EDB Postgres Advanced Server 18 のサポートを導入します。
- **Connection Manager の機能強化** PGD 6.2.0 では、LDAP 認証のサポートや、リーダーの変更中に既存のクライアントの接続がドロップするのを防ぐルーティングの安定性の向上など、Connection Manager の重要な機能強化が導入されています。
- **透過的データ暗号化 TDE の採用の容易化** `pgd node setup` および `pgd node upgrade` コマンドに TDE オプションが含まれるようになりました。これにより、非 TDE ノードから TDE ノードへのシングルコマンド変換など、TDE の採用が簡素化され、古い PGE または EPAS バージョン v15 より前を実行している PGD クラスターをアップグレードする際の TDE の採用が大幅に合理化されます。

### 2.4.2 機能

| Description | Addresses |
|---|---|
| *pgd node upgrade* `および `*pgd node setup* `コマンドに *TDE* オプションを追加しました。<*hr/*> `*pgd node upgrade* `コマンドは、 `*--key-unwrap-command* `および--copy-by-block`オプションを使用して非 TDE PGD ノードから TDE 対応ノードへのインプレース変換をサポートするようになりました。 *pgd node setup* `コマンドには、新しいクラスターノードの初期プロビジョニング中に *TDE* を構成する包括的なオプションが含まれるようになりました。これらの新しいオプションは次のとおりです。--data-encryption`、*--copy-key-from*、*--key-wrap-command*、*--key-unwrap-command*、*--no-key-wrap*、および `--data-encryption-keylen` AES キー長を受け入れます。 | |
| `bdr.stat_receiver_transactions`ビューを追加しました。<hr/>ビューは、レシーバーパイプラインのトランザクションに関するリアルタイム情報を提供します。このビューは、現在受信され、処理のためにライターに送信されたトランザクションをレポートし、ライターのアプリケーションの進行状況を含む、トランザクションがファイルまたはライタにストリーミングされているかどうかを示します。 | |
| Connection Manager に LDAP 認証のサポートを追加しました。<hr/>Connection Manager は、OpenLDAP 構成ファイルと環境変数がサポートされていないことを除き、PostgreSQL と同じ LDAP 認証機能をサポートするようになりました。 | |
| `bdr.writers`ビューに詳細なモニタリング列を追加しました。<hr/>`bdr.writers`ビューには、モニタリングを強化するためのいくつかの新しい列が含まれます。トランザクション適用ソース、トランザクションの合計サイズ、アプリケーションの進行状況、適用された最後に変更、および最後に適用された変更に対応する上流 LSN を報告するようになりました。 | |

### 2.4.3 機能強化

| Description | Addresses |
|---|---|
| EPAS クラスター`setup`コマンドの推定デフォルト`superuser`値。<hr/>EPAS クラスターをプロビジョニングする際、 *pgd node setup* `コマンドの--superuser`オプションは、明示的に指定しない限り、自動的にデフォルトの`enterprisedb`に設定されます。 | |
| *pgd cluster verify --verbose --arch* `出力に *Details* `列を追加しました。<*hr/*>--verbose`オプションを使用する場合、`pgd cluster verify`コマンドのアーキテクチャ検証セクションに`Details`列が含まれるようになりました。この列は、アーキテクチャチェック中に発生した問題に関する具体的な詳細情報を提供します。 | |

次のページに続く

表 11 – 前のページからの続き

| Description | Addresses |
|---|---|
| パーティングノード名の再利用のための `pgd node part`コマンドを強化しました。<hr/>`pgd node part`コマンドの動作が変更され、ノードの名前が変更されるまで待機するだけで元の名前を再利用し、バックグラウンドで複雑なパーティングオペレーションを継続するようになりました。この機能強化は、PGD クラスター内のすべてのノードがバージョン 6.2 以降を実行している場合、 `--no-wait`オプションも非推奨にします。 | |
| `bdr.part_node`で分割ノード名の再利用を許可します。<hr/>PGD ノード管理インターフェイスでは、分割操作がまだ進行中であっても、分割ノードの名前を再利用できるようになりました。これは、 `bdr.part_node`ファンクションを呼び出すときに新しい `concurrent=true`オプションを設定することにより実行されます。 | |
| 暗号化されたサーバーサイド証明書のサポート。<hr/>Connection Manager は、サーバープライベートキーを復号する PostgreSQL `ssl_passhrase_command`オプションをサポートするようになりました。 | 53695 |
| `pgd node setup`の PGD データベースユーザー認証が向上しました。<hr/>コマンドは、パスワードファイルからユーザー認証資格情報を読み取るように強化されました。このファイルの場所は、 `--dsn`グローバルオプション、`PGPASSFILE`環境変数、または`~/.pgpass`ファイルをこの順序でチェックすることにより解決されます。 | |
| `analytics_write_catalog`ノードグループオプションを使用して、`bdr.prefer_analytics_engine`設定が Iceberg REST カタログで機能するようになりました。<hr/>以前は、この設定は PGFS ストレージの場所にのみ適用されていました。 | |
| *--group filter* `を `*pgd nodes list* `および `*pgd raft show* `コマンドに追加しました。<*hr/*>--group`フィルタは、マルチグループ PGD クラスターを管理するときに指定されたデータグループで出力をフィルタリングする機能を提供します。 | |

### 2.4.4 バグ修正

| Description | Addresses |
|---|---|
| PostgreSQL 18 以降にアップグレードするときに `pgd node upgrade`コマンドが `invalid_slots`エラーで失敗した問題を修正しました。<hr/> BDR 拡張機能が更新され、`pg_upgrade`ユーティリティとの互換性が向上し、アップグレードプロセスがクリーンに実行されます。 | |
| 接続マネージャー DSN を使用したスイッチオーバーの矛盾した出力を修正しました。<hr/>`pgd group set-leader`コマンドが成功と失敗の両方の紛らわしいメッセージを表示する問題を解決しました。この問題は、スイッチオーバーが正常に完了した後、書き込みリーダーの変更により、接続マネージャーがすぐに接続をリセットしたために発生しました。 | |

次のページに続く

表 12 – 前のページからの続き

| Description | Addresses |
|---|---|
| サブスクライバ専用ノードと監視ノードに関係する予期しないサブスクリプションとスロット作成を修正しました。<hr/>この修正により、サブスクライバ専用ノードと監視ノードの間でサブスクリプションとレプリケーションスロットの両方が誤って作成されるのを防ぎます。この修正は、PGD 5.9.0 から 6.2.0 へのアップグレード中に発生した、予期しないスロットが監視ノードに残された問題にも対処しています。 | |
| 接続マネージャーのルーティング変更ロジックの改善。<hr/>書き込みリーダーまたは読み取り専用ノードのルーティング変更が発生するたびに、接続マネージャーが既存のクライアント接続をドロップする問題を解決しました。この修正により、クライアントは、現在接続している PGD ノードが残っている限り、ルーティング更新の影響を受けなくなります。 | |
| 接続マネージャーで接続がサイレントにドロップされる問題を修正しました。<hr/>この修正により、接続マネージャーでの認証失敗の処理が向上します。 | |
| *pg_hba.conf* `address`フィールドで完全なホスト名が使用されているときに接続マネージャーがクラッシュする問題を修正しました。<hr/>この修正では、ホスト名サフィックスの処理も修正します。ホスト名のマッチングは接続マネージャーでサポートされていないため、ホスト名またはホスト名サフィックスを使用する HBA ルールが接続マネージャーによって正しく無視されるようになり、不安定な状態が防止されます。 | |
| 再起動後のライタエラー`could not open temporary file`を修正しました。<hr/>この問題は、大規模なトランザクションをストリーミングするときに発生し、ライタは再起動しましたが、一時ファイル名が誤って解放されたことを発見しました。このバグは並列適用で発生したもので、現在解決されています。 | 53341 |
| 別の PGD バージョンでクラスターに参加した後、ノードの`node_uuid`および`node_kind`の値が保持されるように問題を修正しました。<hr/>このバグにより、システムテーブル`bdr.local_node`および`bdr.node`の`node_uuid`が完全なクラスター後に分岐しましたアップグレード。この修正により、参加ノードの UUID と種類がプロセス全体にわたって正しく保持されるようになります。 | |
| galloc シーケンスを含むシーケンスまたはスキーマの名前の変更を許可します。<hr/>galloc シーケンスを含むシーケンスまたはスキーマの名前の変更などの同時操作が Raft 同期に問題を発生させる可能性がある問題を解決しました。この修正では、`bdr.sequence_kind`カタログ内の各シーケンスのグローバル UUID 識別子を導入します。これは、シーケンスまたはスキーマの名前が変更されても一定のままです。 | |

次のページに続く

表 12 – 前のページからの続き

| Description | Addresses |
|---|---|
| `pgd node setup`の物理的参加のバージョンとディストリビューションチェックを追加しました。<hr/>`pgd node setup`コマンドは、物理的参加を許可する前に、より厳密な互換性チェックを実行するようになりました。<ul><li>ローカルノードとリモートノードの BDR バージョンが完全に一致しない場合</li><li>ローカルノードとリモートノードの Postgres メジャーバージョンが同じでない場合、コマンドは物理的参加の代わりに論理的参加を試行します。</li><li> li><li>ローカルノードとリモートノードの PostgreSQL ディストリビューションフレーバーが同じでない場合、コマンドはエラーで終了します。</li></ul> | |
| `pgd group`コマンドの後方互換性の問題を修正しました。<hr/>`pgd group show`などの PGD 6.x CLI コマンドが、古い PGD 5.x クラスターに対して呼び出されたときにエラー `No query found for version`を報告する下位互換性の問題を解決しました。修正により、これらのコマンドは、メジャー PGD バージョンにわたって正しく機能するように保証されます。 | |
| `pgd node setup`の無効なシステム識別子チェックを修正しました。<hr/>指定されたリモートノードの有効なベースバックアップノードを検証するときに`pgd node setup`コマンドが誤って不一致エラーを報告した問題を解決しました。 | |
| ハンドルは、`bdr.run_on_x`ファンクションの `CONTEXT`値を返しました。<hr/>PGD CLI は、データベース接続の冗長度を調整して `CONTEXT`値を無視するようになりました。以前は、特に`bdr.run_on_x`ファンクションからのエラーの場合、出力にノイズを追加していました。 | |
| 接続タイムアウトを導入することにより、`bdr.msgb_connect`のハングを修正しました。<hr/>ネットワークタイムアウト設定が欠落しているために`bdr.msgb_connect`ファンクションが無期限にハングする可能性がある問題を解決しました。この修正により、ファンクションにタイムアウトが追加されます。 | 53442 |
| 既存のデータディレクトリから PGD ノードを作成するように `pgd node setup`を修正しました。<hr/>`pgd node setup`コマンドが既存のデータディレクトリを使用して PGD ノードを正常に作成および構成できない問題を解決しました。 | |
| 接続マネージャーは、現在ルーティングが有効になっているノードグループの一部を含む、すべてのノードの設定済みポートで起動および応答するようになりました。<hr/>これらの場合、トップレベルのノードグループから継承した構成が自動的に適用されます。 | |
| 主キーと一意キーの両方で構成されたテーブルの `UPDATE`オペレーション中に発生する可能性のある一意制約違反エラーを修正しました。<hr/>以前は、`update_missing`解決と `UPDATE`競合新しいレコードを挿入しようとしますが、競合のチェックがありませんでした。既存の主キーまたは一意キーと照合し、エラーが発生します。この問題は、シナリオを`insert_exists`競合として扱い、適切な解決ロジックを適用することにより修正されました。 | 50539, 51417, 54323 |

次のページに続く

表 12 – 前のページからの続き

| Description | Addresses |
|---|---|
| `ALTER TABLE SET pgd.replicate_to_analytics`プロパティを個々のパーティションに適用できるようになりました。<hr/>これにより、どのパーティションが分析ターゲットにデータをレプリケートするかをきめ細かく制御できます。 | |
| 重複する値が `analytics_storage_location`または `analytics_write_catalog`設定のいずれかに割り当てられたときにコンセンサス破壊を引き起こした問題を修正しました。 | |

## 2.5 EDB Postgres Distributed 6.1.2 release notes

リリース日 2025 年 11 月 6 日

EDB Postgres Distributed 6.1.2 には、安定性と信頼性の向上に焦点を当てたバグ修正のみが含まれています。このリリースには、新機能や機能強化は含まれません。詳細は以下をご覧ください。

### 2.5.1 バグ修正

| Description | Addresses |
|---|---|
| テーブルリライトを使用したストリーミングトランザクション中の ALTER TABLE の失敗を修正しました<hr/>`ALTER TABLE`コマンドは、テーブルリライトサポートとファイルまたはライタへのストリーミングを使用する場合、一時テーブルの作成に失敗しました。この修正では、ストリームメッセージのコールバック方法を実装し、これらの操作中の潜在的なデータ損失シナリオを防止します。 | |

## 2.6 EDB Postgres Distributed 6.1.1 release notes

リリース日 2025 年 10 月 9 日

EDB Postgres Distributed 6.1.1 には、安定性と信頼性の向上に焦点を当てたバグ修正のみが含まれています。このリリースには、新機能や機能強化は含まれません。詳細は以下をご覧ください。

### 2.6.1 バグ修正

| Description | Addresses |
|---|---|
| 'pgd node setup' を実行する際の物理的参加を修正する<hr/>PGD CLI の `pgd node setup`コマンドは、物理レプリカとして物理コピーを正しくセットアップするように修正されました。以前は、バグによりコマンドで物理レプリケーションを適切に構成できず、その結果、新しいノードで一部の変更が失われる可能性がありました。 PGD CLI の以前のバージョンで `pgd node setup`を使用して追加されたノードについては、一貫性チェックを実行することをお勧めします。 | |
| テーブルリライトをトリガーする DDL コマンドの処理を修正します<hr/>特定の DDL 操作バイナリ互換でない場合の `ALTER TABLE ... ADD COLUMN ... DEFAULT <volatile expression>`または `ALTER TABLE ... ALTER COLUMN ... SET DATA TYPE`などを実行すると、レプリケーション障害が発生する可能性があります。 PGD 6.1.1 のこの修正により、レプリケーションは、テーブル書き換えをトリガーする DDL コマンドを正しく処理できるようになりました。 | |
| 接続マネージャーは現在のスキーマ JDBC パラメーターを尊重します<hr/>以前は、 PGD Connection Manager を介して接続された JDBC クライアントは `currentSchema`パラメーターを無視していました。その結果、接続の `search_path`は誤ってデフォルトの `public`に設定されます。この問題は現在解決されています。 | |
| prefer_analytics_engine は、Iceberg REST カタログテーブルで動作します<hr/>`prefer_analytics_engine`フラグが更新され、Iceberg REST カタログにレプリケートされるテーブルで正しく機能するようになりました。 `bdr.prefer_analytics_engine = true`を設定すると、`pgaa.replicate_to_analytics = true`を使用してテーブルを参照するクエリーが、可能な場合に分析エンジンによって実行されます。 | |

## 2.7 EDB Postgres Distributed 6.1.0 release notes

リリース日 2025 年 8 月 19 日

これは、 PGD 6 のマイナーリリースであり、 PGD 4.4.0+および PGD 5.9.0+からの直接アップグレードサポートを導入します。これにより、古いバージョンからのシームレスなインプレースアップグレードが可能になる PGD 6 の最初のバージョンになります。この重要な機能に加えて、PGD 6.1.0 には、さまざまな新機能、機能強化、および重要なバグ修正が含まれています。

### 2.7.1 ハイライト

- **PGD 4 および PGD 5 から PGD 6 へのシームレスなアップグレード** PGD 4.4.0+および PGD 5.9.0+ユーザーは、PGD 6.1.0 で始まる PGD 6 への直接アップグレードを実行できるようになり、スムーズな移行を保証します。最新の機能まで。
- **Apache Iceberg へのレプリケーション** PGD は、Analytics Accelerator を使用する際に Apache Iceberg へのレプリケーションをサポートするようになりました。この新機能により、PGD クラスターから Iceberg データレイクにデータを直接ストリーミングし、PGD 内から Iceberg データを直接照会できます。これにより、大規模なデータの分析と統合のオプションが大幅に拡大します。

- **PGD 6 がマテリアライズドビューをサポートするようになりました** CREATE、ALTER、REFRESH、および DROP コマンドはクラスター全体でレプリケートされ、すべてのノードでマテリアライズドビューの一貫性が保証されます。マテリアライズドビューが作成されるか、同時に更新されると、上流ノードにデータが設定され、他のノードにレプリケートされ、この機能の完全なサポートを提供します。
- **PGD 6 が IBM LinuxONE** に登場しました。 PGD のネイティブ機能が IBM LinuxONE の新しいサポートで拡張され、最適化されたアップタイム、スループット、エネルギー効率を提供します。

### 2.7.2 機能

| Description | Addresses |
|---|---|
| Postgres 設定 GUC を取得および更新するための CLI コマンドを追加しました。<hr/>`pgd node get-config`および `pgd node set-config`コマンドを追加して、ユーザーが CLI を介して Postgres 設定 GUC を取得および変更できるようにします。 | |
| `bdr.autopartition`ファンクションに追加された新しい`drop_after_retention_period`引数<hr/>`bdr.autopartition()`は、追加のブール引数 `drop_after_retention_period`を取ります。ユーザーがパーティションをドロップせず、デタッチのみを希望する場合、false に設定できます。 | 82332 |
| PGD CLI を使用したノード構成の簡素化<hr/>pgd コマンドラインツールには、個々のノード構成の管理を簡単にする新しいコマンドが含まれるようになりました。 `pgd node <node-name> get-config`および`pgd node <node-name> set-config`を使用して、クラスター内のノードの特定の Postgres 設定 GUC を CLI から直接取得および更新できるようになりました。 | |
| PGD CLI のコミットスコープの表示<hr/>PGD CLI には、クラスターのコミットスコープ構成を表示できる新しいコマンド `pgd commit-scopes list`が含まれています。このコマンドは、ルールや適用されるグループを含むすべてのコミットスコープの詳細を提供します。これにより、データの一貫性設定の可視性と管理が向上します。 | |
| PGD CLI-クラスターからノードを分離する<hr/>PGD CLI には、新しいコマンド `pgd node part`が含まれ、アクティブなクラスターからノードを削除する合理的な方法を提供します。このコマンドは、 *node_part* SQL ファンクションを呼び出すプロセスを自動化し、その進行状況をレポートし、コマンドラインから直接ノード管理を簡素化します。 | |
| リーダー DDL ロック<hr/>新しいロックタイプのリーダー DDL ロックは、DDL ステートメントのロックにデフォルトで使用されます。このロックは書き込みリーダーのみをロックし、マジョリティノードがロック操作に参加する必要はありません。 `bdr.ddl_locking`構成パラメーターを調整することにより、古い動作を復元できます。 | |
| 監視ノードへのロールのレプリケーション<hr/>ロールは、ノードグループに参加したとき、およびデータノードで作成または変更されたときに監視ノードに複製されるようになりました。これは、主に GUC 権限を処理する場合に役立ちます。 | |

次のページに続く

表 15 – 前のページからの続き

| Description | Addresses |
|---|---|
| PGD クラスターのコミットスコープをリストするコマンドを追加しました。<hr/>コマンド `pgd commit-scopes list`を追加して、PGD クラスターのコミットスコープ情報をリストします。 | |
| クラスターから PGD ノードを分離するコマンドを追加しました。<hr/>コマンド `pgd node part`を追加して、クラスターから PGD ノードを分離します。 | |

### 2.7.3 機能強化

| Description | Addresses |
|---|---|
| デフォルト値を表示するように`pgd group get-option`および`pgd group show`コマンドを更新します。<hr/>`pgd group get-option`および`pgd group show`コマンドは、すべてのオプションのデフォルト値を表示するようになりました。`pgd group show`コマンドは、オプション値のソースも表示します。 | |
| CLI コマンドの PGD Essential コンストレインを適用します。<hr/>CLI は、必要に応じてコマンドの PGD Essential コンストレインをチェックおよび適用します。詳細については、<a href="https://www.enterprisedb.com/docs/pgd/latest/essential-how-to/pgd-cli/#pgd-essential-constraints">PGD Essential CLI コンストレイン</a>を参照してください。詳細。 | |
| Raft スナップショットのエクスポート/インポートを改善して Raft を完全にリセットします<hr/>Raft コンセンサスメカニズムがスナップショットを処理する方法を改善しました。これにより、まれにデータの一貫性の問題が発生した場合、スナップショットを復元するとシステムの状態が完全にリセットされ、一貫した開始ポイントから操作をスムーズに再開できます。 | |
| PGD の EPAS インターバルパーティションのサポート<hr/>新しい機能強化により、EDB Postgres Advanced Server AutoPartition によって作成されたパーティションは、マルチノード PGD 環境内で正常に動作できます。この修正により、パーティションが各ノードで一貫して作成され、すべてのノードで可視されるようになります。 | |
| max_connections エラーレポートの向上<hr/>接続マネージャーは、max_connections 制限に到達したときに、より役立つ具体的なエラーメッセージを提供するようになりました。ユーザーには、「申し訳ありませんが、既に多すぎるクライアントがあります。」のような明確なメッセージが表示されます。接続の問題の診断が非常に簡単になります。 | |
| 明示的なトランザクションブロックでテーブルリライトを生成するコマンドのサポート<hr/>PGD は、トランザクションブロック内でテーブルリライトを生成する ALTER TABLE コマンドの実行をサポートするようになりました。単一のトランザクションブロックで複数のコマンドがサポートされています。これにより、すべての変更を単一のアトミック単位として適用することにより複雑な DDL 操作が簡素化され、複数のコマンドが同じテーブルをターゲットとしている場合にパフォーマンスの最適化を提供します。 | |

次のページに続く

表 16 – 前のページからの続き

| Description | Addresses |
|---|---|
| PGD との Postgres 互換性-フェーズ 5<hr/>これは、 Postgres の互換性に関する将来の更新のためのプレースホルダーです。 | |

### 2.7.4 の変更

| Description | Addresses |
|---|---|
| listen_address グループオプションの削除<hr/>構成を簡素化するために `listen_address`オプションは `pgd group set-option`から削除されました。接続マネージャーは、*listen_addresses* GUC からの値を直接使用するようになり、以前はサポートされていなかったホスト名または複数のアドレスの設定に関する問題を解決します。 | |

### 2.7.5 バグ修正

| Description | Addresses |
|---|---|
| `witness-only`グループの `pgd cluster verify`コマンド警告を修正しました。<hr/>`pgd cluster verify`コマンドは、`witness-only`グループの `Data nodes per group`チェックの警告を表示しました。このチェックは`witness-only`グループには適用されないため、ステータス `Ok`が表示されます。 | |
| setup コマンドのデータディレクトリチェックを修正します。<hr/>`pgd node setup`コマンドは、指定されたデータディレクトリオプションに `PG_VERSION`ファイルの存在をチェックします。ただし、コードは誤って `pgdata`ではなく `bindir`オプションを使用しています。 | |
| 混合バージョンクラスターの修正<hr/>Postgres のローリングアップグレード中に一時的に使用されるアーキテクチャである混合バージョンの PGD クラスター Postgres 12 と Postgres 16 の組み合わせなどでレプリケーションが失敗する可能性のあるバグを解決しました。以前は、新しいノードの DDL ステートメントは、「認識されない構成パラメーター」により古いノードでレプリケーションスロットの切断を発生させる可能性がありました。エラー。この修正により、ローリングアップグレード中または一時的な混合バージョン状態でレプリケーションが正しく機能するようになりました。 | 46373 |
| ルーティングの変更に関する接続マネージャーの動作<hr/>接続マネージャーは、クラスタールーティング構成の変更たとえば、グローバルとルーティングなしの切り替えなど、次の接続を正しく処理するようになりました。これにより、予期しない動作が解決され、構成の更新後にクライアント接続が意図したとおりにルーティングされます。 | |

次のページに続く

表 18 – 前のページからの続き

| Description | Addresses |
|---|---|
| 分割されたノードを再構築する際の node_pkey エラーの修正<hr/>一意コンストレイン`node_pkey`に違反した重複したキー値でコンセンサスワーカーがクラッシュする可能性があるバグに修正が追加されました。分割されたノードが再構築されているときに発生したエラー。この修正により、この問題が発生せずに`bdr_init_physical`を使用して分割されたノードを正常にリビルドできるようになりました。 | 44467, 45084 |
| Witness-Only グループの PGD CLI cluster verify コマンドの修正<hr/>pgd cluster verify コマンドは、データノードを含まない witness-only グループの警告を表示しなくなりました。これにより、コマンドがこれらのグループに誤ってフラグを立てた問題が解決され、検証出力がすべてのクラスタータイプで正確になるように保証されます。 | |
| 遅延ノードを分離するための Raft の修正<hr/>以前にリーダーであった大幅に遅延したノードがクラスターから分離された後、Raft は復旧に失敗しなくなりました。これにより、Raft リーダーが" でクラッシュする可能性があるバグが修正されます。 BDR ノード XXX が見つかりませんでした。エラー。以前は Raft 状態の手動で強制的なリセットが必要でした。現在、Raft はこのシナリオを正しく処理し、自動的に復旧し、クラスターの安定性を保証します。 | 50099 |
| PGD CLI マルチグループセットアップの修正<hr/>マルチグループクラスターの作成を妨げていた PGD CLI setup コマンドのバグが修正されました。以前は、ノードを 2 番目のグループに追加しようとすると、コマンドは役に立たないエラーで失敗していました。この修正により、コマンドが新しいグループのノードの作成を正しく処理するようになり、マルチグループクラスターのセットアップを成功させることができます。 | |
| CLI の setup コマンド GUC Update の修正<hr/>PGD CLI の setup コマンドは、リモートノードの Postgres 設定を更新しようとしても失敗しなくなりました。この修正は、「不正な配列リテラル」を解決します。クラスターのセットアップ中に発生したエラーを修正し、初期化プロセスの信頼性が向上しました。 | |
| 「時計が逆方向に移動しました」の修正スノーフレークシーケンスのエラー<hr/>「クロックが逆方向に移動しました」を発生させる可能性のある競合状態が修正されました。システムクロックが正常に機能している場合でも、Snowflake シーケンスを使用するとエラーが発生します。この修正により、シーケンスが同時アクセスを正しく処理できるようになり、このエラーを防止し、Snowflake シーケンス生成の信頼性が向上します。 | 49376, 43659 |
| pg_failover_slots によるクラッシュを修正します<hr/>リカバリーが進行中である場合、またはバックエンドがデータベースに接続されていない場合、CDC フックが no-op であることを確認します。これは、`pg_failover_slots`がアクティブな場合に必ず発生する可能性があります。 | |
| 接続マネージャー stat の stat ビューのクラッシュを修正します<hr/>ビュー`bdr.stat_connection_manager_hba_file_rules`には、`pg_hba.conf`ファイル、特に`all`の`Address/Mask`列でいくつかの解析問題がありました。値。 | |

次のページに続く

表 18 – 前のページからの続き

| Description | Addresses |
|---|---|
| ノード種類チェック<hr/>PGD CLI には、ノード種類が適切に設定されていないノードの検証チェックが含まれるようになりました。このチェックは、ノード種類の設定を逃した以前のメジャーバージョンのアップグレードからのノードで警告を発します。 | |
| `pgd node setup`コマンドでノード GUC の更新をスキップします。<hr/>`pgd node setup`は、PGD ノードのセットアップ中に Postgres 設定を検証します。CLI が使用する `bdr_superuser`ロールは ALTER SYSTEM クエリーを実行できないため、コマンドはノード GUC の更新をスキップします。コマンドは Postgres 設定の検証を続行し、設定の変更が必要な場合は警告メッセージを表示します。 | |

## 2.8 EDB Postgres Distributed 6.0.2 release notes

リリース日 2025 年 6 月 25 日

これは、 EDB Postgres Distributed 6.0 の最初のパッチリリースであり、バグ修正と PGD 6.0 で導入された新機能の機能強化が含まれています。

### 2.8.1 バグ修正

| Description | Addresses |
|---|---|
| `node setup`コマンドによって生成されたスロット名に関する問題を修正しました。<hr/>コマンドは、PGD スロット名と互換性のないアンダースコア `_`を含む UUID 値としてスロット名を生成しました。 | |
| bdr.stat_connection_manager_get_stats() のセグメンテーションフォールトを修正しました。<hr/>ファンクションによって返されたメモリは、不正なメモリコンテキストに割り当てられました。 | |
| BDR オブジェクトの GRANT チェックが EPAS オブジェクトタイプで動作することを確認します<hr/>ユーザーに BDR 拡張オブジェクトの権限を付与できるかどうかを決定するメカニズムは、EPAS 固有のオブジェクトで失敗していました。これは PGD リリース 5.7.0 からの回帰であり、現在修正されています。 | 49649 |
| `pgd cluster show`コマンドのチェックの名前が `Clock Skew`から `Clock Drift`に変更されました。<hr/>チェックの名前は、PGD の他の場所で使用される用語と一貫性を持って変更されます。 | |
| CLI が `pgd cluster verify`コマンドの GUC の値の取得に失敗した場合の警告メッセージを追加しました。<hr/>CLI が何らかの理由で GUC の値を取得できない場合、コマンドはクラッシュする可能性があります。 | |

## 2.9 EDB Postgres Distributed 6.0.1 release notes

リリース日 2025 年 6 月 9 日

PGD 6 は、Postgres のよりシンプルで復元力の高い高可用性を提供します。従来のストリーミングレプリケーションでは、アップグレードと日常的なメンテナンスのためのダウンタイムが必要になることが多く、複雑なツールに依存しています。PGD は、ビルトインの論理レプリケーションベースのアーキテクチャでこれらの課題を解決し、アプリケーションを中断せずにオンラインアップグレードとメンテナンスを可能にし、チームが運用の変更中でもサービスをスムーズに実行し続けるのを支援します。また、シームレスなフェイルオーバーを提供し、外部プロキシ、ロードバランサー、またはコンセンサスシステムの必要性を排除します。

### 2.9.1 ハイライト

- **新しいビルトイン接続マネージャー**クライアント接続を正しいノードに自動的にルーティングし、アプリケーションアーキテクチャを簡素化し、動的なトポロジの変更をサポートし、ビルトインセッションプーラーと専用の読み取り/書き込みおよび読み取り専用ポートを含みます。外部ソフトウェアや複雑な構成は必要ありません。この新しいコンポーネントは、PGD 6 以降では使用できなくなった PGD プロキシを置き換えます。
- **事前定義されたコミットスコープ**ビルトインのトランザクション持続性プロファイルを使用して一貫性の選択を簡素化します。複雑なセットアップは必要ありません。システムカタログで定義され、すぐに使用できるスコープを使用して、パフォーマンスと保護の適切なバランスを選択します。
- **クラスターセットアップのための新しい CLI コマンド** pgd node setup コマンドにより、コマンドラインから直接初期クラスター作成とノードの追加が可能になりました。これにより、ユーザーは PGD を展開する方法の柔軟性が向上し、展開ツールを一貫した方法で標準化できます。

### 2.9.2 機能

| Description | Addresses |
|---|---|
| ビルトイン接続マネージャー<hr/>接続のルーティングを自動的に処理し、非リーダーへの読み取り専用接続を強制できる新しいビルトイン接続マネージャー。 | |
| CLI クラスターセットアップ<hr/>PGD CLI では、クラスターの初期セットアップと、`pgd node setup`コマンドを使用してコマンドラインからノードを追加できるようになりました。 | |
| group create/join でシーケンス種類を設定する<hr/>古いバージョンで行ったようにノードを作成するときではなく、bdr グループを作成/参加するときに、*bdr.default_sequence_kind* GUC に基づいて分散シーケンスを変換します。 | |

次のページに続く

表 20 – 前のページからの続き

| Description | Addresses |
|---|---|
| 分散シーケンスの start 値を自動的に設定します<hr/>galloc シーケンスの start 値を、ローカルシーケンスで最後に使用された後に次の有効な数値に設定します。この変更により、分散シーケンス、特に galloc を作成するときに、既に使用されているものに基づいて start 値を調整する必要がありません。 | |
| 自動同期と調整の有効化<hr/>ノード状態の自動同期と調整がデフォルトで有効になりました。これは、ノードが自分の状態をリーダーノードと自動的に同期し、手動介入を必要とせずに差異を調整することを意味します。詳細については、<a href="/pgd/latest/reference/node_management/automatic_sync">ドキュメント</a>をご覧ください。 | |
| node_uuid 列を bdr.node および bdr.local_node に追加します<hr/>node_uuid は、指定された名前のノードのインスタンスを一意に識別します。ランダムな node_uuid は、ノードの作成時に生成され、ノードの存続期間を通して一定のままです。 node_id 列が、ノード名ではなく node_uuid から取得されるようになりました。当分の間、同じ名前のノードを再参加できる前に、ノードを完全に分離する必要があります。これは、将来のリリースで緩和され、part_node がプロセスされたらすぐに再参加を許可する可能性があります。古いインスタンスが開始され、完了する前に。当面は、古い PGD バージョンからのアップグレードと、古い PGD ノードを使用するクラスターでの混合バージョンのオペレーションはサポートされていません。この制限は、将来のリリースで解決される予定です。 | |
| レプリケーションオリジンとスロットの命名スキームを変更する<hr/>レプリケーションオリジンとスロット名はノード uuid を使用するようになったため、指定された名前のノードの特定の化身に対応します。同様に、グループ名の代わりにノードグループの uuid が使用されます。データベース名のハッシュは、データベース名の代わりに使用されます。オリジン名とノード名は、ユーザーの観点から不透明な識別子として扱う必要があることに注意してください。これらの名前の構造に依存したり、これらが人間にとって特に意味があると期待しないでください新しい命名スキームは次のとおりです。<h4>スロット命名規則</h4><ul><li>ノードへの通常のスロット => `bdr_node_<targetuuid>_<dbhash>`</li><li>ノードの参加スロット => `bdr_node_<targetuuid>_<dbhash>_tmp`</li><li>トップグループのグループスロット => `bdr_group_<topgroupuuid>_<dbhash>`</li><li>転送用のスロット+リード to リード => `bdr_node_<targetuuid>_<originidhex>_<dbhash>`</li><li>分析スロット => `bdr_analytics_<groupuuid>_<dbhash>`</li><li>デコーディングスロット => `bdr_decoder_<topgroupuuid>_<dbhash>`</li></ul><h4>オリジン命名規則</h4><ul><li>ノードへの通常のオリジン => `bdr_<originuuid>_<dbhash>`</li><li>送信元からソースノードへ =><li> `bdr_<originuuid>_<sourceoidhex>_<dbhash>`</li></ul> | |

次のページに続く

表 20 – 前のページからの続き

| Description | Addresses |
|---|---|
| PGD Essential のシステムで許可されるノードグループの数の制限。<hr/>任意の時点で、3 つ以下のノードグループ 1 つのトップグループと 2 つのサブグループが存在できることを確認します。制限を超えると、エラーが発生します。 | |
| PGD Essential 制限の適用 - データノード数<hr/>PGD Essential クラスターが 4 つを超えるデータノードに参加することを許可しないでください。 | |
| 指定が特定の LSN に到達するまで待機する `bdr.wait_node_confirm_lsn()`ファンクションを追加しました<hr/>`bdr.wait_node_confirm_lsn(`) は、利用可能な場合は指定されたノードの confirmed_flush_lsn を確認します。それ以外の場合は、そのノードの `pg_replication_origin_progress()`を照会し、指定された LSN に到達するのを待機します。ノード。 | |
| サブスクライバ専用ノードをデータノードグループに追加できるようになりました<hr/>以前のバージョンでは、サブスクライバ専用ノードはタイプ“subscriber-only”のノードグループにのみ追加できました。 PGD 6 では、create_node を使用するときに node_kind='subscriber_only' を指定することにより、サブスクライバー専用ノードをデータノードグループに追加することもできます。 join_node_group は、データノードグループを使用して実行できます。 | |
| *bdr.local_analytics_slot_name()* SQL ファンクションを追加します。<hr/>分析スロットの名前を返します。これは、分析機能が使用されているかどうかに関係なく、正しい名前を生成するだけです。 | |
| node_uuid 列を `bdr.node_summary`ビューに追加します。<hr/>bdr.node および bdr.local_node への node_uuid 列の追加を補完するために追加されました。 | |

### 2.9.3 機能強化

| Description | Addresses |
|---|---|
| 複数の競合行の解決<hr/>`pk_exists`と `multiple_unique_conflicts`の両方の競合タイプが、競合の一部である古い行を削除することにより、複数の競合行を解決できるようになりました。 `multiple_unique_conflicts`はデフォルトで`update_if_newer`リゾルバーになるため、デフォルトでエラーをスローしなくなりました。 | |
| `bdr.stat_activity`ビューの改善<hr/>`backend_type`は、追加のプロセス ID なしで PGD ワーカーの一貫したワーカータイプを表示するようになりました。`wait_event_type`および `wait_event`には、“extension”を表示する代わりに、より多くの待機イベントが含まれるようになりました。一部のイベント用。また、接続管理関連の列が追加され、実際のクライアントアドレス/ポート、およびセッションが読み取り専用かどうかを表示します。 | |

次のページに続く

表 21 – 前のページからの続き

| Description | Addresses |
|---|---|
| PARTED ノードは、クラスター内のすべてのノードから自動的に削除されます。<hr/>PGD 6.0.0 から、 bdr.part_node 機能が強化され、クラスター内のすべてのノードから parted ノードのメタデータを自動的に削除します。<ul><li>For </li><li>リモートノードの場合、その特定のノードのメタデータのみを削除します。したがって、このリリースでは</li><li>すべてのノードのグループスロットが PARTED ノードから発生するすべてのトランザクションに追いつくまで、ノードは PART_CLEANUP 状態にあります。</li><li>ノードは PARTED に残されませんノードが PARTED 状態に移動するとすぐに削除されるため、状態。</li></ul> | |
| *pgd node show* CLI コマンドの `--summary`および `--options`フラグ。<hr/>`--summary`および `--options`フラグを `pgd node show`コマンドに追加して、`pgd node show`コマンドの出力をフィルタリングします。これは、他の `show`コマンドとの対称性も維持します。 | |
| *pgd cluster verify* CLI コマンドで検証された GUC が増えました。<hr/>`bdr.lock_table_locking`および `bdr.truncate_locking` GUC を `pgd cluster verify`コマンドで検証された GUC のリストに追加します。 | |
| テーブルの書き換え `ALTER TABLE... ALTER COLUMN`呼び出しがサポートされるようになりました。<hr/>テーブル全体が書き換えられ、変更がバイナリ強制可能ではない列の type コマンドの変更がサポートされるようになりました。<pre><code class="lang-sql"> CREATE TABLE foo (c1 int,c2 int, c3 int, c4 box, UNIQUE(c1, c2) INCLUDE(c3,c4)); ALTER TABLE foo ALTER c1 TYPE bigint; – テーブルリライトへの結果 *</pre>This also includes support for `ALTER TYPE* when using the *USING* clause:<pre><code class="language-sql">CREATE TABLE foo (id serial primary key,data text); ALTER TABLE foo ALTER data TYPE BYTEA USING data::bytea; `</pre>テーブルリライトは、大規模なテーブルで長期間 AccessExclusiveLock を保持する場合があります。 | |
| 不変ではない `ALTER TABLE... ADD COLUMN`呼び出しの制限が削除されました。<hr/>不変ではない `ALTER TABLE... ADD COLUMN`呼び出しの制限が削除されました。 | |
| 論理結合中にロールとテーブルスペースを同期する<hr/>ロールとテーブルスペースは、結合ソースノードからスキーマが復元される前に同期されるようになりました。すでに既存のロールまたはテーブルスペースまたは EPAS プロファイルが存在する場合、それらは結合ソースノードからの設定と同じ設定、パスワードなどを持つように更新されます。システムロールつまり initdb によって作成されたものは同期されません。 | |
| `bdr.node_group_config_summary`ビューの紹介<hr/>新しい `bdr.node_group_config_summary`ビューには、有効値、有効値のソース、デフォルト値、値を継承できるかどうかを含むグループオプションに関する詳細情報が含まれています。これは、 `pg_settings`と同様の精神です。 | |

次のページに続く

表 21 – 前のページからの続き

| Description | Addresses |
|---|---|
| リーダー DML ロック<hr/>新しいロックタイプのリーダー DML ロックは、DML をブロックする必要がある DDL ステートメントをロックするためにデフォルトで使用されます。このロックは書き込みリーダーのみをロックし、すべてのノードがロック操作に参加する必要はありません。 `bdr.ddl_locking`構成パラメーターを調整することにより、古い動作を復元できます。 | |
| run_on_*ファンクションでの bdr.xact_レプリケーションの無効化<hr/>ファンクション`run_on_nodes`、*run_on_all_nodes*、および`run_on_group`は、デフォルトで`bdr.xact_replication`を`off`に設定するようになりました。 | |
| デフォルトでレプリカ ID がいっぱい<hr/>`bdr.default_replica_identity`の`auto`値が REPLICA IDENTITY FULL に変更されました。この設定は、ノードのクラッシュとリカバリーをまたいだ挿入、更新、および削除間の競合検出における一部のエッジケースを防ぎます。 PGD グループが作成され、最初の PGD ノードのデータベースが空でない場合、つまりデータが含まれるテーブルがいくつかある場合、のレプリカ ID すべてのテーブルは`bdr.default_replica_identity`に従って設定されます。 | |
| DDL オペレーションとしてのテーブルスペースレプリケーションがサポートされています。<hr/>テーブルスペース操作 `CREATE/ALTER/DROP TABLESPACE`が DDL オペレーションとしてレプリケートされるようになりました。ユーザーが同じマシンで複数のノードを使用した構成を実行している場合、開発者オプションを有効にする必要があります <a href=" <a href="https://www.postgresql.org/docs/current/runtime-config-developer.html#GUC-ALLOW-IN-PLACE-TABLESPACES">https://www.postgresql.org/docs/cu config-developer.html#GUC -ALLOW-IN-PLACE-TABLESPACES</a> ">`allow_in_place_tablespace`</a>。 | |
| CLI デバッグメッセージを改善します。<hr/>ログメッセージのフォーマットを改善して、読みやすく Postgres ログメッセージと対称的です。 | |
| *pgd cluster verify --settings* CLI コマンド出力の新しい列。<hr/>`pgd cluster verify --settings`コマンドの結果に`recommended_value`列を追加します。列は表形式出力には表示されませんが、JSON 出力には表示されます。 | |
| CLI の並べ替えられた出力を表示します。<hr/>表形式の出力を使用するコマンドの出力がリソース名で並べ替えられるようになりました。複数のリソースを表示するコマンドは、各リソース列で出力を順番にソートします。 | |
| サブスクライバー専用ノードのレプリケーション。<hr/>サブスクライバー専用ノードは、大部分のデータノードにレプリケートされた後にのみデータを受信するようになりました。これには、特別な構成は必要ありません。その後、 bdr.standby_slot_names および bdr.standby_slots_min_confirmed オプションは、同様のフィジカルスタンバイ機能が pg_failover_slots 拡張機能および PG17+で提供されているため、削除されます。 | |

次のページに続く

表 21 – 前のページからの続き

| Description | Addresses |
|---|---|
| ノードの自動同期とリコンシリエーションはデフォルトで有効になっています。<hr/>GUC <a href="/pgd/latest/reference/tables-views-functions/pgd-settings#bdrenable_auto_sync_reconcile">`bdr.enable_auto_sync_reconcile`</a> はデフォルトでオフでした、ただし、6.0 ではデフォルトでオンになっています。この GUC 設定により、ノードがしばらくダウンしているときに、他のすべてのノードがこのノードに関して等しく自動的にキャッチされます。また、ダウンするノードによって孤立した準備されたトランザクションがある場合、それらは解決され、それらを作成したコミットスコープのルールに従ってアボートまたはコミットされます。 | |
| 非推奨の古い CLI コマンドを削除します。<hr/>古い PGD 5 以下の CLI コマンドを削除します。これらは非推奨ですが、下位互換性のためにサポートされています。 | |
| コミットスコープロジックはデータノードでのみ実行されるようになりました。<hr/>以前は、非データノードはコミットスコープロジックを処理しようとしましたが、無害ではありますが、混乱を招く可能性がありましたログメッセージ。 | |
| ダンプとリストアオペレーションの開始と停止を明示的にログに記録します。<hr/>これにより、ノードクローン作成プロセスの可視性が向上し、考えられる問題のデバッグが役立ちます。 | |

### 2.9.4 の変更

| Description | Addresses |
|---|---|
| サブグループでルーティングがデフォルトで有効になるようになりました<hr/>ルーティングおよび拡張子 raft が、データグループデータノードを含むサブグループでデフォルトで有効になるようになりました。 | |
| ファンクション`bdr.join_node_group`は、トランザクションで実行されなくなる場合があります。<hr/>グループ参加をロールバックすることはできないため、べき等トランザクションの一部を形成することはできません。 | |
| 非推奨の `pause_in_standby`パラメーターがファンクション`bdr.join_node_group()`から削除されました。<hr/>`pause_in_standby`は、PGD 5.0.0 以降非推奨になりました。 `bdr.create_node()`を実行するときに論理スタンバイノードを指定する必要があります | |
| BDR グローバルシーケンスは、`UNLOGGED`として作成または設定できなくなりました。ログに記録されていない BDR シーケンスは、サーバーのクラッシュに続いて予期しない動作を表示する場合があります。既存のログのない BDR シーケンスは、ログに記録されたものに変換できます。 | |

### 2.9.5 バグ修正

| Description | Addresses |
|---|---|
| 低下したクラスターでの CLI `pgd cluster show`コマンドの問題を修正します。<hr/>N ノードクラスターで 1 つのノードのみが起動および実行されている場合、`pgd cluster show`コマンドはクロックドリフトのエラーで失敗しました。コマンドは、他のコンポーネント`health`および`summary`の有効な出力を返すようになりましたが、 `clock-drift`の適切なエラーを報告します。 | |
| 存在しないノードが指定された場合の CLI `pgd node show`コマンドの問題を修正します。<hr/>存在しないノードがコマンドに指定されている場合、`pgd node show`コマンドはクラッシュしました。コマンドは、適切なエラーメッセージを表示して正常に失敗するように修正されました。 | |
| `pgd node upgrade`コマンドを使用して Postgres のローリングアップグレード後に、壊れたレプリケーションスロットの問題を修正します。<hr/>ノードのアップグレード中に、ライタの原点位置を親原点にマージします。 PGD 5 以前のライタのオリジン名は、インプレースアップグレード中に変更される可能性がある親オリジン ID にマップされます。 | 46412,48747 |
| *pgd replication show* CLI コマンドのタイムスタンプ解析の問題を修正しました。<hr/>`pgd replication show`コマンドは、EPAS タイムスタンプをフォーマットするときにクラッシュしていました。 | |
| 分割するノードが存在しないグループに属する場合がある問題を修正しました<hr/>特定のノードを分割すると、その同じノードに、オリジンが既に分割されグループがドロップされたサブスクリプションを持つ場合があります。以前は、これにより PGD が中断されていましたが、修正されました。 | |
| num_writers は正または-1 である必要があります<hr/>bdr.alter_node_group_option() および bdr.alter_node_group_config() で使用される num_writers オプションは、正または-1 である必要があります。 | |
| 非一意のインデックスの更新によるレプリケーションの破損を修正します<hr/>非一意のインデックスを使用したテーブルの更新で ERROR `concurrent INSERT when looking for delete rows`が発生し、レプリケーションを中断する場合を修正します。 | 43523,43802,45244,47815 |
| アップグレード後の Raft リーダー選択タイムアウト/失敗を修正します<hr/>非推奨の GUC `bdr.raft_election_timeout`で設定されたカスタム値が置換用の`bdr.raft_global_election_timeout`に適用されることを確認します | |
| サブスクライバ専用ノードの無効なサブスクリプションが再度有効になっていないことを確認します<hr/>サブスクリプションの再構成中に、サブスクリプションに変更が必要ない場合は、ユーザーが明示的に無効にされている可能性があるため、有効にしないでください。リーダーシップの変更がない場合、サブスクリプションの再構成をスキップします。 | 46519 |
| サブスクライバー専用ノードは、DDL の実行時にロックを取得しません<hr/>サブスクライバー専用ノードは、 DDL の実行時にクラスターでロックを取得しようとしなくなりました。 DDL はローカルで実行され、他のノードにレプリケートされません。 | 47233 |

次のページに続く

表 23 – 前のページからの続き

| Description | Addresses |
|---|---|
| データベースシステムのシャットダウンでのハングを修正しました。<hr/>データベースシステムの高速シャットダウン中に WAL 送信者が終了しない原因となる可能性のある、ダウンストリームによる非トランザクション WAL メッセージの確認応答を修正しました。 | 49022 |
| bdr_init_physical のデッドロックの問題を修正しました。<hr/>bdr_init_physical の不要なノードデータのクリーニングと同時モニタリングクエリ間のデッドロックを修正しました。 | 46952 |
| 新しいクラスターノードの一貫性の問題を修正しました。<hr/>クラスターに参加する新しいノードが、すべてのデータノードに対するレプリケーションの進行状況を取得する前に CATCHUP フェーズを終了した場合の問題を修正しました。これにより、新しいノードがクラスターと同期しない場合があります。 | |
| CREATE SEQUENCE 警告に正しいシーケンスタイプが表示されることを確認します<hr/>シーケンスが実際には `snowflakeid`であるときに、警告メッセージが `timeshard`を参照する場合がありました。 | |

# 第3章　Overview

EDB Postgres 分散 PGD は、高度な競合管理、データ損失保護、および throughput up to 5X faster than native logical replication を備えたマルチマスターレプリケーションとデータ分散を提供します。また、最大ファイブナインの高可用性を備えた分散 Postgres クラスターも有効にします。

## 3.1 Known issues and limitations

### 3.1.1 既知の問題

これらは EDB Postgres Distributed 製品バージョン 6.4 の現在既知の問題です。これらの既知の問題は PGD のチケットシステムで追跡され、将来のリリースで解決される予定です。

- マテリアライズドビューのレプリケーションは、新しく作成されたマテリアライズドビューでのみ機能します。製品バージョン 6.4.0 にアップグレードする前に作成されたマテリアライズドビューは、アップグレード後に複製されません。この問題を防ぐには、次のいずれかを行うことができます。
  - PGD 製品バージョン 6.4.0. にアップグレードする前に、各ノードでマテリアライズドビューを手動で作成します。
    * OR
  - PGD にアップグレードした後 {{version.full }}.0.、マテリアライズドビューをドロップして作成して、複製されるようにします。
- `update_origin_change` 競合のリゾルバーが `skip` に設定され、`synchronous_commit=remote_apply` が使用され、同じ行の同時更新が 2 つの異なるノードに繰り返し適用される場合、PGD ライタとのデッドロックが原因で更新ステートメントの 1 つがハングする場合があります。*bdr_read_all_conflicts* で述べたように、 `skip` は `update_origin_change` 競合のデフォルトのリゾルバーではなく、この組み合わせは運用環境で使用することを目的としていません。そのノードへの到着順序に基づいて、2 つの競合する更新のうちの 1 つを破棄します。これにより、発散クラスターが発生する可能性があります。`skip` 競合リゾルバーの使用を選択するまれな状況では、`remote_apply` モードの使用に関する問題に注意してください。
- Decoding Worker 機能は、CAMO/Eager/Group Commit では動作しません。CAMO/Eager/Group コミットを使用したインストールでは、`enable_wal_decoder` を無効のままにする必要があります。
- Lag Control は、完全に分離されたノードではいかなる方法でもコミット遅延を調整しません。これは、他のすべてのノードが到達不能または動作していない場合です。少なくとも 1 つのノードが接続すると、レプリケーションラグコントロールはその作業を取得し、PGD コミット遅延を再度調整します。
- 時間ベースのラグ制御の場合、PGD は現在、履歴適用レートに基づいた推定キャッチアップ時間ではなく、コミットタイムスタンプによって測定されたラグタイムを使用しています。

- CAMO ペアの CAMO パートナーを変更することは現在不可能です。ペアを追加または削除することのみが可能です。ペアを追加または削除するには、Postgres の再起動や構成のリロードも必要ありません。
- グループコミットは *Commit At Most Once* と組み合わせることはできません。
- Eager Replication を使用するトランザクションは、まだ DDL を実行できません。TRUNCATE コマンドが使用できます。
- パラレル適用は、グループコミットとの組み合わせで現在サポートされていません。グループコミットを使用するときは、(a) bdr.alter_node_group_option を使用してノードグループの `num_writers` を 1 に設定するか、(b) GUC bdr.writers_per_subscription を使用して、必ず無効にしてください。

Configuration of generic replication を参照してください。

- 現在、コミットスコープの変更または削除に対する保護はありません。同時に変更または削除されるコミットスコープでトランザクションを実行すると、トランザクションを適用しようとしているダウンストリームノードでのエラーが原因で、トランザクションブロックまたはレプリケーションが完全に停止する可能性があります。変更または削除する前に、特定のコミットスコープを使用するトランザクションが終了していることを確認してください。
- PGD CLI は、クラスターから分離されたノードにまだ接続している場合、クラスターの状態に関する古いデータを返すことができます。

pgd-cli-config.yml ファイルを編集するか、--dsn 設定を変更して、クラスター内のアクティブなノードのみが接続用にリストされるようにします。

コミットスコープを安全に変更するには、 bdr.alter_commit_scope を使用します。

- シリアル化可能なトランザクションで実行される DDL は、エラー `ERROR: could not serialize access due to read/write dependencies among transactions` が発生する場合があります。回避策は、シリアル化可能なトランザクションの外部で DDL を実行することです。
- EDB Postgres Advanced Server 17 データ型

**BFILE は、現在サポートされていません。これは、**
`BFILE` がデータベースに保存されるファイル参照であり、ファイル自分自身はデータベースの外部に保存され、レプリケートされないためです。

- パラレル適用を有効にしてコミュニティ PostgreSQL で PGD を実行している場合、ライタプロセスはロックタイムアウトのため予期せず終了する場合があります。これは、PGD がライタプロセスで 30 秒の `lock_timeout` を設定するために発生します。これは、あるライタが、大規模なトランザクションを処理している別のライタが保持するトランザクションロックで待機する場合、不十分である場合があります。タイムアウトに達すると、待機中のライタが終了し、他のライターとレシーバーも同様に終了し、レプリケーションプロセスがリセットされます。回避策として、`bdr.enable_parallel_apply = off` を設定してパラレル適用を無効にするか、 EDB Postgres バリアントに移行します。この問題は PGD 6.3 で解決されました。
- PGD 6.3.0 では、ユーザーのログイン属性は、 `CREATE USER` ステートメントの後にピアノードにレプリケートされません。1 つのノードで作成されたユーザーは、元のノードにログインできますが、レプリケートされたロール定義にログイン属性が欠落しているため、他のすべてのノードでブロックされます。この問題は、PGD 6.3.1 で解決されました。

影響を受けるロールを特定するには、 psql で次のコマンドを実行します。

```
\du
```

**Attributes** 列では、ログインアクセスがある必要があるが `Cannot login` と表示されるロールが影響を受けます。

または、システムカタログを直接照会します。

```
SELECT rolname AS role_name, rolcanlogin
FROM pg_roles
WHERE rolcanlogin = false;
```

アップグレードせずに問題を解決するには、影響を受ける各ノードでログインアクセスを手動で付与します。

```
ALTER USER <username> WITH LOGIN;
```

### 3.1.2 制限事項

展開を計画するときは、次の EDB Postgres 分散 PGD デザインの制限を考慮してください。

#### ノード

- PGD は、適切なハードウェアとネットワークを想定して、数百のノードを実行できます。ただし、メッシュベースの展開の場合、通常は 1 つのクラスターで 48 を超えるノードを実行することはお勧めしません。48 ノード制限を超えて追加の読み取りスケーラビリティが必要な場合は、メッシュネットワークへの接続を追加せずにサブスクライバ専用ノードを追加できます。
- PGD のコンセンサスメカニズムにフォールトトレランスを提供するために、グループ内のノードの最小推奨数は 3 です。2 つのノードのみの場合、ノードの 1 つが応答しない場合、コンセンサスは失敗します。分散シーケンス生成などの一部の PGD 操作では、コンセンサスが必要です。EDB Postgres Distributed が使用するコンセンサスメカニズムの詳細については、*Architectural details* を参照してください。

#### 単一のインスタンス上の複数のデータベース

同じ Postgres インスタンス上の複数のデータベースで PGD を使用することのサポートは、PGD 5 から **非推奨** であり、PGD 6 ではサポートされなくなります。製品の機能を拡張するにつれて、操作的および機能的に導入される追加の複雑さはありません。マルチデータベースデザインではより長く実行可能です。

これがベストプラクティスであり、PGD インスタンスごとに 1 つのデータベースのみを構成することをお勧めします。

現在、CLI や Connection Manager などのツールがその推奨事項を成文化しています。

単一のインスタンスで最大 10 のデータベースをホストすることは引き続き可能ですが、そうすると多くの差し迫ったリスクと現在の制限が発生します。

- PGD 構成の変更が必要な場合は、データベースごとに管理コマンドを実行する必要があります。これを行うと、潜在的な不一致とエラーのリスクが増加します。
- 各データベースを個別に監視する必要があり、オーバーヘッドを追加します。

- 接続マネージャーは、データベースレベルではなく Postgres インスタンスレベルで動作します。つまり、リーダーノードはすべてのデータベースで同じです。
- データベースを追加するごとに、サーバーのリソース要件が増加します。それぞれが、レプリケーションを維持する独自のワーカープロセス、たとえば、論理ワーカー、WAL センダー、WAL レシーバーなどを必要とします。それぞれは、レプリケーションクラスター内の他のインスタンスへの独自の接続セットも必要です。これらのニーズは、すべてのデータベースのパフォーマンスに重大な影響を与える可能性があります。
- CAMO やグループコミットなどの同期レプリケーション方法は、予想通りに動作しません。Postgres WAL はデータベース間で共有されるため、同期コミット確認は任意のデータベースから取得できますが、必ずしも正しいコミットの順序ではありません。
- CLI 統合は、1 つのデータベースを想定しています。

### 耐久性オプション クォーラムコミット/グループコミット/ CAMO

PGD 耐久性オプションの動作にはさまざまな制限があります。これらの制限は、Quorum Commit、 Group Commit、および CAMO 間の相互作用、およびそれらが *Decoding worker* トランザクションストリーミングでの使用 などの PGD 機能とどのように相互作用するかの製品です。

また、古い同期レプリケーションとの相互運用性、明示的な 2 フェーズコミットとの相互運用性、およびコミットスコープルール内でサポートされていない組み合わせにも制限があります。

次の制限は、コミットスコープの使用と、それらが有効にするさまざまな耐久性オプションに適用されます。

#### 一般的な耐久性の制限

- *Legacy synchronous replication using PGD* は、CAMO、Eager、グループコミット、およびクォーラムコミットで使用されるものとは異なるトランザクション確認のメカニズムを使用します。この 2 つは互換性がないため、一緒に使用しないでください。Quorum Commit、 Group Commit、 CAMO、または Eager を使用するときは常に、 `synchronous_standby_names` で PGD ノードが構成されていないことを確認します。
- Postgres の 2 フェーズコミット 2PC トランザクションつまり

**PREPARE TRANSACTION**
: は、 CAMO、 Group Commit、Quorum Commit、または Eager では使用できません。これらの機能は下で 2 フェーズコミットを使用するためです。

#### グループコミット

*Group Commit (legacy)* は、グループ内のノードを介した構成可能な同期コミットを有効にします。この機能を使用する場合、次の制限を考慮してください。

- グループコミットを使用する場合、すべての DDL を実行できるわけではありません。サポートされていない DDL を使用する場合、警告がログに記録され、トランザクションのコミットスコープがローカルに設定されます。サポートされている DDL 操作は次のとおりです。
  - 非コンカレント `CREATE INDEX`
  - 非コンカレント `DROP INDEX`

  - 個々のテーブルまたはインデックスの非コンカレント REINDEX
  - CLUSTER 単一のリレーションまたはインデックスのみ
  - ANALYZE
  -
  -
  -
  -
  -

- 明示的な 2 フェーズコミットは、既に 2 フェーズコミットを使用しているため、グループコミットではサポートされていません。
- 同じトランザクションでの異なるコミット決定オプションの組み合わせ、または同じトランザクションでの異なる競合解決オプションの組み合わせは、サポートされていません。
- 現在、Raft のコミット決定は非常に遅く、非常に低い TPS を生成します。eager 競合解決設定でのみこれらを使用して、PGD 4 以前の Eager All-Node Replication 動作を取得することをお勧めします。

#### 熱心な

熱心な競合解決 は、グループコミットを介して利用でき、クォーラムコミットで常にアクティブです。競合する可能性のあるトランザクションを積極的に中止することにより、競合を回避します。グループコミットを介して使用される場合、グループコミットと同じ制限の対象となります。

Eager では、 NOTIFY SQL コマンドまたは pg_notify() ファンクションは許可されていません。 LISTEN または UNLISTEN も許可されません。

#### クォーラムコミット

クォーラムコミット は、分散トランザクションの一貫性のためのコミットスコープの種類です。この機能を使用する場合、次の制限を考慮してください。

- クラスター内のすべてのノードは PGD 6.4.0 以降を実行している必要があります。以前のバージョンのノードを含むクラスターは、ローリングアップグレード中のクラスターを含む、クォーラムコミットを使用できません。
- 明示的なグループリストを使用した MAJORITY 、 ANY n 、および NOT はサポートされていません。サポートされているグループは ALL 、MAJORITY ORIGIN GROUP 、MAJORITY CLUSTER 、および MAJORITY FOR EACH GROUP です。
- トランザクションストリーミングは、file または off モードでのみサポートされています。他のストリーミングモードは、クォーラムコミットと互換性がありません。
- Quorum Commit は既に内部で 2PC を使用しているため、明示的な 2 フェーズコミット 2PC は Quorum Commit ではサポートされていません。
- DEGRADE TO はサポートされていません。

### CAMO

*Commit At Most Once* CAMO は、アプリケーションが複数回コミットするのを防ぐことを目的とした機能です。この機能を使用する場合、計画時に次の制限を考慮してください。

- CAMO は、オリジンノードで最近失敗した COMMIT の結果を照会するように設計されています。切断が発生した場合、アプリケーションは CAMO パートナーからトランザクションステータスを要求する必要があります。障害後、ステータスを要求する前に、遅延ができるだけ少ないことを確認します。アプリケーションは、15 分を超えて保存される CAMO 決定に依存してはなりません。
- 再起動などの結果、アプリケーションが割り当てられたグローバル識別子を忘れた場合、それを回復する簡単な方法はありません。したがって、アプリケーションはシャットダウンする前に、未処理のトランザクションが終了するのを待機することをお勧めします。
- クライアントが適切なチェックを適用するには、CAMO によって保護されたトランザクションを、暗黙的なトランザクション制御を備えた単一のステートメントにすることはできません。また、CAMO は、トランザクション制御プロシージャ、またはトランザクションを開始または終了しようとする `DO` ブロックで使用することはできません。
- CAMO はコミットステータスを解決しますが、コミット時の保留中の通知を解決しません。CAMO、`NOTIFY` SQL コマンドまたは `pg_notify()` ファンクションは許可されていません。 `LISTEN` または `UNLISTEN` も許可されません。
- 変更を再生するときに、CAMO トランザクションは他のトランザクションと同様に競合を検出する場合があります。タイムスタンプ競合検出が使用される場合、CAMO トランザクションは、プリペアオンザオリジンノードのタイムスタンプを使用します。これは、トランザクションがオリジンノード自分自身で可視される前です。
- CAMO は現在トランザクションストリーミングと互換性がありません。CAMO を使用する予定がある場合は、必ずトランザクションストリーミングを無効にしてください。このオプションはグローバルにまたは PGD ノードグループで構成できます。*Transaction streaming configuration* を参照してください。
- CAMO は現在デコードワーカーと互換性がありません。CAMO を使用する予定がある場合は、デコードワーカーを有効にしないようにしてください。このオプションは PGD ノードグループで構成できます。*Decoding worker disabling* を参照してください。
- CAMO を使用する場合、すべての DDL を実行できるわけではありません。サポートされていない DDL を使用する場合、警告がログに記録され、トランザクションのコミットスコープはローカルのみに設定されます。サポートされている DDL 操作は次のとおりです。
  - 非コンカレント `CREATE INDEX`
  - 非コンカレント `DROP INDEX`
  - 個々のテーブルまたはインデックスの非コンカレント `REINDEX`
  - `CLUSTER` 単一のリレーションまたはインデックスのみ
  - `ANALYZE`
  -
  -

  -
  -
  -

- 明示的な 2 フェーズコミットは、既に 2 フェーズコミットを使用しているため、CAMO ではサポートされていません。
- CAMO トランザクションのみを `DEGRADE TO` 句と組み合わせて、可用性が低下した場合に非同期オペレーションに切り替えることができます。

#### 混合 PGD バージョン

PGD は、アップグレードプロセス中に PGD の混合バージョンが動作できるようにすることにより、*enable rolling upgrades of PGD* に開発されました。私たちは、ユーザーがアップグレード中にのみ混合バージョンを実行し、アップグレードが開始されたら、そのアップグレードを完了することを想定しています。アップグレード時を除き、PGD の混合バージョンの実行はサポートされていません。

#### その他の制限

この非包括的なリストには、予想されるデザインによる他の制限が含まれています。将来的にそれらを解決する予定はありません。展開を計画するときは、次の制限を考慮してください。

- `galloc` シーケンスは、ロールバックしたトランザクションでシーケンスを作成し、同じ名前で再度作成する場合、一部のチャンクをスキップする場合があります。チャンクのスキップは、DDL レプリケーションがアクティブでないときにシーケンスを作成して削除し、DDL レプリケーションがアクティブなときに再度作成する場合にも発生する可能性があります。順序保証に違反しないため、問題の影響は軽度です。シーケンスは、一部の最初のチャンクのみをスキップします。また、回避策として、シーケンスの開始値を `bdr.alter_sequence_set_kind()` ファンクションの引数として指定できます。

## 3.2 PGD compatibility

### 3.2.1 PGD の PostgreSQL バージョンとの互換性

次の表は、EDB Postgres DistributedPGD と互換性のある PostgreSQL のメジャーバージョンを示しています。

| PGD 6 | Postgres Version |
|---|---|
| *PGD 6 の UUID 関連の変更* | 18.1.0+ |
| *PGD 6 の UUID 関連の変更* | 17.5.0+ |
| *PGD 6 の UUID 関連の変更* | 16.9.0+ |
| *PGD 6 の UUID 関連の変更* | 15.13.0+ |

次のページに続く

表 1 – 前のページからの続き

| PGD 6 | Postgres Version |
|---|---|
| *PGD 6 の UUID 関連の変更* | 14.18.0+ |

EDB では、サポートされている PGD で Postgres メジャーバージョンの最新のマイナーバージョンを使用することをお勧めします。

### 3.2.2 オペレーティングシステムおよびアーキテクチャとの PGD の互換性

次の表は、 EDB Postgres Distributed のバージョンと、さまざまなオペレーティングシステムおよびアーキテクチャとの互換性を示しています。

#### Linux

| Operating System | x86_64 (amd64) | ppc64le | arm64/ aarch64 |
|---|---|---|---|
| RHEL 8 | Yes | Yes | |
| RHEL 9 | Yes | Yes | Yes |
| Oracle Linux 8 | Yes | | |
| Oracle Linux 9 | Yes | | |
| Rocky Linux/AlmaLinux | Yes | | |
| SUSE Linux Enterprise Server 15SP6 | Yes | Yes | |
| SUSE Linux Enterprise Server 15SP7 | Yes | Yes | |
| Ubuntu 22.04 | Yes | | |
| Ubuntu 24.04 | Yes | | |
| Debian 12 | Yes | | Yes |
| IBM Linux One | Yes | | |

**注釈**

PGD の以前のバージョンについては、 *Compatibility changes* を参照してください。

## 3.3 Postgres Distributed terminology

この用語リストには、 EDB Postgres DistributedPGD に関連する馴染みのない用語が含まれています。

### 3.3.1 非同期レプリケーション

オリジナルノードでトランザクションが完了した後、他の PGD クラスターメンバーにデータをコピーするレプリケーションのタイプ。非同期レプリケーションは、*Legacy synchronous replication using PGD* よりも高いパフォーマンスと低いレイテンシーを提供できます。ただし、非同期レプリケーションでは、さまざまなクラスターメンバーに変更が表示されるまでにかかる時間に遅延が発生する場合があります。クラスターは **最終的な一貫性** になりますが、ノードが明らかに互いに同期していない可能性があります。

### 3.3.2 コミットスコープ

PGD クラスターのノードとグループ間でトランザクションがコミットされる方法を管理するためのルール。*Legacy synchronous replication using PGD* 、 **クォーラムコミット** 、 **同期コミットとグループコミット** 、 *CAMO* **またはコミット・アット・モスト・ワンス** 、**熱心な競合解決** 、 **ラグコントロール**、およびその他の PGD 機能を構成するために使用されます。

### 3.3.3 CAMO またはコミット・アット・モスト・ワンス

一部のアプリケーションの高価値トランザクションでは、アプリケーションが正確に 1 回だけ、フェールオーバーと再試行が発生した場合に 1 回だけ正常にコミットする必要があります。PGD でこれが発生するように保証には、CAMO を有効にして、アプリケーションがトランザクションにアクティブに参加できるようにします。

### 3.3.4 競合

データは PGD クラスターのノードにわたってレプリケートされるため、あるソースからの変更が別のソースからの変更と衝突する場合があります。これは競合であり、競合の解決で処理できます。競合の解決は、どのソースが正しいまたは優先されるかを決定する一連のルールです。競合は、競合のないデータ型を使用して回避することもできます。

### 3.3.5 接続マネージャー

クライアントが分散クラスター内の適切なノードに接続できるように、PGD は、クライアントがニーズに基づいて適切なノードに接続できる接続管理システムを提供します。

このシステムは、クライアントがクラスターのパフォーマンスと可用性を維持しながら、必要なデータにアクセスできるように設計されています。プロキシシステムとは異なり、この接続管理システムはデータベースインスタンス自分自身に組み込まれているため、より効率的で信頼性の高い接続が可能になります。

### 3.3.6 コンセンサス

**コミット決定に** *Raft* **を使用する** がグループ全体の決定を行う方法。グループ内のノードの数を指定すると、Raft は、決定に投票するマジョリティのコンセンサスノードの数を 2 プラス 1 で割ったものを探します。たとえば、書き込みリーダーが選択される場合、グループ内のどのノードが書き込みリーダーになるかについて Raft コンセンサスが求められます。投票メンバーのクォーラムがある場合にのみコンセンサスに到達できます。

### 3.3.7 クラスター

一般に、クラスターは、エンドユーザーに 1 つのシステムとして見えるように配置された複数のシステムのグループです。*pgd cluster* および *Postgres* **クラスター** も参照してください。

### 3.3.8 DDL データ定義言語

データベースの構造の定義と管理を処理する SQL コマンドのサブセット。DDL ステートメントは、データベース内のオブジェクトつまりスキーマ、テーブル、インデックスを作成、変更、および削除できます。一般的な DDL コマンドは、CREATE、ALTER、および DROP です。

### 3.3.9 DML データ操作言語

データベースに保持されているデータの操作を処理する SQL コマンドのサブセット。DML ステートメントは、データベース内のテーブルの行を作成、変更、および削除できます。一般的な DML コマンドは、INSERT、UPDATE、および DELETE です。

### 3.3.10 耐久性

持続性は、トランザクションがコミットされると、システム障害が発生した場合でもコミットされたままであることを保証します。PGD は、 *Commit scopes* を介してこれを管理します。

### 3.3.11 熱心な

競合する可能性のある着信トランザクションを検出し、それらのいずれかを「迅速に」中止して一貫性を維持することにより競合を回避する同期コミットモード。

### 3.3.12 最終的な一貫性

異なるクラスターメンバーの同じ項目への変更を記述する分散コンピューティング一貫性モデルは、最終的に同じ値に収束します。競合解決と競合のないレプリケートデータ型を使用した非同期論理レプリケーションは、PGD の最終的一貫性を示します。

### 3.3.13 フェイルオーバー

高可用性データベースクラスターの障害を認識し、一貫性と可用性を維持するためのアクションを実行する自動プロセス。目標は、ダウンタイムとデータの損失を最小限に抑えることです。

### 3.3.14 グループコミット

複数の PGD ノードがコミット時にトランザクションを正常に受信および確認する必要がある同期コミットモード。

### 3.3.15 即時一貫性

すべてのレプリカが同期的かつ同時に更新される分散コンピューティングモデル。このモデルでは、書き込みが完了した後のすべての読み取りが、すべてのノードで同じ値を参照します。このアプローチの欠点は、パフォーマンスへの悪影響です。

### 3.3.16 遅延制御

Lag Control は、セカンダリノードがプライマリノードから大きく遅れないようにするために使用される調整メカニズムです。ノードのレプリケーション遅延が設定されたしきい値を超えると、PGD は元のノードでの取り込みを自動的に調整して、クラスターの残りの部分が追いつくようにします。

### 3.3.17 論理レプリケーション

データベース内の変更をレプリケートするより効率的な方法。物理的ストリーミングレプリケーションは元のデータベースのディスクブロックを複製しますが、論理レプリケーションは代わりに、基になる物理的記憶形式とは無関係に変更を取得し、変更を表示するようにサブスクライブしたすべてのシステムに公開します。次に、各サブスクライバーは変更をローカルに適用します。論理レプリケーションは、ほとんどの DDL コマンドをサポートできません。

### 3.3.18 論理スタンバイノード

論理スタンバイノードは、クラスター内のデータの読み取り専用レプリカを提供するために使用されます。これらは加入者専用ノードに似ていますが、より柔軟になるように設計されており、より幅広いシナリオで使用できます。

### 3.3.19 ノード

分散システムの要素の総称。ノードは任意のサービスのホストとして機能できます。PGD では、*Creating PGD nodes* は Postgres データベース、BDR 拡張機能、および Connection Manager を実行します。

通常、高可用性を実現するために、各ノードは別の物理ハードウェアで実行されますが、常にそうとは限りません。

### 3.3.20 ノードグループ

PGD クラスター内の PGD ノードをグループに編成して、クラスターの論理操作を反映できます。たとえば、特定の物理的な場所にあるデータノードは、その場所の専用ノードグループの一部であることができます。

各グループは、*Monitoring Raft* コンセンサス アルゴリズムを使用してリーダーを選択します。データ固有のグループでは、このリーダーは書き込みリーダーとして機能し、すべての着信書き込み操作を処理して一貫性を維持します。

グループはデータのレプリケーションを管理し、「トップレベル」グループはクラスター全体を表し、他のすべてのサブグループの親として機能します。

### 3.3.21 PGD クラスター

エンドユーザーには 1 つのシステムとして見えながら、単一障害点を回避するように配置された複数の冗長データベースシステムとプロキシのグループ。PGD クラスターは、Docker インスタンス、クラウドインスタンス、または「ベア」Linux ホスト、またはこれらのプラットフォームの組み合わせで実行できます。PGD クラスターには、バックアップノードを含めることもできます。クラスター内のデータノードは、トップレベルグループにグループ化され、さまざまなローカル**ノードグループ** にグループ化されます。

### 3.3.22 PGD ノード

PGD クラスターには、データベースを実行し、PGD クラスターに参加するノードがあります。一般的な PGD ノードは、Postgres データベース、 BDR 拡張機能、および Connection Manager を実行します。 PGD モードは **データノード**とも呼ばれ、データを保存することを示唆しています。ただし、一部の PGD ノード、特に *Witness ノード* はそれを行いません。

### 3.3.23 物理的レプリケーション

物理レプリケーションは、1 つ以上のスタンバイクラスターメンバーに変更されるときにデータベースディスクブロックの正確なコピーを作成することにより、サーバーをレプリケートする簡単に実装された方法を提供します。ただし、使用方法には制限があります。たとえば、1 つのマスターノードのみが書き込みトランザクションを実行できます。また、この方法では、すべてのクラスターメンバーが、同じオペレーティングシステムと CPU アーキテクチャを備えたデータベースソフトウェアの同じメジャーバージョンを使用している必要があります。

### 3.3.24 Postgres クラスター

通常、PostgreSQL では、単一のサーバーで実行されている多数のデータベースをデータベースのクラスターと呼びます。この種類の Postgres クラスターは、可用性が高くありません。高可用性と冗長性を取得するには、*pgd cluster* が必要です。

### 3.3.25 クォーラム

クォーラムとは、分散投票に参加するために必要な投票ノードの最小数です。これにより、行われた決定が有効であることが保証されます。たとえば、PGD クラスターで **コミット決定に** *Raft* **を使用する** *Monitoring Raft* **コンセンサス** が必要な場合、投票に参加する最小数の投票ノードが必要です。5 ノードクラスターの場合、クォーラムはクラスター投票の 3 ノードです。コンセンサスは 5/2+1 ノード、3 つのノードが同じように投票します。投票ノードが 2 つだけの場合、コンセンサスは決して確立されません。PGD では、*DDL locking details* および Raft の決定にクォーラムが必要です。

### 3.3.26 クォーラムコミット

ノードがローカルにコミットする前に、すべての参加ノードにわたってコミット決定を調整することにより、分散トランザクションの一貫性を保証するコミットスコープの種類。すべてのノードは一緒にコミットまたはロールバックします。熱心な競合解決は常にアクティブです。

### 3.3.27 レプリケートされた利用可能なフォールトトレランス Raft

分散クラスター内のマシンのクォーラムからの投票を使用してコンセンサスを確立するコンセンサスアルゴリズム。PGD は、グループトップレベルまたはローカル内で Raft を使用して、書き込みリーダーであるノードを確立します。

### 3.3.28 読み取りスケーラビリティ

増加する読み取りワークロードを処理するシステムの機能。たとえば、PGD は 1 つ以上のリードレプリカノードをクラスターに導入し、アプリケーションにプライマリノードに書き込みを指示し、レプリカノードに読み取りを行わせることができます。読み取りワークロードの増加に応じて、リードレプリカノードの数を増加してパフォーマンスを維持できます。

### 3.3.29 サブスクリプション

PGD ノードは、データに加えられた変更を関心のあるノードに公開します。他の PGD ノードは、これらの変更をサブスクライブするように求めます。この動作により、サブスクリプションが作成され、各ノードが更新されるメカニズムです。PGD ノードは、他の PGD ノードの変更を双方向にサブスクライブします。

### 3.3.30 スイッチオーバー

アプリケーションまたはプロキシとクラスター内のアクティブなデータベースノード間の接続の計画的な変更。通常はメンテナンスのために行われます。

### 3.3.31 同期レプリケーション

変更がすべての参加ノードで同時に更新される場合、通常は 2 フェーズコミットを活用します。このアプローチでは、コミットする前に変更を複製し、競合を解決しますが、ノード全体で必要な調整のため、レイテンシーのパフォーマンスコストが発生します。

### 3.3.32 サブスクライバー専用ノード

PGD クラスターは、双方向レプリケーションをベースとしています。ただし、読み取り専用サーバーが必要な場合など、一部のユースケースでは、双方向レプリケーションは必要ありません。この場合、加入者専用ノードが使用されます。データベース内の変更のみをサブスクライブして自分自身を最新の状態に保ち、ノードで直接実行に対して正しい結果を提供します。この機能を使用して、PGD クラスターでのホリゾンタル読み取りのスケーラビリティを有効にできます。これらは 2 つの方法で構成できます。

- データグループ内これらのノードは、標準のデータノードと並んで読み取り専用レプリカとして機能します。これらはデータへのローカルアクセスを提供しますが、特定のレプリケーションの最適化のメリットはありません。
- 加入者専用グループ内この構成には、書き込みリーダーがありません。すべてのノードは厳密に読み取り専用であり、PGD はレプリケーションプロセスを合理化するオプションの最適化を適用できます。

### 3.3.33 2 フェーズコミット 2PC

複数のデータベースノード間の一貫性を実現するためのマルチステップのプロセス。最初のフェーズでは、元のノードで準備され、すべての参加ノードに送信されるトランザクションが表示されます。各参加ノードは、トランザクションを適用できることを検証し、その準備ができていることを元のノードに通知します。これは準備フェーズです。第 2 フェーズでは、すべての参加ノードが準備ができていることを通知した場合、元のノードはトランザクションのコミットを続行し、参加ノードにもコミットするように通知します。これはコミットフェーズです。準備フェーズで、いずれかのノードが準備ができていないことを通知した場合、トランザクション全体が中止されます。このプロセスにより、すべてのノードが同じ変更を取得します。

### 3.3.34 垂直スケーリングまたはスケールアップ

Oracle Exadata など、そのアーキテクチャの物理的制限に到達するまで、特定のワークロードをサポートするリソース CPU、メモリ、ストレージ、ネットワークを増加する従来のコンピューティングアプローチ。

### 3.3.35 Witness ノード

Witness ノードは、主にクラスターがコンセンサスを確立するのを支援するために機能します。コンセンサスを確立するには、奇数のデータノードが必要です。リソースが制限されている場合、監視ノードを使用してクラスターの決定に参加できますが、データを複製することはできません。データを保持しないことは、スタンバイサーバーとして動作できない、または同期コミットのマジョリティを提供できないことを意味します。

### 3.3.36 ライトリーダー

Always-on アーキテクチャでは、ノードがアプリケーションの正しい接続エンドポイントとして選択されます。このノードは書き込みリーダーと呼ばれます。選択すると、PGD Connection Manager はクエリと更新をルーティングします。1 つのノードのみが書き込みを受信するため、意図しないマルチノードの書き込みを回避できます。書き込みリーダーは、データノードのクォーラムのコンセンサスによって選択されます。書き込みリーダーが使用できなくなると、データノードは別のノードを選択して書き込みリーダーになります。書き込みリーダーではないノードは、**シャドウノード** と呼ばれます。

### 3.3.37 ライター

**無効になったサブスクリプションへの対応** がデータ変更を PGD ノードに配信すると、データベースサーバーはライタと呼ばれるワーカープロセスに、これらの変更を適用するタスクを実行します。

## 3.4 Quickstart guide

このガイドでは、基本的な PGD クラスターのインストール、接続、データのロード、および基本的な SQL 操作の実行をガイドします。

- *Creating your first cluster* 。
- *Working with SQL and the PGD cluster* 。
- *Loading data into your PGD Cluster* 。

- *Using PGD CLI* 。

### 3.4.1 Creating your first cluster

このガイドは、Docker Compose を使用してローカルクラスターを作成するのに役立ちます。これは、EDB Postgres DistributedPGD の特徴と機能に慣れるための優れた方法です。

#### 前提条件

- ローカルマシンにインストールされている Docker および Docker Compose。

#### PGD Docker クイックスタートキットのインストール

最初の PGD クラスターを作成するには、EDB が提供する Docker Compose ファイルを使用できます。これにより、3 つのノードでローカルクラスターがセットアップされます。これは、テストと開発の目的に最適です。

1. ローカルマシンに Docker および Docker Compose がインストールされていることを確認します。Docker installation guide をまだ行っていない場合は、フォローできます。
2. ターミナルを開き、docker をインストールしたマシンで、PGD クラスターの新しいディレクトリを作成します。例

```
mkdir pgd-cluster
cd pgd-cluster
```

3. 次のコマンドを実行して、PGD Docker Compose ファイルをダウンロードします。

```
curl -L https://enterprisedb.com/docs/pgd/latest/overview/quickstart/assets/pgd_
↪quickstart.sh | bash
```

これにより、 PGD Docker クイックスタートキットがダウンロードされます。これには、Docker Compose ファイルおよび PGD Essential を開始するために必要なその他のファイルが含まれています。

4. ダウンロードが完了したら、PGD クラスターの環境を準備する必要があります。これは、次のコマンドを実行して実行します。

```
./qs.sh prepare
```

このコマンドは、PGD クラスターに必要なディレクトリとファイルを作成します。

5. 次に、PGD クラスターの Docker イメージをビルドする必要があります。これは、次のコマンドを実行して行うことができます。

```
export EDB_SUBSCRIPTION_TOKEN=...
./qs.sh build
```

このコマンドは、PGD クイックスタートクラスターに必要な Docker イメージをビルドします。

6. イメージがビルドされたら、Docker Compose を使用して PGD クラスターを起動できます。次のコマンドを実行します。

```
./qs.sh start
```

このコマンドは、Docker コンテナを起動し、バックグラウンドで実行されるデフォルト構成でローカルクラスターを作成します。

### PGD クラスターへのアクセス

1. コンテナが起動して実行されたら、次のコマンドを使用して PGD クラスターにアクセスできます。

```
docker compose exec host-1 psql pgddb
```

このコマンドは、 psql コマンドラインインターフェイスを使用して、クラスターの最初のノードに直接接続します。

これは、メンテナンスおよび管理タスクのためにデータベースに接続する方法です。

アプリケーションとユーザーのアクセスには、通常、クラスター内のすべてのホストの TCP ポート 6432 で実行されている接続マネージャーを使用して接続します。

1. 次のコマンドを使用して、クラスター内の書き込みリーダーノードに接続できます。

```
docker compose exec host-1 psql -h host-1 -p 6432 pgddb
```

-h host-1 は、クラスター内のホストの名前に置き換えることができます。それらはすべて接続マネージャーを実行します。

ローカルマシンに psql クライアントをインストールしている場合は、次のコマンドを使用してクラスターに接続することもできます。

```
export PGPASSWORD=secret
psql -h localhost -p 6432 -U postgres pgddb
```

これは、ポート 6432 の host-3 コンテナで実行されている接続マネージャーに接続します。これは、クラスター内の書き込みリーダーノードにルーティングされます。

```
pgddb=# select node_name from bdr.local_node_summary;
node_name
-----------
node-1
(1 row)
```

1. コンテナの外部から PGD CLI を使用するには、次のコマンドを実行できます。

```
docker compose exec host-1 pgd nodes list
__OUTPUT__
 Node Name | Group Name | Node Kind | Join State | Node Status
```

(次のページに続く)

(前のページからの続き)

```
- ----------+------------+-----------+------------+-------------
 node-1     | group-1    | data      | ACTIVE     | Up
 node-2     | group-1    | data      | ACTIVE     | Up
 node-3     | group-1    | data      | ACTIVE     | Up
```

この pgd コマンドは、クラスター内のノードとそのステータスをリストします。

host-1 コンテナでシェルを取得し、pgd コマンドを直接実行することもできます。

```
docker compose exec host-1 bash
pgd nodes list
__OUTPUT__
 Node Name | Group Name | Node Kind | Join State | Node Status
- ----------+------------+-----------+------------+-------------
 node-1     | group-1    | data      | ACTIVE     | Up
 node-2     | group-1    | data      | ACTIVE     | Up
 node-3     | group-1    | data      | ACTIVE     | Up
```

これにより、 PGD CLI にアクセスでき、host-1 コンテナで直接 PGD コマンドを実行できます。

#### 次のステップ

最初の PGD クラスターを作成したため、次のトピックを検討できます。

- *Working with SQL and the PGD cluster* SQL コマンドを使用してクラスターに接続および対話する方法を理解する。
- `COPY` コマンドまたは `pg_dump` および `pg_restore` を使用してクラスターに *Loading data into your PGD Cluster* をインストールします。
- *Using PGD CLI* クラスターを監視および管理します。

### 3.4.2 Working with SQL and the PGD cluster

PGD クラスターを使用する最初のステップは、SQL を使用して接続することです。これは、 `psql` コマンドラインインターフェイスまたは PostgreSQL をサポートする他の SQL クライアントを使用して実行できます。

クイックスタートキットの Docker Compose セットアップを使用している場合、`psql` コマンドは既にインストールされ構成されています。つまり、ホストの 1 つでシェルを取得するだけで作業が開始できます。`host-1` コンテナにサインインして、クラスター内でコマンドを実行できます。

```
docker compose exec host-1 bash
```

これにより、 `host-1` コンテナ内にシェルが提供され、PGD クラスターに対して PGD および SQL コマンドを実行できます。

### PGD クラスター内の接続

PGD Essential では、メンテナンスタスクを実行する場合を除き、通常、クラスター内のすべてのホストの TCP ポート 6432 で実行されている接続マネージャーを使用してクラスターに接続します。

次のコマンドを使用して、クラスター内の書き込みリーダーノードに接続できます。

```
psql -h <host> -p 6432 -U <username> <database>
```

ユーザーなしで実行されている新しいクラスターは postgres スーパーユーザーを除き 、1 つのレプリケートデータベース pgddb があるため、次のコマンドを使用してクラスターに接続できます。

```
psql -h host-1 -p 6432 -U postgres pgddb
```

これは、ポート 6432 の host-1 コンテナで実行されている接続マネージャーに接続し、クラスター内の書き込みリーダーノードにルーティングされます。host-1 は、クラスター内のホストの名前に置き換えることができます。それらはすべて接続マネージャーを実行します。

次のコマンドを実行すると、クラスター内のどのノードに接続しているかを確認できます。

```
select node_name from bdr.local_node_summary;
__OUTPUT__
 node_name
- ----------
 node-1
```

クラスター内で node-1 ノードを実行している host-1 コンテナに接続したため、これは驚くべきことではありません。

psql を終了し、次のように再接続すると

```
psql -h host-2 -p 6432 -U postgres pgddb
```

クラスター内の node-1 ノードに接続していることがわかります。

```
select node_name from bdr.local_node_summary;
 node_name
- ----------
 node-1
```

これは、クラスター内の書き込みリーダーノード node-1 にルーティングする接続マネージャーです。これを確認するには、次のことを実行できます。

```
\! pgd group group-1 show --summary
__OUTPUT__
 Group Property    | Value
- ------------------+---------
 Group Name        | group-1
 Parent Group Name | pgd
```

(次のページに続く)

(前のページからの続き)

```
Group Type          | data
Write Leader        | node-1
Commit Scope        |
```

psql の\! コマンドを使用して、psql セッション内からシェルコマンドを直接実行できます。

### SQL での作業

クラスターに接続したので、SQL コマンドでの作業を開始できます。通常の PostgreSQL データベースを使用するのと同じように、テーブルを作成し、データを挿入し、クエリを実行できます。

たとえば、テーブルを作成し、いくつかのデータを挿入できます。

```
CREATE TABLE users (
    id SERIAL PRIMARY KEY,
    name VARCHAR(100),
    email VARCHAR(100) UNIQUE
);
INSERT INTO users (name, email) VALUES
(Alice, alice@example.com),
(Bob, bob@example.com);
```

次に、データを照会できます。

```
SELECT * FROM users;
__OUTPUT__
 id |  name  |        email
- ---+--------+---------------------
  2 | Alice  |      alice@example.com
  3 | Bob    |      bob@example.com
(2 rows)
```

より複雑なクエリーを実行し、テーブルを結合し、PostgreSQL のすべての機能を使用することもできます。すべての SQL コマンドをカバーすることはこのガイドの範囲内ではありませんが、SQL 構文とコマンドの詳細については、PostgreSQL documentation を参照してください。

### PGD との違い

重要なのは、これらの SQL コマンドがクラスター全体でレプリケートされることです。PGD がレプリケーションを処理します。たとえば、その serial キーはクラスター全体でグローバル一意キーに自動的に変換されているため、クラスター内の任意のノードにデータを挿入でき、それは他のすべてのノードにレプリケートされます。PGD Essential の場合、書き込みリーダーに接続する必要があるため、これはそれほど重要ではありませんが、PGD Expanded では、クラスター内の任意のノードに接続して SQL コマンドを実行できます。この自動変更により、サービスを心配せずにそれを行うことができます。競合または重複。PGD Essential を使用すると、将来も対応でき、SQL コマンドやアプリケーションコードを変更せずに、後で簡単に PGD Expanded に移行できます。

#### 次のステップ

PGD クラスターに接続し、いくつかの SQL コマンドを実行したため、次のトピックを探索できます。

- *Loading data into your PGD Cluster* 外部ソースからデータをインポートする方法を学習します。
- *Using PGD CLI* コマンドラインから PGD クラスターを管理します。

### 3.4.3 Loading data into your PGD Cluster

PGD は、その中核が Postgres データベースであるため、PostgreSQL データベースを使用するのと同じツールと方法を使用して、PGD クラスターにデータをロードできます。まず、このガイドでは、PGD クラスターにデータをロードするプロセスを説明します。

#### オンライン CSV インポート

最初に、オンライン CSV ファイルから PGD クラスターにデータをインポートする方法を示します。この場合、それは Baseball Databank からの履歴野球データです。 psql の\COPY コマンドを使用して、URL から直接インポートします。 \COPY がしないことの 1 つはテーブルを作成することです。したがって、最初にテーブルを作成する必要があります。

psql を使用して PGD クラスターに接続します。docker compose exec host-1 psql を使用するか、psql がローカルにインストールされている場合は、それを使用してホストマシンのポート 6432 に接続します。

```
CREATE TABLE batters (
                      id SERIAL,
                      playerid VARCHAR(9),
                      yearid INTEGER,
                      stint INTEGER,
                      teamid VARCHAR(3),
                      lgid VARCHAR(2),
                      g INTEGER,
                      ab INTEGER,
                      r INTEGER,
                      h INTEGER,
                      "2b" INTEGER,
                      "3b" INTEGER,
                      hr INTEGER,
                      rbi INTEGER,
                      sb INTEGER,
                      cs INTEGER,
                      bb INTEGER,
                      so INTEGER,
                      ibb INTEGER,
                      hbp INTEGER,
                      sh INTEGER,
                      sf INTEGER,
                      gidp INTEGER,
```

(次のページに続く)

(前のページからの続き)

```
                    PRIMARY KEY (id)
);
```

これで、 \COPY コマンドを使用して CSV データを batters テーブルにインポートできます。

```
\COPY batters(playerid,yearid,stint,teamid,lgid,g,ab,r,h,"2b","3b",hr,rbi,sb,cs,bb,so,
↪ibb,hbp,sh,sf,gidp) FROM PROGRAM curl "https://raw.githubusercontent.com/cbwinslow/
↪baseballdatabank/master/core/Batting.csv" DELIMITER , CSV HEADER
```

このコマンドは、curl を使用して URL から CSV ファイルを取得し、\COPY コマンドに直接パイプします。このコマンドは、 batters テーブルにデータをインポートします。rotations(…) エントリーは、CSV データを行のどのフィールドに移動するかを定義します。DELIMITER ',' CSV HEADER オプションは、ファイルがスキップされるヘッダー行を含む CSV であることを指定します。

コマンドをコピーして、psql セッションに貼り付けます。すべてが正しく設定されている場合、データがエラーなしでインポートされるのが表示されるはずです。次のような、コピーされた行数を示す出力が表示されます。

```
COPY 110495
```

データが正しくロードされたことを確認するために、単純なクエリーを実行できます。

```
SELECT COUNT(*) FROM batters;
```

次のような結果が表示されるはずです。

```
 count
- ------
 110495
(1 row)
```

これは、110,495 行が batters テーブルに正常にインポートされたことを確認します。

すぐにそれを使用して、1998 年のホームランリーダーが誰であったかを考えてみましょう

```
SELECT playerid, yearid, teamid, hr
FROM batters
WHERE yearid = 1998
ORDER BY hr DESC
LIMIT 1;
```

次のような出力が表示されるはずです。

```
 playerid  | yearid | teamid | hr
- ----------+--------+--------+----
 mcgwima01 |   1998 | SLN    | 70
(1 row)
```

そして、それを 1998 年の最高ランクのホームラン打者トップ 5 のコンテキストに置きたい場合は、次のことができます。

```
SELECT playerid, yearid, teamid,
       rank() OVER (PARTITION BY yearid ORDER BY hr desc) hr_rank,
       hr
FROM batters
WHERE yearid = 1998
ORDER BY hr_rank LIMIT 5;
```

次のような出力が表示されるはずです。

```
 playerid  | yearid | teamid | hr_rank | hr
- ----------+--------+--------+---------+----
 mcgwima01 |   1998 | SLN    |       1 | 70
 sosasa01  |   1998 | CHN    |       2 | 66
 griffke02 |   1998 | SEA    |       3 | 56
 vaughgr01 |   1998 | SDN    |       4 | 50
 belleal01 |   1998 | CHA    |       5 | 49
(5 rows)
```

PGD を使用すると、ウィンドウファンクションのような高度な SQL 機能を含む PostgreSQL のフルパワーを活用してデータを分析できますが、ノードがオフラインになった場合でもマルチプルのノードにわたって完全にレプリケートされ、可用性が高いという追加の利点があります。

#### 次のステップ

いくつかのデータを PGD クラスターにロードしたので、次のトピックを探索できます。

- *Using PGD CLI* コマンドラインから PGD クラスターを管理します。

### 3.4.4 Using PGD CLI

PGD CLI は、EDB Postgres 分散 PGD クラスターを管理およびモニタリングするためのコマンドラインインターフェイスです。ノードの作成、ノードの参加、レプリケーションの管理など、クラスターでさまざまな操作を実行するためのコマンドセットを提供します。

*Creating your first cluster* を使用している場合、それは既にインストールと構成されています。

PGD クラスターの最初のホストにログインします。

```
docker compose exec host-1 bash
```

そして、 PGD CLI のバージョンを確認します。

```
pgd --version
__OUTPUT__
pgd-cli version 製品バージョン 6.4.0
```

!!!note

docker compose exec コマンドを使用して、PGD クラスターの最初のホストのコンテキストで実行することにより、コンテナの外部から次のコマンドを実行することもできます。

```
docker compose exec host-1 pgd <command>
```

また、それらはすべて PGD CLI がインストールおよび構成されているため、クラスター内の任意のホストから pgd コマンドを実行できます。

pgd cluster show コマンドを使用して、クラスターの全体的なステータスを表示することから始めます。

```
pgd cluster show
__OUTPUT__

#  Summary

 Group Name | Parent Group | Group Type | Node Name | Node Kind
- ----------+--------------+------------+-----------+-----------
 group-1    | pgd          | data       | node-1    | data
 group-1    | pgd          | data       | node-2    | data
 group-1    | pgd          | data       | node-3    | data
 pgd        |              | global     |           |

#  Health

 Check              | Status | Details
- ------------------+--------+-------------------------------------------------
 Connections        | Ok     | All BDR nodes are accessible
 Raft               | Ok     | Raft Consensus is working correctly
 Replication Slots | Ok     | All PGD replication slots are working correctly
 Clock Skew         | Ok     | Clock drift is within permissible limit
 Versions           | Ok     | All nodes are running the same PGD version

#  Clock Drift

 Reference Node | Node Name | Clock Drift
- ---------------+-----------+-------------
 node-3         | node-2    | *
 node-3         | node-1    | *
```

このコマンドは、クラスター、そのノード、およびそれらの正常性ステータスの概要を提供します。また、ノード間のクロックドリフトも示します。これは、レプリケーションの一貫性にとって重要です。

pgd node show コマンドを使用して、個々のノードのステータスを表示することもできます。

```
pgd node node-1 show
__OUTPUT__
```

(次のページに続く)

(前のページからの続き)

```
#  Summary

 Node Property   | Value
- ----------------+------------
 Node Name       | node-1
 Group Name      | group-1
 Node Kind       | data
 Join State      | ACTIVE
 Node Status     | Up
 Node ID         | 4153941939
 Snowflake SeqID | 1
 Database        | pgddb

#  Options

 Option Name    | Option Value
- ---------------+-------------------------------------------------
 route_dsn      | port=5432 dbname=pgddb host=host-1 user=postgres
 route_fence    | false
 route_priority | -1
 route_reads    | true
 route_writes   | true
```

pgd CLI コマンドの構造は階層的で、コマンドは機能ごとにグループ化されています。次を実行して、使用可能なコマンドとその説明を表示できます。

```
pgd --help
__OUTPUT__
Manages PGD clusters

Usage: pgd [OPTIONS] <COMMAND>

Commands:
  cluster       Cluster-level commands
  group         Group related commands
  groups        Groups listing commands
  node          Node related commands
  nodes         Nodes listing commands
  events        Event log commands
  replication   Replication related commands
  raft          Raft related commands
  commit-scope  Commit scope management commands
  assess        PGD compatibility assessment of Postgres server
  completion    Generate the autocompletion script for pgd for the specified shell
```

(次のページに続く)

(前のページからの続き)

```
Options:
  -V, --version  Print version

Global Options:
  -f, --config-file <CONFIG_FILE>  Sets the configuration file path
      --dsn <DSN>                  Sets the PostgreSQL connection string e.g.
↪"host=localhost port=6000 user=postgres dbname=postgres" [env: PGD_CLI_DSN=]
  -o, --output <OUTPUT_FORMAT>     Sets the output format for tables [env: PGD_CLI_
↪OUTPUT=] [default: psql] [possible values: json, psql, modern, markdown, simple]
      --debug                      Print debug messages, useful while troubleshooting␣
↪[env: PGD_CLI_DEBUG=]
  -h, --help                       Print help
```

group 、node などのコマンドは、次の引数としてグループまたはノード名を取り、特定のコマンドが続きます。cluster 、groups 、nodes などのコマンドは、クラスターレベルで動作するか、すべてのグループまたはノードをリストするため、グループ名またはノード名を必要としません。

次のコマンドを実行して、特定のコマンドのヘルプを取得することもできます。

```
pgd <COMMAND> --help
```

### クラスターステータスの表示

PGD クラスターの全体的なステータスを表示するために、既に pgd cluster show コマンドを使用しています。これは、すべてのクラスター情報を表示します。クラスターの正常性ステータスのみを表示するには、--health オプションを使用できます。

```
pgd cluster show --health
__OUTPUT__
 Check             | Status | Details
- ------------------+--------+-------------------------------------------------
 Connections       | Ok     | All BDR nodes are accessible
 Raft              | Ok     | Raft Consensus is working correctly
 Replication Slots | Ok     | All PGD replication slots are working correctly
 Clock Skew        | Ok     | Clock drift is within permissible limit
 Versions          | Ok     | All nodes are running the same PGD version
```

または、概要ステータスのみを表示する場合は、 --summary オプションを使用できます。

```
pgd cluster show --summary
__OUTPUT__
 Group Name | Parent Group | Group Type | Node Name | Node Kind
- -----------+--------------+------------+-----------+-----------
 group-1    | pgd          | data       | node-1    | data
 group-1    | pgd          | data       | node-2    | data
```

(次のページに続く)

(前のページからの続き)

```
 group-1    | pgd          | data       | node-3    | data
 pgd        |              | global     |           |
```

### グループとグループステータスの表示

クラスター内のすべてのグループのステータスを表示するには、pgd groups list コマンドを使用できます。

```
pgd groups list
__OUTPUT__
 Group Name | Parent Group Name | Group Type | Nodes
- -----------+-------------------+------------+-------
 group-1    | pgd               | data       | 3
 pgd        |                   | global     | 0
```

これで、トップレベルグループ pgd と、3 つのノードを含むデータグループ group-1 が表示されます。すべてのノードは、クラスター全体のすべてのアクティビティを調整するトップレベルグループのメンバーです。データグループ group-1 は、自分自身間でデータを複製し、グループ内の着信クエリーをグループ内の書き込みリーダーノードにルーティングする 3 つのデータノードのグループです。

pgd group show コマンドを使用して、グループの詳細を深く掘り下げることができます。

```
pgd group group-1 show
__OUTPUT__

#  Summary

 Group Property    | Value
- ------------------+---------
 Group Name        | group-1
 Parent Group Name | pgd
 Group Type        | data
 Write Leader      | node-1
 Commit Scope      |

#  Nodes

 Node Name | Node Kind | Join State | Node Status
- ----------+-----------+------------+-------------
 node-1    | data      | ACTIVE     | Up
 node-2    | data      | ACTIVE     | Up
 node-3    | data      | ACTIVE     | Up

#  Options
```

(次のページに続く)

(前のページからの続き)

```
 Option Name                          | Option Value
- -------------------------------------+----------------------
 analytics_storage_location           |  (inherited)
 apply_delay                          | 00:00:00 (inherited)
 check_constraints                    | true (inherited)
 default_commit_scope                 |  (inherited)
 enable_raft                          | true
 enable_routing                       | true
 enable_wal_decoder                   | false (inherited)
 http_port                            |  (inherited)
 location                             |
 num_writers                          | -1 (inherited)
 read_only_consensus_timeout          |  (inherited)
 read_only_max_client_connections     |  (inherited)
 read_only_max_server_connections     |  (inherited)
 read_only_port                       |  (inherited)
 read_write_consensus_timeout         |  (inherited)
 read_write_max_client_connections    |  (inherited)
 read_write_max_server_connections    |  (inherited)
 read_write_port                      |  (inherited)
 route_reader_max_lag                 | -1
 route_writer_max_lag                 | -1
 route_writer_wait_flush              | false
 streaming_mode                       | default (inherited)
 use_https                            | true
```

このコマンドは、グループ、そのノード、およびそれらのステータスの概要を提供します。また、ルーティングが有効になっているかどうか、監視用の HTTP ポート、その他の構成設定などのグループオプションも表示します。

cluster コマンドと同様に、 `--summary` オプションを使用して、グループの概要のみを表示することもできます。

```
pgd group group-1 show --summary
__OUTPUT__
 Group Property    | Value
- ------------------+---------
 Group Name        | group-1
 Parent Group Name | pgd
 Group Type        | data
 Write Leader      | node-1
 Commit Scope      |
```

これで、グループがトップレベルグループ `pgd` の子であり、データグループであり、グループの書き込みリーダーノードが `node-1` であることがわかります。このグループにはコミットスコープが設定されていません。これは、デフォルトのコミットスコープを使用していることを意味します。

--nodes オプションは、グループ内のノードを表示するために使用できます。

```
pgd group group-1 show --nodes
__OUTPUT__
 Node Name | Node Kind | Join State | Node Status
- ----------+-----------+------------+-------------
 node-1    | data      | ACTIVE     | Up
 node-2    | data      | ACTIVE     | Up
 node-3    | data      | ACTIVE     | Up
```

また、同様に、 --options オプションを使用してグループオプションを表示できます。

```
pgd group group-1 show --options
__OUTPUT__
 Option Name                        | Option Value
- -----------------------------------+---------------------
 analytics_storage_location         |  (inherited)
 apply_delay                        | 00:00:00 (inherited)
 check_constraints                  | true (inherited)
 default_commit_scope               |  (inherited)
 enable_raft                        | true
 enable_routing                     | true
 enable_wal_decoder                 | false (inherited)
 http_port                          |  (inherited)
 location                           |
 num_writers                        | -1 (inherited)
 read_only_consensus_timeout        |  (inherited)
 read_only_max_client_connections   |  (inherited)
 read_only_max_server_connections   |  (inherited)
 read_only_port                     |  (inherited)
 read_write_consensus_timeout       |  (inherited)
 read_write_max_client_connections  |  (inherited)
 read_write_max_server_connections  |  (inherited)
 read_write_port                    |  (inherited)
 route_reader_max_lag               | -1
 route_writer_max_lag               | -1
 route_writer_wait_flush            | false
 streaming_mode                     | default (inherited)
 use_https                          | true
```

ご覧のとおり、オプションの多くは親グループであるトップレベルグループ pgd から継承されます。enable_raft および enable_routing オプションは true に設定されています。これは、グループがレプリケーションと書き込みリーダーノードへのルーティングクエリに Raft コンセンサスを使用していることを意味します。接続マネージャーポートを介して作成されます。

pgd group pgd show コマンドを使用して、親グループ pgd を見てみましょう。

```
pgd group pgd show
__OUTPUT__

#  Summary

 Group Property    | Value
- ------------------+--------
 Group Name        | pgd
 Parent Group Name |
 Group Type        | global
 Write Leader      |
 Commit Scope      |
```

これは、トップレベルグループ pgd がグローバルグループであることを示しています。これは、データグループではなく、独自のデータノードを持たないことを意味します。この場合、クラスター内のデータグループのアクティビティを調整するためにのみ使用されます。データノードがないため、書き込みリーダーはありません。

出力の次の部分には、グループ内のノードが表示されますが、空です。

```
#  Nodes

 Node Name | Node Kind | Join State | Node Status
- ----------+-----------+------------+-------------
```

pgd グループのオプションを次に示します。

```
#  Options

 Option Name                        | Option Value
- -----------------------------------+--------------
 analytics_storage_location         |
 apply_delay                        | 00:00:00
 check_constraints                  | true
 default_commit_scope               |
 enable_raft                        | true
 enable_routing                     | false
 enable_wal_decoder                 | false
 http_port                          |
 location                           |
 num_writers                        | -1
 read_only_consensus_timeout        |
 read_only_max_client_connections   |
 read_only_max_server_connections   |
 read_only_port                     |
 read_write_consensus_timeout       |
```

(次のページに続く)

(前のページからの続き)

```
 read_write_max_client_connections |
 read_write_max_server_connections |
 read_write_port                   |
 route_reader_max_lag              | -1
 route_writer_max_lag              | -1
 route_writer_wait_flush           | false
 streaming_mode                    | default
 use_https                         | true
```

これらは、トップレベルグループ pgd のオプションです。これは、group-1 がオプションを継承する場所です。ただし、ここでは enable_routing オプションは false に設定されています。これは、トップレベルグループが独自のデータノードを持たないため、クエリーをデータノードにルーティングしないことを意味します。enable_raft オプションは true に設定されています。これは、トップレベルグループが Raft コンセンサスを使用してクラスターの管理を調整することを意味します。

オプションが設定されていない場合、デフォルト値が使用されます。たとえば、00:00:00 に設定される apply_delay オプション、つまりクラスターへの変更の適用に遅延がないことを意味します。

### ノードとノードのステータスの表示

クラスター内のすべてのノードのステータスを表示するには、pgd nodes list コマンドを使用できます。

```
pgd nodes list
__OUTPUT__
 Node Name | Group Name | Node Kind | Join State | Node Status
- -----------+------------+-----------+------------+-------------
 node-1    | group-1    | data      | ACTIVE     | Up
 node-2    | group-1    | data      | ACTIVE     | Up
 node-3    | group-1    | data      | ACTIVE     | Up
```

pgd node show コマンドを使用して、特定のノードのステータスを表示することもできます。

```
pgd node node-1 show
__OUTPUT__

#  Summary

 Node Property   | Value
- ----------------+------------
 Node Name       | node-1
 Group Name      | group-1
 Node Kind       | data
 Join State      | ACTIVE
 Node Status     | Up
 Node ID         | 4153941939
 Snowflake SeqID | 1
```

(次のページに続く)

(前のページからの続き)

```
 Database        | pgddb

#  Options

 Option Name    | Option Value
- ---------------+-------------------------------------------------
 route_dsn      | port=5432 dbname=pgddb host=host-1 user=postgres
 route_fence    | false
 route_priority | -1
 route_reads    | true
 route_writes   | true
```

ここで、ノード自分自身についての詳細を確認できます。ノードの名前とそれが属しているグループ、それがデータノードであること、グループにアクティブに参加していること、および実行中であることがわかります。ノード ID はノードの一意の識別子であり、Snowflake SeqID は、クラスター内のイベントを順序付けするために使用されます。最後に、そのデータベースが pgddb 、Quickstart Docker Compose キットで作成されたデフォルトデータベースであることがわかります。

ノードのオプションを次に示します。これらはこの特定のノードに固有です。

- route_dsn は、ノードの接続文字列であり、接続マネージャーがクエリーをこのノードにルーティングするために使用されます。
- route_fence は false に設定されています。これは、ノードにクエリのルーティングを防止するフェンス設定がないことを意味します。
- route_priority は-1 に設定されています。これは、ノードがルーティングクエリーに特定の優先度を持たないことを意味します。
- route_reads および route_writes は両方とも true に設定されています。これは、ノードが読み取りクエリと書き込みクエリの両方を処理できることを意味します。

これらは、接続マネージャーがクエリーをノードにルーティングするときに使用します。これらは、ダウンせずにアクティブなノードを制御する方法でもあります。route_fence を true に設定すると、接続マネージャーがクエリをこのノードにルーティングできなくなりますが、クラスターの一部でデータをレプリケートすることはできます。

### ノードオプションの設定

pgd node set コマンドを使用してノードのオプションを設定できます。たとえば、node-1 の route_fence オプションを true に設定するには、次のコマンドを実行できます。

```
pgd node node-1 set-option route_fence true
```

node-1 の接続マネージャーに接続してみると

```
psql -h host-1 -p 6432
```

接続が得られます。ただし、ルーティングクエリーからフェンスされているため、node-1 ノードにルーティングされません。代わりに、グループ内の現在の書き込みリーダー node-2 にルーティングされます。

```
select node_name from bdr.local_node_summary;
 node_name
 -----------
 node-2
(1 row)
```

次を実行してフェンシングを終了して元に戻すと

```
pgd node node-1 set-option route_fence false
```

node-1 ノードの接続マネージャーに再度接続できるようになりました。

```
psql -h host-1 -p 6432
```

そして、node-1 ノードに接続されたことがわかります。

```
select node_name from bdr.local_node_summary;
 node_name
 -----------
 node-1
(1 row)
```

# 第4章　Planning your PGD deployment

EDB Postgres 分散 PGD 展開の計画には、アプリケーションの要件と PGD の機能を理解することが含まれます。このセクションでは、PGD 展開を計画するための重要な考慮事項の概要を提供します。

- Choosing your deployment pattern: PGD がサポートする展開パターンを理解し、要件に最適なものを選択します。
- *Choosing a Postgres distribution* PGD で展開する Postgres ディストリビューションを選択します。
- *Choosing your deployment method* ニーズに合った展開方法を選択します。
- *Comparing replication solutions* EDB Postgres 分散と他のレプリケーションソリューションとの比較。
- *Choosing between Physical Streaming Replication (PSR) and EDB Postgres Distributed (PGD)* 標準の Postgres ストリーミングレプリケーションと PGD 間のトレードオフを評価して、ワークロードに適切な高可用性ソリューションを決定します。
- Other considerations: 展開に影響を与える可能性のある他の要素を検討します。

## 4.1 Deployment patterns

PGD は、単一の高可用 HA グループから世界的に分散されたマルチリージョンクラスターに至るまで、いくつかの展開パターンを提供します。選択したパターンにより、データが存在する場所、書き込みを受け入れる場所、およびクラスターが障害から復旧する方法が決まります。重要なディメンションは、アクティブな書き込み場所の数、必要なディザスターリカバリーのレベル、および規制要件がデータの保存場所を制限するかどうかです。各選択は、書き込みレイテンシー、競合処理、および操作の複雑さに影響を与えます。

### 4.1.1 パターンの選択

ほとんどの PGD 展開は、それぞれが 1 つ以上のパターンにマッピングする 4 つの目標のいずれかから開始します。

#### 継続的な可用性の維持

PGD は、ノード障害、データセンター DC 障害、およびメジャー Postgres バージョンのアップグレードを含むローリングメンテナンス操作を通じてデータベースの利用可能な状態を維持します。*Single data group, global read scaling* パターンは、単一のロケーション内で、2 つ以上のノードにわたって HA を提供します。完全なセカンドアクティブサイトのコストなしでデータセンターレベルの保護を行う場合、*Primary active group, DR group* パターンは、プライマリ DC に障害が発生した場合に引き継ぐ容量の削減された DR サイトを追加します。

### 複数のリージョンにまたがる書き込み

複数の地域のユーザーにサービスを提供するアプリケーションは、それらのユーザーの近くに書き込み容量を配置すると利点があります。*Two data groups, active-active* および *Three data groups, active-active-active* パターンは両方ともアクティブ/アクティブで実行されます。すべての場所が書き込みを受け入れ、相互にレプリケートします。2 つのデータグループパターンは、プライマリおよびセカンダリの場所に適しています。監視専用の 3 番目のロケーションは、完全な 3 番目のデータサイトなしで Raft マジョリティを達成し、3 つの完全なデータグループは、残りのリージョンを中断せずにリージョン全体の障害を乗り切ることにより、Raft マジョリティをさらに取得します。

リージョン間のアクティブ-アクティブレプリケーションにより、書き込み競合が発生する可能性があります。PGD は、構成可能な競合解決ポリシーを使用して競合を自動的に解決しますが、スキーマとアプリケーションの書き込みは、競合を念頭に置いて設計する必要があります。

### データの局所性の適用

GDPR や CCPA などの規制では、個人データや機密データを特定の管轄区域内に留めることが必要です。*Multiple locations, data residency* パターンは、データ分類ごとに選択的レプリケーションを実装します。個人識別情報 PII、財務記録、医療データを含むローカル限定データは、リージョン内でのみ複製されます。製品カタログや構成などのグローバル参照データは、どこでも複製されます。各リージョンは、独自の HA グループで独立して実行されます。

### 読み取りのスケールアウト

読み取りの多いワークロードは、分離されていない場合、書き込みクラスターを過負荷にする可能性があります。*Single data group, global read scaling* パターンは、分析、レポート、および API 読み取りトラフィックをサブスクライバー専用ノードにオフロードします。サブスクライバー専用ノードは、書き込みグループからすべての変更を受信しますが、書き込みまたはコンセンサスに参加しないため、それらを追加しても書き込みパフォーマンスまたはコミットスコープのクォーラムに影響はありません。これらを追加リージョンに配置して、ローカルユーザーの読み取り遅延を削減できます。

| Pattern | Locations | Active write locations | Disaster recovery |
|---|---|---|---|
| *Single data group, global read scaling* | 1 | 1 | Node failure only |
| *Two data groups, active-active* | 2 | 2 | DC failure |
| *Three data groups, active-active-active* | 3+ | 3+ | Region failure |
| *Single data group, global read scaling* | 1+ | 1 | Node failure |
| *Primary active group, DR group* | 2 | 1 | DC failure (manual) |
| *Multiple locations, data residency* | 3+ | 3+ | Region failure |

### 4.1.2 アーキテクチャ要素

PGD クラスター内の各ノードは、PGD 拡張機能を実行している Postgres インスタンスです。ノードはグループに編成され、各グループは書き込みリーダーを選択して書き込みを調整します。接続マネージャーは各ノードで実行され、クライアント接続を現在の書き込みリーダーにルーティングし、アプリケーションの接続文字列を変更せずに自動フェイルオーバーを有効にします。

#### ノードタイプ

PGD クラスターは、それぞれが異なるロールを担う 3 つのノードタイプを使用します。

- **データノード**データを保存および管理し、読み取りおよび書き込みを処理し、レプリケーションとコンセンサスに参加します。
- **サブスクライバー専用ノード**書き込みグループからすべての変更を受信しますが、書き込みを受け入れず、コンセンサスに参加しません。これらは、分析やレポートなどの読み取り中心のワークロードに適しています。
- **Witness ノード** Raft コンセンサスに参加しますが、ユーザデータは保存しません。3 番目の場所の監視は、2 つのデータの場所が相互に接続を失った場合に、スプリットブレインシナリオを解決します。

#### ノードのロール

データノードは、条件の変更に応じてノード間で自動的に転送される一時的なロールを引き受けます。

- **書き込みリーダー**アプリケーションが接続マネージャーを介して接続するときのすべての書き込み操作の現在のターゲット。書き込みリーダーに障害が発生すると、別のノードが数秒で選択されます。
- **Raft リーダー**書き込みリーダー選挙やスキーマ変更の調整を含む、グループ全体のコンセンサス決定を管理します。

#### コミットスコープ

コミットスコープは、PGD が各トランザクションに提供する耐久性保証を制御します。デフォルトは、結果的一貫性を備えた非同期レプリケーションです。マジョリティプロテクトからクォーラムコミットまでのより強力なオプションは、書き込みレイテンシーの追加を犠牲にして同期調整を追加します。クォーラムコミットは、最も厳密な一貫性保証を提供し、ノードがローカルにコミットする前に、すべての参加ノードにわたってコミット決定を調整します。各パターンページには、そのトポロジに推奨されるコミットスコープが記載されています。完全なリファレンスについては、*Commit scopes* を参照してください。

#### Single data group

単一のデータグループは基本的な PGD 展開パターンであり、単一の場所内の高可用性を提供します。3 つのデータノードは Raft コンセンサスグループを形成し、Connection Manager が書き込みリーダールーティングを自動的に処理します。

ノードを複数のアベイラビリティーゾーンに分散して、復元力を高めることができますが、そうすると、マジョリティまたはオール保護のコミットスコープを使用する場合、書き込みレイテンシーが増加します。

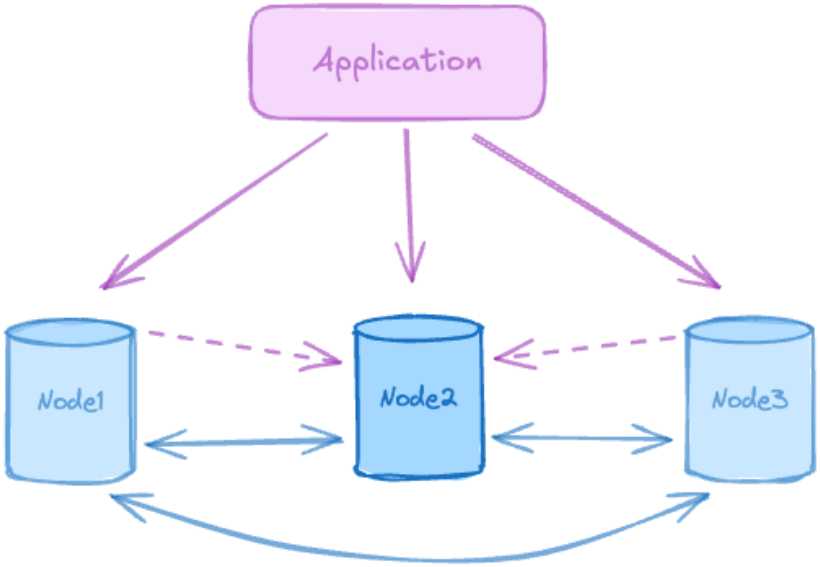


図 1: Single data group topology

### 図の凡例

シンボルは、ノードタイプ、グループコンテナ、および接続タイプを表します。

### このパターンを使用する場合

単一のデータグループは、単一のデータセンター内の標準の高可用性 HA 展開です。単一ノードの障害を許容し、数秒で自動フェイルオーバーを提供するため、単一リージョン内で HA を必要とし、メジャーバージョンのアップグレードまたはメンテナンスタスクのブロック中にダウンタイムを延長する余裕がないアプリケーションに最適です。

このパターンには次のような利点があります。

- 導入、操作、およびトラブルシューティングが最も簡単なトポロジ。
- Raft マジョリティは常に利用可能で、1 つのノードに障害が発生しても生き残ります。
- 単一のデータセンター DC に展開された場合、最小の書き込みレイテンシー。
- 複数のアベイラビリティーゾーンに展開された場合、単一の AZ 障害に耐えます。
- オペレーターの介入なしで数秒の自動フェイルオーバー。
- 最小 3 ノードでの低コスト。

注意すべき制限がいくつかあります。

- ディザスターリカバリーなし。DC レベルの障害は、クラスター全体をダウンさせます。
- 単一の DC に展開された場合、すべてのノードは同じ障害ドメインを共有します。
- 地理的分離を義務付ける規制要件には適していません。
- 2 つのデータノードと監視で使用する場合、耐久性のオプションは制限されます。

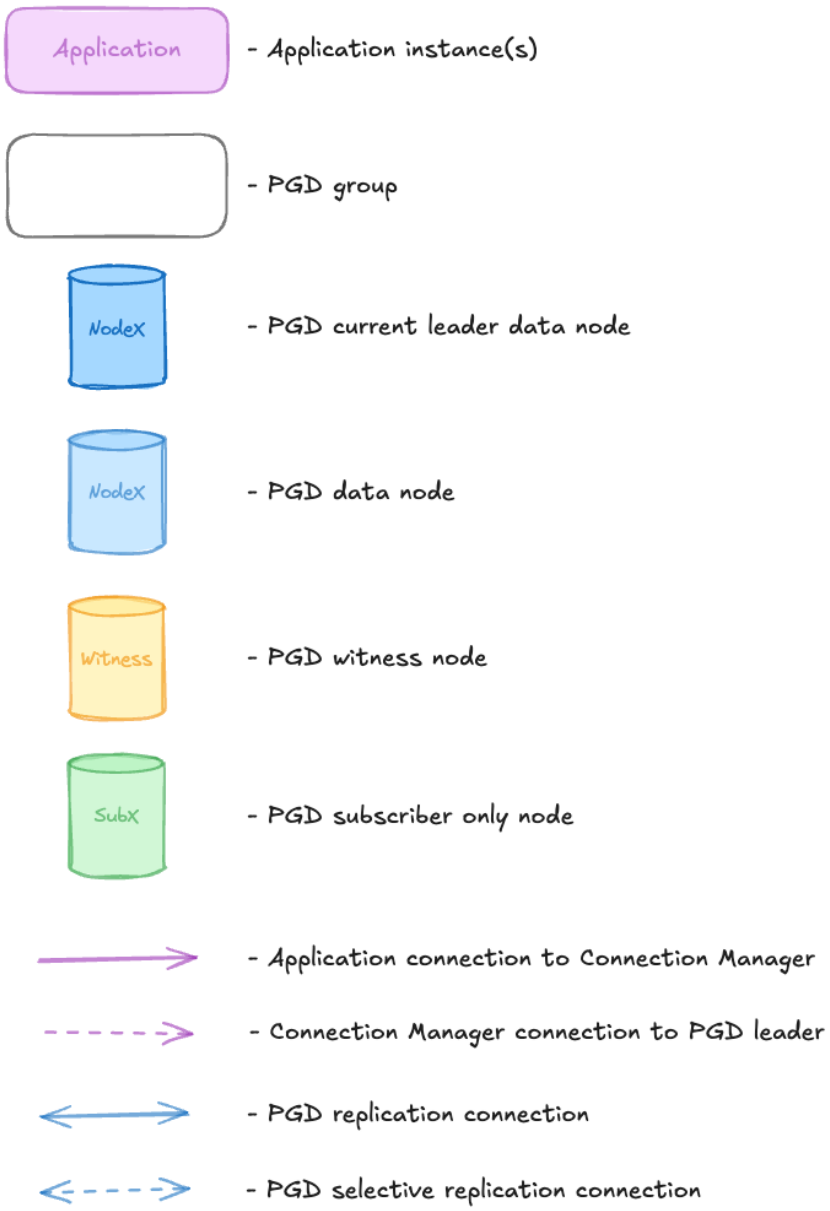


図 2: Diagram legend showing node types, group containers, and connection types

### このパターンに推奨されるコミットスコープ

- **マジョリティプロテクト**マジョリティノードへの同期複製。実稼働使用に最も一般的に推奨されるスコープ。
- **クォーラムコミット**ノードがローカルにコミットする前に、すべての参加ノードにわたってコミット決定を調整します。決済やコアバンキングなど、厳密な分散一貫性を必要とするワークロードに推奨されます。Witness ノードはクォーラムに寄与しないため、3 つのノードすべてがデータノードである必要があります。`MAJORITY ORIGIN GROUP QUORUM COMMIT ABORT ON (timeout = 6s)` を使用します。
- **アダプティブプロテクト**利用可能な場合は大部分のノードへの同期レプリケーション、構成可能なタイムアウトの後、非同期レプリケーションに自動的に低下します。2 つのデータノードと 1 つの監視を使用する場合に推奨されます。
- **オールプロテクト**グループ内のすべてのノードへの同期レプリケーション。最高の耐久性保証を提供しますが、ノードに障害が発生すると、書き込みレイテンシーが増加し、可用性トレランスが低下します。
- **ローカルプロテクト**ローカルノードのみの耐久性を備えた非同期レプリケーション。実稼働環境での使用はお勧めしません。

コミットスコープは、パフォーマンス、耐久性、一貫性、スケーラビリティにわたってさまざまなトレードオフを行います。

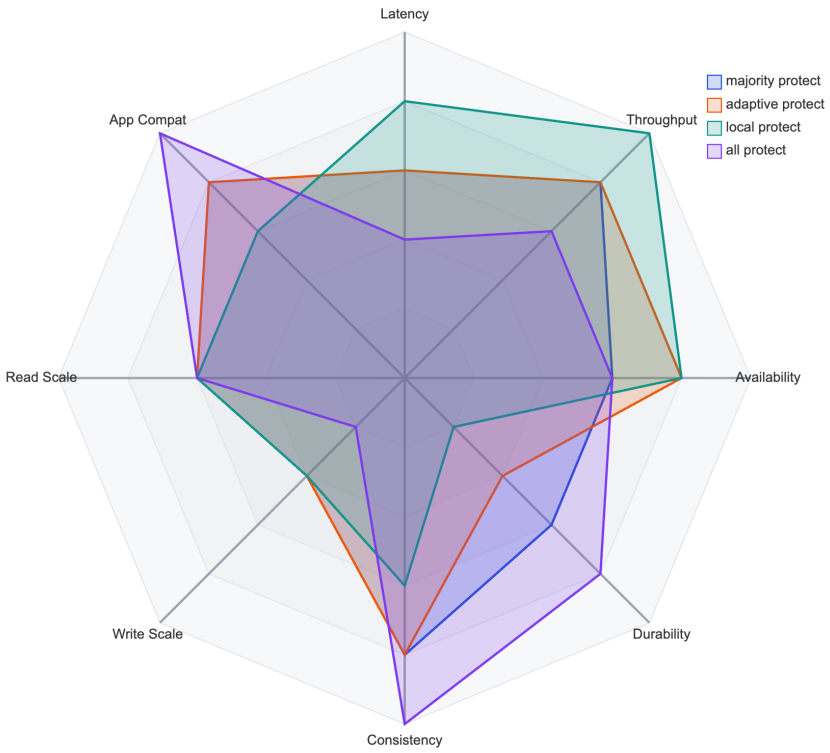


図 3: Single data group behavior diagram

### Two data groups, active-active

2 つのデータグループパターンは、地理的に離れた 2 つのデータセンターにまたがり、それぞれがローカルの高可用性 HA のために独自の独立した Raft コンセンサスグループを実行しています。グループごとに 3 つ以上のデータノードがある場合、各グループは自分のサイト内でマジョリティベースのコミットスコープを使用できます。

両方のデータセンターは、ローカルの読み取りと書き込みを同時に提供し、グループ間のレプリケーションにより、両方の場所を最新の状態に保ちます。

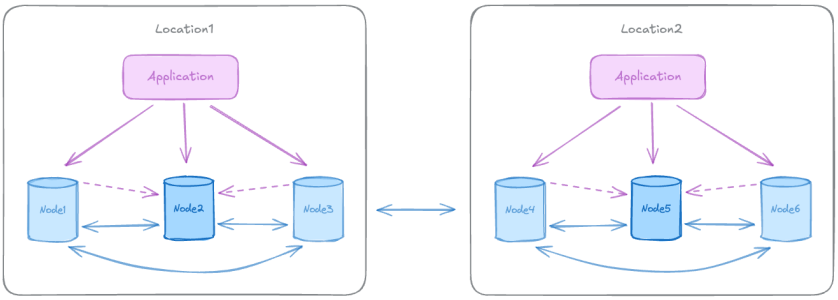


図 4: Two data groups topology

### 図の凡例

シンボルは、ノードタイプ、グループコンテナ、および接続タイプを表します。

### このパターンを使用する場合

このパターンは、金融サービス、医療、および 2 つのアクティブな場所にまたがるデータ損失ゼロのジオレプリケーションを必要とするアプリケーションに適しています。どちらのデータセンターも、競合の解決と同時に書き込みを提供し、場所を超えて真のアクティブ-アクティブオペレーションを提供します。

このパターンには次のような利点があります。

- 真のディザスターリカバリーを提供し、完全なデータセンター DC の障害を乗り切ります。
- アクティブ-アクティブで実行され、両方の DC がローカルレイテンシーの読み取りおよび書き込みを処理します。
- コンプライアンス要件のための地理的冗長性。
- ローカルルーティングでは、各 DC はほとんどの操作で独立して動作します。

注意すべき制限がいくつかあります。

- オプティミスティックなクロスリージョンの一貫性処理では、競合は避けられません。
- DC 障害時の書き込みレイテンシーとゼロデータ損失は緊張しています。コミットスコープが強くなると、レイテンシーが増加します。
- Raft マジョリティ DC 損失がないと、グローバル割り当てシーケンス galloc は、最終的に割り当てられたレンジとブロックを使い果たします。
- グローバルルーティングは使用できません。1 つの DC が失われると、クラスター全体の Raft マジョリティが壊れます。

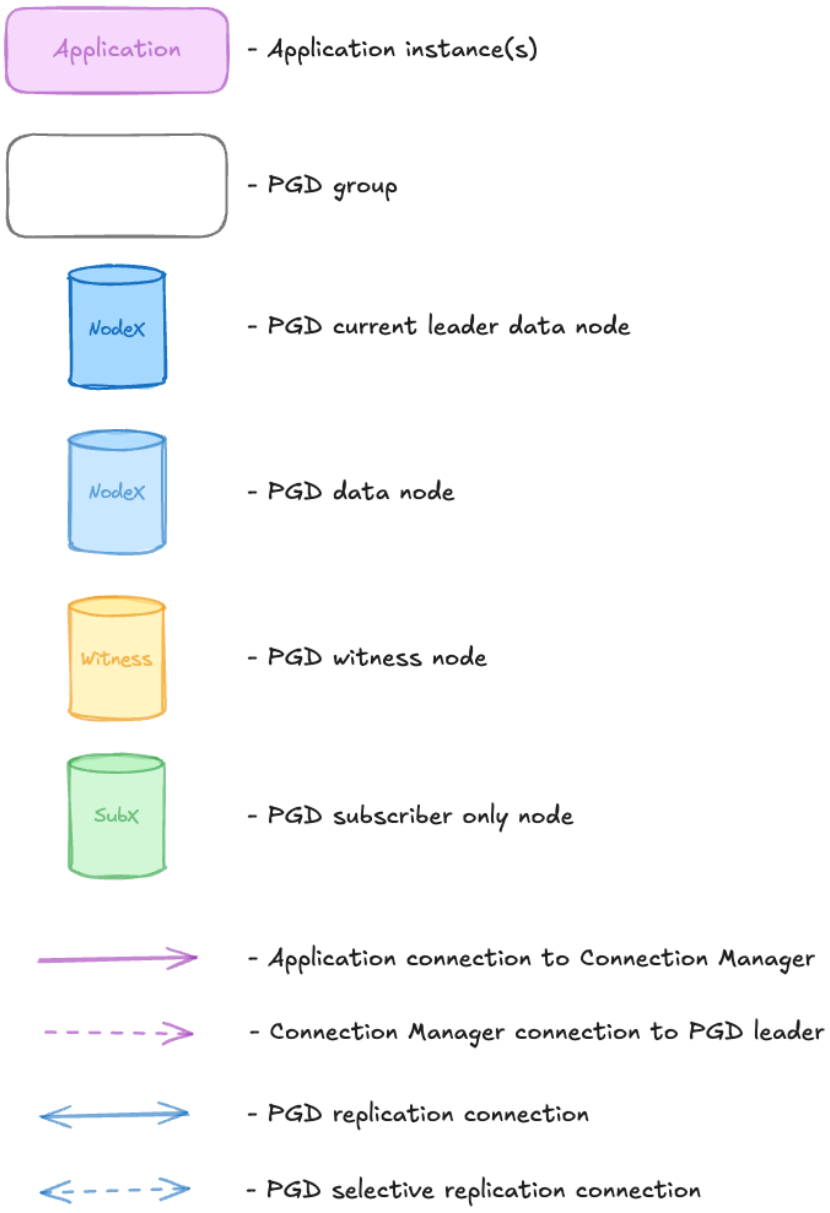


図 5: Diagram legend showing node types, group containers, and connection types

Raft マジョリティと galloc の制限は、以下で説明するように、監視の場所を追加することにより対処できます。

### 監視の場所の追加

単一の監視ノードを含む 3 番目の場所を追加することにより、Raft マジョリティと galloc の制限が解決されます。Witness は Raft コンセンサスに参加しますが、ユーザーデータを保存しません。奇数のクォーラム 7 つの投票者 6 つのデータノードと 1 つの監視により、クラスターは 1 つの DC が失われた場合でも Raft マジョリティを維持し、グローバルルーティングを復元し、galloc シーケンスが割り当てられたレンジを使い果たさないようにします。

その構成は自分自身でローカル HA を実現できないため、グループごとの監視なしでグループごとに 2 つのデータノードのみを使用する場合にも、監視の場所が必要です。

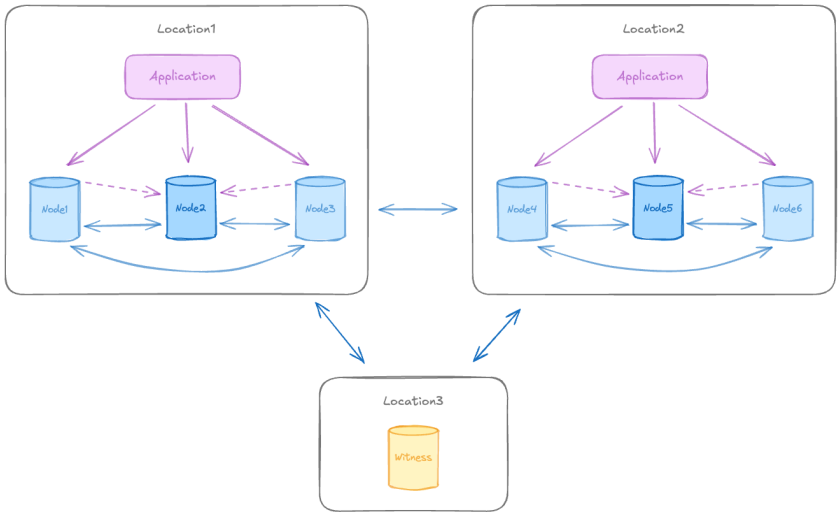


図 6: Two data groups with witness topology

### このパターンに推奨されるコミットスコープ

- **マジョリティプロテクト**グループ内のマジョリティノードへの同期複製。実稼働使用に最も一般的に推奨されるスコープ。
- **クォーラムコミット**ノードがローカルにコミットする前に、すべての参加ノードにわたってコミット決定を調整し、コミットする前に両方のリージョンにわたってトランザクションが持続することを保証します。決済やコアバンキングなど、厳密な分散一貫性を必要とするワークロードに推奨されます。`MAJORITY FOR EACH GROUP IN CLUSTER QUORUM COMMIT ABORT ON (timeout = 6s)` を使用します。これは、新しい書き込みリーダーが、以前のリーダーによって確認された最新のコミットデータを持つことを保証します。
- **CAMO** グループごとに 2 つのデータノードと 1 つの監視を使用する場合、CAMO は同期レプリケーションを介した厳密に 1 回のトランザクション保証を追加します。
- **ローカルプロテクト**ローカルノードのみの耐久性を備えた非同期レプリケーション。実稼働環境での使用はお勧めしません。

コミットスコープは、パフォーマンス、耐久性、一貫性、スケーラビリティにわたってさまざまなトレードオフを行います。

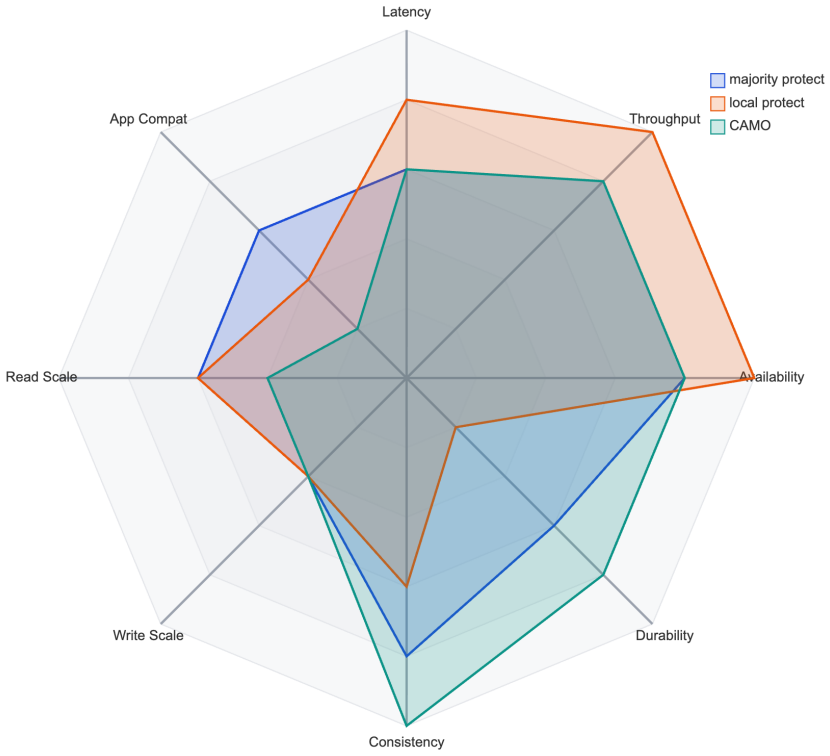


図 7: Two data groups behavior diagram

### Three data groups, active-active-active

3 つのデータグループパターンは 3 つ以上の地理的リージョンにまたがり、それぞれがローカルの高可用性 HA のために独自の独立した Raft コンセンサスグループを実行しています。グループごとに 3 つ以上のデータノードがある場合、各グループは、接続マネージャーが書き込みリーダールーティングをローカルに処理し、独自のリージョン内でマジョリティベースのコミットスコープを使用できます。

アプリケーションは、低遅延の読み取りおよび書き込みのためにローカルリージョンに接続します。リージョン間のレプリケーションは、通常、リージョン間のレイテンシーペナルティを回避するために非同期です。

### 図の凡例

シンボルは、ノードタイプ、グループコンテナ、および接続タイプを表します。

### このパターンを使用する場合

このパターンは、グローバル SaaS アプリケーション、CDN のようなデータ分散、および 3 つ以上のリージョンにわたるデータの局所性と最終的一貫性を備えたジオレプリケーションを必要とする多国籍企業に適しています。

このパターンには次のような利点があります。

- すべてのリージョンでの低レイテンシーの読み取りおよび書き込みにより、真のグローバルプレゼンスを有効にします。
- 他のリージョンが動作を継続している間、リージョン全体の障害は生き残ります。

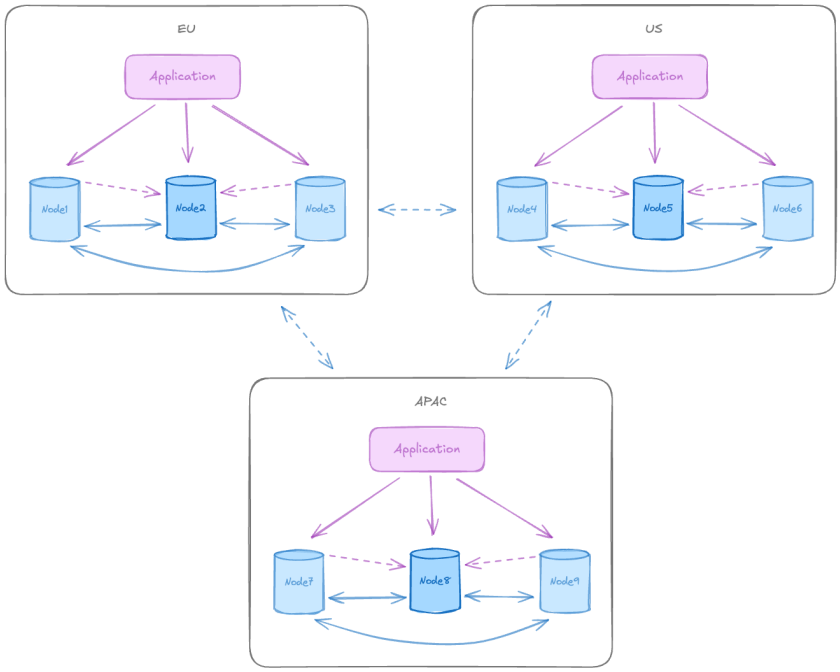


図 8: Three data groups topology

- 3 つ以上のリージョンを持つ Natural Raft マジョリティ。
- グローバルに分散されたアプリケーションのための最高のエンドユーザーエクスペリエンス。

注意すべき制限がいくつかあります。

- オプティミスティックなクロスリージョンの一貫性処理では、競合は避けられません。
- より多くのノード、より多くの場所、より多くの監視による運用の複雑性の向上。
- リージョン間のレイテンシーは、レプリケーションラグと競合検出ウィンドウに影響を与えます。

### グローバルルーティングの追加

3 つ以上のデータグループは、クラスター全体がグループごとに 1 つではなく単一の書き込みリーダーを持つグローバルルーティングで使用することもできます。単一のグローバル書き込みリーダーは、すべての書き込みを 1 つの場所を介してルーティングすることにより、書き込み競合を排除しますが、書き込みリーダーをホストしていないリージョンの書き込みレイテンシーが増加します。書き込みリーダーの場所が変更されると、すべてのリージョンのレイテンシー特性がシフトします。

### このパターンに推奨されるコミットスコープ

- **マジョリティプロテクト**グループ内のマジョリティノードへの同期複製。実稼働使用に最も一般的に推奨されるスコープ。
- **クォーラムコミット**ノードがローカルにコミットする前に、すべての参加ノードにわたってコミット決定を調整し、コミットする前にすべてのリージョンでトランザクションが持続することを保証します。決済やコアバンキングなど、厳密な分散一貫性を必要とするワークロードに推奨されます。`MAJORITY FOR EACH GROUP IN CLUSTER QUORUM COMMIT ABORT ON (timeout = 6s)` を使用します。これは、リージョンに関係なく、新しい書き込みリーダーに、以前のリーダーによって確認され

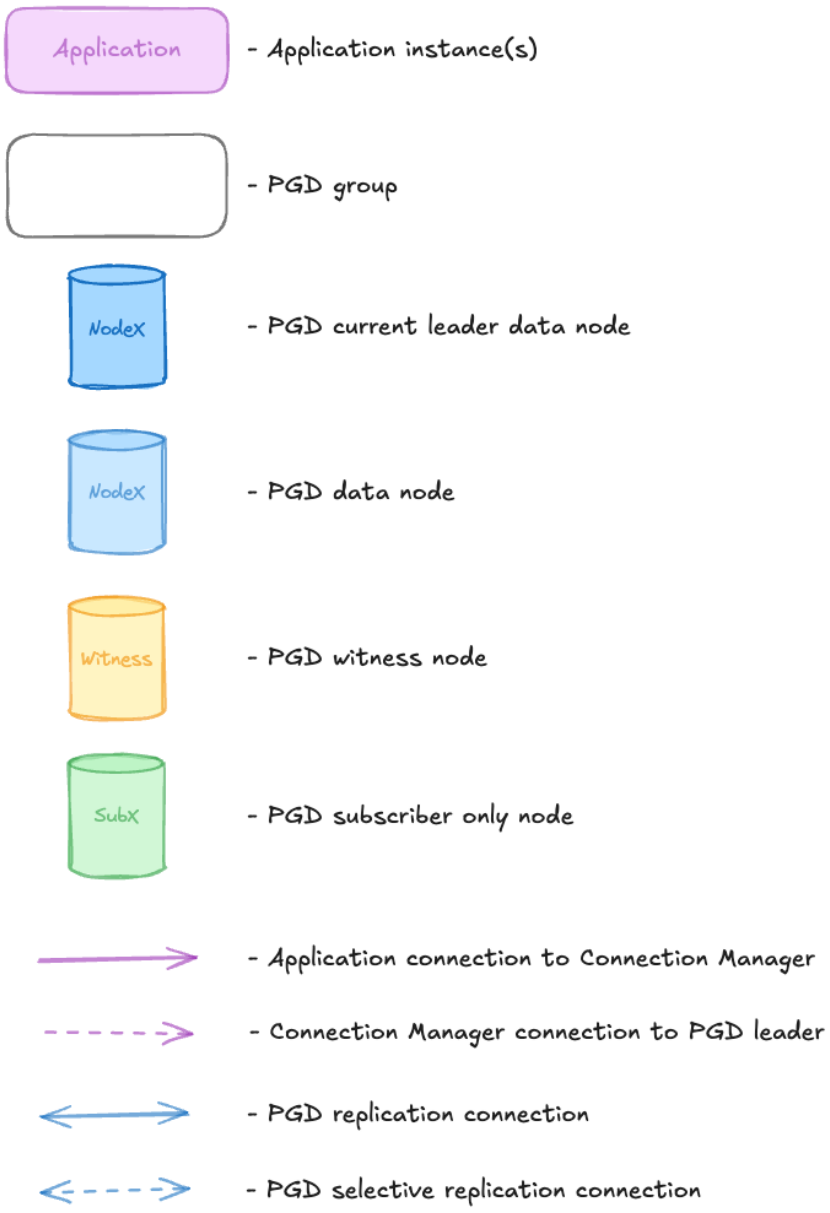


図 9: Diagram legend showing node types, group containers, and connection types

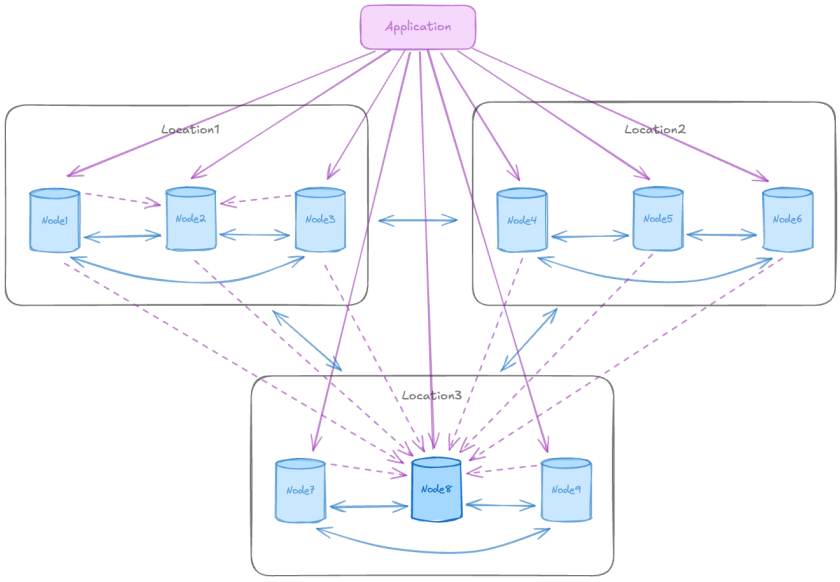


図 10: Three data groups with global routing topology

た最新のコミットデータがあることを保証します。

- **アダプティブプロテクト**利用可能な場合はグループ内のノードの大部分への同期レプリケーション、構成可能なタイムアウトの後、非同期レプリケーションに自動的に低下します。グループごとに 2 つのデータノードと 1 つの監視を使用する場合にお勧めします。
- **ローカルプロテクト**ローカルノードのみの耐久性を備えた非同期レプリケーション。実稼働環境での使用はお勧めしません。

コミットスコープは、パフォーマンス、耐久性、一貫性、スケーラビリティにわたってさまざまなトレードオフを行います。

### Single data group, global read scaling

グローバル読み取りスケーリングパターンは、複数のサブスクライバ専用グループを 3 つ以上のノードで構成される中央の書き込みグループに追加します。接続マネージャーは、書き込みグループ内の書き込みリーダールーティングを自動的に処理しますが、各サブスクライバ専用グループは独自のロードバランサーとスケーリングポリシーを持つことができます。

サブスクライバー専用ノードはすべての変更を受信しますが、コンセンサスを書き込んだり、コンセンサスに参加したりすることはできません。書き込みグループに影響を与えることなくグループは動的に追加および削除でき、書き込みグループ自分自身を複数のリージョンまたはアベイラビリティーゾーンに分散して可用性を高めることができます。

#### 図の凡例

シンボルは、ノードタイプ、グループコンテナ、および接続タイプを表します。

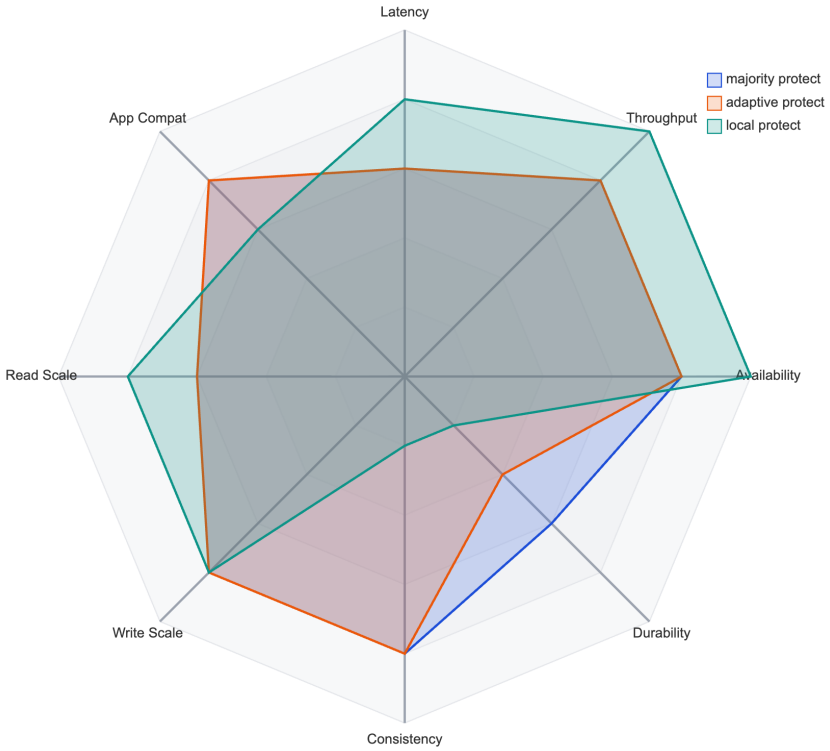


図 11: Three data groups behavior diagram

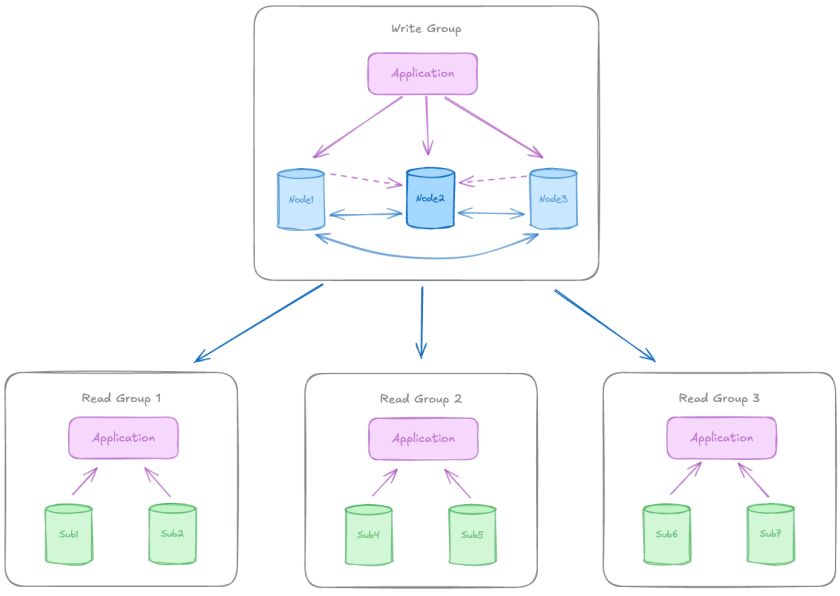


図 12: Single data group global read scaling topology

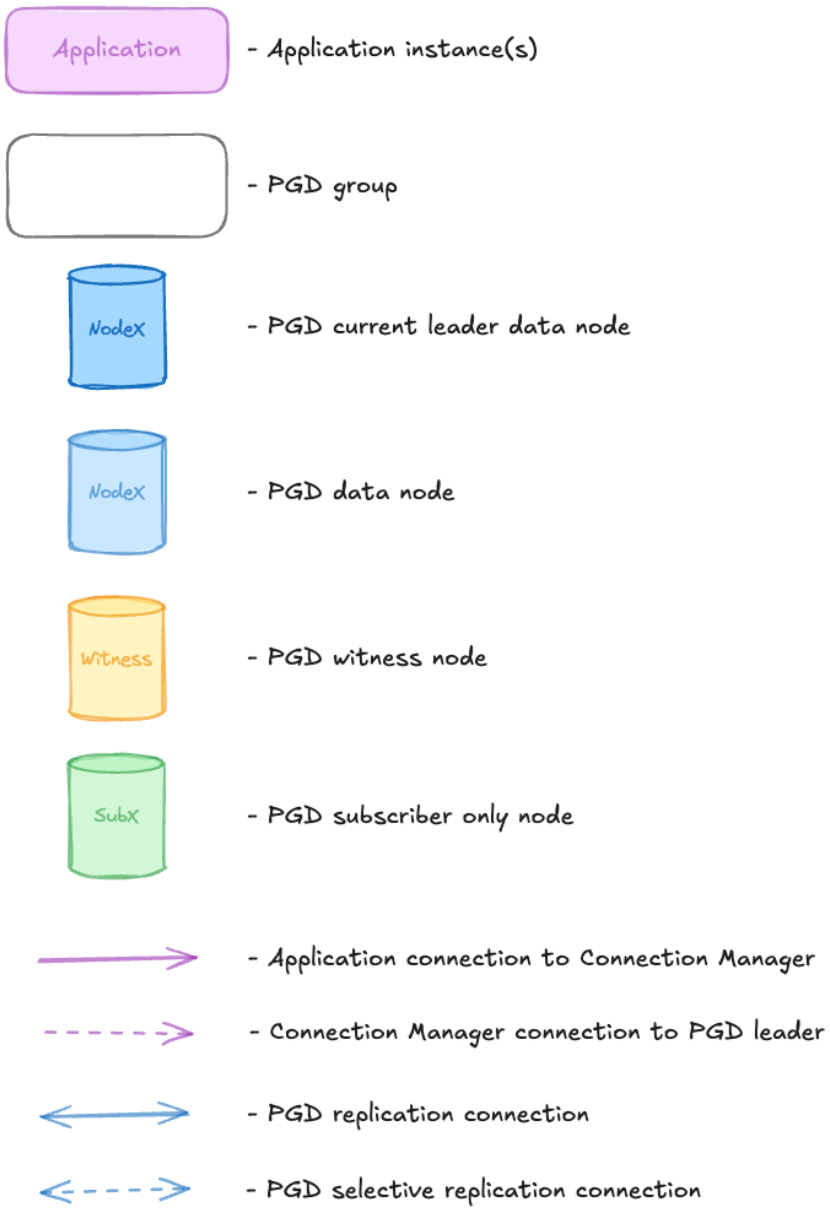


図 13: Diagram legend showing node types, group containers, and connection types

### このパターンを使用する場合

このパターンは、分析とレポート、API 読み取りトラフィック、構成配布、およびリージョンリードレプリカ用に別個のグループを使用した、ワークロードの分離を必要とする読み取り中心のワークロードに適しています。各グループは、ワークロードに基づいて独立してスケーリングします。

このパターンには次のような利点があります。

- 書き込みクラスターに影響を与えることなく、ノードを追加することにより、読み取りを大規模にスケールします。
- レポートと分析を運用書き込みパスからオフロードします。
- サブスクライバー専用ノードは、ローカル読み取りレイテンシーのために追加リージョンに配置できます。
- クラスターを再構成せずに読み取りノードを追加および削除することにより、動的なスケーリングをサポートします。
- 大量の読み取りワークロードをトランザクション処理から分離します。

注意すべき制限がいくつかあります。

- サブスクライバー専用ノードにはレプリケーションラグがあり、読み取り後書き込みの一貫性要件には適していません。
- アプリケーションは、加入者専用ノードとデータノードを区別して、接続を明示的にルーティングする必要があります。
- サブスクライバノードの遅延またはダウンタイムは、データノードでの WAL 保持を増加させ、ディスク使用量に影響を与えます。

### このパターンに推奨されるコミットスコープ

- **マジョリティプロテクト**書き込みグループ内のマジョリティデータノードへの同期レプリケーション。実稼働使用に最も一般的に推奨されるスコープ。
- **アダプティブプロテクト**可能な場合は書き込みグループ内のノードの大部分への同期レプリケーション、構成可能なタイムアウトの後、非同期レプリケーションに自動的に低下します。2 つのデータノードと 1 つの監視を使用する場合に推奨されます。
- **ローカルプロテクト**ローカルノードのみの耐久性を備えた非同期レプリケーション。実稼働環境での使用はお勧めしません。

コミットスコープは、パフォーマンス、耐久性、一貫性、スケーラビリティにわたってさまざまなトレードオフを行います。

### Primary active group, DR group

プライマリアクティブグループ、DR グループパターンは、2 つのデータノードを備えたフルキャパシティのプライマリサイトと、1 つのノードを備えた容量の削減されたディザスターリカバリー DR サイトをペアにします。接続マネージャーは、2 つのプライマリノード間の自動フェイルオーバーを処理し、オペレーターの介入なしで単一ノードの障害が発生してもプライマリサイトを利用可能な状態に保ちます。

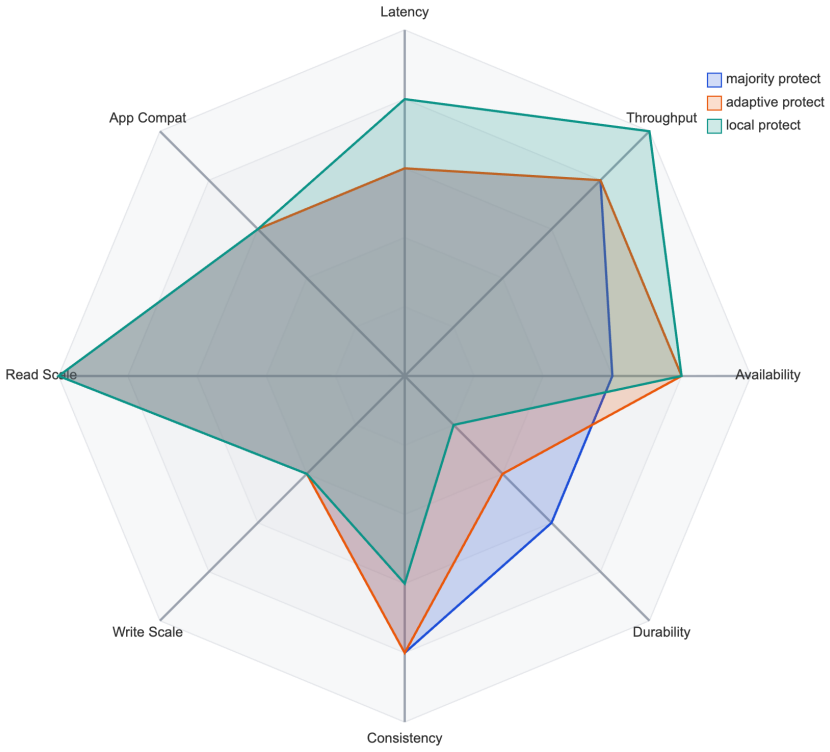


図 14: Single data group global read scaling behavior diagram

DR サイトは非同期レプリケーションを受信し、フェールオーバーのためにスタンバイのままです。DR フェイルオーバー中に、アプリケーションは、プライマリが復元されるまで、容量を削減した単一の DR ノードにスイッチします。

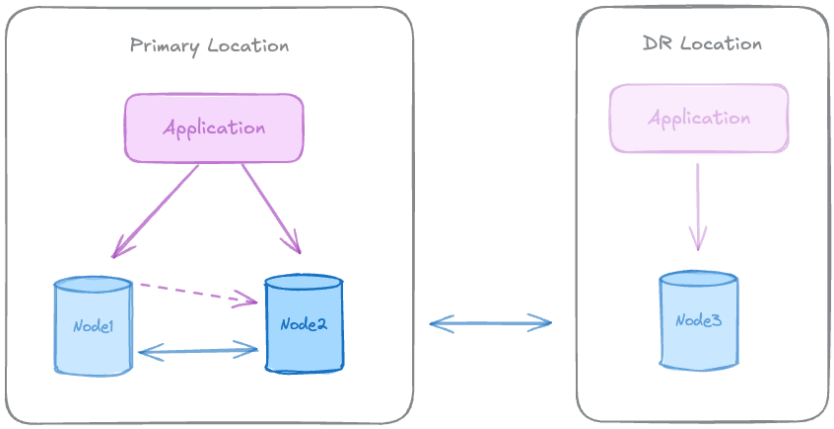


図 15: Primary active group, DR group topology

クラスターは、利用可能なノードに応じて、4 つの状態を移行します。

- 定常状態プライマリで 2 つのノードがアクティブで、1 つのノードが DR ロケーションでレプリケーションを受信します。
- プライマリノードの障害数秒で 2 番目のプライマリノードへの自動フェイルオーバー。
- プライマリデータセンター DC の障害数分での DR ノードへの手動または調整されたフェイルオーバー。

- リカバリー DR ノードからプライマリ DC を再構築し、フェールバックします。

### 図の凡例

シンボルは、ノードタイプ、グループコンテナ、および接続タイプを表します。

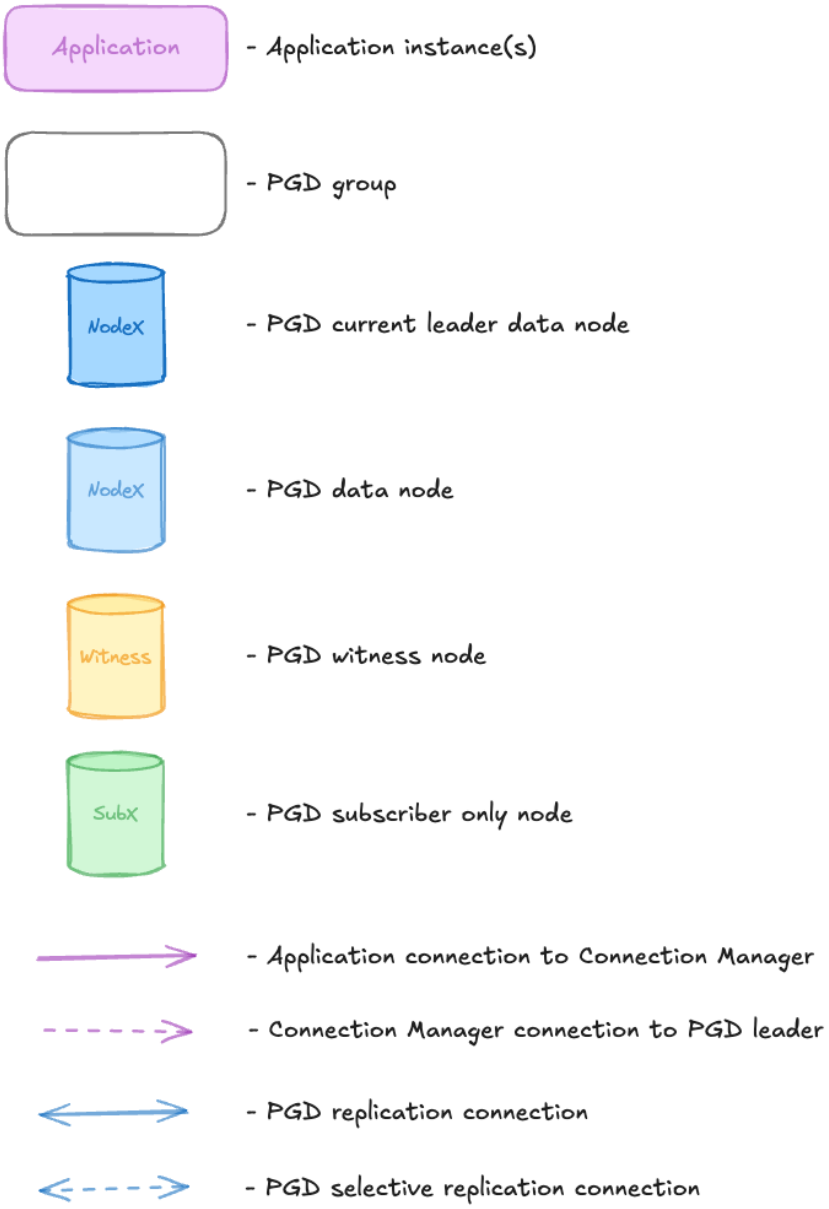


図 16: Diagram legend showing node types, group containers, and connection types

### このパターンを使用する場合

このパターンは、DR インフラストラクチャの予算に制限があるディザスターリカバリーを必要とする組織に適しています。プライマリサイトにはフル HA があり、DR サイトはより低コストでデータ保護と復旧機能を提供します。これは、フルキャパシティのアクティブ/アクティブが正当化されない展開で一般的です。

このパターンには次のような利点があります。

- DR を使用した最も安価なセットアップ、別の場所でデータを保護。

- プライマリデータセンター内のフル HA、単一ノードの障害に耐えます。
- DR ノードは、読み取り専用のレポートと分析に使用できます。
- 完全なマルチ DC アクティブ-アクティブよりも簡単です。

注意すべき制限がいくつかあります。

- DR 容量は単一のノードに制限され、DR サイトに HA はありません。
- DR フェイルオーバーには、手動介入またはオーケストレーションが必要です。
- アクティブ/アクティブワークロードはサポートされていません。
- 高可用性を維持するためにレプリケーションが非同期に低下するため、カスケード障害または復旧時に一貫性の問題が発生する可能性があります。

### このパターンに推奨されるコミットスコープ

- **アダプティブプロテクト**プライマリグループの両方のノードへの同期レプリケーション、ノード損失時の可用性のために非同期レプリケーションに低下します。実稼働使用に最も一般的に推奨されるスコープ。
- **ローカルプロテクト**ローカルノードのみの耐久性を備えた非同期レプリケーション。実稼働環境での使用はお勧めしません。

コミットスコープは、パフォーマンス、耐久性、一貫性、スケーラビリティにわたってさまざまなトレードオフを行います。

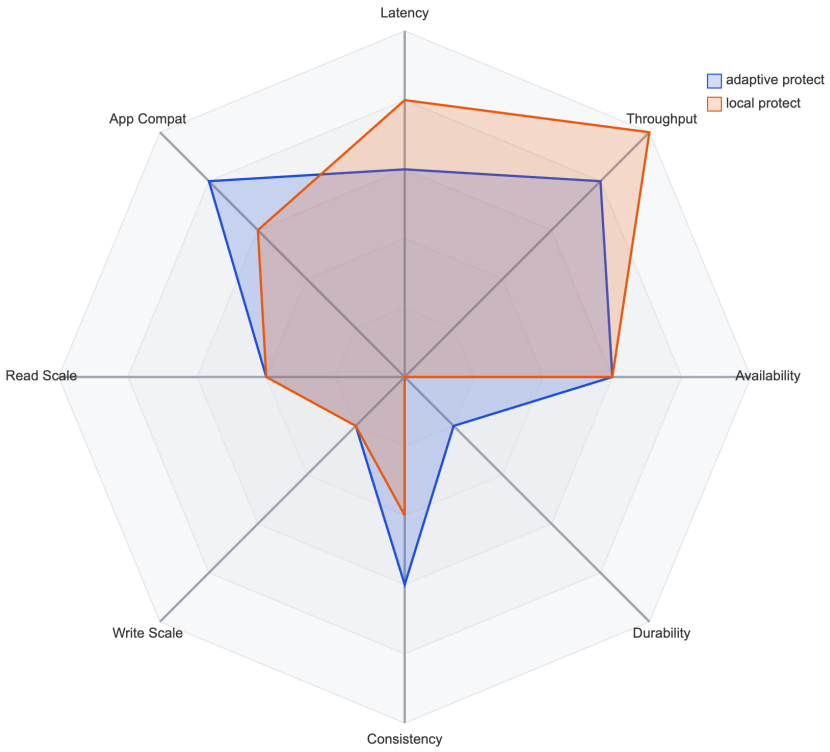


図 17: Primary active group, DR group behavior diagram

### Multiple locations, data residency

複数の場所、データ常駐パターンは、個人データと機密データが元のリージョンに残り、国境を越えて複製されることはないマルチリージョン展開です。各リージョンは、ローカルの高可用性 HA 用に 2 つ以上のデータノードを使用して独自の独立した Raft コンセンサスグループを実行します。

製品カタログ、構成、匿名化された分析を含むグローバル参照データのみがリージョン間で複製され、GDPR、CCPA、およびその他のデータローカライゼーション法へのコンプライアンスが可能になります。

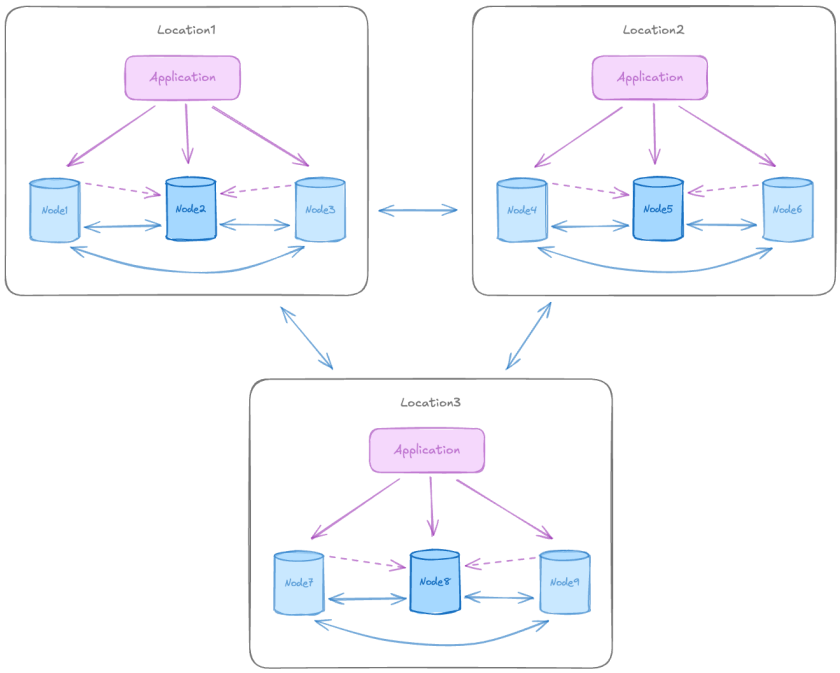


図 18: Multiple locations, data residency topology

選択的レプリケーションは、データを 2 つのカテゴリーに分類することにより機能します。

- **ローカルのみ** リージョン内でのみ複製される個人識別情報 PII、顧客データ、財務記録、および医療データ。
- **グローバル** どこでも複製される製品カタログ、構成、参照データ、および匿名化された分析。

### 図の凡例

シンボルは、ノードタイプ、グループコンテナ、および接続タイプを表します。

### このパターンを使用する場合

このパターンは、銀行データを国内に維持する必要がある金融サービス、患者データの居住地が必要なヘルスケア、政府機関、顧客データを維持する必要がある多国籍 SaaS など、データローカライゼーションの厳しい規制要件を備えたジオレプリケーションを必要とするグローバル企業に適しています。特定の管轄区域。

このパターンには次のような利点があります。

- GDPR、CCPA、およびデータローカライゼーション規制へのコンプライアンスが可能になります。
- 国境を越えたデータ転送の法的な複雑さを回避します。

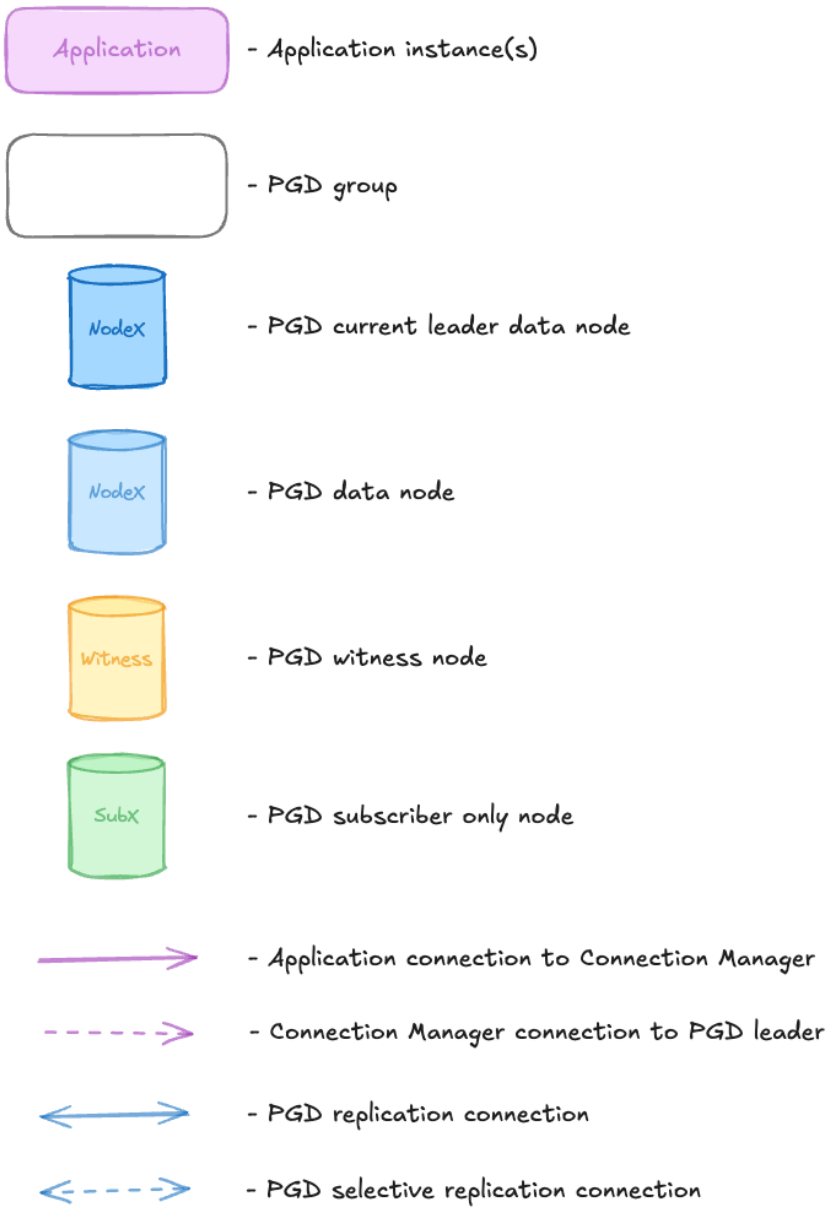


図 19: Diagram legend showing node types, group containers, and connection types

- データをリージョン内に維持することにより、リージョンユーザーに低遅延を提供します。
- リージョンデータルーティングを使用した単一のグローバルアプリケーションコードベース。
- リージョンの障害を分離するので、1 つのリージョンがダウンしても他のリージョンに影響を与えません。

注意すべき制限がいくつかあります。

- すべてのデータのクロスリージョンビューはありません。
- スキーマデザインは、ローカルデータとグローバルデータを明確に分離する必要があります。
- モニタリングは、リージョンごとのコンプライアンスを追跡する必要があります。

### このパターンに推奨されるコミットスコープ

- **マジョリティプロテクト**リージョン内のマジョリティノードへの同期レプリケーション。実稼働使用に最も一般的に推奨されるスコープ。
- **アダプティブプロテクト**利用可能な場合は大部分のノードへの同期レプリケーション、構成可能なタイムアウトの後、非同期レプリケーションに自動的に低下します。2 つのデータノードと 1 つの監視を使用する場合に推奨されます。
- **ローカルプロテクト**ローカルノードのみの耐久性を備えた非同期レプリケーション。実稼働環境での使用はお勧めしません。

コミットスコープは、パフォーマンス、耐久性、一貫性、スケーラビリティにわたってさまざまなトレードオフを行います。

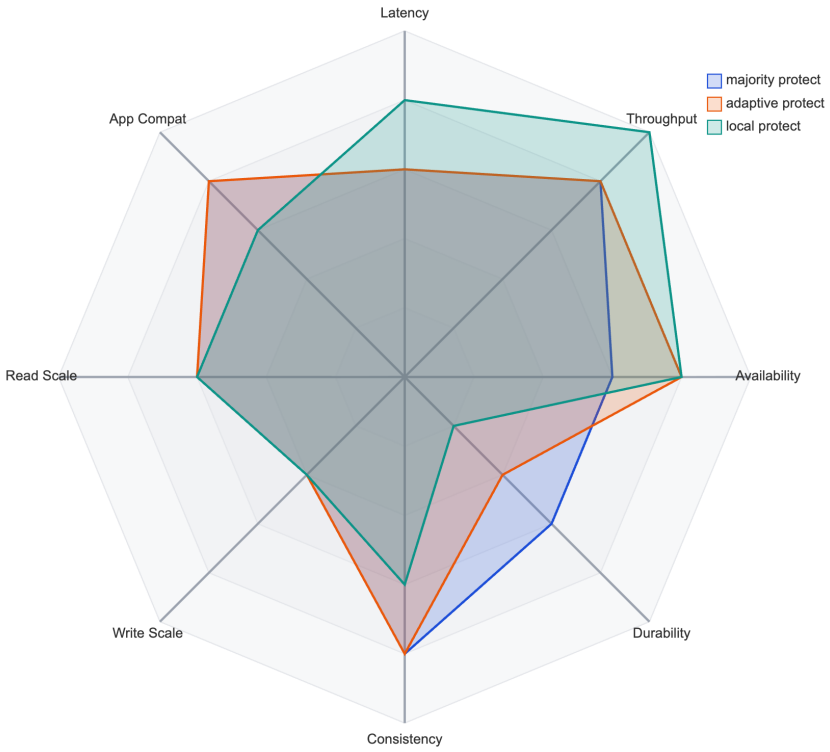


図 20: Multiple locations, data residency behavior diagram

## 4.2 Choosing a Postgres distribution

EDB Postgres Distributed は、3 つの異なる Postgres ディストリビューション PostgreSQL、 EDB Postgres Extended Server、または EDB Postgres Advanced Server で展開できます。特定の EDB Postgres 分散機能の可用性は、使用されている Postgres ディストリビューションによって異なります。したがって、ビジネスニーズに最適な Postgres ディストリビューションを採用することが重要です。たとえば、Commit At Most Once CAMO 機能を使用することがユースケースにとってミッションクリティカルである場合は、CAMO を処理するために必要なコア機能を持たないオープンソースの PostgreSQL を採用しないでください。

次の表は、 Postgres ディストリビューションとバージョンに依存する EDB Postgres Distributed の機能をリストしています。

| Feature | PostgreSQL | EDB Postgres Extended | EDB Postgres Advanced |
|---|---|---|---|
| *Replicated large objects* | Y | Y | Y |
| *Replicated large objects* | Y | Y | Y |
| ステートメント固有の *DDL* レプリケーションの問題 | Y | Y | Y |
| *Replicated large objects* | Y | Y | Y |
| *Replicated large objects* | Y | Y | Y |
| *Replicated large objects* | Y | Y | Y |
| サブスクライバー専用ノード | Y | Y | Y |
| モニタリング パラレル適用 | Y | Y | Y |
| *Parallel Apply* | Y | Y | Y |
| Conflict-free replicated data types (CRDTs) | Y | Y | Y |
| 列レベルの競合解決の特別な問題 | Y | Y | Y |
| *Transform* トリガー | Y | Y | Y |
| 競合トリガー | Y | Y | Y |
| 非同期レプリケーション | Y | Y | Y |
| *Legacy synchronous replication using PGD* | Y | Y | Y |
| 同期コミットとグループコミット | N | Y | Y |
| *Replicated large objects* | N | Y | Y |
| 熱心な競合解決 | N | Y | Y |
| ラグコントロール | N | Y | Y |
| *Decoding worker* | N | Y | Y |
| *Replicated large objects* | N | Y | Y |
| *Replicated large objects* | N | Y | Y |
| *Replicated large objects* | N | Y | Y |
| *Replicated large objects* | N | Y | Y |

次のページに続く

表 2 – 前のページからの続き

| Feature | PostgreSQL | EDB Postgres Extended | EDB Postgres Advanced |
|---|---|---|---|
| 透過的データ暗号化 *TDE* のサポート | N | 15+ | 15+ |

## 4.3 Choosing your deployment method

次の方法を使用して、 EDB Postgres DistributedPGD 製品を展開およびインストールできます。

- Manual deployment and administration は、選択したインフラストラクチャに PGD を直接手動でプロビジョニング、インストール、および構成できることにより、きめの細かい制御を提供します。この方法には、基になる Postgres エンジンを手動でセットアップし、ノードグループ設定を構成することが含まれます。このアプローチは、オーダーメイド環境に最大限の柔軟性を提供しますが、PGD の構成パラメーターとライフサイクル管理を深く理解する必要があります。
- Trusted Postgres Architect TPA は、Ansible を使用して Postgres クラスターをビルドするオーケストレーションツールです。さまざまなシナリオで Postgres をセットアップおよび操作する方法を文書化したリファレンスアーキテクチャセットを使用して Postgres クラスターをビルドします。TPA は、EDB が従うベストプラクティスを表し、その推奨事項はクイックテストベッドおよび実稼働環境に適用されます。TPA の柔軟性により、仮想マシン、AWS クラウドインスタンス、または Linux ホストハードウェアに展開できます。
- EDB Postgres Distributed for Kubernetes は、EDB によって設計、開発、およびサポートされている Kubernetes オペレーターです。PGD レプリケーションを使用した、マルチマスターアーキテクチャと高可用性 Postgres データベースクラスターのライフサイクル全体をカバーします。これは、オープンソースの CloudNativePG オペレーターをベースにしており、EDB Postgres Advanced Server を使用した Oracle との互換性、 EDB Postgres Extended または Advanced Server を使用した透過的データ暗号化 TDE、IBM Power や OpenShift を含む追加のサポート対象プラットフォームなどの追加価値を提供します。
- Hybrid Manager は、ハイブリッドおよびマルチクラウド環境全体で PGD クラスターをオーケストレーションおよび監視するように設計された管理コンポーネントです。さまざまなインフラプロバイダーにまたがる PGD ノードの管理の複雑性を簡素化し、ライフサイクル操作、状態モニタリング、構成管理のための統一インターフェイスを提供します。オンプレミスデータセンターとパブリッククラウド展開のギャップを埋めることにより、ハイブリッドマネージャーは、地理的に分散したクラスターの一貫した運用標準と高可用性を保証します。

## 4.4 Comparing replication solutions

レプリケーションソリューションは、マルチリージョンの展開、競合管理、スキーマの変更、およびメジャーバージョンのアップグレードを処理する方法が大きく異なります。この比較を使用して、これらのディメンションの他のオプションと比較して PGD を評価します。

**最終更新日 2026-04-14**

| Product | Version |
|---|---|
| PGD Standalone | 製品バージョン 6.4 |
| PGD HM-managed | 製品バージョン 6.4 |
| PSR + EFM | 5.3 |

### 4.4.1 展開

| Product | On premis | Cloud | Multi-cloud | Hybrid (on-prem + cloud) | SLA | Native Kubernetes Operator |
|---|---|---|---|---|---|---|
| PGD Standalone | ✓ | ✓ | ✓ | ✓ | 99.999 | ✓ (PGD for Kubernetes (GA Q1 2025), based on CloudNativePG) |
| PGD HM-managed | ✓ | ✓ | ✓ | No | 99.995 | ✓ |
| PSR + EFM | ✓ | ✓ | ✓ | ✓ | 99.99 | Partial (CloudNativePG, no native EFM integration) |

### 4.4.2 パフォーマンス

| Product | Read scalability | Horizontal scalability | Transaction streaming | Parallel apply (vertical scalability) | Logical Change Record (LCR) sharing |
|---|---|---|---|---|---|
| PGD Standalone | ✓ (subscribe only nodes) | No | ✓ | ✓ | ✓ (decodes WAL once, shares to all subscribers) |
| PGD HM-managed | No | No | ✓ | ✓ | ✓ |
| PSR + EFM | ✓ (physical standbys) | No | ✓ | No | No (decodes per subscriber) |

### 4.4.3 耐久性

| Product | Asynchronous replication | Optional RPO limit for asynchronous replication | Synchronous replication | Consensus based replication | Per transaction durability setting |
|---|---|---|---|---|---|
| PGD Standalone | ✓ | ✓ | ✓ | ✓ | ✓ |
| PGD HM-managed | ✓ | ✓ | ✓ | ✓ | ✓ |
| PSR + EFM | ✓ | No | ✓ | No | ✓ |

### 4.4.4 一貫性

| Product | Automatic conflict management | Conflict avoidance types | Conflict avoidance at commit (pessimistic conflict handling) | Builtin distributed sequence | CRDT (Conflict-Free Replicated Data Types) |
|---|---|---|---|---|---|
| PGD Standalone | ✓ | ✓ | ✓ | ✓ (snowflake (bigint), galloc (int/bigint)) | ✓ (multiple nodes update same row simultaneously without global locking) |
| PGD HM-managed | ✓ | ✓ | ✓ | ✓ (snowflake (bigint), galloc (int/bigint)) | ✓ |
| PSR + EFM | N/A | N/A | N/A | N/A | No |

### 4.4.5 データ分散

| Product | Data residency/selective replication | Cluster level Active-Active (writers in different regions) | Regional Active-Active (multiple writers within region) | Automatic partitioning | Offload cold data to cheaper storage |
|---|---|---|---|---|---|
| PGD Standalone | ✓ | ✓ | ✓ (not recommended unless specific setup) | ✓ | ✓ |
| PGD HM-managed | ✓ | ✓ | No | ✓ | ✓ |
| PSR + EFM | No | No | No | ✓ with EPAS | No |

### 4.4.6 メンテナンス

| Product | Near-zero downtime major version upgrades by adding nodes | Near-zero downtime inplace major version upgrades | Rolling schema upgrades/green blue (with application assistance) | Rolling maintenance operations |
|---|---|---|---|---|
| PGD Standalone | ✓ | ✓ | ✓ | ✓ |
| PGD HM-managed | ✓ | ✓ | ✓ | ✓ |
| PSR + EFM | ✓ (using logical replication) | No | No | No |

### 4.4.7 接続管理

| Product | Automatic failover | Automatic connection failover for switchover | Cluster level connection routing | Region level connection routing | pgbouncer support | Built-in connection manager (no external consensus/proxy required) |
|---|---|---|---|---|---|---|
| PGD Standalone | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ (PGD 6.0+, eliminates HARP + etcd dependency) |
| PGD HM-managed | ✓ | ✓ | ✓ | ✓ | No | ✓ |
| PSR + EFM | ✓ | ✓ | ✓ | No | ✓ | No (requires pgbouncer / HAProxy) |

### 4.4.8 DDL サポート

| Product | General DDL replication | Granular (per-object) DDL locking | Create and drop objects | Add columns to table | Change column type | CREATE TABLE AS |
|---|---|---|---|---|---|---|
| PGD Standalone | ✓ | ✓ | ✓ | ✓ | ✓ (rewrite requires permit_unsafe_ | ✓ (with restrictions) |
| PGD HM-managed | ✓ | ✓ | ✓ | ✓ | ✓ (rewrite requires permit_unsafe_ | ✓ (with restrictions) |
| PSR + EFM | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

### 4.4.9 Postgres の互換性

| Product | Latest supported version | Works on standard PG | Supports TDE | Supports custom types (i.e. PostGIS) | Supports extensions | CDC failover support | Large Object support | Multiple DB support |
|---|---|---|---|---|---|---|---|---|
| PGD Standalone | 18 | ✓ | ✓ with EPAS/PGE | ✓ | Many | ✓ (PGD 5.7+) | No | No |
| PGD HM-managed | 18 | No | ✓ with EPAS/PGE | ✓ | Many | ✓ (PGD 5.7+) | No | No |
| PSR + EFM | 17 | ✓ | ✓ with EPAS/PGE | ✓ | All | ✓ | ✓ | ✓ |

### 4.4.10 管理

| Product | CLI | UI | Monitoring options |
|---|---|---|---|
| PGD Standalone | ✓ | ✓ (PEM) | Hybrid Manager, SQL, CLI, PEM |
| PGD HM-managed | ✓ | ✓ | Hybrid Manager, SQL, CLI, edbctl |
| PSR + EFM | ✓ | ✓ (PEM) | Hybrid Manager, SQL, CLI, PEM |

### 4.4.11 ライセンス

| Product | Source available | Open source |
|---|---|---|
| PGD Standalone | No | No |
| PGD HM-managed | No | No |
| PSR + EFM | No | No |

## 4.5 Choosing between Physical Streaming Replication (PSR) and EDB Postgres Distributed (PGD)

EDB は、高可用性と分散高可用性の両方をサポートしています。高可用性は、物理ストリーミングレプリケーション PSR を介して実現されますが、通常は EDB Failover EFM や Patroni などの外部ツールで管理されます。分散高可用性は、EDB Postgres 分散 PGD で実現されます。

Postgres の適切なレプリケーション戦略の選択は、アプリケーション固有の可用性、一貫性、および地理的要件によって異なります。PSR はローカルの高可用性と標準の DR シナリオに効果的ですが、PGD は、高度な一貫性と耐久性の保証とともに継続的な書き込み可用性を必要とする「常時オン」のグローバルアーキテクチャ向けに設計されています。

### 4.5.1 EFM または Patroni を使用した PSR

PSR セットアップでは、ノードは通常、プライマリまたはスタンバイとして分類されます。プライマリは、すべての書き込み操作を処理し、先行書き込みログ WAL ストリームを生成する中央機関として機能します。スタンバイは、継続的な復旧状態を維持する復元力のあるセカンダリノードとして機能します。

PSR はブロックレベルで動作し、1 つ以上のスタンバイにプライマリのバイトごとのコピーを作成します。PostgreSQL は自動フェイルオーバーをネイティブに処理しないため、外部オーケストレーションツールが必要です。

#### PSR 管理ツール

PSR を管理するためのツールは数多くありますが、ここで強調表示されている 2 つのツールは、EDB 顧客または最も広く採用されているオープンソースソリューションによって最も広く展開されているため、EDB によってサポートされています。

- ** EDB Failover Manager (EFM): ** このエンタープライズグレードのエージェントは、クラスターの状態を監視し、仮想 IPVIP アドレスを管理し、スタンバイからプライマリへのプロモーションを自動化します。このソリューションは、ネイティブ PSR を優先し、書き込み操作が単一のノードに集中しているベアメタルおよび仮想化プラットフォームを使用する場合に最適です。このアーキテクチャでは、セカンダリリージョンは通常、ディザスターリカバリー DR ターゲットとして機能します。リージョン間でトラフィックを移動するには、通常、データベースとアプリケーションレイヤーの両方で正式な調整されたプロモーションプロセスが必要です。
- ** Patroni: ** これは、活発な開発コミュニティがある PSR 用の人気のオープンソースオーケストレーションツールです。Patroni は、外部の分散コンセンサスストア DCS、最も一般的には `etcd` に依存して、リーダー選択とクラスター状態を維持します。EFM と同様に、このソリューションは、通常プライマリおよびセカンダリリージョンにサービスを提供し、サイト間で操作をシフトするための調整された手動またはスクリプト化されたプロセスを必要とするアーキテクチャを使用した PSR を既に選択している場合に最適です。

### 4.5.2 EDB Postgres Distributed PGD

PGD は、PostgreSQL を分散マルチマスタープラットフォームに変換する統合拡張スタックです。これは、PGD アーキテクトによって PostgreSQL コアに提供された**論理レプリケーション**機能をもとに構築されていますが、コア機能に欠けている重要なクラスター全体のオーケストレーションを追加します。

#### PGD を選択する場合

PGD は、ダウンタイムがオプションでなく、グローバルなフットプリント全体でデータの完全性を保証する必要があるティア 1 エンタープライズアプリケーション向けに設計されています。要件に次のものが含まれる場合、PGD を検討します。

- **高度な耐久性と分散一貫性**同期性の限定された大まかなオプションを提供する PSR とは異なり、PGD は真の分散一貫性を提供します。commit scopes を介して、ローカルノードとリモートノードの両方の適用動作を正確に定義できます。これにより、トランザクションが永続的であると考えられる前に、確認する必要があるノードの数と特定のリージョンを正確に選択でき、グローバルからトランザクションレベルまでデータ保護に対してパフォーマンスのバランスを保つことができます。

- **常時接続要件 99.999%** PGD は、ゼロ待機フェイルオーバーを提供します。すべてのノードは書き込み可能です。いずれかに障害が発生すると、アプリケーションは昇格プロセスを待たずに別のアクティブノードに再ルーティングします。これにより、PSR に固有の検出とプロモーションの遅延が排除され、目標復旧時間 RTO がネットワークトラフィックの再ルーティングに必要な時間に削減されます。
- **ゼロダウンタイム操作** PGD は、アプリケーションがオンラインになったままのローリングアップグレードと OS パッチを有効にします。クラスターはアクティブ-アクティブであるため、書き込み可用性を失ったり、複雑なスイッチオーバープロセスを必要としたりすることなく、メンテナンスのためにノードを 1 つずつオフラインにすることができます。このアーキテクチャでは、他のノードの可用性やパフォーマンスに影響を与えることなく、特定のノードで `REINDEX` や `VACUUM FULL` などの大量のメンテナンス操作を実行することもできます。
- **グローバル読み取りおよび書き込みスケーリング**大規模な読み取り容量を必要とするワークロードの場合、PGD はサブスクライバ専用ノードをサポートします。書き込み可能なノードと大規模なサブスクライバーをメッシュアーキテクチャ内のユーザーの物理的に近くに配置することにより、クロスリージョンの遅延を排除し、地理に関係なくきびきびとしたアプリケーションの応答時間を保証します。このアプローチでは、プライマリリージョン全体で高性能の書き込みファブリックを維持しながら、読み取りをスケーリングできます。

PostgreSQL の PGD、PSR+EFM、pglogical2、およびコア論理レプリケーション間の詳細な内訳については、*Comparing replication solutions* を参照してください。

現在の環境を移行するには、ステップバイステップガイド Migrating from Postgres Physical Streaming Replication (PSR) to PGD に従ってください。

## 4.6 Other considerations

運用環境に Postgres Distributed を展開する際に、従うべきベストプラクティスが多数あります。これらのプラクティスは、PGD クラスターの信頼性、パフォーマンス、およびセキュリティを保証するのに役立ちます。このセクションでは、運用環境に PGD を展開する際に考慮すべき重要なベストプラクティスのいくつかを概要します。

- *Sizing*
- *Time and PGD*

### 4.6.1 Sizing

#### CPU/コアのサイジング

運用展開の場合、EDB は、Postgres データノードごとに少なくとも 4 コアを推奨します。Witness ノードはデータレプリケーション操作に参加せず、この要件を満たす必要はありません。サブグループ Raft なしで 1 つのコアで十分です。サブグループ Raft を使用する場合、2 つのコアで十分です。

ノードが昇格した場合のパフォーマンスの低下を回避するために、常にデータノードとまったく同じようにロジカルスタンバイのサイズを設定します。

特定のパフォーマンス要件の詳細なベンチマークを作成して、ワークロードに基づいて適切なサイジングを決定することをお勧めします。EDB プロフェッショナルサービスチームが、必要に応じてサポートします。

開発のために、2 コア未満の Postgres データノードを割り当てないでください。Barman ノードのサイズは、データベースのサイズとデータ変更率によって異なります。

Postgres データノードと Barman ノードは、仮想マシンまたはベアメタル展開モードで展開できます。ただし、復元性が低下するため、同じ物理ハードウェア上にある VM に複数のデータノードを展開しないでください。

### 4.6.2 Time and PGD

#### 時計とタイムゾーン

EDB Postgres 分散は、マルチプルのタイムゾーンのノードで動作するように設計されており、真にワールドワイドなデータベースクラスターを実現します。個々のサーバーは一致するタイムゾーンで構成する必要はありませんが、`log_timezone = UTC` を使用して、人間が判読可能なサーバーログにアクセスしやすく比較しやすくすることをお勧めします。

NTP またはその他のソリューションを使用してサーバークロックを同期します。

他の一部のソリューションとは異なり、クロック同期はパフォーマンスにとって重要ではありません。クロックスキューはオリジンの競合検出に影響を与える可能性がありますが、EDB Postgres Distributed は、存在するスキューを報告および管理する制御を提供します。EDB Postgres Distributed は、*bdr_read_all_conflicts* で説明しているように、行バージョンの競合検出も提供します。

# 第5章　Managing lifecycle operations

EDB Postgres DistributedPGD は、マルチマスタークラスターの状態と進化を維持するためのツールと方法論を提供します。ライフサイクル操作を効果的に管理すると、ソフトウェアの更新、移行、構成の調整を通じて高可用性環境のパフォーマンスと安全が維持されます。

- *Deploying and configuring PGD on Hybrid Manager* _ 手動インストール、TPA 自動展開、CloudNativePG 展開、およびハイブリッドマネージャーの統合を含む、さまざまな環境に PGD クラスターをプロビジョニングするための手順。
- *PGD のアップグレード* _ 基になる PostgreSQL エンジンと PGD 拡張機能の両方のマイナーおよびメジャーバージョンのアップグレードを最小限のダウンタイムで実行するための詳細なワークフロー。
- *Testing and tuning PGD clusters* _ `pgd-bench` を使用した分散ワークロードのベンチマークと、スループットを最大化するためにシステムパラメーターを最適化することに関するガイダンス。
- *Backup and recovery* _ 分散マルチマスタートポロジに合わせた物理的および論理的バックアップ戦略を含む、データを保護するためのベストプラクティス。

## 5.1 Deploying and configuring EDB Postgres Distributed

このセクションでは、 EDB Postgres 分散を展開する方法とその構成方法について説明します。

PGD を展開するには、主に 4 つの方法があります。

- Manual deployment and administration は、サーバーセットに EDB Postgres 分散を手動で展開および構成する方法について説明します。
- *Deploying and configuring PGD using TPA* は、TPA を使用して Docker 環境、Linux ホスト、または AWS に分散 EDB Postgres を展開および構成する方法について説明します。
- *Deploying and configuring PGD on Kubernetes* は、Kubernetes 環境に分散 EDB Postgres を展開および構成する方法について説明します。
- *Deploying and configuring PGD on Hybrid Manager* は、ハイブリッドマネージャーに分散した EDB Postgres を展開および構成する方法について説明します。

### 5.1.1 Manually

このセクションでは、EDB Postgres DistributedPGD を手動で展開および構成する方法について説明します。

学習/テスト環境をインストールする場合は、 Quickstart guide を使用することをお勧めします。

PGD クラスターの手動展開は、次の手順で構成されます。

1. **前提条件** を紹介します。

2. *Configuring PGD repositories* 。
3. *Installing the database and pgd* 。
4. *Configuring the cluster* 。
5. *Checking the cluster* 。
6. *Connecting to your PGD cluster* の方法を学びます。

### Prerequisites

#### プロビジョニングホスト

PGD を展開するプロセスの最初のステップは、ホストをプロビジョニングおよび構成することです。

Linux がインストールされているクラウドの仮想マシンインスタンス、Linux がインストールされているオンプレミス仮想マシン、または同じく Linux がインストールされているオンプレミスの物理ハードウェアに展開できます。

どの supported Linux operating system 、およびどの展開プラットフォームを選択しても、マシンのプロビジョニングの結果は、スーパーユーザー、管理者、または sudo 特権を持つユーザーで SSH を使用してアクセスできる Linux システムである必要があります。

プロビジョニングされた各マシンは、クラスターにプロビジョニングしている他のマシンに接続できる必要があります。

クラウド展開では、パブリックネットワークまたは VPC を介してこれを行うことができます。

オンプレミス展開は、ローカルネットワークを介して接続できる必要があります。

| Vendor | Platform | Guide |
|---|---|---|
| Amazon | AWS | Tutorial: Get started with Amazon EC2 Linux instances |
| Microsoft | Azure | *Replicated large objects* |
| Google | GCP | *Replicated large objects* |

#### 管理ユーザーを作成する

プロビジョニングされたインスタンスごとに管理ユーザーを構成することをお勧めします。admin ユーザーは、スーパーユーザーまたは sudo スーパーユーザーへの権限を持っている必要があります。また、証明書を使用してパスワードなしの SSH アクセス用に管理ユーザーを構成することをお勧めします。

#### ネットワーク接続を確認する

作成した管理ユーザーを使用して、各マシンがプロビジョニングしている他のマシンと通信できることを確認します。

特に、 PostgreSQL TCP/IP ポート EDB Postgres Advanced Server の場合は 5444、EDB Postgres Extended およびコミュニティ PostgreSQL の場合は 5432、クラスター内のすべてのマシンで開いている必要があります。PGD 接続マネージャーは、クラスター内のすべてのノードからアクセスできることも必要です。デ

フォルトでは、接続マネージャーはポート 6432 または EDB Postgres Advanced Server の場合は 6444 を使用します。

#### 実際の例

このシリーズの作業例では、Red Hat Enterprise Linux 9 を使用する 3 つのホストがプロビジョニングされました。

- host-1
- host-2
- host-3

これらのホストはクラウドで構成されました。そのため、各ホストにはパブリック IP アドレスとプライベート IP アドレスの両方があります。クラスターのプライベート IP アドレスを使用します。

プライベート IP アドレスは次のとおりです。

- host-1 `192.168.254.166`
- host-2 `192.168.254.247`
- host-3 `192.168.254.135`

クラスター例では、これらのプライベート IP アドレスを使用するように`/etc/hosts` も編集されました。

```
192. 168.254.166 host-1
192. 168.254.247 host-2
192. 168.254.135 host-3
```

### Configuring PGD repositories

PGD データノードとして使用する各ホストで、データベースと PGD ソフトウェアをインストールする必要があります。

#### PGD リポジトリの構成

次の環境変数を設定します。

#### EDB_SUBSCRIPTION_TOKEN

これは、EDB サブスクリプションに登録したときに受け取ったトークンです。EDB リポジトリへのアクセスを認証するために使用されます。

```
export EDB_SUBSCRIPTION_TOKEN=<your-token>
```

#### EDB_SUBSCRIPTION_PLAN

これは、EDB で持っているサブスクリプションのタイプです。 `standard`、`enterprise`、または `community` です。

```
export EDB_SUBSCRIPTION_PLAN=<your-subscription-plan>
```

### EDB_REPO_TYPE

これは、使用するパッケージマネージャーのタイプであり、インストーラーに、どのタイプのパッケージが必要であるかを通知します。これは、Ubuntu / Debian の場合は deb 、CentOS / RHEL の場合は rpm です。

```
export EDB_REPO_TYPE=<your-repo-type>
```

#### リポジトリのインストール

構成する必要があるリポジトリは 2 つあります。1 つはデータベースソフトウェア用、もう 1 つは PGD ソフトウェア用です。

次のコマンドは、データベースの EDB リポジトリを使用するようにパッケージマネージャーを構成するスクリプトをダウンロードして実行します。

```
curl -1sSLf "https://downloads.enterprisedb.com/$EDB_SUBSCRIPTION_TOKEN/$EDB_
↪SUBSCRIPTION_PLAN/setup.$EDB_REPO_TYPE.sh" | sudo -E bash
```

次のコマンドは、PGD の EDB リポジトリを使用するようにパッケージマネージャーを構成するスクリプトをダウンロードして実行します。

```
curl -1sSLf "https://downloads.enterprisedb.com/$EDB_SUBSCRIPTION_TOKEN/postgres_
↪distributed/setup.$EDB_REPO_TYPE.sh" | sudo -E bash
```

#### 実際の例

この例では、エンタープライズサブスクリプションを使用して PGD Expanded で EDB Postgres Advanced Server 17 をインストールできる CentOS / RHEL システムにリポジトリを構成します。

#### 環境変数を設定する

```
export EDB_SUBSCRIPTION_TOKEN=XXXXXXXXXXXXXX
export EDB_SUBSCRIPTION_PLAN=enterprise
export EDB_REPO_TYPE=rpm
```

#### リポジトリのインストール

```
#  For PGD Expanded, there are two repositories to install.

curl -1sSLf " https://downloads.enterprisedb.com/$EDB_SUBSCRIPTION_TOKEN/$EDB_
↪SUBSCRIPTION_PLAN/setup.$EDB_REPO_TYPE.sh" | sudo -E bash
curl -1sSLf " https://downloads.enterprisedb.com/$EDB_SUBSCRIPTION_TOKEN/postgres_
↪distributed/setup.$EDB_REPO_TYPE.sh" | sudo -E bash
```

次のステップは、 *Installing the database and pgd* です。

### Installing the database and pgd

PGD データノードとして使用する各ホストで、データベースと PGD ソフトウェアをインストールする必要があります。

*Configuring PGD repositories* を取得したら、パッケージマネージャーを使用してデータベースと PGD ソフトウェアをインストールできます。

#### データベースと PGD ソフトウェアのインストール

#### Postgres バージョンを設定する

環境変数を設定して、インストールする Postgres のバージョンを指定します。これは通常、Postgres 17 の 17 です。

```
export PG_VERSION=17
```

#### パッケージ名を設定する

環境変数を設定して、データベースと PGD ソフトウェアのパッケージ名を指定します。パッケージ名は、使用しているデータベースと使用しているプラットフォームによって異なります。

#### EDB Postgres Advanced Server

```
export EDB_PACKAGES="edb-as$PG_VERSION-server edb-pgd6-<edition>-epas$PG_VERSION"
```

<edition> は、インストールする PGD エディション essential または expanded です。


```
export EDB_PACKAGES="edb-as$PG_VERSION-server edb-pgd6-<edition>-epas$PG_VERSION"
```

<edition> は、インストールする PGD エディション essential または expanded です。


#### EDB Postgres Extended

```
export EDB_PACKAGES="edb-postgresextended-$PG_VERSION edb-pgd6-<edition>-pgextended
↪$PG_VERSION"
```

<edition> は、インストールする PGD エディション essential または expanded です。

```
export EDB_PACKAGES="edb-postgresextended$PG_VERSION-server edb-postgresextended$PG_
↪VERSION-contrib edb-pgd6-<edition>-pgextended$PG_VERSION"
```

<edition> は、インストールする PGD エディション essential または expanded です。

#### コミュニティ PostgreSQL

```
export EDB_PACKAGES="postgresql-$PG_VERSION edb-pgd6-<edition>-pg$PG_VERSION"
```

<edition> は、インストールする PGD エディション essential または expanded です。

```
export EDB_PACKAGES="postgresql$PG_VERSION-server postgresql$PG_VERSION-contrib edb-
↪pgd6-<edition>-pg$PG_VERSION"
```

<edition> は、インストールする PGD エディション essential または expanded です。

#### インストールコマンドを実行する

プラットフォームに適切なインストールコマンドを実行します。

```
sudo apt install -y $EDB_PACKAGES
```

```
sudo dnf install -y $EDB_PACKAGES
```

このコマンドは、指定されたパッケージとそれに必要な依存関係をインストールします。インストールが完了すると、データベースと PGD ソフトウェアがシステムにインストールされます。

#### 実際の例

この例では、 実際の例 で設定したリポジトリ構成を使用して、CentOS / RHEL システムに PGD Expanded で EDB Postgres Extended Server 17 をインストールします。

```
export PG_VERSION=17
export EDB_PACKAGES="edb-as$PG_VERSION edb-pgd6-expanded-epas$PG_VERSION"
sudo dnf install -y $EDB_PACKAGES
```

### Configuring the cluster

#### クラスターの構成

プロセスの次のステップは、データベースとクラスターを構成することです。

これには、各ホストにログインし、 pgd コマンドを実行してデータベースユーザーとしてクラスターを作成することが含まれます。

これらの手順は、使用しているプラットフォーム、および使用している Postgres のバージョンによって異なります。

#### クラスター名

クラスターの名前を選択する必要があります。これは、PGD CLI およびデータベースでクラスターを識別するために使用される名前です。例では<cluster-name> と呼ばれます。指定しない場合、デフォルト名は pgd です。

#### グループ名

グループの名前を選択する必要があります。これは、PGD CLI およびデータベースでグループを識別するために使用される名前です。例では<group-name> と呼ばれます。

グループ名はクラスター内で一意である必要があります。

### ノード名

また、各ノードの名前を選択する必要があります。これは、PGD CLI およびデータベースでノードを識別するために使用される名前です。例では<node-name> と呼ばれます。これは、ノードが実行されているマシンの名前であるホスト名とは別です。

ノード名は、グループ内およびクラスター内で一意である必要があります。

### パスとユーザー

例で使用されるパスとユーザーは、使用している Postgres のバージョン、および使用しているプラットフォームによって異なります。

Postgres バージョンを選択します

次に、プラットフォームを選択します。

| |
|---|
| *enterprisedb* |
| *5444* |
| */usr/lib/edb-as/$PG_VERSION/bin/* |
| */var/lib/edb-as/$PG_VERSION/main/* |

```
sudo -iu enterprisedb
export PG_VERSION=<version>
export PATH=$PATH:/usr/lib/edb-as/$PG_VERSION/bin/
export PGDATA=/var/lib/edb-as/$PG_VERSION/main/
export PGPORT=5444
```

| |
|---|
| *enterprisedb* |
| *5444* |
| */usr/edb/as$PG_VERSION/bin/* |
| */var/lib/edb/as$PG_VERSION/data/* |

```
sudo -iu enterprisedb
export PG_VERSION=<version>
export PATH=$PATH:/usr/edb/as$PG_VERSION/bin/
export PGDATA=/var/lib/edb/as$PG_VERSION/data/
export PGPORT=5444
```

次に、プラットフォームを選択します。

| |
|---|
| *postgres* |
| *5432* |
| */usr/lib/edb-pge/$PG_VERSION/bin/* |
| */var/lib/edb-pge/$PG_VERSION/main/* |

```
sudo -iu postgres
export PG_VERSION=<version>
export PATH=$PATH:/usr/lib/edb-pge/$PG_VERSION/bin/
export PGDATA=/var/lib/edb-pge/$PG_VERSION/main/
export PGPORT=5432
```

| |
|---|
| *postgres* |
| *5432* |
| */usr/edb/pge$PG_VERSION/bin/* |
| */var/lib/edb-pge/$PG_VERSION/data/* |

```
sudo -iu postgres
export PG_VERSION=<version>
export PATH=$PATH:/usr/edb/pge$PG_VERSION/bin/
export PGDATA=/var/lib/edb-pge/$PG_VERSION/data/
export PGPORT=5432
```

次に、プラットフォームを選択します。

| |
|---|
| *postgres* |
| *5432* |
| */usr/lib/postgresql/$PG_VERSION/bin/* |
| */var/lib/postgresql/$PG_VERSION/main/* |

```
sudo -iu postgres
export PG_VERSION=<version>
export PATH=$PATH:/usr/lib/postgresql/$PG_VERSION/bin/
export PGDATA=/var/lib/postgresql/$PG_VERSION/main/
export PGPORT=5432
```

| |
|---|
| *postgres* |
| *5432* |
| */usr/pgsql-$PG_VERSION/bin/* |
| */var/lib/pgsql/$PG_VERSION/data/* |

```
sudo -iu postgres
export PG_VERSION=<version>
export PATH=$PATH:/usr/pgsql-$PG_VERSION/bin/
export PGDATA=/var/lib/pgsql/$PG_VERSION/data/
export PGPORT=5432
```

### 各ホストで

上記のスクリプト/設定からコマンドを実行して、各ホストで Postgres ユーザーの環境変数とパスを設定します。これにより、pgd コマンドが Postgres 実行可能ファイルとデータディレクトリを見つけられるようになります。

1. 適切なユーザーを使用して、データベースユーザーとしてログインします。

```
sudo -iu <db-user>
```

1. Postgres バージョンの環境変数を設定します。 <version> を、17 など、使用している実際のバージョン番号に置き換えることを忘れないでください。

```
export PG_VERSION=<version>
```

1. Postgres 実行可能ファイルをパスに追加します。

```
export PATH=$PATH:<executable-path>
```

1. Postgres データディレクトリ環境変数を設定します。

```
export PGDATA=<data-directory>
```

1. Postgres パスワード環境変数を設定します。 <db-password> をデータベースユーザーに使用する実際のパスワードに置き換えることを忘れないでください。

```
export PGPASSWORD=<db-password>
```

### 最初のホストで

クラスター内の最初のホストは最初のノードでもあり、クラスターの作成を開始する場所になります。最初のホストで、次のコマンドを実行してクラスターを作成します。

```
pgd node <first-node-name> setup --dsn "host=<first-host> user=<db-user> port=<db-
↪port> dbname=<dbname>" --group-name <group-name>
```

このコマンドは、データディレクトリを作成してデータベースを初期化して、最初のノードにクラスターとグループを作成します。

#### 2 番目のホストで

2 番目のホストで、次のコマンドを実行してクラスターを作成します。

```
pgd node <second-node-name> setup --dsn "host=<second-host> user=<db-user> port=<db-
↪port> dbname=<db-name>" --cluster-dsn "host=<first-host> user=<db-user> port=<db-
↪port> dbname=<db-name>"
```

このコマンドは、2 番目のホストにノードを作成し、cluster-dsn 設定を使用してクラスターに参加し、最初のホストに接続します。

#### 3 番目のホストで

3 番目のホストで、次のコマンドを実行してクラスターを作成します。

```
pgd node <third-node-name> setup --dsn "host=<third-host> user=<db-user> port=<db-
↪port> dbname=<db-name>" --cluster-dsn "host=<first-host> user=<db-user> port=<db-
↪port> dbname=<db-name>"
```

このコマンドは、3 番目のホストにノードを作成し、cluster-dsn 設定を使用してクラスターに参加し、最初のホストに接続します。

#### 実際の例

この例では、前の手順で *Configuring PGD repositories* および *Installing the database and pgd* を作成した CentOS / RHEL システム上の EDB Postgres Extended Server 17 を使用して PGD Essential クラスターを構成します。

次に、 `node-1` 、`node-2` 、および `node-3` と呼ばれる 3 つのノードを使用して `pgd` と呼ばれるクラスターを作成します。

- グループ名は `group-1` になります。ホストは `host-1` 、`host-2` 、および `host-3` です。
- Postgres のバージョンは 17 です。
- データベースユーザーは `postgres` です。
- データベースポートは 5432 です。
- データベース名は `pgddb` です。
- Postgres 実行可能ファイルは`/usr/edb/pge17/bin/` にあります。
- Postgres データディレクトリは`/var/lib/edb-pge/17/main/` にあります。
- Postgres のパスワードは `secret` です。

Postgres バージョン環境変数 PG_VERSION は、前の手順で 17 に設定されていると仮定し、ユーザーを切り替えるときに環境変数を保存していることに注意してください。

#### 最初のホストで

```
sudo  -iu postgres
export PG_VERSION=17
export PATH=$PATH:/usr/edb/pge$PG_VERSION/bin/
export PGDATA=/var/lib/edb-pge/$PG_VERSION/data/
export PGPASSWORD=secret
pgd node node-1 setup --dsn "host=host-1 user=postgres port=5432 dbname=pgddb" --
↪group-name group-1
```

#### 2 番目のホストで

```
sudo -iu postgres
export PG_VERSION=17
export PATH=$PATH:/usr/edb/pge$PG_VERSION/bin/
export PGDATA=/var/lib/edb-pge/$PG_VERSION/data/
export PGPASSWORD=secret
pgd node node-2 setup --dsn "host=host-2 user=postgres port=5432 dbname=pgddb" --
↪cluster-dsn "host=host-1 user=postgres port=5432 dbname=pgddb"
```

#### 3 番目のホストで

```
sudo -iu postgres
export PG_VERSION=17
export PATH=$PATH:/usr/edb/pge$PG_VERSION/bin/
export PGDATA=/var/lib/edb-pge/$PG_VERSION/data/
export PGPASSWORD=secret
pgd node node-3 setup --dsn "host=host-3 user=postgres port=5432 dbname=pgddb" --
↪cluster-dsn "host=host-1 user=postgres port=5432 dbname=pgddb"
```

次のステップは、 *Checking the cluster* です。

### Checking the cluster

#### クラスターの確認

クラスターが起動して実行されたら、いくつかの基本的なチェックを実行して、どの程度効率的に複製されているかを確認することは価値があります。

次の例は、これを行うための 1 つの簡単な方法を示していますが、実行するテストがユースケースに適切であることを確認する必要があります。

インストールおよび構成済みのノードで、ログインして `psql` を実行してデータベースに接続します。 EDB Postgres Advanced Server を使用している場合は、 `enterprisedb` ユーザーを使用します。それ以外の場合は、 `postgres` を使用します。

```
sudo -iu postgres psql pgddb
```

このコマンドは、host-1/node-1 上のデータベースに_直接_接続します。

### クイックテスト

- **準備**
- クラスターの準備ができていることを確認します。
- host-1/node-1 のデータベースにログインします。
- `select bdr.wait_slot_confirm_lsn(NULL, NULL);` を実行します。
- クエリーが返されると、クラスターの準備が整います。
- **データを作成する**クラスターがレプリケートしていることをテストする最も簡単な方法は、1 つのノードにログインし、テーブルを作成し、それに値を設定することです。
- ノード 1 で、テーブル `sql CREATE TABLE quicktest ( id SERIAL PRIMARY KEY, value INT );` を作成します
- ノード 1 で、テーブルにデータを入力します。

```
INSERT INTO quicktest (value) SELECT random()*10000 FROM generate_series(1,10000);
```

- ノード 1 で、パフォーマンスを監視します。

```
select * from bdr.node_replication_rates;
```

- ノード 1 で、値列の合計を取得しますチェック用。

```
select COUNT(*),SUM(value) from quicktest;
```

- **データの確認**
- ノード 2 にログインします。host-2/node-2 のデータベースにログインします。
- ノード 2 で、値列の合計を取得しますチェック用 `sql select COUNT(*),SUM(value) from quicktest;`
- ノード 1 の結果と比較します。
- ノード 3 にログインします。host-3/node-3 のデータベースにログインします。
- ノード 3 で、値列の合計を取得しますチェック用 `sql select COUNT(*),SUM(value) from quicktest;`
- ノード 1 およびノード 2 の結果と比較します。

#### 実際の例

#### 準備

host-1 の Postgres サーバーにログインします。

```
ssh admin@host-1
sudo -iu postgres psql "host=host-1 port=5432 username=postgres dbname=pgddb"
```

これは、PGD のノード 1 への接続です。

#### クラスターの準備ができていることを確認します

クラスターの準備ができていることを確認するには、次のコマンドを実行します。

```
select bdr.wait_slot_confirm_lsn(NULL, NULL)
```

このクエリは、クラスターが初期化でビジーな間はブロックし、クラスターの準備が整うと結果を返します。

別のウィンドウで、host-2 の Postgres サーバーにログインします。

```
ssh admin@host-2
sudo -iu postgres psql "host=host-2 port=5432 username=postgres dbname=pgddb"
```

#### データを作成する

#### ノード 1 で、テーブルを作成します

実行します

```
CREATE TABLE quicktest ( id SERIAL PRIMARY KEY, value INT );
```

#### ノード 1 で、テーブルにデータを設定します

```
INSERT INTO quicktest (value) SELECT random()*10000 FROM generate_series(1,10000);
```

このコマンドは、10000 行のランダム値のテーブルを生成します。

#### ノード 1 で、パフォーマンスを監視します

できるだけ早く実行します。

```
select * from bdr.node_replication_rates;
```

このコマンドは、そのデータが他の 2 つのノードに複製された速度に関する統計を表示します。

```
pgddb=# select * from bdr.node_replication_rates;
__OUTPUT__
 peer_node_id | target_name | sent_lsn   | replay_lsn | replay_lag | replay_lag_bytes␣
```

(次のページに続く)

(前のページからの続き)

```
↪| replay_lag_size | apply_rate | catchup_interv
al
- ------------+------------+----------+-----------+-----------+-----------------
↪+-----------------+-----------+--------------
- --
   1954860017 | node-3   | 0/DDAA908 | 0/DDAA908  | 00:00:00   |               0 | 0␣
↪bytes         |      13682 | 00:00:00
   2299992455 | node-2    | 0/DDAA908 | 0/DDAA908  | 00:00:00   |               0 |␣
↪0 bytes         |      13763 | 00:00:00
(2 rows)
```

そして、それは既に複製されています。

### ノード 1 でチェックサムを取得します

実行します

```
select COUNT(*),SUM(value) from quicktest;
```

このコマンドは、生成されたデータからいくつかの値を取得します。

```
pgddb=# select COUNT(*),SUM(value) from quicktest;
__OUTPUT__
 count  |    sum
- -------+-----------
 100000 | 498884606
(1 row)
```

### データを確認する

### host-2 の Postgres サーバーにログインします

```
ssh admin@host-2
sudo -iu postgres psql "host=host-2 port=5432 username=postgres dbname=pgddb"
```

これは、PGD のノード 2 への接続です。

### ノード 2 で、チェックサムを取得します

実行します

```
select COUNT(*),SUM(value) from quicktest;
```

このコマンドは、生成されたデータのノード 2 の値を取得します。

```
pgddb=# select COUNT(*),SUM(value) from quicktest;
__OUTPUT__
```

(次のページに続く)

(前のページからの続き)

```
 count  |    sum
- -------+-----------
 100000 | 498884606
(1 row)
```

#### ノード 1 の結果と比較します

値は同じです。

ノード 3 でプロセスを繰り返すか、任意のノードで新しいデータを生成し、他のノードにレプリケートされるのを確認できます。

#### host-3 の Postgres サーバーにログインします

```
ssh admin@host-3
sudo -iu enterprisedb psql pgddb
```

これは、PGD のノード 3 への接続です。

#### ノード 3 で、チェックサムを取得します

実行します

```
select COUNT(*),SUM(value) from quicktest;
```

このコマンドは、生成されたデータのノード 3 の値を取得します。

```
pgddb=# select COUNT(*),SUM(value) from quicktest;
__OUTPUT__
 count  |    sum
- -------+-----------
 100000 | 498884606
(1 row)
```

#### ノード 1 およびノード 2 の結果と比較します

値は同じです。

### Connecting to your PGD cluster

EDB Postgres DistributedPGD は、Postgres と同じ接続方法を使用します。違いは、クラスターへの接続のほとんどが、クラスター内のすべてのデータノードに組み込まれている接続マネージャーを経由することです。

データノードに直接接続することもできますが、メンテナンス以外の目的で、特定のノードのデータベースインスタンスで作業する場合はお勧めしません。

通常、接続マネージャーを介してクラスターに接続する必要があります。PGD は、展開と管理が簡単になるように設計されています。つまり、クラスターには、クラスターへのすべての書き込みを処理する書き込みリーダーノードがあることを意味します。接続マネージャーは、読み取り/書き込み接続を書き込みリーダーに転送します。クライアントまたはアプリケーションが行う必要があるのは、接続マネージャーのポートを使用することだけであり、接続マネージャーが残りを処理します。

接続マネージャーは、書き込みを書き込みリーダーに転送し、読み取りがクラスター内の正しいノードに転送されるようにします。データノードに直接接続する場合、これらの機能を利用できない場合があります。データを読み取ることのみが必要なアプリケーションの場合、接続マネージャーは、書き込みリーダーではないノードに読み取りを転送できます。これは、クラスターの負荷のバランスをとり、パフォーマンスを向上させるのに役立ちます。

#### 接続マネージャーを介して接続する

Postgres はポートと接続の構成が非常に柔軟であるため、簡単にするために、この例では Postgres と接続マネージャーのデフォルトのポート設定を使用します。Postgres のデフォルトポートは 5432 で、接続マネージャーのデフォルトポートは 6432 です。

接続文字列でそのポートを使用して、クラスターに接続できます。したがって、たとえば、psql コマンドラインツールを使用している場合、次のようにクラスターに接続できます。

```
psql -h host-1 -p 6432 -U pgdadmin -d pgddb
```

`host-1` は、接続しているノードのホスト名です。接続マネージャーは、接続をクラスター内の書き込みリーダーノードに転送します。

#### データノードに直接接続する

クラスター内のデータノードに直接接続できますが、それはお勧めしません。ただし、データノードに直接接続する必要がある場合は、次のコマンドを使用できます。

```
psql -h host-1 -p 5432 -U pgdadmin -d pgddb
```

### 5.1.2 Deploying and configuring PGD using TPA

Trusted Postgres ArchitectTPA は、自己管理インフラストラクチャでの高可用性 EDB Postgres 分散 PGD クラスターを展開および管理するためのプライマリオーケストレーションツールです。Ansible を使用して、PGD スタック全体のインストール、構成、最適化を自動化します。EDB Postgres Advanced Server を使用した Oracle との互換性、 EDB Postgres Extended または Advanced Server を使用した透過的データ暗号化 TDE などの追加価値を提供します。

TPA を使用して PGD クラスターを展開する方法については、TPA documentation を参照してください。

### 5.1.3 Deploying and configuring PGD on Kubernetes

EDB CloudNativePG Global Cluster は、EDB によって設計、開発、およびサポートされている Kubernetes オペレーターです。PGD レプリケーションを使用した、マルチマスターアーキテクチャと高可用性 Postgres データベースクラスターのライフサイクル全体をカバーします。これは、オープンソースの CloudNativePG オペレーターをベースにしており、EDB Postgres Advanced Server を使用した Oracle との互換性、 EDB

Postgres Extended または Advanced Server を使用した透過的データ暗号化 TDE、IBM Power や OpenShift を含む追加のサポート対象プラットフォームなどの追加価値を提供します。

Kubernetes の EDB Postgres 分散を使用して PGD クラスターを展開するには、次のリソースを参照してください。

- EDB CloudNativePG グローバルクラスタードキュメントの Quick start は、次のために必要な手順を介して機能します。
-

**Installation and upgrade は、EDB CloudNativePG Global**
Cluster のインストールとアップグレードのための詳細な手順を提供します。

### 5.1.4 Deploying and configuring PGD on Hybrid Manager

EDB Postgres AI Hybrid ManagerHM は、 EDB Postgres AI プラットフォームの集中コントロールプレーンであり、 EDB Postgres Advanced Server を介した Oracle 互換性や EDB Postgres Extended または Advanced Server を使用した透過的データ暗号化 TDE などの追加価値を提供します。Advanced High Availability AHA と分散高可用性 DHA の 2 つのプライマリ構成を介して、PGD 搭載クラスターのライフサイクル全体を展開および管理できます。

これらの構成の主な違いは、スケーラビリティとグループアーキテクチャにあります。

- **分散高可用性 DHA** グローバル規模と地理的な冗長性のために設計されたこの構成は、マルチプルのデータグループをサポートします。これにより、さまざまな場所またはリージョンにワークロードを分散できます。
- **高度な高可用性 AHA** 局所的な復元力のために設計されたこの構成は、単一のデータグループで構成されています。単一のリージョンまたはデータセンター内の高可用性を実現するのに最適です。

リソース要件や推奨されるユースケースを含む、これらのアーキテクチャの詳細な比較は、クラスター作成プロセス中に Hybrid Manager コンソール内で提供されます。

これらの高可用性構成を展開するための詳細な手順については、

Creating a database cluster を参照してください。

## 5.2 Upgrading

PGD と Postgres は密接に関連していますが、これらは別のアップグレードパスを持つ別の製品です。このセクションでは、PGD と Postgres の両方をアップグレードする方法について説明します。

### 5.2.1 PGD のアップグレード

EDB Postgres Distributed は、柔軟なプラットフォームです。これは、アップグレードパスが PGD のインストール方法に大きく依存することを意味します。

- ** *Upgrading PGD clusters manually* ** —PGD クラスターを手動でインストールして構成した場合、マイナーとメジャーの両方のバージョン間でクラスターを移動できます。

- ** *Supported PGD upgrade paths* ** —サポートされているいくつかのアップグレードパスを使用できます。
- ** *Compatibility changes* ** — PGD 3.x または 4.x から PGD 5.x 以降にアップグレードする場合、バージョン間の互換性の変更を理解する必要があります。

### 5.2.2 Postgres または Postgres と PGD メジャーバージョンのアップグレード

- **

**In-place Postgres major version upgrades ** — `pgd node upgrade`**
を使用して、1 つ以上のノードで Postgres バージョンまたは Postgres と PGD メジャーバージョンを手動でアップグレードする方法。

- ** *Performing a Postgres major version rolling upgrade on a PGD cluster* ** —クラスターで Postgres のメジャーバージョンのアップグレードを実行する方法。

### 5.2.3 その他のアップグレード

- ** *Application schema upgrades* ** —PGD で複数の分散サーバーを実行している場合、アプリケーションのスキーマを安全にアップグレードするためのガイド。

#### Upgrading PGD clusters manually

EDB Postgres Distributed は複数のソフトウェアコンポーネントで構成されているため、アップグレード戦略はアップグレードされるコンポーネントに部分的に依存します。

通常、**ローリングアップグレード**と呼ばれるアプローチを使用して、ほぼゼロのダウンタイムでクラスターをアップグレードできます。このアプローチを使用して、ノードは 1 つずつアップグレードされ、アプリケーション接続は既にアップグレードされたノードにスイッチオーバーされます。

すべてのノードを停止し、すべてのノードでアップグレードを実行して、クラスター全体を再起動することもできます。このアプローチは、標準の PostgreSQL セットアップと同じです。すべてのノードを同時にアップグレードするこの戦略は、ソフトウェアの混合バージョンでの実行を回避するため、最も簡単です。ただし、ダウンタイムが発生するため、何らかの理由でローリングアップグレードを実行できない場合を除き、お勧めしません。

EDB Postgres 分散クラスターをアップグレードするには

1. アップグレードを計画します。
2. アップグレードの準備をします。
3. サーバーソフトウェアをアップグレードします。
4. アップグレードを確認および検証します。

### アップグレード計画

各ノードをアップグレードするには、主に 2 つの方法があります。

- ノードを新しいソフトウェアバージョンにアップグレードします。**ローリングサーバーソフトウェアのアップグレード** を参照してください。
- ノードを、新しいバージョンがインストールされたものに置き換えます。**ノード参加を使用したローリングアップグレード** を参照してください。

これらのアプローチは両方ともローリング方法で使用できます。

### ローリングアップグレードの考慮事項

クラスターでローリングアップグレードが実行されている間、ソフトウェアの混合バージョンがクラスターで実行されています。たとえば、ノード A の PGD 4.3.6 があり、ノード B およびノード C が 5.6.1 であるとします。この状態では、レプリケーションとグループ管理は最も古いバージョンこの例では 4.3.6 のプロトコルと機能を使用するため、新しいバージョンが提供するプロトコルの変更を必要とする新機能は無効になっています。すべてのノードが同じバージョンにアップグレードされると、新機能が有効になります。

同様に、WAL デコーダー対応ノードを含むクラスターがローリングアップグレードを実行している場合、高いバージョンの PGD ノードの WAL デコーダーは、より高い pg 論理バージョンで *logical change records (LCRs)* を生成します。PGD ノードの下位バージョンの WAL デコーダは、下位の pg 論理バージョンの LCR を生成します。その結果、プロトコルバージョンの不一致のため、PGD ノードの上位バージョンの WAL 送信者は LCR を使用することは想定されていません。PGD ノードの下のバージョンでは、WAL 送信者は引き続き LCR を使用できます。すべての PGD ノードが同じ PGD バージョンになると、WAL 送信者は LCR を使用します。

ローリングアップグレードは、以前のリリースのすべてのノードを含むクラスターで開始されます。次に、すべてのノードが新しいリリースになるまで、一度に 1 つのノードを新しいリリースにアップグレードして続行します。同時に実行されるソフトウェアのバージョンは 3 つまでです。別のアップグレードを開始する前に、すべてのノードを完全にアップグレードした状態でアップグレードを完了する必要があります。

ビジネスリスクを削減するために追加の注意が必要な場合、アップグレードの実行により多くの時間が必要になる場合があります。最大限の注意を払い、実稼働システムのアップグレードに必要な時間を削減するために、最初に別のテスト環境でアップグレードを実行することをお勧めします。

アップグレードを完了するために絶対に必要な時間を超えて、ソフトウェアの混合バージョンで実行しないでください。*pgd commit-scopes list* コマンドを使用して、クラスター内のバージョンを確認できます。

混合バージョンでの実行が長くなると、問題が発生する可能性が高くなり、それらの診断と解決が難しくなります。ビジネスのオフピーク時間に、短期間にアップグレードすることをお勧めします。

ローリングアップグレードを使用してソフトウェアのメジャーバージョンをアップグレードすることはできますが、1 つのクラスターでの PostgreSQL、EDB Postgres Extended、および EDB Postgres Advanced Server の混合はサポートされていません。したがって、このアプローチを使用して Postgres バリアントを変更することはできません。

**警告**

EDB Postgres Distributed のダウングレードはサポートされていません。クラスターを手動で再構築する必要があります。

### ローリングサーバーソフトウェアのアップグレード

ローリングアップグレードとは、クラスターを停止せずにクラスター内の各ノードで **ローリングサーバーソフトウェアのアップグレード** がシークエンシャルにアップグレードされる場合です。各ノードは一時的にクラスターへの参加を停止され、サーバーソフトウェアがアップグレードされます。更新されると、クラスターに返され、不在中のクラスターのアクティビティに追いつきます。

実際の手順は、Postgres コンポーネントが新しいメジャーバージョンにアップグレードされるかどうかによって異なります。

アップグレードプロセス中に、現在アップグレードされていないノードにアプリケーションをスイッチオーバーして、アプリケーションのデータベースの継続的な可用性を提供できます。

### ノード参加を使用したローリングアップグレード

サーバーソフトウェアをアップグレードするもう 1 つの方法は、新しいノードをクラスターに参加させ、後で古いバージョンのソフトウェアを実行している既存のノードのいずれかをドロップすることです。

このアプローチの場合、手順は常に同じです。ただし、ノード参加が含まれるため、大規模なデータ転送が必要になる可能性があります。

すべてのノードが Postgres の新しい同じリリースにアップグレードされるまで、新しい Postgres バージョンでのみ使用可能な機能を使用しないように注意してください。これは、Postgres の新しいリリースに追加された新しい DDL 構文に特に当てはまります。

**注釈**

*bdr_init_physical* は、ソースノードのバイトごとのコピーを作成するため、メジャー Postgres バージョンから別のバージョンへのアップグレード中に使用できません。実際、現在 `bdr_init_physical` では、ソースの PGD バージョンと参加ノードでもまったく同じであることが必要です。新しいノードに参加する方法によるローリングアップグレードには使用できません。代わりに、論理結合を使用します。

### CAMO 対応クラスターのアップグレード

CAMO 対応クラスターをアップグレードするには、CAMO グループを 1 つずつアップグレードする必要がありますが、アップグレードするグループの CAMO 保護を無効にし、新しい *Commit scopes* ベースの設定を使用して再構成する必要があります。

CAMO ペアを構成する 2 つの BDR ノードを PGD 5.0 にアップグレードするには、次のアプローチをお勧めします。

1. 2 つのノードのいずれか上のトランザクションで `bdr.enable_camo` が `off` のままであることを確認するか、クライアントを 2 つのノードからリダイレクトします。CAMO を使用しようとするときに CAMO ペアリングを削除すると、エラーが発生し、それ以上のトランザクションが防止されます。
2. BDR 4.x の場合、`bdr.remove_camo_pair` を使用して CAMO の構成を解除し、ペアを解除します。

3. 2 つのノードを PGD 5.0 にアップグレードします。
4. 2 つのノードの専用ノードグループを作成し、そのノードグループに移動します。
5. このノードグループ、および CAMO を使用するノードのペアの *Commit scopes* を作成します。
6. `default_commit_scope` を設定するか、セッションまたはトランザクションに `bdr.enable_camo` の代わりに `bdr.commit_scope` を明示的に設定するようにクライアントを変更することにより、CAMO 保護を再度アクティブにします。
7. 必要に応じて、クライアントが CAMO 保護ノードに再度接続できるようにします。

#### アップグレードの準備

ソフトウェアの各メジャーリリースには、以前のリリースとの互換性に影響を与える可能性のあるいくつかの変更が含まれています。これらは、Postgres 構成、展開スクリプト、および PGD を使用するアプリケーションに影響を与える可能性があります。アップグレードの前に、これらの変更を検討し、必要な調整を行うことをお勧めします。

release notes およびバージョン固有のアップグレードノートで記載されている個々の変更を参照してください。

#### サーバーソフトウェアのアップグレード

個々のノードに分散された EDB Postgres のアップグレードは、インプレースで発生します。BDR 拡張機能をアップグレードするときにバックアップと復元を行う必要はありません。

#### BDR 拡張機能のアップグレード

BDR 拡張機能のアップグレードプロセスは、次の高レベルの手順で構成されています。

#### ノードをフェンスする

アップグレードが完了するまで、アップグレードされるノードが書き込みリーダーにならないように、アップグレードを開始する前にノードをフェンスする必要があります。

#### Postgres を停止します

バイナリパッケージのアップグレード中に、通常は最初に実行中の Postgres サーバーを停止するのが最善です。そうすることで、アップグレード中に予期せず再起動が発生した場合に、混合バージョンがロードされなくなります。

#### アップグレードパッケージ

アップグレードの最初の手順は、新しいバージョンの BDR パッケージをインストールすることです。このインストールでは、新しいバイナリと拡張 SQL スクリプトの両方をインストールします。この手順はオペレーティングシステムに固有です。

### Postgres を開始する

パッケージがアップグレードされたら、Postgres インスタンスを起動できます。BDR 拡張機能は、新しいバイナリが拡張機能の古いバージョンを検出すると、起動時にアップグレードされます。

### ノードのフェンスを解除します

ノードのアップグレードが完了した後、ノードのフェンスを解除できます。

!!!注

PGD 製品バージョン 6.4 にアップグレードする前に、PGD 4 クラスターは PGD 4.4.1 を実行している必要があります。

PGD 4.4.1 から PGD 製品バージョン 6.4 にアップグレードするには、Harp プロキシから Connection Manager に移動する追加の手順が必要です。詳細については、実際の例 *Upgrade PGD 4 to PGD 製品バージョン 6.4* を参照してください。

PGD 製品バージョン 6.4 にアップグレードする前に、PGD 5 クラスターは PGD 5.9 を実行している必要があります。

PGD 5.9 から PGD 製品バージョン 6.4 にアップグレードするには、PGD プロキシから Connection Manager に移動する追加の手順が必要です。詳細については、実際の例 *Upgrade PGD 5 to PGD 製品バージョン 6.4* を参照してください。

Postgres のインプレースアップグレードのプロセスは、Postgres の新しいマイナーバージョンにアップグレードするか、Postgres の新しいメジャーバージョンにアップグレードするかによって異なります。

### マイナーバージョン Postgres のアップグレード

Postgres の新しいマイナーバージョンへのアップグレードは、*BDR 拡張機能のアップグレード* に似ています。通常、Postgres の停止、パッケージのアップグレード、および Postgres の再起動だけが必要なすべてです。

ただし、特定のマイナーバージョンのアップグレードには、インデックスの再作成などのより多くの手順が推奨される場合があります。アップグレード先の Postgres のバージョンのリリースノートを参照してください。

### メジャーバージョン Postgres のアップグレード

Postgres の新しいメジャーバージョンへのアップグレードは、マイナーバージョンへのアップグレードよりも複雑です。

EDB Postgres Distributed は、 `pgd node upgrade` コマンドラインユーティリティを提供し、

in-place Postgres major version upgrades を実行するために使用できます。

> **注釈**
>
> Postgres、 BDR 拡張機能、その他を含むソフトウェアの新しいメジャーバージョンにアップグレードする場合、アプリケーションがアップグレードするソフトウェアのターゲットバージョンと互換性があることを確認することが常に重要です。

### アップグレードのチェックと検証

PGD ノードをアップグレードした後、バイナリの現在のバージョンを確認できます。

```
SELECT bdr.bdr_version();
```

ノードをアップグレードした後は常に *Monitoring the Connection Manager* をチェックして、アップグレードされたノードが予想通りに動作していることを確認します。

### PGD 5- PGD プロキシから接続マネージャーへの移動

次の手順を使用して、PGD プロキシから接続マネージャーに移動します。

1. PGD 5.9 ノードのいずれかから次のクエリーを実行して、すべてのユーザーパスワードの SCRAM ハッシュがすべてのノードで同じであることを確認します。

- ```
  DO $$
  DECLARE
      rec RECORD;
      command TEXT;    password TEXT;
  BEGIN
      FOR rec IN SELECT rolname,rolpassword FROM pg_authid WHERE rolcanlogin =␣
  ↪true AND rolpassword like SCRAM-SHA%
      LOOP
          password := rec.rolpassword;
          command := ALTER ROLE  || quote_ident(rec.rolname) ||  WITH ENCRYPTED␣
  ↪PASSWORD  || quote_literal(password);
          EXECUTE command;
      END LOOP;
  END;
  $$;
  SELECT wait_slot_confirm_lsn(NULL, NULL);
  ```

- !!!注

**これを実行した後、新しいユーザーは 5.9 に追加されません。追加された場合は、クエリを再度実行します。上記のブロックはパスワードを変更しません。すべてのノードのクラスターで SCRAM ハッシュが同じであることを保証するだけです。 .. ::**

2. 書き込みリーダーにならないように、*pgd node <node-name> set-option route_fence true* を使用してこのクラスター内のノードをフェンスします。 3. GUC *bdr.enable_builtin_connection_manager* を*true*に有効にします。 4. サーバーを再起動します。 5. サーバーで実行されている PGD プロキシを停止します。 6. サーバーを再起動します。デフォルトのポートで実行されている接続マネージャーで起動します。プロキシの読み取りポートと書き込みポートが異なる場合、Connection Manager ポートの読み取りポートと書き込みポートは、 *bdr.alter_node_group_option()* によってプロキシと同じになるように変更できます。 7. ノードのフェンスを解除します。 このノードは、ユーザーからの接続を受け入れ、接続マネージャーを介して書き込みリーダーにルーティングできるようになりました。 8. クラスター内のノードごとにこれを繰り返します。 これにより、すべてのノードが接続マネージャーを介してルーティングされるようになります。 9. PGD
  ```

5 から PGD 製品バージョン 6.4 へのメジャーバージョンのアップグレードを実行する場合も、*rolling upgrade steps* に進みます。

### Supported PGD upgrade paths

#### バージョン製品バージョン 6.4 へのアップグレード

EDB Postgres Distributed は semantic versioning を使用します。同じメジャーバージョン内のすべての変更には下位互換性があるため、アップグレード時のリスクが低下し、アップグレードターゲットとして新しいマイナーまたはパッチリリースを選択できます。

次のアップグレードパスは、PGD 製品バージョン 6.4 でサポートされています。

| Source version | Target version |
|---|---|
| 4.4.1 | 製品バージョン 6.4 |
| 5.9.1 | 製品バージョン 6.4 |
| 6.0.2 | 製品バージョン 6.4 |

### Compatibility changes

#### 概要

PGD 製品バージョン 6.4 には、5.x およびその他の以前のバージョンと互換性のない変更がいくつかあります。4.x のユーザーの場合、version 5.x compatibility の変更は引き続き適用されます。

#### 接続マネージャー

PGD 6 では、 PGD 5 のプロキシソリューションを、バックグラウンドワーカーを使用して緊密に統合されたアプローチで置き換える新しい接続マネージャーを導入し、PGD の読み書き、読み取り専用、および http ステータスネットワークインターフェイスを公開します。これは、管理する個別の PGD-プロキシインスタンスがなくなり、接続マネージャーがグループ構成を使用するため、古いプロキシ機能がすべて削除されたことを意味します。

さらに、 `bdr.proxy_conifg` テーブルと `bdr.proxy_config_summary` ビューが削除されました。接続マネージャーの構成に `bdr.node_group_config_summary` を使用する必要があります。これには、ルーティングを含む他のすべてのグループ設定も含まれます。

詳細については、 *Connection Manager Authentication* を参照してください。

#### DDL ロックの変更

新しいグローバルロックタイプ `ddl_locking = leader` があり、グローバル DML ロックの代わりにデフォルトで PGD 製品バージョン 6.4 によって使用されます。このロックにより、テーブルがすべてのグループリーダーですべてのノードではなくで排他的にロックされます。PGD Essential の場合、これは、リーダーが 1 つだけであるため、実際にはロックがすぐに従事されることを意味します。これは、システムが単一ノードの Postgres セットアップと同じ動作を示すことを意味します。`LOCK TABLE` および `TRUNCATE TABLE` は、デフォルトでこの新しいロックタイプを使用するようになりましたデフォルトの動作がオリジンノードの外部でロックしない前に。

LOCK TABLE は、ACCESS SHARE、ROW SHARE、ROW EXCLUSIVE、SHARE UPDATE EXCLUSIVE に DDL ロックを使用し、SHARE ROW EXCLUSIVE、EXCLUSIVE、および ACCESS EXCLUSIVE に DML ロックを使用するようになりました。これは、SHARE+ は DML をブロックしますが、下位レベルはブロックしないことを意味しますこれは、Postgres でこれらのレベルが動作する方法とより密接に連携しています。LOCK TABLE は、常に DML ロックを使用していました。

#### 結合/部分動作

#### ファンクションとノードの状態の変更

bdr.part_node の場合

- wait_for_completion はローカルノードで動作します
- force オプションは非推奨であり、bdr.drop_node(force:=true) と同じ動作をします。

bdr.drop_node の場合

- cascade オプションは削除されました
- force は、常に即時にノードをローカルにドロップし、以前のカスケード動作も想定するようになりました

bdr.join_node_group の場合

- pause_in_standby オプションは削除されました
- ノードがスタンバイかどうかは、 bdr.create_node の実行時に node_kind によって決定されます

新しいノード状態 PART_CLEANUP が追加されました。これは、グループスロットが分離ノードに関連するレコードを通過するのを待機し、ノードに関連する元のレコードをクリーンアップします。

#### シーケンス変換

PGD クラスターを作成するときに、既存の *Sequences* は分散されたものに自動的に変換されます。

#### ノードを分割するとドロップします

ノードが分離されると、最終的にドロップされ、分離の最後にノードは PARTED 状態では表示されません。したがって、ノードが PARTED 状態に到達するのを待機するスクリプトは変更が必要になる場合があります。また、このオペレーションは、一貫性を保証するために分割ノードからの最後のトランザクションのすべてのクロスのグループスロットまで待機するため、以前のバージョンよりも時間がかかる場合があります。

#### 管理機能の変更

- enable_proxy_routing ノードグループオプションは enable_routing になりましたが、bdr.alter_node_group_option は下位互換性のため古い構文をサポートしています。ビューには、新しい名前のみが表示されます。
- すでに非推奨である bdr.alter_node_group_config() は削除されました。代わりに bdr.alter_node_group_option() を使用してください。

- bdr.autopartition() の managed_locally オプションは削除されました。ローカルに管理されるパーティションのみがサポートされるようになりました。同時に ap_is_local は bdr.autopartition_rules カタログから削除されました。

#### その他

#### レプリカ ID

- レプリカ ID はデフォルトで FULL になりました。
- これにより、ノードのクラッシュと復旧にわたる挿入、更新、および削除間の競合検出におけるいくつかのエッジケースが削除されます。

#### すべての書き込みで実行

- Run on (all) write はレプリケートされなくなりました。
- bdr.xact_replicate はデフォルトで off に設定され、run on (all) nodes コマンドを介して送信されるコマンドのレプリケーションでの偶発的な問題の可能性が減少します。
- 一部のログメッセージを理解しやすいように言い換えました。
- 一部の内部ログメッセージのログレベルが LOG から DEBUG に低下したため、ログに表示されない場合があります。

#### 一般的な UI の改善

- ’pglogical’を含まないように名前変更されたワーカー
- 以前に非推奨だった構成パラメーター GUC が削除されましたこれには、たとえば、pg 論理的なプレフィックスが付いているすべての GUC が含まれます。
- デフォルトでは、競合は Postgres ログまたは conflict_history テーブルのいずれにもログに記録されません

#### デフォルトの変更

#### グローバルロックタイムアウト

bdr.global_lock_timeout は、デフォルトで 1 分に設定されています。

#### 自動同期

bdr.enable_auto_sync_reconcile はデフォルトで ON に設定されています。この機能は 自動同期との関係で文書化されています。これにより、ダウンノードからの変更を他のすべてのノードと同期できます。

### Application schema upgrades

EDB Postgres Distributed のアップグレードと同様に、アプリケーションスキーマをアップグレードするには 2 つのアプローチがあります。より簡単なオプションは、影響を受けるすべてのアプリケーションを停止し、スキーマのアップグレードを実行し、新しいスキーマバリアントを使用するようにアップグレード

されたアプリケーションを再起動することです。このアプローチでは、ある程度のダウンタイムが発生します。

このダウンタイムを排除するために、 EDB Postgres Distributed は、ローリングアプリケーションスキーマのアップグレードを実行するための便利なツールを提供します。

次の推奨事項とヒントは、クラスターに対するアプリケーションスキーマのアップグレードの影響を軽減します。

### ローリングアプリケーションスキーマのアップグレード

デフォルトでは、 DDL はすべてのノードに自動的に送信されます。*Running DDL safely* で説明しているように、この動作は手動で制御できます。このアプローチを使用して、ノード全体にデータベーススキーマ間の差異を作成できます。

PGD は、ノード間のわずかな違いがあってもレプリケーションを続行できるように設計されています。これらの機能は、ダウンタイムなしでアプリケーションスキーマを移行できるように、またはロジカルスタンバイノードでレポートまたはテストをできるように設計されています。

これを実稼働クラスターで正しく動作させるには、注意的なスクリプトが必要です。広範なテストをお勧めします。

詳細は、 差分を含むノード間のレプリケーション を参照してください。

1 つのノードが新しいテーブルを追加する DDL を実行すると、最新の DDL をまだ受け取っていないノードは追加のテーブルを処理する必要があります。これを考慮して、ローリングスキーマアップグレードの適切な設定は、`target_table_missing` 競合が発生した場合に `skip` リゾルバーを適用するようにすべてのノードを構成することです。ノードにテーブルを追加する前に、この構成を実行します。この設定は永続的なものです。

次のクエリーを **各ノードで個別に** 実行します。 `node1` を実際のノード名に置き換えます。

```
SELECT bdr.alter_node_set_conflict_resolver(node1,
       target_table_missing, skip);
```

1 つのノードがテーブルに列を追加する DDL を実行すると、最新の DDL をまだ受け取っていないノードが追加の列を処理する必要があります。これを考慮して、ローリングスキーマアップグレードの適切な設定は、`target_column_missing` 競合が発生した場合に `ignore` リゾルバーを適用するようにすべてのノードを構成することです。1 つのノードに列を追加する前にこれを実行します。この設定は永続的なものです。

次のクエリーを **各ノードで個別に** 実行します。 `node1` を実際のノード名に置き換えます。

```
SELECT bdr.alter_node_set_conflict_resolver(node1,
       target_column_missing, ignore);
```

1 つのノードがテーブルから列を削除する DDL を実行すると、最新の DDL をまだ受け取っていないノードが不足している列を処理する必要があります。この状況により、`use_default_value` リゾルバーを使用する `source_column_missing` 競合が発生します。したがって、NULL を受け入れず、DEFAULT 値を持たない列には、2 段階のプロセスが必要です。

1. NOT NULL 制約を削除するか、すべてのノードの列の DEFAULT 値を追加します。

2. カラムを取り外します。

コンストレインは、ローリング方法で削除できます。現在、一度に 1 ノードずつ、ローリング方法でテーブルコンストレインの追加を処理する方法はありません。

1 つのノードが既存の列の型を変更する DDL を実行すると、現在の型とターゲット型間のバイナリ強制性の存在によっては、オペレーションで基になるテーブルデータが書き換えられない場合があります。その場合、それは基になる列タイプのメタデータの更新だけです。テーブルの書き換えは通常制限されています。ただし、制御された DBA 環境では、すべてのノードの非レプリケート環境でこの列のタイプのローリングアップグレードを採用することにより、列のタイプを自動的にキャスト可能なものに変更できます。詳細は、*ALTER TABLE* を参照してください。

### In-place Postgres or Postgres and PGD major version upgrades

コマンドラインユーティリティ pgd node upgrade を使用して、PGD ノードを Postgres の新しいメジャーバージョン、または Postgres と PGD のメジャーバージョンにアップグレードできます。

!!!注 5.7.0 より前のバージョンでは、インプレースメジャーバージョンのアップグレードに使用されるコマンドは bdr_pg_upgrade でした。ただし、このコマンドには、 pgd node upgrade のように、Postgres メジャーバージョンと PGD バージョンの両方を同時にアップグレードするオプションがありません。

pgd node upgrade は、標準の pg_upgrade のラッパーで、PGD 固有のロジックをプロセスに追加して、スムーズなアップグレードを保証します。

#### 用語

この用語は、アップグレードプロセスと関連するコンポーネントを説明するときに使用されます。

*Postgres cluster* _ データベースファイル実行可能ファイルとデータの両方。実行時にシステム上の Postgres データベースインスタンスを構成します。

*Old Postgres cluster* _ アップグレードする既存の Postgres クラスター、データを移行する既存の Postgres クラスター。

*New Postgres cluster* _ データの移行先の新しい Postgres クラスター。この Postgres クラスターは、古いクラスターの 1 つ前のメジャーバージョンである必要があります。

#### 注意事項

両方の Postgres クラスターが pg_upgrade のすべての要件を満たしている必要があるという事実を含む、標準の Postgres メジャーバージョンのアップグレードの注意事項が適用されます。

さらに、他のツールがレプリケーションスロットとレプリケーション起点を使用している場合、pgd node upgrade を使用しないでください。アップグレード後には、PGD スロットとオリジンのみが復元されます。

pgd node upgrade のいくつかの前提条件を満たす必要があります。

- 古い Postgres クラスターを使用してアプリケーションを切断します。たとえば、PGD クラスター内の別のノードにリダイレクトできます。
- 両方の Postgres クラスターのピア認証を構成します。 bdr_pg_upgrade はピア認証が必要です。

- 新しいバージョンの PGD と同じまたは新しいバージョンを新しいクラスターにインストールする必要があります。
- PGD バージョンは 4.1.0 以降である必要があります。
- 新しいクラスターはシャットダウン状態である必要があります。
- 新しいクラスターに PGD パッケージをインストールする必要があります。
- 新しいクラスターは、古いクラスター構成と一致するように、必要に応じて既に初期化および構成されている必要があります。
- データベース、テーブル、およびその他のオブジェクトは、新しいクラスターに存在してはなりません。

**注釈**

PGD 5.7.0+にアップグレードする場合、両方のクラスターで同じ PGD バージョンを実行する必要はありません。新しいクラスターは 5.7.0 以降で実行されている必要があります。その場合、*pgd node upgrade* は PGD バージョンを 5.7.x にアップグレードし、Postgres メジャーバージョンをアップグレードします。

`pgd node upgrade` を実行する前に、古い Postgres クラスターを起動することをお勧めします。CLI は、古い Postgres クラスターがシャットダウンしている場合、それを起動します。

### 使用法

Postgres または Postgres と PGD の新しいメジャーバージョンにアップグレードするには、最初に新しいバージョンパッケージをインストールする必要があります。

### `pgd node upgrade` コマンドライン

`pgd node upgrade` は、すべてのパラメーターを pg_upgrade に渡します。したがって、pg_upgrade でサポートされているパラメーターを指定できます。

### 概要

```
pgd node <NODE_NAME> upgrade [OPTION] ...
```

### オプション

pg_upgrade のオプションに加えて、次のパラメーターを `pgd node upgrade` に渡すことができます。

### 必須パラメーター

これらのパラメーターは、コマンドラインで指定するか、 `--database` パラメーターを除くすべての場合、同等の環境変数で指定します。これらは `pgd node upgrade` によって使用されます。

- `-b, --old-bindir` _ Old Postgres クラスター bin ディレクトリ。
- `-B, --new-bindir` _ 新しい Postgres クラスターの bin ディレクトリ。

- -d, --old-datadir _ Old Postgres クラスターデータディレクトリ。
- -D, --new-datadir _ 新しい Postgres クラスターデータディレクトリ。
- --database _ PGD データベース名。

### オプショナルパラメーター

これらのパラメーターはオプショナルであり、pgd node upgrade によって使用されます。

- -p, --old-port _ 古いクラスターポート番号。
- -s, --socketdir _ アップグレード中にポストマスターソケットに使用するディレクトリ。
- --check _ チェックのみを実行し、クラスターを変更しないように指定します。

### その他のパラメーター

上記のいずれでもない他のパラメーターは pg_upgrade に渡されます。pg_upgrade は次のパラメーターを受け入れます。

- -j, --jobs _ 同時に使用するプロセスまたはスレッドの数。
- -k, --link _ 新しいクラスターにファイルをコピーする代わりに、ハードリンクを使用します。
- -o, --old-options _ 古い postgres コマンドに渡すオプション。複数の呼び出しが追加されます。
- -O, --new-options _ 新しい postgres コマンドに渡すオプション。複数の呼び出しが追加されます。
- -N, --no-sync _ アップグレードされたクラスター内のすべてのファイルがディスクに書き込まれるのを待たないでください。
- -P, --new-port _ 新しいクラスターポート番号。
- -r, --retain _ 正常に完了した後でも SQL とログファイルを保持します。
- -U, --username _ クラスターのインストールユーザー名。
- --clone _ 効率的なファイルのクローン作成を使用します。

### 環境変数

コマンドラインパラメーターの代わりにこれらの環境変数を使用できます。

- PGBINOLD _ Old Postgres クラスター bin ディレクトリ。
- PGBINNEW _ 新しい Postgres クラスターの bin ディレクトリ。
- PGDATAOLD _ Old Postgres クラスターデータディレクトリ。
- PGDATANEW _ 新しい Postgres クラスターデータディレクトリ。
- PGPORTOLD _ 古い Postgres クラスターポート番号。
- PGSOCKETDIR _ アップグレード中にポストマスターソケットに使用するディレクトリ。

### 例

次のようなシナリオが考えられます。

- アップグレードするクラスターのノード名は kaolin です。
- 古い Postgres クラスターの bin ディレクトリは/usr/lib/postgresql/16/bin です。
- 新しい Postgres クラスターの bin ディレクトリは/usr/lib/postgresql/17/bin です。
- 古い Postgres クラスターデータディレクトリは/var/lib/postgresql/16/main です。
- 新しい Postgres クラスターデータディレクトリは/var/lib/postgresql/17/main です。
- データベース名は bdrdb です。

次のコマンドを使用して、クラスターをアップグレードできます。

```
pgd node kaolin upgrade \
- -old-bindir /usr/lib/postgresql/16/bin \
- -new-bindir /usr/lib/postgresql/17/bin \
- -old-datadir /var/lib/postgresql/16/main \
- -new-datadir /var/lib/postgresql/17/main \
- -database bdrdb
```

### 実行された手順

これらの手順は、pgd node upgrade の実行時に実行されます。

> **注釈**
>
> *--check* が `pgd node upgrade` への引数として指定される場合、 CLI はデータベースを変更する手順をスキップします。

### PGD Postgres チェック

| Steps | *--check* supplied |
|---|---|
| Collecting pre-upgrade new cluster control data | *run* |
| Checking new cluster state is shutdown | *run* |
| Checking PGD versions | *run* |
| Starting old cluster (if shutdown) | *skip* |
| Connecting to old cluster | *skip* |
| Checking if bdr schema exists | *skip* |
| Turning DDL replication off | *skip* |
| Terminating connections to database | *skip* |
| Waiting for all slots to be flushed | *skip* |
| Disconnecting from old cluster | *skip* |
| Stopping old cluster | *skip* |

次のページに続く

表 9 – 前のページからの続き

| Steps | *--check* supplied |
|---|---|
| Starting old cluster with PGD disabled | *skip* |
| Connecting to old cluster | *skip* |
| Collecting replication origins | *skip* |
| Collecting replication slots | *skip* |
| Disconnecting from old cluster | *skip* |
| Stopping old cluster | *skip* |

### pg_upgrade 手順

標準の pg_upgrade 手順が実行されます。

**注釈**

指定された場合、 *--check* は pg_upgrade に渡されます。

### PGD のアップグレード後の手順

| Steps | *--check* supplied |
|---|---|
| Collecting old cluster control data | *skip* |
| Collecting new cluster control data | *skip* |
| Advancing LSN of new cluster | *skip* |
| Starting new cluster with PGD disabled | *skip* |
| Connecting to new cluster | *skip* |
| Creating replication origin, repeated for each origin | *skip* |
| Advancing replication origin, repeated for each origin | *skip* |
| Creating replication slot, repeated for each slot | *skip* |
| Stopping new cluster | *skip* |

## Performing a Postgres major version rolling upgrade on a PGD cluster

### Postgres メジャーバージョンのアップグレード

Postgres データベースのメジャーバージョンをアップグレードして、改善された機能、パフォーマンス強化、およびセキュリティ更新にアクセスすることは、一般的な管理タスクです。EDB Postgres 分散 PGD クラスターに対して同じことを行うことは、基本的に同じプロセスですが、ローリングアップグレードとして実行されます。

ローリングアップグレードプロセスにより、クラスターの可用性と操作の継続性を維持しながら、個々のクラスターノードを新しいメジャー Postgres バージョンに更新できます。このアプローチでは、各ノードがシークエンシャルにアップグレードされるときに、クラスターの残りの部分を操作可能なままにすることにより、ダウンタイムを最小限に抑え、データの整合性を保証します。

次の一般的な手順と実際の例の概要は、スムーズで制御されたアップグレードプロセスを提供するのに役立ちます。

### アップグレードの準備

アップグレードの準備をするには、アップグレードしようとしているサブグループとノードを特定し、最初のアップグレード順序をメモします。

これを行うには、SSH を使用してノードのいずれかに接続し、`pgd nodes list` コマンドを実行します。

```
sudo -u postgres pgd nodes list
```

`pgd nodes list` コマンドは、PGD クラスター内のすべてのノードと、各ノードが属しているサブグループを表示します。次に、各サブグループの書き込みリーダーであるノードを見つけます。

```
sudo -u postgres pgd group <group_name> show --summary
```

このコマンドは、クラスターで実行されている<group_name> によってトークン化された pgd グループに関する情報を表示します。どのノードが書き込みリーダーであるかなど。操作の継続性を維持するには、書き込みリーダーをアップグレードする前に、サブグループ内の別のノードにスイッチする必要があります。計画的なスイッチオーバーの数を最小限に抑えるには、ノードのサブグループをアップグレードするときに、最後にライタリーダーをアップグレードします。

アップグレードが完了するまで、アップグレードされるノードが書き込みリーダーにならないように、アップグレードを開始する前にノードをフェンスし、ノードのアップグレードが完了した後にノードのフェンスを解除する必要があります。

計画のためにどのノードが現在の書き込みリーダーであるかを確認した場合でも、サブグループの書き込みリーダーは、そのノードをアップグレードする前に、操作上の理由でいつでも別のノードに変更される可能性があります。したがって、ノードをアップグレードする前に、ノードが書き込みリーダーでないことを確認する必要があります。

これで、各サブグループで最後に指定された書き込みリーダーノードをアップグレードすることを目的として、一度に 1 サブグループごとにアップグレード順序を決定するための十分な情報が得られました。

### 各ノードでアップグレードを実行する

**重要**

データの損失を防ぐため、アップグレードプロセスを開始する前に、データベースと構成ファイルがバックアップされていることを確認してください。

アップグレードの準備 を使用して、SSH 経由で接続しているときに各ノードで次の手順を実行します。

- **現在の Postgres バージョンの確認**
- PGD からバージョンを表示します。

`sudo -u postgres pgd nodes list --versions .`

- 予想されるメジャーバージョンが実行されていることを確認します。

- **ターゲットノードが書き込みリーダーでないことを確認します**
- ターゲットノードがアップグレードするグループの書き込みリーダーであるかどうかを確認します。

```
sudo -u postgres pgd group <group_name> show --summary
```

- ターゲットノードが、アップグレードしているグループ/サブグループの現在の書き込みリーダーである場合、別のノードへの計画的なスイッチオーバーを実行します。

```
sudo -u postgres pgd group <group_name> set-leader <new_leader_node_name>
```

- **ノードをフェンスする**
- アップグレードが完了するまで、アップグレードされるノードが書き込みリーダーにならないように、アップグレードを開始する前にノードをフェンスする必要があります。
- **ターゲットノードで Postgres を停止します**
- 現在のノードで Postgres サービスを停止します。

```
sudo systemctl stop postgres
```

ターゲットノードは、クラスター内のノードとしてアクティブに参加していません。

- **PGD とユーティリティのインストール**
- アップグレードする Postgres バージョンと互換性のある PGD とそのユーティリティをインストールします。

```
sudo apt install edb-pgd6-<edition>-pg<new_postgres_version_number>
```

- **新しい Postgres インスタンスを初期化します**
- 新しいバージョンの PostgreSQL のデータベースファイルを格納するディレクトリを作成します。

```
sudo mkdir -p /opt/postgres/datanew
```

- chown .
- を使用してユーザー postgres にディレクトリへの所有権権限があることを確認します。作成したディレクトリで新しい PostgreSQL データベースクラスターを初期化します。この手順には、PostgreSQL の新しくインストールされたバージョンが提供する initdb コマンドの使用が含まれます。--data-checksums フラグを含めて、クラスターがデータチェックサムを使用するようにします。

```
sudo -u postgres <path_to_postgres_bin>/initdb -D /opt/postgres/datanew
--data-checksums
```

<path_to_postgres_bin> を、新しくインストールされた PostgreSQL バージョンの bin ディレクトリへのパスに置き換えます。

このコマンドは、postgres ユーザーまたは適切な権限を持つ別のユーザーとして実行する必要がある場合があります。

- **新しい Postgres バージョンに構成を移行する**
- 現在の PostgreSQL データディレクトリで次の構成ファイルを見つけます。
- postgresql.conf _ データベースシステムに関連する設定を含むメイン構成ファイル。

- postgresql.auto.conf _ PostgreSQL によって設定された設定が含まれています、ALTER SYSTEM コマンドで変更したものなど。
- pg_hba.conf _ クライアント認証を管理し、どのユーザーがどのホストからどのデータベースに接続できるかを指定します。
- conf.d ディレクトリ全体存在する場合 _ 構成設定を個別のファイルに整理できます。 better manageability.
- これらのファイルと conf.d ディレクトリを、PostgreSQL のアップグレードバージョン用に作成した新しいデータディレクトリにコピーします。

既存の構成ファイルからコメントを外した行だけを新しい構成ファイルにコピーし、変更するたびに各設定のドキュメントをチェックして、追加の調整が必要でないことを確認する方が効率的であると感じる場合があります。

- **Postgres サービスが非アクティブであることを確認します**
- 続行する前に、古いデータディレクトリと新しいデータディレクトリの両方でアクティブな PostgreSQL プロセスがないことを確認することが重要です。この検証手順により、アップグレードプロセス中のデータの破損または競合が防止されます。

sudo systemctl status postgres コマンドを使用して、Postgres が停止したことを確認します。停止していない場合は、systemctl stop postgres を実行して、停止されたことを再度確認します。

- **バージョンアップグレードのために PGDATA ディレクトリをスワップします**
- /opt/postgres/data を/opt/postgres/dataold に、/opt/postgres/datanew を/opt/postgres/data に名前変更します。

この手順により、システムは次の重要なフェーズ、pgd ノードのアップグレードを実行して PostgreSQL バージョンの移行を完了するための準備が整います。

- **アップグレードの実現可能性の検証**
- pgd node upgrade ツールは、現在のセットアップの予備スキャンを実行し、アップグレードプロセスを妨げる可能性のある潜在的な問題を特定するように設計された--check オプションを提供します。

pgd node upgrade によって作成されたアップグレードログファイルを保存できるように、/home/upgrade/ などのユーザー postgres に所有権が付与されたアップグレードディレクトリからこのチェックを実行する必要があります。安全性チェックを開始するには、--check オプションを pgd node upgrade コマンドに追加します。

この操作は、変更を加えずにアップグレードプロセスをシミュレートし、互換性の問題、非推奨機能、またはアップグレードの成功に必要な構成調整に関する洞察を提供します。

- このチェックで示された警告またはエラーに対処して、新しいバージョンへの平穏な移行を保証します。
- **Postgres メジャーバージョンのアップグレードを実行します**
- --check オプションを指定せずに pgd node <node_name> upgrade コマンドを実行して、アップグレードプロセスを実行します。
- 注意が必要なエラーまたは警告のコマンド出力をモニタすることが重要です。

- アップグレードプロセスにかかる時間データベースのサイズとセットアップの複雑さによって異なります。

- **Postgres サービス構成の更新**

- postgres.service ファイルのバージョン番号を更新することにより、新しい PostgreSQL バージョンを反映するようにサービス構成を更新します。

```
sudo sed -i -e 's/<old_version_number>/<new_version_number>/g' /etc/systemd/system/
postgres.service
```

- システムのサービスマネージャーを更新して、これらの変更を適用します。

```
sudo systemctl daemon-reload
```

- **Postgres を再起動します**

- PostgreSQL サービスの再起動に進みます。

```
systemctl start postgres
```

- **新しい Postgres バージョンを検証する**

- PostgreSQL インスタンスがアップグレードされたことを確認します。

```
sudo -u postgres pgd nodes list --versions
```

- **ノードのフェンスを解除します**

- アップグレードを検証した後、ノードのフェンスを解除できます。

- **アップグレード後のクリーンアップ**

- アップグレードの直後、高い運用負荷を導入する前に、ANALYZE オプションを使用して vacuumdb を実行します。このコマンドを実行すると、即時のパフォーマンスへの影響が最小限に抑えられ、より正確なテストのためにデータベースを準備します。

- 古いバージョンのデータディレクトリ/opt/postgres/dataold を削除します。

### 実際の例 PGD 4 から PGD 6 へのアップグレード

この作業した例では、PGD 4 から PGD 製品バージョン 6.4 へのインプレースメジャーバージョンのローリングアップグレードについて説明します。

#### 概要

HARP プロキシベースのルーティングを使用する PGD 4 クラスターは、クラスター全体が製品バージョン 6.4.0 以降にアップグレードされるまで、すべてのノードにこのルーティング方法を継続します。HARP プロキシベースのルーティングは、混合バージョンクラスター内で同じように機能します。4.x クラスターには書き込みリーダーがないため、HARP は独自のメカニズムを使用してリーダーを選択します。クラスター全体がバージョン製品バージョン 6.4.0 以降にアップグレードされると、Connection Manager は既に有効になっており、準備が整います。アプリケーションは接続マネージャーポート/ノードを指すことができ、ルーティングを開始します。動作していることが確認されたら、プロキシを停止できます。

!!!注 この例では、 PGD アーキテクチャで推奨されているように、Harp プロキシが PGD ノードと同じ場所にないことを前提としています。PGD ノードと同じ場所に Harp プロキシがある場合、アップグレード手順については EDB サポートに連絡してください。

### harp-プロキシリーダーの確認

harp-プロキシリーダーではないノードでアップグレードを開始します。どのノードが harp プロキシリーダーであるかを確認します。

```
test-pgd6major-d1:~ $ harpctl get leader a
```

```
Cluster  Name                     Location Ready Fenced Allow Routing Routing Status␣
↪Role    Type Lock Duration
- ------  ----                      -------- ----- ------ ------------- --------------␣
↪----    ---- --------------
bdrgroup test-pgd6major-d2 a        true  false  true                        primary␣
↪bdr  6
```

### ノードをフェンスする

HARP からアップグレードするノードをフェンスオフし、フェンスされたことを確認して、アップグレードの途中でリーダーにならないようにします。

```
test-pgd6major-d1:~ $ harpctl fence test-pgd6major-d1
```

```
INFO    cmd/fence.go:42 fence node test-pgd6major-d1
```

```
test-pgd6major-d1:~ $ harpctl get nodes
Cluster  Name                     Location Ready Fenced Allow Routing Routing Status␣
↪Role    Type Lock Duration
- ------  ----                      -------- ----- ------ ------------- --------------␣
↪----    ---- --------------
bdrgroup test-pgd6major-d1 a        false true   true          N/A            primary␣
↪bdr  6
bdrgroup test-pgd6major-d2 a        true  false  true          ok             primary␣
↪bdr  6
```

### Postgres サービスを停止します

フェンスされたノードで、Postgres サービスを停止します。

### HARP マネージャーを停止します

フェンスされたノードで、 HARP マネージャーを停止します。

### パッケージの削除とインストール

フェンスされたノードで、 PGD 4.4 および cli パッケージを削除し、 PGD 製品バージョン 6.4 パッケージをインストールします。

### Postgres サービスを開始する

フェンスされたノードで、Postgres サービスを開始します。これにより、接続マネージャーが有効になっている状態で、PGD ローカルノードの PGD 製品バージョン 6.4 へのインプレースアップグレードが実行されます。

### HARP マネージャーをスタートする

フェンスされたノードで、 HARP マネージャーを起動します。

### ノードのフェンスを解除する

アップグレードされたノードのフェンシングを HARP から解除します。

```
test-pgd6major-d1:~ $ harpctl unfence test-pgd6major-d1
```

### すべてのノードに対して手順を繰り返します。

他のすべてのノードで同じ手順を繰り返します。

### クラスターバージョンの確認

bdr.group_raft_details を実行して、更新されたクラスターバージョンがバージョン 6001 であることを確認します。

### SCRAM ハッシュの確認

アップグレードノードから次のクエリーを実行して、SCRAM ハッシュが各ユーザーのすべてのノードで同じであることを確認します。これは、アプリケーションが接続マネージャーに切り替える前に必要です。

```
DO $$
DECLARE
    rec RECORD;
    command TEXT;
BEGIN
    FOR rec IN SELECT rolname,rolpassword FROM pg_authid WHERE rolcanlogin = true AND␣
↪rolpassword like SCRAM-SHA%
    LOOP
        command := ALTER ROLE  || quote_ident(rec.rolname) ||  WITH ENCRYPTED␣
↪PASSWORD  ||  || rec.rolpassword || ;
        EXECUTE command;
    END LOOP;
END;
```

(次のページに続く)

(前のページからの続き)

```
$$;
SELECT wait_slot_confirm_lsn(NULL,NULL);
```

### ルーティングを有効にする

グローバルまたはローカルルーティング要件に従って、ノードグループルーティングを有効にします。ローカルルーティングの場合はサブグループで有効にし、グローバルルーティングの場合はトップグループで有効にします。

```
bdrdb=# SELECT bdr.alter_node_group_option(node_group_name := bdrgroup,config_key :=␣
↪enable_routing, config_value := true::TEXT);
```

出力

```
 alter_node_group_option
- ------------------------
 (1 row)
```

### 接続マネージャーに切り替える

これで、アプリケーションを Connection Manager に安全に切り替えられるはずです。

### harp マネージャーとプロキシサービスを停止します。

これで、実行中の harp マネージャーとプロキシサービスを安全に停止できるはずです。

> **注釈**
>
> これで、PGD 4 から PGD 製品バージョン 6.4 へのインプレースメジャーバージョンのローリングアップグレードの例が完了しました。

### 実際の例 PGD 5 から PGD 6 へのアップグレード

この作業した例では、 `pgd node upgrade` コマンドを使用して、3 ノード PGD 5、EPAS 13 クラスターから PGD 製品バージョン 6.4、EPAS 17 クラスターへのインプレースメジャーバージョンローリングアップグレード PGD & Postgres について説明します。

### 前提条件

3 ノード PGD 5、EPAS 13 クラスターが稼働していることを確認します。これは、TPA 展開されたクラスターです。

```
pgd-a1:/home/rocky $ pgd nodes list
```

出力

```
Node Name  Group Name Node Kind Join State Node Status
- --------- ---------- --------- ---------- -----------
pgd-a1     group-a    data      ACTIVE     Up
pgd-a2     group-a    data      ACTIVE     Up
witness-a1 group-a    witness   ACTIVE     Up
```

#### 新しいサーバーと PGD のパッケージをインストールします

EPAS 17 のパッケージと、対応する PGD 6 パッケージがクラスター内のすべてのノードにインストールされていることを確認します。バイナリの競合を防ぐには、PGD 6 パッケージをインストールする前に、PGD 5 パッケージつまり edb-pgd5-cli および edb-bdr5-epas13 を削除する必要があります。以下のコマンドは、RHEL 8 プラットフォームで使用されました。特定のプラットフォームに適切なコマンドを使用します。

```
dnf remove edb-pgd5-cli
dnf install edb-as17-server edb-pgd6-essential-epas17 -y
```

#### アップグレード前の手順

#### バージョンチェック

クラスターの現在のバージョンを確認しますオプショナル。

```
pgd-a1:/home/rocky $ pgd nodes list --versions
```

出力

```
Node Name  BDR Version  Postgres Version
- --------- ------------ -----------------
pgd-a1     5.9.0        13.22.28
pgd-a2     5.9.0        13.22.28
witness-a1 5.9.0        13.22.28
```

#### 接続マネージャーに移動

PGD 5 は、ルーティングに PGD プロキシを使用します。 PGD 6 では、PGD プロキシは Connection Manager に置き換えられました。 PGD 5 から PGD 6 にアップグレードする場合、次の手順を使用して Connection Manager に移動します。 *PGD 5- PGD* プロキシから接続マネージャーへの移動 を参照してください。

#### 書き込みリーダーノードの検証

アップグレードするノードが書き込みリーダーノードではないことを確認します。

```
pgd-a1:/home/rocky $ pgd group group-a show --summary
```

出力

```
Group Property    Value
- ---------------- -------
Group Name        group-a
Parent Group Name dc-1
Group Type        data
Write Leader      pgd-a2
Commit Scope
```

現在の書き込みリーダーはノード pgd-a2 であるため、ノード pgd-a1 をアップグレードするのが適切です。

アップグレードするノードの場合、書き込みリーダーを別のノードに切り替えます。

必要に応じて、 pgd group set-leader コマンドを使用して書き込みリーダーを切り替えます。

```
witness-a1:/home/rocky $ /usr/edb/as17/bin/pgd group group-a set-leader pgd-a1
```

```
witness-a1:/home/rocky $ /usr/edb/as17/bin/pgd group group-a show --summary
```

出力

```
 Group Property    | Value
- ------------------+---------
 Group Name        | group-a
 Parent Group Name | dc-1
 Group Type        | data
 Write Leader      | pgd-a1
 Commit Scope      |
```

#### ノードをフェンスする

書き込みリーダーにならないように、pgd node <node-name> set-option route_fence true を使用してこのクラスター内のノードをフェンスします。

#### 新しい Postgres インスタンスを初期化する

initdb ユーティリティを実行して、新しいサーバーを初期化します。--data-checksums オプションが設定されていることを確認します。

```
/usr/edb/as17/bin/initdb -D /var/lib/edb/as17/data -E UTF8 --data-checksums
```

デフォルトのデータディレクトリを使用したくない場合は、新しいデータディレクトリを作成します。この例では、簡単にするためにデフォルトを使用します。

#### 構成を新しい Postgres バージョンに移行する

次のファイルとディレクトリ存在する場合は、PostgreSQL のアップグレードバージョン用に作成した新しいデータディレクトリにコピーします。

- postgresql.conf

- postgresql.auto.conf
- pg_hba.conf
- conf.d ディレクトリ

```
cp /opt/postgres/data/postgresql.conf /var/lib/edb/as17/data/
cp /opt/postgres/data/postgresql.auto.conf /var/lib/edb/as17/data/
cp /opt/postgres/data/pg_hba.conf /var/lib/edb/as17/data/
```

TPA 展開されたクラスターがある場合は、 conf.d ディレクトリもコピーします。

```
cp -r /opt/postgres/data/conf.d /var/lib/edb/as17/data/
```

たとえば、 operator_precedence_warning GUC が発生する場合がありますが、新しい構成では無視できます。

```
sudo systemctl stop postgres
sudo systemctl status postgres
```

!!!注

PostgreSQL インスタンスがサービスとして構成されているため、この例では systemctl コマンドが使用されています。セットアップが異なる場合は、pg_ctl ユーティリティの使用が必要になる場合があります。

実際のアップグレードを実行する前に、ドライランを実行して、互換性とアップグレードの実現可能性を確認します。pgd node upgrade ツールには、アップグレードプロセスの一部のドライランを実行する--check オプションがあります。このオプションを使用して、アップグレードをスムーズに行うことができます。--check オプションを指定して upgrade コマンドを実行します。

```
/usr/edb/as17/bin/pgd node pgd-a1 upgrade \
- -database bdrdb -B /usr/edb/as17/bin \
- -socketdir /tmp \
- -old-bindir /usr/edb/as13/bin \
- -old-datadir /opt/postgres/data \
- -new-datadir /var/lib/edb/as17/data \
- -username enterprisedb \
- -old-port 5444 \
- -new-port 5444 \
- -check
```

チェックが成功すると、次のような出力が返されます。

```
Performing BDR Postgres Checks
- ------------------------------
Getting old PG instance shared directory                    ok
Getting new PG instance shared directory                    ok
Collecting pre-upgrade new PG instance control data         ok
Checking new cluster state is shutdown                      ok
```

(次のページに続く)

(前のページからの続き)

```
Checking BDR extension versions                             ok
Checking Postgres versions                                  ok

Finished BDR pre-upgrade steps, calling pg_upgrade
- ------------------------------------------------

Performing Consistency Checks
- ----------------------------
Checking cluster versions                                   ok
Checking database user is the install user                  ok
Checking database connection settings                       ok
Checking for prepared transactions                          ok
Checking for contrib/isn with bigint-passing mismatch       ok
Checking data type usage                                    ok
Checking for user-defined encoding conversions              ok
Checking for user-defined postfix operators                 ok
Checking for incompatible polymorphic functions             ok
Checking for not-null constraint inconsistencies            ok
Checking for presence of required libraries                 ok
Checking database user is the install user                  ok
Checking for prepared transactions                          ok
Checking for new cluster tablespace directories             ok

 *Clusters are compatible*
```

アップグレードを実行します。

ドライランチェックに合格した場合、 --check オプションを指定せずにコマンドを実行して、アップグレードを実行できます。

```
/usr/edb/as17/bin/pgd node pgd-a1 upgrade \
- -database bdrdb -B /usr/edb/as17/bin \
- -socketdir /tmp \
- -old-bindir /usr/edb/as13/bin \
- -old-datadir /opt/postgres/data \
- -new-datadir /var/lib/edb/as17/data \
- -username enterprisedb \
- -old-port 5444 \
- -new-port 5444
```

アップグレードが成功すると、次のような出力が返されます。

```
Performing BDR Postgres Checks
- -----------------------------
Getting old PG instance shared directory                    ok
```

(次のページに続く)

(前のページからの続き)

```
Getting new PG instance shared directory                    ok
Collecting pre-upgrade new PG instance control data         ok
Checking new cluster state is shutdown                      ok
Checking BDR extension versions                             ok
Checking Postgres versions                                  ok

Collecting Pre-Upgrade BDR Information
- ------------------------------------
Collecting pre-upgrade old PG instance control data         ok
Connecting to the old PG instance                           ok
Checking for BDR extension                                  ok
Checking BDR node name                                      ok
Terminating connections to database                         ok
Waiting for all slots to be flushed                         ok
Disconnecting from old cluster PG instance                  ok
Stopping old PG instance                                    ok
Starting old PG instance with BDR disabled                  ok
Connecting to the old PG instance                           ok
Collecting replication origins                              ok
Collecting replication slots                                ok
Disconnecting from old cluster PG instance                  ok
Stopping old PG instance                                    ok

Finished BDR pre-upgrade steps, calling pg_upgrade
- ------------------------------------------------

Performing Consistency Checks
- ----------------------------
Checking cluster versions                                   ok
Checking database user is the install user                  ok
Checking database connection settings                       ok
Checking for prepared transactions                          ok
Checking for contrib/isn with bigint-passing mismatch       ok
Checking data type usage                                    ok
Checking for user-defined encoding conversions              ok
Checking for user-defined postfix operators                 ok
Checking for incompatible polymorphic functions             ok
Checking for not-null constraint inconsistencies            ok
Creating dump of global objects                             ok
Creating dump of database schemas
                                                            ok
Checking for presence of required libraries                 ok
Checking database user is the install user                  ok
Checking for prepared transactions                          ok
Checking for new cluster tablespace directories             ok
```

(次のページに続く)

(前のページからの続き)

```
If `pg_upgrade` fails after this point, you must re-initdb
the new cluster before continuing.

Performing Upgrade
- -----------------
Setting locale and encoding for new cluster                   ok
Analyzing all rows in the new cluster                         ok
Freezing all rows in the new cluster                          ok
Deleting files from new pg_xact                               ok
Copying old pg_xact to new server                             ok
Setting oldest XID for new cluster                            ok
Setting next transaction ID and epoch for new cluster         ok
Deleting files from new pg_multixact/offsets                  ok
Copying old pg_multixact/offsets to new server                ok
Deleting files from new pg_multixact/members                  ok
Copying old pg_multixact/members to new server                ok
Setting next multixact ID and offset for new cluster          ok
Resetting WAL archives                                        ok
Setting frozenxid and minmxid counters in new cluster         ok
Restoring global objects in the new cluster                   ok
Restoring database schemas in the new cluster
                                                              ok
Copying user relation files
                                                              ok
Setting next OID for new cluster                              ok
Sync data directory to disk                                   ok
Creating script to delete old cluster                         ok
Checking for extension updates                                notice

Your installation contains extensions that should be updated
with the ALTER EXTENSION command.  The file
    update_extensions.sql
when executed by psql by the database superuser will update
these extensions.

Upgrade Complete
- ---------------
Optimizer statistics are not transferred by pg_upgrade.
Once you start the new server, consider running:
    /usr/edb/as17/bin/vacuumdb -U enterprisedb --all --analyze-in-stages
Running this script will delete the old clusters data files:
    ./delete_old_cluster.sh

pg_upgrade complete, performing BDR post-upgrade steps
```

(次のページに続く)

(前のページからの続き)

```
- ----------------------------------------------------
Collecting post-upgrade old PG instance control data          ok
Collecting post-upgrade new PG instance control data          ok
Checking LSN of the new PG instance                           ok
Starting new PG instance with BDR disabled                    ok
Connecting to the new PG instance                             ok
Creating replication origin bdr_bdrdb_dc_1_pgd_a2             ok
Advancing replication origin bdr_bdrdb_dc_1_pgd_a2 to 0/3...  ok
Creating replication origin bdr_bdrdb_dc_1_witness_a1         ok
Advancing replication origin bdr_bdrdb_dc_1_witness_a1 to...  ok
Creating replication slot bdr_bdrdb_dc_1                      ok
Creating replication slot bdr_bdrdb_dc_1_witness_a1           ok
Creating replication slot bdr_bdrdb_dc_1_pgd_a2               ok
Stopping new PG instance                                      ok
```

**注釈**

ハードリンクに`--link` オプションを使用できます。このオプションは、両方のデータディレクトリが同じファイルシステム上にある場合に機能します。詳細については、PostgreSQL ドキュメントの pg_upgrade を参照してください。

### アップグレード後の手順

#### Postgres サービスファイルの更新

/etc/systemd/system/postgres.service にある PostgreSQL サービスファイルで、新しいサーバーのサーバーバージョン、データディレクトリ、およびバイナリディレクトリを更新します。

更新されたサービスファイルの例

```
[Unit]
Description=Postgres 17 (TPA)
After=syslog.target
After=network.target

[Service]
Type=simple
User=enterprisedb
Group=enterprisedb
OOMScoreAdjust=-1000
Environment=PG_OOM_ADJUST_VALUE=0
Environment=PGDATA=/var/lib/edb/as17/data
StandardOutput=syslog
ExecStart=/usr/edb/as17/bin/edb-postgres -D ${PGDATA} -c config_file=/var/lib/edb/
↪as17/data/postgresql.conf
```

(次のページに続く)

(前のページからの続き)

```
ExecStartPost=+/bin/bash -c echo 0xff > /proc/$MAINPID/coredump_filter
ExecReload=/bin/kill -HUP $MAINPID
KillMode=mixed
KillSignal=SIGINT
Restart=no
LimitCORE=infinity

[Install]
WantedBy=multi-user.target
```

#### postgres サービスを開始する

daemon-reload を実行し、Postgres サービスを開始します。

```
systemctl daemon-reload
systemctl start postgres
```

> **注釈**
>
> サーバーがサービスとして実行されていない場合は、サービスファイルの更新をスキップし、 pg_ctl ユーティリティを使用してサーバーを起動できます。

#### アップグレードされたクラスターバージョンの確認

次のコマンドを使用して、アップグレードされたクラスターバージョンを確認します。

```
pgd-a1:/home/rocky $ /usr/edb/as17/bin/pgd nodes list --versions
```

出力

```
 Node Name  | BDR Version          | Postgres Version
- -----------+----------------------+------------------
 pgd-a1     | PGD 製品バージョン 6.4.0 Essential | 17.6.0
 pgd-a2     | 5.9.0                | 13.22.28
 witness-a1 | 5.9.0                | 13.22.28
```

ノード **pgd-a1** の BDR バージョンは製品バージョン 6.4.0 にアップグレードされ、Postgres バージョンは 17.6.0 にアップグレードされました。

#### ノードのフェンスを解除する

`pgd node <node-name> set-option route_fence false` を使用してこのクラスター内のノードのフェンスを解除して、書き込みリーダーにならないようにします。

### 接続マネージャーが動作していることを確認する

アップグレードしたノードで接続マネージャーポートデフォルトでは 6444 を介してクエリーを実行します。

```
pgd-a1:/home/rocky $ psql "host=pgd-a1 port=6444 dbname=bdrdb user=enterprisedb " -c
↪"select node_name from bdr.local_node_summary;"
```

出力

```
 node_name
- ----------
 pgd-a2
(1 row)
```

### クリーンアップとバキューム分析

ベストプラクティスとして、この時点でデータベースにバキュームを実行します。アップグレードが実行されると、アップグレード後レポートに次のものが含まれていることに気づいたかもしれません。

```
Upgrade Complete
- ----------------
Optimizer statistics are not transferred by pg_upgrade.
Once you start the new server, consider running:
    /usr/edb/as17/bin/vacuumdb -U enterprisedb --all --analyze-in-stages
Running this script will delete the old clusters data files:
    ./delete_old_cluster.sh
```

これで、バキュームを実行できます。ターゲットノードで、次を実行します。

```
/usr/edb/as17/bin/vacuumdb -U enterprisedb --all --analyze-in-stages
```

このノードを元に戻す必要がない場合は、古いクラスターのデータファイルをクリーンアップすることもできます。

```
./delete_old_cluster.sh
```

### 残りのノードをアップグレードする

クラスター内のすべてのノードに対してこれらの手順を実行する必要があります。唯一の違いは、upgrade コマンドのノード名です。簡単なリファレンスのために、ノード **pgd-a2** および **witness-a1** のコマンドが提供されています。

ノード **pgd-a2**

```
/usr/edb/as17/bin/pgd node pgd-a2 upgrade \
- -database bdrdb -B /usr/edb/as17/bin \
- -socketdir /tmp \
- -old-bindir /usr/edb/as13/bin \
```

(次のページに続く)

(前のページからの続き)

```
- -old-datadir /opt/postgres/data \
- -new-datadir /var/lib/edb/as17/data \
- -username enterprisedb \
- -old-port 5444 \
- -new-port 5444
```

ノード **witness-a1**

```
/usr/edb/as17/bin/pgd node witness-a1 upgrade \
- -database bdrdb -B /usr/edb/as17/bin \
- -socketdir /tmp \
- -old-bindir /usr/edb/as13/bin \
- -old-datadir /opt/postgres/data \
- -new-datadir /var/lib/edb/as17/data \
- -username enterprisedb \
- -old-port 5444 \
- -new-port 5444
```

#### クラスターの最終的な状態を確認する

次のコマンドを使用して、ノードバージョンを確認します。

```
pgd-a2:/home/rocky $ /usr/edb/as17/bin/pgd nodes list --versions
```

出力

```
 Node Name  | BDR Version          | Postgres Version
- ----------+----------------------+------------------
 pgd-a1     | PGD 製品バージョン 6.4.0 Essential  | 17.6.0
 pgd-a2     | PGD 製品バージョン 6.4.0 Essential  | 17.6.0
 witness-a1 | PGD 製品バージョン 6.4.0 Essential  | 17.6.0
```

クラスターのすべてのノードが PGD 製品バージョン 6.4.0 および EPAS 17.6.0 にアップグレードされました。

#### 接続マネージャーの検証

すべてのデータノードについて、次のコマンドを使用して接続マネージャーを確認します。

```
pgd-a2:/home/rocky $ psql "host=pgd-a2 port=6444 dbname=bdrdb user=enterprisedb " -c
↪"select node_name from bdr.local_node_summary;"
```

出力

```
 node_name
- ----------
```

(次のページに続く)

(前のページからの続き)

```
 pgd-a1
(1 row)
```

```
pgd-a2:/home/rocky $ psql "host=pgd-a2 port=6445 dbname=bdrdb user=enterprisedb " -c
↪"select node_name from bdr.local_node_summary;"
```

出力

```
 node_name
- ----------
 pgd-a2
(1 row)
```

**注釈**

これで、3 ノード PGD 5、EPAS 13 クラスターから PGD 6、EPAS 17 クラスターのインプレースメジャーバージョンローリングアップグレード PGD & Postgres の作業例が完了しました。

## 5.3 Testing and tuning PGD clusters

次のアプローチを使用して PGD アプリケーションをテストできます。

### 5.3.1 pgd_bench

Postgres ベンチマークアプリケーション pgbench は、新しいアプリケーション pgd_bench の形式で拡張されました。

pgd_bench は、PostgreSQL bin ディレクトリに追加される通常のコマンドラインユーティリティです。このユーティリティは、PostgreSQL pgbench ツールに基づいていますが、CAMO トランザクションと PGD 固有のワークロードのベンチマークをサポートしています。

pgd_bench の機能は pgbench 機能のスーパーセットですが、正常に動作するには BDR 拡張機能をインストールする必要があります。

主な違いは次のとおりです。

- 特定の場合にグローバルロックタイムアウトを防ぐための標準の pgbench シナリオでの初期化 `-i` フラグを調整。
- 標準シナリオの `VACUUM` コマンドは、すべてのノードで実行されます。
- pgd_bench リリースは BDR 拡張機能のリリースに関連付けられ、対応する Postgres ディストリビューションに対してビルドされます。この情報は、`--version` フラグの出力に反映されます。

現在のバージョンでは、CAMO または通常の PGD 展開を使用しながらフェールオーバーテストを実行できます。

次のオプションが追加されました。

```
- m, --mode=regular|camo|failover
mode in which pgbench should run (default: regular)
```

- -m camo または-m failover を使用して、pgd_bench のモードを指定します。 -m failover 仕様を使用して、通常の PGD 展開でフェイルオーバーをテストできます。

```
- -retry
retry transactions on failover
```

- --retry を使用して、 -m failover モードでフェイルオーバーが発生したときにトランザクションをリトライするかどうかを指定します。このオプションは-m camo モードではデフォルトで有効になっています。

これらのオプションに加えて、DSNform でフェイルオーバーのピアノードに関する接続情報を指定する必要があります。

次に、 CAMO 環境の例を示します。

```
pgd_bench -m camo -p $node1_port -h $node1_host bdrdemo \
    "host=$node2_host user=postgres port=$node2_port dbname=bdrdemo"
```

このコマンドは CAMO モードで実行されます。ノード 1 に接続し、テストを実行します。ノード 1 への接続が失われると、pgd_bench はノード 2 に接続します。ノード 2 を照会して、進行中のトランザクションのステータスを取得します。中止されたトランザクションおよび進行中のトランザクションは、CAMO モードで再試行されます。

フェイルオーバーモードで--retry を指定すると、進行中のトランザクションが再試行されます。このシナリオでは、進行中のトランザクションのステータスを見つける方法はありません。

### 5.3.2 pgd_bench の使用に関する注意事項

- カスタム init スクリプトを使用する場合、DDL コマンドの背後にある意味を理解することが重要です。通常、CREATE INDEX などの DDL 操作を続行する前に、セカンダリノードがデータのロード手順に追いつくのを待機することをお勧めします。後者は、データのロードが完了するまで取得できないグローバルロックを取得するため、タイムアウトになる可能性があります。

- PostgreSQL および/または BDR 拡張機能を含む考えられる拡張機能によって発行される NOTICE および WARNING メッセージなどのクライアントメッセージを抑制するために追加の手順は実行されません。client_min_messages 、bdr.camo_enable_client_warnings などの適切な変数を設定してそれらを抑制するのはあなたの責任です。

- pgd_bench は、カスタムスクリプトの SQL トランザクションを開始しません。SQL トランザクションで実行するスクリプトには、transaction start および end コマンドを含める必要があります。 -m /--mode オプションを camo に設定して pgd_bench が実行された場合、提供されるカスタムスクリプトは SQL コマンドをトランザクションでラップする必要があります。そうしないと、CAMO 機能が予想通りに動作しません。

### パフォーマンスのテストとチューニング

PGD を使用すると、マルチプルのノードに書き込みトランザクションを発行できます。これらの書き込みを各ノードに戻すと、パフォーマンスのコストが発生します。

第一に、別のノードからの変更の再生には、CPU コストと I/O コストが発生し、WAL レコードが生成されます。SQL を再実行する必要がないため、CPU オーバーヘッドが低いため、通常、リソースの使用量は元のトランザクションよりも少なくなります。UPDATE および DELETE トランザクションの場合、データがキャッシュされていない場合、再生時に I/O コストが発生する可能性があります。

次に、変更の再生は、テーブルレベルおよび行レベルのロックを保持するため、ローカルワークロードに対する競合を生成する可能性があります。競合のないレプリケートデータ型 CRDT と列レベルの競合検出 CLCD 機能は、同時更新の場合でも正しい回答を取得しますが、通常のロックオーバーヘッドは削除されません。ロック競合が発生した場合は、競合する更新を回避するか、トランザクションをできるだけ短くしてください。大規模なトランザクションで頻繁に更新された行は、そのトランザクションのパフォーマンスのボトルネックを発生させます。複雑なアプリケーションでは、スケーラビリティを維持するためにいくつかの考慮が必要です。

パフォーマンスに問題があると思われる場合は、ベンチマークツールを使用してパフォーマンステストを開発します。pgd_bench を使用すると、ユースケースに固有のカスタムテストスクリプトを作成できるため、SQL のオーバーヘッドを理解し、同時実行の影響を測定できます。

PGD の実行が遅い場合、次のことをお勧めします。

1. 実稼働システムの問題ケースにできる限り近づけて、pgd_bench のカスタムテストスクリプトを作成します。
2. 1 つのノードでスクリプトを実行して、ベースライン図を取得します。
3. 1 つのノードで実行したのと同じ合計数のセッションを使用して、運用環境で発生するのと同じ数のノードでスクリプトを実行します。このテクニックは、複数のノードに移動した効果を示します。
4. これらの 2 つのテストのセッション数を増やして、アプリケーションでの競合の増加による影響をプロットできます。
5. テストがレプリケーションの遅延を考慮して十分な長さであることを確認します。
6. テスト中にレプリケーション遅延が増加していないことを確認します。

通常の Postgres チューニング機能をすべて使用して、アプリケーションの重要な部分の速度を向上させます。

## 5.4 Backup and recovery

PGD は、分散された高可用システムであるように設計されています。クラスターの 1 つ以上のノードが失われた場合、それらを置き換える最良の方法は、残りのノードから直接新しいノードのクローンを作成することです。

PGD におけるバックアップとリカバリーの役割は、次の状況などのディザスターリカバリー DR を提供することです。

- クラスター内のすべてのノードの損失
- データの破損、アプリケーションエラー、またはセキュリティ侵害の結果、マルチプルのノードにわたる重大な修正不可能なデータの破損

### 5.4.1 論理バックアップとリストア

pg_dump は、 **論理バックアップ**と呼ばれることもあります、通常は PGD で使用できます。

#### postgresql.conf の一時的な設定

最初に、postgresql.conf に次の設定を一時的に設定します。

```
#  Increase from the default of `1GB` to something large, but still a

#  fraction of your disk space since the non-WAL data must also fit.

#  This decreases the frequency of checkpoints.

max_wal_size = 100GB

#  Increase the amount of memory for building indexes. Default is

#  64MB. For example, 1GB assuming 128GB total RAM.

maintenance_work_mem = 1GB

#  Increase the receiver and sender timeout from 1 minute to 1hr to

#  allow large transactions through.

wal_receiver_timeout = 1h
wal_sender_timeout = 1h

#  Increase the number of writers to make better use of parallel

#  apply. Default is 2. Make sure this isnt overriden lower by the

#  node group config num_writers setting.

bdr.writers_per_subscription = 5

#  Increase Raft-related election timeouts with default values of 6s

#  and 3s.

bdr.raft_global_election_timeout = 20s
bdr.raft_group_election_timeout = 10s

#  Increase the size of the shared memory queue used by the receiver to

#  send data to the writer process from the default 1MB.
```

(次のページに続く)

(前のページからの続き)

```
bdr.writer_input_queue_size = 32MB
```

さらに

- トランザクションがストリーミングされるように、デフォルトの bdr.streaming_mode = 'auto' がオーバーライドされていないことを確認します。
- 上記にリストされているセッションまたは postgresql.conf 設定が、一般にノードグループレベルの設定でオーバーライドされていないことを確認します。

次に、 pg_dump および pg_restore に進みます。

### pg_dump / pg_restore

グローバルロックタイムアウトのリスクを軽減するために、事前データ、データ、および事後データを個別にダンプすることをお勧めします。例

```
pg_dump -h $PG_HOST -p $PG_PORT -U $PG_USER -d $PGD_DB -v --exclude-schema="bdr" --
↪exclude-extension="bdr" --section=pre-data -Fc -f pgd-pre-data.dump
pg_dump -h $PG_HOST -p $PG_PORT -U $PG_USER -d $PGD_DB -v --exclude-schema="bdr" --
↪exclude-extension="bdr" --section=data -Fc -f pgd-data.dump
pg_dump -h $PG_HOST -p $PG_PORT -U $PG_USER -d $PGD_DB -v --exclude-schema="bdr" --
↪exclude-extension="bdr" --section=post-data -Fc -f pgd-post-data.dump
```

そして、これらの SQL ファイルをノードで直接実行して復元します。接続マネージャーポートでこれらを実行しないでください。

```
PGOPTIONS="-cbdr.commit_scope=local" pg_restore -h $PG_HOST -p $PG_PORT -U $PG_USER -
↪d $PGD_DB --section=pre-data -f pgd-pre-data.dump
PGOPTIONS="-cbdr.commit_scope=local" pg_restore -h $PG_HOST -p $PG_PORT -U $PG_USER -
↪d $PGD_DB --section=data -f pgd-data.dump
psql -h $PG_HOST -p $PG_PORT -U $PG_USER -d $PGD_DB -c SELECT bdr.wait_slot_confirm_
↪lsn(NULL, NULL)
PGOPTIONS="-cbdr.commit_scope=local" pg_restore -h $PG_HOST -p $PG_PORT -U $PG_USER -
↪d $PGD_DB --section=post-data -f pgd-post-data.dump
psql -h $PG_HOST -p $PG_PORT -U $PG_USER -d $PGD_DB -c SELECT bdr.wait_slot_confirm_
↪lsn(NULL, NULL)
```

この時点で、ダンプはクラスター内のすべてのノードで復元されます。

対照的に、単純な pg_dump および pg_restore でセクションを分割しない場合、リストアはグローバルロックタイムアウトで失敗する可能性があります。

pg_restore でグローバルロックタイムアウトが発生する場合は、`-cbdr.ddl_locking=off` を `PGOPTIONS` に追加します。

`-j /--jobs` で pg_restore を実行する場合、`max_worker_processes` および `max_parallel_maintenance_workers` を同じ量だけ増やす必要があります。

#### 単一ノードクラスターへの復元を優先します

特に、Postgres ダンプからクラスターを最初にセットアップする場合、単一の PGD ノードを使用してクラスターに復元することをお勧めします。次に、クラスター内に必要なノードごとに `pgd node setup` を実行します。これは、内部で `bdr_init_physical` を使用する物理的結合を実行します。

#### シークエンス

pg_dump は、ローカルシーケンスとグローバルシーケンスの両方をローカルシーケンスであるかのようにダンプします。この動作は意図的なもので、PGD スキーマをダンプして他の PostgreSQL データベースに移植できるようにします。これは、ダンプ時にシーケンス種類のメタデータが失われることを意味するため、リストアはすべてのシーケンス種類をリストア時に `bdr.default_sequence_kind` の値に効果的にリセットします。

シーケンスごとに正確なシーケンス種類をリセットする復元後スクリプトを作成するには、次のような SQL スクリプトを使用することができます。

```
SELECT SELECT bdr.alter_sequence_set_kind(||
        nspname||.||relname||,||seqkind||);
FROM bdr.sequences
WHERE seqkind != local;
```

`bdr.crdt_raw_value = on` を使用して pg_dump を実行した場合、`bdr.crdt_raw_value = on` でのみダンプをリロードできます。

テクニカルサポートでは、PGD のバックアップとリカバリーに物理的バックアップ技術を使用することをお勧めします。

### 5.4.2 物理的なバックアップとリストア

Barman などの標準の PostgreSQL ソフトウェアを使用して、 EDB Postgres 分散クラスター内のノードの物理バックアップを取得できます。

PostgreSQL ノードに適用されるのと同じ手順を使用して、PGD ノードの物理バックアップを実行できます。PGD ノードは、 BDR 拡張機能を実行する単なる PostgreSQL ノードです。

PostgreSQL バックアップ技術を PGD に適用するときは、次の特定の点を考慮してください。

- PGD は単一のデータベースのレベルで動作しますが、物理バックアップにはインスタンス内のすべてのデータベースが含まれます。データベースを簡単にバックアップおよび復元できるように計画します。
- バックアップは、1 つのノードのみのコピーを作成します。最も単純なケースでは、すべてのノードにすべてのデータのコピーがあるため、すべてのデータをキャプチャするには 1 つのノードのみをバックアップする必要があります。ただし、その単一のコピーを含むサイトがダウンした場合、PGD の目標は満たされないため、最小はサイトごとに少なくとも 1 つのノードバックアップ多数のコピーなど。
- ただし、各ノードにはレプリケーションされていないローカルデータがある場合、またはレプリケーションセットの定義が複雑な場合があるため、すべてのノードがすべてのレプリケーションセットをサブスクライブしない場合があります。これらの場合、バックアップ計画には、レプリケートされて

いないローカルデータと、各レプリケーションセットをサブスクライブする少なくとも 1 つのノードのバックアップをバックアップする方法に関する計画も含める必要があります。

### リストア

標準の PostgreSQL ノードと同じ手順で物理バックアップを取ることができますが、PGD ノードの物理バックアップを復元するのは少し複雑です。

### EDB Postgres 分散クラスター障害またはバックアップからの新しいクラスターのシード

物理バックアップを復元するための最も一般的なユースケースには、データセンターに障害が発生した場合など、クラスター内のすべての PGD ノードの障害または交換が含まれます。

また、この手順を実行して、EDB Postgres 分散クラスターの現在のコンテンツのクローンを作成して、QA または開発インスタンスをシードすることもできます。

その場合、単一の PGD ノードとオプションで WAL アーカイブの物理バックアップに基づいて、PGD 機能を復元できます。

- いくつかの PGD ノードがまだライブで実行されている場合は、PGD ノードを復元したホストをフェンスオフして、残りの PGD ノードに接続できないようにします。この実践により、新しいノードが既存のクラスターを混乱させないようになります。
- PGD ノードの 1 つの物理バックアップから単一の PostgreSQL ノードを復元します。
- バックアップに関連付けられている WAL アーカイブがある場合は、適切な `postgresql.conf` を作成し、リカバリーで PostgreSQL を起動して最新の状態まで再生します。必要に応じて、ここで代替の `recovery_target` を指定できます。
- 復元されたノードを起動するか、スタンバイ復旧している場合は読み取り/書き込みに昇格させます。生き残ったノードからフェンスしてください。
- 物理バックアップに含まれる残りの PGD メタデータをクリーンアップします。
- PostgreSQL インスタンスを完全に停止して再起動します。
- `bdr.join_node_group()` ファンクション呼び出しに基づいた標準プロシージャで PGD ノードをさらに追加します。

### PGD メタデータのクリーンアップ

残った PGD メタデータをクリーンアップするには

1. bdr.drop_node を使用して PGD ノードを削除します。
2. PostgreSQL を完全に停止して再起動します重要！

### レプリケーション起点のクリーンアップ

レプリケーションオリジンはシステムカタログに永続的に記録されるため、別の手順で明示的に削除する必要があります。したがって、これらはバックアップと復元されたインスタンスに含まれます。これらは

依存関係として明示的に記録されていないため、BDR 拡張機能をドロップするときに自動的に削除されません。

クラッシュセーフな方法で着信レプリケーションの進行状況を追跡するため、PGD はリモートマスターノードごとに 1 つのレプリケーションオリジンを作成します。したがって、以前のクラスターのノードごとに、これを 1 回実行します。

```
SELECT pg_replication_origin_drop(bdr_dbname_grpname_nodename);
```

次のようにして、レプリケーションオリジンをリストできます。

```
SELECT * FROM pg_replication_origin;
```

PGD によって作成されたものは、名前で簡単に認識されます。

#### レプリケーションスロットのクリーンアップ

pg_basebackup で物理バックアップが作成された場合、レプリケーションスロットはバックアップから省略されます。

他のバックアップ方法には、レプリケーションスロットが古いまたは無効な状態で保持される場合があります。バックアップを復元したら、次のコマンドを使用してすべてのレプリケーションスロットをドロップします。

```
SELECT pg_drop_replication_slot(slot_name)
FROM pg_replication_slots;
```

一部のスロットを保持する理由がある場合は、`WHERE slot_name LIKE 'bdr%'` 句を追加できますが、これはほとんど役に立ちません。

> **警告**
>
> これらのコマンドを使用して、ライブ PGD ノードでレプリケーションスロットをドロップしないでください

### 5.4.3 最終的な一貫性

EDB Postgres 分散クラスター内のノードは **結果的に一貫性** ですが、**完全に一貫性**ではありません。特定のノードの物理バックアップは、そのノードが実際に想定する状態に制限されたポイントインタイムリカバリ機能を提供します。

次の例は、同じ EDB Postgres 分散クラスター内の 2 つのノードがどのようにして同じ状態のシーケンスを通過しないか、通常は通過しないかを示しています。

2 つのノード N1 および N2 を含み、最初は S 状態にあるクラスターを検討します。トランザクション W1 がノード N1 に適用され、同時に競合しないトランザクション W2 がノード N2 に適用される場合、ノード N1 は次の状態を経ます。

```
(N1)   S  -->  S + W1  -->  S + W1 + W2
```

ノード N2 は、次の状態を経ます。

```
(N2)   S  -->  S + W2  -->  S + W1 + W2
```

つまり、ノード N1 は決して*状態 S + W2 を引き受けることはなく、ノード N2 も同様に状態 S + W1 を引き受けることはありません。ただし、両方のノードは同じ状態 S + W1 + W2 になります。この状況を考慮すると、バックアップ戦略の決定方法に影響を与える可能性があります。

### ポイントインタイムリカバリー PITR

前の例では、変更も時間的に一貫していないことを示しました。 W1 と W2 は両方とも T1 で発生しますが、変更 W1 は T2 まで N2 に適用されません。

PostgreSQL PITR は、単一のマスターから COMMIT オーダーで変更が到着するという想定に基づいて設計されています。したがって、PITR は、特定の時点 PIT に到達するまで変更をスキャンすることにより可能です。このスキームを使用すると、T1 などの観点から 1 つのノードを単一の PIT に復元できます。ただし、その状態には、その時間近くにコミットしたがノードにまだ到着していない他のノードからの他のデータは含まれません。その結果、回復は部分的に一貫していない、または少なくとも 1 つのレプリケーションオリジンのみについて一貫していると考えられる場合があります。

PostgreSQL PITR では、標準の構文を使用できます。

```
recovery_target_time = T1
```

PGD では、マルチプルのマスターからの変更が可能で、すべては 1 つのノードの WAL ログに記録され、レプリケーションオリジン識別子を使用して個別に識別されます。

PGD は、すべてまたは一部のレプリケーションオリジンの特定の時点への PITR を可能にし、ノードのすべてのサブセットにわたって完全に一貫した観点を提供します。

したがって、マルチオリジンの場合、WAL ストリームを、複数のストリームがすべて 1 つの大きなストリームに混合されたものとして表示できます。PIT はまだ 1 つだけですが、それは各原点の異なるポイントとして個別に到達されます。

WAL ストリームは、要求されたオリジンが PIT を見つけるまで読み取られます。すべての変更は、その時点までに適用されますが、オリジンの PIT に到達した後、トランザクションレコードはオリジンのコミット済みとしてマークされません。

WAL には 1 つの LSN「停止ポイント」がありますが、シングルオリジン PITR の場合と同様に、単一のタイムスタンプが一貫して適用されます。

定義された PIT に到達すると、必要に応じてリカバリーを続行できるように新しい PIT が設定される場合があります。

目的の停止ポイントに到達した後、復旧したサーバーが昇格する場合は、最初にシャットダウンします。pg_resetwal を使用して、このサーバーのタイムラインで使用されるよりも高い LSN 値に LSN を移動します。このアプローチでは、論理デコードによって生成される重複 LSN がないことが保証されます。

この特定の例では、N1 は T1 に復元されます。また、後まで N1 に適用されなかった場合でも、 T1 によってコミットされた他のノードからの変更も含まれます。

マルチオリジン PITR を要求するには、 postgresql.conf ファイルの標準の構文を使用します。

```
recovery_target_time = T1
```

2 つの方法のいずれかで T1 に復元されるレプリケーションオリジンのリストを指定する必要があります。新しいパラメーター recovery_target_origins を介して別の multi_recovery.conf ファイルを使用できます。

```
recovery_target_origins = *
```

または、オリジンのサブセットを recovery_target_origins のリストとして指定できます。

```
recovery_target_origins = 1,3
```

指定された recovery_target_time へのローカル WAL アクティビティのリカバリーは、常に暗黙的に実行されます。recovery_target_origins で指定されていないオリジンの場合、recovery_target_origins で記載されているリストの目標が達成される時期に応じて、任意の時点でリカバリーを停止できます。

multi_recovery.conf ファイルが存在しない場合、リカバリーは、COMMIT オーダーで単一のマスターから到着する変更を想定して設計された元の PostgreSQL PITR 動作がデフォルトになります。

> **注釈**
>
> この機能は、 EDB Postgres Extended でのみ使用できます。 Barman は`multi_recovery.conf` ファイルを作成しません。

### 5.4.4 モニタリング

次のクエリーを使用して、復元プロセスの進行状況を確認します。

```
SELECT pg_size_pretty(pg_database_size(bdrdb));
```

上記のクエリーは、復元ノードのデータベースサイズを示しています。復元が進行し、元のノードのサイズに近づくと、サイズが増加します。ただし、肥大化のため、論理リストアは常に元のサイズより少し小さくなります。

```
SELECT * FROM bdr.node_replication_rates;
```

上記のクエリーは、レプリケーションの速度を示しています。ただし、進行状況情報は、大規模なトランザクションの場合、誤解を招く可能性があります。遅延と進行状況が階段状に表示されます。

```
SELECT
  application_name,
  state,
  wait_event_type,
```

(次のページに続く)

(前のページからの続き)

```
  wait_event,
  now() - state_change AS state_change_ago
FROM
  pg_stat_activity
WHERE
  application_name LIKE %pg_restore%;
```

上記のクエリーは、pg_restore が何をしているか、ブロックされているかどうか、待機しているものについて、または動作しているかどうか、およびステータスを継続的に変更することに関する情報を示します。

次のビューを確認して、レプリケーションスロット、蓄積された遅延、破損したレプリケーションなどの問題を確認します。

- pg_catalog.pg_stat_replication_slots
- pg_catalog.pg_replication_slots
- bdr.node_slots

また、 `bdr.stat_subscription` を使用して、各サブスクリプションの統計を表示します。たとえば、パラレル適用またはトランザクションストリームを確認します。

## 5.5 Migrating to Postgres Distributed (PGD)

従来のアーキテクチャから EDB Postgres 分散 PGD に移行することにより、組織は高可用性、地理的に分散されたワークロード、およびマルチマスター書き込み機能を実現できます。

次のガイドでは、既存のインフラストラクチャを高可用性 PGD クラスターに移行するためのテストおよびサポートされている経路について詳しく説明します。

- Migrate from Postgres Physical Streaming Replication (PSR) to PGD: シードノードアプローチを使用して標準のプライマリスタンバイアーキテクチャを PGD クラスターに移行し、データ転送とバージョンのアップグレード中のダウンタイムを最小限に抑えます。

### 5.5.1 Migrating from Postgres Physical Streaming Replication (PSR) to PGD

Postgres 物理ストリーミングレプリケーション PSR から Postgres DistributedPGD への移行は、シングルプライマリアーキテクチャからマルチプライマリメッシュネットワークへの移行を意味します。この移行は、通常、高可用性と地理的に分散した書き込み機能のニーズによって促進されます。

詳細については、Choosing between PSR and PGD を参照してください。

この移行戦略のコアは **シードノード**であり、標準のフィジカルスタンバイとして既存の PSR クラスターに最初に参加する新しいノードです。同期が完了すると、物理ストリームから分離され、論理サブスクライバに変換されます。

論理レプリケーションにより、Postgres の異なるメジャーバージョン間でデータをフローできます。シードノードがアップグレードして PGD を実行すると、新しいクラスターの「アンカー」として機能し、元のソースプライマリから新しいデータを受信したまま、シードノードから「物理的結合」を使用して追加の PGD ノードを参加できます。

この方法を使用することにより、元のプライマリノードがプロセス全体を通じて「真実のソース」であり、アプリケーションは最後のカットオーバーフェーズまで中断せずに実行され続けます。

#### 前提条件

- 現在物理レプリケーション EFM によって管理されているオプションを利用している PGE、EPAS、または Community Postgres バージョン 14 以降を実行しているノード。
- 十分なストレージを備えた 2 つの追加ノードシードノードと追加ノード。
- すべての既存のノードと新しいノード間のネットワーク接続。

#### 移行の概要

移行は次のフェーズで構成されています。

1. *Preparing your environment* ソースプライマリノードを特定し、PGD クラスターをアンカーするシードノードを選択します。
2. *Transferring data to the seed node* 新しいノードで手動フィジカルスタンバイを確立します。
3. *Converting to logical replication* シードノードを物理ストリーミングから論理ストリーミングに移行します。
4. *Upgrading the seed node and installing PGD* インプレースメジャーバージョンアップグレードを実行し、PGD パッケージをインストールします。
5. *Initializing PGD* BDR 拡張を作成し、最初の PGD ノードグループを初期化します。
6. *Joining additional PGD nodes* クラスターを拡張して高可用性を保証します。
7. *Application cutover* トラフィックを新しい PGD クラスターにスイッチします。
8. *Performing cleanup* クラスターを分離し、古いバージョンを削除します。

### Preparing your environment

物理ストリーミングレプリケーション PSR から Postgres DistributedPGD への移行を開始する前に、ノードの主要なロールを特定し、ターゲットトポロジを確立する必要があります。これにより、環境が移行の一時的なオーバーヘッドを処理でき、クラスター移行の明確なパスを提供できます。

#### ソースノードの選択

既存のクラスターからソースノードを選択して、PGD クラスターの初期データとその後の更新を提供する必要があります。論理レプリケーションはスタンバイノードからカスケードできないため、現在のプライマリノードのみが実行可能なソースになります。

次のことを検討します。

- **リソースオーバーヘッド**移行中に、ソースノードは少なくとも 2 つの PGD ノードに追加の論理デコードを実行する必要があり、追加の CPU 容量が必要です。
- **ストレージ**ソースノードは、レプリカの同期を保つために追加の WAL データを保持する必要があり、標準操作よりも多くのディスク領域が必要です。

- **最適化**現在のプライマリの電力が不足している場合は、移行を開始する前に、より有利なノードへのスイッチオーバープロモーションを検討します。
- **フェールオーバー処理**移行中にプライマリノードに障害が発生した場合、新しいプライマリをソースとして使用してプロセスを中止および再起動する必要があります。

### シードノードの選択

シードノードは、PGD クラスターを形成するグループから選択された新しいノードです。このノードは、移行中に一意の移行的なロールを提供します。

1. ソースノードからデータベースの完全なコピーを受信します。
2. インプレースメジャーバージョンのアップグレードを実行します。
3. 最終的な PGD クラスターを初期化します。

### PGD クラスタートポロジの計画

事前にターゲットトポロジを確立します。続行する前に、場所、ホスト名、PGD ノード名、および接続文字列 DSN を定義します。このガイドでは、シードノード最初のノードと 1 つの追加ノードの 2 ノード PGD スタートを前提としています。さらにノードを追加するには、2 番目のノードと同じプロセスに従います。

次の表を使用して、クラスターに必要な環境変数をマップし、すべての構成スクリプト間の一貫性を保証します。

| Variable | Description | Example |
|---|---|---|
| *${SEED_NODE_PGD_1* | シードノードに割り当てられた PGD 固有の名前。 | node-1 |
| *${SEED_NODE_DSN}* | シードノードに到達するための接続文字列。 | host=node-1 port=5444 dbname=pgddb user=postgres |
| *${ADD_NODE_PGD_N* | 2 番目のノードの PGD 固有の名前。 | node-2 |
| *${ADD_NODE_DSN}* | 追加ノードに到達するための接続文字列。 | host=node-2 port=5432 dbname=pgddb user=postgres |
| *${PGD_CLUSTER_NAM* | すべてのノードを含む PGD ノードグループの名前。 | pgd_cluster_main |

### コマンド環境設定

移行手順では、各コマンドに使用するノードを指定します。

移行コマンドが正しく実行されるように保証には、環境が次の要件を満たしていることを確認します。

- **シェル権限** 指示で `su` を使用することが明示的に指定されていない限り、すべてのコマンドを `root` として実行します。
- **環境変数** `${PGDATA}` が Postgres データディレクトリたとえば、EPAS 14 の`/var/lib/edb/as14/data` に設定されていることを確認します。`${PGPORT}` の設定も必要になる場合があります。
- **ターゲットデータベース** `psql` を実行する前に`${PGDATABASE}` を設定することにより、移行する特定のデータベースデフォルトの `postgres` または `edb` データベースではないに接続します。

次のステップ *Transferring data to the seed node* 。

### Transferring data to the seed node

このフェーズでは、物理レプリケーションを介してソースノードをミラーリングするシードノードにデータベースのスタンバイコピーを作成することにより、シードノードを初期化します。このスタンバイは、最終的な PGD クラスターのシードノードとして機能します。既存のクラスターへの干渉を最小限に抑えるために、特にフェールオーバーマネージャー EFM によって管理されている場合、このレプリカは手動で追加されます。

#### 論理レプリケーションのソースノードの準備

論理レプリケーションに移行するには、シードノードを準備する必要があります。まだ有効になっていない場合は、ソースノードで `wal_level` を `logical` に設定します。この変更にはデータベースの再起動が必要であるため、この手順を事前に実行することをお勧めします。

1. ソースノードで次のコマンドを実行します。

```
ALTER SYSTEM SET wal_level = logical;
```

1. ソースノードの Postgres サービスを再起動して、構成の変更を有効にします。

```
su -u enterprisedb --command "pg_ctl restart"
```

1. 再起動後、`wal_level` が正しいこと、および十分なレプリケーションスロットとワーカープロセスが使用可能なことを確認します。**M** の既存のスタンバイと **N** の将来の PGD ノードを使用したセットアップの場合、これらのパラメーターが少なくとも以下に示されているものであることを確認します。

```
SHOW wal_level;                         -- Expected logical

SHOW max_replication_slots;             -- Must be at least M + N
SHOW max_wal_senders;                   -- Must be at least M + N
SHOW max_logical_replication_workers;   -- Must be at least N
```

#### アウトバウンド論理レプリケーションの許可

外部論理接続を許可するようにソース環境を構成し、移行ストリームに必要な認証資格情報を確立します。

1. ソースノードには、データの移行を促進するためにレプリケーション特権を持つロールが必要です。既存のロールを使用するか、別のロールを作成できます。例

```
CREATE ROLE repl LOGIN NOSUPERUSER REPLICATION PASSWORD your_password_here;
```

EDB は、 `.pgpass` ファイルを使用して、接続文字列 DSN にパスワードを含めずに資格情報を安全に管理することをお勧めします。

1. 次のコマンドを使用して PGD ノードからの接続を検証し、対話型のパスワード入力なしで接続できることを確認します。

```
su -u enterprisedb -c "psql --no-password ${SOURCE_DSN} replication=1 \
--command IDENTIFY SYSTEM"
```

シードノードと将来のすべての PGD ノード間の接続を必ずテストします。これには、ソースノードで pg_hba.conf の更新が必要になる場合があります。

### シードノードの準備

物理的バックアップに対応できるように、シードノードに Postgres の同じディストリビューションとバージョンをインストールします。詳細は、Installing EPAS 、 Installing PGE 、または *Postgres* のインストールを参照してください。

> **注釈**
>
> 移行中に手動で管理する必要があるため、シードノードに EFM をインストールしないでください。

### 物理的バックアップの作成

ソースの構成が完了したら、シードノードから次のコマンドを実行して、ソースノードからデータベースの物理スナップショットを取得します。転送を開始する前に、宛先の${PGDATA} ディレクトリが存在しないことを確認してください。

```
su -u enterprisedb -c "pg_basebackup ${SOURCE_DSN}    \
  --pgdata=$(PGDATA)                                   \
  --write-recovery-conf                                \
  --wal-method=stream                                  \
  --create-slot                                        \
  --slot=migration_phy_slot                            \
  --progress                                           \
  --verbose"
```

このコマンドは、デフォルトでスプレッドチェックポイントを使用して、ソースノードへの I/O インパクトを最小限に抑えます。より高いディスク I/O を犠牲にして転送を高速化するには、--checkpoint=fast を追加できます。

転送時間は、データの合計量と、ソースノードとシードノード間のネットワーク帯域幅によって異なります。帯域幅が制限されている環境の場合、詳細な構成オプションに `–compress-level=

*0-9]* flag to optimize the transfer speed; refer to the official [pg_basebackup documentation <https://www.postgresql.org/docs/14/app-pgbasebackup.html>`_ を使用して圧縮を有効にできます。

### シードノードの構成と起動

1. 転送が完了したら、シードノードの postgresql.conf および pg_hba.conf を調整して、ローカル環境ホスト名、IP アドレス、パスを反映します。シードノードのレプリケーション設定が PGD の要件と一致することを検証します。

```
SHOW wal_level;                          -- Expected: logical
SHOW max_replication_slots;              -- Should be at least M + N
SHOW max_wal_senders;                    -- Should be at least M + N
SHOW max_logical_replication_workers;    -- Should be at least N
```

1. シードノードでデータベースを起動して、ストリーミングを開始します。

```
su -u enterprisedb --command "pg_ctl start"
```

#### 物理的レプリケーションの検証

ソースノードとシードノードの両方からレプリケーションステータスを監視して、それらが同期されていることを確認します。

ソースノードで

1. 移行スロットのステータスを確認します。 `lag_size` が低いまたは減少しているシードノードに割り当てられたスロットを表示する必要があります。

```
SELECT slot_name, active, restart_lsn, confirmed_flush_lsn,
       pg_size_pretty(pg_wal_lsn_diff(pg_current_wal_lsn(), restart_lsn)) AS lag_size
FROM pg_replication_slots
WHERE slot_name = migration_phy_slot;
```

1. アクティブな接続を確認します。シードノードの IP アドレスがレプリケーション統計に表示されることを確認します。

```
SELECT client_addr, state, sent_lsn, write_lsn, flush_lsn, replay_lsn,
       pg_size_pretty(pg_wal_lsn_diff(sent_lsn, replay_lsn)) AS replay_lag
FROM pg_stat_replication;
```

シードノードで

1. リカバリモードを確認します。ノードはスタンバイとして正しく動作している必要があります。

```
SELECT pg_is_in_recovery();              -- Must return true
```

1. WAL レシーバーのステータスを確認します。プライマリソースノードへの接続の状態を確認します。

```
SELECT slot_name, sender_host || : || sender_port AS sender, status,
now() - last_msg_send_time    AS last_msg_send_age,
now() - last_msg_receipt_time AS last_msg_receipt_age,
now() - latest_end_time       AS last_end_age,
pg_size_pretty(pg_wal_lsn_diff(latest_end_lsn, written_lsn)) AS write_lag_size,
pg_size_pretty(pg_wal_lsn_diff(latest_end_lsn, flushed_lsn)) AS flush_lag_size
FROM pg_stat_wal_receiver;
```

1. 受信位置と再生位置を比較します。これは、再生ラグ_受信したがデータベースにまだ適用されていないデータの量を特定します。

```
SELECT pg_size_pretty(pg_wal_lsn_diff(pg_last_wal_receive_lsn(),
                                      pg_last_wal_replay_lsn())) AS lag_bytes;
```

1. 時間ごとに遅延を監視します。現在の時間と最後に再生されたトランジションの時間差を測定します。

```
SELECT EXTRACT(EPOCH FROM (now() - pg_last_xact_replay_timestamp())) AS lag_seconds;
```

次のステップ *Converting to logical replication* 。

### Converting to logical replication

Postgres ディストリビューションのメジャーバージョンのアップグレードと、その後の Postgres Distributed のインストールを促進するには、シードノードへのレプリケーションストリームを論理レプリケーションに変換する必要があります。データ損失のないシームレスな移行を行うには、物理レプリケーションが中断されたソースノードの WAL ストリーム内の正確なログシーケンス番号 LSN で論理レプリケーションを再開する必要があります。これは、スイッチオーバー LSN と呼ばれます。

#### ソースノードでのパブリケーションの作成

レプリケートするデータを定義し、WAL ファイルを保持する専用スロットを作成することにより、ソースノードを準備します。

1. データベース内のすべてのテーブルのパブリケーションを作成します。

```
CREATE PUBLICATION migration_seed_pub FOR ALL TABLES;
```

1. パブリケーションを確認します。

```
SELECT * FROM pg_publication;
SELECT * FROM pg_publication_tables;
```

1. 論理レプリケーションスロットを作成します。

```
SELECT pg_create_logical_replication_slot(migration_node_${SEED_NODE_NAME}, pgoutput);
```

この段階でスロットを作成すると、ソースノードは、スイッチオーバー操作の前に必要な WAL データの保持を開始します。ソースノードは最終的にマルチプルの PGD ノードをサポートするため、各スロットには、サービスを提供するターゲットノードを識別するように特に名前が付けられます。

#### シードノードの昇格

シードノードを昇格させると、着信物理レプリケーションストリームが停止され、ノードがスタンドアロンの書き込み可能なインスタンスに変換されます。

> **⚠ 警告**
>
> データの整合性を維持するには、この段階でシードノードへのアプリケーションの書き込みを防止する必要があります。アプリケーションは、ソースノードに排他的に書き込みを続ける必要があります。

シードノードで、次のコマンドを実行します。

```
su -u enterprisedb --command "pg_ctl promote"
```

### シードノードへの論理レプリケーションの有効化

昇格後、新しい論理レプリケーションスロットを物理レプリケーションが終了したポイントに合わせる必要があります。

1. シードノードで、物理ストリームから再生された最後の LSN を特定します。

```
SELECT pg_last_wal_replay_lsn();
```

1. ソースノードで、論理スロットをその LSN に手動で移動します。${SWITCH_OVER_LSN} を、上記の手順で取得した値に置き換えます。

```
SELECT pg_replication_slot_advance(migration_node_${SEED_NODE_NAME}, ${SWITCH_OVER_
↪LSN});
```

1. シードノードで論理サブスクリプションを作成します。

```
CREATE SUBSCRIPTION migration_seed_sub
        CONNECTION ${SOURCE_DSN}
        PUBLICATION migration_seed_pub
            WITH (
                enabled = false,
                copy_data = false,
                create_slot = false,
                slot_name = migration_node_${SEED_NODE_NAME}
                );
```

1. サブスクリプションがレプリケーションのテーブルを正しく識別したことを確認します。

```
SELECT n.nspname AS schemaname,
    c.relname AS tablename,
    sr.srsubstate AS state,
    s.subname AS subscription_name
FROM pg_subscription_rel sr
JOIN pg_class c ON sr.srrelid = c.oid
JOIN pg_namespace n ON c.relnamespace = n.oid
JOIN pg_subscription s ON sr.srsubid = s.oid
WHERE s.subname = migration_seed_sub
ORDER BY n.nspname, c.relname;
```

1. 論理レプリケーションを有効にします。

```
ALTER SUBSCRIPTION migration_seed_sub ENABLE;
```

#### 物理レプリケーションのクリーンアップ

1. 論理ストリームが検証されアクティブになったら、ソースノードで次のコマンドを実行して、古い物理レプリケーションコンポーネントを削除し、リソースを解放します。

```
SELECT pg_drop_replication_slot(migration_phy_slot);
```

1. シードノードで、次のコマンドを実行して、物理レプリケーションの残りの構成パラメーターを削除します。

```
ALTER SYSTEM RESET primary_conninfo;
ALTER SYSTEM RESET primary_slot_name;
```

次のステップ *Upgrading the seed node and installing PGD* 。

### Upgrading the seed node and installing PGD

このフェーズでは、シードノードで Postgres ディストリビューションのメジャーバージョンのアップグレードを実行し、Postgres Distributed 用に準備します。

#### アップグレードのシードノードの準備

メジャーバージョンのアップグレードを実行するには、シードノードをオフラインにする必要があります。アップグレード後にサブスクリプションのメタデータを手動で更新する必要があるため、サービスを停止する前に論理サブスクリプションを無効にすることを強くお勧めします。

1. シードノードで、論理レプリケーションを無効にします。

```
ALTER SUBSCRIPTION migration_seed_sub DISABLE;
```

1. シードノードで既存の Postgres インスタンスを停止します。

```
su -u enterprisedb --command "pg_ctl stop"
```

1. ターゲットバージョンのパッケージをシードノードにインストールします。pg_upgrade は、移行を実行するために両方のバージョンのバイナリを必要とするため、古いバージョンがインストールされたままであることを確認します。詳細は、Installing EPAS 、 Installing PGE 、または *Postgres* **のインストール** を参照してください。
2. 新しいバイナリたとえば /usr/edb/as17/bin/ を指すように${PATH} 環境変数を調整し、バージョンを確認します。

```
which pg_ctl
```

#### インプレースアップグレードの実行

データファイルをコピーする代わりにハードリンクを使用するインプレースアップグレードには、--link フラグを pg_upgrade とともに使用します。これにより、必要なディスク領域と時間が大幅に削減されますが、古い${OLD_PGDATA} と新しい ${NEW_PGDATA} ディレクトリの両方が同じファイルシステム上に存在する必要があります。

1. シードノードで、新しいデータディレクトリを準備します。

```
OLD_PGDATA = ${PGDATA}
unset PGDATA
mkdir ${NEW_PGDATA}
chown enterprisedb: ${NEW_PGDATA}
chmod 700 ${NEW_PGDATA}
su enterprisedb -c "initdb -D ${NEW_PGDATA} -E utf-8"
```

1. シードノードで、最初にチェックモードで pg_upgrade を実行します。

```
mkdir /tmp/pg_upgrade
cd /tmp/pg_upgrade
su postgres -c "pg_upgrade \
--old-datadir ${OLD_PGDATA} \
--new-datadir ${NEW_PGDATA} \
--old-bindir /usr/edb/as14/bin/ \
--new-bindir /usr/edb/as17/bin/ \
--link \
--check"
```

1. チェックで肯定的な結果が返された場合は、 --check フラグを使用せずに同じコマンドを実行して、実際のアップグレードを実行します。

```
cd /tmp/pg_upgrade
su postgres -c "pg_upgrade \
--old-datadir ${OLD_PGDATA} \
--new-datadir ${NEW_PGDATA} \
--old-bindir /usr/edb/as14/bin/ \
--new-bindir /usr/edb/as17/bin/ \
--link"
```

1. 完了したら、Postgres ディストリビューションの新しいバージョンを起動します。

```
su -u enterprisedb --command "pg_ctl start"
```

#### サブスクリプションの再有効化

メジャーバージョンのアップグレードでは、サブスクリプションの内部テーブルマッピング情報は引き継がれません。特に、カタログテーブル pg_subscription_rel は空になります。サブスクリプションをすぐに有効にすると、テーブルデータを同期せずにレプリケーションスロットが前方に移動され、データ損失が発生します。

これを回避するには、サブスクリプションを「偽構成」して、テーブルメタデータの更新中に実際のスロットを進めないようにする必要があります。存在しないスロットについてシードノードログで報告されるエラーは無視して安全です。

1. シードノードで、偽のスロットをポイントしてメタデータを安全に更新します。

```
ALTER SUBSCRIPTION migration_seed_sub SET (slot_name = bogus);
```

1. パブリケーションを有効にして更新します。

```
ALTER SUBSCRIPTION migration_seed_sub ENABLE;
ALTER SUBSCRIPTION migration_seed_sub REFRESH PUBLICATION WITH (copy_data = false);
ALTER SUBSCRIPTION migration_seed_sub DISABLE;
```

1. サブスクリプションのテーブルリストが**シードノードへの論理レプリケーションの有効化** で取得した構成と一致することを確認します。

```
SELECT n.nspname AS schemaname,
    c.relname AS tablename,
    sr.srsubstate AS state,
    s.subname AS subscription_name
FROM pg_subscription_rel sr
JOIN pg_class c ON sr.srrelid = c.oid
JOIN pg_namespace n ON c.relnamespace = n.oid
JOIN pg_subscription s ON sr.srsubid = s.oid
WHERE s.subname = migration_seed_sub
ORDER BY n.nspname, c.relname;
```

1. サブスクリプションを正しいスロットに戻し、適切な LSN でレプリケーションを再開します。

```
ALTER SUBSCRIPTION migration_seed_sub
    SET (slot_name = migration_node_${SEED_NODE_NAME});
ALTER SUBSCRIPTION migration_seed_sub ENABLE;
```

この段階で、シードノードは新しい Postgres バージョンを実行し、ソースプライマリから論理更新を受信しています。PGD の初期化に進む前に、システムが安定するまで待ち、シードノードが完全に追いついたことを確認します。

次のステップ *Initializing PGD* 。

### Initializing PGD

シードノードがターゲット Postgres バージョンにアップグレードされたら、双方向レプリケーション BDR 拡張機能をインストールし、クラスターを初期化してノードを分散環境に移行する必要があります。

### PGD のインストールと構成

シードノードには、Postgres ディストリビューションとバージョンたとえば、EPAS 17 の PGD Essential の `edb-pgd6-essential-epas17` と特に一致する PGD パッケージが必要です。続行する前に、適切な PGD リポジトリとインストールする PGD エディション Essential または Expanded を構成していることを確認してください。詳細は、Installing the database and pgd を参照してください。

1. BDR 拡張機能を有効にするには、Postgres 構成を更新してライブラリをプリロードし、シードノードでコミットタイムスタンプを有効にします。

```
cat >> ${PGDATA}/postgresql.conf <<EOF
shared_preload_libraries = bdr
track_commit_timestamp = on
EOF
```

1. Postgres を再起動して BDR 拡張機能をロードします。

```
su -u enterprisedb --command "pg_ctl restart"
```

### PGD クラスターを初期化する

拡張機能がロードされると、レプリケートするデータベース内に BDR 拡張機能を作成できます。

1. シードノードで、 BDR 拡張機能を作成します。

```
CREATE EXTENSION bdr;
```

1. BDR がロードされていることを検証し、バージョンを確認します。

```
SELECT bdr.bdr_version();
```

1. シードノードを BDR ノードとして定義し、クラスターを形成する初期ノードグループを確立します。

```
SELECT bdr.create_node(
    node_name := ${SEED_NODE_PGD_NAME},
    dsn := ${SEED_NODE_DSN_DSN}
);
```

1. ノードグループを作成します。

```
SELECT bdr.create_node_group(
    node_group_name := ${PGD_CLUSTER_NAME}
);
```

### PGD CLI の構成

PGD コマンドラインインターフェイス CLI を使用して、クラスターを管理およびモニタします。**all** の現在および将来の PGD ノードで、クラスター全体の接続エンドポイントをリストする構成ファイルを作成する必要があります。

```
mkdir -p /etc/edb/pgd-cli
cat <<EOF > tee /etc/edb/pgd-cli/pgd-cli-config.yml
cluster:
  name: ${PGD_CLUSTER_NAME}
  endpoints:
    - ${SEED_NODE_DSN}
    - ${ADD_NODE_DSN}
```

(次のページに続く)

(前のページからの続き)

```
      - ...
EOF
```

### PGD インストールの検証

PGD CLI を使用して、シングルノードセットアップが正常であり、CLI がクラスターと通信するように正しく構成されていることを確認します。シードノードで、次を実行します。

```
pgd nodes list -- Expected to list exactly one node (itself).
pgd cluster show --health -- Expected to show OK for all checks.
```

次のステップ *Joining additional PGD nodes* 。

### Joining additional PGD nodes

このフェーズでは、PGD クラスターを単一ノードを超えて拡張して、高可用機能を有効にし、最終的なアプリケーションのカットオーバーの前に冗長性を確立します。BDR は外部論理レプリケーションストリームから到着したデータを自動的に転送しないため、ソースノードは独自の専用論理レプリケーションストリームを介して各 PGD ノードに個別にフィードする必要があります。このアーキテクチャにより、移行中にクラスター全体でデータの一貫性が保証されます。

この手順には、 **物理的結合** と呼ばれる、内部で pg_basebackup を使用する物理コピーを介した高速データ転送が含まれます。このプロセスは、PGD CLI を介して自動化されています。

### PGD パッケージのインストール

追加ノードには、Postgres ディストリビューションとバージョンたとえば、EPAS 17 の PGD Essential の edb-pgd6-essential-epas17 と一致する PGD パッケージが必要です。続行する前に、適切な PGD リポジトリとインストールする PGD エディション Essential または Expanded を構成していることを確認してください。詳細は、Installing the database and pgd を参照してください。

1. BDR 拡張機能を有効にするには、Postgres 構成を更新してライブラリをプリロードし、シードノードでコミットタイムスタンプを有効にします。

```
cat >> ${PGDATA}/postgresql.conf <<EOF
shared_preload_libraries = bdr
track_commit_timestamp = on
EOF
```

1. Postgres を再起動して BDR 拡張機能をロードします。

```
su -u enterprisedb --command "pg_ctl restart"
```

### ソースノードでの追加のレプリケーションスロットの作成

新しいノードを追加するには、ソースノードに専用のレプリケーションスロットを作成する必要があります。新しいノードのデータを調整するには、シードノードをリファレンスポイントとして使用し、そのレプリケーションを一時的に一時停止して、一貫したスイッチオーバー LSN を確立します。

1. ソースノードで、追加の PGD ノードの追加レプリケーションスロットを作成します。

```
SELECT pg_create_logical_replication_slot(migration_node_${ADD_NODE_PGD_NAME},
    pgoutput);
```

1. シードノードで、ソースノードからの論理レプリケーションを一時停止します。

```
ALTER SUBSCRIPTION migration_seed_sub DISABLE;
```

1. レプリケーションが一時停止しているときに、新しいスロットをシードノードの LSN と合わせて、正確なスイッチオーバーポイントをマークします。ソースノードで、次を実行してスロットを進めます

```
SELECT pg_replication_slot_advance(migration_node_${ADD_NODE_PGD_NAME},
    (SELECT confirmed_flush_lsn
        FROM pg_replication_slots
        WHERE slot_name = migration_node_${SEED_NODE_PGD_NAME}));
```

1. ソースノードで、どちらのスロットも進んでいないことを確認します LSN が静的なままで遅延が増加するはずです。

```
SELECT slot_name, active, restart_lsn, confirmed_flush_lsn,
    pg_size_pretty(pg_wal_lsn_diff(pg_current_wal_lsn(), restart_lsn)) AS lag_size
FROM pg_replication_slots;
```

1. ソースノードで、スタンバイからのアクティブな接続を確認します。シードノードも追加ノードのエントリを表示してはなりません。

```
SELECT client_addr, state, sent_lsn, write_lsn, flush_lsn, replay_lsn,
    pg_size_pretty(pg_wal_lsn_diff(sent_lsn, replay_lsn)) AS replay_lag
FROM pg_stat_replication;
```

### 追加の PGD ノードへの参加

PGD CLI は、物理的結合を介してデータ転送プロセスを自動化します。これは、pg_basebackup を内部で使用して、新しいノードにシードノードのクローンを作成します。

1. 追加の PGD ノードで、次を実行します。

```
su postgres -c pgd node ${ADD_NODE_PGD_NAME} setup      \
    --listen-addr="*"                                    \
    --initial-node-count=3                               \
    --pgdata=${PGDATA}                                   \
    --group-name ${PGD_CLUSTER_NAME}                     \
```

(次のページに続く)

(前のページからの続き)

```
    --cluster-dsn "${SEED_NODE_DSN}"                    \
    --dsn "${ADD_NODE_DSN}"                             \
    --verbose
```

1. 完了したら、新しいノードが ACTIVE であり、シードノードから可視であることを確認します。

```
SELECT * FROM bdr.node_summary;
```

> **警告**
>
> PGD クラスターがアクティブになり、ほとんどの DDL 操作を含むすべての変更をレプリケートします。ノード固有の操作を実行する前に DDL レプリケーション`SET bdr.ddl_replication = off` を無効にして、意図しない伝播を防ぎます。アプリケーションがスイッチオーバーするまで、PGD ノードに DML 変更を直接適用しないでください。

### ソースノードからのレプリケーションの再開

両方のノードが同じ LSN に同期すると、論理データフローを再開できます。

1. シードノードで、ソースノードからの論理レプリケーションを再度有効にします。

```
SET bdr.ddl_replication = off;
ALTER SUBSCRIPTION migration_seed_sub ENABLE;
```

1. 追加ノードでは、サブスクリプションは物理的参加中に意図的にコピーされません。これを手動で作成し、ソースノードの専用スロットを指す必要があります。

```
SET bdr.ddl_replication = off;
CREATE SUBSCRIPTION migration_add_sub
    CONNECTION ${SOURCE_NODE_DSN}
    PUBLICATION migration_seed_pub    -- same publication, can be used multiple times
    WITH (
        enabled = false,
        copy_data = false,
        create_slot = false,
        slot_name = migration_node_${ADD_NODE_PGD_NAME}
    );
```

1. 追加ノードで、サブスクリプションのテーブルリストがシードノードへの論理レプリケーションの有効化 で取得した構成と一致することを確認します。

```
SELECT n.nspname AS schemaname,
    c.relname AS tablename,
    sr.srsubstate AS state,
    s.subname AS subscription_name
```

(次のページに続く)

(前のページからの続き)

```
FROM pg_subscription_rel sr
JOIN pg_class c ON sr.srrelid = c.oid
JOIN pg_namespace n ON c.relnamespace = n.oid
JOIN pg_subscription s ON sr.srsubid = s.oid
WHERE s.subname = migration_seed_sub
ORDER BY n.nspname, c.relname;
```

1. 追加ノードで、サブスクリプションを有効にしてソースノードからレプリケーションを再開します。

```
SET bdr.ddl_replication = off;
ALTER SUBSCRIPTION migration_seed_sub ENABLE;
```

#### すべてのレプリケーションストリームの検証

すべてのノードがソースからデータを受信し、PGD クラスター内で通信していることを確認します。データは PGD からソースノードに逆流しないことに注意してください。

1. ソースノードで、次のクエリーを実行して、非 PGD 論理サブスクリプションラグを確認します。

```
SELECT slot_name, active, confirmed_flush_lsn, pg_size_pretty(pg_wal_lsn_diff(pg_
↪current_wal_lsn(), confirmed_flush_lsn)) as lag FROM pg_replication_slots;
```

PGD ノードごとに 1 つのエントリ、および遅延は最小限であることが予想されます。

1. 各 PGD ノードで、PGD メッシュとレプリケーションのステータスを確認します。PGD クラスターのアイドル状態のため、トラフィックは非常に少ないと予想されますが、遅延はありません。

```
SELECT * FROM bdr.node_summary;
SELECT * FROM bdr.subscription_summary;
SELECT * FROM bdr.node_replication_rates;
```

次のステップ *Application cutover* 。

### Application cutover

PGD クラスターが完全に同期したら、このフェーズでは、Postgres ストリーミングレプリケーション PSR 環境から Postgres DistributedPGD クラスターへの最終的な移行について詳しく説明します。

#### アプリケーションの書き込みを停止する

データの相違を防ぐには、アプリケーションレベルで古い PSR クラスターへのすべての書き込みトラフィックを停止します。ソースノードで新しいトランザクションが開始されていないことを確認します。

#### 遅延検証の実行

トラフィックをシフトする前に、PGD クラスターがソースノードからすべての未処理の変更を取り込んだことを確認します。シードノードで次のコマンドを実行して、移行サブスクリプションのレプリケーションラグがゼロであることを確認します。

```
DO $$
DECLARE
   wait_lsn pg_lsn := pg_current_wal_lsn();
   current_flush_lsn pg_lsn;
BEGIN
  LOOP
    current_flush_lsn := (SELECT MAX(flush_lsn) FROM pg_stat_replication WHERE␣
↪application_name IN (migration_seed_sub, migration_add_sub));
    IF wait_lsn <= current_flush_lsn THEN
      EXIT;
    END IF;
    RAISE NOTICE Waiting for replication to catch up, current lag: %,  pg_wal_lsn_
↪diff(wait_lsn, current_flush_lsn);
    PERFORM pg_sleep(1);
  END LOOP;
END;$$;
```

### サブスクリプションの無効化

遅延がゼロになったら、元の PSR ソースに接続するすべての PGD ノードで論理サブスクリプションを無効にします。すべての PGD ノードで、次を実行します。

```
SET bdr.ddl_replication = off;
ALTER SUBSCRIPTION migration_seed_sub DISABLE;
```

### アプリケーショントラフィックのスイッチング

PGD クラスターが完全に最新の状態であることが確認され、古いソースから分離されると、トラフィックをリダイレクトし、オペレーションを再開できるようになりました。

接続文字列を更新して、アプリケーションが Connection Manager を指すようにします。特定の方法たとえば、DNS、Kubernetes ConfigMaps、または.env ファイルの更新などは、特定のアプリケーションアーキテクチャによって異なります。

アプリケーショントラフィックを再開します。 Connection Manager は、要求を正しい PGD ノードにルーティングします。

### 検証の実行

アプリケーションが PGD クラスターでライブになった後、最終的な検証チェックを実行して、PGD メッシュが正常で、アプリケーションからデータを受信し、PGD クラスター内のすべてのノードに伝播していることを確認します。これには、レプリケーションストリームを再検証して、書き込みがすべてのノードに伝播していることを確認することが含まれます。

1. すべてのノードが’アクティブ’状態であり、内部サブスクリプションが正しくストリーミングされていることを確認します。任意の PGD ノードで、次を実行します。

```
SELECT * FROM bdr.node_summary;
SELECT * FROM bdr.subscription_summary;
```

1. レプリケーションレートを確認して、クラスターが内部遅延を蓄積することなく新しい負荷を処理していることを確認します。任意の PGD ノードで、次を実行します。

```
SELECT * FROM bdr.node_replication_rates;
```

次のステップ *Performing cleanup* 。

### Performing cleanup

この最終フェーズでは、古いレプリケーションインフラストラクチャを廃止します。

#### ソースレプリケーションの分離

アプリケーションが PGD クラスターに正常にスイッチされた後、移行リンクを削除して環境を完全に分離します。

データがソースノードを介して流れないことを確認するには、元の PSR ソースに接続するすべての PGD ノードで論理サブスクリプションを無効にする必要があります。すべての PGD ノードで、次を実行します。

```
DROP SUBSCRIPTION `subscription_name`;
```

`subscription_name` は、各ノードのサブスクリプションの名前です。

#### シードノードのクリーンアップ

移行が完了し、クラスターが安定すると、シードノードは Postgres の古いバージョンを必要としません。対応する古いパッケージをノードから削除して、リソースを解放し、将来のメンテナンスを簡素化できるようになりました。

# 第6章　Configuration and Management

## 6.1 Node types and capabilities

PGD クラスターには、それぞれに独自のロールを持つ複数の異なるタイプのノードを含めることができます。このセクションでは、PGD クラスターで構成できるさまざまなタイプのノードについて説明します。

- *Overview of durability options* は、PGD クラスターに存在できるノードの種類と、それに関連するロールの概要です。
- *Witness ノード* は、グループでコンセンサスを確立する専用の PGD ノードの特別なクラスである監視ノードを調べます。
- *Logical standby nodes* は、スタンバイのノードを効率的に同期させ、障害が発生した場合にプライマリとしてステップインする準備をする方法を示します。
- *Subscriber-only nodes and groups* は、加入者専用ノードが加入者専用グループと連携する方法、読み取りスケーラビリティを向上させる方法、およびそれらを構成するためのさまざまなオプションを検討します。

### 6.1.1 An overview of PGD Node types

#### データノード

PGD のデータノードは、Postgres インスタンスを実行するノードです。他のすべてのデータノードにデータを複製します。また、ロックとリーダーシップを中心にクラスター全体の Raft の意思決定にも参加します。1 つ以上のグループのメンバーになることができ、デフォルトでは、クラスター内のすべてのデータノードにまたがる「トップレベル」グループのメンバーです。

データノードは、他の 3 つのノードが構築される基盤でもあります。

#### Witness ノード

監視ノードは、ロックとリーダーシップを中心にクラスター全体の Raft の意思決定に参加するという点でデータノードのように動作します。ただし、データを複製または保存しません。監視ノードの目的は、クラスターがコンセンサスを求めるマジョリティを達成できるように利用できることです。*Witness ノード* に詳細があります。

#### ロジカルスタンバイノード

論理スタンバイノードは、別のノードから論理データの変更を受信し、ローカルに複製するノードです。PGD は、ロジカルスタンバイノードを使用して、複製しているノードが利用できなくなった場合にそのノードを置き換えることができますが、いくつかの注意事項があります。詳細については、*Logical standby nodes* を参照してください。

#### サブスクライバー専用ノード

サブスクライバー専用ノードは、名前が示すように、クラスター内の変更をサブスクライブするのみが、他のノードに変更をレプリケートしないデータノードです。サブスクライバ専用ノードは、アプリケーションの読み取り専用ノードとして使用できます。ノードを作成するときにデータノードが `subscriber-only` を指定し、サブスクライバー専用グループに追加することにより、サブスクライバー専用ノードを作成します。詳細は、*Subscriber-only nodes and groups* を参照してください。

### 6.1.2 Witness nodes

監視ノードは、データノードとして機能する軽量ノードですが、データを保存または複製しません。監視ノードを使用して、Raft コンセンサスを使用する PGD クラスターが奇数の投票ノードを持つようにできるため、決定を行うときにマジョリティを達成できます。

#### PGD グループまたはリージョン内の Witness ノード

監視ノードの一般的な使用の 1 つは、PGD グループに 2 つのデータノードがあるが、推奨される 3 つのデータノードでリソースが利用できない場合です。この場合、監視ノードを PGD グループに追加して、ローカルの Raft の意思決定に 3 番目の投票ノードを提供できます。これらの決定は、主に使用するプロキシの書き込みリーダーを誰が選択するかに関するものです。2 つのノードのみでは、どのデータノードが書き込みリーダーであるかについてコンセンサスがない可能性があります。2 つのデータノードと監視者がある場合、2 人の候補者データノードと 3 人の有権者データノードと監視者が存在します。データノードがダウンしている場合、書き込みリーダーを選択できる投票者はまだ 2 人います。

#### リージョン外の Witness ノード

より高いレベルでは、複数の PGD グループが異なるリージョンにマッピングされている場合、監視ノードを使用できます。たとえば、2 つのリージョンのリージョンごとに 3 つのデータノードがある場合、通常に実行されている間、6 つのデータノードすべてが Raft 決定に参加し、DDL および DML グローバルロックを取得できます。データノードがダウンしている場合でも、コンセンサスを取得するための十分なデータノードが存在します。しかし、ネットワーク分割が発生し、他のリージョンとの接続が失われると、6 つのノードのうち 3 つのノードのみが利用できますが、これはコンセンサスには十分ではありません。このシナリオを回避するには、PGD クラスターの一部として 3 番目のリージョンに監視ノードを展開できます。この監視ノードにより、リージョンが利用できないときに PGD クラスターのほとんどの操作要件でコンセンサスを達成できます。

### 6.1.3 Logical standby nodes

PGD では、オフロードノード、読み取り専用ノード、受信専用ノード、または論理リードレプリカとしても知られる**ロジカルスタンバイノード**を作成できます。マスターノードには、0、1 つ、またはそれ以上のロジカルスタンバイノードを持つことができます。

> **注釈**
>
> 論理スタンバイノードは、データセンター間のネットワークトラフィックが懸念される環境で使用できます。それ以外の場合、ロケーションごとにより多くのデータノードを持つことが常に優先されます。

論理スタンバイノードは、継続的リカバリの状態で保持されるノードであり、必要になるまで常に更新されます。この動作は、パフォーマンスを向上させるために論理レプリケーションを使用しながら、Postgres フィジカルスタンバイが動作する方法に似ています。論理スタンバイノードは変更を受信しますが、ローカルで行われた変更を他のノードに送信しません。

ロジカルスタンバイは、bdr.create_node でノードを作成するときに node_kind を standby として指定することにより作成されます。

後で、必要に応じて、 bdr.promote_node を使用します

ロジカルスタンバイを完全な通常の送信/受信ノードに移動します。

ロジカルスタンバイは、bdr.join_node_group の DSN で定義された 1 つのソースノードからデータが送信されます。他のすべてのノードからの変更はこの 1 つのソースノードから受信され、マルチプルのサイト間の帯域幅が最小化されます。

高可用性を実現するために、ソースノードが停止した場合、1 つのロジカルスタンバイをフルノードに昇格させ、シングルマスター操作と同様のフェールオーバー操作でソースを置き換えることができます。複数のロジカルスタンバイノードが存在する場合、他のノードは新しいマスターをフォローできないため、この手法の有効性は 1 つのロジカルスタンバイに制限されます。

既存の PGD ノードから新しいスタンバイが作成される場合、グループスロットが最後に進められてから少なくとも 16MB の LSN が経過するまで、操作に必要なレプリケーションスロットは新しいスタンバイに同期されません。極端な場合、これには、スロットがストリーミングレプリカで同期または作成される前に、完全な 16 MB が必要になる場合があります。この間隔でフェールオーバーまたはスイッチオーバーが発生した場合、グループスロットおよびその他の依存スロットがまだ存在していないため、ストリーミングスタンバイを PGD ノードに代わって昇格させることはできません。

スタンバイのスロット同期プロセスは、上流のファンクションを呼び出すことによりこれを解決します。この機能は、WAL スイッチを実行し、すべての PGD ピアノードに進行状況の更新を再生するように要求することにより、EDB Postgres 分散クラスター全体のグループスロットを移動します。この動作により、グループスロットが短期間で前進します。これにより、スタンバイが初期スロットの同期に必要な時間が削減され、必要に応じてより迅速なフェイルオーバーが可能になります。

PostgreSQL では、スロットを昇格する前に、スタンバイでスロットの同期が完了することを確認することが重要です。ターゲットデータベースのスタンバイで次のクエリを実行して、スロットがアップストリームと同期していることを監視および確認できます。このクエリーが true を返すと、プロモーションを続行できます。

```
SELECT true FROM pg_catalog.pg_replication_slots WHERE
    slot_type = logical AND confirmed_flush_lsn IS NOT NULL;
```

WAL スイッチを手動で実行し、すべての PGD ピアノードに進行状況の更新を再生するように要求することにより、PGD クラスター全体のスロット同期プロセスをナッジすることもできます。このアクティビティにより、グループスロットが短時間で前進し、スタンバイでのスロット同期アップアクティビティも高速化されます。このターゲットデータベース内の任意の PGD ピアノードで次のクエリーを実行できます。

```
SELECT bdr.run_on_all_nodes(SELECT pg_catalog.pg_switch_wal());
SELECT bdr.run_on_all_nodes(SELECT bdr.request_replay_progress_update());
```

スタンバイでモニタリングクエリを使用して、これらのクエリがそのスタンバイでのスロットの同期の高速化に役立つことを確認します。

ロジカルスタンバイでは、書き込みトランザクションが許可されます。これにより、ロジカルスタンバイに追加のインデックス、データのより長い保持期間、中間作業テーブル、LISTEN/NOTIFY、一時テーブル、マテリアライズドビュー、およびその他の違いを持つことができるため、大きなメリットがあります。

プロモーション前にコミットするロジカルスタンバイにローカルに加えられた変更は、他のノードに送信されません。プロモーション後にコミットするすべてのトランザクションは以降に送信されます。ロジカルスタンバイへの書き込みを実行する場合、昇格する前にデータベースを静止するように注意してください。

ロジカルスタンバイノードに DDL 変更を行う場合がありますが、それらはレプリケートされず、グローバル DDL ロックを取得しようとしません。DDL と同様に動作する PGD ファンクションもレプリケートされません。*Running DDL safely* を参照してください。ロジカルスタンバイに互換性のない DDL 変更を行った場合、データベースは**ダイバージェントノード**です。現在、ダイバージェントノードのプロモーションはレプリケーションに失敗します。その結果、ロジカルスタンバイノードをスタンバイとして使用する場合は、ロジカルスタンバイノードに発散的な変更がないようにするか、発散ノードが昇格しないように計画します。

### 6.1.4 Subscriber-only nodes and groups

サブスクライバー専用ノードとグループは、PGD クラスターに読み取りスケーリングを構築する強力な方法を提供します。

- *Overview of durability options* は、PGD でサブスクライバ専用ノードとグループがどのように動作するかを紹介します。
- *Creating Subscriber-only groups and nodes* は、サブスクライバー専用のグループとノードを作成する方法について説明します。
- *Joining nodes to a Subscriber-only group* は、メンバーがいる既存のサブスクライバー専用グループにノードを参加させる方法について説明します。
- *Optimizing subscriber-only groups* は、より効率的なレプリケーションのためにグループリーダーを使用する PGD サブスクライバのみの最適化トポロジ機能の構成方法の詳細を提供します。

#### An overview of Subscriber-only nodes

##### 概要

多くのユースケースは、クエリと更新を処理できるデータベースノードへのアクセスに依存していますが、読み取り専用のデータベースクエリを処理できるノードへのアクセスのみを必要とするユースケースもあります。このような読み取りスケーリングは、クラスター内のアクティブなデータベースノードから読み取り専用トラフィックを移動することにより、コアクラスターのパフォーマンスを向上させながら、データベースアクセスをより広く利用できるようにします。

#### サブスクライバー専用ノード

サブスクライバー専用ノードの基本的なアイデアは、メインクラスターから読み取り専用クエリをオフロードするために使用できる読み取り専用ノードを提供することです。PGD クラスターのデフォルトのトポロジは、いわゆるフルメッシュトポロジであり、すべてのノードが他のすべてのノードに接続します。これは、ノードを接続する最も堅牢でフォールトトレラントな方法ですが、一部のユースケースでは非効率である場合があります。

サブスクライバ専用ノードは、サブスクライバ専用グループのメンバーになることができますが、PGD 6 以降では、データグループの一部になることができます。

#### 購読者専用グループ

PGD の加入者専用グループは、加入者専用ノードを収集します。各グループは、さまざまなリージョンまたはさまざまなアプリケーションの要求に対応できます。

データグループとは異なり、購読者専用グループには、独自の raft コンセンサスメカニズムがありません。これは、加入者専用グループが必要な数の加入者専用ノードを持つことができることも意味します。

PGD 6 より前は、サブスクライバー専用グループの存在によってレプリケーショントポロジが変更されませんでした。デフォルトでは、サブスクライバ専用グループ内のすべてのノードは、クラスター内の他のすべてのノードからレプリケートされた変更を個別に受信します。

#### 加入者専用グループの最適化

PGD 6 以降では、オプションで加入者専用グループのトポロジを最適化できます。

データノードにプロキシと raft 対応グループを使用するクラスターの場合、サブスクライバー専用グループは、レプリケートされた変更を受信するためのより効率的なモデルを使用できます。

最適化されたトポロジオプションは、PGD プロキシの書き込みリーダーと同様に、各サブスクライバ専用グループにグループリーダーを作成します。グループリーダーは、クラスターからすべての変更を受信し、グループ内の他のノードにそれらを複製します。この機能の詳細については、*Optimizing subscriber-only groups* を参照してください。

#### サブスクライバー専用ノードと DDL

サブスクライバー専用ノードは、ローカルに発行された DDL コマンドを実行できますが、これらの変更はクラスター内の他のノードにレプリケートされず、クラスターでロックを取得しようとしません。

### Creating Subscriber-only groups and nodes

サブスクライバー専用ノードを作成するプロセスは、ノードを含めるサブスクライバー専用グループを作成することから始まります。PGD クラスター内の既存の完全に参加したノードでこの手順を実行します。

#### 購読者専用グループの手動作成

購読者専用グループを作成するには、グループの作成時に `node_group_type` を `subscriber-only` として指定する必要があります。たとえば、ここでは、「host-one」で実行されているノード「node-one」にログインしています。これは、独自のデータグループのメンバーであり、すべてのノードと同様に、ここでは

topgroup と呼ばれるトップレベルグループのメンバーです。このノードに直接ログインして、次の SQL コマンドを使用して sogroup という名前の新しいサブスクライバー専用グループを作成します。

```
select bdr.create_node_group(sogroup, topgroup, false, subscriber-only);
```

またはパラメーター名を使用してより明示的に

```
select bdr.create_node_group(node_group_name:=sogroup,
        parent_group_name:=topgroup,
        join_node_group:=false,
        node_group_type:=subscriber-only);
```

これにより、 topgroup グループの子である sogroup という名前のサブスクライバー専用グループが作成されます。join_node_group の false パラメーターは、このコマンドを実行するノードが、新しく作成されたグループに参加しないことを示します。グループへの自動的な参加はデフォルトの動作であり、この場合、抑制する必要があります。

#### 新しいサブスクライバー専用グループへのノードの手動追加

新しいデータノードを初期化して、サブスクライバー専用グループに追加できるようになりました。データノードを作成し、他のデータノードの場合と同様に bdr 拡張機能を構成します。

次に、この新しいノードを subscriber-only ノードとして作成する必要があります。これを行うには、新しいノードにログインし、次の SQL コマンドを実行します。

```
select bdr.create_node(so-node-1, host=so-host-1 dbname=pgddb port=5444, subscriber-
↪only);
```

次に、その新しいノードにログインし、次の SQL コマンドを使用して sogroup グループに追加します。

```
select bdr.join_node_group(host=host-one dbname=pgddb port=5444,sogroup);
```

またはパラメーター名を使用してより明示的に

```
select bdr.join_node_group(dsn:=host=host-one dbname=pgddb port=5444,
        node_group_name:=sogroup);
```

これは、新しいノードに sogroup グループに参加するように指示します。クラスタートポロジの知識がないため、DSN で指定されたノードに接続して、グループに参加するために必要な情報を受信します。この例では、これはたまたまサブスクライバー専用グループの作成に使用したノードと同じノードですが、クラスターに完全に参加しているノードである可能性があります。

### Joining nodes to a Subscriber-only group

PGD クラスターにサブスクライバー専用グループがない場合、 *Creating Subscriber-only groups and nodes* のプロセスに従ってグループを作成する必要があります。加入者専用グループを作成したら、加入者専用ノードをそれに参加できます。

### 既存の加入者専用グループへのノードの参加

*Creating Subscriber-only groups and nodes* へのノードの参加とは異なり、既存の加入者専用グループへのノードの参加はより簡単なプロセスです。

最初に、新しいノードを加入者専用ノードとして作成します。新しいノードで次の SQL コマンドを実行します。

```
select bdr.create_node(so-node-2, host=so-host-2 dbname=pgddb port=5444, subscriber-
↪only);
```

またはパラメーター名を使用してより明示的に

```
select bdr.create_node(node_name:=so-node-2,
        dsn:=host=so-host-2 dbname=pgddb port=5444,
        node_type:=subscriber-only);
```

このコマンドは、ホスト so-host-2 に so-node-2 という名前の新しいノードを作成し、サブスクライバ専用ノードとして構成します。ノードは、グループに参加するまでクラスターに参加できません。

*Creating Subscriber-only groups and nodes* では、 sogroup という名前のグループを作成し、ホスト shost-1 に so-node-1 というサブスクライバ専用ノードを追加しました。既存のデータグループのノードを使用して、その結合を促進します。ただし、この新しい加入者専用ノードを使用して、別の加入者専用ノードを追加することはできません。クラスターに完全に参加しているアクティブなデータノードを使用する必要があります。*Creating Subscriber-only groups and nodes* の例では、このタスクのクラスターのデータグループで host-one を使用します。 shost-2 で次の SQL コマンドを使用して、sogroup グループに参加できます。

```
select bdr.join_node_group(host=host-one dbname=pgddb port=5444,sogroup);
```

またはパラメーター名を使用してより明示的に

```
select bdr.join_node_group(dsn:=host=host-one dbname=pgddb port=5444,
        node_group_name:=sogroup);
```

このコマンドは、新しいノードに sogroup グループに参加するように指示します。クラストポロジの知識がないため、DSN で指定されたノードに接続して、グループに参加するために必要な情報を受信します。そのノードは、新しいノードをグループに参加する要求のソースとして機能するため、クラスターに完全に参加する必要があります。

新しいノードがグループに参加すると、最初に同期することから開始され、クラスター内の他のノードからレプリケーション変更の受信を開始します。

> **注釈**
>
> グループが *Optimizing subscriber-only groups* を使用している場合を除き、参加しているサブスクライバ専用グループのサブスクライバ専用グループリーダーから変更をレプリケートします。

### Optimizing subscriber-only groups

PGD 6 以降では、 *Overview of durability options* のトポロジを最適化できるようになりました。

この最適化されたトポロジでは、少数の完全にアクティブなノード_データグループの書き込みリーダー_が、加入者専用グループのグループリーダーに変更をレプリケートします。これらのグループリーダーは、加入者専用グループの他のメンバーに変更を複製します。

#### 最適化されたトポロジの要件

クラスターに次のいずれかがある場合、このモデルを有効にすることはできません。

- トップレベルグループの直接メンバーであるデータノード
- データノードサブグループなし
- プロキシルーティングが有効になっているデータノードサブグループはありません

これらのいずれかに該当する場合、加入者専用グループのノードはフルメッシュトポロジに戻ります。

新しい SO グループとノードレプリケーションの利点を得るには、データノードをサブグループに配置し、サブグループでプロキシルーティングを有効にする必要があります。

#### 最適化されたトポロジの仕組み

データノードにグループを使用するクラスターの場合、サブスクライバー専用グループは、より効率的なモデルを使用できます。このモデルは、PGD プロキシの書き込みリーダーと同様に、サブスクライバ専用のグループリーダーを使用します。

各加入者専用グループは、そのグループリーダーを使用して、グループ内の他の加入者専用ノードに変更を複製します。グループリーダーは、入ってくる変更のレプリケーションプロキシとして機能します。

データグループの書き込みリーダーノードは、加入者専用グループのグループリーダーに変更を複製します。データグループ内の他のノードは、自分のデータグループ内のノードおよび他のデータグループのデータノードとのみ複製します。変更を加入者専用グループに直接複製しません。

#### 加入者専用グループリーダー

PGD 6 以降では、各加入者専用グループに独自のグループリーダーが割り当てられます。これは、加入者専用グループには、独自のグループ Raft コンセンサスメカニズムがないためです。代わりに、クラスターのトップレベルグループは Raft コンセンサスメカニズムを使用して、各加入者専用グループのグループリーダーの選択を処理します。このグループリーダーの選択は、トポロジ最適化設定に関係なく、PGD 6 ではデフォルトでオンになっています。

加入者専用グループのグループリーダーは定期的に接続性がテストされ、使用できない場合、トップレベルグループの投票ノードは、加入者専用グループから新しい加入者専用ノードを選択して、グループリーダーになります。次に、新しいグループリーダーが選択されます。

最適化されたテクノロジーがオフになっているため、この選択はレプリケーショントポロジに影響を与えません。最適化されたトポロジがないと、すべてのデータノードがクラスター内の他のすべてのノードに変更をレプリケートします。

### 最適化されたトポロジのグループリーダー

最適化されたトポロジが有効になると、サブスクライバ専用グループのグループリーダーのみがクラスター内の他のデータグループの書き込みリーダーから変更を受け取ります。グループリーダーは、これらの変更を加入者専用グループ内の他のノードに複製する責任を負います。

他の投票ノードは、加入者専用グループのノードからグループリーダーを選択します。グループリーダーが選択されると、クラスター全体が変更を認識し、データグループの書き込みリーダーは、この新しく選択されたグループリーダーノードにのみデータを複製します。データグループ内の他のデータノードは、加入者専用グループのノードにデータを複製しません。

このアプローチは、多数の SO ノードが存在する場合に発生する可能性のあるアクティブな接続の爆発を回避し、レプリケーショントラフィックの量を削減します。

サブスクライバ専用ノードとグループは、PGD ツリートポロジのビルディングブロックを形成します。

### 最適化されたモデルの有効化

デフォルトでは、 PGD 6 はフルメッシュトポロジを強制します。これは、ここで説明する最適化がオフであることを意味します。最適化されたトポロジを有効にするには、データノードをサブグループに配置し、サブグループでプロキシルーティングを有効にする必要があります。次に、GUC bdr.force_full_mesh を `off` に設定して、最適化をアクティブにできるようにします。

> **注釈**
>
> この GUC は、各データノードと各ノードの `postgresql.conf` ファイルで設定する必要があります、変更を有効にするには再起動します。

最適化されたトポロジの要件が満たされない場合、加入者専用グループのノードはフルメッシュトポロジに戻ります。これが発生すると、クラスターメッセージ内のノードのログで、次のような最適化が不可能な理由を見つけます。

データノードがトップレベルのノードグループの一部である場合

```
node: <nodename> is part of top-level nodegroup: <toplevelgroupname>: changing to␣
↪full mesh".
```

データグループでプロキシルーティングが有効になっていない場合

```
node: <nodename> is in nodegroup: <nodegroupname> that does not have proxy routing:␣
↪changing to full mesh.
```

## 6.2 Node management

PGD クラスター内のすべてのデータノードは、1 つ以上のグループのメンバーです。デフォルトでは、すべてのデータノードは、PGD クラスター内のすべてのデータノードにまたがるトップレベルグループのメンバーです。ノードは、PGD クラスターの論理または地理的組織を反映するように構成できるサブグループに属することもできます。

ノードとサブグループで使用可能なさまざまなオプションを使用して、ノードとグループを管理できます。

- *Creating PGD nodes* は、PGD クラスターに新しいノードを作成するために必要な手順をカバーしています。
- *Groups and subgroups* は、PGD でグループとサブグループがどのように動作するかについて詳しく説明します。
- *Creating and joining PGD groups* は、新しい PGD グループを作成する方法と、PGD ノードをそれらに参加する方法を検討します。
- *Viewing PGD topology* は、PGD クラスターのノードとグループの構造を表示できるコマンドと SQL クエリーの詳細を示しています。
- *Removing nodes and groups* は、グループからノード、またはクラスターからグループを安全に削除するために実行するプロセスを示します。
- *Node restart and down node recovery* は、障害またはスケジュールされたダウンタイムの後にノードをサービスに戻すために必要な手順と、それが返すクラスターに与える影響について詳しく説明します。
- **自動同期との関係** は、PGD での自動同期機能がどのように動作し、それを使用してノード間の同期を維持する方法を検討します。
- *Node UUIDs* は、ノードの UUID が PGD で使用される方法と、それらがどのように生成されるかについて説明します。
- *Replication slots created by PGD* は、PGD が動作しているときに Postgres レプリケーションスロットがどのように消費されるかを調べます。

### 6.2.1 Creating PGD nodes

#### それはただの Postgres です

PGD ノードは、 BDR 拡張機能がインストールされた Postgres インスタンスです。BDR 拡張機能は、ノード間の双方向レプリケーションを有効にし、PGD の基盤です。

つまり、最も一般的な言葉では、Postgres と BDR 拡張機能をインストールし、PGD グループ内の他のノードに接続するようにノードを構成することにより、PGD ノードを作成できます。ただし、考慮すべき詳細がいくつかあります。

#### Postgres のバージョンはどれですか?

PGD は Postgres の上に構築されているため、PGD ノードに使用する Postgres のディストリビューションとバージョンが重要です。使用する Postgres のバージョンは、使用している PGD のバージョンと互換性がある必要があります。互換性マトリックスは release notes で見つけることができます。PGD の機能は、使用している Postgres のディストリビューションに依存する場合があります。EDB Postgres Advanced Server は、PGD の推奨ディストリビューションです。 PGD は、EDB Postgres Extended Server および Community Postgres もサポートしています。

### Postgres のインストール

構成する各ノードに選択した Postgres ディストリビューションをインストールする必要があります。EDB Postgres Advanced Server documentation 、 EDB Postgres Extended Server documentation 、および *Postgres installation documentation* で、各ディストリビューションのインストール手順を見つけることができます。Postgres のインストールをカバーする *Installing the database and pgd* を参照することもできます。

### BDR 拡張機能のインストール

BDR 拡張機能は、PGD の分散アーキテクチャのキーとなります。PGD クラスター内の各ノードに BDR 拡張機能をインストールする必要があります。BDR 拡張機能は、 EDB Postgres 分散リポジトリから入手できます。EDB リポジトリを Linux のパッケージ管理システムに追加し、PGD パッケージをインストールする必要があります。リポジトリー構成手順は *Configuring PGD repositories* にあります。

リポジトリの構成が完了したら、パッケージマネージャーを使用して PGD パッケージをインストールできます。パッケージ名は `edb-pgd6-<edition>-pg<postgres_distribution><version>` で、ここで `<postgres_distribution>` は使用している Postgres ディストリビューションであり、`<version>` は使用している Postgres のバージョンです。たとえば、コミュニティ Postgres 14 を使用しており、PGD Essential をインストールする場合、パッケージ名は `edb-pgd6-essential-pg14` です。パッケージ名前の詳細については、*Installing the database and pgd* を参照してください。

### PGD のデータベースの構成

このプロセスは PGD に固有であり、BDR 拡張機能と連携するように Postgres インスタンスを構成し、PGD クラスターと連携するようにさまざまな設定の調整が含まれます。手順は次のとおりです。

- `postgresql.conf` の `shared_preload_libraries` 設定の先頭に BDR 拡張`$libdir/bdr` を追加します。
- `postgresql.conf` で `wal_level` GUC 変数を `logical` に設定します。
- `postgresql.conf` で `track_commit_timestamp` を`'on'` に設定することにより、コミットタイムスタンプの追跡をオンにします。
- `postgresql.conf` で `max_worker_processes` ～`'16'` を設定することにより、最大ワーカープロセスを 16 以上に増やします。
- ルールを `pg_hba.conf` に追加して、ノードが相互に接続できるようにします。 - これらの行が `pg_hba.conf` に存在することを確認します。

```
host all all all md5
host replication all all md5
```

- `.pgpass` ファイルを追加して、ノードが相互に認証できるようにします。
- 他のノードにログインするための十分な特権を持つユーザーを構成します。
- `.pgpass` ファイルの詳細については、Postgres ドキュメントの

The Password File を参照してください。

これらの手順が完了したら、Postgres インスタンスを再起動して変更を適用します。

#### レプリケートデータベースのセットアップ

Postgres 内にデータベースを作成して、PGD のレプリケートデータベースとして使用します。これは、CREATE DATABASE コマンドで行うことができます。作成されるデータベースは、PGD クラスター内の他のノードがレプリケートするデータベースの名前である必要があります。通常、データベースの名前は pgddb です。

BDR 拡張機能を構成およびセットアップした特定のデータベースインスタンスにログインします。これは、スーパーユーザーとして CREATE EXTENSION bdr; コマンドを実行して行うことができます。このコマンドは、そのデータベース内に BDR 拡張機能を作成します。

#### 次のステップ

ノードの構成が完了し、PGD クラスターでグループに参加またはグループを開始する準備が整います。*Creating and joining PGD groups* セクションで、ノードをグループに参加させる手順を見つけます。

### 6.2.2 Groups and subgroups

#### グループ

PGD クラスターのノードはグループに集められます。「トップレベル」グループは常に存在し、すべてのデータノードが自動的に属するグループです。「トップレベル」グループは、サブグループの直接の親になることもできます。

#### サブグループ

グループには、0 個以上のサブグループを含めることもできます。サブグループは、データセンターまたは場所を表すために使用でき、コミットスコープが特定のリージョンのノードを全体として参照できます。接続マネージャーは、サブグループを利用して、書き込みリーダーになることができるノードを描写することもできます。

node_group_type 値は、サブグループの作成時のタイプを指定します。一部のサブグループタイプは、グループ内のノードの動作を変更します。たとえば、**サブスクライバー専用ノード** サブグループは、グループ内のすべてのノードをサブスクライバ専用ノードにします。

#### サブグループごとの接続文字列の構成

Kubernetes およびクラウド環境では、同じサブグループ内のノードは、ノードのプライマリ node_dsn として公開されていない、より高速または安価なローカルネットワークパスにアクセスすることがよくあります。ノードで node_group_dsn を設定すると、同じサブグループ内の他のノードに、代わりにその DSN を使用するように指示します。他のサブグループのノードは、引き続き node_dsn を使用します。

*bdr.alter_node_option* を使用して node_group_dsn を設定します。

```
SELECT bdr.alter_node_option(node_eu1, group_dsn, host=foo dbname=bdrdb port=5432);
```

NULL を渡してリセットします。

```
SELECT bdr.alter_node_option(node_eu1, group_dsn, NULL);
```

node_dsn エンドポイントは、初期セットアップのためにすべてのノードからアクセス可能なままである必要があります。ノードごとの現在構成されている有効な値については、*bdr.node_config_summary* を参照してください。

### 6.2.3 Creating and joining PGD groups

#### PGD グループの作成と参加

PGD の場合、すべてのノードは他のすべてのノードに接続する必要があります。構成を簡単にするために、新しいノードが参加すると、既存のすべてのノードがそれに接続するように構成します。このため、最初に作成した PGD ノードを含むすべてのノードは、

PostgreSQL connectionstring

他のノードが接続するために使用できるもの。この接続文字列は、データソース名 DSN と呼ばれることがあります。

両方の形式の接続文字列がサポートされています。したがって、`host=myhost port=5432 dbname=mydb` のようなキー-値形式、または `postgresql://myhost:5432/mydb` のような URI 形式を使用できます。

SQL ファンクション bdr.create_node_group()

ローカルノードから PGD グループを作成します。これにより、そのノードで PGD がアクティブになり、他のノードがその時点で 1 つのノードのみで構成される PGD グループに参加できるようになります。作成時に、このノードへの接続に使用する他のノードの接続文字列を指定する必要があります。

最初の PGD ノードグループが作成されると、デフォルトのレプリケーションセットも作成され、既存のすべてのテーブルがセットに追加されます。その後、すべてのテーブルの `bdr.default_replica_identity` 値がテーブルの `REPLICA IDENTITY` として構成されます。

ノードグループが作成されると、

bdr.create_node() を使用してさらにノードを作成し、

bdr.join_node_group() ファンクションを使用して PGD グループに参加できます。

または、コマンドラインユーティリティ bdr_init_physical を使用して、pg_basebackup を使用して新しいノードを作成します。 pg_basebackup を使用する場合、bdr_init_physical ユーティリティはオプションでターゲットデータベースのみのベースバックアップを指定できます。以前の動作は、データベースクラスター全体をバックアップすることでした。このユーティリティを使用すると、アクティビティがより速く完了し、不要なデータベースを除外するため、使用する領域が少なくなります。ターゲットデータベースのみを指定した場合、除外されたデータベースは新しいノードでクリーンアップおよび削除されます。

新しい PGD ノードが既存の PGD グループに参加するか、ノードが上流ピアをサブスクライブする場合、レプリケーションを開始する前に、システムは既存のデータをピアノードからローカルノードにコピーする必要があります。このコピーは、ローカルデータとリモートデータが同一に始まるように、注意深く調整する必要があります。pg_dump を自分自身で使用するだけでは十分ではありません。BDR 拡張機能は、この初期コピーを作成するためのビルトイン機能を提供します。

参加プロセス中に、BDR 拡張機能は提供されたソースノードをベースとして使用して既存のデータを同期し、PGD グループのメッシュトポロジで自分自身を確立するために必要なすべてのメタデータ情報を作成します。この初期コピー中にソースと新しいノード間の接続が切断された場合、参加プロセスを最初からリスタートします。

クラスターに参加するノードには、PGD グループのデータベースに既に存在するスキーマまたはデータが含まれていない必要があります。新しく参加するデータベースは、BDR 拡張機能を除き空であることをお勧めします。ただし、必要なデータベースユーザーとロールがすべて作成されていることが重要です。また、非スーパーユーザーが参加操作を実行している場合、スーパーユーザーの権限を必要とする拡張機能を手動で作成する必要があります。詳細については、**接続とロール** を参照してください。

オプションで、 bdr.join_node_group の `synchronize_structure` パラメーターを使用して、スキーマの同期をスキップできます。

ファンクション。この場合、スキーマは新しく参加するノードに既に存在する必要があります。

最適な接続論理的に近く、理想的には低遅延、高帯域幅を持つソースノードを、参加するソースノードとして選択することをお勧めします。そうすることで、結合が完了するまでに必要な時間が短縮されます。

Raft コンセンサスアルゴリズムを使用して参加手順を調整します。これには、既存のノードのほとんどがオンラインで到達可能である必要があります。

論理結合プロシージャ bdr.join_node_group を使用します

ファンクションは、`COPY` 操作を実行してデータ同期を実行し、マルチプルのライターパラレル適用が有効になっている場合を使用します。

ノード参加は、参加にかかる時間の大部分で他のノード参加と同時に実行できます。ただし、一度に 1 つのレギュラーノードのみが PROMOTE または PROMOTING の状態になることができます。他のすべてのノードが起動して実行されている場合、これらの状態は通常かなり短いです。それ以外の場合、結合はこの段階でシリアル化されます。サブスクライバ専用ノードはこのルールの例外であり、それらは同時に PROMOTE および PROMOTING 状態になることもできるため、それらの参加プロセスは完全にコンカレントです。

参加プロセスはソースとして 1 つのノードのみを使用するため、ノードの大部分が利用可能な場合、ノードがダウンしているときに実行できます。このアプローチでは、論理結合を実行するときに複雑さが発生する可能性があります。論理結合中に、ソースノードからコピーされた行のコミットタイムスタンプは、ソースノードの最新のコミットタイムスタンプに設定されます。これより前のコミットタイムスタンプを持つノードでコミットされた変更は、ノードがダウンしているか、大幅な遅延があるため、他のノードからの変更と競合する可能性があります。この場合、新しく参加したノードは他のノードとは異なる方法で解決され、発散が発生する可能性があります。その結果、ノード間に重大なレプリケーションラグが存在する場合は、ノード参加を実行しないことをお勧めします。これが必要に応じて、すべてのノードが利用可能になって追いついた後、新しく参加したノードで LiveCompare を実行してデータの相違を修正します。

`pg_dump` は、ソースノードに同時の DDL アクティビティがある場合、キャッシュルックアップの失敗が原因で失敗する場合があります。bdr.join_node_group は内部で pg_dump を使用するため、ソースノードに同時の DDL アクティビティがある場合、失敗する可能性があります。その場合、結合を再試行すると機能します。

### 6.2.4 Viewing PGD topology

#### PGD グループのリスト

#### pgd-cli の使用

pgd-cli `groups list` コマンドを使用して、PGD クラスター内のすべてのグループをリストします。

```
pgd groups list
```

```
Group Name  Parent Group Name Group Type     Nodes
----------- ----------------- --------------- -----
bdrgroup    bdrgroup          global          0
group_a     bdrgroup          data            4
group_b     bdrgroup          data            4
group_c     bdrgroup          data            1
group_so    bdrgroup          subscriber-only 1
```

#### SQL を使用する

次の単純なクエリーは、現在のノードがメンバーであるすべての PGD ノードグループをリストします。現在、bdr.local_node_summary から 1 行のみを返します。

```
SELECT node_group_name
FROM bdr.local_node_summary;
```

より複雑なクエリーを使用して、各ノードグループの構成を表示できます。

```
SELECT g.node_group_name
, ns.pub_repsets
, ns.sub_repsets
, g.node_group_default_repset     AS default_repset
, node_group_check_constraints    AS check_constraints
FROM bdr.local_node_summary ns
JOIN bdr.node_group g USING (node_group_name);
```

### PGD グループ内のノードのリスト

#### pgd-cli の使用

nodes list コマンドを使用して、PGD クラスター内のすべてのノードをリストします。

```
pgd nodes list
```

```
Node Name           Group Name Node Kind       Join State Node Status
------------------- ---------- --------------- ---------- -----------
bdr-a1              group_a    data            ACTIVE     Up
bdr-a2              group_a    data            ACTIVE     Up
logical-standby-a1  group_a    standby         ACTIVE     Up
witness-a           group_a    witness         ACTIVE     Up
bdr-b1              group_b    data            ACTIVE     Up
bdr-b2              group_b    data            ACTIVE     Up
logical-standby-b1  group_b    standby         ACTIVE     Up
```

(次のページに続く)

(前のページからの続き)

```
witness-b           group_b    witness         ACTIVE     Up
witness-c           group_c    witness         ACTIVE     Up
subscriber-only-c1  group_so   subscriber-only ACTIVE     Up
```

グループ名で grep を使用して、リストを特定のグループにフィルタリングします。

```
pgd nodes list | grep group_b
```

```
bdr-b1             group_b  data      ACTIVE     Up
bdr-b2             group_b  data      ACTIVE     Up
logical-standby-b1 group_b  standby   ACTIVE     Up
witness-b          group_b  witness   ACTIVE     Up
```

#### SQL を使用する

bdr.node_summary ビューから、特定のノードグループ mygroup`などのすべてのノードのリストを抽出できます。例

```
SELECT node_name          AS name
, node_seq_id             AS ord
, peer_state_name         AS current_state
, peer_target_state_name AS target_state
, interface_connstr       AS dsn
FROM bdr.node_summary
WHERE node_group_name = mygroup;
```

### 6.2.5 Removing nodes and groups

#### PGD グループからのノードの削除

PGD は、長期のノードの停止から復旧するように設計されているため、ノードを完全に削除するかどうかをシステムに明示的に指示する必要があります。ノードを完全にシャットダウンし、他のノードに通知しない場合、パフォーマンスが低下し、最終的にシステム全体が動作を停止します。

ノードの削除は *parting* とも呼ばれ、

bdr.part_node() を使用して実行されます

ファンクション。ノードを削除するには、ノードの作成中に渡されるノード名を指定する必要があります。

bdr.part_node() ファンクションは、削除するノードを含む、PGD グループのアクティブなノードから呼び出すことができます。

参加手順と同様に、分割は Raft コンセンサスを使用して実行され、動作するにはノードの大部分がオンラインである必要があります。

分離プロセスはすべてのノードに影響を与えます。Raft リーダーはノード間の投票を管理して、どのノードが分離ノードからの最新のデータを持つかを確認します。次に、残りのすべてのノードが最新のノードへのセカンダリの一時的な接続を作成して、不足しているデータをキャッチアップできるようにします。

分離されたノードは引き続き PGD に知られていますが、リソースを消費しません。ノードは、分割されたノードと同じ名前で再度追加される場合があります。まれに、ファンクション

bdr.drop_node() を使用して、分割されたノードのすべてのメタデータをクリアすることが必要になる場合があります。

#### PGD グループ全体の削除

通常、PGD グループは場所にマッピングされます。ロケーションが展開されない場合、そのロケーションの PGD グループも削除する必要がある可能性があります。

削除される PGD グループは空である必要があります。グループを削除する前に、グループ内のすべてのノードを分割する必要があります。

### 6.2.6 Node restart and down node recovery

PGD は、ノードの再起動またはノードの切断から復旧するように設計されています。切断されたノードは、各ピアノードに再接続し、そのノードから欠落しているデータを複製することにより、グループに再参加します。

ノードが起動すると、各接続は `bdr.node_slots.state = catchup` を使用して bdr.node_slots に表示され始め、欠落データの複製を開始します。キャッチアップは、各ピアノードからの欠落データの量に応じて継続され、サーバーのワークロードに応じて、時間の経過とともに増加する可能性があります。

各ノードの書き込みアクティビティの量が均一でない場合、より多くのデータを持つノードからのキャッチアップ期間は他のノードよりも大幅に長くかかる場合があります。最終的に、スロットの状態は `bdr.node_slots.state = streaming` に変更されます。

時間や日などの長期間オフラインになっているノードは、さまざまな理由でリソースの問題を発生させる可能性があります。次の問題を理解することなく、長期の停止を計画しないでください。

各ノードは変更情報を保持します 1 つの replication slot を使用

ピアノードごとに) したがって、後で一時的に到達不能なノードへの変更を再生できます。ピアノードが無期限にオフラインになったままである場合、この蓄積された変更情報により、最終的にノードは PostgreSQL トランザクションログ `pg_wal` の *WAL* 用の記憶領域が不足し、これに似たエラーでデータベースサーバーがシャットダウンする可能性があります。

```
PANIC: could not write to file "pg_wal/xlogtemp.559": No space left on device
```

または、他のディスク不足関連の症状を報告する場合があります。

さらに、オフラインノードのスロットもカタログ xmin を抑制し、カタログテーブルのバキュームを防止します。

EDB Postgres Extended Server および EDB Postgres Advanced Server では、オフラインノードもデータのフリーズを抑制して、競合解決データの損失を防ぎます *bdr_read_all_conflicts* を参照してください。

管理者は、ノードの停止を監視し *Monitoring the Connection Manager* を参照して、ノードに十分な空きディスク領域があることを確認する必要があります。ワークロードが予測可能な場合は、時間の経過とともに使用される領域の量を計算できるため、重要な問題が発生する前にノードがダウンできる最大時間を予測できます。

PGD によって作成されたレプリケーションスロットを手動で削除しないでください。そうした場合、クラスターは損傷し、*Replication slots created by PGD* で説明しているように、スロットを使用していたノードをクラスターから分離する必要があります。

ノードがオフラインである間、他のノードはオフラインノードから同じデータのセットをまだ受け取っていない可能性があるため、これはノード全体でのわずかな相違として現れる場合があります。分離プロセスは、ノード間のこの不均衡を修正します。

パートキャッチアップと呼ばれる分離のフェーズ中に、オフラインノードに関して他のすべてのノードから最も遠いノードが選択されます。他のノードがこの最も遠いノードに関して等しく追いついていない場合、最も先のノードをソースとして、オフラインノードをオリジンとして、等しく追いついていない各ノードをターゲットとして同期が開始されます。同期は、基本的にターゲットノードからソースノードの最前方ノードへのサブスクリプションであり、オフラインノードオリジンからターゲットノードに変更を転送します。

他のノードがどれだけ遅れているかによっては、この同期は別れるときに時間がかかる場合があります。同期が完了し、すべてのノードが等しく追いつくと、分割はノードの分割に進みます。この同期がなければ、強制的な部分が実行されると、クラスターの状態は一貫していない場合があります。これは、データが分岐する可能性があることを意味します。自動同期機能により、ノードがオフラインになると、これが検出され、すべてのノードが同期プロセスによってこのオフラインノードに関して等しくキャッチアップされます。これにより、ノードが分離するまで待機してデータの一貫性を保証する必要がありません。

ノードが長期間ダウンしているか、永久に失われている場合、推奨される戦略は、障害が発生したノードを分離し、その場所に新しいノードに再参加することです。新しいノードは、元のノードと同じ名前を使用できます。

### 6.2.7 Automatic synchronization

#### 同期の自動トリガー

BDR マネージャープロセスは、同期要求の自動トリガーを実行します。*bdr.raft_global_election_timeout* デフォルト 6 秒を超えてノードからの更新がない場合、ダウンしているとみなされます。

ノードは互いの近さについてチェックされます。すべてのノードが等しく追い付いている場合、同期は必要ありません。そうでない場合、「ダウン」ノードから最も前にあるノードがソースとして選択されます。ソースが決定されると、各ターゲット-オリジンとソース以外のノード-に対して、同期要求が設定されます。Witness ノードとスタンバイノードは、同期のターゲットである必要はありません。

ビュー bdr.sync_node_requests_summary は、同期要求を追跡します。

- `Origin` オリジンノードはダウンノードです。
- `Source` ソースノードは、オリジンから最も遠いノードです。
- `Target` オリジンに関してソースの背後にある他の各ノード。
- `Sync_start_lsn` 同期が開始されたときにオリジンからターゲットが受信した最大 LSN。
- `Sync_end_lsn` 同期が終了したときのオリジンからのターゲットノードのターゲット LSN。
- `Sync_status` 同期のステータス。
- `Sync_start_ts` 同期が開始された時刻。

同期要求がカタログに入力されると、完了まで転送されます。

### キャンセル

選択したソースノードがダウンしていることが判明した場合、マネージャーは同期操作をキャンセルします。これは、他のノードが起動している可能性があるため、最も遠いとは言えませんが、少なくとも一部のターゲットよりも前にあります。また、ノードを同期するために使用される場合があります。したがって、マネージャーはダウンしたノードをソースとして持つすべての同期操作をキャンセルし、ダウンしていない別のノードを同期のソースとして選択します。成功した同期とキャンセルされた同期のステートマシンを以下に説明します。

同期キャンセル API

bdr.sync_node_cancel() は、手動でのみ使用され、同期要求が何らかの理由でスタックし、クラスターの通常の機能をブロックしている場合にのみ使用されます。

```
select bdr.sync_node_cancel(origin, source)
```

これにより、指定されたオリジンとソースを持つすべてのターゲットのすべての同期ノード要求がキャンセルされます。これは、書き込みリードからのみ呼び出すことができます。

### 同期要求のライフサイクル

単一の同期要求には、オリジン、ソース、ターゲット、および到達する sync_end_lsn があります。同期要求はさまざまな状態を経て、各状態は別のノードで実行されます。

状態は次のとおりです。 **setup** 書き込みリードで実行します。sync_end_lsn を除く同期のフィールドを設定します。 **setup_source** ソースで実行します。 sync_end_lsn を設定し、同期サブスクリプションのスロットを作成します。**setup_target** これはターゲットノードで実行されます。この状態では、オリジンへの元のサブスクリプションは無効になっています。同期サブスクリプションはターゲットで設定され、オリジンの変更をソースノードからターゲットに転送します。**start** これはターゲットで実行されます。ターゲットの進行状況を監視して、sync_end_lsn に到達したかどうかを確認し、到達した場合は同期状態に移動します。**synced** サブスクリプションは sync_end_lsn に同期しました。この状態では、スロットはドロップされます。 **complete** この状態はターゲットで実行されます。この状態では、同期サブスクリプションがターゲットでドロップされ、元のサブスクリプションが有効になります。次に、done 状態に移行します。**done** これは、同期が成功したことを意味します。 **cancel start** これはターゲットノードで実行されます。この状態では、同期サブスクリプションは後でドロップする準備として無効になり、オリジンへの元のサブスクリプションが再度有効になります。**cancel resume** これはターゲットノードで実行されます。この状態では、同期サブスクリプションがドロップされます。**cancel denied** これはソースノードで実行されます。この状態では、スロットはドロップされます。**failed** キャンセルが発生し、すべてのクリーンアップが完了した場合、同期はこの状態で終了します。これは、同期が発生せず、再試行する必要があることを意味します。

同期のキャンセルは、選択したソースノードがダウンしていることが判明した場合に、自動的に発生する場合があります。キャンセル中に、サブスクリプションとスロットをクリーンアップし、元のサブスクリプションを有効にする必要があります。ソースまたはターゲットノードがダウンしている場合、同期要求はストールする可能性があります。

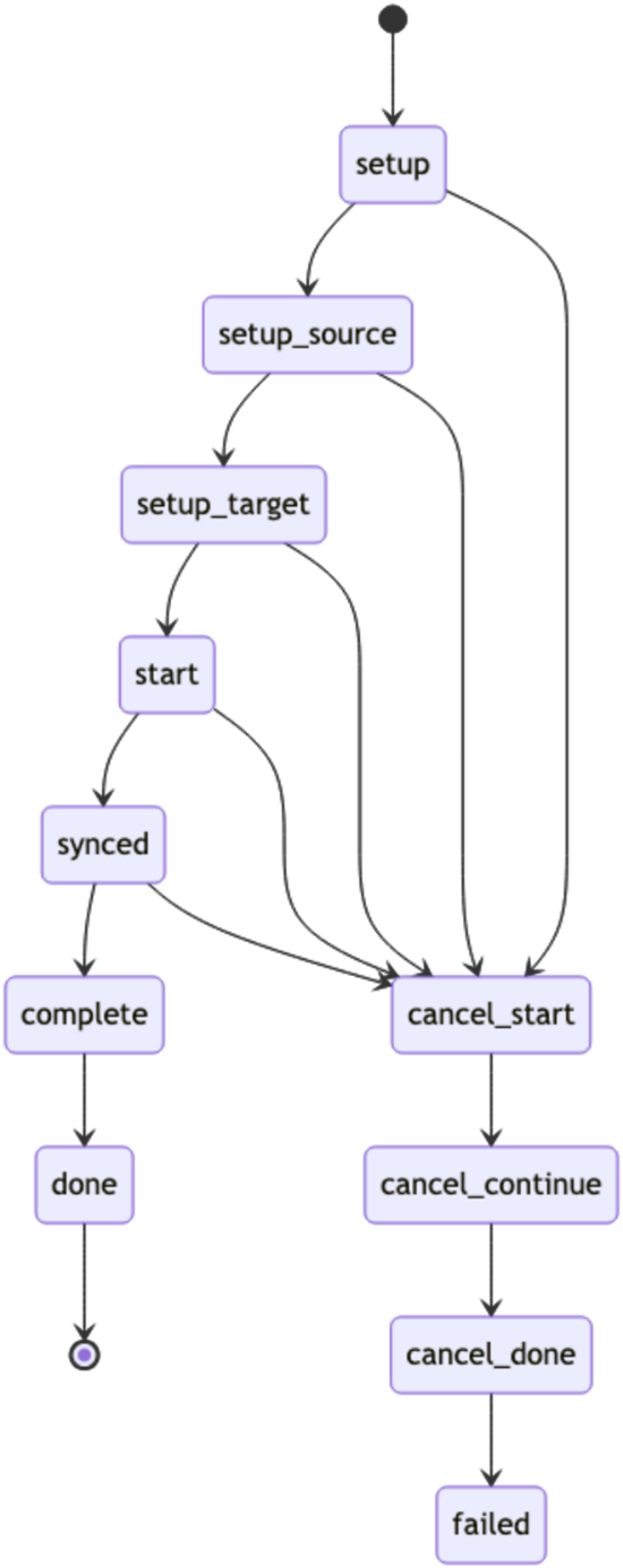


図 1: Autosync Request Life Cycle

#### GUC

自動同期を制御する GUC は bdr.enable_auto_sync_reconcile で、デフォルトで true に設定されています。オフにするには、すべてのノードで false に設定し、サーバーを再起動する必要があります。

### 6.2.8 Node UUIDs

PGD 6 では、各ノードに、クラスター内のノードを識別するために使用される UUID があります。この UUID は、ノードの作成時に生成され、そのノードに一意です。UUID は、PGD のさまざまな場所で見つけることができます。

- bdr.node テーブル。クラスター内の各ノードに関する情報が含まれています。
- bdr.node_summary ビュークラスター内のノードの人が判読可能なビューを提供します。
- bdr.local_node テーブル、ローカルノードに関する情報が含まれています。
- uuid 値は、ノードごとに作成されるレプリケーションスロットの名前付けにも表示されます。

PGD のノード管理で使用されますが、UUID の使用は PGD の既存の機能に影響を与えません。UUID は、ノードとグループを識別するために内部で使用され、ユーザーが PGD と対話する方法を変更しません。

#### なぜ UUID なのか?

UUID は PGD で使用され、クラスター内の各ノードに一意の識別子を提供します。PGD の以前のバージョンは、ノード名を識別子として使用していましたが、2 つのノードの名前が同じである場合、競合が発生する可能性がありました。UUID を使用することにより、PGD は各ノードに時間の経過とともに変更されない一意の識別子を持つことを保証できます。これは、クラスターにノードが頻繁に追加または削除される可能性がある PGD のような分散システムでは特に重要です。UUID により、新しいノードは既存のノードと同じ名前を持つ場合がありますが、異なる UUID を持ち、既存のノードと競合しないことが保証されます。

#### UUID はどのように生成されますか?

新しいノードが作成されると、そのノードの UUID が生成されます。この UUID は、カーネルの強力な乱数ジェネレーターを使用して作成され、一様にランダムであることが保証されています。この保証により、UUID は一意であり、簡単に推測できないことが保証されます。生成された UUID は bdr.node テーブルに保存され、クラスター内のノードを識別するために使用されます。

#### ノードが削除され、置換が追加された場合はどうなりますか?

ノードがクラスターから削除され、置換ノードが追加されると、置換ノードには新しい UUID が割り当てられます。これにより、置換ノードがクラスター内の別のエンティティとして扱われ、既存のノードと競合しないことが保証されます。ただし、PGD は、新しいノードを受け入れる前に、古いノードをクラスターから完全に分離する必要があります。置換ノードの UUID は、クラスター内の他のノードの UUID と同じ方法で使用されます。

#### PGD 6 の UUID 関連の変更

- 以前にノードを区別するために使用されていた bdr.node テーブルの generation フィールドは、使用されなくなりました。すべてのノードで 0 のままです。
- bdr.node テーブルの node_uuid フィールドは、PGD 6 で null になることはありません。将来、混合バージョンクラスターで null になる可能性があります。

### 6.2.9 Replication slots created by PGD

PGD の以前のバージョンでは、レプリケーションスロットには人間が判読できる名前がありました。PGD 6 は、ノードとグループの UUID の使用に切り替えて、より良い識別を保証します。

レプリケーションスロットは、PostgreSQL がレプリケーションの進行状況を追跡するために使用します。これらは、レプリケートされるデータが失われず、レプリケーションプロセスの一貫性を保証するために使用されます。PGD では、レプリケーションスロットを使用して、そのノードからのレプリケーションの進行状況を追跡します。ダウンストリームノードごとに 1 つのスロットがあります。特定のノードからすべてのダウンストリームノードにわたってグローバルにレプリケーションの進行状況を追跡するために使用される特別なレプリケーションスロットもあります。

- bdr_<topgroupuuid>_<dbhash> という名前のグループスロット 1 つ
- bdr_node_<targetnodeuuid>_<dbhash> という名前の N-1 ノードスロット。N は、クラスター内のノードの合計数です。ダイレクトロジカルスタンバイがある場合は、

topgroupuuid は、トップレベルグループの UUID - 文字未満の文字列表現であり、 dbhash はデータベース名のハッシュです。次を使用して、トップレベルグループの UUID を取得できます。

```
select node_group_uuid from bdr.node_group where node_group_parent_id=0;
```

dbhash は、データベース名のハッシュです。次を使用してハッシュを取得できます。

```
select to_hex(hashtext(pgddb));
```

targetnodeuuid は、ターゲットノードの UUID の文字列表現です- 文字未満です。次を使用してターゲットノードの UUID を取得できます。

```
select node_uuid from bdr.node where node_name=<target_node_name>;
```

完全なグループスロット名は、ファンクション bdr.local_group_slot_name() によって返されます。

> **警告**
>
> これらのスロットはドロップしないでください。 PGD はそれらを作成および管理し、必要に応じて削除します。

- bdr_ スロットで始まるスロットに直接触れないでください。
- スロット名を接頭辞 bdr_ で始めないでください。

グループスロットは、ノードからレプリケートされている PGD クラスター内のノードのレプリケーションの進行状況を追跡するために使用されます。PGD クラスター内の各ノードには、独自のグループスロットがあり、そのノードからのレプリケーションの進行状況を追跡するために使用されます。

グループスロットは次の目的で使用されます。

- 既存のノードをすべて起動して実行せずに、新しいノードを PGD グループに参加させますが、ノードの大部分は起動している必要があります。参加中にダウンしたノードが再び複製を開始した場合、このプロセスではデータの損失は発生しません。
- 一部のノードが分離されたノードに完全に追いついていない場合でも、クラスターからノードを一貫して分離します。
- フリーズポイントを抑えて、いくつかの競合の見逃しを回避します。

グループスロットは通常非アクティブで、他のノードからの Raft 進行状況メッセージに応答して定期的にのみ早送りされます。

**警告**

グループスロットをドロップしないでください。通常は非アクティブですが、 EDB Postgres 分散クラスターの適切なオペレーションには依然として重要です。これをドロップすると、PGD の機能の一部またはすべてが動作を停止するか、誤った結果が発生する可能性があります。

#### その他のスロット名

PGD 内の他の機能は、レプリケーションスロットを利用します。

たとえば、ノードがグループに追加されると、そのノードのスロットが作成され、レプリケーションプロセスの進行状況を追跡します。

このスロットの名前は `bdr_node_<targetnodeuuid>_<dbhash>_tmp` です。

PGD の分析およびデコード機能用に作成されたスロットもあります。これらのスロットには次の名前があります。

| Slot type | Slot name |
|---|---|
| Forwarding slot, leader-to-leader slot | *bdr_node_<targetnodeuuid>_<originidhex>_<dbhash>* |
| Analytics slot | *bdr_analytics_<groupuuid>_<dbhash>* |
| Decoding slot | *bdr_decoder_<topgroupuuid>_<dbhash>* |

## 6.3 Connection Manager

PGD 6.0 では、 PGD 5 のプロキシソリューションを、バックグラウンドワーカーを使用して緊密に統合されたアプローチで置き換える新しい接続マネージャーを導入し、PGD の読み書き、読み取り専用、および http ステータスネットワークインターフェイスを公開します。

- *Overview of durability options* は、Connection Manager の機能と利点をカバーしています。

- *Connection Manager Authentication* は、Connection Manager での認証がどのように動作するかを説明します。
- *Connection pooling* では、ビルトイン接続プーリングモードとその構成方法について説明します。
- *Load Balancing with Connection Manager* は、 Connection Manager でロードバランシングを使用する方法をカバーしています。
- *Configuring Connection Manager* は、使用可能な構成オプションとそれらの設定方法について詳しく説明します。
- *Monitoring the Connection Manager* は、モニタリングに使用可能なテーブルと HTTP エンドポイントをカバーします。

### 6.3.1 Connection Manager overview

#### 接続マネージャーについて

Connection Manager は、 EDB Postgres Distributed PGD 6.0 の新しいバックグラウンドワーカーであり、クライアントアプリケーションの単一のエントリポイントを提供することにより、PGD クラスターへの接続プロセスを簡素化します。PGD 5.x プロキシソリューションを、PGD の読み書き、読み取り専用、および HTTP ステータスネットワークインターフェイスを公開する緊密に統合されたアプローチに置き換えます。

Connection Manager は PGD に完全に統合されており、既存の PGD アーキテクチャとシームレスに動作するように設計されています。すべての PGD データノードには、着信接続をリッスンし、クラスター内の適切なノード、特にクラスター内の現在の書き込みリーダーにルーティングする Connection Manager インスタンスがあります。また、クラスターからデータを読み取ることのみが必要なアプリケーションに読み取り専用インターフェイスも提供します。

PGD 6.3.0 から、接続マネージャーはサブスクライバー専用ノードでデフォルトで起動し、主にレプリケートデータの受信専用のノードでも、高度な接続プーリングおよびセッション管理機能にアクセスできます。

Connection Manager は、以前は別個のツールを必要とした 3 つの機能を統合します。書き込みを現在のリーダーにルーティングし、読み取りトラフィックをノード全体に分散し、バックエンド接続をプールします。PGD 内で実行されるため、認証は `pg_hba.conf` で一度構成され、プロキシレイヤーとデータベースの両方に適用され、維持する個別のユーザーリストはありません。

#### 接続マネージャーを使用する

接続マネージャーは、デフォルトで Postgres サーバーの構成に従います。読み取り/書き込みポート、読み取り専用ポート、および HTTP ポートの 3 つのポートがあります。読み取り/書き込みポートは書き込み操作に使用され、読み取り専用ポートは読み取り操作に使用されます。HTTP ポートは、監視と管理の目的で使用されます。

読み取り/書き込みポートは、デフォルトで、Postgres ポート+1000 通常は 6432 に設定されます。読み取り専用ポートは、Postgres ポート+1001 通常は 6433 に設定されます。HTTP ポートは Postgres ポート+1002 通常は 6434 に設定されます。

接続マネージャーを使用するには、データノードの読み取り書き込みまたは読み取り専用ポートに接続するようにクライアントアプリケーションを構成します。接続マネージャーは、クラスター内の適切なノードに接続をルーティングします。

読み取り専用ポートに接続すると、接続が読み取り専用操作に制限されます。読み取り/書き込みに変更できないことを除き、

SET TRANSACTION READ ONLY 。transaction_read_only 構成パラメーターは、これらの接続で on を正しくレポートします。

#### 認証

接続マネージャーは、Postgres と同じ pg_hba.conf 構成を使用して TLS 終端と事前認証を実行し、維持する別のユーザーリストはありません。サポートされているメソッドとコンストレインについては、*Connection Manager Authentication* を参照してください。

#### 接続プーリング

接続マネージャーにはビルトインの接続プーリングが含まれており、ほとんどの展開で pgBouncer などの別個のプーラーの必要がありません。プーリングは、多くのクライアント接続で共有することによりバックエンドの Postgres 接続の数を減らし、データベースサーバーのメモリとプロセスのオーバーヘッドを削減します。リファレンスの詳細については *Connection pooling* を、DBA セットアップガイダンスについては *Configuring connection pooling* を、開発者ガイダンスについては *Using connection pooling in your application* を参照してください。

#### ロードバランシング

接続マネージャーは、ロードバランサーに代わるものではありません。外部ロードバランサーを使用して、別のノードで実行されている Connection Manager インスタンス間でクライアントトラフィックを分散します。構成の詳細と HAProxy の例については、*Load Balancing with Connection Manager* を参照してください。

#### 構成

接続マネージャーは、ポート、接続制限、コンセンサスタイムアウト、プールモードを含む PGD グループオプションを介して構成されます。完全なオプションのリファレンスについては、*Configuring Connection Manager* を参照してください。

#### モニタリング

接続マネージャーは、カタログビューと HTTP エンドポイントを介して接続数とプールの使用率を公開します。使用可能なメトリックについては、*Monitoring the Connection Manager* を参照してください。

### 6.3.2 Connection Manager Authentication

Connection Manager の認証は、Postgres 独自の pg_hba.conf ファイルを介して構成されます。 Connection Manager は、Postgres と同じ認証方法を使用します。

### Connection Manager 接続タイプ

接続マネージャーは、pg_hba.conf で次の接続タイプをサポートしています。

- host - TCP/IP 接続
- hostssl - SSL を使用した TCP/IP 接続
- hostnossl - SSL を使用しない TCP/IP 接続

### Connection Manager の認証方法

接続マネージャーは、pg_hba.conf で次の認証方法をサポートしています。

- trust - 認証なし
- reject - 接続を拒否する
- md5 - MD5 パスワード認証
- scram-sha-256 - SCRAM-SHA-256 パスワード認証
- cert - SSL 証明書認証
- ldap - LDAP サーバーを使用して認証する

> **注釈**
>
> Connection Manager は、クライアントユーザーとして PGD ノードに対して認証できる必要があります。レプリケーションと内部通信のために他の PGD ノードから発生した接続を受け入れるように、各 PGD ノードで `pg_hba.conf` ファイルを構成します。

証明書認証方法を使用する場合、接続マネージャーはサーバーキーを提示します。この証明書を Connection Manager アドレスから受け入れるように PGD ノードを構成する必要があります。

Connection Manager は、pg_hba.conf の user および database フィールドの正規表現マッチングもサポートしています。これにより、ユーザー名とデータベース名を照合するためのパターンを指定できるため、複数のユーザーとデータベースの認証の管理が簡単になります。

グループメンバーシップチェックもサポートされています。これにより、各ユーザーを個別に指定するのではなく、データベースに接続できるユーザーのグループを指定できます。

Connection Manager は、 cert 認証方法を使用する場合のユーザー ID マッピングに pg_ident.conf をサポートします。

### サポートされていない pg_hba.conf ルール

ルールが接続マネージャーでサポートされていない場合、警告としてログに記録され、無視されます。

#### LDAP 認証の注意事項と制約

接続マネージャーは、次のキー制約付きで、PostgreSQL と同じ LDAP 認証機能をサポートします。

- Connection Manager は、LDAP サーバーとの通信に OpenLDAP ライブラリを使用しません。その結果、標準の OpenLDAP 構成ファイル `ldap.conf` または `.ldaprc` など および環境変数は無視されます。
- Connection Manager は、 LDAPS または LDAP StartTLS を使用する場合、デフォルトで LDAP サーバー証明書を検証します。LDAP サーバーの SSL 証明書に必要な CA 証明書は、システムのトラストストアにある必要があります。
- `LDAPTLS_REQCERT=never` 環境変数を設定することにより、LDAP サーバーの SSL 証明書の検証を抑制します。この変数の他の値は無視されます。

### 6.3.3 Connection pooling

クライアント全体でバックエンド接続を再利用するように接続プーリングを構成し、サーバーメモリのオーバーヘッドを削減し、多くの短期間のセッションを含むワークロードのスループットを向上させます。Connection Manager のビルトインプーリングにより、ほとんどの展開で pgBouncer などの外部ツールの必要性がなくなります。

DBA セットアップと移行ガイダンスについては *Configuring connection pooling* を、開発者ガイダンスについては *Using connection pooling in your application* を参照してください。

#### プールモードの構成

`bdr.alter_node_group_option` を使用してノードグループのプールモードを設定します。

```
SELECT bdr.alter_node_group_option(mygroup, server_pool_mode, transaction);
```

または PGD CLI を使用して

```
pgd group mygroup set-option server_pool_mode transaction
```

`server_pool_mode` グループオプションは、プーリング動作を制御します。

| Mode | Behavior |
|---|---|
| *none* | プーリングはありません。各クライアント接続は、その有効期間全体にわたって専用のバックエンド接続を取得します。クライアントが切断すると、バックエンド接続が閉じます。このモードはデフォルトです。 |
| *session* | バックエンド接続は、最初に使用するときにクライアントに割り当てられ、クライアントが切断するとプールに返されます。接続マネージャーは、再利用のために接続を返し、セッション状態をリセットする前に *Replicated large objects* を実行します。 |

次のページに続く

表 2 – 前のページからの続き

| Mode | Behavior |
| --- | --- |
| *transaction* | バックエンド接続は、トランザクションの開始時に割り当てられ、トランザクションが終了するとプールに返されます。トランザクション間では、クライアントはバックエンド接続を保持しないため、他のクライアントが利用できるようにします。接続マネージャーは、再利用のために接続を返す前に *Replicated large objects* を実行し、セッション状態をリセットし、次のトランザクションでクライアントがサポートされている接続パラメーターを再適用します。 |

現在のプールモードは、 bdr.node_group_summary ビューの server_pool_mode 列に表示されます。

### セッションパラメーターの管理

接続マネージャーは、すべてのプールモードでバックエンドに特定の接続パラメーターのセットを転送します。認識するパラメーターは client_encoding 、DateStyle 、TimeZone 、standard_conforming_strings 、application_name 、search_path 、および extra_float_digits です。 -c name=value 構文を使用して、options 接続パラメーターを介して追加パラメーターを含めることができます。

トランザクション間でバックエンド接続が変更される可能性がある transaction モードでは、接続マネージャーは新しいバックエンドを割り当てるたびにこれらのパラメーターを再適用します。接続パラメーターのみが再適用されるため、SET コマンドを使用してトランザクションモードでセッションパラメーターを構成することは避けてください。SET で行われた変更は、バックエンド接続が変更されたときに保持されません。接続文字列の例については、*Using connection pooling in your application* を参照してください。

### 準備されたステートメントを使用する

拡張クエリプロトコルを介して送信されるプリペアドステートメントは、すべてのプールモードのトランザクションにわたってシームレスに動作します。接続マネージャーは、バックエンドで欠落しているプリペアドステートメントを自動的に検出し、オンデマンドで再準備し、クライアントとバックエンド間でステートメントの名前付けを透過的に管理します。

transaction モードでは、各トランザクションの終了にバックエンド接続がプールに返されたときに DISCARD ALL が実行され、PREPARE /EXECUTE SQL ステートメントで作成されたプリペアドステートメントの割り当てが解除されます。PREPARE /EXECUTE SQL ステートメントを使用するトランザクション内で発行するか、代わりに拡張照会プロトコルを使用します。具体的な手順については、開発者ガイドの **準備されたステートメントを使用する** を参照してください。

### 接続の再利用

session および transaction モードでは、クライアントが切断されるか、トランザクションが終了すると、接続マネージャーはバックエンド接続をプールに返します。クライアントがサポートされていない機能をトリガーするか、保留中のコマンドで接続を閉じると、接続マネージャーはそのバックエンド接続をプールに返すのではなく、破棄します。他のクライアントは影響を受けません。

#### 未サポートの機能

一部の Postgres 機能は、接続プーリングと互換性がないため、サポートされていません。

- セッション中に行われた SET 変更は、 transaction モードでは持続しません。代わりに、接続文字列でセッションパラメーターを構成します。詳細は、セッションパラメーターの管理 を参照してください。
- PREPARE /EXECUTE /DEALLOCATE SQL ステートメントは transaction モードではサポートされていません。 DISCARD ALL は、バックエンド接続がプールに返されたときに実行され、準備されたステートメントの割り当てが解除されます。これらのステートメントを使用するトランザクション内で発行するか、代わりに拡張照会プロトコルを使用します。詳細は、準備されたステートメントを使用する を参照してください。
- ホールダブルカーソル WITH HOLD は、transaction モードではサポートされていません。これらはトランザクション境界を超えて持続するため、トランザクションの終了時にバックエンド接続をプールに返すことと競合します。
- LISTEN は transaction モードではサポートされていません。LISTEN サブスクリプションは、単一のトランザクションを超えて持続します。これは、バックエンド接続をプールに返すことと競合します。
- トランザクション境界を越えて保持されるアドバイザリロックは、transaction モードではサポートされていません。
- トランザクションを超えてアクセスされる一時テーブルは、 transaction モードではサポートされていません。
- レプリケーション接続はすべてのプールモードで拒否されます。

### 6.3.4 Load Balancing with Connection Manager

外部ロードバランサーを使用して、複数の Connection Manager インスタンスにトラフィックを分散できます。接続マネージャーポートを他の Postgres ポートと同様に扱いますが、書き込みトラフィックには読み取り書き込みポートを使用し、読み取りトラフィックには読み取り専用ポートを使用します。

ロードバランサーによっては、構成方法について多くのオプションがあります。

#### Connection Manager ルーティング

Connection Manager でトラフィックを書き込みリーダーにルーティングする場合は、Connection Manager の読み取り/書き込みポートに接続します。接続マネージャーは、ノードがメンバーであるノードグループの書き込みリーダーにトラフィックをルーティングします。

#### HAProxy の例 Connection Manager ルーティング

HAProxy では、次の例のような構成を使用して、それぞれが異なるホストとポートで実行されている 3 つの Connection Manager インスタンス間でロードバランシングができます。haproxy ポート 5010 への読み取り書き込みトラフィックは、いずれかのノード上の Connection Manager の読み取り書き込みポートデフォルト 6432 に転送されます。haproxy ポート 5011 への読み取り専用トラフィックは、Connection Manager の読み取り専用ポートデフォルト 6433 に転送されます。ポート 5011 は読み取り専用トラフィックに使用され、ラウンドロビン方式でノード全体に分散されます。

```
global
    maxconn 100

defaults
    log     global
    mode    tcp
    retries 2
    timeout client 30m
    timeout connect 4s
    timeout server 30m
    timeout check 5s

listen stats
    mode http
    bind 127.0.0.1:7000
    stats enable
    stats uri /

listen read-write
    bind *:5010
    option pgsql-check user postgres
    default-server inter 3s fall 3 rise 2 on-marked-down shutdown-sessions
    server node1 host1:6432 maxconn 100 check port 6432
    server node2 host2:6432 maxconn 100 check port 6432
    server node3 host3:6432 maxconn 100 check port 6432

listen read-only
    balance roundrobin
    bind *:5011
    option pgsql-check user postgres
    default-server inter 3s fall 3 rise 2 on-marked-down shutdown-sessions
    server node1 host1:6433 maxconn 100 check port 6433
    server node2 host2:6433 maxconn 100 check port 6433
    server node3 host3:6433 maxconn 100 check port 6433
```

### ダイレクトルーティング

接続マネージャーをバイパスして、クラスター内のノードに直接接続することができます。このシナリオで接続マネージャーを活用するには、その HTTP エンドポイントを参照して、クラスター内のノードに関する情報を参照してください。どのノードが書き込みリーダーであるかなど。次に、そのノードに直接接続します。`/node/is-read-write` エンドポイントを使用して、ノードが書き込みリーダーであるかどうかを確認し、`/node/is-read-only` エンドポイントを使用して読み取り専用プールに関する情報を取得できます。

このようにロードバランサーを構成する主な理由は、サーバーで TLS 終端と認証が発生する従来の Postgres 接続に似たレイヤー 4 TCP プロキシが必要な場合です。接続マネージャーは、プロキシレベル、つまり接

続マネージャー内で TLS 終端と認証を処理するレイヤー 7 プロキシを提供します。

このアプローチの欠点は、接続マネージャーの機能、特に読み取り専用プールへの読み取り専用接続を強制する機能の恩恵を受けられないことです。接続は、書き込みリーダーではなく、非書き込みリーダーノードに転送されます。

使用可能なエンドポイントの詳細については、

Monitoring Connection Manager を参照してください。

### HAProxy の例 ダイレクトルーティング

この例は、HAProxy を使用して、正常性ステータスに基づいてトラフィックをノードに直接ルーティングする構成を示しています。構成は、/node/is-read-write および/node/is-read-only エンドポイントを照会することによりノードの状態を確認し、それに応じてトラフィックをルーティングします。前の例と同様に、この例では、ポート 5010 の読み取り書き込みトラフィックを書き込みリーダーであるノードにルーティングし、ポート 5011 の読み取り専用トラフィックを読み取り専用のノードにルーティングします。

```
global
    maxconn 100

defaults
    log     global
    mode    tcp
    retries 2
    timeout client 30m
    timeout connect 4s
    timeout server 30m
    timeout check 5s

listen stats
    mode http
    bind 127.0.0.1:7000
    stats enable
    stats uri /

listen read-write
    bind *:5010
    option httpchk GET /node/is-read-write
    http-check expect string true
    default-server inter 3s fall 3 rise 2 on-marked-down shutdown-sessions
    server node1 host1:5432 maxconn 100 check port 6434
    server node2 host2:5432 maxconn 100 check port 6434
    server node3 host3:5432 maxconn 100 check port 6434

listen read-only
    balance roundrobin
```

(次のページに続く)

(前のページからの続き)

```
    bind *:5011
    option httpchk GET /node/is-read-only
    http-check expect string true
    default-server inter 3s fall 3 rise 2 on-marked-down shutdown-sessions
    server node1 host1:5432 maxconn 100 check port 6434
    server node2 host2:5432 maxconn 100 check port 6434
    server node3 host3:5432 maxconn 100 check port 6434
```

この構成は、`/node/is-read-write` および`/node/is-read-only` エンドポイントを照会することにより、ノードの状態を確認します。正常で、それぞれ読み取り書き込みまたは読み取り専用操作に使用可能なノードにのみトラフィックをルーティングします。

### 6.3.5 Configuring Connection Manager

#### 接続マネージャーの構成

Connection Manager は、ノードがメンバーであるグループの PGD Group オプションから構成を取得します。

これらは、pgd group set-option コマンドまたは *bdr.alter_node_group_option()* を使用したグループ構成 SQL ファンクションを使用して構成できます。

接続マネージャーの構成には、次のオプションが使用できます。

| Option | Default | Description |
|---|---|---|
| read_write_port | Postgres's port + 1000 (usually 6432) | 読み取り/書き込み接続をリッスンするポート |
| read_only_port | read_write_port + 1 (usually 6433) | 読み取り専用接続をリッスンするポート |
| http_port | read_write_port + 2 (usually 6434) | REST API 呼び出しをリッスンする HTTP ポート統合の目的 |
| use_https | false | HTTP リスナーが HTTPS を使用するかどうか。有効になっている場合、サーバー証明書は TLS に使用されます |
| server_pool_mode | none | 接続プーリングモード。有効な値は`none`プーリングなし、*session*、および`transaction`です。 *Connection pooling* を参照してください。 |
| read_write_max_client_cor | max_connection | 許可される最大読み取り/書き込みクライアント接続、デフォルトは max_connections |
| read_write_max_server_co | max_connections | サーバーに対して開かれる最大読み取り/書き込み接続 |
| read_only_max_client_con | max_connections | 許可される最大読み取り専用クライアント接続 |
| read_only_max_server_cor | max_connections | サーバーに対して開かれる読み取り専用接続の最大数 |
| read_write_consensus_time | 0 (immediate action) | 読み取り/書き込み接続が受け入れられなくなるまでにコンセンサスが失われるまで待機する時間 |
| read_only_consensus_time | 0 (immediate action) | 読み取り専用接続が受け入れられなくなるまでに、コンセンサス損失を待機する時間。 |

次のページに続く

表 3 – 前のページからの続き

| Option | Default | Description |
|---|---|---|

**注釈**

接続マネージャーで HTTPS を有効にするには、*use_https* オプションを`true` に設定し、*postgresql.conf* またはコマンドラインを介して`ssl_cert_file` および`ssl_key_file` を構成する必要があります。グローバルサーバー SSL *ssl = on* が無効になっている場合でも、HTTPS はアクティブにできます。いずれかの証明書ファイルが欠落している場合、Connection Manager は`use_https` 設定に関係なく、デフォルトのプレーンテキスト HTTP を使用します。

### 6.3.6 Monitoring the Connection Manager

SQL クエリーと HTTP エンドポイントを介して、接続マネージャーとその接続のステータスを表示できます。

#### 使用可能な SQL テーブルとビュー

接続マネージャーは、接続マネージャーとその接続のステータスを監視するために使用できる多数のテーブルとビューを提供します。これらには次のものが含まれます。

- bdr.stat_activity _ これは、`connection_manager_client_addr` および `connection_manager_client_port` に関する追加列で強化された `pg_stat_activity` からの情報であり、接続が接続マネージャーを介して来ており、読み取り専用ポートを介して接続している場合は `session_read_only` です。
- bdr.stat_connection_manager _ これは、接続マネージャーのステータスに関する統計を提供するビューです。
- bdr.stat_connection_manager_connections _ これは、Connection Manager の接続に関する統計を提供するビューです。
- bdr.stat_connection_manager_node_stats _これは、各データノードの接続マネージャーに関する統計を提供するビューです。
- bdr.stat_connection_manager_hba_file_rules _これは、このノードで使用されている接続マネージャーの HBA ファイルルールを示すビューです。

#### 使用可能な HTTP / HTTPS エンドポイント

接続マネージャーは、HTTP API を介して監視できます。

true/false を返すエンドポイントは、true の場合 200 ステータスコード、false の場合 503 ステータスコードも結果ます。

次のエンドポイントが使用できます。

| Endpoint | Description |
|---|---|
| /connection/is-live | 接続マネージャーはライブリッスンしていますか。常に「true」を返します。マネージャーが実行されていない場合、クライアントは単に接続/URL を開くことに失敗します。 |
| /connection/is-ready | 接続マネージャーは接続を受け入れる準備ができていますか。true(200)/false(503) を返します。 |
| /connection/is-ready-ro | 接続マネージャーは読み取り専用接続を受け入れる準備ができていますか。true(200)/false(503) を返します。 |
| /connection/is-ready-rw | 接続マネージャーは読み取り/書き込み接続を受け入れる準備ができていますか。true(200)/false(503) を返します。 |
| /node/is-read-write | この PGD ノードは、接続マネージャーではなく PGD ノード自分自身、読み取り/書き込みノードですか、書き込みリーダーですか、true(200)/false(503) を返します。 |
| /node/is-read-only | この PGD ノードは、接続マネージャーではなく PGD ノード自分自身であり、読み取り専用ノード書き込みリーダーではなく、true(200)/false(503) を返します。 |
| /group/read-write-info | 接続マネージャーのこのインスタンスの読み取り/書き込みプールに関する情報を返します。ノード ID、ノード名、ノードホスト、ノードポート、ノード dbname を含む JSON 形式のプール内のノードのリスト。読み取り/書き込みプールの場合、プールには 1 つのエントリのみが含まれます。 |
| /group/read-only-info | 接続マネージャーのこのインスタンスの読み取り専用プールに関する情報を返します。ノード ID、ノード名、ノードホスト、ノードポート、ノード dbname を含む JSON 形式のプール内のノードのリスト。 |

以下は、`/group/read-write-info` エンドポイントからの応答本文の例です。

```
[
  {
    "id": 683485707,
    "name": "node-1",
    "host": "host-1",
    "port": 5432,
    "dbname": "pgddb"
  }
]
```

### ロギング

すべての Connection Manager ログメッセージは PostgreSQL ログに書き込まれます。

log_line_prefix の%r および%h エスケープシーケンスの動作が、それぞれ「proxy_address/client_address」および「proxy_port/client_port」をログに記録するように変更されました。

これは、使用するサーバー接続の GUC をプロキシ設定することにより実現されます。ユーザーはこの GUC をオーバーライドできるため、client_address から派生したセキュリティコンテキストは、完全なセッションログを参照して検証する必要があります。

## 6.4 Postgres configuration

いくつかの Postgres 構成パラメーターは PGD ノードに影響を与えます。ノードごとにこれらのパラメーターを異なるように設定できますが、一般的にはお勧めしません。

PGD 独自の設定については、PGD settings reference を参照してください。

### 6.4.1 Postgres 設定

PGD を正しく実行するには、次の Postgres 設定が必要です。

- `wal_level` _ PGD は論理デコードに依存するため、 `logical` に設定する必要があります。
- `shared_preload_libraries` _ 拡張機能を有効にするには `bdr` を含める必要があります。他のほとんどの拡張機能は、コンマ区切りリストの `bdr` エントリーの前後に表示できます。このリストには `pglogical` を含めないでください。
- `track_commit_timestamp` _ 競合する各行のタイムスタンプを取得するには、競合解決のために `on` に設定する必要があります。

PGD では、これらの PostgreSQL 設定を適切な値に設定する必要があります。これは、クラスターのサイズとスケールによって異なります。

- `logical_decoding_work_mem` _ 論理デコードによって使用されるメモリバッファサイズ。このサイズよりも大きいトランザクションはバッファをオーバーフローし、ローカルディスクに一時的に保存されます。デフォルトは 64MB ですが、かなり高く設定できます。
- `max_worker_processes` _ PGD はレプリケーションとメンテナンスタスクにバックグラウンドワーカーを使用するため、正しく動作するには十分なワーカースロットが必要です。各データベースのワーカーの正しい最小数の式は、次の値を合計することです。
- PostgreSQL インスタンスごとに 1 つ
- そのインスタンスのデータベースごとに 1 つ
- PGD 対応データベースごとに 4 つ
- PGD グループのピアノードごとに 1 つ
- ピアノードの数 回 the (ライターの数 bdr.num_writers) プラス 1 ノードが PGD グループから削除されるときに、一時的により多くのワーカープロセスが必要になる場合があります。
- `max_wal_senders` _ 各ピアノードに 2 つ必要です。
- `max_replication_slots` _ 各ピアノードに 2 つ必要です。
- `max_active_replication_origins` _ 各ピアノードに 3 つが必要です。PostgreSQL 18 以降で利用できます。
- `wal_sender_timeout` および `wal_receiver_timeout` _ 通常はデフォルトの 1 分で十分ですが、大規模なトランザクションの場合、処理にこの時間よりも長い時間が必要になる場合があります。WAL 送信者は、待機しているレプリケーション接続に送信する前に、トランザクションのフルサイズを処理する必要があるため、Postgres はそれをタイムアウトとして認識できます。問題が実際に大規模なトランザクションによるものである場合は、wal_sender_timeout を `3600s` 以上のような高い値に引き上げ、サーバーをリロードすることにより問題が解決される可能性があります。さらに、この設

定は、ノードが CAMO パートナーを切断または再接続したとみなすまでの時間を決定します。詳細は、*CAMO failure scenarios* を参照してください。

N 個のピアノードを持つグループの通常の実行では、PGD には N 個のスロットと WAL 送信機が必要です。同期中に、PGD は別の N-1 スロットと WAL センダーを一時的に使用するため、この時折のピーク要求に十分な高いパラメーターを設定するように注意してください。

Parallel apply がオンの場合、スロット数はフォーミュラ*ライターから N スロットに増やす必要があります。これは、`max_replication_slots` もレプリケーションオリジンの最大数を設定し、Parallel apply の一部の機能はライタごとに追加のオリジンを使用するためです。

*Decoding worker* が有効になっている場合、このプロセスには PGD グループごとに 1 つの追加のレプリケーションスロットが必要です。

`max_worker_processes` 、`max_wal_senders` 、および `max_replication_slots` パラメーターを変更するには、ローカルノードを再起動する必要があります。

古い同期レプリケーションモードは、次のパラメーターを使用してサポートされています。詳細と制限については、*Commit scopes* を参照してください。

- `synchronous_commit` および `synchronous_standby_names` _ PGD レプリケーションの耐久性とパフォーマンスに影響します。physical replication と同様の方法。

### 6.4.2 準備されたトランザクションの最大数

#### max_prepared_transactions

明示的な 2 フェーズコミット、CAMO、または Eager トランザクションのため、クラスター全体で同時に準備されたトランザクションの最大数に対処するために十分に高く設定する必要があります。制限を超過すると、ノードはローカル 2 フェーズコミットまたは CAMO トランザクションを実行できず、クラスター上のすべての Eager トランザクションを防止します。このパラメーターは、Postgres サーバーの起動時にのみ設定できます。

### 6.4.3 グローバル構成に関する考慮事項

特定の PostgreSQL 構成パラメーターはスタンドアロンインスタンスに役立ちますが、PGD クラスターの安定性、バックグラウンドワーカー、およびレプリケーションプロセスに悪影響を与える可能性があります。

次のパラメーターは、`postgresql.conf` でグローバルに有効にしたり、PGD ノードデータソース名 DSN にオプションとして含めたりする必要はありません。

- `idle_session_timeout`
- `transaction_timeout`

代わりに、これらのパラメーターをセッションごとに設定して、システムレベルの操作を干渉しないことを確認します。

## 6.5 Monitoring

レプリケーションセットアップの監視は、システムが次のことを保証するために重要です。

- 最適なパフォーマンス
- ディスク領域が不足しない
- オペレーションを停止する可能性のある他の障害に遭遇しない

自動監視を導入して、問題が発生したときに管理者が警告を受け取り、プロアクティブなアクションを実行できるようにすることが重要です。たとえば、レプリケーションスロットが大幅に遅延し始めた場合、管理者に警告を与えることができます。

EDB は、バージョン 8.1 から PGD をサポートする Postgres Enterprise ManagerPEM を提供します。詳細については、*Monitoring EDB PostgresDistributed* を参照してください。

または、ツールまたはユーザーは、BDR 拡張機能が提供する情報ビューと機能を独自に呼び出すことができます。*Monitoring through SQL* を参照

詳細については。

### 6.5.1 Monitoring through SQL

EDB Postgres Distributed は、分散の性質に固有のいくつかのモニタリングおよび統計ビューを提供します。標準の Postgres モニタリングは、EDB Postgres 分散をモニタリングするのにも役立ちます。

#### モニタリングの概要

PGD グループは、複数のサーバーで構成されており、多くの場合ノードと呼ばれます。すべてのノードを監視して、グループ全体の状態を確認します。

bdr_monitor ロールは、`bdr.monitor` ファンクションを実行して、3 つのレベルのいずれかを使用して PGD の状態の評価を提供できます。

- `OK` — 通常、緑色で表示されます。
- `WARNING` — 通常、黄色で表示されます。
- `CRITICAL` — 通常、赤で表示されます。
- `UNKNOWN` — 認識されない状況の場合、多くの場合赤で表示されます。

PGD は、さまざまな内部メトリックの瞬間的な状態を表示する動的なカタログビューも提供します。また、ユーザーが要求した構成のデフォルトと構成の変更を保存するメタデータカタログも提供します。これらのビューとテーブルの一部は、bdr_monitor または bdr_read_all_stats からアクセスできますが、一部には、より高いセキュリティ要件が必要なユーザーまたは内部情報が含まれています。

PGD を使用すると、各ノードを個別に監視したり、単一のノードへのアクセスでグループ全体を監視したりできます。各ノードを個別に監視する場合は、各ノードに接続し、監視要求を発行します。これらの要求は他のノードを呼び出してグループレベルの情報セットをアセンブルするため、単一のノードからグループをモニタする場合は、`bdr.group` で始まるビューを使用します。

bdr_superuser によって bdr.run_on_all_nodes() ファンクションへのアクセスが付与されている場合、すべてのノードに独自の呼び出しを行うことができます。

### 監視ノードの参加と削除

デフォルトでは、ノード管理機能は join または part 操作が完了するまで待機します。それぞれの wait_for_completion ファンクション引数を使用して、待機をオフにできます。待機がオフになっている場合、join または part オペレーションが終了したときを確認するには、bdr.node_summary および bdr.event_summary を使用して間接的にノードの状態を確認します。

ヘルパーファンクション bdr.wait_for_join_completion() は、呼び出されると、すべての未処理のノード結合操作領域が完了するまで PostgreSQL セッションを一時停止します。

この例は、 bdr.node_summary からの SELECT クエリーの出力を示しています。これは、2 つのノードがアクティブで、別の 1 つが参加していることを示します。

```
#  SELECT node_name, interface_connstr, peer_state_name,

#      node_seq_id, node_local_dbname

#  FROM bdr.node_summary;

- [ RECORD 1 ]-----+----------------------------------------
node_name          | node1
interface_connstr | host=localhost dbname=postgres port=7432
peer_state_name    | ACTIVE
node_seq_id        | 1
node_local_dbname | postgres
- [ RECORD 2 ]-----+----------------------------------------
node_name          | node2
interface_connstr | host=localhost dbname=postgres port=7433
peer_state_name    | ACTIVE
node_seq_id        | 2
node_local_dbname | postgres
- [ RECORD 3 ]-----+----------------------------------------
node_name          | node3
interface_connstr | host=localhost dbname=postgres port=7434
peer_state_name    | JOINING
node_seq_id        | 3
node_local_dbname | postgres
```

また、テーブル bdr.node_catchup_info は、ノードへの参加または分離に関連する可能性のあるキャッチアップ状態に関する情報を提供します。

ノードが分離されると、クラスター内の一部のノードがその分離したノードからすべてのデータを受信しない場合があります。したがって、ノードを分離すると、そのデータを既に受信し、転送できるノードから一時的なスロットが作成されます。

catchup_state は、次のいずれかです。

```
10 = setup
20 = start
30 = catchup
40 = done
```

### マネージャーワーカーの監視

マネージャーワーカーは、他のすべてのワーカーの管理を含む多くのバックグラウンドタスクを担当します。そのため、特に行き詰まっているように見える場合、それが何をしているのかを知ることが重要です。

したがって、bdr.stat_worker ビューは、マネージャーワーカーを含む PGD ワーカーのワーカーごとの統計を提供します。マネージャーワーカーがスタックしないように、実行中の現在のタスクは、“pgd manager:” の接頭辞が付いている `query` フィールドで報告されます。

マネージャーワーカーの `worker_backend_state` フィールドは、マネージャーがアイドル状態かビジー状態かを報告します。

### モニタリングルーティング

ルーティングは、シームレスなアプリケーションエクスペリエンスと競合の回避を保証するための PGD の重要な部分です。ルーティングの変更は、障害の検出を含めて、迅速に発生する必要があります。同時に、中断はできるだけ少なくしたいと考えています。また、サポートされているユースケースで、適切なロードバランシングを保証したいと考えています。

これらをすべて監視することは、問題に気づき、問題をデバッグし、より最適な構成を通知するために重要です。したがって、ルーティングに関連する監視統計には、2 つの主なビューがあります。

- bdr.stat_routing_state 接続マネージャーを使用した接続ルーティングの状態をモニタリングするための使用接続のルーティングに使用します。
- Raft リーダーの観点からのルーティング候補ノードに関する情報については bdr.stat_routing_candidate_state ビューは他のノードでは空です。

### レプリケーションピアのモニタリング

レプリケーションアクティビティの監視には、2 つのメインビューを使用します。

- bdr.node_slots 送信レプリケーションを監視するため
- bdr.subscription_summary 着信レプリケーションを監視するため

標準の PostgreSQL レプリケーション監視ビューを照会することにより、`bdr.node_slots` が提供する情報のほとんどを取得することもできます

pg_catalog.pg_stat_replication

および pg_catalog.pg_replication_slots 。

各ノードには 1 つの PGD グループスロットがあります。これらのスロットには接続してはならず、アクティブとしてマークされることはほとんどありません。これは正常であり、何かがダウンまたは切断されていることを意味するものではありません。ノード管理の *Replication slots created by PGD* を参照してください。

### 送信レプリケーションの監視

送信レプリケーションアクティビティの監視に別のビューを使用できます。

- bdr.node_replication_rates 送信レプリケーションを監視するため

bdr.node_replication_rates ビューは、送信レプリケーションアクティビティの全体像と、特にピアノードのキャッチアップ推定を提供します。

```
#  SELECT * FROM bdr.node_replication_rates;

- [ RECORD 1 ]----+-----------
peer_node_id     | 112898766
target_name      | node1
sent_lsn         | 0/28AF99C8
replay_lsn       | 0/28AF99C8
replay_lag       | 00:00:00
replay_lag_bytes | 0
replay_lag_size  | 0 bytes
apply_rate       | 822
catchup_interval | 00:00:00
- [ RECORD 2 ]----+-----------
peer_node_id     | 312494765
target_name      | node3
sent_lsn         | 0/28AF99C8
replay_lsn       | 0/28AF99C8
replay_lag       | 00:00:00
replay_lag_bytes | 0
replay_lag_size  | 0 bytes
apply_rate       | 853
catchup_interval | 00:00:00
```

apply_rate は、1 秒あたりのバイト数のレートを参照します。これは、ピアがローカルノードからデータを消費している速度です。ノードがクラスターに再接続したときの replay_lag は、すぐにゼロに設定されます。この情報は将来のリリースで修正される予定です。回避策として、ピアノードがローカルノードデータに追いつくまでに必要な時間を参照する catchup_interval 列を使用することをお勧めします。他のフィールドは、bdr.node_slots から利用できます。

ビュー。

管理者は、ローカルノードからの送信レプリケーションについて bdr.node_slots を照会できます。これは、現在のノードで知られているグループ内の他のすべてのノードのレプリケーションステータスと、現在のノードで PGD によって作成された追加のレプリケーションスロットに関する情報を表示します。

```
#  SELECT node_group_name, target_dbname, target_name, slot_name, active_pid,

#      catalog_xmin, client_addr, sent_lsn, replay_lsn, replay_lag,

#      replay_lag_bytes, replay_lag_size
```

(次のページに続く)

(前のページからの続き)

```
#  FROM bdr.node_slots;

- [ RECORD 1 ]---+---------------------------
node_group_name | bdrgroup
target_dbname   | postgres
target_name     | node3
slot_name       | bdr_postgres_bdrgroup_node3
active_pid      | 15089
catalog_xmin    | 691
client_addr     | 127.0.0.1
sent_lsn        | 0/23F7B70
replay_lsn      | 0/23F7B70
replay_lag      | [NULL]
replay_lag_bytes| 120
replay_lag_size | 120 bytes
- [ RECORD 2 ]---+---------------------------
node_group_name | bdrgroup
target_dbname   | postgres
target_name     | node2
slot_name       | bdr_postgres_bdrgroup_node2
active_pid      | 15031
catalog_xmin    | 691
client_addr     | 127.0.0.1
sent_lsn        | 0/23F7B70
replay_lsn      | 0/23F7B70
replay_lag      | [NULL]
replay_lag_bytes| 84211
replay_lag_size | 82 kB
```

PGD はメッシュネットワークであるため、クラスター内の遅延の完全なビューを取得するには、参加しているすべてのノードでこのクエリを実行する必要があります。

replay_lag_bytes は、ローカルサーバーの現在の WAL 書き込み位置と、ピアノードによって再生確認された最後の位置である replay_lsn 間の WAL 位置の差を報告します。 replay_lag_size は、同じの人が判読可能な形式です。通常、WAL には、レプリケートされないが replay_lag_bytes でカウントされる大量の書き込みが含まれることを理解することが重要です。

- VACUUM アクティビティ
- インデックスの変更
- 同じノード上の他のデータベースに関連付けられた書き込み
- レプリケーションセットの一部ではないテーブルの書き込み

したがって、ここで報告されるバイト単位の遅延は、ピアノードを最新の状態にするためにワイヤで複製する必要があるデータの量ではなく、処理する必要があるサーバーサイドの WAL の量のみです。

同様に、replay_lag は、ピアノードが追いつくまでに要した時間、または現在位置から bdr.node_slots が照会された時点の書き込み位置まで再生に要した時間を表すものではありません。ピアが最新のコミットを確認したときと現在の実測時間との間の遅延を測定します。この列は、ノードが再接続した直後にゼロに設定されるため、replay_lag_bytes と replay_lag_size または bdr.node_replication_rates の catchup_interval をモニタすることをお勧めします。

論理レプリケーションがトランザクションをストリーミングしている間、バイトと時間の両方の遅延は進行しません。コミットがレプリケートされた場合にのみ変更されます。したがって、遅延は「のこぎり波」の傾向があり、トランザクションがストリーミングされると上昇し、ピアノードがコミット、フラッシュ、確認を送信すると再び低下します。報告された LSN は、同様の理由で、スムーズに進むのではなく、「階段」を位置づけています。

レプリケーションが切断されている場合 active = 'f'、active_pid 列は NULL、client_addr およびアクティブな接続でのみ意味をなすその他のフィールドと同様に、NULL です。 state フィールドは'disconnected' です。_lsn フィールドは、クライアントが再生および保存したことが確認されている最後の位置であるため、confirmed_flush_lsn と同じです。 _lag フィールドは、クライアントで確認された最新のフラッシュから現在までの経過時間を示します。_lag_size および_lag_bytes フィールドは、confirmed_flush_lsn とローカルサーバーの現在の WAL 挿入位置間の距離をレポートします。

> **注釈**
>
> *restart_lsn* が他の `lsn` 列の後ろにあるのは通常です。これは、レプリケーションまたはピアノードの遅延に関する問題を示しているわけではありません。 *restart_lsn* は、PostgreSQL の内部論理デコードが中断された場合に WAL を読み取る位置です。これは、通常、まだレプリケートおよびフラッシュされていない最も古いトランザクションの位置を反映します。非常に古い `restart_lsn` は、切断後のレプリケーションの再起動を遅くし、望ましいよりも多くの WAL の保持を強制する場合がありますが、それ以外の場合は無害です。心配する場合は、非常に長時間実行されているトランザクションと忘れられた準備済みトランザクションを探してください。

### 着信レプリケーションの監視

bdr.subscription_summary ビューを照会することにより、受信レプリケーションサブスクリプションとも呼ばれますを高レベルでモニタできます。このクエリは、EDB Postgres 分散クラスター内の他のノードへの既知のサブスクリプションのリストとレプリケーションワーカーの状態を表示します。

```
#  SELECT node_group_name, origin_name, sub_enabled, sub_slot_name,

#      subscription_status

#  FROM bdr.subscription_summary;

- [ RECORD 1 ]-------+---------------------------
node_group_name     | bdrgroup
origin_name         | node2
sub_enabled         | t
sub_slot_name       | bdr_postgres_bdrgroup_node1
subscription_status | replicating
```

(次のページに続く)

(前のページからの続き)

```
- [ RECORD 2 ]-------+--------------------------
node_group_name     | bdrgroup
origin_name         | node3
sub_enabled         | t
sub_slot_name       | bdr_postgres_bdrgroup_node1
subscription_status | replicating
```

bdr.stat_subscription を介してサブスクリプションの概要統計をモニタリングし、それぞれ bdr.stat_receiver および bdr.stat_writer を使用してサブスクリプションレプリケーションレシーバーとサブスクリプションレプリケーションライターをモニタリングすることにより、サブスクリプションをさらにモニタできます。

### LCR を使用した WAL 送信者のモニタリング

*Decoding worker* が有効になっている場合、ファンクション bdr.wal_sender_stats() を使用して、各 WAL センダーの現在の論理変更レコード LCR ファイルに関する情報をモニタできます。例

```
postgres=# SELECT * FROM bdr.wal_sender_stats();
   pid   | is_using_lcr |       decoder_slot_name       |              lcr_file_name
- -------+--------------+-------------------------------+---------------------------
↪--------------
 2059904 | f            |                               |
 2059909 | t            | bdr_postgres_bdrgroup_decoder |␣
↪000000000000000000000140000000000000000
 2059916 | t            | bdr_postgres_bdrgroup_decoder |␣
↪000000000000000000000140000000000000000
(3 rows)
```

is_using_lcr が FALSE の場合、decoder_slot_name /lcr_file_name は NULL です。これは、デコードワーカーが有効になっていない場合、または WAL 送信者が *Logical standby nodes* を提供している場合に当てはまります。

また、ファンクション

bdr.get_decoding_worker_stat() を使用して、デコードワーカーに関する情報をモニタできます。例

```
postgres=# SELECT * FROM bdr.get_decoding_worker_stat();
   pid   | decoded_upto_lsn | waiting | waiting_for_lsn
- -------+------------------+---------+-----------------
 1153091 | 0/1E5EEE8        | t       | 0/1E5EF00
(1 row)
```

### PGD レプリケーションワーカーのモニタリング

すべての PGD ワーカーは、

pg_stat_activity と同じ列と情報コンテンツがあるシステムビュー bdr.stat_activity に表示されます。したがって、このビューは、PGD システムの状態に関する次の洞察を提供します。

- 待機の理由が PGD に関連している場合、wait_event 列に強化された情報があります。
- ライタプロセスが DDL を実行している場合、またはマネージャーワーカーがアクティブな場合を除き、PGD ワーカーの query 列は空白です。この場合、 query 列のエントリには「pgd manager: 」の接頭辞が付けられます。

bdr.workers ビューには、 bdr.stat_activity からは使用できない PGD ワーカー固有の詳細が表示されます。

bdr.event_summary ビューは、作業の続行に問題があるワーカーによって報告された最後のエラーがある場合を表示します。この情報は永続的なため、エラーの存在だけでなく、エラーの時間を記録することが重要です。ほとんどのエラーは一時的なもので、PGD ワーカーは失敗した操作を再試行します。

### PGD ライターのモニタリング

別のシステムビュー bdr.writers は、ライタアクティビティを監視します。このビューには、ライタワーカーの現在のステータスのみが表示されます。それには以下が含まれます

- sub_name ライタが属しているサブスクリプションを識別する
- ライタプロセスの pid
- streaming_allowed ライタが進行中のストリーミングトランザクションの適用をサポートしているかどうかを知るため
- is_streaming ライタが現在ストリーミングトランザクションを適用しているかどうかを知るため
- commit_queue_position コミットキュー内のライタの位置を確認する

PGD は、オリジンで発生したのと同じコミット順序に従って、コミット順序を尊重します。並列ライターの場合、複数のライターが異なるトランザクションを同時に適用する場合があります。commit_queue_position は、コミットする順序を示します。値 0 は、ライタが最初にコミットすることを意味します。値-1 は、コミット位置がまだ不明であることを意味します。これは、ストリーミングトランザクション、またはライタが現在トランザクションを適用していない場合に発生します。

### コミットスコープのモニタリング

コミットスコープは、耐久性と一貫性の構成フレームワークです。そのため、それらはトランザクションのパフォーマンスに影響を与えるため、それらに関する統計を取得することが重要です。さらに、障害シナリオでは、コミットスコープの構成が原因でトランザクションがスタックしているように見える場合があるため、どのコミットスコープが使用されているか、何を待機しているかなどに関する洞察が必要です。

したがって、これらの 2 つのビューは、コミットスコープに関する関連統計を示します。

- bdr.stat_commit_scope デグレードイベントカウンター ndegrades 、nconfig_degrades や最後の構成状態変更のタイムスタンプ last_state_change_time を含む各コミットスコープの累積統計。
- bdr.stat_commit_scope_state バックエンドプロセスによるコミットスコープの現在の使用に関する情報。

デグレードイベントのモニタリングの詳細については、**機能低下イベントの監視** を参照してください。

### グローバルロックの監視

グローバルロックは、現在 DDL レプリケーションにのみ使用されていますが、PGD グループ全体に存在する重量ロックです。

現在、2 種類のグローバルロックがあります。

- DDL ロック、データベース内の一時的なものではない永続的なオブジェクトつまりテーブルに対するすべての DDL 操作をシリアル化するために使用されます
- DML リレーションロック、リレーション定義を変更する DDL 操作中にリレーションへの書き込みをロックアウトするために使用されます

DDL 操作のタイプと bdr.ddl_locking 設定の値に応じて、同じトランザクションに一方または両方のエントリータイプを作成できます。

ローカルノードで保持されているグローバルロックは、bdr.global_locks ビューに表示されます。このビューには、ロックのタイプが表示されます。リレーションロックの場合、ロックされている関係、ロックを保持している PID ローカルの場合、およびロックがグローバルに付与されたかどうかを示します。グローバルアドバイザリロックの場合、lock_type 列は GLOBAL_LOCK_ADVISORY を示し、 relation 列は、ロックが取得されるアドバイザリキーを示します。

この例は、 bdr.ddl_locking = 'all' を使用して ALTER TABLE ステートメントを実行中の bdr.global_locks の出力を示しています。

```
#  SELECT lock_type, relation, pid FROM bdr.global_locks;

- [ RECORD 1 ]--------------
lock_type | GLOBAL_LOCK_DDL
relation  | [NULL]
pid       | 15534
- [ RECORD 2 ]--------------
lock_type | GLOBAL_LOCK_DML
relation  | someschema.sometable
pid       | 15534
```

ロックタイミング情報を含むすべてのフィールドの詳細については、Catalogs を参照してください。

### 競合のモニタリング

レプリケーション *bdr_read_all_conflicts* は、複数のノードが相互作用できる方法で同じ行に影響を与える変更を行う場合に発生する可能性があります。PGD システムをモニタして競合を特定し、可能であればアプリケーションを変更して競合を排除するか、競合の頻度を減らします。

デフォルトでは、すべての競合は bdr.conflict_history ログに記録されます。このログには、競合するデータの完全な詳細が含まれているため、行は行レベルのセキュリティで保護され、レプリケートテーブルの所有者にのみ表示されます。所有者は競合を予想し、それらを分析して、存在する場合、どれが解決すべき問題と考えられるかを確認する必要があります。

モニタリングには、ユーザーデータを含まない bdr.conflict_history_summary を使用します。この例は、効率的なクエリプランを使用して、その日に発生した競合の数をカウントするクエリを示しています。

```
SELECT count(*)
FROM bdr.conflict_history_summary
WHERE local_time > date_trunc(day, current_timestamp)
  AND local_time < date_trunc(day, current_timestamp + 1 day);
```

### 統計を適用する

PGD は、サブスクリプションごととテーブルごとの両方で、レプリケーションの適用に関する統計を収集します。

2 つのモニタリングビューが存在します。サブスクリプション統計の場合は bdr.stat_subscription 、リレーション統計の場合は bdr.stat_relation です。これらのビューは両方とも次を提供します。

- レプリケートされた INSERT / UPDATE / DELETE / TRUNCATE の数
- ブロックアクセスとキャッシュヒット率
- 読み取り/書き込みの合計 I/O 時間
- ファイルにストリーミングされた進行中のトランザクションの数
- ライターにストリーミングされる進行中のトランザクションの数
- コミット/中止された進行中のストリーミングトランザクションの数

リレーションのみの場合、 bdr.stat_relation には次のものも含まれます。

- リレーションのレプリケーションの処理に費やされた合計時間
- リレーションのみのロックがある場合を取得するまでの合計ロック待機時間

サブスクリプションの場合のみ、 bdr.stat_subscription には以下が含まれます。

- サブスクリプションに対してレプリケートされた COMMIT/ DDL の数
- このサブスクリプションがアップストリームに接続した回数

これらの統計の追跡は、それぞれ PGD GUC

bdr.track_subscription_apply および bdr.track_relation_apply によって制御されます。

次に、これらの出力例を示します。

```
#  SELECT sub_name, nconnect, ninsert, ncommit, nupdate, ndelete, ntruncate, nddl

FROM bdr.stat_subscription;
- [ RECORD 1 ]---------------------------------
sub_name  | bdr_regression_bdrgroup_node1_node2
nconnect  | 3
ninsert   | 10
ncommit   | 5
nupdate   | 0
ndelete   | 0
```

(次のページに続く)

(前のページからの続き)

```
ntruncate | 0
nddl      | 2
```

この場合、サブスクリプションは上流に 3 回接続し、10 行を挿入し、5 つのトランザクション内で 2 つの DDL コマンドを実行しました。

ファンクション bdr.reset_subscription_stats および bdr.reset_relation_stats を使用して、これらのビューの統計カウンターをゼロにリセットできます。

PGD は、それぞれ bdr.stat_receiver および bdr.stat_writer を使用して、各サブスクリプションのサブスクリプションレプリケーションレシーバーとサブスクリプションレプリケーションライターに関する統計も監視します。

### 標準の PostgreSQL 統計ビュー

テーブルとインデックスの使用状況に関する統計は、通常、ダウンストリームマスターによって更新されます。これは、

autovacuum の正しい機能のために重要です。ダウンストリームマスターにローカル書き込みがなく、統計がリセットされていない場合、これらの 2 つのビューは、アップストリームとダウンストリームで対応する結果を表示します。

- `pg_stat_user_tables`
- `pg_statio_user_tables`

> **注釈**
>
> 上流のテーブル統計がダウンストリームのテーブル統計と*同様*であることを必ずしも期待しているわけではありません。私たちはそれらが同じ量だけ*変更*されることを期待しています。統計が 100 万の挿入と 100 万の更新を示すテーブルの例を考えます。新しいノードが PGD グループに参加すると、新しいノードの同じテーブルの統計には、1M の挿入とゼロの更新が示されます。ただし、その瞬間から、一方の側ですべての変更がもう一方の側にレプリケートされるため、上流と下流のテーブル統計は同じ量だけ変更されます。

インデックスは変更を適用するために使用されるため、ダウンストリーム側の識別インデックスは、非識別インデックスよりも `UPDATE` および `DELETE` を実行するワークロードで頻繁に使用される場合があります。

ビルトインインデックスモニタリングビューは次のとおりです。

- `pg_stat_user_indexes`
- `pg_statio_user_indexes`

これらのビューはすべて、PostgreSQL documentation on the statistics views で詳細に説明されています。

### PGD バージョンのモニタリング

PGD では、同じクラスター内のノード全体で異なる Postgres バージョンと異なる BDR 拡張バージョンを実行できます。この機能は、アップグレードに役立ちます。

ビュー bdr.group_versions_details は、ファンクション bdr.run_on_all_nodes() を使用して、すべてのノードから Postgres および BDR 拡張バージョンを同時に取得します。例

```
pgddb=# SELECT node_name, postgres_version, bdr_version
        FROM bdr.group_versions_details;
 node_name | postgres_version | bdr_version
- ----------+------------------+-------------
 node1     | 15.2.0           | 5.0.0
 node2     | 15.2.0           | 5.0.0
```

推奨されるセットアップは、すべてのノードでできるだけ早く同じ最新のバージョンを実行するようにすることです。クラスターは、BDR 拡張機能のさまざまなバージョンを長時間実行しないことをお勧めします。

モニタリングのために、次のアラートレベルをお勧めします。

- status = UNKNOWN、message =このノードは PGD グループの一部ではありません
- status = OK、message =すべてのノードが同じ PGD バージョンを実行しています
- status = WARNING、message =アクセスできないノードが少なくとも 1 つあります
- status = WARNING、message =他のノードと比較して、異なる PGD バージョンを実行しているノードがあります

説明した動作は、

bdr.monitor_group_versions() ファンクションに実装されています。このファンクションは、ビュー bdr.group_version_details から返された PGD バージョン情報を使用して、クラスター全体のバージョンチェックを提供します。例

```
pgddb=# SELECT * FROM bdr.monitor_group_versions();
 status |                message
- -------+---------------------------------------
 OK     | All nodes are running same BDR versions
```

### Monitoring Raft コンセンサス

Raft コンセンサスは、常にクラスター全体で動作している必要があります。Raft コンセンサスが機能せずに EDB Postgres 分散クラスターを実行した場合の影響は次のとおりです。

- PGD データ変更のレプリケーションは、引き続き正常に動作する場合があります。
- グローバル DDL/DML ロックは機能しません。
- Galloc シーケンスは、最終的にチャンクを使い果たします。
- Eager Replication が機能しない。

- クラスターメンテナンス操作ノード参加、部分ノード、プロモートスタンバイは引き続き許可されますが、完了しない代わりにハングする場合があります。
- PGD ノード間でノードのステータスが正しく同期されない場合があります。
- PGD グループレプリケーションスロットは LSN を進歩させないため、ディスク上に WAL ファイルを保持します。

ビュー bdr.group_raft_details は、ファンクション bdr.run_on_all_nodes() および bdr.get_raft_status() を使用して、すべてのノードから同時に Raft コンセンサスステータスを取得します。例

```
pgddb=# SELECT node_id, node_name, state, leader_id
FROM bdr.group_raft_details;
  node_id   | node_name | node_group_name |     state      | leader_id
- -----------+-----------+-----------------+---------------+------------
 1148549230 | node1     | top_group       | RAFT_LEADER   | 1148549230
 3367056606 | node2     | top_group       | RAFT_FOLLOWER | 1148549230
```

これらの条件がすべて満たされる場合、Raft コンセンサスは正常に動作します。

- 有効な状態 RAFT_LEADER または RAFT_FOLLOWER がすべてのノードで定義されています。
- ノードのうちの 1 つだけが RAFT_LEADER です。
- leader_id はすべての行で同じであり、state = RAFT_LEADER が存在する行の node_id と一致する必要があります。

時々、Raft コンセンサスは新しい選挙を開始して、新しい RAFT_LEADER を定義します。選挙中に、RAFT_LEADER が存在せず、一部のノードが自分自身を RAFT_CANDIDATE と見なす中間状況が発生する場合があります。選挙全体には bdr.raft_global_election_timeout より長くかかることはありませんデフォルトでは 6 秒に設定されています。上記のクエリーが選挙中の状況を結果た場合、bdr.raft_global_election_timeout を待って、クエリーを再度実行します。bdr.raft_global_election_timeout が経過した後、リストされた条件の一部がまだ満たされていない場合、Raft コンセンサスは機能していません。

Raft コンセンサスは、単一のノードでのみ正常に動作していない可能性があります。たとえば、ノードの 1 つが現在のリーダーを認識せず、自分自身を RAFT_CANDIDATE と見なします。この場合、次のことを確認することが重要です。

- すべての PGD ノードは、レギュラー接続とレプリケーション接続の両方を介して相互にアクセスできますチェックファイル pg_hba.conf 。
- PGD バージョンはすべてのノードで同じです。
- bdr.raft_global_election_timeout はすべてのノードで同じです。

場合によっては、特にノードが地理的に離れている場合、またはネットワーク遅延が高い場合、bdr.raft_global_election_timeout のデフォルト値 6 秒では十分でない場合があります。すべてが正しいことを確認した後でも Raft コンセンサスがまだ機能していない場合は、すべてのノードで bdr.raft_global_election_timeout を 30 秒に増加することを検討します。

Raft コンセンサスがクラスター操作タスクにどのような影響を与えるか、また Raft コンセンサスがグループスロットを前進させることに直接責任があるため、モニタリングアラートレベルは次のように定義され

ます。

- status = UNKNOWN、message =このノードは PGD グループの一部ではありません
- status=OK、message=Raft コンセンサスは正常に動作しています
- status = WARNING、message =アクセスできないノードが少なくとも 1 つあります
- status = WARNING、 message = RAFT_CANDIDATE としてノードがあります、選択が進行中の可能性があります
- status = WARNING、message = RAFT_LEADER がありません、選挙が進行中の可能性があります
- status=CRITICAL、message=Raft コンセンサスには単一のノードがあります
- status = CRITICAL、message = RAFT_CANDIDATE としてノードがありますが、RAFT_LEADER が定義されています
- status=CRITICAL、message=RAFT_LEADER として設定されたノードとは異なるリーダーをフォローしているノードがあります。

説明した動作はファンクション bdr.monitor_group_raft() に実装されます。これは、ビュー bdr.group_raft_details から返された Raft コンセンサスステータス情報を使用します

クラスター全体の Raft チェックを提供します。例

```
pgddb=# SELECT * FROM bdr.monitor_group_raft();
node_group_name | status |               message
- ---------------/--------+-------------------------------------
mygroup          | OK     | Raft Consensus is working correctly
```

Raft コンセンサスの状態をより詳細に与えることができるさらに 2 つのビューは、ローカルノードでの Raft コンセンサスの状態を提供する bdr.stat_raft_state と、Raft リーダー上のときのビューを提供する bdr.stat_raft_followers_state です。他のノード) その Raft リーダーのフォロワーの状態に関するもの。

### レプリケーションスロットの監視

各 PGD ノードは以下を保持します。

- アクティブな PGD ピアごとに 1 つのレプリケーションスロット
- 1 つのグループレプリケーションスロット

例

```
pgddb=# SELECT slot_name, database, active, confirmed_flush_lsn
FROM pg_replication_slots ORDER BY slot_name;
        slot_name         | database | active | confirmed_flush_lsn
- ------------------------+----------+--------+---------------------
 bdr_pgddb_bdrgroup       | pgddb    | f      | 0/3110A08
 bdr_pgddb_bdrgroup_node2 | pgddb    | t      | 0/31F4670
 bdr_pgddb_bdrgroup_node3 | pgddb    | t      | 0/31F4670
 bdr_pgddb_bdrgroup_node4 | pgddb    | t      | 0/31F4670
```

ピアスロット名は bdr_<DATABASE>_<GROUP>_<PEER> 規則に従いますが、PGD グループスロット名は bdr_<DATABASE>_<GROUP> 規則に従います。ファンクション bdr.local_group_slot_name() を使用してグループスロットにアクセスできます。

ピアレプリケーションスロットは、すべてのノードで常にアクティブである必要があります。ピアレプリケーションスロットがアクティブでない場合、それは次のいずれかを意味する可能性があります。

- 対応するピアはシャットダウンしているか、アクセスできません。
- PGD レプリケーションが壊れています。

ERROR または FATAL のログファイルを grep し、すべてのノードで bdr.event_summary を確認します。根本的な原因は、たとえば、ノードの 1 つで DDL レプリケーションを無効にして、互換性のない DDL が実行されたことが考えられます。

ただし、PGD グループレプリケーションスロットは、ほとんどの場合非アクティブです。他のすべてのピアが対応するトランザクションを既に消費している場合、PGD はこのスロットを維持し、その LSN を進めます。したがって、グループスロットのステータスを監視する必要はありません。

ファンクション bdr.monitor_local_replslots() は、すべての PGD ノードレプリケーションスロットが予想通りに動作しているかどうかの概要を提供します。この概要は、*Optimizing subscriber-only groups* が有効になっているときに PGD クラスターでサブスクライバ専用グループリーダーとして動作しているサブスクライバ専用ノードでも利用できます。例

```
pgddb=# SELECT * FROM bdr.monitor_local_replslots();
 status |                    message
- -------+-------------------------------------------------
 OK     | All BDR replication slots are working correctly
```

次のステータスの概要のいずれかが返されます。

| Status | Message |
|---|---|
| UNKNOWN | This node is not part of any BDR group |
| OK | All BDR replication slots are working correctly |
| OK | This node is part of a subscriber-only group |
| CRITICAL | There is at least 1 BDR replication slot which is inactive |
| CRITICAL | There is at least 1 BDR replication slot which is missing |

### トランザクションの COMMIT の監視

デフォルトでは、PGD トランザクションはローカルノードにのみコミットされます。その場合、トランザクションの COMMIT は迅速に処理されます。

PGD の *Commit scopes* 機能は、特定のクエリーの耐久性、一貫性、パフォーマンスのバランスを可能にするさまざまな同期トランザクションコミットスコープを提供します。bdr.stat_activity カタログを調べることにより、これらのトランザクションをモニタできます。トランザクションがコミットされると、プロセスはさまざまな wait_event 状態を報告します。このモニタリングは、進行中のトランザクションのみをカバーし、履歴タイミング情報は提供しません。

## 6.6 AutoPartition in PGD

PGD AutoPartition を使用すると、テーブルを複数のパーティション、他のテーブルに分割し、必要に応じてパーティションを作成および削除できます。自動パーティション化は、データを管理可能なチャンクに分割できるため、時間の経過とともに増加する大規模なテーブルの管理に役立ちます。新しいパーティションを定期的に作成し、データ保持期間の有効期限が切れたら削除できます。

この機能は、パーティションの作成と削除の低競合ロックなどの PGD の機能を使用します。

PGD 管理は、主に SQL で呼び出すことができるファンクションを使用して実行します。PGD のすべてのファンクションは bdr スキーマで公開されています。search_path に配置しない限り、各ファンクションの名前をスキーマ修飾する必要があります。

### 6.6.1 パーティションの自動作成

PGD AutoPartition は bdr.autopartition() を使用します

テーブルの自動レンジパーティションの定義を作成または変更するファンクション。定義が存在しない場合、作成されます。それ以外の場合、後の実行で定義が変更されます。

PGD 5.5 以降の PGD AutoPartition は、基になる Postgres 機能を活用して、テーブルの残りの部分をロックせずにパーティションをアタッチまたはデタッチ/ドロップできます。5.5 より前のバージョンの PGD はこの機能をサポートしておらず、テーブルをロックします。

テーブルが RANGE パーティション化されていない場合、または複数列パーティションキーが使用されている場合、エラーが発生します。

デフォルトでは、AutoPartition はパーティションをローカルに管理します。パーティションのローカルでの管理は、パーティション化されたテーブルがレプリケートテーブルでない場合に役立ちます。その場合、すべてのノードにすべてのパーティションを必要としないか、したくない場合があります。例、ビルトイン bdr.conflict_history

テーブルはレプリケートテーブルではありません。これは AutoPartition によってローカルに管理されます。各ノードはこのテーブルのパーティションをローカルに作成し、十分に古くなると削除します。

次のことも考慮します。

- アクティビティは、エントリーが enabled = on とマークされている場合にのみ実行されます。
- AutoPartition によって管理されるテーブルのパーティションを手動で作成または削除しないことをお勧めします。そうすると、AutoPartition メタデータの一貫性がなくなり、障害が発生する可能性があります。

### 6.6.2 AutoPartition の例

日次パーティション、データを 1 か月間保持します

```
CREATE TABLE measurement (
logdate date not null,
peaktemp int,
unitsales int
```

(次のページに続く)

(前のページからの続き)

```
) PARTITION BY RANGE (logdate);

bdr.autopartition(measurement, 1 day, data_retention_period := 30 days);
```

パーティションがあと 2 つだけ残っている場合、5 つの事前パーティションを作成します。各パーティションには、10 億の注文を保持できます。

```
bdr.autopartition(Orders, 1000000000,
        partition_initial_lowerbound := 0,
        minimum_advance_partitions := 2,
        maximum_advance_partitions := 5
    );
```

### 6.6.3 RANGE パーティションテーブル

PGD 自動パーティション化は、 RANGE キーワードを使用したレンジパーティション化をサポートしています。レンジパーティション化を使用すると、列の値の範囲に基づいてテーブルをパーティション化できます。たとえば、日付でテーブルをパーティション化できます。各パーティションには、特定の日付範囲のデータが含まれます。これは、データを管理可能なチャンクに分割できるため、時間の経過とともに増加する大規模なテーブルの管理に役立ちます。

partition_increment 値の範囲ごとに新しいパーティションが追加されます。下限と上限は partition_increment 離れています。 timestamp または date タイプのパーティションキーを持つテーブルの場合、partition_increment は interval タイプの有効な定数である必要があります。たとえば、1 Day を指定すると、新しいパーティションが毎日追加され、パーティション境界は 1 日ごとに追加されます。

パーティション列が snowflakeid 、timeshard 、または ksuuid シーケンスに接続されている場合、partition_increment を interval タイプとして指定する必要があります。それ以外の場合、パーティションキーが整数または数値の場合、partition_increment は同じデータ型の有効な定数である必要があります。たとえば、 1000000 を指定すると、100 万の値ごとに新しいパーティションが追加されます。

テーブルに既存のパーティションがない場合、指定された partition_initial_lowerbound が最初のパーティションの下限として使用されます。partition_initial_lowerbound を指定しない場合、システムはパーティション列タイプと指定された partition_increment からその値を導き出そうとします。たとえば、 partition_increment が 1 Day として指定されている場合、partition_initial_lowerbound は CURRENT DATE に設定されます。partition_increment が 1 Hour として指定される場合、partition_initial_lowerbound は現在の日付の現在時間に設定されます。後続のパーティションの境界は、partition_increment 値を使用して設定されます。

システムは常に特定の最小数の事前パーティションを持とうとします。新しいパーティションを作成するかどうかを決定するために、指定された partition_autocreate_expression を使用します。これは、チェックが実行されるたびに評価される SQL によって評価できる式です。たとえば、列タイプ date のパーティションテーブルの場合、partition_autocreate_expression が DATE_TRUNC('day',CURRENT_DATE) として指定され、partition_increment が 1 Day として指定され、minimum_advance_partitions が 2 として指定されるとします。次に、最後のパーティションの上限が DATE_TRUNC('day', CURRENT_DATE) + '2 Days'::interval 未満になるまで、新しいパーティションが作成されます。

式は、システムが新しいパーティションをチェックするたびに評価されます。

列タイプ integer のパーティションテーブルの場合、partition_autocreate_expression を SELECT max(partcol) FROM schema.partitioned_table として指定できます。次に、システムは、パーティション化された列の最大値が最後のパーティションの上限の minimum_advance_partitions * partition_increment の距離内にあるかどうかを定期的にチェックします。クエリーが効率的に実行されるように、partcol にインデックスを作成します。カラムタイプ integer 、smallint 、または bigint のパーティションテーブルに partition_autocreate_expression を指定しない場合、システムは max(partcol) に設定します。

data_retention_period が設定されている場合、パーティションはこの期間の後にドロップされます。ロックを最小限に抑えるために、新しいパーティションが追加されると同時にパーティションが削除されます。この値を設定しない場合、パーティションを手動で削除する必要があります。

data_retention_period パラメーターは、タイムスタンプベースおよび関連するパーティションでのみサポートされています。期間は、パーティションの上限を考慮して計算されます。上限と比較して、指定された期間が満了すると、パーティションはドロップされます。

### 6.6.4 パーティションの自動作成の停止

bdr.drop_autopartition() を使用する

指定されたリレーションの自動パーティションルールを削除します。リレーションの保留中の作業項目はすべて削除され、新しい作業項目は作成されません。

### 6.6.5 パーティションの作成を待機しています

パーティションの作成は非同期プロセスです。AutoPartition は、ローカルまたはすべてのノードでパーティションが作成されるのを待機する一連の機能を提供します。

bdr.autopartition_wait_for_partitions() を使用する

ローカルノードでパーティションの作成を待機します。ファンクションは、パーティションテーブル名とパーティションキー列値を取得し、その値を保持するパーティションが作成されるまで待機します。

このファンクションは、パーティションがローカルに作成されるのをのみ待機します。パーティションがリモートノードにも存在することは保証されません。

すべての PGD ノードでパーティションが作成されるのを待機するには、

bdr.autopartition_wait_for_partitions_on_all_nodes() を使用します

ファンクション。このファンクションは、ローカルノードとすべてのリモートノードを内部的にチェックし、パーティションがどこでも作成されるまで待機します。

### 6.6.6 パーティションの検索

bdr.autopartition_find_partition() を使用します

指定されたパーティションキー値のパーティションを見つけるファンクション。その値を保持するパーティションが存在しない場合、ファンクションは NULL を結果ます。それ以外の場合、パーティションの Oid を結果ます。

### 6.6.7 自動パーティショニングの有効化または無効化

bdr.autopartition_enable() を使用する

指定されたテーブルで自動パーティション化を有効にします。自動パーティション化が既に有効になっている場合、アクションは発生しません。同様に、

bdr.autopartition_disable() を使用します

指定されたテーブルで自動パーティショニングを無効にします。

### 6.6.8 パーティションの削除またはデタッチ

デフォルトでは、 `data_retention_period` の有効期限が切れると、`autopartition()` によって管理されるテーブルのパーティションがデタッチおよびドロップされます。`drop_after_retention_period` パラメーターはこの動作を制御できるため、これらのパーティションを保持することを選択できます。`false` に設定すると、パーティションは親テーブルからデタッチされるだけで、ドロップされません。これにより、ユーザーは最終的に削除する前にパーティションを他のテーブルに再アタッチしたり、処理したりできる可能性があります。パラメーターのデフォルト値は `true` であるため、既存または新しい自動パーティションテーブルでは、`data_retention_period` の有効期限が切れるとパーティションはドロップされます。デタッチされたがドロップされていないパーティションを処理するには、これらのパーティションを他のテーブルに手動で再アタッチするか、ドロップします。

## 6.7 Commit Scopes

完全に管理および構成可能なコミットスコープは、PGD Expanded の機能です。

PGD Expanded は、デフォルトの非同期レプリケーションを補完するさまざまな同期モードを提供します。コミットスコープを使用して、これらの同期モードを構成します。コミットスコープは、PGD が同期操作を処理する方法と、システムがトランザクションがコミットされたと判断するタイミングを定義するルールです。

PGD Essential は、事前定義され変更できない限定的なコミットスコープを提供します。

### 6.7.1 のご紹介

- *Overview of durability options* は、同期コミットを議論する際に使用される概念といくつかの重要な用語を紹介します。
- *Durability terminology* は、レプリケーションでノードを参照する方法を含む、PGD の耐久性オプションの周辺で使用される用語をリストします。
- *Commit scopes* は、コミットスコープの構造と、ニーズに合わせてそれらを定義する方法をより詳細に示しています。
- *Predefined commit scopes* は、 PGD Essential で使用可能な事前定義されたコミットスコープをリストします。
- *Commit scope groups* は、オリジングループの概念と、コミットスコープルールを定義するときにこれらを活用する方法を紹介します。

- *Commit scope rules* は、コミットスコープルールの構文と定式化方法を検討します。
- *Comparing durability options* は、耐久性に関してコミットスコープオプションがどのように動作するかを比較します。
- *Degrading commit scope rules* は、より厳密な設定でタイムアウトが発生した場合に、より低い設定に正常に低下できるコミットスコープルールを設定する方法を示します。

### 6.7.2 コミットスコープの種類

- **同期コミット** は、従来の同期レプリケーションと同様の方法で動作するコミットスコープメカニズムですが、コミットスコープフレームワーク内からです。
- **クォーラムコミット** は、分散トランザクションの一貫性が必要なワークロードのコミットスコープの種類です。クォーラムコミットは、すべての参加ノードが一緒にコミットまたはロールバックすることを保証し、競合が発生する前に防止します。
- *Commit At Most Once* は、 Commit At Most Once オプションに焦点を当てています。このオプションでは、アプリケーションが再試行する前にトランザクションがコミットされたことを確認します。これにより、コミットが多くても 1 回のみ発生することが保証されます。
- **ラグコントロール** は、最も遅いノードに応じてノードを動的に調整し、データベースノードがサービス停止になったときに同期ノードがどの程度非同期になるかを調整するコミットスコープメカニズムを調べます。
- *Group Commit (legacy)* は、 Group Commit オプションに焦点を当てており、ノードのグループが完了したことに同意したときにトランザクションを完了として定義できます。

### 6.7.3 コミットスコープを使用した作業

- *Administering* は、使用中のコミットスコープを使用して PGD クラスターを管理する方法に対処します。
- *Legacy synchronous replication using PGD* は、PGD で従来の Postgres 同期操作にアクセスする方法を示しています。
- *Internal timing of operations* は、古いレプリケーションと PGD の非同期および同期操作、特にトランザクションがディスクにフラッシュされるまたは可視される順序の違いを比較します。

#### Overview of durability options

##### 概要

EDB Postgres 分散 PGD では、 **コミットスコープ**を使用して、耐久性、一貫性、可用性、パフォーマンスのニーズに基づいて、いくつかのレプリケーション構成から選択できます。

基本構成では、PGD は非同期レプリケーションを使用します。ただし、コミットスコープは、デフォルトとトランザクションごとの動作の両方を変更できます。

ビルトインの物理または論理レプリケーションと同じ方法で、標準の `synchronous_standby_names` を使用して古い Postgres 同期レプリケーションを構成することもできます。ただし、コミットスコープは、レプリケーション動作の柔軟性と制御を提供します。

さまざまな同期設定は、関連しているがすべて個別に実装できるアプリケーションの 3 つのプロパティに影響を与えます。

- 耐久性複数のノードに書き込むと、クラッシュ耐性が向上し、クラッシュと再起動後にデータを復旧できます。
- 可視性クライアントへのコミット確認により、データベースは、一部のノードセットでコミットされたトランザクションの即時の可視性を保証します。
- 競合処理競合はコミット後に楽観的に処理でき、コミットタイムスタンプに基づいてトランザクションがレプリケートされるときに競合が解決されます。または、プリコミット時に悲観的に処理できます。クライアントは、最終的にすべてのノードに適用されるトランザクションに依存して、さらなる競合または中止を取得して、クライアントにエラーを直接通知できます。

コミットスコープでは、トランザクションの 4 種類の制御耐久性が許可されます。

- **同期コミット**この種のコミットスコープでは、オリジンノードが大部分のノードの確認を待機する動作が可能になり、ネイティブの Postgres 同期コミットのように動作します。
- **クォーラムコミット**クォーラムコミットは、内部的に 2 フェーズコミットを使用して、クラスター全体でトランザクションがコミットまたはロールバックを保証し、クラスター全体で競合するトランザクションを効果的にシリアル化します。
- *Commit At Most Once* この種のコミットスコープはクライアントと連携して、再試行前にトランザクションがコミットされたことを確認し、フェールオーバーシナリオでの重複コミットを防ぎます。
- **ラグコントロール**この種のコミットスコープは、コミットを続行する前にレプリケーションの観点からノードがどれだけ遅れていてもよいかを制御します。

同期コミットと CAMO は、それぞれ、耐久性のさらなる制御のために *Degrading commit scope rules* をサポートしています。

### Durability terminology

#### 耐久性の用語

このページでは、PGD の耐久性オプションに直接関連する用語と定義をカバーします。他の用語については、*Postgres Distributed terminology* を参照してください。

#### ノード

PGD ノードは、トランザクションのレプリケーション中にさまざまなロールを引き受けます。これらはトランザクションごとに暗黙的に割り当てられ、同時トランザクションの場合でも関係ありません。

- *origin* は、クライアントまたはアプリケーションからトランザクションを受信するノードです。最初にトランザクションを処理するノードであり、他の PGD ノードへのレプリケーションを開始し、確認またはエラーでクライアントに応答します。
- **オリジンノードグループ** は、オリジンを含む PGD グループです。
- **パートナー**ノードは、グループコミット要件に従ってトランザクションを確認することが想定されている PGD ノードです。

- **コミットグループ**は、コミットに関係するすべての PGD ノードのグループです。つまり、オリジンとそのすべてのパートナーノード。いくつかのまたはすべてのピアノードにすることができます。

### Commit scopes

コミットスコープにより、アプリケーションは EDB Postgres Distributed の耐久性と一貫性に関する詳細な制御を提供します。

コミットスコープは、トランザクションがコミットされるときのシステムの動作を記述する一連のルールです。実際の動作は、コミットスコープのルールがどの種類のコミットスコープを使用するかによって異なります**同期コミット** 、 **クォーラムコミット** 、 *Commit At Most Once* 、 **ラグコントロール**、またはこれらの組み合わせを使用します。

ほとんどのコミットスコープの種類はトランザクションの処理を制御しますが、Lag Control は例外で、遅いまたはキューアップされているレプリケーション操作に応じてシステムのパフォーマンスを動的に調整します。ただし、通常は他のコミットスコープ種類と組み合わせて使用されます

### コミットスコープ構造

すべてのコミットスコープには名前 `commit_scope_name` があります。

各コミットスコープには 1 つ以上のルールがあります。

コミットスコープ内の各ルールには、コミットスコープルールを一意に識別する `origin_node_group` があります。

`origin_node_group` は PGD グループであり、トランザクションのオリジネーターであるときにこのルールを適用するノードを定義します。

最後に、これらのトランザクションに適用されるコミットスコープまたはコミットスコープ種類の組み合わせを定義するルールがあります。

したがって、コミットスコープに次のルールがある場合

origin_node_group :=‘example_bdr_group’、rule :=‘MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT’,

次に、ルールは、`example_bdr_group` のノードがトランザクションを発行したときに適用されます。

ルール自分自身は、変更を確認するために指定されたグループのノードの数を指定します。`MAJORITY ORIGIN GROUP` はコミットスコープグループであり、`SYNCHRONOUS COMMIT` はコミットスコープの種類です。これは、変更がコミットされたと見なされる前に、元のノードのグループ内のノードの大部分が変更を確認する必要があることを意味します。

### コミットスコープの選択方法

変更が発生すると、PGD はトランザクションまたはノードに使用するコミットスコープを検索します。

トランザクションがコミットスコープを指定している場合、そのスコープが使用されます。

指定しない場合、システムはデフォルトのコミットスコープを検索します。デフォルトのコミットスコープはグループレベルの設定です。システムはグループツリーを参照します。ノードのグループを含むグルー

プツリーの下部から始めて上へ、default_commit_scope 設定が定義されているグループを検索します。このコミットスコープが使用されます。

default_commit_scope が見つからない場合、ノードの GUC、bdr.commit_scope が使用されます。そして、それが設定されていないか、local に設定されている場合、コミットスコープは適用されず、PGD の非同期レプリケーションが使用されます。

コミットスコープはローカルでなく、コミットが実行されているノードが origin_node_group に直接的または間接的に関連していない場合、使用されません。

### コミットスコープの作成

bdr.create_commit_scope を使用して、ルール例をコミットスコープに追加します。例

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := example_bdr_group,
    rule := MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT,
    wait_for_ready := true
);
```

これにより、example_bdr_group から発生したトランザクションのルール MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT が example_scope と呼ばれるスコープに追加されます。

example_scope に以前にルールが存在しなかった場合、このルールを追加するとスコープが存在します。

ルールが追加されるときは、origin_node_group が既に存在する必要があります。そうしない場合、追加操作全体がエラーで破棄されます。

ルールが評価されます。ルールが存在しないグループを記載しているか、グループの設定がグループのノードの他の構成設定と互換性がない場合、警告が表示されますが、ルールは追加されます。

ルールが追加されると、コミットスコープが使用できるようになります。

wait_for_ready は、ルールが関連するノードに追加されるまで、bdr.create_commit_scope() 呼び出しをブロックするかどうかを制御します。設定はデフォルトで true であり、省略できます。

### コミットスコープを使用する

この例のスコープを使用するには、トランザクション内で bdr.commit_scope を設定できます

```
BEGIN;
SET LOCAL bdr.commit_scope = example_scope;
...
COMMIT;
```

トランザクションがデータを書き込む前に、コミットスコープを設定する必要があります。

bdr.alter_node_group_option を使用して、グループまたはサブグループのデフォルトとしてコミットスコープを設定できます。

```
SELECT bdr.alter_node_group_option(
  node_group_name := example_bdr_group,
  config_key := default_commit_scope,
  config_value := example_scope
);
```

グループまたはサブグループのデフォルトを完全にクリアするには、default_commit_scope 値を local に設定します。

```
SELECT bdr.alter_node_group_option(
  node_group_name := example_bdr_group,
  config_key := default_commit_scope,
  config_value := local
);
```

PGD CLI を使用してこの変更を行うこともできます。

```
pgd set-group-options example-bdr-group --option default_commit_scope=example_scope
```

また、PGD CLI を使用して値を local に設定することにより、デフォルトをクリアできます。

```
pgd set-group-options example-bdr-group --option default_commit_scope=local
```

最後に、次を使用してノードのデフォルトの commit_scope を設定できます。

```
SET bdr.commit_scope = example_scope;
```

PGD のデフォルトの非同期レプリケーションを使用するには、bdr.commit_scope から local を設定します。

### Predefined commit scopes

EDB Postgres DistributedPGD は、使用可能な次の事前定義されたコミットスコープを提供します。

- ローカルプロテクト
- ラグプロテクト
- マジョリティプロテクト
- アダプティブプロテクト

#### ローカルプロテクト

```
ASYNCHRONOUS COMMIT
```

local protect コミットスコープは、PGD Essential のデフォルトのコミットスコープです。耐久性保証のない非同期コミットを提供します。これは、トランザクションがローカルノードの WAL に書き込まれるとすぐに、クラスター内の他のノードからの確認を待たずにコミットされたと考えられることを意味します。

このコミットスコープは、高可用性と低遅延がデータの耐久性よりも重要であるシナリオに適しています。ただし、ノードに障害またはネットワークの問題が発生した場合のデータ損失に対する保証はありません。

### ラグプロテクト

```
MAJORITY ORIGIN GROUP LAG CONTROL (max_lag_time = 30s, max_commit_delay = 10s)
```

lag protect コミットスコープは、マジョリティオリジングループのラグタイムに基づいて、耐久性保証を提供します。これにより、遅延時間が指定された制限内この場合 30 秒であり、コミット遅延も指定された制限この場合 10 秒以内である場合にのみトランザクションがコミットされたと見なされます。これは、ネットワークの問題またはノードに障害が発生した場合のデータの損失を防ぐのに役立ちます。

このコミットスコープは、データの一貫性と耐久性が重要であるが、ある程度の遅延が許容されるシナリオで役立ちます。指定された遅延とコミット遅延制限内でオリジングループの大部分のノードによって確認されるまで、トランザクションがコミットされたとみなされないことを保証することにより、パフォーマンスとデータの安全性のバランスが可能です。

### マジョリティプロテクト

```
MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT
```

majority protect コミットスコープは、マジョリティオリジングループに基づいて耐久性保証を提供します。これにより、トランザクションは、オリジングループ内のノードの大部分によって確認された場合にのみコミットされたと見なされます。これは、ノードに障害またはネットワークの問題が発生した場合に、データの一貫性と耐久性を保証するのに役立ちます。

このコミットスコープは、データの一貫性と耐久性が重要なシナリオに適しており、local protect コミットスコープと比較してデータ損失に対する高レベルの保護を提供します。ただし、トランザクションがコミットされたものと見なされる前にマルチプルのノードからの確認が必要であるため、多少のレイテンシーが発生する場合があります。

### アダプティブプロテクト

```
MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT DEGRADE ON (timeout = 10s, require_write_
↪lead = true) TO ASYNCHRONOUS COMMIT
```

adaptive protect コミットスコープは、より柔軟な耐久性保証を提供します。これにより、トランザクションはマジョリティオリジングループの同期コミットに基づいてコミットされたと考えられますが、指定されたタイムアウトこの場合 10 秒内にトランザクションを確認できない場合、非同期コミットに低下する可能性があります。これは、ネットワーク遅延またはノードの障害により、トランザクションの確認に遅延が発生する可能性があるシナリオで役立ちます。

このコミットスコープは、データの一貫性と耐久性が重要であるが、潜在的な遅延に対処するためにある程度の柔軟性が必要であるシナリオに適しています。local protect コミットスコープと比較してより高いレベルの保護を提供しながら、指定されたタイムアウト内にノードの大部分が確認できない場合でもトランザクションがコミットされたと見なすことにより、パフォーマンスとデータの安全性のバランスを提供します。

例

- SET LOCAL コマンドを使用して、トランザクションの lag protect コミットスコープを設定します。

```
BEGIN;
SET LOCAL bdr.commit_scope = lag protect;
-- Your transaction statements here
COMMIT;
```

これにより、トランザクションが指定されたコミットスコープでコミットされることが保証され、目的のレベルの耐久性と可用性が提供されます。

- SET コマンドを使用して、セッション全体の majority protect コミットスコープを設定します。

```
SET bdr.commit_scope = majority protect;
```

これにより、セッションで実行されるすべてのトランザクションが指定されたコミットスコープを使用することが保証され、目的のレベルの耐久性と可用性が提供されます。

- bdr.alter_node_group_option() ファンクションを使用して、PGD グループの adaptive protect コミットスコープを設定します。

```
SELECT bdr.alter_node_group_option(
        node_group_name:=mygroup,
        config_key:=default_commit_scope,
        config_value:=adaptive protect);
```

これにより、指定された PGD グループで実行されたすべてのトランザクションが指定されたコミットスコープを使用し、セッションまたはトランザクションレベルの設定でオーバーライドされない限り、目的のレベルの耐久性と可用性を提供します。

### Commit scope groups

PGD でトランザクションがコミットする前に、1 つ以上のノードがそれを確認する必要があります。コミットスコープグループは、それらがどのノードであるかを指定するコミットスコープルールの一部です。

例、次のコミットスコープ定義の場合。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := top_group,
    rule := MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT,
    wait_for_ready := true
);
```

ルール MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT では、SYNCHRONOUS COMMIT はコミットスコープの種類であり、確認の動作を制御します。MAJORITY ORIGIN GROUP はコミットスコープグループであり、元のノードと同じサブグループ内のノードの大部分が確認する必要があることを意味します。コミットスコープルールの構文の完全な説明については、*Commit scope rules* を参照してください。

コミットスコープグループは、しきい値を設定する量子子 MAJORITY などの 2 つの部分で構成され、もう 1 つはカウントするノードを識別するグループ指定子 ORIGIN GROUP などです。

### 数量子

量指定子は確認しきい値を設定し、トランザクションをコミットする前にグループ指定子のノードの数を確認する必要があります。

- ALL では、グループ内のすべてのノードを確認する必要があります。単一ノードに障害が発生すると、ノードが復旧するかタイムアウトが満了するまで、新しいコミットが妨げられます。
- MAJORITY では、ノードの半分を超えて 3 つ中 2 つ、または 5 つ中 3 つを確認する必要があります。この量子は、クォーラムを維持しながら半分未満のノードの障害を許容するため、高可用性構成の標準の選択です。
- ANY n で確認するには、少なくとも n ノードが必要です。ANY 1 は、少なくとも 1 つのノードを意味します。 ANY 2 は、少なくとも 2 つを意味します。これにより、きめの細かい制御が可能になり、完全なマジョリティを必要とせずにリモートリージョンから少なくとも 1 つの確認が必要な場合に役立ちます。

### グループ指定子

グループ指定子は、どのノードがトランザクションを確認する資格があるかを識別します。

### 明示的なグループ名

明示的なグループ名を使用して、クラスター内に存在する特定の名前付けの PGD グループをターゲットにします。この動作は修正され、どのノードがトランザクションを開始したかに関係なく、予測可能です。例、MAJORITY (region_a) 、ALL (region_a) 、または ANY 2 (region_a) 。

### クラスター

CLUSTER は、クラスター内のすべてのサブグループにわたって、クラスター内のすべてのデータノードから選択します。クォーラムを単一のサブグループではなくクラスター全体にまたがる場合に、CLUSTER を使用します。例、MAJORITY CLUSTER または ALL CLUSTER 。

### ORIGIN_GROUP

ORIGIN_GROUP は、トランザクション元のノードを含むサブグループを動的に解決し、固定グループに名前を付けるのではなくトランザクションが書き込まれる場所に適応します。例、MAJORITY ORIGIN GROUP または ALL ORIGIN GROUP 。

> **注釈**
>
> *ORIGIN_GROUP* および `ORIGIN GROUP` は互換性があります。

ORIGIN_GROUP は、ローカルルーティングを使用する展開の標準の選択です。各サブグループには独自の書き込みリーダーがあり、アプリケーションはローカルリージョンに接続します。サブグループごとに個別のルールを記述するのではなく、ORIGIN_GROUP を使用する単一のルールは、どのサブグループがオリジンであるかに関係なく、正しく適用されます。

たとえば、トランザクションが region_a のノードから発生した場合、ORIGIN_GROUP は region_a に解決されます。 region_b から発生した場合、region_b に解決されます。同じコミットスコープは、どちらの場合も正しく動作します。

### ORIGIN_GROUP ではありません

NOT ORIGIN_GROUP は、起点ノードを含まないすべてのサブグループに動的に解決されます。これは ORIGIN_GROUP を補完するものであり、マルチリージョンルールでそれと一緒に使用されるように設計されています。例、ANY 1 NOT ORIGIN_GROUP または ALL NOT ORIGIN_GROUP 。

NOT ORIGIN_GROUP は、通常、単独で使用されることはありません。その目的は、MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT AND ANY 1 NOT ORIGIN GROUP SYNCHRONOUS COMMIT と同様に、ORIGIN_GROUP を介してローカルリージョンを既にカバーしているルールにクロスリージョンの確認要件を追加することです。

### グループごとに

明示的なリストを使用する FOR EACH GROUP では、各名前付けグループ内で独立して量指定子が満たされる必要があります。たとえば、MAJORITY FOR EACH GROUP (region_a, region_b) は、 region_a のノードの大部分と region_b のノードの大部分が、トランザクションがコミットする前に両方とも確認する必要があることを意味します。

FOR EACH GROUP は、両方のグループのノードをプールし、合計の大部分を必要とする MAJORITY (region_a, region_b) より厳密です。 FOR EACH GROUP では、すべてのグループ内で個別にしきい値を満たす必要があります。

規制要件または可用性保証で、コミットが確認される前にすべてのリージョンでデータの耐久性が必要な場合など、各リージョンがトランザクションを独立して確認する必要がある場合に、FOR EACH GROUP を使用します。

### クラスター内のグループごとに

FOR EACH GROUP IN CLUSTER は、明示的なリストを使用した FOR EACH GROUP の動的同等です。サブグループに個々に名前を付けるのではなく、クラスター内のすべてのサブグループに量指定子を自動的に適用します。例、MAJORITY FOR EACH GROUP IN CLUSTER または ANY 1 FOR EACH GROUP IN CLUSTER 。

サブグループが追加または削除されると、ルールはコミットスコープ定義を変更せずに適応します。FOR EACH GROUP IN CLUSTER は、サブグループのセットが時間の経過とともに変更される可能性がある 3 つ以上のリージョンを持つクラスター、またはクラスターが成長しても単一のコミットスコープ定義を有効なままにしておきたい場合に、FOR EACH GROUP IN CLUSTER を使用します。

### グループの結合

グループ式を AND と組み合わせて、複数のグループからの確認を必要とします。各式は独立して評価され、トランザクションがコミットする前にすべてが満たされる必要があります。

#### ローカルマジョリティとリモート確認が必要

最も一般的なマルチリージョンパターンでは、ローカルリージョン内のノードの大部分に加えて、他のリージョンからの少なくとも 1 つの確認が必要です。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := ha_scope,
    origin_node_group := top_group,
    rule := MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT AND ANY 1 NOT ORIGIN GROUP␣
↪SYNCHRONOUS COMMIT,
    wait_for_ready := true
);
```

どのサブグループがトランザクションを発生したかに関係なく、ORIGIN GROUP はそのサブグループに解決され、NOT ORIGIN GROUP は他のすべてのサブグループに解決されます。このルールでは、常にローカルマジョリティとローカルサブグループの外部からの少なくとも 1 つの確認が必要で、コミットされたすべてのトランザクションが少なくとも 2 つのリージョンで確認されていることを保証します。リモート確認により、他の場所でのトランザクションの耐久性が保証されるため、単一リージョンの停止がデータ損失を引き起こす可能性はありません。

#### すべてのローカルノードとリモート確認を必要とする

サブグループが固定され、どのグループが参加するかを明示したい場合

```
SELECT bdr.create_commit_scope(
    commit_scope_name := explicit_scope,
    origin_node_group := region_a,
    rule := ALL (region_a) SYNCHRONOUS COMMIT AND ANY 1 (region_b) SYNCHRONOUS COMMIT,
    wait_for_ready := true
);
```

region_a のすべてのノードと region_b の任意の 1 つのノードを確認する必要があります。明示的なグループ ORIGIN_GROUP ではなくで ALL を使用する場合、これは通常、すべてのコミットで両方のセンターが表される必要がある 2 データセンター構成に適切な構造です。

これらのグループを使用する対応するコミットスコープの種類については、クォーラムコミット および 同期コミット を参照してください。

### Commit scope rules

コミットスコープルールは、コミットスコープメカニズムの中核です。これらは、コミットスコープが適用する内容を定義します。

コミットスコープルールは、組み合わせて動作する 1 つ以上の操作で構成されます。ルール間で AND を使用します。

各オペレーションは、コミットスコープグループ、オプションの確認レベル、および独自のパラメーターを持つことができるコミットスコープの種類の 2 つまたは 3 つの部分で構成されています。

```
commit_scope_group [ confirmation_level ] commit_scope_kind
```

完全な正式な構文図は、 Commit scopes リファレンスで入手できます。

MAJORITY ORIGIN GROUP QUORUM COMMIT ABORT ON (timeout = 6s) などの一般的なコミットスコープルールは、そのコンポーネントに分割できます。MAJORITY ORIGIN GROUP は、コミットスコープグループであり、どのノードが応答して、トランザクションを処理したことを確認する必要があるかを指定します。この例では、元のノードのグループ内のノードの大部分が確認する必要があります。

確認レベルが指定されていません。これは、デフォルト visible が使用されることを意味します。 visible を使用すると、ノードは、すべてのトランザクションの変更がディスクにフラッシュされ、他のトランザクションから可視されると確認します。

このオペレーションの最後の部分はコミットスコープの種類です。この例では QUORUM COMMIT です。QUORUM COMMIT は、ノードがコミットする前にすべての参加ノードを調整する同期 2 フェーズコミットです。ABORT ON 句は必須で、トランザクションを中止する前にクォーラムを待機する時間を設定します。

### コミットスコープグループ

コミットスコープグループには、 ANY 、ALL 、および MAJORITY の 3 種類があります。これらのすべての後には、かっこ内の 1 つ以上のグループのリストが続きます。グループのこのリストを組み合わせて、この操作が適用されるノードのプールを作成します。このリストには、NOT を付けることができます。これは、プールをリストにない他のすべてのグループに反転します。

- ANY n の後には、整数値 n が続きます。これは、リストされたグループのノード内の n ノードに変換されます。
- ALL の後にはグループが続き、リストされたグループのノード内のすべてのノードに変換されます。
- MAJORITY はグループが続き、マジョリティを与えるには、リストされたグループのノードの半分に 1 つを加えたものを必要とすることを意味します。
- ANY n NOT の後には、整数値 n が続きます。これは、リストされたグループのノードにない n ノードに変換されます。
- ALL NOT の後にはグループが続き、リストされたグループのノードにないすべてのノードに変換されます。
- MAJORITY NOT はグループが続き、マジョリティを与えるには、リストされたグループのノードにないノードの半分と 1 を必要とすることを意味します。

上記の式はすべて、評価時にグループ内のデータノードのみを考慮します。Witness ノードおよびその他の非データノードは無視されます。

### 確認レベル

PGD ノードは、さまざまなタイミングでトランザクションの確認を送信できます。保護レベルが増加すると、確認ノードの観点から、これらは次のとおりです。

- received _ リモート PGD ノードは、ローカルアプリケーションを開始する前に、トランザクションを受け取ったらすぐに確認します。
- replicated _ トランザクションの変更を適用した後、ディスクにフラッシュする前に確認します。

- durable _ すべての変更がディスクにフラッシュされた後、トランザクションを確認します。
- visible デフォルト_ すべての変更がディスクにフラッシュされ、同時トランザクションに可視された後、トランザクションを確認します。

コミットスコープのルールでは、次のように ON のかっこ内のノードグループ定義にこれらの確認レベルを追加できます。

- ANY 2 (right_dc) ON replicated
- ALL (left_dc) ON visible デフォルト
- ALL (left_dc) ON received AND ANY 1 (right_dc) ON durable

> **注釈**
>
> PostgreSQL の `synchronous_standby_names` 機能に慣れている場合は、 *synchronous_standby_names* とコミットスコープの文法は似ているように見えますが、微妙な違いがあることに注意してください。前者は起点ノードを考慮しませんが、後者は考慮します。たとえば、 *synchronous_standby_names = 'ANY 1 (..)'* は、*ANY 2 (...)* のコミットスコープと同等です。この違いにより、マジョリティに関する推論が簡単になり、起点ノードがトランザクションの耐久性と可視性にも貢献することを反映しています。

### コミットスコープの種類

現在、4 つのコミットスコープ種類があります。以下は概要と、詳細へのリンクです。

#### 同期コミット

同期コミットは、ネイティブの Postgres synchronous_commit オプションのように動作するように設計されたコミットスコープオプションですが、コミットスコープ環境内から使用できます。これは、オプションの DEGRADE ON 句をサポートする同期非 2 フェーズコミットオペレーションです。このオプションの前にあるコミットスコープグループは、SYNCHRONOUS COMMIT が使用するグループと確認要件を制御します。

詳細については、 同期コミット を参照してください。

#### クォーラムコミット

クォーラムコミットは、ノードがコミットする前にすべての参加ノードを調整する同期 2 フェーズコミットです。すべての参加ノードは一緒にコミットまたはロールバックします。ノードがローカルにコミットできる前に、クォーラムコミットは熱心な競合解決を使用して、事実後解決するのではなく不一貫性を防止します。サポートされているコミットスコープグループは、ALL 、MAJORITY ORIGIN GROUP 、MAJORITY CLUSTER 、および MAJORITY FOR EACH GROUP です。

詳細については、 クォーラムコミット を参照してください。

### CAMO

Commit At Most Once または CAMO により、クライアント/アプリケーション、オリジンノード、およびパートナーノードは、トランザクションが最大 1 回データベースにコミットされることを保証できます。クライアントはプロセスに関係するため、CAMO プロセスに参加するにはアプリケーションを変更する必要があります。

詳細については、 *Commit At Most Once* を参照してください。

#### ラグコントロール

ラグ制御を使用すると、システムのレプリケーションパフォーマンスが指定された制限を超えると、データベースとのクライアントの相互作用にコミット遅延を自動的に挿入し、クライアントにバックプレッシャーを提供できます。Lag Control には、適用できる最大コミット遅延を設定するパラメーターがあります。また、そのコミット遅延の増加をトリガーするキューサイズまたは処理時間の面でも制限があります。

詳細については、 ラグコントロール を参照してください。

コミットスコープルールは、組み合わせて動作する 1 つ以上の操作で構成されます。AND を使用して 1 つのルールを形成します。例

```
MAJORITY (Region_A) SYNCHRONOUS COMMIT AND ANY 1 (Region_A) LAG CONTROL (MAX_LAG_SIZE␣
↪= 50MB)
```

最初の操作は、Region_A の大部分に対する同期コミットを設定します。2 番目の操作は、 Region_A のノードのいずれかが 50MB を超える遅延がある場合に、コミット遅延の押し上げを開始する遅延制御を追加します。この操作の組み合わせにより、ノードに遅延が発生しているときに遅延制御が動作できます。

### Comparing durability options

#### 比較

PGD で使用可能な同期レプリケーションのオプションのほとんどは、さまざまなレベルの同期を許可し、パフォーマンスとノードまたはネットワークの停止に対する保護の間のさまざまなトレードオフを提供します。

次の 確認レベル リストは、その確認レベルが有効なときにユーザーが表示することを予期すべきものと、それがパフォーマンス、耐久性、一貫性にどのように影響を与える可能性があるかを説明します。

!!!注 クォーラムコミット は、ON DURABLE および ON VISIBLE のみをサポートしています。 ON VISIBLE がデフォルトです。

#### 受信時

**Expect** ピアノードは変更を受信しました。ピアノードのタプルストアでは何も更新されず、ストレージに書き込まれません。

受信時の確認は、通常に動作しているピアが、完全または部分的なネットワークの停止に直面しても、最終的にそれ以上の通信を必要とせずにトランザクションを適用できることを意味します。この確認には永続ストレージが関係しないため、ピアノードがクラッシュしてもトランザクションの再送信が必要になる場合があります。

**の場合**トランザクションの起点ノードは、トランザクションの受信を待機するだけです。トランザクションが大きい場合、システムの TPS パフォーマンスが向上する場合があります。

**対** 古い読み取りの可能性の増加。Postgres サーバーまたはオペレーティングシステムのクラッシュが発生すると、データが失われる可能性があるため、全体的に ON RECEIVED は堅牢ではありません。

#### レプリケートについて

**Expect** ピアノードは変更を受信し、タプルストアに適用しました。変更はストレージに書き込まれましたが、ストレージはディスクにフラッシュされていません。

レプリケーションの確認は、ピアが変更を受信して適用したことを意味します。これらの変更はストレージに書き込まれますが、オペレーティングシステムのキャッシュとバッファーに残っています。システムはそれらをディスクにまだ永続化していません。

**For** このチェックポイントは、トランザクション処理のタイムラインのさらに下方にあります。オリジンノードは、トランザクションが適用されるのを待機するだけで、永続化されません。

**対** ON RECEIVED よりも古い読み取りの可能性がわずかに低くなります。また、 ON REPLICATED では、データは Postgres のクラッシュから生き残ることができますが、オペレーティングシステムのクラッシュからは生き残れません。

### ON DURABLE

**Expect** ピアノードは変更を受信し、タプルストアに適用し、変更をストレージに永続化しました。変更はまだ他のセッションで利用できるようにしていません。

永続的な確認は、トランザクションが書き込まれ、ピアノードのストレージにフラッシュされたことを意味します。これにより、ピアノードのクラッシュと復旧後のデータの損失から保護されます。ただし、セッションが切断して再接続する前に ON DURABLE ルールを使用してトランザクションをコミットした場合、トランザクションの変更は再接続されたセッションに可視される保証はありません。

Quorum Commit コミットスコープ種類と使用する場合、これは変更が可視されることも意味します。

**の場合**より堅牢で、クラッシュが発生した場合に再送信なしで回復できます。

**Against** フェイルオーバーが発生した場合の一貫性を保証しません。

#### オンビジブル

**Expect** ピアノードは変更を受信して適用し、それらの変更を永続化してストレージにフラッシュしました。

可視性の確認は、トランザクションが完全にリモートで適用されたことを意味します。切断して再接続する前に、セッションが ON VISIBLE ルールを使用してトランザクションをコミットした場合、トランザクションの変更は再接続されたセッションに可視されることが保証されます。

**対象** 堅牢で一貫性があります。

**対** パフォーマンスが低下します。

### Degrading commit scope rules

`SYNCHRONOUS COMMIT` および `CAMO` には、それぞれ、特定のパフォーマンスしきい値を超えたときにトランザクションの要件を低下させるオプション機能があります。`GROUP COMMIT` は機能を低下させることはできませんが、タイムアウトで中断することはできます。

ノードがトランザクションを適用しており、そのトランザクションがタイムアウトした場合、単にロールバックするのではなく、完了するトランザクションの要件を低下させるプロセスをトリガーすると役立つ場合があります。

DEGRADE ON は、トランザクションのコミットスコープルールを正常に低下させるためのルートを提供します。最も単純では、DEGRADE ON は、タイムアウトと、コミットスコープが正常に低下できるコミットスコープ操作の 2 番目のセットを取得します。

たとえば、20 ミリ秒または 30 ミリ秒のタイムアウト後、コミットスコープを満たすための要件は ALL (node_group_name) SYNCHRONOUS COMMIT から MAJORITY (node_group_name) SYNCHRONOUS COMMIT に低下し、トランザクションがより安定して適用されるようになります。

degrade 句がトリガーされるように、書き込みリーダーがトランザクションのオリジネーターであることを必要とすることもできます。これは、たとえば 2 つのデータノードと 1 つの監視ノードがある「スプリットブレインシナリオ」で役立つ場合があります。2 つのデータノードの間でネットワークが分割され、両方のデータノードへの接続があると仮定すると、監視ノードとの raft 選挙を介してリーダーに選ばれるのは一方だけであるため、一方のみが機能低下することは許可されます。

### 動作

一般化された DEGRADE 句がトランザクションに適用されたときにどのように動作するかについては、2 つの部分があります。

コミット中に 1 回、処理中のコミットがコミットスコープルールを満たす応答を待機している間、PGD はタイムアウトをチェックし、タイムアウトが満了している場合、処理中のコミットは DEGRADE のコミットスコープルールを待機するように再構成されます。条項。実際、この時点で、DEGRADE 句のコミットスコープルールは既に満たされている可能性があります。

このメカニズムだけでは、目的の動作には不十分です。これだけでは、すべてのトランザクション接続の問題により低下することが確実なトランザクションも含めて、低下モードがキックインする前にタイムアウトが満了するのを待機する必要があり、パフォーマンスに重大な影響を与えるためです。このような劣化クラスターシナリオ。

これを回避するために、PGD マネージャープロセスは定期的に 5 秒ごとに接続と適用レート bdr.node_replication_rates のものをチェックし、レプリケーションの現在の状態に基づいて、その時点で低下するコミットスコープがある場合、それらは自動的に低下—その後の処理時にそのコミットスコープを使用するトランザクションは、タイムアウトを待機する代わりに低下ルールを使用します—レプリケーションが再び十分に迅速に移動していることをマネージャープロセスが検出するまで。

### 同期コミットとグループコミット

SYNCHRONOUS COMMIT と GROUP COMMIT は両方とも timeout および require_write_lead パラメーターを持ち、デフォルトはそれぞれ 0 および false です。デフォルトの 0 では即時低下が発生するため、おそらく常に timeout を設定する必要があります。デグレードモードに切り替えるために、書き込みリーダーがトランザクションのオリジネーターであることを要求することもできます。ここでも、デフォルトは false です。 SYNCHRONOUS COMMIT の場合、 timeout および require_write_lead はデグレードに適用され、GROUP COMMIT の場合、これらのパラメーターはアボートに適用されます。他のノードからの確認を待機する前にプライマリで既にコミットされているため、GROUP COMMIT コミットスコープはデグレードできず、SYNCHRONOUS COMMIT コミットスコープはアボートできません。

SYNCHRONOUS COMMIT には、どのルールに低下できるかに関するオプションもあります。これは、どのルールから低下するかによって異なります。

まず、非同期操作に低下させることができます。

```
ALL (left_dc) SYNCHRONOUS COMMIT DEGRADE ON (timeout=20s) TO ASYNC
```

同じコミットスコープ種類のより制限の低いコミットグループに低下することもできます。これも種類が SYNCHRONOUS_COMMIT または GROUP COMMIT のいずれかである場合に限ります。たとえば、次のように機能を低下させることができます。

```
ALL (left_dc) SYNCHRONOUS COMMIT DEGRADE ON (timeout=20s) TO MAJORITY (left_dc)␣
↪SYNCHRONOUS COMMIT
```

または次のように。

```
ANY 3 (left_dc) SYNCHRONOUS COMMIT DEGRADE ON (timeout=20s) TO ANY 2 (left_dc)␣
↪SYNCHRONOUS COMMIT
```

ただし、 SYNCHRONOUS COMMIT から GROUP COMMIT に低下することはできません。

### CAMO

*Commit At Most Once* は、同じ timeout および require_write_lead パラメーター同じデフォルトでそれぞれ 0 および false をサポートしますが、オプションは非同期オペレーションに低下するだけであるという点で簡単です。

```
ALL (left_dc) CAMO DEGRADE ON (timeout=20ms, require_write_lead=true) TO ASYNC
```

繰り返しになりますが、デフォルトは 0 であるため、 timeout パラメーターを設定する必要があります。

#### 機能低下イベントの監視

bdr.stat_commit_scope を使用して、機能低下イベントを追跡します。3 つの主要なメトリックは、機能低下動作の可視性を提供します。

- ndegrades は、トランザクションごとの機能低下イベントを追跡します。コミット中にバックエンドが低下タイムアウトに達し、その特定のトランザクションを正常に低下させるたびにインクリメントされます。このメトリックは、機能の低下が発生した個々のトランザクションの合計数を示します。
- nconfig_degrades は、構成レベルの状態の縮退モードへの変更を追跡します。バックグラウンドワーカーマネージャープロセスが、ノードの可用性チェックに基づいて、コミットスコープの共有状態を低下に切り替えるときにインクリメントします。これは、コミットスコープ構成自分自身が縮退状態になった回数を示し、復旧するまでの後続のすべてのトランザクションに影響を与えます。
- last_state_change_time は、最後の構成レベルの状態変更のタイムスタンプを記録します劣化した状態への移行または回復。このメトリックを使用して、コミットスコープが通常と機能低下の間で最後に移行したときを正確に特定し、他のシステムイベントまたは停止との相関を追加します。

#### モニタリングクエリの例

機能低下イベントや最新の構成レベルの状態変更を含むコミットスコープの統計を表示します。

```
SELECT
  commit_scope_name,
  ndegrades,
  nconfig_degrades,
  last_state_change_time,
  CASE
    WHEN last_state_change_time IS NULL THEN never degraded
    WHEN age(now(), last_state_change_time) < interval 5 seconds THEN recently changed
    ELSE stable for  || age(now(), last_state_change_time)::text
  END AS state_change_info
FROM bdr.stat_commit_scope
WHERE ndegrades > 0 OR nconfig_degrades > 0
ORDER BY last_state_change_time DESC NULLS LAST;
```

これらのメトリックは、トラブルシューティングの時間的コンテキストとともに、トランザクションごとの即時フォールバックと長期的な状態管理の低下動作の両方に対する完全な可視性を提供します。

### Synchronous Commit

コミットスコープの種類 SYNCHRONOUS COMMIT

#### 概要

PGD の SYNCHRONOUS COMMIT は、基になるオペレーションでの PostgreSQL の synchronous_commit オプションに似た方法で動作するコミットスコープの種類です。ただし、PostgreSQL オプションとは異なり、コミットスコープとして構成され、PGD での構成と対話が簡単です。

GROUP COMMIT や CAMO などの他のコミットスコープの種類とは異なり、SYNCHRONOUS COMMIT オペレーションのトランザクションは 2 フェーズコミット 2PC トランザクションに変換されません。これらは Postgres synchronous_commit のように動作します。

#### 例

この例では、このコミットスコープが使用されている場合、left_dc グループ内のノードは SYNCHRONOUS COMMIT を使用して、left_dc グループ内の他のノードに変更をレプリケートします。 left_dc グループ内のノードの大部分を検索して、トランザクションをコミットしたことを確認します。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_sc_scope,
    origin_node_group := left_dc,
    rule := MAJORITY (left_dc) SYNCHRONOUS COMMIT,
    wait_for_ready := true
);
```

#### 構成

SYNCHRONOUS COMMIT は、オプションの DEGRADE ON 句をサポートしています。特定の構成パラメーターについては SYNCHRONOUS COMMIT コミットスコープリファレンスを参照してください。または、オプションの低下に関しては *Degrading commit scope rules* を参照してください。

#### 確認

| Confirmation level | PGD Synchronous Commit handling |
|---|---|
| *received* | A remote PGD node confirms the transaction once it's been fully received and is in the in-memory write queue. |
| *replicated* | Same behavior as *received*. |
| *durable* | Confirms the transaction after all of its changes are flushed to disk. Analogous to *synchronous_commit* = *on* in legacy synchronous replication. |
| *visible* (default) | Confirms the transaction after all of its changes are flushed to disk and it's visible to concurrent transactions. Analogous to *synchronous_commit* = *remote_apply* in legacy synchronous replication. |

#### 詳細

現在、SYNCHRONOUS COMMIT は、コミットスコープルール構文の確認レベルを使用しません。

コミットスコープルールでは、元のキーワード SYNCHRONOUS_COMMIT は SYNCHRONOUS COMMIT にエイリアスになりました。アンダースコアの代わりにスペースを使用すると、Postgres のネイティブ SYNCHRONOUS_COMMIT と区別するのに役立ちます。

### Quorum Commit

コミットスコープの種類 QUORUM COMMIT

#### 概要

標準の同期レプリケーションは、リモートノードがデータを受信したことを確認しますが、2 つのノードが独立して競合する変更をコミットすることを妨げるものではありません。2 つのノードはそれぞれ同じ行を読み取り、ローカル検証チェックを通過し、両方とも正常にコミットできます。クラスターは一貫性のない状態で、どちらの側でもエラーは報告されません。たとえば、決済システムでは、それぞれがローカルで口座残高をチェックする 2 つのノードは、両方ともアカウント制限を超える同時出金を承認できます。

クォーラムコミットは、ノードがコミットする前にノードの大部分の同意を必要とすることにより、この発散を防ぎます。クラスターが合意に達できない場合、トランザクションはどこでもロールバックされます。その結果、すべてのノードにわたって単一の一貫した状態が得られ、事実後に解決する発散や競合はありません。

クォーラムコミットは、下位互換性のためのみに保持されているグループコミットを置き換えます。不一貫を許可するように構成を誤る可能性のある柔軟な構文を提供したグループコミットとは異なり、クォーラムコミットは一貫性を保証できるものに構文を制限します。

### クォーラムコミットを使用する場合

不一貫が発生した後に解決できないワークロードには、クォーラムコミットを使用します。決済とコアバンキングが最も明確に適合し、クラスター全体で確認できないトランザクションはどこにもコミットしません。通信料金請求、厳しい監査とコンプライアンス要件を伴う規制業界、およびグローバル一貫性保証に依存する独自のデータベースからの移行も、同様に良い候補です。これらのワークロードには、従来、独自のデータベースが必要でした。Quorum Commit は、Postgres に同じ保証を提供します。

### How Quorum Commit works

クォーラムコミットでは、ノードがコミットする前に、すべての参加ノードが同意すること、熱心な競合解決を使用して、事実後解決するのではなく不一貫性を防ぐ必要があります。すべての書き込みは、PGD の *Overview of durability options* によって管理される単一の書き込みリーダーを介する必要があります。利用可能なノードが大部分よりも少ない場合、クラスターはクォーラムに到達できず、手動介入が必要です。

### 一貫性の保証

すべての保証が適用されるように、Postgres 17 以降をお勧めします。以前のバージョンもサポートされていますが、すべての保証が適用されるわけではありません。ルーティンググループ内のノードの大部分以上をカバーするコミットスコープで使用される場合、クォーラムコミットは次の保証を提供します。

- **コミットとは、どこでもコミットされることを意味します。**トランザクションは、必要な数のノードがコミットした場合にのみコミットします。アプリケーションがコミット確認を受信すると、必要なすべてのノードが同じ結果のデータを保持します。
- **書き込みリーダーは、常に最新のコミットされた状態を反映します。**コミット後の書き込みリーダーからの読み取りは、常に最新の結果を結果ます。この保証は、書き込みリーダーの変更全体にわたって保持されます。古い書き込みリーダーで表示されているトランザクションは、新しい書き込みリーダーで表示されます。
- **Aborted は、どこでも中断されることを意味します。**トランザクションが中断されると、その効果はすべてのノードでロールバックされます。クラスター内のどこにも部分的な状態は表示されません。
- **書き込みリーダー変更のゼロリカバリ時間 RTO0。**書き込みリーダーが変更されると、新しい書き込みリーダーはすぐにトランザクションを受け入れることができます。古い書き込みリーダーからの進行中のトランザクションは、調整によって解決され、新しいリーダーの操作をブロックしません。
- **データの相違なし。**クラスターのコミットされた状態から分岐するノードはありません。コミットの競合によるレプリケーションの中断は発生しません。

!!!警告上記の保証を可能にするために、クォーラムコミットは *Degrading commit scope rules* 句または `MAJORITY` より弱い *Commit scope groups* をサポートしていません。

### 熱心な競合解決

クォーラムコミットは、事後に競合を検出するのではなく、クォーラムフェーズ中に、競合するトランザクションのいずれかをコミットする前に、競合するトランザクションの 1 つを中止します。熱心な競合解決は、クォーラムコミットで常にアクティブであり、無効にすることはできません。コミット時にシリアル化エラーを処理し、失敗したトランザクションを再試行するようにアプリケーションを構成します。再試行ループでコミットをラップし、シリアル化エラーを処理するトランザクションテンプレートについては、開発者ガイドの クォーラムコミット を参照してください。

### トランザクションのライフサイクル

クォーラムコミットは内部で *Explicit two-phase commit (2PC)* を使用します。クォーラムに到達するまで、ノードはコミットしません。準備フェーズでは、オリジンノードがトランザクションをローカルに準備し、データの耐久性を高めますが、他のトランザクションからはまだ可視されていません。準備されたトランザクションは、すべての参加ノードに複製され、各ノードは独自のローカルコピーを準備します。

次に、起点ノードは、必要なコミットスコープグループから確認を収集します。必要な数を受け取ると、コミット決定を行い、すべての参加ノードに伝播し、すべてが一緒にコミットします。Raft コンセンサスがコミット決定に使用される場合、PGD のビルトイン Raft がオリジンノードではなく決定を行います。

ノードごとに発生するすべてのクォーラムコミットトランザクションを処理するには、*max_prepared_transactions* を `max_connections` と少なくとも同じ高さに構成します。

### 自動調整

クォーラムコミットトランザクションでは、オリジンノードは、必要な確認を受信するとコミット決定を行い、レプリケーションを介してその決定をすべてのノードに伝播します。トランザクションを準備した後、その決定がすべてのノードに到達する前にオリジンが失敗した場合、残りのノードは結果のない準備されたトランザクションを保持したままになります。

これらの準備されたトランザクションは、ロックおよびその他のリソースを保持するため、クラスターが進行する前に解決する必要があります。新しい書き込みリーダーはコーディネーターとして機能し、準備された各トランザクションに対して正しいコミットまたは中止の決定に到達する必要があります。オリジンが既にアボートしたトランザクションをコーディネーターがコミットするか、オリジンが既にコミットしたトランザクションをアボートすると、クラスターは分岐します。

PGD は、リコンシリエーションを介してこれらのトランザクションを自動的に解決することにより、この発散を防ぎます。*bdr.enable_auto_sync_reconcile* を介してデフォルトで有効になっています。

### 調整フェーズ

1. ノードが `bdr.raft_global_election_timeout` デフォルト 6s から聞いていない場合、PGD マネージャープロセスは調整要求を作成し、状態が `STARTED` であるそのオリジンの `bdr.reconcile_2pc_requests` にエントリーを追加します。
2. 新しい書き込みリーダーは Raft コンセンサスを介して選出され、通常は確認された LSN でオリジンから最も前方のノードであり、準備されたトランザクションとプリコミット決定を既に保持している可能性が最も高いノードです。この書き込みリーダーは、調整のためのコーディネーターとして機能します。
3. コーディネーターは、孤立した各準備済みトランザクションのコミットまたは中止の決定を決定します。

**``commit_decision = group`` デフォルトの場合**コーディネーターはコミット決定 CD キャッシュを調べます。

- CD キャッシュでプリコミット決定が見つかった場合、オリジンは失敗する前に既にコミットを決定していました。一部のノードは既にローカルにコミットしている場合があります。コーディネーターはトランザクションをコミットして、すべてのノードを契約させます。
- CD キャッシュでコミットまたはアボート決定が見つかった場合、最終的な決定は、以前の調整ラウンドまたは Raft を介して既に記録されています。コーディネーターはその決定に従い、示されたと

おりにローカルでコミットまたは中止します。

- 決定が見つからない場合、コーディネーターはすべてのノードを照会し、コミット決定または準備されたトランザクションを持つノードがあるかどうかを確認します。どこかでコミット決定が見つかった場合、トランザクションはコミットされます。それ以外の場合、コーディネーターは確認をカウントし、コミットスコープルールに従ってクォーラムに到達した場合、トランザクションをコミットします。

**``commit_decision = raft``を使用する場合** Raft は唯一のアービターとして機能します。コーディネーターはすべての未解決のトランザクションを中止し、以前の Raft の決定が優先されます。決定が記録されなかった場合、トランザクションはどこでもきれいに中止されます。

1. コーディネーターは Raft を介して決定を送信し、各ノードはそれに応じてトランザクションをコミットまたは中止します。
2. トランザクションはすべてのノードで解決されます。準備されたトランザクションが保持しているロックとリソースが解放されます。

リコンシリエーション中にコーディネーター自分自身に障害が発生すると、次の書き込みリーダーが引き継ぎます。このプロセスはべき等であり、決定の相違を引き起こさないことが保証されています。決定が Raft を介して既に送信されている場合、新しいコーディネーターは CD キャッシュでそれを見つけて、それにに従います。

### モニタリング リコンシリエーション

ノードに障害が発生した後、*bdr.reconcile_2pc_requests_summary* を照会して、調整が進行中であるかどうか、およびどのノードがコーディネーターとして機能しているかを確認します。進行中のヘルスモニタリングの場合、*bdr.stat_commit_scope* はスコープごとに `nreconciliations`、`ncommitted`、および `naborted` を追跡します。

### 自動同期との関係

Reconciliation は、automatic synchronization とともにバックグラウンドで自動的に実行されます。リコンシリエーションは最初にトランザクション結果を解決し、次に同期はこれらの結果を逃したノードに伝播します。

### コミットレイテンシー

コミット前の調整では、非同期レプリケーションと比較して、約 2 つのネットワークラウンドトリップが追加されます。1 つ以上のノードに遅延が発生している場合、確認の取得の遅延が大幅に発生する可能性があります。ピアノードが確認する前に、トランザクションをローカルに適用する必要があります。これにより、トランザクションサイズに比例して遅延が追加されます。

論理スタンバイノード、サブスクライバ専用ノード、およびまだ参加またはキャッチアップしているノードは、クォーラムに含まれませんが、最終的に変更を受信します。Witness ノードはクォーラムに参加できず、データの変更を受信しません。

### タイムアウト処理

構成されたタイムアウト内にクォーラムに到達しない場合、トランザクションは中止され、すべての参加ノードでロールバックされます。部分的なコミットは不可能です。すべてのノードがコミットするか、どれもコミットしません。トランザクションは、中止されると、ネットワークが復旧しても再送信されません。アプリケーションは再試行する必要があります。

クォーラムフェーズ中にノードに障害が発生すると、残りのノードはタイムアウトが満了するまで待機してから中止します。障害が発生したノードがタイムアウト前に復旧した場合、クォーラム試行は成功する可能性があります。切断されたノードはトランザクションを準備ができた状態に保ち、再接続するとマジョリティ決定と調整します。

## Using Quorum Commit

クォーラムコミットスコープは bdr.create_commit_scope で作成され、トポロジとルーティングモードに合わせて構成されます。スコープは作成されたら、ノードグループのデフォルトとして適用したり、セッションごとまたはトランザクションごとに設定したりできます。

!!!注クラスター内のすべてのノードは、ローリングアップグレード中を含めて、PGD 6.4.0 以降を実行している必要があります。

### コミットスコープの作成

クォーラムコミットの構文は次のとおりです。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := <scope_name>,
    origin_node_group := <node_group>,
    rule := <commit_scope_group> [ ON { durable | visible } ] QUORUM COMMIT
        [ ( commit_decision = { group | raft } ) ]
        ABORT ON ( timeout = <value> ),
    wait_for_ready := true
);
```

rule パラメーターは、次のオプションを受け入れます。

- **``<commit_scope_group>`` :** コミットする前にトランザクションを確認する必要があるノードを定義します。サポートされている値は ALL 、MAJORITY ORIGIN GROUP 、MAJORITY CLUSTER 、MAJORITY CLUSTER WITH LEADER 、MAJORITY FOR EACH GROUP (subgroup1, subgroup2, ...) 、または MAJORITY FOR EACH GROUP IN CLUSTER です。コミットスコープグループの詳細については、*Commit scope groups* を参照してください。

明示的なグループリストを使用した MAJORITY 、 ANY n 、および NOT はサポートされていません。

- **``commit_decision`` :** オプショナル。コミット決定に到達する方法を制御し、ノード障害後の 自動調整 の動作に影響を与えます。
- group デフォルトオリジンノードは、コミットスコープグループから必要な確認を受信すると、コミットすることを決定します。

- raft コミット決定は通過します PGD のビルトイン Raft コンセンサス。group より遅いですが、特定のマルチグループトポロジに必要です。
- **``ON { durable | visible }``：** オプショナル。確認レベルを設定します。リモートノードがトランザクションを処理したことを確認するポイントです。
- durable すべての変更がディスクにフラッシュされた後の確認
- visible デフォルトすべての変更がディスクにフラッシュされ、表示された後の確認同時トランザクションに。
- **``ABORT ON ( timeout = <value> )``：** 必須。アボートする前にクォーラムを待機する時間を設定します。この期間内にコンセンサスに到達しない場合、トランザクションはすべてのノードで中止およびロールバックされ、ネットワークが復旧しても再送信されません。<value> をインターバル、例、10s に設定します。値の選択に関するガイダンスについては、**タイムアウトの選択** を参照してください。

### トポロジ別の構成

適切なクォーラムコミット構成は、クラスタートポロジとルーティングモードによって異なります。

#### 単一ロケーションクラスター 3 つのデータノード

単一のサブグループ内に 3 つのデータノードがあり、サブグループレベルでルーティングが有効になっているクラスター。このトポロジは、一度に 1 つのノードの障害を許容します。MAJORITY ORIGIN GROUP をお勧めします。これは、トランザクションがコミットする前にマジョリティ 3 つのうち 2 つを確認する必要があります。書き込みをブロックせずに 1 つのノードに障害が発生する場合があります。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := top_group,
    rule := MAJORITY ORIGIN GROUP QUORUM COMMIT ABORT ON (timeout = 6s),
    wait_for_ready := true
);
```

#### 2 リージョンクラスター 3+3 データノード

それぞれ 3 つのデータノードからなる 2 つのサブグループを含むクラスター。クラスターは、リージョンごとに 1 つずつ、最大 2 つのノードの障害を許容します。

ローカルルーティングが有効になっている場合、MAJORITY FOR EACH GROUP IN CLUSTER は、各サブグループ内の多数決の確認が必要です。リージョン間のネットワーク分割により、トランザクションは設定されたタイムアウトまで待機してから中止されます。パーティションが修復されると、書き込みが自動的に再開されます。耐久性は、オリジンリージョンだけでなく、トランザクションがコミットする前に両方のリージョンで保証されます。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := top_group,
```

(次のページに続く)

(前のページからの続き)

```
    rule := MAJORITY FOR EACH GROUP IN CLUSTER QUORUM COMMIT ABORT ON (timeout = 6s),
    wait_for_ready := true
);
```

グローバルルーティングが有効になっている場合、MAJORITY CLUSTER WITH LEADER は、トランザクションがコミットする前に、書き込みリーダーを含むクラスター全体のマジョリティの確認を必要とします。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := top_group,
    rule := MAJORITY CLUSTER WITH LEADER QUORUM COMMIT ABORT ON (timeout = 6s),
    wait_for_ready := true
);
```

### 3 リージョンクラスター 3+3+3 データノード

それぞれ 3 つのデータノードからなる 3 つのサブグループを持つクラスター。

グローバルルーティングでは、MAJORITY CLUSTER WITH LEADER は、書き込みリーダーを含む 9 つのノードすべてのマジョリティの確認が必要です。このトポロジは、リージョン全体の損失を含む、最大 4 つのノードの障害を許容します。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := top_group,
    rule := MAJORITY CLUSTER WITH LEADER QUORUM COMMIT ABORT ON (timeout = 6s),
    wait_for_ready := true
);
```

ローカルルーティングでは、 MAJORITY FOR EACH GROUP IN CLUSTER は各リージョン内のマジョリティの確認を必要とするため、コミットされたと見なされる前に 3 つのリージョンすべてでトランザクションを耐久性のあるものにします。このトポロジは、リージョンごとに 1 つずつ、最大 3 つのノードの障害を許容します。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := top_group,
    rule := MAJORITY FOR EACH GROUP IN CLUSTER QUORUM COMMIT ABORT ON (timeout = 6s),
    wait_for_ready := true
);
```

!!!注クラスタートポロジが許容するように設計された障害シナリオのみをテストします。たとえば、3 リージョンクラスター内の 2 つのリージョンの障害、または単一ロケーションクラスター内の複数のノードの障害をテストしないでください。PGD は、機能が低下したクラスターについては保証しません。

### コミット決定に Raft を使用する

上記の構成に commit_decision = raft を追加すると、オリジンノードに直接ではなく、PGD のビルトイン Raft コンセンサスを介してコミット決定がルーティングされます。必要な確認を受け取った後、Raft がすべてのノードにわたって決定を確認した後にのみトランザクションはコミットします。commit_decision の選択は、ノードに障害が発生した後の 自動調整 の動作にも影響を与えます。group では、コーディネーターはノードを照会し、確認をカウントして、結果を判断する必要があります。raft では、Raft が唯一のアービターとして機能し、コーディネーターはすべての未解決のトランザクションをすぐに解決できます。これは、以前の Raft の決定が優先されるためです。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := top_group,
    rule := MAJORITY ORIGIN GROUP QUORUM COMMIT (commit_decision = raft) ABORT ON␣
↪(timeout = 6s),
    wait_for_ready := true
);
```

### タイムアウトの選択

ABORT ON (timeout = <value>) 句はクォーラムコミットには必須です。タイムアウト内にクォーラムに到達しない場合、トランザクションはすべてのノードで中止およびロールバックされ、接続が回復しても再送信されません。

ローカルルーティングを使用する場合、タイムアウトを bdr.raft_group_election_timeout デフォルト 3s より短く設定します。書き込みリーダーが選択タイムアウトを超えて応答しない場合、新しい書き込みリーダーが選択されます。そのポイントの前にトランザクションをタイムアウトすると、飛行中のトランザクションと着信書き込みリーダー間の競合が回避されます。

グローバルルーティングを使用する場合、タイムアウトを bdr.raft_global_election_timeout デフォルト 6s より短く設定します。グローバル選択タイムアウトの間ノードが応答されない場合、トランザクションの調整が開始されます。タイムアウトを 6 秒以下に設定すると、書き込みリーダーリコンシリエーションが開始される前にトランザクションが完全に中止されます。

### デフォルトのコミットスコープの設定

bdr.alter_node_group_option を使用して、コミットスコープをノードグループのデフォルトとして設定できます。そのグループ内のノード上のすべてのトランザクションはコミットスコープを自動的に使用し、すべてのワークロードにわたって一貫した保護を保証します。

```
SELECT bdr.alter_node_group_option(sg1, default_commit_scope, majority_qc);
```

!!!重要クォーラムコミットは、それを使用するトランザクションが使用しないトランザクションと同じテーブルで同時に実行される場合、一貫性を保証しません。ノードグループにデフォルトのコミットスコープを設定すると、すべてのトランザクションが一貫性の利点を得ることができます。

#### クォーラムコミット用のアプリケーションの作成

クォーラムコミットには、すべてのサブスクリプションで有効になっているファーストトラックライタが必要です。トランザクションでクォーラムコミットスコープを使用する前に、トップレベルグループで:ref:``enable_fast_track_writer` ノート <*enable_fast_track_writer* ノート>` オプションを有効にして、クラスター全体で有効にします。

```
SELECT bdr.alter_node_group_option(topgroup, enable_fast_track_writer, true);
```

クォーラムコミットは熱心な競合解決を使用するため、アプリケーションは非同期レプリケーションの場合よりも頻繁にシリアル化エラーが発生します。トランザクションを retry ループでラップしてシリアル化エラーを処理します。

```
BEGIN;
SET LOCAL bdr.commit_scope = banking_scope;
- - ... application queries ...
COMMIT;
```

COMMIT がシリアル化エラーを返した場合、BEGIN からトランザクションを再実行します。再試行率はワークロードによって異なります。多くの行を変更するか、同時トランザクションと重なり合う大規模なトランザクションは、より頻繁に再試行されます。

グローバルルーティングを使用する場合は bdr.raft_global_election_timeout デフォルト 6 秒未満、ローカルルーティングを使用する場合は bdr.raft_group_election_timeout デフォルト 3 秒未満にタイムアウトを設定します。詳細については、**タイムアウトの選択** を参照してください。

#### 互換性

クォーラムコミットは *Parallel Apply* と互換性があります。 Quorum Commit を使用する場合、Parallel apply を無効にする必要はありません。

クォーラムコミットは、 file および off モードでのみ **トランザクションストリーミングでの使用** と互換性があります。

#### 制限事項

*Known issues and limitations* の Quorum Commit セクションを参照してください。

### Commit At Most Once

コミットスコープの種類 CAMO

#### 概要

Commit At Most Once CAMO 機能の目的は、アプリケーションが複数回コミットしないようにすることです。

CAMO がないと、 COMMIT が送信された後にクライアントが接続を失うと、アプリケーションはサーバーから応答を受け取らない場合があるため、トランザクションがコミットしたかどうかがわかりません。

アプリケーションは、次の 2 つのオプションの間を簡単に決定できません。

- 場合によっては、データが 2 回入力される可能性があるため、同じデータでトランザクションを再試行します。
- トランザクションを再試行せず、データがまったく処理されないリスク

これらはいずれも、高値のデータに関する重大なエラーです。

この状況を回避する 1 つの方法は、トランザクションの一意のインデックスを持つテーブルに少なくとも 1 つの INSERT を含めることです。ただし、これはアプリケーションのデザインに依存し、アプリケーション固有のエラー処理ロジックが必要であるため、すべての場合に効果的ではありません。

PGD の CAMO 機能は、より一般的なソリューションを提供し、INSERT を必要としません。 bdr.commit_scope によってアクティブ化されると、アプリケーションはトランザクション識別子が既に割り当てられている場合を含むメッセージを受信します。それ以外の場合、トランザクションの最初の write ステートメントはその情報をクライアントに送信します。

アプリケーションが明示的な COMMIT を送信する場合、プロトコルは、COMMIT が送信される前にアプリケーションがトランザクション識別子の通知を受信することを保証します。サーバーが COMMIT に応答しない場合、アプリケーションはトランザクション識別子を使用して別の PGD ノードからトランザクションの最終ステータスを要求することにより、このエラーを処理できます。以前のトランザクションステータスがわかっている場合、アプリケーションはトランザクションを再試行するかどうかを安全に決定できます。

CAMO は、同じ PGD グループからの 2 つの PGD ノードであるパートナーノードのペアを作成することにより機能します。この操作モードでは、ペアの各ノードは、他のピアで実行された最近のトランザクションの結果を知っており、特にニーズに応じて、COMMIT 中に切断されたトランザクションの結果を知っています。アプリケーションからトランザクションを受信するノードは「オリジン」と呼ばれ、これらのトランザクションを確認するノードは「パートナー」と呼ばれます。ただし、CAMO ペアのノードの CAMO 構成に違いはありません。ペアは対称です。

!!!警告

CAMO、高度なエラー処理を利用するには、ユーザーのアプリケーションを変更する必要があります。パラメーターを有効にするだけでは保護を得るには十分ではありません。リファレンスクライアント実装は、要求に応じてお客様に提供されます。

CAMO を使用するには、アプリケーションはマルチステートメント要求の一部としてではなく、別の要求として明示的な COMMIT メッセージを発行する必要があります。CAMO は、プロシージャーまたは暗黙的なコミットを使用する単一ステートメントトランザクションから発行されたトランザクションのステータスを提供できません。

### 構成

構成パラメーターについては、CAMO コミットスコープリファレンスを参照してください。

### 確認

| Confirmation | Level | CAMO handling |
|---|---|---|
| *received* | | Not applicable, only uses the default, *VISIBLE*. |
| *replicated* | | Not applicable, only uses the default, *VISIBLE*. |
| *durable* | | Not applicable, only uses the default, *VISIBLE*. |

次のページに続く

表 7 – 前のページからの続き

| Confirmation Level | CAMO handling |
|---|---|
| *visible* (default) | Confirms the transaction after all of its changes are flushed to disk and it's visible to concurrent transactions. |

#### 制限事項

制限事項 の CAMO セクションを参照してください。

#### 障害シナリオ

さまざまな構成では、さまざまな障害シナリオが発生します。

#### レシーバー側のデータの永続性

デフォルトでは、PGL ライタはリモートノードからトランザクションを適用するときに `bdr.synchronous_commit = off` モードで動作します。これは CAMO にも当てはまります。つまり、トランザクションは、CAMO パートナーのディスクに到達する前に、オリジンノードに確認されます。クラッシュまたはハードウェア障害が発生した場合、確認されたトランザクションは CAMO パートナーで自分自身では回復できない場合があります。これは、CAMO パートナーノードが復旧するとトランザクションを再配布するため、CAMO 起点ノードが動作している限り問題ではありません。

これは、CAMO が単一ノードの障害から保護できることを意味します。これは、ローカルモードおよびリモート書き込みと同様に、またはそれと組み合わせて正しいです。

CAMO ペアの両方のノードの停止をカバーするには、`bdr.synchronous_commit = local` を使用して、コミット前の確認の前にフラッシュを強制できます。これは、リモート書き込みモードでもローカルモードでも動作せず、レイテンシーの影響を受けやすいコミットパスにある CAMO パートナーの I/O 要件により、パフォーマンスに影響を与えます。

#### 非同期モード

コミットスコープで `DEGRADE ON ... TO ASYNC` 句が使用される場合、ノードは CAMO パートナーの準備ができているかどうかを検出します。そうでない場合、一時的に非同期ローカルモードに切り替わります。このモードの場合、ノードは CAMO モードに戻すまで、ローカルにトランザクションをコミットします。

これでは、COMMIT ステータスを取得することはできませんが、一貫性よりも可用性を選択できます。このモードは、単一ノードの障害を許容できます。CAMO ペアの両方のノードに障害が発生した場合、可用性を維持するために一致しないコミット決定を選択する可能性があり、データの不一貫性が発生します。

CAMO パートナーが準備ができているように切り替えるには、接続する必要があり、推定キャッチアップインターバルは `TO ASYNC` の `timeout` 値を下回る必要があります。

bdr.is_camo_partner_ready() を使用して CAMO パートナーの現在の準備状況を確認できます。bdr.node_replication_rates は、キャッチアップ時間の現在の推定値を提供します。

CAMO 保護から非同期モードへの切り替えは、実際の CAMO トランザクションによってのみトリガーされます。これは、コミットが `TO ASYNC` の `timeout` 値を超えているため、または CAMO パートナーが既に知られている場合、コミット時に切断されたためです。このスイッチは、推定キャッチアップインターバルとは無関係です。CAMO ペアが、現在のノードが `enable_routing` ノードグループオプションで構成

されたグループの書き込みリードであることを必要とするように構成されている場合。構文については、*Commit scopes* を参照してください。これにより、分離されたノードが非同期モードに切り替わらないことによるスプリットブレイン状況を防ぐことができます。CAMO グループに enable_routing が設定されていない場合、オリジンノードはすぐに非同期モードに切り替わります。

非同期モードから CAMO モードへの切り替えは、接続を開始する CAMO パートナーノードに依存します。CAMO パートナーは、少なくとも 30 秒ごとに再接続を試行します。したがって、接続が再確立された後、CAMO パートナーが元のノードに接続するまでに最大 30 秒かかる場合があります。CAMO パートナーに蓄積された遅延は、CAMO 保護モードへの切り替えをさらに遅延させます。

通常の CAMO オペレーション中とは異なり、非同期モードでは、コミットオーバーヘッドの追加はありません。これは、CAMO ペアが通常処理できるよりも多くのトランザクションをノードが継続的に処理できるため、問題になる可能性があります。CAMO パートナーが最終的に再接続してトランザクションを適用した場合でも、このような状況では遅延が増加するだけであり、CAMO 保護の再確立を妨げます。トランザクションのスループットを人為的に調整するために、PGD は bdr.camo_local_mode_delay 設定を提供します。これにより、ローカルモードで COMMIT を任意の時間遅延できます。予想されるワークロード中に通常の CAMO モードでコミット時間を測定し、それに応じてこの遅延を構成することをお勧めします。デフォルトは 5 ミリ秒で、これは非同期ネットワークと比較的迅速な CAMO パートナーの応答を反映しています。

アーキテクチャと可用性要件を考慮して、非同期モードを許可するかどうかの選択を検討します。次の例は、詳細を提供します。

#### 例

この例では、互いに CAMO パートナーである 2 つの PGD ノードを使用したセットアップを検討します。

```
- - create a CAMO commit scope for a group over
- - a definite pair of nodes
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := camo_dc,
    rule := ALL (left_dc) CAMO DEGRADE ON (timeout=500ms) TO ASYNC
);
```

この CAMO コミットスコープが正当であるためには、グループ内のノードの数が正確に 2 である必要があります。複数のノードで構成されるグループで ALL または ANY 2 を使用すると、非定量化グループ式が明確なペアに解決されないため、エラーになります。ノード。

#### 非同期モードの場合

非同期モードが許可されている場合、単一障害点はありません。1 つのノードに障害が発生した場合

- 他のノードは、障害が発生したノードで COMMIT 中に切断されたすべてのトランザクションのステータスを判断できます。
- 新しい書き込みトランザクションが許可されます。2 番目のノードにも障害が発生した場合、その時点でコミットされていたトランザクションの結果は不明です。

#### 非同期モードなし

非同期モードが許可されない場合、各ノードはトランザクションをコミットするために他のノードを必要とします。つまり、各ノードは単一障害点です。1 つのノードに障害が発生した場合

- 他のノードは、障害が発生したノードで COMMIT 中に切断されたすべてのトランザクションのステータスを判断できます。
- ノードが復旧するまで、新しい書き込みトランザクションは防止されます。

### アプリケーションの使用

#### 概要と要件

CAMO は、再試行ループとクライアント側での特定のエラー処理に依存しています。それには 3 つの側面があります。

- トランザクションの COMMIT の結果を確認する必要があり、一時的なエラーが発生した場合、クライアントはトランザクションを再試行する必要があります。
- COMMIT の前に、クライアントはノード ID とトランザクション ID 両方 32 ビット整数で構成されるトランザクションのグローバル識別子を取得する必要があります。
- トランザクションの COMMIT を試行中に現在のサーバーに障害が発生した場合、アプリケーションは CAMO パートナーに接続し、そのトランザクションのステータスを取得し、応答に応じて再試行する必要があります。

アプリケーションは、 COMMIT 中に切断された場合にトランザクションステータスを確認する目的でのみグローバルトランザクション識別子を保存する必要があります。特に、アプリケーションは別の永続レイヤーを必要としません。アプリケーションに障害が発生した場合、データベースの情報のみを再起動します。

これを説明するために、この例では、C のような擬似コードで書かれた CAMO 対応クライアントアプリケーションの再試行ループを示します。接続情報を提供する 2 つの DSN origin_dsn および partner_dsn を想定しています。これらは通常、bdr.create_node への最初の呼び出しに使用される DSN と同じであり、 bdr.node_summary 、interface_connstr 列でルックアップできます。

```
PGconn *conn = PQconnectdb(origin_dsn);
```

プロセスは、オリジンノードへの接続を開始します。次に、ループに入ります。

```
loop {
    PQexec(conn, "BEGIN");
```

次に、トランザクションを開始し、変更の入力を開始します。

```
PQexec(conn, "INSERT INTO ...");
...
```

完了したら、ローカルノード ID とトランザクション ID の記録を作成する必要があります。どちらもパラメーターとして使用できます。

```
node_id = PQparameterStatus(conn, "bdr.local_node_id");
xid = PQparameterStatus(conn, "transaction_id");
```

これで、コミットする準備が整いました。

```
PQexec(conn, "COMMIT");
if (PQresultStatus(res) == PGRES_COMMAND_OK)
    return SUCCESS;
```

結果が PGRES_COMMAND_OK であれば、それは問題ありません。次に進んでください。しかし、そうでない場合は、CAMO を使用してトランザクションを完了まで追跡する必要があります。最初に尋ねる質問は「接続が悪かったですか?」

```
else if (PQstatus(res) == CONNECTION_BAD)
{
```

接続が悪い場合は、CAMO パートナーノードをチェックして、トランザクションがそこに到達したかどうかを確認できます。

```
conn = PQconnectdb(partner_dsn);
if (!connectionEstablished())
    panic();
```

パートナーノードに接続できない場合、できることはあまりありません。この場合、パニックになるか、同様の行動をとります。

ただし、接続できる場合は、

bdr.logical_transaction_status() を使用して、トランザクションがどのように発生したかを確認できます。コードは、トランザクションをコミットする直前に、必要な値 node_id および xid トランザクション ID を記録しました。

```
    sql = "SELECT bdr.logical_transaction_status($node_id, $xid)";
    txn_status = PQexec(conn, sql);
    if (txn_status == "committed")
        return SUCCESS;
    else
        continue;   // to retry the transaction on the partner
}
```

トランザクションがコミットされたと報告する場合、このトランザクションは成功したと言えます。これ以上のアクションは必要ありません。一方、コミットされたことを報告しない場合は、ループを続行して、パートナーノードでトランザクションを再試行できます。

```
    else
    {
        if (isPermanentError())
            return FAILURE;
```

(次のページに続く)

(前のページからの続き)

```
        else
        {
            sleep(increasing_retry_delay);

            continue;
        }
    }
}
```

トランザクションのステータスが成功または不正な接続ではなかった場合、問題が永続的なエラーであったかどうかを確認します。その場合、トランザクションの失敗を報告します。そうでない場合は、再試行できます。再試行のたびに増加する期間コードをスリープさせて、トランザクションを再試行します。

### CAMO パートナーとの連携

彼らに割り当てられたロール。

bdr.is_camo_partner_connected() を使用すると、CAMO パートナーノードの接続ステータスを確認できます。

CAMO パートナーの準備ができていることを確認するには、ファンクション bdr.is_camo_partner_ready を使用します。その下で、これはローカルモードへの、またはローカルモードからのスイッチをトリガーします。

設定された CAMO パートナーの詳細を確認するには、

bdr.get_configured_camo_partner() を使用します。このファンクションは、ローカルノードの CAMO パートナーを結果ます。

ファンクション bdr.wait_for_camo_partner_queue() を使用して、CAMO パートナーがキューを処理するのを待機できます。このファンクションは、bdr.wait_for_apply_queue のラッパーです。違いは bdr.wait_for_camo_partner_queue()

デフォルトでは CAMO パートナーノードを照会します。ローカルノードが CAMO ペアの一部でない場合、エラーを返します。

ノードに障害が発生したときにコミットされていたトランザクションのステータスを確認するには、アプリケーションは

bdr.logical_transaction_status() ファンクションを使用する必要があります。

このファンクションに、チェックするトランザクションの node_id と transaction_id を渡します。この機能は、CAMO ペアの一部であるノードでのみ使用できます。

すべての場合、コミット発行後 15 分以内にファンクションを呼び出す必要があります。CAMO パートナーはこのようなメタ情報を定期的にパージする必要があるため、古いトランザクションに正しい回答を提供できません。

トランザクションのステータスを照会する前に、このファンクションは、受信キューが消費され完全に適用されるまで待機します。このメカニズムは、受信されたがまだ適用されていないトランザクションの早期の否定的な回答を防ぎます。

その名前にもかかわらず、それは常に読み取り専用操作であるわけではありません。ステータスが不明な場合、CAMO パートナーはトランザクションをコミットするか中止するかを決定し、その決定をローカルに保存して、今後の一貫性を保証します。

クライアントは、オリジンでコミットする前にこのファンクションを呼び出してはなりません。そうしないと、トランザクションが強制的にロールバックされる可能性があります。

### 接続プールとプロキシ

CAMO クラスターをデザインするときは、接続プールとプロキシの効果を考慮してください。プロキシは、コミットグループ内のすべてのノード、つまり CAMO ペアの両方のノードにトランザクションを自由に分散することができます。

アプリケーションが適切なノード ID を取得するように注意してください。セッションプーリングを使用する場合、クライアントは同じノードに接続されたままであるため、ノード ID はクライアントセッションの有効期間を通じて一定のままです。ただし、より詳細なトランザクションプーリングでは、次の例のように、クライアントはすべてのトランザクションのノード ID を取得する必要があります。

PGD ノードに直接接続されていないクライアントは、フェールオーバーまたはスイッチオーバーに気付かない場合があります。ただし、`bdr.local_node_id` パラメーターを使用して、現在接続しているノードを判断することができます。COMMIT 中の切断という重要な状況では、プロキシはその切断をエラーとして CAMO プロトコルを適用するクライアントに適切に転送する必要があります。

`received` モードの CAMO の場合、CAMO ペア間を切り替える可能性のあるプロキシは、`bdr.wait_for_camo_partner_queue` ファンクションを使用して、古い読み取りを防ぐ必要があります。

### CAMO の制限

CAMO の制限については *Known issues and limitations* でカバーされています。

### パフォーマンスへの影響

CAMO は、コミット時にメッセージのラウンドトリップを追加することにより、Postgres レプリケーションプロトコルを拡張します。アプリケーションのコミットレイテンシーは非同期レプリケーションの場合よりも高く、これは主に、関係するノード間のラウンドトリップ時間によって決まります。同時セッションの数を増やすと、並列処理が増加して、合理的なトランザクションスループットを取得できます。

トランザクションを確認する CAMO パートナーは、トランザクションの状態を保存する必要があります。非 CAMO オペレーションと比較して、これは、オリジンから適用される各トランザクションの追加シークが必要になる場合があります。

### クライアントアプリケーションのテスト

クライアント側で CAMO を適切に使用することは簡単ではありません。ノードのクラッシュやネットワークの停止などの障害シナリオに対して、PGD クラスターを使用してアプリケーションの動作をテストすることを強くお勧めします。

### Lag Control

コミットスコープの種類 `LAG CONTROL`

#### 概要

Lag Control は、レプリケーションが設定された制限の外部で実行されている場合、レプリカブル更新を行うトランザクションを処理した後、元のノードのクライアント接続に遅延が注入されるメカニズムを提供します。この遅延は、着信トランザクションを遅くし、定義された制限内にレプリケーションを戻すように設計されています。

#### 背景

PGD 起点ノード上のデータベースアプリケーションのデータスループットは、コミットされたデータをダウンストリームのピアノードに複製できる速度を超える場合があります。

この不均衡が持続すると、RPO、RCO、GEO などの組織目標の充足が危険にさらされる可能性があります。

- **目標復旧ポイント RPO** は、計画外のイベントによって失われる可能性のあるデータの最大許容量を指定します。通常は時間として表されます。PGD では、RPO は、1 つ以上のピアノードに適用されていないコミットされたデータの許容量を決定します。
- **リソース制約目標 RCO** は、使用可能なストレージが有限であることを確認します。PGD では、遅延が増加すると、これらのストレージリソースの要求が増加します。
- **グループ弾性目標 GEO** は、ノードがピアノードに保存できないレートで新しいデータを生成していないことを保証します。

組織が目的を達成できるように、PGD はラグコントロールを提供します。この機能は、アプリケーションに侵入することなく、潜在的な不均衡を正確に調整する手段を提供します。これは、データを変更する READ WRITE トランザクションに遅延を透過的に導入することにより実現します。この遅延 PGD コミット遅延は 0 ミリ秒で始まります。

LAG CONTROL コミットスコープ種類を使用して、グループ内のノード間でコミットを遅延できる最大時間、最大遅延時間、または最大遅延サイズ WAL のサイズに基づいて、設定できます。

ノードが十分な数のノードで指定された最大値内でトランザクションを処理できる場合、PGD コミット遅延は 0 ミリ秒に留まるか、0 ミリ秒に向かって削減されます。ただし、十分な数のノードで最大値を超過する場合、元のノードでの PGD コミット遅延が増加します。十分な数のノードで Lag Control コンストレインが再び満たされるまで、増加し続けます。

PGD コミット遅延は、トランザクションが完了して、すべてのロックとリソースを解放した後に発生します。この遅延のタイミングにより、同時アクティブなトランザクションは、遅延されたトランザクション値の監視と変更、およびそのリソースの取得を続行できます。

厳密には、PGD コミット遅延はトランザクションごとの遅延ではありません。これは、特定のクライアント接続のトランザクションストリームにおけるコミット遅延の平均値です。この技術により、コミット遅延と値のきめの細かい調整が、OS スケジューラー、クロック割り込み、およびシステム負荷による変動の粗い粒度をエスケープできます。また、PGD ランタイムコミット遅延を可能な限り最小の期間のマイクロ秒以内に安定させて、遅延測定しきい値を維持できます。

Postgres commit delay を混同しないでください。これらは無関係であり、異なる機能を実行します。一方を他方に置き換えないでください。

### 要件

Lag Control の使用を開始するには

- すべてのデータベースアプリケーションが許容できる最大許容コミット遅延時間 max_commit_delay を決定します。
- 使用する遅延測定を決定します。ラグサイズ max_lag_size またはラグタイム max_lag_time のいずれかを選択します。
- 関係するグループまたはサブグループ、および確認を満たすために必要な各コレクションのノードの最小数を決定します。この情報は、コミットスコープルールの定義の基礎を形成します。

### 構成

コミットスコープでラグ制御を指定すると、コミットスコープルールがスパンするノード全体で一貫した調整されたパラメーター設定ができます。ラグコントロール仕様は、トップグループのデフォルトのコミットスコープに、またはオリジングループのコミットスコープの一部として含めることができます。

例のように、サブグループとして表される 2 つのデータセンター left_dc および right_dc を含む構成を取得します。

```
SELECT bdr.create_node_group(
    node_group_name := left_dc,
    parent_group_name := top_group,
    join_node_group := false
);
SELECT bdr.create_node_group(
    node_group_name := right_dc,
    parent_group_name := top_group,
    join_node_group := false
);
```

次のコードは、サブグループごとに個々のルールを使用して、これらの 2 つのデータセンターのラグ制御ルールを追加します。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := left_dc,
    rule := ALL (left_dc) LAG CONTROL (max_commit_delay=500ms, max_lag_time=30s) AND␣
↪ANY 1 (right_dc) LAG CONTROL (max_commit_delay=500ms, max_lag_time=30s),
    wait_for_ready := true
);
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := right_dc,
    rule := ANY 1 (left_dc) LAG CONTROL (max_commit_delay=0.250ms, max_lag_
↪size=100MB) AND ALL (right_dc) LAG CONTROL (max_commit_delay=0.250ms, max_lag_
↪size=100MB),
```

(次のページに続く)

(前のページからの続き)

```
    wait_for_ready := true
);
```

Lag Control コミットスコープルールを、グループコミットおよび CAMO ルール仕様を含む既存のコミットスコープルールに追加できます。

`max_commit_delay` は間隔であり、通常はミリ秒 1ms 単位で指定されます。ミリ秒未満の精度での小数値の使用がサポートされています。

`max_lag_size` は、WAL バイト単位で最大許容遅延を指定する整数です。

`max_lag_time` は、通常秒単位で指定される間隔で、時間に関して最大許容遅延を指定します。

最大コミット遅延 `max_commit_delay` は、ハード制限を表す天井値であり、コミット遅延が構成値を超えないことを意味します。

最大ラグサイズと時間 `max_lag_size` および `max_lag_time` は、超過できるソフトリミットです。最大コミット遅延に到達すると、遅延対策に追加のバックプレッシャーはなく、継続的な増加を防ぎます。

#### 確認

| Confirmation level | Lag Control handling |
|---|---|
| *received* | Not applicable, only uses the default, *VISIBLE*. |
| *replicated* | Not applicable, only uses the default, *VISIBLE*. |
| *durable* | Not applicable, only uses the default, *VISIBLE*. |
| *visible* (default) | Not applicable, only uses the default, *VISIBLE*. |

#### トランザクションアプリケーション

PGD コミット遅延は、ユーザーアプリケーションのデータを変更するすべての READ WRITE トランザクションに適用されます。この動作は、宣言された READ WRITE トランザクションを含む、データを変更しないトランザクションがコミット遅延から免除されることを意味します。

非同期トランザクションのコミットは、PGD コミット遅延も実行します。これは直観に反しているように見えるかもしれませんが、非同期コミットは、そのパフォーマンスのおかげで、レプリケーション遅延の最大のソースの 1 つになる可能性があります。

Postgres および PGD 補助プロセスは、トランザクションのコミット時に遅延しません。最も注目的なのは、PGD ライターは、ローカルノードでリモートトランザクションを適用するときにコミット遅延を実行しないことです。PGD ライターは発信レプリケーションラグに何も寄与せず、遅延によってトランザクションコミットが調整されないことにより、着信レプリケーションラグを最も削減できるため、これはデザインです。

### 制限事項

最大コミット遅延は、ハード制限を表す天井値であり、コミット遅延が構成された値を超えないことを意味します。逆に、最大遅延はサイズと時間の両方で測定され、超過できるソフト制限です。最大コミット遅延に到達すると、遅延対策に追加のバックプレッシャーはなく、継続的な増加を防ぎます。

PGD レプリケーションセットを変更しないオリジントランザクションをコミット遅延から免除する方法はありません。これらのトランザクションの場合、最大トランザクション遅延を 0 に SET LOCAL にすると役立ちます。

### 注意事項

アプリケーション TPS は、レプリケーション遅延に影響を与える可能性のある多くの要因の 1 つです。他の要素には、PGD コミット遅延の効果があまりない可能性があるトランザクションの平均サイズが含まれます。特に、大量のロード操作はレプリケーション遅延の増加を引き起こす可能性があり、これは、最大許容遅延を下回っていますが、通常のアプリケーションで合理的に予想されるレベルを超えて PGD ランタイムコミット遅延の付随的な上昇をトリガーする可能性があります。

同様に、非常に高い OLTP 要件と適度なデータ変更があるアプリケーションは、許容可能な PGD コミット遅延設定によって不当に抑制される可能性があります。

これらの場合、 `SET [SESSION|LOCAL]` コマンドを使用して、これらのアプリケーションのラグ制御設定をカスタム構成するか、これらのアプリケーションを変更すると便利です。たとえば、大量のロード操作は複数の小さなトランザクションに分割されて、トランザクションのスナップショット期間と WAL 保持サイズを制限したり、大量のロードが失敗した場合にリスタートポイントを確立したりする場合があります。ラグ制御を参照して、これらのトランザクションコミットは、非常に長い PGD コミット遅延をスケジュールして、以前の部分的な大量ロードによって発生した遅延をダイジェストすることもできます。

### 組織目標の達成

前にリストした目標の例では

- 適切な最大ラグタイムを設定することにより、RPO を満たすことができます。
- 適切な最大遅延サイズを設定することにより、RCO を満たすことができます。
- GEO は、PGD ランタイムコミット遅延と PGD ランタイムラグ対策をモニタリングすることにより満たすことができます。

前述のように、PGD ランタイムコミット遅延の最大値が PGD 構成のコミット遅延制限でペッグされ、遅延測定値が一貫して PGD 構成の最大レベルを超える場合、このシナリオは PGD グループ拡張のマーカーになる可能性があります。

## Lag Control と拡張機能

PGD コミット遅延は、コミット後の遅延です。これは、トランザクションがコミットされた後、トランザクションによってロックまたは取得されたすべての Postgres リソースが解放された後に発生します。したがって、遅延は、同時アクティブなトランザクションがその値を監視または変更する、またはそのリソースを取得することを妨げるものではありません。Postgres 拡張機能によって管理される外部リソースについては、同じ保証はできません。拡張機能の依存関係に関係なく、postgresql.conf で拡張機能ベースのリソースマネージャーの前に PGD 拡張機能がリストされている場合、同じ保証を行うことができます。

### Group Commit (legacy)

コミットスコープの種類 GROUP COMMIT

#### 概要

グループコミットの目標は、単一ノードの障害または一時的な停止が発生した場合のデータの損失から保護することです。これは、COMMIT 時にトランザクションを正常に確認するために複数の PGD ノードを必要とすることにより実現します。確認はトランザクション処理のさまざまなポイントで送信できますが、トランザクションがディスクにフラッシュされ、他のすべてのトランザクションから見える場合、デフォルトの「可視」になります。

#### 例

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := left_dc,
    rule := ALL (left_dc) GROUP COMMIT(commit_decision=raft) AND ANY 1 (right_dc)␣
↪GROUP COMMIT,
    wait_for_ready := true
);
```

この例では、 left_dc グループのすべてのノードと right_dc グループのいずれかのノードがコミットされたトランザクションを受信して正常に確認する必要があるコミットスコープを作成します。

#### 要件

通常のオペレーションでは、グループコミットはアプリケーションに対して透過的です。フェールオーバー中に進行していたトランザクションには、アプリケーションまたは間のプロキシによってトリガーまたは統合される調整フェーズが必要です。このアクティビティは現在、元のノードが復旧した場合、またはクラスターから分離された場合にのみ発生します。この動作は、Postgres の古いビルトイン同期レプリケーションの場合と同じです。

グループコミットでコミットされたトランザクションは、その下で *two-phasecommit* を使用します。したがって、ノードごとに発生するこのようなすべてのトランザクションを処理できるように十分に高い max_prepared_transactions を構成します。

#### 制限事項

*Known issues and limitations* の Group Commit セクションを参照してください。

#### 構成

GROUP_COMMIT は、オプショナルの GROUP COMMIT パラメーター、ABORT ON および DEGRADE ON 句をサポートしています。構成パラメーターの完全な説明については、

GROUP_COMMIT コミットスコープリファレンスを参照してください。または、DEGRADE ON オプション一般の詳細については、*Degrading commit scope rules* セクションを参照してください。

#### 確認

| Confirmation level | Group Commit handling |
|---|---|
| *received* | A remote PGD node confirms the transaction immediately after receiving it, prior to starting the local application. |
| *replicated* | Confirms after applying changes of the transaction but before flushing them to disk. |
| *durable* | Confirms the transaction after all of its changes are flushed to disk. |
| *visible* (default) | Confirms the transaction after all of its changes are flushed to disk and it's visible to concurrent transactions. |

#### 動作

Group Commit の動作は、コミットスコープによって適用される構成によって異なります。

#### 決定をコミットする

`group` 、`partner` 、および `raft` の 3 つの方法でコミットを決定するようにグループコミットを構成できます。

`group` 決定がデフォルトです。コミットスコープグループで必要な数の確認を受信したときに、元のノードによってコミットが確認されることを指定します。違いは、コミットの決定が PREPARE レプリケーションに基づいて行われるのに対し、耐久性は COMMIT (PREPARED) レプリケーションをチェックすることです。

`partner` 決定は、 *Commit At Most Once* CAMO が使用するものです。このアプローチは、ノードグループに 2 つのデータノードが存在する場合にのみ機能します。これらの 2 つのノードはお互いのパートナーであり、オリジンではなくレプリカが何かをコミットするかどうかを決定します。このアプローチでは、アプリケーションが何らかの形でコンセンサスの一部であるため、CAMO トランザクションプロトコルを使用して正常に動作するには、アプリケーションの変更が必要です。このアプローチの詳細については、*Commit At Most Once* を参照してください。

`raft` 決定は、コミット決定に PGD ビルトインの Raft コンセンサスを使用します。`raft` 決定を使用すると、パフォーマンスが低下する可能性があります。現在、ALL コミットスコープグループで `GROUP COMMIT` を使用する場合にのみ必要です。

ALL コミットスコープグループを使用するには、コミット決定を `raft` に設定して、 **トランザクション調整** 問題を回避する必要があります。

#### 競合の解決

競合の解決は `async` または `eager` です。

非同期とは、PGD が、特定のノードに構成されている行レベルの解決を使用して、レプリケーション中にオプティミスティック競合解決を行うことを意味します。これは、元のトランザクションがコミットされたか、まだ進行中であるかに関係なく発生します。非同期競合解決の仕組みの詳細については、*bdr_read_all_conflicts* を参照してください。

Eager とは、競合が COMMIT 契約の一部として熱心に解決され、競合するトランザクションがシリアル化エラーで中止されることを意味します。このアプローチは、パフォーマンスを犠牲にして非同期解決より

も高い分離を提供します。

ALL コミットスコープグループを使用するには、リコンシリエーションの問題を回避するために、**決定をコミットする** を `raft` に設定する必要があります。

Eager 競合解決の仕組みの詳細については、**熱心な競合解決** を参照してください。

### アボート

COMMIT でコンセンサスを取得できないトランザクションが永久にハングしないように、`ABORT ON` 句でタイムアウトを指定できます。タイムアウトの後、トランザクションのアボートが要求されます。アボート要求が送信されたときにトランザクションがコミットされることが既に決定されている場合、クライアントがアボートメッセージを受信する場合でも、トランザクションは最終的に COMMIT します。

**制限事項** も参照してください。

### トランザクション調整

元ノードでのグループコミットトランザクションのコミットは、暗黙的に 2 フェーズコミットに変換されます。

最初のフェーズ準備では、トランザクションがローカルで準備され、コミットの準備が整います。データは永続化されますが、この段階ではコミットされていないため、他のトランザクションはこのトランザクションによって行われた変更を認識できません。この準備されたトランザクションは、通常の論理レプリケーションを介して残りのすべてのノードにコピーされます。

オリジンノードは、グループコミット文法のルールに従って、他のノードからの確認を求めます。クラスター内の最小限必要なノードから確認を取得した場合、このトランザクションをコミットすることを決定し、第 2 フェーズコミットに進みます。コミットフェーズでは、この決定をレプリケーションを介して他のノードに送信します。これらのノードは、最終的にこのメッセージの取得にコミットします。

さまざまな段階で失敗する可能性があります。たとえば、オリジンノードは、トランザクションの準備後にクラッシュする場合があります。または、オリジンと 1 つ以上のレプリカがクラッシュする場合があります。

これにより、準備されたトランザクションがシステムに残ります。Postgres の `pg_prepared_xacts` ビューは、システムで準備されたトランザクションを表示できます。準備されたトランザクションは、ロックおよびその他のリソースを保持している可能性があります。これらのロックとリソースを解放するには、トランザクションを中止またはコミットします。その決定は、ノードのコンセンサスで行う必要があります。

`commit_decision` が `raft` の場合、Raft はコーディネーターとして機能し、これらのトランザクションは最終的に自動的にリコンサイルされます。

`commit_decision` が `group` の場合、トランザクションは Raft を使用しません。代わりに、クラスター内の書き込みリードがコーディネーターの役割を実行します。これは、サブグループ内の変更に関して最も進んでいるノードであるためです。ノードがダウンしていることを検出し、ダウンしたノードをオリジンとして持っている準備されたトランザクションを探すことにより、このようなノードの調整を開始します。

このようなすべてのトランザクションについて、コミットスコープのルールに従ってノードに準備されたトランザクションがあるかどうかを確認し、決定を行います。この決定は Raft を介して伝えられ、ノードの大部分が調整を行うために立ち上がっている必要があります。

このプロセスはバックグラウンドで発生します。これを制御または発行するために使用するコマンドはありません。

### 熱心な競合解決

Eager 競合解決 Eager Replication としても知られ、COMMIT 決定プロセス中にシリアル化可能なエラーで互いに競合するトランザクションをアボートすることにより、競合を防止します。

*DDL overview* の競合解決オプションの 1 つとして *Commit scopes* を使用して構成します。

### 使用法

迅速な競合解決を有効にするには、クライアントをコミットスコープに切り替える必要があります。ここで示すように、セッションレベルまたは個々のトランザクションでそれを使用します。

```
BEGIN;

SET LOCAL bdr.commit_scope = eager_scope;

... other commands possible...
```

クライアントは、トランザクションの最後に COMMIT の発行を継続し、PGD に 2 つのフェーズを管理させることができます。

```
COMMIT;
```

この場合、 eager_scope コミットスコープは次のように定義されます。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := eager_scope,
    origin_node_group := top_group,
    rule := ALL (top_group) GROUP COMMIT (conflict_resolution = eager, commit_
↪decision = raft) ABORT ON (timeout = 60s),
    wait_for_ready := true
);
```

Eager 競合解決ルールのコミットスコープグループは、ALL または MAJORITY のみです。 ALL を使用する場合、commit_decision 設定も raft に設定する必要があります。

### エラー処理

PGD がトランザクションを管理することを考えると、クライアントは COMMIT の結果のみを確認する必要があります。これは、シングルノードの Postgres を含む、いかなる場合にもお勧めします。

オリジンノードに障害が発生した場合、残りのノードは最終的に少なくとも ABORT ON timeout の後、グローバルに準備されたトランザクションをロールバックすることを決定します。Raft は、一貫性のないコミットとロールバックの決定を防ぎます。ただし、これには接続されたノードの大部分が必要です。切断されたノードは、必要に応じて最終的にコミットまたはロールバックするようにトランザクションを準備し続け、決定してさらに進行した可能性のあるノードの大部分と調整します。

### Eager Replication の効果一般

#### アボート率の増加

シングルノード Postgres、またはデフォルトの非同期レプリケーションモードの PGD でも、COMMIT 時のエラーはほとんどありません。競合解決に eager を使用するコミットスコープの使用により追加された同期ステップも、エラーのソースも追加します。アプリケーションは、通常は再試行ループを適用することにより、このようなエラーを適切に処理するように準備する必要があります。

アボートの率は、ワークロードにのみ依存します。多くの行を変更する大規模なトランザクションは、他の同時トランザクションと競合する可能性が非常に高くなります。

### MAJORITY および ALL ノードレプリケーション一般の影響

#### コミットレイテンシーの増加

コミットスコープの使用により同期ステップを追加すると、ノード間の通信が増加し、コミット時のレイテンシーが増加することを意味します。ALL がコミットスコープで使用される場合、これは、ノードがダウンするとトランザクションが失敗するため、システムの可用性が低下することを意味します。

1 つ以上のノードが遅延している場合、確認を取得する際の往復遅延が大きくなり、高遅延が発生する可能性があります。ALL または MAJORITY ノードレプリケーションでは、約 2 つのネットワークラウンドトリップ、最悪の場合、最も遠いピアノードまでが追加されます。論理スタンバイノードと、まだ参加またはキャッチアップ中のノードは含まれませんが、最終的に変更を受け取ります。

ピアノードがトランザクションのローカルでの準備を確認する前に、それをローカルに適用する必要があります。これにより、トランザクションのサイズによっては、コミットレイテンシーがさらに増加します。この設定は synchronous_commit 設定とは無関係です。

## Administering

グループコミットを使用して PGD クラスターを実行する場合、ノードを安全にシャットダウンおよび再起動する方法など、システムを管理するときにいくつかのことに注意する必要があります。

### 計画的なシャットダウンと再起動

受信確認でグループコミットを使用する場合、計画的なシャットダウンまたは再起動に注意してください。デフォルトでは、適用キューはシャットダウンする前に処理されます。ただし、immediate シャットダウンモードでは、シャットダウン時にキューは破棄されるため、停止したノードはキュー内のトランザクションを「忘れ」ます。元のノードに同時に障害が発生すると、両方のノードに障害が発生したかのように、データの損失が発生する可能性があります。

適用キューがディスクにフラッシュされるようにするには、メンテナンスタスクに smart のいずれかを使用します。このアプローチでは、必要な同期レベルを維持し、データの損失を防ぎます。

## Legacy synchronous replication using PGD

Postgres は physical streamingreplication を提供します

PSR、これは一方向ですが synchronousvariant を提供します。

下位互換性を維持するために、PGD は synchronous_commit および synchronous_standby_names を使用した同期レプリケーションの構成を引き続きサポートしています。クォーラムコミット または 同期コミット の使用を検討します

代わりに。

PGD レプリケーションオプションとは異なり、PSR 同期は最初に持続し、コミットレコードの WAL フラッシュの後に複製します。

### 使用法

PGD を使用した同期レプリケーションを有効にするには、関連する PGD ピアノードのアプリケーション名を synchronous_standby_names に追加する必要があります。 FIRST x または ANY x を使用すると、非 PGD スタンバイノードの要件と競合しない場合、ある程度の柔軟性が提供されます。

追加したら、 synchronous_commit を使用してトランザクションごとの同期レベルを構成できます。デフォルトは on です。この設定は、アプリケーション名を synchronous_standby_names に追加することにより、同期レプリケーションが既に有効になっていることを意味します。synchronous_commit を local または off に設定すると、同期レプリケーションがオフになります。

PGD は永続化する前にトランザクションを適用するため、値 on および remote_apply は論理レプリケーションでは等価です。

### 比較

次の表は、トランザクションが発行されたオリジンノードから COMMIT 確認を受信した後、クライアントがレプリケートされたピアノードから期待できることをまとめています。Mode 列は、バリアントに応じて異なる意味を持ちます。PSR および PGD を使用した古い同期レプリケーションの場合、synchronous_commit 設定を参照します。

| Variant | Mode | Received | Visible | Durable |
|---|---|---|---|---|
| PSR Async | off (default) | no | no | no |
| PSR Sync | remote_write (2) | yes | no | no (3) |
| PSR Sync | on (2) | yes | no | yes |
| PSR Sync | remote_apply (2) | yes | yes | yes |
| PGD Legacy Sync (1) | remote_write (2) | yes | no | no |
| PGD Legacy Sync (1) | on (2) | yes | yes | yes |
| PGD Legacy Sync (1) | remote_apply (2) | yes | yes | yes |

*(1)* **代わりにクォーラムコミットの使用を検討します。**

- (2) synchronous_replication_availability を async' に設定してローカルモードに切り替えない限り、許可されている場合。それ以外の場合、非同期 PGD デフォルトの値が適用されます。

*(3) OS* **に書き込まれ、** *OS* **が実行されたままで** *Postgres* **のみがクラッシュした場合に持続します。**

### Postgres 構成パラメーター

次の表は、デフォルト以外の値 req に設定する必要がある構成設定とオプショナルですが特定のバリアントに影響を与える構成設定の概要を提供します。

| Setting (GUC) | Group Commit | Lag Control | PSR | Legacy Sync |
|---|---|---|---|---|
| synchronous_standby_names | n/a | n/a | req | req |
| synchronous_commit | n/a | n/a | opt | opt |
| synchronous_replication_availability | n/a | n/a | opt | opt |

### コミットスコープへの移行

`synchronous_commit` および `synchronous_standby_names` とは無関係に、PGD のグループコミット機能を構成します。代わりに、`bdr.commit_scope` GUC を使用すると、トランザクションごとにスコープを選択できます。また、各ノードで `synchronous_standby_names` を個別に構成するのではなく、グループコミットはグローバルに同期したコミットスコープを使用します。

> **注釈**
>
> *synchronous_standby_names* およびコミットスコープの文法は似ていますが、前者はオリジンノードを考慮していませんが、後者は考慮します。したがって、たとえば *synchronous_standby_names = 'ANY 1 (..)'* は、*ANY 2 (...)* のコミットスコープと同等です。この選択により、マジョリティに関する推論が簡単になり、起点ノードがトランザクションの耐久性と可視性にも貢献することを反映しています。

### Internal timing of operations

さまざまなモードがどのように動作するかをより深く理解するために、従来の物理的ストリーミングレプリケーション PSR と PGD はさまざまな方法でトランザクションを適用することを知っておくと役立ちます。これらの順序の違いは、パフォーマンス、耐久性、および一貫性に実際的な意味を持ちます。PSR より PGD を選択する場合のより広範な比較については、

Choosing between PSR and PGD を参照してください。

レガシー PSR では、操作の順序は次のとおりです。

1. オリジンはコミットレコードを WAL にフラッシュし、トランザクションをローカルに認識できるようにします。
2. ピアノードは変更を受信し、書き込みを発行します。
3. ピアは、受信した変更をディスクにフラッシュします。
4. ピアは変更を適用し、ピアでトランザクションを認識できるようにします。

変更は、変更を適用する前にディスクに書き込まれることに注意してください。

PGD では、デフォルトでと Lag Control を使用すると、操作の順序が異なります。これらの場合、トランザクションがピアのディスクにフラッシュされる前に、変更はピアで可視されるようになります。

1. オリジンはコミットレコードを WAL にフラッシュし、トランザクションをローカルに認識できるようにします。
2. ピアノードは、メモリ内の適用キューに変更を受信します。
3. ピアは変更を適用し、ピアでトランザクションを認識できるようにします。
4. ピアはディスクにフラッシュすることによりトランザクションを永続化します。

PGD のクォーラムコミット、グループコミット、および CAMO の場合、オリジンノードはトランザクションをローカルに表示する前に特定の数の確認を待機します。操作の順序は次のとおりです。

1. Origin は、準備または事前コミットレコードを WAL にフラッシュします。
2. ピアノードは、メモリ内の適用キューに変更を受信します。
3. ピアは変更を適用し、ピアでトランザクションを認識できるようにします。
4. ピアはディスクにフラッシュすることによりトランザクションを永続化します。
5. オリジンはコミットして、トランザクションをローカルに可視化します。

次の表は、違いをまとめたものです。

| Variant | Order of apply vs persist | Replication before or after commit |
|---|---|---|
| PSR | persist first | after WAL flush of commit record |
| PGD Async | apply first | after WAL flush of commit record |
| PGD Lag Control | apply first | after WAL flush of commit record |
| PGD Quorum Commit | apply first | before COMMIT on origin |
| PGD Group Commit | apply first | before COMMIT on origin |
| PGD CAMO | apply first | before COMMIT on origin |

## 6.8 Conflict Management

EDB Postgres Distributed は、アクティブ/アクティブまたはマルチマスター DBMS です。非同期で使用される場合、標準のデータ型を使用する場合、複数の異なるノードから同じまたは関連する行への書き込みにより、データが *bdr_read_all_conflicts* になる場合があります。

競合はエラーではありません。ほとんどの場合、これらは発生時に PGD が検出および解決できるイベントです。解決策はアプリケーションの性質とデータの意味によって異なるため、PGD がアプリケーションにそれらを解決する方法に関して幅広い選択肢を提供することが重要です。

デフォルトでは、競合は行レベルで解決されます。2 つのノードからの変更が競合する場合、ローカルまたはリモートのタプルのいずれかが選択され、他方は破棄されます。たとえば、コミットタイムスタンプは、2 つの競合する変更と新しい方が保持されると比較される場合があります。このアプローチでは、すべてのノードが同じ結果に収束し、クラスター全体でコミットオーダーのようなセマンティクスを確立します。

列レベルの競合の検出と解決は、*CLCD* で説明している PGD で利用できます。

競合を回避したい場合は、*Group Commit (legacy)* を **熱心な競合解決** または競合のないデータ型 CRDT とともに使用できます。*CRDT* **データ型を使用した列レベルの競合の処理** で説明しています。接続マネー

ジャーを使用して、すべての書き込みを 1 つの書き込みリーダーにルーティングし、ノード間の競合の可能性を排除することもできます。

### 6.8.1 Conflicts

EDB Postgres Distributed は、アクティブ/アクティブまたはマルチマスター DBMS です。非同期で使用される場合、標準のデータ型を使用するときに、複数の異なるノードから同じまたは関連する行への書き込みにより、データの競合が発生する可能性があります。

競合はエラーではありません。ほとんどの場合、これらは発生時に PGD が検出および解決できるイベントです。このセクションでは、その検出と解決を管理できる PGD 機能を紹介します。

- *DDL overview* は、PGD での競合のアイデアを紹介し、それらがどのように発生する可能性があるかを説明します。
- *Types of Conflict* は、PGD で発生する可能性のあるさまざまな種類の競合をリストおよび議論します。
- *Column-level conflict detection* は、PGD が競合検出のために提供するメカニズムを導入します。
- **欠落列の競合の解決** は、PGD が競合を解決する方法と、デフォルトの動作を変更する方法を説明します。
- *Conflict logging* は、PGD が競合ログを保持する場所を指摘し、競合レポートを実行する方法を説明します。
- *Data verification with LiveCompare* は、LiveCompare が発生した競合を指摘することにより、データの一貫性を保つ方法を説明します。

#### Overview

EDB Postgres Distributed は、アクティブ/アクティブまたはマルチマスター DBMS です。非同期で使用される場合、標準のデータ型を使用するときに、複数の異なるノードから同じまたは関連する行への書き込みにより、データの競合が発生する可能性があります。

競合はエラーではありません。ほとんどの場合、これらは発生時に PGD が検出および解決できるイベントです。それらの解決は、アプリケーションの性質とデータの意味によって異なります。したがって、PGD はアプリケーションに競合を解決する方法についての幅広い選択肢を提供することが重要です。

デフォルトでは、競合は行レベルで解決されます。2 つのノードからの変更が競合する場合、PGD はローカルまたはリモートタプルのいずれかを選択し、他方は破棄します。たとえば、コミットタイムスタンプは、2 つの競合する変更と新しい方が保持されると比較される場合があります。このアプローチでは、すべてのノードが同じ結果に収束し、クラスター全体でコミットオーダーのようなセマンティクスを確立します。

**欠落列の競合の解決** で説明しているように、競合処理は構成可能です。 PGD は、*Stream triggers* で説明しているように、競合トリガーを使用してテーブルごとに競合を検出し、それらをさまざまな方法で処理できます。

*CLCD* で説明しているように、列レベルの競合の検出と解決は、PGD で利用できます。

デフォルトでは、すべての競合は bdr.conflict_history にログに記録されます。競合の可能性がある場合、テーブル所有者は競合を監視して回避方法を分析するか、アプリケーションタスクとして定期的に処理する計画を立てる必要があります。LiveCompare ツールは、発散がないか定期的にスキャンすることもできます。

一部のクラスタリングシステムは、分散ロックメカニズムを使用して、データへの同時アクセスを防止します。これらはサーバーが互いに非常に近い場合に合理的なパフォーマンスを発揮しますが、許容可能なパフォーマンスには非常に低い遅延が重要である地理的に分散したアプリケーションをサポートできません。

分散ロックは本質的に悲観的なアプローチです。PGD は、可能な限り競合を回避するが、一部の種類の競合の発生を許可し、発生した場合にそれらを解決する楽観的なアプローチを提唱しています。

### 競合が発生する仕組み

ノード間の競合は、関係するすべてのトランザクションが同じノードで同時に発生した場合に発生することはできない一連のイベントの結果として発生します。ノードはトランザクションがコミットした後にのみ変更を交換するため、各トランザクションはコミットしたノードで個別に有効です。他の競合する作業を同時に実行した別のノードに適用された場合、有効ではありません。

PGD レプリケーションは基本的に他のノードでトランザクションを再生するため、適用されるトランザクションと受信ノードでコミットされたトランザクションとの間に競合がある場合、再生操作は失敗する可能性があります。

Postgres にはそれを防ぐトランザクション間通信メカニズムがあるため、すべてのトランザクションが単一のノードで実行されている場合、ほとんどの競合は発生しません。これらのメカニズムの例は、`UNIQUE` インデックス、`SEQUENCE` 操作、行およびリレーションロック、`SERIALIZABLE` 依存関係追跡などです。これらのメカニズムはすべて、進行中のトランザクション間で通信して、望ましくない同時実行の問題を防ぐ方法です。

PGD には、分散トランザクションマネージャーもロックマネージャーもありません。これが、レイテンシーとネットワークパーティションで良好なパフォーマンスを発揮する理由の一部です。その結果、異なるノード上のトランザクションは、デフォルトの遅延レプリケーションを使用する場合、互いに完全に独立して実行されます。ノード間の独立性が低いほど、競合を完全に回避できます。これが、これが重要な場合に PGD が Eager Replication も提供する理由です。

### 競合の回避または許容

ほとんどの場合、競合を回避または許容するようにアプリケーションをデザインできます。

競合は、マルチプルのノードで同時に発生している場合にのみ発生します。競合を回避する最も簡単な方法は、1 つのノードにのみ書き込むか、一度に 1 つの特定のノードから特定の方法で特定の行のみに書き込むことです。

この回避は、多くのアプリケーションで自然に発生します。たとえば、多くのコンシューマアプリケーションでは、所有ユーザーのみがアカウントのデフォルトの請求先住所の変更など、データを変更できます。このようなデータの変更では、更新の競合が発生することはほとんどありません。

ノードがダウンする直前に変更を加える可能性があるため、変更は失われるようです。次に、同じ変更を再度行うと、別のノードで 2 つの更新が発生する場合があります。ダウンしたノードが復旧すると、古い変更を他のノードに送信しようとします。データの最終更新が保持されるため、リジェクトされます。

`INSERT` / `INSERT` 競合の場合、*global sequences* を使用してこのタイプの競合を防止します。

部屋予約アプリケーションなどのオブジェクト間の関係を割り当てるアプリケーションの場合、`update_if_newer` を適用すると、許容できるビジネス結果が得られない場合があります。つまり、同じ部屋を予約したことを 2 人に個別に確認することは役に立ちません。最も簡単な解決策は、Eager Replication を使用して、1 つの予約のみが成功するようにすることです。アプリケーションによっては、より複雑な

方法が可能である場合があります。たとえば、各ノードに 100 席を割り当て、そのノードのライタがそれらを予約できるようにします。ただし、ローカルで利用できるものがない場合は、ほとんどのシートが予約された後に分散ロックスキームまたは Eager Replication を使用します。

特定の種類の更新が特定のノードからのみ発生するようにする別の手法は、異なるノードを介して異なる種類のトランザクションをルーティングすることです。例

- 1 つのノードで小包を受信するが、別のノードを使用して小包を配送する
- 1 つのノードで注文が入力され、2 番目のノードで作業が準備され、別のノードで顧客に提供されるサービスアプリケーション

多くの場合、最良のコースは、競合の発生を許可し、競合に対処するために PGD の競合解決メカニズムと連携するようにアプリケーションをデザインすることです。

### Types of Conflict

#### 主キーまたは一意の競合

最も一般的な競合は行の競合であり、2 つの操作が単一のノードではできない方法で同じキーを持つ行に影響を与えます。PGD はこれらのほとんどを検出でき、`update_if_newer` 競合リゾルバーを適用します。

行の競合には次のものが含まれます。

- `INSERT` vs `INSERT`
- `UPDATE` vs `UPDATE`
- `UPDATE` vs `DELETE`
- `INSERT` vs `UPDATE`
- `INSERT` vs `DELETE`
- `DELETE` vs `DELETE`

`bdr.node_conflict_resolvers` ビューは、すべての既知の競合タイプに対して競合解決が現在どのように構成されているかに関する情報を提供します。

#### INSERT/INSERT の競合

最も一般的な競合 `INSERT` /`INSERT` は、2 つの異なるノードでの `INSERT` 操作が、同じ `PRIMARY KEY` 値または `PRIMARY KEY` が存在しない場合、単一の `UNIQUE` コンストレインの同じ値を持つタプルを作成する場合に発生します。

PGD は、元のノードのタイムスタンプに従って、2 つのうちの最後に挿入されたタプルを保持することにより、この状況を処理します。ユーザー定義の競合ハンドラーは、この動作をオーバーライドできます。

この競合は `insert_exists` 競合タイプを生成します。これは、デフォルトで解決されます。コミット時間に基づいて、新しい行を選択し、その 1 つだけ `update_if_newer` リゾルバー。他のリゾルバーを構成できます。詳細は、欠落列の競合の解決 を参照してください。

この競合タイプを解決するには、列レベルの競合解決とユーザー定義の競合トリガーを使用することもできます。

*global sequences* を使用して、このタイプの競合を効果的に排除できます。

### UNIQUE または EXCLUDE 制約に違反する INSERT 操作

INSERT /INSERT 競合は、複数の UNIQUE 制約のうちの 1 つが PRIMARY KEY である可能性がある、または 1 つ以上の EXCLUDE 制約に違反する可能性があります。

次の場合、レプリケーション変更を適用すると、multiple_unique_conflicts 競合が生成されます。これらの場合は両方とも、複数の他の行との競合が発生します。

- 新しい行が複数の UNIQUE 制約に違反し、その結果、他の複数の行と競合する場合。
- 新しい行が複数の EXCLUDE 制約または単一の EXCLUDE 制約に違反する場合。

このような競合が発生した場合、レプリケーションを続行するには、いくつかの行を削除する必要があります。multiple_unique_conflicts のリゾルバー設定によっては、適用プロセスはエラーで終了するか、着信行をスキップするか、一部の行を削除します。削除では、正しい PRIMARY KEY を含む行を保持し、他の行を削除しようとします。

!!!警告

この方法で複数の行が競合している場合、競合解決の結果、挿入操作を続行すると、データの一部が常に削除されます。

### UPDATE/UPDATE の競合

異なるノードでの 2 つの同時 UPDATE 操作が同じタプルを変更するが、その PRIMARY KEY を変更しない場合、再生時に UPDATE /UPDATE 競合が発生する可能性があります。

これらは、構成と状況に基づいて、さまざまな種類の競合を生成する可能性があります。テーブルが 行バージョンの競合検出 で構成されている場合、元のキー行がローカル行と比較されます。それらが異なる場合、update_differing 競合が生成されます。 オリジンの競合検出を使用する場合、行の起点がチェックされます。オリジンは、現在のローカル行が発生したノードです。これが変更されると、update_origin_change 競合が生成されます。その他の場合、UPDATE は通常、競合を生成せずに適用されます。

これらの競合は両方とも、*INSERT/INSERT の競合* で説明している insert_exists と同じ方法で解決されます。

### 主キーの UPDATE の競合

PGD は、現在、 PRIMARY KEY が UPDATE オペレーションによって変更される競合解決を実行できません。主キーを更新できますが、既存の値と競合しないことを確認する必要があります。

主キーの更新の競合は 多様な競合 であり、手動介入が必要です。

Postgres で主キーの更新は可能ですが、Postgres と PGD の両方に問題があります。

これらの更新がどのように動作するかを確認するには、主キーを使用して基本的なテーブルを作成し、最初の行を挿入します。

```
CREATE TABLE pktest (pk integer primary key, val integer);
INSERT INTO pktest VALUES (1,1);
```

主キー列の更新は技術的に可能です。新しい値は既存のデータと競合しないため、このコマンドは成功します。

```
UPDATE pktest SET pk=2 WHERE pk=1;
```

ただし、テーブルに複数の行があるとします。

```
INSERT INTO pktest VALUES (3,3);
```

一部の UPDATE 操作は成功します。

```
UPDATE pktest SET pk=4 WHERE pk=3;
```

すべての行を選択して、テーブルへの変更を確認します。主キー値が正常にシフトされたことに注意してください。

```
SELECT * FROM pktest;
 pk | val
- ---+-----
  2 |   1
  4 |   3
(2 rows)
```

他の UPDATE 操作は制約エラーで失敗します。

```
UPDATE pktest SET pk=4 WHERE pk=2;
```

テーブルからすべての行を選択し、標準の制約違反エラーを観察します。

```
ERROR:  duplicate key value violates unique constraint "pktest_pkey"
DETAIL:  Key (pk)=(4) already exists
```

Postgres は主キーの一意性を厳密に適用するため、PGD がなくても、これらのランタイムエラーを防ぐようにアプリケーションをデザインする必要があります。

PGD のような分散環境では、主キーの更新の複雑さが増加します。複数のノードがキーを同時に変更しようとすると、結果は、変更がオーバーラップまたは競合するかどうかによって異なります。

2 つのノードが異なる行を異なる新しい主キー値に更新する場合、PGD は変更を正常にレプリケートできます。

```
- -node1:
UPDATE pktest SET pk=pk+1 WHERE pk = 2;

After replication, both nodes reflect the incremented values, maintaining a␣
↪consistent state across the cluster:

- -node2:
UPDATE pktest SET pk=pk+1 WHERE pk = 4;
```

```
SELECT * FROM pktest;

 pk | val
- ---+-----
  3 |   1
  5 |   3
(2 rows)
```

両方の変更が受け入れられるため、これらの次の 2 つの変更を同時に実行すると、発散エラーが発生します。両方のノードは自分自身のローカル変更を受け入れますが、他のノードに変更を適用すると、update_missing 競合が発生します。

```
- -node1:
UPDATE pktest SET pk=1 WHERE pk = 3;
- -node2:
UPDATE pktest SET pk=2 WHERE pk = 3;
```

このシナリオでは、各ノードで異なるデータのままになります。

```
- -node1:
SELECT * FROM pktest;
 pk | val
- ---+-----
  1 |   1
  5 |   3
(2 rows)

- -node2:
SELECT * FROM pktest;
 pk | val
- ---+-----
  2 |   1
  5 |   3
(2 rows)
```

LiveCompare を使用して、この状況を特定して解決できます。

同時競合により問題が発生します。これら 2 つの変更を同時に実行することは、解決するのが簡単ではありません。

```
- -node1:
UPDATE pktest SET pk=6, val=8 WHERE pk = 5;
- -node2:
UPDATE pktest SET pk=6, val=9 WHERE pk = 5;
```

どちらの変更もローカルに適用されるため、ノード間の相違が発生します。しかし、ターゲットへの適用は、両方のノードで失敗し、重複キー値違反エラーが発生します。このエラーにより、レプリケーション

が停止され、手動の解決が必要です。

conflict_type update_pkey_exists を skip 、update 、または update_if_newer に設定すると、この重複キー違反エラーを回避でき、レプリケーションは中断されません。このアプローチでも、更新の性質によっては、発散が発生する可能性があります。

update_pkey_exists から update_if_newer を設定することにより、同じ古いキーが同じ新しいキーによって同時に更新されている場合、発散を回避できます。ただし、特定の状況では、update_if_newer でも発散が発生します。つまり、2 つの異なる行が両方とも同じ新しい主キーに同時に更新される場合。

その結果、アプリケーション、特に PGD で主キーの UPDATE 操作を許可しないことを強くお勧めします。アプリケーションの一部が主キーを変更する場合、同時変更を回避するには、クォーラムコミット を使用してこれらの変更を行います。これは、熱心な競合解決を使用して、ノードがコミットする前に同時競合するコミットを防止します。

**警告**

*update_pkey_exists* 競合の競合解決の結果が更新された場合、常にいずれかの行が削除されます。

### UNIQUE または EXCLUDE 制約に違反する UPDATE 操作

*UNIQUE または EXCLUDE 制約に違反する INSERT 操作* と同様に、着信 UPDATE が複数の UNIQUE /EXCLUDE インデックス PRIMARY KEY を含むに違反するか、複数の行が競合するように単一の EXCLUDE インデックスに違反する場合、PGD は multiple_unique_conflicts 競合を発生させます。

PGD は、遅延一意コンストレインをサポートしています。トランザクションがソースでコミットできる場合、競合が発生しない限り、ターゲットにクリーンに適用されます。ただし、遅延主キーをレプリカ ID として使用することはできないため、ユースケースはそれと複数の一意の制約の使用に関する警告によって既に制限されています。

### UPDATE/DELETE の競合

1 つのノードは、別のノードが同時に削除する行を更新できます。この場合、再生時に UPDATE /DELETE 競合が発生する可能性があります。

削除された行がまだ検出可能な場合 削除された行が VACUUM によって削除されなかった場合、update_recently_deleted 競合が生成されます。デフォルトでは、UPDATE はスキップされますが、この解像度を構成できます。詳細は、欠落列の競合の解決 を参照してください。

ローカルノードのレプリケーションが遅延している場合、データベースは、UPDATE を受信するまでに削除された行をクリーンアップできます。この場合、PGD は UPDATE /DELETE 競合と *INSERT / UPDATE の競合* を区別できません。update_missing 競合が生成されます。

競合する DELETE および UPDATE の別のタイプは、行がローカルに更新された後に来る DELETE です。この状況では、結果は使用される競合検出の種類によって異なります。デフォルトのオリジンの競合検出 を使用する場合、競合は検出されないため、 DELETE が適用され、行が削除されます。 行バージョンの競合検出 を有効にすると、delete_recently_updated 競合が生成されます。delete_recently_updated 競合のデフォルトの解決策は、skip 削除です。ただし、解決を構成するか、それを処理するように競合トリガーを構成できます。

### INSERT / UPDATE の競合

デフォルトの非同期オペレーションモードを使用している場合、ノードは元の INSERT を受信する前に行の UPDATE を受信する場合があります。これは、3 つ以上のノードがアクティブな場合にのみ発生する可能性があります *3* つ以上のノードとの競合 を参照してください。

これが発生すると、update_missing 競合が生成されます。デフォルトの競合リゾルバーは insert_or_skip ですが、代わりに insert_or_error または skip を使用できます。insert-or-action を実行するリゾルバーは、可能な場合行全体が受信されたとき、UPDATE からのデータに基づいて、新しい行を INSERT 試行します。行の再構成を可能にするには、テーブルに REPLICA IDENTITY FULL が必要であるか、行にトーストされたデータが含まれていない必要があります。

トーストデータの詳細については、 *TOAST* サポートの詳細 を参照してください。

### INSERT/DELETE の競合

INSERT /UPDATE 競合と同様に、ノードは INSERT をまだ受け取っていない行で DELETE オペレーションを受信する場合があります。これも、3 つ以上のノードがセットアップされている場合にのみ可能です *3* つ以上のノードとの競合 を参照してください。

PGD は現在この競合タイプを検出できません。 INSERT オペレーションは競合タイプを生成せず、 INSERT が適用されます。

DELETE オペレーションは常に delete_missing 競合を生成します。これは、デフォルトでオペレーションをスキップすることにより解決されます。

### DELETE/DELETE の競合

DELETE / DELETE 競合は、2 つの異なるノードが同じタプルを同時に削除すると発生します。

このシナリオでは、常に delete_missing 競合が生成されます。これは、デフォルトで、オペレーションをスキップすることにより解決されます。

両方の DELETE 操作は同じ効果があるため、この競合は無害です。それらのいずれかを無視しても安全です。

### 3 つ以上のノードとの競合

1 つのノードが行を挿入し、それが 2 番目のノードに再生され、そこで更新されると、3 番目のノードは、最初のノードから INSERT を受信する前に、2 番目のノードから UPDATE を受信できます。このシナリオは、INSERT /UPDATE 競合です。

これらの競合は、UPDATE を破棄することにより処理されます。これにより、異なるノードに異なるデータが発生する可能性があります。これらは多様な競合 です。

この競合タイプは、3 つ以上のマスターでのみ発生する可能性があります。少なくとも 2 人のマスターがアクティブに書き込みを行っている必要があります。

また、ノード 1 からノード 3 へのレプリケーションラグは、次の一連のアクションを許可するように十分に高い必要があります。

1. ノード 2 はノード 1 から INSERT を受信します
2. ノード 2 は UPDATE を実行します

3. ノード 3 はノード 2 から UPDATE を受信します
4. ノード 3 はノード 1 から INSERT を受信します

insert_or_error または場合によっては、 update_missing 競合タイプの insert_or_skip 競合リゾルバーを使用することは、これらの競合の実行可能な緩和戦略です。ただし、このオプションを有効にすると、INSERT /DELETE 競合の扉が開きます。

1. ノード 1 が UPDATE を実行します
2. ノード 2 が DELETE を実行します
3. ノード 3 はノード 2 から DELETE を受信します
4. ノード 3 はノード 1 から UPDATE を受信し、INSERT に変換します

これらが問題である場合は、テーブルまたはデータベースのフリーズ設定を調整して、update_recently_deleted として正しく検出されるようにすることをお勧めします。

**熱心な競合解決** を使用して、これらの競合を防ぐこともできます。熱心な競合解決は **クォーラムコミット** で常にアクティブで、いずれかがコミットする前に競合するトランザクションを中止します。

INSERT /DELETE 競合は、3 つ以上のノードで発生する場合もあります。このような競合は、UPDATE が DELETE に置き換えられることを除き、 INSERT /UPDATE と同じです。これにより、delete_missing 競合が発生する可能性があります。

PGD は、 update_missing 競合で発生するように、各 INSERT を最近削除されたチェックにすることを選択できます。ただし、これを行うコストは大部分のユーザーにペナルティを与えるため、現時点では代わりに delete_missing ログに記録されます。

将来のリリースでは、delete_missing 競合が発生したときに LiveCompare を使用した再チェックにより、INSERT /DELETE 異常を自動的に解決します。アプリケーションは、bdr.conflict_history_summary ビューを確認することにより、これらを手動で実行できます。

これらの競合は、2 つの主な問題のユースケースで発生する可能性があります。

- INSERT の後にすぐに DELETE が続きます、キューイングアプリケーションで使用できます
- テーブルの主キー識別子が再利用される場合

これらのケースはいずれも一般的なものではありません。これらの問題のあるユースケースが発生した場合、影響を受けるテーブルをレプリケートしないことをお勧めします。

PGD は識別子の一意に依存してレプリケーションが正しく動作するため、後者の場合に問題があります。

同じ一意の識別子を挿入、削除、および後で再利用するアプリケーションは、問題を発生する可能性があります。これは、ABA problem と呼ばれます。PGD は、行が現在の行、最後の行、またはかなり古い行であるかを知る方法がありません。

一意の識別子の再利用は、時間の経過とともに一意の識別を妨げ、監査、トレーサビリティ、および賢明なデータ品質を妨げるため、ビジネス上の問題でもあります。アプリケーションは一意の識別子を再利用する必要はありません。

変更がシステム全体に渡されるまでにかかる時間間隔で発生する識別子の再利用は、問題を発生させます。この時間は通常のオペレーションでは短いかもしれませんが、ダウンノードではその間隔を数時間または数日に延長できます。

アプリケーションは一意の識別子を再利用しないことをお勧めします。その場合、1 年未満は再利用を避けるための措置を講じます。

この問題は、シーケンスまたは UUID を使用するアプリケーションでは発生しません。

### 外部キーコンストレインの競合

適用されるリモートトランザクションと既存のローカルデータの競合は、FOREIGN KEY FK コンストレインでも発生する可能性があります。

PGD は session_replication_role = 'replica' を使用して変更を適用するため、変更を適用するときに外部キーは再チェックされません。アクティブ/アクティブ環境では、参照テーブルへの挿入と同時に参照先テーブルで削除が発生すると、この状況により FK 違反が発生する可能性があります。このシナリオは、INSERT /DELETE 競合に似ています。

シングルマスター Postgres では、参照先テーブルの値を参照する INSERT /UPDATE は、行レベルのロックを取得する前に、DELETE 操作が完了するまで待機する必要があります。 DELETE が参照値を削除する場合、INSERT /UPDATE は FK チェックに失敗します。

マルチマスター PGD において。ノード間の行レベルのロックはありません。参照テーブルの INSERT は、参照テーブルの DELETE の背後で待機しないため、両方のアクションが同時に発生できます。したがって、参照テーブルのあるノードの INSERT /UPDATE は、別のノードの参照先テーブルの DELETE として同時に値を使用できます。したがって、結果は、参照先テーブルに存在しない参照テーブルの値です。

実際、この状況は、 DELETE 操作が、一般的な操作ではない参照テーブルでの DELETE 操作とは別のトランザクションの参照対象テーブルで発生した場合に発生します。

Orders -> OrderItems のような親子関係では、これは一般的ではありません。OrderItem を完全に削除するよりも、キャンセル済みとしてマークする可能性が高くなります。参照/ルックアップデータの場合、新しいファクトデータに同じ値を使用すると同時にエントリを完全に削除することはまれです。

ぶら下がった FK が可能ですが、一般的にこのリスクは非常に低いです。したがって、PGD は、この場合をカバーする汎用ソリューションを課しません。これが発生する状況を理解すると、2 つの解決策が考えられます。

最初の解決策は、一般的に一度に 1 つのノードのみから変更される、変更の頻度が低い、または変更のコンカレンシーがアプリケーションを介する場合に、密接に関連するエンティティに FK の使用を制限することです。このアプローチでは、アプリケーションレベルでの FK 違反を回避します。

2 番目の解決策は、 PGD 提供のファンクション bdr.ri_fkey_trigger() および bdr.ri_fkey_on_del_trigger() を使用して、この場合から保護するトリガーを追加することです。BEFORE トリガーとして呼び出された場合、これらのファンクションは FOREIGN KEY 情報を使用して、SET NULL 制約があるかのように参照列を NULL に設定することにより FK 異常を回避します。このアプローチでは、1 つのトリガーですべての FK を再チェックするため、FK 違反を防ぐには、テーブルごとにトリガーを 1 つだけ追加する必要があります。

例として、Fact と RefData の 2 つのテーブルがあるとします。Fact には、RefData を参照する FK があります。Fact は参照テーブルであり、RefData は参照されるテーブルです。各テーブルにトリガーを 1 つ追加する必要があります。

RefData で参照されている行が既に削除されている場合、Fact に列を NULL に設定するトリガーを追加します。

```
CREATE TRIGGER bdr_replica_fk_iu_trg
    BEFORE INSERT OR UPDATE ON fact
    FOR EACH ROW
    EXECUTE PROCEDURE bdr.ri_fkey_trigger();

ALTER TABLE fact
    ENABLE REPLICA TRIGGER bdr_replica_fk_iu_trg;
```

RefData テーブルで DELETE が発生したときに、Fact の列を NULL に設定するトリガーを追加します。

```
CREATE TRIGGER bdr_replica_fk_d_trg
    BEFORE DELETE ON refdata
    FOR EACH ROW
    EXECUTE PROCEDURE bdr.ri_fkey_on_del_trigger();

ALTER TABLE refdata
    ENABLE REPLICA TRIGGER bdr_replica_fk_d_trg;
```

両方のトリガーを追加すると、ダングリング外部キーが回避されます。

### TRUNCATE の競合

TRUNCATE は、すべての行の DELETE と同様に動作しますが、行ごとに削除するのではなく、テーブルデータを物理的に削除することによりこのアクションを実行します。その結果、行レベルの競合処理は利用できないため、明確な競合が存在する場合でも、TRUNCATE コマンドは他の DML アクションとの競合を生成しません。

その結果、TRUNCATE への他のノードで別の DML が同時に実行される場合、再生の順序により発散的な変更が発生する可能性があります。

次のいずれかのアクションを実行できます。

- TRUNCATE が他のコンカレント DML とともに実行されていないことを確認します。LiveCompare を利用して、このような矛盾を強調表示します。
- TRUNCATE を、 WHERE 句のない DELETE ステートメントに置き換えます。このアプローチは、大規模なテーブルではパフォーマンスが低下する可能性があります。
- Set bdr.truncate_locking = 'on' TRUNCATE コマンドのロック動作を設定します。この設定は、TRUNCATE が bdr.ddl_locking 設定に従うかどうかを決定します。すべてのノードが起動している必要があるため、これは TRUNCATE のデフォルト動作ではありません。この構成は、すべての場合に使用できないか、必要な場合があります。

### ロールとテーブルスペースの違いのデータの競合

競合は、ノードにロールのようなグローバル Postgres システム全体のデータが異なる場合にも発生する可能性があります。この競合により、操作主に DDL が発生し、1 つのノードで正常に実行してコミットできますが、他のノードに適用されません。

たとえば、ノード 1 には fred という名前のユーザーがあり、そのユーザーはノード 2 で作成されたもので

はない場合があります。ノード 1 の fred がテーブルを作成すると、テーブルは所有者を fred に設定して複製されます。DDL コマンドがノード 2 に適用されると、fred という名前のユーザーが存在しないため、DDL は失敗します。この障害により、Postgres ログにエラーが生成されます。

PGD が実行されているデータベースにユーザー fred を作成することにより、この競合を解決するには、管理者の介入が必要です。将来的にこれを解決するように `bdr.role_replication = on` を設定できます。

### ロックの競合とデッドロックアボート

PGD ライタプロセスは通常のユーザーセッションと同じように動作するため、行とテーブルのロックに関する通常のルールの対象となります。これにより、PGD ライタプロセスがユーザートランザクションまたは相互に保持するロックで待機する場合があります。

関連するロックには次のものが含まれます。

- ユーザーセッションによる明示的なテーブルレベルのロック `LOCK TABLE ...`
- ユーザーセッションによる明示的な行レベルのロック `SELECT ... FOR UPDATE/FOR SHARE`
- ローカルアクティビティまたは他のノードからのレプリケーションからの行 `UPDATE` 、`INSERT` 、または `DELETE` 操作のための暗黙的ロック

PGD ライタプロセスは、ユーザートランザクションとデッドロックになる場合があります。ユーザートランザクションは、ライタプロセスが保持するロックで待機しています。その逆も可能です。2 つのライタプロセスが互いにデッドロックする場合もあります。Postgres のデッドロック検出機能は、問題のあるトランザクションのいずれかに介入して終了します。PGD ライタプロセスが終了すると、再試行され、通常は成功します。

これらの問題はすべて一時的なもので、通常、管理者のアクションは必要ありません。ライタプロセスがアイドルなユーザーセッションのロックの背後で長時間スタックしている場合、管理者はユーザーセッションを終了して、レプリケーションフローを再度取得できます。ただし、これは、ユーザーが別のユーザーセッションに影響を与える長いロックを保持していることと変わりません。

Postgres の log_lock_waits 機能を使用すると、ロック関連の再生ストールを特定するのに役立ちます。

### 多様な競合

発散競合は、別のノードで同じであるはずのデータが予期せず異なる場合に発生します。多様な競合は発生すべきではありませんが、執筆時点でこのような競合すべてを確実に防止できるわけではありません。

すべてのノードが変更を再生する前に別のノードが同じ行のキーを変更した場合、行の `PRIMARY KEY` を変更すると、発散競合が発生する可能性があります。主キーの変更を避けるか、1 つの指定ノードでのみ変更します。

行データを含む多様な競合では、一般に、一方のノードのデータを他方のノードと一貫性があるように手動で調整する管理者のアクションが必要です。文書に従って PGD を使用し、安全でないとマークされた設定または機能を回避する限り、このような競合は発生しません。

管理者は、このような競合を手動で解決する必要があります。競合の性質によっては、`bdr.ddl_replication` や `bdr.ddl_locking` などの詳細オプションの使用が必要になる場合があります。ただし、これらのオプションを不注意に使用すると、状況をさらに悪化させ、汎用命令では対処できない競合が発生する可能性があります。

### TOAST サポートの詳細

Postgres は、TOAST と呼ばれる大きな列にアウトオブラインストレージを使用します。

論理デコーディング PGD がその上に構築されるおよび論理レプリケーションでの TOAST 値の処理は、テーブルのメイン行の一部として保存されるインラインデータとは異なります。

TOAST 値は、値が変更された場合にのみトランザクションログ WAL に記録されます。トーストされた列の値を変更しなかった `UPDATE` ステートメントは、その列のない行を生成するため、特に UPDATE の競合を処理する場合に問題が発生する可能性があります。*INSERT / UPDATE の競合* で述べたように、PGD は、`insert_or_error` を使用して `update_missing` 競合が解決され、欠落している TOAST 列がある場合にエラーを報告します。

ただし、非同期レプリケーションと同時ワークロードの場合、これよりも微妙な問題が発生します。Eager トランザクションは影響を受けません。たとえば、A、B、および C と呼ばれる 3 つのノードを持つ EDB Postgres 分散クラスターでの次のワークロードを想像してください。

1. ノード A で txn A1 は UPDATE SETcol1 = ’ toast data…’ を実行し、最初にコミットします。
2. ノード B で txn B1 は UPDATE SET other_column = ‘anything else’ ; そして、A1 の後にコミットします。
3. ノード C では、ノード A への接続が遅れます。
4. ノード C では txn B1 が最初に適用されますが、col1 の TOASTed 列を見逃しますが、競合せずに適用されます。
5. ノード C では、txn A1 は update_origin_change で競合し、スキップされます。
6. ノード C は、A1 からのトーストされたデータを永久に逃します。

PGD を使用する場合、このシナリオは通常問題になりません。PGD は、マルチマスターにビルトイン論理レプリケーションまたはプレーン pglogical のいずれかを使用する場合です。PGD は、 TOAST 列変更を最近レプリケートし、ローカル `UPDATE` がそうでない行にローカル `UPDATE` を検出すると、TOAST 列の独自のロギングを追加します。 t TOAST を変更するしたがって、PGD は、異なるノード間でトーストされたデータの不一貫性を防ぎます。この状況により、マルチプルのノードで更新が発生した場合、つまりタプルのオリジンが変更された場合、WAL ログの増加が発生します。通常、PGD AlwaysOn アーキテクチャの場合と同様に、すべての更新が単一のノードから行われる場合、追加の WAL オーバーヘッドはゼロです。

> **注釈**
>
> メインテーブルで同じことをせずに TOAST テーブルのみで `VACUUM FULL` または `CLUSTER` を実行すると、追加のロギングが機能するために必要なメタデータが削除されます。これは、このようなステートメントの後の短期間、これらの同時実行問題に対する保護が存在しないことを意味します。

> **警告**
>
> TOAST の追加 WAL ロギングは、標準の Postgres の `BEFORE UPDATE` トリガーを使用して実行されます。このトリガーは、テーブル上のすべての `BEFORE UPDATE` トリガーの中で、トリガー名に基づいて、アルファベット順に最後にソートする必要があります。これを簡単にするために `zzzz_bdr_`

という接頭辞が付いていますが、その後にソートされる名前でトリガーを作成しないようにしてください。そうしないと、同時実行の問題から保護されません。

insert_or_error 競合解決の場合、引き続き REPLICA IDENTITY FULL の使用が必要です。

トーストされた列に関連するこれらの問題は、 REPLICA IDENTITY FULL を使用するテーブルには影響しません。この設定では、行全体がキーの一部と考えられるため、キーの一部としてトースト値が常にログに記録されます。PGD は、新しい行を再構成し、キー行から欠落しているデータを埋めることができます。その結果、REPLICA IDENTITY FULL を使用すると、WAL サイズが大幅に増加する可能性があります。

### Conflict detection

PGD は、競合検出のための次のメカニズムを提供します。

- オリジンの競合検出 デフォルト
- 行バージョンの競合検出
- *Column-level conflict detection*

#### オリジンの競合検出

オリジン競合検出は、トランザクションの発生元のノードに記録されたコミットタイムスタンプを使用および依存します。これには、クロックが正しく動作するか、2 つのノード間の最速のメッセージの許容範囲内に同期している必要があります。これが当てはまらない場合、競合解決はより前にあるノードを優先する傾向があります。パラメーター bdr.maximum_clock_skew および bdr.maximum_clock_skew_action を使用して、ノード間のクロックスキューを管理できます。

**注釈**

ローカルノードのクロックがリモートノードのコミットタイムスタンプより遅れている場合、グループコミットスコープを使用するトランザクションではコミットレイテンシーが増加します。ローカルノードは、ローカルのタイムスタンプがリモートコミットタイムスタンプより古くなるまで待機する必要があるため、*bdr.maximum_clock_skew* または`bdr.maximum_clock_skew_action` の設定に関係なく発生します。 最適なパフォーマンスを保証し、遅延を最小限に抑えるには、システムクロックを厳密に同期する必要があります。

競合は、レプリケーション元が変更されたかどうかに基づいて最初に検出されるため、競合と判明しない可能性のある状況で競合トリガーが呼び出されます。したがって、このメカニズムは正確ではありません、偽陽性の競合を生成する可能性があります。

元の情報は、行がフリーズされた時点まででのみ利用できます。凍結された後に行に到着する更新は競合を発生させないため、すべての場合に適用されます。これは bdr_init_physical によって新しいノードを追加する場合の通常のケースであるため、競合が発生すると、その場合に多くの偽陽性の結果が発生します。

オフラインだったノードが再接続してデータ変更の送信を開始すると、新しく到着した更新が更新する凍結行よりも古い場合、発散エラーが発生する可能性があります。挿入と削除は、この状況の影響を受けません。

*Node restart and down node recovery* で説明しているように、長時間の停止のためにノードをダウンしたままにしないことをお勧めします。

EDB Postgres Extended Server および EDB Postgres Advanced Server では、ノードがダウンしている間、PGD は行のフリーズを抑制します。このメカニズムはこの状況を適切に処理するため、パラメーター設定を変更する必要はありません。

Postgres の他のバリアントでは、この状況をある程度注意して管理する必要がある場合があります。

フリーズは、通常、バキュームされる行が現在の xid から vacuum_freeze_min_age xid より古い場合に発生します。これは、これらのパラメーターに適切に高い値を構成する必要があることを意味します。

- vacuum_freeze_min_age
- vacuum_freeze_table_age
- autovacuum_freeze_max_age

トランザクションレートに基づいて値を選択し、データベースノードから競合データを削除する前にダウンタイムの猶予期間を与えます。たとえば、vacuum_freeze_min_age が 5 億に設定されている場合、1000TPS を実行するノードは、競合データが削除されるまで 5.5 日強の間ダウンする可能性があります。CommitTS データ構造は、その設定で 5 GB のディスク上の領域を必要とするため、トランザクションレートの低いシステムはより低い設定からメリットを得ることができます。

最初に、推奨される設定は次のとおりです。

```
#  1 billion = 10GB

autovacuum_freeze_max_age = 1000000000

vacuum_freeze_min_age = 500000000

#  90% of autovacuum_freeze_max_age

vacuum_freeze_table_age = 900000000
```

次のことに注意してください。

- autovacuum_freeze_max_age はノード起動時にのみ設定できます。
- vacuum_freeze_min_age を設定できるため、低い値を使用すると行が早期にフリーズし、競合が無視される場合があります。個々のテーブルに autovacuum_freeze_min_age および toast.autovacuum_freeze_min_age を設定することもできます。
- CLUSTER または VACUUM FREEZE コマンドを実行しても、行が早期にフリーズされ、競合が無視される場合があります。

#### 行バージョンの競合検出

PGD は、行バージョン管理を使用し、ノードのシステムクロックとは無関係に競合検出を行うオプションを提供します。

行バージョンの競合検出には、3 つのことを有効にする必要があります。これらの手順のいずれかが正しく実行されない場合、**オリジンの競合検出** が使用されます。

- 行バージョンの競合検出を使用するすべてのテーブルで REPLICA IDENTITY FULL を有効にします。
- bdr.alter_table_conflict_detection を使用して、テーブルの行バージョンの追跡を有効にします。このファンクションは、指定した名前の列と、新しい列値を管理する UPDATE トリガーを追加します。列は INTEGER タイプとして作成されます。

カウンタは UPDATE でのみインクリメントされますが、この技術により、UPDATE と DELETE の両方の競合検出が可能になります。

このアプローチはランポートタイムスタンプに似ており、競合検出の ABA 問題を完全に防止します。

> **注釈**
>
> 行レベルの競合解決は、行バージョン管理を使用しても **欠落列の競合の解決** 構成に基づいて処理されます。行バージョンの生成方法は、競合を検出する場合にのみ役立ちます。行のどのバージョンが新しいかに関する信頼できる情報として依存しないでください。

特定のテーブルに使用されている現在の競合検出ストラテジーを判断するには、ビュー bdr.tables の列 conflict_detection を参照してください。

現在の競合検出ストラテジーを変更するには、 bdr.alter_table_conflict_detection を使用します。

### Conflict resolution

ほとんどの競合は自動的に解決できます。PGD は、last-update-wins メカニズム、より正確には update_if_newer 競合リゾルバーにデフォルトで設定されます。このメカニズムでは、競合検出に使用される同じコミットタイムスタンプに基づいて、2 つの競合する行の最後に挿入または変更された行を保持します。特定のコーナーケースのシナリオでの動作は、bdr.create_node_group または bdr.alter_node_group に使用される設定によって異なります。

PGD では、 bdr.alter_node_set_conflict_resolver を使用して競合解決のデフォルト動作をオーバーライドできます。

### Conflict logging

PGD 6.0 以降、競合ログはデフォルトで有効になっていません。ノードごとに競合ログを有効にできます。これにより、特定のノードで発生した競合のログを記録できます。これは、デバッグやモニタリングの目的に役立ちます。

名前付けノードでこのコマンドを実行して、その特定のノードでのすべての競合のログを有効にします。すべてのノードでロギングを有効にする場合は、PGD グループの各ノードでこのコマンドを実行します。

```
SELECT bdr.alter_node_set_log_config(`nodename`, false, true, NULL, NULL);
```

ノード名コントロールの後の他のパラメーターは、それぞれ、競合を PGD ログファイルに記録するかどうか、および競合を bdr.conflict_history テーブルに記録するかどうかを指定します。最後の 2 つのパラメーターは、ログに記録する競合タイプと競合の解決を指定する文字列の配列を受け取ります。両方の完

全なリストについては、リファレンスセクションの Conflicts を参照してください。これらのパラメーターを NULL に設定すると、PGD はすべての競合タイプと解決策をログに記録します。

PGD は、すべての競合のログを bdr.conflict_history テーブルに記録します。bdr.alter_node_set_log_config を使用して、この動作をより詳細に変更できます。

#### 競合報告

レポートのテーブルに記録された競合を要約できます。レポートにより、アプリケーションの所有者は競合を特定、理解、解決し、競合を防ぐためのアプリケーションの変更を導入できます。

```
SELECT nspname, relname
, date_trunc(day, local_time) :: date AS date
, count(*)
FROM bdr.conflict_history
WHERE local_time > date_trunc(day, current_timestamp)
GROUP BY 1,2,3
ORDER BY 1,2;
```

```
 nspname | relname |    date    | count
- --------+---------+------------+-------
 my_app  | test    | 2019-04-05 |     1
(1 row)
```

### Data verification with LiveCompare

LiveCompare は、2 つのデータベースを比較して、それらが同一であることを確認するように設計されたユーティリティプログラムです。

LiveCompare は PGD スタックの一部として含まれ、PGD ノードの任意のペアを対象とできます。デフォルトでは、レプリケートされたすべてのテーブルを比較し、違いを報告します。LiveCompare は、Postgres や Oracle などの非 PGD データソースでも動作します。

LiveCompare を使用して、着信行を継続的に監視することもできます。コンテキスト情報を失うことなく停止および開始できるため、都合の良いときに実行できます。

LiveCompare では、複数のテーブルを同時にチェックできます。いくつかのテーブル、またはテーブル内の行のセクションのみのチェックを許可するように構成できます。チェックは、最初に行全体のハッシュを比較することにより実行されます。異なる場合、LiveCompare は行全体を比較します。LiveCompare は、有用なサイズのバッチで行を比較することにより、オーバーヘッドを回避します。

違いが見つかった場合は、時間の経過とともに再チェックできるため、最終的な一貫性の遅延が発生します。

詳細については、 LiveCompare ドキュメントを参照してください。

### 6.8.2 Column-level conflict detection

デフォルトでは、競合は行レベルで解決されます。2 つのノードからの変更が競合する場合、ローカルまたはリモートのタプルのいずれかが選択され、他方は破棄されます。たとえば、2 つの競合する変更のコミットタイムスタンプを比較し、新しい方が保持される場合があります。このアプローチでは、すべてのノードが同じ結果に収束し、クラスター全体でコミットオーダーのようなセマンティクスを確立します。

ただし、少なくとも一部の場合、行レベルではなく列レベルで競合を解決することが適切な場合があります。

- *DDL overview* は、行レベルの競合解決とは対照的に列レベルの競合解決を導入し、行レベルの競合解決よりも適している場合を提案します。
- *Enabling and disabling column-level conflict resolution* は、列レベルの競合解決を有効にする例を提供し、列レベルの競合解決を有効にしてテーブルをリストする方法について説明します。
- *Timestamps in column-level conflict resolution* は、`column_modify_timestamp` と `column_commit_timestamp` の使用の違いを説明し、列レベルの競合解決に関連付けられたタイムスタンプを選択および検査する方法を示します。

#### Overview

デフォルトでは、競合は行レベルで解決されます。2 つのノードからの変更が競合する場合、ローカルまたはリモートのタプルのいずれかが選択され、他方は破棄されます。たとえば、2 つの競合する変更のコミットタイムスタンプを比較し、新しい方が保持される場合があります。このアプローチでは、すべてのノードが同じ結果に収束し、クラスター全体でコミットオーダーのようなセマンティクスを確立します。

ただし、少なくとも一部の場合、行レベルではなく列レベルで競合を解決することが適切な場合があります。

##### 列レベルで解決する場合

テーブル t に 2 つの整数列 a および b と、単一の行 (1,1) がある単純な例を考えます。 1 つのノードで次を実行します。

```
UPDATE t SET a = 100
```

別のノードでは、前述の `UPDATE` を受信する前に、次のことを同時に実行します。

```
UPDATE t SET b = 100
```

> **注釈**
>
> *UPDATE* によって変更される属性は、トリガー内の古い行と新しい行を比較することにより決定されます。これは、属性が値を変更しない場合、明示的に設定されていても変更として検出されないことを意味します。たとえば、 *UPDATE t SET a = a* は、 *a* をどの行でも変更されたものとしてマークしません。同様に、*UPDATE t SET a = 1* は、既に `1` に設定されている行の `a` を変更済みとしてマークしません。

このシーケンスにより、`UPDATE-UPDATE` 競合が発生します。

update_if_newer 競合解決では、コミットタイムスタンプが比較され、新しい行バージョンが保持されます。2 番目のノードが最後にコミットしたと仮定すると、結果は (1,100) であり、列 a への変更を効果的に破棄します。

多くのユースケースでは、この動作は望ましく、予想されます。ただし、一部のユースケースでは、これが問題になる場合があります。たとえば、アプリケーションの各部分が別のノードに接続され、共有テーブルの列の専用サブセットを更新するマルチノードクラスターを考えます。その場合、さまざまなコンポーネントが競合し、変更を上書きする可能性があります。

このようなユースケースでは、特定のテーブルの競合を列レベルで解決することがより適切である場合があります。これを行うために、PGD は各列の最後の変更のタイムスタンプを個別に追跡し、それを使用して最新の値を選択し、基本的に update_if_newer を実行します。

前の例に適用すると、どのノードもそのような行を認識しないにもかかわらず、結果は両方のノードで (100,100) になります。

列レベルの競合の解決について考えるとき、テーブルを垂直にパーティション化して、各更新が 1 つのスライスのデータのみに影響を与えるようにすると便利です。このアプローチにより、列のさまざまなサブセットへの変更間の競合が排除されます。実際、垂直パーティショニングは、列レベルの競合解決の実用的な代替品です。

列レベルの競合解決には、テーブルに REPLICA IDENTITY FULL が必要です。

bdr.alter_table_conflict_detection() ファンクションはそれを確認し、この設定が存在しない場合はエラーで失敗します。

#### 列レベルの競合解決の特別な問題

列を独立して扱うことにより、すべての変更が同じノードで発生する場合、不可能な方法でコンストレインに違反するのが簡単になります。例、次のようなテーブルを考えます。

```
CREATE TABLE t (id INT PRIMARY KEY, a INT, b INT, CHECK (a > b));
INSERT INTO t VALUES (1, 1000, 1);
```

1 つのノードが次のことを行うと仮定します。

```
UPDATE t SET a = 100;
```

別のノードは同時に次のことを行います。

```
UPDATE t SET b = 500;
```

これらの更新はそれぞれ、最初の行で実行されたときに有効であるため、各ノードに渡されます。しかし、他のノードにレプリケートする場合、結果の行は CHECK (a > b) 制約に違反し、問題が手動で解決されるまでレプリケーションは停止します。

### CRDT データ型を使用した列レベルの競合の処理

デフォルトでは、列レベルの競合解決は、より高いタイムスタンプを持つ値を選択し、他の値を破棄します。ただし、さまざまなより複雑な方法で競合を調整することはできます。たとえば、情報を破棄せずに競合する値をマージできる オペレーションベースの *CRDT* タイプ *CmCRDT* を使用できます。

### Enabling and disabling column-level conflict resolution

bdr.alter_table_conflict_detection() ファンクションは、列レベルの競合解決を管理します。

### bdr.alter_table_conflict_detection を使用して列レベルの競合解決を有効にする

bdr.alter_table_conflict_detection ファンクションは、テーブル名と列名を引数として受け取ります。列は column_modify_timestamp 列としてテーブルに追加されます。また、ファンクションは、各変更の前に新しい列のタイムスタンプを維持する 2 つのトリガー BEFORE INSERT および BEFORE UPDATE も追加します。

```
db=# CREATE TABLE my_app.test_table (id SERIAL PRIMARY KEY, val INT);
CREATE TABLE

db=# ALTER TABLE my_app.test_table REPLICA IDENTITY FULL;
ALTER TABLE

db=# SELECT bdr.alter_table_conflict_detection(
db(# my_app.test_table::regclass,
db(# column_modify_timestamp, cts);
 alter_table_conflict_detection
- --------------------------------
 t

db=# \d my_app.test_table
                                      Table "my_app.test_table"
 Column |         Type          | Collation | Nullable |                      Default
- -------+-----------------------+-----------+----------+-----------------------------
↪---------------------
 id     | integer               |           | not null | nextval(my_app.test_table_id_
↪seq::regclass)
 val    | integer               |           |          |
 cts    | bdr.column_timestamps |           | not null | s 1 775297963454602 0 0::bdr.
↪column_timestamps
Indexes:
    "test_table_pkey" PRIMARY KEY, btree (id)
Triggers:
    bdr_clcd_before_insert BEFORE INSERT ON my_app.test_table FOR EACH ROW EXECUTE␣
↪FUNCTION bdr.column_timestamps_current_insert()
    bdr_clcd_before_update BEFORE UPDATE ON my_app.test_table FOR EACH ROW EXECUTE␣
↪FUNCTION bdr.column_timestamps_current_update()
```

新しい列は、デフォルト値で NOT NULL を指定します。これは、ALTER TABLE ... ADD COLUMN がテーブルリライトを実行しないことを意味します。

> **注釈**
>
> *bdr.column_timestamps* データ型の列を他の目的で使用することは避けてください。悪影響を与える可能性があります。たとえば、テーブルを列レベルの競合解決に切り替えますが、これはトリガーがなしでは正しく機能しません。

#### 列レベルの競合解決を使用したテーブルのリスト

次のクエリーを使用して、列レベルの競合解決が有効になっているテーブルをリストできます。

```
SELECT nc.nspname, c.relname
FROM pg_attribute a
JOIN (pg_class c JOIN pg_namespace nc ON c.relnamespace = nc.oid)
  ON a.attrelid = c.oid
JOIN (pg_type t  JOIN pg_namespace nt ON t.typnamespace = nt.oid)
  ON a.atttypid = t.oid
WHERE NOT pg_is_other_temp_schema(nc.oid)
  AND nt.nspname = bdr
  AND t.typname = column_timestamps
  AND NOT a.attisdropped
  AND c.relkind IN (r, v, f, p);
```

このクエリーは、bdr.column_timestamp 型の列の存在を検出します。

### Timestamps in column-level conflict resolution

列レベルの競合の解決は、テーブルに含まれるタイムスタンプ列によって異なります。

#### column_modify_timestamp と column_commit_timestamp の比較

2 つの列レベルの競合検出方法のいずれかを選択すると、変更された列とタイムスタンプのマッピングを含む列がテーブルに追加されます。

タイムスタンプマッピングを格納する列は自動的に管理されます。結果は予測できない場合があるため、クエリーで値を指定またはオーバーライドしないでください。可能な場合、ユーザーが値をオーバーライドしようとしても無視されます。

テーブルの列タイムスタンプを有効または無効にする場合、コードは DDL ロックを使用して、スイッチ前からの保留中の変更がないことを確認します。このアプローチでは、両方のタプルのタイムスタンプと競合のみが表示されないようにします。そうしないと、コードは予期せずローカルタプルのタイムスタンプとリモートタプルの NULL を認識する場合があります。また、変更がすべてのノードで同じ方法列レベルまたは行レベルで解決されることも保証します。

### column_modify_timestamp

競合検出方法として column_modify_timestamp が選択されている場合、変更された列に割り当てられたタイムスタンプは、現在のタイムスタンプであり、select_clock_timestamp() を実行して取得する値と似ています。

このアプローチは簡単で、多くの場合、たとえば、競合する行が列の重複しないサブセットを変更する場合、それは正しいです。ただし、そのシンプルさは予期しない効果をもたらす可能性があります。

たとえば、 UPDATE が複数の行に影響を与える場合、UPDATE の実行中クロックは刻み続けます。したがって、各行は、UPDATE によって同時に変更されている場合でも、わずかに異なるタイムスタンプを取得します。この動作は、さまざまなノードで実行される変更がどのようにインターリーブされるかに応じて、同時の変更のエフェクトがさまざまな方法で「混合」される可能性があることを意味します。

もう 1 つ考えられる問題は、クロックスキューです。別のノードのクロックがドリフトすると、それらのノードによって生成されたタイムスタンプもドリフトします。このクロックスキューは、タイムスタンプが明らかに切り替わっているため、新しい変更が破棄されるなどの予期しない動作を誘発する可能性があります。ただし、パラメーター bdr.maximum_clock_skew および bdr.maximum_clock_skew_action を使用してノード間のクロックスキューを管理できます。

現在のタイムスタンプはコミットタイムスタンプと無関係であるため、競合の解決にそれを使用することは、結果がコミット順序と等価ではないこと、つまりシリアル化できないことを意味します。

現在のタイムスタンプを使用して変更またはコミットを注文する場合、2 つ以上のノードがたまたま同じタイムスタンプを生成したため、競合する変更のタイムスタンプがまったく同じになる場合があります。このリスクは、通常の行レベルの競合解決でも発生する可能性があるため、列レベルの競合解決に固有ではありません。この状況では、ノード ID がタイブレーカーとして使用されます。ノード ID が高い方が優先します。このアプローチでは、すべてのノードに同じ変更が適用されます。

### column_commit_timestamp

column_commit_timestamp で指定された実際のコミットタイムスタンプを競合検出方法として使用することもできます。このアプローチには、UPDATE で行われたすべての変更で同じであるコミット時間を使用する利点があります。

> **注釈**
>
> 将来的にステートメントトランザクションが追加される可能性があります。これにより、コンカレントステートメントまたはトランザクションの混合効果に関する問題が解決されます。それでも、これらのオプションはいずれも、コミットオーダーと同等の結果を生成することはできません。

### カラムのタイムスタンプの検査

変更された列のタイムスタンプを格納する列は、トリガーによって維持されます。直接変更しないでください。たとえば、競合がどのように解決されたかを調査しているときに、現在のタイムスタンプの値を検査すると役立つ場合があります。

> **注釈**

タイムスタンプマッピングはトリガーによって維持され、トリガーの実行順序が重要です。カスタムトリガーがタプルを変更し、 pgl_clcd_トリガーの後に実行される場合、変更された列は正しく検出されません。これにより、誤った競合解決が発生する可能性があります。トリガー内のタプルを変更する必要がある場合は、 pgl_clcd_トリガーの前に実行されていることを確認します。

次の機能は、タイムスタンプを検査するために役立ちます。

### bdr.column_timestamps_to_text(bdr.column_timestamps)

このファンクションは、タイムスタンプマッピングの人間が判読可能な表現を返し、値をテキストにキャストするときに使用されます。

```
db=# select cts::text from test_table;
                                                    cts
- ---------------------------------------------------------------------------------------
↪----------------
 {source: current, default: 2018-09-23 19:24:52.118583+02, map: [2 : 2018-09-23␣
↪19:25:02.590677+02]}
(1 row)
```

### bdr.column_timestamps_to_jsonb(bdr.column_timestamps)

このファンクションは、タイムスタンプマッピングの JSONB 表現を変換し、値を jsonb にキャストするときに使用されます。

```
db=# select jsonb_pretty(cts::jsonb) from test_table;
                   jsonb_pretty
- ------------------------------------------------
 {                                                +
     "map": {                                     +
         "2": "2018-09-23T19:24:52.118583+02:00"  +
     },                                           +
     "source": "current",                         +
     "default": "2018-09-23T19:24:52.118583+02:00"+
 }
(1 row)
```

### bdr.column_timestamps_resolve(bdr.column_timestamps, xid)

このファンクションは、最新のトランザクションが既にコミットしている場合、それによって変更された属性のコミットタイムスタンプを使用してマッピングを更新します。これは、コミットタイムスタンプを使用する場合にのみ重要です。たとえば、この場合、最後のトランザクションは 2 番目の属性を更新しました attnum = 2

```
test=# select cts::jsonb from test_table;
                                                                    cts
```

(次のページに続く)

(前のページからの続き)

```
- ----------------------------------------------------------------------------
↪-------------------------------------------------
 {"map": {"2": "2018-09-23T19:29:55.581823+02:00"}, "source": "commit", "default":
↪"2018-09-23T19:29:55.581823+02:00", "modified": [2]}
(1 row)

db=# select bdr.column_timestamps_resolve(cts, xmin)::jsonb from test_table;
                                            column_timestamps_resolve
- ----------------------------------------------------------------------------
↪---------------------------------
 {"map": {"2": "2018-09-23T19:29:55.581823+02:00"}, "source": "commit", "default":
↪"2018-09-23T19:29:55.581823+02:00"}
(1 row)
```

### 6.8.3 Conflict-free replicated data types

競合のないレプリケートデータ型 CRDT は、従来の解決のように行の 1 つを破棄するのではなく、同時に変更された行からの値のマージをサポートします。

- *CRDTs Overview* は、テーブルで CRDT を使用する方法、構成オプション、CRDT の例を含む CRDT の概要を提供します。
- *Using CRDTs* は、テーブルで CRDT を使用する方法を調査し、いくつかの構成オプションをレビューし、CRDT のいくつかの例とその動作をレビューします。
- *Operation-based and state-based CRDTs* は、オペレーションベースと状態ベースの CRDT の違いをレビューします。
- *CRDT Disk-space requirements* は、CRDT、特に状態ベースの CRDT のディスクサイズの考慮事項をカバーしています。
- *CRDTs vs conflict handling/reporting* は、CRDT での競合処理と報告がどのように動作するかについて説明します。
- *Resetting CRDT values* は、CRDT 値をリセットする際の課題について説明し、これを正常に行うためのガイダンスを提供します。
- *Implemented CRDTs* は、PGD で使用可能な 6 つの CRDT のそれぞれを実装例とともに詳細に説明します。

#### CRDTs Overview

#### CRDT の概要

競合のないレプリケートデータ型 CRDT は、従来の解決のように行の 1 つを破棄するのではなく、同時に変更された行からの値のマージをサポートします。各 CRDT タイプは、`bdr.crdt_handlers` カタログに追加される追加のコールバックを備えた別の PostgreSQL データタイプとして実装されます。マージプロセスは、適用側の PGD ライタ内で発生し、ユーザーアクションは必要ありません。

CRDT では、列レベルの競合解決の特別な問題 で説明しているように、テーブルで列レベルの競合解決を有効にする必要があります。

PostgreSQL の## CRDT

*Implemented CRDTs* は、 bdr の一部として bdr スキーマにインストールされます。便宜上、基本的な演算子+ 、# 、および! と多くの一般的な集計関数 min 、max 、sum 、および avg が pg_catalog で作成されます。したがって、search_path を調整せずに使用できます。

### Using CRDTs

#### テーブルで CRDT を使用する

CRDT を使用するには、 integer などの標準のビルトインデータ型ではなく、 CREATE/ALTER TABLE で特定のデータ型を使用する必要があります。たとえば、1 つの通常の整数カウンターと単一の行を含む次のテーブルを検討します。

#### 非 CRDT の例

```
CREATE TABLE non_crdt_example (
    id      integer             PRIMARY KEY,
    counter integer             NOT NULL DEFAULT 0
);

INSERT INTO non_crdt_example (id) VALUES (1);
```

2 つの異なるノードで次の SQL を同時に発行するとします。

```
UPDATE non_crdt_example
   SET counter = counter + 1    -- "reflexive" update
 WHERE id = 1;
```

両方の更新を適用した後、次のクエリを使用して結果の値を確認できます。

```
SELECT * FROM non_crdt_example WHERE id = 1;
   id |   counter
 -----+-----------
    1 |         1
(1 row)
```

このコードは、 update_if_newer 競合リゾルバーが原因でインクリメントの 1 つが失われたことを示しています。

#### CRDT の例

代わりに CRDT カウンターデータ型を使用するには、次の手順を実行します。

テーブルを作成しますが、カウンターデータ型として CRDT bdr.crdt_gcounter を使用します。

```
CREATE TABLE crdt_example (
    id      integer             PRIMARY KEY,
    counter bdr.crdt_gcounter   NOT NULL DEFAULT 0
);
```

列レベルの競合解決のためにテーブルを構成します。

```
ALTER TABLE crdt_example REPLICA IDENTITY FULL;

SELECT bdr.alter_table_conflict_detection(crdt_example,
            column_modify_timestamp, cts);
```

次に、この例の値を含む行を挿入します。

```
INSERT INTO crdt_example (id) VALUES (1);
```

ここで、以前と同様に、2 つのノードで同じ SQL を同時に発行するとします。

```
UPDATE crdt_example
   SET counter = counter + 1    -- "reflexive" update
 WHERE id = 1;
```

変更が適用されると、カウンターが更新を同時に実行できることがわかります。

```
SELECT id, counter FROM crdt_example WHERE id = 1;
   id |   counter
 -----+-----------
    1 |         2
(1 row)
```

この例は、CRDT が他の方法で競合する非同期同時更新に直面しても、アキュムレータ列が動作できることを示しています。

### 構成オプション

bdr.crdt_raw_value 構成オプションは、クエリが現在の値または CRDT タイプの完全な内部状態を返すかどうかを決定します。デフォルトでは、現在の数値のみが返されます。true に設定されている場合、クエリーは完全な状態の表現を結果ます。特殊なハッシュ演算子# を使用して、特殊な演算子デフォルト動作を使用せずに現在の数値のみを要求できます。bdr.crdt_raw_value = on を使用して完全な状態がダンプされた場合、値は bdr.crdt_raw_value = on でのみリロードできます。

> **注釈**
>
> *bdr.crdt_raw_value* は、クライアントに返されたデータ、つまり選択リストの単純な列参照のみのフォーマットを適用します。クエリーの他の部分の列参照たとえば`WHERE` 句または select リストの式などは、 # 演算子を使用する必要がある場合があります。

### さまざまな種類の CRDT

crdt_gcounter タイプは、例が示すように、 x = x + 1 などの再帰的 UPDATE SQL でのみ動作する状態ベースの CRDT タイプの例です。

CRDT の別のクラスは、*delta CRDT* タイプです。これらは **オペレーションベースの** *CRDT* **タイプ** *CmCRDT* の特別なサブクラスです。

デルタ CRDT では、値の更新は、同じノード上の以前の値と比較されます。次に、変更は他のすべてのノードにデルタとして適用されます。

```
CREATE TABLE crdt_delta_example (
    id       integer            PRIMARY KEY,
    counter  bdr.crdt_delta_counter NOT NULL DEFAULT 0
);

ALTER TABLE crdt_delta_example REPLICA IDENTITY FULL;

SELECT bdr.alter_table_conflict_detection(crdt_delta_example,
            column_modify_timestamp, cts);

INSERT INTO crdt_delta_example (id) VALUES (1);
```

2 つのノードで次の SQL を同時に発行するとします。

```
UPDATE crdt_delta_example
   SET counter = 2          -- notice NOT counter = counter + 2
 WHERE id = 1;
```

両方の更新を適用した後、次のクエリを使用して結果の値を確認できます。

```
SELECT id, counter FROM crdt_delta_example WHERE id = 1;
   id | counter
 -----+---------
    1 |       4
(1 row)
```

通常の integer 列の場合、結果は 2 になります。ただし、デルタ CRDT カウンターで行を更新する場合、古い行バージョンで開始し、新しい行バージョンを作成し、両方をリモートノードに送信します。そこで、そこにあるバージョンたとえばローカルバージョンなどと比較します。標準 CRDT は NEW と LOCAL バージョンをマージしますが、デルタ CRDT は OLD と NEW バージョンを比較し、デルタをローカルバージョンに適用します。

### クエリのプランニングと最適化

重要な問題は、クエリのプランニングと最適化がこれらの新しいデータ型でどのように機能するかです。CRDT タイプは透過的に処理されます。 ANALYZE とオプティマイザーの両方が動作するため、他に何もせずに推定とクエリプランニングは正常に動作します。

### Operation-based and state-based CRDTs

#### オペレーションベースの CRDT タイプ CmCRDT

操作は明示的に転送されないが、リモートノードから受け取った古い行と新しい行から計算されるため、操作ベースの型の実装は簡単です。

現在、次のオペレーションベースの CRDT が実装されています。

- デルタカウンター `crdt_delta_counter <デルタカウンター *crdt_delta_counter*>` _ `bigint` カウンターインクリメント/デクリメント
- デルタサム `crdt_delta_sum <デルタサム *crdt_delta_sum*>` _ `numeric` sum インクリメント/デクリメント

これらの型は、デルタを計算する少しのコードで既存のデータ型を活用します。たとえば、`crdt_delta_counter` は `bigint` 上のドメインです。

このアプローチは、デルタを計算する方法が知られている型でのみ可能ですが、結果はシンプルで安価、スペースと CPU の両方の点で、いくつかの追加の利点があります。たとえば、基になるデータ型の演算子/構文を活用できます。

主な欠点は、非同期コンカレント環境ではこの値を確実にリセットできないことです。

> **注釈**
>
> カスタムデータタイプを作成し、状態と最後のオペレーションを保存することにより、より複雑なオペレーションベースのタイプを実装できます。すべての変更はデコードされて転送されるため、複数の操作は必要ありません。しかし、その時点で、主な利点単純さ、既存のデータ型の再利用などは、領域要件を除き、状態ベースの型たとえば、まだリセット機能と比較して利点を得ることなく失われます。ノードごとの状態は必要ありません。

#### 状態ベースの CRDT タイプ CvCRDT

状態ベースの型は、より複雑な内部状態を必要とするため、操作ベースの型のように通常のデータ型を直接使用できません。

現在、4 つの状態ベースの CRDT が実装されています。

- Grow 専用カウンター `crdt_gcounter <Grow 専用カウンター *crdt_gcounter*>` _ `bigint` カウンターインクリメント専用
- Grow-only sum `crdt_gsum <Grow-only sum *crdt_gsum*>` _ `numeric` sum/counter インクリメントのみ
- ポジティブネガカウンター `crdt_pncounter <ポジティブネガカウンター *crdt_pncounter*>` _ `bigint` カウンターインクリメント/デクリメント
- 正負の合計 `crdt_pnsum <正負の合計 `crdt_pnsum`>` _ `numeric` sum/counter インクリメント/デクリメント

内部状態には通常、ノードごとの情報が含まれるため、ディスク上のサイズは増加しますが、追加の利点が得られます。カスタムデータ型を実装する必要があるということは、より多くのコード in/out ファンクションと演算子を意味します。

利点は、値を確実にリセットする機能、変更が失われた場合のある程度の自己修復性正常に動作するクラスターでは発生しません、およびソースノード以外のノードから変更を受信できる機能です。

たとえば、ノード A で値が変更され、変更は B にレプリケートされますが、A と C 間のネットワークの問題により C にレプリケートされません。B が値を変更し、この変更が C にレプリケートされる場合、それには A からの元の変更。オペレーションベースの CRDT を使用すると、A-C ネットワーク接続が再び動作し始めるまで、ノード C は変更を受信しません。

CvCRDT の主な欠点は、状態ベースの *CRDT* ディスク領域要求 の点でコストが高いことです。クラスターから既に削除されたノードを含む、ノードごとに少しの情報が必要です。状態 varlena タイプにシリアル化の複雑な性質は、CPU 使用量の増加を意味します。

### CRDT Disk-space requirements

重要な考慮事項は、CRDT タイプ、特にディスク上のサイズに関連するオーバーヘッドです。

#### オペレーションベースの CRDT ディスク領域要求

*operation-based types* の場合、タイプは他のタイプの単なるドメインであるため、これは簡単です。ノードの数に関係なく、必要なディスク領域は同じです。

- `crdt_delta_counter` _ `bigint` と同じ 8 バイト
- `crdt_delta_sum` _ `numeric` と同じ変数、精度とスケールに依存します

操作ベースの CRDT タイプはノードごとの情報を保存しないため、ノードの数への依存関係はありません。

#### 状態ベースの CRDT ディスク領域要求

*state-based types* の場合、状況はより複雑です。すべてのタイプは可変長であり、基本的に `bytea` 列として保存され、ヘッダーと、値を変更した各ノードの特定の量のノードごとの情報で構成されています。

`bigint` バリアントの場合、おおよそのサイズを計算する式は次のとおりです。

- `crdt_gcounter` _ `32B (header) + N * 12B (per-node)`
- `crdt_pncounter` - _ `48B (header) + N * 20B (per-node)`

`N` は、この値を変更したノードの数を示します。

`numeric` バリアントの場合、ヘッダー部分とノードごとの部分の両方に `numeric` 可変長値が含まれるため、正確な式はありません。保持する必要があるこのような値の数を知るには

- `crdt_gsum`
- 固定 `20B (header) + N * 4B (per-node)`
- 変数 `(2 + N)` `numeric` 値
- `crdt_pnsum`
- 固定 `20B (header) + N *4B (per-node)`
- 変数 `(4 + 2* N)` `numeric` 値

> **注釈**
>
> 複数のノードで値が更新されない場合、クラスター内のノードの数は関係ありません。また、更新が同時実行されたかどうか、競合が発生したかどうかも関係ありません。

さらに、これらのノードのうちのいくつがクラスターから既に削除されたかは関係ありません。状態を圧縮する方法はまだありません。

### CRDTs vs conflict handling/reporting

#### CRDT タイプと競合処理

テーブルには CRDT 列と非 CRDT 列の両方を含めることができるため、ほとんどの列は非 CRDT であることが予想されるため、通常の競合解決と CRDT マージの両方を行う必要があります。

競合の解決は最初に発生し、保持するタプル applytuple と破棄するタプルを決定します。次にマージフェーズが発生し、CRDT 列のデータを破棄タプルから適用タプルにマージします。

> **注釈**
>
> この処理により、毎回マージが発生する必要があるため、単純な競合解決に比べて CRDT タイプのコストが多少高くなります。これは、競合解決で高速パスのいずれか現在のトランザクションで変更されたパスなどを使用できる場合でも、当てはまります。

#### CRDT タイプと競合レポート

デフォルトでは、検出された競合は個別に報告されます。CRDT タイプがなければ、競合解決は本質的に使用可能な情報の半分を破棄するため、これは理にかなっています構成に応じてローカルまたはリモート行。これにより、データの損失が発生します。

CRDT タイプを使用すると、何も捨てずに情報の両方の部分を組み合わせることができ、データ損失の問題を排除します。このアプローチにより、競合報告が不要になります。

このため、CRDT マージによって競合を完全に解決できる場合、競合レポートはスキップされます。各列は、次の 2 つの条件の少なくとも 1 つを満たす必要があります。

- ローカルタプルとリモートタプルの値が同じ NULL または等しい。
- CRDT データ型を使用するため、マージできます。

> **注釈**
>
> CRDT 列はないが、ローカル/リモートタプルのすべての値が等しい場合にも、競合レポートはスキップされます。

### Resetting CRDT values

CRDT 値のリセットは可能ですが、特別な処理が必要です。クラスターの非同期の性質は、実装方法に関係なく、さまざまなノードが変更ストリームのさまざまな場所でリセット操作を認識する可能性があることを意味します。別のノードが同時にリセットを開始する場合があります。つまり、他のノードからリセットを観察する前に。

言い換えれば、リセット操作を正しく動作させるには、通常の操作に対して可換である必要があります。単一ノードでうまく機能する可能性のある値をリセットする多くの単純な方法は、この理由で失敗します。

#### CRDT 値をリセットする際の課題

たとえば、値をリセットする最も簡単なアプローチは次のとおりです。

```
UPDATE crdt_table SET cnt = 0 WHERE id = 1;
```

状態ベースの CRDT では、これは機能しません。他のノードの状態を破棄しますが、ローカルでのみです。リモートノードのマージファンクションによって追加され、値の発散を引き起こし、最終的に他のノードでの変更によりそれを受信します。

オペレーションベースの CRDT では、更新が-cnt の減算として解釈されるため、これは機能するように見える場合があります。ただし、それは同時リセットがない場合にのみ機能します。2 つのノードが同時にリセットを実行しようとすると、デルタが 2 回適用され、負の値を取得しますリセットからは予想されません。

DELETE + INSERT をリセットとして使用できるように見えるかもしれませんが、このアプローチにはいくつかの弱点もあります。行が同じキーで再挿入された場合、他のノードからの変更に関して操作ストリームの同じ位置ですべてのノードがそれを見るという保証はありません。PGD は、コンカレントケースでデータ異常を引き起こす可能性があるため、同じ主キー値の再利用を特にお勧めしません。

#### CRDT 値のリセットを確実に処理する方法

状態ベースの CRDT タイプは、次のような特別な! 演算子を使用してリセットを確実に処理できます。

```
UPDATE tab SET counter = !counter WHERE ...;
```

「確実に」とは、値にマルチプルの同時リセットと発散という 2 つの問題がないことを意味します。

オペレーションベースの CRDT タイプは、複数の同時リセットが回避されるため、クォーラムコミット などの Eager Replication を実行するコミットスコープを使用してのみ確実にリセットできます。クォーラムコミットスコープを使用して、いずれかの種類の CRDT を特定の値に設定することもできます。

### Implemented CRDTs

現在、6 つの CRDT データ型が実装されています。

- Grow 専用のカウンターと合計
- ポジティブ/ネガティブカウンタと合計
- デルタカウンターとサム

カウンタと合計は、カウンタタイプが整数ベース bigint であるのに対し、sum タイプが 10 進数ベース numeric であることを除き、ほとんど同じ動作をします。

次のクエリーを使用して、現在実装されている CRDT データ型をリストできます。

```
SELECT n.nspname, t.typname
FROM bdr.crdt_handlers c
JOIN (pg_type t JOIN pg_namespace n ON t.typnamespace = n.oid)
  ON t.oid = c.crdt_type_id;
```

#### Grow 専用カウンター crdt_gcounter

- 非負の値のインクリメントのみをサポート value + int および counter + bigint 演算子。
- # 演算子を使用するか、bigint にキャストすることにより、カウンターの現在値を取得できます。
- counter = value のような単純な割り当てとは互換性がありません。これは、新しい値がアプリケーションのどこかで計算される場合の一般的なパターンです。
- ! 演算子 counter = !counter を使用してカウンターを単純にリセットできます。
- crdt_gcounter_to_text を使用して内部状態を検査できます。

```
CREATE TABLE crdt_test (
    id       INT PRIMARY KEY,
    cnt      bdr.crdt_gcounter NOT NULL DEFAULT 0
);

INSERT INTO crdt_test VALUES (1, 0);      -- initialized to 0
INSERT INTO crdt_test VALUES (2, 129824); -- initialized to 129824
INSERT INTO crdt_test VALUES (3, -4531);  -- error: negative value

- - enable CLCD on the table
ALTER TABLE crdt_test REPLICA IDENTITY FULL;
SELECT bdr.alter_table_conflict_detection(crdt_test, column_modify_timestamp, cts);

- - increment counters
UPDATE crdt_test SET cnt = cnt + 1 WHERE id = 1;
UPDATE crdt_test SET cnt = cnt + 120 WHERE id = 2;

- - error: minus operator not defined
UPDATE crdt_test SET cnt = cnt - 1 WHERE id = 1;

- - error: increment has to be non-negative
UPDATE crdt_test SET cnt = cnt + (-1) WHERE id = 1;

- - reset counter
UPDATE crdt_test SET cnt = !cnt WHERE id = 1;
```

(次のページに続く)

(前のページからの続き)

```
- - get current counter value
SELECT id, cnt::bigint, cnt FROM crdt_test;

- - show internal structure of counters
SELECT id, bdr.crdt_gcounter_to_text(cnt) FROM crdt_test;
```

### Grow-only sum crdt_gsum

- 負でない値のインクリメントのみをサポート sum + numeric 。
- # 演算子を使用するか、numeric にキャストすることにより、合計の現在値を取得できます。
- sum = value のような単純な割り当てとは互換性がありません。これは、新しい値がアプリケーションのどこかで計算される場合の一般的なパターンです。
- ! 演算子 sum = !sum を使用して合計を単純にリセットできます。
- crdt_gsum_to_text を使用して内部状態を検査できます。

```
CREATE TABLE crdt_test (
    id       INT PRIMARY KEY,
    gsum     bdr.crdt_gsum NOT NULL DEFAULT 0.0
);

INSERT INTO crdt_test VALUES (1, 0.0);      -- initialized to 0
INSERT INTO crdt_test VALUES (2, 1298.24); -- initialized to 1298.24
INSERT INTO crdt_test VALUES (3, -45.31);  -- error: negative value

- - enable CLCD on the table
ALTER TABLE crdt_test REPLICA IDENTITY FULL;
SELECT bdr.alter_table_conflict_detection(crdt_test, column_modify_timestamp, cts);

- - increment sum
UPDATE crdt_test SET gsum = gsum + 11.5 WHERE id = 1;
UPDATE crdt_test SET gsum = gsum + 120.33 WHERE id = 2;

- - error: minus operator not defined
UPDATE crdt_test SET gsum = gsum - 15.2 WHERE id = 1;

- - error: increment has to be non-negative
UPDATE crdt_test SET gsum = gsum + (-1.56) WHERE id = 1;

- - reset sum
UPDATE crdt_test SET gsum = !gsum WHERE id = 1;

- - get current sum value
SELECT id, gsum::numeric, gsum FROM crdt_test;
```

(次のページに続く)

(前のページからの続き)

```
- - show internal structure of sums
SELECT id, bdr.crdt_gsum_to_text(gsum) FROM crdt_test;
```

### ポジティブネガカウンター crdt_pncounter

- 正と負の値の両方のインクリメントをサポート counter + int および counter + bigint 演算子を介して。
- # 演算子を使用するか、bigint にキャストすることにより、カウンターの現在値を取得できます。
- counter = value のような単純な割り当てとは互換性がありません。これは、新しい値がアプリケーションのどこかで計算される場合の一般的なパターンです。
- ! 演算子 counter = !counter を使用してカウンターを単純にリセットできます。
- crdt_pncounter_to_text を使用して内部状態を検査できます。

```
CREATE TABLE crdt_test (
    id       INT PRIMARY KEY,
    cnt      bdr.crdt_pncounter NOT NULL DEFAULT 0
);

INSERT INTO crdt_test VALUES (1, 0);      -- initialized to 0
INSERT INTO crdt_test VALUES (2, 129824); -- initialized to 129824
INSERT INTO crdt_test VALUES (3, -4531);  -- initialized to -4531

- - enable CLCD on the table
ALTER TABLE crdt_test REPLICA IDENTITY FULL;
SELECT bdr.alter_table_conflict_detection(crdt_test, column_modify_timestamp, cts);

- - increment counters
UPDATE crdt_test SET cnt = cnt + 1      WHERE id = 1;
UPDATE crdt_test SET cnt = cnt + 120    WHERE id = 2;
UPDATE crdt_test SET cnt = cnt + (-244) WHERE id = 3;

- - decrement counters
UPDATE crdt_test SET cnt = cnt - 73    WHERE id = 1;
UPDATE crdt_test SET cnt = cnt - 19283 WHERE id = 2;
UPDATE crdt_test SET cnt = cnt - (-12) WHERE id = 3;

- - get current counter value
SELECT id, cnt::bigint, cnt FROM crdt_test;

- - show internal structure of counters
SELECT id, bdr.crdt_pncounter_to_text(cnt) FROM crdt_test;
```

(次のページに続く)

(前のページからの続き)

```
- - reset counter
UPDATE crdt_test SET cnt = !cnt WHERE id = 1;

- - get current counter value after the reset
SELECT id, cnt::bigint, cnt FROM crdt_test;
```

#### 正負の合計 crdt_pnsum

- sum + numeric を介して正と負の値の両方のインクリメントをサポートします。
- then # 演算子を使用するか、numeric にキャストすることにより、合計の現在値を取得できます。
- sum = value のような単純な割り当てとは互換性がありません。これは、新しい値がアプリケーションのどこかで計算される場合の一般的なパターンです。
- ! 演算子 sum = !sum を使用して合計を単純にリセットできます。
- crdt_pnsum_to_text を使用して内部状態を検査できます。

```
CREATE TABLE crdt_test (
    id       INT PRIMARY KEY,
    pnsum    bdr.crdt_pnsum NOT NULL DEFAULT 0
);

INSERT INTO crdt_test VALUES (1, 0);       -- initialized to 0
INSERT INTO crdt_test VALUES (2, 1298.24); -- initialized to 1298.24
INSERT INTO crdt_test VALUES (3, -45.31);  -- initialized to -45.31

- - enable CLCD on the table
ALTER TABLE crdt_test REPLICA IDENTITY FULL;
SELECT bdr.alter_table_conflict_detection(crdt_test, column_modify_timestamp, cts);

- - increment sums
UPDATE crdt_test SET pnsum = pnsum + 1.44     WHERE id = 1;
UPDATE crdt_test SET pnsum = pnsum + 12.20    WHERE id = 2;
UPDATE crdt_test SET pnsum = pnsum + (-24.34) WHERE id = 3;

- - decrement sums
UPDATE crdt_test SET pnsum = pnsum - 7.3      WHERE id = 1;
UPDATE crdt_test SET pnsum = pnsum - 192.83   WHERE id = 2;
UPDATE crdt_test SET pnsum = pnsum - (-12.22) WHERE id = 3;

- - get current sum value
SELECT id, pnsum::numeric, pnsum FROM crdt_test;

- - show internal structure of sum
SELECT id, bdr.crdt_pnsum_to_text(pnsum) FROM crdt_test;
```

(次のページに続く)

(前のページからの続き)

```
- - reset sum
UPDATE crdt_test SET pnsum = !pnsum WHERE id = 1;

- - get current sum value after the reset
SELECT id, pnsum::numeric, pnsum FROM crdt_test;
```

### デルタカウンター crdt_delta_counter

- bigint ドメインとして定義されているため、 bigint 列とまったく同様に動作します。
- 正と負の値の両方のインクリメントをサポートします。
- counter = value のような単純な割り当てと互換性があります。これは、新しい値がアプリケーションのどこかで計算される場合に一般的です。
- 値を確実にリセットする簡単な方法はありません。

```
CREATE TABLE crdt_test (
    id       INT PRIMARY KEY,
    cnt      bdr.crdt_delta_counter NOT NULL DEFAULT 0
);

INSERT INTO crdt_test VALUES (1, 0);      -- initialized to 0
INSERT INTO crdt_test VALUES (2, 129824); -- initialized to 129824
INSERT INTO crdt_test VALUES (3, -4531);  -- initialized to -4531

- - enable CLCD on the table
ALTER TABLE crdt_test REPLICA IDENTITY FULL;
SELECT bdr.alter_table_conflict_detection(crdt_test, column_modify_timestamp, cts);

- - increment counters
UPDATE crdt_test SET cnt = cnt + 1      WHERE id = 1;
UPDATE crdt_test SET cnt = cnt + 120    WHERE id = 2;
UPDATE crdt_test SET cnt = cnt + (-244) WHERE id = 3;

- - decrement counters
UPDATE crdt_test SET cnt = cnt - 73    WHERE id = 1;
UPDATE crdt_test SET cnt = cnt - 19283 WHERE id = 2;
UPDATE crdt_test SET cnt = cnt - (-12) WHERE id = 3;

- - get current counter value
SELECT id, cnt FROM crdt_test;
```

#### デルタサム crdt_delta_sum

- numeric ドメインとして定義されるため、 numeric 列とまったく同様に動作します。
- 正と負の値の両方のインクリメントをサポートします。
- sum = value のような単純な割り当てと互換性があります。これは、新しい値がアプリケーションのどこかで計算される場合に一般的です。
- 値を確実にリセットする簡単な方法はありません。

```
CREATE TABLE crdt_test (
    id       INT PRIMARY KEY,
    dsum     bdr.crdt_delta_sum NOT NULL DEFAULT 0
);

INSERT INTO crdt_test VALUES (1, 0);       -- initialized to 0
INSERT INTO crdt_test VALUES (2, 129.824); -- initialized to 129824
INSERT INTO crdt_test VALUES (3, -4.531);  -- initialized to -4531

- - enable CLCD on the table
ALTER TABLE crdt_test REPLICA IDENTITY FULL;
SELECT bdr.alter_table_conflict_detection(crdt_test, column_modify_timestamp, cts);

- - increment counters
UPDATE crdt_test SET dsum = dsum + 1.32   WHERE id = 1;
UPDATE crdt_test SET dsum = dsum + 12.01  WHERE id = 2;
UPDATE crdt_test SET dsum = dsum + (-2.4) WHERE id = 3;

- - decrement counters
UPDATE crdt_test SET dsum = dsum - 7.33   WHERE id = 1;
UPDATE crdt_test SET dsum = dsum - 19.83  WHERE id = 2;
UPDATE crdt_test SET dsum = dsum - (-1.2) WHERE id = 3;

- - get current counter value
SELECT id, cnt FROM crdt_test;
```

# 第7章　Advanced functionality

## 7.1 DDL replication

DDL はデータ定義言語の略で、データベースオブジェクトを作成、変更、および削除する SQL 言語のサブセットです。

PGD は自動 DDL レプリケーションを提供し、特定の DDL 変更を簡単にします。自動レプリケーションを使用すると、DDL 変更をすべてのノードに手動で配布し、一貫性を確認する必要はありません。

このセクションでは、PGD で DDL レプリケーションがどのように処理されるかについて説明します。

- *DDL overview* は、PGD の DDL レプリケーションでできることの一般的な概要を提供します。
- *DDL locking details* は、DDL レプリケーションがロックを使用して DDL を安全にレプリケートする方法を調べます。
- *Managing DDL with PGD replication* は、クラスターのスムーズな実行に過度に影響を与えないように、DDL 変更のインパクトを制限する理由と方法に関するベストプラクティスアドバイスを提供します。
- *DDL and role manipulation statements* は、クラスター内の複数のデータベースにわたるロールの操作に関する問題をノートします。
- *Workarounds for DDL restrictions* は、DDL レプリケーションが列、制約、タイプの変更などの制限を提示する可能性がある状況を処理するためのさまざまなオプションを提供します。
- *PGD functions that behave like DDL* は、DDL と同様に動作するため、同様の方法で動作し、同様の制限の対象となる PGD ファンクションについて詳しく説明します。

### 7.1.1 DDL overview

DDL はデータ定義言語の略で、データベースオブジェクトを作成、変更、および削除する SQL 言語のサブセットです。

#### レプリケートされた DDL

操作の利便性と正確性を維持するために、PGD はほとんどの DDL アクションをレプリケートしますが、次の例外を除きます。

- 一時的な関係
- 特定の DDL ステートメントほとんどの場合、長時間実行
- ロックコマンド `LOCK`

・テーブルメンテナンスコマンド VACUUM 、ANALYZE 、CLUSTER

- オートバキュームのアクション

・操作コマンド CHECKPOINT 、ALTER SYSTEM

- データベースに関するアクション

自動 DDL レプリケーションにより、DDL 変更をすべてのノードに手動で配布し、一貫性を確認する必要がなく、特定の DDL 変更が簡単になります。

デフォルトのレプリケーションセットでは、DDL はデフォルトですべてのノードにレプリケートされます。

### PostgreSQL との違い

DDL レプリケーションに関して、PGD はスタンドアロンの PostgreSQL とは大きく異なります。それを同じに扱うことは、PGD で最も一般的な問題です。

テーブルレプリケーションとの主な違いは、DDL レプリケーションが DDL の結果をレプリケートしないことです。代わりに、ステートメントをレプリケートします。これは、ほとんどの場合で非常にうまく機能しますが、DDL をすべてのノードで同様に実行する必要があるという要件が導入されます。より微妙な点は、DDL は、拡張機能ビルトインではなくによって導入されたデータ型を含む、すべてのデータ型固有のパラメーター設定に関して不変である必要があることです。たとえば、DDL ステートメントは、各ノードで使用されるデフォルトのエンコーディングで正しく実行される必要があります。

### PGD システムでの DDL の実行

PGD グループは、スタンドアロンの PostgreSQL サーバーと同じではありません。これは、集中ロックやトランザクションコーディネーターを使用しない非同期マルチマスターレプリケーションに基づいています。これは、DDL の実行時に重要な意味を持ちます。

パラレルに実行される DDL は、PGD で引き続き実行されます。DDL の実行は、実行時に各ノードの並列オペレーションに影響を与えるパラメーターを尊重するため、ノード間の設定の違いに気づく場合があります。

競合する DDL ステートメントの実行を防ぐことが重要です。それ以外の場合、DDL レプリケーションの競合が発生し、レプリケーションが停止します。

PGD は、 DDL の競合に対して次のレベルの保護を提供します。

`bdr.ddl_locking = 'all'` は最も厳密なオプションであり、DDL が任意のノードから同時に実行される可能性があり、正確性を保証したい場合に最適です。

`bdr.ddl_locking = 'leader'` は、書き込みリーダーのみをロックし、ノードの大部分がロック操作に参加する必要はありません。

`bdr.ddl_locking = 'auto'` デフォルトは、可能であれば `leader` ロックを自動的に選択し、それ以外の場合は `all` にフォールバックします。

`bdr.ddl_locking = 'dml'` は、任意の時点で 1 つのノードから DDL を実行する場合にのみ安全なオプションです。DDL を実行する場所を完全に制御できる場合にのみこの設定を使用します。単一のノードから DDL を実行すると、ノード間の競合が保証されません。ノード内の競合は PostgreSQL によって既に処理されています。

bdr.ddl_locking = 'off' は最も厳密でないオプションであり、一般的な使用は危険です。このオプションは、ロックを完全にスキップし、パフォーマンスのオーバーヘッドを回避するため、新しい空のデータベーススキーマを作成するときに役立つオプションです。

これらのオプションは、bdr_superuser、スーパーユーザー、または postgres.conf 構成ファイルでのみ設定できます。

bdr.replicate_ddl_command を使用する場合、そのファンクションに渡される DDL コマンドにのみ指定された bdr.ddl_locking 設定を使用して、3 番目の引数でこのパラメーターを直接設定できます。

#### DDL および混合 PostgreSQL バージョン

PGD は、クラスター内の異なるメジャー Postgres バージョン間の DDL レプリケーションをサポートしていません。クラスターは Postgres の同じメジャーバージョンを実行するため、ほとんどの場合、これは問題ではありません。ただし、これは、メジャーバージョンから別のメジャーバージョンへのクラスターのローリングアップグレードを実行する場合には当てはまりません。この場合、DDL レプリケーションは、すべてのノードが同じメジャーバージョンにアップグレードされるまでサポートされないため、使用しないでください。

スクリプトが DDL レプリケーションをトリガーする場合があるため、拡張機能を更新するときはアップグレードプロセスで特別な注意を払う必要があります。この場合、アップグレードが完了する前にスクリプトを実行する必要がある場合は、スクリプトの実行中に bdr.ddl_replication 設定を off に設定する必要があります。

### 7.1.2 DDL replication options

bdr.ddl_replication パラメーターは、レプリケーション動作を指定します。

bdr.ddl_replication = on がデフォルトです。この設定は、DDL をデフォルトのレプリケーションセット、デフォルトではすべてのノードを意味します。デフォルト以外のレプリケーションセットは、*Replication sets* がない限り DDL をレプリケートしません

彼らのために定義されました。

ファンクション bdr.replicate_ddl_command() を使用して、DDL を特定のレプリケーションセットにレプリケートすることもできます。この機能は、ノードがダウンしているときに DDL コマンドを実行する場合に役立ちます。ノードまたはレップセットのサブセット、たとえば、site1 のすべてのノードに存在するインデックスまたはパーティションが必要な場合にも役立ちます。

```
SELECT bdr.replicate_ddl_command(
                CREATE INDEX CONCURRENTLY ON foo (col7);,
                ARRAY[site1],     -- the replication sets
                all);              -- ddl_locking to apply
```

お勧めしませんが、bdr.ddl_replication 構成パラメーターを使用して、自動 DDL レプリケーションをスキップし、各ノードで手動で実行できます。

```
SET bdr.ddl_replication = off;
```

設定すると、PGD は実行された DDL コマンドのグローバルロックとレプリケーションの両方をスキップします。次に、すべてのノードで DDL を手動で実行する必要があります。

> **警告**
>
> グローバルロックを使用せずに各ノードで DDL を手動で実行すると、競合する DDL または DML が同時に実行される場合、PGD グループ全体のレプリケートが停止される場合があります。

bdr_superuser またはスーパーユーザーのみが `bdr.ddl_replication` パラメーターを設定できます。`postgres.conf` 構成ファイルで設定することもできます。

### 7.1.3 DDL locking details

グローバル DDL ロックとグローバル DML ロックの 2 種類のロックは、PGD でレプリケートされた DDL の正しさを強制します。

#### グローバル DDL ロック

グローバル DDL ロックは、`ddl_locking = 'all'` の場合にのみ使用されます。この種類のロックは、各 DDL ステートメントの実行中にクラスターで他の DDL が実行されないようにします。この動作は、一般的な場合に完全な正しさを保証しますが、多くの単純な場合には厳密すぎます。PGD は、スキーマ変更が行われるトランザクションで初めて DDL 操作のグローバルロックを取得します。これにより、クラスター内の DDL 実行トランザクションが効果的にシリアル化されます。つまり、DDL の実行中は、ノード上の他の接続は、別のテーブルに影響を与える場合でも、別の DDL コマンドを実行できません。

DDL 操作のロックを取得するために、DDL を実行する PGD ノードは PGD グループ内の他のノードに接続し、DDL を実行する排他的権限を付与するように求めます。

ロック要求は通常のレプリケーションストリームによって送信され、ノードは同様にレプリケーションストリームによって応答します。したがって、ノードまたは少なくともノードの大部分が、レプリケーション遅延をほとんど発生せずに実行されることが重要です。そうしないと、ノードが DDL ロックを取得するまでに時間がかかる場合があります。ノードの大部分が同意すると、DDL 実行が実行されます。

DDL ロックの順序は、Raft プロトコルを使用して決定されます。1 つのノードで実行された DDL ステートメントは、他のすべてのノードで同じシーケンスで実行されます。

DDL を実行しているノードがクラスター内で実行された以前のすべての DDL の効果を確認したことを確認するために、以前の DDL を実行したノードに追いつくまで待機します。現在の DDL を実行しているノードが、以前の DDL を実行したノードと比較してレプリケーションが遅れている場合、ロックの取得に時間がかかる場合があります。したがって、単一のノード、または他のノードで発生したレプリケーションの変更にほぼ追いついたノードから DDL を実行することをお勧めします。

グローバル DDL ロックは、データおよび監視ノードの大部分によって付与される必要があります。ここで、マジョリティは適格ノードの N/2+1 です。サブスクライバー専用ノードは参加する資格がありません。

### グローバル DML ロック

グローバル DML ロックまたはリレーション DML ロックとして知られるこの種類のロックは、ddl_locking = all または ddl_locking = dml のいずれか、および DDL ステートメントがフライト中の DML ステートメントが失敗する可能性がある場合に使用されます。これらの障害は、一意制約、チェック制約、NOT NULL 制約などの制約を追加または変更するときに発生する可能性があります。リレーション DML ロックは、一度に 1 つのリレーションにのみ影響します。これらのロックにより、変更がキュー内にあるときに DDL が実行されず、エラーでレプリケーションが停止する可能性があります。

テーブルのグローバル DML ロックを取得するために、DDL を実行する PGD ノードは PGD グループ内の他のすべてのノードに接続し、書き込みに対してテーブルをロックするように依頼し、そのテーブルの保留中の変更がすべてドレインされるまで待機します。すべてのノードが完全に追いつくと、DML ロックの発信者は自由にテーブルのスキーマ変更を実行し、他のノードにレプリケートできます。

グローバル DML ロックは、各ノードのテーブルの EXCLUSIVE LOCK を保持するため、実行中にそのテーブルに対する DML、他の DDL、VACUUM、および index コマンドをブロックします。これは、通常 EXCLUSIVE LOCK 以上を使用しないコマンドにグローバル DML ロックが保持されている場合にも当てはまります。

保留中の DML 操作がドレインされるまでに長い時間がかかる場合があり、レプリケーションが現在遅延している場合はさらに長くなります。これは、データの変更とは異なり、行表現と制約に影響を与えるスキーマの変更は、構成されているすべてのノードに到達でき、現在の書き込みレートに適度に追随しているときにのみ実行できることを意味します。ノードがダウンしているときにこのような DDL コマンドを実行する必要がある場合は、最初にダウンしたノードを構成から削除します。

ロックが付与される前に、資格のあるすべてのデータノードがグローバル DML ロックを付与することに同意する必要があります。Witness およびサブスクライバ専用ノードは参加する資格がありません。

DDL ステートメントがレプリケートされない場合、グローバルロックは取得されません。

*PGD システムでの DDL の実行* で説明しているように、bdr.ddl_locking パラメーターでロック動作を指定します。

- ddl_locking = all は、グローバル DDL ロックを取得し、必要に応じてリレーション DML ロックを取得します。
- ddl_locking = leader は、リーダーベースのグローバル DML ロックを有効にします。
- ddl_locking = auto は、可能であれば leader を選択し、それ以外の場合は all を選択します。
- ddl_locking = dml は、グローバル DDL ロックをスキップし、必要に応じてリレーション DML ロックを取得します。
- ddl_locking = off は、グローバル DDL ロックとリレーション DML ロックの両方をスキップします。

一部の PGD ファンクションは DDL 変更を行います。これらのファンクションには、DDL ロック動作が適用されます。これについては、各ファンクションのドキュメントに記載されています。

したがって、ddl_locking = dml は、競合する DDL が他のノードから実行されないことを保証できる場合にのみ安全です。この設定では、グローバル DDL ロックのみを必要とするステートメントは、グローバルロックをまったく使用しません。

`ddl_locking = off` は、DDL が実行されるデータベースオブジェクトで競合する DDL と競合する DML 操作がないことを保証できる場合にのみ安全です。ロックをオフにした後に問題が発生すると、データの進行中の変更が失われる可能性があります。ユーザーアプリケーションチームは、発生した問題を解決する必要があります。

場合によっては、同時に実行される DDL を適切にシリアル化できる場合があります。これらのシリアル化障害が発生した場合、DDL が再実行される場合があります。

DDL レプリケーションは、昇格するまでロジカルスタンバイノードでアクティブになりません。

一部の PGD 管理機能は DDL のように動作します。つまり、グローバルロックを取得しようとし、DDL レプリケーションがアクティブな場合、アクションがレプリケートされます。レプリケートファンクションの完全なリストは、*PGD functions that behave like DDL* にリストされています。

一時テーブルで実行される DDL は、グローバルロックを必要としません。

現在のトランザクションで作成されたオブジェクトの ALTER または DROP には、グローバル DML ロックは必要ありません。

グローバル DDL ロックとグローバル DML ロックのモニタリングは *Monitoring the Connection Manager* に表示されます。

### 7.1.4 Managing DDL with PGD replication

#### DDL のインパクトの最小化

DDL の影響を最小限に抑えることは、データベースにとって良い運用上のアドバイスです。PGD では、これらの点がさらに重要になります。

- DDL の影響を最小限に抑えるには、DDL を実行するトランザクションを短くします。大量の行変更と組み合わせず、長時間実行される外部キーまたはその他の制約の再チェックを避けてください。
- `ALTER TABLE` の場合、`ADD CONSTRAINT` を単独で使用するのではなく、`ADD CONSTRAINT NOT VALID` を使用し、その後に `VALIDATE CONSTRAINT` との別のトランザクションを使用します。 `VALIDATE CONSTRAINT` は、すべてのノードで再生されるまで待機するため、確認の受信に顕著な遅延が発生します。
- インデックスを作成するときは、可能な限り `CONCURRENTLY` オプションを使用します。

長時間実行 DDL を実行する別の方法は、DDL レプリケーションを無効にして、各ノードで個別に DDL ステートメントを実行することです。次の例に示すように、単一の SQL ステートメントを使用してこれを行うことができます。グローバルロックルールは引き続き適用されるため、このタイプの使用法で自分自身をロックアウトしないように注意してください。これは回避策です。

```
SELECT bdr.run_on_all_nodes($ddl$
        CREATE INDEX CONCURRENTLY index_a ON table_a(i);
$ddl$);
```

`CREATE INDEX CONCURRENTLY` はマルチトランザクションコマンドであるため、セッション全体で DDL レプリケーションを無効にする必要があることに注意して、`CREATE INDEX CONCURRENTLY` で bdr.run_on_all_nodes() テクニックを使用することをお勧めします。実行中の書き込みを防止するため、

実稼働システムでは CREATE INDEX を避けてください。 REINDEX は、AccessExclusiveLocks を保持するため、使用を避けてください。

代わりに、REINDEX CONCURRENTLY または reindexdb --concurrently を使用します。

次のようなコマンドラインユーティリティを使用する場合、DDL レプリケーションを無効にできます。

```
$ export PGOPTIONS="-c bdr.ddl_replication=off"
$ pg_restore --section=post-data
```

複数の DDL ステートメントは、個々のステートメントとして起動されるよりも単一のトランザクションに束ねたほうがメリットがある場合があるため、DDL ロックは 1 回のみを取ります。テーブルレベルのロックが通常の操作を妨げる場合、これは望ましくない場合があります。

DDL がシステムを長時間保持している場合は、 psql の **Control-C** または pg_cancel_backend() を使用して、元のノードで DDL を安全にキャンセルできます。他のノードから DDL ロックをキャンセルすることはできません。

オプションのグローバルロックタイムアウト設定を使用して、グローバルロックにかかる時間を制御できます。bdr.global_lock_timeout は、グローバルロックの取得がキャンセルされるまでの待機時間を制限します。bdr.global_lock_statement_timeout は、グローバルロックを保持するトランザクション内のステートメントの実行タイム長を制限し、bdr.global_lock_idle_timeout は、グローバルロックを保持するトランザクションの最大許可アイドル時間ステートメント間の時間を設定します。値をゼロに設定することにより、これらのタイムアウトをすべて無効にできます。

元のノードで DDL 操作がコミットされると、キャンセルまたは中止することはできません。PGD グループは、グローバルロックを確認した他のノードに正常に適用され、それらが再生を確認するまで待機する必要があります。このため、DDL トランザクションは短くて速いままにしてください。

### ダウンしたノードでの DDL の処理

グローバル DDL ロックを開始するノードがグローバルロック DDL または DML を取得した後にダウンした場合、ロックはアクティブなままです。タイムアウトが設定されていても、グローバルロックはタイムアウトしません。ノードが復旧した場合、ノードは保持しているすべてのグローバルロックを解放します。

長時間または無期限にダウンしたままになる場合は、PGD グループからノードを削除してグローバルロックを解放します。これが、bdr_superuser として SET コマンドを使用して緊急 DDL を実行し、bdr.ddl_locking 値を更新する理由の 1 つです。

グローバルロックを確認した後、それを取得するコマンドが実行される前に、他のノードのいずれかがダウンした場合、ノードがアップしているかのように、ロックを要求するそのコマンドの実行は続行されます。

前述のように、グローバル DDL ロックはノードの大部分のみが応答する必要があるため、大部分が実行され到達可能である限り、クラスターの一部がダウンしている場合に機能します。ただし、クラスター全体が利用できない限り DML ロックは取得できません。

グローバル DDL またはグローバル DML ロックでは、別のノードがダウンすると、コマンドは通常に続行され、ロックが解放されます。

#### ステートメント固有の DDL レプリケーションの問題

すべてのコマンドを自動的に複製できるわけではありません。bdr.ddl_replication をオフにして DDL レプリケーションをオフにしない限り、このようなコマンドは許可されません。

PGD は、データベースでアクティブなときに一部の DDL ステートメントの実行を防止します。これにより、正しくレプリケートできないステートメント、またはレプリケーションがまだサポートされていないステートメントを禁止することにより、システムの一貫性が保護されます。

ステートメントが PGD で許可されない場合、多くの場合、同じことを行う別の方法を見つけることができます。たとえば、揮発性のデフォルト値を含む列を追加する `ALTER TABLE` は実行できません。しかし、一般的には、機能する一連の独立した `ALTER TABLE` および `UPDATE` ステートメントとして言い換えることができます。

通常、サポートされていないステートメントは実行されず、`feature_not_supported` SQLSTATE `0A000` エラーが発生します。

一時オブジェクトを参照または依存する DDL は、PGD でレプリケートできず、DDL レプリケーションを有効にして実行するとエラーがスローされます。

### 7.1.5 DDL command handling matrix

次の表では、許可されるユーティリティまたは DDL コマンド、レプリケートされるもの、およびレプリケート時に使用されるグローバルロックの種類について説明します。

`ALTER TABLE` のようなより複雑なステートメントの場合、これらは実行されるサブコマンドによって異なる場合があります。このようなすべてのコマンドには、次の表の下に詳細な説明があります。

#### コマンドマトリックス

| Command | Allowed | Replicated | Lock |
|---|---|---|---|
| ALTER AGGREGATE | Y | Y | DDL |
| ALTER CAST | Y | Y | DDL |
| ALTER COLLATION | Y | Y | DDL |
| ALTER CONVERSION | Y | Y | DDL |
| ALTER DATABASE | Y | N | N |
| ALTER DATABASE LINK | Y | Y | DDL |
| ALTER DEFAULT PRIVILEGES | Y | Y | DDL |
| ALTER DIRECTORY | Y | Y | DDL |
| ALTER DOMAIN | Y | Y | DDL |
| ALTER EVENT TRIGGER | Y | Y | DDL |
| ALTER EXTENSION | Y | Y | DDL |
| ALTER FOREIGN DATA WRAPPER | Y | Y | DDL |

次のページに続く

表 1 – 前のページからの続き

| Command | Allowed | Replicated | Lock |
|---|---|---|---|
| ALTER FOREIGN TABLE | Y | Y | DDL |
| ALTER FUNCTION | Y | Y | DDL |
| ALTER INDEX | Y | Y | DDL |
| ALTER LANGUAGE | Y | Y | DDL |
| ALTER LARGE OBJECT | N | N | N |
| ALTER MATERIALIZED VIEW | Y | 少なくとも *1* 回の配信保証 | DML |
| ALTER OPERATOR | Y | Y | DDL |
| ALTER OPERATOR CLASS | Y | Y | DDL |
| ALTER OPERATOR FAMILY | Y | Y | DDL |
| ALTER PACKAGE | Y | Y | DDL |
| ALTER POLICY | Y | Y | DDL |
| ALTER PROCEDURE | Y | Y | DDL |
| ALTER PROFILE | Y | Y | *DDL locking details* |
| ALTER PUBLICATION | Y | Y | DDL |
| ALTER QUEUE | Y | Y | DDL |
| ALTER QUEUE TABLE | Y | Y | DDL |
| ALTER REDACTION POLICY | Y | Y | DDL |
| ALTER RESOURCE GROUP | Y | N | N |
| ALTER ROLE | Y | Y | DDL |
| ALTER ROUTINE | Y | Y | DDL |
| ALTER RULE | Y | Y | DDL |
| ALTER SCHEMA | Y | Y | DDL |
| ALTER SEQUENCE | *DDL locking details* | Y | DML |
| ALTER SERVER | Y | Y | DDL |
| ALTER SESSION | Y | N | N |
| ALTER STATISTICS | Y | Y | DDL |
| ALTER SUBSCRIPTION | Y | Y | DDL |
| ALTER SYNONYM | Y | Y | DDL |
| ALTER SYSTEM | Y | N | N |
| ALTER TABLE | *DDL locking details* | Y | *DDL locking details* |
| ALTER TABLESPACE | Y | Y | DDL |
| ALTER TEXT SEARCH CONFIGURATION | Y | Y | DDL |
| ALTER TEXT SEARCH DICTIONARY | Y | Y | DDL |

次のページに続く

表 1 – 前のページからの続き

| Command | Allowed | Replicated | Lock |
|---|---|---|---|
| ALTER TEXT SEARCH PARSER | Y | Y | DDL |
| ALTER TEXT SEARCH TEMPLATE | Y | Y | DDL |
| ALTER TRIGGER | Y | Y | DDL |
| ALTER TYPE | Y | Y | DDL |
| ALTER USER MAPPING | Y | Y | DDL |
| ALTER VIEW | Y | Y | DDL |
| ANALYZE | Y | N | N |
| BEGIN | Y | N | N |
| CHECKPOINT | Y | N | N |
| CLOSE | Y | N | N |
| CLOSE CURSOR | Y | N | N |
| CLOSE CURSOR ALL | Y | N | N |
| CLUSTER | Y | N | DML |
| COMMENT | Y | *DDL locking details* | DDL |
| COMMIT | Y | N | N |
| COMMIT PREPARED | Y | N | N |
| COPY | Y | N | N |
| COPY FROM | Y | N | N |
| CREATE ACCESS METHOD | Y | Y | DDL |
| CREATE AGGREGATE | Y | Y | DDL |
| CREATE CAST | Y | Y | DDL |
| CREATE COLLATION | Y | Y | DDL |
| CREATE CONSTRAINT | Y | Y | DDL |
| CREATE CONVERSION | Y | Y | DDL |
| CREATE DATABASE | Y | N | N |
| CREATE DATABASE LINK | Y | Y | DDL |
| CREATE DIRECTORY | Y | Y | DDL |
| CREATE DOMAIN | Y | Y | DDL |
| CREATE EVENT TRIGGER | Y | Y | DDL |
| CREATE EXTENSION | Y | Y | DDL |
| CREATE FOREIGN DATA WRAPPER | Y | Y | DDL |
| CREATE FOREIGN TABLE | Y | Y | DDL |
| CREATE FUNCTION | Y | Y | DDL |

次のページに続く

表 1 – 前のページからの続き

| Command | Allowed | Replicated | Lock |
|---|---|---|---|
| CREATE INDEX | Y | Y | DML |
| CREATE LANGUAGE | Y | Y | DDL |
| CREATE MATERIALIZED VIEW | Y | 少なくとも *1* 回の配信保証 | DML |
| CREATE OPERATOR | Y | Y | DDL |
| CREATE OPERATOR CLASS | Y | Y | DDL |
| CREATE OPERATOR FAMILY | Y | Y | DDL |
| CREATE PACKAGE | Y | Y | DDL |
| CREATE PACKAGE BODY | Y | Y | DDL |
| CREATE POLICY | Y | Y | DML |
| CREATE PROCEDURE | Y | Y | DDL |
| CREATE PROFILE | Y | Y | *DDL locking details* |
| CREATE PUBLICATION | Y | Y | DDL |
| CREATE QUEUE | Y | Y | DDL |
| CREATE QUEUE TABLE | Y | Y | DDL |
| CREATE REDACTION POLICY | Y | Y | DDL |
| CREATE RESOURCE GROUP | Y | N | N |
| CREATE ROLE | Y | Y | DDL |
| CREATE ROUTINE | Y | Y | DDL |
| CREATE RULE | Y | Y | DDL |
| CREATE SCHEMA | Y | Y | DDL |
| CREATE SEQUENCE | *DDL locking details* | Y | DDL |
| CREATE SERVER | Y | Y | DDL |
| CREATE STATISTICS | Y | Y | DDL |
| CREATE SUBSCRIPTION | Y | Y | DDL |
| CREATE SYNONYM | Y | Y | DDL |
| CREATE TABLE | Y | Y | DDL |
| CREATE TABLE AS | *DDL locking details* | Y | DDL |
| CREATE TABLESPACE | Y | Y | DDL |
| CREATE TEXT SEARCH CONFIGURATION | Y | Y | DDL |
| CREATE TEXT SEARCH DICTIONARY | Y | Y | DDL |

次のページに続く

表 1 – 前のページからの続き

| Command | Allowed | Replicated | Lock |
|---|---|---|---|
| CREATE TEXT SEARCH PARSER | Y | Y | DDL |
| CREATE TEXT SEARCH TEMPLATE | Y | Y | DDL |
| CREATE TRANSFORM | Y | Y | DDL |
| CREATE TRIGGER | Y | Y | DDL |
| CREATE TYPE | Y | Y | DDL |
| CREATE TYPE BODY | Y | Y | DDL |
| CREATE USER MAPPING | Y | Y | DDL |
| CREATE VIEW | Y | Y | DDL |
| DEALLOCATE | Y | N | N |
| DEALLOCATE ALL | Y | N | N |
| DECLARE CURSOR | Y | N | N |
| DISCARD | Y | N | N |
| DISCARD ALL | Y | N | N |
| DISCARD PLANS | Y | N | N |
| DISCARD SEQUENCES | Y | N | N |
| DISCARD TEMP | Y | N | N |
| DO | Y | N | N |
| DROP ACCESS METHOD | Y | Y | DDL |
| DROP AGGREGATE | Y | Y | DDL |
| DROP CAST | Y | Y | DDL |
| DROP COLLATION | Y | Y | DDL |
| DROP CONSTRAINT | Y | Y | DDL |
| DROP CONVERSION | Y | Y | DDL |
| DROP DATABASE | Y | N | N |
| DROP DATABASE LINK | Y | Y | DDL |
| DROP DIRECTORY | Y | Y | DDL |
| DROP DOMAIN | Y | Y | DDL |
| DROP EVENT TRIGGER | Y | Y | DDL |
| DROP EXTENSION | Y | Y | DDL |
| DROP FOREIGN DATA WRAPPER | Y | Y | DDL |
| DROP FOREIGN TABLE | Y | Y | DDL |
| DROP FUNCTION | Y | Y | DDL |
| DROP INDEX | Y | Y | DDL |
| DROP LANGUAGE | Y | Y | DDL |

次のページに続く

表 1 – 前のページからの続き

| Command | Allowed | Replicated | Lock |
|---|---|---|---|
| DROP MATERIALIZED VIEW | Y | 少なくとも *1* 回の配信保証 | DML |
| DROP OPERATOR | Y | Y | DDL |
| DROP OPERATOR CLASS | Y | Y | DDL |
| DROP OPERATOR FAMILY | Y | Y | DDL |
| DROP OWNED | Y | Y | DDL |
| DROP PACKAGE | Y | Y | DDL |
| DROP PACKAGE BODY | Y | Y | DDL |
| DROP POLICY | Y | Y | DDL |
| DROP PROCEDURE | Y | Y | DDL |
| DROP PROFILE | Y | Y | DDL |
| DROP PUBLICATION | Y | Y | DDL |
| DROP QUEUE | Y | Y | DDL |
| DROP QUEUE TABLE | Y | Y | DDL |
| DROP REDACTION POLICY | Y | Y | DDL |
| DROP RESOURCE GROUP | Y | N | N |
| DROP ROLE | Y | Y | DDL |
| DROP ROUTINE | Y | Y | DDL |
| DROP RULE | Y | Y | DDL |
| DROP SCHEMA | Y | Y | DDL |
| DROP SEQUENCE | Y | Y | DDL |
| DROP SERVER | Y | Y | DDL |
| DROP STATISTICS | Y | Y | DDL |
| DROP SUBSCRIPTION | Y | Y | DDL |
| DROP SYNONYM | Y | Y | DDL |
| DROP TABLE | Y | Y | DML |
| DROP TABLESPACE | Y | Y | DDL |
| DROP TEXT SEARCH CONFIGURATION | Y | Y | DDL |
| DROP TEXT SEARCH DICTIONARY | Y | Y | DDL |
| DROP TEXT SEARCH PARSER | Y | Y | DDL |
| DROP TEXT SEARCH TEMPLATE | Y | Y | DDL |
| DROP TRANSFORM | Y | Y | DDL |
| DROP TRIGGER | Y | Y | DDL |
| DROP TYPE | Y | Y | DDL |
| DROP TYPE BODY | Y | Y | DDL |

次のページに続く

表 1 – 前のページからの続き

| Command | Allowed | Replicated | Lock |
|---|---|---|---|
| DROP USER MAPPING | Y | Y | DDL |
| DROP VIEW | Y | Y | DDL |
| EXECUTE | Y | N | N |
| EXPLAIN | Y | *DDL locking details* | *DDL locking details* |
| FETCH | Y | N | N |
| GRANT | Y | *DDL locking details* | DDL |
| GRANT ROLE | Y | Y | DDL |
| IMPORT FOREIGN SCHEMA | Y | Y | DDL |
| LISTEN | Y | N | N |
| LOAD | Y | N | N |
| LOAD ROW DATA | Y | Y | DDL |
| LOCK TABLE | Y | N | *DDL locking details* |
| MOVE | Y | N | N |
| NOTIFY | Y | N | N |
| PREPARE | Y | N | N |
| PREPARE TRANSACTION | Y | N | N |
| REASSIGN OWNED | Y | Y | DDL |
| REFRESH MATERIALIZED VIEW | Y | 少なくとも *1* 回の配信保証 | DML |
| REINDEX | Y | Y | DDL |
| RELEASE | Y | N | N |
| RESET | Y | N | N |
| REVOKE | Y | *DDL locking details* | DDL |
| REVOKE ROLE | Y | Y | DDL |
| ROLLBACK | Y | N | N |
| ROLLBACK PREPARED | Y | N | N |
| SAVEPOINT | Y | N | N |
| SECURITY LABEL | Y | *DDL locking details* | DDL |
| SELECT INTO | *DDL locking details* | Y | DDL |
| SET | Y | N | N |
| SET CONSTRAINTS | Y | N | N |
| SHOW | Y | N | N |
| START TRANSACTION | Y | N | N |
| TRUNCATE TABLE | Y | *DDL locking details* | *DDL locking details* |
| UNLISTEN | Y | N | N |
| VACUUM | Y | N | N |

### コマンドに関する注意事項

#### ALTER SEQUENCE

PGD 6.3.0 以降、ALTER SEQUENCE が完全にサポートされています。

PGD 6.2.0 は、galloc シーケンスに RENAME および SET SCHEMA 機能を導入しましたが、この機能は、6.2.0 へのアップグレード後に 6.2.0 より前の PGD バージョンで作成されたシーケンスでは動作しませんでした。失敗の理由は、PGD 6.2.0 が galloc シーケンスを追跡するためのグローバル UUID 識別子を導入したことです。これはシーケンス作成時に生成されましたが、以前の PGD バージョンで作成されたシーケンスには欠落していました。

PGD 6.3.0 では、BDR マネージャーがシーケンスの UUID が欠落している場合は生成し、以前の PGD バージョンで作成されたシーケンスの機能を有効にするため、その制限を削除します。

#### ALTER TABLE

通常、 ALTER TABLE コマンドは許可されます。ただし、いくつかのサブコマンドはサポートされていません。

#### ALTER TABLE は許可されないコマンド

ALTER TABLE の一部のバリアントは、現在 PGD ノードで使用できません。

- ALTER COLUMN ... SET STORAGE external _ 列がテーブルのレプリカ ID の列の 1 つである場合、拒否されます。コマンドが安全であることが確認できる場合は、bdr.permit_unsafe_commands を使用してこの動作をオーバーライドできます。
- RENAME _ 自動パーティションテーブルの名前を変更することはできません。
- SET SCHEMA _ 自動パーティションテーブルのスキーマは設定できません。
- ALTER TABLE ... ADD FOREIGN KEY _ 現在のユーザーが参照されるテーブルを読み取る権限がない場合、または参照されるテーブルで現在のユーザーがバイパスできない RLS 制限が有効になっている場合、サポートされません。

次の例は、timestamp 型の定数値を timestamptz 型の列に追加しようとするため、失敗します。 timestamp と timestamptz 間のキャストは、セッションのタイムゾーンに依存するため、不変ではありません。

```
ALTER TABLE foo
  ADD expiry_date timestamptz DEFAULT timestamp 2100-01-01 00:00:00 NOT NULL;
```

DML ロックを取得せずに、 CHECK や FOREIGN KEY 制約などの特定のタイプの制約を追加できます。ただし、これには、最初に NOT VALID 制約を作成し、次に ALTER TABLE ... VALIDATE CONSTRAINT コマンドを使用して別のトランザクションで制約を検証する 2 段階のプロセスが必要です。コンストレインの追加 を参照

詳細については、

### ALTER TABLE ロック

ALTER TABLE の次のバリアントは、DDL ロックのみを取得し、DML ロックは取得しません。

- ALTER TABLE ... ADD COLUMN ... (immutable) DEFAULT
- ALTER TABLE ... ALTER COLUMN ... SET DEFAULT expression
- ALTER TABLE ... ALTER COLUMN ... DROP DEFAULT
- ALTER TABLE ... ALTER COLUMN ... TYPE リライトが必要ない場合
- ALTER TABLE ... ALTER COLUMN ... SET STATISTICS
- ALTER TABLE ... VALIDATE CONSTRAINT
- ALTER TABLE ... ATTACH PARTITION
- ALTER TABLE ... DETACH PARTITION
- ALTER TABLE ... ENABLE TRIGGER ENABLE REPLICA TRIGGER は引き続き DML ロックを取得します
- ALTER TABLE ... CLUSTER ON
- ALTER TABLE ... SET WITHOUT CLUSTER
- ALTER TABLE ... SET ( storage_parameter = value [, ... ] )
- ALTER TABLE ... RESET ( storage_parameter = [, ... ] )
- ALTER TABLE ... OWNER TO

ALTER TABLE の他のすべてのバリアントは、変更されるテーブルで DML ロックを取得します。ALTER TABLE の一部のバリアントには、以下に説明する制限があります。

### ALTER TABLE の例

さまざまなタイプの ALTER TABLE ... ALTER COLUMN TYPE ATCT 操作が可能です。一部の ATCT 操作は、基になる列タイプのメタデータのみを更新し、基になるテーブルデータの書き換えは必要ありません。これは通常、既存の列タイプとターゲットタイプがバイナリ強制可能な場合に当てはまります。

```
CREATE TABLE foo (id BIGINT PRIMARY KEY, description VARCHAR(20));
ALTER TABLE foo ALTER COLUMN description TYPE VARCHAR(128);
```

ただし、VARCHAR(128) から VARCHAR(20) に変更されたコマンドを元に戻すためにこの変更を行うことは、バイナリ強制不可であり、テーブルの書き換えが発生します。

```
CREATE TABLE sample (col1 BIGINT PRIMARY KEY, col2 VARCHAR(128), col3 INT);
ALTER TABLE sample ALTER COLUMN col2 TYPE VARCHAR(256);
```

バイナリの強制性のため、列タイプを VARCHAR または TEXT データ型に変更することもできます。繰り返しますが、これは基になる列タイプの単なるメタデータの更新です。

```
ALTER TABLE sample ALTER COLUMN col2 TYPE VARCHAR;
ALTER TABLE sample ALTER COLUMN col2 TYPE TEXT;
```

### TABLE REWRITE の例

ユーザーは列のデータ型を変更したり、ヌル以外のデフォルトで列を追加したり、メンテナンスジョブを実行したりする場合があります。これらの操作は、テーブルの書き換えをトリガーする可能性がある PostgreSQL の DDL コマンドでサポートされています。テーブル書き換え操作には、ディスク上にテーブルの新しいバージョンの作成が含まれます。CLUSTER や VACUUM などの DDL 操作は、テーブル定義を変更せずにテーブルの書き換えを直接発生させます。ALTER TABLE…ALTER COLUMN TYPE のような DDL 操作は、列を変更し、テーブルの書き換えをトリガーする場合があります。既存のすべてのデータを古いバージョンから新しいものにコピーし、新しいバージョンの準備ができて一貫性があると、古いテーブルを新しいものに置き換えます。

典型的なテーブル書き換え ATCT 操作の例

次のテーブルを検討します。

```
CREATE TABLE foo(
    c1 int,
    c2 int,
    c3 int,
    name text,
    time_now timestamptz,
    dollar numeric,
    indian_currency int GENERATED ALWAYS AS (dollar * 90.50) STORED
);
```

そのテーブルを指定すると、次の ALTER TABLE コマンドはテーブルの書き換えをトリガーします。

列のデータ型を変更します。

```
ALTER TABLE foo ALTER c1 TYPE bigint;
```

USING 句を使用して列タイプを互換性のないタイプに変更します。

```
ALTER TABLE foo ALTER COLUMN name TYPE int USING name::integer;
```

揮発性列タイプ

```
ALTER TABLE foo ALTER COLUMN time_now TYPE timestamp;
```

変更可能なデフォルトを使用して列を追加します。

```
ALTER TABLE foo ADD COLUMN time_clock timestamptz NOT NULL DEFAULT clock_timestamp();
```

生成された列

```
ALTER TABLE foo ALTER COLUMN indian_currency TYPE numeric;
```

単一のトランザクションブロックで複数の ALTER TABLE DDL 操作を実行します。

```
BEGIN;
ALTER TABLE foo ALTER c2 TYPE bigint;
ALTER TABLE foo ALTER c3 TYPE bigint;
COMMIT;
```

PGD クラスターでは、これには書き換えが含まれるため、書き換えアクティビティは DML タイプのグローバルリーダーロックを取得し、したがって、クラスター内のすべてのグループリーダーが利用できる必要があります。

!!!注 提供されている ALTER TABLE 例は、使用可能な ATCT 操作の完全なリストではありません。

### ALTER TYPE

ALTER TYPE はレプリケートされますが、PostgreSQL はこれらの依存関係を記録しないため、グローバル DML ロックはそのデータ型を使用するすべてのテーブルに適用されません。*Workarounds for DDL restrictions* を参照してください。

### コメントオン

COMMENT ON のすべてのバリアントが許可されますが、COMMENT ON TABLESPACE/DATABASE/LARGE OBJECT はレプリケートされません。

### CREATE PROFILE または ALTER PROFILE

ファンクションが SECURITY DEFINER の場合、プロファイルに関連付けられている PASSWORD_VERIFY_FUNCTION は IMMUTABLE である必要があります。このような CREATE PROFILE または ALTER PROFILE コマンドはレプリケートされますが、このプロファイルを使用する後続の CREATE USER または ALTER USER コマンドは、 writer ワーカーが cannot change current role within security-restricted operation エラーをスローするため、レプリケーションを中断します。

### CREATE SEQUENCE

一般に CREATE SEQUENCE がサポートされていますが、グローバルシーケンスを使用する場合、一部のオプションは効果がありません。

### CREATE TABLE AS と SELECT INTO

CREATE TABLE AS および SELECT INTO は、すべてのサブコマンドも許可される場合にのみ許可されます。

### EXPLAIN

一般に EXPLAIN は許可されますが、 EXPLAIN ANALYZE はデータベースに副作用を与える可能性があるため、いくつかの制限があります。

#### EXPLAIN ANALYZE レプリケーション

EXPLAIN ANALYZE は、分析されたステートメントのレプリケーションルールに従います。

#### EXPLAIN ANALYZE ロック

EXPLAIN ANALYZE は、分析されたステートメントのロックルールに従います。

#### GRANT と REVOKE

通常、GRANT および REVOKE ステートメントはサポートされていますが、GRANT/REVOKE ON TABLESPACE/LARGE OBJECT はレプリケートされません。

#### ロックテーブル

LOCK TABLE はレプリケートされませんが、bdr.lock_table_locking が on に設定されている場合、グローバル DML ロックを取得する場合があります。

bdr.global_lock_table() ファンクションを使用して、グローバル DML ロックを明示的に要求することもできます。

#### マテリアライズドビュー

CREATE MATERIALIZED VIEW は、上流ノードにのみデータを入力します。データは複製されます。REFRESH MATERIALIZED VIEW も上流ノードでのみ実行されます。結果は下流ノードに複製されます。

マテリアライズドビューレプリケーションは、BDRPG、EDB Advanced Server 、または 2QPG でのみ利用できます。PGD のコミュニティ Postgres ビルドは、4 つのコマンドすべてを非レプリケートとして扱います。

#### セキュリティラベル

SECURITY LABEL のすべてのバリアントが許可されますが、SECURITY LABEL ON TABLESPACE/DATABASE/LARGE OBJECT はレプリケートされません。

#### TRUNCATE レプリケーション

TRUNCATE コマンドは、DDL ステートメントではなく DML としてレプリケートされます。テーブルの TRUNCATE がレプリケートされるかどうかは、影響を受ける各テーブルのレプリケーション設定によって異なります。

#### TRUNCATE ロック

TRUNCATE は他の DDL と同じ方法でレプリケートされませんが、bdr.truncate_locking が on に設定されている場合、グローバル DML ロックを取得できます。

### 7.1.6 DDL and role manipulation statements

ユーザーは PostgreSQL インスタンスのグローバルオブジェクトです。つまり、ユーザーは複数のデータベースにまたがりますが、PGD は個々のデータベースレベルで動作します。この動作のため、ロール操作ステートメントの処理には追加の検討が必要です。

PGD では、レプリケートされた DDL によって参照されるロールがすべてのノードに存在する必要があります。ロールの付与、パスワードなどは同じである必要はありませんが、存在する必要があります。

bdr.role_replication が有効デフォルトであり、ロール操作ステートメントが PGD 対応データベースで実行されている場合、PGD はロール操作ステートメントをレプリケートします。

ロール操作ステートメントには次のものが含まれます。

- CREATE ROLE
- ALTER ROLE
- DROP ROLE
- GRANT ROLE
- CREATE USER
- ALTER USER
- DROP USER
- CREATE GROUP
- ALTER GROUP
- DROP GROUP

一般に、次のいずれかです。

- bdr.role_replication = off でシステムを構成し、Ansible、Puppet、Chef などの外部オーケストレーションツールまたは bdr.replicate_ddl_command() によって明示的にレプリケートされて、すべてのロール変更ユーザーとグループを展開します。
- PostgreSQL インスタンス上の 1 つの PGD 対応データベースに bdr.role_replication = on があるようにシステムを構成し、そのデータベースですべてのロール管理 DDL を実行します。

すべてのロール管理コマンドを 1 つのデータベースで実行することをお勧めします。

ロールレプリケーションがオフになっている場合、管理者は、あるノードで DDL が使用するロールが他のノードにも存在することを確認する必要があります。それ以外の場合、PGD apply は、他のノードでロールが作成されるまで、エラーでストールします。

### 7.1.7 Workarounds for DDL restrictions

PGD DDL オペレーション処理の制限の一部をワークアラウンドできます。多くの場合、操作をより小さな変更に分割すると、単一のステートメントとして許可されないか、過剰なロックが必要な目的の結果が生成される場合があります。

#### コンストレインの追加

DML ロックを必要とせずに CHECK および FOREIGN KEY 制約を追加できます。これには、2 段階のプロセスが含まれます。

- ALTER TABLE ... ADD CONSTRAINT ... NOT VALID

- `ALTER TABLE ... VALIDATE CONSTRAINT`

2 つの異なるトランザクションでこれらの手順を実行します。これらの手順は両方とも、テーブルでのみ DDL ロックを取得するため、1 つ以上のノードがダウンしている場合でも実行できます。ただし、コンストレインを検証するには、PGD は次のことを確認する必要があります。

- クラスター内のすべてのノードは `ADD CONSTRAINT` コマンドを参照します。
- コンストレインを検証するノードは、これらのノードに NOT VALID コンストレインを作成する前に、他のすべてのノードからレプリケーション変更を適用します。

したがって、新しいメカニズムでは、制約の検証中にすべてのノードがアップしている必要はありませんが、すべてのノードが `ALTER TABLE .. ADD CONSTRAINT ... NOT VALID` コマンドを適用し、十分に進行していることは必要です。PGD は、コンストレインを検証する前に、一貫した状態に到達するのを待機します。

新しい機能では、クラスターが Raft プロトコルバージョン 24 以降で実行されている必要があります。Raft プロトコルがまだアップグレードされていない場合、古いメカニズムが使用され、DML ロック要求が発生します。

### 7.1.8 PGD functions that behave like DDL

次の PGD 管理機能は DDL のように動作します。これは、DDL レプリケーションがアクティブで、DDL フィルター設定で許可されている場合、グローバルロックを取得しようとし、アクションがレプリケートされることを意味します。詳細については、各機能のドキュメントを参照してください。

レプリケーションセットの管理

- bdr.create_replication_set
- bdr.alter_replication_set
- bdr.drop_replication_set
- bdr.replication_set_add_table
- bdr.replication_set_remove_table
- bdr.replication_set_add_ddl_filter
- bdr.replication_set_remove_ddl_filter

競合管理

- bdr.alter_table_conflict_detection
- `bdr.column_timestamps_enable` 非推奨。`bdr.alter_table_conflict_detection()` を使用します
- `bdr.column_timestamps_disable` 非推奨。`bdr.alter_table_conflict_detection()` を使用します

シーケンス管理

- bdr.alter_sequence_set_kind

ストリームトリガー

- bdr.create_conflict_trigger
- bdr.create_transform_trigger
- bdr.drop_trigger

## 7.2 Decoding worker

PGD は、データが送信されるノードの数に関係なく、デコードを 1 回実行するデコードワーカープロセスを有効にするオプションを提供します。このオプションは、各 PGD ノードに新しいプロセス WAL デコーダーを導入します。接続ごとに 1 つの WAL 送信者プロセスがまだ存在しますが、これらのプロセスはデータの送受信のタスクを実行するだけです。まとめると、これらの変更は、大規模な PGD グループの CPU オーバーヘッドを削減し、WAL 送信プロセスが通信により多くの時間を費やすようになったため、より高いレプリケーションスループットを可能にします。

### 7.2.1 有効にする

`enable_wal_decoder` は、各 PGD グループのオプションであり、現在デフォルトで無効になっています。bdr.alter_node_group_option() を使用して、PGD グループのデコードワーカーを有効または無効にできます。

デコードワーカーが有効になっている場合、PGD は論理変更レコード LCR ファイルを保存して、デコードの間、およびすべてのサブスクライブノードがデータを受信したときに変更のバッファリングを許可します。LCR ファイルは、各ローカルノードのデータディレクトリの `pg_logical` ディレクトリの下に保存されます。レプリケーション遅延が増加すると、LCR ファイルの数とサイズが変化するため、このプロセスも監視が必要です。どの PGD ノードでも必要とされない LCR は、定期的にクリーニングされます。2 つの連続するクリーンアップ間の間隔は、

bdr.lcr_cleanup_interval によって制御されます。デフォルトは 3 分です。

bdr.lcr_cleanup_interval が 0 の場合、クリーンアップは無効になります。

### 7.2.2 無効化

無効にすると、論理デコードは、各ノードをサブスクライブする各ノードの WAL 送信プロセスによって実行されます。この場合、LCR ファイルは書き込まれません。

PGD グループでデコードワーカーが有効になっている場合でも、次の GUC はノードごとに LCR の生成と使用を制御します。デフォルトでは、これらは `false` です。 LCR を生成および使用するには、PGD グループのデコードワーカーを有効にし、PGD グループの各ノードでこれらの GUC を `true` に設定します。

- bdr.enable_wal_decoder _ `false` の場合、LCR を使用するすべての WAL 送信者は WAL を直接使用するように再起動します。`true` と PGD グループ構成を使用すると、デコードワーカープロセスが開始され、LCR および LCR を使用する WAL センダーが生成されます。
-

**bdr.receive_lcr _ サブスクライブノードで `true`**
すると、パブリッシャーノードの WAL 送信者に LCR を使用するように要求します利用可能な場合。

サブスクライバ専用ノードは、各ノードから変更を直接受信します。したがって、加入者専用ノードにサービスを提供する WAL 送信者は、LCR を使用できます。

LCR が生成されても、対応する WAL は、デコードワーカーが有効になっていない場合と同様に保持されます。将来的には、LCR が必要でない場合、LCR に対応する WAL を削除することが可能になる可能性があります。

参考までに、LCR ファイル名の最初の 24 文字は、WAL ファイル名のものと似ています。名前の最初の 8 文字は現在すべて「0」です。将来的には、WAL セグメントファイル名の最初の 8 文字と同様の TimeLineId を表すことが予想されます。次の 16 文字の名前のシーケンスは、WAL ストリームに対して LCR 変更を追跡するために使用される WAL セグメント番号に似ています。

ただし、論理変更は、それらが属しているトランザクションのコミット順序に従って並べ替えられます。したがって、LCR セグメント内のそれらの配置は、WAL セグメント内の対応する WAL の配置と一致しません。

最後の 16 文字のセットは、LCR セグメントのサブセグメント番号を表します。各 LCR ファイルはサブセグメントに対応します。LCR ファイルはバイナリで可変サイズです。`bdr.max_lcr_segment_file_size` デフォルトの 1GB を調整することにより、LCR ファイルの最大サイズを制御できます。

#### トランザクションストリーミングでの使用

トランザクションストリーミングでの使用 とデコードワーカーを同時に有効にすることが可能です。トランザクションストリーミングは、WAL 送信者がコミットする前に部分的なトランザクションを送信でき、レプリケーションラグを削減できることを意味します。WAL デコーダーが部分トランザクションのデコードをサポートするようになったため、デコードワーカーは部分トランザクションをデコードし、LCR ファイルに保存できます。次に、LCR ファイルを使用して、サブスクライバノードにトランザクションを適用します。これにより、遅延が減少することにより CPU 使用率が削減され、「.spill」ファイルが生成されないため、ディスク領域の使用量が削減されます。

WAL デコーダーは常にトランザクションを LCR にストリーミングしますが、ダウンストリーム要求に基づいて、WAL 送信者はトランザクションをストリーミングするか、通常の `BEGIN..COMMIT` シナリオを模倣します。

この機能をサポートするために、システムは追加のストリーミングファイルを作成します。これらのファイルは、`STR_TXN_<file-name-format>` および `CAS_TXN_<file-name-format>` で始まる名前があり、ストリーミングされた各トランザクションは独自のペアを作成します。

WAL デコーダーを使用したトランザクションストリーミングを有効にするには、bdr.alter_node_group_option を使用して PGD グループの `bdr.streaming_mode` セットを `default' に設定します。

## 7.3 CDC Failover support

### 7.3.1 背景

PGD の以前のバージョンでは、データベース内のデータに発生した論理変更のフィードを提供できるノードに論理レプリケーションスロットを作成できました。これらの論理レプリケーションスロットはノードに対してローカルであり、レプリケートされていません。特定のノードで変更をレプリケートするだけでなく、この動作は、クラスター内のノードフェールオーバーが発生した場合に課題を提示しました。その

シナリオでは、障害が発生したノードの論理レプリケーションのコンシューマには、別のノードに消費を継続するスロットのレプリカがありません。

この問題へのソリューションは、サブスクライバ専用ノードを中間として使用して設計できますが、論理レプリケーションのコストが大幅に増加します。

### 7.3.2 CDC フェイルオーバーのサポート

このニーズに対応するために、PGD は CDC フェールオーバーサポートを導入しました。これは、クラスター全体で自動論理スロットレプリケーションをアクティブにするオプションで有効な機能です。これにより、論理スロットのレプリケーションのコンシューマは、障害が発生したときに任意のノードから変更データを受信できます。

#### CDC フェールオーバーの仕組み

CDC フェールオーバーサポートが有効になっているノードで論理スロットが作成されると、スロットはクラスター全体でレプリケートされます。これは、クラスター内の任意のノードでスロットを使用できることを意味します。ノードに障害が発生すると、クラスター内の別のノードからスロットが消費される可能性があります。これにより、論理レプリケーションストリームを中断せずに継続できます。

ただし、スロットのコンシューマがクラスター内の別のノードに接続する場合、コンシューマが以前に持っていた接続は PGD によって閉じられます。この動作は、スロットがマルチプルのノードから同時に消費されないようにするためです。バックグラウンドで、PGD は Raft コンセンサスプロトコルを使用して、スロットが一度に 1 つのノードのみから消費されるようにしています。これは、一度に 1 つのスロットのみが消費されるという保証が、スプリットブレインシナリオでは成立しないことを意味します。

現在、CDC フェールオーバーのサポートは、トップグループオプションによって制御されるグローバルオプションです。`failover_slot_scope` トップグループオプションは現在、論理スロットのレプリケーションを無効にする `local` または `global` に設定できます。デフォルト。 `global` 設定は、PGD データベースで作成されたすべての非一時論理スロットのレプリケーションを有効にします。

一時論理スロットは、有効期間がそれらを作成したセッションにスコープされ、そのセッションが終了すると消失するため、レプリケートされません。

#### 少なくとも 1 回の配信保証

CDC フェールオーバーサポートは、コンシューマーがすべての変更を少なくとも 1 回受信できるようにするための手順を実行します。これは、配信が確認されるまでスロットを抑制することにより行われ、その時点でスロットは非同期方法ですべてのノードで進められます。スロットが消費されていたノードに障害が発生した場合、スロットは、コンシューマがクラスター内のノードに接続するまで保持されます。これにより、スロットが進行できます。

#### 1 回限りの配信

現在、正確に 1 回だけの配信を保証する方法はなく、使用するアプリケーションが、以前に完了したトランザクションの破棄を管理することを期待しています。

### 7.3.3 CDC フェールオーバーサポートの有効化

CDC フェールオーバーのサポートを有効にするには、SQL コマンドを実行し、次のパラメーターを使用して bdr.alter_node_group_option ファンクションを呼び出します。

```
select bdr.alter_node_group_option(<top-level group name>,
     failover_slot_scope,
     global);
```

<top-level group name> は、クラスターのトップレベルグループの名前に置き換えます。名前がわからない場合は、bdr.node_group の node_group_parent_id が 0 に等しいグループです。

名前がわからない場合は、bdr.node_group で 0 に等しい node_group_parent_id を持つグループです。次のものも使用できます。

```
SELECT bdr.alter_node_group_option(
     node_group_name,
     failover_slot_scope,
     global)
    from bdr.node_group
    where node_group_parent_id=0;
```

このコマンドにより、正しいトップレベルグループのオプションを設定できます。

CDC フェールオーバーが有効になったら、新しいグローバルにレプリケートされたスロットを作成するには、以下を使用できます。

```
SELECT pg_create_logical_replication_slot(myslot,
                                          test_decoding);
```

オプションが global に設定される前に作成された論理レプリケーションスロットはレプリケートされません。新しいスロットのみが複製されます。

フェールオーバースロットは、レプリケーション接続で CREATE_REPLICATION_SLOT コマンドを使用して作成することもできます。

フェイルオーバースロットのステータスは、 bdr.failover_replication_slots テーブルで追跡されます。

### 7.3.4 Postgres 17+での CDC フェールオーバーのサポート

Postgres 17 以降では、フェールオーバーのサポートが追加され、スタンバイを再開できるようになりました。このためには、failover という名前の pg_create_logical_replication_slot のオプションを使用します。この新しい設定では、 failover_slot_scope の設定に関係なく、 failover を true に設定する必要があります。

```
SELECT pg_create_logical_replication_slot(myslot,
                                          test_decoding,
                                          failover=>true);
```

### 7.3.5 一貫した初期スナップショットの取得

論理レプリケーションスロットが作成されると、一貫したスナップショットが Postgres によってエクスポートされます。このスナップショットを使用して、データの一貫した初期コピーを取得できます。PGD のフェイルオーバースロットメカニズムも同じ手順に従います。ただし、コンシューマは、スロットが最初に作成されたノードと同じノードからスナップショットを取得する必要があります。さらに、同じノードから初期レプリケーションを開始する必要があります。コンシューマがレプリケーションストリームを介して十分な変更を受信すると、フェールオーバースロットは `failover_safe` としてマークされます。スロットが `failover_safe` としてマークされると、コンシューマは PGD クラスター内の他のノードに安全にフェイルオーバーできます。ただし、他の考慮事項が適用されます。以下を参照してください。

スロットが `failover_safe` かどうかを確認するには、`bdr.failover_replication_slots` カタログを照会し、指定されたスロットの `failover_safe` 列の値を確認できます。

スロットが `failover_safe` とマークされる前に、コンシューマが他の PGD ノードに接続し、レプリケーションを開始しようとすると、PGD によって適切なエラーが発生します。

### 7.3.6 新しく参加したノードへのフェイルオーバー

新しいノードが PGD クラスターに参加すると、CDC フェールオーバースロットのデコードターゲットとして機能する準備がすぐに整わない場合があります。新しく参加したノードには、コンシューマがまだ使用していない変更をデコードするためのすべての WAL ファイルがない場合があります。このようなノードから消費すると、データの損失が発生する可能性があります。PGD は、レプリケーションの進行状況を内部で追跡し、安全なまで新しいノードがフェールオーバーターゲットになるのを防ぐことにより、このような状況を検出して防止します。コンシューマが、デコードターゲットとして機能する準備がまだ整っていないノードに接続しようとすると、適切なエラーが発生します。

### 7.3.7 オリジンごとの進行状況の追跡

トランザクションは、PGD クラスター内の任意のノードから発生できます。コンシューマが PGD ノードに接続し、トランザクションのデコードを開始すると、そのノードで発生したトランザクションと、クラスター内の他のノードからレプリケートされたトランザクションの変更を受信する場合があります。コンシューマは、このようなすべての PGD ノードまたはオリジンにわたってレプリケーションの進行状況を追跡し、重複したトランザクションが正しく処理されることを確認することが求められます。これを促進するために、Postgres-Extended および EnterpriseDB Advance Server の `test_decoding` プラグインが機能強化され、トランザクションの起点情報が含まれるようになりました。コンシューマは、論理レプリケーションの開始時に `include-origin` オプションを `on` に設定することにより、オリジン情報を受信することを選択できます。

以下は、起点情報を含む `test_decoding` プラグインのサンプル出力です。

```
BEGIN 1723654 (origin 2) (origin_name bdr_bdrdemo_bdrgroup_node2) (origin_lsn 0/
↪1D948910)
table public.pgbench_accounts: UPDATE: old-key: aid[integer]:39958 bid[integer]:1␣
↪abalance[integer]:0 filler[character]:
                                       new-tuple: aid[integer]:39958 bid[integer]:1␣
↪abalance[integer]:-1783 filler[character]:
table public.pgbench_tellers: UPDATE: old-key: tid[integer]:6 bid[integer]:1␣
```

(次のページに続く)

(前のページからの続き)

```
↪tbalance[integer]:0 new-tuple: tid[integer]:6 bid[integer]:1 tbalance[integer]:-
↪1783 filler[character]:null
table public.pgbench_branches: UPDATE: old-key: bid[integer]:1 bbalance[integer]:0␣
↪new-tuple: bid[integer]:1 bbalance[integer]:-1783 filler[character]:null
table public.pgbench_history: INSERT: tid[integer]:6 bid[integer]:1␣
↪aid[integer]:39958 delta[integer]:-1783 mtime[timestamp without time zone]:2025-01-
↪31 16:51:21.511571 filler[character]:null
COMMIT 1723654 (origin 2) (origin_name bdr_bdrdemo_bdrgroup_node2) (origin_lsn 0/
↪1D948910)
```

コンシューマはこの情報を利用して、オリジンごとの進行状況を追跡できます。

PGD は、bdr.logical_checkpoints カタログ内のすべてのノードにわたるレプリケーションの進行状況も記録し、コンシューマはカタログのデコードされた変更を受信し、その情報を使用してレプリケーションの進行状況を知ることができます。

```
⊠⊠ BEGIN 65720
id[name]:370098259-0-6056978 origin_node[oid]:370098259 origin_lsn[pg_lsn]:0/6056978␣
↪local_node[oid]:370098259 local_lsn[pg_lsn]:0/6056978 peer_count[integer]:2 peer_
↪nodes[oid[]]:{2228531844,4052927809} peer_lsns[pg_lsn[]]:{0/4836758,0/67F0AF8}
COMMIT 65720
```

この例では、ノード 370098259 がレプリケーションの進行状況を報告しています。コンシューマがこの変更レコードを受信すると、ノード 2228531844 および 4052927809 からそれぞれ 0/4836758 および 0/67F0AF8 までのすべてを受信したことを確認できます。

### 7.3.8 制限事項

CDC フェールオーバースロットのサポートには、特定の制限があります。

- CDC フェールオーバースロットのサポートには、 EDB Postgres DistributedPGD 5.7+の最新バージョンと Postgres Extended または EDB Postgres Advanced Server の最新のマイナーリリース 2025 年 2 月利用可能が必要です。
- CDC フェールオーバーのサポートはグローバルオプションであり、スロットごとに設定はできません。CDC フェールオーバーの有効なステータスの変更は、以前にプロビジョニングされたスロットに影響を与えないため、これを有効にして global に設定し、レプリケートスロットを作成してから無効にして local に設定して、単一のレプリケートスロットを作成することが可能です。
- CDC フェールオーバーのサポートは、一時的なスロットではサポートされていません。
- CDC フェールオーバーのサポートは、 failover オプションを false に設定して作成されたスロットではサポートされていません。
- CDC フェールオーバーのサポートは、EDB Postgres Advanced Server および EDB Postgres Extended Server でのみ動作します。コミュニティ Postgres インストールではサポートされていません。
- オプションが有効になっている場合、既存のスロットはフェールオーバースロットに変換されません。

- pg_logical_slot_get_changes() などの Postgres のビルトイン機能は使用できますが、スロットが他の場所でデコードされていないことは保証されず、クラスター全体のレプリケーションの進行状況を正確に更新できません。したがって、デコードされた変更を受信するファンクションに依存しないことをお勧めします。

## 7.4 Parallel Apply

### 7.4.1 パラレル適用とは何ですか?

Parallel apply は、PGD ノードがサブスクリプションごとにマルチプルのライターを使用できるようにする PGD の機能です。この動作は、一般にサブスクリプションのスループットを向上させ、レプリケーションパフォーマンスを向上させます。

#### パラレル適用の構成

2 つの変数は PGD のパラレル適用を制御します bdr.max_writers_per_subscription

デフォルト 8 および bdr.writers_per_subscription

デフォルト 2。

```
bdr.max_writers_per_subscription = 8
bdr.writers_per_subscription = 2
```

この構成では、各サブスクリプションに 2 つのライターを与えます。ただし、状況によっては、システムはサブスクリプションに最大 8 つのライターを割り当てる場合があります。

bdr.max_writers_per_subscription の変更

有効にするにはサーバーを再起動する必要があります。

bdr.writers_per_subscription を変更できます

再起動なしの特定のサブスクリプションの場合

1. bdr.alter_subscription_disable を使用してサブスクリプションを停止します。
2. 新しい値の設定。
3. bdr.alter_subscription_enable を使用してサブスクリプションを再開します。

ただし、最初に、 select * from bdr.subscription を使用してサブスクリプションの名前を確立します。この例では、サブスクリプション名は bdr_pgddb_bdrgroup_node2_node1 です。

```
SELECT bdr.alter_subscription_disable (bdr_pgddb_bdrgroup_node2_node1);

UPDATE bdr.subscription
SET num_writers = 4
WHERE sub_name = bdr_pgddb_bdrgroup_node2_node1;

SELECT bdr.alter_subscription_enable (bdr_pgddb_bdrgroup_node2_node1);
```

### Parallel apply を使用する場合

Parallel apply はデフォルトで常にオンになっており、ほとんどの操作では、オンのままにすることをお勧めします。

### モニタリング パラレル 適用

Parallel apply のデッドロックの緩和をサポートするために、PGD は bdr.stat_subscription に列を追加します。新しい列は `nprovisional_waits` 、`ntuple_waits` 、および `ncommmit_waits` です。これらは、Parallel apply が、以前はデッドロックであったものをどの程度適切に管理しているかを示すメトリックです。これらはシステム全体のパフォーマンスを反映するものではありません。

`nprovisional_waits` 値は、同時適用トランザクションによって実行されている同じタプルの操作の数を反映しています。これらは暫定的な待機であり、実際にはまだ待機していませんが、待機を開始する可能性があります。

タプルの書き込みが安全に適用できるまで待機する必要がある場合、`ntuple_waits` でカウントされます。コミットされる前に待機した完全に適用されたトランザクションは、`ncommit_waits` でカウントされます。

### パラレル適用の無効化

Parallel apply を無効にするには、 bdr.writers_per_subscription を 1 に設定します。

### デッドロックの軽減

Parallel apply が動作している場合、サブスクリプションからのトランザクションの変更は、マルチプルのライターによって書き込まれます。ただし、各ライタは、トランザクションの最終コミットが、オリジンノードで実行されたコミット順序に違反していないことを保証します。違反がある場合は、エラーが生成され、トランザクションをロールバックできます。

このメカニズムは、以前は、次のことがすべて true の場合、結果のエラーがデッドロックとして現れる可能性があることを意味しました。

- トランザクションはコミット保留中で、別のトランザクションが変更する必要がある行を変更します。
- 保留中のトランザクションが実行する前に、オリジンノードで他のトランザクションが実行されました。
- 保留中のトランザクションはロック要求を取り出しました。

さらに、エラーを処理すると、にかかった時間の組み合わせにより、レプリケーションラグが増加する可能性があります。

- デッドロックを検出するには
- クライアントがトランザクションをロールバックする場合
- すでに適用された変更の間接的なガベージコレクションの場合
- 作業をやり直すには

ここで、Parallel apply のデッドロックの緩和が役立ちます。いかなるトランザクションでも、Parallel apply は、現在のトランザクションが書き込む行タプルに既にスケジュールされているトランザクションを調べます。見つかった場合、行は、行に変更を適用する前に、他のトランザクションがコミットされるまで待機する必要があるとしてマークされます。このアプローチにより、行が正しい順序で書き込まれます。

#### パラレル適用のサポート

PGD 6 では、Parallel apply は CAMO および Quorum Commit で動作します。グループコミットとは互換性がありません。グループコミットを使用している場合、パラレル適用を無効にする必要があります。

## 7.5 Replication sets

レプリケーションセットは、PGD ノードがサブスクライブできるテーブルのグループです。レプリケーションセットを使用して、各ノードが他のノードの正確なコピーである通常の対称マルチマスタートポロジよりも複雑なレプリケーショントポロジを作成できます。

すべての PGD グループは、グループと同じ名前のレプリケーションセットを作成します。このレプリケーションセットは、すべてのユーザーテーブルと DDL レプリケーションに使用されるデフォルトのレプリケーションセットです。すべてのノードがそれにサブスクライブされます。つまり、デフォルトでは、すべてのユーザーテーブルがすべてのノード間でレプリケートされます。

### 7.5.1 レプリケーションセットを使用する

bdr.create_replication_set を使用して、挿入、更新、削除、または切り捨てアクションを含めるかどうかを指定して、レプリケーションセットを作成できます。1 つのオプションでは既存のテーブルをセットに追加でき、2 番目のオプションでは作成時にテーブルを追加するかどうかを定義します。

テーブルを手動で定義して、レプリケーションセットに追加または削除することもできます。

レプリケーションセットに含まれるテーブルは、ノードがクラスターに参加したときおよびその後に維持されます。

ノードが参加したら、レプリケーションセットからテーブルを削除できますが、再同期操作を使用して新しいテーブルを追加する必要があります。

デフォルトでは、新しく定義されたレプリケーションセットは DDL または PGD 管理ファンクション呼び出しをレプリケートしません。bdr.replication_set_add_ddl_filter を使用する

複製するコマンドを定義します。

PGD は、すべてのノードでレプリケーションセット定義を作成します。次に、bdr.alter_node_replication_sets を使用して各レプリケーションセットをパブリッシュまたはサブスクライブするように各ノードを定義できます。

ファンクションを使用して、これらの定義を後で変更したり、レプリケーションセットを削除したりできます。

> **注釈**
>
> 選択的レプリケーションにデフォルトのレプリケーションセットを使用しないでください。クラスター内の PGD ノードにあるデフォルトのレプリケーションセットは、 DDL レプリケーションおよび管理ファンクション呼び出しにもデフォルトで使用されるため、ドロップまたは変更しないでください。

### 7.5.2 パーティション分割テーブルの動作

PGD は、パーティション化テーブルを透過的にサポートします。つまり、パーティション化されたテーブルをレプリケーションセットに追加できます。

パーティションに関係する変更は、ダウンストリームにレプリケートされます。

> **注釈**
>
> パーティションがパーティションテーブルを介してレプリケートされる場合、パーティションで直接実行されるステートメントは、親テーブルで実行されたときと同じようにレプリケートされます。例外は `TRUNCATE` コマンドであり、これは常に影響を受けるテーブルまたはパーティションのリストを使用して複製します。

個々のパーティションをレプリケーションセットに追加できます。この場合、それらは通常のテーブルと同様に、つまり、ダウンストリームのパーティションと同じ名前のテーブルにレプリケートされます。パーティショニング定義がプロバイダーとサブスクライバーの両方で同じである場合、パーティショニングロジックを実行する必要がないため、この動作にはいくつかのパフォーマンスの利点があります。

> **注釈**
>
> ルートパーティションテーブルがレプリケーションセットの一部である場合、個々のパーティションのメンバーシップは無視されます。そのルートテーブルのメンバーシップのみが考慮されます。

### 7.5.3 外部キーを使用した動作

外部キー制約は、参照テーブルの各行が参照先テーブルの行と一致することを保証します。したがって、参照テーブルがレプリケーションセットのメンバーである場合、参照されるテーブルも同じレプリケーションセットのメンバーである必要があります。

PGD の現在のバージョンは、この条件をチェックまたは適用しません。レプリケーションセットにテーブルを追加するとき、データベース管理者は、外部キーによって参照されるすべてのテーブルも追加されていることを確認する必要があります。

次のクエリを使用して、この要件を満たさないすべての外部キーとレプリケーションセットをリストできます。参照テーブルはレプリケーションセットのメンバーですが、参照されるテーブルはレプリケーションセットのメンバーではありません。

```
SELECT t1.relname,
       t1.nspname,
       fk.conname,
       t1.set_name
  FROM bdr.tables AS t1
  JOIN pg_catalog.pg_constraint AS fk
    ON fk.conrelid = t1.relid
   AND fk.contype = f
 WHERE NOT EXISTS (
  SELECT *
    FROM bdr.tables AS t2
   WHERE t2.relid = fk.confrelid
     AND t2.set_name = t1.set_name
);
```

このクエリの出力は次のようになります。

```
 relname | nspname |  conname   | set_name
- --------+---------+------------+----------
 t2      | public  | t2_x_fkey | s2
(1 row)
```

この出力は、テーブル t2 がレプリケーションセット s2 のメンバーであるが、外部キー t2_x_fkey によって参照されるテーブルがそうでないことを意味します。

TRUNCATE CASCADE コマンドは、コマンドをレプリケートする前にレプリケーションセットのメンバーシップを考慮します。例

```
TRUNCATE table1 CASCADE;
```

これは、レプリケーションセットのみの一部であるすべてのテーブルでカスケードなしの TRUNCATE になります。

```
TRUNCATE table1, referencing_table1, referencing_table2 ...
```

### 7.5.4 レプリケーションセットのメンバーシップ

1 つ以上のレプリケーションセットにテーブルを追加または削除できます。そうすると、これらのテーブルの変更 DML のレプリケーションにのみ影響します。スキーマ変更 DDL は、DDL レプリケーションセットフィルターによって処理されます *DDL* レプリケーションフィルタリング を参照してください。

レプリケーションは、ノードレプリケーションセット構成を使用してレプリケーションセットのテーブルメンバーシップを使用して、レプリケートするアクションとそれらのレプリケート先のノードを決定します。決定は、すべてのメンバーシップとレプリケーションセットオプションの和集合を使用して行われます。テーブルが、INSERT アクションのみをレプリケートするレプリケーションセット A と、UPDATE アクションのみをレプリケートするレプリケーションセット B のメンバーであるとします。ターゲットノード

がレプリケーションセット A および B の両方にサブスクライブされている場合、INSERT および UPDATE アクションは両方ともレプリケートされます。

bdr.replication_set_add_table および bdr.replication_set_remove_table を使用してメンバーシップを制御できます。

### 7.5.5 レプリケーションセットのリスト

次のクエリーを使用して、既存のレプリケーションセットをリストできます。

```
SELECT set_name
FROM bdr.replication_sets;
```

このクエリを使用して、特定のレプリケーションセット内のすべてのテーブルをリストできます。

```
SELECT nspname, relname
FROM bdr.tables
WHERE set_name = myrepset;
```

外部キーを使用した動作 は、参照されるテーブルが参照テーブルと同じレプリケーションセットに含まれていないすべての外部キーをリストするクエリーを示しています。

次の SQL を使用して、現在のノードが発行およびサブスクライブするレプリケーションセットを表示します。

```
SELECT node_id,
       node_name,
       pub_repsets,
       sub_repsets
  FROM bdr.local_node_summary;
```

このコードは、次のような出力を生成します。

```
  node_id   | node_name |    pub_repsets     |   sub_repsets
- ----------+-----------+--------------------------------------
 1834550102 | s01db01   | {bdrglobal,bdrs01} | {bdrglobal,bdrs01}
(1 row)
```

クラスター内のすべてのノードに対して同じクエリーを実行するには、次のクエリーを使用できます。このアプローチでは、すべてのノードに関連付けられたレプリケーションセットを同時に取得します。

```
WITH node_repsets AS (
  SELECT jsonb_array_elements(
    bdr.run_on_all_nodes($$
        SELECT
          node_id,
          node_name,
          pub_repsets,
```

(次のページに続く)

(前のページからの続き)

```
          sub_repsets
        FROM bdr.local_node_summary;
    $$)::jsonb
  ) AS j
)
SELECT j->response->command_tuples->0->>node_id AS node_id,
       j->response->command_tuples->0->>node_name AS node_name,
       j->response->command_tuples->0->>pub_repsets AS pub_repsets,
       j->response->command_tuples->0->>sub_repsets AS sub_repsets
FROM node_repsets;
```

これは、例を示します。

```
  node_id   | node_name |    pub_repsets     |    sub_repsets
- ----------+-----------+--------------------+--------------------
 933864801  | s02db01   | {bdrglobal,bdrs02} | {bdrglobal,bdrs02}
 1834550102 | s01db01   | {bdrglobal,bdrs01} | {bdrglobal,bdrs01}
 3898940082 | s01db02   | {bdrglobal,bdrs01} | {bdrglobal,bdrs01}
 1102086297 | s02db02   | {bdrglobal,bdrs02} | {bdrglobal,bdrs02}
(4 rows)
```

### 7.5.6 DDL レプリケーションフィルタリング

デフォルトでは、サポートされているすべての DDL のレプリケーションは、デフォルトの PGD グループレプリケーションセットを介して発生します。このレプリケーションは、PGD グループと同じ名前の DDL フィルターを使用して実現されます。このフィルタは、PGD グループの作成時に、デフォルトの PGD グループレプリケーションセットに追加されます。

既存のすべてのレプリケーションセットの DDL レプリケーションフィルターを変更することにより、この動作を調整できます。これらのフィルターは、レプリケーションセットのテーブルメンバーシップとは独立しています。データ変更と同様に、各 DDL ステートメントは、複数のレプリケーションセットの複数のフィルターと一致する場合でも、1 回だけレプリケートされます。

次のクエリーを使用して既存の DDL フィルターをリストできます。これは、フィルターごとに、コマンドタグとロール名に適用された正規表現を示します。

```
SELECT * FROM bdr.ddl_replication;
```

bdr.replication_set_add_ddl_filter および bdr.replication_set_remove_ddl_filter を使用して、DDL フィルターを操作できます。これらは DDL であると考えられるため、DDL レプリケーションとグローバルロックの対象になります。

### 7.5.7 選択的レプリケーションの例

この例では、テーブルを特定のノードのグループに選択的にレプリケートするように EDB Postgres 分散を構成します。

#### クラスター構成

この例では、6 つのデータノード、data-a1 ～data-a3 および data-b1 ～data-b3 のクラスターが 2 つの場所にあることを前提としています。これらがメンバーである 2 つのロケーションは、region_a および region_b グループとして表されます。

また、お勧めのように、この例では記載されていませんが、 region-c に witness という名前の監視ノードがあります。クラスターの名前は sere です。

#### 応募要項

この例は、音楽作品の公演に出席した人々の意見を記録するアプリケーションで動作します。出席者用のテーブル、作品用のテーブル、そしてオピニオンテーブルがあります。オピニオンテーブルには、各出席者が見た各作品、どこで、いつ見たか、および作品をどのように採点したかが記録されます。データ規制のため、この例では、意見データは、意見が記録されたリージョンにのみ存在する必要があることを前提としています。

#### テーブルの作成

最初のステップは、適切なテーブルを作成することです。

```
CREATE TABLE attendee (
   id bigserial PRIMARY KEY,
   email text NOT NULL
);

CREATE TABLE work (
    id int PRIMARY KEY,
    title text NOT NULL,
    author text NOT NULL
);

CREATE TABLE opinion (
    id bigserial PRIMARY KEY,
    work_id int NOT NULL REFERENCES work(id),
    attendee_id bigint NOT NULL REFERENCES attendee(id),
    country text NOT NULL,
    day date NOT NULL,
    score int NOT NULL
);
```

### グループとレプリケーションセットの表示

デフォルトでは、 EDB Postgres 分散は、各テーブル全体を各すべてのノードにレプリケートするように構成されます。これは、レプリケーションセットを介して管理されます。

初期構成のデフォルトレプリケーションセットを表示するには、次のコマンドを実行します。

```
SELECT node_group_name, default_repset, parent_group_name
FROM bdr.node_group_summary;
```

```
 node_group_name | default_repset | parent_group_name
- ----------------+----------------+-------------------
 sere            | sere           |
 region_a        | region_a       | sere
 region_b        | region_b       | sere
 region_c        | region_c       | sere
```

出力には、トップレベルのグループ sere があり、 sere という名前のデフォルトのレプリケーションセットがあることがわかります。3 つのサブグループには、それぞれ、サブグループと同じ名前のレプリケーションセットがあります。region_a グループには、region_a デフォルトレプリケーションセットがあります。

デフォルトでは、既存のすべてのテーブルと新しいテーブルは、トップレベルグループのレプリケーションセットのメンバーになります。

### レプリケーションセットへのテーブルの追加

次の手順は、リージョンを表すグループに属するレプリケーションセットにテーブルを追加することです。前述のように、新しいテーブルはすべて sere レプリケーションセットに自動的に追加されます。次を実行して、それを確認できます。

```
SELECT relname, set_name FROM bdr.tables ORDER BY relname, set_name;
```

```
 relname  | set_name
- ---------+----------
 attendee | sere
 opinion  | sere
 work     | sere
(3 rows)
```

opinion テーブルは、region_a でのみ、別に region_b でのみレプリケートする必要があります。これを行うには、各リージョンのレプリカセットにテーブルを追加します。

```
SELECT bdr.replication_set_add_table(opinion, region_a);
SELECT bdr.replication_set_add_table(opinion, region_b);
```

しかし、opinion はまだ sere レプリケーションセットのメンバーであるため、これで完了ではありません。テーブルが複数のレプリケーションセットのメンバーである場合、各テーブルでレプリケートされま

す。ただし、各行は各ターゲットノードで 1 回のみレプリケートされるため、これはパフォーマンスに影響しません。opinion をすべてのノードでレプリケートしたくないため、トップレベルグループのレプリケーションセットから削除する必要があります。

```
SELECT bdr.replication_set_remove_table(opinion, sere);
```

これらの変更を確認できるようになりました。

```
SELECT relname, set_name FROM bdr.tables ORDER BY relname, set_name;
```

```
 relname  | set_name
- ---------+-------------------
 attendee | sere
 opinion  | region_a
 opinion  | region_b
 work     | sere
(4 rows)
```

このプロセスにより、目的の選択的レプリケーションが提供されます。行われたかどうかを確認するには、次の手順を使用してテストします。

### 選択的レプリケーションのテスト

最初にいくつかのテストデータ 2 つの作品と出席者を作成します。 data-a1 に直接接続して、次のコードを実行します。

```
INSERT INTO work VALUES (1, Aida, Verdi);
INSERT INTO work VALUES (2, Lohengrin, Wagner);
INSERT INTO attendee (email) VALUES (gv@example.com);
```

これらのテーブルにデータが追加されたため、外部キー制約に違反せずに opinion テーブルに挿入できます。

```
INSERT INTO opinion (work_id, attendee_id, country, day, score)
SELECT work.id, attendee.id, Italy, 1871-11-19, 3
  FROM work, attendee
 WHERE work.title = Lohengrin
   AND attendee.email = gv@example.com;
```

挿入が完了したら、同じノードでデータベースの内容を検証できます。

```
SELECT a.email
, o.country
, o.day
, w.title
, w.author
, o.score
```

(次のページに続く)

(前のページからの続き)

```
FROM opinion o
JOIN work w ON w.id = o.work_id
JOIN attendee a ON a.id = o.attendee_id;
```

```
    email       | country |    day     |   title   | author | score
- ---------------+---------+------------+-----------+--------+-------
 gv@example.com | Italy   | 1871-11-19 | Lohengrin | Wagner |     3
(1 row)
```

ここで、ノード data-a2 および data-a3 に接続し、同じクエリーを実行すると、同じ結果が得られます。データは region_a で複製されています。 data-b1 、data-b2 、または data-b3 に接続する場合、クエリーは行を返しません。これは、 attendee および work テーブルにデータが入力されていますが、選択する opinion 行がないためです。これは、 region_a での opinion のレプリケーションがそのリージョンでのみ発生するためです。

次に、 data-b1 に接続し、そこに意見を挿入します。

```
INSERT INTO attendee (email) VALUES (fb@example.com);

INSERT INTO opinion (work_id, attendee_id, country, day, score)
SELECT work.id, attendee.id, Germany, 1850-08-27, 9
  FROM work, attendee
 WHERE work.title = Lohengrin
   AND attendee.email = fb@example.com;
```

この意見は region_b でのみ複製されます。 data-b1 、data-b2 、および data-b3 では、次を実行できます。

```
SELECT a.email
, o.country
, o.day
, w.title
, w.author
, o.score
FROM opinion o
JOIN work w ON w.id = o.work_id
JOIN attendee a ON a.id = o.attendee_id;
```

```
    email       | country |    day     |   title   | author | score
- ---------------+---------+------------+-----------+--------+-------
 fb@example.com | Germany | 1850-08-27 | Lohengrin | Wagner |     9
(1 row)
```

各 region_b データノードで同じ結果が表示されます。 region_a ノードでクエリーを実行すると、この特定のエントリが表示されません。

最後に、 attendee テーブルがすべてのノードで同様に共有されていることに注意してください。任意の

ノードで、クエリを実行します。

```
SELECT * FROM attendee;
```

```
         id         |    email
- -------------------+----------------
 904252679641903104 | gv@example.com
 904261037006536704 | fb@example.com
(2 rows)
```

## 7.6 Security and roles

EDB Postgres Distributed を使用すると、デザインで保存されたデータへのアクセスを付与せずに PGD クラスターを管理できます。これは、ロールとシステムオブジェクトへの制御されたアクセスを使用して実現します。

- *Security and roles* は、PGD 機能へのアクセスを制御するために PGD が事前定義するロールを導入します。
- *Role management* は、マルチデータベースノードと新しいノードでロールが管理される方法について説明します。
- *PGD predefined roles* は、 PGD ロールの特定の特権を詳細に説明します。
- *Roles and replication* は、PGD レプリケーションがロールおよび特権とどのように相互作用するかについて説明します。
- *Access control* は、テーブル、ファンクション、カタログオブジェクト、トリガーが PGD ロールおよび Postgres 属性とどのように相互作用するかについて説明します。

### 7.6.1 Roles

PGD の構成と管理にはスーパーユーザーアクセスは必要ありません。スーパーユーザーアクセスを使用しないことをお勧めします。代わりに、PGD の管理に必要な特権は、次の事前定義されたロールに分割されます。

| Role | Description |
|---|---|
| *bdr_superuser* | すべての PGD テーブルとファンクションにアクセスする最高の特権ロール。 |
| *bdr_read_all_stats* | PGD の状態を理解するために十分な、テーブル、ビュー、およびファンクションへの読み取り専用アクセスを持つロール。 |
| *bdr_monitor* | bdr_read_all_stats の特権と、モニタリングのための追加の特権が含まれます。 |
| *bdr_application* | PGD を実行するアプリケーションに必要な最小限の特権。 |
| *Replicated large objects* | :ref:``bdr.conflict_history` 列<*bdr.conflict_history* 列>` のすべての競合を表示できます。 |

これらのロールは、 PostgreSQL の pg_ predefinedroles に似た名前が付けられます。

PGD bdr_ ロールは、BDR 拡張機能のインストール時に作成されます。各ロールが持つ特権の詳細については、*PGD predefined roles* を参照してください。

PGD の管理には、管理者がユーザーデータにアクセスする必要はありません。

競合に関する情報を保護するための取り決めについては、Logging conflicts to a table で説明されています。

bdr.conflict_history_summary ビューを使用して競合をモニタできます。

### 7.6.2 Role management

ユーザーは、PostgreSQL インスタンスのグローバルオブジェクトです。CREATE ROLE コマンドまたはそのエイリアス CREATE USER は、PGD レプリケートデータベースで実行される場合、自動的にレプリケートされます。ロールまたはユーザーが非 PGD、非レプリケートデータベースで作成された場合、ロールはその PostgreSQL インスタンスにのみ存在します。GRANT ROLE および DROP ROLE は同じように動作し、PGD レプリケートデータベースに適用される場合にのみレプリケートします。

> **注釈**
>
> Postgres 用語でのユーザーは、単にログイン権限を持つロールであることに注意してください。

#### ロールルール - 未複製のロールはありません

非 PGD の非レプリケートデータベースにロールまたはユーザーを作成する場合、PGD レプリケートデータベースのオブジェクトをそのロールに依存させないことが特に重要です。PGD は、PGD レプリケートデータベースにないロールをレプリケートできないため、レプリケーションプロセスが中断します。

bdr.role_replication 設定をオフにすることにより、この自動レプリケーション動作を無効にできますが、それはお勧めしません。

#### 新しいノードのロール

bdr_init_physical を使用して追加された新しい PGD ノードは、PGD クラスターの他のノードからロールを自動的にレプリケートします。

PGD 6.0.1 以降、 bdr_init_physical を使用せずに PGD ノードが PGD グループに手動で参加すると、既存のロールが新しく参加したノードに自動的にコピーされます。これは、グループに参加する前に、新しいノードでロールを手動で作成する必要がなくなることを意味します。

ロールが新しいノードにコピーされるときに、同じ名前の既存のロールまたはテーブルスペースが存在する場合、新しいノードの既存のロールまたはテーブルスペースは更新されて、ロールまたはテーブルスペースと同じ設定パスワードを共有します) 参加操作のソースノードで。

#### 接続とロール

新しい PGD ノードを割り当てるときに、 bdr.create_node の local_dsn 引数と bdr.join_node_group の join_target_dsn の DSN で指定されたユーザー

データベースオブジェクトを参照、作成、および管理するために頻繁に使用されます。

PGD は、これらの DSN で SUPERUSER 権限を持つロールを使用する場合でも、特権エスカレーション攻撃を防ぐように注意深く書き込まれています。

攻撃対象領域をさらに削減するために、これらの DSN でより制限されたユーザーを指定できます。少なくとも、このようなユーザーには、次の条件が満たされるように、すべてのノードで権限を付与する必要があります。

- ユーザは REPLICATION 属性を持っています。
- データベースに対する CREATE 権限が付与されています。
- bdr_superuser ロールを継承します。
- 複製するすべてのデータベースオブジェクトを直接、または所有者ロールの権限から所有します。

また、デフォルト以外の拡張機能 BDR 拡張機能を除くがソースノードに存在し、スーパーユーザーのみがインストールできる場合、スーパーユーザーは参加ターゲットノードでこれらの拡張機能を手動で作成する必要があります。それ以外の場合、参加プロセスは失敗します。

PostgreSQL 13 以降では、スーパーユーザーの権限を必要とし、手動でインストールする必要がある拡張機能を特定できます。ソースノードで、次を実行します。

```
SELECT name, (trusted IS FALSE AND superuser) AS superuser_only
  FROM pg_available_extension_versions
 WHERE installed AND name != bdr;
```

すべてのノードが参加したら、DML および DDL レプリケーションを引き続き許可するために、次の権限をさらに削減できます。

- ユーザは REPLICATION 属性を持っています。
- bdr_superuser ロールを継承します。

### 7.6.3 PGD predefined roles

PGD の事前定義されたロールは、BDR 拡張機能のインストール時に作成されます。BDR 拡張機能がデータベースから削除された後、ロールは存在し続けます。ドロップが必要な場合は、手動でドロップする必要があります。

#### bdr_superuser

このロールは、PGD 関連のすべてを管理できる管理ユーザー用です。データベースとテーブルアクセスを分離して管理できます。これを使用すると、PostgreSQL のスーパーユーザー権限を付与せずに PGD クラスターを管理できるユーザーができます。

##### 特権

- スキーマ BDR のすべてのテーブルのすべての特権
- スキーマ BDR のすべてのルーチンのすべての特権

### bdr_read_all_stats

このロールは、ユーザーまたはアプリケーションが PGD クラスターの統計と状態を観察するために必要なほとんどのテーブル、ビュー、およびファンクションへの読み取りアクセスを提供します。

#### 特権

次の SELECT 特権

- bdr.autopartition_partitions
- bdr.autopartition_rules
- bdr.ddl_epoch
- bdr.ddl_replication
- bdr.global_consensus_journal_details
- bdr.global_lock
- bdr.global_locks
- bdr.group_camo_details
- bdr.local_consensus_state
- bdr.local_node_summary
- bdr.node
- bdr.node_catchup_info
- bdr.node_catchup_info_details
- bdr.node_conflict_resolvers
- bdr.node_group
- bdr.node_local_info
- bdr.node_peer_progress
- bdr.node_replication_rates
- bdr.node_slots
- bdr.node_summary
- bdr.replication_sets
- bdr.replication_status
- bdr.sequences
- bdr.stat_activity
- bdr.stat_relation
- bdr.stat_subscription *deprecated*

- 

bdr.state_journal_details

- bdr.subscription
- bdr.subscription_summary
- bdr.tables
- bdr.taskmgr_local_work_queue
- bdr.taskmgr_work_queue
- 

bdr.worker_errors *deprecated*

- bdr.workers
- bdr.writers
- `bdr.xid_peer_progress`

以下の EXECUTE 特権

- `bdr.bdr_edition` *deprecated*
- bdr.bdr_version
- bdr.bdr_version_num
- bdr.decode_message_payload
- bdr.get_consensus_status
- bdr.get_decoding_worker_stat
- bdr.get_global_locks
- bdr.get_min_required_replication_slots
- bdr.get_min_required_worker_processes
- bdr.get_raft_status
- bdr.get_relation_stats
- bdr.get_slot_flush_timestamp
- `bdr.get_sub_progress_timestamp`
- bdr.get_subscription_stats
- bdr.lag_control
- bdr.lag_history
- bdr.node_catchup_state_name
- bdr.node_kind_name

- bdr.peer_state_name
- bdr.show_subscription_status
- bdr.show_workers
- bdr.show_writers
- bdr.stat_get_activity
- bdr.wal_sender_stats
- bdr.worker_role_id_name

### bdr_monitor

このロールは、ユーザーまたはアプリケーションが PGD クラスターを監視するために必要なテーブル、ビュー、および機能への読み取りアクセスを提供します。これには、*bdr_read_all_stats* ロールのすべての特権が含まれています。

#### 特権

*bdr_read_all_stats* のすべての特権に加えて、次の追加特権。

次の SELECT 特権

- bdr.group_raft_details
- bdr.group_replslots_details
- bdr.group_subscription_summary
- bdr.group_versions_details
- bdr.raft_instances

次の EXECUTE 特権

- bdr.get_raft_instance_by_nodegroup
- bdr.monitor_camo_on_all_nodes
- bdr.monitor_group_raft
- bdr.monitor_group_versions
- bdr.monitor_local_replslots
- bdr.monitor_raft_details_on_all_nodes
- bdr.monitor_replslots_details_on_all_nodes
- bdr.monitor_subscription_details_on_all_nodes
- bdr.monitor_version_details_on_all_nodes
- bdr.node_group_member_info

### bdr_application

このロールは、PGD 機能、オブジェクト、シーケンス、CRDT データ型、CAMO ステータス機能、トリガー管理機能などの機能へのアクセスを必要とするアプリケーション向けに設計されています。

#### 特権

次の EXECUTE 特権

- column_timestamps データ型のすべてのファンクション
- CRDT データ型のすべてのファンクション
- bdr.alter_sequence_set_kind
- bdr.create_conflict_trigger
- bdr.create_transform_trigger
- bdr.drop_trigger
- bdr.get_configured_camo_partner
- bdr.global_lock_table
- bdr.is_camo_partner_connected
- bdr.is_camo_partner_ready
- bdr.logical_transaction_status
- bdr.ri_fkey_trigger
- bdr.seq_nextval
- bdr.seq_currval
- bdr.seq_lastval
- bdr.trigger_get_committs
- bdr.trigger_get_conflict_type
- bdr.trigger_get_origin_node_id
- bdr.trigger_get_row
- bdr.trigger_get_type
- bdr.trigger_get_xid
- bdr.wait_for_camo_partner_queue
- bdr.wait_slot_confirm_lsn
- bdr.wait_node_confirm_lsn

これらの機能の多くは、使用する前に追加の特権を必要とします。たとえば、bdr.alter_sequence_set_kind を正常に実行するには、テーブルの所有者である必要があります。これらの追加ルールは、特定の機能ごとに説明します。

#### bdr_read_all_conflicts

PGD は競合を bdr.conflict_history にログ記録します

テーブル。競合はテーブルの所有者にのみ可視されるため、所有者が競合履歴を読み取るために追加の特権は必要ありません。

ただし、すべてのテーブルの競合を表示できるユーザーが便利な場合は、オプションでそのユーザーにロール bdr_read_all_conflicts を付与できます。

#### 特権

bdr.conflict_history に明示的なポリシーが設定され、このロールが bdr.conflict_history テーブルを読み取ることができます。

### 7.6.4 Roles and replication

#### DDL および DML レプリケーションとユーザー

ユーザーが実行した DDL 変更は、各ノードでその同じユーザーとして適用されます。

テーブルへの DML 変更は、ターゲットノード上のテーブル所有ユーザーとして複製されます。

デフォルトでは、PGD はノード全体に同じ所有者を持つ新しいテーブルをレプリケートします。

#### テーブルの所有権の違い

同じユーザーが各ノードのテーブルを所有することをお勧めします。これはデフォルトの動作ですが、オーバーライドできます。その場合、考慮すべきことがいくつかあります。

テーブル A がノード 1 のユーザー X が所有し、ノード 2 のユーザー Y が所有している状況を考えます。ユーザー Y がユーザー X よりも高い特権を持っている場合、これは特権エスカレーションと見なされる場合があります。

ノードにはさまざまなユースケースがある可能性があるため、このシナリオを許可します。しかし、私たちはそれに対して警告します。テーブルの所有者が異なるノードに異なる場合、セキュリティ管理者にこの構成の計画と監査を手伝ってもらうことをお勧めします。

#### レプリケーションと行レベルのセキュリティ

行レベルのセキュリティポリシーが有効になっているテーブルでは、適用時にポリシーを強化せずに変更がレプリケートされます。この動作は、FORCE ROW LEVEL SECURITY が指定されている場合でも、NO FORCE ROW LEVEL SECURITY として適用される変更と同等です。これは希望するものでない場合は、すべての行のレプリケートを回避する row_filter を指定します。すべてのノードの行セキュリティポリシーは同一であるか、少なくとも互換性があることをお勧めしますが、これは強制しません。

#### bdr_superuser のロールとレプリケーション

ユーザー bdr_superuser は、PGD のレプリケーションを制御し、レプリケーションセットから任意のテーブルを追加または削除できます。bdr_superuser は、個々のテーブルに対する特権を必要とせず、お勧めしません。レプリケーションセットファンクションへのアクセスを制限する必要がある場合は、これらのファンクションの制限バージョンを SECURITY DEFINER ファンクションとして実装し、適切なユーザーに付与できます。

#### 特権制限

PGD は追加の制限を適用し、TRIGGER または REFERENCES 特権のみに依存する DDL の使用を効果的に防止します。

GRANT ALL は引き続き TRIGGER と REFERENCES 特権の両方を付与するため、特権を明示的に記述することをお勧めします。たとえば、ALL の代わりに GRANT SELECT, INSERT, UPDATE, DELETE, TRUNCATE を使用します。

#### 外部キー特権

ALTER TABLE ... ADD FOREIGN KEY は、ユーザーが参照されるテーブルに対する SELECT 特権を持っている場合、または参照先テーブルで現在のユーザーがバイパスできない RLS 制限が有効になっている場合にのみサポートされます。

これは、REFERENCES 特権だけでは PGD で外部キーを作成するのに十分でないことを意味します。REFERENCES 特権のみに依存すると、テーブルスキャンではなくトリガーを使用して検証チェックが実行されるため、通常は役立ちません。通常、コストが高すぎるため、正常に使用できません。

### 7.6.5 Access control

#### カタログテーブル

システムカタログと情報スキーマテーブルは、PGD によってレプリケーションから常に除外されます。

さらに、拡張機能が所有するテーブルはレプリケーションから除外されます。

#### PGD ファンクションと演算子

すべての PGD ファンクションは bdr スキーマで公開されています。これらのファンクションへの呼び出しは、search_path に bdr を配置するのではなく、スキーマ修飾する必要があります。

すべての PGD 演算子は pg_catalog スキーマを介して使用できるため、ユーザーは search_path から public スキーマを問題なく除外できます。

#### カタログオブジェクトに対する特権の付与

管理者は、テーブル、ビュー、ファンクションなどのカタログオブジェクトに対する明示的な特権を付与してはなりません。*PGD predefined roles* で説明されているロールのいずれかを付与して、これらのオブジェクトへのアクセスを管理します。

この要件は、結合のどちらの側のノードも全く同じバージョンの PGD、したがって PGD カタログを持たない場合でも、ノードグループに参加できる柔軟性の結果です。

より正確には、個々のカタログオブジェクトの特権が明示的に付与された場合、参加するノードから抽出された対応する GRANT ステートメントが参加しているノードに適用されない可能性があるため、`bdr.join_node_group()` プロシージャーは失敗する可能性があります。

### トリガー

PostgreSQL では、テーブルの所有者と TRIGGER 特権を付与された人は誰でもトリガーを作成できます。非テーブル所有者によって付与されたトリガーは、PGD でテーブル所有者として実行されます。これは、セキュリティ問題を発生させる可能性があります。TRIGGER 特権はめったに使用されず、PostgreSQL コアチームは、「個別の TRIGGER 権限は時代遅れだと考えられるものです。」と述べています。

PGD は、テーブルにトリガーを作成できる人に関するより厳密なルールを使用することにより、この問題を緩和します。

- スーパーユーザー トリガーを作成できます。
- bdr_superuser トリガーを作成できます。
- テーブルの所有者 PostgreSQL と同じルールに従ってトリガーを作成できます。トリガーによって使用されるファンクションの EXECUTE 特権が必要です。
- テーブルに対する TRIGGER 特権を持つユーザーテーブルと同じ所有者が所有するファンクションを使用し、標準の PostgreSQL ルールを満たす場合にのみトリガーを作成できます。具体的には、ファンクションに対する EXECUTE 特権が必要です。テーブルとファンクションの両方の所有者が同じで、所有者がユーザーにテーブルの TRIGGER 特権とファンクションの EXECUTE 特権の両方を付与することにした場合。そのユーザーがこのファンクションを使用してそのテーブルにトリガーを作成してもよいと想定されています。
- テーブルに対する TRIGGER 特権を持つユーザー

SECURITY DEFINER clause で定義されたファンクションを使用してトリガーを作成することもできます

EXECUTE 権限がある場合。 SECURITY DEFINER 句により、ファンクションは常に標準の PostgreSQL と PGD の両方でファンクションの所有者として実行されます。

このロジックは、PostgreSQL では、トリガーの所有者がそれを作成したユーザーではなく、そのトリガーが使用するファンクションの所有者であるという事実に基づいて構築されています。

同じルールが既存のテーブルに適用され、既存のテーブルに、テーブルの所有者が所有せず、SECURITY DEFINER ファンクションを使用しないトリガーがある場合、レプリケーションセットに追加することはできません。

PGD レプリケーションは変更を適用する場合、システムレベルのデフォルトの search_path のみを使用します。他の search_path 設定を想定しているレプリカトリガー、ストリームトリガー、およびインデックス式ファンクションは、適用時に実行すると失敗します。これが発生しないようにするには、デフォルトの search_path のみを使用してオブジェクト参照を明確に解決するか、影響を受けるファンクションに `ALTER FUNCTION ... SET search_path = ...` を使用してファンクションの検索パスを設定します。デフォルトの search_path を使用する場合、schema.objectname などのオブジェクトへの完全修飾参照を常に使用します。

## 7.7 Sequences

多くのアプリケーションでは、一意のサロゲート ID をデータベースエントリに割り当てる必要があります。多くの場合、データベース SEQUENCE オブジェクトはこれらを生成するために使用されます。PostgreSQL では、これらは次のいずれかです。

- CREATE SEQUENCE コマンドを使用して手動で作成し、nextval() ファンクションを呼び出して取得したシーケンス
- serial および bigserial 列または GENERATED BY DEFAULT AS IDENTITY 列

ただし、PostgreSQL の標準シーケンスはマルチノード対応ではなく、ローカルノードでのみ一意の値を生成します。このようなシーケンスによって生成された一意の ID は、マルチマスターレプリケーションで破棄された INSERT アクションによる競合とデータ損失を発生させるため、これは重要です。

彼らに割り当てられたロール。

このため、 PGD は、 _global sequences_と呼ばれる、PGD グループ全体の bigint または bigserial データ型のシーケンスを使用して、一意の ID を生成するアプリケーション透過的な方法を提供します。

PGD グローバルシーケンスは、アプリケーションがデータベースを使用して、ほとんどの場合ただし必ずしもすべての場合ではない非同期分散システムで一意の合成キーを生成する簡単な方法を提供します。

PGD グローバルシーケンスを使用すると、挿入の競合に関する問題を回避できます。グローバルシーケンスを使用している列に PRIMARY KEY または UNIQUE 制約を定義すると、ノードは他のノードと同じ値を取得できません。PGD がノード間で挿入を同期すると、競合することはできません。

PGD グローバルシーケンスは PostgreSQL シーケンスを拡張するため、クラッシュセーフです。これらを使用するには、bdr_application ロールを付与される必要があります。

グローバルシーケンスにはさまざまなアルゴリズムが考えられます。

- SnowflakeId シーケンス
- グローバルに割り当てられたレンジシーケンス

SnowflakeId シーケンスは、どの時点でもノード間通信を必要としないアルゴリズムを使用して値を生成します。より速く、より堅牢で、値が作成されたときにタイムスタンプを記録する便利な特性があります。

SnowflakeId シーケンスには、64 ビット BIGINT データ型でのみ動作し、最大 19 桁の長さの値を生成するという制限があります。これは、JavaScript Number タイプなどの一部のホスト言語データ型で使用するには長すぎる場合があります。グローバルに割り当てられたシーケンスは、ノード間コンセンサスによって必要に応じて補充できるローカルな値の範囲を割り当て、BIGINT または INTEGER シーケンスに適しています。

bdr.alter_sequence_set_kind() を使用してグローバルシーケンスを作成できます

ファンクション。このファンクションは、標準の PostgreSQL シーケンスを取得し、PGD グローバルシーケンスとしてマークします。シーケンスを標準の PostgreSQL シーケンスに変換して戻すこともできます。

PGD は、構成変数 bdr.default_sequence_kind も提供します。この変数は、CREATE SEQUENCE コマンドが実行されたとき、または serial 、bigserial 、または GENERATED BY DEFAULT AS IDENTITY 列が作成されたときに作成するシーケンスの種類を決定します。有効な設定は次のとおりです。

- local —新しく作成されたシーケンスは、標準の PostgreSQL ローカルシーケンスです。

- galloc —常にグローバルに割り当てられたレンジシーケンスを作成します。
- snowflakeid —時間、ノード ID、およびカウンターコンポーネントで構成される BIGINT シーケンスのグローバルシーケンスを作成します。INTEGER シーケンスでは使用できませんので、 bigserial には使用できますが、 serial には使用できません。
- timeshard —SnowflakeId シーケンスの古いバージョン。下位互換性のためのみに提供されています。SnowflakeId が優先されます。
- distributed デフォルト — bdr.default_sequence_kind にのみ使用できる特別な値。int8 シーケンス つまり bigserial の場合は snowflakeid を選択し、int4 つまり serial および int2 シーケンスの場合は galloc シーケンスを選択します。

bdr.sequences ビューには、個々のシーケンス種類に関する情報が表示されます。

currval() および lastval() ファンクションは、すべてのタイプのグローバルシーケンスで正しく動作します。

### 7.7.1 自動シーケンス変換

PGD 6.0 以降では、ノードを PGD グループに参加するか、新しいグループを作成する行為も、ローカルシーケンスのグローバルシーケンスへの変換をトリガーします。bdr.default_sequence_kind を distributed に設定します。この設定は、ローカルシーケンスの変換に最適な種類のシーケンスを選択します。bdr.default_sequence_kind が local に設定されている場合、シーケンスはローカルシーケンスとして残されます。galloc への変換は、シーケンスがグループ内の他のシーケンスと競合しないことを保証する方法で実行されます。

ローカルシーケンスで開始し、後で galloc シーケンスに切り替える場合は、bdr.default_sequence_kind を galloc に設定し、変換する各シーケンスで bdr.alter_sequence_set_kind() ファンクションを実行します。ただし、シーケンスの開始値を手動で設定して、テーブル内の既存の値と競合しないようにする必要があることに注意してください。これ全般、特に How to set a new start value for a sequence の詳細については、**ローカルシーケンスの *galloc* シーケンスへの変換** を参照してください。

#### SnowflakeId シーケンス

SnowflakeId シーケンスによって生成された ID は、緩やかな時間順序であるため、これらを使用して、標準の PostgreSQL シーケンスのように、データ挿入のおおよその順序を取得できます。1 つのノードでも、同じミリ秒以内に生成された値は順序が崩れる場合があります。緩い時間順序付けの特性は、レンジパーティションキーとしての使用に適していることを意味します。

SnowflakeId シーケンスは 1 つ以上のノードで動作し、ノード参加プロセスが完了した後にノード間通信を必要としません。したがって、拡張ネットワーク分割のリスクがある場合でも、それらを使用し続けることができます。これらは、レプリケーション遅延やノード間レイテンシーの影響を受けません。

SnowflakeId シーケンスは、標準のシーケンスとは異なる方法で一意の ID を生成します。このアルゴリズムは、シーケンス番号に 3 つのコンポーネントを使用します。シーケンスの最初のコンポーネントは、シーケンス番号生成時のタイムスタンプです。シーケンス番号の 2 番目のコンポーネントは、各 PGD ノードに割り当てられる一意の ID であり、これにより、異なるノードからの ID が常に異なることが保証されます。3 番目の成分は、ローカルシーケンスによって生成された数です。

競合がないことを保証には一意のノード ID をシーケンス番号に追加するだけで十分ですが、シーケンスの別の便利なプロパティも保持したいと思います。シーケンス番号の順序は、データがテーブルに挿入された順序にほぼ対応します。最初にタイムスタンプを配置することにより、これが保証されます。

SnowflakeId シーケンスには、いくつかの制限と注意事項が適用されます。

SnowflakeId シーケンスは 64 ビット幅で、`bigint` または `bigserial` が必要です。生成される値は最大 19 桁の長さです。実用的な 32 ビット `integer` バージョンがないため、 `serial` シーケンスでは使用できません。代わりに、グローバルに割り当てられたレンジシーケンスを使用します。

SnowflakeId の場合、特定のノードでミリ秒ごとに生成されるシーケンス値 4096 の制限があります 1 秒あたり約 400 万シーケンス値。シーケンス値の生成が特定のミリ秒内でラップアラウンドする場合、SnowflakeId シーケンスは次のミリ秒まで待機し、そのミリ秒の新しい値を取得します。

SnowflakeId シーケンスはタイムスタンプをシーケンス値にエンコードするため、指定された時間フレーム内でのみ新しいシーケンス値を生成できますシステムクロックによって異なります。使用できる最も古いタイムスタンプは 2016-10-07 で、これは SnowflakeId のエポック時間です。値は 2086 年に負の値にラップし、2156 年までに数値が完全になくなります。

タイムスタンプは SnowflakeId シーケンスの重要な部分であるため、Postgres プロセスの有効期間中に使用される最新のタイムスタンプより古いシーケンスを生成しないための追加の保護がありますただし、Postgres の再起動の間ではありません。

SnowflakeId シーケンスの入力として使用されるシーケンスの `INCREMENT` オプションは、事実上無視されます。これは、多くのオブジェクトリレーショナルマッパー ORM ツール、特に Hibernate と同様に、シーケンス ID キャッシュを行うアプリケーションに関連する可能性があります。シーケンスは時間ベースであるため、アプリケーションがキャッシュされた値で何かをできるようになるまでにシーケンスは新しい非衝突値に進むため、実際的な効果はほとんどありません。

同様に、基になるシーケンスの `START` 、`MINVALUE` 、`MAXVALUE` 、および `CACHE` 設定を変更することもできますが、変更するメリットはありません。シーケンスの下位 14 ビットが使用され、残りは破棄されるため、値範囲制限はファンクションの結果に影響を与えません。同じ理由で、`setval()` は SnowflakeId シーケンスには役立ちません。

### タイムシャードシーケンス

タイムシャードシーケンスは、既存のインストールとの下位互換性のために提供されていますが、新しいアプリケーションの使用はお勧めしません。代わりに SnowflakeId シーケンスを使用することをお勧めします。

Timeshard は SnowflakeId に非常に似ていますが、制限が異なり、保護が少なく、パフォーマンスが低下します。

Timeshard と SnowflakeId の違いは次のとおりです。

- Timeshard は、ミリ秒ごとに最大 16384 秒ごとに約 1600 万を生成できます。これは、SnowflakeId を超えています。ただし、特定のミリ秒内のラップアラウンドに対する保護はありません。Timeshard シーケンスを使用するスキーマは、特定の列に timeshard 値を使用する場合、`UNIQUE` 制約の使用を保護する必要があります。
- Timeshard シーケンスのタイムスタンプコンポーネントは、2050 年に値がなくなり、bigint と組み合わせて使用すると、値は 2033 年に負の数にラップします。これは、2033 年以降に生成されたシー

ケンスが負の値を持つことを意味します。これは SnowflakeId よりもかなり短い時間であり、これが SnowflakeId が優先される主な理由です。

- Timeshard シーケンスは、時折のディスク書き込みが必要です標準のローカルシーケンスと同様に。SnowflakeId はメモリで計算されるため、SnowflakeId シーケンスは、一般に timeshard シーケンスよりも少し高速です。

### ログに記録されていないシーケンスと PGD

Postgres 15 以降、ログに記録されないシーケンスを作成できるようになりました。これらは関連しており、WAL に書き込まれず、レプリケートされないログ非記録テーブルに似ています。PGD とログなしシーケンスのコンテキストでは、ログなしの PGD シーケンスを持つことは合理的な構成ではなく、ノードに障害が発生した場合に予期しない問題を引き起こす可能性があります。したがって、ログのない PGD 配列の作成、または PGD 配列のログのない配列への変換を防止します。

### グローバルに割り当てられたレンジシーケンス

グローバルに割り当てられたレンジまたは `galloc` シーケンスは、値のレンジチャンクを各ノードに割り当てます。ローカルレンジが使い果たされると、他のノード間のコンセンサスによって、新しいレンジがグローバルに割り当てられます。この動作はキー空間を効率的に使用しますが、現在割り当てられているローカルレンジが使い果たされたときにシーケンスジェネレーターが進行するには、ローカルノードがクラスター内のノードの大部分に接続している必要があります。

SnowflakeId シーケンスとは異なり、`galloc` シーケンスは、PostgreSQL が提供するすべてのシーケンスデータ型 `smallint` 、`integer` 、および `bigint` をサポートします。これは、64 ビットシーケンスに問題がある環境で `galloc` シーケンスを使用できることを意味します。例には、53 ビット値のみをサポートするため、またはシーケンスが制限されたスペースで出力に表示される場合、JavaScript での整数の使用が含まれます。

各投票ノードによって割り当てられる範囲は、シーケンスが使用しているデータ型に基づいて、現在事前に決定されています。

- smallint — 1 000 の数字
- integer — 1 000 000 の数字
- bigint — 1 000 000 000 数値

各ノードは seq_chunk_size の 2 つのチャンク—現在使用するものと将来の使用のために予約されたチャンク—を割り当てるため、1 つのノードから生成される値は単調増加します。ただし、グローバルに見ると、生成された値はまったく順序付けられていません。これは、b ツリーインデックスへの影響によりパフォーマンスの低下を引き起こす可能性があり、通常は、生成された値がレンジパーティションキーとして役に立たないことを意味します。

`galloc` シーケンスの主な欠点は、割り当てられたレンジが使い果たされると、シーケンスジェネレーターがノード間通信を必要とするローカルノードの次のレンジに関するコンセンサスを求めなければならないことです。これにより、PGD グループの大部分がアクセスできない場合、遅延または操作の問題が発生する可能性があります。(これは、後のリリースで回避される可能性があります。)

`CACHE` 、`START` 、`MINVALUE` 、および `MAXVALUE` オプションは、`galloc` シーケンスで正常に動作します。ただし、シーケンスを `galloc` 種類に変換する前に、それらを設定する必要があります。 `INCREMENT BY` オ

プションも正常に動作します。ただし、各シーケンスデータ型に割り当てられた上記の範囲以上の増分値を割り当てることはできません。setval() は、galloc シーケンスのグローバル状態をリセットしません。使用しないでください。

PGD は、 bdr.sequence_alloc カタログの galloc シーケンスを追跡して、現在割り当てられているチャンクを管理します。Raft はチャンクの割り当てを管理しますが、シーケンスの名前または名前空間の変更は、伝統的にレプリケーションストリームを介して処理されます。PGD 6.3.0 以降、ALTER SEQUENCE ... RENAME および ALTER SEQUENCE ... SET SCHEMA は、すべてのシーケンスタイプで完全にサポートされています。 galloc シーケンスの場合、このサポートは bdr.sequence_kind catalog のグローバル UUID 識別子 sequuid に依存しています。これにより、PGD は名前または名前空間の変更にわたって一貫してシーケンスを追跡できます。

これらのオブジェクトの管理の詳細については、*ALTER SEQUENCE* を参照してください。

### ローカルシーケンスの galloc シーケンスへの変換

ローカルシーケンスを galloc に変換する前に、いくつかの前提条件に注意する必要があります。

#### 1. シーケンスと列のデータ型が一致することを確認します

シーケンスのデータ型が、使用される列のデータ型と一致することを確認します。たとえば、bigint シーケンスを作成し、 integer 列のデフォルトをそのシーケンスによって返される nextval() に割り当てることができます。 bigint の場合 1 000 000 000 のブロックで割り当てられる galloc シーケンスでは、3 つ以上のノードが使用されている場合、nextval() によって返される値がすぐに int4 範囲を超えます。

この例は、何が起こり得るかを示しています。

```
CREATE SEQUENCE int8_seq;

SELECT sequencename, data_type FROM pg_sequences;
 sequencename | data_type
- -------------+-----------
 int8_seq     | bigint
(1 row)

CREATE TABLE seqtest (id INT NOT NULL PRIMARY KEY);

ALTER SEQUENCE int8_seq OWNED BY seqtest.id;

SELECT bdr.alter_sequence_set_kind(public.int8_seq::regclass, galloc, 1);
 alter_sequence_set_kind
- ------------------------

(1 row)

ALTER TABLE seqtest ALTER COLUMN id SET DEFAULT nextval(int8_seq::regclass);
```

2 つのノードで INSERT INTO seqtest VALUES(DEFAULT) を実行した後、テーブルには次の値が含まれます。

```
SELECT * FROM seqtest;
     id
- ----------
          2
 2000000002
(2 rows)
```

ただし、3 番目のノードで同じ操作を試行すると、シーケンスが値 4000000002 を生成したため、integer out of range エラーで失敗します。

**Tip**

PostgreSQL からシーケンスの現在のデータ型を取得できます

pg_sequences

ビュー。現在の値が新しいデータ型の最大値を超えない限り、ALTER SEQUENCE ... AS ... たとえば ALTER SEQUENCE public.sequence AS integer を使用してシーケンスのデータ型を変更できます。

### 2. シーケンスの新しい開始値を設定します

シーケンス種類が galloc に変更されると、書き換えられ、ローカルシーケンスの定義された開始値からリスタートします。これが実稼働データベースの既存のシーケンスで発生した場合、現在の値を照会し、開始値を適切に設定する必要があります。このユースケースを支援するために、PGD ではファンクション

bdr.alter_sequence_set_kind() で開始値を渡すことができます。すでにオフセットを使用しており、複数のノードからの書き込みがある場合は、最大使用値を確認し、少なくとも次の値までシーケンスをリスタートする必要があります。

```
- - determine highest sequence value across all nodes
SELECT max((x->response->command_tuples->0->>nextval)::bigint)
    FROM jsonb_array_elements(
        bdr.run_on_all_nodes(
            ESELECT nextval(\public.sequence\);
            )::jsonb) AS x;

- - turn into a galloc sequence
SELECT bdr.alter_sequence_set_kind(public.sequence::regclass, galloc, $MAX + $MARGIN);
```

ユーザーはシーケンスをロックできないため、 $MARGIN 値を残して、max() 値が照会されている間操作を続行できるようにする必要があります。

bdr.sequence_alloc テーブルは、チャンクサイズと、クラスター全体に割り当てられたレンジに関する情報を提供します。

この例では、シーケンスは 333 で始まり、クラスターには 2 つのノードがあります。割り当て数は 4 で、ノードごとに 2 であり、チャンクサイズは 1000000 で、整数シーケンスに関連しています。

```
SELECT * FROM bdr.sequence_alloc
    WHERE seqid = public.categories_category_seq::regclass;
          seqid          | seq_chunk_size | seq_allocated_up_to | seq_nallocs |      ␣
↪seq_last_alloc
- ------------------------+----------------+---------------------+-------------+------
↪-----------------------
 categories_category_seq |        1000000 |             4000333 |           4 | 2020-
↪05-21 20:02:15.957835+00
(1 row)
```

各ノードの特定のシーケンスに現在割り当てられているレンジを表示するには、ファンクション bdr.galloc_chunk_info を実行します。

- ノード Node1 は、 333 〜2000333 のレンジを使用しています。

```
SELECT * FROM bdr.galloc_chunk_info(categories_category_seq);
 chunk_start | chunk_end
- ------------+-----------
         334 |   1000333
     1000334 |   2000333
(2 rows)
```

- ノード Node2 は、 2000334 〜4000333 のレンジを使用しています。

```
SELECT * FROM bdr.galloc_chunk_info(categories_category_seq);
 chunk_start | chunk_end
- ------------+-----------
     2000334 |   3000333
     3000334 |   4000333
```

ノードはチャンクを終了すると、新しいチャンクのコンセンサスを求め、最初に利用可能なものを取得します。この例では、4000334〜5000333 です。これは新しい予約チャンクであり、古い予約チャンクの消費を開始します。

### 7.7.2 UUID、KSUUID、およびその他のアプローチ

PGD で使用できるグローバルシーケンスを使用せずに、他の方法でグローバルに一意の ID を生成できます。例

- UUID とその PGD バリアント KSUUID
- ノードごとに異なるオフセットを持つローカルシーケンスつまり手動
- 外部調整されたナチュラルキー

PGD アプリケーションは、他の方法を安全に使用できません。PGD はノード間で行ロックを取得しないため、シーケンス生成に SELECT ... FOR UPDATE 、UPDATE ... RETURNING ... または同様のものに依存するカウンターテーブルベースのアプローチは、PGD では正しく動作しません。同じ値が複数のノードで生成されます。同じ理由で、「ギャップレス」シーケンス生成の通常の戦略は PGD では機能しません。ほ

とんどの場合、アプリケーションは、2 フェーズコミットを使用して外部ソースからギャップレスである必要があるシーケンスの生成を調整します。または、PGD グループ内の 1 つのノードでのみ生成されます。

### 7.7.3 KSUUID v2 ファンクション

PGD アプリケーションは、他の方法を安全に使用できません。PGD はノード間で行ロックを取得しないため、シーケンス生成に `SELECT ... FOR UPDATE`、`UPDATE ... RETURNING ...` または同様のものに依存するカウンターテーブルベースのアプローチは、PGD では正しく動作しません。同じ値が複数のノードで生成されます。同じ理由で、「ギャップレス」シーケンス生成の通常の戦略は PGD では機能しません。ほとんどの場合、アプリケーションは、2 フェーズコミットを使用して外部ソースからギャップレスである必要があるシーケンスの生成を調整します。または、PGD グループ内の 1 つのノードでのみ生成されます。

#### UUID

`UUID` キーは、代わりにシーケンスを完全に回避し、128 ビットのユニバーサル一意識別子を使用します。これらはランダムまたは擬似乱数値であり、非常に大きいため、同じ値を 2 回生成することはほぼ不可能です。`UUID` キーを使用する場合、ノードが継続的に通信する必要はありません。

万が一、衝突が発生した場合、競合検出は、挿入された 2 つのレコードの新しい方を選択して保持します。競合ログが有効になっている場合、このようなイベントが記録されます。ただし、約 2^64 キーが生成された後にのみ衝突が実際に可能になるため、発生する可能性は非常に低いです。

`UUID` キーの主な欠点は、領域とネットワークの点でやや非効率であることです。これらは、主キーとしてだけでなく、外部キーで参照される場合、および回線上で送信される場合にも、より多くの領域を消費します。また、すべてのアプリケーションが `UUID` キーにうまく対応できるわけではありません。

#### KSUUID

PGD は、`UUID` データの K ソート可能なバリアントを操作するためのファンクションを提供します。KSUUID として知られ、PostgreSQL 標準の `UUID` データ型を使用して保存できる値を生成します。 `KSUUID` 値は、`UUID` 標準に従って、タイムスタンプとランダムデータの両方を保存するという点で `UUIDv1` に似ています。違いは、 `KSUUID` が K ソート可能であること、つまりタイムスタンプで弱くソート可能であることを意味します。これにより、よりコンパクトな `btree` インデックスが生成されるため、データベースキーとしてより役立ちます。この動作により、検索の効率が向上し、結果データの自然な時間ソートが可能になります。`UUIDv1` とは異なり、`KSUUID` 値には、生成されたコンピューターの MAC が含まれないため、それらの使用にセキュリティ上の懸念はありません。

あらゆる場合に `KSUUID` v2 を推奨するようになりました。通常の比較演算子で生成された値を直接並べ替えることができます。

PGD には `KSUUID` の 2 つのバージョン v1 および v2 があります。古い `KSUUID` v1 は非推奨ですが、既存のインストールをサポートするために維持されます。新しいインストールには使用しないでください。v1 と v2 の内部コンテンツには互換性がありません。そのため、それらを操作する機能も互換性がありません。`KSUUID` の v2 も、`UUID` バージョン番号を保存しなくなりました。

PGD リファレンスの KSUUID v2 functions および KSUUID v1 functions を参照してください。

### ステップシーケンスとオフセットシーケンス

オフセットステップシーケンスでは、各ノードで通常の PostgreSQL シーケンスが使用されます。各シーケンスは同じ量だけインクリメントし、異なるオフセットで開始します。たとえば、ステップ 1000 では、ノード 1 のシーケンスは 1001、2001、3001 などを生成します。node2 のシーケンスは、1002、2002、3002 などを生成します。このスキームは、ノードが長期間通信できない場合でもうまく機能します。ただし、デザイナーはスキーマを確立するときにノードの最大数を指定する必要があり、ノードごとの構成が必要です。間違いにより、シーケンスのオーバーラップが簡単に発生する可能性があります。

次のように 1 つのノードで目的のシーケンスを作成することにより、PGD でこのアプローチを構成するのは比較的簡単です。

```
CREATE TABLE some_table (
    generated_value bigint primary key
);

CREATE SEQUENCE some_seq INCREMENT 1000 OWNED BY some_table.generated_value;

ALTER TABLE some_table ALTER COLUMN generated_value SET DEFAULT nextval(some_seq);
```

次に、 `setval()` を呼び出す各ノードで、各ノードに異なるオフセット開始値を与えます。例

```
- - On node 1
SELECT setval(some_seq, 1);

- - On node 2
SELECT setval(some_seq, 2);

 -- ... etc
```

将来的に変更することは困難であり、中断を伴うため、追加するすべてのノードのための領域を残すように十分な大きさの `INCREMENT` を許可してください。

`bigint` 値を使用する場合、10000 以上のオフセットを使用する場合でも、キーの枯渇に関する実質的な懸念はありません。枯渇に近づく可能性があるまでには、数百のマシンが毎秒数百万の挿入を実行して、数百年かかります。

PGD は、現在、このようなステップ/オフセットシーケンスでノードごとのオフセットを構成するための自動化を提供していません。

### 複合キー

ステップ/オフセットシーケンスのバリアントは、`PRIMARY KEY (node_number, generated_value)` で構成される複合キーを使用することです。ノード番号は通常、各ノードで異なる番号を返すファンクションから取得されます。DDL レプリケーションを一時的に無効にし、定数 SQL ファンクションを作成することにより、このようなファンクションを作成できます。または、レプリケーションセットの一部ではない 1 行テーブルを使用して、各ノードに異なる値を保存できます。

### 7.7.4 関連項目

- Global Sequence management interfaces
- KSUUID v2 functions
- KSUUID v1 functions

## 7.8 Replicated large objects

PGD は、

PostgreSQL large object interface をミラーするレプリケートされたラージオブジェクト実装を提供します。論理レプリケーションによってレプリケートされない標準の PostgreSQL ラージオブジェクトとは異なり、PGD ラージオブジェクトは、クラスター内のすべてのノードに論理的にレプリケートされます。galloc シーケンスを使用してオブジェクト識別子 OID が生成され、すべてのノードにわたって一意性が保証されます。

アプリケーションまたはドライバーが PostgreSQL ラージオブジェクトインターフェイスに依存しており、すべてのノードにわたってそのデータをレプリケートする必要がある場合に、レプリケートされたラージオブジェクトを使用します。レプリケートされたラージオブジェクトは、既存のアプリケーションを既にラージオブジェクトを使用している PGD に移行する場合、またはラージオブジェクト API を介してバイナリデータを管理するドライバーまたは ORM を使用する場合に一般的に使用されます。

新しいアプリケーションの場合、`BYTEA` の方が簡単で、特別な構成は必要ありません。適切なアプローチを選択するためのガイダンスについては、*Preparing your application for PGD* を参照してください。

### 7.8.1 レプリケートされたラージオブジェクトを使用する

PGD は、レプリケートされたラージオブジェクトを操作する 2 つの方法を提供します。`bdr.lo_*` ファンクションを直接呼び出すか、透過的リダイレクトを有効にして既存のアプリケーションがコードを変更せずに動作するようにします。

#### bdr.lo_*ファンクションの呼び出し

PGD は、`bdr` スキーマのすべての PostgreSQL ラージオブジェクトサーバーサイドファンクションと同等のものを公開します。たとえば、`lo_create()` の代わりに `bdr.lo_create()` 、`lo_open()` の代わりに `bdr.lo_open()` を公開します。完全なファンクションリファレンスについては、*Replicated large objects* を参照してください。

#### PostgreSQL ラージオブジェクト呼び出しのリダイレクト

標準の PostgreSQL `lo_ *` ファンクションへの呼び出しを同等の `bdr.lo_*` に自動的にリダイレクトするように `bdr.override_large_objects = on` を設定するため、既存のアプリケーションはコードを変更せずにレプリケートされたラージオブジェクトを使用します。

リダイレクトは SQL レベルの呼び出しに適用されます。libpq ベースのクライアントとドライバーは、SQL をバイパスして、プロトコルレベルでラージオブジェクト呼び出しも発行します。これらもリダイレクトするには、Postgres に直接接続するのではなく、*Connection Manager Authentication* 読み取り/書き込みポートデフォルト 6432 に接続します。

### 7.8.2 大きなオブジェクトの保存方法

レプリケートされたラージオブジェクトは *bdr.pgd_largeobject* に保存され、権限は *bdr.pgd_largeobject_metadata* に保存されます。これらのテーブルは pg_catalog.largeobject および pg_catalog.largeobject_metadata と同じ目的を果たしますが、その内容はすべてのノードにレプリケートされます。

### 7.8.3 制限事項

- ALTER LARGE OBJECT はサポートされていません。
- この機能を有効にする前に作成されたラージオブジェクトは pg_catalog.largeobject に保存され、bdr.lo_* インターフェイスからアクセスできません。
- psql の\lo_list は、 pg_catalog.largeobject を直接読み取るため、PGD ラージオブジェクトを表示しません。

## 7.9 Stream triggers

PGD は、ダウンストリーム/ターゲットノードでの追加データ処理に使用できる新しいタイプのトリガーを導入しました。

- 競合トリガー
- トランスフォームトリガー

これらのタイプのトリガーは、まとめて_stream Trigger_と呼ばれます。

ストリームトリガーは、構文がトリガーのようなものになるように設計されています。これらは PostgreSQL BEFORE トリガーアーキテクチャを活用し、PostgreSQL BEFORE トリガーと同様のパフォーマンス特性を持つ可能性があります。

通常の PostgreSQL トリガーと同様に、複数のトリガー定義で 1 つのトリガーファンクションを使用できます。トリガーファンクションは、CREATE FUNCTION ... RETURNS TRIGGER 形式で定義されたプログラムです。トリガーの作成に、 CREATE TRIGGER コマンドを使用する必要はありません。代わりに、特別な PGD ファンクション

bdr.create_conflict_trigger() および bdr.create_transform_trigger() を使用してストリームトリガーを作成します。

トリガーは、作成されると、カタログテーブル pg_trigger に表示されます。ストリームトリガーは tgisinternal = true および tgenabled = 'D' としてマークされ、名前サフィックス’_bdrc’または’_bdrt’が付いています。bdr.triggers ビューは、テーブルに関連するトリガー、実行されているプロシージャーの名前、それをトリガーするイベント、およびトリガータイプに関する情報を提供します。

ストリームトリガーは、通常の SQL 処理では有効になっていません。このため、ALTER TABLE ... ENABLE TRIGGER は、特定の名前バリアントと ALL バリアントの両方でストリームトリガーに対してブロックされます。このメカニズムにより、トリガーは通常の SQL トリガーとして実行されません。

これらのトリガーは、ダウンストリームまたはターゲットノードで実行されます。元のノードで実行するオプションはありません。ただし、オリジンで row_filter 式の使用を検討することができます。

また、ストリームトリガーの実行中に適用される DML は、他の PGD ノードにレプリケートされず、標準のローカルトリガーの実行をトリガーしません。これは意図的なものです。たとえば、これを使用して、ストリームトリガーによってキャプチャされた変更または競合を、クラッシュセーフでそのノードに固有のテーブルにログ記録できます。実用的な例については、ストリームトリガーの例 を参照してください。

### 7.9.1 適用中のトリガー実行

変換トリガーは、トリガーテーブルに着信する変更ごとに、最初に_ 1 回実行されます。これらのトリガーは、一致するターゲット行を見つけようとする前に起動するため、非常に幅広い変換を効率的かつ一貫して適用できます。

次に、UPDATE および DELETE 変更の場合、ターゲット行を見つけます。ターゲット行がない場合、これらの変更タイプに対してそれ以上の処理は発生しません。

次に、テーブルレベルでレプリカトリガーとして明示的に有効にされていたノーマルトリガーを実行します。

```
ALTER TABLE tablename
ENABLE REPLICA TRIGGER trigger_name;
```

次に、潜在的な競合が存在するかどうかを判断します。その場合、そのテーブルに存在する競合トリガーを呼び出します。

#### 欠落列の競合の解決

変換トリガーが実行される前に、PostgreSQL は着信タプルをターゲットテーブルの行タイプと照合しようとします。

入力行には存在するがターゲットテーブルには存在しない列は、target_column_missing タイプの競合をトリガーします。逆に、ターゲットテーブルには存在するが、入力行にない列は、source_column_missing 競合をトリガーします。これらの 2 つの競合タイプのデフォルトの解決策は、それぞれ ignore_if_null および use_default_value です。

これは、ローリングスキーマのアップグレードのコンテキストで関連しますたとえば、スキーマの新しいバージョンが新しい列を導入する場合。古いバージョンのスキーマから新しいバージョンにレプリケートする場合、ソース列が欠落しており、新しく導入された列にデフォルト値を移入する use_default_value ストラテジーが適切です。

ただし、新しいスキーマバージョンを持つノードから古いスキーマバージョンを持つノードに複製する場合、ターゲットテーブルから列が失われます。ignore_if_null リゾルバーは、ユーザーがアップグレードされたノードの新しい列に NULL 以外の値を含むタプルを挿入するとすぐにレプリケーションを中断するため、ローリングアップグレードには適していません。

この例を考慮すると、ローリングスキーマアップグレードの適切な設定は、target_column_missing 競合が発生した場合に ignore リゾルバーを適用するように各ノードを構成することです。

これは、次のクエリーで行うことができます。これは、各ノードで個別に実行する必要があります。node1 を実際のノード名に置き換えます。

```
SELECT bdr.alter_node_set_conflict_resolver(node1,
    target_column_missing, ignore);
```

#### データ損失と発散リスク

競合リゾルバーを ignore に設定すると、データの損失とクラスターの発散が発生する可能性があります。

次の例を考えます。テーブル t はノード 1 および 2 に存在しますが、その列 col はノード 1 にのみ存在します。

競合リゾルバーが ignore に設定されている場合、ノード 1 に c が null でない行、たとえば (pk=1, col=100) が存在する場合があります。その行はノード 2 にレプリケートされ、列 c の値は破棄されますたとえば、(pk=1) 。

列 c がノード 2 のテーブルに追加される場合、最初に既存のすべての行で NULL に設定され、上記で検討される行は (pk=1, col=NULL) になります。 pk=1 を持つ行はすべてのノードで同一ではなくなり、したがって、クラスターは発散します。

ノード 2 にレプリケートされた行には col=NULL があるため、デフォルトの ignore_if_null リゾルバーはこのリスクの影響を受けません。

この例に基づいて、 ignore リゾルバーが使用されたローリングスキーマアップグレードの最後に、クラスター全体に対して LiveCompare を実行することをお勧めします。この実践は、発散を確実に検出および修正するのに役立ちます。

### 7.9.2 行型の用語

PGD は次の行タイプを使用します。

- SOURCE_OLD は更新前の行、つまりキーです。
- SOURCE_NEW は、別のノードからの新しい行です。
- TARGET は、ノードに既に存在する行、つまり競合する行です。

### 7.9.3 競合トリガー

競合トリガーは、PGD によって競合が検出されると実行されます。彼らは、紛争が発生したときに何が起こるかを決定します。

- トリガーファンクションが行を返した場合、アクションはターゲットに適用されます。
- トリガーファンクションが NULL 行を返した場合、アクションはスキップされます。

たとえば、DELETE に対してトリガーが呼び出された場合、DELETE をスキップする場合、トリガーは NULL を返します。 DELETE を続行する場合は、行値を返します。SOURCE_OLD または TARGET のいずれかが機能します。競合するオペレーションが INSERT または UPDATE で、選択された解決が競合する行を削除することである場合、トリガーは明示的に削除を実行し、NULL を返す必要があります。トリガーファンクションは、選択した他の SQL アクションを実行できますが、これらのアクションはローカルにのみ適用され、レプリケートされません。

2 つ以上のノード間で実際のデータの競合が発生すると、2 つ以上の変更が同時に発生します。変更が適用されると、競合の解決は各ノードで独立して発生します。これは、競合の解決が各ノードで 1 回発生し、それらの間で大幅な時間差で発生する可能性があることを意味します。その結果、競合トリガーの複数の実行間の通信は不可能です。トリガーが関連するすべてのイベントに対してまったく同じ結果を与えることを確認するのは、競合トリガーの作成者の責任です。そうしないと、データの相違が発生します。

同じ競合トリガーが複数のイベントと一致する場合、トリガーで TG_OP 変数を使用して、競合を生成した操作を特定できます。

デフォルトでは、PGD は行のレプリケーションオリジンの変更を監視することにより競合を検出します。したがって、1 つの変更のみが発生している場合でも、競合トリガーを呼び出すことができます。この場合、実際の競合は存在しないため、この競合検出メカニズムは偽陽性の競合を生成する可能性があります。競合トリガーは、これらをすべて同様に処理する必要があります。

場合によっては、タイムスタンプ競合検出は競合をまったく検出しない場合があります。たとえば、UPDATE の直後に DELETE が発生するコンカレント UPDATE /DELETE では、最初に UPDATE を、次に DELETE を参照するノードは競合を認識しません。競合が発生しない場合、競合トリガーは呼び出されません。同じ状況で、行バージョンの競合検出を使用すると、競合が表示され、競合トリガーで処理できます。

トリガーファンクションは、操作タイプに応じて、追加の状態情報と競合に関係するデータ行にアクセスします。

- INSERT では、競合トリガーはソースから SOURCE_NEW 行と TARGET 行にアクセスできます。
- UPDATE では、競合トリガーはソースから SOURCE_OLD および SOURCE_NEW 行と TARGET 行にアクセスできます。
- DELETE では、競合トリガーはソースから SOURCE_OLD 行と TARGET 行にアクセスできます。

そのオペレーションに値が存在する場合、ファンクション bdr.trigger_get_row() を使用して、SOURCE_OLD 、SOURCE_NEW 、または TARGET 行を取得できます。

競合トリガーの変更はトランザクション的に発生し、構成変更のレプリケーション中にグローバル DML ロックによって保護されます。この動作は、ALTER TABLE の一部のバリアントの処理方法に似ています。

競合トリガー内で主キーが更新されると、実行タイミングの違いにより一意制約違反エラーが発生する場合があります。したがって、競合トリガーでの主キーの更新は避けてください。

### 7.9.4 Transform トリガー

これらのトリガーは、特定のテーブルに対するデータストリームのすべての行に対して実行される点を除き、競合トリガーに似ています。戻り値と公開変数の動作は似ていますが、変換トリガーはターゲット行が識別される前に実行されるため、TARGET 行はありません。

PGD の各テーブルで複数の変換トリガーを指定できます。変換トリガーはアルファベット順に実行されます。

変換トリガーは行をフィルターして取り除くことができ、必要に応じて追加の操作を実行できます。列の値を変更したり、NULL に設定したりできます。戻り値によって、実行される次のアクションが決まります。

- トリガーファンクションが行を返した場合、ターゲットに適用されます。
- トリガーファンクションが NULL 行を返した場合、それ以上に実行するアクションはありません。未実行のトリガーは実行されません。
- トリガー機能は、自分の選択に応じて他のアクションを実行できます。

トリガーファンクションは、追加の状態情報と、競合に関係する行にアクセスします。

- INSERT では、変換トリガーはソースから SOURCE_NEW 行にアクセスできます。

- UPDATE では、変換トリガーはソースから SOURCE_OLD および SOURCE_NEW 行にアクセスできます。
- DELETE では、変換トリガーはソースから SOURCE_OLD 行にアクセスできます。

ファンクション bdr.trigger_get_row() を使用して、SOURCE_OLD または SOURCE_NEW 行を取得できます。このタイプのトリガーは、このようなターゲット行が存在する場合が識別される前に実行されるため、TARGET 行は使用できません。

変換トリガーは通常の BEFORE 行トリガーに非常に似ていますが、次の重要な違いがあります。

- 変換トリガーは、入ってくる変更ごとに呼び出されます。 UPDATE および DELETE では、BEFORE トリガーはまったく呼び出されず、テーブル内で一致する行が見つからない場合、DELETE が変更されます。
- 変換トリガーは、パーティションテーブルルーティングが発生する前に呼び出されます。
- 変換トリガーは、SOURCE_OLD を介してルックアップキーにアクセスできます。これは、通常の SQL トリガーでは使用できません。

### 7.9.5 行の内容

SOURCE_NEW 、SOURCE_OLD 、および TARGET の内容は、オペレーション、テーブルの REPLICA IDENTITY 設定、およびターゲットテーブルの内容によって異なります。

TARGET 行は、競合トリガーでのみ使用できます。TARGET 行には、ターゲットテーブルで UPDATE または DELETE を適用するときに行が見つかった場合にのみデータが含まれます。行が見つからない場合、TARGET は NULL です。

### 7.9.6 実行オーダー

トリガーの実行順序

- Transform トリガー _ ターゲットへの入力行ごとに 1 回実行します。
- 通常トリガー _ 行ごとに 1 回実行します。
- 競合トリガー _ 競合が存在する行ごとに 1 回実行します。

### 7.9.7 ストリームトリガーの例

update_if_newer 競合リゾルバーと同様の動作を提供する競合トリガー

```
CREATE OR REPLACE FUNCTION update_if_newer_trig_func
RETURNS TRIGGER
LANGUAGE plpgsql
AS $$
BEGIN
    IF (bdr.trigger_get_committs(TARGET) >
        bdr.trigger_get_committs(SOURCE_NEW)) THEN
    RETURN TARGET;
    ELSIF
```

(次のページに続く)

(前のページからの続き)

```
        RETURN SOURCE;
    END IF;
END;
$$;
```

カウンター列にデルタ変更を適用し、他のすべての列に SOURCE_NEW を使用する競合トリガー。

```
CREATE OR REPLACE FUNCTION delta_count_trg_func
RETURNS TRIGGER
LANGUAGE plpgsql
AS $$
DECLARE
    DELTA bigint;
    SOURCE_OLD record;
    SOURCE_NEW record;
    TARGET record;
BEGIN
    SOURCE_OLD := bdr.trigger_get_row(SOURCE_OLD);
    SOURCE_NEW := bdr.trigger_get_row(SOURCE_NEW);
    TARGET := bdr.trigger_get_row(TARGET);

    DELTA := SOURCE_NEW.counter - SOURCE_OLD.counter;
    SOURCE_NEW.counter = TARGET.counter + DELTA;

    RETURN SOURCE_NEW;
END;
$$;
```

すべての変更を適用する代わりにログテーブルにログを記録する変換トリガー

```
CREATE OR REPLACE FUNCTION log_change
RETURNS TRIGGER
LANGUAGE plpgsql
AS $$
DECLARE
    SOURCE_NEW record;
    SOURCE_OLD record;
    COMMITTS timestamptz;
BEGIN
    SOURCE_NEW := bdr.trigger_get_row(SOURCE_NEW);
    SOURCE_OLD := bdr.trigger_get_row(SOURCE_OLD);
    COMMITTS := bdr.trigger_get_committs(SOURCE_NEW);

    IF (TG_OP = INSERT) THEN
        INSERT INTO log SELECT I, COMMITTS, row_to_json(SOURCE_NEW);
```

(次のページに続く)

(前のページからの続き)

```
    ELSIF (TG_OP = UPDATE) THEN
        INSERT INTO log SELECT U, COMMITTS, row_to_json(SOURCE_NEW);
    ELSIF (TG_OP = DELETE) THEN
        INSERT INTO log SELECT D, COMMITTS, row_to_json(SOURCE_OLD);
    END IF;

    RETURN NULL; -- do not apply the change
END;
$$;
```

この例は、信頼できるサイト、優先ノード、または Always Wins 解決としても知られる、信頼できるソースの競合検出を実装する競合トリガーを示しています。bdr.trigger_get_origin_node_id() ファンクションを使用して、3 つ以上のノードで動作するソリューションを提供します。

```
CREATE OR REPLACE FUNCTION test_conflict_trigger()
RETURNS TRIGGER
LANGUAGE plpgsql
AS $$
DECLARE
    SOURCE  record;
    TARGET  record;

    TRUSTED_NODE    bigint;
    SOURCE_NODE     bigint;
    TARGET_NODE     bigint;
BEGIN
    TARGET := bdr.trigger_get_row(TARGET);
    IF (TG_OP = DELETE)
        SOURCE := bdr.trigger_get_row(SOURCE_OLD);
    ELSE
        SOURCE := bdr.trigger_get_row(SOURCE_NEW);
    END IF;

    TRUSTED_NODE := current_setting(customer.trusted_node_id);

    SOURCE_NODE := bdr.trigger_get_origin_node_id(SOURCE_NEW);
    TARGET_NODE := bdr.trigger_get_origin_node_id(TARGET);

    IF (TRUSTED_NODE = SOURCE_NODE) THEN
        RETURN SOURCE;
    ELSIF (TRUSTED_NODE = TARGET_NODE) THEN
        RETURN TARGET;
    ELSE
        RETURN NULL; -- do not apply the change
    END IF;
```

(次のページに続く)

(前のページからの続き)

```
END;
$$;
```

## 7.10 Transaction streaming

論理レプリケーションでは、トランザクションはパブリッシャーで同時にデコードされますが、トランザクションがコミットされるまでサブスクライバーに送信されません。変更が `logical_decoding_work_mem` PostgreSQL 13 以降を超える場合、それらはディスクにあふれます。これは、特に大規模なトランザクションの場合、サブスクライバーに到達するまでにある程度の遅延が発生し、パブリッシャーで追加の I/O を必要とする場合があることを意味します。

PostgreSQL 14 以降、パブリッシャーでコミットされる前に、オプションでトランザクションをデコードして、サブスクライバーに送信できます。サブスクライバは、着信した変更をステージングファイルまたはファイルセットに保存し、トランザクションがコミットするときに適用します、またはトランザクションがアボートする場合は破棄します。これにより、トランザクションがコミットされるとすぐに、サブスクライバーにトランザクションを適用できるようになります。

### 7.10.1 PGD の機能強化

PostgreSQL のビルトイントランザクションストリーミングには、次の制限があります。

- パブリッシャーのディスクに変更をスピルする必要はありませんが、各サブスクライバーのディスクに変更を書き込む必要があります。
- トランザクションが中止される場合、作業各サブスクライバが受け取った変更および関連するストレージ I/O は無駄になります。

ただし、PGD はパラレル適用をサポートし、各サブスクライバーでマルチプルのライタプロセスを有効にします。この機能を活用して、次の機能強化を提供します。

- デコードされたトランザクションは、サブスクライバーのライタに直接ストリーミングできます。
- デコードされたトランザクションは、サブスクライバーのディスクに保存する必要はありません。
- サブスクライバーへのトランザクションの適用を開始する前に、トランザクションがコミットされるのを待機する必要はありません。

#### 注意事項

- Parallel apply を有効にする必要があります。
- 多数の小規模で競合するトランザクションで構成されるワークロードは、ライター間で頻繁なデッドロックを発生させる可能性があります。

**注釈**

ライタへの直接ストリーミングはまだ実験的な機能です。注意して使用してください。具体的には、ストリーミングのコミットタイムスタンプを利用できない可能性があるため、競合解決ではうまく機能しない可能性があります。トランザクションはオリジンでまだコミットされていない可能性があります。

### 7.10.2 構成

2 つの場所でトランザクションストリーミングを構成します。

- ノードレベルで、 GUC bdr.default_streaming_mode を使用
- グループレベルで、ファンクション bdr.alter_node_group_option() を使用

#### bdr.default_streaming_mode を使用したノード構成

許可される値は次のとおりです。

- `off`
- `writer`
- `file`
- `auto`

デフォルト値は `auto` です。

この設定の変更を有効にするには、サブスクリプションごとに pglogical レシーバープロセスを再起動します。

これは、サーバーを再起動することで実現できます。

`bdr.default_streaming_mode` が `off` 以外の値に設定されている場合、サブスクライバーはパブリッシャーにトランザクションストリーミングを要求します。これが提供される方法は、グループ構成設定によっても異なります。詳細は、*bdr.default_streaming_mode* **を使用したノード構成** を参照してください。

#### bdr.alter_node_group_option() を使用したグループ構成

ファンクションでパラメーター `streaming_mode` を使用できます bdr.alter_node_group_option()

グループトランザクションストリーミング構成を設定します。

許可される値は次のとおりです。

- `off`
- `writer`
- `file`
- `auto`
- `default`

デフォルト値は default です。

現在の設定の値は、ビュー bdr.node_group の列 node_group_streaming_mode に含まれています。返される値は単一の char 型で、使用可能な値は D default 、W writer 、 F file 、 A auto 、および O off です。

### 構成設定の効果

トランザクションストリーミングは、GUC bdr.default_streaming_mode によってサブスクライバーレベルで制御されます。トランザクションストリーミングを無効にする off に設定されていない限り、サブスクライバーはトランザクションストリーミングを要求します。

パブリッシャーがトランザクションストリーミングを提供できる場合、トランザクションサイズが logical_decoding_work_mem で設定されたしきい値を超えるたびに、トランザクションをストリーミングします。通常、パブリッシャーは、トランザクションがファイルにストリーミングされるか、ライタにストリーミングされるかを制御できません。COPY などの一部の状況を除き、サブスクライバーがトランザクションをライタにストリーミングすることを示唆する場合があります可能であれば。

サブスクライバーは、パブリッシャーから受け取ったトランザクションをライタまたはファイルにストリーミングできます。決定はいくつかの要素に基づいて行われます。

- Parallel apply がオフ num_writers = 1 の場合、ファイルにストリーミングされます。ライタ 0 は常に非ストリーミングトランザクション用に予約されています。
- Parallel apply がオンになっているが、すべてのライターがすでにビジー状態である場合、新しいトランザクションはファイルにストリーミングされます。PGD ライタのステータスを確認するには、*PGD* ライターのモニタリング を参照してください。

ライタへのストリーミングが可能である場合、つまりフリーライタが利用可能な場合、トランザクションをライタにストリーミングするかファイルにストリーミングするかは、次の表にあるグループとノード設定の組み合わせに基づいて決定されます。

| Group | Node | Streamed to |
|---|---|---|
| off | (any) | (none) |
| (any) | off | (none) |
| writer | file | file |
| file | writer | file |
| default | writer | writer |
| default | file | file |
| default | auto | writer |
| auto | (any) | writer |

グループ構成が auto に設定されている場合、またはグループ構成が default でノード構成が auto である場合、パブリッシャーがこれを行うことをヒントにした場合にのみ、トランザクションはライタにストリーミングされます。

現在、パブリッシャーは、サブスクライバーが次のトランザクションタイプのライタにストリーミングすることを示唆しています。これらは競合しないことが知られており、ライタが安全に処理できます。

- COPY

- CREATE INDEX CONCURRENTLY

### 7.10.3 モニタリング

bdr.stat_subscription を使用して、トランザクションストリーミングの使用をモニタできます。

サブスクライバノード上のファンクション。

- nstream_writer _ライタにストリーミングされたトランザクションの数。
- nstream_file _ ファイルにストリーミングされたトランザクションの数。
- nstream_commit _ コミットされたストリーミングトランザクションの数。
- nstream_abort _ 中止されたストリーミングトランザクションの数。
- nstream_start _ 開始されたストリーミングトランザクションの数。
- nstream_stop _ 完全に受信されたストリーミングトランザクションの数。

## 7.11 Explicit two-phase commit (2PC)

!!!注

2 フェーズコミットは、 Quorum Commit、 Group Commit、または CAMO では使用できません。 *Commit scope limitations* を参照してください。

X/Open 分散トランザクション処理 DTP モデルは、3 つのソフトウェアコンポーネントを想定しています。

- トランザクション境界を定義し、トランザクションを構成するアクションを指定するアプリケーションプログラム AP
- 共有リソースへのアクセスを提供するデータベースやファイルアクセスシステムなどのリソースマネージャー RM
- トランザクションに識別子を割り当て、その進行状況を監視し、トランザクションの完了と障害からの回復に対して責任を負うトランザクションマネージャー TM と呼ばれる別のコンポーネント

PGD は、 PREPARE TRANSACTION および COMMIT PREPARED /ROLLBACK PREPARED コマンドを使用して明示的な外部 2PC をサポートします。外部的には、 EDB Postgres 分散クラスターは、単一のセッションのトランザクションマネージャーに対する単一のリソースマネージャーであるように見えます。

bdr.commit_scope が local の場合、トランザクションはローカルノードでのみ準備されます。コミットされると、変更がレプリケートされ、PGD がコミット後の競合解決を適用します。

明示的な 2 フェーズコミットでは、 local に設定して bdr.commit_scope を使用するのは意味がないと思われる場合があります。ただし、このオプションは、トランザクションのレイテンシーと堅牢性の間のトレードオフを制御できるように提供されています。

これらのコミットスコープ種類は内部的に 2 フェーズコミットを使用するため、明示的な 2 フェーズコミットは Quorum Commit、Group Commit、または CAMO では機能しません。明示的な 2 フェーズコミットは、*Parallel Apply* または **トランザクションストリーミングでの使用** でもうまく機能しません。明示的な 2 フェーズコミットを使用する前に、両方を無効にします。

### 7.11.1 使用する

ローカルコミットスコープの 2 フェーズコミットは、標準の PostgreSQL とまったく同じように動作します。ローカルコミットスコープを使用します。

```
BEGIN;
SET LOCAL bdr.commit_scope = local;

... other commands possible...
```

コミットの最初のフェーズを開始するには、クライアントはグローバルトランザクション ID を割り当てる必要があります。これは、トランザクションを識別する一意の文字列です。

```
PREPARE TRANSACTION some-global-id;
```

最初のフェーズが成功した後、すべてのノードが変更を適用し、トランザクションをコミットする準備が整います。次に、クライアントは同じノードから第 2 フェーズを呼び出す必要があります。

```
COMMIT PREPARED some-global-id;
```

# 第8章　Guidance for DBAs

## 8.1 Encrypting columns in PGD

列の暗号化は、テーブルの機密フィールドをスクランブルされた読み取り不能形式の暗号テキストに変換することにより保護しますが、レコードの残りの部分は元の人間が判読可能な形式のままにします。暗号化は `INSERT` または `UPDATE` 中に SQL レベルで発生するため、暗号文のみが書き込み先行ログ WAL に到達し、PGD ノード全体でレプリケートされ、保存、バックアップ、および転送中の機密データが保護されたままになります。

分散クラスターでは、列暗号化は次の利点も提供します。

- **ノードの独立性**アプリケーションは、一貫した復号化のために任意のノードにキーを提供できます。
- **攻撃面の削減**キーが平文で WAL に入力されることはありません。暗号文のみが複製されます。
- **メモリ分離** セッションが終了すると、キーはクリアされます。

PGD は、次を介して列レベルの暗号化をサポートしています。

- ** pgcrypto ** コミュニティ Postgres 拡張機能。EDB パッケージとしては出荷されませんが、PostgreSQL、EDB Postgres ExtendedPGE、および EDB Postgres Advanced Server で使用する場合にサポートされています。
- ** DBMS_CRYPTO ** EPAS でのみ入手可能な Oracle 互換パッケージ。

どちらも PGD レプリケーションと完全な互換性があります。 `pgcrypto` または `DBMS_CRYPTO` を使用する場合、データはヒープに書き込まれる前に暗号化ファンクションを使用して変換されます。pgcrypto は、結果の暗号文を `bytea` として保存し、DBMS_CRYPTO は `RAW` 、`BLOB` 、または `CLOB` を使用します。これらはすべて PGD がネイティブにレプリケートできます。

### 8.1.1 暗号化方法の比較

列の暗号化は、PGD のデータを保護するためのいくつかのアプローチの 1 つです。それぞれがスタックの異なるレベルで動作し、さまざまな脅威から保護します。

Transparent Data Encryption (TDE) は、ストレージレベルでデータを暗号化する PGE および EPAS のオプション機能であり、暗号化キーがなしではディスク上のデータベースファイルを読み取ることができません。TDE は PGD レプリケーションと並行して透過的に動作しますが、実行中のデータベースにアクセスし、すべてのデータをプレーンテキストで照会できる DBA に対する保護はありません。

Data redaction は、ポリシーに基づいてクエリ結果の機密値をマスクする EPAS の機能です。ディスク上の基になるデータは変更されないため、TDE または列暗号化と組み合わせて、意味のある保護を提供する必要があります。

列の暗号化はフィールドレベルで保護します。これは、上記のいずれよりも詳細です。セッションレベルの復号キーを持たない DBA は、ライブクエリーであっても、暗号文のみを参照します。トレードオフは、アプリケーションがキーを管理し、暗号化と復号化を明示的に処理する必要があることです。

| Technique | Protection level | Key managed by | Protects against | Visible to DBAs |
|---|---|---|---|---|
| **TDE** | Block/storage-level | Database/KMIP | Physical disk theft or OS-level access | Yes |
| **Data redaction** | Presentation-level | Database policy | Unauthorized app users | Yes |
| **Column encryption** | Field-level | Application | Compromised DBAs or system admins | No (ciphertext only) |

### 8.1.2 分散クラスターでの暗号化の計画

PGD で列を暗号化すると、シングルノード設定では発生しない考慮事項が発生します。列の暗号化を実装する前に、インデックス作成、CPU オーバーヘッド、およびストレージとレプリケーションへの影響を考慮してください。

#### 暗号化された列のインデックス作成と検索の処理

pgcrypto と DBMS_CRYPTO は両方とも、デフォルトで確率的ランダム化暗号化を使用します。つまり、同じ平文を複数回暗号化すると、毎回異なる暗号文が生成されます。pgcrypto は `pgp_sym_encrypt()` を介してこの動作を実現し、DBMS_CRYPTO は CBC モードの AES を介してこれを行います。

ランダム化された暗号化には 2 つの重要な意味があります。

- **インデックスの非互換性**暗号テキスト列の標準の B ツリーインデックスは、プレーンテキストの等価ルックアップをサポートできません。
- **クエリーの失敗** 次の一般的なパターンは、`pgp_sym_encrypt()` への各呼び出しが列に格納されているものと一致しない異なる暗号文を生成するため、レコードが存在する場合でも結果ゼロを返します。エラーは発生しないため、失敗は存在しないレコードと間違われやすくなります。

```
SELECT * FROM customers
WHERE email = pgp_sym_encrypt(alice@example.com, your_key);
```

特に pgcrypto では、列タイプが `bytea` ではなく `text` である場合、Postgres は代わりに型不一致エラーを発生させます。

```
ERROR: operator does not exist: text = bytea
```

正確な値で暗号化された列を検索するには、次のいずれかの戦略を使用します。

- **決定的暗号化**特定の平文に対して常に同じ暗号文を生成するように暗号化アルゴリズムを構成します。
- **ブラインドインデックス推奨** pgcrypto の `hmac()` のようなファンクションを使用して、平文の一方通行暗号ハッシュを含む追加列を保存し、代わりにその列にインデックスを付けます。クエリは検索値をハッシュし、暗号化された列ではなくインデックスと比較します。

両方のアプローチの実装の詳細については、

pgcrypto documentation または *DBMS_CRYPTO reference* を参照してください。

> **警告**
>
> どちらのアプローチも等価パターンを明らかにします。格納された値を観察できる攻撃者は、2 つの行に同じデータが含まれる場合を推論でき、状態フラグや国コードのような低品種フィールドでは、平文を完全に推定できます。

#### CPU オーバーヘッドの考慮

暗号化と復号化は、データベースサーバーで実行される計算量の多い操作です。ノードに十分な CPU ヘッドルームがあることを確認します。AES-NI 命令セットをサポートするプロセッサーの AES など、ハードウェアアクセラレーションを活用するアルゴリズムを使用することにより、この影響を大幅に削減できます。

#### ストレージと WAL の成長の計画

暗号文は元の平文より大きいため、テーブルサイズ、WAL ボリューム、および PGD がノードを同期するために必要なレプリケーション帯域幅が増加します。それに応じて容量を計画します。

### 8.1.3 pgcrypto を使用したカラムの暗号化

pgcrypto は、`pgp_sym_encrypt()` や `pgp_sym_decrypt()` などの暗号化機能を提供するコミュニティ Postgres 拡張機能です。PostgreSQL、PGE、および EPAS で動作します。ビューは透過的な復号を処理し、トリガーは書き込み時に自動暗号化を処理するため、アプリケーションは暗号化機能を直接呼び出さずに平文の読み取りおよび書き込みができます。

1. 拡張機能を作成します。

```
CREATE EXTENSION IF NOT EXISTS pgcrypto;
```

1. `bytea` として定義された暗号化列を使用して `customers` テーブルを作成します。

```
CREATE TABLE customers (
    customer_id   uuid    DEFAULT uuidv7() PRIMARY KEY,
    username      TEXT    NOT NULL,
    email         BYTEA   NOT NULL
);
```

1. 透過的な復号化のためのビューを作成します。

```
-- Note: requires SET app.pgc_key = your_key to be run in the session first
CREATE OR REPLACE VIEW customers_view AS
SELECT
    customer_id,
    username,
    pgp_sym_decrypt(email, current_setting(app.pgc_key)) AS email
FROM customers;
```

1. データが書き込まれる前に自動的に暗号化するトリガーを作成します。

```
CREATE OR REPLACE FUNCTION encrypt_customer_pii_trigger()
RETURNS TRIGGER AS $$
BEGIN
    IF current_setting(app.pgc_key, true) IS NULL OR current_setting(app.pgc_key) = ␣
↪THEN
        RAISE EXCEPTION Encryption key not set. Please run SET app.pgc_key = ...;
    END IF;

    IF (TG_OP = INSERT OR NEW.email != OLD.email) THEN
        NEW.email := pgp_sym_encrypt(
            convert_from(NEW.email, utf8),
            current_setting(app.pgc_key),
            cipher-algo=aes256
        );
    END IF;
    RETURN NEW;
END;
$$ LANGUAGE plpgsql;

CREATE TRIGGER trg_encrypt_customer_pii
BEFORE INSERT OR UPDATE ON customers
FOR EACH ROW EXECUTE FUNCTION encrypt_customer_pii_trigger();
```

1. ビューとトリガーを配置したら、セッションキーを設定し、通常のようにテーブルと対話します。

```
-- 1. Initialize session key
SET app.pgc_key = your_secret_key;

-- 2. Trigger handles encryption automatically
INSERT INTO customers (username, email) VALUES (dolores, dolores@example.com);
```

生の customers テーブルを照会すると、 email 列が暗号文を格納するのに対し、 customers_view は復号化された平文を返すことがわかります。

```
SELECT customer_id, username, email FROM customers;
__OUTPUT__
 customer_id                          | username | email
- -------------------------------------+----------+-----------------------------------
↪------------
 01966e42-1af7-7a2a-b4b9-9cf8e0d3c2a | dolores  | \xc30d04070302e4a7f23b8c12d956aa...
(1 row)
```

```
SELECT customer_id, username, email FROM customers_view;
__OUTPUT__
 customer_id                          | username | email
- -------------------------------------+----------+--------------------
```

(次のページに続く)

(前のページからの続き)

```
 01966e42-1af7-7a2a-b4b9-9cf8e0d3c2a | dolores   | dolores@example.com
(1 row)
```

### 8.1.4 DBMS_CRYPTO を使用した列の暗号化

DBMS_CRYPTO は、EPAS でのみ使用可能な Oracle 互換パッケージです。 Oracle DBMS_CRYPTO API をミラーする暗号化および復号化機能を提供します。pgcrypto と同様に、ビューは透過的な復号化を処理し、トリガーは書き込み時に自動暗号化を処理します。

1. RAW として定義された暗号化列を使用して customers テーブルを作成します。

```
CREATE TABLE customers (
    customer_id   uuid     DEFAULT uuidv7() PRIMARY KEY,
    username      TEXT     NOT NULL,
    email         RAW(128)
);
```

1. 透過的な復号化のためのビューを作成します。

```
CREATE OR REPLACE VIEW customers_view AS
SELECT
    customer_id,
    username,
    UTL_I18N.RAW_TO_CHAR(
        DBMS_CRYPTO.DECRYPT(
            src => email,
            typ => 258 + 256 + 4096, -- AES256 + CBC + PKCS7
            key => UTL_I18N.STRING_TO_RAW(current_setting(app.pgc_key), AL32UTF8)
        ),
        AL32UTF8
    ) AS email
FROM customers;
```

1. データが書き込まれる前に自動的に暗号化するトリガーを作成します。

```
CREATE OR REPLACE FUNCTION epas_encrypt_customer_pii_trigger()
RETURNS TRIGGER AS $$
DECLARE
    l_typ INTEGER := 258 + 256 + 4096; -- AES256 + CBC + PKCS7
    l_key RAW(32) := UTL_I18N.STRING_TO_RAW(current_setting(app.pgc_key), AL32UTF8);
BEGIN
    IF (TG_OP = INSERT OR NEW.email IS DISTINCT FROM OLD.email) THEN
        NEW.email := DBMS_CRYPTO.ENCRYPT(
            src => UTL_I18N.STRING_TO_RAW(convert_from(NEW.email, utf8), AL32UTF8),
            typ => l_typ,
```

(次のページに続く)

(前のページからの続き)

```
            key => l_key
        );
    END IF;
    RETURN NEW;
END;
$$ LANGUAGE plpgsql;

CREATE TRIGGER trg_epas_encrypt_customer_pii
BEFORE INSERT OR UPDATE ON customers
FOR EACH ROW EXECUTE FUNCTION epas_encrypt_customer_pii_trigger();
```

1. ビューとトリガーを配置したら、セッションキーを設定し、通常のようにテーブルと対話します。

```
-- 1. Initialize session key
SET app.pgc_key = my_secure_32_byte_key_1234;

-- 2. Trigger handles encryption automatically
INSERT INTO customers (username, email) VALUES (dolores, dolores@example.com);
```

生の customers テーブルを照会すると、 email 列が暗号文を RAW 16 進文字列として保存するのに対し、customers_view は復号化された平文を返すことがわかります。

```
SELECT customer_id, username, email FROM customers;
__OUTPUT__
 customer_id                          | username | email
- -------------------------------------+----------+-----------------------------------
↪------------
 01966e42-1af7-7a2a-b4b9-9cf8e0d3c2a | dolores  | A3F2C8D14E7B0591...
(1 row)
```

```
SELECT customer_id, username, email FROM customers_view;
__OUTPUT__
 customer_id                          | username | email
- -------------------------------------+----------+--------------------
 01966e42-1af7-7a2a-b4b9-9cf8e0d3c2a | dolores  | dolores@example.com
(1 row)
```

### 8.1.5 暗号化キーを安全に管理する

PGD はデータを複製しますが、キーは複製しません。つまり、すべてのノードが同じキーマテリアルにアクセスする必要があります。キーを安全に管理するには、次のプラクティスに従ってください。

- **ハードコーディングのキーを避ける** キーは pg_catalog に保存され、システムカタログへのアクセスを持つすべてのユーザーに表示されるため、SQL スクリプトまたは CREATE VIEW 定義に含めないでください。

- **セッションレベルの構成パラメーターを使用して、キーを渡します。**セッションレベルで設定すると、キーがメモリにのみ存在し、セッションが閉じるときにクリアされることを意味します。パラメーター名には、app.* 名前空間内の任意の名前を指定できます。例 app.pgc_key

```
SET app.pgc_key = your_ultra_secure_aes_256_key_here;
```

**重要**

構成パラメーターは、コマンドが実行された特定のノードとセッションにローカルです。 PGD はマルチノードクラスターとして動作するため、アプリケーションは、現在の接続を処理する特定のノードに`SET` コマンドが発行されていることを確認する必要があります。アプリケーションがノードを切り替える場合たとえば、接続マネージャーを介して、暗号化されたデータの読み取りまたは書き込みを試行する前に、新しいノードでパラメーターを再設定する必要があります。 !!!

- **プロダクション展開には外部キー管理サービスを使用します。** HashiCorp Vault や AWS KMS などのサービスを使用すると、アプリケーションは実行時にキーを取得し、セッション中に PGD に提供できるため、キーがローカルに永続化されることはありません任意の PGD ノード上のディスク。

## 8.2 Using tiered storage

PGD は、Postgres のベクトル化クエリエンジンおよびクラウドネイティブストレージ拡張機能である Postgres Analytics Accelerator (PGAA) との統合を介して、階層化ストレージをサポートします。PGAA は、Apache Iceberg ®や Delta Lake のような列指向形式を使用してオブジェクトストレージのデータを管理し、ETL パイプラインを必要とせずに照会可能な Postgres テーブルとして公開します。

PGD と PGAA は一緒に、統合アーキテクチャ内で完全なデータライフサイクルを管理し、2 つの異なるストレージ階層を作成します。

- ホットティア PGD ミリ秒未満のレイテンシーと完全な ACID コンプライアンスを必要とする現在の操作データ用のローカル PostgreSQL ストレージ。
- コールド層 PGAA 高速分析クエリ用に最適化された Apache Iceberg ®または Delta Lake のようなカラムナフォーマットを使用した費用対効果の高いオブジェクトストレージ。

これらを組み合わせることで、統合された Postgres インターフェイスを介した透過的なアクセスを維持しながら、高性能トランザクションストレージから費用対効果の高いデータレイクにデータをシームレスに移行できます。このアーキテクチャは、インパクトの高いさまざまなユースケースをサポートしています。

- **規制へのコンプライアンス**古いデータと監査証跡をオフロードして、長期的な保持を保証しながら、無駄のない高性能システムを維持します。
- **時系列と IoT** センサーとイベントストリームの費用対効果の高いコールド履歴分析と、ホットな最近のデータを使用したリアルタイムダッシュボードのバランスをとります。
- **金融サービス**トランザクションストレージコストを高騰させることなく、必要な大量の履歴取引レコードを管理します。
- **小売および季節的なビジネスサイクル**適応ポリシーを使用して、古いレコードをコールドストレージに階層化しながら、カスタマーサービスのためにピークシーズンの注文データをホットに保ちます。

### 8.2.1 テーブルタイプを理解する

このアーキテクチャは、2 つの異なるストレージ層ホット層ローカル SSD、ミリ秒未満のレイテンシーとコールド層オブジェクトストレージ、カラムナーフォーマットを実装しています。データは、次の 3 つの異なるテーブルタイプのいずれかに存在します。

| Table type | Storage location | Behavior |
|---|---|---|
| Heap | Local disk only | ホット層の標準のトランザクションテーブル。 |
| Hybrid Transactional and Analytics Processing (HTAP) | Local + object storage | 分析コピーにマッピングされたヒープテーブル。PGD は、すべての変更をほぼリアルタイムで複製します。 |
| PGAA | Object storage only | デルタまたはアイスバーグ形式で利用可能な純粋な分析コールドテーブル。 |

### 8.2.2 階層化ストレージ戦略

運用ニーズに基づいて、層間でデータを移行するには 3 つの使用可能な方法があります。

#### 階層テーブルの実装

この方法では、 **PGD AutoPartition** を活用して、大容量データのライフサイクルを自動化します。ヒープテーブルは、次のパーティション構造に変換されます。

- **ホットパーティション**最近のデータは、アクティブな処理のヒープまたは `enable_replication` が有効になっている場合は HTAP として残ります。
- **コールドパーティション**パーティションが経過期間のしきい値を超えると、PGD はそれらをオブジェクトストレージに大量コピーし、PGAA テーブルに変換し、ローカルディスク領域を再利用して、データを照会可能に維持します。

#### アナリティクスへの複製

このメソッドは、テーブルをオブジェクトストレージ内の分析コピーと継続的に同期します。ヒープテーブルは HTAP テーブルに変換されます。ユーザーには、標準のトランザクションテーブルのように見えますが、分析コピーはバックグラウンドで維持されます。

トランザクションのスループットに負担を与えることなく、ほぼリアルタイムの同期を保証するために、PGAA は Iceberg Merge-on-ReadMoR を使用します。これは、データブロック全体を書き換えるのではなく、小さな「削除ファイル」で変更を記録し、最小限のオーバーヘッドでデータレイクを最新の状態に保ちます。

#### アナリティクスへのオフロード

この方法は、HTAP テーブルのローカルディスク領域をすぐに再利用し、オブジェクトストレージ PGAA テーブルにコピーのみを残す 1 回限りの外科的方法を提供します。ローカルヒープコピーは切り捨てられ、テーブルアクセス方法は `PGAA` に変更されます。これはオブジェクトストレージをポイントするようになりました。

トランザクション書き込み用にデータを戻す必要がある場合は、オブジェクトストレージからローカルの Postgres ヒープにテーブルを復元できます。

| Method | Purpose | Use case |
|---|---|---|
| Tiered tables | Automated lifecycle management | High-volume time-series data where historical retention is critical and automate |
| Replication | Continuous synchronization | Replicating a transactional table in near real-time for general analytics. |
| Offload | Selective archiving | One-time, surgical removal and archival of a specific table. |

### 8.2.3 メタデータ管理

PGAA は、PGD と統合するときにオブジェクトストアのメタデータを管理する 2 つの方法をサポートしています。

- **カタログオフロード** storage locations を使用してメタデータをデータファイルとともに保存し、オブジェクトストレージに直接接続します。
- **カタログ管理オフロード** Iceberg REST などのメタデータ using a centralized catalog service からデータを分離します。

#### Setting up PGAA

PGD の階層化ストレージは、Postgres Analytics AcceleratorPGAA 拡張機能を利用しています。PGAA は、ベクトル化クエリエンジンとクラウドネイティブストレージハンドラーを提供し、PGD がデータパーティションをオブジェクトストレージにオフロードし、インプレースで照会できるようにします。

階層化テーブルの構成、アナリティクスへのレプリケート、またはデータのオフロードを行う前に、クラスター内のすべてのノードに PGAA をインストールし、オブジェクトストレージの宛先を定義する必要があります。ノードグループレベルでストレージターゲットを構成すると、すべてのノードが同じ場所に書き込みおよび読み取ります。

#### 前提条件

- **クラスター** PGAA および PGFS 拡張機能がインストールされた PGD バージョン 6.1 以降。
- **ストレージの場所** Iceberg または Delta Lake 形式を使用するローカル、S3、GCP、または Azure ストレージ。
- **カタログ** 外部カタログサービスを使用する場合、Iceberg REST カタログのみがサポートされます。
- **権限**データベースユーザーには、PGD および PGAA ファンクションに対する `CREATE` 、`ALTER` 、および `EXECUTE` 特権が必要です。

> **注釈**
>
> 保管場所またはカタログサービスに関連付けられている資格情報には、宛先バケットの読み取りおよび書き込み権限が必要です。次のファンクションを実行して、権限を確認できます。成功した場合は `NULL` を結果ます。

- 保存場所の場合

```
SELECT pgaa.test_storage_location (my_storage_location, true);
```

- カタログサービスの場合

```
SELECT pgaa.test_catalog(my_iceberg_catalog, test_writes := FALSE);
```

### PGAA 拡張機能のダウンロードとインストール

PGD クラスター内のすべてのノードに PGAA 拡張機能をインストールします。

1. EDB リポジトリからパッケージをダウンロードします。

```
export EDB_SUBSCRIPTION_TOKEN=<your-token>
export EDB_SUBSCRIPTION_PLAN=<your-subscription-plan>
curl -1sSLf "https://downloads.enterprisedb.com/$EDB_SUBSCRIPTION_TOKEN/$EDB_
↪SUBSCRIPTION_PLAN/setup.deb.sh" | sudo -E bash
sudo apt-get install -y <postgresql-distribution>-pgaa
```

```
export EDB_SUBSCRIPTION_TOKEN=<your-token>
EDB_SUBSCRIPTION_PLAN=<your-subscription-plan>
curl -1sSLf "https://downloads.enterprisedb.com/$EDB_SUBSCRIPTION_TOKEN/$EDB_
↪SUBSCRIPTION_PLAN/setup.rpm.sh" | sudo -E bash
sudo dnf install -y <postgresql-distribution>-pgaa
```

ここで、- $EDB_SUBSCRIPTION_TOKEN は、EDB サブスクリプションの登録時に受け取ったトークンです。 - $EDB_SUBSCRIPTION_PLAN は、standard PGE または Community Postgres の場合または enterprise EPAS の場合。 - <postgresql-distribution> は、Postgres ディストリビューション edb-postgresextended 、edb-as 、または postgres 、およびバージョンを指定します。

1. postgresql.conf ファイルを更新します。 pgaa を shared_preload_libraries に追加すると、Postgres はバックグラウンドワーカーとして Seafowl クエリエンジンのライフサイクルを管理できます。

```
shared_preload_libraries = pgaa
pgaa.autostart_seafowl = on
```

1. Postgres サービスを再起動して、新しいライブラリをロードします。

### オブジェクトストレージをポイントする

コールドデータの宛先を定義します。分散 PGD クラスター内で作業しているため、これらの構成呼び出しを bdr.replicate_ddl_command でラップして、設定がグループ内のすべてのノードに伝播されるようにする必要があります。

!!!ヒント

ノードグループを表示するには、SELECT node_group_id FROM bdr.node_group; を使用します。

1. *PGFS extension* を使用してストレージの場所を作成し、ノードグループの分析ストレージとして設定します。

```
SELECT bdr.replicate_ddl_command($$
  SELECT pgfs.create_storage_location(
  my-storage-location,
  protocol://my-bucket-name/path/,
  {"region": "region-name"},
  {"access_key_id": "...", "secret_access_key": "..."}
  )
$$);
```

1. analytics_write_catalog をクリアします。analytics_storage_location を割り当てる前に、その値が NULL である必要があります。

```
SELECT bdr.alter_node_group_option(my-pgd-group, analytics_write_catalog, NULL);
```

1. ノードグループの分析ストレージの場所を設定します。

```
SELECT bdr.alter_node_group_option(my-pgd-group, analytics_storage_location, my-
↪storage-location);
```

1. external catalog への接続を構成します。

```
SELECT bdr.replicate_ddl_command($$
  SELECT pgaa.add_catalog(
    my-catalog-name,
    iceberg-rest,
    {
      "url": "https://my-catalog-rest-endpoint.com",
      "token": "MY_AUTH_TOKEN",
    }
  )
);
```

1. analytics_storage_location をクリアします。analytics_write_catalog を割り当てる前に、その値が NULL である必要があります。

```
SELECT bdr.alter_node_group_option(my-pgd-group, analytics_storage_location, NULL);
```

1. カタログをデフォルトの分析の書き込み場所として設定します。

```
SELECT bdr.replicate_ddl_command($$
  SELECT bdr.alter_node_group_option(my-pgd-group, analytics_write_catalog, my-
↪catalog-name)
);
```

### 次のステップ

PGD クラスターのストレージターゲットを定義したら、パフォーマンスと容量の要件に基づいて、テーブルがそのストレージとどのように相互作用するかを決定します。

- **階層テーブルの実装**ローカルヒープストレージから経過期間のしきい値に基づいてターゲットとされる分析ストレージにパーティションを自動的に移行することにより、ゼロタッチデータライフサイクルを確立します。
- **アナリティクスへの複製**ローカルトランザクションヒープテーブルを維持すると同時に、大量の分析処理のためにカラムナコピーをストレージターゲットに同期します。
- **アナリティクスへのオフロード**アナリティクスにレプリケートされる既存のテーブルからヒープディスク領域を再利用します。

## Implementing tiered tables

階層テーブルは、 Postgres Distributed PGD AutoPartition と Postgres Analytics Accelerator PGAA を組み合わせて、自動化されたデータライフサイクルを作成することにより、ユニファイドストレージアーキテクチャを提供します。標準ヒープテーブルが階層構造に変換されると、システムは手動介入なしでトランザクションストレージから分析ストレージへのデータの移行を管理し、ストレージコストを最適化するゼロタッチデータライフサイクル管理を提供します。

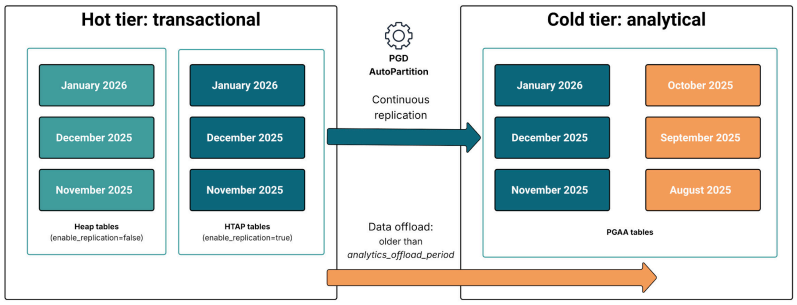


図 1: tiered_tables

このアーキテクチャでは、クエリはホットとコールドの両方のデータ層にシームレスにまたがり、物理的な場所に関係なくデータセット全体への単一アクセスポイントを提供します。最近のデータは、高性能のトランザクション処理のためにローカルヒープパーティションに残りますが、古いコールドパーティションは、定義された経過期間のしきい値を超えると、オブジェクトストレージに透過的にオフロードされます。

`pgaa.convert_to_tiered_table()` を呼び出すことにより、標準ヒープテーブルはハイブリッドパーティション構造に変換されます。

- **ホット層**最近のデータパーティションはローカルヒープパーティションに残るか、レプリケーションが有効になっている場合はハイブリッドトランザクションおよび分析処理 HTAP テーブルになります。HTAP の場合、PGAA は Iceberg Merge-on-ReadMoR を使用して、軽量の「削除ファイル」として更新と削除をキャプチャします。これにより、大規模なデータブロックの書き換えのオーバーヘッドなしで、クラウド層の同期が維持されます。
- **コールド層** パーティションが `analytics_offload_period` で定義された年齢しきい値を超えると、データレイクの PGAA テーブルに自動的に変換されます。ローカルストレージは再利用されますが、

データは同じテーブルインターフェイスを介して透過的に照会可能なままです。

### プロセスの概要

ヒープテーブルを階層テーブルに変換するには、次の手順を実行します。

1. *Setting up PGAA* 保存場所またはカタログサービスのいずれかを選択します。
2. pgaa.convert_to_tiered_table() ファンクションを使用して**ヒープテーブルの階層テーブルへの変換** 。
3. **階層化テーブルのモニタリング** 自動オフロード。
4. ストレージニーズの進化に応じた **階層化テーブルの変更と無効化** 設定。

### ヒープテーブルの階層テーブルへの変換

pgaa.convert_to_tiered_table() ファンクションは、自動オフロードプロセスを開始します。enable_replication が true の場合、継続的レプリケーションのためにアクティブなパーティションを HTAP テーブルに変換します。例

```
CALL pgaa.convert_to_tiered_table(
    relation := my-transactional-table::regclass,
    range_partition_column := date,
    partition_increment := 1 month,
    analytics_offload_period := 1 year,
    initial_lower_bound := 2010-01-01,
    retention_period := 5 years,
    enable_replication := true,
    minimum_advance_partitions := 2,
    maximum_advance_partitions := 5,
    drop_after_retention_period := true,
    purge_analytics_target := true
);
```

そこで

- relation 階層化テーブルに変換するヒープテーブルの名前。
- range_partition_column レンジパーティション化に使用される列。DATE または TIMESTAMP であり、主キーに含まれる必要があります。
- partition_increment 各パーティションのサイズを定義する時間間隔。
- analytics_offload_period パーティションがローカルストレージからデータレイクのコールド層に移動される経過期間のしきい値。
- initial_lower_bound 最初のパーティションの下限の開始値。
- retention_period データがコールド層からパージされる期間。
- enable_replication true の場合、アクティブなパーティションのオブジェクトストレージへのリアルタイムレプリケーションを HTAP テーブルとして有効にします。false の場合、ホットパーティ

ションはヒープテーブルとして残ります。

- minimum_advance_partitions 維持する将来のパーティションの最小数。
- maximum_advance_partitions 維持する将来のパーティションの最大数。
- drop_after_retention_period 保持期間の有効期限が切れた後、パーティションをローカルにドロップするかどうか。
- purge_analytics_target true の場合、オブジェクトストアディレクトリまたはカタログエントリに既存のテーブルが削除されます。false に設定されており、既存のテーブルが検出された場合、コマンドはエラーで失敗します PGD 6.3 以降でサポートされています。

> **注釈**
>
> レンジパーティションに使用される列は、主キーである必要があります。

シーケンスまたは外部キー制約を含むテーブルは、階層テーブルではサポートされていません。階層テーブルに変換する前に、シーケンスまたはコンストレインを削除する必要があります。

階層化テーブルは、すべてのパーティションの UNION を実行する統合ビューを作成し、データライフサイクル全体で単一のクエリインターフェイスを提供します。bdr.prefer_analytics_engine パラメーターを有効にして、階層テーブル全体のクエリーを Seafowl エンジンにルーティングできます。

```
SET bdr.prefer_analytics_engine = TRUE;
```

enable_replication が true の場合、この設定はパーティションテーブル全体の DirectScan を有効にし、Seafowl は単一のパスでオフロードされたコールドデータとレプリケートされたホットデータの両方に対してクエリを実行できます。

> **注釈**
>
> ストレージ場所ベースのレプリケーションを使用する場合、パーティション分割テーブルのユニオンビューは Seafowl エンジンでのみサポートされ、Spark を介してアクセスできません。カタログサービスを使用する場合、Seafowl と Spark の両方を介してユニオンビューにアクセスできます。 !!!

### 階層化テーブルのモニタリング

pgaa.list_tiered_tables() ファンクションは、現在自動化されたライフサイクル管理が行われているすべてのテーブルの高レベルの概要を提供します。その出力は、どのヒープテーブルが階層化テーブルに正常に変換されたか、どのパーティション化ロジックが適用されているか、およびローカルストレージとオブジェクトストレージ間の現在のバランス階層化および非階層化データサイズを特定するのに役立ちます。

```
SELECT * FROM pgaa.list_tiered_tables();
```

bdr.analytics_table ビューを照会して、Postgres スキーマと基になる分析エンジン間の低レベルのマッピングを表示します。このテーブルは、ローカルノードグループとリモート分析ターゲットの関係を追跡します。

```
SELECT * FROM bdr.analytics_table;
```

次のクエリを使用して、ホットパーティションとコールドパーティションを区別します。

```
SELECT
    relname AS partition_name,
    amname AS access_method
FROM pg_class c
JOIN pg_am am ON c.relam = am.oid
WHERE relname LIKE my-transactional-table%
  AND c.relkind = r
ORDER BY relname;
```

### 階層化テーブルの変更と無効化

PGD と PGAA コマンドを組み合わせて使用して、ライフサイクルポリシーを変更したり、標準のパーティションテーブルに戻すことができます。

#### ライフサイクル設定の調整

bdr.autopartition() ファンクションを実行する既存の階層テーブルのパーティション化とオフロードロジックを更新できます。

これは、オフロードしきい値を短縮または拡張する必要がある場合に特に役立ちます。たとえば、analytics_offload_period を 1 か月から 3 か月に変更するには、次のコマンドを実行します。

```
SELECT bdr.replicate_ddl_command($$
  SELECT bdr.autopartition(
    relation := public.my-transactional-table,
    partition_increment := 1 month,
    partition_initial_lowerbound := 2020-01-01,
    analytics_offload_period := 3 months
  )
$$);
```

詳細は、 AutoPartition in PGD を参照してください。

!!!警告

これらの設定の変更は、新しいパーティションにのみ影響し、既存のデータは再編成されません。パーティションサイズを増やしても、システムは現在の小さなパーティションをマージしません。現在のセットが使い果たされると、より大きなパーティションの作成が開始されます。新しいオフロードタイミングはすべてのパーティションに適用されますが、以前の設定で作成された物理構造は変更されません。

enable_replication の現在の設定を変更するには、pgaa.enable_analytics_replication() または pgaa.disable_analytics_replication() ファンクションを使用します。これらのファンクションは、親テーブルで呼び出されると、変更を既存および将来のすべての子パーティションに伝播します。

```
CALL pgaa.enable_analytics_replication(my-transactional-table::regclass);
```

個々のパーティションでレプリケーションを直接有効または無効にすることもできます。ただし、親テーブルで既に有効になっている場合、子パーティションでレプリケーションを無効にすることはできません。また、親で既に有効にされている場合、子パーティションで有効にすることはできません。

#### 階層テーブルの無効化

標準のパーティションヒープテーブルに戻すには

1. enable_replication がアクティブであった場合、親テーブルのレプリケーションを停止します。

```
CALL pgaa.disable_analytics_replication(my-transactional-table::regclass);
```

1. コールドパーティションを復元します。コールドパーティションごとに、次を実行します。

```
pgaa.restore_from_analytics(partition_name);
```

または ALTER TABLE partition_name SET ACCESS METHOD heap; を使用します。

1. 親テーブルから階層テーブルフラグを削除します。

```
ALTER TABLE my-transactional-table SET (pgaa.tiered_table = false);
```

1. AutoPartition を停止しますオプショナル。テーブルが新しいパーティションを完全に作成しないようにするには、次のコマンドを実行します。

```
bdr.drop_autopartition(my-transactional-table);
```

1. テーブルが SELECT * FROM pgaa.list_tiered_tables() を使用して階層テーブルとしてリストされていないことを確認します。
2. すべてのパーティションがアクセス方法として heap をリストしていることを確認します。

```
SELECT
    c.relname AS partition_name,
    a.amname AS access_method
FROM pg_class c
JOIN pg_am a ON c.relam = a.oid
WHERE c.relname LIKE my-transactional-table%
  AND c.relkind = r;
```

### Replicating to analytics

Postgres DistributedPGD クラスター内で Postgres アナリティクスアクセラレーター PGAA を使用して、ローカルヒープテーブルをハイブリッドトランザクションおよび分析処理 HTAP テーブルに変換することにより、分析データレイクと同期します。

この状態では、すべての INSERT 、UPDATE 、および DELETE オペレーションがキャプチャされ、ほぼリアルタイムでオブジェクトストレージにストリーミングされます。テーブルはアプリケーションの標準のト

ランザクションエンドポイントであり続けますが、PGD はバックグラウンドで高忠実度の分析レプリカを維持し、運用パフォーマンスに影響を与えることなく BI ツールとクエリエンジンでデータをすぐに利用できるようにします。

このプロセスの主な利点は次のとおりです。

- **ニアリアルタイム分析**バッチ ETL プロセスを待たずに、データレイクの最新のトランザクションデータにアクセスします。
- **デュアルステートのアクセシビリティ**トランザクション用のローカルの高性能ストレージを維持しながら、大量の分析クエリ用のカラムナバージョンを提供します。
- **統合管理**標準の SQL および PGAA 機能を介して、データのライフサイクルとレプリケーション設定を直接管理します。
- **マージオンリードの効率** PGAA は、Iceberg Merge-on-ReadMoR を使用してデータを同期します。行が変更されるたびに大きなファイルを書き換える代わりに、単に削除ファイルに変更を記録します。これにより、レプリケーションの高速性が維持され、オブジェクトストレージの負荷が大幅に削減されます。

### プロセスの概要

テーブルの継続的レプリケーションを有効にするには、次の手順を実行します。

1. *Setting up PGAA* 保存場所またはカタログサービスのいずれかを選択します。
2. **レプリケーションの有効化** データストリームを開始します。
3. **レプリケーションピアのモニタリング** ステータスとヘルス。
4. **レプリケーションの無効化** 同期が必要でなくなった場合。

### レプリケーションの有効化

`pgaa.enable_analytics_replication()` ファンクションを使用して、既存のヒープテーブルを構成して、データをオブジェクトストレージにストリーミングします。ファンクションは、ヒープテーブルを分析 HTAP テーブルに変換し、論理レプリケーションワーカーを起動します。

```
CALL pgaa.enable_analytics_replication(my_table::regclass, purge_analytics_target :=␣
↪TRUE);
```

レプリケーションが開始される前に、ターゲットの場所にある既存のデータをパージするように `purge_analytics_target:= TRUE` を設定します PGD 6.3 以降が必要です。

または、テーブルオブジェクトでレプリケーション状態を直接切り替えます。

```
ALTER TABLE my_table SET (pgd.replicate_to_analytics = TRUE);
```

`bdr.prefer_analytics_engine` パラメーターを有効にして、クエリを PGAA エンジン Seafowl または Spark にルーティングできます。

```
SET bdr.prefer_analytics_engine = TRUE;
```

無効にすると、クエリーは標準の Postgres エンジンによってローカルヒープに対して実行されます。

### モニタリングレプリケーション

レプリケーションが有効になると、テーブルがローカルリレーションリストと分析メタデータの両方に表示されることを確認することにより、HTAP ステータスを確認できます。

ローカルテーブルと分析ターゲット間のマッピングを表示します。

```
SELECT * FROM bdr.analytics_table;
```

現在 PGAA によって管理されているすべてのテーブルをリストします。

```
SELECT * FROM pgaa.list_analytics_tables();
```

### レプリケーションの無効化

レプリケーションを無効にすると、論理ワーカーによる変更のキャプチャが停止されます。ローカルテーブルは標準のヒープテーブルに戻り、オブジェクトストレージのリモートコピーはデータの静的スナップショットになります。

`pgaa.disable_analytics_replication()` ファンクションを使用して同期を停止します。

```
CALL pgaa.disable_analytics_replication(my_table::regclass);
```

または、 `ALTER TABLE` を使用します。

```
ALTER TABLE my_table SET (pgd.replicate_to_analytics = FALSE);
```

```
SET bdr.prefer_analytics_engine= FALSE;
```

**注釈**

以前に無効になったテーブルでレプリケーションを再度有効にすると、システムは完全なリロードを実行します。このプロセスは、オブジェクトストレージ内の既存の分析データを効果的に削除し、現在のローカルヒープデータを全体的に再アップロードし、継続的同期を開始します。 !!!

### Offloading to analytics

ハイブリッドトランザクションおよび分析処理 HTAP テーブルは、Postgres アナリティクスアクセラレーター PGAA を使用して、 Postgres DistributedPGD クラスターとオブジェクトストレージの分析コピーを継続的に同期し、両方の冗長ストレージを犠牲にして、トランザクションパフォーマンスに影響を与えることなく、大量の分析を有効にしますローカルヒープとデータレイク。

HTAP テーブルの分析へのオフロードは、ローカルヒープコピーを削除し、ディスク領域を再利用するために使用される手動の外科的操作です。テーブルをオフロードすると、システムは次のことを行います。

- ローカルヒープテーブルを切り捨て、PGD ノードのディスク領域をすぐに解放します。

- データレイクへのリアルタイムレプリケーションを停止します。
- テーブルアクセス方法を PGAA に変更し、Postgres テーブル定義をオブジェクトストレージ内の既存のデータを直接ポイントします。

その結果、Postgres を介して完全に照会可能なままですが、ローカルストレージを消費しないコールドテーブルが生成されます。将来的に再びトランザクション書き込みが必要な場合は、テーブルを復元し、データをレイクからローカルヒープにストリーミングして戻すことができます。

### プロセスの概要

テーブルをデータレイクに外科的にオフロードするには、次の手順を実行します。

1. *Setting up PGAA* 保存場所またはカタログサービスのいずれかを選択します。
2. テーブルのオフロード 。
3. トランザクションアクセスが再度必要な場合は、 アナリティクスからの復元ヒープに戻します。

### テーブルのオフロード

pgaa.convert_to_analytics() ファンクションを使用して、同期された HTAP テーブルをコールドの分析テーブルに変換し、ローカルディスクからデータを削除します。

```
CALL pgaa.convert_to_analytics(my-table::regclass);
```

ファンクションはローカルデータを削除し、テーブルのアクセス方法を PGAA に再定義して、外部の Iceberg または Delta ファイルのみを指すようにします。

または、ALTER TABLE コマンドでアクセス方法を切り替えます。pgaa.auto_truncate オプションを使用して、ローカルディスク領域がすぐに解放されるようにします。

保存場所の場合

```
ALTER TABLE public.my_table
  SET ACCESS METHOD pgaa,
  SET (
    pgaa.auto_truncate = true,
    pgaa.format = iceberg,
    pgaa.storage_location = my-storage-location,
    pgaa.path = public.my_table
  );
```

カタログ管理テーブルの場合

```
ALTER TABLE public.my_catalog_managed_table
  SET ACCESS METHOD pgaa,
  SET (
      pgaa.auto_truncate = true,
      pgaa.format = iceberg,
      pgaa.managed_by = my_iceberg_catalog,
```

(次のページに続く)

(前のページからの続き)

```
    pgaa.catalog_namespace = public,
    pgaa.catalog_table = my_remote_table_name
);
```

#### アナリティクスからの復元

トランザクション書き込みを再開するか、データをローカルの高性能トランザクションストレージに戻す必要がある場合は、pgaa.restore_from_analytics() ファンクションを使用します。

```
CALL pgaa.restore_from_analytics(public.transactional_table);
```

ファンクションは、PGAA テーブルの構成を確認し、オブジェクトストレージ内の外部 Iceberg/Delta パスを特定し、データをテーブルのローカルストレージ構造に書き換えます。データがローカルに存在すると、ファンクションはテーブルのアクセス方法をローカルトランザクションエンドポイントに再度設定します。

ALTER TABLE コマンドを使用して、テーブルをデータレイクのコールド状態からローカルトランザクションヒープテーブルに復元することもできます。この手動方法は、ローカルストレージへの移行をきめ細かく制御します。

```
ALTER TABLE public.my_table
SET ACCESS METHOD heap;
```

> **警告**
>
> テーブルをヒープアクセス方法に復元しても、オブジェクトストレージからファイルは削除されません。オブジェクトストレージプロバイダーを介してデータを手動で削除するか、PyIceberg のようなツールを使用してカタログからテーブルを削除します。

## 8.3 Recovering a failed node

PGD は、再起動や一時的なネットワークブリップなどの短期間の停止から障害が発生したノードを自動的に復旧します。障害が発生したノードは再接続し、見逃した変更を追いつき、介入なしでクラスターに再参加します。pgd nodes list を実行することにより、ノードが CATCHUP を介して ACTIVE に戻ることを確認できます。

自動復旧が成功しない場合、ノードを手動で復旧する必要があります。これは通常、次の場合に発生します。

- ノードのデータディレクトリが破損しており、Postgres が起動しません。
- ノードがオフラインになっている時間が長いため、正常なノードのレプリケーションスロットが無効になっているか、ログ先行書き込み WAL が危険なほど蓄積されています。
- 基になるハードウェアに障害が発生したため、ノードを再構築する必要があります。
- ノードは JOINING など、自分自身で進行することができない状態にあります。

障害が発生したノードを分離し、新しいメンバーとして戻すことにより、リカバリを実行します。このプロシージャは、 *Commit scopes* を使用する PGD クラスターに適用されます。以下をカバーします。

1. クラスターの状態の評価
2. 障害が発生したノードのクラスターからの削除
3. 障害が発生したノードの復旧
4. 復旧したノードの検証
5. クラスターの状態の検証

!!!警告障害が発生したノードを長期間無人のままにしないでください。正常な各ノードは、オフラインピアの WAL とレプリケーションデータを保持します。これにより、ディスク領域が枯渇し、カタログのバキュームがブロックされる可能性があります。ノードが完全に失われるか、ダウンしている時間が長すぎる場合は、ノードを分割し、その場所に新しいノードに再参加します。代替品は、元のノードの名前を再利用できます。

クラスターに損傷を与える可能性があるため、PGD によって作成されたレプリケーションスロットを手動で削除しないでください。

開始する前に、次のことを確認します。

- *Configuring PGD CLI* へのアクセス。
- 正常なノードの接続文字列 DSN。
- *bdr_superuser* または同等のクラスターで pgd および SQL コマンドを実行します。

### 8.3.1 クラスターの状態の評価

アクションを実行する前に、すべてのノードとレプリケーションの現在の状態を確認してください。

1. どのノードが `ACTIVE` で、どのノードがダウンまたは遅延しているかを確認します。

```
pgd nodes list
```

出力を使用して、障害の種類を判断します。ダウンしているノードの半分未満がダウンしている場合、 マイノリティ障害クォーラム維持 があり、クラスターはクォーラムを保持します。半分以上がダウンしている場合、 マジョリティ障害クォーラムの損失 があり、クラスターはクォーラムを失います。これにより、次の手順でどの復旧アプローチを使用するかが決まります。

1. 非アクティブまたは遅延しているレプリケーションスロットを確認します。

```
pgd replication show --slots
```

正常な各ノードは、オフラインピアのレプリケーションスロットを保持するため、WAL が蓄積されます。スロットの遅延は、ディスクの圧力を示します。無効なスロットは、WAL が破棄され、ノードがきれいに再参加できない可能性があることを意味します。

1. 現在の書き込みリーダーを特定します。

```
pgd group <group_name> show
```

### 8.3.2 障害が発生したノードのクラスターからの削除

このアプローチは、障害が発生したノードの数によって異なります。

#### マイノリティ障害クォーラム維持

マイノリティ障害とは、半分未満のノードに障害が発生した場合です。残りのノードはまだマジョリティを形成しているため、クラスターはクォーラムを保持します。つまり、書き込みに関して合意に達し、通常の動作を継続できます。

PGD CLI を使用して、障害が発生した各ノードを分離します。これは、少数の障害に推奨される方法です。<healthy_node_dsn> は正常なノードの接続文字列に置き換え、<failed_node_name> は削除するノードの名前に置き換えます。複数のノードに障害が発生している場合は、各ノードに対してコマンドを繰り返します。

```
pgd --dsn "<healthy_node_dsn>" node <failed_node_name> part
```

例

```
pgd --dsn "host=node2.example.com dbname=bdrdb user=enterprisedb" node node1 part
```

または、正常なノードから bdr.part_node ファンクションを使用し、<failed_node_name> を削除するノードの名前に置き換えます。複数のノードに障害が発生している場合は、各ノードに対してコマンドを繰り返します。

```
SELECT bdr.part_node(<failed_node_name>,
                     wait_for_completion => true,
                     force => false,
                     concurrent => true);
```

分離操作が開始され、ノードが安全になると自動的に削除されます。ノードは次の状態を遷移します。

ACTIVE → PART_START → PARTING → PART_CATCHUP → PART_TX_RESOLVE → PART_CLEANUP → PARTED → 自動的にドロップされます。

PART_CATCHUP フェーズでは、残りのノードは、操作が完了する前に、分割されたノードからの欠落データを同期します。これにより、データの損失が保証され、クラスターの一貫性が維持されます。

自動ドロップは、残りのすべてのノードが分割されたノードからのすべての変更を消費すると、発生します。手動でカタログのクリーンアップを実行する必要はありません。

#### マジョリティ障害クォーラムの損失

マジョリティ障害とは、ノードの半分以上に障害が発生した場合です。残りのノードはマジョリティを形成できなくなるため、クラスターはクォーラムを失います。これは、書き込みに関する合意に達できなくなり、障害が発生したノードが削除されるまで書き込み操作がブロックされることを意味します。

ノードの大部分に障害が発生している場合、クラスターコンセンサスが必要なため pgd node part は使用できません。 代わりに、bdr.drop_node() と force => true を使用し、ローカルコミットスコープのトランザクションでラップします。残りのすべての正常なノードでこれを実行します。

```
BEGIN;
SET LOCAL bdr.commit_scope = local;
SELECT bdr.drop_node(<failed_node_1>, force => true);
SELECT bdr.drop_node(<failed_node_2>, force => true);
COMMIT;
```

標準の分割プロセスとは異なり、強制ドロップは即時です。つまり、コンセンサスを待たずにノードが削除されます。

!!!注 強制ドロップに相当する PGD CLI はありません。SQL を使用する必要があります。 bdr.part_node(force => true) は PGD 6 で非推奨であり、bdr.drop_node(force => true) のエイリアスです。

#### ノードがクラスターから削除されたことの確認

<healthy_node_dsn> を正常なノードの接続文字列に置き換えます。

```
pgd --dsn "<healthy_node_dsn>" nodes list
```

マイノリティ障害の場合、ノードがなくなったことを確認する前に、自動クリーンアップが完了するのを待ってください。強制ドロップを使用するマジョリティ障害の場合、ノードはすぐに削除されます。

### 8.3.3 障害が発生したノードの復旧

クラスターに再参加する必要がある障害が発生したノードごとに、次のプロセスを繰り返します。PGD では部分化またはドロップされたノード名を再利用できるため、復旧したノードは元の名前を再利用できます。

#### データベースの停止とデータディレクトリの準備

障害が発生したノードで

1. Postgres サービスを停止し、停止していることを確認します。

```
sudo systemctl stop postgresql.service
sudo systemctl status postgresql.service
```

1. 既存のデータディレクトリをクリアする前に、バックアップまたは移動します。

```
cd /path/to/postgres/
mv data data_old
```

1. テーブルスペースディレクトリを含むデータディレクトリをクリーンアップします。

```
rm -rf <pgdata>/*
rm -rf /path/to/tablespace/*
```

#### pgd ノードセットアップの実行

```
pgd node <new_node_name> setup \
  --pgdata <pgdata> \
  --dsn "host=localhost port=<port> dbname=<dbname> user=<username>" \
  --cluster-dsn "host=<healthy_node_host> port=5432 dbname=<dbname> user=<username>" \
  --log-file <log_file> \
  --group-name <group_name> \
  --initial-node-count <initial_node_count>
```

そこで

- <new_node_name> 復旧したノードの名前。PGD では、ノードが PARTED 状態にある場合、元のノード名を再利用できます。
- --pgdata 復旧したノードのデータディレクトリ。
- --dsn 復旧するノードの接続文字列。
- --cluster-dsn クラスター内の正常なノードへの接続文字列。
- --log-file Postgres ログファイルへのパス。
- --group-name ノードグループ名。グループの最初のノードに必須。指定しない場合、ノードはアクティブノードのグループに参加します。
- --initial-node-count クラスターまたは計画中のノードの数。リソース設定を計算するために使用されます。

このコマンドは、データディレクトリを自動的に初期化し、クラスターに参加し、Postgres サーバーを起動します。出力で Node <new_node_name> added to the cluster successfully を探して、完了を確認します。

### 8.3.4 復旧したノードの検証

復旧したノードがクラスターに完全に再参加したことを確認します。ノードの参加状態が ACTIVE であることを確認します。

```
pgd --output-format psql --dsn "<recovered_node_dsn>" node <new_node_name> show
```

Raft コンセンサスが確立されていることを確認します。すべてのノードは RAFT_LEADER または RAFT_FOLLOWER を表示する必要があります。

```
pgd --output-format psql --dsn "<recovered_node_dsn>" raft show
```

予想されるすべてのレプリケーションスロットが存在し、アクティブであることを確認します。

```
pgd --output-format psql --dsn "<recovered_node_dsn>" replication show --slots
```

復旧したノードと他の正常なノードの両方からクラスターの可視性を確認します。すべてのノードは ACTIVE または CATCHUP として表示される必要があります。

```
pgd --output-format psql --dsn "<node_dsn>" nodes list
```

### 8.3.5 クラスターの状態の検証

ノードがクラスターに追いつきながら、レプリケーション遅延をモニタします。

```
pgd --output-format psql --dsn "<node_dsn>" replication show --slots
```

ルーティングが有効になっている場合は、復旧したノードでルーティング情報が利用可能で正しいことを確認します。

```
pgd --output-format psql --dsn "<node_dsn>" node <new_node_name> show
```

リカバリが完了すると、コミットスコープの動作が自動的に再開され、手動での再構成は必要ありません。

ノードが追いついている間、レプリケーション遅延、ディスク領域、Postgres ログ、およびノード間のネットワーク接続に注意してください。

### 8.3.6 障害対応

以下は、ノードの復旧中に発生する可能性のある一般的な問題とその解決方法です。

#### リカバリーはタイムアウトで失敗します

ノード間のネットワーク接続を確認し、ソースノードが応答していて正常であることを確認します。

#### 復旧したノードは ``ACTIVE`` ではなく ``JOINING`` として表示されます

復旧したノードの Postgres ログでエラー、特に復元フェーズに関する問題を確認します。正常なノードから `pgd replication show --slots` を実行して、他のノードにレプリケーションスロットが存在することを確認します。

#### 復旧後にレプリケーションラグが増加します

リソース制約 CPU、ディスク I/O、メモリを確認し、ノード間のネットワーク帯域幅を確認し、Postgres 構成で十分なワーカーとレプリケーションスロットが許可されていることを確認します *Postgres* **構成パラメーター** を参照してください。

### 8.3.7 リカバリー後のチェックリスト

通常の操作に戻る前に、次のチェックリストを使用して、クラスターが完全に正常な状態であることを確認します。

- [ ] 復旧したすべてのノードは、`join_state` を `ACTIVE` として表示します。
- ☒ Raft コンセンサスが確立される — すべてのノードが `RAFT_LEADER` または `RAFT_FOLLOWER` を表示します。
- [ ] レプリケーションスロットはすべてのノードでアクティブです。
- ☒ すべてのノードは `pgd nodes list` でお互いを認識します。

- [ ] どのノードの Postgres ログにもエラーはありません。
- ☒ レプリケーション遅延は許容範囲内です。

## 8.4 Developing applications for geo-replication

アプリケーションを地域分散 PGD クラスターに対して実行するときに考慮すべきことを学びます。PGD はレプリケーションとフェイルオーバーを透過的に処理しますが、ジオレプリケーションはアプリケーションレベルでの動作を導入し、接続、書き込み、および再試行方法の変更が必要になる場合があります。

### 8.4.1 geo レプリケーションがアプリケーションに影響を与える場合

geo レプリケーションでアプリケーションの変更が必要なシナリオを特定します。

- **アプリケーションは書き込みの直後に読み取ります**。レプリケーションは非同期であるため、別のノードまたは場所からの読み取りは、最近のコミットをまだ反映していない場合があります。
- **書き込みを失うことはできません**。デフォルトのコミットは非同期です。書き込みがレプリケートされる前にノードに障害が発生すると、書き込みは失われます。
- **失敗したコミットは安全に再試行できません**。このリスクはどのデータベースにも存在しますが、地域分散クラスターでは、書き込みリーダー選択中により多くの飛行中の障害が発生します。書き込みはノード間で複製されるため、中断されたコミットは一部のノードで既に有効になっている場合があり、ブラインド再試行はデータを破損する可能性があります。
- **複数の場所が同じ行に書き込みます**別の場所から同じ行への同時書き込みにより、競合が発生します。PGD はそれらを自動的に解決しますが、持続的な競合率は、再設計が必要なアクセスパターンを示します。
- **アプリケーションコードから DDL を実行します**。スキーマの変更は、クラスター全体のロックを取得します。ローカルクラスターでミリ秒単位で完了する DDL は、マルチリージョンクラスターでは数秒を要し、実行中に他のトランザクションをブロックする場合があります。

### 8.4.2 アプリケーションのジオレプリケーションの準備をする

新しいアプリケーションを構築するか、既存のアプリケーションを適応するかに同じ考慮事項が適用されます。

#### 接続マネージャーを介して接続する

個々のノードに直接接続するのではなく、接続マネージャーエンドポイントを使用します。接続マネージャーは、書き込み接続を現在の書き込みリーダーにルーティングし、フェールオーバーを自動的に処理します。復元力を高めるために複数のホストで接続文字列を構成します。

*Connecting through Connection Manager* を参照してください。

### オペレーションごとにコミットスコープを選択します

DBA は、使用可能なコミットスコープを定義し、グループまたはサブグループレベルでデフォルトを設定します。アプリケーションは、セッションごとまたはトランザクションごとにそのデフォルトをオーバーライドできます。ほとんどの書き込みにグループのデフォルトを使用し、耐久性が重要な場合はマジョリティ保護にオーバーライドし、支払いや在庫の減少などの非冪等操作に CAMO を使用します。

*Choosing commit scopes* を参照してください。

### 競合を最小限に抑えるようにスキーマをデザインする

競合は、複数の場所が同じデータに同時に書き込むときに発生します。デザイン時のスキーマとアクセスパターンの選択は、事後の構成変更よりも競合頻度に大きな影響を与えます。

*Preventing conflicts* を参照してください。

### 選択的レプリケーションの構成

レプリケーションセットを使用して、どのテーブルがどのノードにレプリケートするかを制御し、リージョン内に機密または場所固有のデータを保持します。

*Configuring selective replication* を参照してください。

### 再試行を正しく処理する

冪等操作、1 回適用しても複数回適用しても同じ結果を生成する操作は、一時的なエラーで自由に再試行できます。重複したアプリケーションが誤った結果を引き起こす支払いや在庫の減少などの非冪等操作では、CAMO ワークフローで再試行する前にトランザクション運命を決定する必要があります。指数バックオフとサーキットブレーカーを使用して、フェールオーバー中にクラスターに負荷がかかるのを回避します。

*Handling retries and failures* を参照してください。

### DDL を安全に実行する

地理的分散クラスターのスキーマ変更は、書き込みリーダー全体でロックを取得するため、単一場所のセットアップよりも遅くなり、混乱が発生します。DDL と DML の混合を避け、 DDL トランザクションを小さくしてください。

*Running DDL safely* を参照してください。

### Connecting through Connection Manager

地域分散 PGD クラスターでは、 *Connection Manager Authentication* がルーティングを自動的に処理し、書き込み接続を現在の書き込みリーダーに転送し、読み取り接続をレプリカに転送します。個々のノードに直接ではなく接続マネージャーのエンドポイントに接続するようにアプリケーションをデザインし、復元のために複数のホストをリストするように接続文字列を構成します。

### 読み取りと書き込みの分割

接続マネージャーは、各ノードで 2 つのポートを公開します。ワークロードに基づいて、トラフィックを適切なポートに転送します。

- **読み取り書き込みポート**接続をノードグループの現在の書き込みリーダーにルーティングします。このポートは、すべての書き込み操作と、最新のデータを必要とする読み取りに使用します。

```
host=node1.<location>,node2.<location>,node3.<location> port=6432 dbname=<dbname>␣
↪user=<username>
```

- **読み取り専用ポート**接続をレプリカにルーティングします。このポートは、わずかなレプリケーションラグを許容できるクエリと読み取りを報告するために使用して、読み取りトラフィックを書き込みリーダーから遠ざけます。

```
host=node1.<location>,node2.<location>,node3.<location> port=6433 dbname=<dbname>␣
↪user=<username>
```

DBA はポート番号を構成し、それらはクラスター内のすべてのノードで一貫しています。正しい値については DBA に問い合わせてください。

両方の接続文字列に複数のノードをリストして、1 つのノードが使用できない場合、クライアントドライバーが次のノードを自動的に試行します。

サブスクライバ専用ノードは書き込みを決して受け入れないため、常に読み取り専用ポートを介して接続します。

### 接続文字列の構成

地域分散展開の場合、次のことに注意して接続文字列を構成します。

- **ローカルノードに接続します。**リモートノードにルーティングするのではなく、アプリケーションサーバーと同じ場所にある Connection Manager インスタンスに接続すると、アプリケーションの実行が速くなります。接続マネージャーは、必要に応じて書き込みリーダーへのルーティングを処理します。
- **適切なタイムアウトを設定します。**クロスリージョンのネットワーク遅延は、単一のデータセンター内よりも長くなります。これを説明するには、接続文字列で標準の Postgres `connect_timeout` パラメーターを使用します。地理分散クラスターの場合、5〜10 秒が一般的です。

### フェイルオーバーの処理

書き込みリーダーに障害が発生すると、接続マネージャーは障害を検出し、選択が完了した後に新しい書き込みリーダーに接続をルーティングします。選挙中に飛行中の接続がドロップするため、アプリケーションは以下を考慮する必要があります。

- **接続のドロップが予想されます**フェールオーバー中に進行中の開いているトランザクションはロールバックされます。アプリケーションは接続エラーをキャッチし、再試行する必要があります。
- **トランザクションの結果を想定しないでください** `COMMIT` 中の接続ドロップはどのデータベースでも発生する可能性がありますが、書き込みリーダー選択により、地域分散クラスターではより頻繁に発生します。接続がドロップしたときにレプリケーションが進行中の場合、トランザクションは一部の

ノードで既にコミットされている可能性があります。標準トランザクションの場合、べき等な操作で再試行します。非冪等クリティカルトランザクションの場合、*CAMO パートナーとの連携* を使用して正確な結果を判断します。

- **バックオフを使用した再試行ロジックを実装します。**接続障害の直後に再試行すると、フェールオーバーの進行中のノードにヒットする危険があります。再試行ストラテジーについては、*Handling retries and failures* を参照してください。
- **ロケーションレベルのフェールオーバーの計画を立てる**ロケーションレベルのフェールオーバーは、多くの場合、DNS またはロードバランサー構成を介してインフラストラクチャレベルで処理されます。接続を直接管理するアプリケーションの場合、接続文字列にセカンダリフォールバックとしてリモートノードを含めることをお勧めします。

### Choosing commit scopes

DBA *Commit scopes* し、グループまたはサブグループレベルでデフォルトを設定します。セッションごとまたはトランザクションごとにスコープを設定して、より強いまたは弱い耐久性保証が必要な操作のグループのデフォルトをオーバーライドします。適切なスコープを選択することは、それぞれのレイテンシーのコストと、レプリケーションが完了する前にノードに障害が発生した場合にどうなるかを理解することを意味します。

DBA が各ノードグループに対して構成したデフォルトのコミットスコープを表示するには

```
SELECT node_group_name, default_commit_scope
FROM bdr.node_group_summary;
```

#### 適切なスコープの選択

各スコープは、書き込みコミット後に何が起こるかについて異なる約束をします。正しい選択は、レプリケーションが完了する前にノードに障害が発生した場合にアプリケーションが何を許容できるかによって異なります。

#### ローカルプロテクト

高スループットのイベントストリーム、分析の取り込み、または耐久性よりも書き込みレイテンシーが重要なワークロードなど、最近の書き込みの損失が許容されるデータにはローカル保護を使用します。トランザクションはローカルでコミットし、他のノードに非同期でレプリケートするため、レプリケーションが完了する前にノードに障害が発生すると、書き込みが失われることを意味します。レプリケーションが追いつくまでに、別のノードでの読み取りは、わずかに古いデータを返す場合があります。

#### マジョリティプロテクト

アカウントの更新、顧客レコード、ほとんどのトランザクションデータなど、失われることはできないが、追加の遅延が許容される場合の書き込みにはマジョリティ保護を使用します。コミットは、オリジングループ内のノードの大部分が返す前に確認を待機するため、アプリケーションが続行する前にマルチプルのノードにわたって書き込みが持続します。コストは、書き込みごとの追加往復レイテンシーであり、マルチリージョンクラスターの場合、通常は数十〜数百ミリ秒です。

### アダプティブプロテクト

1 つのデータノードが到達不能になった場合にストリクトマジョリティ保護が失敗する 2 ノードプラス Witness グループには、アダプティブ保護を使用します。アダプティブ保護は、ノードの大部分が利用可能な場合、マジョリティ保護のように動作しますが、一時的に耐久性が低下するというコストで、一時的なネットワークの問題中に構成可能なタイムアウトの後、非同期レプリケーションに正常に低下します。

## CAMO

支払い、資金移動、および同じトランザクションを 2 回適用すると不正確な結果が発生するオペレーションに CAMO を使用します。コミットは専用のパートナーノードを待機し、トランザクションが 1 回だけ適用されることを保証します。コミット中に接続がドロップした場合でも、アプリケーションはトランザクションの運命を照会し、再試行するかどうかを決定できます。CAMO、スコープの設定を超えてアプリケーションコードの変更が必要です。再試行パターンについては、*CAMO* **再試行ワークフロー** を参照してください。

CAMO には、CAMO グループ内の正確に 2 つのデータノードが必要で、マジョリティ保護よりも高いコミットレイテンシが追加されます。これは、その遅延が受け入れられない高スループットのワークロードには適切ではありません。

### クォーラムコミット

分散一貫性が必要で、事後に競合を解決できない決済、コアバンキング、および台帳システムにクォーラムコミットを使用します。耐久性を確認するマジョリティ保護とは異なり、クォーラムコミットは、ノードがローカルにコミットする前に、すべての参加ノードにわたってコミット決定を調整します。2 つのトランザクションが競合する場合、一方はコミットフェーズ中に、いずれかがコミットする前にアボートされます。

クォーラムコミットは熱心な競合解決を使用するため、アプリケーションは他のスコープよりもコミット時にシリアル化エラーに遭遇します。再試行ループの例を含むセットアップ、構成、およびアプリケーションパターンについては、*Using Quorum Commit* を参照してください。

### スコープの設定

ほとんどの操作でグループのデフォルトを使用し、それが必要な書き込みの場合にのみより強力なスコープにオーバーライドします。

接続上のすべての操作が同じ耐久性要件を共有する場合、支払い処理のみに使用される専用接続など、セッションごとにスコープを設定します。単一の接続が混合ワークロードを処理し、特定の操作のみがより強力な保証を必要とする場合、トランザクションごとに設定します。

- **セッションごと** セッションレベルで `bdr.commit_scope` を設定して、その接続内のすべてのトランザクションにスコープを適用します。

```
SET bdr.commit_scope = majority protect;
```

セッション内の後続のすべてのトランザクションは、変更するかセッションが終了するまで、このスコープを使用します。

- **トランザクションごと** トランザクション内で `SET LOCAL` を使用して、そのトランザクションのみのスコープをオーバーライドします。セッションレベルの設定は、トランザクションが終了すると復元

されます。

```
BEGIN;
SET LOCAL bdr.commit_scope = majority protect;
-- critical write operations
COMMIT;
```

!!!重要クォーラムコミットは、それを使用するトランザクションが使用しないトランザクションと同じテーブルで同時に実行される場合、一貫性を保証しません。ノードグループのデフォルトがクォーラムコミットスコープの場合、同じテーブルにタッチするトランザクションの弱いスコープにオーバーライドすることは避けてください。

現在のセッションでアクティブなコミットスコープを確認するには

```
SHOW bdr.commit_scope;
```

#### その他の考慮事項

##### 大量操作

高スループットの大量操作には同期スコープの使用を避けてください。トランザクションごとのラウンドトリップレイテンシーの追加は、大規模なバッチ全体で増加し、大量のロードが大幅に遅くなる場合があります。大量操作をスコープごとに分割する重要なレコードには強力なスコープを使用し、残りの部分にはデフォルトを使用します。

##### タイムアウトしたコミット

コミットスコープがリモート確認を必要とし、リモートの場所に到達できない場合、トランザクションは失敗します。再試行ロジックは一時的なネットワーク障害を処理しますが、タイムアウトしたコミットには特別な注意が必要です。コミットがタイムアウトした場合、結果は不明です。トランザクションは、タイムアウト前に一部のノードでコミットされた可能性があります。重複したアプリケーションが受け入れられない操作には、CAMO を使用します。 *CAMO* **再試行ワークフロー** を参照してください。

##### 自分の書き込みを読む

アプリケーションが 1 つの接続に書き込み、すぐに別の接続を読み取る場合、読み取りはレプリケーションをまだ受け取っていないノードに到達する可能性があります。後続の読み取りが以前の書き込みを反映することを保証するには、書き込みトランザクションで同期コミットスコープを使用します。書き込みが必要なノードにレプリケートされるまでコミットは返されないため、これらのノードでの後続の読み取りは書き込みを参照します。

```
BEGIN;
SET LOCAL bdr.commit_scope = majority protect;

- - Insert a new transaction for a known account
INSERT INTO transactions (transaction_id, account_id, direction, amount, balance_
↪after, description, channel, initiated_at)
VALUES (
```

(次のページに続く)

(前のページからの続き)

```
    uuidv7(),
    <account_id>,
    D,
    250.00,
    <balance_after>,
    Online purchase,
    ONLINE,
    NOW()
)
RETURNING account_id, transaction_id, balance_after;

COMMIT;

-- This read now reflects the insert on any node in the majority
SELECT balance_after
FROM transactions
WHERE transaction_id = <returned-transaction-id>;
```

### Preventing conflicts

地域分散 PGD クラスターでは、すべての場所が同時に書き込みを受け入れます。競合は、レプリケーションが変更を伝播する前に、別のノードで同じデータが変更されると発生します。各ノードは変更をローカルにコミットし、レプリケーション中に他のノードから競合する変更を受信します。

PGD は、これらの衝突を検出して、自動的に解決しますが、自動解決はセーフティネットであり、良いデザインの代替ではありません。スキーマとアクセスパターンを設計して、競合が発生する頻度を最小限に抑えます。

### PGD が競合を解決する方法を理解する

PGD は、アプリケーションが関与せずに、ほとんどの競合を自動的に解決します。例外は、error リゾルバーセットとの競合であり、レプリケーションを停止し、手動介入が必要です。他のすべての競合タイプの場合、それぞれに決定的なデフォルトリゾルバーがあります。どのリゾルバーが適用されるかを知ることは、自動解決が予想される結果を生成する場合と、生成しない場合を特定するのに役立ちます。すべての競合は bdr.conflict_history で追跡されます。

- update_update 同じ行が 2 つの異なるノードで更新されました。last-update-wins によって解決されました。
- insert_insert 同じ主キーが 2 つの異なるノードに挿入されました。新しいタイムスタンプの行を維持することにより解決されました。
- update_delete 行はあるノードで更新され、別のノードで削除されました。デフォルトでは、削除が優先されます。
- delete_delete 行は両方のノードで同時に削除されました。データには関係ありませんが、ログに記録されます。

より難解なケースを含む競合タイプの完全なリストについては、*bdr_read_all_conflicts* を参照してください。

### 競合を最小限に抑えるデザイン

競合は、複数の場所が同じデータに同時に書き込まれている場合の症状です。それらを削減する最も効果的な方法は、各データの明確な所有場所を持つようにデータモデルをデザインすることです。

- **クラスター全体の一意のキーを使用します**自動インクリメントシーケンスはノードローカルであり、別のノードで同じ値を生成するため、insert_insert 競合が発生します。クラスター全体で一意の主キーには、PGD のグローバルシーケンスタイプ SnowflakeId または galloc を使用します。クライアントサイドで生成された UUID も競合セーフですが、領域効率は低くなります。PGD シーケンスオプションについては、 *Sequences* を参照してください。
- **場所ごとにデータを分割する**各ユーザーまたはテナントが常に 1 つの場所に書き込む場合、そのデータに競合が発生することはできません。各ユーザー、テナント、またはデータパーティションをホームの場所に割り当て、そこに一貫して書き込みトラフィックをルーティングします。アカウント残高や在庫レベルなど、高い書き込み速度と厳密な一貫性要件を持つデータは、ロケーションベースのパーティショニングの良い候補です。製品カタログやルックアップテーブルなど、ほとんどが読み取られ、ほとんど変更されないデータは、競合のリスクなしでどこにでも書き込むことができます。
- **可能であれば、追加専用パターンを使用します。**挿入のみを受信するテーブル監査ログ、イベントストリーム、時系列データは、更新または削除の競合を持つことはできません。これらのテーブルを挿入専用にデザインし、集計が必要な場合は別の概要テーブルを使用します。
- **共有カウンターを回避します**複数の場所によってインクリメントされるカウンターは、発生を待機している競合です。代わりに、インクリメントを個々の行として記録し、読み取り時に集計するか、同時更新で正しくマージするように設計された PGD の CRDT データ型を使用します。*CRDT* **データ型を使用した列レベルの競合の処理** を参照してください。
- **長時間実行されるトランザクションを回避します**長期間ロックを保持するか、スナップショットを読み取るトランザクションは、競合する書き込みが別の場所から到着できるウィンドウを増加させます。トランザクションは短くしてください。

### 競合のモニタリング

すべての競合タイプは bdr.conflict_history で追跡されます。bdr.conflict_history_summary を定期的にチェックして、競合率の高いテーブルがないかどうかを確認します。

```
SELECT nspname, relname, conflict_type, count(*)
FROM bdr.conflict_history_summary
GROUP BY nspname, relname, conflict_type
ORDER BY count(*) DESC;
```

特定のテーブルでの高い競合率が持続する場合は、再検討する価値があるアクセスパターンを示しています。上記のデザインパターンを確認するか、アクセスパターンの変更が実現可能でない場合は、カスタム競合リゾルバーを検討してください。最近の競合に関係する実際の行値を検査するには

```
SELECT local_time, nspname, relname, conflict_type, conflict_resolution,
       key_tuple, local_tuple, remote_tuple, apply_tuple
FROM bdr.conflict_history
ORDER BY local_time DESC
LIMIT 20;
```

last-update-wins が誤った結果を生成し、アクセスパターンを変更できないテーブルの場合、DBA にカスタム競合リゾルバーを構成するように依頼してください。競合リゾルバーの構成 を参照してください。

### Configuring selective replication

*Replication sets* を使用して、どのテーブルがどのノードにレプリケートするかを制御し、リージョン内に機密または場所固有のデータを保持します。

各ノードグループには、同じ名前の事前構成されたレプリケーションセットがあります。デフォルトでは、すべてのテーブルはトップレベルグループのレプリケーションセットに属し、クラスター内のすべてのノードにレプリケートされます。

グループ階層とそれに関連するレプリケーションセットを表示するには

```
SELECT node_group_name, default_repset, parent_group_name
FROM bdr.node_group_summary;
```

テーブルが現在どのレプリケーションセットに属しているかを確認するには

```
SELECT relname, set_name FROM bdr.tables ORDER BY relname, set_name;
```

テーブルを特定の場所に制限するには、それらの場所のレプリケーションセットに追加し、トップレベルグループのレプリケーションセットから削除します。

```
SELECT bdr.replication_set_add_table(<table_name>, <location_a>);
SELECT bdr.replication_set_add_table(<table_name>, <location_b>);
SELECT bdr.replication_set_remove_table(<table_name>, <top_level_group>);
```

この変更後、 <location_a> のノードのテーブルへの書き込みは<location_a> の他のノードにのみレプリケートされ、 <location_b> ノードの書き込みは<location_b> 内でのみレプリケートされます。

トップレベルレプリケーションセットに残ったテーブルは、どこでもレプリケートし続けます。製品カタログや構成など、すべての場所で必要なグローバル参照データのトップレベルのレプリケーションセットを使用します。

### Handling retries and failures

地域分散クラスターへの接続は、書き込みリーダー選択、ネットワークパーティション、またはコミットスコープのタイムアウト中にドロップする場合があります。同じ操作を 2 回適用せずにこれらの障害を処理する再試行ロジックを実装し、サーキットブレーカーを使用して、復旧中のクラスターの過負荷を回避します。

#### リトライロジックの実装

地理的分散クラスターへの接続は、書き込みリーダー選択またはネットワークイベント中に失敗する場合があります。次の原則を使用して、データベース操作を再試行ループでラップします。

- **一時的なエラーのみを再試行します** シリアル化障害 `40001` 、デッドロックエラー `40P01` 、および接続エラーが再試行の候補です。アプリケーションレベルのエラーまたは制約違反はそうではありません。

- **指数バックオフを使用します**障害の直後に再試行すると、復旧中のクラスターがヒットする可能性があります。短い遅延で開始し、再試行ごとに 2 倍にし、小さなランダムジッターを追加して、雷のような群れの動作を回避します。
- **最大再試行回数を設定します**無期限に再試行しないでください。適切な回数の試行の後、失敗し、呼び出し元にエラーが表示されます。
- **可能であれば、再試行される操作を冪等にします。**冪等操作は、一度適用されても複数回適用されても同じ結果を生成します。誤った結果なしで同じトランザクションを 2 回安全に適用できる場合は、以前の結果を確認せずに再試行します。それができない場合は、*CAMO* **再試行ワークフロー** を使用します。

### CAMO 再試行ワークフロー

CAMO は、DBA によって設定されたコミットスコープです。グループのデフォルトでない場合は、*CAMO* **パートナーとの連携** の説明に従って、セッションごとまたはトランザクションごとに設定します。

支払い、在庫の減少、および同じトランザクションを 2 回適用すると不正確な結果が発生するものなどの非冪等操作の場合、標準の再試行ロジックは安全ではありません。CAMO は、トランザクションが最大 1 回適用されることを保証し、アプリケーションに、再試行するかどうかを決定する前に、切断されたトランザクションの最終的な結果を判断する方法を提供します。

CAMO トランザクションを発行する前に、CAMO パートナーの準備ができていることを確認します。

```
SELECT bdr.is_camo_partner_connected();
SELECT bdr.is_camo_partner_ready();
```

パートナーの準備ができていない場合、CAMO は非同期モードに低下する場合があります。その場合、厳密に 1 回保証は利用できません。

完全な再試行パターンとトランザクションステータスの照会については、*Commit At Most Once* を参照してください。

### サーキットブレーカーを使用する

サーキットブレーカーは、アプリケーションがフェールオーバー中のデータベースを再試行要求であふれさせるのを防ぎます。これがないと、リーダー選挙中の大量の再試行により、回復が遅れ、新しく選ばれたリーダーを圧倒する可能性があります。

ローリングウィンドウで障害率を追跡するサーキットブレーカーを実装します。

- **クローズ状態** 要求は通常に通過します。
- **オープン状態**要求はデータベースにヒットせずに失敗します。障害がウィンドウ内のしきい値を超えるとトリガーされます。
- **ハーフオープン状態**リカバリタイムアウトの後、単一のプローブ要求が通過します。成功すると、回路が閉じます。障害が発生すると、回路は開いたままになります。

地域分散クラスターのフェールオーバーのリカバリタイムアウトは、構成に応じて 5〜30 秒です。サーキットブレーカーのオープン期間を少なくともその長さに設定します。

### Running DDL safely

地域分散クラスターでの DDL 操作は、書き込みリーダー全体でロックを取得し、ステートメントの間保持されます。DDL トランザクションを小さくし、DML との混合を避け、トラフィックの少ない時間帯にスケジュールして、アプリケーションのスループットへの影響を最小限に抑えます。

ロックの取得時間は、書き込みリーダー間の往復レイテンシーに比例するため、ローカルクラスターでミリ秒を要する DDL は、マルチリージョンセットアップでは数秒かかる場合があります。これらのプラクティスを適用して、影響を最小限に抑えます。

- **追加的な変更を行う**列の追加、インデックスの追加、または新しいテーブルの作成は低リスクです。列の削除、列の名前の変更、およびデータ型の変更は、スキーマ変更をまだ受け取っていないノードで実行中のアプリケーションコードを中断する可能性があるため、より注意が必要です。ローリング展開では、古いスキーマに依存するすべてのアプリケーションバージョンがリタイアするまで、下位互換性を維持します。
- **最初に Nullable 列を追加します** 新しい列に `NOT NULL` 制約がある場合、クラスター全体の既存の行は、制約を適用する前にデフォルト値またはバックフィルが必要です。最初に列を Nullable として追加し、データをバックフィルして、後続の移行に制約を追加します。
- **アプリケーションコードの前にスキーマ変更を展開します**最初に追加的なスキーマ変更をロールアウトし、次にそれらを使用するアプリケーションコードを展開します。スキーマは、ノードがそれに依存するクエリの送信を開始する前に、すべてのノードに配置されます。
- **DDL を分離したままにする**各 DDL ステートメントは、クラスター全体の DDL ロックを保持し、他の同時実行 DDL を防ぎます。DDL ステートメントを単一のトランザクションで組み合わせるのではなく、個別に実行し、DDL と DML を決して混合しないでください。スキーマの変更とデータ操作を組み合わせると、ロックの期間が延長されてトランザクション全体をカバーします。
- **DDL を実行する前に長時間実行トランザクションを解決します**長時間実行トランザクションは、DDL がクラスター全体のロックを取得することをブロックします。DDL を実行する前に、長時間実行されているクエリまたはトランザクションを確認および解決します。
- **ロールバック計画を立ててください**すべての DDL をコミット後にロールバックできるわけではありません。`DROP TABLE` や `DROP COLUMN` のような破壊的な変更の場合、本番で変更を実行する前に、テストされたロールバックプロシージャを作成します。

## 8.5 Setting up access controls

PGD クラスター内のすべてのポイントで接続を暗号化および認証して、レプリケーショントラフィックを保護し、アプリケーションアクセスを制御し、無許可のノードの参加を防止します。アクセス制御は 3 つのレベルで動作します。

- ** *Securing inter-node connections* ** は、レプリケーションと Raft コンセンサスのために PGD ノードが相互に作成する接続です。これらの接続を保護すると、レプリケーショントラフィックが暗号化され、許可されたノードのみがクラスターに参加できます。
- ** *Securing client application connections* ** は、アプリケーションがクラスターを認証および接続する方法を対象としています。これらの接続を保護することにより、アプリケーションの認証方法、接続するポート、および HTTP 管理エンドポイントが TLS を使用するかどうかが制御されます。

- ** ノード接続への接続マネージャーの保護 ** は、Connection Manager がクライアントに代わってバックエンドノードに接続する方法をカバーしています。各ノードは、これらの接続を受け入れるように構成する必要があります。

### 8.5.1 Securing inter-node connections

PGD ノードはネットワークを介して互いに直接接続して、変更をレプリケートします。SSL がないと、レプリケーショントラフィックは暗号化されず認証されないため、傍受にさらされたままになり、未認証のピアがレプリケーションに参加できます。各ノードで SSL を構成して、レプリケーショントラフィックを暗号化し、無許可のノードがクラスターに参加するのを防ぎます。

#### 認証方法の設定

PGD は、ノード間接続を認証するための 2 つのアプローチをサポートしています。

> **注釈**
>
> 接続マネージャーは、ノード間 SSL 用に構成されたものと同じ `server.crt` を読み取ります。接続マネージャーの要件を満たさない証明書を使用すると、接続マネージャーが起動できません。詳細は、*SSL 証明書要件を満たす* を参照してください。

ベストプラクティスとして、ノード間接続には、`postgres` スーパーユーザーではなく専用のデータベースロールを使用します。ノードをクラスターに参加させる前に、各ノードで次を実行してロールを作成します。

```
CREATE ROLE pgd_replication LOGIN IN ROLE bdr_superuser;
```

このロールは、以降のセクションで示す接続文字列と `pg_hba.conf` ルールで使用します。

証明書認証は、暗号化とともに相互認証を提供するため、お勧めします。環境で証明書が使用できない場合、認証局 CA 検証を備えた SCRAM-SHA-256 は、接続を暗号化しながらパスワードベースの認証を提供します。

#### 証明書を使用した認証

証明書認証が推奨されるアプローチです。クライアント証明書は、暗号化と相互認証の両方を提供します。各ノードはその ID を証明する証明書を提示し、すべてのトラフィックは転送中に暗号化されます。

#### プロビジョニング証明書

各 PGD ノードは着信レプリケーション接続を受け入れ、発信接続を開始するため、すべてのノードがサーバーとクライアントの両方として機能します。各ノードには、接続を受け入れるためのサーバーサイドの証明書ファイルと、接続を開始するためのクライアントサイドの証明書ファイルが必要です。

次のファイルが各ノードのデータディレクトリに存在することを確認します。それらを生成するためのガイダンスについては、PostgreSQL ドキュメントの SSL certificate setup を参照してください。

- `server.crt` および `server.key` —ノードのサーバー証明書とプライベートキー。

- bdr_client.crt —ノードのクライアント証明書。
- bdr_client.key —クライアント証明書のプライベートキー。このファイルは、誰でも読み取り可能であってはなりません。
- root.crt - クライアントとサーバー証明書の両方を検証するために使用される CA 証明書。

証明書は次の要件も満たす必要があります。

- サーバーの SSL 証明書は、root.crt の CA によって直接または間接的に署名される必要があります。
- 接続のホスト名または IP アドレスは、証明書のサブジェクト代替名 SAN または共通名 CN と一致する必要があります。
- クライアント証明書の CN は、PGD 接続を行う Postgres ユーザーと一致する必要があります。

> **注釈**
>
> Postgres は、現在 IP アドレスのサブジェクトの代替名をサポートしていません。ノードがホスト名ではなく IP アドレスで接続する場合、IP アドレスは証明書の CN フィールドと一致する必要があります。

### 証明書認証の構成

1. 各ノードの postgresql.conf で SSL を有効にします。

```
ssl = on
ssl_cert_file = server.crt
ssl_key_file = server.key
```

各ノードで Postgres を再起動して、変更を適用します。

```
pg_ctl restart -D "$PGDATA"
```

1. 各ノードの現在の接続文字列を確認します。

```
SELECT node_name, interface_connstr FROM bdr.node_summary;
```

1. *bdr.alter_node_interface()* を使用して、各ノードの接続文字列を SSL パラメーターで更新します。このファンクションは接続文字列全体を置き換えるため、新しい SSL パラメータと並んで既存のすべてのパラメータを含めます。変更するノードを含む、グループ内のすべてのノードで実行します。例

```
SELECT bdr.alter_node_interface(
  node-1,
  host=host-1 port=5432 dbname=pgddb user=pgd_replication sslmode=verify-full␣
↪sslcert=bdr_client.crt sslkey=bdr_client.key sslrootcert=root.crt
);
```

1. 各ノードの pg_hba.conf にエントリを追加して、ノード間接続を許可します。

```
hostssl all         pgd_replication <node_subnet>/24 cert
hostssl replication pgd_replication <node_subnet>/24 cert
```

<node_subnet> を、PGD ノードの実際のサブネットまたは IP 範囲に置き換えます。

1. 各ノードで Postgres をリロードして、pg_hba.conf の変更を適用します。

```
pg_ctl reload -D "$PGDATA"
```

### SCRAM-SHA-256 での認証と CA 検証

クライアント証明書が使用できない場合は、パスワード認証を使用しながら、接続を暗号化し、CA に対してサーバーの証明書を検証します。

#### 証明書と資格情報のプロビジョニング

各ノードに以下が設定されていることを確認します。

- データディレクトリの server.crt および server.key —ノードのサーバー証明書とプライベートキー。それらを生成するためのガイダンスについては、PostgreSQL ドキュメントの

SSL certificate setup を参照してください。

- データディレクトリの root.crt —サーバーの SSL 証明書の検証に使用される CA 証明書。
- Postgres プロセスユーザーのホームディレクトリにある

**.pgpass ファイル、各ノードの正しいパスワードが含まれています。ファイルには、0600**
以上の権限が必要です。アクセス許可がオープンすぎる場合、ファイルはサイレントに無視されます。

#### SCRAM-SHA-256 認証の構成

1. 各ノードの postgresql.conf で SSL を有効にします。

```
ssl = on
ssl_cert_file = server.crt
ssl_key_file = server.key
```

各ノードで Postgres を再起動して、変更を適用します。

```
pg_ctl restart -D "$PGDATA"
```

1. *Node management interfaces* を使用して、各ノードの接続文字列を更新します。このファンクションは接続文字列全体を置き換えるため、新しい SSL パラメータと並んで既存のすべてのパラメータを含めます。変更するノードを含む、グループ内のすべてのノードで実行します。

```
SELECT bdr.alter_node_interface(
  node-1,
  host=host-1 port=5432 dbname=pgddb user=pgd_replication sslmode=verify-ca␣
```

(次のページに続く)

(前のページからの続き)

```
↪sslrootcert=root.crt
);
```

1. ノード間接続用に各ノードの pg_hba.conf にエントリを追加します。

```
hostssl all         pgd_replication <node_subnet>/24 scram-sha-256
hostssl replication pgd_replication <node_subnet>/24 scram-sha-256
```

<node_subnet> を、PGD ノードの実際のサブネットまたは IP 範囲に置き換えます。

1. 各ノードで Postgres をリロードして、pg_hba.conf の変更を適用します。

```
pg_ctl reload -D "$PGDATA"
```

### 構成の検証

SSL の構成後、レプリケーション接続が暗号化され、すべてのノードがアクティブであることを確認します。クラスター内の任意のノードでこれらのクエリを実行します。

- レプリケーション接続が SSL を使用していることを確認します。

```
SELECT pid, ssl, version, cipher, bits
FROM pg_stat_ssl
JOIN pg_stat_replication USING (pid);
__OUTPUT__
pid | ssl | version |         cipher         | bits
-----+-----+---------+------------------------+------
51 | t   | TLSv1.3 | TLS_AES_256_GCM_SHA384 |  256
50 | t   | TLSv1.3 | TLS_AES_256_GCM_SHA384 |  256
(2 rows)
```

成功した結果は、ssl が t に設定された場合、レプリケーション接続ごとに 1 行を表示します。この出力は、暗号化がアクティブであることを確認しますが、使用されている認証方法は表示されません。

- pg_hba.conf で設定された認証方法の確認

```
SELECT type, database, user_name, address, auth_method
FROM pg_hba_file_rules
WHERE auth_method IN (cert, scram-sha-256);
__OUTPUT__
type   |   database    | user_name  |  address   |  auth_method
--------+---------------+------------+------------+--------------
hostssl | {all}         | {pgd_replication} | 172.20.0.0 | scram-sha-256
hostssl | {replication} | {pgd_replication} | 172.20.0.0 | scram-sha-256
(2 rows)
```

追加した行は、設定した auth_method たとえば cert または scram-sha-256 とともに表示されます。

- すべてのノードがまだ正常に複製していることを確認します。

```
SELECT node_name, peer_state_name, peer_target_state_name
FROM bdr.node_summary;
__OUTPUT__
node_name | peer_state_name | peer_target_state_name
----------+-----------------+-----------------------
node-1    | ACTIVE          | ACTIVE
node-3    | ACTIVE          | ACTIVE
node-2    | ACTIVE          | ACTIVE
(3 rows)
```

すべてのノードは、両方の状態列に ACTIVE として表示されます。

### 8.5.2 Securing client application connections

PGD へのクライアントアプリケーションの接続は、読み取り/書き込みおよび読み取り専用のルーティングエンドポイントを提供する接続マネージャーを介して通過します。このパスを保護するには、2 つのポイントでの構成が必要です。クライアント側で、アプリケーションの認証と接続の方法を制御し、バックエンド側で、Connection Manager がクライアントに代わって Postgres ノードを認証する方法を制御します。

#### クライアントから Connection Manager への接続の保護

SSL を有効にし、認証を構成してアプリケーショントラフィックを暗号化し、クラスターに接続できるクライアントを制御します。

#### SSL の有効化

クライアント接続に SSL を使用するには、次のファイルが各ノードのデータディレクトリに存在する必要があります。それらを生成するためのガイダンスについては、PostgreSQL ドキュメントの SSL certificate setup を参照してください。

- server.crt および server.key —ノードのサーバー証明書とプライベートキー。

各ノードの postgresql.conf で SSL を有効にします。

```
ssl = on
ssl_cert_file = server.crt
ssl_key_file = server.key
```

各ノードで Postgres を再起動して、変更を適用します。

```
pg_ctl restart -D "$PGDATA"
```

### SSL 証明書要件を満たす

接続マネージャーは TLS に rustl を使用します。これにより、OpenSSL よりも厳密な証明書要件が適用されます。これらの要件を満たさない証明書は、接続マネージャーを起動できません。要件は、Postgres が使用する同じ `server.crt` ファイルに適用されます。一部の EDB パッケージは、FIPS サポートを有効にしてビルドされます。その場合、Connection Manager は FIPS 140-3 暗号化要件も適用します。

サーバー証明書は次のことを行う必要があります。

- Be X.509 v3
- インクルード `basicConstraints:  CA:FALSE`
- ノードのホスト名または IP アドレスを使用して `subjectAltName` を含めます
- インクルード `extKeyUsage:  serverAuth`

証明書には、SHA-256 以上の RSA または ECDSA キーを使用する必要があります。DSA キーと MD5 または SHA-1 署名はサポートされていません。

OpenSSL とは異なり、rustls はホスト名検証のためにコモンネーム CN にフォールバックしません。`subjectAltName` 拡張機能が必要です。

### 認証の構成

接続マネージャーは、Postgres で使用可能な認証方法のほとんどをサポートしています。各ノードの pg_hba.conf にルールを追加して認証を構成し、Postgres をリロードしてそれらを適用します。

サポートされている認証方法については、 *Connection Manager* **の認証方法** を参照してください。

各ノードの `pg_hba.conf` にルールを追加して、接続できるクライアントと認証方法を制御します。例、アプリケーションサーバーからのパスワード認証を要求するには

```
hostssl   all   all   10.0.0.0/8   scram-sha-256
```

`pg_hba.conf` を編集した後、Postgres をリロードして変更を適用します。

```
pg_ctl reload -D "$PGDATA"
```

> **注釈**
>
> *pg_hba.conf* ルールが接続マネージャーでサポートされていない場合、警告としてログに記録され無視されます。接続マネージャーは、サポートしているルールを使用して動作を継続します。 *pg_hba.conf* 変更の後、Postgres サーバーログを確認して、サポートされていないルールがないかを確認します。

### エンドポイントの構成

接続マネージャーは、プライマリノードにトラフィックをルーティングする読み取り/書き込みエンドポイント、レプリカ全体でロードバランシングを行う読み取り専用エンドポイント、および管理 API の HTTP エンドポイントの 3 つのエンドポイントを公開します。デフォルトが環境内の他のサービスと競合する場合、エンドポイントポートを構成します。接続数を制限してデータベースを保護したり、HTTP エンドポイントで HTTPS を有効にしたりすることもできます。

| Option | Default | Purpose |
|---|---|---|
| *read_write_port* | Postgres port + 1000 (usually 6432) | Read-write connections |
| *read_only_port* | *read_write_port* + 1 (usually 6433) | Read-only connections |
| *http_port* | *read_write_port* + 2 (usually 6434) | HTTP management API (returns JSON) |

`pgd group set-option` を使用してこれらを構成します。

```
pgd group set-option --group-name <group_name> read_write_port <port>
```

### HTTP エンドポイントでの HTTPS の有効化

デフォルトでは、HTTP 管理エンドポイントはプレーン HTTP を使用します。HTTPS を有効にするには

1. `use_https` を `true` に設定します。

```
pgd group set-option --group-name <group_name> use_https true
```

2. Connection Manager は、*SSL の有効化* セクションで構成されたものと同じ `server.crt` および `server.key` ファイルを使用します。これらのファイルが各ノードのデータディレクトリに存在することを確認します。

> **注釈**
>
> Postgres で `ssl = on` がグローバルに設定されていない場合でも、HTTPS を有効にできます。いずれかの証明書ファイルが欠落している場合、Connection Manager は `use_https` 設定に関係なくプレーン HTTP にフォールバックします。

### 接続制限の設定

接続マネージャーは、各エンドポイントでクライアントとサーバーの接続に個別の制限を適用し、デフォルトで `postgresql.conf` から `max_connections` に設定されます。制限をオーバーライドするには

```
pgd group set-option --group-name <group_name> read_write_max_client_connections
↪<value>
```

| Option | Default | Description |
|---|---|---|
| *read_write_max_client_* | *max_connections* | 読み取り/書き込みポートでの最大クライアント接続 |
| *read_write_max_server_* | *max_connections* | 読み取り/書き込みトラフィックの最大サーバー接続 |

次のページに続く

表 5 – 前のページからの続き

| Option | Default | Description |
|---|---|---|
| *read_only_max_client_c* | *max_connections* | 読み取り専用ポートでの最大クライアント接続 |
| *read_only_max_server_* | *max_connections* | 読み取り専用トラフィックの最大サーバー接続 |

**注釈**

サーバー接続制限 *read_write_max_server_connections* および`read_only_max_server_connections` は、現在ユーザーごとに適用されています。これは既知の制限であり、将来のリリースで解除される予定です。

### コンセンサスタイムアウトの設定

デフォルトでは、接続マネージャーはクォーラムを失うとすぐに接続を拒否します。これにより、一貫性を保証できないクラスターへのルーティングトラフィックから保護されます。コンセンサスタイムアウトを設定すると、クラスターに接続が拒否される前に回復する猶予期間が与えられ、クラスターの状態が不確実なときに接続を受け入れるというコストで、一時的な障害中の可用性が向上します。アプリケーションが猶予期間中に不確実性を許容できる場合にのみこの値を設定します。

```
pgd group set-option --group-name <group_name> read_write_consensus_timeout <seconds>
```

### ノード接続への接続マネージャーの保護

接続マネージャーはクライアント接続を受け入れると、クライアントの資格情報を使用して適切なバックエンドノードに接続します。各バックエンドノードには、接続マネージャーからのこの接続を許可する `pg_hba.conf` ルールが必要です。

SCRAM-SHA-256 は、これらのバックエンド接続に推奨される認証方法です。証明書認証は可能ですが、接続マネージャーはユーザーごとのクライアント証明書をサーバーに提示できないため、構成と管理が複雑です。

1. 各ノードの `pg_hba.conf` にルールを追加して、Connection Manager がクライアントユーザーとして接続できるようにします。

```
hostssl all all <node_address>/32 scram-sha-256
```

<node_address> を、Connection Manager がそのノード上の Postgres に接続するアドレスに置き換えます。

1. `pg_hba.conf` を編集した後、Postgres をリロードして変更を適用します。

```
pg_ctl reload -D "$PGDATA"
```

## 8.6 Handling replication apply errors

レプリケーションの適用に障害が発生すると、PGD はデフォルトで失敗したトランザクションを無期限に再試行します。永続的な障害は、適用の繰り返しによるシステムの過負荷と、繰り返しのエラーメッセージでログを埋める可能性があります。

*apply_error_policy* ノードグループオプションを使用して、トランザクションの適用に失敗した場合の PGD の動作を制御します。設定した回数の試行後に再試行を停止し、レビューのために失敗した変更を記録し、サブスクリプションを無効にするか、失敗したトランザクションをスキップしてレプリケーションを続行するように設定します。

### 8.6.1 エラーポリシーの選択

4 つのエラーポリシーが使用でき、ノードグループレベルで構成されます。

- **``keep_retrying``** は、失敗したトランザクションを無期限に再試行します。PGD はサイレントに再試行し続け、何も記録されず、サブスクリプションが無効になることはありません。一時的なロック競合など、自己修復が予想される一時的な障害に使用します。`keep_retrying` がデフォルトです。
- **``disable_after_retries``** `apply_error_max_retries` 回まで再試行し、サブスクリプションを無効にして、理由を記録します。障害が無期限に持続することはできないが、一時的なものである可能性がある場合に使用します。
- **``record_all_and_disable``** は、失敗したトランザクションを完全に記録し、サブスクリプションを無効にします。レプリケーションを停止する前に、何が失敗したかの完全な記録が必要な場合に使用します。
- **``record_all_and_skip``** は、失敗したトランザクションを完全に記録し、スキップして、レプリケートを続行します。サブスクリプションは有効なままですが、失敗した変更は `bdr.pending_failed_changes` にサイレントに蓄積されます。すべてのトランザクションが適用されることを保証するよりも継続性が重要な高可用性展開に使用します。

`record_all_and_disable` および `record_all_and_skip` ポリシーは、アクションを実行する前にキャプチャ専用モードで再生することにより完全なトランザクションをキャプチャし、後で検査および解決するための完全なレコードを提供します。

### 8.6.2 エラーポリシーの構成

`bdr.alter_node_group_option` を使用してノードグループにポリシーを設定します。ノードグループ名を見つけるには

```
SELECT node_group_name FROM bdr.local_node_summary;
```

選択したポリシーを適用します。

```
SELECT bdr.alter_node_group_option(
    top_group,
    apply_error_policy,
    record_all_and_disable
);
```

必要に応じて、サポートパラメーターを構成します。例

```
- - Retry attempts before the policy takes action (default: 3)
SELECT bdr.alter_node_group_option(
```

(次のページに続く)

(前のページからの続き)

```
    top_group,
    apply_error_max_retries,
    5
);

- - Maximum transactions record_all_and_skip can auto-skip before disabling.
- - Set to -1 for unlimited. Default is NULL (unlimited).
SELECT bdr.alter_node_group_option(
    top_group,
    apply_error_max_skips,
    100
);

- - Maximum bytes to capture per failed transaction (default: 10485760, which is 10␣
↪MB)
SELECT bdr.alter_node_group_option(
    top_group,
    apply_error_max_record_size,
    52428800
);
```

適用エラーグループオプションの完全なリストについては、*bdr.alter_node_group_option()* を使用したグループ構成 を参照してください。

### 8.6.3 無効になったサブスクリプションへの対応

サブスクリプションが disable_after_retries または record_all_and_disable ポリシーで無効になると、レプリケーションが停止します。次の手順に従って、障害を調査および解決します。

1. 影響を受けるサブスクリプションとその理由を確認します。

```
SELECT sub_name, disable_reason, effective_error_policy, pending_changes
FROM bdr.subscription_error_status
WHERE NOT sub_enabled;
```

disable_reason 列は、PGD がサブスクリプションを無効にした理由を記録し、pending_changes は、レビューを待機している失敗した変更の数を示します。完全な列リファレンスについては、*bdr.subscription_error_status* を参照してください。

1. 失敗した変更を確認します。

```
SELECT id, failed_txn_id, subscription_name, remote_xid, remote_commit_lsn,
       operation, qualified_table, sql_statement, error_message
FROM bdr.pending_failed_changes_sql
WHERE change_status = pending
ORDER BY id;
```

1. オプションで、スキップしても安全かどうかを分析します。bdr.format_skip_analysis は、スキップを SAFE 、UNSAFE 、または UNKNOWN として分類し、その理由を説明します。

```
SELECT bdr.format_skip_analysis(id)
FROM bdr.pending_failed_changes
WHERE change_status = pending
LIMIT 1;
```

1. bdr.resolve_failed_transaction を使用して解像度をプレビューします。デフォルトは dry_run := true です。

```
SELECT bdr.resolve_failed_transaction(
    bdr_mydb_mygroup_node1_node2,  -- subscription_name
    12345::xid,                    -- remote_xid
    0/1234567::pg_lsn,             -- remote_commit_lsn
    skipped,                       -- resolution: skipped, applied, or discarded
    true                             -- re_enable
);
```

または、3 部分の識別子の代わりに failed_txn_id bdr.pending_failed_changes からの UUID を渡します。

1. dry_run := false を渡して解決を実行します。

```
SELECT bdr.resolve_failed_transaction(
    bdr_mydb_mygroup_node1_node2,
    12345::xid,
    0/1234567::pg_lsn,
    skipped,
    true,    -- re_enable
    false    -- dry_run
);
```

すぐに再有効化せずに解決するには、準備ができたら re_enable := false を渡し、手動で再有効化します。

```
SELECT bdr.alter_subscription_enable(bdr_mydb_mygroup_node1_node2);
```

1. リトライカウンターとスキップカウンターをリセットして、ポリシーしきい値が将来の障害に新しく適用されるようにします。

```
SELECT bdr.alter_subscription_reset_error_counters(bdr_mydb_mygroup_node1_node2);
```

1. 解像度を確認したら、解決された行を bdr.pending_failed_changes から削除して、領域を再利用します。解像度の概要は bdr.resolved_transactions に保存されます。

```
DELETE FROM bdr.pending_failed_changes
WHERE change_status IN (skipped, applied);
```

#### 個々の変更の解決

失敗したトランザクションにさまざまな変更が含まれ、安全に解決できるものとそうでないものは、bdr.skip_failed_change および bdr.apply_failed_change を使用して、トランザクション全体ではなく個々の行を操作します。すべての変更を個別に解決したら、bdr.resolve_failed_transaction を呼び出してサブスクリプションの位置を進め、再度有効にします。

#### 切り捨てられたキャプチャの処理

失敗したトランザクションが apply_error_max_record_size を超えると、PGD は可能な限り多くの変更を記録し、bdr.pending_failed_changes に capture_truncated = true を設定します。レコードが不完全であるため、切り捨てられたトランザクションは bdr.apply_failed_transaction を使用して再適用できません。

切り捨てられたトランザクションを解決するには、bdr.reconstruct_failed_transaction を使用して部分 SQL を確認し、その出力を参照として使用してターゲットノードに完全なトランザクションを手動で適用してから、bdr.skip_failed_transaction を使用してキャプチャされた変更をスキップとしてマークし、サブスクリプションを再度有効にします。

大規模なトランザクションでの切り捨てを防ぐには、次の障害の前に apply_error_max_record_size を増やします。

### 8.6.4 蓄積された障害の監視

record_all_and_skip では、サブスクリプションは有効なままでレプリケーションは継続されますが、失敗したトランザクションはサイレントに蓄積されます。bdr.subscription_error_status を定期的にチェックしてそれらをキャッチします。

```
SELECT sub_name, total_skips, pending_changes
FROM bdr.subscription_error_status;
```

pending_changes がゼロ以外の場合、無効になったサブスクリプションへの対応 と同じ手順を使用して累積障害を確認および解決します。サブスクリプションがまだアクティブであるため、再有効化の手順をスキップします。

過去の決議を確認するには

```
SELECT subscription_name, origin_node, remote_xid, resolution,
       insert_count, update_count, delete_count, error_message, resolved_at
FROM bdr.resolved_transactions_info
ORDER BY resolved_at DESC;
```

## 8.7 Configuring connection pooling

PGD ノードグループの *Connection Manager Authentication* の接続プーリングを有効にして、クライアント間でサーバー接続をリサイクルすることによりデータベースのスループットを最大化し、多くの短期間の接続を使用するアプリケーションからの接続スパイクを防ぎます。Connection Manager のビルトイントランザクションプーリングでは、追加のソフトウェアを展開または保守する必要はありません。

各プールモードで正常に動作するアプリケーションを作成するためのガイダンスについては、*Using connection pooling in your application* を参照してください。

### 8.7.1 前提条件

- 各データノード PGD 6.4 以降で実行されている接続マネージャー。*Monitoring the Connection Manager* を参照して、実行されていることを確認します。
- bdr_superuser ロールまたは run と同等のロール bdr.alter_node_group_option

### 8.7.2 接続プーリングの有効化

接続プーリングはデフォルトで無効になっており、server_pool_mode は none に設定されています。これを有効にするには、ノードグループのモードを session または transaction に設定します。

SQL を使用する

```
SELECT bdr.alter_node_group_option(mygroup, server_pool_mode, transaction);
```

PGD CLI を使用する

```
pgd group mygroup set-option server_pool_mode transaction
```

プーリングを有効にしても、既存の接続は影響を受けず、新しい接続のみが更新されたモードを使用します。transaction モードから切り替える場合、オープンなトランザクションのない既存のクライアント接続は、次回トランザクションを開始したときに閉じられます。

接続は長くなりますが、トランザクションは短く、間にアイドル時間がある場合、たとえば、クライアントアプリケーションが独自の接続プーリングを使用する場合、transaction モードを使用します。

持続期間の短い接続、またはアプリケーションがトランザクション全体で持続するセッション状態に依存している場合 SET 、一時テーブル、またはアドバイザリロックなど session モードを使用します。各モードの詳細については、*Connection pooling* を参照してください。

!!!警告トランザクションモードは、各トランザクションの終了にほとんどのセッション状態をリセットします。有効にする前に、WITH HOLD カーソル、LISTEN 、または SET 変更など、トランザクション全体でセッション状態を持続する必要がある機能に依存するアプリケーションがあるかどうかを開発チームに確認してください。影響を受ける機能の完全なリストについては、*Using connection pooling in your application* を参照してください。

アクティブなプールモードを確認します。

```
SELECT node_group_name, server_pool_mode FROM bdr.node_group_summary;
```

### 8.7.3 プールのモニタリング

*Monitoring the Connection Manager* は、アクティブな接続とプールの使用率を追跡するために使用可能なメトリックと HTTP エンドポイントを説明します。

### 8.7.4 pgBouncer からの移行

Connection Manager はビルトインの接続プーリングを提供し、pgBouncer などの別個のプーラーの必要を排除します。そのプーリングモードは、pgBouncer のセッションおよびトランザクションプーリングモードに似ています。移行するには、次の手順に従ってください。

1. 現在の pgBouncer プールモード `session` または `transaction` と一致するように `server_pool_mode` を設定します。

2. `read_write_max_server_connections` を pgBouncer の `pool_size` と同様の値に設定し、`read_write_max_client_connections` を pgBouncer の `max_client_conn` と同様の値に設定します。最初に、`read_only_max_server_connections` および `read_only_max_client_connections` を同じ値に設定します。実際のワークロードに基づいて、4 つの設定すべてを調整します。

3. pgBouncer の代わりに Connection Manager の読み取り/書き込みポートを指すようにアプリケーションの接続文字列を更新します。

4. アプリケーションの動作を確認します。

   - `session` モードでは、アプリケーションの動作は直接 Postgres 接続から変更ありません。
   - `transaction` モードでは、`SET` コマンド、プリペアドステートメント、または `WITH HOLD` カーソルの使用に特に注意を払い、必要に応じて調整します。*Using connection pooling in your application* を参照してください。

1. 移行が確認されたら、インフラストラクチャから pgBouncer を削除します。

認証構成は、各 PGD ノードの `pg_hba.conf` に残ります。個別のユーザーリストまたはプーラー構成ファイルは必要ありません。接続マネージャーを介して Postgres に接続できるユーザーはプールを使用でき、アクセスを許可または取り消すための 2 番目の場所はありません。

接続マネージャーは、一部のシナリオでの直接 Postgres 接続と比較して、追加の制限を課します。移行を完了する前に、アプリケーションを徹底的にテストします。

# 第9章　Guidance for Developers

## 9.1 Designing applications for PGD

シングルノードデータベースでは発生しないレプリケーション動作、書き込み競合、および一貫性のトレードオフを考慮して、PGD 用のアプリケーションをデザインします。

- *Preparing your application for PGD* は、新しいアプリケーションと既存のアプリケーションの両方に適用され、ワークロードパターン、競合回避、接続ルーティング、および PGD に固有のレプリケーション動作に対処します。
- *Assessing an existing application for PGD* は、`pgd assess` によってフラグされたスキーマパターンからバックアップ戦略の変更および互換性のないレプリケーション構成に至るまで、既存の Postgres 展開と PGD 間の互換性ギャップを特定します。

### 9.1.1 Preparing your application for PGD

PGD の分散的な性質により、既存のアプリケーションを移植するか、スクラッチから構築するかにかかわらず、シングルノードの Postgres では発生しないデザインの決定が発生します。このガイダンスでは、完全なデータベーススキーマとすべてのデータがすべてのノードにレプリケートされることを前提としています。

> **注釈**
>
> PGD は、選択的レプリケーションもサポートしているため、どのテーブルをどのノードにレプリケートするかを制御できます。 *Configuring selective replication* を参照してください。

#### ワークロードの理解

デザインを決定する前に、アプリケーションの読み取りと書き込みの比率を理解します。読み取りを別のノードにルーティングして、負荷を分散し、書き込みリーダーノードから完全にオフロードできます。

アプリケーションがオブジェクトリレーショナルマッパー ORM を使用している場合、発行された DDL を確認して、生成されたテーブルに主キーがあることを確認し、接続マネージャーの読み取り/書き込みおよび読み取り専用ポートにマップするように個別の接続プールを構成し、生成された SQL を監査します以下で説明するパターン。

### 書き込み戦略を計画する

主な目標として、書き込み競合を回避するようにアプリケーションをデザインします。書き込みトランザクションは、クラスター内のすべてのデータノードにレプリケートされるため、競合によるコストがシングルノードデータベースよりも高くなります。

アプリケーションが、パーティション間でオーバーラップする書き込みなしで、地域またはユーザーグループごとに書き込みトラフィックをきれいにパーティション化できる場合、別のリージョンにマルチプルの書き込みリーダーを使用するローカルルーティングが最適です。書き込みをリージョンのローカルに保つことで、リージョン間の競合が排除され、スループットが向上します。そのきれいな分離が保証できない場合、書き込み量の規模が拡大するにつれて競合率が増加します。その場合、すべての書き込みが単一の書き込みリーダーに送られるグローバルルーティングの方が安全な選択です。ルーティング構成自分自身はデータベース管理者 DBA のタスクですが、アプリケーションの書き込みパターンを理解することにより、どのモデルが適合するかが決まります。

PGD はデフォルトでグローバルシーケンスを使用します。 `int4` および `int2` シーケンスは、ノードごとに値の範囲を割り当てる `galloc` を使用します。 `int8` シーケンスは `snowflakeid` を使用します。どちらも ID にギャップを生成します。アプリケーションでギャップのないシーケンスを想定しないでください。シーケンス種類とそのトレードオフの詳細については、*Sequences* を参照してください。競合せずにマージする必要がある可換更新の場合、conflict-free replicated data types (CRDTs) を使用します。

競合の検出と解決の詳細については、 *Conflict Management* を参照してください。

### 読み取り戦略を計画する

読み取りトラフィックを接続マネージャーの読み取り専用ポートにルーティングします。これにより、接続が書き込みリーダーに転送されるのではなく、すべてのノードに分散されます。HAProxy は、Connection Manager の読み取り専用ポートを使用して、ノード全体で読み取りトラフィックのバランスを保つこともできます。構成例については、*Load Balancing with Connection Manager* を参照してください。

最近行った注文や更新された口座残高など、ユーザーが自分の書き込みをすぐに確認する必要がある読み取りの場合、これらの読み取りを書き込みリーダーにルーティングするか、より厳密な *Commit scopes* を使用して、完了する前に十分なノードで書き込みが可視されるようにします。集計数やソーシャルメトリックなどの古いデータが受け入れられる読み取りの場合、読み取り専用ポートを使用します。

### レプリケーション用のスキーマのデザイン

`UPDATE` または `DELETE` 操作を行うすべてのテーブルに `PRIMARY KEY` または `UNIQUE` 制約を定義します。PGD は、レプリケーションの適用中の高速行ルックアップのために、制約の背後にある一意インデックスを使用します。

主キーと一意制約がないテーブルの場合、PGD は、テーブルが作成されるか、レプリケーションに追加されるときに警告を発します。`REPLICA IDENTITY FULL` がデフォルトであるため、テーブルは引き続きレプリケートされます。更新と削除は、古い行全体と一致することによりレプリケートされますが、適用側のルックアップは非常に遅くなります。PGD は、適切なインデックスが存在する場合は使用するか、シークエンシャルスキャンにフォールバックします。主キーの欠落は通常見落としであり、書き込み負荷がかかるとスキャンコストが複合します。実稼働テーブルのデフォルトに依存しないでください。

PGD は、更新可能なビューを透過的に処理します。更新可能なビューを介したデータ操作言語 DML は、オリジンの基になるベーステーブルに解決され、ターゲット上の同じベーステーブルに適用されます。

PGD は、パーティション化されたテーブルを透過的にサポートしますパーティションへの変更は、ダウンストリームに自動的にレプリケートされます。*AutoPartition in PGD* は、時間の経過とともに増加する大規模なテーブルを管理するため、低競合のロックを使用してパーティションの作成と削除を自動化します。

スキーマで使用されるすべての照合順序がすべてのノードで使用できること、およびデフォルトの照合順序がノード間で同じであることを確認します。PGD はレプリカ ID 値で等価検索を使用するため、行レプリケーション自分自身は影響を受けません。ただし、一致しない照合修飾子で定義された一意インデックスは予期しない動作をする場合があり、照合可能な式を使用する行フィルタは、ノード間で照合が異なる場合に一貫性のない結果を生成する可能性があります。

### レプリケーション動作のアカウンティング

PGD は、すべてのトリガーとルールが処理された後、各ステートメントの最終結果のみをレプリケートします。`INSERT ... ON CONFLICT UPDATE` は、オリジンで発生した挿入または更新のいずれかを送信します。更新または削除が影響を与える行がない場合、何もレプリケートされません。

次の動作はシングルノードデータベースとは異なるため、開発中に注意が必要です。

- **トリガー** は、シングルノード Postgres と同様に、SQL が最初に実行された元のノードでのみ起動します。PGD がターゲットノードにレプリケートされた変更を適用すると、そのテーブルのトリガーは起動されません。代わりにそれらのエフェクトが複製されます。
- ターゲットノードに変更が適用されたときにトリガーを起動するには、`ALTER TABLE ... ENABLE ALWAYS TRIGGER` を使用します。適用時にのみ起動するには、`ALTER TABLE ... ENABLE REPLICA TRIGGER` .
- を使用します。ステートメントレベルのトリガー `FOR EACH STATEMENT` と列ごとの `UPDATE` トリガー `UPDATE OF column_name` は、設定に関係なく、適用時に実行されることはありません。
- 適用時に起動するトリガーファンクションは、オブジェクト参照を完全修飾する必要があります、または `ALTER FUNCTION ... SET search_path = ...` で検索パスを明示的に設定します。 PGD apply は、システムレベルのデフォルト `search_path` を使用し、別のパスを想定するファンクションは適用時に失敗します。
- ルールは、シングルノード Postgres と同様に、元のノードでのみ実行されます。PGD は、ルール自分自身ではなく、結果の変更をレプリケートします。
- `SELECT ... FOR UPDATE/FOR SHARE` およびアドバイザリロックからの行レベルのロックはレプリケートされません。`SERIALIZABLE` 分離はサポートされていますが、トランザクションはノード間でシリアル化されず、マルチプルのノード上の同時トランザクションは引き続き競合する可能性があります。
- バイナリデータの場合は `BYTEA` 列を使用します。これらは、基になる `TOAST` ストレージを含む最大 1GB を含む完全にレプリケートされます。アプリケーションがストリーム指向のラージオブジェクトインターフェイスを必要とする場合、PGD は *Replicated large objects* もサポートします。
- `CHECK` コンストレインでは揮発性ファンクションを回避します。PGD は、デフォルトで適用時に `CHECK` 制約を再実行し、揮発性ファンクションがオリジンでの結果とは異なる結果を返す場合、レプリケーションがブレークする可能性があります。この動作は、*pgd group set-option* グループオプションによって制御されます。
- `REPLICA IDENTITY` 列を外部としてマークしないでください。

- カスケードキーを含む外部キーは、書き込みノードでのみ適用され、適用時に再チェックされません。ローカルの一貫性のためにそれらを使用しますが、競合中のクロスノードの整合性違反を防止するとは考えないでください。

#### データ異常の回避

PGD は、レプリカ ID 値を使用して行を識別し、テーブル名を使用してテーブルを識別します。削除後の識別子、またはドロップ後のテーブル名を再利用すると、PGD が解決できないあいまい性が発生します。このあいまいさは ABA problem であり、PGD は、識別子が現在のオブジェクトを参照するか、以前のオブジェクトを参照するか、古いオブジェクトを参照するかを判断できません。

データの異常と相違を回避するには

- `INSERT` の行に一意の識別子を使用します。
- `UPDATE` の一意の識別子は変更しないでください。
- 削除された一意の識別子は再利用しないでください。
- ドロップされたオブジェクト名を再利用しないでください。

違反は、ノード全体でデータ異常を引き起こす可能性があります。ありそうもないことですが、不可能ではありません。削除された行の識別子は、ダウンしたノードを含むすべてのノードで `DELETE` が再生した後にのみ安全に再利用できます。良好な条件下では、これには 1 秒未満かかりますが、ノードが著しく低下している場合は数日かかる場合があります。

### 9.1.2 Assessing an existing application for PGD

既存のアプリケーションを PGD に移行する前に、その互換性を評価して、変更が必要なスキーマパターン、バックアップギャップ、およびレプリケーション構成を特定します。`pgd assess` コマンドは、シングルインスタンスセットアップや物理レプリケーションクラスターを含む、PGD がまだ実行されていない Postgres 展開用に設計されています。

#### pgd promote を使用する

アプリケーションを提供する実行中の Postgres インスタンスがある場合は、`pgd assess` を使用して PGD との互換性を評価します。このコマンドは、Postgres インスタンスに接続し、最も一般的なアプリケーションパターンのスキーマとデータベースの内部状態を確認し、それらを PGD に適応する方法に関する推奨事項を提供します。アプリケーションのライブトラフィックまたは実際のクエリは分析しません。

```
pgd assess --dsn <connection_string>
```

コマンドとその出力の詳細については、 *pgd promote* を使用する を参照してください。

#### バックアップ戦略の確認

バックアップに現在使用されているツールを特定します。PGD クラスターは複数のノードと場所にまたがるため、ディスク領域とバックアップ帯域幅の要件が増加します。すべてのノードを個別にバックアップするのではなく、グループまたは場所ごとに少なくとも 1 つのバックアップを計画します。

別の戦略は、単一のノードをバックアップし、そこからクラスターの残りの部分を再構築することです。最初にシングルノード PGD クラスターに復元し、次に追加ノードごとに `pgd node setup` を実行して、物理的結合を介して参加します。詳細は、*Backup and recovery* を参照してください。

#### レプリケーションと CDC セットアップの確認

デフォルトでは、データウェアハウスまたはイベントストリーミングパイプラインへの送信論理レプリケーションは、単一の PGD ノードに関連付けられます。CDC フェールオーバーのサポートが明示的に有効になっていない限り、論理レプリケーションスロットはクラスター全体でレプリケートされません。CDC フェールオーバーのサポートが有効になっている場合、スロットはノード全体でレプリケートされますが、フェールオーバーは完全に透過的ではありません。コンシューマは、新しいノードへの再接続をサポートする必要があります。詳細は、*CDC Failover support* を参照してください。

2 フェーズコミット 2PC プロトコルまたは XA トランザクションを介した他のデータベースとの同期は、PGD と互換性がありません。

## 9.2 Managing geo-replication

PGD は、データセンター、アベイラビリティーゾーン、クラウドリージョンを含む複数の地理的場所にまたがるクラスターをサポートし、各場所は独立して書き込みを受け入れることができます。geo レプリケーションを使用すると、ロケーションの障害から保護し、データをユーザーの近くに保つ、復元力の高いグローバル分散データベースシステムを構築できます。

### 9.2.1 geo レプリケーションを使用する場合

geo レプリケーションは、次の 1 つ以上が必要な場合に適切な選択です。

- **ロケーション障害保護**データセンターまたはアベイラビリティーゾーンがオフラインになると、別のロケーションはデータを復元せずに読み取りおよび書き込みを提供し続けます。
- **データの所在地**規制要件により、特定のデータは特定の地域または国内に留まることが義務付けられています。
- **リージョンを超えた低遅延の読み取り**さまざまなリージョンのユーザーは、中央の場所へのラウンドトリップなしでデータに高速アクセスする必要があります。
- **低い目標復旧時間 RTO を使用したディザスターリカバリー**バックアップから手動で復元せずに、別の場所へのほぼ瞬時のフェイルオーバーが必要です。

### 9.2.2 導入ワークフロー

新しい地域分散 PGD クラスターを設定するときは、次の手順に従ってください。

#### 1. 展開パターンを選択します

何かを構成する前に、アクティブな書き込み場所の数、ルーティングモード、およびコミットスコープを決定します。展開パターンは、後続のすべての決定に影響を与えます。

*Choosing a geo-replication deployment pattern* を参照してください。

### 2. ノード分散を計画し、サブグループを作成します

ノードをプロビジョニングする前に、ロケーション間のネットワーク遅延を測定し、単一のロケーションが投票ノードのマジョリティを保持しないようにクォーラムを計画します。監視の場所が必要かどうかを判断します。次に、トップレベルグループの下に各場所のサブグループを作成し、ノードをそのサブグループに割り当てます。

*Setting up your geo-replication topology* を参照してください。

### 3. ルーティングを構成する

ローカルルーティングロケーションごとの書き込みリーダーとグローバルルーティング単一のクラスター全体の書き込みリーダーのいずれかを決定します。ローカルルーティングはデフォルトであり、変更の必要はありません。グローバルルーティングの場合、トップレベルグループで有効にし、サブグループで無効にします。

ルート優先度を設定して、フェールオーバー中に書き込みリーダーとして優先されるノードを制御します。

**ルーティングしきい値の設定** を参照してください。

### 4. コミットスコープの構成

耐久性要件に基づいてコミットスコープを選択し、場所のサブグループごとに構成します。

*Commit scopes* を参照してください。

### 5. レプリケーションとシーケンスを調整する

並列適用とストリーミング設定を調整して、書き込み量が多い場合のレプリケーション遅延を削減します。クラスター全体の一意の値を生成するようにシーケンスが構成されていることを確認します。

*Replication and sequence tuning* を参照してください。

### 6. 遅延制御の構成

クロスリージョンレプリケーション遅延が懸念される場合は、ラグ制御ルールをコミットスコープに追加して、リモートノードが遅延した場合の書き込みスループットを調整します。

*Controlling replication lag* を参照してください。

### 7. クラスターを検証し、操作の準備をする

構成が完了したら、運用に行く前に、サブグループ、ルーティング、およびコミットスコープが正しく設定されていることを確認します。メンテナンスページでは、ローリングアップグレード、DDL の考慮事項、ネットワークパーティションの処理、競合モニタリングなどの進行中の運用タスクもカバーしています。

*Performing cluster maintenance* を参照してください。

### Choosing a geo-replication deployment pattern

geo レプリケーションを構成する前に、展開パターンを決定します。選択は、ルーティングの構成方法、ロケーション障害時のクラスターの動作、および操作の複雑さの程度に影響を与えます。

各パターンの完全な説明は *Deployment patterns* にあります。このページでは、特にジオレプリケーションに重要な決定事項について説明します。

#### アクティブな書き込み場所の数の決定

アクティブな書き込み場所の数は、最も基本的な決定です。アクティブな書き込み場所が多いということは、各リージョンのユーザーの書き込みレイテンシが低いことを意味しますが、クロスリージョンの書き込み競合は避けられず、競合の解決はアプリケーションデザインの一部である必要があります。

PGD は、Raft コンセンサスアルゴリズムを使用して、書き込みリーダーの選択や DDL 操作の実行など、ノード全体の決定を調整します。Raft は、これらの決定を行う前に、クォーラムノードの大部分を利用できる必要があります。地理分散クラスターでは、これは、アクティブなロケーションの数が、ロケーション障害中にクラスターがクォーラムを維持できるかどうかに直接影響を与えることを意味します。

- **2 つのアクティブな書き込み場所**最も単純なジオレプリケーションセットアップ。各ロケーションはローカルの読み取りと書き込みを提供しますが、1 つのロケーションを失うと Raft マジョリティが破損し、ロケーションが復旧するまで DDL およびグローバルシーケンスをブロックします。3 番目の場所に監視ノードを追加すると、3 番目の完全なアクティブなロケーションのコストなしでこの制限が解決されます。関連するパターンは、 *Two data groups, active-active* とその監視バリアントです。
- **3 つ以上のアクティブな書き込み場所** 1 つの場所に障害が発生した場合でも、ナチュラル Raft マジョリティを維持し、クォーラムの懸念を排除します。トレードオフは、操作の複雑さです。関連する展開パターンは、機密データをオリジンリージョン内に留める必要があるバリアントの *Three data groups, active-active-active* および *Multiple locations, data residency* です。

#### ローカルルーティングとグローバルルーティングの選択

PGD は、書き込みトラフィックの 2 つのルーティングモードをサポートしています。ノードグループの *enable_routing* 構成オプションを設定することにより、優先ルーティングモードを構成します。

- **ローカルルーティング**各場所には独自の書き込みリーダーがあり、書き込みをローカルにルーティングします。このアプローチでは、書き込み遅延が最小限に抑えられ、ほとんどの操作で各場所が独立して動作することを意味します。ただし、リージョンを超えた競合の可能性があり、2 つのグループのみでロケーションを失うことは、グローバルコンセンサスを破ることです。
- **グローバルルーティング**クラスター全体には、場所に関係なく単一の書き込みリーダーがあります。すべての書き込みトラフィックはそのリーダーにルーティングされ、他のリージョンからの書き込みのレイテンシーの追加と引き換えに、クロスリージョンの競合が排除されます。複数の場所が書き込みを受け入れることができる場合でも、グローバルルーティングは、これらの場所にわたる単一のリーダーへの書き込みを制限します。このモードは、3 つ以上のアクティブな書き込み場所、または監視場所を持つ 2 つのデータグループで動作します。プライマリロケーションに障害が発生した場合、残りの 2 つのロケーションはまだ Raft マジョリティを持っており、新しい書き込みリーダーを選択できます。

構成手順については、 ルーティングしきい値の設定 を参照してください。

#### データ常駐要件の適用

GDPR、CCPA、および各国のデータローカライゼーション法などの規制では、個人データまたは機密データが元のリージョン内にとどまることが必要です。

データ常駐の適用は、*Multiple locations, data residency* 展開パターンに組み込まれています。ローカルルーティングでは、各リージョンのサブグループに独自の書き込みリーダーがあり、アプリケーションはローカルエンドポイントに接続するため、機密データはデフォルトでリージョン内に残ります。クロスリージョンレプリケーションは選択的です。グローバル参照データ製品カタログ、構成、匿名化された分析のみがグループ間でレプリケートされます。個人データ、財務記録、およびその他の規制データは、元のグループにローカルなままです。

ローカルデータとグローバルデータを明確に分離するようにスキーマをデザインし、常にリージョンデータを正しいローカルエンドポイントに書き込むようにアプリケーションを構成する必要があります。

#### 比較

各ジオレプリケーション展開パターンは、クォーラム復元力、ルーティングオプション、競合処理、およびデータの常駐性にわたってさまざまなトレードオフを行います。

| Deployment pattern | Active write locations | Quorum during location failure | Supports global routing | Conflict resolution required | Data residency enforcement |
|---|---|---|---|---|---|
| Two data groups, active-active | 2 | No (Yes with witness) | No (Yes with witness) | Yes | No |
| Three data groups, active-active-active | 3+ | Yes | Yes | Yes | No |
| Multiple locations, data residency | 2+ | Yes | Yes | Yes | Yes |

### Setting up your geo-replication topology

展開パターンを選択したら、次のステップはクラスタートポロジを構成することです。これには、ノード分散の計画、場所サブグループの作成、およびこれらのサブグループへのノードの割り当てが含まれます。

#### ノード分散の計画

クラスターを構成する前に、ロケーション間のネットワーク遅延と帯域幅を測定します。クロスロケーションレイテンシーは、レプリケーションラグ、同期コミットパフォーマンス、および DDL 期間に直接影響を与えます。

クォーラムプランニングの場合、単一の場所が投票ノードのマジョリティを保持していないことを確認します。たとえば、5 つのノードのうち 3 つが場所 A にあり、2 つが場所 B にある場合、場所 A のみがマジョリティを保持します。ロケーション A を失うと、ロケーション B がまだ実行されている場合でも、クラスターは読み取り専用になります。

2 か所の展開の場合、3 番目の場所の監視ノードが標準のソリューションです。ロケーションごとに 2 つのデータノードと 1 つの監視がある場合、投票者は合計 5 人になるため、いずれかのロケーションが失敗す

る可能性があり、残りのノードがマジョリティを保持したままになります。ローカルの高可用性を実現するには、各場所に少なくとも 2 つのデータノードが必要です。

### ロケーションサブグループの作成

PGD は、グループ階層を使用して物理的トポロジを表します。各地理的場所はノードサブグループにマッピングされ、すべてのロケーションサブグループは、クラスター全体にまたがる単一のトップレベルグループの下にあります。場所ごとにサブグループを作成します。

```
SELECT bdr.create_node_group(
    node_group_name := <location_a>,
    parent_group_name := <top_group>,
    join_node_group := false
);

SELECT bdr.create_node_group(
    node_group_name := <location_b>,
    parent_group_name := <top_group>,
    join_node_group := false
);
```

join_node_group := false パラメーターは、現在のノードを参加させずにグループを作成します。ノードは次のステップで個別に結合されます。

<location_a> および<location_b> をロケーションの説明的な名前たとえば us_east および eu_west などに置き換え、 <top_group> をトップレベルの PGD グループの名前に置き換えます。

サブグループが作成されたことを確認するには

```
SELECT node_group_name, parent_group_name
FROM bdr.node_group_summary
ORDER BY parent_group_name;
```

### ノードのサブグループへの割り当て

ノードをクラスターに参加させるときに、それが属するサブグループを指定します。

```
SELECT bdr.join_node_group(
    node_group_name := <location_a>
);
```

これを各ノードで実行して、その場所のサブグループに関連付けます。接続マネージャーは、サブグループメンバーシップを使用して、書き込みトラフィックを各場所の正しい書き込みリーダーにルーティングします。

サブグループが配置されたら、展開パターンのルーティングとルート優先度を構成します。ルーティングしきい値の設定 を参照してください。

### Configuring routing

地域分散 PGD クラスターでは、ルーティングによって、どのノードが書き込みを受け入れるかが決まります。PGD 6 は、各 Postgres インスタンス内で実行される Connection Manager を使用して、これを処理します。書き込む必要があるアプリケーションは、接続マネージャーの読み取り/書き込みポートに接続する必要があります。これは、接続をノードグループの現在の書き込みリーダーにルーティングします。

場所全体で書き込みリーダーシップがどのように構造化されるかは、ローカルルーティングを構成するかグローバルルーティングを構成するかによって異なります。

- **ローカルルーティング**各場所に独自の書き込みリーダーを与えます。アプリケーションが場所 A のノードに接続すると、接続マネージャーはそれを場所 A の書き込みリーダーにルーティングし、書き込み遅延を低く保ちます。すべてのロケーションは同時に書き込みを受け入れます。これは、クロスロケーションの競合が可能であることを意味します。
- **グローバルルーティング**クラスター全体の単一の書き込みリーダーを指定します。ノードの接続マネージャーは、アプリケーションがどの場所にあるかに関係なく、そのクラスター全体のリーダーに書き込み接続をルーティングします。これにより、クロスロケーションの競合が排除されますが、リモート場所からの書き込みがクロスロケーションの遅延が発生することを意味します。

2 つのノードグループ *Postgres* **設定** 制御書き込みリーダー選択

- `enable_raft` —グループが Raft コンセンサスに参加するかどうか
- `enable_routing` —グループが独自の書き込みリーダーを選択するかどうか

PGD では、クラスターには、すべての場所のすべてのノードにまたがる単一のトップレベルグループと、ロケーションごとにサブグループがあります。`enable_routing` 設定は、ルーティングがローカルかグローバルかを決定します。

| | Top-level group *enable_routing* | Subgroup *enable_routing* |
|---|---|---|
| Local routing (default) | *off* | *on* |
| Global routing | *on* | *off* |

`enable_raft` は、トップレベルグループとサブグループの両方で常に `on` であり、構成間で変更する必要はありません。

ローカルルーティングはデフォルトであり、明示的な構成は必要ありません。グローバルルーティングの場合、トップレベルグループに `enable_routing=on` を設定し、各サブグループに `enable_routing=off` を設定します。

グローバルルーティングは、監視ロケーションを含む *Three data groups, active-active-active* パターン、*Multiple locations, data residency* パターン、または *Two data groups, active-active* パターンでのみ使用できます。監視がないと、2 つのデータグループ設定で 1 つの場所を失うと、Raft マジョリティがブレイクし、クラスター全体の書き込みリーダーを選択または維持することが不可能になります。

### 構成例

#### ローカルルーティング

各ロケーションサブグループは、独自の書き込みリーダーを選択します。トップレベルグループは、書き込みリーダー選挙に参加しません。

```
- - Top-level group does not elect a write leader
SELECT bdr.alter_node_group_option(top_group, enable_routing, false);

- - Each subgroup elects its own write leader
SELECT bdr.alter_node_group_option(location_a, enable_routing, true);
SELECT bdr.alter_node_group_option(location_b, enable_routing, true);
```

接続マネージャーは、書き込みをローカル書き込みリーダーにルーティングし、各場所の書き込み遅延を低く保ちます。

#### グローバルルーティング

トップレベルグループは、クラスター全体の単一の書き込みリーダーを選択します。場所サブグループは、自分自身を選択しません。

```
- - Top-level group elects the cluster-wide write leader
SELECT bdr.alter_node_group_option(top_group, enable_routing, true);

- - Subgroups do not elect their own write leaders
SELECT bdr.alter_node_group_option(location_a, enable_routing, false);
SELECT bdr.alter_node_group_option(location_b, enable_routing, false);
```

すべての書き込みは、単一のクラスター全体の書き込みリーダーにルーティングされます。リモートロケーションでは、クロスロケーションの書き込みレイテンシーが発生しますが、クロスロケーションの競合はありません。

#### 接続マネージャーの調整

接続マネージャーには、地域分散展開のために確認する重要な設定がいくつかあります。

- `read_write_port` および `read_only_port` —接続マネージャーがそれぞれ書き込みトラフィックと読み取りトラフィックをリッスンするポート。これらがすべての場所で一貫して構成されていることを確認します。
- `read_write_max_client_connections` および `read_only_max_client_connections` —受け入れられるクライアント接続の数を制限します。ロケーションごとの接続数が多い地理分散クラスターでは、これらを調整することによりリソースの枯渇を防ぎます。
- `read_write_consensus_timeout` —コンセンサスを失った後、書き込み接続の受け入れを停止するまでに接続マネージャーが待機する時間。デフォルトは `0` で、すぐに停止することを意味します。地理的分散クラスターでは、クロスロケーションのレイテンシーにより、真の障害ではない短いコンセンサス中断が発生する場合があります。一時的なネットワークイベント中の不要な書き込みの中断を回避するために、短いタイムアウトを設定することを検討します。

- read_only_consensus_timeout — 読み取り専用接続の場合、上記と同じ。

オプションの完全なリストについては、 *Configuring Connection Manager* を参照してください。

### ルート優先度の設定

ルート優先度は、グループ内の書き込みリーダーとしてどのノードを優先するかを制御します。値が大きいほど、優先度が高いことを意味します。書き込みリーダーシップが通常の条件でアプリケーショントラフィックの近くにあり、必要な場合にのみリモートノードに移動するように、ローカルノードに高い優先度を割り当てます。

```
SELECT bdr.alter_node_option(
    node_name := <node_name>,
    config_key := route_priority,
    config_value := 1
);
```

優先度スキームを文書化して、運用チームが場所を超えて予想されるフェイルオーバー動作を理解できるようにします。

### 現在の設定の確認

すべてのグループの現在の Raft とルーティング設定を確認します。

```
SELECT node_group_name,
       node_group_enable_raft,
       node_group_enable_routing
FROM bdr.node_group;
```

現在の書き込みリーダーを表示します。

```
SELECT node_group_name, write_lead
FROM bdr.node_group_routing_summary;
```

## Configuring commit scopes for geo-replication

コミットスコープは、PGD がトランザクションを確認する方法と、コミットされたとみなされる前に書き込みを確認する必要があるノードの数を制御します。geo レプリケーションでは、選択は書き込みレイテンシーと、ロケーションに障害が発生した場合にどの程度のデータが失われるリスクに直接影響します。

geo レプリケーションをサポートする展開パターンは、次のコミットスコープを使用します。

| Commit scope | Description | Latency | Data loss risk | When to use | Supported topology patterns | Notes |
|---|---|---|---|---|---|---|
| **Majority protect** | コミットする前に、オリジングループ内のノードの大部分が確認するのを待機します。 | Medium | Low | Zero data loss required and write latency is acceptable. Recommended for most production workloads. | Any group with 3+ nodes | Survives minority node failures. |
| **Local protect** | 他のノードを待たずにローカルにコミットします。変更は非同期的に複製されます。 | Lowest | Highest | Lowest possible write latency and some data loss on failure is acceptable. | Any topology | Not recommended for production use. Data loss possible if a node fails before replication. |
| **Adaptive protect** | 利用可能な場合はマジョリティに同期し、タイムアウトデフォルト10秒後に非同期に低下します。 | Variable | Low | Durability by default, but cluster must stay writable during partial failures. | Groups with 2 data nodes + witness | Balances durability and availability. Graceful degradation during node failures. |
| **CAMO** | 専用のパートナーノードを使用した1回だけのセマンティクスによる同期レプリケーション。 | Higher | None | Exactly-once transaction requirements where duplicate detection isn't possible at the application level. | Groups with exactly 2 data nodes | Prevents duplicate transactions after failover. Higher latency. Requires application changes. |

次のページに続く

表 3 – 前のページからの続き

| Commit scope | Description | Latency | Data loss risk | When to use | Supported topology patterns | Notes |
|---|---|---|---|---|---|---|
| **Quorum Commit** | ノードがローカルにコミットする前に、2 フェーズコミットを使用して、すべての参加ノードにわたってコミット決定を調整します。競合するトランザクションは、コミットフェーズ中に検出され、中止されます。 | High | None | Workloads requiring distributed transaction consistency where conflicts can't be resolved after the fact, such as payments, core banking, and ledger systems. | Any group with 3+ nodes (majority required) | No *DEGRADE TO* support. Requires write leader routing. Application must handle serialization errors and retry. |

### コミットスコープの選択

コミットスコープの選択は、レイテンシー、耐久性、障害中の動作に直接影響を与える重要な決定です。

次の質問を使用して、展開に適切なコミットスコープを特定します。答えは、トポロジ、目標復旧ポイント RPO 要件、およびワークロードに混合の耐久性のニーズがあるかどうかによって異なります。

- ワークロードは、支払いやコアバンキングなど、競合するトランザクションが両方ともコミットできない厳密な分散一貫性を必要としますか?
- **Yes** - Quorum Commit
- **No** - continue
- ロケーションに障害が発生した場合、アプリケーションはデータ損失を許容できますか
- **Yes** —リモート確認なしのマジョリティ保護で十分です
- **No** —リモートノード確認をルールに追加
- グループごとに 2 つのデータノード+監視がありますか?
- **Yes** —アダプティブ保護を使用します b_tran_3 **No** —マジョリティ保護を使用
- クロスロケーションの書き込みレイテンシーはすべてのトランザクションで許容されますか?
- **Yes** —ルールにリモート確認を追加します
- **No** —遅延制御でローカル耐久性を使用
- 異なる耐久性ニーズを持つワークロードが混合していますか

- **Yes** —セッションまたはトランザクションをオーバーライドして複数のスコープを使用します
- **No** — 単一のデフォルトスコープを使用します

1 回だけのトランザクションセマンティクスを必要とするワークロードの場合、CAMO は追加の保証を提供しますが、実装にはアプリケーションレベルの変更が必要です。詳細は、*Commit At Most Once* を参照してください。

### コミットスコープの作成

コミットスコープを選択したら、ルールが適用されるオリジンノードグループごとに `bdr.create_commit_scope()` を使用します。ほとんどの地域分散展開では、 `ORIGIN_GROUP` が推奨されるアプローチです。トランザクションを開始したノードのサブグループに動的に解決されるため、クラスター全体に 1 つのルールのみが必要です。例、マジョリティ保護スコープを作成するには

```
SELECT bdr.create_commit_scope(
    commit_scope_name := majority_protect,
    origin_node_group := <top_group>,
    rule := MAJORITY ORIGIN_GROUP SYNCHRONOUS COMMIT,
    wait_for_ready := true
);
```

ロケーションに非対称の耐久性要件がある場合は、代わりにロケーションごとに別個のルールを作成できます。詳細については、*Using Quorum Commit* を参照してください。

マルチグループ展開でマジョリティ保護を使用する場合、クロスロケーションオプションで `bdr.create_commit_scope()` の `rule` パラメーターを拡張して、リモートノードがコミットに参加する方法を制御できます。

- **ローカル持続性のみ** - オリジングループ内のノードの大部分が確認するとすぐにコミットします。変更はリモートの場所に非同期に複製されます。

```
MAJORITY ORIGIN_GROUP SYNCHRONOUS COMMIT
```

- **リモート確認** —コミットする前に確認する少なくとも 1 つのリモートノードの要件を追加します。書き込みごとにクロスロケーションのラウンドトリップ時間を犠牲にして、完全なロケーション障害が発生してもデータが生き残ることを保証します。

```
MAJORITY ORIGIN_GROUP SYNCHRONOUS COMMIT AND ANY 1 NOT ORIGIN_GROUP SYNCHRONOUS COMMIT
```

- **遅延制御** - フルスピードでローカルにコミットしますが、リモート遅延が構成されたしきい値を超える場合、トランザクションを調整します。同期確認を必要とせずに、リモート場所が大幅に遅延するのを防ぎます。

```
MAJORITY ORIGIN_GROUP SYNCHRONOUS COMMIT AND ALL NOT ORIGIN_GROUP LAG CONTROL (max_
↪lag_time = 30s)
```

### クォーラムコミットスコープの作成

クォーラムコミットでは、ノードがコミットする前にすべての参加ノードが同意する必要があります。他のコミットスコープの種類とは異なり、DEGRADE TO はサポートされておらず、熱心な競合解決は常にアクティブです。ABORT ON 句は必須です。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := quorum_scope,
    origin_node_group := <top_group>,
    rule := MAJORITY ORIGIN GROUP QUORUM COMMIT ABORT ON (timeout = 6s),
    wait_for_ready := true
);
```

マルチリージョンクラスターの場合、MAJORITY FOR EACH GROUP IN CLUSTER を使用して、各サブグループ内のマジョリティを必要とします。トポロジに適切な構成を選択するためのガイダンスについては、クォーラムコミット を参照してください。

### CAMO スコープの作成

CAMO を使用している場合、ルールは SYNCHRONOUS COMMIT の代わりに CAMO キーワードを使用します。CAMO、トランザクションの結果を追跡する同じサブグループ内の指定されたパートナーノードが必要です。したがって、コミット中に元のノードが失敗した場合、アプリケーションはパートナーを照会して、トランザクションが適用されたかどうかを判断できます。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := camo_scope,
    origin_node_group := <location_a>,
    rule := ALL (<location_a>) CAMO DEGRADE ON (timeout=30s, require_write_lead=true)␣
↪TO ASYNC,
    wait_for_ready := true
);
```

require_write_lead=true オプションは、ノードが現在の書き込みリーダーでない限り、非同期への低下を防ぎます。これがないと、非リーダーノードはタイムアウトで低下してローカルにコミットする可能性があり、書き込みリーダーがまだ起動していて、同じグループ内の書き込みを受け入れている場合、データ損失のリスクがあります。

CAMO は、他のコミットスコープ種類よりもオーバーヘッドが高く、フェールオーバー後のトランザクションの解決を処理するにはアプリケーションの変更が必要です。詳細については、*Commit At Most Once* を参照してください。

### デフォルトのコミットスコープの設定

各サブグループのデフォルトとしてコミットスコープを設定して、その場所から発生するすべてのトランザクションに自動的に適用します。

```
SELECT bdr.alter_node_group_option(
    node_group_name := <location_a>,
```

(次のページに続く)

(前のページからの続き)

```
    config_key := default_commit_scope,
    config_value := majority_protect
);

SELECT bdr.alter_node_group_option(
    node_group_name := <location_b>,
    config_key := default_commit_scope,
    config_value := majority_protect
);
```

!!!重要クォーラムコミットは、それを使用するトランザクションが使用しないトランザクションと同じテーブルで同時に実行される場合、一貫性を保証しません。ノードグループにデフォルトのコミットスコープを設定すると、すべてのトランザクションが一貫性の利点を得ることができます。

グループ構成を変更せずに、特定のセッションまたはトランザクションのデフォルトをオーバーライドできます。

```
- - Session-level override
SET bdr.commit_scope = majority_protect;

- - Transaction-level override
BEGIN;
SET LOCAL bdr.commit_scope = camo_scope;
- - ...
COMMIT;
```

#### 構成の検証

- 各サブグループに正しいデフォルトのコミットスコープがあることを確認します。

```
SELECT node_group_name, default_commit_scope
FROM bdr.node_group_summary;
```

- コミットスコープルールを確認します。

```
SELECT commit_scope_name, commit_scope_origin_node_group, commit_scope_rule
FROM bdr.commit_scopes;
```

- 現在のセッションのアクティブなコミットスコープを確認します。

```
SHOW bdr.commit_scope;
```

- CAMO を使用している場合、パートナーノードに依存する前に、パートナーノードが接続され準備ができていることを確認します。

```
SELECT bdr.is_camo_partner_connected();
SELECT bdr.is_camo_partner_ready();
```

### よくある落とし穴の回避

- **運用ワークロードにはマジョリティ保護を使用します。**ローカル保護は、他のノードを待機せずにコミットします。レプリケーションが完了する前にノードに障害が発生すると、データが損失します。
- **高遅延の接続には、遅延制御を備えたローカル耐久性を使用します。**すべてのトランザクションでリモート確認を待機すると、クロスロケーションの遅延が高い場合、書き込みタイムアウトが発生します。
- **混合ワークロードには複数のスコープを使用します。**非クリティカルな書き込みは、クリティカルな書き込みと同じ遅延コストを支払うべきではありません。強力な耐久性保証を必要としないワークロードのセッションまたはトランザクションレベルでオーバーライドします。
- **2 場所展開の場合、3 番目の場所に監視ノードを追加します。**これがないと、ロケーションを失うと Raft クォーラムが破損し、クラスターは書き込みリーダーを選択できなくなります。
- **重複したトランザクションを許容できないワークロードには、CAMO を使用します。**これがないと、コミット中にノードに障害が発生すると、フェールオーバー後にトランザクションが 2 回適用される場合があります。
- **クォーラムコミットでのシリアル化エラーを予期します。**迅速な競合解決は、コミットフェーズ中に 2 つの競合するトランザクションのいずれかを中止します。シリアル化エラーをキャッチして再試行するようにアプリケーションを構成します。

## Controlling replication lag

地理的な場所間のネットワーク遅延は、リージョン間のレプリケーションが単一の場所内よりも本質的に遅いことを意味します。管理されないまま放置されると、レプリケーションラグが蓄積し、目標復旧ポイント RPO をリスクにさらし、正常なノードのディスク領域を消費し、クラスターパフォーマンスを低下させます。

PGD は、クロスリージョン遅延を管理するための 2 つの補完ツールを提供します。書き込みスループットを調整してリモートノードが遅延しないようにする Lag Control と、レプリケーションメッシュを追加せずにリージョンの読み取りトラフィックをオフロードするサブスクライバ専用ノードです。

### 遅延制御を使用する

Lag Control は、レプリケーションラグが定義された制限を超えると、オリジンノードの書き込みトランザクションにわずかな遅延を自動的に挿入するコミットスコープの種類です。これにより、着信書き込み速度が遅くなり、ダウンストリームノードが追いつく時間が与えられます。

PGD コミット遅延は 0 ミリ秒で始まります。十分なノードの遅延が構成された制限内にある場合、遅延は 0 またはそれに近いままです。制限を超過すると、クラスターが追いつくまで遅延が増加します。遅延は、トランザクションがコミットしてロックを解放した後に適用されるため、同時トランザクションはブロックされません。

### 遅延制御の構成

ほとんどのジオレプリケーション展開では、オリジングループ用に同期コミットスコープが既に構成されています。`AND` を使用して同じルールにリモートグループに `LAG CONTROL` を追加して、同期耐久性がローカルに適用され、遅延調整がクロスリージョンに適用されるようにします。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := geo_scope_lag,
    origin_node_group := <top_group>,
    rule := ALL ORIGIN_GROUP SYNCHRONOUS COMMIT AND ANY 1 NOT ORIGIN_GROUP LAG␣
↪CONTROL (max_commit_delay=500ms, max_lag_time=30s),
    wait_for_ready := true
);
```

同期コミット要件なしでラグ調整のみが必要な場合は、両方で LAG CONTROL を使用します。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := geo_scope_lag_only,
    origin_node_group := <top_group>,
    rule := ALL ORIGIN_GROUP LAG CONTROL (max_commit_delay=500ms, max_lag_time=30s)␣
↪AND ANY 1 NOT ORIGIN_GROUP LAG CONTROL (max_commit_delay=500ms, max_lag_time=30s),
    wait_for_ready := true
);
```

主要なパラメーターは次のとおりです。

- max_commit_delay クライアントの書き込みトランザクションに注入できる最大遅延。これは難しい天井です。遅延がこの値を超えることはありません。
- max_lag_time 時間で表される許容可能なレプリケーション遅延。これはソフトリミットです。
- max_lag_size WAL バイトとして表される許容可能なレプリケーションラグ。これもソフトリミットです。

#### 遅延制御では十分でない場合

コミット遅延が一貫して構成された上限にあり、遅延がまだ制限を超えている場合は、書き込みスループットが使用可能なクロスリージョン帯域幅に対して基本的に高すぎることを示しています。次を検討します。

- クロスリージョンネットワーク帯域幅の増加。
- 大規模な大量ロード操作を、より長いコミット遅延を持つ小さなトランザクションに分割します。
- すべてのテーブルをクロスリージョンにレプリケートする必要があるかどうかの確認。

#### ルーティングしきい値の設定

ルーティングしきい値は、遅延ノードが書き込みおよび読み取りルーティングから削除されるタイミングを制御します。ノードの遅延が設定されたしきい値を超えると、接続マネージャーは追いつくまでトラフィックのルーティングを停止します。どちらのしきい値のデフォルト値は-1 で、チェックを無効にします。値はバイト単位です。

遅延がしきい値を超えたときにノードを書き込みリーダー資格から削除するように route_writer_max_lag を設定します。

```
SELECT bdr.alter_node_group_option(
    node_group_name := <group_name>,
    config_key := route_writer_max_lag,
    config_value := 16777216
);
```

遅延がしきい値を超えたときに読み取りルーティングからノードを削除するように route_reader_max_lag を設定します。

```
SELECT bdr.alter_node_group_option(
    node_group_name := <group_name>,
    config_key := route_reader_max_lag,
    config_value := 16777216
);
```

現在のしきい値を確認します。

```
SELECT node_group_name, route_writer_max_lag, route_reader_max_lag
FROM bdr.node_group_routing_config_summary;
```

#### チューニングレプリケーションの適用

変更をより速く適用すると、書き込み量が多い場合の遅延が削減されます。ノードグループレベルで並列適用ワーカーの数 num_writers 、ストリーミングモード streaming_mode 、および意図的な適用遅延 apply_delay を構成します。詳細は、*Replication and sequence tuning* を参照してください。

#### レプリケーションラグのモニタリング

次のビューを使用して、クラスター全体の現在の遅延を確認します。

- ノードごとの現在のレプリケーションレートと推定キャッチアップ時間を確認します。

```
SELECT target_name, replay_lag, replay_lag_bytes, apply_rate, catchup_interval
FROM bdr.node_replication_rates;
```

- ノードごとにサブスクリプションステータスと最後に適用されたトランザクションタイムスタンプを確認します。

```
SELECT origin_name, target_name, subscription_status, last_xact_replay_timestamp
FROM bdr.subscription_summary;
```

- 詳細な LSN レベル情報については、スロットラグを確認します。

```
SELECT target_name, state, replay_lag_bytes, replay_lag_size
FROM bdr.node_slots
ORDER BY replay_lag_bytes DESC;
```

遅延情報の精度は、各ノードがクラスターの残りの部分にレプリケーション位置をブロードキャストする頻度を制御する構成パラメーターである bdr.replay_progress_frequency に依存します。デフォルトは 60 秒です。遅延モニタリングの応答性を高めるには、この値を低くします。更新の頻度が高いと、少量のネットワークオーバーヘッドが追加されることに注意してください。

### サブスクライバー専用ノード

リージョンのユーザーがデータの読み取りがほとんどで、書き込みがほとんどない場合、そのリージョンにサブスクライバ専用ノードを追加すると、完全なアクティブロケーションをコストせずに読み取りレイテンシを大幅に削減できます。

サブスクライバー専用ノードはクラスターからレプリケートされた変更を受け取りますが、書き込みを発生せず、レプリケーションメッシュまたは Raft コンセンサスに参加しないため、軽量で追加が簡単です。詳細は、 *Subscriber-only nodes and groups* を参照してください。

### サブスクライバー専用ノードの追加

1. トップレベルグループの下に加入者専用のサブグループを作成します。

```
SELECT bdr.create_node_group(
    node_group_name := <read_region>,
    parent_group_name := <top_group>,
    node_group_type := subscriber-only
);
```

1. サブスクライバ専用ノードとして新しいノードを作成します。

```
SELECT bdr.create_node(
    node_name := <so_node>,
    dsn := host=<so_host> dbname=<dbname> port=5444,
    node_type := subscriber-only
);
```

1. アクティブなデータノード別の加入者専用ノードではないの DSN をエントリポイントとして使用して、新しいノードを加入者専用サブグループに参加させます。

```
SELECT bdr.join_node_group(
    dsn := host=<data_node_host> dbname=<dbname> port=5444,
    node_group_name := <read_region>
);
```

### サブスクライバー専用ノードへのルーティング読み取り

接続マネージャーは、読み取り専用トラフィックをサブスクライバー専用ノードにルーティングできます。読み取りリージョン内のノードの読み取り専用エンドポイントを使用するようにアプリケーションの接続文字列を構成します。

読み取り専用ルーティングの構成の詳細については、*Connection Manager Authentication* を参照してください。

### 加入者専用ノードの制限

- サブスクライバー専用ノードは書き込みを受け入れることができません。書き込もうとすると失敗します。
- 彼らは Raft コンセンサスに参加しないため、コミットスコープの確認のクォーラムにカウントされません。
- ネットワークの状況によっては、サブスクライバ専用ノードのレプリケーションラグがアクティブノードよりも高くなる場合があります。

## Replication and sequence tuning

地理分散クラスターは、高遅延のネットワークリンクを介して変更をレプリケートするため、レプリケーションのスループットとシーケンスの一意性が単一ロケーションのセットアップよりも重要になります。ノードグループレベルでこれらの設定を構成して、リモートノードに変更を適用する方法と、シーケンスが場所を超えて一意の値を生成する方法を制御します。

### パラレル適用

*Parallel Apply* により、各ノードは、マルチプルのライタプロセスをシークエンシャルではなく同時に使用して、着信変更を適用できるため、書き込み量が多い場合の適用ラグが削減されます。bdr.writers_per_subscription がノードごとに 2 つのライターに設定されている場合、デフォルトで有効になっています。ノードグループレベルで num_writers を使用して、bdr.max_writers_per_subscription で設定された制限内で、サブグループごとのライター数を設定します。

1. ロケーションサブグループごとの num_writers の数を変更することにより、地域分散クラスターのスループットを向上させます。値が高いほどスループットは向上しますが、メモリ使用量が増加します。場所のサブグループごとに繰り返します。

```
SELECT bdr.alter_node_group_option(
    node_group_name := <location_a>,
    config_key := num_writers,
    config_value := 4
);
```

1. 大規模なトランザクションが最初にディスクに書き込まれるのではなく、ライタプロセスに直接送信されるように streaming_mode を writer に設定し、受信ノードの I/O オーバーヘッドを削減します。

```
SELECT bdr.alter_node_group_option(
    node_group_name := <location_a>,
    config_key := streaming_mode,
    config_value := writer
);
```

デフォルトは auto で、PGD は使用可能なリソースに基づいて file と writer のいずれかを選択します。書き込み量が一貫して多い地域分散クラスターには、writer を明示的に設定することをお勧めします。トレードオフは、追加のリソース使用量です。writer モードでは、大規模なトランザクションごとに専用のライタプロセスが必要です。file モードでは、PGD は適用する前に一時ディスクファイルに書き込みます。

`streaming_mode = 'writer'` は、*Commit At Most Once* と互換性がありません。クラスターが CAMO を使用している場合、`streaming_mode` をデフォルトの `auto` のままにします。

### 遅延を適用する

サブスクライバ専用のサブグループで `apply_delay` を使用して遅延リードレプリカを作成し、偶発的なデータ損失または論理的破損に対するリカバリウィンドウを提供します。遅延は、オリジンでのトランザクションの元のコミット時間から測定されます。デフォルトは `0s` 遅延なしです。

```
SELECT bdr.alter_node_group_option(
    node_group_name := <delayed_read_replica_group>,
    config_key := apply_delay,
    config_value := 30min
);
```

**注釈**

データノードグループまたはトップレベルグループで `apply_delay` を設定しないでください。設定すると、アクティブなクラスターに意図的なレプリケーションラグが発生します。

### シークエンス

標準の Postgres シーケンスは各ノードにローカルであり、複数の場所に同時に書き込むと主キーの競合が発生します。すべてのノードにわたって一意性を保証する PGD グローバルシーケンスを使用するように地理分散クラスターを構成します。`bdr.default_sequence_kind` 構成パラメーターは、新しく作成されたシーケンスとシリアル列に適用されるタイプを制御します。

- `snowflakeid` —タイムスタンプベースのアルゴリズムを使用して、ノード間通信なしで一意の値を生成します。64 ビットの `bigint` または `bigserial` 列でのみ動作します。ネットワークパーティションとクロスリージョンのレイテンシーに対する耐性があります。
- `galloc` —クラスター全体のカウンターから値の範囲を割り当てます。新しい範囲が必要な場合はクォーラムが必要です。`smallint` 、`integer` 、および `bigint` をサポートしますが、クォーラムの可用性の影響を受けます。
- `distributed` デフォルト — `bigint` シーケンスの場合は `snowflakeid` を、`smallint` および `integer` シーケンスの場合は `galloc` を自動的に使用します。

地域分散クラスターの場合、デフォルトの `distributed` 設定を使用することをお勧めします。可能な場合、大規模なシーケンスの `galloc` のクォーラム依存関係を回避しながら、より小さい整数型をサポートします。

**注釈**

複数の書き込み場所がある地域分散クラスターでは *bdr.default_sequence_kind = local* の設定を避けてください。ローカルシーケンスは、クラスター全体の一意性を保証せず、マルチマスターレプリケーションで挿入競合の原因となります。

グローバルシーケンスタイプと動作の詳細については、*Sequences* を参照してください。

### Performing cluster maintenance

実稼働に移行する前にクラスター構成を確認し、継続的なメンテナンスを注意深く計画します。単一ロケーションのクラスターでの簡単な操作は、複数のロケーションがアクティブな場合、クロスロケーションフェールオーバーをトリガーし、レプリケーションラグを導入し、競合を発生させる可能性があります。

#### クラスター構成の検証

セットアップが完了したら、運用に行く前に、サブグループ、ルーティング、およびコミットスコープが正しく構成されていることを確認します。

- サブグループとそのルーティング設定を確認します。

```
SELECT node_group_name, node_group_enable_raft, node_group_enable_routing
FROM bdr.node_group;
```

- 現在の書き込みリーダーを確認します。

```
SELECT node_group_name, write_lead
FROM bdr.node_group_routing_summary;
```

- コミットスコープ構成を確認します。

```
SELECT commit_scope_name, commit_scope_origin_node_group, commit_scope_rule
FROM bdr.commit_scopes;
```

#### データベースの変更の管理

地理分散クラスターでは、スキーマの変更と日常的なメンテナンスの動作が異なります。DDL はクラスター全体のロックを取得しますが、 `VACUUM` や `ANALYZE` のような操作は各ノードでローカルに実行されますが、レプリケーションスロットのラグの影響を受ける可能性があります。

#### DDL の実行

DDL 操作はクラスター全体のロックを取得しますが、ロックのタイプとタイミングは操作によって異なります。`ALTER TABLE` や `CREATE INDEX` などのスキーマ変更は、すべてのノードでグローバル DDL ロックを同時に取得し、ロック取得時間は最も遠いノードへの往復時間とほぼ等しくなります。`DROP TABLE` などのオペレーションは、レプリケーションもフラッシュするグローバル DML ロックを取得し、完了時間はネットワーク遅延のみではなく現在のレプリケーションラグに依存します。

どちらの場合も、長時間実行 DDL は、ロックが保持されている間レプリケーションをブロックし、失敗した DDL は場所間で一貫性のない状態を残す場合があります。

リスクを軽減するには、DDL トランザクションを最小限に抑え、単一のトランザクションで複数の DDL 変更を組み合わせないようにします。実際のクロスロケーションのレイテンシーをシミュレートするステージング環境で DDL タイミングをテストします。メンテナンス時間帯に DDL をスケジュールします。

### VACUUM と ANALYZE の実行

VACUUM および ANALYZE は、各ノードでローカルに実行され、レプリケートされません。Autovacuum は各ノードで独立して動作するため、場所を超えて特別な調整は必要ありません。

PGD 固有の懸念事項は、レプリケーションスロットの遅延です。オフラインまたは遅延ノードのスロットは catalog_xmin を抑制し、VACUUM がカタログテーブルをクリーニングできず、元のノードに肥大化が蓄積します。スロットラグを定期的にチェックして、これを早期にキャッチします。

```
SELECT target_name, state, replay_lag_bytes, replay_lag_size, xmin, catalog_xmin
FROM bdr.node_slots
ORDER BY replay_lag_bytes DESC;
```

### 書き込みリーダーシップの管理

地理的分散クラスターでは、ノードの障害または計画的な操作が原因で、書き込みリーダーシップが場所間を移動する場合があります。レイテンシーに合わせてフェールオーバー検出を構成し、ノードをオフラインにする前にクォーラムを維持し、計画的なスイッチオーバーを使用して書き込みリーダーシップを安全に転送します。

### フェイルオーバーの処理

書き込みリーダーの障害が発生すると、PGD は TCP キープアライブと接続タイムアウトを介してそれを検出し、Raft コンセンサスを介して新しい書き込みリーダーを選択します。クロスロケーションのレイテンシーと許容可能なフェイルオーバー時間と一致するように次のパラメーターを調整します。

- bdr.global_keepalives_idle 、bdr.global_keepalives_interval 、および bdr.global_keepalives_count は、障害が発生した接続が検出される速さを制御します。
- bdr.global_connection_timeout は、接続を確立するときに待機する最大時間を設定します。高遅延の展開では、この値を小さくして、完全な TCP タイムアウトを待たずに障害をより迅速に検出します。
- bdr.raft_global_election_timeout は、新しい書き込みリーダーがトップレベルグループで選択されるまでの時間を制御します。bdr.raft_group_election_timeout は、ロケーションごとのサブグループ選択に適用されます。予想されるクロスロケーション遅延がネットワークジッターからの誤選択を引き起こしている場合、これらの値を増やします。

### クォーラムの維持

ノードをオフラインにする前に、クラスターがクォーラム投票ノードのマジョリティを維持していることを確認します。クォーラムがないと、新しい書き込みリーダーは選択されず、クラスターは読み取り専用状態になり、手動介入が必要です。続行する前に、現在の投票ノードの数を確認してください。

```
SELECT nvoting_nodes, is_voting FROM bdr.stat_raft_state;
```

残りの投票ノードの数がマジョリティを下回る場合、ノードをフェンスまたは停止しないでください。2 つの場所の展開では、3 番目の場所の監視ノードは、1 つの場所に障害が発生した場合にクォーラムを維持するために必要な追加の投票を提供します。詳細は、ノード分散の計画 を参照してください。

### 計画的なスイッチオーバーの実行

計画的なスイッチオーバーの前に、ターゲットノードが正常であり、追い付いていることを確認します。現在のリーダーでアクティブな長時間実行トランザクションがないことを確認し、可能な場合はトラフィックの少ない時間帯にスイッチオーバーを実行します。

1. ターゲットノードのレプリケーションスロットがアクティブでストリーミングで、遅延がほぼゼロであることを確認します。

```
SELECT target_name, active, state, replay_lag_bytes, replay_lag_size
FROM bdr.node_slots
WHERE target_name = <target_node_name>;
```

スロットには、ゼロに近い `active = true` 、`state = 'streaming'` 、および `replay_lag_bytes` が表示されるはずです。状態が `catchup` または `disconnected` である場合、または遅延が高い場合、続行する前に待機します。

1. スイッチオーバーを開始します。

```
SELECT bdr.routing_leadership_transfer(
    node_group_name := <group_name>,
    leader_name := <target_node_name>
);
```

1. クラスターが正常な状態であることを確認します。
2. 新しい書き込みリーダーが選択されたことを確認します。

```
SELECT node_group_name, write_lead
FROM bdr.node_group_routing_summary;
```

1. プロキシルーティングが新しいリーダーに更新されたことを確認します。

```
SELECT node_group_name, write_lead_name, previous_write_lead_name
FROM bdr.stat_routing_state;
```

1. 新しいリーダーから他のノードへのレプリケーションが低い遅延でストリーミングされていることを確認します。

```
SELECT target_name, active, state, replay_lag_bytes, replay_lag_size
FROM bdr.node_slots
WHERE origin_name = <new_leader_node_name>;
```

すべてのスロットで、ゼロに近い `state = 'streaming'` および `replay_lag_bytes` が表示されるはずです。

### ノードのメンテナンスの実行

ローリング操作の場所を認識した順序に従い、ノードをオフラインにする前にフェンスします。

### ロケーションをまたいだローリングオペレーション

複数のノードに影響を与えるアップグレードまたはメンテナンスを実行する場合、次の順序に従って、クロスロケーションフェイルオーバーを最小限に抑え、できるだけ長く書き込みをローカルに保ちます。この順序は、各場所に独自の書き込みリーダーがあるローカルルーティングを想定しています。グローバルルーティングでは、クラスター全体に単一の書き込みリーダーがあるため、手順 2 をスキップして、最後にグローバル書き込みリーダーを処理します。

1. セカンダリロケーションの非リーダーノードから開始します。
2. セカンダリの場所の書き込みリーダーに移動しますこれにより、そのロケーション内でローカルフェイルオーバーが発生します。グローバルルーティングを使用する場合、手順 2 をスキップします。
3. プライマリロケーションの非リーダーノードに進みます。
4. プライマリ書き込みリーダーを最後に処理します。プライマリ書き込みリーダーのダウンは、クロスロケーションフェイルオーバーを発生させる可能性が最も高い操作です。

### フェンシングノード

フェンシングは、クラスターから削除せずにノードを書き込みリーダー資格から削除します。ノードの再起動が必要なメンテナンスの前、アップグレードの前、およびノードの状態に疑問がある場合に、フェンシングを使用します。

ノードをフェンスするには

```
SELECT bdr.alter_node_option(
    node_name := <node_name>,
    config_key := route_fence,
    config_value := true
);
```

フェンスされたノードはクォーラムに参加し、レプリケーションの受信を継続するため、クラスターのコンセンサスを維持する能力に影響はありません。ただし、複数のノードを同時にフェンシングすると、可用性が低下する可能性があります。メンテナンス完了後は常にアンフェンス。

フェンスを解除するには

```
SELECT bdr.alter_node_option(
    node_name := <node_name>,
    config_key := route_fence,
    config_value := false
);
```

### クラスターの監視とインシデントへの対応

操作の一部として、レプリケーションスロットの状態と競合率を定期的に確認します。以下のビューを使用して、インシデントが発生した場合の影響を評価し、復旧を確認します。

### ネットワークパーティションの処理

場所がパーティション化されると、クォーラムを保持する側が動作を継続します。クォーラムのない側は読み取り専用になります。分離された側のレプリケーションスロットは一時停止し、パーティションの期間中に遅延が増加します。

パーティションが修復したら、操作を再開する前に復旧を確認します。

1. bdr.node_slots のラグメトリックを使用してキャッチアップの進行状況を監視し、すべてのスロットが streaming に戻るまで待機します。
2. パーティションウィンドウの競合率を確認します。競合のモニタリング を参照してください。

### 競合のモニタリング

ジオレプリケーションでは、同じ行が異なる場所で同時に更新される可能性があるため、競合の可能性が増加します。デフォルトでは、PGD は last-update-wins を使用して競合を解決します。タイムスタンプベースのリゾルバーは、ノード間のクロック同期に依存しています。すべてのノードで NTP または chronyd を実行することにより、クロックスキューを最小限に抑えます。

bdr.conflict_history_summary を定期的にチェックして、競合率の高いテーブルがないかどうかを確認します。

```
SELECT nspname, relname, conflict_type, count(*)
FROM bdr.conflict_history_summary
GROUP BY nspname, relname, conflict_type
ORDER BY count(*) DESC;
```

特定のテーブルでの高い競合率が持続する場合は、アプリケーションチームと再検討する価値があるアクセスパターンを示しています。永続的な競合は通常、デザインの問題です。ロケーションベースのデータの分割、追加専用パターン、または CRDT データ型への切り替えは、競合解決を調整するよりも長期的に効果的です。

### 競合リゾルバーの構成

アクセスパターンの変更が現実的でなく、last-update-wins が特定のテーブルに対して誤った結果を生成する場合、bdr.alter_node_set_conflict_resolver() を使用してカスタムリゾルバーを構成します。使用可能なストラテジーは update_if_newer デフォルト、skip 、および error です。ファンクション呼び出しはレプリケートされないため、クラスター内の各ノードで実行します。最初に SELECT node_name FROM bdr.node_summary を使用してノード名を確認します。

```
SELECT bdr.alter_node_set_conflict_resolver(
    node_name         => node1,
    conflict_type     => update_update,
```

(次のページに続く)

(前のページからの続き)

```
    conflict_resolver => error
);
```

すべての競合タイプにわたって現在のリゾルバー構成を確認するには

```
SELECT * FROM bdr.node_conflict_resolvers;
```

競合タイプとリゾルバーの完全なリストについては、**欠落列の競合の解決** を参照してください。

## 9.3 Using connection pooling in your application

*Connection Manager Authentication* の接続プーリングで動作するようにアプリケーションを調整します。

プーリングの有効化と操作に関する DBA ガイダンスについては、*Configuring connection pooling* を参照してください。

### 9.3.1 アクティブプールモードの理解

`bdr.node_group_summary` を照会して、アクティブなプールモードを確認します。

```
SELECT node_group_name, server_pool_mode FROM bdr.node_group_summary;
```

プールモードは、トランザクション間でどのセッション状態を持続するかを決定します。

| Feature | *none*/*session* | *transaction* |
|---|---|---|
| *SET* changes persist | Yes | No |
| *PREPARE*/*EXECUTE* statements persist | Yes | No |
| Temporary tables persist | Yes | No |
| Advisory locks across transactions | Yes | No |
| Holdable cursors (*WITH HOLD*) | Yes | No |
| *LISTEN* subscriptions | Yes | No |

`none` および `session` モードでは、アプリケーションは、ほとんどの場合 Postgres 直接接続の場合と同じように動作します。例外は、接続マネージャーが特定の接続パラメーターのセットのみをバックエンドに転送することです。詳細は、**セッションパラメーターの管理** を参照してください。 `none` と `session` の違いはインフラレベルです。`session` モードでは、クライアントが切断した後にバックエンド接続が再利用され、サーバーオーバーヘッドが削減されますが、これはアプリケーションには見えません。

`transaction` モードでは、各トランザクションが終了し、ほとんどのセッション状態がクリアされると、バックエンド接続はプールに返されます。次のトランザクションは別のバックエンドで実行できます。一部の Postgres 機能は、トランザクション間で持続せず、結果として動作しないセッション状態に依存しています。完全なリストについては、*Connection pooling's unsupported features* を参照してください。

アプリケーションがこれらの機能に依存している場合は、DBA に session または none モード *Configuring connection pooling* を参照してください、または影響を受けるコードパスを再設計して、単一のトランザクション内のすべての状態を維持します。

### 9.3.2 アプリケーションをトランザクションモードに適応させる

Postgres の直接接続で動作する一部のアプリケーションパターンは、トランザクションモードでの調整が必要です。

#### セッションパラメーターの管理

接続マネージャーは、特定の接続パラメーターのセットをバックエンドに転送し、トランザクションモードでの新しいバックエンドの割り当てごとにそれらを再適用します。認識するパラメーターは client_encoding 、DateStyle 、TimeZone 、standard_conforming_strings 、application_name 、search_path 、および extra_float_digits です。

これらを SET コマンドではなく接続文字列で設定して、プールが割り当てるすべてのバックエンドに適用されるようにします。

```
host=myhost dbname=mydb options=-c search_path=myschema -c work_mem=256MB
```

または URI で。

```
postgresql://myhost/mydb?options=-c%20search_path%3Dmyschema%20-c%20work_mem%3D256MB
```

認識されるリストの外部のパラメーターは、 options 接続パラメーターを介して渡されない限り、バックエンドの変更に再適用されません。

#### プリペアドステートメントを使用する

接続マネージャーは、バックエンドで欠落しているプリペアドステートメントを自動的に検出し、オンデマンドで再準備するため、拡張クエリプロトコルを介して送信されるプリペアドステートメントは、トランザクション間でシームレスに動作します。ほとんどのクライアントライブラリは、デフォルトで拡張クエリプロトコルを使用します。

DISCARD ALL は、各トランザクションの最後にバックエンド接続がプールに返されたときに実行され、PREPARE /EXECUTE SQL ステートメントで作成されたプリペアドステートメントの割り当てを解除します。これらのエラーを回避するには、次のプラクティスに従ってください。

- PREPARE 、EXECUTE 、および DEALLOCATE SQL ステートメントは避けます。代わりに拡張クエリープロトコルを使用します。ほとんどのクライアントライブラリはデフォルトで使用します。
- PREPARE /EXECUTE /DEALLOCATE SQL ステートメントを使用する必要がある場合は、使用するトランザクション内で発行します。

## 9.4 Application use

PGD でのアプリケーションの開発は、PostgreSQL データベースでの作業とほぼ同じです。ただし、異なるのは、アプリケーションがレプリケーションとどのように相互作用するかを認識する必要があることです。PGD がアプリケーションでどのように動作するか、レプリケートされる SQL 、さまざまなノードがどのように処理されるか、およびその他の重要な情報を知る必要があります。

- *DML and DDL replication and nonreplication* は、SQL ステートメントの 2 つのクラスの違いと、PGD がそれらのレプリケートを処理する方法を示します。また、PGD がまったくレプリケートしないコマンドも調べます。
- *Nodes with differences* は、レプリケートノードに異なるテーブル構造とスキーマが存在する構成で PGD がどのように動作するかを調べます。また、このようなノードを LiveCompare で比較する方法、およびノードで実行されている PostgreSQL バージョンの違いを処理する方法についても説明します。
- *Internal timing of operations* は、非同期/同期レプリケーションがアプリケーションのデータのビューにどのような影響を与える可能性があるかを示し、古い読み取りを軽減する機能を示します。
- *Using extensions with PGD* は、PGD で拡張機能を選択、インストール、および構成する方法を説明します。
- *Use of table access methods (TAMs) in PGD* TAMs は、PGD で使用可能な TAM とそれらを有効にする方法をノートします。
- *Feature compatibility* は、どのサーバー機能がどのコミットスコープで動作するか、およびどのコミットスコープをデイジーチェーン接続できるかを示します。

### 9.4.1 DML and DDL replication and nonreplication

SQL ステートメントの 2 つのメジャークラスは、 DML と DDL です。

- DML はデータ変更言語であり、テーブルに格納されているデータを変更する SQL ステートメントに関係します。これには、UPDATE、DELETE、および INSERT が含まれます。
- DDL はデータ定義言語であり、データの保存方法を変更する SQL ステートメントに関係します。これには、CREATE、ALTER、および DROP が含まれます。

PGD は、各クラスを個別に処理します。

#### DML レプリケーション

PGD は DML ステートメントをレプリケートしません。DML ステートメントによって発生した変更をレプリケートします。たとえば、2 つの行を変更した UPDATE は 2 つの変更をレプリケートしますが、行を削除しなかった DELETE は何もレプリケートしません。これは、volatile ステートメントの実行結果がレプリケートされ、ステートメントベースのレプリケーションで発生する可能性のあるノード間の相違がないことを意味します。

#### DDL レプリケーション

DDL レプリケーションは DML とは異なります。DDL の場合、PGD はステートメントをレプリケートし、すべてのノードで実行します。したがって、`DROP TABLE IF EXISTS` はローカルノードで何もレプリケートしない場合がありますが、DDL レプリケーションが有効になっている場合、ステートメントは他のノードに送信されて実行されます。詳細については、*Running DDL safely* を参照してください。

PGD は、同じトランザクション内でも、混合された DML および DDL ステートメントが正しく動作するように機能します。

#### レプリケートされていないステートメント

これらの 2 つのクラスの外部には、PGD がデザインでレプリケートしない SQL コマンドがあります。次のユーザーコマンドは PGD によってレプリケートされないため、その効果はローカル/オリジンノードでのみ発生します。

- カーソル操作 DECLARE、CLOSE、FETCH
- 実行コマンド DO、CALL、PREPARE、EXECUTE、EXPLAIN
- セッション管理 DEALLOCATE、DISCARD、LOAD
- パラメーターコマンド SET、SHOW

・コンストレイン操作 SET CONSTRAINTS

- ロックコマンド LOCK
- テーブルメンテナンスコマンド VACUUM、ANALYZE、CLUSTER
- 非同期操作 NOTIFY、LISTEN、UNLISTEN

`NOTIFY` SQL コマンドと `pg_notify()` ファンクションはレプリケートされないため、フェイルオーバーの場合の通知は信頼できません。これは、サーバーがクラッシュしたときにトランザクションがコミットされた場合、フェールオーバー時に通知が簡単に失われる可能性があることを意味します。`LISTEN` を実行しているアプリケーションは、フェイルオーバーが発生した場合の通知を見逃す場合があります。

これは標準の PostgreSQL レプリケーションに当てはまりますが、PGD はこれをまだ改善していません。

CAMO および Eager Replication オプションでは、`NOTIFY` SQL コマンドまたは `pg_notify()` ファンクションは使用できません。

### 9.4.2 Nodes with differences

#### 差分を含むノード間のレプリケーション

デフォルトでは、 DDL はすべてのノードに送信されます。*Running DDL safely* で説明しているように、この動作を制御でき、それを使用して、ノード間のデータベーススキーマ間の差異を作成できます。PGD は、ノード間のわずかな違いがあってもレプリケーションを続行できるように設計されています。これらの機能は、ダウンタイムなしでアプリケーションスキーマを移行できるように、またはロジカルスタンバイノードでレポートまたはテストをできるように設計されています。

現在、レプリケーションには、すべてのノードで同じテーブル名が必要です。将来の機能では、異なるテーブル名間のマッピングが可能になる可能性があります。

ターゲット上のパーティションを変更する更新のサポートを含む、パーティションテーブルにレプリケートする通常のテーブルのソースなど、異なるパーティション定義を持つテーブル間でレプリケートが可能です。パーティショニング定義がソースとターゲットで同じである場合、適用時に動的パーティションルーティングを実行する必要がないため、高速になります。詳細については、*Replication sets* を参照してください。

デフォルトでは、すべての列がレプリケートされます。

PGD は、列名に基づいてデータ列をレプリケートします。列の名前が同じで異なるデータ型がある場合、PGD はソースタイプからターゲットタイプにキャストを試行します（それを許可するキャストが定義されている場合）。

PGD は、列数が異なるテーブル間のレプリケーションをサポートしています。

ターゲットにソースからの欠落列がある場合、PGD は `target_column_missing` 競合を発生させます。デフォルトの競合リゾルバーは `ignore_if_null` です。これは、NULL 以外の値が到着した場合にエラーをスローします。または、`ignore` の競合リゾルバーを使用してノードを構成することもできます。この設定はエラーをスローしませんが、追加の列をサイレントに無視します。

ターゲットにソースレコードで見られない追加の列がある場合、PGD は `source_column_missing` 競合を発生させます。デフォルトの競合リゾルバーは `use_default_value` です。追加列にデフォルト NULL ヌル可能な場合またはデフォルト式がある場合、レプリケーションは続行されます。そうでない場合、エラーをスローし、レプリケーションを停止します。

変換トリガーはテーブルで使用して、デフォルト値を提供したり、適用前にさまざまな方法で受信データを変更したりすることもできます。

ソースとターゲットに異なる制約がある場合、レプリケーションが試行されますが、ソースからの行をターゲットに適用できない場合、失敗する場合があります。ここで行フィルタが役立ちます。

1 つのスキーマからより緩和されたスキーマにデータを複製しても、障害は発生しません。スキーマからより制限的なスキーマへのデータの複製は、潜在的な障害のソースになる可能性があります。これを解決する正しい方法は、より緩和な側にコンストレインを配置して、悪いデータを入力できないようにすることです。このようにして、レプリケーションによって不正なデータが到着することはなく、より制限的なスキーマへの変換に失敗することはありません。たとえば、あるスキーマに TEXT タイプの列があり、別のスキーマが XML と同じ列を定義している場合、TEXT 列に CHECK 制約を追加して、テキストが XML であることを強制します。

ノードごとに異なるインデックスを使用してテーブルを定義できます。デフォルトでは、インデックス定義はレプリケートされます。ノードのサブセットのみまたはローカルにインデックスを作成する方法を指定するには、*Running DDL safely* を参照してください。

`fillfactor` や `toast_tuple_target` などのストレージパラメーターは、テーブルのノード間で問題なく異なる場合があります。その動作の例外は、テーブルのストレージパラメーター `user_catalog_table` の値がすべてのノードで同一である必要があることです。

レプリケートされるテーブルは、各ノードで同じユーザー/ロールが所有している必要があります。詳細は、*Security and roles* を参照してください。

ロールは、各ノードの接続に対して異なるパスワードを持つことができますが、デフォルトでは、ロールの変更は各ノードにレプリケートされます。ノードのサブセットのみまたはローカルでロールパスワードを変更する方法を指定するには、*Running DDL safely* を参照してください。

#### 違いがあるノード間の比較

LiveCompare は、PGD ノードおよび非 PGD ノードとデータベースのデータを比較するためのツールです。比較して最終結果に到達するには、少なくとも 2 つの接続が必要です。

LiveCompare 1.3 以降、`all_bdr_nodes` セットを使用して構成できます。この設定により、クラスター内の個別のノードごとに関連するすべての DSN を明確にする必要がありません。EDB Postgres 分散クラスターには、接続情報を備えた N 個のノードがありますが、LiveCompare 1.3 以降がジョブを完了するために必要な初期接続と出力接続のみです。`logical_replication_mode` 設定は、すべてのノードがどのように通信しているかを示します。

すべての構成は、たとえば、 `bdrLC.ini` という名前の`.ini` ファイルで実行されます。 `/etc/2ndq-livecompare/` でこの構成ファイルのテンプレートを見つけます。

LiveCompare の実行中、N+1 の進行状況バーが表示されます。N はプロセス数です。すべてのテーブルがソースされると、1 秒あたりのトランザクション tps が測定されたとして時間が表示されます。このメカニズムは時間をカウントし続け、推定を提供して、最後に合計実行時間を提供します。

このツールは、テーブル、スキーマ、replication_sets など、多くのカスタマイズとフィルターを提供します。LiveCompare は、コンテキスト情報を失うことなく stop-start を使用できるため、都合の良いときに実行できます。比較の後、概要と DML スクリプトが生成されるので、確認できます。DML を適用して、見つかった差異を修正します。

#### 異なるリリースレベル間のレプリケーション

ノード間のその他の違いは、ノードに PostgreSQL の異なるメジャーバージョンが存在するかどうかです。PGD は、異なるメジャーリリースバージョン間で複製するように設計されています。この機能は、ダウンタイムなしでメジャーバージョンのアップグレードを可能にするように設計されています。

PGD は、PGD ソフトウェアの異なるバージョンを持つノード間で複製するように設計されています。この機能は、ダウンタイムなしでバージョンのアップグレードとメンテナンスを可能にするように設計されています。

ただし、クラスター内のメジャーバージョンのノードを参加することは可能ですが、クラスターが新しいプロトコルバージョンを使用している場合、マイナーバージョンのノードを追加することはできません。そうするとエラーが返されます。

これらの機能は両方とも、特定の制限の影響を受ける可能性があります。既知の非互換性については、Releasenotes を参照してください。

### 9.4.3 Timing considerations and synchronous replication

デフォルトでは非同期であるため、ピアノードは遅延する場合があります。この動作により、複数の PGD ノードに接続したクライアント、またはそれらをスイッチングしたクライアントが古いデータを読み取ることができます。

*queue waitfunction* は、クライアントまたはプロキシに提供され、このような古い読み取りを防止します。

Postgres の同期レプリケーション機能は、PGD でも利用できます。さらに、PGD は、より同期したレプリケーションのための複数のバリアントを提供します。使用可能なすべてのバリアントとそのさまざまなモードの概要と比較については、*Commit scopes* を参照してください。

### 9.4.4 Using extensions with PGD

#### PGD およびその他の PostgreSQL 拡張機能

PGD は、PostgreSQL 拡張機能として実装されています

Supported Postgres database servers を参照してください。PostgreSQL の拡張性と柔軟性を利用して、低レベルのシステムの動作を変更して、マルチマスターレプリケーションを提供します。

原則として、コミュニティ PostgreSQL、 EDB Postgres Advanced Server が提供する拡張機能、およびサードパーティの拡張機能は PGD で使用できます。ただし、PGD の分散的な性質は、選択してインストールする拡張機能を注意深く検討および計画する必要があることを意味します。

#### 論理デコーディングを提供する拡張機能

wal2json などの論理デコードを提供する拡張機能は、理論的には PGD で動作すると考えられます。ただし、フェイルオーバーはサポートされていないため、このような拡張機能から読み取られる WAL ストリームが中断される可能性があることを意味します。

#### レプリケーションまたは HA 機能を提供する拡張機能

レプリケーションまたは HA /フェールオーバーに関連する機能で PostgreSQL を拡張する拡張機能は、PGD でうまく動作する可能性が低く、PGD クラスターの状態に悪影響を与える可能性があります。これらは避けることをお勧めします。

#### サポートされている拡張機能

これらの拡張機能は PGD によって明示的にサポートされています。

#### EDB Advanced Storage テーブルのアクセス方法

*EDB Advanced Storage Pack* は、拡張機能として実装されるテーブルアクセス方法 TAM の選択を提供します。次の TAM は、PGD での使用が認定されています。

- *Autocluster*
- *Refdata*

詳細については、 *Use of table access methods (TAMs) in PGD* を参照してください。

#### pgaudit

PGD は、 pgaudit 拡張機能との互換性を確保するために変更されました。構成情報については、*Postgres 設定* を参照してください。

#### 拡張機能のインストール

PostgreSQL 拡張機能は、ファンクション、データ型、およびオプションで 1 つ以上の共有ライブラリなどの SQL オブジェクトを提供します。拡張機能をインストールして使用する前に、これらを PostgreSQL バックエンドにロードする必要があります。

> **警告**
>
> 関連する拡張機能パッケージは、クラスター内のすべてのノードで利用できる必要があります。そうしないと、拡張機能のインストールが失敗し、クラスターの安定性に影響を与える可能性があります。

Trusted Postgres Architect を使用して PGD が展開されている場合、そのツールを使用して拡張機能を構成します。詳細については、Adding Postgres extensions を参照してください。

以下は、手動で構成された PGD インストールに関連します。

#### shared_preload_libraries の構成

拡張機能が共有ライブラリを提供する場合、拡張機能をインストールする前にこのライブラリを shared_preload_libraries 構成パラメーターに含めます。

`shared_preload_libraries` は、拡張機能名のコンマ区切りのリストで構成されます。 `bdr` を含める必要があります。通常、他の拡張機能を指定する順序は重要ではありません。

`CREATE EXTENSION` で拡張機能をインストールする前に、クラスター内のすべてのノードで `shared_preload_libraries` を構成します。新しい構成をアクティブにするには、PostgreSQL を再起動する必要があります。

*Postgres* **設定** も参照してください。

#### 拡張機能のインストール

`CREATE EXTENSION` コマンドを使用して拡張機能をインストールします。これは、クラスター内の 1 つのノードでのみ行う必要があります。PGD の DDL レプリケーションは、他のすべてのノードに伝播します。

> **警告**
>
> *CREATE EXTENSION* を実行する前に DDL レプリケーションを無効にするなど、各ノードに拡張機能を手動でインストールしようとしないでください。

`bdr.replicate_ddl_command()` などのコマンドを使用して `CREATE EXTENSION` を実行しないでください。

### 9.4.5 Use of table access methods (TAMs) in PGD

*EDB Advanced Storage Pack* は、 EDB Postgres 15.0 から利用可能なテーブルアクセス方法 TAM を提供します。

次の TAM は、PGD 6.0 での使用が認定されています。

- *Autocluster*
- *Refdata*

EDB によって認定されるまで、他の TAM の使用は制限されます。

PGD クラスターでこれらの TAM のいずれかを使用するには、適切な拡張ライブラリ `autocluster` および/または `refdata` を各ノードの `shared_preload_libraries` パラメーターに追加し、PostgreSQL サーバーを再起動する必要があります。

クラスター内のすべてのノードの `shared_preload_libraries` に拡張ライブラリが存在すると、 `CREATE EXTENSION autocluster;` または `CREATE EXTENSION refdata;` を使用して拡張機能自分自身を作成できます。 `CREATE EXTENSION` コマンドは、1 つのノードでのみ実行する必要があります。クラスター内の他のノードに複製されます。

拡張機能を作成した後、 `CREATE TABLE test USING autocluster;` または `CREATE TABLE test USING refdata;` を使用して、指定された TAM でテーブルを作成します。これらのコマンドは、クラスター内のすべての PGD ノードに複製します。

これらのテーブルアクセス方法の詳細については、以下を参照してください。

- *Autocluster example*
- *Refdata example*

### 9.4.6 Feature compatibility

#### サーバー機能/コミットスコープの相互運用性

すべてのサーバー機能がすべてのコミットスコープで動作するわけではありません。この表は、相互運用するものを示しています。

| | Async | Parallel Apply | Transaction Streaming | Single Decoding Worker |
|---|---|---|---|---|
| *Group Commit* | N/A | Does not interoperate | Does not interoperate Attempting to use Group Commit and Transaction | Interoperates |
| *CAMO* | N/A | Interoperates | Does not interoperate | Does not interoperate |
| *Lag Control* | Interoperates | Interoperates | Interoperates | Interoperates |
| *Synchronous Commit* | N/A | Interoperates | Interoperates | Interoperates |

#### 注

N/A : 表の Async 列は、使用中の同期コミットスコープのない PGD を表します。Lag Control は同期コミットスコープではありません。これは制御コミットスコープであるため、非同期操作で使用できます。

グループコミットとトランザクションストリーミングを使用しようとすると、警告が表示されます。警告は、トランザクションストリーミングを無効にすることを示唆しており、トランザクションが発生しているように見えます。バックグラウンドでは、グループコミットが無効になって、トランザクションが発生できるようになりました。

### コミットスコープ/コミットスコープの相互運用性

コミットスコープを混合することはできませんが、 `AND` 演算子を使用して *combine rules* を行うことはできます。次の表は、コミットスコープを結合できる場所を示しています。

| | *Group Commit* | *CAMO* | *Lag Control* | *Synchronous Commit* |
|---|---|---|---|---|
| *Group Commit* | N/A | Does not combine | Combines | Combines |
| *CAMO* | Does not combine | N/A | Combines | Does not combine |
| *Lag Control* | Combines | Combines | N/A | Combines |
| *Synchronous Commit* | Combines | Does not combine | Combines | N/A |

#### 注

各コミットスコープは暗黙的に自分自身を操作します。

# 第10章　Reference

## 10.1 PGD Reference

リファレンスセクションは、 EDB Postgres Distributed で使用可能なすべてのファンクション、ビュー、およびコマンドの最終的なリストです。

- 

Tables, Views, and Functions

- Command Line Interface (CLI)

### 10.1.1 Tables, views and functions reference

リファレンスセクションは、 EDB Postgres Distributed で使用可能なすべてのファンクション、ビュー、およびコマンドの最終的なリストです。

#### User visible catalogs and views

- *bdr.camo_decion_journal*
- *bdr.commit_Decisions*
- *bdr.commit_scopes*
- *bdr.conflict_history*
- *bdr.conflict_history_summary*
- *bdr.consensus_kv_data*
- *bdr.crdt_handlers*
- *bdr.ddl_replication*
- *bdr.depend*
- *bdr.failover_replication_slots*
- *bdr.global_consensus_journal*
- *bdr.global_consensus_journal_details*
- *bdr.global_consensus_response_journal*
- *bdr.global_lock*
- *bdr.global_locks*

- *bdr.group_camo_details*
- *bdr.group_lease*
- *bdr.group_raft_details*
- *bdr.group_replslots_details*
- *bdr.group_subscription_summary*
- *bdr.group_versions_details*
- *bdr.leader*
- *bdr.local_consensus_snapshot*
- *bdr.local_consensus_state*
- *bdr.local_node*
- *bdr.local_node_summary*
- *bdr.local_sync_status*
- *bdr.node_config*
- *bdr.node_catchup_info*
- *bdr.node_catchup_info_details*
- *bdr.node_conflict_resolvers*
- *bdr.node_group_config*
- *bdr.node_group_replication_sets*
- *bdr.node_group_summary*
- *bdr.node_local_info*
- *bdr.node_log_config*
- *bdr.node_peer_progress_mapping*
- *bdr.node_replication_rates*
- *bdr.node_slots*
- *bdr.node_summary*
- *bdr.parted_origin_catchup_info*
- *bdr.parted_origin_catchup_info_details*
- *bdr.pending_failed_changes*
- *bdr.pending_failed_changes_sql*
- *bdr.pgd_largeobject*
- *bdr.pgd_largeobject_metadata*

- *bdr.queue*
- *bdr.replication_set*
- *bdr.replication_set_table*
- *bdr.replication_set_ddl*
- *bdr.replication_sets*
- *bdr.resolved_transactions*
- *bdr.resolved_transactions_info*
- *bdr.schema_changes*
- *bdr.sequence_alloc*
- *bdr.sequences*
- *bdr.stat_activity*
- *bdr.stat_commit_scope*
- *bdr.stat_commit_scope_state*
- *bdr.stat_connection_manager*
- *bdr.stat_connection_manager_connections*
- *bdr.stat_connection_manager_node_stats*
- *bdr.stat_connection_manager_hba_file_rules*
- *bdr.stat_raft_followers_state*
- *bdr.stat_raft_state*
- *bdr.stat_receiver*
- *bdr.stat_receiver_transactions*
- *bdr.stat_reconcile_commit_scope*
- *bdr.stat_relation*
- *bdr.stat_routing_candidate_state*
- *bdr.stat_routing_state*
- *bdr.stat_subscription*
- *bdr.stat_worker*
- *bdr.stat_writer*
- *bdr.subscription*
- *bdr.subscription_summary*
- *bdr.subscription_error_status*

- *bdr.tables*
- *bdr.taskmgr_work_queue*
- *bdr.taskmgr_workitem_status*
- *bdr.taskmgr_local_work_queue*
- *bdr.taskmgr_local_workitem_status*
- *bdr.trigger*
- *bdr.triggers*
- *bdr.workers*
- *bdr.* ライター
- *bdr.worker_tasks*

### System functions

#### バージョン情報ファンクション

- *bdr.bdr_version*
- *bdr.bdr_version_num*

#### システム情報ファンクション

- *bdr.get_relation_stats*
- *bdr.get_subscription_stats*

#### システムおよび進行状況情報パラメーター

- *bdr.local_node_id*
- *bdr.last_committed_lsn*
- *transaction_id*

#### ノードステータスファンクション

- *bdr.is_node_connected*
- *bdr.is_node_ready*

#### コンセンサスファンクション

- *bdr.consensus_disable*
- *bdr.consensus_enable*
- *bdr.consensus_proto_version*
- *bdr.consensus_snapshot_export*

- *bdr.consensus_snapshot_import*
- *bdr.consensus_snapshot_verify*
- *bdr.get_consensus_status*
- *bdr.get_raft_status*
- *bdr.raft_leadership_transfer*

### ユーティリティファンクション

- *bdr.transfer_all_leaderships*
- *bdr.alter_node_kind*
- *bdr.alter_subscription_skip_changes_upto*
- *bdr.connection_manager_refresh_pools*
- *bdr.get_node_sub_apply_lsn*
- *bdr.get_node_sub_receive_lsn*
- *bdr.global_lock_table*
- *bdr.group_lease_override*
- *bdr.local_group_slot_name*
- *bdr.node_group_type*
- *bdr.replicate_ddl_command*
- *bdr.run_on_all_nodes*
- *bdr.run_on_group*
- *bdr.run_on_nodes*
- *bdr.stat_reset*
- *bdr.wait_for_apply_queue*
- *bdr.wait_for_xid_progress*
- *bdr.wait_node_confirm_lsn*
- *bdr.wait_slot_confirm_lsn*

### グローバルアドバイザリーロック

- *bdr.global_advisory_lock*
- *bdr.global_advisory_unlock*

### モニタリング機能

- *bdr.monitor_group_versions*
- *bdr.monitor_group_raft*
- *bdr.monitor_local_replslots*
- *bdr.wal_sender_stats*
- *bdr.get_decoding_worker_stat*
- *bdr.lag_control*

### ルーティング機能

- *bdr.routing_leadership_transfer*

### CAMO ファンクション

- *bdr.is_camo_partner_connected*
- *bdr.is_camo_partner_ready*
- *bdr.get_configured_camo_partner*
- *bdr.wait_for_camo_partner_queue*
- *bdr.camo_transactions_resolved*
- *bdr.logical_transaction_status*

### コミットスコープファンクション

- *bdr.add_commit_scope*
- *bdr.create_commit_scope*
- *bdr.alter_commit_scope*
- *bdr.drop_commit_scope*
- *bdr.remove_commit_scope*

### エラーポリシーファンクションの適用

- *bdr.skip_failed_change*
- *bdr.skip_failed_transaction*
- *bdr.apply_failed_change*
- *bdr.apply_failed_transaction*
- *bdr.modify_failed_change*
- *bdr.reconstruct_change_sql*

- *bdr.reconstruct_failed_transaction*
- *bdr.resolve_failed_transaction*
- *bdr.analyze_skip_impact*
- *bdr.analyze_transaction_skip_impact*
- *bdr.format_skip_anarise*
- *bdr.alter_subscription_reset_error_counters*

### PGD settings

### CRDTs vs conflict handling/reporting

- *bdr.default_conflict_detection*

#### グローバルシーケンスパラメーター

- *bdr.default_sequence_kind*

#### DDL 処理

- *bdr.default_replica_identity*
- *bdr.ddl_replication*
- *bdr.role_replication*
- *bdr.ddl_locking*
- *bdr.truncate_locking*

#### グローバルロック

- *bdr.global_lock_max_locks*
- *bdr.global_lock_timeout*
- *bdr.global_lock_statement_timeout*
- *bdr.global_lock_idle_timeout*
- *bdr.lock_table_locking*
- *bdr.predictive_checks*

### Node management

- *bdr.replay_progress_frequency*

#### 汎用レプリケーション

- *bdr.writers_per_subscription*
- *bdr.max_writers_per_subscription*
- *bdr.xact_replication*
- *bdr.permit_unsafe_commands*
- *bdr.batch_inserts*
- *bdr.maximum_lock_skew*
- *bdr.maximum_lock_skew_action*
- *bdr.accept_connections*
- *bdr.writer_input_queue_size*
- *bdr.writer_output_queue_size*
- *bdr.min_worker_backoff_migrate*

### CRDTs Overview

- *bdr.crdt_raw_value*

### Replicated large objects

- *bdr.override_large_objects*

### Commit scope rules

- *bdr.commit_scopes*

### Commit At Most Once

- *bdr.camo_local_mode_deploy*
- *bdr.camo_enable_client_warnings*

#### トランザクションストリーミングでの使用

- *bdr.default_streaming_mode* を使用したノード構成

#### ラグコントロール

- *bdr.lag_control_max_commit_deploy*
- *bdr.lag_control_max_lag_size*
- *bdr.lag_control_max_lag_time*
- *bdr.lag_control_min_conforming_nodes*
- *bdr.lag_control_commit_deploy_adjust*

- *bdr.lag_control_sample_interval*
- *bdr.lag_control_commit_deploy_start*

#### モニタリングとロギング

- *bdr.debug_level*
- *bdr.trace_level*
- *bdr.track_subscription_apply*
- *bdr.track_relation_apply*
- *bdr.track_apply_lock_timing*

### Decoding worker

- *bdr.enable_wal_decoder*
- *bdr.receive_lcr*
- *bdr.lcr_cleanup_interval*

#### 接続設定

- *bdr.global_connection_timeout*
- *bdr.global_keepalives*
- *bdr.global_keepalives_idle*
- *bdr.global_keepalives_interval*
- *bdr.global_keepalives_count*
- *bdr.global_tcp_user_timeout*

#### トポロジ設定

- *bdr.force_full_mesh*

#### アナリティクス設定

- *bdr.prefer_analytics_engine*

#### 内部設定 - Raft タイムアウト

- *bdr.raft_global_election_timeout*
- *bdr.raft_group_election_timeout*
- *bdr.raft_response_timeout*

#### 内部設定 - その他の Raft 値

- *bdr.raft_keep_min_entries*
- *bdr.raft_log_min_apply_duration*
- *bdr.raft_log_min_message_duration*
- *bdr.raft_group_max_connections*

#### 内部設定 - その他の値

- *bdr.backwards_compatibility*
- *bdr.track_replication_estimates*
- *bdr.lag_tracker_apply_rate_weight*
- *bdr.enable_auto_sync_reconcile*
- *bdr.allow_bdr_schema_changes*

### Node management

- ノードの状態のリスト
- ノード管理コマンド
- *bdr_init_physical*
- *bdr_config*

### Node management interfaces

- *bdr.alter_node_group_option()* を使用したグループ構成
- *bdr.alter_node_interface*
- *bdr.alter_node_option*
- *bdr.alter_subscription_enable*
- *bdr.alter_subscription_disable*
- *bdr.create_node*
- *bdr.create_node_group*
- *bdr.drop_node_group*
- *bdr.join_node_group*
- *bdr.part_node*
- *bdr.promote_node*
- *bdr.switch_node_group*

- *bdr.sync_node_cancel*
- *bdr.wait_for_join_completion*

ルーティングしきい値の設定

**Commit scopes**

- コミットスコープの構文
- *commit_scope_degrade_operation*
- コミットスコープターゲット
- *ORIGIN_GROUP*
- *Commit scope groups*
- 任意
- *ANY NOT*
- *MAJORITY* および *ALL* ノードレプリケーション一般の影響
- マジョリティではない
- *ALTER TABLE* は許可されないコマンド
- すべてではない
- 確認レベル
- 受信時
- レプリケートについて
- *ON DURABLE*
- オンビジブル
- コミットスコープの種類
- 同期コミット
- *DEGRADE ON* パラメーター
- *commit_scope_degrade_operation*
- クォーラムコミット
- *QUORUM COMMIT* パラメーター
- *QUORUM COMMIT ABORT ON* パラメーター
- *commit_*決める設定
- *CAMO* パートナーとの連携
- *DEGRADE ON* パラメーター
- ラグコントロール

- *LAG CONTROL* パラメーター
- 同期コミットとグループコミット
- *GROUP COMMIT* パラメーター
- *QUORUM COMMIT ABORT ON* パラメーター
- *DEGRADE ON* パラメーター
- *transaction_tracking* 設定
- *fence_resolution* 設定
- *GROUP COMMIT commit_*決断設定
- *commit_scope_degrade_operation* 設定

### bdr_read_all_conflicts

- *Column-level conflict detection*
- 競合タイプのリスト
- 欠落列の競合の解決
- 競合リゾルバーのリスト
- デフォルトの競合リゾルバー
- *List of conflict resolutions*
- *Conflict logging*

### Column-level conflict functions

- *bdr.alter_table_conflict_detection* を使用して列レベルの競合解決を有効にする
- *bdr.alter_node_set_conflict_resolver*
- *bdr.alter_node_set_log_config*

### Replication set management

- *bdr.create_replication_set*
- *bdr.alter_replication_set*
- *bdr.drop_replication_set*
- *bdr.alter_node_replication_sets*

#### レプリケーションセットのメンバーシップ

- *bdr.replication_set_add_table*
- *bdr.replication_set_remove_table*

### DDL レプリケーションフィルタリング

- *bdr.replication_set_add_ddl_filter*
- *bdr.replication_set_remove_ddl_filter*

### Testing and tuning commands

- *pgd_bench*

### Sequences

#### シーケンスファンクション

- *bdr.alter_sequence_set_kind*
- *bdr.extract_timestamp_from_snowflakeid*
- *bdr.extract_nodeid_from_snowflakeid*
- *bdr.extract_localseqid_from_snowflakeid*
- *bdr.timestamp_to_snowflakeid*
- *bdr.extract_timestamp_from_timeshard*
- *bdr.extract_nodeid_from_timeshard*
- *bdr.extract_localseqid_from_timeshard*
- *bdr.timestamp_to_timeshard*
- *bdr.galloc_chunk_info*

### KSUUID v2 ファンクション

- *bdr.gen_ksuuid_v2*
- *bdr.ksuuid_v2_cmp*
- *bdr.extract_timestamp_from_ksuuid_v2*

### KSUUID v1 ファンクション

- *bdr.gen_ksuuid_v2*
- *bdr.uuid_v1_cmp*
- *bdr.extract_timestamp_from_ksuuid_v2*

### Replicated large objects

#### ラージオブジェクトインターフェイス

- *bdr.lo_creat*
- *bdr.lo_create*
- *bdr.lo_open*
- *bdr.lo_close*
- *bdr.loread*
- *bdr.lowwrite*
- *bdr.lo_lseek*
- *bdr.lo_lseek64*
- *bdr.lo_tell*
- *bdr.lo_tell64*
- *bdr.lo_truncate*
- *bdr.lo_truncate64*
- *bdr.lo_unlink*
- *bdr.lo_import*
- *bdr.lo_import_with_oid*
- *bdr.lo_export*
- *bdr.lo_get*
- *bdr.lo_get_fragment*
- *bdr.lo_from_bytea*
- *bdr.lo_put*

### AutoPartition in PGD

- *bdr.autopartition*
- *bdr.drop_autopartition*
- *bdr.autopartition_wait_for_partitions*
- *bdr.autopartition_wait_for_partitions_on_all_nodes*
- *bdr.autopartition_find_partition*
- *bdr.autopartition_enable*
- *bdr.autopartition_disable*

- 内部ファンクション
- *bdr.autopartition_create_partition*
- *bdr.autopartition_drop_partition*

### ストリームトリガーリファレンス

#### Stream triggers manipulation interfaces

- *bdr.create_conflict_trigger*
- *bdr.create_transform_trigger*
- *bdr.drop_trigger*

#### Stream triggers row functions

- *bdr.trigger_get_row*
- *bdr.trigger_get_committs*
- *bdr.trigger_get_xid*
- *bdr.trigger_get_type*
- *bdr.trigger_get_conflict_type*
- *bdr.trigger_get_origin_node_id*
- *bdr.ri_fkey_on_del_trigger*

#### Stream triggers row variables

- *TG_NAME*
- *TG_WHEN*
- *TG_LEVEL*
- *TG_OP*
- *TG_RELID*
- *TG_TABLE_NAME*
- *TG_TABLE_SCHEMA*
- *TG_NARGS*
- *TG_ARGV[]*

### Internal catalogs and views

- *bdr.autopartition_partitions*
- *bdr.autopartition_rules*
- *bdr.ddl_epoch*
- *bdr.event_history*
- *bdr.event_summary*
- *bdr.local_leader_change*
- *bdr.node_config*
- *bdr.node_config_summary*
- *bdr.node_group_config*
- *bdr.node_group_routing_config_summary*
- *bdr.node_group_routing_info*
- *bdr.node_group_routing_summary*
- *bdr.node_peer_progress_mapping*
- *bdr.node_routing_config_summary*
- *bdr.prepared_xacts_summary*
- *bdr.reconcile_2pc_requests*
- *bdr.reconcile_2pc_requests_summary*
- *bdr.sequence_kind*
- *bdr.sync_node_requests*
- *bdr.sync_node_requests_summary*

### Internal system functions

#### 一般的な内部ファンクション

- *bdr.bdr_get_commit_decions*
- *bdr.bdr_track_commit_Decision*
- *bdr.consensus_kv_fetch*
- *bdr.consensus_kv_store*
- *bdr.decode_message_payload*
- *bdr.decode_message_response_payload*
- *bdr.difference_fix_origin_create*

- *bdr.difference_fix_session_reset*
- *bdr.difference_fix_session_setup*
- *bdr.difference_fix_xact_set_avoid_conflict*
- *bdr.drop_node*
- *bdr.get_global_locks*
- *bdr.get_node_conflict_resolvers*
- *bdr.get_slot_flush_timestamp*
- *bdr.internal_alter_sequence_set_kind*
- *bdr.internal_replication_set_add_table*
- *bdr.internal_replication_set_remove_table*
- *bdr.internal_submit_join_request*
- *bdr.isolation_test_session_is_blocked*
- *bdr.local_node_info*
- *bdr.msgb_connect*
- *bdr.msgb_deliver_message*
- *bdr.node_catchup_state_name*
- *bdr.node_kind_name*
- *bdr.peer_state_name*
- *bdr.pg_xact_origin*
- *bdr.reconcile_2pc_status_name*
- *bdr.request_replay_progress_update*
- *bdr.reset_relation_stats*
- *bdr.reset_subscription_stats*
- *bdr.resynchronize_table_from_node*
- *bdr.seq_currval*
- *bdr.seq_lastval*
- *bdr.seq_nextval*
- *bdr.show_subscription_status*
- *bdr.show_workers*
- *bdr.show_writers*
- *bdr.sync_status_name*

### タスクマネージャー機能

- *bdr.taskmgr_set_leader*
- *bdr.taskmgr_get_last_completed_workitem*
- *bdr.taskmgr_work_queue_check_status*
- *bdr.get_min_required_replication_slots*
- *bdr.get_min_required_worker_processes*
- *bdr.stat_last_reset_time*
- *bdr.stat_get_activity*
- *bdr.stat_get_commit_scope*
- *bdr.stat_get_commit_scope_proc_wait*
- *bdr.stat_get_reconcile_commit_scope*
- *bdr.worker_role_id_name*
- *bdr.lag_history*
- *bdr.get_raft_instance_by_nodegroup*
- *bdr.monitor_camo_on_all_nodes*
- *bdr.monitor_raft_details_on_all_nodes*
- :ref:``bdr.monitor_replslots_details_on_all_nodes` <*bdr.monitor_replslots_details_on_all_nodes*>`
- *bdr.monitor_subscription_details_on_all_nodes*
- *bdr.monitor_version_details_on_all_nodes*
- *bdr.node_group_member_info*

### Column-level conflict functions

- *bdr.alter_table_conflict_detection* を使用して列レベルの競合解決を有効にする
- *bdr.alter_node_set_conflict_resolver*
- *bdr.alter_node_set_log_config*

### Column-level conflict functions

- *bdr.column_timestamps_create*

### bdr_read_all_conflicts

- *Column-level conflict detection*
- 競合タイプのリスト
- 欠落列の競合の解決
- 競合リゾルバーのリスト
- デフォルトの競合リゾルバー
- *List of conflict resolutions*
- *Conflict logging*

### User visible catalogs and views

カタログとビューは、アルファベット順にここにリストされています。

### bdr.camo_decion_journal

`bdr.logical_transaction_status` が呼び出された場合、フェイルオーバー後に CAMO パートナーノードによって解決された決定の永続ジャーナル。`bdr.commit_decisions` とは異なり、これは通常の動作条件つまり、CAMO ペアの両方のノードが実行され接続されている下で処理されたトランザクションをカバーしません。このジャーナルのエントリは、自動的にクリーンアップされることはありません。これは、システムが依存しない診断ツールです。

#### bdr.camo_decision_journal 列

| Name | Type | Description |
|---|---|---|
| origin_node_id | oid | トランザクションが実行されたノードの OID |
| origin_xid | oid | リモートオリジンノードのトランザクション ID |
| decision | char | コミットの場合は'c'、アボートの場合は'a' |
| decision_ts | timestamptz | ディシジョンタイム |

### bdr.commit_Decisions

Commit At Most Once CAMO パートナーとして構成されたノードで内部的に使用されるビュー。過去 15 分間に CAMO パートナーがトランザクションに対して行った決定を示します。

#### bdr.commit_decisions 列

| Name | Type | Description |
|---|---|---|
| origin_node_id | oid | トランザクションが実行されたノードの OID |
| origin_xid | xid | リモートオリジンノードのトランザクション ID |
| decision | char | 決定が下されました。`c`コミット、または`a`アボートです。 |

次のページに続く

表 2 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| local_xid | xid | ローカルノードのトランザクション ID |
| commit_ts | timestamptz | トランザクションのコミットタイムスタンプ |
| decision_ts | timestamptz | ディシジョンタイム |

### bdr.commit_scopes

bdr.commit_scope でグループコミットを有効にするために使用できるすべてのコミットスコープを保存するカタログ。

#### bdr.commit_scopes 列

| Name | Type | Description |
|---|---|---|
| commit_scope_id | oid | 参照するスコープの ID |
| commit_scope_name | name | 参照するスコープの名前 |
| commit_scope_origin_node_gr | oid | ID で参照される、ルールが適用されるノードグループ |
| sync_scope_rule | text | スコープの定義 |

### bdr.conflict_history

このテーブルは、競合がログに記録されるデフォルトのテーブルです。テーブルは列 local_time で RANGE パーティション化されており、Autopartition によって管理されます。デフォルトのデータ保持期間は 30 日です。

このテーブルへのアクセスは、テーブルの所有者が可能であり、行レベルのセキュリティによって制限されている、自分が所有するテーブルのすべての競合を表示できます。

詳細については、 *bdr_read_all_conflicts* を参照してください。

#### bdr.conflict_history 列

| Name | Type | Description |
|---|---|---|
| sub_id | oid | この競合を生成したサブスクリプション。`bdr.subscription`テーブルに結合できます |
| origin_node_id | oid | 競合する変更を生成したノードの OID *Replicated large objects* カタログで見られる |
| local_xid | xid | 競合時のレプリケーションプロセスのローカルトランザクション |
| local_lsn | pg_lsn | 競合時のローカル LSN |
| local_time | timestamp with time zone | 紛争の現地時間 |

次のページに続く

表 4 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| remote_xid | xid | リモートノードオリジンで競合する変更を生成したトランザクション |
| remote_change_nr | oid | トランザクション内の変更のインデックス |
| remote_commit_lsn | pg_lsn | リモートノードオリジンで競合する変更を生成したトランザクションのコミット LSN |
| remote_commit_time | timestamp with time zone | リモートノードオリジンで競合する変更を生成したトランザクションのコミットタイムスタンプ |
| conflict_type | text | 検出された競合のタイプ |
| conflict_resolution | text | 選択された競合解決 |
| conflict_index | regclass | 競合するインデックスインデックスが削除されていない場合にのみ有効です |
| reloid | oid | 競合するリレーションインデックスがドロップされていない場合にのみ有効 |
| nspname | text | 競合時に競合が発生したリレーションのスキーマの名前 名前の変更に従いません |
| relname | text | 競合時に競合が発生したリレーションの名前名前の変更に従いません。 |
| key_tuple | json | 行のマッチングに使用されるキーの Json 表現 |
| remote_tuple | json | 競合する着信行の Json 表現 |
| local_tuple | json | ローカル競合行の Json 表現 |
| apply_tuple | json | 結果の適用された行の Json 表現 |
| local_tuple_xmin | xid | ローカル競合行を生成したトランザクション `local_tuple`が設定され、行が固定されていない場合 |
| local_tuple_node_id | oid | ローカル競合行を生成したノード `local_tuple`が設定され、行が固定されていない場合 |
| local_tuple_commit_time | timestamp with time zone | ローカル競合行の最後の既知の変更タイムスタンプ `local_tuple`が設定され、行が固定されていない場合。このコミットタイムスタンプは、このタプルを生成したノードに属します。 |

### bdr.conflict_history_summary

行の競合に関するユーザーが読み取り可能な詳細を含むビュー。

#### bdr.conflict_history_summary 列

| Name | Type | Description |
|---|---|---|
| nspname | text | スキーマの名前 |
| relname | text | テーブルの名前 |
| origin_node_id | oid | 競合する変更を生成したノードの OID *Replicated large objects* カタログで見られる |

次のページに続く

表 5 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| remote_commit_lsn | pg_lsn | リモートノードオリジンで競合する変更を生成したトランザクションのコミット LSN |
| remote_change_nr | oid | トランザクション内の変更のインデックス |
| local_time | timestamp with time zone | 紛争の現地時間 |
| local_tuple_commit_time | timestamp with time zone | ローカルコミットの時間 |
| remote_commit_time | timestamp with time zone | リモートコミットの時間 |
| conflict_type | text | 競合の種類 |
| conflict_resolution | text | 採択された決議 |

### bdr.consensus_kv_data

`bdr.consensus_kv_store()` および `bdr.consensus_kv_fetch()` インターフェイスによって使用される内部 Raft ベースの KV ストアの永続ストレージ。

#### `bdr.consensus_kv_data` 列

| Name | Type | Description |
|---|---|---|
| kv_key | text | 一意のキー |
| kv_val | json | json 形式の任意の値 |
| kv_create_ts | timestamptz | 最終書き込みタイムスタンプ |
| kv_ttl | int | ミリ秒単位の値の生存時間 |
| kv_expire_ts | timestamptz | 有効期限タイムスタンプ *kv_create_ts* + *kv_ttl* |

### bdr.crdt_handlers

この表は、すべての CRDT データ型のマージハンドラーファンクションをリストしています。

#### `bdr.crdt_handlers` 列

| Name | Type | Description |
|---|---|---|
| crdt_type_id | regtype | CRDT データ型 ID |
| crdt_merge_id | regproc | このデータ型のマージファンクション |

### bdr.ddl_replication

このビューには、現在の *DDL filters* によってセットアップされた DDL レプリケーション構成がリストされます。

#### bdr.ddl_replication 列

| Name | Type | Description |
|---|---|---|
| set_ddl_name | name | DDL フィルタの名前 |
| set_ddl_tag | text | 適用するコマンドタグ正規表現 |
| set_ddl_role | text | 適用されるロール正規表現 |
| set_name | name | このフィルタが定義されているレプリケーションセットの名前 |

### bdr.depend

このテーブルは、PGD カタログ内の内部オブジェクトの依存関係を追跡します。

### bdr.failover_replication_slots

このテーブルは、フェールオーバーサポートで使用されている論理レプリケーションスロットのステータスを追跡します。フェイルオーバーレプリケーションスロットの詳細については、*CDC Failover support* を参照してください。

#### bdr.failover_replication_slots 列

| Name | Type | Description |
|---|---|---|
| slot_name | name | レプリケーションスロットの名前 |
| slot_id | oid | レプリケーションスロットの ID |
| node_group_id | oid | ノードグループの ID |
| plugin | name | プラグインの名前 |
| twophase | boolean | スロットは 2 フェーズコミットに使用されますか |
| active_node | oid | アクティブノードの ID |
| active_pid | int | 現在スロットをデコードしているプロセスの PID |
| prev_node | oid | 以前のノードの ID |

### bdr.global_consensus_journal

このカタログテーブルは、グローバルコンセンサスを管理中に送信されたすべての Raft メッセージのログを記録します。

`bdr.global_consensus_response_journal` カタログと同様、ペイロードはバイナリエンコードされた形式で保存され、`bdr.decode_message_payload()` ファンクションでデコードできます。詳細については、*bdr.global_consensus_journal_details* ビューを参照してください。

#### bdr.global_consensus_journal 列

| Name | Type | Description |
|---|---|---|
| log_index | int8 | ジャーナルエントリの ID |
| term | int8 | Raft 用語 |
| origin | oid | 要求が発生したノードの ID |
| req_id | int8 | リクエストの ID |
| req_payload | bytea | 要求のペイロード |
| trace_context | bytea | 要求のトレースコンテキスト |

### bdr.global_consensus_journal_details

このビューは、 `bdr.decode_message_payload()` ファンクションを使用してペイロードをデコードして、送信された Raft メッセージと対応する応答を表示します。

#### `bdr.global_consensus_journal_details` 列

| Name | Type | Description |
|---|---|---|
| node_group_name | name | ノードグループの名前 |
| log_index | int8 | ジャーナルエントリの ID |
| term | int8 | Raft 用語 |
| request_id | int8 | リクエストの ID |
| origin_id | oid | 要求が発生したノードの ID |
| req_payload | bytea | 要求のペイロード |
| origin_node_name | name | 要求が発生したノードの名前 |
| message_type_no | oid | 要求の PGD メッセージタイプの ID |
| message_type | text | 要求の PGD メッセージタイプの名前 |
| message_payload | text | 要求の PGD メッセージペイロード |
| response_message_type_no | oid | 応答の PGD メッセージタイプの ID |
| response_message_type | text | 応答の PGD メッセージタイプの名前 |
| response_payload | text | 応答の PGD メッセージペイロード |
| response_errcode_no | text | 応答の SQLSTATE |
| response_errcode | text | 応答のエラーコード |
| response_message | text | 応答のエラーメッセージ |

### bdr.global_consensus_response_journal

このカタログテーブルは、グローバルコンセンサスを管理中に受信した Raft メッセージへのすべての応答を収集します。

`bdr.global_consensus_journal` カタログと同様、ペイロードはバイナリエンコードされた形式で保存され、`bdr.decode_message_payload()` ファンクションでデコードできます。詳細については、*bdr.global_consensus_journal_details* ビューを参照してください。

#### bdr.global_consensus_response_journal 列

| Name | Type | Description |
|---|---|---|
| log_index | int8 | ジャーナルエントリの ID |
| res_status | oid | 応答のステータスコード |
| res_payload | bytea | 応答のペイロード |
| trace_context | bytea | 応答のトレースコンテキスト |

### bdr.global_lock

このカタログテーブルには、サーバー再起動時にグローバルロック状態を回復するために必要な情報が保存されます。

bdr.global_lock の表示行は必ずしもすべてのグローバルロックアクティビティを反映しているわけではないため、使用状況をモニタリングする場合は、*bdr.global_locks* ビューが適しています。

このテーブルの内容は変更しないでください。重要な PGD カタログです。

#### bdr.global_lock 列

| Name | Type | Description |
|---|---|---|
| ddl_epoch | int8 | ロックの DDL エポック |
| origin_node_id | oid | グローバルロックが発生したノードの OID |
| lock_type | oid | ロックのタイプ DDL または DML |
| nspname | name | ロックされたリレーションのスキーマ名 |
| relname | name | ロックされたリレーションのリレーション名 |
| groupid | oid | トップレベルグループの OID アドバイザリーロックの場合 |
| key1 | integer | 最初の 32 ビットキーまたは 64 ビットキーの下位 32 ビットアドバイザリーロック用 |
| key2 | integer | 64 ビットキーの 2 番目の 32 ビットキー以上の 32 ビットアドバイザリロック用 |
| key_is_bigint | boolean | 64 ビット整数キーを使用する場合に true アドバイザリロックの場合 |

### bdr.global_locks

このノードのアクティブなグローバルロックを含むビュー。*bdr.global_locks* ビューは、PGD の共有メモリロック状態追跡を公開し、管理者に PGD のグローバルロックアクティビティと進行状況についてのより深い洞察を提供します。

グローバルロックの監視 を参照

グローバルロックの詳細については、

### bdr.global_locks 列

| Name | Type | Description |
|---|---|---|
| origin_node_id | oid | グローバルロックが発生したノードの OID |
| origin_node_name | name | グローバルロックが発生したノードの名前 |
| lock_type | text | ロックのタイプ DDL または DML |
| relation | text | ロックされたリレーション名 DML ロックの場合またはキーアドバイザリロックの場合 |
| pid | int4 | ロックを保持しているプロセスの PID |
| acquire_stage | text | ロック取得プロセスの内部状態 |
| waiters | int4 | 同じグローバルロックを待機するバックエンドのリスト |
| global_lock_request_time | timestamptz | このグローバルロックの取得がオリジンノードによって開始された時間 |
| local_lock_request_time | timestamptz | ローカルノードがローカルロックの取得を試行し始めた時間 |
| last_state_change_time | timestamptz | `acquire_stage`が最後に変更された時間 |

カラムの詳細

- `relation` DML ロックの場合、`relation` は、DML ロックが取得されたリレーションを示します。グローバルアドバイザリロックの場合、`relation` 列は、実際に、ロックが取得された 2 つの 32 ビット整数または 1 つの 64 ビット整数を表示します。
- `origin_node_id` および `origin_node_name` これらがローカルノードの ID および名前と同じである場合、ローカルノードはグローバル DDL ロックのイニシエーター、つまり、取得トランザクションを実行しているノードです。これらのフィールドが別のノードを指定している場合、ローカルノードは代わりにローカル DDL ロックを取得して、リモートノードからのグローバル DDL ロック要求を満たそうとします。
- `pid` グローバル DDL ロックを要求したプロセスのプロセス ID。ローカルノードが要求ノードの場合。他のノードでは NULL。オリジンノードを照会して、ロッカー pid を特定します。
- `global_lock_request_time` グローバルロック要求イニシエーターがグローバルロックを取得するプロセスを開始したタイムスタンプ。現在のノードで不明な場合は null にすることができます。この時間は DDL ロック要求の先頭にスタンプされ、DDL エポック管理および保留中レプリケーションキューの必要なフラッシュに要した時間が含まれます。現在、オリジンノードでのみ知られています。
- `local_lock_request_time` ローカルノードがこのグローバルロックのローカルロックを取得しようと開始したタイムスタンプ。これには、重量セッションロックの取得に要した時間が含まれますが、DDL エポックまたはキューのフラッシュにかかる時間は含まれません。ローカルノードの再起動後にロックが再取得された場合、それがノードの再起動時間になります。
- `last_state_change_time` このグローバルロックエントリーの `bdr.global_locks.acquire_stage` フィールドが最後に変更されたタイムスタンプ。

### bdr.group_camo_details

bdr.run_on_all_nodes を使用して、すべてのノードから CAMO 関連の情報を収集します。

#### bdr.group_camo_details 列

| Name | Type | Description |
|---|---|---|
| node_id | text | 内部ノード ID |
| node_name | text | ノードの名前 |
| camo_partner | text | カモフラージュパートナーのノード名 |
| is_camo_partner_connecte | text | 接続状態 |
| is_camo_partner_ready | text | 準備状況 |
| camo_transactions_resolve | text | 保留中および未解決の CAMO トランザクションはありますか |
| apply_lsn | text | 再生されたとして報告された最新の位置可視 |
| receive_lsn | text | 受信した変更またはメッセージの最新の LSN 再起動の場合に遡ることができます |
| apply_queue_size | text | apply_lsn と receive_lsn のバイト差 |

### bdr.group_lease

ノードグループのリースに関する情報を保存するカタログテーブル。特定の種類のリースを保持できるノードは、一度に 1 つのノードのみです。現在、リース機能は、kind = 'a' を指定して Postgres アナリティクスアクセラレーター PGAA を使用してアナリティクスに複製する場合にのみ使用されます。

#### bdr.group_lease 列

| Name | Type | Description |
|---|---|---|
| node_group_id | oid | リースに関連付けられたノードグループの識別子。 |
| node_id | oid | 現在リースを保持しているノードの識別子。 |
| received_ts | timestamp with time zone | 最後のリース更新要求を受信したローカルノード時間。 |
| request_ts | timestamp with time zone | 更新要求が行われたときのリース所有者の時計からのタイムスタンプ。 |
| duration | integer | 秒単位のリース期間。 |
| kind | char | 特定のリースのタイプたとえば、アナリティクスの「a」。 |

### bdr.group_raft_details

bdr.run_on_all_nodes を使用して、すべてのノードから Raft コンセンサスステータスを収集します。

#### bdr.group_raft_details 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | 内部ノード ID |
| node_name | name | ノードの名前 |
| node_group_name | name | グループの名前は次の一部です |
| state | text | ノード上の Raft ワーカーの状態 |
| leader_id | oid | RAFT_LEADER のノード ID |
| current_term | int | Raft 選挙内部 ID |
| commit_index | int | Raft スナップショットの内部 ID |
| nodes | int | アクセス可能なノードの数 |
| voting_nodes | int | 投票ノードの数 |
| protocol_version | int | このノードのプロトコルバージョン |

### bdr.group_replslots_details

bdr.run_on_all_nodes を使用して、すべてのノードから PGD スロット情報を収集します。

#### bdr.group_replslots_details 列

| Name | Type | Description |
|---|---|---|
| node_group_name | text | PGD グループの名前 |
| origin_name | text | オリジンノードの名前 |
| target_name | text | ターゲットノードの名前 |
| slot_name | text | このサブスクリプションで使用されるオリジンノードのスロット名 |
| active | text | スロットはアクティブですか?接続は接続されていますか? |
| state | text | レプリケーションの状態キャッチアップ、ストリーミングなど、またはオフラインの場合は「切断」 |
| write_lag | interval | 報告された書き込みのおおよその遅延時間 |
| flush_lag | interval | 報告されたフラッシュのおおよその遅延時間 |
| replay_lag | interval | 報告されたリプレイのおおよそのラグタイム |
| sent_lag_bytes | int8 | send_lsn と現在の WAL 書き込み位置のバイト差 |
| write_lag_bytes | int8 | write_lsn と現在の WAL 書き込み位置のバイト差 |
| flush_lag_bytes | int8 | lush_lsn と現在の WAL 書き込み位置のバイト差 |
| replay_lag_byte | int8 | recovery_lsn と現在の WAL 書き込み位置のバイト差 |

### bdr.group_subscription_summary

bdr.run_on_all_nodes を使用して、すべてのノードからサブスクリプションステータスを収集します。

#### bdr.group_subscription_summary 列

| Name | Type | Description |
|---|---|---|
| origin_node_name | text | サブスクリプションの元の名前 |
| target_node_name | text | サブスクリプションのターゲットの名前 |
| last_xact_replay_timestamp | text | 最後に再生されたトランザクションのタイムスタンプ |
| sub_lag_seconds | text | now と last_xact_replay_timestamp 間のラグ |

### bdr.group_versions_details

bdr.run_on_all_nodes を使用して、すべてのノードから PGD 情報を収集します。

#### bdr.group_versions_details 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | 内部ノード ID |
| node_name | name | ノードの名前 |
| postgres_version | text | ノード上の PostgreSQL バージョン |
| bdr_version | text | ノード上の PGD バージョン |

### bdr.leader

クラスター内のサブグループ全体でリーダーノードを追跡します。クラスター内のすべての書き込みリーダーとサブスクライバー専用グループリーダーのステータスを表示します最適化されたトポロジが有効になっている場合。

#### bdr.leader 列

| Name | Type | Description |
|---|---|---|
| node_group_id | oid | ノードグループの ID。 |
| leader_node_id | oid | リーダーノードの ID。 |
| generation | int | リーダーノードの生成。 Leader_kind は、セマンティクスを設定します。 |
| leader_kind | "char" | リーダーノードの種類。 |

Leader_kind 値は次のとおりです。

| Value | Description |
|---|---|
| W | プロキシルーティングごとに、リーダーを書き込みます。この場合、リーダーはサブグループの Raft インスタンスによって維持されます。 `generation`は、各 `bdr.node_group_routing_info`レコードの `write_leader_version`に対応します。 |
| S | 購読者専用グループのリーダー。 SO グループのこの指定されたメンバーは、上流データノードをサブスクライブし、残りの SO グループメンバーに上流の変更を公開する任務を負います。リーダーは、トップレベルの Raft インスタンスによって維持されます。 `generation`は、リーダーが変更されると順次更新されます。 |

### bdr.local_consensus_snapshot

このカタログテーブルには、ローカルノードによって作成または受信されたコンセンサススナップショットが含まれています。

#### bdr.local_consensus_snapshot 列

| Name | Type | Description |
|---|---|---|
| log_index | int8 | ジャーナルエントリの ID |
| log_term | int8 | Raft 用語 |
| snapshot | bytea | Raft スナップショットデータ |

### bdr.local_consensus_state

このカタログテーブルには、ローカルノード上の Raft の現在の状態が保存されます。

#### bdr.local_consensus_state 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | ノードの ID |
| current_term | int8 | Raft 用語 |
| apply_index | int8 | Raft はインデックスを適用します |
| voted_for | oid | 今期にこのノードによって投じられた投票 |
| last_known_leader | oid | 最後に知られている Raft リーダーの node_id |

### bdr.local_node

このテーブルは、現在の Postgres インスタンスの現在のデータベース内のローカルノードを識別します。

#### bdr.local_node 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | ノードの ID |
| pub_repsets | text[] | 公開されたレプリケーションセット |
| sub_repsets | text[] | サブスクライブされたレプリケーションセット |
| node_uuid | uuid | ノードの UUID |

### bdr.local_node_summary

*bdr.node_summary* および pub_repsets および sub_repsets と同じ情報を含むビューですが、ローカルノードのみを対象とします。

### bdr.local_sync_status

サブスクリプションまたはテーブル同期プロセスのステータスに関する情報。

#### bdr.local_sync_status 列

| Name | Type | Description |
|---|---|---|
| sync_kind | char | 完了した同期の種類 |
| sync_subid | oid | 同期を実行するサブスクリプションの ID |
| sync_nspname | name | 同期テーブルのスキーマ名存在する場合 |
| sync_relname | name | 同期テーブルの名前存在する場合 |
| sync_status | char | 同期の現在の状態 |
| sync_remote_relid | oid | 上流に同期されたテーブルがある場合の ID |
| sync_end_lsn | pg_lsn | 同期状態が最後に変更された位置 |

### bdr.node

このテーブルは、クラスター内のすべての PGD ノードをリストしています。

ビュー *bdr.node_summary* は、bdr.node のほとんどの列の人が判読可能なバージョンを提供します。

#### bdr.node 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | ノードの ID |
| node_name | name | ノードの名前 |
| node_group_id | oid | ノードグループの ID |
| source_node_id | oid | ソースノードの ID |
| synchronize_structure | "char" | 結合中に実行されるスキーマの同期 |
| node_state | oid | ノードの一貫した状態 |

次のページに続く

表 27 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| target_state | oid | ノードが到達しようとしている状態参加または昇格中に |
| seq_id | int4 | 一意のシーケンス番号を生成するために使用されるノードのシーケンス識別子 |
| dbname | name | ノードのデータベース名 |
| node_dsn | char | ノードの接続文字列 |
| proto_version_ranges | int[] | ノードによってサポートされているプロトコルバージョンレンジ |
| generation | smallint | ノードが以前のノードと同じ名前で参加したときにインクリメントされるカウンター |
| node_kind | oid | ノード種類の ID |
| node_join_finished | boolean | 結合が完了したかどうかを確認する |
| node_uuid | uuid | ノードの UUID 一意 |

### bdr.node_catchup_info

このカタログテーブルは、各ノードに関連するキャッチアップ情報を記録します。結合または部分プロシージャに関連している場合。

#### bdr.node_catchup_info 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | ノードの ID |
| node_source_id | oid | データのソースとして使用されるノードの ID |
| slot_name | name | このソースに使用されるスロット |
| min_node_lsn | pg_lsn | ノードがピアノードからの直接再生に切り替えることができる最小 LSN |
| catchup_state | oid | キャッチアップ状態のステータスコード |
| origin_node_id | oid | トランザクションが必要なノードの ID |

ノード node_id は、分割ノード origin_node_id からの欠損データが必要な場合、転送によって既にそれ node_source_id を持つノードからそれを取得できます。このテーブルのレコードは、ノード node_id が EDB Postgres 分散クラスターのメンバーになるまで持続します。

### bdr.node_catchup_info_details

情報をより分かりやすく表示する bdr.node_catchup_info カタログのビュー

### bdr.node_conflict_resolvers

すべての既知の競合タイプの現在構成されている競合解決。

#### bdr.node_conflict_resolvers 列

| Name | Type | Description |
|---|---|---|
| conflict_type | text | 競合の種類 |
| conflict_resolver | text | この競合タイプに使用されるリゾルバー |

### bdr.node_group

このカタログテーブルには、すべての PGD ノードグループがリストされています。ユーザーが読み取り可能な詳細を含むビューについては、*bdr.node_group_summary* も参照してください。

#### bdr.node_group 列

| Name | Type | Description |
|---|---|---|
| node_group_id | oid | ノードグループの ID。 |
| node_group_name | name | ノードグループの名前。 |
| node_group_default_repset | oid | このノードグループのデフォルトレプリケーションセット。 |
| node_group_default_repset | oid | このノードグループのデフォルトレプリケーションセット。 |
| node_group_parent_id | oid | 親グループの ID これがルートグループの場合は 0。 |
| node_group_flags | int | グループフラグ。 |
| node_group_uuid | uuid | グループの uuid。 |
| node_group_apply_delay | interval | プロバイダーからの変更を適用するまでに加入者が待機する時間。 |
| node_group_check_constra | bool | 適用プロセスがデータを適用するときにコンストレインをチェックするかどうか。 |
| node_group_num_writers | int | このノードグループをバッキングするサブスクリプションに使用するライターの数。 |
| node_group_enable_wal_d | bool | グループに enable_wal_decoder が設定されているかどうか。 |
| node_group_streaming_mo | char | トランザクションストリーミング設定 'O'-オフ、'F'-ファイル、 'W'-ライタ、 'A'-自動、 'D'-デフォルト。 |
| node_group_default_comm | oid | ノードグループのデフォルトのコミットスコープの ID。 |
| node_group_location | char | ノードグループに関連付けられた場所の名前。 |
| node_group_enable_routin | char | ノードグループが接続マネージャーを介したルーティングを許可するかどうか。 |
| node_group_enable_raft | bool | ノードグループが Raft コンセンサスを許可するかどうか。 |

### bdr.node_group_replication_sets

PGD グループ用に作成されるデフォルトのレプリケーションセットを示すビュー。bdr.replication_sets も参照してください。

#### bdr.node_group_replication_sets 列

| Name | Type | Description |
|---|---|---|
| node_group_name | name | PGD グループの名前 |
| def_repset | name | デフォルトの repset の名前 |
| def_repset_ops | text[] | デフォルトの repset によってレプリケートされたアクション |
| def_repset_ext | name | デフォルトの「外部」repset の名前通常は def_repset と同じです |
| def_repset_ext_ops | text[] | デフォルトの「外部」rep セットによってレプリケートされたアクション通常は def_repset_ops と同じです |

### bdr.node_group_summary

ノードグループに関するユーザーが読み取り可能な詳細を含むビュー。*bdr.node_group_config* も参照してください。

#### bdr.node_group_summary 列

| Name | Type | Description |
|---|---|---|
| node_group_name | name | ノードグループの名前 |
| default_repset | name | このノードグループのデフォルトレプリケーションセット |
| parent_group_name | name | 親グループの名前これがルートグループの場合は NULL |
| node_group_type | text | ノードグループのタイプ"global"、"data"、"shard"、または"subscriber-only"のいずれか |
| apply_delay | interval | プロバイダーからの変更を適用するまでに加入者が待機する時間 |
| check_constraints | boolean | 適用プロセスがデータの適用時にコンストレインをチェックするかどうか |
| num_writers | integer | このノードグループをバッキングするサブスクリプションに使用するライターの数 |
| enable_wal_decoder | boolean | グループに enable_wal_decoder セットが設定されているかどうか |
| streaming_mode | text | トランザクションストリーミング設定 "off"、"file"、"ライタ"、"auto"または"default" |
| default_commit_scope | name | ノードグループのデフォルトのコミットスコープの名前 |

次のページに続く

表 32 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| location | name | ノードグループに関連付けられた場所の名前 |
| server_pool_mode | text | このノードグループの *Connection pooling* モード *none*、*session*、または`transaction` |
| enable_routing | boolean | ノードグループが接続マネージャーを介したルーティングを許可するかどうか |
| enable_raft | boolean | ノードグループが Raft コンセンサスを許可するかどうか |
| route_writer_max_lag | bigint | 受け入れられる最大書き込み遅延 |
| route_reader_max_lag | bigint | 受け入れられる最大読み取り遅延 |
| route_writer_wait_flush | boolean | フラッシュを待機する必要がある場合にスイッチ |

### bdr.node_local_info

ローカルノードに固有のノードごとの構成を保存するために使用されるカタログテーブルノードごとの構成のグローバルビューではなく。

#### bdr.node_local_info 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | ノードの OID ローカルノードを含む |
| applied_state | oid | ノード状態の内部 ID |
| ddl_epoch | int8 | ノードによって処理された最後のエポック番号 |
| slot_name | name | そのノードへの接続に使用されるスロットの名前ローカルノードの場合は NULL |
| origin_name | name | そのノードの レプリケーション起点のクリーンアップ の名前。ローカルノード、またはサブスクライバ専用ノードやスタンバイなどのデータノードではないノードの場合、NULL になります。 |

### bdr.node_log_config

競合ログ構成に関する情報を保存するカタログビュー。

#### bdr.node_log_config 列

| Name | Description |
|---|---|
| log_name | ロギング構成の名前 |
| log_to_file | サーバーログファイルにログを記録するかどうか |
| log_to_table | テーブルにログを記録するかどうか、およびどのテーブルがターゲットであるか |
| log_conflict_type | NULL がすべてを意味する場合、ログに記録する競合タイプ |

次のページに続く

表 34 – 前のページからの続き

| Name | Description |
|---|---|
| log_conflict_res | NULL がすべてを意味する場合、どの競合解決をログに記録しますか |

### bdr.node_peer_progress

レプリケーションストリーム内のすべてのノードの進行状況を追跡するために使用されるカタログ。クラスター内のすべてのノードは、その進行状況を bdr.replay_progress_frequency ミリ秒ごとに他のすべてのノードに定期的にブロードキャストしますデフォルトは 60000 ミリ秒、つまり 1 分です。このリレーションには N * (N-1) 行が必要です。

モニタリングには、 bdr.node_slots ビューの方が興味深いかもしれません。 *Monitoring the Connection Manager* も参照してください。

#### bdr.node_peer_progress 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | この位置情報を報告した元ノードの OID |
| peer_node_id | oid | この位置情報が報告されたノードのピアリモートノードの OID |
| last_update_sent_time | timestamptz | レポートが元ノードによって送信された時刻 |
| last_update_recv_time | timestamptz | ローカルサーバーがレポートを受信した時刻 |
| last_update_node_lsn | pg_lsn | レポート時の元のノードの LSN |
| peer_position | pg_lsn | 元のノードから見たノードのピアの最新の LSN |
| peer_replay_time | timestamptz | 報告ノードによって認識されたピアの最新の再生時間 |
| last_update_horizon_xid | oid | 内部解像度ホライズン すべての下位 xid は報告ノードで解決されることが知られています |
| last_update_horizon_lsn | pg_lsn | 内部解像度ホライズン レポートノードの LSN に関して同じ |

### bdr.node_replication_rates

このビューには、特定のノードからの送信レプリケーションアクティビティに関する情報が含まれています。

#### bdr.node_replication_rates 列

| Column | Type | Description |
|---|---|---|
| peer_node_id | oid | この情報が報告されたノードのピアリモートノードの OID |
| target_name | name | ターゲットピアノードの名前 |
| sent_lsn | pg_lsn | 最新の送信位置 |
| replay_lsn | pg_lsn | 再生されたとして報告された最新の位置可視 |
| replay_lag | interval | 報告されたリプレイのおおよそのラグタイム |

次のページに続く

表 36 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| replay_lag_bytes | int8 | attempts_lsn とオリジンの現在の WAL 書き込み位置のバイト差 |
| replay_lag_size | text | repair_lsn と現在の WAL 書き込み位置間の人間が判読可能なバイトの差 |
| apply_rate | bigint | ピアノードで 1 秒ごとに適用される LSN |
| catchup_interval | interval | ピアノードがまだ適用されていないすべての変更に追いつくまでに必要なおおよその時間 |

**注釈**

*replay_lag* は、再接続後すぐにゼロに設定されます。回避策として、*replay_lag_bytes* 、*replay_lag_size* 、または `catchup_interval` を使用します。

### bdr.node_slots

このビューには、PGD が現在のデータベースで使用するレプリケーションスロットに関する情報が含まれています。

送信レプリケーションの監視 を参照

このビューのフィールドの使用と解釈に関するガイダンスについては。

#### bdr.node_slots 列

| Name | Type | Description |
|---|---|---|
| target_dbname | name | ターゲットノード上のデータベース名 |
| node_group_name | name | PGD グループの名前 |
| node_group_id | oid | PGD グループの OID |
| origin_name | name | オリジンノードの名前 |
| target_name | name | ターゲットノードの名前 |
| origin_id | oid | 元ノードの OID |
| target_id | oid | ターゲットノードの OID |
| local_slot_name | name | PGD に従ったレプリケーションスロットの名前 |
| slot_name | name | Postgres によるスロットの名前上記と同じ |
| is_group_slot | boolean | スロットがこのノードのノードグループのクラッシュリカバリスロットである場合は true ["Group Replication Slot"](nodes#Group Replication Slot) を参照してください。 |
| is_decoder_slot | boolean | このスロットはデコードワーカー機能によって使用されますか |
| plugin | name | このスロットを使用する論理デコードプラグイン pg-logical_output または bdr である必要があります |

次のページに続く

表 37 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| slot_type | text | スロットのタイプ論理的である必要があります。 |
| datoid | oid | 現在のデータベースの OID |
| database | name | 現在のデータベースの名前 |
| temporary | bool | スロットは一時的なものですか |
| active | bool | スロットはアクティブですか?接続は接続されていますか? |
| active_pid | int4 | スロットに接続されたプロセスの PID |
| xmin | xid | スロットに必要な XID |
| catalog_xmin | xid | スロットに必要なカタログ XID |
| restart_lsn | pg_lsn | スロットがデコードをリスタートできる LSN |
| confirmed_flush_lsn | pg_lsn | 最新に確認された複製位置 |
| usesysid | oid | レプリケーションセッションが実行されているユーザーの sysid |
| usename | name | レプリケーションセッションが実行されているユーザーのユーザー名 |
| application_name | text | クライアント接続のアプリケーション名 `synchronous_standby_names`によって使用されます |
| client_addr | inet | クライアント接続の IP アドレス |
| client_hostname | text | クライアント接続のホスト名 |
| client_port | int4 | クライアント接続のポート |
| backend_start | timestamptz | 接続が開始されたとき |
| state | text | レプリケーションの状態キャッチアップ、ストリーミングなど、またはオフラインの場合は「切断」 |
| sent_lsn | pg_lsn | 最新の送信位置 |
| write_lsn | pg_lsn | 最新の立場を書面どおりに報告 |
| flush_lsn | pg_lsn | 最新の位置はディスクにフラッシュされたとして報告されました |
| replay_lsn | pg_lsn | 再生されたとして報告された最新の位置可視 |
| write_lag | interval | 報告された書き込みのおおよその遅延時間 |
| flush_lag | interval | 報告されたフラッシュのおおよその遅延時間 |
| replay_lag | interval | 報告されたリプレイのおおよそのラグタイム |
| sent_lag_bytes | int8 | send_lsn と現在の WAL 書き込み位置のバイト差 |
| write_lag_bytes | int8 | write_lsn と現在の WAL 書き込み位置のバイト差 |
| flush_lag_bytes | int8 | lush_lsn と現在の WAL 書き込み位置のバイト差 |
| replay_lag_bytes | int8 | recovery_lsn と現在の WAL 書き込み位置のバイト差 |
| sent_lag_size | text | send_lsn と現在の WAL 書き込み位置の間の人間が判読可能なバイトの差 |
| write_lag_size | text | write_lsn と現在の WAL 書き込み位置間の人間が判読可能なバイトの差 |
| flush_lag_size | text | flash_lsn と現在の WAL 書き込み位置の間の人間が判読可能なバイトの差 |
| replay_lag_size | text | repair_lsn と現在の WAL 書き込み位置間の人間が判読可能なバイトの差 |

> **注釈**
>
> *replay_lag* は、再接続後すぐにゼロに設定されます。回避策として、*replay_lag_bytes* または `replay_lag_size` を使用します。

### bdr.node_summary

このビューには、ローカルノードに知られているすべての PGD ノードに関する概要情報が含まれています。

#### bdr.node_summary 列

| Name | Type | Description |
|---|---|---|
| node_name | name | ノードの名前 |
| node_group_name | name | ノードが一部である PGD グループの名前 |
| interface_connstr | text | ノードへの接続文字列 |
| peer_state_name | text | 人間が読める形式のノードの一貫した状態 |
| peer_target_state_name | text | ノードが到達しようとしている状態参加または昇格中に |
| node_seq_id | int4 | 一意のシーケンス番号を生成するために使用されるノードのシーケンス識別子 |
| node_local_dbname | name | ノードのデータベース名 |
| node_id | oid | ノードの OID |
| node_group_id | oid | PGD ノードグループの OID |
| node_kind_name | oid | ノード種類名 |
| node_uuid | uuid | ノードの UUID |

### bdr.parted_origin_catchup_info

このテーブルは、分割されたオージンに関連する各ノードの関連キャッチアップ情報を記録します。

#### bdr.parted_origin_catchup_info 列

| Name | Type | Description |
|---|---|---|
| parting_peer_node_id | oid | 分割されたノードの ID |
| node_id | oid | ノードの ID |
| node_group_id | oid | ノードグループの ID |
| origin_catchup_lsn | pg_lsn | ノードがグループスロットが追いつくのを待機し、状態を DONE に移動する LSN |
| origin_catchup_state | oid | 分割されたオリジンキャッチアップのステータスコード |

ノード `node_id` は、そのグループスロットが記録された LSN `origin_catchup_lsn` に追いつくのを待機します。これは、そのグループスロットが PARTED ノード `parting_peer_node_id` から発生するすべてのトランザクションに追いつくようにするためです。

このテーブルのレコードは、パーティングノード parting_peer_node_id が自動的に削除されるまで持続します。

### bdr.parted_origin_catchup_info_details

このテーブルは、bdr.parted_origin_catchup_info のわかりやすいビューであり、分割されたオージンに関連する各ノードの関連キャッチアップ情報をこの場合はテキスト形式で示しています。

#### bdr.parted_origin_catchup_info_details 列

| Name | Type | Description |
|---|---|---|
| target_node_id | oid | ターゲットノードの ID |
| target_node_name | text | ターゲットノードの名前 |
| parting_node_id | oid | 分割されたノードの ID |
| parting_node_name | text | 分割されたノードの名前 |
| node_group_id | oid | ノードグループの ID |
| node_group_name | text | ノードグループの名前 |
| parting_catchup_lsn | pg_lsn | ノードがグループスロットが追いつくのを待機し、状態を DONE に移動する LSN |
| parting_catchup_state | oid | Parted Origin のキャッチアップステータスコード |
| parting_catchup_state_name | text | Parted Origin のキャッチアップステータステキスト |

ノード target_node_id は、そのグループスロットが記録された LSN parting_catchup_lsn に追いつくのを待機します。これは、そのグループスロットが PARTED ノード parting_node_id から発生するすべてのトランザクションに追いつくようにするためです。

このテーブルのレコードは、パーティングノード parting_node_id が自動的に削除されるまで持続します。

### bdr.pending_failed_changes

レプリケーション中に適用できなかったトランザクションからキャプチャされた変更を保存するテーブル。各行は、失敗したトランザクションからの 1 つの DML または DDL 操作を表します。ポリシーが record_all_and_disable または record_all_and_skip に設定されている場合、apply error ポリシーによって作成されます。このテーブルの内容の確認と解決に関するガイダンスについては、*Handling replication apply errors* を参照してください。

#### bdr.pending_failed_changes 列

| Name | Type | Description |
|---|---|---|
| *id* | *bigint* | 主キー。 |

次のページに続く

表 41 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| *failed_txn_id* | *uuid* | 失敗したトランザクションの一意の識別子。レゾリューションファンクションを呼び出すときに、`(subscription_name, remote_xid, remote_commit_lsn)`の便利な代替として使用します。 |
| *sub_id* | *oid* | 失敗したトランザクションを受け取ったサブスクリプションの OID。 |
| *origin_node_id* | *oid* | トランザクションが発生したノードの OID。 |
| *remote_xid* | *xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | *pg_lsn* | 元ノードで失敗したトランザクションのコミット LSN。 |
| *remote_commit_end_lsn* | *pg_lsn* | コミットレコードの終わりの LSN。サブスクリプションの位置を進めるために使用されます。 |
| *remote_commit_time* | *timestamptz* | 元のノードでのコミットタイムスタンプ。 |
| *change_seq* | *integer* | トランザクション内の変更の順序。 |
| *schema_name* | *name* | ターゲットテーブルのスキーマ。 |
| *table_name* | *name* | ターゲットテーブルの名前。 |
| *operation* | *text* | 操作タイプ *INSERT*、*UPDATE*、*DELETE*、*SQL*、*DDL*、または `TRUNCATE` |
| *key_columns* | *jsonb* | 主キー列の値。 |
| *old_values* | *jsonb* | `UPDATE`および `DELETE`操作の変更前イメージ列値。 |
| *new_values* | *jsonb* | `INSERT`および`UPDATE`操作のアフターイメージ列値、または`DDL`および `SQL`操作の SQL テキスト。 |
| *change_status* | *text* | 解決ステータス *pending*、*skipped*、*applied*、または`modified`。 |
| *modified_values* | *jsonb* | `bdr.modify_failed_change`で設定されたユーザー編集の列値。変更が再試行されるときに使用されます。 |
| *error_message* | *text* | 失敗した適用からのエラーメッセージ。失敗したトランザクションの最後の変更に設定されます。 |
| *error_detail* | *text* | 追加のエラーの詳細。 |
| *error_sqlstate* | *text* | エラーの SQLSTATE コード。 |
| *capture_truncated* | *boolean* | トランザクションが `apply_error_max_record_size`を超えたために変更キャプチャが切り捨てられたかどうか。 `true`の場合、トランザクションの一部の変更が欠落する場合があり、 `bdr.apply_failed_transaction`はトランザクションの適用を拒否します。デフォルトは `false`です。 |
| *recorded_at* | *timestamptz* | 変更が記録されたときのタイムスタンプ。 |
| *resolved_at* | *timestamptz* | 変更が解決されたときのタイムスタンプ。 |
| *resolved_by* | *text* | 変更が解決された方法 *user*、*auto_skip*、または`retry`。 |

### bdr.pending_failed_changes_sql

キー列をサブスクリプションおよびオリジンノード名と組み合わせ、bdr.reconstruct_change_sql によって生成された変更ごとに再構成された SQL ステートメントを含む bdr.pending_failed_changes のビュー。ベーステーブルとは異なり、ビューは new_values の代わりに modified_values 列が存在する場合は使用するため、再構成された SQL は bdr.modify_failed_change によって行われた編集を反映します。

#### bdr.pending_failed_changes_sql 列

| Name | Type | Description |
|---|---|---|
| *id* | *bigint* | `bdr.pending_failed_changes`の行の主キー。 |
| *failed_txn_id* | *uuid* | 失敗したトランザクションの UUID。 |
| *subscription_name* | *name* | サブスクリプションの名前。 |
| *origin_node* | *name* | 元ノードの名前。 |
| *remote_xid* | *xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | *pg_lsn* | 元ノードで失敗したトランザクションのコミット LSN。 |
| *change_seq* | *integer* | トランザクション内の変更の順序。 |
| *qualified_table* | *text* | スキーマ修飾テーブル名 `schema_name.table_name` 。 |
| *operation* | *text* | 操作タイプ *INSERT*、*UPDATE*、*DELETE*、*SQL*、*DDL*、または `TRUNCATE` |
| *sql_statement* | *text* | `bdr.reconstruct_change_sql`によって生成された、キャプチャされた変更を表す実行可能 SQL 。 |
| *change_status* | *text* | 解決ステータス *pending*、*skipped*、*applied*、または`modified`。 |
| *error_message* | *text* | 失敗した適用からのエラーメッセージ。失敗したトランザクションの最後の変更に設定されます。 |
| *recorded_at* | *timestamptz* | 変更が記録されたときのタイムスタンプ。 |

### bdr.pgd_largeobject

レプリケートされたラージオブジェクトのデータページを保存します。このテーブルは、

pg_catalog.largeobject の PGD 同等であり、*Replicated large objects* によって使用されます。その内容はすべてのノードに複製されます。

#### bdr.pgd_largeobject 列

| Name | Type | Description |
|---|---|---|
| *loid* | *oid* | このページが属するラージオブジェクトの OID。 |
| *pageno* | *int* | ラージオブジェクト内の 0 から始まるページ番号。 |
| *data* | *bytea* | このページに保存されているデータ。スパースページの場合、長さはゼロになる場合があります。 |

### bdr.pgd_largeobject_metadata

レプリケートされたラージオブジェクトの所有権とアクセス制御情報を保存します。このテーブルは、

pg_catalog.largeobject_metadata の PGD 同等であり、*Replicated large objects* によって使用されます。その内容はすべてのノードに複製されます。

#### bdr.pgd_largeobject_metadata 列

| Name | Type | Description |
|---|---|---|
| *oid* | *oid* | ラージオブジェクトの OID。 |
| *lomowner* | *oid* | ラージオブジェクトの所有者の OID。 |
| *lomacl* | *aclitem[]* | ラージオブジェクトのアクセス権限。 |

### bdr.queue

このテーブルには、レプリケートされた DDL ステートメントの履歴レコードが保存されます。

#### bdr.queue 列

| Name | Type | Description |
|---|---|---|
| queued_at | timestamptz | ステートメントがキューに入れられたとき |
| role | name | ステートメントを実行したロール |
| replication_sets | text[] | ステートメントはどのレプリケーションセットに公開されましたか |
| message_type | char | メッセージの種類。考えられる値 A-テーブル同期 D-DDL S-シークエンス T-Truncate Q- SQL ステートメント |
| message | json | ステートメントのレプリケーションに必要なメッセージのペイロード |

### bdr.replication_set

レプリケーションセット構成を保存するテーブル。ユーザークエリーの場合、代わりに bdr.replication_sets ビューを確認することをお勧めします。

#### bdr.replication_set 列

| Name | Type | Description |
|---|---|---|
| set_id | oid | レプリケーションセットの OID |
| set_nodeid | oid | ノードの OID 現在は常にローカルノード oid |

次のページに続く

表 46 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| set_name | name | レプリケーションセットの名前 |
| replicate_insert | boolean | レプリケーションセットが INSERT をレプリケートするかどうかを示します |
| replicate_update | boolean | レプリケーションセットが UPDATE をレプリケートするかどうかを示します |
| replicate_delete | boolean | レプリケーションセットが DELETE をレプリケートするかどうかを示します |
| replicate_truncate | boolean | レプリケーションセットが TRUNCATE をレプリケートするかどうかを示します |
| set_isinternal | boolean | 予約済み |
| set_autoadd_tables | boolean | 新しいテーブルがこのレプリケーションセットに自動的に追加されるかどうかを示します |
| set_autoadd_seqs | boolean | 新しいシーケンスがこのレプリケーションセットに自動的に追加されるかどうかを示します |

### bdr.replication_set_table

レプリケーションセットテーブルのメンバーシップを保存するテーブル。ユーザークエリーの場合、代わりに `bdr.tables` ビューを確認することをお勧めします。

#### bdr.replication_set_table 列

| Name | Type | Description |
|---|---|---|
| set_id | oid | レプリケーションセットの OID |
| set_reloid | regclass | テーブルのローカル ID |
| set_att_list | text[] | 予約済み |
| set_row_filter | pg_node_tree | コンパイルされた行フィルタリング式 |

### bdr.replication_set_ddl

レプリケーションセット ddl レプリケーションフィルターを保存するテーブル。ユーザークエリーの場合、代わりに `bdr.ddl_replication` ビューを確認することをお勧めします。

#### bdr.replication_set_ddl 列

| Name | Type | Description |
|---|---|---|
| set_id | oid | レプリケーションセットの OID |
| set_ddl_name | name | DDL フィルタの名前 |
| set_ddl_tag | text | DDL フィルタのコマンドタグ |
| set_ddl_role | text | DDL を実行するロール |

### bdr.replication_sets

現在ノードで使用されていない場合でも、PGD グループで定義されたレプリケーションセットを表示するビュー。

#### bdr.replication_sets 列

| Name | Type | Description |
|---|---|---|
| set_id | oid | レプリケーションセットの OID |
| set_name | name | レプリケーションセットの名前 |
| replicate_insert | boolean | レプリケーションセットが INSERT をレプリケートするかどうかを示します |
| replicate_update | boolean | レプリケーションセットが UPDATE をレプリケートするかどうかを示します |
| replicate_delete | boolean | レプリケーションセットが DELETE をレプリケートするかどうかを示します |
| replicate_truncate | boolean | レプリケーションセットが TRUNCATE をレプリケートするかどうかを示します |
| set_autoadd_tables | boolean | 新しいテーブルがこのレプリケーションセットに自動的に追加されるかどうかを示します |
| set_autoadd_seqs | boolean | 新しいシーケンスがこのレプリケーションセットに自動的に追加されるかどうかを示します |

### bdr.resolved_transactions

`bdr.resolve_failed_transaction` によって解決された、失敗した各トランザクションの概要レコードを保存するテーブル。`bdr.pending_failed_changes` を切り詰めて領域を再利用した後、解決アクションの履歴としてテーブルを使用します。

#### bdr.resolved_transactions 列

| Name | Type | Description |
|---|---|---|
| *id* | *bigint* | 主キー。 |
| *sub_id* | *oid* | サブスクリプションの OID。 |
| *origin_node_id* | *oid* | 元ノードの OID。 |
| *remote_xid* | *xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | *pg_lsn* | 元ノードで解決されたトランザクションのコミット LSN。 |
| *remote_commit_end_lsn* | *pg_lsn* | コミットレコードの終わりの LSN。 |
| *remote_commit_time* | *timestamptz* | 元のノードでのコミットタイムスタンプ。 |
| *resolution* | *text* | 適用される解像度 *skipped*、*applied*、または`discarded`。 |
| *insert_count* | *integer* | トランザクション内の INSERT 操作の数。 |
| *update_count* | *integer* | トランザクション内の UPDATE 操作の数。 |
| *delete_count* | *integer* | トランザクション内の DELETE 操作の数。 |

次のページに続く

表 50 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| *ddl_count* | *integer* | トランザクション内の DDL 操作の数。 |
| *error_message* | *text* | 元の障害からのエラーメッセージ。 |
| *error_detail* | *text* | 追加のエラーの詳細。 |
| *error_sqlstate* | *text* | 元のエラーの SQLSTATE コード。 |
| *recorded_at* | *timestamptz* | 障害が最初にキャプチャされたときのタイムスタンプ。 |
| *resolved_at* | *timestamptz* | トランザクションが解決されたときのタイムスタンプ。 |

### bdr.resolved_transactions_info

人間が判読できるサブスクリプションとオリジンノード名、および計算された合計変更数を追加する `bdr.resolved_transactions` のビュー。

#### `bdr.resolved_transactions_info` 列

| Name | Type | Description |
|---|---|---|
| *id* | *bigint* | 主キー。 |
| *subscription_name* | *name* | サブスクリプションの名前。 |
| *origin_node* | *name* | 元ノードの名前。 |
| *remote_xid* | *xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | *pg_lsn* | 元ノードで解決されたトランザクションのコミット LSN。 |
| *remote_commit_time* | *timestamptz* | 元のノードでのコミットタイムスタンプ。 |
| *resolution* | *text* | 適用される解像度 *skipped*、*applied*、または`discarded`。 |
| *insert_count* | *integer* | トランザクション内の INSERT 操作の数。 |
| *update_count* | *integer* | トランザクション内の UPDATE 操作の数。 |
| *delete_count* | *integer* | トランザクション内の DELETE 操作の数。 |
| *ddl_count* | *integer* | トランザクション内の DDL 操作の数。 |
| *total_changes* | *integer* | 変更の合計数 `insert_count + update_count + delete_count + ddl_count`。 |
| *error_message* | *text* | 元の障害からのエラーメッセージ。 |
| *error_sqlstate* | *text* | 元のエラーの SQLSTATE コード。 |
| *recorded_at* | *timestamptz* | 障害が最初にキャプチャされたときのタイムスタンプ。 |
| *resolved_at* | *timestamptz* | トランザクションが解決されたときのタイムスタンプ。 |

### bdr.schema_changes

スキーマ win PGD へのすべての変更を表示するシンプルなビュー。

#### bdr.schema_changes 列

| Name | Type | Description |
|---|---|---|
| schema_changes_ts | timestampstz | トリガーの ID |
| schema_changes_change | char | 変更タイプのフラグ |
| schema_changes_classid | oid | クラス ID |
| schema_changes_objectid | oid | オブジェクト ID |
| schema_changes_subid | smallint | サブスクリプション |
| schema_changes_descr | text | オブジェクトが変更されました |
| schema_changes_addrname | text[] | スキーマ変更の場所 |

### bdr.sequence_alloc

galloc シーケンスの割り当ての詳細を表示するビュー。

#### bdr.sequence_alloc 列

| Name | Type | Description |
|---|---|---|
| seqid | regclass | シーケンスの ID |
| seq_chunk_size | bigint | 値内のチャンクのシーケンス番号 |
| seq_allocated_up_to | bigint | NA |
| seq_nallocs | bigint | NA |
| seq_last_alloc | timestamptz | 最後に割り当てられたシーケンス |

### bdr.sequences

このビューには、内部 PGD 簿記のシーケンスを除き、すべてのシーケンスとその種類がリストされます。

#### bdr.sequences 列

| Name | Type | Description |
|---|---|---|
| nspname | name | シーケンスを含む名前空間 |
| relname | name | シーケンスの名前 |
| seqkind | text | シーケンスのタイプ'local'、'timeshard'、'galloc' |

### bdr.stat_activity

各バックエンドまたはワーカープロセスの動的アクティビティ。

これには、

pg_stat_activity と同じ情報が含まれていますが、待機が PGD に関連し、次の接続マネージャー関連フィールドが追加されるときに `wait_event` が正しく設定される場合を除きます。

#### bdr.stat_activity 列

| Name | Type | Description |
|---|---|---|
| connection_manager_client_addr | inet | クライアント接続の IP アドレス |
| connection_manager_client_port | int | 接続マネージャーに接続したクライアントのソースポート接続マネージャーを介して接続が行われる場合 |
| connection_manager_client_hostn | text | クライアント接続のホスト名接続が接続マネージャーを介して行われる場合 |
| session_read_only | boolean | セッションが読み取り専用かどうか。接続マネージャーの読み取り専用ポートに接続されています |

### bdr.stat_commit_scope

各コミットスコープの統計を含むビュー。

#### bdr.stat_commit_scope 列

| Column | Type | Description |
|---|---|---|
| commit_scope_name | name | コミットスコープの名前 |
| group_name | name | コミットスコープが定義されているグループの名前 |
| ncalls | bigint | コミットスコープが使用された回数 |
| ncommits | bigint | コミットスコープで行われたコミットの成功した数 |
| naborts | bigint | 使用されるコミットスコープが最終的に中止された回数 |
| ndegrades | bigint | コミット中にバックエンドが低下タイムアウトにヒットしたトランザクションごとの低下イベントの数 |
| nconfig_degrades | bigint | バックグラウンドワーカーによって低下モードに変更された構成レベルの状態の数 |
| last_state_change_time | timestamp with time zone | 最後に構成レベルの状態変更のタイムスタンプ縮退状態への移行または回復 |
| total_commit_time | double precision | コミットスコープを使用したコミットに費やされた合計時間ミリ秒単位 |
| min_commit_time | double precision | コミットスコープを使用してコミットに費やされた最小時間ミリ秒単位 |
| max_commit_time | double precision | コミットスコープを使用してコミットに費やした最大時間ミリ秒単位 |
| mean_commit_time | double precision | コミットスコープを使用したコミットに費やされた平均時間ミリ秒単位 |
| stats_reset | timestamp with time zone | ビュー内のすべての統計が最後にリセットされた時間 |

### bdr.stat_commit_scope_state

バックエンドによるコミットスコープの現在の使用に関する情報のビュー。

#### bdr.stat_commit_scope_state 列

| Column | Type | Description |
|---|---|---|
| pid | integer | バックエンドのプロセス ID |
| commit_scope_name | name | 使用されているコミットスコープの名前 |
| group_name | name | コミットスコープが定義されているグループの名前 |
| waiting_op_num | integer | まだ満たされていないコミットスコープ内の最初のオペレーションのインデックス |
| waiting_prepare_confirmations | integer | オペレーションに必要な PREPARE 確認の数 |
| waiting_commit_confirmations | integer | オペレーションに必要な COMMIT 確認の数 |
| waiting_lsn_confirmations | integer | オペレーションでまだ必要な LSN 確認の数 |

### bdr.stat_connection_manager

このノードの接続マネージャーの統計を含むビュー。

#### bdr.stat_connection_manager 列

| Column | Type | Description |
|---|---|---|
| ntotal_rw_conns | bigint | 読み取り/書き込み接続の合計数 |
| ntotal_ro_conns | bigint | 読み取り専用接続の合計数 |
| nactive_rw_conns | int | アクティブな読み取り/書き込み接続の数 |
| nactive_ro_conns | int | アクティブな読み取り専用接続の数 |

### bdr.stat_connection_manager_connections

接続マネージャーへの接続に関する情報を含むビュー。

#### bdr.stat_connection_manager_connections 列

| Column | Type | Description |
|---|---|---|
| connection_manager_client_addr | text | 接続マネージャーに接続したクライアントの IP アドレス。 |
| connection_manager_client_port | int | クライアントが接続マネージャーとの通信に使用している TCP ポート番号。 |

次のページに続く

表 59 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| connection_manager_addr | text | 接続マネージャーノードの IP アドレス。 |
| connection_manager_port | int | 接続マネージャーが Postgres ノードとの通信に使用している TCP ポート番号。 |
| session_read_only | boolean | セッションが読み取り専用かどうか。 |
| client_uses_tls | boolean | クライアントが TLS を使用して接続マネージャーノードに接続しているかどうか。 |

### bdr.stat_connection_manager_node_stats

このノードの接続マネージャーのサーバー接続統計に関する情報を含むビュー。

#### bdr.stat_connection_manager_node_stats 列

| Column | Type | Description |
|---|---|---|
| node_id | oid | ノードの OID |
| node_name | name | ノードの名前 |
| route_rw_connections | boolean | 読み取り/書き込み接続をこのノードにルーティングするかどうか |
| route_ro_connections | boolean | 読み取り専用接続をこのノードにルーティングするかどうか |
| ntotal_rw_conns | bigint | 読み取り/書き込み接続の合計数 |
| ntotal_ro_conns | bigint | 読み取り専用接続の合計数 |
| nactive_rw_conns | int | アクティブな読み取り/書き込み接続の数 |
| nactive_ro_conns | int | アクティブな読み取り専用接続の数 |

### bdr.stat_connection_manager_hba_file_rules

接続マネージャーが HBA ファイル pg_hba.conf から使用している有効なサポートされているルールのみを表示するビューと、それらのルールに関する情報。

#### bdr.stat_connection_manager_hba_file_rules 列

| Column | Type | Description |
|---|---|---|
| rule_number | integer | ルール番号。これは、認証中に一致が見つかるまで各ルールが検討される順序を示します。 |
| file_name | text | このルールを含むファイルの名前。 |
| line_number | integer | file_name で参照されるファイル内のこのルールの行番号。 |
| type | text | 接続の種類。 |
| database | text[] | このルールが適用されるデータベース名のリスト。 |

次のページに続く

表 61 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| user_name | text[] | このルールが適用されるユーザー名のリスト |
| address | text | ホスト名または IP アドレス、またはすべてのいずれか、samehost または samenet、またはローカル接続の場合は null。 |
| netmask | text | IP アドレスマスク、または該当しない場合は null。 |
| auth_method | text | 認証方法。 |
| auth_options | text | 認証方法に指定されたオプションがある場合。 |

### bdr.stat_raft_followers_state

Raft リーダーノード上の raft リーダーのフォロワーの状態のビュー他のノードでは空。

#### bdr.stat_raft_followers_state 列

| Column | Type | Description |
|---|---|---|
| group_name | name | この情報の対象となるグループ各グループには、個別のコンセンサスを構成できます。 |
| node_name | name | フォロワーノードの名前。 |
| sent_commit_index | bigint | 最新の Raft インデックスがフォロワーノードに送信されます。 |
| match_index | bigint | Raft インデックスは、フォロワーノードからの次の応答と一致すると予想されます。 |
| last_message_time | timestamp with time zone | フォロワーノードから見た最後のメッセージ任意の要求を含む。 |
| last_heartbeat_send_time | timestamp with time zone | リーダーが最後にハートビートをフォロワーノードに送信した時間。 |
| last_heartbeat_response_tir | timestamp with time zone | リーダーが最後にフォロワーノードからのハートビート応答を見た時間。 |
| approx_clock_drift_ms | bigint | ミリ秒単位のフォロワーノードに対してリーダーによって見られたおおよそのクロックドリフト。 |

### bdr.stat_raft_state

ローカルノードでの Raft コンセンサスの状態を説明するビュー。

#### bdr.stat_raft_state 列

| Column | Type | Description |
|---|---|---|
| group_name | name | この情報の対象となるグループ各グループには、個別のコンセンサスを構成できます。 |

次のページに続く

表 63 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| raft_stat | text | Raft のローカルノードの状態 'LEADER'、'CANDIDATE'、'FOLLOWER'、'STOPPED' |
| leader_name | name | Raft リーダーの名前がある場合 |
| voted_for_name | name | ローカルノードがリーダーとして最後に投票したノード |
| is_voting | boolean | Raft のローカルノード部分が投票しています |
| heartbeat_timeout_ms | bigint | ローカルノードのハートビートタイムアウト |
| heartbeat_elapsed_ms | bigint | ローカルノードがリーダーからのハートビートを確認してから経過したミリ秒数 |
| current_term | bigint | ローカルノードの現在の Raft ターム |
| commit_index | bigint | ローカルノードの現在の Raft コミットインデックス |
| apply_index | bigint | ローカルノードがカタログに適用された Raft コミットインデックス |
| last_log_term | bigint | 要求ログの最後の Raft ターム |
| last_log_index | bigint | 要求ログの最後の Raft インデックス |
| oldest_log_index | bigint | 要求ログに残っている最も古い Raft インデックス |
| newest_prunable_log_index | bigint | 要求ログから安全に削除できる最新の Raft インデックス |
| snapshot_term | bigint | 最後のスナップショットの Raft ターム |
| snapshot_index | bigint | 最後のスナップショットの Raft インデックス |
| nnodes | integer | Raft コンセンサス内のノードの数通常はグループ内のノードの数と同じである必要があります。 |
| nvoting_nodes | integer | Raft コンセンサスでの投票ノードの数 |

### bdr.stat_receiver

レプリケーションサブスクリプションレシーバープロセスに関する必要な情報をすべて含むビュー。

#### bdr.stat_receiver 列

| Column | Type | Description |
|---|---|---|
| worker_role | text | BDR ワーカーのロール常に'receiver' |
| worker_state | text | レシーバーワーカーの状態'running'、'down'、または 'disabled' のいずれかです。 |
| worker_pid | integer | レシーバーワーカーのプロセス ID |
| sub_name | name | レシーバーが属しているサブスクリプションの名前 |
| sub_slot_name | name | レシーバーが使用するレプリケーションスロット名 |
| source_name | name | このレシーバーのソースノード接続先。これは通常起点ノードと同じですが、フォワードモードサブスクリプションの場合は異なります。 |

次のページに続く

表 64 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| origin_name | name | このレシーバーのオリジンノード転送された変更を受信するノード。これは通常ソースノードと同じですが、フォワードモードサブスクリプションの場合は異なります。 |
| subscription_mode | char | サブスクリプションのモード、詳細については :ref:``bdr.subscription_summary` 列 <*bdr.subscription_summary* 列>` を参照してください。 |
| sub_replication_sets | text[] | このレシーバーがサブスクライブしているレプリケーションセット |
| sub_apply_delay | interval | 適用遅延インターバル |
| receive_lsn | pg_lsn | これまでに受け取った最後の変更の LSN |
| receive_commit_lsn | pg_lsn | これまでに受け取った最後のコミットの LSN |
| xact_apply_lsn | pg_lsn | 最後に適用されたトランザクション LSN |
| xact_flush_lsn | pg_lsn | 最後にフラッシュされたトランザクション LSN |
| xact_apply_timestamp | timestamp with time zone | 最後に適用されたトランザクションコミットのタイムスタンプ |
| worker_start | timestamp with time zone | 受信機が起動した時刻 |
| worker_xact_start | timestamp with time zome | レシーバーがローカル db トランザクションを開始した時間現在ローカルトランザクションを処理している場合、通常は NULL、詳細については、*Replicated large objects* の `xact_start`を参照してください。 |
| worker_backend_state_chai | timestamp with time zone | バックエンド状態変更のタイムスタンプ、詳細については、 *Replicated large objects* の `state_change`を参照してください |
| worker_backend_state | text | 現在のバックエンド状態、詳細については、*Replicated large objects* の `state`を参照してください。 |
| wait_event_type | text | レシーバーが現在待機している待機イベントのタイプある場合、詳細については、*Replicated large objects* の `wait_event_type`を参照してください |
| wait_event | text | レシーバーが現在待機している正確なイベント存在する場合、詳細については *Replicated large objects* の `wait_event`を参照してください。 |
| connection_dsn | text | レプリケーションレシーバーがソースノードに接続するために現在使用している DSN。同じサブグループ内のノードに設定されている場合は *Replicated large objects* を反映し、それ以外の場合は `node_dsn`を反映します。 |

### bdr.stat_receiver_transactions

リアルタイムのトランザクションステータスを含むビュー、レシーバーがアクティブに処理しているトランザクション、ライターに送信されるトランザクション、およびトランザクションがファイルまたはライタにストリーミングされているかどうかを詳細に示します。また、現在適用されているトランザクションに関するライターの進行状況も報告します。

#### bdr.stat_receiver_transactions 列

| Column | Type | Description |
|---|---|---|
| sub_name | name | サブスクリプションの名前 |
| remote_node | bdr.regnode | トランザクションを送信する上流ノードの名前 |
| remote_xid | xid | 上流ノードからのトランザクション識別子 |
| remote_commit_timestamp | timestamp with time zone | 上流ノードからのトランザクションコミットタイムスタンプ |
| remote_commit_lsn | pg_lsn | 上流ノードからのトランザクションコミット LSN |
| origin_node_name | bdr.regnode | トランザクションが最初にコミットされたオリジンノード |
| origin_lsn | pg_lsn | オリジンノード上のトランザクションの LSN |
| txn_start_received_at | timestamp with time zone | トランザクションがダウンストリームノードで最初に受信されたときのタイムスタンプ |
| txn_last_change_received_ | timestamp with time zone | トランザクションの最後の変更がダウンストリームノードで受信されたときのタイムスタンプ |
| txn_last_change_received_l | pg_lsn | 受け取った最後に変更に対応する LSN |
| txn_source | text | このトランザクションのソース リプレイ/ファイル/ストリーム |
| is_streaming | text | トランザクションが現在上流からストリーミング中であるかどうか |
| txn_status | text | トランザクションのステータス |
| commit_position | integer | コミットキュー内のこのトランザクションの位置 |
| writer_pid | integer | トランザクションを適用するライタプロセスの PID 該当する場合 |
| txn_size | text | トランザクションのサイズトランザクションがまだストリーミング中の場合は変更される可能性があります |
| txn_apply_progress | text | トランザクションの進行状況ライタによって現在適用されている場合 |
| last_change_applied_type | text | ライタによって適用された最後の変更のタイプ |
| last_change_applied_lsn | pg_lsn | ライタによって適用された最後の変更の LSN |

### bdr.stat_reconcile_commit_scope

コミットスコープごとの統計を表示します。*bdr.stat_reset* ファンクションを使用して、統計をリセットできます。

#### bdr.stat_reconcile_commit_scope 列

| Column | Type | Description |
|---|---|---|
| commit_scope_name | name | 監視されるコミットスコープの名前。 |
| origin_name | name | 調整が実行される元のノードの名前。 |
| nreconciles | bigint | 調整フェーズで処理されたトランザクションの合計数。 |
| ncommits | bigint | このスコープとオリジンノードのリコンシリエーション中に完了した成功したコミットの合計数。 |
| naborts | bigint | 調整フェーズ中に中止されたトランザクションの合計数。 |
| ntimeouts | bigint | 特に調整タイムアウトによって発生した中止の数。 |
| total_reconcile_time | bigint | 調整フェーズで費やされた累積時間ミリ秒単位。 |
| stats_reset | timestamptz | このビューの統計が最後にリセットされたときのタイムスタンプ。 |

### bdr.stat_relation

各リレーションの適用統計を表示します。*bdr.track_relation_apply* で追跡が有効になっている場合にのみデータが含まれます

および特定のリレーションに対してデータがレプリケートされたかどうか。

`lock_acquire_time` は、 *bdr.track_apply_lock_timing* の場合にのみ更新されます

は on デフォルト `off` に設定されます。

*bdr.reset_relation_stats()* を呼び出すことにより、保存されたリレーション統計をリセットできます。

#### bdr.stat_relation 列

| Column | Type | Description |
|---|---|---|
| nspname | name | リレーションのスキーマの名前 |
| relname | name | リレーションの名前 |
| relid | oid | リレーションの OID |
| total_time | double precision | リレーションのレプリケーションの処理に費やされた合計時間ミリ秒単位 |
| ninsert | bigint | リレーションに対してレプリケートされた挿入の数 |
| nupdate | bigint | リレーションに対して複製された更新の数 |
| ndelete | bigint | リレーションに対してレプリケートされた削除の数 |
| ntruncate | bigint | リレーションに対してレプリケートされたトランケートの数 |

次のページに続く

表 67 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| shared_blks_hit | bigint | リレーションの共有ブロックキャッシュヒットの合計数 |
| shared_blks_read | bigint | リレーションに対して読み取られた共有ブロックの合計数 |
| shared_blks_dirtied | bigint | リレーション用にダーティな共有ブロックの合計数 |
| shared_blks_written | bigint | リレーション用に書き込まれた共有ブロックの合計数 |
| blk_read_time | double precision | リレーションのブロックの読み取りに費やされた合計時間ミリ秒単位 `track_io_timing`が有効な場合、それ以外の場合はゼロ |
| blk_write_time | double precision | リレーションのブロックの書き込みに費やされた合計時間ミリ秒単位 `track_io_timing`が有効な場合、それ以外の場合はゼロ |
| lock_acquire_time | double precision | リレーションのロックの取得に費やされた合計時間ミリ秒単位 `bdr.track_apply_lock_timing`が有効な場合、それ以外の場合はゼロ |
| stats_reset | timestamp with time zone | 最後に統計をリセットした時間 `bdr.reset_relation_stats()`によって実行 |

### bdr.stat_routing_candidate_state

Raft リーダーのルーティング候補ノードに関する情報のビュー他のノードでは空。

#### bdr.stat_routing_candidate_state 列

| Column | Type | Description |
|---|---|---|
| node_group_name | name | この情報の対象となるグループ各グループは別個のルーティングプロキシを持つことができます。 |
| node_name | name | 候補ノード名 |
| node_route_fence | boolean | ノードはフェンスされています true の場合、リーダーまたは読み取り専用の接続ターゲットになることはできません。 |
| node_route_reads | boolean | ノードは読み取り専用の接続ターゲットとして考えられています |
| node_route_writes | boolean | ノードは書き込みリード候補として検討されています。 |
| last_message_time | timestamp with time zone | このノードによって見られた最後の Raft メッセージ任意の要求を含む時間ノードの活性度を確認するために使用されます |

### bdr.stat_routing_state

接続マネージャーが接続をルーティングするために使用する接続ルーティングの状態のビュー。

#### bdr.stat_routing_state 列

| Column | Type | Description |
|---|---|---|
| node_group_name | name | これはグループの情報です各グループは個別のルーティングプロキシを持つことができます。 |
| write_lead_name | name | 書き込みリードノードの名前 |
| previous_write_lead_name | name | 以前の書き込みリードノードの名前 |
| read_names | name[] | 読み取り専用接続がルーティングされるノードの配列 |
| write_candidate_names | name[] | フェールオーバーの場合に書き込みリードになるために必要なすべての基準と一致するノード |
| read_candidate_names | name[] | フェールオーバーの場合に読み取り専用の接続ターゲットになるために必要なすべての基準と一致するノード |

### bdr.stat_subscription

各サブスクリプションの適用統計を表示します。*bdr.track_subscription_apply* で追跡が有効になっている場合にのみデータが含まれます。

*bdr.reset_subscription_stats()* を呼び出すことにより、保存されたサブスクリプション統計をリセットできます。

#### bdr.stat_subscription 列

| Column | Type | Description |
|---|---|---|
| sub_name | name | サブスクリプションの名前 |
| subid | oid | サブスクリプションの OID |
| mean_apply_time | double precision | 適用トランザクションごとの平均時間ミリ秒 |
| nconnect | bigint | このサブスクリプションがアップストリームに接続した回数 |
| ncommit | bigint | このサブスクリプションが行ったコミット数 |
| nabort | bigint | ライタがこのサブスクリプションに対して実行した中止の数 |
| nerror | bigint | このサブスクリプションでライタがヒットしたエラーの数 |
| nskippedtx | bigint | このサブスクリプションのライタによってスキップされたトランザクションの数`skip_transaction`競合リゾルバーのため |
| ninsert | bigint | このサブスクリプションが行った挿入の数 |

次のページに続く

表 70 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| nupdate | bigint | このサブスクリプションが行った更新の数 |
| ndelete | bigint | このサブスクリプションが行った削除数 |
| ntruncate | bigint | このサブスクリプションが行った切り捨ての数 |
| nddl | bigint | このサブスクリプションが実行した DDL 操作の数 |
| ndeadlocks | bigint | デッドロックが発生したエラーの数 |
| nretries | bigint | ライタが実行した再試行数完全な再起動/再接続を行わずに |
| nstream_writer | bigint | ライタにストリーミングされたトランザクションの数 |
| nstream_file | bigint | ファイルにストリーミングされたトランザクションの数 |
| nstream_commit | bigint | コミットされたストリーミングトランザクションの数 |
| nstream_abort | bigint | アボートされたストリーミングトランザクションの数 |
| nstream_start | bigint | 処理された STREAM START メッセージの数 |
| nstream_stop | bigint | 処理された STREAM STOP メッセージの数 |
| nstream_commit | bigint | コミットされたストリーミングトランザクションの数 |
| nstream_abort | bigint | アボートされたストリーミングトランザクションの数 |
| nstream_prepare | bigint | 準備されたストリーミングトランザクションの数 |
| nstream_insert | bigint | 処理されたストリーミング挿入の数 |
| nstream_update | bigint | 処理されたストリーミング更新の数 |
| nstream_delete | bigint | 処理されたストリーミング削除の数 |
| nstream_truncate | bigint | 処理されたストリーミングトランケートの数 |
| shared_blks_hit | bigint | サブスクリプションによる共有ブロックキャッシュヒットの合計数 |
| shared_blks_read | bigint | サブスクリプションによって読み取られた共有ブロックの合計数 |
| shared_blks_dirtied | bigint | サブスクリプションによってダーティされた共有ブロックの合計数 |
| shared_blks_written | bigint | サブスクリプションによって書き込まれた共有ブロックの合計数 |
| blk_read_time | double precision | サブスクリプションがブロックの読み取りに費やした合計時間ミリ秒単位`track_io_timing`が有効な場合、それ以外の場合はゼロ |
| blk_write_time | double precision | サブスクリプションがブロックの書き込みに費やした合計時間ミリ秒単位`track_io_timing`が有効な場合、それ以外の場合はゼロ |
| connect_time | timestamp with time zone | 現在の上流接続が確立された時間、接続されていない場合は NULL |
| last_disconnect_time | timestamp with time zone | 最後のアップストリーム接続がドロップされた時刻 |
| start_lsn | pg_lsn | このサブスクリプションが上流からレプリケーションの開始を要求した LSN |
| retries_at_same_lsn | bigint | サブスクリプションが同じ LSN 値からリスタートされた試行回数 |

次のページに続く

表 70 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| curr_ncommit | bigint | 現在の接続が確立された後にこのサブスクリプションが実行したコミット数 |
| npre_commit_confirmation | bigint | CAMO パートナーによるプリコミット確認の数 |
| npre_commit | bigint | プリコミット数 |
| ncommit_prepared | bigint | 準備されたトランザクションのコミット数 |
| nabort_prepared | bigint | 準備されたトランザクションアボートの数 |
| nprovisional_waits | bigint | 同時適用トランザクションによる同じタプルの更新/削除操作の数。これらは暫定的な待機です。 *Parallel Apply* を参照 |
| ntuple_waits | bigint | 安全に適用されるまで待機した更新/削除操作の数。 *Parallel Apply* を参照 |
| ncommit_waits | bigint | コミットされる前に待機する必要があった、完全に適用されたトランザクションの数。 *Parallel Apply* を参照 |
| neager_lock_conflicts | bigint | 別のライタまたはバックエンドがロックを保持しているために、クォーラムコミットトランザクションの適用中にライタが待機し、場合によっては再起動が必要な回数。 `ncommit_prepared`と比較して高い値は、議論の多いワークロードを示します。 |
| neager_conflicts | bigint | `neager_lock_conflicts`を除く、発生した熱心な競合の数。これらは、ライタがトランザクションを中止し、オリジンに負のフィードバックを送信する原因となる競合です。 |
| stats_reset | timestamp with time zone | 最後に統計をリセットした時間 `bdr.reset_subscription_stats()`によって実行 |

### bdr.stat_worker

PGD マネージャーワーカーの概要情報とワーカーごとの統計を含むビュー。

#### bdr.stat_worker 列

| Column | Type | Description |
|---|---|---|
| worker_role | text | BDR ワーカーのロール |
| worker_pid | integer | ワーカーのプロセス ID |
| sub_name | name | ワーカーが関連しているサブスクリプションの名前がある場合 |
| worker_start | timestamp with time zone | ワーカーが開始した時刻 |
| worker_xact_start | timestamp with time zone | ワーカーがローカル db トランザクションを開始した時間。詳細については、 *Replicated large objects* の `xact_start`を参照してください |

次のページに続く

表 71 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| worker_xid | xid | ワーカーのトランザクション ID、詳細については、*Replicated large objects* の `backend_xid`を参照してください。 |
| worker_xmin | xid | ワーカーが必要とする最も古いトランザクション ID。詳細については、 *Replicated large objects* の `backend_xmin`を参照してください。 |
| worker_backend_state_chai | timestamp with time zone | バックエンド状態変更のタイムスタンプ詳細については、 *Replicated large objects* の `state_change`を参照してください |
| worker_backend_state | text | 現在のバックエンド状態、詳細については、*Replicated large objects* の `state`を参照してください |
| wait_event_type | text | ワーカーが現在待機している待機イベントのタイプ存在する場合、詳細については、*Replicated large objects* の `wait_event_type`を参照してください。 |
| wait_event | text | ワーカーが待機している正確なイベント存在する場合、詳細については、 *Replicated large objects* の `wait_event`を参照してください。 |
| blocked_by_pids | integer[] | ワーカーをブロックする PID のリスト 存在する場合 |
| query | text | ワーカーによって現在実行中のクエリ |
| worker_query_start | timestamp with time zone | ワーカーによって実行された現在のクエリが開始されたタイムスタンプ |

### bdr.stat_writer

各サブスクリプションレプリケーションライタの概要情報と統計を含むビュー。サブスクリプションごとに複数のライターが存在する場合があります。

#### bdr.stat_writer 列

| Column | Type | Description |
|---|---|---|
| worker_role | text | BDR ワーカーのロール常に' ライタ' |
| worker_state | text | ワーカーの状態'running'、'down'、または'disabled' のいずれかです。 |
| worker_pid | integer | ライタのプロセス ID |
| sub_name | name | ライタが属しているサブスクリプションの名前 |
| writer_nr | integer | 同じサブスクリプションのライタグループのライターインデックス |
| nxacts | bigint | ライタが開始以降に処理したトランザクションの数 |
| ncommits | bigint | start 以降にライタが処理したコミットの数 |
| naborts | bigint | start 以降にライタが処理した中止の数 |

次のページに続く

表 72 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| commit_queue_position | integer | 同じライタグループ内の他のライターに対してトランザクションをシリアル化する場合、コミットキュー内の位置 |
| xact_source_xid | xid | ソースノードで現在処理されているトランザクションのトランザクション ID |
| xact_source_commit_lsn | pg_lsn | ソースノードで現在処理されているトランザクションの LSN |
| xact_nchanges | bigint | 現在処理されているトランザクションで書き込まれた変更の数 1000 変更ごとに更新 |
| xact_origin_node_name | name | 現在処理されているトランザクションのオリジンノード |
| xact_origin_lsn | pg_lsn | 現在処理されているトランザクションのオリジン LSN |
| xact_origin_timestamp | timestamp with time zone | 現在処理されているトランザクションのオリジンコミットタイムスタンプ |
| streaming_allowed | boolean | ライタは大規模なトランザクションの直接ストリームを受信できます |
| is_streaming | boolean | ライタは現在、大規模なトランザクションの直接ストリームを受信しています |
| nstream_file | bigint | ライタが処理したストリームファイルの数 |
| nstream_writer | bigint | ライタが処理した直接ストリーミングトランザクションの数 |
| worker_start | timestamp with time zone | ライタが書き始めた時期 |
| worker_xact_start | timestamp with time zone | ライタがローカル db トランザクションを開始した時間詳細については、 *Replicated large objects* の xact_start を参照してください。 |
| worker_xid | xid | ワーカーのトランザクション ID 詳細については、*Replicated large objects* の `backend_xid`を参照してください。 |
| worker_xmin | xid | ワーカーが必要とする最も古いトランザクション ID 詳細については、 *Replicated large objects* の `backend_xmin`を参照してください。 |
| worker_backend_state_chai | timestamp with time zone | バックエンド状態変更のタイムスタンプ詳細については、 *Replicated large objects* の `state_change`を参照してください。 |
| worker_backend_state | text | 現在のバックエンド状態詳細については、 *Replicated large objects* の `state`を参照してください。 |
| wait_event_type | text | ライタが現在待機している待機イベントのタイプある場合、詳細については、 *Replicated large objects* の `event_type`を参照してください。 |
| wait_event | text | ライタが待機している正確なイベント存在する場合、詳細については *Replicated large objects* の `wait_event`を参照してください。 |

次のページに続く

表 72 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| blocked_by_pids | integer[] | ライタをブロックする PID のリスト 存在する場合 |
| query | text | ライタによって現在実行中のクエリ通常は DDL にのみ設定されます |
| worker_query_start | timestamp with time zone | ワーカーによって実行された現在のクエリが開始されたタイムスタンプ |
| command_progress_cmdtag | text | 進行状況追跡のあるコマンドの場合、ライタによって現在処理されているコマンドを示します。 'CREATE INDEX'、'CREATE INDEX CONCURRENTLY'、'REINDEX'、'REINDEX CONCURRENTLY'、'CLUSTER'、および'VACUUM FULL' のいずれかです。 |
| command_progress_relation | text | 進行状況追跡を使用するコマンドの場合、コマンドがどの関係を作業しているかを特定します |
| command_progress_phase | text | 進行状況追跡を含むコマンドの場合、コマンドが行っている現在のフェーズの名前、詳細については、Postgres ドキュメントの *Replicated large objects* を参照してください |
| command_progress_count | integer | 進行状況追跡を使用するコマンドの場合、このコマンドが通過したフェーズの数 |
| command_progress_phase_nr | integer | 進行状況追跡を使用するコマンドの場合、`command_progress_count`のフェーズの数 |
| command_progress_phase_tupl | real | 進行状況追跡を使用するコマンドの場合、コマンドの現在のフェーズで処理する必要がある行数フェーズが行を処理する場合 |
| command_progress_tuples_don | bigint | 進行状況追跡を使用するコマンドの場合、コマンドの現在のフェーズが既に処理した行数フェーズが処理行の場合 |

### bdr.subscription

このカタログテーブルには、ローカル PGD ノードが所有するすべてのサブスクリプションとそのモードがリストされています。

#### bdr.subscription 列

| Name | Type | Description |
|---|---|---|
| sub_id | oid | サブスクリプションの ID |
| sub_name | name | サブスクリプションの名前 |
| nodegroup_id | oid | ノードグループの ID |
| origin_node_id | oid | オリジンノードの ID |
| source_node_id | oid | ソースノードの ID |
| target_node_id | oid | ターゲットノードの ID |
| subscription_mode | char | サブスクリプションのモード |

次のページに続く

表 73 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| sub_enabled | bool | サブスクリプションが有効かどうかレプリケーションである必要があります。 |
| apply_delay | interval | このサブスクリプションの変更の適用はどのくらい遅れますか通常 0 |
| slot_name | name | このサブスクリプションで使用されるアップストリームのスロット |
| origin_name | name | このサブスクリプションで使用されるローカルオリジン |
| num_writers | int | このサブスクリプションが使用するライタプロセスの数 |
| streaming_mode | char | サブスクリプションのストリーミング構成 |
| replication_sets | text[] | このサブスクリプションによってレプリケートされたレプリケーションセット NULL = all |
| forward_origin | text[] | このサブスクリプションによって転送されたオリジン NULL = all |

### bdr.subscription_summary

このビューには、ローカルノードが他のノードに対して持っているすべての PGD サブスクリプションに関する概要情報が含まれています。

#### bdr.subscription_summary 列

| Name | Type | Description |
|---|---|---|
| node_group_name | name | ノードが一部である PGD グループの名前 |
| sub_name | name | サブスクリプションの名前 |
| origin_name | name | オリジンノードの名前 |
| target_name | name | ターゲットノードの名前通常はローカルノード |
| sub_enabled | bool | サブスクリプションは有効になっていますか |
| sub_slot_name | name | このサブスクリプションで使用されるオリジンノードのスロット名 |
| sub_replication_sets | text[] | サブスクライブされたレプリケーションセット |
| sub_forward_origins | text[] | サブスクリプションは、オリジン以外のノードから転送された変更を受け入れますか |
| sub_apply_delay | interval | オリジンと比較して、これだけの遅延トランザクション |
| sub_origin_name | name | このサブスクリプションで使用されるレプリケーション元の名前 |
| bdr_subscription_mode | char | サブスクリプションモード |
| subscription_status | text | サブスクリプションワーカーのステータス |
| node_group_id | oid | ノードが一部である PGD グループの OID |
| sub_id | oid | サブスクリプションの OID |

次のページに続く

表 74 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| origin_id | oid | 元ノードの OID |
| target_id | oid | ターゲットノードの OID |
| receive_lsn | pg_lsn | 受信した変更またはメッセージの最新の LSN これは、再起動の場合に後退する可能性があります |
| receive_commit_lsn | pg_lsn | 受け取った最後に COMMIT の最新の LSN これは、再起動の場合に後退する可能性があります |
| last_xact_replay_lsn | pg_lsn | このサブスクリプションで再生された最後のトランザクションの LSN |
| last_xact_flush_lsn | timestamptz | このサブスクリプションで再生され、永続的にディスクにフラッシュされる最後のトランザクションの LSN |
| last_xact_replay_timestamp | timestamptz | このサブスクリプションで再生された最後のトランザクションのタイムスタンプ |

### bdr.subscription_error_status

各サブスクリプションの現在のエラーポリシー構成とエラーカウンターの状態を示すビュー。

#### bdr.subscription_error_status 列

| Name | Type | Description |
|---|---|---|
| *sub_name* | *name* | サブスクリプションの名前。 |
| *sub_enabled* | *boolean* | サブスクリプションが現在有効になっているかどうか。 |
| *disable_reason* | *text* | サブスクリプションが無効になった理由、または有効になっている場合は `NULL`。 |
| *effective_error_policy* | *text* | サブスクリプションのアクティブなエラーポリシー。ノードグループ構成から派生します。 |
| *max_retries* | *integer* | ポリシーがアクションを実行するまでの構成された最大再試行数。 |
| *current_retries* | *integer* | 現在または最近の失敗の再試行回数。 |
| *max_skips* | *integer* | 設定された最大スキップ数、または無制限の場合は `NULL`。 |
| *total_skips* | *integer* | `record_all_and_skip`ポリシーによってスキップされたトランザクションの合計数。 |
| *pending_changes* | *bigint* | このサブスクリプションのステータスが `pending`である `bdr.pending_failed_changes`の変更数。 |

### bdr.tables

このビューは、レプリケーションセットのテーブルメンバーシップに関する情報をリストします。テーブルが複数のレプリケーションセットに存在する場合、このテーブルに複数回表示されます。

#### bdr.tables 列

| Name | Type | Description |
|---|---|---|
| relid | oid | リレーションの OID |
| nspname | name | スキーマリレーションの名前は次のとおりです |
| relname | name | リレーションの名前 |
| set_name | name | レプリケーションセットの名前 |
| set_ops | text[] | レプリケートされた操作のリスト |
| rel_columns | text[] | レプリケートされた列のリスト NULL = すべての列* |
| row_filter | text | 行フィルタリング式 |
| conflict_detection | text | 使用される競合検出方法 row_origin デフォルト、row_version または column_level |

(*) これらの列は将来の使用のために予約されており、現在 NULL である必要があります。

### bdr.taskmgr_work_queue

タスクマネージャーによって作成および処理された作業項目が含まれています。作業項目は 1 つのノードでのみ作成され、別のノードで処理されます。

#### bdr.taskmgr_work_queue 列

| Column | Type | Description |
|---|---|---|
| ap_wq_workid | bigint | ワークアイテムの一意の ID |
| ap_wq_ruleid | int | autopartition_rules にリストされているルールの ID。ルールは bdr.autopartition コマンドを使用して指定されます |
| ap_wq_relname | name | タスクが属するリレーションの名前 |
| ap_wq_relnamespace | name | この作業項目のルールで指定されたテーブルスペースの名前 |
| ap_wq_partname | name | 作業項目によって作成されたパーティションの名前 |
| ap_wq_work_category | char | 作品カテゴリー; *c* パーティションの作成、*m* パーティションの移行 、 *d* パーティションのドロップ 、または *a* パーティションの変更 です。 |
| ap_wq_work_sql | text | 作業項目の SQL クエリー |
| ap_wq_work_depends | Oid[] | ワークアイテムが依存するノードの OID |

### bdr.taskmgr_workitem_status

各ノードでローカルに更新される作業項目のステータス。

#### bdr.taskmgr_workitem_status 列

| Column | Type | Description |
|---|---|---|
| ap_wi_workid | bigint | ワークアイテムの ID |
| ap_wi_nodeid | Oid | ワークアイテムが処理されているノードの OID |
| ap_wi_status | char | ステータス`q`キューイング、`c`完全、`f`failed、または`u`unknown のいずれかです |
| ap_wi_started_at | timestamptz | 作業項目の開始タイムスタンプ |
| ap_wi_finished_at | timestamptz | 作業項目の終了タイムスタンプ |

### bdr.taskmgr_local_work_queue

タスクマネージャーによって作成および処理された作業項目が含まれています。これは bdr.taskmgr_work_queue に似ていますが、これらの作業項目はローカル管理テーブル用である点が異なります。各ノードは、クラスター内の他のノードから独立して、独自のローカル作業項目を作成および処理します。

#### bdr.taskmgr_local_work_queue 列

| Column | Type | Description |
|---|---|---|
| ap_wq_workid | bigint | ワークアイテムの一意の ID |
| ap_wq_ruleid | int | autopartition_rules にリストされているルールの ID。ルールは bdr.autopartition コマンドを使用して指定されます |
| ap_wq_relname | name | タスクが属するリレーションの名前 |
| ap_wq_relnamespace | name | この作業項目のルールで指定されたテーブルスペースの名前。 |
| ap_wq_partname | name | 作業項目によって作成されたパーティションの名前 |
| ap_wq_work_category | char | カテゴリ; *c* パーティションの作成 、 *m* パーティションの移行 、 *d* パーティションのドロップ 、または *a* パーティションの変更 です。 |
| ap_wq_work_sql | text | 作業項目の SQL クエリー |
| ap_wq_work_depends | Oid[] | 常に NULL |

### bdr.taskmgr_local_workitem_status

ローカルに管理されるテーブルの作業項目のステータス。

#### bdr.taskmgr_local_workitem_status 列

| Column | Type | Description |
|---|---|---|
| ap_wi_workid | bigint | ワークアイテムの ID |

次のページに続く

表 80 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| ap_wi_nodeid | Oid | ワークアイテムが処理されているノードの OID |
| ap_wi_status | char | ステータス`q`キューイング、`c`完全、`f`failed、または`u`unknown のいずれかです |
| ap_wi_started_at | timestamptz | 作業項目の開始タイムスタンプ |
| ap_wi_finished_at | timestamptz | 作業項目の終了タイムスタンプ |

### bdr.trigger

このビューでは、作成されたすべてのストリームトリガーを表示できます。多くの場合、ここでのトリガーは `bdr.create_conflict_trigger` から作成されます。

#### `bdr.trigger` 列

| Name | Type | Description |
|---|---|---|
| trigger_id | oid | トリガーの ID |
| trigger_reloid | regclass | 関連するファンクションの名前 |
| trigger_pgtgid | oid | Postgres トリガー ID |
| trigger_type | char | トリガー呼び出しの種類 |
| trigger_name | name | トリガーの名前 |

### bdr.triggers

読みやすい列を含む `bdr.trigger` の展開ビュー。

| Name | Type | Description |
|---|---|---|
| trigger_name | name | トリガーの名前 |
| event_manipulation | text | オペレーション |
| trigger_type | bdr.trigger_type | トリガーの種類 |
| trigger_table | bdr.trigger_reloid | トリガーを呼び出すテーブル |
| trigger_function | name | 使用するファンクション |

### bdr.workers

実行中の PGD ワーカープロセスに関する情報。

これを pid を使用して `bdr.stat_activity` と結合して、PGD ワーカーの状態についてのさらに多くの洞察を得ることができます。

### bdr.workers 列

| Name | Type | Description |
|---|---|---|
| worker_pid | int | ワーカープロセスのプロセス ID |
| worker_role | int | ワーカーロールの数値表現 |
| worker_role_name | text | ワーカーロールの名前 |
| worker_subid | oid | ワーカーが関連付けられている場合のサブスクリプション ID |

## bdr. ライター

PGD ライタプロセスに関する特定の情報。

### bdr.writers 列

| Name | Type | Description |
|---|---|---|
| sub_name | name | サブスクリプションの名前 |
| writer_nr | int | 現在のトランザクションでライタによって適用された変更の数 |
| pid | int | ワーカープロセスのプロセス ID |
| syncing_rel | int | 同期される関係の OID 存在する場合 |
| streaming_allowed | text | このライタはライタへの直接ストリーミングのターゲットになりますか |
| is_streaming | bool | このライタにストリーミングされているトランザクションはありますか |
| remote_xid | xid | 処理中のトランザクションのリモートトランザクション ID 存在する場合 |
| remote_commit_lsn | pg_lsn | 最後に処理されたコミットの LSN |
| commit_queue_position | int | 内部コミットキューの位置 |
| nxacts | bigint | このライタによって処理されたトランザクションの数 |
| ncommits | bigint | このライタがコミットしたトランザクションの数 |
| naborts | bigint | このライタによって中止されたトランザクションの数 |
| nstream_file | bigint | このライタによって処理されたストリームからファイルへのトランザクションの数 |
| nstream_writer | bigint | このライタによって処理されたライターへのストリーミングトランザクションの数 |
| xact_nchanges | bigint | このライタによって処理された変更の数 1000 行ごとに更新 |
| xact_origin_node_id | oid | トランザクションが発生したノード |
| xact_origin_lsn | pg_lsn | オリジンでのトランザクションの LSN |
| xact_origin_ts | timestamp with time zone | オリジンでのトランザクションのコミットタイムスタンプ |

次のページに続く

表 84 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| xact_apply_source | text | トランザクションの適用に使用されるメソッド *replay/file/stream* |
| xact_apply_size | bigint | トランザクションのサイズ既知の場合 |
| xact_apply_progress | bigint | これまでの取引の進行状況 |
| remote_change_type | "char" | ライタによって適用された最後の変更のタイプ |
| remote_change_lsn | pg_lsn | ライタによって適用された最後の変更の LSN |

### bdr.worker_tasks

bdr.worker_tasks ビューは、PGD の現在のワーカー起動レート制限状態と、バックグラウンドでのワーカーの起動と登録アクティビティに関するいくつかの基本的な統計を表示します。

ここにリストされている他のビューとは異なり、現在のデータベースと PGD ノードに固有ではありません。現在の PostgreSQL インスタンスのすべての PGD ノードの状態が表示されます。現在のデータベースに参加してフィルタリングします。

bdr.worker_tasks は、walsender と出力プラグインを追跡しません。

#### bdr.worker_tasks 列

| Column | Type | Description |
|---|---|---|
| task_key_worker_role | integer | ワーカーロール識別子 |
| task_key_worker_role_nam | text | ワーカーロール名 |
| task_key_dboid | oid | データベース識別子、使用可能な場合 |
| datname | name | データベースの名前使用可能な場合 |
| task_key_subid | oid | サブスクリプション識別子、可能な場合 |
| sub_name | name | サブスクリプションの名前使用可能な場合 |
| task_key_ext_libname | name | ライブラリの名前おそらく bdr |
| task_key_ext_funcname | name | ファンクションエントリーポイントの名前 |
| task_key_ext_workername | name | ワーカーに割り当てられた名前 |
| task_key_remoterelid | oid | リモート同期関係の識別子使用可能な場合 |
| task_pid | integer | ワーカーのプロセス ID |
| task_registered | timestamp with time zone | ワーカー登録タイムスタンプ |
| since_registered | interval | ワーカーが登録されてからのインターバル |
| task_attached | timestamp with time zone | ワーカー接続タイムスタンプ |
| since_attached | interval | ワーカーが接続してからのインターバル |
| task_exited | timestamp with time zone | ワーカー終了タイムスタンプ |
| since_exited | interval | ワーカーが終了してからのインターバル |
| task_success | boolean | ワーカーはまだ実行されていますか？ |

次のページに続く

表 85 – 前のページからの続き

| Column | Type | Description |
|---|---|---|
| task_next_launch_not_befo | timestamp with time zone | ワーカーが再度再起動されるときのタイムスタンプ |
| until_launch_allowed | interval | 次の起動までの残り時間 |
| task_last_launch_requestor | integer | 起動を要求したプロセス ID |
| task_last_launch_request_ti | timestamp with time zone | 要求が行われたときのタイムスタンプ |
| since_last_request | interval | 最後の要求からのインターバル |
| task_last_launch_request_a | boolean | 最後の要求は成功しましたか？ |
| task_nrequests | integer | 要求の数 |
| task_nregistrations | integer | 登録数 |
| task_prev_pid | integer | 前世代のプロセス ID |
| task_prev_registered | timestamp with time zone | 以前に登録されたタスクのタイムスタンプ |
| since_prev_registered | interval | 前回の登録からのインターバル |
| task_prev_launched | timestamp with time zone | 前回の起動のタイムスタンプ |
| since_prev_launched | interval | 前回の起動からのインターバル |
| task_prev_exited | timestamp with time zone | 前のタスクが終了した時のタイムスタンプ |
| since_prev_exited | interval | 前のタスクが終了してからの間隔 |
| task_first_registered | timestamp with time zone | 最初に登録が発生したときのタイムスタンプ |
| since_first_registered | interval | 最初の登録からのインターバル |

### System functions

主に SQL から呼び出すファンクションを使用して、PGD 管理を実行します。PGD のすべてのファンクションは bdr スキーマで公開されています。search_path に bdr を配置する代わりに、これらのファンクションへの呼び出しをスキーマ修飾します。

#### バージョン情報ファンクション

#### bdr.bdr_version

このファンクションは、現在使用している BDR 拡張機能のバージョンのテキスト表現を取得します。

#### bdr.bdr_version_num

このファンクションは、現在使用されている BDR 拡張機能のバージョン番号を取得します。バージョン番号は単調増加であるため、この値をより小さいおよびより大きい比較に使用できます。

次の式は、メジャーバージョン、マイナーバージョン、およびパッチリリースで構成されるバージョン番号を単一の数値に結果ます。

```
MAJOR_VERSION  *10000 + MINOR_VERSION*  100 + PATCH_RELEASE
```

### システム情報ファンクション

#### bdr.get_relation_stats

リレーション情報を返します。

#### bdr.get_subscription_stats

現在のサブスクリプション統計を結果ます。

### システムおよび進行状況情報パラメーター

PGD は、 SHOW または current_setting() ファンクションなどを使用して SQL で直接照会できるいくつかのパラメーターを公開します。クライアントアプリケーションから PQparameterStatus または同等のものを使用することもできます。

#### bdr.local_node_id

セッションを初期化すると、これはクライアントが接続しているノード ID に設定されます。これにより、アプリケーションは、透明なプロキシの背後にある場合でも、接続されているノードを認識できます。

接続プールとプロキシ でも使用されます。

#### bdr.last_committed_lsn

非同期トランザクションの COMMIT ごとに、このパラメーターは更新されて、オリジンノードのコミットレコードの最後を指します。bdr.wait_for_apply_queue と組み合わせると、アプリケーションはマルチプルのノードにわたってカウサル読み取りを実行できます。つまり、トランザクションがリモートで可視されるまで待機します。

#### transaction_id

CAMO トランザクションが進行中の場合、transaction_id が更新されて、割り当てられたトランザクション ID が表示されます。このパラメーターは、PQparameterStatus または同等の使用を使用してのみ照会でき、SQL ではアクセスできません。 アプリケーションの使用 を参照

使用例

### ノードステータスファンクション

#### bdr.is_node_connected

#### 概要

```
bdr.is_node_connected(node_name name)
```

指定されたピアのワルセンダーがこのノードでアクティブかどうかを確認することにより、ブール値を返します。

### bdr.is_node_ready

#### 概要

```
bdr.is_node_ready(node_name name, span interval DEFAULT NULL)
```

ラグが指定されたスパンより低いか、それ以外の場合は TO ASYNC の timeout より低いかどうかをチェックすることにより、ブール値を返します。

### コンセンサスファンクション

### bdr.consensus_disable

#### 概要

```
bdr.consensus_disable(node_group_name text DEFAULT NULL)
```

サーバーが再起動するか、bdr.consensus_enable を使用して再有効化されるまで、ローカルノードのコンセンサスワーカーを無効にします。最初に発生した方。

node_group_name パラメーターは、インターフェイスの一貫性のために受け入れられますが、現在プレースホルダーです。このファンクションは、データベース全体でコンセンサスワーカーを無効にし、渡された値に関係なく、ノード上のすべての Raft インスタンスに影響を与えます。

> **警告**
>
> コンセンサスを無効にすると、PGD の一部の機能が無効になり、長時間無効になっていると EDB Postgres 分散クラスターの可用性に影響します。この機能は、テクニカルサポートと協力する場合にのみ使用します。

### bdr.consensus_enable

#### 概要

```
bdr.consensus_enable(node_group_name text DEFAULT NULL)
```

ローカルノードで無効になったコンセンサスワーカーを再度有効にします。

node_group_name パラメーターは、インターフェイスの一貫性のために受け入れられますが、現在プレースホルダーです。このファンクションは、データベース全体でコンセンサスワーカーを再度有効にし、渡された値に関係なく、ノード上のすべての Raft インスタンスに影響を与えます。

### bdr.consensus_proto_version

ローカルノードが現在使用しているコンセンサスプロトコルのバージョンを返します。

PGD グループの再構成の内部メカニズムに必要です。

### bdr.consensus_snapshot_export

#### 概要

```
bdr.consensus_snapshot_export(version integer DEFAULT NULL, raft_instance_id integer␣
↪DEFAULT NULL)
```

ローカルノードの現在コミットおよび適用された状態から新しい PGD コンセンサススナップショットを生成し、bytea として結果ます。

デフォルトでは、サポートされている最新の Raft バージョンのスナップショットがエクスポートされます。ただし、明示的な version 番号を渡すことにより、それをオーバーライドできます。

raft_instance_id パラメーターは、スナップショットをエクスポートする Raft インスタンスを識別します。NULL デフォルトの場合、トップレベルの Raft インスタンスが使用されます。特定のノードグループの Raft インスタンスをターゲットにするには、そのインスタンス ID を渡します。これは、bdr.raft_instances を介してルックアップできます。

エクスポートするノードは現在の Raft リーダーである必要はなく、リーダーの最新の状態で完全に最新の状態である必要はありません。ただし、bdr.consensus_snapshot_import() は、このようなスナップショットを受け入れない場合があります。

新しいスナップショットは、ローカルノードの bdr.local_consensus_snapshot テーブルに自動的に保存されません。呼び出し元にのみ返されます。

生成されたスナップショットは、エクスポートノードの Raft ログ位置の背後にある同じ PGD ノードグループ内の他のノードの bdr.consensus_snapshot_import() に渡される場合があります。

この機能が動作するには、ローカル PGD コンセンサスワーカーを無効にする必要があります。一般的な使用法は次のとおりです。

```
SELECT bdr.bdr_consensus_disable();
COPY (SELECT * FROM bdr.consensus_snapshot_export()) TO /var/lib/postgresql/my_node_
↪consensus_snapshot.data;
SELECT bdr.bdr_consensus_enable();
```

PGD コンセンサスワーカーが無効になっている間

- ノードでの DDL ロックの試行は失敗するか、タイムアウトします。
- galloc シーケンスは新しい値を取得しません。
- Eage および CAMO トランザクションの一時停止またはエラー。
- 分散コンセンサスシステムを必要とする他の機能が中断されます。必要なダウンタイムは、一般に非常に短時間です。

ユースケースによっては、 bdr.local_consensus_snapshot テーブルの snapshot フィールドから既に存在するスナップショットを抽出し、代わりにそれを使用することが現実的である場合があります。その際、コンセンサスワーカーを停止する必要はありません。

### bdr.consensus_snapshot_import

#### 概要

```
bdr.consensus_snapshot_import(snapshot bytea, raft_instance_id integer DEFAULT NULL,␣
↪node_name bdr.regnode DEFAULT NULL)
```

bdr.consensus_snapshot_export() によって、通常は同じ PGD ノードグループ内の別のノードからエクスポートされたコンセンサススナップショットをインポートします。

別のノード上の bdr.local_consensus_snapshot テーブルの snapshot フィールドから直接抽出されたスナップショットを使用することもできます。

この機能は、破損またはユーザーエラーが発生した場合に、PGD ノードのカタログ状態を既知の良好な状態にリセットするのに役立ちます。

raft_instance_id パラメーターは、スナップショットをインポートする Raft インスタンスを指定します。NULL デフォルトの場合、トップレベルの Raft インスタンスが使用されます。特定のノードグループの Raft インスタンスをターゲットにするには、そのインスタンス ID を渡します。これは、bdr.raft_instances を介してルックアップできます。

node_name パラメーターは、スナップショットを特定のソースノード、通常はスナップショットがエクスポートされたリーダーノードに関連付けます。NULL デフォルトの場合、ソースノードは属性されません。

スナップショットが生成されたときに、インポートノードの apply_index がスナップショットエクスポートノードの commit_index 以下の場合、スナップショットをインポートできます。bdr.get_raft_status() を参照してください。ログが既に進みすぎているためにスナップショットを受け入れることができないノードは、エラーを発生させ、変更は行いません。スナップショットがインポートされると、ノードは現在のリーダーから残りの変更を取得するため、インポートされたスナップショットは完全に最新の状態である必要はありません。

この機能が動作するには、インポートノードで PGD コンセンサスワーカーを無効にする必要があります。詳細については、bdr.consensus_snapshot_export() のノートを参照してください。

この機能を使用して、次のことを実行してローカルノードに新しい Raft スナップショットを生成させることができます。

```
SELECT bdr.consensus_snapshot_import(bdr.consensus_snapshot_export());
```

このアプローチでは、現在適用されているログの位置まで Raft ログが切り捨てられる場合があります。

### bdr.consensus_snapshot_verify

#### 概要

```
bdr.consensus_snapshot_verify(snapshot bytea)
```

bdr.consensus_snapshot_export() によってエクスポートされた指定されたコンセンサススナップショットを確認します。スナップショットヘッダーには、生成されたバージョンが含まれており、ノードは同じバージョンに対して検証しようとします。

スナップショットは、クラスター内の同じノードまたは他のノードでエクスポートされた可能性があります。スナップショットを検証するノードがエクスポートされたスナップショットのバージョンをサポートしていない場合、エラーが発生します。

### bdr.get_consensus_status

現在のコンセンサス Raft ワーカーに関するステータス情報を結果ます。

### bdr.get_raft_status

現在のコンセンサス Raft ワーカーに関するステータス情報を結果ます。bdr.get_consensus_status のエイリアス。

### bdr.raft_leadership_transfer

#### 概要

```
bdr.raft_leadership_transfer(node_name text,
                             wait_for_completion boolean,
                             node_group_name text DEFAULT NULL)
```

node_name で指定されたノードに Raft リーダーになるように要求します。要求は任意の PGD ノードから開始でき、現在のリーダーに内部転送されて、指定されたノードにリーダーシップを転送します。指定されたノードは、完全な投票権を持つ ACTIVE PGD ノードである必要があります。

wait_for_completion が false の場合、要求はベストエフォートベースで提供されます。ノードが bdr.raft_global_lection_timeout 期間にリーダーになることができない場合、他の有効なノードが再びリーダーになります。また、リーダーシップは Raft プロトコルごとに時間の経過とともに変更される場合があります。true の戻り結果は、要求が正常に送信されたことのみを示します。

wait_for_completion が true の場合、ファンクションは指定されたノードが新しいリーダーになるまで待機し、要求されたノードが Raft リーダーになれなかった場合たとえば、ネットワークの問題のため、無限に待機します。したがって、無限ループを防ぐために、常に statement_timeout と wait_for_completion を設定することをお勧めします。

node_group_name はオプショナルであり、リーダーシップの移転が発生するノードグループの名前を指定するために使用できます。指定しない場合、デフォルトの NULL になり、クラスター内のトップレベルのグループとして解釈されます。node_group_name を指定すると、ファンクションは指定されたノードグループ内でのみリーダーシップを譲渡します。

#### ユーティリティファンクション

### bdr.transfer_all_leaderships

ローカルノードが現在書き込みまたは Raft リーダーシップを保持しているすべてのグループで、書き込みおよび Raft リーダーシップをローカルノードから移転します。この機能を使用して、ノードをオフラインにする前に、ノードからすべてのリーダーシップをドレインします。

#### 概要

```
bdr.transfer_all_leaderships(writeleader_timeout interval DEFAULT 10s)
```

#### パラメーター

| Name | Type | Default | Description |
|---|---|---|---|
| *writeleader_timeout* | interval | *'10s'* | 各書き込みリーダーシップ転送のタイムアウト。書き込みリーダーは自動選択され、このタイムアウトに到達するまで再試行されます。 |

#### 戻り値

すべての転送が成功した場合、またはノードがリーダーシップを保持していない場合は `true` を返し、書き込みリーダー転送がタイムアウトした場合は `false` を返します。

書き込みリーダー転送がタイムアウトすると、ファンクションは残りの書き込みリーダーと Raft 転送の処理を継続し、ベストエフォートベースで、他の転送が成功した場合でも `false` を結果ます。 Raft 転送は、`statement_timeout` 構成パラメーターを尊重して、リーダーシップが変更されるか、クエリがキャンセルされるまで、無期限に再試行されます。

ファンクションは `bdr.routing_leadership_transfer` および `bdr.raft_leadership_transfer` とロジックを共有し、1 回の呼び出しで両方の操作を完了して、結果が速くなります。

### bdr.alter_node_kind

PGD5 は、タスクマネージャーリーダーノードの概念を導入しました。ノードは PGD によって選択されますが、アップグレードされたクラスターの場合、クラスター内のすべてのノードで `node_kind` を適切に設定することが重要です。これは、最新の PGD バージョンにアップグレードした後、ノードごとに `bdr.alter_node_kind()` SQL ファンクションを呼び出すことにより手動で実行します。

#### 概要

```
bdr.alter_node_kind(node_name text,
                    node_kind text);
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_name* | 種類を変更するノードの名前。 |
| *node_kind* | ノードの種類。 |

### bdr.alter_subscription_skip_changes_upto

論理レプリケーションはバージョンをまたがってレプリケートでき、ロールのようなグローバル変更をレプリケートせず、選択的にレプリケートできるため、論理レプリケーション適用プロセスでエラーが発生し、変更の適用を停止する場合があります。

可能な場合は、ターゲット側に変更を加えてこのような問題を修正します。CREATE レプリケーションをブロックしている欠落テーブル、CREATE 必要なロール、GRANT 必要な権限など。ただし、その方法では問題を修正できない場合があり、トランザクションを完全にスキップする必要がある場合があります。変更はトランザクション全体としてスキップされますオールオアナッシング。スキップする場所を決定するには、次の例のように、ログ出力を使用してコミット LSN を見つけるか、論理デコード機能を使用して変更ストリームをピークにします。

トランザクションが 1 つの変更のみを行った場合を除き、多くの場合、ターゲット側でトランザクションのエフェクトを手動で適用する必要があります。そのため、以下の例に示すように、可能な限り問題のあるトランザクションを保存することが重要です。

pg_catalog.pg_logical_slot_get_binary_changes を使用して目的の LSN にスキップすることにより、bdr.alter_subscription_skip_changes_upto なしで変更をスキップすることが可能であるため、これは便利なファンクションです。論理デコードにおけるある種のエラーをバイパスする場合がありますが、より速いスキップを行います。

この機能は、無効になっているサブスクリプションでのみ動作します。

通常の手順のシーケンスは次のとおりです。

1. 問題のあるサブスクリプションと問題のあるコミットの LSN を特定します。
2. サブスクリプションを無効にします。
3. 可能であれば、ソースノードで pg_catalog.pg_logical_slot_peek_changes を使用してトランザクションのコピーを保存します。
4. ターゲットノード上の bdr.alter_subscription_skip_changes_upto 。
5. 必要に応じて、修復または同等の変更をターゲットに手動で適用します。
6. サブスクリプションを再度有効にします。

> **警告**
>
> このファンクションを使用すると、問題を悪化させるのは簡単です。それが唯一のオプションであると確信できない限り、何もしないでください。

#### 概要

```
bdr.alter_subscription_skip_changes_upto(
  subname text,
  skip_upto_and_including pg_lsn
);
```

### 例

トランザクションの適用はエラーで失敗し、ターゲット側の変更などのインパクトの低い修正ではこの問題を解決できないと判断しました。トランザクションをスキップする必要があると判断します。

エラーログで、次の例のように、スキップするコミットレコード LSN を見つけます。

```
ERROR:  XX000: CONFLICT: target_table_missing; resolver skip_if_recently_dropped␣
↪returned an error: table does not exist
CONTEXT:  during apply of INSERT from remote relation public.break_me in xact with␣
↪commit-end lsn 0/300AC18 xid 131315
committs 2021-02-02 15:11:03.913792+01 (action #2) (effective sess origin id=2 lsn=0/
↪300AC18)
while consuming I message from receiver for subscription bdr_regression_bdrgroup_
↪node1_node2 (id=2667578509)
on node node2 (id=3367056606) from upstream node node1 (id=1148549230, reporiginid=2)
```

ログのこの部分には、必要な情報があります。the_target_lsn: **0/300AC18** the_subscription: **bdr_regression_bdrgroup_node1_node2**

次に、サブスクリプションを無効にして、適用ワーカーがレプリケーションスロットに接続しようとしないようにします。

```
SELECT bdr.alter_subscription_disable(the_subscription);
```

トランザクションの一部だけをスキップすることはできません。それはオール・オア・ナッシングです。したがって、サブスクリプションのスロット名を使用して、最初にプロバイダー側でコピーして、レコードを保存することを強くお勧めします。

```
\copy (SELECT * FROM pg_catalog.pg_logical_slot_peek_changes(the_slot_name,
    the_target_lsn, NULL, min_proto_version, 1, max_proto_version, 1,
    startup_params_format, 1, proto_format, json))
 TO transaction_to_drop.csv WITH (FORMAT csv);
```

この例は、読みやすいように複数行に分割されていますが、単一行で発行します。\copy は、複数行のコマンドをサポートしていません。

peek を get に変更することにより、変更をスキップできますが、bdr....skip_changes_upto は、すべてのデータのデコードと出力を回避する高速なスキップを行います。

```
SELECT bdr.alter_subscription_skip_changes_upto(subscription_name,
    the_target_lsn);
```

ダンプされたトランザクションの内容をガイドとして使用して、同じ変更またはその修復バージョンをターゲットノードに手動で適用できます。

最後に、サブスクリプションを再度有効にします。

```
SELECT bdr.alter_subscription_enable(the_subscription);
```

### bdr.connection_manager_refresh_pools

このファンクションは、PGD 接続マネージャーに接続プールを更新するように指示します。

この機能は、他のツールとの統合、または接続プールの即時の再評価が必要な場合に役立ちます。この機能はローカルノードでのみ動作します。クラスター全体の接続マネージャーを更新するには、各ノードでファンクションを個別に実行する必要があります。実行には bdr_superuser ロールが必要です。コメントを展開する コメント オンライン R861 解決済み

#### 概要

```
SELECT bdr.connection_manager_refresh_pools();
```

> **注釈**
>
> このファンクションは、ルーティングの変更が検出された場合たとえば、書き込みリーダーの変更、ルーティングターゲットからのノードの追加/削除、またはコンセンサスが失われた場合、既存のクライアント接続を終了する場合があります。

### bdr.get_node_sub_apply_lsn

サブスクライバーでこのファンクションを使用して、特定のオリジンから受信および適用された最後の LSN を取得できます。

#### 概要

```
bdr.get_node_sub_apply_lsn(node_name name)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_name* | LSN が取得されるレプリケーションストリームのソースであるノードの名前。 |

### bdr.get_node_sub_receive_lsn

サブスクライバーでこのファンクションを使用して、特定のオリジンから受信した最後の LSN を取得できます。適用されるトランザクションに関連する LSN インクリメントのみを考慮して、フィルタリングしないか、フィルタリングすることができます。

このファンクションの出力と bdr.get_node_sub_apply_lsn() の出力の差は、対応する適用キューのサイズを測定します。

#### 概要

```
bdr.get_node_sub_receive_lsn(node_name name, committed bool default true)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_name* | LSN が取得されるレプリケーションストリームのソースであるノードの名前。 |
| *committed* | デフォルト true では、このファンクションは、最後の LSN 全体ではなく、受信したトランザクションのコミットのみを考慮します。これには、サブスクライバノードに影響を与えないアクションが含まれます。 |

### bdr.global_lock_table

このファンクションは、指定されたテーブルのグローバル DML ロックを取得します。*DDL locking details* を参照

グローバル DML ロックの詳細については、

#### 概要

```
bdr.global_lock_table(relation regclass)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *relation* | ロックへの関係の名前または oid。 |

#### 注

このファンクションは、ddl_locking 設定とは無関係にグローバル DML ロックを取得します。

bdr.global_lock_table ファンクションには、bdr.backwards_compatibility が 30618 以下に設定されていない限り、ロックされた relation に対する UPDATE 、DELETE 、または TRUNCATE 特権が必要です。

#### 例

```
pgddb=# SELECT bdr.local_group_slot_name();
 local_group_slot_name
- ----------------------
 bdr_pgddb_bdrgroup
```

### bdr.group_lease_override

ノードグループのリース所有権を手動で譲渡またはキャンセルします。通常の動作条件では、PGD ノードはリースを自動的に転送します。ただし、リース所有者がクラッシュ、停止、またはコンセンサスから切断された場合、他のノードは、現在のリースの有効期限が切れるまでリースを取得できません。このファンクションは、その待機期間をバイパスします。

> **危険**
>
> このファンクションは、以前のリース所有者がリソースにアクセスしていないことが確実な場合にのみ使用します。たとえば、古いリーダーがコンセンサスから切断されているがまだアクティブなときに分析リースを手動で転送すると、両方のノードが同じストアに書き込み、データの重複が発生する可能性があります。

#### パラメーター

| Parameter | Description |
|---|---|
| *node_group* | リースオーバーライドが適用されるグループの名前。 |
| *node* | リースを保持するノードの名前。 |
| *duration* | 秒単位のリース期間。`0`以下の値を指定すると、現在のリースがキャンセルされます。 |
| *kind* | リース種類たとえば *analytics* 。 |
| *local_only* | `false`デフォルトの場合、オーバーライドはコンセンサスを介してすべてのノードに送信されます。`true`の場合、オーバーライドはローカルノードにのみ影響します。コンセンサスが失われた場合にのみ`true`を使用します。 |

### bdr.local_group_slot_name

ローカルノードのグループスロットの名前を返します。

### bdr.node_group_type

指定されたノードグループのタイプを返します。戻り値は、ノードグループの作成時に `bdr.create_node_group()` に渡されたものと同じですが、グループの作成時に `node_group_type` が NULL として渡された場合、 `global` が返されます。

#### 例

```
pgddb=# SELECT bdr.node_group_type(bdrgroup);
 node_group_type
- ----------------
 global
```

### bdr.replicate_ddl_command

ノードのグループに DDL コマンドを複製する機能。

#### 概要

```
bdr.replicate_ddl_command(ddl_cmd text,
                          replication_sets text[],
                          ddl_locking text,
                          execute_locally bool)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *ddl_cmd* | 実行する DDL コマンド。 |
| *replication_sets* | `ddlcommand`を適用するレプリケーションセット名の配列。 NULL の場合、またはファンクションに `ddlcommand`のみが渡される場合、このパラメーターはアクティブな PGD グループのデフォルトレプリケーションセットに設定されます。 |
| *ddl_locking* | レプリケート中に *Replicated large objects* 値を設定する文字列。 `replicate_ddl_command`を実行しているローカルシステム上の `bdr.ddl_locking`の GUC 値のデフォルト。 |
| *execute_locally* | DDL コマンドをローカルで実行するかどうかを決定するブール値。デフォルト true。 |

#### 注

このファンクションの唯一の必須パラメーターは ddl_cmd です。

bdr.replicate_ddl_command() は常にコマンドをレプリケートし、*bdr.ddl_replication* の設定の影響を受けません。

。

### bdr.run_on_all_nodes

すべてのノードでクエリを実行するファンクション。

> **警告**
>
> このファンクションは、ノードの DSN で指定されたノード間接続に使用されるユーザーの権限を使用して、リモートノードで任意のクエリーを実行します。この機能に特権を付与するときは注意してください。

### 概要

```
bdr.run_on_all_nodes(query text)
```

### パラメーター

| Parameter | Description |
|---|---|
| *query* | 実行する任意のクエリー。 |

### 注

このファンクションは、他のノードに接続してクエリーを実行し、各ノードから JSON 形式で結果を返します。複数の行が各ノードから返され、JSON 配列としてエンコードされる場合があります。ノードがダウンしているため接続できないなどのエラーは、応答フィールドに表示されます。明示的な statement_timeout またはその他のランタイムパラメーターが設定されていないため、デフォルトが使用されます。

このファンクションは、通常のレプリケーションを経由しません。すべての既知のノードへの直接クライアント接続を使用します。デフォルトでは、コマンドはクラスター内のすべてのノードに既に送信されているため、`bdr.ddl_replication = off` で接続が作成されます。

PGD 6 以降では、このファンクションは接続で `bdr.xact_replication=off` も設定して、コマンドが別のノードで実行された場合にのみトランザクションがローカルで実行されるようにします。

レプリケーションが中断され、ノード間で不一致が発生する危険があるため、この機能を使用する場合は注意してください。透過的 DDL レプリケーションまたは `bdr.replicate_ddl_command()` を使用して DDL をレプリケートします。DDL は、将来のリリースでブロックされる可能性があります。

### 例

例、次のクエリのように、モニタリングでこのファンクションを使用すると便利です。

```
SELECT bdr.run_on_all_nodes($$
    SELECT local_slot_name, origin_name, target_name, replay_lag_size
      FROM bdr.node_slots
     WHERE origin_name IS NOT NULL
$$);
```

このクエリは、2 ノードクラスターで次のような結果を返します。

```
[
    {
        "dsn": "host=node1 port=5432 dbname=pgddb user=postgres ",
        "node_id": "2232128708",
        "response": {
            "command_status": "SELECT 1",
            "command_tuples": [
                {
```

(次のページに続く)

(前のページからの続き)


```
                    "origin_name": "node1",
                    "target_name": "node2",
                    "local_slot_name": "bdr_pgddb_bdrgroup_node2",
                    "replay_lag_size": "0 bytes"
                }
            ]
        },
        "node_name": "node1"
    },
    {
        "dsn": "host=node2 port=5432 dbname=pgddb user=postgres ",
        "node_id": "2058684375",
        "response": {
            "command_status": "SELECT 1",
            "command_tuples": [
                {
                    "origin_name": "node2",
                    "target_name": "node1",
                    "local_slot_name": "bdr_pgddb_bdrgroup_node1",
                    "replay_lag_size": "0 bytes"
                }
            ]
        },
        "node_name": "node2"
    }
]
```


### bdr.run_on_group

ノードのグループでクエリを実行する機能。

> **警告**
>
> このファンクションは、ノードの DSN で指定されたノード間接続に使用されるユーザーの権限を使用して、リモートノードで任意のクエリーを実行します。この機能に特権を付与するときは注意してください。

#### 概要

```
bdr.run_on_group(node_group_name text, query text)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_group_name* | クエリが実行されるノードグループの名前。 |
| *query* | 実行する任意のクエリー。 |

#### 注

このファンクションは、他のノードに接続してクエリーを実行し、各ノードから JSON 形式で結果を返します。複数の行を各ノードから返すことができ、JSON 配列としてエンコードされます。ノードがダウンしているため接続できないなどのエラーは、応答フィールドに表示されます。明示的な statement_timeout またはその他のランタイムパラメーターが設定されていないため、デフォルトが使用されます。

このファンクションは、通常のレプリケーションを経由しません。すべての既知のノードへの直接クライアント接続を使用します。デフォルトでは、接続は `bdr.ddl_replication = off` で作成され、同じレプリケートされた DDL コマンドが複数のノードに送信される場合のレプリケーションの問題を回避します。

PGD 6 以降では、このファンクションは接続で `bdr.xact_replication=off` も設定して、コマンドが別のノードで実行された場合にのみトランザクションがローカルで実行されるようにします。

レプリケーションが中断され、グループ内のノード間で不一致が発生する危険があるため、この機能を使用する場合は注意してください。グローバルスキーマ変更の場合、DDL をレプリケートするには、透過的 DDL レプリケーションまたは `bdr.replicate_ddl_command()` のいずれかを使用します。

### bdr.run_on_nodes

指定されたノードのリストでクエリを実行するファンクション。

> **警告**
>
> このファンクションは、ノードの DSN で指定されたノード間接続に使用されるユーザーの権限を使用して、リモートノードで任意のクエリーを実行します。この機能に特権を付与するときは注意してください。

#### 概要

```
bdr.run_on_nodes(node_names text[], query text)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_names* | クエリが実行されるノード名のテキスト ARRAY。 |
| *query* | 実行する任意のクエリー。 |

#### 注

このファンクションは、他のノードに接続してクエリーを実行し、各ノードから JSON 形式で結果を返します。複数の行を各ノードから返すことができ、JSON 配列としてエンコードされます。ノードがダウンしているため接続できないなどのエラーは、応答フィールドに表示されます。明示的な statement_timeout またはその他のランタイムパラメーターが設定されていないため、デフォルトが使用されます。

このファンクションは、通常のレプリケーションを経由しません。すべての既知のノードへの直接クライアント接続を使用します。デフォルトでは、接続は `bdr.ddl_replication = off` で作成され、同じレプリケートされた DDL コマンドが複数のノードに送信される場合のレプリケーションの問題を回避します。

PGD 6 以降では、このファンクションは接続で `bdr.xact_replication=off` も設定して、コマンドが別のノードで実行された場合にのみトランザクションがローカルで実行されるようにします。

レプリケーションが中断され、ノード間で不一致が発生する危険があるため、この機能を使用する場合は注意してください。グローバルスキーマ変更の場合、DDL をレプリケートするには、透過的 DDL レプリケーションまたは `bdr.replicate_ddl_command()` のいずれかを使用します。

### bdr.stat_reset

指定されたカテゴリの統計をリセットします。このファンクションを実行すると、*bdr.stat_reconcile_commit_scope* ビューの `stats_reset` 列の値がクリアされます。

`bdr_superuser` 権限が必要です。

#### パラメーター

| Parameter | Description |
|---|---|
| stats_tupe | リセットする統計のカテゴリ。有効な値は `commit_scope`および `reconcile_commit_scope`です。 `NULL`の場合、ファンクションのデフォルトは `commit_scope`になります。 |

### bdr.wait_for_apply_queue

ファンクション `bdr.wait_for_apply_queue` を使用すると、PGD ノードは、特定の PGD ノードから発生した特定のトランザクションのローカルアプリケーションを待機できます。そのピアノードからのすべてのトランザクションがローカルに適用された後にのみ返されます。アプリケーションまたはプロキシはこの機能を使用して、古い読み取りを防止できます。

便宜上、PGD は CAMO および CAMO パートナーノードにこのファンクションのバリアントを提供します。*bdr.wait_for_camo_partner_queue* を参照してください

。

特定の LSN が指定された場合、それは、ピアが待機するリカバリストリームのポイントです。これは、以前のまたは同時接続でそのピアノードから取得した bdr.last_committed_lsn と使用できます。

指定された target_lsn が NULL の場合、このファンクションはローカル受信バッファをチェックし、指定されたピアノードから受け取った最後のトランザクションの LSN を使用し、既に受信したすべてのトランザクションが適用されるのを効果的に待機します。これは、ピアノードに障害が発生し、どのトランザクションが送信されたかが不明な場合に特に役立ちます。この場合、送信者側でまだ転送中またはバッファリングされているトランザクションは待機されません。

#### 概要

```
bdr.wait_for_apply_queue(peer_node_name TEXT, target_lsn pg_lsn)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *peer_node_name* | 着信トランザクションがキューに入れられて待機するピアノードの名前。NULL の場合、すべてのピアノードの適用キューが使用されるまで待機します。 |
| *target_lsn* | 待機するピアノードからのレプリケーションストリームの LSN。通常、ピアノードから `bdr.last_committed_lsn`を介して学習します。 |

### bdr.wait_for_xid_progress

このファンクションを使用して、特定のノードノード ID で識別される特定のトランザクション XID で識別されるが、クラスターで十分な進行を行うのを待機できます。進行状況は、ノードに適用されているトランザクションとして定義され、このノードは、トランザクションが適用される前に行われた他のすべてのレプリケーション変更を確認しています。

#### 概要

```
bdr.wait_for_xid_progress(origin_node_id oid, origin_topxid int4, allnodes boolean␣
↪DEFAULT true)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *origin_node_id* | トランザクションが発生したノードのノード ID。 |
| *origin_topxid* | トランザクションの XID。 |
| *allnodes* | `true`の場合、すべてのノードでトランザクションが進行するまで待機します。それ以外の場合、現在のノードのみを待機します。 |

#### 注

現在、これらのトランザクションのみが追跡されるため、DDL コマンドをレプリケートしたトランザクションにのみファンクションを使用できます。誤った origin_node_id または origin_topxid が指定された場合、ファンクションは永久にまたは statement_timeout が発生するまで待機する場合があります。

### bdr.wait_node_confirm_lsn

ノードが特定の LSN を通過するまで待機します。

この機能を使用すると、このセッションの最後の書き込みが 1 つまたはすべてのノードで再生されるまで待機できます。

ファンクションは、呼び出されると、ノードが特定の LSN を渡すのを待機します。LSN が指定されない場合、現在の wal_flush_lsn pg_current_wal_flush_lsn() ファンクションを使用しての位置がローカルノードで使用されます。ノード名パラメーターを指定すると、そのノードが LSN を渡すのを待機するようにファンクションに指示します。ノード名が NULL を渡すことにより指定されない場合、ファンクションはすべてのノードが LSN を渡すまで待機します。

*Optimizing subscriber-only groups* を使用している場合、このファンクションの使用をお勧めします

*bdr.wait_slot_confirm_lsn* の代わりに

。

これは、最適化されたトポロジでは、すべてのノードにレプリケーションスロットがあるわけではないため、ファンクション bdr.wait_slot_confirm_lsn が予想通りに動作しない場合があるためです。bdr.wait_node_confirm_lsn は、レプリケーションスロットのないノードで動作するように設計されており、代替戦略を使用してノードの進行状況を判断します。

ノードが現在ダウンしているか、更新していない場合、または単に接続できない場合、待機は無期限に継続されます。この状態を回避するには、statement_timeout を待機する準備ができている最大時間に設定します。

#### 概要

```
bdr.wait_node_confirm_lsn(node_name text DEFAULT NULL, target_lsn pg_lsn DEFAULT NULL)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_name* | 待機するノードの名前。 NULL の場合、すべてのノードを待機します。 |
| *target_lsn* | 待機する LSN。 NULL の場合、ローカルノードで現在の wal_flush_lsn を使用します。 |

#### 注

使用するには bdr_application 権限が必要です。

### bdr.wait_slot_confirm_lsn

このセッションの最後の書き込みが 1 つまたはすべてのノードで再生されるまで待機できます。

スロットが特定の LSN を通過するまで待機します。位置が指定されない場合、現在の書き込み位置がローカルノードで使用されます。

スロット名が渡されない場合、すべての PGD スロットが LSN を通過するまで待機します。

ファンクションは、他のノードからの変更を 1000 ミリ秒ごとにポーリングします。

スロットが同時にドロップされた場合、そのスロットの待機は終了します。ノードが現在ダウンしており、スロットを更新していない場合、待機は継続されます。その場合、statement_timeout をより早く完了するように設定することができます。

*Optimizing subscriber-only groups* を使用している場合

、 bdr.wait_node_confirm_lsn の使用をお勧めします

代わりに。

#### 概要

```
bdr.wait_slot_confirm_lsn(slot_name text DEFAULT NULL, target_lsn pg_lsn DEFAULT NULL)
```

#### 注

使用するには bdr_application 権限が必要です。

#### パラメーター

| Parameter | Description |
|---|---|
| *slot_name* | 待機するレプリケーションスロットの名前。 NULL の場合、すべての PGD スロットを待機します。 |
| *target_lsn* | 待機する LSN。 NULL の場合、ローカルノードの現在の書き込み LSN を使用します。 |

### グローバルアドバイザリーロック

PGD は、グローバルアドバイザリーロックをサポートしています。これらのロックは、PGD でサポートされている勧告ロックがグローバルであることを除き、PostgreSQL で使用可能な勧告ロックと似ています。これらは DDL ロックと同様のセマンティクスに従います。したがって、アドバイザリロックはマジョリティコンセンサスによって取得され、すべてのノードの大部分が連携できる限り、1 つ以上のノードがダウンまたは遅延している場合でも使用できます。

現在、EXCLUSIVE ロックのみがサポートされています。したがって、別のノードまたは同じノード上の別のバックエンドがオブジェクトの勧告ロックを既に取得している場合、他のノードまたはバックエンドはロックが解放されるまで待機する必要があります。

アドバイザリーロックはトランザクション的な性質を持っています。したがって、トランザクションが終了する前に明示的に解放しない限り、トランザクションの終了時にロックは解放されます。この場合、リリースされるとすぐに利用可能になります。セッションレベルの勧告ロックは現在サポートされていません。

グローバル勧告ロックは再入可能です。したがって、同じリソースが 3 回ロックされている場合、3 回ロックを解除して、他のセッションで使用できるように解放する必要があります。

### bdr.global_advisory_lock

このファンクションは、提供されたオブジェクトの EXCLUSIVE ロックを取得します。ロックが利用できない場合、ロックが利用可能になるか、bdr.global_lock_timeout に到達するまで待機します。

#### 概要

```
bdr.global_advisory_lock(key bigint)
```

#### パラメーター

- key — アドバイザリーロックが取得されるオブジェクト。

#### 概要

```
bdr.global_advisory_lock(key1 integer, key2 integer)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *key1* | 複合キーの最初の部分。 |
| *key2* | 複合キーの 2 番目の部分。 |

### bdr.global_advisory_unlock

このファンクションは、アプリケーション定義のソースで以前に取得したロックを解放します。ロックは、アプリケーションによって同じトランザクションで取得されている必要があります。それ以外の場合、エラーが発生します。

#### 概要

```
bdr.global_advisory_unlock(key bigint)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *key* | アドバイザリロックが取得されるオブジェクト。 |

#### 概要

```
bdr.global_advisory_unlock(key1 integer, key2 integer)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *key1* | 複合キーの最初の部分。 |
| *key2* | 複合キーの 2 番目の部分。 |

### モニタリング機能

#### bdr.monitor_group_versions

クラスター全体のバージョンチェックを提供するために、このファンクションはビュー bdr.group_version_details から返された PGD バージョン情報を使用します。

#### 概要

```
bdr.monitor_group_versions()
```

#### 注

モニタリング パラレル 適用 で説明しているように、このファンクションは、フィールド status および message を持つレコードを結果ます。

このファンクションは bdr.run_on_all_nodes() を呼び出します。

#### bdr.monitor_group_raft

クラスター全体の Raft チェックを提供するために、このファンクションはビュー bdr.group_raft_details から返された PGD Raft 情報を使用します。

#### 概要

```
bdr.monitor_group_raft()
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_group_name* | 確認するノードグループ名。 |

#### 注

モニタリング パラレル 適用 で説明しているように、このファンクションは、フィールド `status` および `message` を持つレコードを結果ます。

このファンクションは `bdr.run_on_all_nodes()` を呼び出します。

### bdr.monitor_local_replslots

このファンクションは、ビュー `pg_replication_slots` スロットアクティブまたは非アクティブから返されたレプリケーションスロットステータス情報を使用して、PGD グループスロットを除くすべてのレプリケーションスロットを考慮したローカルチェックを提供します。

この機能は、*Optimizing subscriber-only groups* が有効になっているときに PGD クラスターでサブスクライバ専用グループリーダーとして動作しているサブスクライバ専用ノードのステータス情報も提供します。

#### 概要

```
bdr.monitor_local_replslots()
```

#### 注

このファンクションは、フィールド `status` および `message` を持つレコードを結果ます。

| Status | Message |
|---|---|
| UNKNOWN | This node is not part of any BDR group |
| OK | All BDR replication slots are working correctly |
| OK | This node is part of a subscriber-only group |
| CRITICAL | There is at least 1 BDR replication slot which is inactive |
| CRITICAL | There is at least 1 BDR replication slot which is missing |

レプリケーションスロットの監視 で詳細な説明が得られます。

### bdr.wal_sender_stats

*Decoding worker* が有効になっている場合、このファンクションは、デコーダースロットと、各 WAL 送信者によって読み取られている現在の論理変更レコード LCR セグメントファイルに関する情報を表示します。

#### 概要

```
bdr.wal_sender_stats()
```

#### 出力列

| Column name | Description |
|---|---|
| *pid* | WAL 送信者の PID。 `pg_stat_replication`の `pid`列に対応します。 |
| *is_using_lcr* | WAL 送信者が LCR ファイルを送信しているかどうか。 |
| *decoder_slot_name* | デコーダーレプリケーションスロットの名前。 |
| *lcr_file_name* | 現在の LCR ファイルの名前。 |

### bdr.get_decoding_worker_stat

*Decoding worker* が有効になっている場合、このファンクションは、現在のデータベースに関連付けられているデコードワーカーの状態に関する情報を表示します。これは、ワーカーの進行状況のデコードに関するより詳細な情報も提供します `pg_replication_slots` を介して利用可能な情報。

#### 概要

```
bdr.get_decoding_worker_stat()
```

#### 出力列

| Column name | Description |
|---|---|
| *pid* | デコードワーカーの PID。 `pg_replication_slots`の `active_pid`列に対応します。 |
| *decoded_upto_lsn* | デコードワーカーがトランザクションログを読み取る LSN。 |
| *waiting* | デコードワーカーが新しい WAL を待機しているかどうか。 |
| *waiting_for_lsn* | 次に予想される WAL の LSN。 |

#### 注

詳細については、 *LCR を使用した WAL 送信者のモニタリング* を参照してください。

### bdr.lag_control

ラグコントロール が有効になっている場合、このファンクションは、ローカルノードと現在のデータベースの構成された遅延対策に適合するノードの数とコミット遅延に関する情報を表示します。

### 概要

```
bdr.lag_control()
```

### 出力列

| Column name | Description |
|---|---|
| *commit_scope_id* | コミットスコープの OID :ref:``bdr.commit_scopes` 列<*bdr.commit_scopes* 列>` を参照してください。 |
| *sessions* | 遅延制御エントリを参照するセッションの数。 |
| *current_commit_delay* | 現在のランタイムコミット遅延ミリ秒単位。 |
| *maximum_commit_delay* | 構成された最大コミット遅延ミリ秒単位。 |
| *commit_delay_adjust* | サンプルインターバル中に発生する可能性のあるランタイムコミット遅延の変更ミリ秒単位。 |
| *current_conforming_nodes* | 遅延対策に適合しているノードの現在の実行時数。 |
| *minimum_conforming_nodes* | 遅延対策に適合するために必要なノードの構成済みの最小数。これを下回ると、コミット遅延調整が適用されます。 |
| *lag_bytes_threshold* | コミット遅延が適用される遅延サイズキロバイト単位。 |
| *maximum_lag_bytes* | キロバイト単位の構成された最大遅延サイズ。 |
| *lag_time_threshold* | ミリ秒単位のコミット遅延が適用される遅延時間。 |
| *maximum_lag_time* | 設定された最大遅延時間ミリ秒単位。 |
| *sample_interval* | ミリ秒単位の遅延サンプル間の構成された最小時間と考えられるコミット遅延調整。 |

## ルーティング機能

### bdr.routing_leadership_transfer

ルーティングリーダーを変更すると、ノードグループのリーダーシップが別のノードに転送されます。`leader_name` が `NULL` の場合、システムは書き込みリーダーの進行状況に基づいて、最も適切なノードを自動的に選択します。

### 概要

```
bdr.routing_leadership_transfer(node_group_name text,
               leader_name text DEFAULT NULL,
               transfer_method text DEFAULT strict,
               transfer_timeout interval DEFAULT 10s);
```

### パラメーター

| Name | Type | Default | Description |
|---|---|---|---|
| *node_group_name* | text | | リーダーシップの譲渡が要求されたグループの名前。 |
| *leader_name* | text | *NULL* | 書き込みリーダーになるノードの名前。 NULL の場合、システムは書き込みリーダーの進行状況に基づいて、最も適切な候補を自動的に選択します。 |
| *transfer_method* | text | *'strict'* | 転送の種類。 *'fast'* `または**デフォルトの**'strict'`で、最大遅延をチェックします。 |
| *transfer_timeout* | interval | *'10s'* | リーダーシップ転送のタイムアウト。デフォルトは 10 秒です。自動選択を使用する場合 `leader_name`が NULL の場合、システムはタイムアウトに達するまで、最も進行しているノードの選択を再試行します。 |

### 注

leader_name が NULL の場合、ファンクションは、書き込みリーダーが最も遠いノードを特定し、適切な候補が利用できるかタイムアウトに到達するまで選択を再試行することにより、最適な候補ノードを選択します。自動選択は、手動ノード選択が必要ない自動フェイルオーバーシナリオに役立ちます。コミットスコープ種類としてクォーラムコミットを使用する場合、leader_name に NULL を渡して、最も適切なノードの自動選択をトリガーします。

### 候補者の自動選択

自動リーダー選択時 leader_name が NULL の場合

- ルートフェンスされたノード *bdr.alter_node_option* を介して route_fence が有効になっているノードは、リーダーシップ譲渡の候補とは見なされません。
- route_writes が false に設定されているノードは、リーダーシップ移行の候補とは見なされません。
- route_priority オプションは自動選択に影響を与えません。選択は、適格ノード間の書き込みリーダーの進行状況にのみ基づいて行われます。

### 混合バージョンのクラスターでの実行

自動リーダー選択 leader_name の NULL を渡すには、PGD 6.4 および Raft プロトコルバージョン 6004 以上が必要です。

6. 4 より前のノードでファンクションを呼び出すと、古いファンクションシグネチャでは非 NULL leader_name パラメーターが必要であるため、ファンクションはアクションを実行せずに終了します。
7. 4 ノードでファンクションを呼び出す場合、動作は Raft プロトコルバージョンによって異なります。

- Raft プロトコルバージョンが 6004 未満で、PGD 6.4 以上のノードが Raft リーダーである場合、ファンクションは `ERROR: leadership auto transfer is supported from raft protocol version 6004 and above` を結果ます。
- 6.4 より前のノードが Raft リーダーの場合、オペレーションはタイムアウトし、`ERROR: switchover timed out` を結果ます。

信頼性の高い自動リーダー選択を行うには、クラスター内のすべてのノードが Raft プロトコルバージョン 6004 以上で PGD 6.4 以降を実行していることを確認します。

### CAMO ファンクション

CAMO、クライアントがトランザクションの進行状況を追跡することにより、トランザクションのコミットに積極的に参加する必要があります。ここにリストされているファンクションはそのために使用され、*Commit At Most Once* で説明されています。

### bdr.is_camo_partner_connected

ペアモードで構成された CAMO パートナーノードの接続ステータスを確認できます。現在、熱心なレプリケーションで使用される CAMO に相当するものはありません。

#### 概要

```
bdr.is_camo_partner_connected()
```

#### 戻り値

CAMO パートナーが現在ローカルノード上の WAL 送信プロセスに接続しているため、トランザクションデータを受信し、確認を送り返すことができるかどうかを示すブール値。

### bdr.is_camo_partner_ready

ペアモードで構成された CAMO パートナーノードの準備状況を確認できます。その下で、これはローカルモードへの、またはローカルモードからのスイッチをトリガーします。

#### 概要

```
bdr.is_camo_partner_ready()
```

#### 戻り値

CAMO パートナーがタイムリーに、つまり `TO ASYNC` の `timeout` の有効期限が切れる前に、ローカルノードから発生したトランザクションを確認することが合理的に期待できるかどうかを示すブール値。

> **注釈**
>
> このファンクションは、過去または現在の状態を照会します。正の戻り値は、 CAMO パートナーが将来のトランザクションを確認できるかどうかを示しません。

### bdr.get_configured_camo_partner

この機能は、ローカルノードの CAMO パートナーペアモードで構成を表示します。

#### 概要

```
bdr.get_configured_camo_partner()
```

### bdr.wait_for_camo_partner_queue

このファンクションは、CAMO パートナーノードを照会することがデフォルトの bdr.wait_for_apply_queue のラッパーです。ローカルノードが CAMO ペアの一部でない場合、エラーを返します。

#### 概要

```
bdr.wait_for_camo_partner_queue()
```

### bdr.camo_transactions_resolved

このファンクションは、CAMO トランザクションが完全に解決されるまでの待機を開始します。

#### 概要

```
bdr.camo_transactions_resolved()
```

### bdr.logical_transaction_status

ノードに障害が発生したときにコミットされていたトランザクションのステータスを確認するには、アプリケーションがこのファンクションを使用し、トランザクションが発生したノードのノード ID と元ノードのトランザクション ID をパラメーターとして渡す必要があります。

#### 概要

```
bdr.logical_transaction_status(node_id OID, xid OID,
                               require_camo_partner boolean DEFAULT true)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_id* | トランザクションが発生する PGD ノードのノード ID。通常、*Replicated large objects* `bdr.local_node_id`から `COMMIT`の前にクライアントによって取得されます。 |

次のページに続く

表 110 – 前のページからの続き

| Parameter | Description |
|---|---|
| *xid* | オリジンノードのトランザクション ID。通常、*COMMIT* から *:ref: `Replicated large objects <Monitoring through SQL>* `transaction_id`の前にクライアントによって取得されます。 |
| *require_camo_partner* | デフォルトで true で、構成チェックが有効になります。これらのチェックを無効にし、CAMO トランザクションではないトランザクションのステータスを照会するには、false に設定します。 |

### 戻り値

ファンクションは、次のいずれかの結果を返します。

- `'committed'::TEXT` — トランザクションはコミットされ、CAMO ペアの両方のノードで可視され、最終的に他のすべての PGD ノードにレプリケートされます。クライアントが再試行する必要はありません。
- `'aborted'::TEXT` —トランザクションは中止され、他の PGD ノードにレプリケートされません。クライアントは再試行するか、トランザクションのコミットの失敗をエスカレートする必要があります。
- `'in progress'::TEXT` —トランザクションはこのローカルノードでまだ進行中であり、まだコミットまたは中止されていません。トランザクションは COMMIT フェーズにあり、CAMO パートナーがコミットを確認または拒否するのを待機している可能性があります。推奨されるクライアントの反応は、元のノードから切断し、CAMO パートナーに再接続して、代わりにそれを照会することです。間にロードバランサーまたはプロキシがあり、クライアントがどのノードを照会するかを制御できない場合、クライアントはステータスが`'committed'` または`'aborted'` に切り替わるまで、繰り返しポーリングのみできます。

熱心な全ノードレプリケーションの場合、ピアノードはまだコミットまたは中止されていないトランザクションに対してこの結果を生成します。この場合、まだレプリケートされていないトランザクション、またはオリジンノードで開始されていないトランザクションでも、ピア PGD ノードで `in progress` 結果を生成する場合があります。ただし、クライアントは、オリジンでコミットする前にトランザクションステータスを照会してはなりません。

- `'unknown'::TEXT` —指定されたトランザクションは、将来であるか、その特定のノードにまだレプリケートされていない、または過去すぎるため不明です。このようなトランザクションのステータスはまだ知られていない、または不明です。この戻り値は、クライアントによる不適切な使用の兆候です。

クライアントは、エラーが発生した場合にファンクション呼び出しを再試行する準備をする必要があります。

### コミットスコープファンクション

#### bdr.add_commit_scope

**非推奨**。代わりに bdr.create_commit_scope を使用してください。以前は、この機能を使用してコミットスコープをノードグループに追加しました。現在は非推奨であり、将来のリリースで削除されるまで警告を発し、その時点でエラーが発生します。

### bdr.create_commit_scope

bdr.create_commit_scope は、指定されたコミットスコープ名とオリジンノードグループのルールを作成します。ルールが EDB Postgres 分散クラスター内のすべてのノードで同じである場合、トップレベルのノードグループに対してこのファンクションを 1 回呼び出すだけで、コミットスコープを完全に定義できます。

または、同じ commit_scope_name ではあるが、トランザクションのオリジンによって異なるオリジンノードグループとコミットスコープのルールを使用して、複数回呼び出すこともできます。

#### 概要

```
bdr.create_commit_scope(
    commit_scope_name NAME,
    origin_node_group NAME,
    rule TEXT,
    wait_for_ready boolean DEFAULT true)
```

#### 注

bdr.create_commit_scope は、非推奨の bdr.add_commit_scope ファンクションを置き換えます。add_commit_scope とは異なり、同じ名前が使用される場合、既存のコミットスコープをサイレントに上書きしません。代わりに、エラーが報告されます。

### bdr.alter_commit_scope

bdr.alter_commit_scope を使用すると、コミットスコープ内のシングルオリジンノードグループの特定のルールを変更できます。

#### 概要

```
bdr.alter_commit_scope(
    commit_scope_name NAME,
    origin_node_group NAME,
    rule TEXT)
```

### bdr.drop_commit_scope

コミットスコープ内の単一のルールを削除します。コミットスコープに複数のルールを定義する場合、ルールごとに 1 回このファンクションを呼び出して、コミットスコープ全体を完全に削除する必要があります。

#### 概要

```
bdr.drop_commit_scope(
    commit_scope_name NAME,
    origin_node_group NAME)
```

> **注釈**
>
> ノードグループによってデフォルトとして使用されているコミットスコープを削除することはできません。

### bdr.remove_commit_scope

**非推奨**。代わりに bdr.drop_commit_scope を使用してください。以前は、このファンクションはノードグループからコミットスコープを削除するために使用されていました。現在は非推奨であり、将来のリリースで削除されるまで警告を発し、その時点でエラーが発生します。

### エラーポリシーファンクションの適用

これらの機能を使用すると、適用エラーポリシーによって記録されたレプリケーション障害を確認、解決、および分析できます。構成の詳細については、*bdr.alter_node_group_option() を使用したグループ構成* の `apply_error_policy`、`apply_error_max_retries`、`apply_error_max_skips`、および `apply_error_max_record_size` オプションを参照してください。ポリシーの構成と障害を解決するためのステップバイステップガイドについては、*Handling replication apply errors* を参照してください。

### bdr.skip_failed_change

`bdr.pending_failed_changes` でキャプチャされた単一の変更をスキップしたとしてマーキングします。変更ステータスは `skipped` に更新されますが、行は削除されません。

#### 概要

```
bdr.skip_failed_change(change_id bigint)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *change_id* | `bdr.pending_failed_changes`のスキップする行の主キー。 |

### bdr.skip_failed_transaction

失敗したトランザクション内のすべての保留中の変更をスキップとしてマークし、オプションでサブスクリプションを再度有効にします。

#### 概要

```
bdr.skip_failed_transaction(
    subscription_name name,
    remote_xid xid,
    remote_commit_lsn pg_lsn,
    dry_run boolean DEFAULT false,
```

(次のページに続く)

(前のページからの続き)

```
    re_enable boolean DEFAULT false
)

bdr.skip_failed_transaction(
    failed_txn_id uuid,
    dry_run boolean DEFAULT false,
    re_enable boolean DEFAULT false
)
```

### パラメーター

| Parameter | Description |
|---|---|
| *subscription_name* | サブスクリプションの名前。 |
| *remote_xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | 元ノードで失敗したトランザクションのコミット LSN。 |
| *failed_txn_id* | *bdr.pending_failed_changes* `からの失敗したトランザクションの *UUID*。 `*subscription_name*、*remote_xid*、および `remote_commit_lsn`を指定するための便利な代替として使用します。 |
| *dry_run* | `true`の場合、何も変更せずにスキップされる変更の数を結果ます。デフォルトは `false`です。 |
| *re_enable* | `true`の場合、スキップした後にサブスクリプションを再度有効にします。デフォルトは `false`です。 |

### 戻り値

スキップされた変更の数を返します。トランザクションが既に解決されている場合、エラーは発生しない場合、`0` を返します。

## bdr.apply_failed_change

1 つのキャプチャされた変更を適用して再試行します。

### 概要

```
bdr.apply_failed_change(change_id bigint)
```

### パラメーター

| Parameter | Description |
|---|---|
| *change_id* | 適用する `bdr.pending_failed_changes`の行の主キー。 |

### bdr.apply_failed_transaction

失敗したトランザクションの保留中の変更をすべて再試行します。デフォルトでドライランモードになるため、コミットする前に結果をプレビューできます。

#### 概要

```
bdr.apply_failed_transaction(
    subscription_name name,
    remote_xid xid,
    remote_commit_lsn pg_lsn,
    dry_run boolean DEFAULT true,
    re_enable boolean DEFAULT false
)

bdr.apply_failed_transaction(
    failed_txn_id uuid,
    dry_run boolean DEFAULT true,
    re_enable boolean DEFAULT false
)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *subscription_name* | サブスクリプションの名前。 |
| *remote_xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | 元ノードで失敗したトランザクションのコミット LSN。 |
| *failed_txn_id* | *bdr.pending_failed_changes* からの失敗したトランザクションの *UUID*。 `*subscription_name*、*remote_xid*、および `remote_commit_lsn`を指定するための便利な代替として使用します。 |
| *dry_run* | `true`デフォルトの場合、実際に変更を適用せずに何が起こるかを示します最後にロールバックします。 `false`を渡して実行します。 |
| *re_enable* | `true`および `dry_run`が `false`の場合、適用後にサブスクリプションを再度有効にします。デフォルトは `false`です。 |

#### 戻り値

次の列を使用して、変更ごとに 1 行を結果ます。

| Column | Type | Description |
|---|---|---|
| *change_num* | *integer* | トランザクション内の変更の順序。 |
| *change_id* | *bigint* | `bdr.pending_failed_changes`の行の主キー。 |
| *operation* | *text* | 操作タイプ *INSERT*、*UPDATE*、*DELETE*、*SQL*、*DDL*、または`TRUNCATE` |
| *schema_name* | *text* | ターゲットテーブルのスキーマ。 |
| *table_name* | *text* | ターゲットテーブルの名前。 |
| *status* | *text* | 適用試行の結果。 Dry-run モードでは、値は`would_apply`、*would_skip*、または `would_error`です。実行時の値は`apply`、*skip*、または`error`です。 |
| *message* | *text* | 変更が失敗した場合はエラーメッセージ、成功した場合は`NULL`。 |

適用中にエラーが発生した場合、ファンクションは、失敗した変更を含むすべての変更の結果を返します。トランザクションが既に解決されている場合、ファンクションはエラーを発生させずに 0 行を返します。

### bdr.modify_failed_change

変更を再試行する前に、キャプチャされた単一の変更の列値を編集します。`bdr.pending_failed_changes` の `modified_values` 列を更新し、変更ステータスを `modified` に設定します。

#### 概要

```
bdr.modify_failed_change(change_id bigint, new_values jsonb)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *change_id* | 変更する`bdr.pending_failed_changes`の行の主キー。 |
| *new_values* | 変更が再試行されたときに使用する更新された列値を含む JSONB オブジェクト。 |

### bdr.reconstruct_change_sql

キャプチャされた単一の変更の実行可能 SQL を生成します。bdr.pending_failed_changes_sql ビューは、このファンクションを内部で使用します。

#### 概要

```
bdr.reconstruct_change_sql(
    operation text,
    schema_name name,
    table_name name,
    key_columns jsonb,
    old_values jsonb,
    new_values jsonb
)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *operation* | 操作タイプ。有効な値は`INSERT`、*UPDATE*、*DELETE*、*SQL*、*DDL*、および`TRUNCATE`です。 |
| *schema_name* | ターゲットテーブルのスキーマ。 |
| *table_name* | ターゲットテーブルの名前。 |
| *key_columns* | 主キー列値を含む JSONB オブジェクト。 |
| *old_values* | `UPDATE`および`DELETE`操作に使用される、ビフォアイメージ列値を含む JSONB オブジェクト。 |
| *new_values* | `INSERT`および `UPDATE`操作の after-image 列値、または `DDL`および`SQL`操作の SQL テキストを含む JSONB オブジェクト。 |

#### 戻り値

再構成された SQL ステートメントを含む text 文字列を結果ます。元の実行中に検索パスが有効であった場合、 DDL 操作には SET search_path 接頭辞が含まれます。

### bdr.reconstruct_failed_transaction

失敗したトランザクションでキャプチャされたすべての変更を対象に、BEGIN および COMMIT でラップされた完全な SQL スクリプトを生成します。

#### 概要

```
bdr.reconstruct_failed_transaction(
    subscription_name name,
```

(次のページに続く)

(前のページからの続き)

```
    remote_xid xid,
    remote_commit_lsn pg_lsn
)

bdr.reconstruct_failed_transaction(failed_txn_id uuid)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *subscription_name* | サブスクリプションの名前。 |
| *remote_xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | 元ノードで失敗したトランザクションのコミット LSN。 |
| *failed_txn_id* | *bdr.pending_failed_changes* からの失敗したトランザクションの *UUID*。 `*subscription_name*、*remote_xid*、および `remote_commit_lsn`を指定するための便利な代替として使用します。 |

#### 戻り値

トランザクション内のすべての変更を含む `text` SQL スクリプトを結果ます。スクリプトを使用して、完全なトランザクションを確認または手動で再生します。

### bdr.resolve_failed_transaction

失敗したトランザクションのすべての保留中の変更を解決済みとしてマークし、失敗したトランザクションを超えてサブスクリプションのポジションを進め、オプションでサブスクリプションを再度有効にします。デフォルトでドライランモードになるため、コミットする前に結果をプレビューできます。

#### 概要

```
bdr.resolve_failed_transaction(
    subscription_name name,
    remote_xid xid,
    remote_commit_lsn pg_lsn,
    resolution text DEFAULT skipped,
    re_enable boolean DEFAULT false,
    dry_run boolean DEFAULT true
)

bdr.resolve_failed_transaction(
    failed_txn_id uuid,
    resolution text DEFAULT skipped,
    re_enable boolean DEFAULT false,
    dry_run boolean DEFAULT true
)
```

### パラメーター

| Parameter | Description |
|---|---|
| *subscription_name* | サブスクリプションの名前。 |
| *remote_xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | 元ノードで失敗したトランザクションのコミット LSN。 |
| *failed_txn_id* | *bdr.pending_failed_changes* からの失敗したトランザクションの *UUID*。 `*subscription_name*、*remote_xid*、および `remote_commit_lsn`を指定するための便利な代替として使用します。 |
| *resolution* | 適用する解決アクション。有効な値は`skipped`、*applied*、および`discarded`です。デフォルトは`skipped`です。 |
| *re_enable* | `true`の場合、解決後にサブスクリプションを再度有効にします。デフォルトは`false`です。 |
| *dry_run* | `true`デフォルトの場合、変更を加えずにオペレーションが何を行うかの詳細なプレビューを結果ます。 `false`を渡して実行します。 |

### 戻り値

ドライランモードでは、計画的な変更を説明する `text` レポートを結果ます。 `dry_run` が `false` の場合、適用された解像度の `text` 確認を結果ます。

### 注

このファンクションを呼び出すと、概要レコードが `bdr.resolved_transactions` に挿入されます。解決したら、`bdr.pending_failed_changes` を切り詰めて、解決の履歴を保持しながら領域を再利用できます。

トランザクションが既に解決されている場合、ファンクションは以前の解決に関する詳細を記載したエラーを発生させます。

## bdr.analyze_skip_impact

単一のキャプチャされた変更をスキップした場合の安全性と潜在的なデータへの影響を分析します。エラータイプ、操作、およびオプションで現在のローカルデータベースの状態に基づいて、リスク評価を結果ます。

### 概要

```
bdr.analyze_skip_impact(
    change_id bigint,
    verify_state boolean DEFAULT true,
    include_transaction_analysis boolean DEFAULT true
)
```

### パラメーター

| Parameter | Description |
|---|---|
| *change_id* | 分析する `bdr.pending_failed_changes`の行の主キー。 |
| *verify_state* | *true* `デフォルトの場合、安全性を評価する前にローカルデータベースを照会して現在の状態を確認します。エラーコードと操作タイプのみに基づいて、より高速な静的分析を行うには *false* を渡します。 |
| *include_transaction_analysi* | `true`デフォルトの場合、同じトランザクション内の他のすべての変更を調べて、依存関係を検出し、警告を生成します。 |

### 戻り値

次の列を含む行を結果ます。

| Column | Type | Description |
|---|---|---|
| *skip_safety* | *text* | 安全性分類 *safe*、*caution*、または `risky` |
| *recommendation* | *text* | 人間が読める推奨事項。 |
| *explanation* | *text* | エラーコンテキストと状態検証結果を含む詳細な説明。 |
| *data_impact* | *text* | データインパクトレベル *none*、*possible_divergence*、または `certain_divergence`。 |
| *suggested_action* | *text* | 推奨されるアクション *skip*、*apply_modified*、または `manual_review`。 |
| *dependency_warnings* | *text[]* | 影響を受けるトランザクションの関連変更に関する警告の配列、またはない場合は `NULL`。 |

### 安全分類

| Safety | Description |
|---|---|
| *safe* | 変更をスキップしても、ノード間でデータの相違が発生する可能性は低いです。 |
| *caution* | 軽微な潜在的な問題が存在します。スキップする前に変更を確認します。 |
| *risky* | 変更をスキップすると、ノード間でデータの相違が発生する可能性があります。 |

### bdr.analyze_transaction_skip_impact

失敗したトランザクションの保留中のすべての変更のスキップ安全性を分析し、変更ごとに 1 行を結果ます。

#### 概要

```
bdr.analyze_transaction_skip_impact(
    subscription_name name,
    remote_xid xid,
    remote_commit_lsn pg_lsn
)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *subscription_name* | サブスクリプションの名前。 |
| *remote_xid* | 元ノードのトランザクション ID。 |
| *remote_commit_lsn* | 元ノードで失敗したトランザクションのコミット LSN。 |

#### 戻り値

トランザクション内の変更ごとに 1 行、次の列を返します。

| Column | Type | Description |
|---|---|---|
| *change_seq* | *integer* | トランザクション内の変更の順序。 |
| *operation* | *text* | 操作タイプ *INSERT*、*UPDATE*、*DELETE*、*SQL*、*DDL*、または `TRUNCATE` |
| *qualified_table* | *text* | スキーマ修飾テーブル名。 |
| *skip_safety* | *text* | 安全性分類 *safe*、*caution*、または `risky` |
| *recommendation* | *text* | 人間が読める変更の推奨事項。 |
| *has_dependents* | *boolean* | トランザクションの他の変更がこの変更の結果によって異なるかどうか。 |

### bdr.format_skip_anarise

キャプチャされた単一の変更のスキップセーフティのフォーマット化された人間が判読可能な分析を結果ます。

#### 概要

```
bdr.format_skip_analysis(change_id bigint)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *change_id* | 分析する `bdr.pending_failed_changes`の行の主キー。 |

#### 戻り値

安全性分類、推奨事項、説明、および状態検証結果を含む `text` レポートを返します。

### bdr.alter_subscription_reset_error_counters

サブスクリプションのエラーリトライおよびスキップカウンターを初期値にリセットします。レプリケーション障害を解決した後にリセットを使用して、サブスクリプションのエラーポリシーしきい値を新しく適用できるようにします。

#### 概要

```
bdr.alter_subscription_reset_error_counters(subscription_name name)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *subscription_name* | サブスクリプションの名前。 |

### PGD settings

PGD 固有の構成設定を設定できます。特に断りのない限り、値はいつでも設定できます。

#### 競合処理

### bdr.default_conflict_detection

新しく作成されるテーブルのデフォルトの競合検出方法を設定します。*bdr.alter_table_conflict_detection()* と同じ値を受け入れます。

#### グローバルシーケンスパラメーター

### bdr.default_sequence_kind

デフォルトの *Sequences* を設定します。

デフォルトは distributed で、これは、 snowflakeid が int8 シーケンスつまり bigserial に使用され、int4 つまり serial および int2 シーケンスの galloc シーケンスに使用されることを意味します。

### DDL 処理

#### bdr.default_replica_identity

新しく作成されるテーブルに REPLICA IDENTITY のデフォルト値を設定します。 REPLICA IDENTITY は、更新または削除された行を識別するためにログ先行書き込みに書き込まれる情報を定義します。

受け入れられる値は次のとおりです。

| Value | Description |
|---|---|
| *default* | 主キーの列の古い値がある場合は、それを記録しますこれはデフォルトの PostgreSQL 動作です。 |
| *full* | 行内のすべての列の古い値を記録します。 |
| *nothing* | 古い行に関する情報は記録しません。 |
| *auto* | テーブルは REPLICA IDENTITY FULL で作成されます。これはデフォルトの PGD 動作です。 |

詳細は、PostgreSQL documentation を参照してください。

PGD は、 PRIMARY KEY または UNIQUE 制約のないテーブルで UPDATE および DELETE オペレーションをレプリケートできません。例外は、テーブルのレプリカ ID がテーブル固有の構成または bdr.default_replica_identity による FULL の場合です。

bdr.default_replica_identity が default で、テーブル定義に UNIQUE 制約が含まれている場合、REPLICA IDENTITY として自動的に選択されません。 ALTER TABLE ... REPLICA IDENTITY ... を使用して REPLICA IDENTITY を明示的に設定する必要があります。

テーブルのレプリカ ID を full に設定すると、テーブルのワイヤに書き込まれる WAL の量とレプリケートされるデータの量が増加します。

!!!警告 bdr.default_replica_identity をデフォルトに設定する場合

ALTER SYSTEM を使用して bdr.default_replica_identity を default に設定する場合、次のように常に値を引用符で囲みます。

```
ALTER SYSTEM SET bdr.default_replica_identity="default";
```

デフォルトの引用符なしは、構成ファイルから設定の削除をトリガーする ALTER SYSTEM コマンドの特別な値であるため、引用符を含める必要があります。設定が削除されると、システムは PGD のデフォルト設定である auto を使用します。

ノード間で DDL を自動的にレプリケートしますデフォルトは on です。

このパラメーターは、bdr_superuser または superuser ロールによってのみ設定できます。

DDL を実行するか、 bdr.ddl_replication = off を使用して PGD 管理機能を呼び出すと、管理者が介入できるまでレプリケーションが停止する状況が発生する場合があります。詳細は、*Running DDL safely* を参照してください。

bdr.ddl_replication が off に設定されている場合は常に、 LOG レベルのログメッセージが PostgreSQL サーバーログに出力されます。さらに、キャプチャされた DDL コマンドまたは PGD レプリケーションファンクションのレプリケーションがこの設定によりスキップされるたびに、WARNING-level メッセージが書き込まれます。

### bdr.role_replication

ノード全体で ROLE コマンドを自動的にレプリケートしますデフォルトは on です。スーパーユーザーのみがこのパラメーターを設定できます。この設定は、bdr.ddl_replication もオンになっている場合にのみ機能します。

外部方法を使用せずにこのパラメーターをオフにして、すべてのノードでロールが同期していることを確認すると、管理者が介入するまでレプリケート DDL がレプリケーションを中断する場合があります。

*DDL and role manipulation statements* を参照してください

詳細については。

### bdr.ddl_locking

DDL のグローバルロックの操作モードを構成します。

このパラメーターは、bdr_superuser または superuser ロールによってのみ設定できます。

考えられるオプションは次のとおりです。

| Value | Description |
|---|---|
| *all* | すべての DDL 操作にグローバルロックを使用します。 |
| *leader* | リーダーベースのグローバル DML ロックを使用します。 |
| *auto* | 可能であれば`leader`を選択し、それ以外の場合は`all`を選択します。デフォルト |
| *dml* | リレーションのグローバル DML ロックを取得することにより書き込みを防止する必要がある DDL 操作にのみグローバルロックを使用します。 |
| *off* | DDL 操作にはグローバルロックを使用しないでください。 |

デフォルトは auto です。

bdr.ddl_locking が off に設定されている場合は常に、 LOG レベルのログメッセージが PostgreSQL サーバーログに出力されます。さらに、この設定によりグローバルロックステップがスキップされるたびに、WARNING メッセージが書き込まれます。一部のステートメントは、2 つの WARNING メッセージを生成します。1 つは DML ロックをスキップし、もう 1 つは DDL ロックをスキップします。

下位互換性のため、bdr.ddl_locking はエイリアスをサポートしています。 on および true は、all のエイリアスです。 false は、off のエイリアスです。

グローバルロック も参照してください。

### bdr.truncate_locking

TRUNCATE コマンドのロック動作を設定しますデフォルトは on/true です。 on/true の場合、TRUNCATE は bdr.ddl_locking 設定に従います。

### グローバルロック

DDL ロックは、*bdr.ddl_locking* によって制御されます。その他のグローバルロック設定には次のものが含まれます。

### bdr.global_lock_max_locks

ノードで保持できるグローバルロックの最大数を設定しますデフォルトは 1000 です。Postgres サーバー起動時にのみ設定できます。

### bdr.global_lock_timeout

グローバルロックの待機の最大許可時間を設定しますデフォルトは 1 分です。値 0 はこのタイムアウトを無効にします。

### bdr.global_lock_statement_timeout

グローバルロックを保持するステートメントの最大許可期間を設定しますデフォルトは 60 分です。値 0 はこのタイムアウトを無効にします。

### bdr.global_lock_idle_timeout

グローバルロックを保持するトランザクションのアイドル時間の最大許可期間を設定しますデフォルトは 10 分です。値 0 はこのタイムアウトを無効にします。

### bdr.lock_table_locking

LOCK TABLE ステートメントのロック動作を設定しますデフォルトはオンです。LOCK TABLE ステートメントは有効にすると、クラスターでグローバル DML ロックを取得し、他のロックステートメントをブロックします。

| Value | Description |
|---|---|
| *on* | すべてのテーブルロックにグローバルロックを使用します。 デフォルト |
| *off* | テーブルロックにグローバルロックを使用しないでください。 |

### bdr.predictive_checks

予測チェックのログレベルを設定します現在グローバルロックによってのみ使用されます。DEBUG 、LOG 、WARNING デフォルト、または ERROR です。予測チェックは、特定の操作を実行するときに予想されるクラスター状態の早期検証です。タイムアウトを待機するのではなく、早期フェイルのこれらの操作に使用できます。グローバルロック用語では、PGD は、接続されている十分なノードがあり、グローバルロックに必要なクォーラムを取得するための妥当な遅延制限内であることをチェックします。

### ノード管理

#### bdr.replay_progress_frequency

レプリケーション位置情報をクラスターの残りの部分に送信する間隔を設定しますデフォルトは 1 分です。

### 汎用レプリケーション

#### bdr.writers_per_subscription

サブスクリプションごとのライターのデフォルト数を設定します。PGD では、グループの `bdr.alter_node_group_option` を使用してこれを変更することもできます。

#### bdr.max_writers_per_subscription

サブスクリプションごとのライターの最大数 `bdr.writers_per_subscription` 設定の上限を設定します。

#### bdr.xact_replication

現在のトランザクションをレプリケートしますデフォルトは on です。

これをオフにすると、トランザクション全体がローカルのみになります。つまり、トランザクションは、PGD による論理デコードおよび論理デコードのその他のすべてのダウンストリームターゲットから可視されません。データは、ロジカルスタンバイノードを含む他のノードに転送されません。

このパラメーターは、bdr_superuser または superuser ロールによってのみ設定できます。

このパラメーターは、 `bdr.permit_unsafe_commands = on` を除き `SET LOCAL` コマンドを使用して現在のトランザクション内でのみ設定できます。

> **注釈**
>
> トランザクションレプリケーションを無効にしても WAL は生成されますが、これらの変更はオリジンでフィルタリングされます。

> **警告**
>
> *bdr.xact_replication* をオフにすると、ノード間でデータの不一貫性が発生します。ノード間のデータの相違からリカバリーするため、またはレプリケーションを続行するために単一ノードの変更が必要なレプリケーション状況にのみ使用します。自分の責任で使用してください。

#### bdr.permit_unsafe_commands

一般的な使用には安全でないと考えられるコマンドの安全性チェックをオーバーライドします。

bdr_superuser または PostgreSQL スーパーユーザーが必要です。

> **警告**
>
> 通常安全だと考えられていないコマンドは、一貫性のない結果を生成するか、レプリケーションを完全に中断する可能性があります。自分の責任で使用してください。

### bdr.batch_inserts

1 つのテーブルの挿入のバッチ処理をオンにする単一のトランザクションでの 1 つのテーブルへの連続した挿入の数。

この設定では、大規模なデータのレプリケーションを、挿入のセットとしてではなく、内部の COPY としてロードできます。これは、ノード参加中の初期データのコピー方法でもあります。

### bdr.maximum_lock_skew

`bdr.maximum_clock_skew_action` をトリガーする前に、受信トランザクションコミットタイムスタンプとサブスクライバーの現在時間の最大差を指定します。

現在再生されているトランザクションのタイムスタンプが、サブスクライバーの現在時刻と比較して将来的であるかどうかを確認します。その場合、その差が `bdr.maximum_clock_skew` より大きい場合、`bdr.maximum_clock_skew_action` 設定で指定されたアクションを実行します。

デフォルトは-1 で、これはクロックスキューを無視するチェックがオフになることを意味します。すべてのサーバーの時計が同期している場合、0 を設定するのが有効です。トランザクションが再生されているという事実は、過去にコミットされたことを意味します。

### bdr.maximum_lock_skew_action

`bdr.maximum_clock_skew` より高いクロックスキューが検出された場合に実行するアクションを指定します。

この設定には 2 つの値があります。

| Value | Description |
|---|---|
| *WARN* | このファクトに関する警告をログに記録します。警告は、サーバーログの氾濫を防ぐために、最大 1 分に 1 回ログに記録されます。 |
| *WAIT* | 現在のローカルタイムスタンプがリモートコミットタイムスタンプから`bdr.maximum_clock_skew`を引いたものより古くなくなるまで待機します。 |

### bdr.accept_connections

PGD への接続を有効または無効にしますデフォルトは on です。

bdr_superuser または PostgreSQL スーパーユーザーが必要です。

### bdr.writer_input_queue_size

レシーバーがライタプロセスにデータを送信するために使用する共有メモリキューのサイズを指定します。ライタプロセスが停止しているか、進行が遅い場合、キューがいっぱいになり、受信側プロセスも停止する可能性があります。したがって、このキューに十分な共有メモリを提供することが重要です。デフォルトは 1MB で、最大許可サイズは 1GB です。GUC の設定には任意のストレージサイズ指定子を使用できますが、デフォルトは KB です。

### bdr.writer_output_queue_size

レシーバーがライタプロセスからデータを受信するために使用する共有メモリキューのサイズを指定します。ライタは大量のデータを送信することは想定されていないため、比較的小さいサイズのキューで十分です。デフォルトは 32KB で、最大許可サイズは 1MB です。GUC の設定には任意のストレージサイズ指定子を使用できますが、デフォルトは KB です。

### bdr.min_worker_backoff_migrate

特定のワーカーが `bdr.min_worker_backoff_delay` ミリ秒ごとよりも頻繁に再起動されないようにすることにより、PGD バックグラウンドワーカー起動のレート制限が可能です。エラーが繰り返されると、バックオフが指数関数的に増加し、最大 `bdr.max_worker_backoff_delay` までジッターが追加されます。

時間単位のサフィックスがサポートされています。

> **注釈**
>
> この設定は現在レシーバーワーカーにのみ影響します。つまり、主にエラーまたは接続失敗時にサブスクリプションが再接続を試行する速度に影響を与えます。

`bdr.min_worker_backoff_delay` のデフォルトは 1 秒です。`bdr.max_worker_backoff_delay` の場合、1 分です。

バックオフ遅延設定が変更され、PostgreSQL 構成がリロードされると、現在のすべてのバックオフがリセットを待機します。さらに、`bdr.worker_task_reset_backoff_all()` ファンクションは、管理者がすべてのバックオフインターバルを強制的にすぐに有効期限が切れるように提供されています。

共有メモリ内の追跡テーブルは、各種類のワーカーの最終起動時刻を記憶するために維持されます。この追跡テーブルは永続的ではありません。これは、汚れたバックエンド終了後のクラッシュリカバリー中のソフトリスタートを含む、PostgreSQL の再起動によってクリアされます。

ビュー *bdr.worker_tasks* を使用してこの状態を検査できるため、管理者は、現在有効になっているバックオフレート制限を確認できます。

レート制限の目的で、ワーカーはタスクによって分類されます。このキーは、ワーカーロール、データベース OID、サブスクリプション ID、サブスクリプションライタ ID、拡張ライブラリ名とファンクション名、拡張機能提供のワーカー名、および同期ライターのリモートリレーション ID で構成されています。`NULL` は、特定の分類子が適用されない場合、たとえば、マネージャーワーカーがサブスクリプション ID がなく、レシーバーがライタ ID がない場合に使用されます。

### CRDT

#### bdr.crdt_raw_value

*CRDT* データ型を使用した列レベルの競合の処理 の出力フォーマットを設定します。

デフォルト出力この設定が `off` の場合、ベース CRDT タイプの現在値のみを結果ことができます、たとえば、`crdt_pncounter` の bigint。 on に設定されている場合、戻り値は CRDT 値の完全な表現を表します。これには、たとえば、マルチプルのノードからの状態を含めることができます。

### ラージオブジェクト

#### bdr.override_large_objects

on に設定すると、PostgreSQL の built-in large object functions `lo_create` 、`lo_open` 、`lo_read` などを同等の `bdr.lo_*` に透過的にリダイレクトするため、既存のアプリケーションはコードを変更せずにレプリケートされたラージオブジェクトを使用します。

リダイレクトは SQL レベルの呼び出しに適用されます。libpq ベースのクライアントとドライバーは、SQL をバイパスして、プロトコルレベルでラージオブジェクト呼び出しも発行します。これらもリダイレクトするには、Postgres に直接接続するのではなく、 *Connection Manager Authentication* 読み取り/書き込みポートデフォルト 6432 に接続します。

デフォルトは `off` です。詳細については、*Replicated large objects* を参照してください。

### コミットスコープ

#### bdr.commit_scope

現在またはデフォルトの *Commit scopes* を設定しますデフォルトは空の文字列です。

### Commit At Most Once

#### bdr.camo_local_mode_deploy

CAMO の非同期モードで適用され、CAMO パートナーがトランザクションを確認する必要があるときに通常発生するオーバーヘッドをエミュレートするコミット遅延デフォルトは 5 ミリ秒です。この機能を無効にするには、`0` に設定します。

#### bdr.camo_enable_client_warnings

CAMO プロパティを保証できないデータベースでアクティビティが実行された場合、警告を発しますデフォルトが有効になっています。十分な情報のあるユーザーは、この設定を無効にして、ログに記録される警告の量を減らすことができます。

### トランザクションストリーミング

#### bdr.default_streaming_mode

サブスクライバノードによるトランザクションストリーミングを制御します。使用可能な値は `off` 、`writer` 、`file` 、および `auto` です。デフォルト `auto` 。 `off` に設定されている場合、サブスクライバーはトランザクションストリーミングを要求しません。他の値に設定されている場合、サブスクライバーはトランザ

クションストリーミングを要求し、パブリッシャーはそれをサポートしている場合、およびグループレベルで構成されている場合にそれを提供します。詳細については、トランザクションストリーミングでの使用 を参照してください。

### 遅延制御

#### bdr.lag_control_max_commit_deploy

ミリ秒単位での許容できる最大許容コミット後遅延。

#### bdr.lag_control_max_lag_size

許容できる最大許容遅延サイズキロバイト単位。

#### bdr.lag_control_max_lag_time

ミリ秒単位の許容できる最大許容遅延時間。

#### bdr.lag_control_min_conforming_nodes

許容可能な遅延測定値を下回るように必要なノードの最小数。

#### bdr.lag_control_commit_deploy_adjust

最大コミット遅延時間の一部として測定されるコミット遅延のマイクロ調整。デフォルト値の 0.01%では、最大コミット遅延に到達するまでに 100 ネットインクリメントが必要です。

#### bdr.lag_control_sample_interval

ミリ秒単位の遅延サンプルとコミット遅延マイクロ調整間の最小時間。

#### bdr.lag_control_commit_deploy_start

コミット遅延インクリメントが適用され始める遅延しきい値。許容可能な遅延尺度の割合で表されます。デフォルト値の 1.0%では、許容可能な遅延測定値に違反するまでコミット遅延のインクリメントは開始されません。

より小さい分数を設定することにより、許容可能な遅延尺度に漸近するように早期に「遅延曲線を曲げる」ことにより、違反を防止できる場合があります。

### モニタリングとロギング

#### bdr.debug_level

PGD がデバッグメッセージを書き込むために使用するログレベルを定義します。デフォルト値は `debug2` です。詳細な PGD デバッグ出力を表示する場合は、`bdr.debug_level = 'log'` を設定します。

### bdr.trace_level

bdr.debug_level と同様に、PGD トレースメッセージに使用するログレベルを定義します。 EDB Postgres 分散クラスターのすべてのノードでトレースを有効にすると、EDB サポートが問題を診断するのに役立つ場合があります。このパラメーターは、Postgres サーバーの起動時にのみ設定できます。

!!!警告

bdr.debug_level または bdr.trace_level を値>= log_min_messages に設定すると、非常に大量のログ出力が生成される場合があります。ディスク領域の枯渇を防ぐためにログのフィルタリング、アーカイブ、およびローテーションの計画が策定されていない限り、運用環境では長期的に有効にしないでください。

トラックは、 bdr.stat_subscription デフォルトは on を使用した各サブスクリプションの統計を適用します。

### bdr.track_relation_apply

トラックは、 bdr.stat_relation デフォルトは off とのリレーションごとに統計を適用します。

### bdr.track_apply_lock_timing

bdr.stat_relation デフォルトは off とのリレーションの統計を追跡するときにロックタイミングを追跡します。

## デコードワーカー

### bdr.enable_wal_decoder

*Decoding worker* デフォルトは false を使用して、単一ノードで論理変更レコード LCR 送信を有効にします。true に設定すると、デコードワーカープロセスが開始され、WAL 送信側は生成した LCR を送信します。false に設定し直すと、LCR を使用する WAL 送信者が再起動され、WAL を直接使用します。

!!!注

また、PGD グループ内のすべてのノードでこの設定を有効にし、グループで enable_wal_decoder オプションを true に設定する必要があります。

別のノードをサブスクライブする場合、この設定により、ノードはサブスクリプションの論理変更レコード LCR の使用を要求できますデフォルトは false です。下流ノードでこの設定が true の場合、ノードは上流ノードに送信するときに LCR を使用することを要求します。ノードで bdr.enable_wal_decoder を true に設定する場合、この設定も true に設定します。

!!!注

また、PGD グループ内のすべてのノードでこの設定を有効にし、グループで enable_wal_decoder オプションを true に設定する必要があります。

論理変更レコード LCR ファイルのクリーンアップ間隔デフォルト 3 分。*Decoding worker* が有効になっている場合、デコードワーカーは LCR ファイルをバッファとして保存します。これらのファイルは定期的にクリーンアップされ、この設定は 2 つの連続するクリーンアップ間の間隔を制御します。ゼロに設定すると、クリーンアップが無効になります。

### 接続設定

以下は、すべてのクロスノード libpq 接続に影響を与える接続設定のセットです。デフォルトはかなり保守的な値に設定されており、ほとんどの運用ニーズをカバーします。すべての変数には SIGHUP コンテキストがあります。これは、リロード時に変更が適用されることを意味します。

#### bdr.global_connection_timeout

接続中に待機する最大時間秒単位でデフォルトは 15 秒です。10 進の整数、たとえば、10 として書き込みます。ゼロ、負の数、または指定なしは、無期限に待機することを意味します。最小許可タイムアウトは 2 秒であるため、値 1 は 2:raw-latex: **として解釈されます**。

#### bdr.global_keepalives

TCP キープアライブを使用するかどうかを制御しますデフォルトは 1、オンを意味します。キープアライブが不要な場合は、これを 0、つまりオフに変更できます。このパラメーターは、Unix ドメインソケットによって作成された接続では無視されます。

#### bdr.global_keepalives_idle

TCP がキープアライブメッセージをサーバーに送信するまでの非アクティブな秒数を制御しますデフォルトは 1 秒です。値 0 は、システムのデフォルトを使用します。このパラメータは、Unix ドメインソケットによって作成された接続、またはキープアライブが無効になっている場合に無視されます。TCP_KEEPIDLE または同等のソケットオプションが使用可能なシステムでのみサポートされています。他のシステムでは、効果はありません。

#### bdr.global_keepalives_interval

サーバーによって確認されなかった TCP キープアライブメッセージを再送信する秒数を制御しますデフォルトは 2 秒です。値 0 は、システムのデフォルトを使用します。このパラメータは、Unix ドメインソケットによって作成された接続、またはキープアライブが無効になっている場合に無視されます。TCP_KEEPINTVL または同等のソケットオプションが使用可能なシステムでのみサポートされています。他のシステムでは、効果はありません。

#### bdr.global_keepalives_count

サーバーへのクライアントの接続がデッドと見なされる前に、失われる可能性のある TCP キープアライブの数を制御しますデフォルトは 3 です。値 0 は、システムのデフォルトを使用します。このパラメータは、Unix ドメインソケットによって作成された接続、またはキープアライブが無効になっている場合に無視されます。TCP_KEEPCNT または同等のソケットオプションが使用可能なシステムでのみサポートされています。他のシステムでは、効果はありません。

#### bdr.global_tcp_user_timeout

接続が強制的に閉じられるまでに、送信されたデータが未確認のままでいられるミリ秒数を制御しますデフォルトは 5000、つまり 5 秒です。値 0 は、システムのデフォルトを使用します。このパラメーターは、Unix ドメインソケットによって作成された接続では無視されます。TCP_USER_TIMEOUT が利用可能なシステムでのみサポートされています。他のシステムでは、効果はありません。

### トポロジ設定

#### bdr.force_full_mesh

フルメッシュトポロジを強制しますデフォルトは on です。 `off` に設定されている場合、PGD はサブスクライバ専用グループに最適化されたトポロジを使用しようとします。この設定は、最適化されたトポロジの要件が満たされた場合にのみ有効です。詳細については、*Optimizing subscriber-only groups* を参照してください。

### アナリティクス設定

#### bdr.prefer_analytics_engine

データのローカルコピーが Postgres ヒープに存在する場合でも、システムがクエリを Postgres Analytics AcceleratorPGAA エンジン Seafowl または Spark にルーティングするかどうかを決定します。デフォルトは `false` です。

### 内部設定 - Raft タイムアウト

#### bdr.raft_global_election_timeout

ネットワーク障害を説明するために、Raft コンセンサスプロトコルは選択と要求のタイムアウトを実装します。この値は、要求がグローバルトップレベルグループに送信されるときに使用されます。デフォルトは 6 秒 6s です。

#### bdr.raft_group_election_timeout

ネットワーク障害を説明するために、Raft コンセンサスプロトコルは選択と要求のタイムアウトを実装します。この値は、要求がサブグループに送信されるときに使用されます。デフォルトは 3 秒 3s です。

#### bdr.raft_response_timeout

レスポンスの場合、 *bdr.raft_global_election_timeout* および *bdr.raft_group_election_timeout* の設定が必要に応じて使用されます。この変数を設定することにより、この動作をオーバーライドできます。`bdr.raft_response_timeout` の設定は、いずれの選択タイムアウト値よりも小さい必要があります。オーバーライドを無効にするには、この変数を-1 に設定します。デフォルトは-1 です。

### 内部設定 - その他の Raft 値

#### bdr.raft_keep_min_entries

ログ圧縮を行うときに Raft ログに保持するエントリの最小数デフォルトは `1000` 、PGD 5.3 以前は `100` です。 `0` の値は、ログ圧縮を無効にします。このパラメーターは、Postgres サーバーの起動時にのみ設定できます。

!!!警告

ログ圧縮が無効になっている場合、ログのサイズは永久に増加します。

ステートマシンを前進させるために、Raft は内部ログにエントリを追加します。通常オペレーションでは、追加には数ミリ秒しかかかりません。これにより、その追加アクションの期間に上限しきい値が設定され、

それを超えると INFO メッセージがログに記録されます。これは問題を示している可能性があります。デフォルトは 3000 ミリ秒です。

#### bdr.raft_log_min_message_duration

コンセンサス要求をログに記録する場合。PGD コンセンサス要求の往復時間を測定し、時間がこのパラメーターデフォルト 5000 ミリ秒を超えた場合、INFO メッセージをログに記録します。

#### bdr.raft_group_max_connections

Postgres サーバーのすべての PGD グループの最大接続数デフォルトは 100 接続です。これらの接続は、グループのノード間で PGD コンセンサス要求を伝送します。このパラメーターは、Postgres サーバーの起動時にのみ設定できます。

### 内部設定 - その他の値

#### bdr.backwards_compatibility

bdr.bdr_version_num 、30618 などで使用されるのと同じ数値形式で、下位互換性のあるバージョンを指定します。デフォルトは現在の PGD バージョンです。一般に不要な効果がある場合でも、以前の PGD バージョンの正確な動作を有効にします。これはリリースごとに異なるため、値が現在のバージョンと異なる場合を除き、構成ファイルで明示的に使用しないことをお勧めします。

#### bdr.track_replication_estimates

ピアノードの適用レートとキャッチアップ間隔の観点からレプリケーション推定を追跡します。CAMO のようなプロトコルは、この情報を使用して、ピアノードの準備状況を推定できます。このパラメーターはデフォルトで有効になっています。

#### bdr.lag_tracker_apply_rate_weight

PGD は、ローカルノードから WAL を適用するという点でピアノードがどれだけ遅れているかを監視し、遅延追跡の適用レートの移動平均を計算します。このパラメーターは、この移動平均計算で新しい計算値がどの程度寄与するかを指定します。デフォルトは 0.1 です。

#### bdr.enable_auto_sync_reconcile

PGD は、有効にすると、ノードがダウンしたときに最も先のノードからのデータを自動的に同期し、障害が発生したオリジンノードが残した未解決の準備済みトランザクションを自動的にリコンサイルします。デフォルト 6.0.1 からがオンになっています。

#### bdr.allow_bdr_schema_changes

bdr スキーマのデータベースオブジェクトの手動変更を許可します。デフォルトオフ。このパラメーターは、明示的に指示された場合にのみ変更します。

### Node management

#### ノードの状態のリスト

| State | Description |
|---|---|
| *NONE* | ワーカーが起動したときにノードの状態は設定解除され、現在の既知の状態にすぐに設定されることが予想されます。 |
| *CREATED* | `bdr.create_node()`が実行されましたが、ノードはまだ EDB Postgres 分散クラスターのメンバーではありません。 |
| *JOIN_START* | `bdr.join_node_group()`は、ローカルノードを既存の EDB Postgres 分散クラスターへの参加を開始します。 |
| *JOINING* | ノード参加が開始され、現在初期同期フェーズにあり、ノードにスキーマとデータを作成しています。 |
| *CATCHUP* | 初期同期フェーズが完了しました。現在、結合は、結合が開始されてから上流のピアノードで実行されたトランザクションを取得および適用する最後のステップにあります。 |
| *STANDBY* | ノード参加は完了しましたが、変更のブロードキャストはまだ開始されていません。すべての結合はこの状態でしばらく時間を費やしますが、ロジカルスタンバイとして定義されている場合、ノードはこの状態を継続します。 |
| *PROMOTE* | ノードはロジカルスタンバイであり、`bdr.promote_node`が呼び出されてノード状態を `ACTIVE`に移動したところです。これらの 2 つの `PROMOTE`状態は、1 つのノードのみが `STANDBY`より高いが `ACTIVE`より低い状態になることができるという事実と一貫している必要があります。 |
| *PROMOTING* | ロジカルスタンバイからフル PGD ノードへの昇格が進行中です。 |
| *ACTIVE* | ノードは完全な PGD ノードであり、現在 `ACTIVE`です。これは、最も一般的なノードのステータスです。 |
| *PART_START* | ノードは `ACTIVE`または `STANDBY`であり、EDB Postgres 分散クラスターからノードを削除するために `bdr.part_node`が呼び出されたばかりです。 |
| *PARTING* | ノードは他のノードから切断し、コンセンサスまたはレプリケーションでそれ以上の役割を果たしません。 |
| *PART_CATCHUP* | 分割されていないノードは、最近分割されたノードから欠落しているデータを同期します。 |
| *PART_CLEANUP* | 非分割ノードは、すべてのノードのグループスロットが PARTED ノードから発生したすべてのトランザクションに追いつくまで待機します。 |
| *PARTED* | これで、すべてのノードでノード分離操作が完了しました。 |
| *PARTED_CONCURRENT* | ノード分割操作がコンカレントオプションで開始されました。 |

`PROMOTE` または `PROMOTING` のいずれかのノードになることは、一度に 1 つのノードのみです。

`PARTED_CONCURRENT` 状態はターゲット状態として使用され、`concurrent = true` オプションで分割が行われたことを示します。

### ノード管理コマンド

PGD は、既存のノードの物理コピー pg_basebackup を使用してノードを PGD グループに追加するためのコマンドラインユーティリティも提供します。

### bdr_init_physical

!!!警告非推奨

このコマンドは、PGD グループ内のノードを作成および管理するより柔軟で強力な方法を提供する pgd CLI コマンド pgd node setup の使用を優先して非推奨です。bdr_init_physical は、将来的にバグ修正のみを受け取ります。新規インストールにはお勧めできません。

bdr_init_physical では、ソースノードと参加ノードの両方がまったく同じ PGD バージョンであることを必要とします。このコマンドを使用して、別の PGD バージョンのノードを既存のクラスターに参加させることはできません。これは、ノード参加方法を使用したローリングアップグレードに bdr_init_physical を使用できないことを意味します。ローリングアップグレードの場合、代わりに論理結合を使用します。詳細については、ノード参加を使用したローリングアップグレード を参照してください。

データディレクトリを指定する必要があります。このデータディレクトリが空の場合、pg_basebackup -X stream を使用して、高速なブロックレベルのコピー操作を使用してディレクトリを埋めます。

指定されたデータディレクトリが空でない場合、新しいノードのベースとして使用されます。最初は、キャッチアップを待機し、マスターノードに昇格してから、PGD グループに参加します。--standby オプションを使用する場合、ロジカルスタンバイノードに変わります。

このコマンドは、すべての PostgreSQL ネイティブ論理レプリケーションサブスクリプションをデータベースから削除または-S オプションを使用する場合を無効にします。また、レプリケーションオリジンとスロットも削除します。

#### 概要

```
bdr_init_physical [OPTION] ...
```

#### オプション

##### 一般オプション

- -D, --pgdata=DIRECTORY —新しいノードに使用するデータディレクトリ。空または存在しないディレクトリ、または pg_basebackup -X stream コマンド必須を使用して設定されたディレクトリのいずれかです。
- -l, --log-file=FILE — ロギングに FILE を使用します。デフォルトは bdr_init_physical_postgres.log です。
- -n, --node-name=NAME — 新しく作成したノードの名前必須。
- --replication-sets=SETS —使用するレプリケーションセット名のコンマ区切りリストの名前。指定しない場合、すべてのレプリケーションセットが使用されます。
- --standby —完全な送信/受信ノードではなく、ロジカルスタンバイ受信専用ノードを作成します。
- --node-group-name —参加するグループ。デフォルトはソースノードと同じグループです。

- -s, --stop — 初期化が完了したら、サーバーを停止します。
- -v — ロギングの詳細度を高めます。
- -L —空または存在しないデータディレクトリ -D オプションで使用する場合、選択的な pg_basebackup を実行します。これは EDB Postgres Extended Server のみの機能です。
- -S —論理レプリケーションサブスクリプションをドロップする代わりに、無効にします。

#### 接続オプション

- -d, --remote-dsn=CONNSTR — リモートノードの接続文字列必須。
- --local-dsn=CONNSTR — ローカルノードの接続文字列必須。

#### 構成ファイルのオーバーライド

- --hba-conf — 新しい pg_hba.conf へのパス。
- --postgresql-conf — 新しい postgresql.conf へのパス。
- --postgresql-auto-conf — 新しい postgresql.auto.conf へのパス。

#### 注

コマンドで指定されたレプリケーションセット名は、ノードが PGD グループに参加する前にデータディレクトリに存在するデータに影響を与えません。これは、bdr_init_physical が独自のベースバックアップを作成するか、既存のベースバックアップが新しい PGD ノードに昇格するかにかかわらず当てはまります。したがって、--replication-sets オプションは、ノードが PGD ノードグループに参加した後にパブリッシュおよびサブスクライブされたデータにのみ影響します。この動作は、bdr.join_node_group() を使用する場合など、論理結合でレプリケーションセットを使用する方法とは異なります。

オペレーターは、結合が完了した後、不要なテーブルを切り詰めることができます。bdr.tables カタログを参照して、レプリケーションセットのメンバーシップを判断し、サブスクライブされたレプリケーションセットのメンバーではないテーブルを特定します。次の理由から、テーブルを削除するのではなく、切り詰めることを強くお勧めします。

- DDL レプリケーションセットは行 DML レプリケーションセットと同じではないため、誤って他のノードでテーブルをドロップする可能性があります。
- 後でテーブルをレプリケーションセットに追加し、ノードの一部のサブセットにドロップした場合、レプリケーションセットに追加する前に、DDL 競合を作成せずにこれらのノードでのみ再作成する必要があります。

非レプリケートテーブルを切り捨て、存在するが空のままにする方が簡単かつ安全です。

### bdr_config

このコマンドラインユーティリティを使用すると、PGD インストールの構成を確認できます。これは、PostgreSQL に付属する pg_config ユーティリティに似ています。トラブルシューティングとサポートを支援するために使用できます。

### 概要

```
bdr_config [OPTION] ...
```

### オプション

| Option | Description |
|---|---|
| *--all* | 構成内のすべてのキーと値を表示します。 |
| *--version* | BDR バージョン関連のキーと値のみを表示します。これには、BDR 拡張機能のフルバージョン、それが実行されている Postgres バージョンとフレーバー、および BDRPG および BDR プラグイン API バージョンが含まれます。 |
| *--debug* | ビルド情報や機能の有効化を含む、 BDR デバッグキーと値のみを表示します。 |

### 例

```
$ /usr/lib/edb-as/16/bin/bdr_config --all
__OUTPUT__
BDR_VERSION_COMPLETE=5.6.0
BDR_VERSION_NUM=50600
PG_VERSION=16.4.1 (Debian 16.4.1~~snapshot11329862135.2980.1.88fbec6-1.bookworm)
PG_VERSION_NUM=160004
PG_FLAVOR=EPAS
BDRPG_API_VERSION_NUM=202309131
BDR_PLUGIN_API_VERSION=7011
USE_ASSERT_CHECKING=false
USE_VALGRIND=false
EXT_ENABLE_DTRACE=false
HAVE_LAG_CONTROL=true
HAVE_ASSESS_UPDATE_RI_HOOK=false
HAVE_BDRPG_PROBES=false
HAVE_CAMO=true
HAVE_DEADLOCK_DETECTOR_HOOK=true
HAVE_HEAP_UPDATE_HOOK=true
HAVE_LAG_TRACKER=true
HAVE_LCR=true
HAVE_LOG_TOAST_COLUMNS=false
HAVE_MISC_HOOKS=true
HAVE_MISSING_PARTITION_CONFLICT=true
HAVE_MULTI_PITR=false
HAVE_SELECTIVE_BASEBACKUP=false
HAVE_STREAMING_XACTS=true
HAVE_SYNC_COMMIT_HOOK=true
HAVE_TWOPHASE_DATA_HOOKS=true
```

(次のページに続く)

(前のページからの続き)

```
HAVE_XLOG_FIND_NEXT_RECORD=true
HAVE_DETACH_CONCURRENTLY=true
HAVE_ANALYTICS=true
```

### Node management interfaces

SQL インターフェイスを使用して、ノードを動的に追加および削除できます。

### bdr.alter_node_group_option

PGD ノードグループ構成を変更します。

#### 概要

```
bdr.alter_node_group_option(node_group_name text,
    config_key text,
    config_value text)
```

#### パラメーター

| Name | Description |
|---|---|
| *node_group_name* | 変更するグループの名前。 |
| *config_key* | 変更するノードグループのオプションのキー。 |
| *config_value* | 指定されたキーに設定する新しい値。 |

`config_value` は、オプションに適切なデータ型に解析されます。

この表は、この機能を使用して変更できるグループオプションを示しています。

| Name | Type | Description |
|---|---|---|
| *analytics_storage_format* | *text* | PGAA が分析データに使用する列状テーブル形式。サポートされている値は `iceberg`デフォルトまたは `delta`です。 |
| *analytics_storage_location* | *text* | 分析データファイルが保存される PGFS パス。主に PGAA 拡張機能を介してオブジェクトストレージに直接書き込む場合に使用されます。 |
| *analytics_write_catalog* | *text* | PGAA 拡張機能がテーブルメタデータを管理するために使用するカタログサービスの名前。 |

次のページに続く

表 134 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| *apply_delay* | *interval* | 着信変更を適用するまでノードが待機する時間。このオプションは、主に遅延加入者専用ノードを使用して特別なサブグループを設定する場合に役立ちます。これは、データノードを含むグループまたはトップレベルのグループでは設定しないでください。デフォルトは`0s`です。 |
| *apply_error_max_record_size* | *bigint* | *record_all_and_disable*`または`*record_all_and_skip*`ポリシーを使用する場合に、失敗したトランザクションごとに記録する変更データの最大バイト数。トランザクションがこの制限を超える場合、制限までででキャプチャされた変更のみが保存され、警告がログに記録されます。デフォルトは`*10485760* 10MB です。 |
| *apply_error_max_retries* | *integer* | エラーポリシーがアクションを実行するまでの再試行回数。*disable_after_retries*、*record_all_and_disable*、および`record_all_and_skip`ポリシーに適用されます。デフォルトは`3`です。 |
| *apply_error_max_skips* | *integer* | サブスクリプションを無効にする前に`record_all_and_skip`ポリシーがスキップできるトランザクションの最大数。無制限にスキップするには、`-1`または`NULL`に設定します。デフォルトは`NULL`です。 |
| *apply_error_policy* | *text* | このグループのサブスクリプションのレプリケーション適用エラーが処理される方法。有効な値は`keep_retrying`無期限に再試行し、*disable_after_retries* *apply_error_max_retries*`回まで再試行し、サブスクリプションを無効にする`*record_all_and_disable* キャプチャ専用モードで再生することによりトランザクション全体を記録し、サブスクリプションを無効にする、および`record_all_and_skip`全体を記録トランザクション、スキップして、レプリケーションを続行します。デフォルトは`keep_retrying`です。 |
| *check_constraints* | *boolean* | 適用プロセスがレプリケートデータを書き込むときに制約をチェックするかどうか。デフォルト値をそのまま使用することをお勧めします。そうしないと、データが失われる危険があります。有効な値は`on`または`off`です。デフォルトは`on`です。 |
| *default_commit_scope* | *text* | デフォルトで使用するコミットスコープ、最初は`local`コミットスコープです。このオプションは、トップレベルのノードグループにのみ適用されます。同じコミットスコープの異なるオリジングループに個々のルールを使用できます。詳細については、*Commit scope groups* を参照してください。 |

次のページに続く

表 134 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| *enable_fast_track_writer* | *boolean* | すべてのサブスクリプションの追加ライタが、クォーラムコミットによって生成された準備されたトランザクションを処理できるようにします。 Quorum Commit コミットスコープで正しく機能するために必要です。クォーラムコミットスコープが`default_commit_scope`として設定されている場合、自動的に有効になります。デフォルトは`off`です。 *Replicated large objects* を参照してください。 |
| *enable_routing* | *boolean* | ここで、グループ書き込みリーダーを介した :ref:``Connection Manager` </pgd/current/connection-manager/>` は、特定のグループで有効になっています。有効な値は`on`または`off`です。デフォルトは、サブグループの場合は`on`、クラスターグループの場合は`off`です。 |
| *enable_raft* | *boolean* | グループが独自の Raft コンセンサスを持つかどうか。このオプションは`enable_routing`〜`on`を設定するために必要です。トップレベルグループのこのオプションは常に`on`です。有効な値は`on`または`off`です。サブグループのデフォルトは`on`です。 |
| *enable_wal_decoder* | *boolean* | デコードワーカープロセスを有効/無効にします。`streaming_mode`が既に有効になっている場合、デコードワーカープロセスを有効にすることはできません。有効な値は`on`または`off`です。デフォルトは`off`です。 |
| *location* | *text* | グループの場所に関する情報。このオプションは、モニタリングのための純粋なメタデータです。デフォルトは`'"`空の文字列です。 |
| *num_writers* | *integer* | このノードグループをバッキングするサブスクリプションの並列ライターの数。有効な値は`-1`または正の整数です。 *-1* `は、*GUC :ref: `Replicated large objects <PGD settings>* で指定された値が使用されることを意味します。 `-1`がデフォルトです。 |
| *route_reader_max_lag* | *integer* | ノードが実行可能な読み取り専用ノードと考えられるバイト単位の最大遅延。将来の使用のために現在予約されています。 |
| *route_writer_max_lag* | *integer* | 書き込みリーダーとして選択される新しい書き込み候補のバイト単位の最大遅延。これを通過する候補者がない場合、ライタは選択されません。デフォルトは`-1`です。 |
| *route_writer_wait_flush* | *boolean* | PGD がフラッシュを待機する必要があるかどうかを切り替えるかどうか。将来の使用のために現在予約されています。 |

次のページに続く

表 134 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| *server_pool_mode* | *text* | 接続マネージャーの *Connection pooling* モード。有効な値は `none` プーリングなし、デフォルト、*session*、および `transaction` です。 |
| *streaming_mode* | *text* | 大規模なトランザクションのストリーミングを有効/無効にします。 *off* **`に設定すると、ストリーミングが無効になります。他の値に設定されている場合、大規模なトランザクションは進行中にデコードされ、変更がダウンストリームに送信されます。値が** *`file`* **に設定されている場合、ストリーミングトランザクションの着信変更はファイルに保存され、上流でトランザクションがコミットされた後にのみ適用されます。値が** *`writer`* **に設定されている場合、着信変更はライターのいずれか利用可能な場合に直接送信されます。** *:ref: `Parallel Apply<Parallel Apply>* が無効になっている場合、またはストリーミングトランザクションを処理できるライタがない場合、変更はファイルに書き込まれ、トランザクションがコミットされた後に適用されます。値が `auto` に設定されている場合、 PGD は、トランザクションプロパティと使用可能なリソースに応じて、*file* **`と** *`writer`* **`の間でインテリジェントに選択しようとします。** *WAL* **デコーダーが既に有効になっている場合、** *`streaming_mode`* **`を有効にすることはできません。デフォルトは** *`auto`* **`です。詳細については、** *:ref: `トランザクションストリーミングでの使用<トランザクションストリーミングでの使用>* を参照してください。 |
| *use_https* | *boolean* | HTTP リスナーが HTTPS を使用するかどうか。有効にすると、TLS にサーバー証明書が使用されます。 *postgresql.conf* **`またはコマンドラインを介して設定された** *`ssl_cert_file`* **`および** *`ssl_key_file`* **`が必要です。変更は、** *Connection Manager* **プロセスを再起動した後にのみ有効になります。デフォルトは** *`false`* **`です。詳細については、** *:ref: `Configuring Connection Manager<Configuring Connection Manager>* を参照してください。 |
| *failover_slot_scope* | *text* | PGD 5.7 以降のみ。論理スロットフェールオーバーサポートのスコープを設定します。有効な値は `global` または `local` です。デフォルトは `local` です。詳細については、 *CDC Failover support* を参照してください。 |

`apply_error_*` オプションの構成例と解決ワークフローについては、*Handling replication apply errors* を参照してください。

#### enable_fast_track_writer ノート

enable_fast_track_writer がオフで、アプリケーションが bdr.commit_scope 構成パラメーターを介してクォーラムコミットを使用する場合、トップレベルグループでこのオプションを明示的に有効にして、クラスター全体で有効にします。

```
SELECT bdr.alter_node_group_option(topgroup, enable_fast_track_writer, true);
```

高速トラックライタを追加すると、ノードごとに必要なレプリケーションスロットの数が増加します。*Postgres* **構成パラメーター** 、PostgreSQL 18 以降の場合は max_active_replication_origins の max_replication_slots 計算を確認します。このオプションを有効にした後、*bdr.get_min_required_replication_slots()* を使用して、値が十分に高いことを確認します。

#### 戻り値

bdr.alter_node_group_option() は、成功すると VOID を返します。

提供されたパラメーターのいずれかが無効な場合、 ERROR が発生します。

#### 注

ビュー *bdr.node_group_summary* を介して、ノードグループオプションの現在の状態を調べることができます。

このファンクションは、デフォルトを変更するように要求をグループコンセンサスメカニズムに渡します。行われた変更は、コンセンサスメカニズムを使用してグローバルに複製されます。

ファンクションはトランザクションではありません。要求はバックグラウンドで処理されるため、ファンクション呼び出しをロールバックすることはできません。また、変更は現在のトランザクションにすぐに表示されない場合があります。

このファンクションはロックを保持しません。

### bdr.alter_node_interface

指定されたノードの接続文字列 DSN を変更します。

#### 概要

```
bdr.alter_node_interface(node_name text, interface_dsn text)
```

#### パラメーター

| Name | Description |
|---|---|
| *node_name* | 変更する既存のノードの名前。 |
| *interface_dsn* | ノードの新しい接続文字列。 |

#### 注

このファンクションを実行し、ローカルノードでのみ変更を加えます。これは、通常、変更されるノードを含む、PGD グループ内のすべてのノードで実行することを意味します。

このファンクションはトランザクションです。ロールバックできます。変更は現在のトランザクションに表示されます。

ファンクションはローカルノードでロックを保持します。

### bdr.alter_node_option

ノードごとの構成オプションを変更します。

#### 概要

```
bdr.alter_node_option(node_name text,
    config_key text,
    config_value text);
```

#### パラメーター

| Name | Description |
|---|---|
| *node_name* | 変更するノードの名前。 |
| *config_key* | 変更するノードのオプションのキー。 |
| *config_value* | 指定されたキーに設定する新しい値。 |

この機能を使用して変更できるノードオプションは次のとおりです。

| Config Key | Description |
|---|---|
| *route_priority* | 同じノードグループ内の他のノードに対するノードの相対ルーティング優先度。デフォルトは `'-1'`です。 |
| *route_fence* | ノードがルーティングからフェンスされているかどうか。 true の場合、ノードは接続マネージャーから接続を受信できません。レプリケーションは影響を受けません。デフォルトは `'f'`false です。 |
| *route_writes* | 書き込みをこのノードにルーティングできるかどうか、つまりノードが書き込みリーダーになることができるかどうか。デフォルトは、データノードの場合は `'t'*true* **であり、他のノードタイプの場合は**'f'`false です。 |
| *route_reads* | 読み取り専用接続をこのノードにルーティングできるかどうか。将来の使用のために現在予約されています。デフォルトは、データおよびサブスクライバ専用ノードの場合は `'t'*true*、**監視ノードおよびスタンバイノードの場合は**'f'`false です。 |
| *route_dsn* | このノードへの接続に使用するプロキシの dsn。このオプションはオプショナル。設定されていない場合、デフォルトでノードの `node_dsn`値が使用されます。 |

次のページに続く

表 137 – 前のページからの続き

| Config Key | Description |
|---|---|
| *group_dsn* | 同じサブグループ内のノードの優先接続文字列。設定すると、同じサブグループ内のノードは、ノードの`node_dsn`の代わりにこの DSN を使用して接続します。デフォルトにリセットするには、NULL に設定します。 :ref:``bdr.node_config`列<*bdr.node_config* 列>` に `node_group_dsn`として保存されます。 |

### bdr.alter_subscription_enable

指定されたサブスクリプションまたはローカル PGD ノードのすべてのサブスクリプションを有効にします。これは、再開サブスクリプションとも呼ばれます。サブスクリプションが既に有効になっている場合、エラーはスローされません。このオペレーションの影響を受けるサブスクリプションの数を返します。

#### 概要

```
bdr.alter_subscription_enable(
    subscription_name name DEFAULT NULL,
    immediate boolean DEFAULT false
)
```

#### パラメーター

| Name | Description |
|---|---|
| *subscription_name* | 有効にするサブスクリプションの名前。 NULL の場合デフォルト、ローカルノードのすべてのサブスクリプションが有効になります。 |
| *immediate* | アクションをすぐに強制するために使用され、有効なサブスクリプションに関連付けられているすべてのワーカーを起動します。このオプションが`true`の場合、トランザクションブロック内でこのファンクションを実行できません。 |

#### 注

この機能はレプリケートされず、ローカルノードサブスクリプション特定のノードまたはすべてのノードにのみ影響します。

このファンクションはトランザクションです。ロールバックできます。現在のトランザクションは、カタログの変更を確認できます。サブスクリプションワーカーは、トランザクションがコミットされた後、バックグラウンドプロセスによって起動されます。

### bdr.alter_subscription_disable

指定されたサブスクリプションまたはローカル PGD ノードのすべてのサブスクリプションを無効にします。オプションで、無効なサブスクリプションに関連付けられているすべてのワーカーをすぐに停止することもできます。これは、サブスクリプションの一時停止とも呼ばれます。サブスクリプションが既に無効になっている場合、エラーはスローされません。このオペレーションの影響を受けるサブスクリプションの数を返します。

#### 概要

```
bdr.alter_subscription_disable(
    subscription_name name DEFAULT NULL,
    immediate boolean DEFAULT false,
    fast boolean DEFAULT true
)
```

#### パラメーター

| Name | Description |
|---|---|
| *subscription_name* | 無効にするサブスクリプションの名前。 NULL の場合デフォルト、ローカルノードのすべてのサブスクリプションが無効になります。 |
| *immediate* | アクションをすぐに強制し、無効になったサブスクリプションに関連付けられているすべてのワーカーを停止するために使用されます。このオプションが`true`の場合、トランザクションブロック内でこのファンクションを実行できません。 |
| *fast* | この引数は、`immediate`の動作に影響を与えます。 `true`デフォルトに設定されている場合、現在の作業が完了するのを待たずに、無効なサブスクリプションに関連付けられているすべてのワーカーを停止します。 |

#### 注

このファンクションはレプリケートされず、ローカルノードサブスクリプション特定のサブスクリプションまたはすべてのサブスクリプションにのみ影響します。

このファンクションはトランザクションです。ロールバックできます。現在のトランザクションは、カタログの変更を確認できます。ただし、サブスクリプションワーカーが停止するタイミングは、immediate の値に依存します。 true に設定されている場合、ワーカーは COMMIT を待たずに stop を受け取ります。fast 引数が true に設定されている場合、ワーカーの中断は、現在の作業が完了するまで待機しません。

### bdr.create_node

ノードを作成します。

#### 概要

```
bdr.create_node(node_name text,
                local_dsn text,
                node_kind DEFAULT NULL)
```

### パラメーター

| Name | Description |
|---|---|
| *node_name* | 新しいノードの名前。データベースごとに 1 つのノードのみが許可されます。有効なノード名は、小文字、数字、ハイフン、アンダースコアで構成されます。 |
| *local_dsn* | ノードへの接続文字列。 |
| *node_kind* | *data* `デフォルト、`*standby*、*subscriber-only*、または `witness`のいずれか。このパラメーターを設定しない場合、または `NULL`を提供する場合、デフォルトの `data`ノード種類が使用されます。 |

### 注

このファンクションは、関連するパブリック接続文字列を使用してローカルノードのレコードを作成します。ローカルレコードは 1 つだけであるため、作成されると、ファンクションは再度実行されるとエラーを報告します。

このファンクションはトランザクションファンクションです。ロールバックできます。これによる変更は、現在のトランザクションに表示されます。

ファンクションは、トランザクションが終了するまで、新しく作成されたノードのロックを保持します。

## bdr.create_node_group

PGD ノードグループを作成します。デフォルトでは、ローカルノードは唯一のメンバーとしてグループに参加します。bdr.join_node_group() を使用してグループにノードを追加できます。

### 概要

```
bdr.create_node_group(node_group_name text,
                      parent_group_name text DEFAULT NULL,
                      join_node_group boolean DEFAULT true,
                      node_group_type text DEFAULT NULL)
```

### パラメーター

| Name | Description |
|---|---|
| *node_group_name* | 新しい PGD グループの名前。ノード名と同様に、有効なグループ名は小文字、数字、アンダースコアのみで構成されます。 |
| *parent_group_name* | ノードサブグループを作成する場合、これは親グループの名前である必要があります。クラスターのメインノードグループを作成するときに `NULL`デフォルトを提供します。 |

次のページに続く

表 141 – 前のページからの続き

| Name | Description |
|---|---|
| *join_node_group* | ノードが作成中のグループに参加するかどうかを決定します。デフォルト値は`true`です。サブグループを作成するときに`false`を提供することは、たとえば、独立したリモートグループを作成する場合、ローカルノードが新しいグループに参加しないことを意味します。この場合、`parent_group_name`を指定する必要があります。 |
| *node_group_type* | 有効な値は`NULL`または`subscriber-only`です。 *NULL* **`デフォルトは、通常の汎用ノードグループを作成するためです。** *`subscriber-only`* **は、** *:ref:`Creating Subscriber-only groups and nodes<Creating Subscriber-only groups and nodes>*を作成するためのもので、そのメンバーはクラスター内の完全に参加したノードからのみ変更を受信しますが、他のノードに変更を送信することはありません。 |

#### 注

このファンクションは、ローカルノードで実行されているローカルコンセンサスワーカーに要求を渡します。

ファンクションはトランザクションではありません。グループの作成はバックグラウンドプロセスであるため、機能が完了すると、変更をロールバックすることはできません。また、変更は現在のトランザクションにすぐに表示されない場合があります。`bdr.wait_for_join_completion` を呼び出して、そうなるまで待機できます。

グループの作成では、ロックは保持されません。

### bdr.drop_node_group

空の PGD ノードグループを削除します。グループ内に参加したノードがある場合、ファンクションは失敗します。

#### 概要

```
bdr.drop_node_group(node_group_name text)
```

#### パラメーター

| Name | Description |
|---|---|
| *node_group_name* | ドロップする PGD グループの名前。 |

#### 注

このファンクションは、グループをドロップするように要求をグループコンセンサスメカニズムに渡します。ファンクションはトランザクションではありません。ドロッププロセスはバックグラウンドで発生し、ロールバックすることはできません。

### bdr.join_node_group

ローカルノードを既存の PGD グループに参加させます。

#### 概要

```
bdr.join_node_group (
    join_target_dsn text,
    node_group_name text DEFAULT NULL,
    wait_for_completion boolean DEFAULT true,
    synchronize_structure text DEFAULT all
)
```

#### パラメーター

| Name | Description |
|---|---|
| *join_target_dsn* | ローカルノードを追加する PGD グループ内の既存のソースノードへの接続文字列を指定します。 |
| *node_group_name* | PGD グループのオプションの名前。デフォルトは NULL で、ソースノードに存在する情報からグループ名を検出しようとします。 |
| *wait_for_completion* | 戻る前に、参加プロセスが完了するのを待ちます。デフォルト `true` |
| *synchronize_structure* | 結合中にデータベース構造スキーマの同期を実行するかどうかを指定します。デフォルト設定の `all` は、完全なデータベース構造を同期します。 `none` は、構造体を同期しません。ただし、データは引き続き同期されます。つまり、データベース構造が参加ノードに既に存在する必要があります。デザインとして、スキーマもデータも監視ノードに同期されないことに注意してください。 |

`wait_for_completion` が `false` として指定された場合、ファンクション呼び出しは、参加プロシージャが開始されるとすぐに返されます。結合の進行状況はログファイルと bdr.event_summary で確認できます。

情報ビュー。ファンクション bdr.wait_for_join_completion() を呼び出すことができます

`bdr.join_node_group()` の後、結合操作が完了するまで待機します。`verbose_progress` を `true` に設定して bdr.wait_for_join_completion() が呼び出された場合、進行状況情報を出力できます。

#### 注

このファンクションは、`join_target_dsn` 接続文字列がポイントするノードを介して、グループコンセンサスメカニズムに要求を渡します。行われた変更は、コンセンサスメカニズムによってグローバルに複製されます。

ファンクションはトランザクション的ではなく、トランザクションで実行されるとエラーを生成します。参加プロセスはバックグラウンドで発生し、ロールバックすることはできません。`wait_for_completion` が `true` に設定されている場合、または後で `bdr.wait_for_join_completion` を呼び出した場合、変更はローカルセッションにのみ可視されます。

ノードは 1 つのグループのみに所属できるため、このファンクションは各ノードで 1 回のみ呼び出すことができます。

ノード参加は、PGD グループのロックを保持しません。

分割処理の進行中のノード名を再利用するには、 bdr.part_node Raft プロトコルバージョン 6003 以降が必要を使用する場合に concurrent = true を設定します。このパラメーターを使用すると、分離ノードが PARTING 状態に到達すると、元の名前を再利用できます。この時点で、ノードの名前は自動的にその UUID に変更されます。

既存のノードが CAMO パートナーとして構成されているノードグループに追加のデータノードを参加させることはできません。ノードを追加するには、最初に CAMO コミットスコープをドロップします。

### bdr.part_node

PGD グループからノードを削除し、最終的にクラスター内のすべてのノードから分離されたノードのメタデータを削除します。

- ローカルノードの場合、リモートノードに関する情報を含むすべてのノードメタデータを削除します。
- リモートノードの場合、その特定のノードのメタデータのみを削除します。

この操作では、ノードからデータは削除されません。

削除するノードを含む、PGD グループ内の任意のアクティブノードからファンクションを呼び出すことができます。

グループの default_commit_scope またはローカルノードの bdr.commit_scope 構成パラメーターが CAMO コミットスコープに設定されている場合、データノードを分割することはできません。ノードを削除するには、最初に CAMO コミットスコープをドロップします。

削除されるノードから分割を実行すると、削除のステータスが誤って報告される、または全く報告されないリスクがあります。これは、削除の過程で、通信がクラスターの残りの部分から切断されるためです。削除は成功する場合がありますが、コマンドを発行したノードに、他のノードで失敗または成功したことを通知する方法はありません。同じ理由で、ファンクションは完了を待機するように設定できません。

ノードが自分自身を分離すると、クラスターの一部ではなくなるため、クラスター内の他のノードを分離することはできません。

ノードを使用してクラスターから自分自身を分離することは避けることをお勧めします。代わりに、完了を待機できるノードからノード分割操作を実行し、操作が完了した後にクラスターステータスを確認します。

> **注釈**
>
> ローカルノードを分離する場合、*wait_for_completion* を `false` に設定する必要があります。それ以外の場合、エラーが報告されます。

> **警告**
>
> このアクションは永続的です。ノードへのレプリケーションを一時的に停止する場合は、*bdr.alter_subscription_disable()* を使用します。

### 概要

```
bdr.part_node (
    node_name text,
    wait_for_completion boolean DEFAULT true,
    force boolean DEFAULT false,
    concurrent boolean DEFAULT false
)
```

### パラメーター

| Name | Description |
|---|---|
| *node_name* | パートへの既存のノードの名前。 |
| *wait_for_completion* | `true`の場合、ノードがクラスターから完全に分離されるまで、ファンクションは返されません。それ以外の場合、ファンクションは分割プロシージャを開始し、待機せずにすぐに戻ります。ローカルノードで実行する場合、または`force`を使用する場合、常に`false`に設定します。 |
| *force* | ローカルノードでノードを強制的に削除します。これにより、コンセンサスに到達できない場合、またはノード分割プロセスがスタックしている場合、ノードの状態をローカルに設定します。 |
| *concurrent* | `true`の場合、ノードが`PARTED`状態に到達するのを待って最終的にドロップされることなく、ノードが`PARTING`状態以上になるとすぐに、分割ノード名の再利用ができます。デフォルトは`false`です。このパラメーターは、クラスターが Raft プロトコルバージョン 6003 以上の場合に使用できます。 |

**警告**

*force* = *true* を使用すると、PGD グループが一貫性のない状態になる場合があります。他の方法でノードを削除できない障害から回復するためにのみ使用します。

### 注

このファンクションは、グループコンセンサスメカニズムに要求を渡して、特定のノードを分割します。行われた変更は、コンセンサスメカニズムによってグローバルに複製されます。分離プロセスはバックグラウンドで発生し、ロールバックすることはできません。`wait_for_completion` が `true` に設定されている場合、分割プロセスによって行われた変更はローカルトランザクションにのみ可視されます。

`force` が `true` に設定されている場合、コンセンサスが失敗すると、このファンクションはローカルノードでのみ特定のノードの状態を設定します。このような場合、ファンクションはトランザクションであり、ファンクションがノードの状態を変更するため、ロールバックできます。`force` が `true` に設定された状態で既に分離処理中のノードでファンクションが呼び出された場合、指定されたノードをローカルに分離されたものとしてマークして、終了します。これは、クラスターでコンセンサスに到達できない場合つまりノードの大部分がダウンしている場合、または分割プロセスがスタックしている場合にのみ役立ちます。

ただし、書き込みを受信していた分割ノードが書き込みを行った場合、実際にはスタックせずに分割プロ

セスに時間がかかる場合があることを考慮することが重要です。他のノードは、特定のノードからの欠落データを再同期する必要があります。他のノードは、すべてのノードのグループスロットが PARTED ノードから発生したすべてのトランザクションに追いつくまで、待機する必要があります。

強制的な分離はこの再同期を完全にスキップし、他のノードを一貫性のない状態のままにする可能性があります。

分離プロセスはロックを保持しません。

concurrent = true の場合、パーティングノードのターゲット状態は PARTED_CONCURRENT に設定されます。 PARTING 状態に到達すると、ノードの名前がその UUID に変更されます。プロセスは、ノードが PARTED 状態に到達し、ドロップされるまで続行されます。したがって、ノードのすべての追跡は、PART_START 状態を通過すると、その UUID を使用する必要があります。

concurrent = true を使用する場合は、wait_for_completion = false を設定することをお勧めします。この構成により、元のノードが PARTING 状態に到達すると、ファンクションは Notice で戻ることができます。この瞬間、ノードの名前はその UUID に変更され、元の名前が解放され、クラスター内の即時の再割り当てまたは再利用が可能です。

### bdr.promote_node

ローカルロジカルスタンバイノードを PGD グループの完全なメンバーに昇格させます。

#### 概要

```
bdr.promote_node(wait_for_completion boolean DEFAULT true)
```

#### 注

このファンクションは、デフォルトを変更するように要求をグループコンセンサスメカニズムに渡します。行われた変更は、コンセンサスメカニズムによってグローバルに複製されます。

ファンクションはトランザクションではありません。プロモーションプロセスはバックグラウンドで発生し、ロールバックすることはできません。wait_for_completion が true に設定されている場合、または後で bdr.wait_for_join_completion を呼び出した場合、変更はローカルトランザクションにのみ可視されます。

プロモーションプロセスは、他のプロモーションに対してロックを保持します。このロックは他の bdr.promote_node 呼び出しをブロックしませんが、プロモーションのバックグラウンドプロセスが一度に複数のノードで進行することを防ぎます。

### bdr.switch_node_group

ローカルノードを現在のサブグループから同じ既存の PGD ノードグループ内の別のサブグループに切り替えます。

### 概要

```
bdr.switch_node_group (
    node_group_name text,
    wait_for_completion boolean DEFAULT true
)
```

### パラメーター

| Name | Description |
|---|---|
| *node_group_name* | PGD グループまたはサブグループの名前。 |
| *wait_for_completion* | 戻る前に、スイッチプロセスが完了するのを待ちます。デフォルト `true` |

`wait_for_completion` が `false` に設定されている場合、これは非同期呼び出しであり、スイッチングプロシージャが開始するとすぐに返されます。スイッチの進行状況は、ログと `bdr.event_summary` 情報ビューで、または `bdr.switch_node_group()` が戻った後に `bdr.wait_for_join_completion()` ファンクションを呼び出すことで確認できます。

### 注

このファンクションは、要求をグループコンセンサスメカニズムに渡します。行われた変更は、コンセンサスメカニズムによってグローバルに複製されます。

ファンクションはトランザクションではありません。スイッチングプロセスはバックグラウンドで発生し、ロールバックすることはできません。`wait_for_completion` が `true` に設定されている場合、または後で `bdr.wait_for_join_completion` を呼び出した場合、変更はローカルトランザクションにのみ可視されます。

ローカルノードは、クラスターを分割する必要なく、現在のサブグループから同じ PGD ノードグループ内の別のサブグループにメンバーシップを変更します。ノードの種類は、ターゲットサブグループ内の既存のノードの種類と一致する必要があります。

ノードスイッチングは、PGD グループのロックを保持しません。

制限現在、この機能では、サブグループとその PGD ノードグループ間のスイッチングのみが許可されています。サブグループ間の移動を有効にするには、2 つの別個の呼び出しを行う必要があります。1) サブグループからノードグループにスイッチし、2) ノードグループから他のサブグループにスイッチします。

## bdr.sync_node_cancel

このファンクションは、指定されたオリジンおよびソースノードの同期要求をキャンセルします。

### 概要

```
bdr.sync_node_cancel(origin text, source text)
```

#### パラメーター

| Name | Description |
|---|---|
| *origin* | 元ノードの名前。 |
| *source* | ソースノードの名前。 |

#### 注

このファンクションは、指定されたオリジンとソースを持つすべてのターゲットのすべての同期ノード要求をキャンセルします。書き込みリードからのみ呼び出すことができます。

### bdr.wait_for_join_completion

このファンクションは、ローカルノードの参加手順が完了するのを待機します。

#### 概要

```
bdr.wait_for_join_completion(verbose_progress boolean DEFAULT false)
```

#### パラメーター

| Name | Description |
|---|---|
| *verbose_progress* | オプションで、参加プロシージャ中に実行される個々のステップに関する情報を出力します。 |

#### 注

このファンクションは、ローカルノードのチェック状態が `bdr.create_node_group`、`bdr.join_node_group`、または `bdr.promote_node` で設定されたターゲット状態に到達するまで待機します。

### Routing functions

### Commit scopes

コミットスコープは、PGD システム内でトランザクションのコミットと競合を処理する方法を決定するルールです。それらの詳細については、*Commit scopes* を参照してください。

次のファンクションを使用して、コミットスコープを操作できます。

- bdr.create_commit_scope
- bdr.alter_commit_scope
- bdr.drop_commit_scope

### コミットスコープの構文

コミットスコープルールの全体的な文法は次のように構成されます。

```
commit_scope:
    commit_scope_operation [AND ...]

commit_scope_operation:
    commit_scope_group confirmation_level commit_scope_kind

commit_scope_target:
    { (node_group [, ...])
      | ORIGIN_GROUP }

commit_scope_group:
{ ANY num [NOT] commit_scope_target
  | MAJORITY [NOT] commit_scope_target
  | ALL [NOT] commit_scope_target }

confirmation_level:
    [ ON { received|replicated|durable|visible } ]

commit_scope_kind:
{ GROUP COMMIT [ ( group_commit_parameter = value [, ... ] ) ] [ ABORT ON ( abort_on_
↪parameter = value ) ] [ DEGRADE ON (degrade_on_parameter = value [, ... ] ) TO␣
↪commit_scope_degrade_operation ]
  | QUORUM COMMIT [ ( commit_decision = { group | raft } ) ] ABORT ON ( timeout =␣
↪value )
  | CAMO [ DEGRADE ON ( degrade_on_parameter = value [, ... ] ) TO ASYNC ]
  | LAG CONTROL [ ( lag_control_parameter = value [, ... ] ) ]
  | SYNCHRONOUS COMMIT [ DEGRADE ON (degrade_on_parameter = value ) TO commit_scope_
↪degrade_operation ] }

commit_scope_degrade_operation:
   commit_scope_group confirmation_level commit_scope_kind
```

node_group は、PGD データノードグループの名前です。

### commit_scope_degrade_operation

commit_scope_degrade_operation は、定義されるルール全体としてより制限の少ないコミットスコープグループと同じコミットスコープ種類であるか、非同期 ASYNC のいずれかです。

例、 *Degrading commit scope rules* ALL SYNCHRONOUS COMMIT から MAJORITY SYNCHRONOUS COMMIT 、または MAJORITY SYNCHRONOUS COMMIT から ANY 3 SYNCHRONOUS COMMIT 、または ANY 3 SYNCHRONOUS COMMIT から ANY 2 SYNCHRONOUS COMMIT 。 SYNCHRONOUS COMMIT から ASYNC に低下することもできます。ただし、関係するコミットスコープグループに関係なく、SYNCHRONOUS COMMIT から GROUP COMMIT またはその逆に低下することはできません。

AND を使用して、それぞれが独自の低下句を持つルールを組み合わせることもできます。

```
ALL ORIGIN GROUP SYNCHRONOUS COMMIT DEGRADE ON (timeout = 10s) TO MAJORITY ORIGIN␣
↪GROUP SYNCHRONOUS COMMIT AND ANY 1 NOT ORIGIN GROUP SYNCHRONOUS COMMIT DEGRADE ON␣
↪(timeout = 20s) TO ASYNC
```

#### コミットスコープターゲット

##### ORIGIN_GROUP

特定のグループをターゲットとする代わりに、トランザクションが発生する最下位グループを動的に参照する ORIGIN_GROUP を使用することもできます。したがって、トップレベルグループ top_group と、子として 2 つのサブグループ left_dc および right_dc がある場合、次のようなコミットスコープを追加します。

```
SELECT bdr.create_commit_scope(
    commit_scope_name := example_scope,
    origin_node_group := top_level_group,
    rule := MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT,
    wait_for_ready := true
);
```

left_dc のノードで発生したトランザクションの場合、 left_dc のノードの大部分は、トランザクションがコミットされる前に同期的に確認する必要があることを意味します。さらに、同じルールは、right_dc のノードから発生したトランザクションの場合、 right_dc のノードの大部分が、コミットされる前にトランザクションを同期的に確認する必要があることも意味します。これにより、left_dc 用と right_dc 用の 2 つの別個のルールをコミットスコープに追加する必要がなくなります。

#### コミットスコープグループ

##### 任意

例 ANY 2 (left_dc)

このコミットスコープグループのトランザクションは、 left_dc グループ内の 2 つのノードがトランザクションを処理したことを確認した後、コミットされたとみなされます。

##### ANY NOT

例 ANY 2 NOT (left_dc)

left_dc グループにない 2 つのノードがトランザクションを処理したことを確認した場合、このコミットスコープグループのトランザクションはコミットされたとみなされます。

#### マジョリティ

例 MAJORITY (left_dc)

left_dc グループ内のノードの大部分がトランザクションを処理したことを確認した場合、このコミットスコープグループのトランザクションはコミットされたとみなされます。

#### マジョリティではない

例 `MAJORITY NOT (left_dc)`

`left_dc` グループにないノードの大部分がトランザクションを処理したことを確認した場合、このコミットスコープグループのトランザクションはコミットされたとみなされます。

#### すべて

例 `ALL (left_dc)`

`left_dc` グループ内のすべてのノードがトランザクションを処理したことを確認した場合、このコミットスコープグループのトランザクションはコミットされたとみなされます。

`ALL` を 同期コミットとグループコミット とともに使用する場合、リコンシリエーションの問題を回避するには、*bdr.commit_Decisions* 設定を `raft` に設定する必要があります。

#### すべてではない

例 `ALL NOT (left_dc)`

`left_dc` グループにないすべてのノードがトランザクションを処理したことを確認した場合、このコミットスコープグループのトランザクションはコミットされたとみなされます。

#### 確認レベル

確認レベルは、リモート PGD ノードがトランザクションの処理中に特定のポイントに到達したことを確認する時点を設定します。`QUORUM COMMIT` は、`durable` および `visible` のみをサポートしています。

#### ON を受信しました

トランザクションは、ローカルアプリケーションを開始する前に、受信後すぐに確認されます。

#### ON が複製されました

トランザクションは、トランザクションの変更を適用した後、ディスクにフラッシュする前に確認されます。

#### ON 耐久性

トランザクションは、すべての変更がディスクにフラッシュされた後に確認されます。

#### ON 可視

これはデフォルトの可視性です。トランザクションは、すべての変更がディスクにフラッシュされ、同時トランザクションに可視された後に確認されます。

### コミットスコープの種類

コミットスコープの種類の詳細とそのパラメーターの詳細

- 同期コミット
- クォーラムコミット

- *CAMO (Commit At Most Once)*
- ラグコントロール
- 同期コミットとグループコミット

Postgres GUC parameter value conventions を使用して、Boolean、enum、int、および interval 値を指定します。

```
SYNCHRONOUS COMMIT [ DEGRADE ON (degrade_on_parameter = value ) TO commit_scope_
↪degrade_operation ]
```

### DEGRADE ON パラメーター

| Parameter | Type | Default | Description |
|---|---|---|---|
| *timeout* | interval | 0 | ミリ秒単位のタイムアウト他の単位を受け入れます。その後、オペレーションが低下します。 0 は設定されていないことを意味します。 |
| *require_write_lead* | Boolean | False | ノードが縮退動作に切り替えることができるように書き込みリードである必要があるかどうかを指定します。 |

これらは、コミットスコープルールがより制限の低いオペレーションモードに低下する条件を設定します。

### commit_scope_degrade_operation

`commit_scope_degrade_operation` は、より制限の少ないコミットスコープグループの `SYNCHRONOUS COMMIT` であるか、非同期 `ASYNC` である必要があります。

### クォーラムコミット

ノードがコミットする前に、すべての参加ノードにわたってコミット決定を調整することにより、分散トランザクションの一貫性を提供します。

詳細については、 クォーラムコミット を参照してください。

```
QUORUM COMMIT [ ( commit_decision = { group | raft } ) ] ABORT ON ( timeout = value )
```

サポートされているコミットスコープグループは、`ALL` 、`MAJORITY ORIGIN GROUP` 、`MAJORITY CLUSTER` 、および `MAJORITY FOR EACH GROUP` です。明示的なグループリストを使用した `MAJORITY` 、 `ANY n` 、および `NOT` はサポートされていません。

`QUORUM COMMIT` では、 `durable` および `visible` 確認レベルのみがサポートされています。 `received` または `replicated` を使用するとエラーになります。

### QUORUM COMMIT パラメーター

| Parameter | Type | Default | Description |
|---|---|---|---|
| *commit_decision* | enum | group | コミットの決定がどのように行われるかを指定します。 *group* I *raft* 。 *commit_決める設定* を参照してください。 |

### QUORUM COMMIT ABORT ON パラメーター

`ABORT ON` は、クォーラムコミットには必須です。

| Parameter | Type | Default | Description |
|---|---|---|---|
| *timeout* | interval | 0 | アボートする前にクォーラムを待機する時間。 0 は設定されていないことを意味します。 |

### commit_決める設定

値 `group` は、オリジンノードがコミットスコープグループから必要な確認を受信すると、コミットすることを決定することを意味します。

値 `raft` は、コミット決定が PGD のビルトイン Raft コンセンサスを介して行われることを意味します。これは `group` より遅いですが、特定のマルチグループトポロジで必要になる場合があります。

### CAMO

クライアントの協力により、フェールオーバーシナリオで同じトランザクションの複数の挿入を防止する保護を有効にします。

詳細は、 *Commit At Most Once* を参照してください。

```
CAMO [ DEGRADE ON ( degrade_on_parameter = value ) TO ASYNC ]
```

### DEGRADE ON パラメーター

タイムアウト時に非同期操作に低下することを許可します。

| Parameter | Type | Default | Description |
|---|---|---|---|
| *timeout* | interval | 0 | ミリ秒単位のタイムアウト他の単位を受け入れますこの後、操作は非同期になります。 0 は設定されていないことを意味します。 |

次のページに続く

表 151 – 前のページからの続き

| Parameter | Type | Default | Description |
|---|---|---|---|
| *require_write_lead* | Boolean | False | ノードが非同期モードに切り替えることができるように書き込みリードである必要があるかどうかを指定します。 |

### ラグコントロール

レプリケーションラグによって制御される動的なレート制限の構成が可能になります。

詳細は、 ラグコントロール を参照してください。

```
LAG CONTROL [ ( lag_control_parameter = value [, ... ] ) ]
```

### LAG CONTROL パラメーター

| Parameter | Type | Default | Description |
|---|---|---|---|
| *max_lag_size* | int | 0 | 特定のノードが別のノードへのレプリケーション接続で持つことができる KB 単位の最大遅延。遅延が `max_commit_delay`によってスケールされたこの最大値を超えると、遅延制御はコミット遅延を調整します。 |
| *max_lag_time* | interval | 0 | 特定のダウンストリームノードへのレプリケーション接続に関して、特定のオリジンが持つことができるミリ秒単位の最大レプリケーションラグ。 |
| *max_commit_delay* | interval | 0 | 各コミットにかかる最大遅延をミリ秒の小数単位で構成します。0 に設定すると、Lag Control が無効になります。各コミット遅延調整の後たとえば、レプリケーションが `max_lag_size`または `max_lag_time`よりも遅延している場合、コミット遅延は `bdr.lag_control_commit_delay_adjust` GUC の重みで再計算されます。`max_commit_delay`は、コミット遅延の上限です。 |

- `max_lag_size` および `max_lag_time` が 0 に設定されている場合、LAG CONTROL は無効になります。
- `max_commit_delay` が設定されていないか、0 に設定されている場合、LAG CONTROL は無効にな

ります。

遅延サイズは、walsender の send_ptr から受信者の apply_ptr までのデルタから導き出されます。

遅延時間は次の式に従って計算されます。

```
lag_time = (lag_size / apply_rate) * 1000;
```

lag_size は、send_ptr と apply_ptr max_lag_size に使用されるデルタであり、 apply_rate は、単純化された式に従う加重指数移動平均です。

```
apply_rate = prev_apply_rate * (1 - apply_rate_weight) +
             ((apply_ptr_diff * apply_rate_weight) / diff_secs);
```

そこで

- prev_apply_rate は、新しいレートを再計算する前に、以前に構成された apply_rate でした。
- apply_rate_weight は、GUC bdr.lag_tracker_apply_rate_weight の値です。
- apply_ptr_diff は、現在の apply_ptr と、適用レートが最後に計算された時点での apply_ptr の差です。
- diff_secs は、最後に適用レートが計算されてからの秒単位のデルタです。

### グループコミット

ノードのコンセンサスによってコミットを確認でき、競合解決設定を制御し、**同期コミット** と同様に、オプションのルール低下パラメーターがあります。

```
GROUP COMMIT [ ( group_commit_parameter = value [, ...] ) ] [ ABORT ON ( abort_on_
↪parameter = value ) ] [ DEGRADE ON (degrade_on_parameter = value ) TO commit_scope_
↪degrade_operation ]
```

### GROUP COMMIT パラメーター

| Parameter | Type | Default | Description |
|---|---|---|---|
| *transaction_tracking* | Boolean | Off/False | トランザクションのステータスを追跡するかどうかを指定します。 *transaction_tracking* **設定** を参照してください。 |
| *conflict_resolution* | enum | async | 競合を処理する方法を指定します。 *async`I`eager* 。 *fence_resolution* **設定** を参照してください。 |

次のページに続く

表 153 – 前のページからの続き

| Parameter | Type | Default | Description |
|---|---|---|---|
| *commit_decision* | enum | group | COMMIT 決定を行う方法を指定します。 *group`I`partner`I`raft* 。*commit_*決める設定 を参照してください。 |

### ABORT ON パラメーター

| Parameter | Type | Default | Description |
|---|---|---|---|
| *timeout* | interval | 0 | ミリ秒単位のタイムアウト他の単位を受け入れます。 0 は設定されていないことを意味します。 |
| *require_write_lead* | Boolean | False | CAMO のみ。設定されている場合、トランザクションをローカル非同期モードに切り替えるには、コンセンサス要求が必要です。 |

### DEGRADE ON パラメーター

| Parameter | Type | Default | Description |
|---|---|---|---|
| *timeout* | interval | 0 | ミリ秒単位のタイムアウト他の単位を受け入れます。その後、オペレーションが低下します。 0 は設定されていないことを意味します。 |
| *require_write_lead* | Boolean | False | ノードが縮退動作に切り替えることができるように書き込みリードである必要があるかどうかを指定します。 |

### transaction_tracking 設定

true に設定されている場合、2 フェーズコミットトランザクション

- ライタが PREPARE メッセージを処理しているときにコミット決定をルックアップします。
- 中断から回復するときは、中断前に準備されたトランザクションを検索します。見つかると、トランザクションのコミットスコープと、対応する RAFT コミット決定を検索します。ノードがトランザクションのオリジンであり、RAFT コミット決定がなく、コミットスコープで `transaction_tracking` がオンであると仮定します。その場合、コミットまたは中止されるまで、この未解決のトランザクションの RAFT コミット決定を定期的に検索します。

#### fence_resolution 設定

値 async は、レプリケーション中に競合解決ポリシーを使用して競合を非同期的に解決することを意味します。

値 eager は、COMMIT 中に、競合するトランザクションのいずれかがアボートすることにより、競合が熱心に解決されることを意味します。

Eager は、 MAJORITY または ALL コミットスコープグループでのみ使用できます。

ALL コミットスコープグループで使用する場合、リコンシレーションの問題を回避するには、*bdr.commit_Decisions* を raft に設定する必要があります。

欠落列の競合の解決 を参照してください。

#### GROUP COMMIT commit_決断設定

値 group は、前述の commit_scope_group 仕様が耐久性だけでなく COMMIT 決定にも影響を与えることを意味します。

値 partner は、パートナーノードがトランザクションをコミットできるかどうかを決定することを意味します。この値は、2 つのデータノードを持つグループでのみ許可されます。

値 raft は、決定が PGD のビルトイン Raft コンセンサスを利用することを意味します。選択したコミットスコープグループのすべてのノードがトランザクションを確認したら、PGD クラスター内のすべてのノードがトランザクションを記録したことを確認するために、全ノードの Raft で記録されます。

ALL コミットスコープグループが使用されている場合、このオプションは使用して、決定を巡ってノード間で相違がないことを保証する必要があります。このオプションはパフォーマンスが低い場合があります。

決定をコミットする を参照してください。

#### commit_scope_degrade_operation 設定

commit_scope_degrade_operation は、より制限の少ないコミットスコープグループの GROUP_COMMIT であるか、非同期 ASYNC である必要があります。

### Conflicts

#### 競合検出

#### 競合タイプのリスト

PGD は、次の競合タイプを認識し、 conflict_type パラメーターとして使用できます。

| Conflict type | Description |
|---|---|
| *insert_exists* | 着信挿入は、主キーまたは一意キー/インデックスを介して既存の行と競合します。 |
| *update_differing* | 着信更新のキー行はローカル行とは異なります。これは、 行バージョンの競合検出 を使用する場合にのみ発生する可能性があります。 |
| *update_origin_change* | 着信更新は、別のノードによって最後に変更された行を変更しています。 |

次のページに続く

表 156 – 前のページからの続き

| Conflict type | Description |
|---|---|
| *update_missing* | 着信更新は、存在しない行を変更しようとしています。 |
| *update_recently_deleted* | 着信更新は、最近削除された行を変更しようとしています。 |
| *update_pkey_exists* | 着信更新により、`PRIMARY KEY`が変更され、変更を適用しているノードに既に存在する値が変更されました。 |
| *multiple_unique_conflicts* | 受信行は、ターゲットテーブルの UNIQUE / EXCLUDE インデックスごとに複数の行と競合します。 |
| *delete_recently_updated* | 現在のノードまたは 行バージョンの競合検出 を使用している場合、行の最新の更新よりも古いコミットタイムスタンプを含む着信削除。 |
| *delete_missing* | 着信削除は、存在しない行を削除しようとしています。 |
| *target_column_missing* | ターゲットテーブルには、受信行に存在する 1 つ以上の列がありません。 |
| *source_column_missing* | 着信行には、ターゲットテーブルに存在する 1 つ以上の列がありません。 |
| *target_table_missing* | ターゲットテーブルがありません。 |
| *apply_error_ddl* | レプリケートされた DDL コマンドを適用するときに Postgres によってエラーがスローされました。 |

### 競合の解決

ほとんどの競合は自動的に解決できます。PGD は、last-update-wins メカニズム、より正確には `update_if_newer` 競合リゾルバーにデフォルトで設定されます。このメカニズムでは、競合検出に使用される同じコミットタイムスタンプに基づいて、2 つの競合する行の最後に挿入または変更された行を保持します。特定のコーナーケースのシナリオでの動作は、`bdr.create_node_group` または `bdr.alter_node_group` に使用される設定によって異なります。

PGD では、次のファンクションを使用して競合解決のデフォルト動作をオーバーライドできます。

### 競合リゾルバーのリスト

PGD ではいくつかの競合リゾルバーが使用でき、処理できる競合タイプのカバレッジが異なります。

| Resolver | Description |
|---|---|
| *error* | エラーをスローし、レプリケーションを停止します。 |
| *skip* | リモート変更の処理をスキップし、次の変更でレプリケーションを続行します。 *insert_exists*、*update_differing*、*update_origin_change*、*update_missing*、*update_recently_deleted*、*update_pkey_exists*、*delete_recently_updated*、*delete_missing*、*target_table_missing*、*target_column_missing*、および `source_column_missing`競合タイプに使用できます。 |

次のページに続く

表 157 – 前のページからの続き

| Resolver | Description |
|---|---|
| *skip_if_recently_dropped* | 最近ダウンストリームで 1 日以内にドロップされたため、ダウンストリームが存在しないテーブル用の場合、リモート変更をスキップします。それ以外の場合はエラーをスローします。 `target_table_missing`競合タイプに使用できます。この競合リゾルバーは、同じ名前のテーブルが削除された後すぐに再作成される場合、課題を引き起こす可能性があります。その場合、ノードのいずれかが、テーブルを再作成する DDL を参照する前に、再作成されたテーブルの DML を参照する場合があります。次に、テーブルが最近ドロップされたと仮定して、リモートデータを誤ってスキップし、データ損失を引き起こします。このリゾルバーを使用する場合、ドロップされた直後にオブジェクト名を再利用しないことをお勧めします。 |
| *skip_transaction* | 競合を生成したトランザクション全体をスキップします。 |
| *update_if_newer* | リモート行が競合するローカル行よりも後にコミットされた場合に更新されます元ノードの壁時計によって決定されます。タイムスタンプが同じ場合、ノード ID はタイブレーカーとして使用され、すべてのノードで同じ行が選択されるように保証されます。より高いノード ID が勝ちます。 *insert_exists*、*update_differing*、*update_origin_change*、および `update_pkey_exists`競合タイプに使用できます。 |
| *update* | レプリケートされたアクションを常に実行します。 *insert_exists INSERT*`を `*UPDATE*`に変更します 、`*update_differing*、*update_origin_change*、*update_pkey_exists*、および `delete_recently_updated` 削除を実行します。 |
| *insert_or_skip* | オリジンから送信された利用可能な情報から新しい行を構築し、 INSERT しようとします。完全な行を作成するための十分な情報がない場合、変更をスキップします。 `update_missing`および `update_recently_deleted`競合タイプに使用できます。 |
| *insert_or_error* | オリジンから送信された利用可能な情報から新しい行を構築し、挿入しようとします。完全な行を構築するために利用できる十分な情報がない場合、エラーをスローし、レプリケーションを停止します。完全な行を構築するために利用できる十分な情報がない場合、エラーをスローし、レプリケーションを停止します。 `update_missing`および `update_recently_deleted`競合タイプに使用できます。 |
| *ignore* | 欠落しているターゲット列を無視して、処理を続行します。 `target_column_missing`競合タイプに使用できます。 |
| *ignore_if_null* | リモート行の追加列に NULL 値が含まれる場合、欠落しているターゲット列を無視します。それ以外の場合は、エラーをスローし、レプリケーションを停止します。 `target_column_missing`競合タイプに使用できます。 |
| *use_default_value* | 欠落している列値をデフォルト列のデフォルトの場合は NULL を含むで埋めて、処理を続行します。デフォルトまたは制約違反つまり、 NOT NULL 列の NULL デフォルトの処理中にエラーが発生すると、レプリケーションが停止されます。 `source_column_missing`競合タイプに使用できます。 |

`insert_exists` 、`update_differing` 、`update_origin_change` 、`update_missing` 、`multiple_unique_conflicts` 、`update_recently_deleted` 、`update_pkey_exists` 、`delete_recently_updated` 、および `delete_missing` 競合タイプは、*Stream triggers* を使用するユーザー定義ロジックでも解決できます。

このマトリックスは、各競合リゾルバーが処理できる競合タイプを示しています。

| | in-sert_e | up-date_( | up-date_( | up-date_r | up-date_r | up-date_ | delete | delete | tar-get_cc | source | tar-get_ta | mul-ti-ple_unique_conflicts | ing<br>ssing |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| error | X | X | X | X | X | X | X | X | X | X | X | X | |
| skip | X | X | X | X | X | X | X | X | X | X | X | X | |
| skip_if_rec | | | | | | | | | | | X | | |
| up-date_if_new | X | X | X | | | X | | | | | | X | |
| update | X | X | X | | | X | X | | | | | X | |
| in-sert_or_ski | | | | X | X | | | | | | | | |
| in-sert_or_err | | | | X | X | | | | | | | | |
| ignore | | | | | | | | | X | | | | |
| ig-nore_if_nul | | | | | | | | | X | | | | |
| use_default_ | | | | | | | | | | X | | | |
| con-flict_trigger | X | X | X | X | X | X | X | X | | | | X | |

### デフォルトの競合リゾルバー

| Conflict type | Resolver |
|---|---|
| insert_exists | update_if_newer |
| update_differing | update_if_newer |
| update_origin_change | update_if_newer |
| update_missing | insert_or_skip |
| update_recently_deleted | skip |
| update_pkey_exists | update_if_newer |
| multiple_unique_conflicts | error |
| delete_recently_updated | skip |
| delete_missing | skip |
| target_column_missing | ignore_if_null |
| source_column_missing | use_default_value |
| target_table_missing (see note) | skip_if_recently_dropped |
| apply_error_ddl | error |

競合解決は、競合リゾルバーが選択した解決の種類を表し、競合を解決するために実行された特定のアクションに対応します。

次の競合解決は、`conflict_resolution` パラメーターで現在サポートされています。

| Resolution | Description |
|---|---|
| *apply_remote* | リモート着信行が適用されました。 |
| *skip* | 行の処理はスキップされましたローカルでの変更は行われませんでした。 |
| *merge* | リモート行とローカル行からの情報をマージして、新しい行が作成されました。 |
| *user* | ユーザーコード競合トリガーは、ターゲットテーブルに書き込まれる行を生成しました。 |

### 競合ログ

バージョン 6.0 以降、PGD はデフォルトで bdr.conflict_history テーブルに競合をログに記録しません。これは、テーブルが大きくなり、パフォーマンスの問題を発生させる可能性があるためです。bdr.alter_node_set_log_config ファンクションを使用して、競合ログを有効にできます。この機能を使用すると、ログに記録する競合を詳細に制御できます。または、次のようにすべての競合をログに記録するように設定できます。

```
SELECT bdr.alter_node_set_log_config(`nodename`, false, true, NULL, NULL);
```

名前付けノードでこのコマンドを実行して、その特定のノードでのすべての競合のログを有効にします。すべてのノードでロギングを有効にする場合は、PGD グループの各ノードでこのコマンドを実行します。

## Conflict functions

### bdr.alter_table_conflict_detection

テーブルの所有者は、特定のテーブルの競合検出がどのように動作するかを変更できます。

#### 概要

```
bdr.alter_table_conflict_detection(relation regclass,
                                   method text,
                                   column_name name DEFAULT NULL)
```

#### パラメーター

- relation — 新しい競合検出方法を設定するリレーションの名前。
- method — 使用する競合検出方法。
- column_name —列検出データの保存に使用する列。これはスキップできます。その場合、列名は競合検出方法に基づいて選択されます。row_origin メソッドは、メタデータストレージに追加の列を必要としません。

競合検出として認識されている方法は次のとおりです。

- row_origin — タプルに加えられた以前の変更のオリジン オリジンの競合検出 を参照してください。これは、テーブルに追加の列を必要としないサポートされている唯一の方法です。
- row_version — 行バージョン列 行バージョンの競合検出 を参照してください。

- column_commit_timestamp — 列ごとのコミットタイムスタンプ *CLCD* で説明。
- column_modify_timestamp — 列ごとの変更タイムスタンプ *CLCD* で説明しています。

#### 注

column_commit_timestamp と column_modify_timestamp の競合検出方法の違いの詳細については、*Timestamps in column-level conflict resolution* を参照してください。

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *ddl filters* 構成の影響を受けることを意味します。

ファンクションは、列レベルの競合解決が有効になっているリレーションで DML グローバルロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに可視されます。

bdr.backwards_compatibility が 30618 以下に設定されていない限り、relation の所有者のみが bdr.alter_table_conflict_detection ファンクションを実行できます。

**警告**

追加の列を使用してメタデータを保存する方法から競合検出方法を変更すると、その列がドロップされます。

**警告**

このファンクションは、*bdr.camo_enable_client_warnings* で警告が無効になっていない限り、CAMO を無効にし、警告を発します。

### bdr.alter_node_set_conflict_resolver

この機能は、特定のノードでの競合解決の動作を設定します。

#### 概要

```
bdr.alter_node_set_conflict_resolver(node_name text,
                                     conflict_type text,
                                     conflict_resolver text)
```

#### パラメーター

- node_name — 変更されるノードの名前。
- conflict_type — 設定を適用する競合タイプ競合タイプのリスト を参照してください。
- conflict_resolver — 指定された競合タイプに使用するリゾルバー競合リゾルバーのリスト を参照してください。

#### 注

現在、ローカルノードのみを変更できます。ファンクション呼び出しはレプリケートされません。複数のノードの設定を変更する場合は、各ノードでファンクションを実行する必要があります。

このファンクションによる構成変更は、bdr.create_node_group または bdr.alter_node_group で指定された競合解決のデフォルト動作をオーバーライドします。

このファンクションはトランザクションです。変更はロールバックできます。それらは現在のトランザクションに表示されます。

### bdr.alter_node_set_log_config

ノードの競合ログ構成を設定します。

#### 概要

```
bdr.alter_node_set_log_config(node_name text,
                              log_to_file bool DEFAULT true,
                              log_to_table bool DEFAULT true,
                              conflict_type text[] DEFAULT NULL,
                              conflict_resolution text[] DEFAULT NULL)
```

#### パラメーター

- node_name — 変更されるノードの名前。
- log_to_file — ノードログファイルに記録するかどうか。
- log_to_table — bdr.conflict_history テーブルにログを記録するかどうか。
- conflict_type — ログに記録する競合タイプ。NULL デフォルトはすべてを意味します。
- conflict_resolution — ログへの競合の解決。NULL デフォルトはすべてを意味します。

#### 注

ローカルノードのみを変更できます。ファンクション呼び出しはレプリケートされません。複数のノードの設定を変更する場合は、各ノードでファンクションを実行する必要があります。

このファンクションはトランザクションです。変更はロールバックできます。それらは現在のトランザクションに表示されます。

#### 競合ログ構成のリスト

bdr.node_log_config ビューには、すべてのロギング構成が表示されます。ここには、ロギング構成の名前、ログの場所、およびログに記録される競合のタイプと解決がリストされています。

#### 競合のテーブルへのロギング

log_to_table が true に設定されている場合、競合はテーブルにログに記録されます。競合ログのターゲットテーブルは bdr.conflict_history です。

このテーブルは列 local_time でレンジパーティション化されています。テーブルはオートパーティションで管理されています。デフォルトでは、新しいパーティションは毎日作成され、過去 1 か月の競合が維持されます。その後、古いパーティションは削除されます。自動パーティションは、事前に 7〜14 のパーティションを作成します。bdr_superuser はこれらのデフォルトを変更できます。

PGD によって管理されるすべてのテーブルで生成された競合はこのテーブルに記録されるため、正当なユーザーのみが競合したデータを読み取ることができるようにすることが重要です。PGD は、 bdr.conflict_history テーブルで ROW LEVEL SECURITY ポリシーを定義することによりこれを行います。テーブルの所有者のみが、各テーブルの競合を読み取ることができます。基になるテーブルに RLS ポリシーが定義、有効化、および適用されている場合、所有者でも競合を読み取ることはできません。FORCE オプションで作成された RLS ポリシーは、テーブルの所有者にも適用されます。その場合、基になるテーブルの一部またはすべての行は、所有者でも読み取ることができない場合があります。したがって、PGD は、競合ログテーブルにより厳密なポリシーも適用します。

事前定義されたロール bdr_read_all_conflicts は、bdr_superuser ロールを付与せずに、bdr.conflict_history テーブルに記録された競合の詳細をすべて表示する必要があるユーザーに付与できます。

デフォルトのロール bdr_read_all_stats は、bdr.conflict_history_summary というカタログビューにアクセスします。このビューにはユーザーデータが含まれないため、ログに記録された競合を監視できます。

### Replication set management

#### bdr.create_replication_set

この機能は、レプリケーションセットを作成します。

このコマンドのレプリケーションは、DDL フィルタリング設定を含む DDL レプリケーション構成の影響を受けます。

#### 概要

```
bdr.create_replication_set(set_name name,
                           replicate_insert boolean DEFAULT true,
                           replicate_update boolean DEFAULT true,
                           replicate_delete boolean DEFAULT true,
                           replicate_truncate boolean DEFAULT true,
                           autoadd_tables boolean DEFAULT false,
                           autoadd_existing boolean DEFAULT true)
```

#### パラメーター

- set_name _ 新しいレプリケーションセットの名前。PGD グループ全体で一意である必要があります。
- replicate_insert _ このレプリケーションセットのテーブルへの挿入をレプリケートするかどうかを示します。

- replicate_update _ このレプリケーションセット内のテーブルの更新をレプリケートするかどうかを示します。
- replicate_delete _ このレプリケーションセット内のテーブルからの削除をレプリケートするかどうかを示します。
- replicate_truncate _ このレプリケーションセット内のテーブルのトランケートをレプリケートするかどうかを示します。
- autoadd_tables _ 新しく作成される将来のテーブルをこのレプリケーションセットにレプリケートするかどうかを示します
- autoadd_existing _ 既存のすべてのユーザーテーブルをこのレプリケーションセットに追加するかどうかを示します。このパラメーターは、autoadd_tables が true に設定されている場合にのみ効果があります。

#### 注

デフォルトでは、新しいレプリケーションセットは DDL または PGD 管理ファンクション呼び出しをレプリケートしません。レプリケーションセットの DDL レプリケーションを設定する方法については、*DDL filters* を参照してください。既存の DDL フィルターは、すべての DDL および管理ファンクション呼び出しをレプリケートするデフォルトのグループレプリケーションセットに設定されます。これは、グループの作成時に作成されますが、PGD グループのデフォルトレプリケーションセットで DDL または PGD 管理ファンクション呼び出しをレプリケートしたくない場合にドロップできます。

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *DDL filters* の影響を受けることを意味します

構成。

ファンクションは DDL グローバルロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

### bdr.alter_replication_set

この機能は、既存のレプリケーションセットのオプションを変更します。

このコマンドのレプリケーションは、DDL フィルタリング設定を含む DDL レプリケーション構成の影響を受けます。

#### 概要

```
bdr.alter_replication_set(set_name name,
                          replicate_insert boolean DEFAULT NULL,
                          replicate_update boolean DEFAULT NULL,
                          replicate_delete boolean DEFAULT NULL,
                          replicate_truncate boolean DEFAULT NULL,
                          autoadd_tables boolean DEFAULT NULL)
```

#### パラメーター

- set_name _ 既存のレプリケーションセットの名前。
- replicate_insert _ このレプリケーションセットのテーブルへの挿入をレプリケートするかどうかを示します。
- replicate_update _ このレプリケーションセット内のテーブルの更新をレプリケートするかどうかを示します。
- replicate_delete _ このレプリケーションセット内のテーブルからの削除をレプリケートするかどうかを示します。
- replicate_truncate _ このレプリケーションセット内のテーブルのトランケートをレプリケートするかどうかを示します。
- autoadd_tables _ 新しく作成される将来のテーブルをこのレプリケーションセットに追加するかどうかを示します。

NULL デフォルトに設定されているオプションは、以前と同じままです。

#### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *DDL filters* の影響を受けることを意味します

構成。

ファンクションは DDL グローバルロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

### bdr.drop_replication_set

この機能は、既存のレプリケーションセットを削除します。

このコマンドのレプリケーションは、DDL フィルタリング設定を含む DDL レプリケーション構成の影響を受けます。

#### 概要

```
bdr.drop_replication_set(set_name name)
```

#### パラメーター

- set_name _ 既存のレプリケーションセットの名前。

### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *ddl filters* の影響を受けることを意味します

構成。

ファンクションは DDL グローバルロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

> **警告**
>
> 少なくとも別のノードで使用されているレプリケーションセットを削除しないでください。削除すると、そのノードでレプリケーションが停止します。これが発生した場合は、影響を受けるノードをそのレプリケーションセットからサブスクライブ解除します。同じ理由で、参加されるノードがそのレプリケーションセットのメンバーである場合、参加操作が進行中のレプリケーションセットをドロップしないでください。レプリケーションセットのメンバーシップは、結合の開始時にのみチェックされます。これは、レプリケーションセットの使用状況に関する情報が各ノードにローカルであるため、グループに参加する前にノードで構成できるために発生します。

alter_node_replication_sets を使用して、ノードのレプリケーションセットのサブスクリプションを管理できます。

## bdr.alter_node_replication_sets

この機能は、ノードが公開およびサブスクライブするレプリケーションセットを変更します。

### 概要

```
bdr.alter_node_replication_sets(node_name name,
                                set_names text[])
```

### パラメーター

- node_name _ 変更するノード。現在はローカルノードである必要があります。
- set_names _ 指定されたノードにレプリケートするレプリケーションセットの配列。空の配列を使用すると、グループのデフォルトレプリケーションセットが使用されます。

### 注

このファンクションはローカルノードでのみ実行され、どのような方法でもレプリケートされません。

このファンクションはノードが参加する前に実行するように設計されているため、リストされたレプリケーションセットの存在はチェックされません。エラーを回避するには、レプリケーションセット名を正しく指定するように注意してください。

この動作により、PGD グループの一部であるノードでだけでなく、まだグループに参加していないノードでもファンクションを呼び出すことができます。このアプローチでは、結合中に同期されるデータを制限します。ただし、スキーマは、レプリケーションセット設定に関係なく、常に完全に同期されます。レプリケーションセットで指定されたテーブルだけでなく、すべてのテーブルがコピーされます。bdr.tables カタログテーブルを参照して、不要なテーブルを削除できます。(これらは PGD の後のバージョンでは自動的に削除される可能性があります。これは現在 true です。 *DDL filters*

それ以外の場合、構成は DDL のレプリケーションを防止します。

この呼び出しの後にノードがサブスクライブするレプリケーションセットは、他のノードからこのファンクションが実行されるノードに変更を実際にレプリケートするために他のノードによって公開されます。

### Replication set membership

#### bdr.replication_set_add_table

このファンクションは、テーブルをレプリケーションセットに追加します。

このファンクションは、テーブルをレプリケーションセットに追加し、ファンクションの呼び出しを含むトランザクションのコミットから変更のレプリケートを開始します。ノード上のテーブルが持っている既存のデータは同期されません。このコマンドのレプリケーションは、DDL フィルタリング設定を含む DDL レプリケーション構成の影響を受けます。

##### 概要

```
bdr.replication_set_add_table(relation regclass,
                              set_name name DEFAULT NULL,
                              columns text[] DEFAULT NULL,
                              row_filter text DEFAULT NULL)
```

##### パラメーター

- relation _ テーブルの名前または Oid。
- set_name _ レプリケーションセットの名前。NULL デフォルトの場合、PGD グループのデフォルトレプリケーションセットが使用されます。
- columns _ 将来の使用のために予約されています現在は何もせず、NULL である必要があります。
- row_filter _ レプリケートされた行のフィルタリングに使用する SQL 式。この式が定義されていない場合つまり、デフォルトの NULL に設定されている場合、すべての行が送信されます。

row_filter は、NULL を使用したブール結果を生成する式を指定します。True または Unknown と評価される式は、行をレプリケートします。False 値は行をレプリケートしません。式にサブクエリーを含めたり、レプリケートされる現在の行の列以外の変数を参照したりすることはできません。システム列を参照することはできません。

row_filter は、ターゲットノードではなく、オリジンノードで実行されます。これにより、この特定のテーブルのレプリケーションに追加の CPU オーバーヘッドがかかりますが、フィルターされた行のデータの送信が完全に回避されます。したがって、ネットワーク帯域幅が削減され、ターゲットノードのオーバーヘッドが適用されます。

row_filter は、特定のテーブルの TRUNCATE コマンドを削除することはありません。レプリケーションセットレベルで TRUNCATE コマンドをフィルタリングできます。

テーブルの一部の列のみをレプリケートできます。 差分を含むノード間のレプリケーションを参照してください。

#### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *DDL filters* の影響を受けることを意味します

構成。

row_filter が NULL でない場合、ファンクションはレプリケーションセットに追加されるリレーションで DML グローバルロックを取得します。それ以外の場合は、DDL グローバルロックのみを使用します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

### bdr.replication_set_remove_table

このファンクションは、レプリケーションセットからテーブルを削除します。

このコマンドのレプリケーションは、DDL フィルタリング設定を含む DDL レプリケーション構成の影響を受けます。

#### 概要

```
bdr.replication_set_remove_table(relation regclass,
                                 set_name name DEFAULT NULL)
```

#### パラメーター

- relation _ テーブルの名前または Oid。
- set_name _ レプリケーションセットの名前。NULL デフォルトの場合、PGD グループのデフォルトレプリケーションセットが使用されます。

#### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *DDL filters* の影響を受けることを意味します

構成。

ファンクションは DDL グローバルロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

### DDL replication filtering

*DDL* レプリケーションフィルタリング も参照してください。

### bdr.replication_set_add_ddl_filter

このファンクションは、 DDL フィルタをレプリケーションセットに追加します。

指定されたフィルターに一致する DDL は、そのセットをサブスクライブしているノードにレプリケートされます。この機能は、PGD 管理機能のレプリケーションにも影響します。

この機能は、ノードでの DDL の実行を妨げるものではありません。DDL を他のノードにレプリケートするかどうかのみを変更します。2 つのノードの間に、すべてのインデックスコマンドを除外するレプリケーションフィルタがあるとします。各ノードに直接接続し、そのノードで目的の DDL を実行することにより、インデックスコマンドを自由に実行できます。

DDL フィルターは command_tag および role_name を指定して、一部の DDL ステートメントのみのレプリケーションを許可できます。command_tag は、event triggers で使用されるものと同じです

通常の PostgreSQL コマンドの場合。典型的な例は、ロジカルスタンバイ上の追加のインデックスコマンドが他のすべてのノードにレプリケートされないようにフィルタを作成することです。

修飾されたファンクション名と一致するタグ名を使用して、使用される PGD 管理機能をフィルタリングできます。たとえば、bdr.replication_set_add_table は、同じ名前のファンクションのコマンドタグです。この場合、このタグにより、bdr.* を使用してすべての PGD ファンクションをフィルタリングできます。

role_name は、コマンドを実行している現在のロールとの照合に使用されます。command_tag と role_name は両方とも正規表現として評価され、大文字と小文字が区別されます。

#### 概要

```
bdr.replication_set_add_ddl_filter(set_name name,
                                   ddl_filter_name text,
                                   command_tag text,
                                   role_name text DEFAULT NULL,
                                   base_relation_name text DEFAULT NULL,
                                   query_match text DEFAULT NULL,
                                   exclusive boolean DEFAULT FALSE)
```

#### パラメーター

- set_name _ レプリケーションセットの名前。NULL の場合、PGD グループのデフォルトのレプリケーションセットが使用されます。
- ddl_filter_name _ DDL フィルターの名前。この名前は、PGD グループ全体で一意である必要があります。
- command_tag _ コマンドタグと一致するための正規表現。NULL は、すべてと一致することを意味します

- role_name _ ロール名と一致するための正規表現。NULL は、すべてのロールと一致することを意味します
- base_relation_name _ 将来の使用のために予約されています。NULL である必要があります。
- query_match _ クエリーにマッチするための正規表現。NULL は、すべてのクエリーと一致することを意味します。
- exclusive _ true の場合、他の一致したフィルタは考慮されませんつまり、排他的フィルタのみが適用されます。複数の排他的フィルタが一致する場合、エラーがスローされます。このパラメーターは、メインレプリケーションセットを介してデフォルトのレプリケーションを維持しながら、特定のコマンドを特定のレプリケーションセットにルーティングするのに役立ちます。

#### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *DDL filters* の影響を受けることを意味します

構成。これは、DDL フィルタ構成の変更のレプリケーションが既存の DDL フィルタ構成の影響を受けることも意味します。

ファンクションは DDL グローバルロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

定義されたレプリケーションフィルタを表示するには、ビュー bdr.ddl_replication を使用します。

#### 例

PGD 管理機能のみを含めるには、次のようなフィルターを定義します。

```
SELECT bdr.replication_set_add_ddl_filter(mygroup, mygroup_admin, $$bdr\..*$$);
```

インデックス DDL を除くすべてを除外するには

```
SELECT bdr.replication_set_add_ddl_filter(mygroup, index_filter,
               ^(?!(CREATE INDEX|DROP INDEX|ALTER INDEX)).*);
```

テーブルとインデックスに対するすべての操作を含め、その他のすべてを除外するには、テーブル用とインデックス用の 2 つのフィルターを追加します。この例は、複数のフィルタが、許可されたすべての DDL コマンドの和集合を提供することを示しています。

```
SELECT bdr.replication_set_add_ddl_filter(bdrgroup,index_filter, ^((?!INDEX).)*$);
SELECT bdr.replication_set_add_ddl_filter(bdrgroup,table_filter, ^((?!TABLE).)*$);
```

### bdr.replication_set_remove_ddl_filter

このファンクションは、レプリケーションセットから DDL フィルターを削除します。

このコマンドのレプリケーションは、DDL フィルタリング設定を含む DDL レプリケーション構成の影響を受けます。

#### 概要

```
bdr.replication_set_remove_ddl_filter(set_name name,
                                      ddl_filter_name text)
```

#### パラメーター

- set_name _ レプリケーションセットの名前。NULL の場合、PGD グループのデフォルトのレプリケーションセットが使用されます。
- ddl_filter_name _ 削除する DDL フィルターの名前。

#### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *DDL filters* 構成の影響を受けることを意味します。これは、DDL フィルタ構成の変更のレプリケーションが既存の DDL フィルタ構成の影響を受けることも意味します。

ファンクションは DDL グローバルロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

### Testing and tuning commands

EDB Postgres Distributed には、PGD クラスターのテストと調整に役立つツールが含まれています。バックグラウンドについては、*Postgres Distributed terminology* を参照してください。

### pgd_bench

#### 概要

EDB Postgres 分散展開のベンチマークツール。

```
pgd_bench [OPTION]... [DBNAME] [DBNAME2]
```

DBNAME は、次のフォーマットの conninfo 文字列です。"host=10.1.1.2 user=postgres dbname=master"

pgd_bench のオプションと使用例については、*pgd_bench in Testing and tuning* を参照してください。

#### オプション

pgd_bench コマンドは、pgbench コマンドのラッパーとして実装されています。これは、同じオプションの多くを共有し、テストを実行するときに pgbench という名前のテーブルを作成したことを意味します。

pgd_bench に固有のオプションには次のものが含まれます。

### 設定モード

-m または--mode

モードは regular 、camo 、または failover に設定できます。デフォルトは regular です。

- regular — pgd_bench を実行するには 1 つのノードのみが必要です。
- camo —CAMO パートナーとして機能するように 2 番目のノードを指定する必要があります。CAMO の設定が必要です。
- failover —フェールオーバーとして機能するように 2 番目のノードを指定する必要があります。

-m failover を使用する場合、追加オプション--retry を使用できます。このオプションは、フェイルオーバーが発生した場合にトランザクションを再試行するように pgd_bench に指示します。-m camo を使用する場合、 --retry オプションは自動的に有効になります。

-m camo を使用し、カスタムスクリプトを提供する場合、スクリプト内の SQL コマンドを SQL トランザクションコマンドでラップする必要があります。つまり、最初の SQL コマンドは BEGIN であり、最後の SQL コマンドは COMMIT である必要があります。

### GUC 変数の設定

-o または--set-option

このオプションの後には、NAME=VALUE エントリが続きます。これらは、pgd_bench が接続する各サーバー、およびそれらのサーバーにのみ Postgres SET コマンドを使用して適用されます。

他のオプションは Postgres pgbench コマンドと同じです。詳細については、PostgreSQL pgbench ドキュメントを参照してください。

オプション pgd_bench および pgbench の完全なリストが以下にあります。

### 初期化オプション

- -i, --initialize — 初期化モードを呼び出します。
- -I, --init-steps=[dtgGvpf]+ デフォルト "dtgvp" —選択した初期化ステップを実行します。
  - d —既存の pgbench テーブルを削除します。
  - t —標準の pgbench シナリオで使用するテーブルを作成します。
  - g —クライアントサイドでデータを生成し、標準テーブルにロードします。既に存在するデータを置き換えます。
  - G —サーバーサイドでデータを生成し、標準テーブルにロードし、既に存在するデータを置き換えます。
  - v — 標準テーブルで VACUUM を呼び出します。
  - p — 標準テーブルに主キーインデックスを作成します。
  - f — 標準テーブル間に外部キー制約を作成します。
- -F, --fillfactor=NUM — フィルファクターを設定します。

- -n, --no-vacuum — 初期化中に VACUUM を実行しないでください。
- -q, --quiet — クワイエットロギング 5 秒ごとに 1 メッセージ。
- -s, --scale=NUM — スケーリング係数。
- --foreign-keys — テーブル間に外部キーコンストレインを作成します。
- --index-tablespace=TABLESPACE —指定されたテーブルスペースにインデックスを作成します。
- --partition-method=(range|hash) —このメソッドで pgbench_accounts をパーティション化します。デフォルトは range です。
- --partitions=NUM — pgbench_accounts を NUM パーツにパーティション化します。デフォルトは 0 です。
- --tablespace=TABLESPACE —指定されたテーブルスペースにテーブルを作成します。
- --unlogged-tables —ログのないテーブルとしてテーブルを作成します。注ログに記録されていないテーブルはレプリケートされません。

#### 実行するものを選択するオプション

- -b, --builtin=NAME[@W] —W で重み付けされたビルトインスクリプト NAME を追加します。デフォルトは 1 です。-b list を使用して、使用可能なスクリプトをリストします。
- -f, --file=FILENAME[@W] — W で重み付けスクリプト FILENAME を追加します。デフォルトは 1 です。
- -N, --skip-some-updates —pgbench_tellers および pgbench_branches の更新。 -b simple-update と同じ。
- -S, --select-only — SELECT のみのトランザクションを実行します。-b select-only と同じ。

#### ベンチマークオプション

- -c, --client=NUM —同時データベースクライアントの数。デフォルトは 1 です。
- -C, --connect — トランザクションごとに新しい接続を確立します。
- -D, --define=VARNAME=VALUE —カスタムスクリプトで使用する変数を定義します。
- -j, --jobs=NUM — スレッドの数。デフォルトは 1 です。
- -l, --log — トランザクション時間をログファイルに書き込みます。
- -L, --latency-limit=NUM —NUM ミリ秒を超えて続いたトランザクションを遅くまでカウントします。
- -m, --mode=regular|camo|failover —pgbench を実行するモード。デフォルトは regular です。
- -M, --protocol=simple|extended|prepared —クエリーを送信するためのプロトコル。デフォルトは simple です。
- -n, --no-vacuum — テストの前に VACUUM を実行しないでください。
- -o, --set-option=NAME=VALUE — ランタイム SET オプションを指定します。

- -P, --progress=NUM —NUM 秒ごとにスレッドの進行状況レポートを表示します。
- -r, --report-per-command —コマンドごとのレイテンシー、失敗、および再試行。
- -R, --rate=NUM — 1 秒ごとのトランザクションのターゲットレート。
- -s, --scale=NUM —このスケール係数を出力で報告します。
- -t, --transactions=NUM —各クライアントが実行するトランザクションの数。デフォルトは 10 です。
- -T, --time=NUM — ベンチマークテストの時間秒単位。
- -v, --vacuum-all —テストの前に 4 つの標準テーブルすべてをバキュームします。
- --aggregate-interval=NUM — NUM 秒以上のデータ。
- --failures-detailed —基本的なタイプごとにグループ化された障害をレポートします。
- --log-prefix=PREFIX —トランザクション時間ログファイルの接頭辞。デフォルトは pgbench_log です。
- --max-tries=NUM —トランザクションを実行しようとする最大試行回数。デフォルトは 1 です。
- --progress-timestamp —進行状況に Unix エポックタイムスタンプを使用します。
- --random-seed=SEED — ランダムシード time 、rand 、integer を設定します。
- --retry — フェイルオーバー時にトランザクションを再試行します。 -m とともに使用されます。
- --sampling-rate=NUM —ログに記録するトランザクションの割合、たとえば、1%の場合 0.01。
- --show-script=NAME —ビルトインスクリプトコードを表示して、終了します。
- --verbose-errors — すべてのエラーのメッセージを出力します。

#### 共通オプション

- -d, --debug — デバッグ出力を印刷します。
- -h, --host=HOSTNAME —データベースサーバーホストまたはソケットディレクトリ。
- -p, --port=PORT — データベースサーバーのポート番号。
- -U, --username=USERNAME —指定されたデータベースユーザーとして接続します。
- -V, --version —バージョン情報を出力して、終了します。
- -?, --help — ヘルプを表示して、終了します。

### Global sequence management interfaces

PGD は、標準の PostgreSQL シーケンスと PGD グローバルシーケンスの間で変換するためのインターフェイスを提供します。

次のファンクションは DDL であると考えられるため、DDL レプリケーションとグローバルロックが適用されます。

### シーケンスファンクション

#### bdr.alter_sequence_set_kind

シーケンスの所有者がシーケンスの種類を設定できます。設定すると、seqkind は bdr.sequences ビューを介してのみ可視されます。他のすべての方法では、シーケンスは通常のシーケンスとして表示されます。

PGD はこのファンクションを DDL として扱うため、現在アクティブな場合、DDL レプリケーションとグローバルロックが適用されます。 *Running DDL safely* を参照してください。

##### 概要

```
bdr.alter_sequence_set_kind(seqoid regclass, seqkind text, start bigint DEFAULT NULL)
```

##### パラメーター

- seqoid — 変更するシーケンスの名前または Oid。
- seqkind — 標準の PostgreSQL シーケンスの場合は local 、グローバルに一意の PGD シーケンスの場合は snowflakeid または galloc 、または古いグローバル一意シーケンスの場合は timeshard 。
- start —galloc およびローカルシーケンスの新しい開始点を指定できます。

##### 注

シーケンス種類を galloc に変更すると、そのシーケンスの最初に割り当てられたレンジは、開始点としてシーケンス start 値を使用します。galloc に変更される前にシーケンスで使用された既存の値が存在する場合、次のコマンドを使用して、新しく生成された値が既存の値と競合しないように、開始ポイントを移動することをお勧めします。

```
ALTER SEQUENCE seq_name START starting_value RESTART
```

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *DDL filters* の影響を受けることを意味します

構成。

ファンクションはグローバル DDL ロックを取得します。また、シーケンスをローカルにロックします。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

bdr.backwards_compatibility が 30618 以下に設定されていない限り、シーケンスの所有者のみが bdr.alter_sequence_set_kind ファンクションを実行できます。

#### bdr.extract_timestamp_from_snowflakeid

このファンクションは、snowflakeid シーケンスのタイムスタンプコンポーネントを抽出します。戻り値の型は timestamptz です。

#### 概要

```
bdr.extract_timestamp_from_snowflakeid(snowflakeid bigint)
```

#### パラメーター

- snowflakeid — snowflakeid シーケンスの値。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.extract_nodeid_from_snowflakeid

このファンクションは、snowflakeid シーケンスの nodeid コンポーネントを抽出します。

#### 概要

```
bdr.extract_nodeid_from_snowflakeid(snowflakeid bigint)
```

#### パラメーター

- snowflakeid — snowflakeid シーケンスの値。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.extract_localseqid_from_snowflakeid

このファンクションは、snowflakeid シーケンスのローカルシーケンス値コンポーネントを抽出します。

#### 概要

```
bdr.extract_localseqid_from_snowflakeid(snowflakeid bigint)
```

#### パラメーター

- snowflakeid — snowflakeid シーケンスの値。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.timestamp_to_snowflakeid

このファンクションは、タイムスタンプ値をダミーの snowflakeid シーケンス値に変換します。

これは、 snowflakeid 列および特定のタイムスタンプのインデックス付き検索または値の比較を行う場合に役立ちます。

たとえば、 snowflakeid シーケンスを使用している列 id を持つテーブル foo があると、次のように昨日の午前 0 時からの変更数を取得できます。

```
SELECT count(1) FROM foo WHERE id > bdr.timestamp_to_snowflakeid(yesterday)
```

この方法で定式化された照会は、列 id のインデックススキャンを使用します。

#### 概要

```
bdr.timestamp_to_snowflakeid(ts timestamptz)
```

#### パラメーター

- ts — snowflakeid シーケンス生成に使用するタイムスタンプ。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.extract_timestamp_from_timeshard

このファンクションは、timeshard シーケンスのタイムスタンプコンポーネントを抽出します。戻り値の型は timestamptz です。

#### 概要

```
bdr.extract_timestamp_from_timeshard(timeshard_seq bigint)
```

#### パラメーター

- timeshard_seq — timeshard シーケンスの値。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.extract_nodeid_from_timeshard

このファンクションは、timeshard シーケンスの nodeid コンポーネントを抽出します。

#### 概要

```
bdr.extract_nodeid_from_timeshard(timeshard_seq bigint)
```

#### パラメーター

- timeshard_seq — timeshard シーケンスの値。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.extract_localseqid_from_timeshard

このファンクションは、timeshard シーケンスのローカルシーケンス値コンポーネントを抽出します。

#### 概要

```
bdr.extract_localseqid_from_timeshard(timeshard_seq bigint)
```

#### パラメーター

- timeshard_seq — timeshard シーケンスの値。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.timestamp_to_timeshard

このファンクションは、タイムスタンプ値をダミーの timeshard シーケンス値に変換します。

これは、 timeshard 列および特定のタイムスタンプのインデックス付き検索または値の比較を行う場合に役立ちます。

たとえば、 timeshard シーケンスを使用している列 id を持つテーブル foo があると、次のように昨日の午前 0 時からの変更数を取得できます。

```
SELECT count(1) FROM foo WHERE id > bdr.timestamp_to_timeshard(yesterday)
```

この方法で定式化された照会は、列 id のインデックススキャンを使用します。

#### 概要

```
bdr.timestamp_to_timeshard(ts timestamptz)
```

#### パラメーター

- ts — timeshard シーケンス生成に使用するタイムスタンプ。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.galloc_chunk_info

このファンクションは、ローカルノードで galloc シーケンスに割り当てられたレンジを取得します。

シーケンスがローカルノードでまだアクセスされていない場合、空の結果セットが返されます。

指定されたシーケンス名が galloc シーケンスでない場合、 ERROR が発生します。

#### 概要

```
bdr.galloc_chunk_info(seqname regclass)
```

#### パラメーター

- seqname - 照会する galloc シーケンスの名前

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### KSUUID v2 ファンクション

KSUUID v2 データ、K-Sortable UUID データを操作するためのファンクション。 *KSUUID in the sequences documentation* も参照してください。

### bdr.gen_ksuuid_v2

このファンクションは、引数として渡されたタイムスタンプの値、または NULL が渡された場合は現在のシステム時間を使用して、新しい KSUUID v2 値を生成します。システム時間を使用して KSUUID を自動的に生成する場合は、NULL 引数を渡します。

戻り値の型は UUID です。

#### 概要

```
bdr.gen_ksuuid_v2(timestamptz)
```

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.ksuuid_v2_cmp

このファンクションは、KSUUID v2 値を比較します。

最初の値の方が新しい場合は 1、2 番目の値が小さい場合は-1、または等しい場合は 0 を返します。

#### 概要

```
bdr.ksuuid_v2_cmp(uuid, uuid)
```

#### パラメーター

- UUID — KSUUID 比較する v2。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.extract_timestamp_from_ksuuid_v2

このファンクションは、KSUUID v2 のタイムスタンプコンポーネントを抽出します。戻り値の型は timestamptz です。

#### 概要

```
bdr.extract_timestamp_from_ksuuid_v2(uuid)
```

#### パラメーター

- UUID — タイムスタンプを抽出する KSUUID v2 値。

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### KSUUID v1 ファンクション

KSUUID v1 データ、K-Sortable UUID データ v1 を操作するためのファンクション。非推奨 - 詳細については、*KSUUID in the sequences documentation* を参照してください。

### bdr.gen_ksuuid

このファンクションは、現在のシステム時間を使用して、新しい KSUUID v1 値を生成します。戻り値の型は UUID です。

#### 概要

```
bdr.gen_ksuuid()
```

#### 注

このファンクションは、ローカルノードでのみ実行されます。

### bdr.uuid_v1_cmp

このファンクションは、KSUUID v1 値を比較します。

最初の値の方が新しい場合は 1、2 番目の値が小さい場合は-1、または等しい場合は 0 を返します。

#### 概要

```
bdr.uuid_v1_cmp(uuid, uuid)
```

#### 注

このファンクションは、ローカルノードでのみ実行されます。

#### パラメーター

- UUID — KSUUID 比較する v1。

### bdr.extract_timestamp_from_ksuuid

このファンクションは、KSUUID v1 または UUIDv1 値のタイムスタンプコンポーネントを抽出します。戻り値の型は timestamptz です。

#### 概要

```
bdr.extract_timestamp_from_ksuuid(uuid)
```

#### パラメーター

- UUID — タイムスタンプを抽出する KSUUID v1 値。

#### 注

このファンクションはローカルノードで実行されます。

### Large object functions

bdr.lo_* ファンクションは、

PostgreSQL large object server-side functions とレプリケートされた同等のファンクションです。各ファンクションは pg_catalog.largeobject ではなく bdr.pgd_largeobject で動作します。特に記載のない限り、動作とパラメーターのセマンティクスは、同等の PostgreSQL と同じです。

レプリケートされたラージオブジェクトの概要については、*Replicated large objects* を参照してください。

### ラージオブジェクトインターフェイス

#### bdr.lo_creat

新しいラージオブジェクトを作成します。 mode 引数は無視されますが、互換性のために受け入れられます。新しいラージオブジェクトの OID を結果ます。

##### 概要

```
bdr.lo_creat(mode int4) returns oid
```

##### パラメーター

- mode — 無視されます。 PostgreSQL lo_creat インターフェイスとの互換性のために受け入れられます。

#### bdr.lo_create

指定された OID で新しいラージオブジェクトを作成します。ラージオブジェクトの OID を結果ます。

##### 概要

```
bdr.lo_create(lobjId oid) returns oid
```

##### パラメーター

- lobjId — 新しいラージオブジェクトに割り当てる OID。 0 を渡して、OID が自動的に割り当てられます。

#### bdr.lo_open

既存のラージオブジェクトを読み取りまたは書き込みのために開きます。後続の操作で使用するラージオブジェクトファイル記述子を返します。

##### 概要

```
bdr.lo_open(lobjId oid, mode int4) returns int4
```

##### パラメーター

- lobjId — 開くラージオブジェクトの OID。
- mode — アクセスモード。読み取りアクセスの場合は INV_READ 0x40000、書き込みアクセスの場合は INV_WRITE 0x20000、または両方を組み合わせて使用します。

### bdr.lo_close

ラージオブジェクトファイル記述子を閉じます。

#### 概要

```
bdr.lo_close(fd int4) returns int4
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。

### bdr.loread

現在のシーク位置にあるラージオブジェクトから最大 len バイトを読み取ります。

#### 概要

```
bdr.loread(fd int4, len int4) returns bytea
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。
- len — 読み取る最大バイト数。

### bdr.lowwrite

現在のシーク位置でラージオブジェクトにデータを書き込みます。書き込まれたバイト数を返します。

#### 概要

```
bdr.lowrite(fd int4, buf bytea) returns int4
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。
- buf — 書き込むデータ。

### bdr.lo_lseek

ラージオブジェクト記述子の現在のシーク位置を移動します。

#### 概要

```
bdr.lo_lseek(fd int4, offset int4, whence int4) returns int4
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。
- offset — シークするバイトオフセット。
- whence — シークモード最初から SEEK_SET 0、現在位置から SEEK_CUR 1、または最後から SEEK_END 2。

### bdr.lo_lseek64

bdr.lo_lseek の 64 ビットバージョン。

#### 概要

```
bdr.lo_lseek64(fd int4, offset int64, whence int4) returns int64
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。
- offset — シークするバイトオフセット。
- whence — シークモード SEEK_SET (0)、SEEK_CUR (1)、または SEEK_END (2)。

### bdr.lo_tell

ラージオブジェクト記述子の現在のシーク位置を返します。

#### 概要

```
bdr.lo_tell(fd int4) returns int4
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。

### bdr.lo_tell64

bdr.lo_tell の 64 ビットバージョン。

#### 概要

```
bdr.lo_tell64(fd int4) returns int64
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。

### bdr.lo_truncate

ラージオブジェクトを指定されたバイト単位の長さに切り捨てます。

#### 概要

```
bdr.lo_truncate(fd int4, len int4) returns int4
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。
- len — ラージオブジェクトの新しい長さのバイト単位。

### bdr.lo_truncate64

bdr.lo_truncate の 64 ビットバージョン。

#### 概要

```
bdr.lo_truncate64(fd int4, len int64) returns int4
```

#### パラメーター

- fd — bdr.lo_open によって返されるファイル記述子。
- len — ラージオブジェクトの新しい長さのバイト単位。

### bdr.lo_unlink

指定された OID を持つラージオブジェクトを削除します。

#### 概要

```
bdr.lo_unlink(lobjId oid) returns int4
```

#### パラメーター

- lobjId — 削除するラージオブジェクトの OID。

### bdr.lo_import

サーバーのファイルシステムからファイルを新しいラージオブジェクトとしてインポートします。新しいオブジェクトの OID を結果ます。スーパーユーザー権限が必要です。

#### 概要

```
bdr.lo_import(filename text) returns oid
```

#### パラメーター

- filename —サーバーのファイルシステムにインポートするファイルの絶対パス。

### bdr.lo_import_with_oid

特定の OID を持つラージオブジェクトとしてサーバーのファイルシステムからファイルをインポートします。スーパーユーザー権限が必要です。

#### 概要

```
bdr.lo_import_with_oid(filename text, lobjId oid) returns oid
```

#### パラメーター

- filename —サーバーのファイルシステムにインポートするファイルの絶対パス。
- lobjId — 新しいラージオブジェクトに割り当てる OID。

### bdr.lo_export

ラージオブジェクトをサーバーのファイルシステム上のファイルにエクスポートします。スーパーユーザー権限が必要です。

#### 概要

```
bdr.lo_export(lobjId oid, filename text) returns int4
```

#### パラメーター

- lobjId — エクスポートするラージオブジェクトの OID。
- filename — サーバーのファイルシステム上の宛先ファイルの絶対パス。

### bdr.lo_get

ラージオブジェクトの内容全体を bytea 値として結果ます。

#### 概要

```
bdr.lo_get(lobjId oid) returns bytea
```

#### パラメーター

- lobjId — 読み取るラージオブジェクトの OID。

### bdr.lo_get_fragment

ラージオブジェクトの一部を bytea 値として結果ます。

#### 概要

```
bdr.lo_get_fragment(lobjId oid, offset int64, length int4) returns bytea
```

#### パラメーター

- lobjId — 読み取るラージオブジェクトの OID。
- offset —読み取りを開始するラージオブジェクト内のバイトオフセット。
- length — 返される最大バイト数。

### bdr.lo_from_bytea

bytea 値から新しいラージオブジェクトを作成します。新しいオブジェクトの OID を結果ます。

#### 概要

```
bdr.lo_from_bytea(lobjId oid, data bytea) returns oid
```

#### パラメーター

- lobjId — 新しいラージオブジェクトに割り当てる OID。 0 を渡して、OID が自動的に割り当てられます。
- data — ラージオブジェクトの初期コンテンツ。

### bdr.lo_put

指定されたオフセットで開始してラージオブジェクトにデータを書き込み、必要に応じてオブジェクトを拡張します。

#### 概要

```
bdr.lo_put(lobjId oid, offset int64, data bytea) returns void
```

#### パラメーター

- lobjId — 書き込むラージオブジェクトの OID。
- offset —書き込みを開始するラージオブジェクト内のバイトオフセット。
- data — 書き込むデータ。

### Autopartition

オートパーティションを使用すると、テーブルを複数のパーティションに分割できます。詳細については、*AutoPartition in PGD* を参照してください。

### bdr.autopartition

bdr.autopartition ファンクションは、テーブルの自動 RANGE パーティションを構成します。

#### 概要

```
bdr.autopartition(relation regclass,
        partition_increment text,
        partition_initial_lowerbound text DEFAULT NULL,
        partition_autocreate_expression text DEFAULT NULL,
        minimum_advance_partitions integer DEFAULT 2,
        maximum_advance_partitions integer DEFAULT 5,
        data_retention_period interval DEFAULT NULL,
        enabled boolean DEFAULT on,
        analytics_offload_period interval DEFAULT NULL,
        drop_after_retention_period boolean DEFAULT true);
```

#### パラメーター

- relation _ テーブルの名前または Oid。
- partition_increment _ 次のパーティション作成までのインターバルまたはインクリメント。
- partition_initial_lowerbound _ テーブルにパーティションがない場合、この下限と partition_increment 離れた上限を持つ最初のパーティションが作成されます。
- partition_autocreate_expression _ 新しいパーティションを作成する時期が来たかどうかを検出するために使用される式。
- minimum_advance_partitions _ システムは、常に少なくとも minimum_advance_partitions パーティションを持とうとします。
- maximum_advance_partitions _ 事前パーティションの数が minimum_advance_partitions を下回った後に、一度に作成するパーティションの数。
- data_retention_period _ 古いパーティションが削除されるまでの間隔定義されている場合。この値は migrate_after_period より大きい必要があります。
- enabled _ アクティビティを無効または一時停止し、後で再開または再有効化できます。
- analytics_offload_period _ パーティションのオフロードのサポートを提供します。将来の使用のために予約されています。
- drop_after_retention_period _ パーティションをドロップする代わりにデタッチできるようにします。ドロップする代わりにデタッチするには、これを false に設定します。

#### 例

日次パーティション、データを 1 か月間保持します

```
CREATE TABLE measurement (
logdate date not null,
peaktemp int,
unitsales int
) PARTITION BY RANGE (logdate);

bdr.autopartition(measurement, 1 day, data_retention_period := 30 days);
```

あと 2 つのパーティションのみが残っている場合、5 つの高度なパーティションを作成します。各パーティションには、10 億の注文を保持できます。

```
bdr.autopartition(Orders, 1000000000,
        partition_initial_lowerbound := 0,
        minimum_advance_partitions := 2,
        maximum_advance_partitions := 5
    );
```

### bdr.drop_autopartition

bdr.drop_autopartition() を使用して、特定のリレーションの自動パーティションルールを削除します。リレーションの保留中の作業項目はすべて削除され、新しい作業項目は作成されません。

```
bdr.drop_autopartition(relation regclass);
```

#### パラメーター

- relation _ テーブルの名前または Oid。

### bdr.autopartition_wait_for_partitions

パーティションの作成は非同期プロセスです。AutoPartition は、ローカルまたはすべてのノードでパーティションが作成されるのを待機する一連の機能を提供します。

bdr.autopartition_wait_for_partitions() を使用して、ローカルノードでパーティションの作成を待機します。ファンクションは、パーティションテーブル名とパーティションキー列値を取得し、その値を保持するパーティションが作成されるまで待機します。

このファンクションは、パーティションがローカルに作成されるのをのみ待機します。パーティションがリモートノードにも存在することは保証されません。

すべての PGD ノードでパーティションが作成されるのを待機するには、bdr.autopartition_wait_for_partitions_on_all_nodes() ファンクションを使用します。このファンクションは、ローカルノードとすべてのリモートノードを内部的にチェックし、パーティションがどこでも作成されるまで待機します。

#### 概要

```
bdr.autopartition_wait_for_partitions(relation regclass, upperbound text);
```

#### パラメーター

- relation _ テーブルの名前または Oid。
- upperbound _ パーティションキー列の値。

### bdr.autopartition_wait_for_partitions_on_all_nodes

#### 概要

```
bdr.autopartition_wait_for_partitions_on_all_nodes(relation regclass, upperbound␣
↪text);
```

#### パラメーター

- relation _ テーブルの名前または Oid。
- upperbound _ パーティションキー列の値。

### bdr.autopartition_find_partition

bdr.autopartition_find_partition() ファンクションを使用して、指定されたパーティションキー値のパーティションを見つけます。その値を保持するパーティションが存在しない場合、ファンクションは NULL を結果ます。それ以外の場合、パーティションの Oid が返されます。

#### 概要

```
bdr.autopartition_find_partition(relname regclass, searchkey text);
```

#### パラメーター

- relname _ パーティションテーブルの名前。
- searchkey _ 検索するパーティションキー値。

### bdr.autopartition_enable

bdr.autopartition_enable を使用して、指定されたテーブルで AutoPartitioning を有効にします。AutoPartitioning が既に有効になっている場合、アクションは発生しません。特定のテーブルで自動パーティションを無効にするには、*bdr.autopartition_disable* を参照してください。

#### 概要

```
bdr.autopartition_enable(relname regclass);
```

#### パラメーター

- relname _ AutoPartitioning を有効にするリレーションの名前。

### bdr.autopartition_disable

bdr.autopartition_disable を使用して、指定されたテーブルで自動パーティションを無効にします。AutoPartitioning が既に無効になっている場合、アクションは発生しません。

#### 概要

```
bdr.autopartition_disable(relname regclass);
```

#### パラメーター

- relname _ AutoPartitioning を無効にするリレーションの名前。

#### 内部ファンクション

### bdr.autopartition_create_partition

AutoPartition は、内部ファンクション bdr.autopartition_create_partition を使用して、親テーブルにスタンドアロンの AutoPartition を作成します。

#### 概要

```
bdr.autopartition_create_partition(relname regclass,
                                   partname name,
                                     lowerb text,
                                     upperb text,
                                     nodes oid[]);
```

#### パラメーター

- relname _ アタッチする親テーブルの名前または Oid。
- partname _ 新しい AutoPartition の名前。
- lowerb _ パーティションの下限。
- upperb _ パーティションの上限。
- nodes _ 新しいパーティションが存在するノードのリスト。このパラメーターは PGD の内部であり、将来の使用のために予約されています。

#### 注

これは、AutoPartition がパーティション管理に使用する内部機能です。このファンクションを直接使用しないことをお勧めします。

### bdr.autopartition_drop_partition

AutoPartition は、内部ファンクション bdr.autopartition_drop_partition を使用して、データ保持ポリシーに従って、不要になったパーティションを削除します。パーティションテーブルのドロップが成功した場合、ファンクションは true を結果ます。

#### 概要

```
bdr.autopartition_drop_partition(relname regclass)
```

#### パラメーター

- relname _ ドロップするパーティションテーブルの名前。

#### 注

このファンクションは、選択したパーティションテーブルで DROP TABLE を使用する前に、親テーブルに DDL ロックを配置します。この機能は、AutoPartition がパーティション管理に使用する内部ファンクションです。このファンクションを直接使用しないことをお勧めします。

### Stream triggers reference

!!!関連項目

ストリームトリガーの概要については、 *Stream triggers* を参照してください。

PL/pgSQL では、次の事前定義された変数とファンクションを使用できます。

- *Stream triggers row variables*
- *Stream triggers row functions*
  - *bdr.trigger_get_row*
  - *bdr.trigger_get_committs*
  - *bdr.trigger_get_xid*
  - *bdr.trigger_get_type*
  - *bdr.trigger_get_conflict_type*
  - *bdr.trigger_get_origin_node_id*
  - edb_21

ストリームトリガーの作成とドロップは、操作インターフェイスを介して管理されます。

- *Stream triggers manipulation interfaces*

- *bdr.create_conflict_trigger*
- *bdr.create_transform_trigger*
- *bdr.drop_trigger*

### Stream triggers manipulation interfaces

ストリームトリガーは、 REPLICA IDENTITY FULL があるテーブル、または TOAST が適用される列のないテーブルにのみ作成できます。

### bdr.create_conflict_trigger

このファンクションは、新しい競合トリガーを作成します。

#### 概要

```
bdr.create_conflict_trigger(trigger_name text,
                            events text[],
                            relation regclass,
                            function regprocedure,
                            args text[] DEFAULT {})
```

#### パラメーター

- trigger_name _ 新しいトリガーの名前。
- events _ このトリガーを起動するイベントの配列。有効な値は’INSERT ‘、’UPDATE ‘、および’DELETE ’です。
- relation _ このトリガーを起動する関係。
- function _ 実行するファンクション。
- args _オプショナル。トリガーファンクションが実行時に受け取るパラメーターの配列 TG_ARGV 変数の内容を指定します。

#### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *ddl filters* 構成の影響を受けることを意味します。

ファンクションは、トリガーが作成されているリレーションでグローバル DML ロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

通常の PostgreSQL トリガーと同様に、bdr.create_conflict_trigger ファンクションには、relation の TRIGGER 特権とファンクションの EXECUTE 特権が必要です。これは、bdr.backwards_compatibility が 30619 以上の場合に適用されます。PGD では、追加のセキュリティルールが競合トリガーを含むすべてのトリガーに適用されます。*Security and roles* を参照してください。

### bdr.create_transform_trigger

このファンクションは、変換トリガーを作成します。

#### 概要

```
bdr.create_transform_trigger(trigger_name text,
                             events text[],
                             relation regclass,
                             function regprocedure,
                             args text[] DEFAULT {})
```

#### パラメーター

- trigger_name _ 新しいトリガーの名前。
- events _ このトリガーを起動するイベントの配列。有効な値は’INSERT ‘、’UPDATE ‘、および’DELETE ’です。
- relation _ このトリガーを起動する関係。
- function _ 実行するファンクション。
- args _オプショナル。トリガーファンクションが実行時に受け取るパラメーターの配列 TG_ARGV 変数の内容を指定します。

#### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *ddl filters* 構成の影響を受けることを意味します。

ファンクションは、トリガーが作成されているリレーションでグローバル DML ロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

通常の PostgreSQL トリガーと同様に、bdr.create_transform_trigger ファンクションには、relation の TRIGGER 特権とファンクションの EXECUTE 特権が必要です。PGD では、追加のセキュリティルールが変換トリガーを含むすべてのトリガーに適用されます。*Security and roles* を参照してください。

### bdr.drop_trigger

このファンクションは、既存のストリームトリガー競合と変換の両方を削除します。

#### 概要

```
bdr.drop_trigger(trigger_name text,
                 relation regclass,
                 ifexists boolean DEFAULT false)
```

#### パラメーター

- trigger_name _ 既存のトリガーの名前。
- relation _ トリガーが定義されているリレーション。
- ifexists _ true に設定されている場合、このファンクションは不足しているトリガーを無視します。

#### 注

このファンクションは、 DDL ステートメントと同じレプリケーションメカニズムを使用します。これは、レプリケーションが *ddl filters* 構成の影響を受けることを意味します。

ファンクションは、トリガーが作成されているリレーションでグローバル DML ロックを取得します。

このファンクションはトランザクションです。トランザクションの ROLLBACK を使用してエフェクトをロールバックできます。変更は現在のトランザクションに表示されます。

relation の所有者のみが bdr.drop_trigger ファンクションを実行できます。

### Stream triggers row functions

### bdr.trigger_get_row

このファンクションは、識別子で指定されたトリガー行の内容を RECORD として結果ます。このファンクションは、不適切に呼び出された場合、つまり操作タイプ TG_OP が DELETE のときに SOURCE_NEW で呼び出された場合、NULL を結果ます。

#### 概要

```
bdr.trigger_get_row(row_id text)
```

#### パラメーター

- row_id —行の識別子。トリガータイプと操作に応じて、SOURCE_NEW 、SOURCE_OLD 、および TARGET のいずれかです。個々のトリガータイプの説明を参照してください。

### bdr.trigger_get_committs

このファンクションは、識別子で指定されたトリガー行のコミットタイムスタンプを結果ます。行が固定されているか使用できないために使用できない場合、NULL を返します。行識別子 SOURCE_OLD の場合は、常に NULL を結果ます。

#### 概要

```
bdr.trigger_get_committs(row_id text)
```

#### パラメーター

- row_id —行の識別子。トリガータイプと操作に応じて、SOURCE_NEW 、SOURCE_OLD 、および TARGET のいずれかです。個々のトリガータイプの説明を参照してください。

### bdr.trigger_get_xid

このファンクションは、識別子で指定された TARGET 行のローカルトランザクション ID を結果ます。行が固定されているか使用できないために使用できない場合、NULL を返します。 SOURCE_OLD および SOURCE_NEW 行識別子の場合は、常に NULL を結果ます。

競合トリガーでのみ使用できます。

#### 概要

```
bdr.trigger_get_xid(row_id text)
```

#### パラメーター

- row_id —行の識別子。トリガータイプと操作に応じて、SOURCE_NEW 、SOURCE_OLD 、および TARGET のいずれかです。個々のトリガータイプの説明を参照してください。

### bdr.trigger_get_type

このファンクションは、現在のトリガータイプ、CONFLICT または TRANSFORM を結果ます。ストリームトリガーの外部で呼び出された場合、null を返します。

#### 概要

```
bdr.trigger_get_type()
```

### bdr.trigger_get_conflict_type

このファンクションは、競合トリガー内で呼び出された場合、現在の競合タイプを結果ます。それ以外の場合は、NULL を返します。

*Types of Conflict* を参照

このファンクションの考えられる戻り値については、

#### 概要

```
bdr.trigger_get_conflict_type()
```

### bdr.trigger_get_origin_node_id

このファンクションは、引数として渡されたトリガー row_id の起点に対応するノード ID を結果ます。オリジンが有効でない場合、行がローカルに生成されたことを意味します、トリガー行引数に応じて、ソースまたはターゲットノードのノード ID を結果ます。行識別子 SOURCE_OLD の場合は、常に NULL を結果ます。この機能を使用して、競合トリガーを定義して、常に信頼できるソースノードを優先できます。

#### 概要

```
bdr.trigger_get_origin_node_id(row_id text)
```

#### パラメーター

- row_id —行の識別子。トリガータイプと操作に応じて、SOURCE_NEW 、SOURCE_OLD 、および TARGET のいずれかです。個々のトリガータイプの説明を参照してください。

### bdr.ri_fkey_on_del_trigger

BEFORE トリガーとして呼び出される場合、このファンクションは FOREIGN KEY 情報を使用して FK 異常を回避します。

#### 概要

```
bdr.ri_fkey_on_del_trigger()
```

### Stream triggers row variables

#### TG_NAME

データ型の名前。この変数には、実際に起動されたトリガーの名前が含まれています。実際のトリガー名には、トリガーの作成中に指定された名前と比較して、_bdrt または_bdrc サフィックストリガータイプにより異なります。

#### TG_WHEN

データ型テキスト。この変数は、競合トリガーと変換トリガーの両方で BEFORE を示します。 bdr.trigger_get_type() 情報ファンクションを呼び出すことにより、ストリームトリガータイプを取得できます。*bdr.trigger_get_type* を参照してください。

#### TG_LEVEL

データ型 text ROW の文字列。

#### TG_OP

データタイプ text トリガーが起動されたオペレーションを識別する INSERT 、UPDATE 、または DELETE の文字列。

#### TG_RELID

データ型 oid トリガー呼び出しの原因となったテーブルのオブジェクト ID。

#### TG_TABLE_NAME

データ型名 トリガーの呼び出しを発生させたテーブルの名前。

#### TG_TABLE_SCHEMA

データ型名トリガー呼び出しの原因となったテーブルのスキーマの名前。パーティションテーブルの場合、これはルートテーブルの名前です。

#### TG_NARGS

データ型整数 `bdr.create_conflict_trigger()` または `bdr.create_transform_trigger()` ステートメントのトリガーファンクションに指定された引数の数。

#### TG_ARGV[]

テキストのデータ型配列 `bdr.create_conflict_trigger()` または `bdr.create_transform_trigger()` ステートメントの引数。インデックスは 0 からカウントします。無効なインデックス 0 未満または `TG_NARGS` 以上の場合、 `NULL` 値になります。

### Internal catalogs and views

カタログとビューは、アルファベット順にここにリストされています。

#### bdr.autopartition_partitions

自動パーティション機能によって作成されたパーティションに関する情報を保存する内部カタログテーブル。

##### bdr.autopartition_partitions 列

| Name | Type | Description |
|---|---|---|
| ap_parent_relid | oid | リレーションの OID |
| ap_part_relname | name | 作成したリレーションの名前 |
| ap_part_created_at | timestamp with time zone | 作成タイムスタンプ |
| ap_part_migrated_at | timestamp with time zone | 移行のタイムスタンプ |
| ap_part_dropped_at | timestamp with time zone | ドロップされたときのタイムスタンプ |

#### bdr.autopartition_rules

自動パーティションルールに関する情報を保存する内部カタログテーブル。

#### bdr.autopartition_rules 列

| Name | Type | Description |
|---|---|---|
| ap_partition_relid | oid | NA |
| ap_partition_relname | name | NA |
| ap_partition_schemaname | name | NA |
| ap_partition_increment_kir | "char" | NA |
| ap_secondary_tablespace | oid | NA |
| ap_maximum_advance_par | integer | NA |
| ap_is_autoscaled | boolean | NA |
| ap_latest_partitions | integer | NA |
| ap_enabled | boolean | NA |
| ap_migrate_after_period | interval | NA |
| ap_data_retention_period | interval | NA |
| ap_last_triggered | timestamp with time zone | NA |
| ap_partition_increment_val | text | NA |
| ap_partition_autocreate_ex | text | NA |
| ap_partition_initial_lowerb | text | NA |
| ap_partition_last_upperbou | text | NA |
| ap_partition_min_upperbou | text | NA |

### bdr.ddl_epoch

DDL エポックごとの状態を保持する内部カタログテーブル。

#### bdr.ddl_epoch 列

| Name | Type | Description |
|---|---|---|
| ddl_epoch | int8 | 単調増加のエポック番号 |
| origin_node_id | oid | このエポックの作成を要求したノードの内部ノード ID |
| epoch_consume_timeout | timestamptz | このエポックのタイムアウト |
| epoch_consumed | boolean | ローカルノードがエポックを完全に処理するとすぐに true に切り替えます |
| epoch_consumed_lsn | boolean | ローカルノードがエポックを処理した LSN |

### bdr.event_history

特定の PGD ノードのクラスターメンバーシップイベントを追跡する内部カタログテーブル。具体的には、以下を追跡します。

- ノードがクラスターに参加します
- Raft の状態が変化するときは常に、ノードがコンセンサスプロトコルでのロールを変更するたびに、リーダー、フォロワー、または候補からリーダーへ。*Monitoring Raft* コンセンサス を参照

- ワーカーがエラーアウトしたとき bdr.workers を参照してください

および *PGD レプリケーションワーカーのモニタリング* )

#### bdr.event_history 列

| Name | Type | Description |
|---|---|---|
| event_node_id | oid | イベントが参照するノードの ID |
| event_type | int | イベントのタイプノード、raft、またはワーカー関連イベント |
| event_sub_type | int | イベントのサブタイプ、つまり、参加、状態変更、またはエラーであるかどうか |
| event_source | text | イベントがソースされたワーカープロセスの名前 |
| event_time | timestamptz | イベントが発生したタイムスタンプ |
| event_text | text | イベントのテキスト表現たとえば、ワーカーのエラー |
| event_detail | text | イベントのより詳細な説明今のところ、ワーカーエラーにのみ関連します |

### bdr.event_summary

より人に優しい形式で情報を表示する bdr.event_history カタログのビュー。具体的には、イベントタイプとサブタイプを整数ではなくテキスト表現として表示します。

### bdr.local_leader_change

これは、リーダー変更履歴の最近の部分のローカルキャッシュです。これは、bdr.leader と同じフィールドがありますが、現在のバージョンのみを追跡するマップではなく、node_group_id、leader_kind、generate の順序セットである点が異なります。

### bdr.node_config

ノードごとの構成オプションを含む内部カタログテーブル。

#### bdr.node_config 列

| Name | Type | Description |
|---|---|---|
| node_id | oid | ノード ID |
| node_route_priority | int | このノードに割り当てられた優先度 |
| node_route_fence | boolean | このノードをフェンスするように切り替えます |
| node_route_writes | boolean | 書き込みを許可するスイッチ |
| node_route_reads | boolean | 読み取りを許可するようにスイッチ |
| node_route_dsn | text | このノードのインターフェイス |

次のページに続く

表 165 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| node_group_dsn | text | 同じサブグループ内のノードの優先接続文字列。設定すると、同じサブグループ内のノードは`node_dsn`の代わりにこの DSN を使用して接続します。デフォルトは空です。 *Replicated large objects* キーと`group_dsn`キーを使用して設定します。 |

### bdr.node_config_summary

より人間が読みやすい形式で情報を表示する bdr.node_config カタログのビュー。

#### bdr.node_config_summary 列

| Name | Type | Description |
|---|---|---|
| node_name | text | このノードの名前 |
| node_id | oid | ノード ID |
| node_route_priority | int | このノードに割り当てられた優先度 |
| node_route_fence | boolean | このノードをフェンスするように切り替えます |
| node_route_writes | boolean | 書き込みを許可するスイッチ |
| node_route_reads | boolean | 読み取りを許可するようにスイッチ |
| node_route_dsn | text | このノードのインターフェイス |
| node_group_dsn | text | 同じサブグループ内のノードの優先接続文字列。設定すると、同じサブグループ内のノードは`node_dsn`の代わりにこの DSN を使用して接続します。デフォルトは空です。 |
| effective_route_dsn | text | このノードの完全な DSN |
| effective_group_dsn | text | :ref:``node_group_dsn` <#bdrnode_config>` から派生した、このノードの実効グループ DSN。 |

### bdr.node_group_config

ノードごとのグループ構成オプションを含む内部カタログテーブル。

#### bdr.node_group_config 列

| Name | Type | Description |
|---|---|---|
| node_group_id | oid | ノードグループ ID。 |
| route_writer_max_lag | bigint | 最大書き込み遅延が受け入れられます。 |
| route_reader_max_lag | bigint | 受け入れられる最大読み取り遅延。 |
| route_writer_wait_flush | boolean | フラッシュを待機する必要がある場合に切り替えます。 |

次のページに続く

表 167 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| leader_progress_tracking_e | boolean | 各データノードの Raft リーダーへの定期的なレプリケーションの進行状況レポートを有効にします。これらのレポートは、書き込みリーダー選択中に最新のノードが選択されることを保証します。デフォルト `true` |
| leader_progress_tracking_p | interval | レプリケーションの進行状況レポートが Raft リーダーに送信される頻度。デフォルト `1s` |

### bdr.node_group_routing_config_summary

ノードグループごとのルーティング構成オプション。

#### bdr.node_group_routing_config_summary 列

| Name | Type | Description |
|---|---|---|
| node_group_name | name | ノードグループ名 |
| location | name | ノードグループのロケーション |
| enable_routing | boolean | グループルーティングは有効になっていますか？ |
| node_group_type | text | ノードグループタイプ"global"、"data"、または"subscriber-only"のいずれか |
| route_writer_max_lag | bigint | 受け入れられる最大書き込み遅延 |
| route_reader_max_lag | bigint | 受け入れられる最大読み取り遅延 |
| route_writer_wait_flush | boolean | フラッシュを待機する |

### bdr.node_group_routing_info

接続マネージャーの現在のルーティング情報を保持する内部カタログテーブル。

#### bdr.node_group_routing_info 列

| Name | Type | Description |
|---|---|---|
| node_group_id | oid | ノードグループ ID。 |
| write_node_id | oid | 現在の書き込みノード。 |
| prev_write_node_id | oid | 以前の書き込みノード。 |
| read_node_ids | oid[] | 読み取り専用ノード ID のリスト。 |
| record_version | bigint | レコードバージョン。ルーティングレコードが変更されるたびに 1 増加します。 |
| record_ts | timestamptz | Record_version の最後の更新のタイムスタンプ。 |
| write_leader_version | bigint | リーダーバージョンを書き込みます。 write_node_id が変更されるたびに record_version からコピーされます。 |

次のページに続く

表 169 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| write_leader_ts | timestamptz | 書き込みリーダーのタイムスタンプ。 write_node_id が変更されるたびに record_ts からコピーされます。 |
| read_nodes_version | bigint | ノードバージョンの読み取り。 read_node_ids リストが変更されるたびに、record_version からコピーされます。 |
| read_nodes_ts | timestamptz | ノードのタイムスタンプを読み取ります。 read_node_ids リストが変更されるたびに record_tw からコピーされます。 |

### bdr.node_group_routing_summary

情報をより分かりやすく表示する bdr.node_group_routing_info カタログのビュー。

#### bdr.node_group_routing_summary 列

| Name | Type | Description |
|---|---|---|
| node_group_name | name | ノードグループ名 |
| write_lead | name | 現在の書き込みリード |
| previous_write_lead | name | 以前の書き込みリード |
| read_nodes | name[] | 現在の読み取り専用ノード |

### bdr.node_peer_progress_mapping

グループスロットの移動を支援するために、他の各ノードに対するクラスター内の各ノードの相対的な進行状況を追跡するために使用される内部カタログ。これは、ノードの進行状況の追跡を支援するために PGD が使用する内部関係です。その内容は、主にシステム使用のためであり、レプリケーションの進行状況の正確な状態または正常性を推測するためのスタンドアロン参照として使用しないでください。

#### bdr.node_peer_progress_mapping 列

| Name | Type | Description |
|---|---|---|
| peer_node_id | oid | この進行状況情報が報告されたリモートピアノードの OID。 |
| local_lsn | pg_lsn | 進行状況が記録された時点のローカルノードの LSN。 |
| remote_lsn | pg_lsn | 進行状況が報告された時点の元のノードの LSN。 |

### bdr.node_routing_config_summary

ノードごとのルーティング構成オプションのわかりやすいビュー。oid および短いフィールド名ではなく、ノード名を表示します。

#### bdr.node_routing_config_summary 列

| Name | Type | Description |
|---|---|---|
| gid | name | ノード名 |
| route_priority | int | このノードに割り当てられた優先度 |
| route_fence | boolean | このノードをフェンスするように切り替えます |
| route_writes | boolean | 書き込みを許可するスイッチ |
| route_reads | boolean | 読み取りを許可するようにスイッチ |
| route_dsn | text | このノードのインターフェイス |

### bdr.prepared_xacts_summary

グループコミットスコープを使用するときにオリジンノードによって暗黙的に作成された準備されたトランザクションの診断情報を提供する内部ローカルノードビュー。

#### bdr.prepared_xacts_summary 列

| Name | Type | Description |
|---|---|---|
| gid | name | 暗黙的に準備されたトランザクションのグローバル識別子。 |
| origin_node_id | oid | トランザクションが発生したノードの OID。 |
| origin_xid | xid | 元ノードで準備されたトランザクションのトランザクション ID。 |
| origin_name | name | 元ノードの名前。 |
| commit_scope_name | name | トランザクションに使用されるコミットスコープの名前。 |
| origin_node_group | name | コミットスコープの元のノードグループに関連付けられたノードグループの名前。 |
| cs_id | oid | コミットスコープの内部 ID。 |
| transaction_state | name | 準備されたトランザクションの現在の状態。例、*waiting for prepared confirmation*。 |
| local_prepared_ts | timestamp with time zone | トランザクションがローカルノードで準備されたときのタイムスタンプ。 |
| local_xid | xid | ローカルノードで準備されたトランザクションのトランザクション ID。 |
| pid | int | 準備されたトランザクションを処理するプロセスの PID がアクティブな場合。 |
| decision | char | 準備されたトランザクションのコミット決定に達した場合。 |
| | | NA |

### bdr.reconcile_2pc_requests

2 フェーズコミット調整要求の現在の状態を保存する内部状態テーブル。ビュー *bdr.reconcile_2pc_requests_summary* は、このテーブルの人間が判読可能な表現形式を提供します。

#### bdr.reconcile_2pc_requests 列

| Name | Type | Description |
|---|---|---|
| origin_id | oid | 調整要求が作成されたオリジンノードの OID。 |
| coord_id | oid | リコンサイラーとして機能するコーディネーターノードの OID。コーディネーターが選択されるまで、空のままです。 |
| reconcile_status | integer | 調整要求のステータスコード。 |
| origin_lsn | pg_lsn | 割り当て時にコーディネーターから見た起点ノードの LSN。コーディネーターが選択されるまで、空のままです。 |
| origin_commit_ts | timestamptz | 割り当て時にコーディネーターが知っているオリジンのコミットタイムスタンプ。コーディネーターが選択されるまで、空のままです。 |
| reconcile_request_ts | timestamptz | リコンシリエーション要求がオリジン ID に登録されたときのタイムスタンプ。 |

### bdr.reconcile_2pc_requests_summary

*bdr.reconcile_2pc_requests* に保存されている調整要求の人が判読可能な概要を提供するビュー。

#### bdr.reconcile_2pc_requests_summary 列

| Name | Type | Description |
|---|---|---|
| origin | text | 調整要求が作成されたオリジンノードの名前。 |
| coordinator | text | 調整を実行するコーディネーターノードの名前。コーディネーターが選択されるまで、空のままです。 |
| origin_commit_ts | timestamptz | 割り当て時にコーディネーターが知っているオリジンのコミットタイムスタンプ。 |
| origin_lsn | pg_lsn | 割り当て時にコーディネーターから見た起点ノードの LSN。コーディネーターが選択されるまで、空のままです。 |
| reconcile_request_ts | timestamptz | リコンシリエーション要求がオリジンに登録されたときのタイムスタンプ。コーディネーターが選択されるまで、空のままです。 |
| reconcile_status | text | 調整要求のテキストステータス。例 *leader elect*、*start*、*done*、*failed* |

### bdr.sequence_kind

各非ローカルシーケンスのタイプを格納する内部状態テーブル。診断にはビュー bdr.sequences をお勧めします。

#### bdr.sequence_kind 列

| Name | Type | Description |
|---|---|---|
| seqid | oid | シーケンスの内部 OID |
| seqkind | char | 内部シーケンス種類 `l`=local,`t`=timeshard,`s`=snowflakeid,`g`=galloc |
| sequuid | name | シーケンスのグローバル識別子。 |

### bdr.sync_node_requests

ノード同期要求の状態を保存する内部状態テーブル。ビュー *bdr.sync_node_requests_summary*

は、このテーブルの人間が判読可能な表現形式を提供します。

#### bdr.sync_node_requests 列

| Name | Type | Description |
|---|---|---|
| sn_origin_node_id | oid | 同期する変更がある使用不可ノード |
| sn_target_node_id | oid | 元のノードの変更を含むノード |
| sn_source_node_id | oid | 同期要求のターゲットノード |
| sn_sync_start_lsn | pg_lsn | 同期要求の開始 LSN |
| sn_sync_start_ts | timestamptz | 同期要求の開始タイムスタンプ |
| sn_sync_end_lsn | pg_lsn | 同期要求の終了 LSN |
| sn_sync_end_ts | timestamptz | 同期要求の終了タイムスタンプ |
| sn_sync_status | text | 同期要求のステータス |

### bdr.sync_node_requests_summary

基になる *bdr.sync_node_requests* の人が判読可能なバージョンを提供するビュー

テーブル。

#### bdr.sync_node_requests_summary 列

| Name | Type | Description |
|---|---|---|
| origin | text | 同期する変更がある使用不可ノード |
| source | text | 元のノードの変更を含むノード |

次のページに続く

表 178 – 前のページからの続き

| Name | Type | Description |
|---|---|---|
| target | text | 同期要求のターゲットノード |
| sync_start_lsn | pg_lsn | 同期要求の開始 LSN |
| sync_start_ts | timestamptz | 同期要求の開始タイムスタンプ |
| sync_end_lsn | pg_lsn | 同期要求の終了 LSN |
| sync_end_ts | timestamptz | 同期要求の終了タイムスタンプ |
| sync_status | text | 同期要求のステータス |

### Internal system functions

以下は内部システム機能です。多くは、さまざまなビューを作成するときに使用されます。ファンクションを直接使用せず、代わりにそれらが提供するビューを使用することをお勧めします。

#### 一般的な内部ファンクション

#### bdr.bdr_get_commit_decions

共有メモリの状態を検査する便利なルーチン。

#### 概要

```
bdr.bdr_get_commit_decisions(dbid OID,
            origin_node_id OID,
            origin_xid xid,
            local_xid xid,
            decision "char",
            decision_ts timestamptz,
            is_camo boolean)
```

#### bdr.bdr_track_commit_Decision

トランザクションのコミットステータスを共有メモリハッシュテーブルに保存します。このファンクションは、アップグレードスクリプトによって使用され、bdr.commit_decions カタログに保存されたコミット決定を共有メモリハッシュテーブルに転送します。トランザクションのコミットステータスは WAL に記録されるため、WAL からリロードできます。

#### 概要

```
bdr.bdr_track_commit_decision(OID, xid, xid, "char", timestamptz, boolean);
```

#### bdr.consensus_kv_fetch

一貫した KV ストアから JSON 形式で値を取得します。

#### 概要

```
bdr.consensus_kv_fetch(IN key text) RETURNS jsonb
```

#### パラメーター

| Parameter | Description |
|---|---|
| *key* | 取得する任意のキー。 |

#### 注

このファンクションは内部ファンクションであり、主に HARP によって使用されます。

> **警告**
>
> ユーザアプリケーションではこのファンクションを使用しないでください。

### bdr.consensus_kv_store

一貫した KV ストアに値を保存します。

値の有効期限のタイムスタンプを返します。この機能は `ttl` に依存しています。 `ttl` が `NULL` の場合、このファンクションは `infinity` を結果ます。値が削除された場合、`-infinity` を返します。

#### 概要

```
bdr.consensus_kv_store(key text, value jsonb,
        prev_value jsonb DEFAULT NULL, ttl int DEFAULT NULL)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *key* | 挿入、更新、または削除する任意の一意のキー。 |
| *value* | 保存する JSON 値。 NULL の場合、既存のレコードは削除されます。 |
| *prev_value* | 設定されている場合、書き込み操作は、現在の値が `prev_value` と等しい場合にのみ実行されます。 |
| *ttl* | 新しい値の有効期間ミリ秒単位。 |

#### 注

これは内部ファンクションであり、主に HARP によって使用されます。

> **警告**
>
> ユーザアプリケーションではこの機能を使用しないでください。

### bdr.decode_message_payload

コンセンサスメッセージのペイロードをより人間が読みやすい出力にデコードする PGD メッセージペイロードファンクション。主に bdr.global_consensus_journal_details デバッグビューによって使用されます。

### bdr.decode_message_response_payload

コンセンサスメッセージへの応答のペイロードをより人間が読みやすい出力にデコードする PGD メッセージペイロードファンクション。主に bdr.global_consensus_journal_details デバッグビューによって使用されます。

### bdr.difference_fix_origin_create

引数として渡された指定された名前を使用して、レプリケーションオリジンを作成しますが、bdr_ 接頭辞を追加します。オリジンの内部 ID を結果ます。このファンクションは、pg_replication_origin_create() と同じ機能がありますが、このファンクションは postgres スーパーユーザー権限ではなく bdr_superuser を必要とする点を除きます。

### bdr.difference_fix_session_reset

現在のセッションをオリジンから再生しないものとしてマークし、本質的に bdr.difference_fix_session_setup() のエフェクトをリセットします。void を返します。このファンクションは、pg_replication_origin_session_reset() と同じ機能がありますが、このファンクションは postgres スーパーユーザー権限ではなく bdr_superuser を必要とする点を除きます。

#### 概要

```
bdr.difference_fix_session_reset()
```

### bdr.difference_fix_session_setup

現在のセッションを現在のオリジンから再生するものとしてマークします。ファンクションは、事前に作成された bdr_local_only_origin ローカルレプリケーションオリジンをセッションに暗黙的に使用します。これにより、再生の進行状況を報告でき、void を返します。このファンクションは、pg_replication_origin_session_setup() と同じ機能がありますが、このファンクションは postgres スーパーユーザー権限ではなく bdr_superuser を必要とする点が異なります。ファンクションの以前の形式 bdr.difference_fix_session_setup(text) は非推奨であり、将来のリリースでは削除される予定です。

#### 概要

```
bdr.difference_fix_session_setup()
```

### bdr.difference_fix_xact_set_avoid_conflict

現在のトランザクションを、LSN ' 0/0' およびタイムスタンプ' 2000-01-01' でコミットしたトランザクションを再生するものとしてマークします。このファンクションは、pg_replication_origin_xact_setup('0/0', '2000-01-01') と同じ機能がありますが、このファンクションは postgres スーパーユーザー権限ではなく bdr_superuser を必要とする点を除きます。

#### 概要

```
bdr.difference_fix_xact_set_avoid_conflict()
```

### bdr.drop_node

ノードのメタデータを削除します。

ノードが PARTED になると、そのメタデータはクラスターのノードに存在したままになります。たとえば、ノードがドロップされるまで、ノードは PARTED としてマークされ、bdr.node_sumary 結果に残ります。

bdr.drop_node('some node', force := true) の呼び出しは、分離中にノードがスタックした場合に必要かつ適切です。そのノードにまだ同期する必要があるデータがある場合、ドロップされるノードからのデータの同期をスキップすることに注意してください。ノードスタックパーティングが既に再イメージ化または削除されている場合、それに対して bdr.drop_node を呼び出して害はありません。これは、クラスター内のすべてのノードでこのスタックしたノードに対して呼び出す必要があることに注意して、ノードがドロップされたという一貫したビューを持つようにします。

このファンクションは、特定のノードのメタデータをローカルデータベースから削除します。ノードは次のいずれかです。

- ローカルノード。この場合、リモートノードに関する情報を含むすべてのノードのメタデータを削除します。
- リモートノード。その場合、その特定のノードのメタデータのみを削除します。

この内部機能の使用は次の場合に制限されます。

- EDB テクニカルサポートから指示された場合。
- ドキュメントで特に指示されている場所。

bdr.part_node を使用して、PGD グループからノードを削除します。このファンクションは、ノードを PARTED 状態に設定し、ノード名の再利用を有効にします。

```
bdr.drop_node(node_name text, cascade boolean DEFAULT false, force boolean DEFAULT␣
↪false)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_name* | 既存のノードの名前。 |
| *cascade* | 非推奨。将来のリリースで削除される予定です。 |
| *force* | 不一致を引き起こす可能性のある危険にもかかわらず、すべてのサニティチェックを回避し、指定された PGD ノードのすべてのメタデータの削除を強制します。テクニカルサポートのみが、分離に関連する緊急の場合に強制ノードドロップを使用します。 |

#### 注

このファンクションを実行する前に、`bdr.part_node()` を使用してノードを分割します。

このファンクションは、特定のノードのメタデータをローカルデータベースから削除します。ノードはローカルノードになることができ、その場合、リモートノードに関する情報を含むすべてのノードのメタデータが削除されます。または、リモートノードの場合、その特定のノードのメタデータのみが削除されます。

**注釈**

各ノードには、スノーフレーク ID および timeshard シーケンスで使用する一意のシーケンス番号が割り当てられているため、一度に最大 1024 のノードレコード ACTIVE と PARTED の両方を持つことができます。 PARTED ノードは自動的にクリーンアップされません。これが問題になる場合は、この機能を使用してこれらのレコードを削除できます。

### bdr.get_global_locks

ローカルノードで保持されているグローバルロックに関する情報を表示します。

`bdr.global_locks` ビューを実装して、ロックのより詳細な概要を提供するために使用されます。

### bdr.get_node_conflict_resolvers

ローカルノード上のすべての競合リゾルバーのテキスト文字列を表示します。

### bdr.get_slot_flush_timestamp

特定のレプリケーションスロットの最後のフラッシュ位置確認のタイムスタンプを取得します。

`bdr.node_slots` ビューを実装するために内部で使用されます。

### bdr.internal_alter_sequence_set_kind

さまざまなファンクション呼び出しのレプリケーションのために内部で使用されていたファンクション。PGD の現在のバージョンでは使用されなくなりました。ローリングアップグレード中の下位互換性のためのみに存在します。

### bdr.internal_replication_set_add_table

さまざまなファンクション呼び出しのレプリケーションのために内部で使用されていたファンクション。PGD の現在のバージョンでは使用されなくなりました。ローリングアップグレード中の下位互換性のためのみに存在します。

### bdr.internal_replication_set_remove_table

さまざまなファンクション呼び出しのレプリケーションのために内部で使用されていたファンクション。PGD の現在のバージョンでは使用されなくなりました。ローリングアップグレード中の下位互換性のためのみに存在します。

### bdr.internal_submit_join_request

新しいノードに参加するためのコンセンサス要求を送信します。

PGD グループの再構成の内部メカニズムに必要です。

### bdr.isolation_test_session_is_blocked

グローバルロック上のブロックのチェックを追加して、元の `pg_isolation_test_session_is_blocked` を拡張および実際に呼び出すヘルパーファンクション。

分離/同時実行テストに使用されます。

### bdr.local_node_info

PGD グループの再構成の内部メカニズムに必要なローカルノードの情報を表示します。

ビュー `bdr.local_node_summary` は、ユーザーの使用に役立つ同様の情報を提供します。

### bdr.msgb_connect

別のノードの接続プーラーに接続します。コンセンサスプロトコルによって使用されます。

### bdr.msgb_deliver_message

メッセージを別のノードの接続プーラーに送信します。コンセンサスプロトコルによって使用されます。

### bdr.node_catchup_state_name

name のキャッチアップ状態コードを変換します。

#### 概要

```
bdr.node_catchup_state_name(catchup_state oid);
```

#### パラメーター

| Parameter | Description |
|---|---|
| *catchup_state* | キャッチアップ状態の OID コード。 |

### bdr.node_kind_name

ノードの種類の人にわかりやすい名前データ|スタンバイ|ウィットネス|サブスクライバーのみを返します。

### bdr.peer_state_name

ノード状態 `node_state` をテキスト表現に変換します。主に `bdr.node_summary` ビューを実装するために使用されます。

### bdr.pg_xact_origin

指定されたトランザクションのオリジン ID を返します。

#### 概要

```
bdr.pg_xact_origin(xmin xid)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *xid* | オリジンが返されるトランザクション ID。 |

### bdr.reconcile_2pc_status_name

数値調整ステータスコードを、対応するテキスト表現に変換します。このファンクションは、主に *bdr.reconcile_2pc_requests_summary* ビューにデータを設定するために使用されます。

#### 概要

```
bdr.reconcile_2pc_status_name(reconcile_status oid) RETURNS text
```

#### パラメーター

| Parameter | Description |
|---|---|
| *reconcile_status* | 変換する特定のリコンシリエーションステータスを表す OID コード。 |

### bdr.request_replay_progress_update

’リプレイ進行状況の更新’Raft メッセージの即時書き込みを要求します。主にテスト目的で使用されますが、コンセンサスメカニズムが機能しているかどうかをテストするために使用することもできます。

### bdr.reset_relation_stats

bdr.stat_relation によって表示される、リレーション統計をリセットした後、ブール結果を返します。

### bdr.reset_subscription_stats

bdr.stat_subscription によって表示される、サブスクリプションによって作成された統計をリセットした後、ブール結果を返します。

### bdr.resynchronize_table_from_node

リモートノードから関係を再同期します。

#### 概要

```
bdr.resynchronize_table_from_node(node_name name, relation regclass)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *node_name* | リレーションデータをコピーまたは再同期するノード。 |
| *relation* | リモートノードからのコピーに関する関係。 |

#### 注

このファンクションは、リレーションのグローバル DML ロックを取得し、リレーションをローカルで切り捨て、リモートノードからデータをコピーします。

リレーションは、同じ名前と定義を持つ両方のノードに存在する必要があります。

以下がサポートされています。

- 同一のパーティション定義を使用したパーティションテーブルの再同期
- パーティション化されたテーブルと非パーティションテーブルへの再同期、およびその逆の再同期
- 外部キー制約を一時的に削除および再作成することによる参照先テーブルの再同期

参照されるテーブルでファンクションを実行した後、参照される列データが参照列値と一致しない場合、ファンクションはエラーをスローします。参照テーブルデータを再同期した後、ファンクションを再実行します。

さらに、リモートノードからデータをコピーした後、生成された列の値をローカルに計算することにより、生成された列とテーブルの再同期をサポートします。

現在、 row_filters はこのファンクションによって無視されます。

bdr.resynchronize_table_from_node ファンクションは、テーブルの所有者が bdr_superuser 権限を持っている場合にのみ実行できます。

### bdr.seq_currval

グローバルシーケンス操作の内部実装の一部。

galloc または snowflakeid シーケンスで currval() が呼び出されるときに、自動的に呼び出されます。

### bdr.seq_lastval

グローバルシーケンス操作の内部実装の一部。

galloc または snowflakeid シーケンスで lastval() が呼び出されるときに、自動的に呼び出されます。

### bdr.seq_nextval

グローバルシーケンスインクリメントの内部実装の一部。

galloc または snowflakeid シーケンスで nextval() が呼び出されるときに自動的に呼び出されます

### bdr.show_subscription_status

サブスクリプションのステータスに関する情報を取得します。主に bdr.subscription_summary ビューを実装するために使用されます。

### bdr.show_workers

bdr ワーカーに関連する情報。

#### 概要

```
bdr.show_workers(
    worker_pid int,
    worker_role int,
    worker_role_name text,
    worker_subid oid)
```

### bdr.show_writers

bdr.writers ビューで使用されるファンクション。

### bdr.sync_status_name

同期状態コードをテキスト表現に変換します。主に *bdr.sync_node_requests_summary* を実装するために使用されます

ビュー。

#### 概要

```
bdr.sync_status_name(sync_state oid)
```

#### パラメーター

| Parameter | Description |
|---|---|
| *sync_state* | 同期状態の OID コード。 |

### タスクマネージャー機能

#### bdr.taskmgr_set_leader

指定された node をタスクマネージャーリーダーノードになるように要求します。リーダーノードは、新しいタスクを作成します。現在、自動パーティションのみがこの機能を利用します。監視ノード、ロジカルスタンバイ、またはサブスクライバ専用ノードはリーダーになることはできません。このような要求はエラーで失敗します。

#### 概要

```
bdr.taskmgr_set_leader(node name, wait_for_completion boolean DEFAULT true);
```

#### bdr.taskmgr_get_last_completed_workitem

クラスター内のすべてのノードで正常に完了した最後の作業項目の id を返します。

#### 概要

```
bdr.taskmgr_get_last_completed_workitem();
```

#### bdr.taskmgr_work_queue_check_status

タスクを生成および終了するジョブを実行しているバックグラウンドワーカーのステータスを確認できます。

ステータスは次のビューから確認できます。

- bdr.taskmgr_work_queue_local_status
- bdr.taskmgr_work_queue_global_status

#### 概要

```
bdr.taskmgr_work_queue_check_status(workid bigint
                            local boolean DEFAULT false);
```

#### パラメーター

| Parameter | Description |
|---|---|
| *workid* | タスクのキー。 |
| *local* | ローカルステータスのみを確認します。 |

#### 注

Taskmgr ワーカーは、`bdr.autopartition` ファンクションが初めて呼び出される前でも、常にバックグラウンドで実行されています。無効な `workid` が使用される場合、ファンクションは `unknown` を結果ます。`In-progress` は一般的なステータスです。

### bdr.get_min_required_replication_slots

PGD-CLI によって使用される内部ファンクション。

### bdr.get_min_required_worker_processes

PGD-CLI によって使用される内部ファンクション。

### bdr.stat_last_reset_time

特定の統計カテゴリの最新のリセットタイムスタンプを取得するために使用される内部ファンクション。標準のモニタリングの場合、 *bdr.stat_reset* をサポートする各統計には、 *bdr.stat_commit_scope* ビューの `stats_reset` 列が含まれ、ユーザーが使用しやすい形式でこの情報を提供します。

### bdr.stat_get_activity

ビュー `bdr.stat_activity` の基になる内部ファンクション。直接使用しないでください。代わりに bdr.stat_activity ビューを使用します。

### bdr.stat_get_commit_scope

コミットスコープの統計を取得する内部ファンクション。標準のモニタリングとトラブルシューティングには、*bdr.stat_commit_scope* ビューを使用します。このビューは、同じ情報をよりアクセスしやすい形式で提供します。

### bdr.stat_get_commit_scope_proc_wait

特定のコミットスコープに関連付けられているプロセスの待機状態統計を取得するために使用される内部ファンクション。標準のモニタリングとトラブルシューティングには、*bdr.stat_commit_scope* ビューを使用します。これは、このデータのよりアクセスしやすい表現を提供します。

### bdr.stat_get_reconcile_commit_scope

コミットスコープのリコンシリエーション統計を取得するために使用される内部ファンクション。標準のモニタリングとトラブルシューティングには、*bdr.stat_reconcile_commit_scope* ビューを使用します。このビューは、このデータのよりアクセスしやすい使いやすい表現を提供します。

### bdr.worker_role_id_name

ビュー `bdr.worker_tasks` を生成するときに使用される内部ヘルパーファンクション。直接使用しないでください。代わりに bdr.worker_tasks ビューを使用します。

### bdr.lag_history

ビュー `bdr.node_replication_rates` を生成するときに使用される内部ファンクション。直接使用しないでください。代わりに bdr.node_replication_rates ビューを使用します。

### bdr.get_raft_instance_by_nodegroup

ビュー `bdr.group_raft_details` を生成するときに使用される内部ファンクション。直接使用しないでください。代わりに bdr.group_raft_details ビューを使用します。

### bdr.monitor_camo_on_all_nodes

ビュー `bdr.group_camo_details` を生成するときに使用される内部ファンクション。直接使用しないでください。代わりに bdr.group_camo_details ビューを使用します。

### bdr.monitor_raft_details_on_all_nodes

ビュー `bdr.group_raft_details` を生成するときに使用される内部ファンクション。直接使用しないでください。代わりに bdr.group_raft_details ビューを使用します。

### bdr.monitor_replslots_details_on_all_nodes

ビュー `bdr.group_replslots_details` を生成するときに使用される内部ファンクション。直接使用しないでください。代わりに bdr.group_replslots_details ビューを使用します。

### bdr.monitor_subscription_details_on_all_nodes

ビュー `bdr.group_subscription_summary` を生成するときに使用される内部ファンクション。直接使用しないでください。代わりに bdr.group_subscription_summary ビューを使用します。

### bdr.monitor_version_details_on_all_nodes

ビュー `bdr.group_versions_details` を生成するときに使用される内部ファンクション。直接使用しないでください。代わりに bdr.group_versions_details ビューを使用します。

#### bdr.node_group_member_info

ビュー bdr.group_raft_details を生成するときに使用される内部ファンクション。直接使用しないでください。代わりに bdr.group_raft_details ビューを使用します。

#### Column-level conflict functions

#### bdr.column_timestamps_create

このファンクションは、列レベルの競合解決を作成します。column_timestamp_enable 内で呼び出されます。

#### 概要

```
bdr.column_timestamps_create(p_source cstring, p_timestamp timestampstz)
```

#### パラメーター

- p_source _ 2 つのオプションは、current または commit です。
- p_timestamp _ タイムスタンプは選択したソースによって異なります。commit の場合は TIMESTAMP_SOURCE_COMMIT 。 current の場合は TIMESTAMP_SOURCE_CURRENT 。

### 10.1.2 EDB Postgres Distributed Command Line Interface (PGD CLI)

EDB Postgres 分散コマンドラインインターフェイス PGD CLI は、 EDB Postgres 分散クラスターを管理するためのツールです。これは、クラスターリソースを検査および管理するための重要なツールです。

EDB Postgres 分散クラスターに対してコマンドを実行して、次のことを行うことができます。

- クラスターの状態を判断し、クラスターの構成を検査し、クラスターのリソースを管理します。
- クラスターのノードとグループを検査および管理します。
- グループに対して書き込みリーダー変更操作を実行します。

PGD クラスターに接続できる Linux および macOS システムに手動でインストールすることもできます。

- HCP の高度な分散高可用性クラスター。
- CloudNative Postgres Global Clusters オペレーターを使用して展開された PGD クラスター。
- 手動で展開された PGD クラスター。

#### PGD エッセンシャルコンストレイン

PGD Essential クラスターには、特定の CLI コマンドに適用される制約があります。

- **クラスターアーキテクチャ** PGD Essential クラスターには、最大 3 つのデータノードがあり、グローバルグループのみにルーティングを有効にする必要があります。pgd cluster verify --arch を使用して、コンプライアンスを確認します。

- **互換性評価** pgd assess コマンドは、複数の一意のインデックス、行レベルのロックの使用率、ロックテーブルの使用率、および listen/notify の使用率を持つテーブルのチェックをスキップします。
- **コミットスコープコマンド** pgd commit-scope create 、pgd commit-scope update 、および pgd commit-scope drop はサポートされていません。
- **グループオプション** pgd group set-option は、enable_raft 、enable_routing 、enable_wal_decoder 、または streaming_mode を更新できません。
- **ノードのセットアップ** pgd node setup は、global グループのルーティングを有効にし、サブグループに対して無効にします。

### Installing PGD CLI

PGD CLI は、PGD クラスターに接続できる任意のシステムにインストールできます。Linux および macOS は、PGD CLI のインストールが現在サポートされているプラットフォームです。

### Installing PGD CLI on Linux

PGD CLI は、ほとんどの Linux ディストリビューションで利用できます。EDB アカウントでアクセスできる EDB リポジトリからインストールできます。PGD ユーザーと EDB Cloud Service ユーザー無料トライアルを含むユーザーは、EDB アカウントと PGD CLI にアクセスします。

#### EDB サブスクリプショントークンを取得する

これらのリポジトリからのダウンロードを有効にするには、トークンが必要です。トークンを取得するには、EDB Repos 2.0 にログインします。 EDB Repos 2.0 ページに初めてアクセスする場合は、 **Request Access** を選択してトークンを生成する必要があります。生成されたトークンが利用できるようになったら、**Copy** アイコンを選択してクリップボードにコピーするか、目のアイコンを選択して表示します。

#### EDB_SUBSCRIPTION_TOKEN 環境変数を設定する

トークンを取得したら、オペレーティングシステムに表示されているコマンドを実行して、<your-token> トークンを置き換えます。

```
export EDB_SUBSCRIPTION_TOKEN=<your-token>
```

次に、オペレーティングシステムに適切なコマンドを実行します。

#### Debian または Ubuntu

Debian または Ubuntu では、 apt パッケージマネージャーを使用して PGD CLI をインストールできます。

```
curl -1sSLf "https://downloads.enterprisedb.com/$EDB_SUBSCRIPTION_TOKEN/postgres_
↪distributed/setup.deb.sh" | sudo -E bash
```

このコマンドが curl: (22) The requested URL returned error: 404 のようなエラーを返した場合、正しいトークンを入力したことを確認してください。

コマンドが成功すると、次のような出力が表示されます。

```
Executing the  setup script for the enterprisedb/postgres_distributed repository ...
...
```

次のコマンドを使用して PGD CLI パッケージをインストールできるようになりました。

```
sudo apt-get install edb-pgd6-cli
```

### RHEL、Rocky、AlmaLinux、または Oracle Linux

RHEL、Rocky、AlmaLinux、または Oracle Linux では、 yum パッケージマネージャーを使用して PGD CLI をインストールできます。 RHEL 8 以降のデフォルトパッケージマネージャーである dnf パッケージマネージャーを使用することもできます。

```
curl -1sSLf "https://downloads.enterprisedb.com/$EDB_SUBSCRIPTION_TOKEN/postgres_
↪distributed/setup.rpm.sh" | sudo -E bash
```

このコマンドが curl:  (22) The requested URL returned error:  404 のようなエラーを返した場合、正しいトークンを入力したことを確認してください。

コマンドが成功すると、次のような出力が表示されます。

```
Executing the  setup script for the enterprisedb/postgres_distributed repository ...
...
```

次のコマンドを使用して PGD CLI パッケージをインストールできるようになりました。

```
sudo dnf install edb-pgd6-cli
```

```
sudo yum install edb-pgd6-cli
```

### Installing PGD CLI on macOS

PGD CLI は、macOS で Homebrew フォーミュラとして利用できます。インストールするには、次のコマンドを実行します。

```
brew tap enterprisedb/tap
brew install pgd-cli
```

> **注釈**
>
> macOS 用のスタンドアロン PGD CLI バイナリは、Apple Silicon ARM64 アーキテクチャでのみ利用できます。 !!!

インストールを確認するには、次を実行します。

```
pgd --version
```

*Using PGD CLI*

## Using PGD CLI

### PGD CLI とは何ですか?

PGD CLI は、PGD クラスターに接続して管理する便利な方法です。これを使用するには、PGD スーパーユーザー権限または同等のユーザーが必要です。スーパーユーザー権限を持つ PGD ユーザーは *bdr_superuser のロールとレプリケーション* です。同等のユーザーの例は、EDB Cloud Service 分散高可用性クラスターの edb_admin です。

- 

.pgpass にエントリを追加します。これには、ホスト、ポート、データベース名、ユーザー名、パスワードが含まれます。

- PGPASSWORD 環境変数にパスワードを設定する。
- 接続文字列にパスワードを含めます。

他のオプションは複数のデータベースでうまく拡張できないか、パスワードの機密性が危険にさらされるため、最初のオプションをお勧めします。

installed pgd-cli を取得したら、 pgd コマンドを実行して PGD コマンドラインインターフェイスにアクセスします。pgd コマンドには、ユーザー名とパスワードとともに、接続するホスト、ポート、データベースに関する詳細が必要です。

### データベース接続文字列を渡す

--dsn フラグを使用して、データベース接続文字列を pgd コマンドに渡します。 --dsn フラグを使用して接続文字列を渡す場合、構成ファイルは必要ありません。構成ファイルが存在する場合でも、フラグが優先されます。例

```
pgd nodes list --dsn "host=bdr-a1 port=5432 dbname=pgddb user=enterprisedb"
```

データベース接続文字列 DSN は、 PGD_CLI_DSN 環境変数を使用して設定することもできます。例

```
export PGD_CLI_DSN="host=bdr-a1 port=5432 dbname=pgddb user=enterprisedb"
pgd nodes list
```

--dsn フラグは環境変数より優先されるため、両方が設定されている場合、--dsn 値が使用されます。

コマンドオプションの説明については、 PGD CLI Command reference を参照してください。

### 構成ファイルの指定

pgd-cli-config.yml ファイルが/etc/edb/pgd-cli または$HOME/.edb/pgd-cli にある場合、pgd はそれを使用します。オプションの-f または--config-file フラグを使用して、この動作をオーバーライドできます。例

```
pgd nodes list -f /opt/my-config.yml
__OUTPUT__
Node Name Group Name   Node Kind Join State Node Status
- -------- ------------ --------- ---------- -----------
kaftan    dc1_subgroup data      ACTIVE     Up
kaolin    dc1_subgroup data      ACTIVE     Up
kaboom    dc1_subgroup data      ACTIVE     Up
```

#### 出力形式の指定

-o または--output フラグを使用して、デフォルトの出力形式を JSON に変更します。例

```
pgd nodes list -o json
[
  {
    "node_name": "kaftan",
    "node_group_name": "dc1_subgroup",
    "node_kind_name": "data",
    "join_state": "ACTIVE",
    "node_status": "Up",
    "node_id": 3490219809,
    "node_seq_id": 2,
    "node_local_dbname": "pgddb"
  },
  {
    "node_name": "kaolin",
    "node_group_name": "dc1_subgroup",
    "node_kind_name": "data",
    "join_state": "ACTIVE",
    "node_status": "Up",
    "node_id": 2111777360,
    "node_seq_id": 1,
    "node_local_dbname": "pgddb"
  },
  {
    "node_name": "kaboom",
    "node_group_name": "dc1_subgroup",
    "node_kind_name": "data",
    "join_state": "ACTIVE",
    "node_status": "Up",
    "node_id": 2710197610,
    "node_seq_id": 3,
    "node_local_dbname": "pgddb"
  }
]
```

PGD CLI は、次の出力形式をサポートしています。

| Setting | Format | Description |
|---|---|---|
| *simple* | Tabular | 単純な表形式ビュー。 (デフォルト)。 |
| *json* | JSON | フォーマットなしの生データを表示します。一部のコマンドでは、JSON 出力には、追加のフィールドや詳細なメッセージなど、表形式出力より多くのデータが表示される場合があります。 |
| *psql* | PSQL | PSQL 出力のスタイルの表形式ビュー。フォーマット。 |
| *modern* | Tabular | ボックス文字を使用してテーブルを区切る表形式ビュー。 |
| *markdown* | Markdown | `pgd assess`などの一部のコマンドの長い形式の非表形式出力を生成するマークダウンスタイルの出力。 |

### コマンドラインヘルプへのアクセス

サポートされているコマンドをリストするには、次を入力します。

```
pgd --help
```

特定のコマンドとそのパラメーターのヘルプが必要な場合は、`pgd <command_name> --help` と入力してください。例

```
pgd nodes list --help
```

### Configuring PGD CLI

PGD CLI は、PGD クラスターに接続できる任意のシステムにインストールできます。PGD CLI を使用するには、PGD スーパーユーザー権限または同等のユーザーが必要です。スーパーユーザー権限を持つ PGD ユーザーは *bdr_superuser* のロールとレプリケーション です。同等のユーザーの例は、EDB Cloud Service 分散高可用性クラスターの edb_admin です。

### PGD CLI とデータベース接続文字列

データベース接続文字列は必要ない場合があります。たとえば、Trused Postgres Architect はシステムに PGD CLI をインストールすると、PGD クラスターへの接続も構成します。これは、 PGD CLI が実行時にクラスターに接続できることを意味します。

PGD CLI を手動でインストールする場合は、 PGD CLI にデータベース接続文字列を指定して、どの PGD クラスターに接続するかを認識する必要があります。

- 

**.pgpass にエントリを追加します。これには、ホスト、ポート、データベース名、ユーザー名、パスワードが含まれます。**

- PGPASSWORD 環境変数にパスワードを設定する。 -

接続文字列にパスワードを含めます。

他のオプションは複数のデータベースでうまく拡張できないか、パスワードの機密性が危険にさらされるため、最初のオプションをお勧めします。

その情報を入手したら、続行できます。

#### 接続するデータベースの構成

PGD CLI は、 PGD CLI 構成ファイルまたはコマンドラインからデータベース接続情報を取得します。

#### コマンドラインでデータベース接続文字列を使用する

--dsn オプションを使用して、接続文字列を pgd に直接渡すことができます。詳細については、データベース接続文字列を渡す を参照してください。例

```
pgd --dsn "host=kaboom port=5432 user=enterprisedb dbname=pgddb" nodes show --versions
```

#### 環境変数でデータベース接続文字列を使用する

コマンドラインで接続文字列を渡す別の方法として、PGD_CLI_DSN 環境変数を接続文字列に設定できます。例

```
export PGD_CLI_DSN="host=kaboom port=5432 user=enterprisedb dbname=pgddb"
pgd nodes show --versions
```

#### 構成ファイルを使用する

pgd-cli-config.yml 構成ファイルを使用して、クラスターのデータベース接続文字列を指定します。構成ファイルには、クラスター内の少なくとも 1 つの PGD ノードのデータベース接続文字列が含まれている必要があります。クラスター名はオプショナルであり、検証されません。

例

```
cluster:
  name: cluster-name
  endpoints:
  - "host=host-1 port=5432 dbname=pgddb user=postgres"
  - "host=host-2 port=5432 dbname=pgddb user=postgres"
  - "host=host-3 port=5432 dbname=pgddb user=postgres"
```

デフォルトでは、pgd-cli-config.yml は/etc/edb/pgd-cli ディレクトリにあります。 PGD CLI は、次の場所で pgd-cli-config.yml を検索します。優先順位は高から低です。

1. /etc/edb/pgd-cli デフォルト 2.$HOME/.edb/pgd-cli

構成ファイルがこれらのディレクトリのいずれにもない場合は、 pgd コマンドでオプションの-f または--config-file フラグを使用して、構成として読み取るようにファイルを設定できます。データベース接続文字列を渡す を参照してください。

### Discovering connection strings

PGD CLI は、PGD クラスターに接続できる任意のシステムにインストールできます。PGD CLI を使用するには、PGD スーパーユーザー権限または同等のユーザーが必要です。スーパーユーザー権限を持つ PGD ユーザーは *bdr_superuser* のロールとレプリケーション です。同等のユーザーの例は、EDB Cloud Service 分散高可用性クラスターの edb_admin です。

### PGD CLI とデータベース接続文字列

データベース接続文字列は必要ない場合があります。たとえば、Trused Postgres Architect はシステムに PGD CLI をインストールすると、PGD クラスターへの接続も構成します。これは、 PGD CLI が実行時にクラスターに接続できることを意味します。

### データベース接続文字列の取得

PGD はさまざまな構成をサポートするため、すべての展開方法には、接続文字列を取得する異なる方法があります。通常、展開の構成から必要な情報を取得できます。次に、その情報を接続文字列にアセンブルできます。

### EDB CloudNative Postgres グローバルクラスターで展開されたクラスターの場合

EDB CloudNative Postgres グローバルクラスター CNPG-GC を使用している場合、接続文字列は展開の構成から取得されます。非常に柔軟なため、接続文字列を取得する方法は複数あります。これは、主に、デプロイメントの services の構成に依存します。

・ノードサービステンプレートを使用する場合、各ノードとプロキシサービスへの直接接続が利用できます。

- グループサービステンプレートを使用する場合、各グループへのゲートウェイサービスがあります。
- プロキシサービステンプレートを使用する場合、単一のプロキシがすべてのアプリケーションのクラスターへのエントリポイントを提供します。

**TODO [DOCS-1499] CNPG-GC が PGD6 CM を使用するように更新されたときにプロキシ参照を削除します**

構成ファイルを参照して、この情報を判断します。

ホスト名または IP アドレス、ポート、データベース名、およびユーザー名を確立します。デフォルトのデータベース名は pgddb です。デフォルトのユーザー名は、EDB Postgres Advanced Server の場合は enterprisedb であり、PostgreSQL および EDB Postgres Extended Server の場合は postgres です。

次に、その情報に基づいて接続文字列をアセンブルできます。

```
"host=<hostnameOrIPAddress> port=<portnumber> dbname=<databasename> user=<username>"
```

デプロイメントの構成で必要な場合、 sslmode=<sslmode> を追加します。

### Command reference

PGD コマンドラインインターフェイスのコマンド名は pgd です。

#### 概要

EDB Postgres 分散コマンドラインインターフェイス PGD CLI は、 EDB Postgres 分散クラスターを管理するツールです。 EDB Postgres 分散クラスターに対してコマンドを実行できます。これを使用して、クラスターリソースを検査および管理できます。

#### コマンド

- **クラスター構成** クラスターを管理するためのクラスターレベルのコマンド
- *pgd cluster show* クラスターレベルの情報を表示します。
- *pgd cluster verify* クラスターレベルの情報を確認します。
- **グループとレプリケーションセットの表示** グループを管理するためのグループレベルのコマンド
- *pgd cluster show* グループレベルの情報を表示します
- *pgd group set-option* グループレベルのオプションを設定します
- *pgd group get-option* グループレベルのオプションを取得します
- *pgd group set-leader* グループの書き込みリーダーを設定しますスイッチオーバー)。
- **グループとレプリケーションセットの表示** グループをリストするためのグループ関連コマンド。 - *pgd commit-scopes list* グループをリストします。
- **新しく参加したノードへのフェイルオーバー** ノードを管理するためのノードレベルのコマンド。
- *pgd* **ノードセットアップの実行** クラスター内のノードをセットアップします。
- *pgd cluster show* ノードレベルの情報を表示します。
- *pgd node part* アクティブなクラスターから PGD ノードを分割します。
- *pgd group set-option* オプション設定ノードレベル.
- *pgd group get-option* ノードレベルのオプションを取得します。
- *pgd node set-config* ノードレベルの構成を設定します。
- *pgd node get-config* ノードレベルの構成を取得します。
- *pgd node upgrade* PGD Postgres ノードのメジャーバージョンのアップグレードを実行します。
- *Node management* ノードをリストするためのノード関連のコマンド。 - *pgd commit-scopes list* ノードをリストします。
- **機能低下イベントの監視** イベントを表示するためのイベントログコマンド。 - *pgd cluster show* イベントを表示します。
- *EXPLAIN ANALYZE* **レプリケーション**レプリケーション関連、レプリケーションを管理するためのコマンド。 - *pgd cluster show* レプリケーション情報を表示します。

- コミット決定に *Raft* を使用する Raft コンセンサスを管理するための Raft 関連コマンド。
- *pgd cluster show* Raft state.
- に関する情報を表示します *pgd raft enable* 1 つ以上のノードで Raft コンセンサスワーカーを有効にします。
- *pgd raft disable* 1 つ以上のノードで Raft コンセンサスワーカーを無効にします。edb_0ran_3asis_29 1 つ以上のノードで Raft コンセンサスワーカーを無効にします。edb_0ran_3asis_28 : Raft リーダーシップを指定されたノードに転送します。
- *pgd raft sync-snapshot* すべてのノードで Raft の状態を同期します。
- *pgd commit-scope create* PGD コミットスコープを管理するためのコミットスコープ関連コマンド。
- *pgd cluster show* コミットスコープに関する情報を表示します。
- *pgd commit-scope create* コミットスコープを作成します。
- *pgd commit-scope update* コミットスコープを更新します。
- *pgd commit-scope drop* コミットスコープをドロップします.
- *pgd commit-scopes list* コミットスコープをリストします。
- クラスターの状態の評価 Postgres サーバーの PGD 互換性を評価します。
- *pgd completion* シェル完了スクリプトを生成します。

### グローバルオプション

すべてのコマンドは、次のグローバルオプションを受け入れます。

| Short | Long | Description |
|---|---|---|
| *-f* | *--config-file* | 構成ファイルへの名前/パス。これは--dsn フラグが存在する場合は無視されますデフォルト「/etc/edb/pgd-cli/pgd-cli-config.yml」 |
| | *--dsn* | データベース接続文字列 例 "host=bdr-a1 port=5432 db-name=pgddb user=postgres" `PGD_CLI_DSN`環境変数によっても設定されます。 |
| *-h* | *--help* | Help for pgd - 使用されるコマンドの特定のヘルプを表示します |
| *-o* | *--output* | 出力フォーマット *json*、*psql*、*modern*、*markdown*、*simple* 出力フォーマット を参照してください。 |

### 追加オプション

`pgd -V` を実行して、pgd CLI のバージョン情報を確認します。

#### 出力フォーマット

-o /--output オプションとともに使用します。

| Format | Description |
|---|---|
| simple | 単純なフォーマット-単純な ASCII テーブルとして出力しますデフォルト。 |
| json | JSON 形式 - JSON ドキュメントとして出力します、非表形式 |
| psql | PSQL フォーマット-PSQL のスタイルの ASCII テーブルとして出力 |
| modern | 最新のフォーマット-ボックスキャラクターを使用して表として出力 |
| markdown | マークダウンテーブルフォーマット - マークダウン互換の ASCII テーブルとして出力 |

### pgd assess

#### 概要

pgd assess コマンドは、EDB Postgres 分散クラスターへの移行に対する Postgres サーバーインスタンスの適合性を評価します。BDR 拡張機能がまだインストールされていない標準の Postgres インスタンスに対して実行するように設計されています。

コマンドは、評価する Postgres サーバーインスタンスに接続する DSN で実行する必要があります。--dsn オプションで渡すか、 PGD_CLI_DSN 環境変数を設定します。いずれも提供しないとエラーが発生します。

このコマンドは、Postgres インスタンスの EDB Postgres Distributed との互換性をチェックし、互換性レポートを生成します。チェック数は、ターゲットインスタンスで検出されたエディションによって異なります。標準の Postgres および PGD Expanded インスタンスは 11 個のチェックを取得しますが、 PGD Essential インスタンスは 7 個の Expanded 固有のチェックは省略されます。

#### 構文

```
pgd assess [OPTIONS]
```

#### オプション

asset コマンドには、コマンド固有のオプションはありません。ただし、評価する Postgres サーバーインスタンスに DSN を接続する必要があります。これは、--dsn オプションの引数として、または PGD_CLI_DSN 環境変数を設定して渡すことができます。

Global Options も参照してください。

#### 例

```
pgd assess
__OUTPUT__
Assessment                          | Result                     | Details
- -----------------------------------+----------------------------+------------------
 ↪----------------------------------------
```

(次のページに続く)

(前のページからの続き)

```
 Multiple Databases                  | Compatible                 | Found only one␣
↪user database
 Materialized Views                  | Compatible                 | No materialized␣
↪views found
 EPAS Queue Tables                   | Compatible                 | No EPAS Queue␣
↪Tables found
 DDL Command Usage                   | Requires workload analysis | Cannot be checked␣
↪automatically at this time
 Advisory Lock Usage                 | Potentially compatible     | No advisory lock␣
↪commands found in pg_stat_statements
 Large Objects                       | Compatible                 | No large objects␣
↪found
 Trigger/Reference Privileges        | Compatible                 | No triggers with␣
↪incompatible privileges found
 Tables with Multiple Unique Indexes | Compatible                 | No tables with␣
↪multiple unique indexes found
 LOCK TABLE Usage                    | Potentially compatible     | No LOCK TABLE␣
↪usage found in pg_stat_statements
 LISTEN/NOTIFY Usage                 | Compatible                 | No LISTEN/NOTIFY␣
↪usage found in pg_stat_statements
 Row-Level Lock Usage                | Potentially compatible     | No row-level␣
↪locking commands found in pg_stat_statements
```

### pgd cluster

`pgd cluster` コマンドは、 EDB Postgres 分散クラスターを管理するために使用されます。

#### サブコマンド

- *pgd cluster show* クラスターレベルの情報を表示します。
- *pgd cluster verify* クラスターレベルの情報を確認します。

### pgd cluster show

#### 概要

`pgd cluster show` コマンドは、 EDB Postgres 分散クラスターのクラスターレベルの情報を表示するために使用されます。

#### 構文

```
pgd cluster show [OPTIONS]
```

### オプション

次の表は、`pgd cluster show` コマンドで使用可能なオプションをリストしています。

| Short | Long | Description |
|---|---|---|
| | --clock-drift | 詳細なクロックドリフト情報のみを表示します。 |
| | --summary | クラスターの概要情報のみを表示します。 |
| | --health | クラスターの状態に関する情報のみを表示します。 |

一度に指定できるのは、上記のオプションのうちの 1 つだけです。

Global Options も参照してください。

### クロックドリフト

クロックドリフトの現在の実装は、このコマンドの実行中にクラスターが高負荷になっている場合、または多数のノードが含まれている場合、不正確な値を返す場合があることに注意してください。

| Symbol | Meaning |
|---|---|
| * | ok |
| ~ | warning (drift > 2 seconds) |
| *!* | critical (drift > 5 seconds) |
| *x* | down / unreachable |
| *?* | unknown |
| - | not applicable |

### 例

#### クラスター情報の表示

```
pgd cluster show
__OUTPUT__

#  Summary

Group Name    Parent Group  Group Type Node Name Node Kind
- ------------ ------------- ---------- --------- ---------
democluster                 global
dc1_subgroup  democluster   data       kaboom    data
dc1_subgroup  democluster   data       kaftan    data
dc1_subgroup  democluster   data       kaolin    data

#  Health

```

(次のページに続く)

(前のページからの続き)

```
Check             Status Details
- ---------------- ------ ----------------------------------------------
Connections       Ok     All BDR nodes are accessible
Raft              Ok     Raft Consensus is working correctly
Replication Slots Ok     All PGD replication slots are working correctly
Clock Drift       Ok     Clock drift is within permissible limit
Versions          Ok     All nodes are running the same PGD version

#  Clock Drift

Reference Node Node Name Clock Drift
- ------------- --------- -----------
kaftan         kaboom    *
kaftan         kaolin    *
```

### pgd cluster verify

#### 概要

`pgd cluster verify` コマンドは、 EDB Postgres 分散クラスターの構成を確認するために使用されます。

#### 構文

```
pgd cluster verify [OPTIONS]
```

#### オプション

次の表は、`pgd cluster verify` コマンドで使用可能なオプションをリストしています。

| Short | Long | Description |
|---|---|---|
| | --settings | クラスター内の Postgres 設定を確認する |
| | --arch | クラスターアーキテクチャの検証 |
| -v | --verbose | 詳細な出力を表示する |

オプションを設定しない場合、コマンドはデフォルトで `settings` と `arch` の両方を検証し、出力は詳細ではありません。

`--verbose` を`--arch` または`--settings` とともに使用して、クラスターステータスに関する詳細情報と具体的なアドバイスを含めます。

結果をインラインマークダウン形式で表示するには、グローバルオプション`-o markdown` を含めます。

Global Options も参照してください。

例

クラスター設定とアーキテクチャの確認

```
pgd cluster verify
__OUTPUT__

##  Architecture

Check                     Status Groups
- ----------------------- ------ ------
Cluster has data nodes    Ok
Witness nodes per group   Ok
Witness-only groups       Ok
Data nodes per group      Ok
Empty groups              Ok
Nodes have node kind set Ok

#  Settings

Setting Name                       Status
- ------------------------------ ------
bdr.accept_connections            Ok
bdr.ddl_locking                   Ok
bdr.max_writers_per_subscription Ok
bdr.raft_group_max_connections    Ok
bdr.replay_progress_frequency     Ok
bdr.role_replication              Ok
bdr.start_workers                 Ok
bdr.writers_per_subscription      Ok
bdr.xact_replication              Ok
max_connections                   Ok
max_prepared_transactions         Ok
max_replication_slots             Ok
max_wal_senders                   Ok
max_worker_processes              Ok
shared_preload_libraries          Ok
track_commit_timestamp            Ok
wal_level                         Ok
```

詳細な出力を使用してクラスターアーキテクチャを検証する

```
pgd cluster verify --verbose --arch
__OUTPUT__
 Check                        | Status  | Groups | Details
- ----------------------------+---------+--------+-----------------------------------
```

(次のページに続く)

(前のページからの続き)

```
↪----------------
 Cluster has data nodes      | Ok      |         |
 Max data nodes in a cluster | Ok      |         |
 Witness nodes per group     | Ok      |         |
 Witness-only groups         | Ok      |         |
 Data nodes per group        | Ok      |         |
 Routing enabled groups      | Warning | dc-1    | Recommended to enable routing only␣
↪on global-group dc-1
 Empty groups                | Ok      |         |
 Nodes have node kind set    | Ok      |         |
```

### クラスターアーキテクチャを確認し、出力マークダウン形式で表示

```
pgd cluster verify --arch -o markdown
__OUTPUT__
Cluster has data nodes
- ----------------------
Cluster must have at least one data node

Result: Ok

Max data nodes in a cluster
- --------------------------
A PGD Essential cluster must have at most three data nodes

Result: Ok

Witness nodes per group
- ----------------------
Each data group has at most one witness node

Result: Ok

Witness-only groups
- ------------------
Cluster has at most one witness-only group

Result: Ok

Data nodes per group
- -------------------
Each data group has at least one data node

Result: Ok
```

(次のページに続く)

(前のページからの続き)

```
Routing enabled groups
- ---------------------
A PGD Essential cluster must have routing enabled for global group only

Following groups have problems:

 - dc-1

Details:

 - Recommended to enable routing only on global-group dc-1

Result: Warning

Empty groups
- -----------
There are no empty groups

Result: Ok

Nodes have node kind set
- ------------------------
All nodes should have node_kind properly set

Result: Ok
```

### pgd commit-scope

pgd commit-scope コマンドは、 EDB Postgres 分散クラスターのコミットスコープを表示および管理するために使用されます。

#### サブコマンド

- *pgd cluster show* コミットスコープに関する情報を表示します。
- *pgd commit-scope create* コミットスコープを作成します。
- *pgd commit-scope update* コミットスコープを更新します。
- *pgd commit-scope drop* コミットスコープを削除します。
- *pgd commit-scopes list* コミットスコープをリストします。

### pgd commit-scope create

#### 概要

pgd commit-scope create コマンドは、 EDB Postgres 分散クラスターにコミットスコープを作成するために使用されます。

#### 構文

```
pgd commit-scope <COMMIT_SCOPE> create [OPTIONS] <RULE_DEFINITION> [GROUP_NAME]
```

<COMMIT_SCOPE> は、作成するコミットスコープの名前です。

<RULE_DEFINITION> は、コミットスコープを定義するルールです。ルールは、トランザクションがコミットされたとみなされるために満たされる必要がある条件を指定します。ルール構文の詳細については、*Commit scopes* および *Commit scope rules* を参照してください。

オプションの [GROUP_NAME] は、コミットスコープが属するグループの名前です。省略すると、デフォルトのトップレベルグループが設定されます。

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

##### コミットスコープの作成

次の例では、 dc1_subgroup グループにルール ANY 2 (dc1) on replicated group commit を使用して abc1 という名前のコミットスコープを作成します。

```
pgd commit-scope abc1 create "ANY 2 (dc1_subgroup) SYNCHRONOUS COMMIT" dc1_subgroup
__OUTPUT__
Command executed successfully
```

コミットスコープを確認します。

```
pgd commit-scope abc1 show
__OUTPUT__
Commit Scope Group Name   Rule Definition
- ----------- ------------ ---------------------------------------
abc1         dc1_subgroup ANY 2 (dc1_subgroup) SYNCHRONOUS COMMIT
```

##### トップレベルグループでのコミットスコープの作成

次の例では、トップレベルグループにルール ANY 2 (dc1_subgroup) SYNCHRONOUS COMMIT を使用して abc2 という名前のコミットスコープを作成します。

```
pgd commit-scope abc2 create "ANY 2 (dc1_subgroup) SYNCHRONOUS COMMIT"
__OUTPUT__
Command executed successfully
```

コミットスコープを確認します。

```
pgd commit-scope abc2 show
__OUTPUT__
Commit Scope Group Name  Rule Definition
- ----------- ----------- ---------------------------------------
abc2         democluster ANY 2 (dc1_subgroup) SYNCHRONOUS COMMIT
```

### pgd commit-scope drop

#### 概要

pgd commit-scope drop コマンドは、 EDB Postgres 分散クラスターからコミットスコープをドロップするために使用されます。

#### 構文

```
pgd commit-scope <COMMIT_SCOPE> drop [OPTIONS] [GROUP_NAME]
```

<COMMIT_SCOPE> は、ドロップするコミットスコープの名前です。

オプションの [GROUP_NAME] は、コミットスコープが属するグループの名前です。省略すると、デフォルトのトップレベルグループが設定されます。グループの名前は、コミットスコープが作成されたグループ名と一致する必要があることに注意してください。

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

#### コミットスコープのドロップ

次の例では、abc2 という名前のコミットスコープをトップレベルグループから削除します。

```
pgd commit-scope abc2 drop
__OUTPUT__
Command executed successfully
```

#### グループからコミットスコープを削除する

次の例は、 dc1_subgroup グループから abc1 という名前のコミットスコープを削除します。

```
pgd commit-scope abc1 drop dc1_subgroup
__OUTPUT__
Command executed successfully
```

### pgd commit-scopes list

#### 概要

pgd commit-scopes list コマンドは、EDB Postgres 分散クラスターのコミットスコープ情報をリストするために使用されます。

#### 構文

```
pgd commit-scopes list [OPTIONS] [GROUP_NAME]
```

GROUP_NAME は、コミットスコープをリストするグループの名前です。指定しない場合、クラスター内のすべてのコミットスコープがリストされます。

#### オプション

コマンド固有のオプションはありません。共通のグローバルオプションについては、グローバルオプション を参照してください。

#### 例

##### すべてのコミットスコープをリストする

```
pgd commit-scopes list
__OUTPUT__
 Commit Scope     | Group Name | Rule Definition
- -----------------+------------+-----------------------------------------------------
↪-------------------------------------------------------------------
 local protect    | pgdx6      | ASYNCHRONOUS COMMIT
 lag protect      | pgdx6      | MAJORITY ORIGIN GROUP LAG CONTROL (max_lag_time =␣
↪30s, max_commit_delay = 10s)
 majority protect | pgdx6      | MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT
 adaptive protect | pgdx6      | MAJORITY ORIGIN GROUP SYNCHRONOUS COMMIT DEGRADE ON␣
↪(timeout = 10s, require_write_lead = true) TO ASYNCHRONOUS COMMIT
 subgroup protect | group-a    | ALL (group-a) SYNCHRONOUS COMMIT
```

##### 指定されたグループのすべてのコミットスコープをリストする

```
pgd commit-scopes list group-a
__OUTPUT__
pgd commit-scopes list group-a
 Commit Scope     | Group Name | Rule Definition
```

(次のページに続く)

(前のページからの続き)

```
- -----------------+------------+---------------------------------
 subgroup protect | group-a    | ALL (group-a) SYNCHRONOUS COMMIT
```

### pgd commit-scope show

#### 概要

pgd commit-scope show コマンドは、 EDB Postgres 分散クラスターのコミットスコープに関する情報を表示するために使用されます。

#### 構文

```
pgd commit-scope <COMMIT_SCOPE> show [OPTIONS]
```

<COMMIT_SCOPE> は、情報を表示するコミットスコープの名前です。

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

#### コミットスコープの表示

次の例は、コミットスコープ abc1 に関する情報を示しています。

```
pgd commit-scope abc1 show
__OUTPUT__
Commit Scope Group Name   Rule Definition
- ----------- ------------ ---------------------------------------
abc1         dc1_subgroup ANY 2 (dc1_subgroup) SYNCHRONOUS COMMIT
```

Group Name 列は、コミットスコープが属するグループの名前を示します。この場合、コミットスコープは dc1_subgroup グループに属します。

Rule Definition 列は、コミットスコープを定義するルールを示します。この場合、ルールは ANY 2 (dc1) SYCHRONOUS COMMIT です。 dc1_subgroup グループはレプリケートされたグループであるため、コミットはグループ内の少なくとも 2 つのノードにレプリケートされ、その内の 2 つのノードはコミットされたとみなされる前にコミットを確認する必要があります。

### pgd commit-scope update

#### 概要

pgd commit-scope update コマンドは、 EDB Postgres 分散クラスターのコミットスコープを更新するために使用されます。

#### 構文

```
pgd commit-scope <COMMIT_SCOPE> update [OPTIONS] <RULE_DEFINITION>[GROUP_NAME]
```

<COMMIT_SCOPE> は、更新するコミットスコープの名前です。

<RULE_DEFINITION> は、コミットスコープを定義するルールです。ルールは、トランザクションがコミットされたとみなされるために満たされる必要がある条件を指定します。ルール構文の詳細については、*Commit scopes* および *Commit scope rules* を参照してください。

オプションの [GROUP_NAME] は、コミットスコープが属するグループの名前です。省略すると、デフォルトのトップレベルグループが設定されます。

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

#### コミットスコープの更新

次の例は、コミットスコープ abc1 をルール ANY 1 (dc1_subgroup) SYNCHRONOUS COMMIT で更新します。

```
pgd commit-scope abc1 update "ANY 1 (dc1_subgroup) SYNCHRONOUS COMMIT" dc1_subgroup
__OUTPUT__
Command executed successfully
```

#### トップレベルグループのコミットスコープの更新

次の例は、トップレベルグループのルール ANY 1 (dc1_subgroup) SYNCHRONOUS COMMIT でコミットスコープ abc2 を更新します。

```
pgd commit-scope abc2 update "ANY 1 (dc1_subgroup) SYNCHRONOUS COMMIT"
__OUTPUT__
Command executed successfully
```

### pgd completion

#### 概要

pgd completion コマンドは、 EDB Postgres 分散 CLI の完了設定を管理するために使用されます。

#### 構文

```
pgd completion <SHELL>
```

<SHELL> は、オートコンプリートスクリプトを生成するシェルです。

シェルの可能な値は bash 、 fish 、 zsh 、および powershell です。

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

```
pgd completion zsh
```

このコマンドは、通常、シェルセッションの起動ファイルの一部として評価されます。Zsh シェルの完了スクリプトを生成し、標準出力に書き込みます。したがって、 .zshrc ファイルに追加します。

```
eval "$(pgd completion zsh)"
```

### pgd events

pgd events コマンドは、 EDB Postgres 分散クラスターでイベントを表示するために使用されます。

#### サブコマンド

- *pgd cluster show* イベントを表示します。

### pgd events show

#### 概要

pgd events show コマンドは、 EDB Postgres 分散クラスターでイベントを表示するために使用されます。追加フラグがない場合、このコマンドは、すべてのノードとグループの最新の 20 件のイベントを表示します。

#### 構文

```
pgd events show [OPTIONS]
```

#### オプション

次の表は、pgd events show コマンドで使用可能なオプションをリストしています。

| Short | Long | Description |
|---|---|---|
| | *--node <NODE_NAME>* | 指定された名前のノードのイベントのみを表示します。 |
| | *--group <GROUP_NAME>* | 指定された名前のグループのイベントのみを表示します。 |
| *-n* | *--limit <LIMIT>* | 表示するイベントの数を制限します。デフォルト 20 |

Global Options も参照してください。

### ノードの状態

| State | Description |
|---|---|
| NONE | ワーカーが起動したときにノードの状態は設定解除され、現在の既知の状態にすぐに設定されることが予想されます。 |
| CREATED | bdr.create_node() は実行されましたが、ノードはまだ EDB Postgres 分散クラスターのメンバーではありません。 |
| JOIN_START | bdr.join_node_group() は、ローカルノードの既存の EDB Postgres 分散クラスターへの参加を開始します。 |
| JOINING | ノード参加が開始され、現在初期同期フェーズにあり、ノードにスキーマとデータを作成しています。 |
| CATCHUP | 初期同期フェーズが完了しました。現在、結合は、結合が開始されてから上流のピアノードで実行されたトランザクションを取得および適用する最後のステップにあります。 |
| STANDBY | ノードの参加は完了しましたが、変更のブロードキャストはまだ開始されていません。すべての結合はこの状態でしばらく時間を費やしますが、論理スタンバイとして定義されている場合、ノードはこの状態で継続します。 |
| PROMOTE | ノードはロジカルスタンバイであり、bdr.promote_node を呼び出してノードの状態を ACTIVE に移動したところです。これらの 2 つの PROMOTE 状態は、1 つのノードのみが STANDBY より高いが ACTIVE より低い状態になることができるという事実と一貫している必要があります。 |
| PROMOTING | ロジカルスタンバイからフル BDR ノードへの昇格が進行中です。 |
| ACTIVE | ノードは完全な BDR ノードであり、現在アクティブです。これは、最も一般的なノードのステータスです。 |
| PART_START | ノードがアクティブまたはスタンバイで、 bdr.part_node を呼び出して EDB Postgres 分散クラスターからノードを削除したところです。 |
| PARTING | ノードは他のノードから切断し、コンセンサスまたはレプリケーションでそれ以上の役割を果たしません。 |
| PART_CATCHUP | 非分割ノードは、最近分割されたノードからの欠落データを同期します。 |
| PARTED | これで、すべてのノードでノード分離操作が完了しました。 |

一度に 1 つのノードのみが PROMOTE または PROMOTING の状態になります。STANDBY は、ノードが読み取り専用状態にあることを示します。

### 例

#### 最新 5 件のイベントを表示します

```
$ pgd events show -n 5
__OUTPUT__
Event Time                       Event Observer Event Subject Event Source Event Type␣
↪Event Subtype Event Text    Event Detail
- ---------------------------- -------------- ------------- ------------ ---------- -
↪------------ ------------- ------------------------------------------------------------
↪---------------------------------------------
```

(次のページに続く)

(前のページからの続き)

```
2025-02-21 17:44:00.444902 UTC kaolin         kaftan        consensus    ROUTING   ␣
↪STATE_CHANGE  WRITE_LEADER  dc1_subgroup
2025-02-21 17:44:00.445080 UTC kaolin         kaolin        consensus    ROUTING   ␣
↪STATE_CHANGE  RAFT_LEADER   {"raft_leader":"kaolin","group_name":"dc1_subgroup",
↪"read_nodes_version":1,"read_nodes":"kaboom,kaolin"}
2025-02-21 17:44:00.452029 UTC kaftan         kaftan        consensus    ROUTING   ␣
↪STATE_CHANGE  LEADER_UPDATE
2025-02-21 17:44:00.456483 UTC kaboom         kaboom        consensus    ROUTING   ␣
↪STATE_CHANGE  LEADER_UPDATE
2025-02-21 17:44:00.456667 UTC kaolin         kaolin        consensus    ROUTING   ␣
↪STATE_CHANGE  LEADER_UPDATE
```

### pgd group

pgd group コマンドは、 EDB Postgres 分散クラスター内のグループを管理するために使用されます。

#### サブコマンド

- *pgd cluster show* グループレベルの情報を表示します。
- *pgd group set-option* グループレベルのオプションを設定します。
- *pgd group get-option* グループレベルのオプションを取得します。
- *pgd group set-leader* グループの書き込みリーダーを設定しますスイッチオーバーを実行します。

### pgd group show

#### 概要

pgd group show コマンドは、 EDB Postgres 分散クラスターでグループレベルの情報を表示するために使用されます。

#### 構文

```
pgd group <GROUP_NAME> show [OPTIONS]
```

<GROUP_NAME> は、情報を表示するグループの名前です。

#### オプション

次の表は、pgd cluster show コマンドで使用可能なオプションをリストしています。

| Long | Description |
|---|---|
| --summary | グループの概要のみを表示します。 |
| --options | グループのオプションのみを表示します。 |

次のページに続く

表 196 – 前のページからの続き

| Long | Description |
|---|---|
| --nodes | グループのノードのみを表示します。 |

一度に指定できるのは、上記のオプションのうちの 1 つだけです。

Global Options を参照してください。

### 例

#### グループ情報を表示する

```
pgd group group-a show
__OUTPUT__

#  Summary

 Group Property    | Value
- ------------------+---------
 Group Name        | group-a
 Parent Group Name | pgdx6
 Group Type        | data
 Write Leader      | pgd-a3
 Commit Scope      |

#  Nodes

 Node Name | Node Kind | Join State | Node Status
- ----------+-----------+------------+-------------
 pgd-a2    | data      | ACTIVE     | Up
 pgd-a3    | data      | ACTIVE     | Up
 pgd-a1    | data      | ACTIVE     | Up

#  Options

 Option Name                       | Option Value  | Option Source
- ----------------------------------+---------------+---------------
 analytics_autoadd_tables          | false         | group
 analytics_storage_location        |               | default
 apply_delay                       | 00:00:00      | inherited
 check_constraints                 | true          | inherited
 default_commit_scope              | local protect | group
 enable_raft                       | true          | group
 enable_routing                    | true          | group
 enable_wal_decoder                | false         | inherited
 http_port                         | 6434          | pg_config
```

(次のページに続く)

(前のページからの続き)

```
 location                          | London        | group
 num_writers                       | 2             | pg_config
 read_only_consensus_timeout       | 00:00:00      | pg_config
 read_only_max_client_connections  | 250           | pg_config
 read_only_max_server_connections  | 250           | pg_config
 read_only_port                    | 6433          | pg_config
 read_write_consensus_timeout      | 00:00:00      | pg_config
 read_write_max_client_connections | 250           | pg_config
 read_write_max_server_connections | 250           | pg_config
 read_write_port                   | 6432          | pg_config
 route_reader_max_lag              | -1            | default
 route_writer_max_lag              | -1            | default
 route_writer_wait_flush           | false         | group
 streaming_mode                    | null          | group
 use_https                         | true          | group
```

### グループの概要情報を表示する

```
pgd group group-a show --summary
__OUTPUT__
 Group Property    | Value
- ------------------+---------
 Group Name        | group-a
 Parent Group Name | pgdx6
 Group Type        | data
 Write Leader      | pgd-a3
 Commit Scope      |
```

### グループノード情報の表示

```
pgd group group-a show --nodes
__OUTPUT__
 Node Name | Node Kind | Join State | Node Status
- ----------+-----------+------------+-------------
 pgd-a1    | data      | ACTIVE     | Up
 pgd-a2    | data      | ACTIVE     | Up
 pgd-a3    | data      | ACTIVE     | Up
```

### グループオプション情報を表示する

```
pgd group group-a show --options
__OUTPUT__
 Option Name                       | Option Value  | Option Source
- ----------------------------------+---------------+---------------
```

(次のページに続く)

(前のページからの続き)

```
analytics_autoadd_tables          | false         | group
analytics_storage_location        |               | default
apply_delay                       | 00:00:00      | inherited
check_constraints                 | true          | inherited
default_commit_scope              | local protect | group
enable_raft                       | true          | group
enable_routing                    | true          | group
enable_wal_decoder                | false         | inherited
http_port                         | 6434          | pg_config
location                          | London        | group
num_writers                       | 2             | pg_config
read_only_consensus_timeout       | 00:00:00      | pg_config
read_only_max_client_connections  | 250           | pg_config
read_only_max_server_connections  | 250           | pg_config
read_only_port                    | 6433          | pg_config
read_write_consensus_timeout      | 00:00:00      | pg_config
read_write_max_client_connections | 250           | pg_config
read_write_max_server_connections | 250           | pg_config
read_write_port                   | 6432          | pg_config
route_reader_max_lag              | -1            | default
route_writer_max_lag              | -1            | default
route_writer_wait_flush           | false         | group
streaming_mode                    | null          | group
use_https                         | true          | group
```

### グループ情報を JSON として表示する

```
pgd group group-a show -o json
__OUTPUT__
[
  {
    "Summary": [
      {
        "info": "Group Name",
        "value": "group-a"
      },
      {
        "info": "Parent Group Name",
        "value": "pgdx6"
      },
      {
        "info": "Group Type",
        "value": "data"
      },
```

(次のページに続く)

(前のページからの続き)


```
      {
        "info": "Write Leader",
        "value": "pgd-a3"
      },
      {
        "info": "Commit Scope",
        "value": ""
      }
    ]
  },
  {
    "Nodes": [
      {
        "join_state": "ACTIVE",
        "node_kind_name": "data",
        "node_name": "pgd-a2",
        "node_status": "Up"
      },
      {
        "join_state": "ACTIVE",
        "node_kind_name": "data",
        "node_name": "pgd-a3",
        "node_status": "Up"
      },
      {
        "join_state": "ACTIVE",
        "node_kind_name": "data",
        "node_name": "pgd-a1",
        "node_status": "Up"
      }
    ]
  },
  {
    "Options": [
      {
        "option_name": "analytics_autoadd_tables",
        "option_source": "group",
        "option_value": "false"
      },
      {
        "option_name": "analytics_storage_location",
        "option_source": "default",
        "option_value": ""
      },
      {
```


(次のページに続く)

(前のページからの続き)


```
      "option_name": "apply_delay",
      "option_source": "inherited",
      "option_value": "00:00:00"
    },
    {
      "option_name": "check_constraints",
      "option_source": "inherited",
      "option_value": "true"
    },
    {
      "option_name": "default_commit_scope",
      "option_source": "group",
      "option_value": "local protect"
    },
    {
      "option_name": "enable_raft",
      "option_source": "group",
      "option_value": "true"
    },
    {
      "option_name": "enable_routing",
      "option_source": "group",
      "option_value": "true"
    },
    {
      "option_name": "enable_wal_decoder",
      "option_source": "inherited",
      "option_value": "false"
    },
    {
      "option_name": "http_port",
      "option_source": "pg_config",
      "option_value": "6434"
    },
    {
      "option_name": "location",
      "option_source": "group",
      "option_value": "London"
    },
    {
      "option_name": "num_writers",
      "option_source": "pg_config",
      "option_value": "2"
    },
    {
```


(次のページに続く)

(前のページからの続き)


```
      "option_name": "read_only_consensus_timeout",
      "option_source": "pg_config",
      "option_value": "00:00:00"
    },
    {
      "option_name": "read_only_max_client_connections",
      "option_source": "pg_config",
      "option_value": "250"
    },
    {
      "option_name": "read_only_max_server_connections",
      "option_source": "pg_config",
      "option_value": "250"
    },
    {
      "option_name": "read_only_port",
      "option_source": "pg_config",
      "option_value": "6433"
    },
    {
      "option_name": "read_write_consensus_timeout",
      "option_source": "pg_config",
      "option_value": "00:00:00"
    },
    {
      "option_name": "read_write_max_client_connections",
      "option_source": "pg_config",
      "option_value": "250"
    },
    {
      "option_name": "read_write_max_server_connections",
      "option_source": "pg_config",
      "option_value": "250"
    },
    {
      "option_name": "read_write_port",
      "option_source": "pg_config",
      "option_value": "6432"
    },
    {
      "option_name": "route_reader_max_lag",
      "option_source": "default",
      "option_value": "-1"
    },
    {
```


(次のページに続く)

(前のページからの続き)


```
        "option_name": "route_writer_max_lag",
        "option_source": "default",
        "option_value": "-1"
      },
      {
        "option_name": "route_writer_wait_flush",
        "option_source": "group",
        "option_value": "false"
      },
      {
        "option_name": "streaming_mode",
        "option_source": "group",
        "option_value": "null"
      },
      {
        "option_name": "use_https",
        "option_source": "group",
        "option_value": "true"
      }
    ]
  }
]
```


### pgd group set-option

#### 概要

`pgd group set-option` コマンドは、EDB Postgres 分散クラスターでグループレベルのオプションを設定するために使用されます。

#### 構文

```
pgd group <GROUP_NAME> set-option [OPTIONS] <OPTION> <VALUE>
```

<GROUP_NAME> は、オプションを取得するグループの名前です。

<OPTION> は、取得する特定のグループオプションの名前であり、<VALUE> は、それに設定する値です。次のオプションが使用できます。

#### グループオプション

| Option | Description |
|---|---|
| apply_delay | グループに変更を適用する際の遅延。 |
| check_constraints | グループ内のコンストレインをチェックするかどうか。 |

次のページに続く

表 197 – 前のページからの続き

| Option | Description |
|---|---|
| default_commit_scope | グループのデフォルトのコミットスコープ。 |
| enable_routing | グループ内のルーティングを有効にするかどうか。 |
| enable_raft | グループで Raft を有効にするかどうか。 |
| enable_wal_decoder | グループで WAL デコーダーを有効にするかどうか。 |
| location | グループの場所。 |
| num_writers | グループ内のライターの数。 |
| route_reader_max_lag | グループ内のリーダーの最大遅延。 |
| route_writer_max_lag | グループ内のライタの最大遅延。 |
| streaming_mode | グループのストリーミングモード。 |
| route_writer_wait_flush | グループ内のライタをフラッシュするための待機時間。 |
| default_seqkind | グループのデフォルトのシーケンスの種類。 |
| default_replica_identity | グループのデフォルトのレプリカ ID。 |
| conflict_detection_method | グループの競合検出方法。 |
| replay_progress_frequency | グループの再生進行頻度。 |
| batch_inserts | グループ内のバッチ挿入を有効にするかどうか。 |
| analytics_storage_location | グループ内の分析の保存場所。 |
| analytics_autoadd_tables | グループ内の分析にテーブルを自動的に追加するかどうか。 |

### グループ接続マネージャーオプション

| Option | Description |
|---|---|
| read_write_port | 読み取り/書き込み接続をリッスンするポート |
| read_only_port | 読み取り専用接続をリッスンするポート |
| http_port | REST API 呼び出しをリッスンする http ポート統合の目的 |
| use_https | http リスナーが HTTPS を使用するかどうか。有効になっている場合、サーバー証明書は TLS に使用されます |
| server_pool_mode | *Connection pooling* モード *none* プーリングなし *session*、または `transaction` |
| read_write_max_client_co | 許可される最大読み取り/書き込みクライアント接続、デフォルトは max_connections |
| read_write_max_server_cc | サーバーに対して開かれる最大読み取り/書き込み接続 |
| read_only_max_client_cor | 許可される最大読み取り専用クライアント接続 |
| read_only_max_server_co | サーバーに対して開かれる読み取り専用接続の最大数 |
| read_write_consensus_tim | 読み取り/書き込み接続が受け入れられなくなるまでにコンセンサスが失われるまで待機する時間 |
| read_only_consensus_time | 読み取り専用接続が受け入れられなくなるまでに、コンセンサス損失を待機する時間。 |

#### グループプロキシオプション PGD 5.x のみ

| Option | Description |
|---|---|
| proxy_listen_address | グループ内のプロキシのリッスンアドレス。 |
| proxy_listen_addresses | グループ内のプロキシのリッスンアドレス。 |
| proxy_listen_port | グループ内のプロキシのリッスンポート。 |
| proxy_max_client_conn | グループ内のプロキシのクライアントの最大接続数。 |
| proxy_max_server_conn | グループ内のプロキシのサーバー接続の最大数。 |
| proxy_server_conn_timeou | グループ内のプロキシのサーバー接続タイムアウト。 |
| proxy_server_conn_keepal | グループ内のプロキシのサーバー接続キープアライブ。 |
| proxy_fallback_node_grou | グループ内のプロキシのフォールバックノードグループ。 |
| proxy_fallback_node_grou | グループ内のプロキシのフォールバックノードグループタイムアウト。 |
| proxy_consensus_grace_pe | グループ内のプロキシのコンセンサス猶予期間。 |
| proxy_read_listen_address | グループ内の読み取りプロキシのリッスンアドレス。 |
| proxy_read_listen_address | グループ内の読み取りプロキシのリッスンアドレス。 |
| proxy_read_listen_port | グループ内の読み取りプロキシのリッスンポート。 |
| proxy_read_max_client_cc | グループ内の読み取りプロキシのクライアントの最大接続数。 |
| proxy_read_max_server_c | グループ内の読み取りプロキシのサーバー接続の最大数。 |
| proxy_read_server_conn_k | グループ内の読み取りプロキシのサーバー接続キープアライブ。 |
| proxy_read_server_conn_t | グループ内の読み取りプロキシのサーバー接続タイムアウト。 |
| proxy_read_consensus_gra | グループ内の読み取りプロキシのコンセンサス猶予期間。 |

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

#### グループの場所を設定する

```
pgd group dc1_subgroup set-option location London
__OUTPUT__
Command executed successfully
```

#### スペースを含む値にオプションを設定する

```
pgd group dc1_subgroup set-option location "New York"
__OUTPUT__
Command executed successfully
```

### pgd group get-option

#### 概要

`pgd group get-option` コマンドは、 EDB Postgres 分散クラスターでグループレベルのオプションを取得するために使用されます。

### 構文

```
pgd group <GROUP_NAME> get-option [OPTIONS] <OPTION>
```

<GROUP_NAME> は、オプションを取得するグループの名前です。

<OPTION> は、取得する特定のグループオプションの名前です。オプションを指定しない場合、コマンドはすべてのグループオプションをリストします。次のオプションが使用できます。

### グループオプション

| Option | Description |
|---|---|
| apply_delay | グループに変更を適用する際の遅延。 |
| check_constraints | グループ内のコンストレインをチェックするかどうか。 |
| default_commit_scope | グループのデフォルトのコミットスコープ。 |
| enable_routing | グループ内のルーティングを有効にするかどうか。 |
| enable_raft | グループで Raft を有効にするかどうか。 |
| enable_wal_decoder | グループで WAL デコーダーを有効にするかどうか。 |
| location | グループの場所。 |
| num_writers | グループ内のライターの数。 |
| route_reader_max_lag | グループ内のリーダーの最大遅延。 |
| route_writer_max_lag | グループ内のライタの最大遅延。 |
| streaming_mode | グループのストリーミングモード。 |
| route_writer_wait_flush | グループ内のライタをフラッシュするための待機時間。 |
| analyt-ics_storage_location | グループ内の分析の保存場所。 |
| analytics_autoadd_tables | グループ内の分析にテーブルを自動的に追加するかどうか。 |

### グループ接続マネージャーオプション

| Option | Description |
|---|---|
| read_write_port | 読み取り/書き込み接続をリッスンするポート |
| read_only_port | 読み取り専用接続をリッスンするポート |
| http_port | REST API 呼び出しをリッスンする http ポート統合の目的 |
| use_https | http リスナーが HTTPS を使用するかどうか。有効になっている場合、サーバー証明書は TLS に使用されます |
| read_write_max_client_co | 許可される最大読み取り/書き込みクライアント接続、デフォルトは max_connections |
| read_write_max_server_cc | サーバーに対して開かれる最大読み取り/書き込み接続 |
| read_only_max_client_cor | 許可される最大読み取り専用クライアント接続 |
| read_only_max_server_co | サーバーに対して開かれる読み取り専用接続の最大数 |
| read_write_consensus_tim | 読み取り/書き込み接続が受け入れられなくなるまでにコンセンサスが失われるまで待機する時間 |

次のページに続く

表 201 – 前のページからの続き

| Option | Description |
|---|---|
| read_only_consensus_time | 読み取り専用接続が受け入れられなくなるまでに、コンセンサス損失を待機する時間。 |

### グループプロキシオプション PGD 5.x のみ

| Option | Description |
|---|---|
| proxy_listen_address | グループ内のプロキシのリッスンアドレス。 |
| proxy_listen_addresses | グループ内のプロキシのリッスンアドレス。 |
| proxy_listen_port | グループ内のプロキシのリッスンポート。 |
| proxy_max_client_conn | グループ内のプロキシのクライアントの最大接続数。 |
| proxy_max_server_conn | グループ内のプロキシのサーバー接続の最大数。 |
| proxy_server_conn_timeou | グループ内のプロキシのサーバー接続タイムアウト。 |
| proxy_server_conn_keepal | グループ内のプロキシのサーバー接続キープアライブ。 |
| proxy_fallback_node_grou | グループ内のプロキシのフォールバックノードグループ。 |
| proxy_fallback_node_grou | グループ内のプロキシのフォールバックノードグループタイムアウト。 |
| proxy_consensus_grace_pe | グループ内のプロキシのコンセンサス猶予期間。 |
| proxy_read_listen_address | グループ内の読み取りプロキシのリッスンアドレス。 |
| proxy_read_listen_address | グループ内の読み取りプロキシのリッスンアドレス。 |
| proxy_read_listen_port | グループ内の読み取りプロキシのリッスンポート。 |
| proxy_read_max_client_cc | グループ内の読み取りプロキシのクライアントの最大接続数。 |
| proxy_read_max_server_c | グループ内の読み取りプロキシのサーバー接続の最大数。 |
| proxy_read_server_conn_k | グループ内の読み取りプロキシのサーバー接続キープアライブ。 |
| proxy_read_server_conn_t | グループ内の読み取りプロキシのサーバー接続タイムアウト。 |
| proxy_read_consensus_gra | グループ内の読み取りプロキシのコンセンサス猶予期間。 |

`(inherited)` の後に値が表示されている場合、これは値がグループに特に設定されていないが、親グループから継承されることを意味します。

### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

### 例

#### すべてのグループオプションをリストする

```
pgd group dc1_subgroup get-option
__OUTPUT__
 Option Name                        | Option Value
- -----------------------------------+---------------
 analytics_autoadd_tables           | false
 analytics_storage_location         |
```

(次のページに続く)

(前のページからの続き)

```
 apply_delay                       | 00:00:00
 check_constraints                 | true
 default_commit_scope              | local protect
 enable_raft                       | true
 enable_routing                    | true
 enable_wal_decoder                | false
 http_port                         | 6434
 location                          |
 num_writers                       | 2
 read_only_consensus_timeout       | 00:00:00
 read_only_max_client_connections  | 250
 read_only_max_server_connections  | 250
 read_only_port                    | 6433
 read_write_consensus_timeout      | 00:00:00
 read_write_max_client_connections | 250
 read_write_max_server_connections | 250
 read_write_port                   | 6432
 route_reader_max_lag              | -1
 route_writer_max_lag              | -1
 route_writer_wait_flush           | false
 streaming_mode                    | null
 use_https                         | true
```

#### 特定のグループオプションを取得する

```
pgd group dc1_subgroup get-option location
__OUTPUT__
 Option Name | Option Value
- ------------+--------------
 location    | London
```

#### 特定のグループオプションを json として取得する

```
pgd group dc1_subgroup get-option location -o json
__OUTPUT__
[
  {
    "option_name": "location",
    "option_value": "London"
  }
]
```

### pgd group set-leader

#### 概要

pgd group set-leader コマンドは、 EDB Postgres 分散クラスター内のグループの書き込みリーダーを設定するために使用されます。

このコマンドは、スイッチオーバー操作を実行します。

#### 構文

```
pgd group <GROUP_NAME> set-leader [OPTIONS] <LEADER>
```

<GROUP_NAME> は、書き込みリーダーを設定するグループの名前であり、<LEADER> は、書き込みリーダーとして設定するノードの名前です。

#### オプション

次の表は、pgd group set-leader コマンドで使用可能なオプションをリストしています。

| Short | Long | Description |
|---|---|---|
| | *--strict* | Strict メソッドデフォルト。 |
| | *--timeout* | メソッドが厳密な場合のタイムアウト期間。 (デフォルト 30s30 秒) |
| | *--fast* | 速い方法。 |

Strict メソッドがデフォルトのメソッドです。厳密な方法は、リーダーを切り替える前に、新しいリーダーが古いリーダーと同期するのを待機します。高速な方法は、route_write_max_lag を無視して、リーダーを切り替える前に新しいリーダーが古いリーダーと同期するのを待機しないため、即時です。

Global Options も参照してください。

#### 例

##### グループの書き込みリーダーの設定

```
pgd group dc1_subgroup set-leader kaboom
__OUTPUT__
Command executed successfully
```

##### ノードが既にリーダーである場合の書き込みリーダーの設定

```
pgd group dc1_subgroup set-leader kaboom
__OUTPUT__
Node kaboom is already the write leader
```

### pgd groups

pgd groups コマンドは、 EDB Postgres 分散クラスター内のグループを表示するために使用されます。

#### サブコマンド

- *pgd commit-scopes list* グループをリストします。

### pgd groups list

#### 概要

pgd groups list コマンドは、EDB Postgres 分散クラスター内のグループを表示するために使用されます。

#### 構文

```
pgd groups list [OPTIONS]
```

#### オプション

pgd groups list コマンドでは、次のオプションが使用できます。

| Short | Long | Description |
|---|---|---|
| -v | --verbose | グループに関する詳細情報を表示します。 |

共通のグローバルオプションについては、 グローバルオプション を参照してください。

#### 例

##### すべてのグループをリストする

```
pgd groups list
__OUTPUT__
Group Name    Parent Group Name Group Type Nodes
- ------------ ----------------- ---------- -----
democluster                     global     0
dc1_subgroup  democluster       data       3
```

##### すべてのグループと詳細情報をリストする

```
pgd groups list --verbose
__OUTPUT__
Group Name    Parent Group Name Group Type Nodes Raft Leader Write Leader Commit␣
↪Scope Node Group ID
- ------------ ----------------- ---------- ----- ----------- ------------ -----------
↪- -------------
```

(次のページに続く)

(前のページからの続き)

```
democluster                    global     0     kaftan
 150732310
dc1_subgroup  democluster      data       3     kaftan      kaboom
 1302278103
```

### pgd node

pgd node コマンドは、 EDB Postgres 分散クラスターのノードを管理するために使用されます。

#### サブコマンド

- *pgd* **ノードセットアップの実行** クラスター内の PGD データノードを構成します。
- *pgd cluster show* ノードレベルの情報を表示します。
- *pgd node part* アクティブなクラスターから PGD ノードを分離します。
- *pgd group set-option* ノードレベルのオプションを設定します。
- *pgd group get-option* ノードレベルのオプションを取得します。
- *pgd node set-config* ノードレベルの構成を設定します。
- *pgd node get-config* ノードレベルの構成を取得します。
- *pgd node upgrade* PGD Postgres ノードのメジャーバージョンのアップグレードを実行します。

### pgd node get-config

#### 概要

pgd node get-config コマンドは、 EDB Postgres 分散クラスターのノードレベルの構成設定 GUC を取得するために使用されます。

#### 構文

```
pgd node <NODE_NAME> get-config [OPTIONS] [CONFIG]
```

<NODE_NAME> は、オプションを取得するノードの名前です。

[CONFIG] は、取得する特定の GUC の名前です。オプションを指定しない場合、すべての GUC が表示されます。

#### オプション

pgd node get-config コマンドでは、次のオプションが使用できます。

| Short | Long | Description |
|---|---|---|
| -v | --verbose | GUC の最小値、最大値、説明などの追加情報を表示します |

Global Options を参照してください。

### 例

#### 特定のノード構成を取得する

```
pgd node pgd-a1 get-config bdr.batch_inserts
__OUTPUT__
 Config Name       | Config Value
- ------------------+--------------
 bdr.batch_inserts | 15
```

#### 詳細な出力を使用して特定のノード構成を取得する

```
pgd node pgd-a1 get-config max_wal_senders --verbose
__OUTPUT__
 Config Name     | Config Value | Min Value | Max Value | Description
- ----------------+--------------+-----------+-----------+----------------------------
↪-----------------------------------------------
 max_wal_senders | 7            | 0         | 262143    | Sets the maximum number of␣
↪simultaneously running WAL sender processes.
```

#### 特定のノード構成を json として取得する

```
 pgd node pgd-a1 get-config max_wal_senders -o json
__OUTPUT__
[
  {
    "config_name": "max_wal_senders",
    "config_value": "6",
    "min_value": "0",
    "max_value": "262143",
    "description": "Sets the maximum number of simultaneously running WAL sender␣
↪processes."
  }
]
```

### pgd node get-option

#### 概要

`pgd node get-option` コマンドは、 EDB Postgres 分散クラスターのノードレベルのオプションを取得するために使用されます。

#### 構文

```
pgd node <NODE_NAME> get-option [OPTIONS] [OPTION]
```

`<NODE_NAME>` は、オプションを取得するノードの名前です。

`[OPTION]` は、取得する特定のグループオプションの名前です。オプションを指定しない場合、すべてのオプションが表示されます。

次のオプションが使用できます。

#### ノードオプション

| Option | Description | Type |
|---|---|---|
| route_priority | 同じノードグループ内の他のノードに対するノードの相対ルーティング優先度。書き込みリーダーを選択する場合にのみ使用されます。 | integer |
| route_fence | ノードをフェンスするように設定します。フェンスされると、ノードは接続マネージャーから接続を受信できません。したがって、書き込みリーダーになることも、読み取り専用ノードプールで使用することもできません。レプリケーションは影響を受けません。 | bool |
| route_writes | 書き込みをこのノードにルーティングできるかどうか、つまりノードが書き込みリーダーになることができるかどうかを決定します。 | bool |
| route_reads | 読み取り専用接続をこのノードにルーティングできるかどうかを決定します PGD 5.5.0 以降。 | bool |
| route_dsn | 接続マネージャーがこのノードに接続するために使用する dsn。 | string |

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

#### すべてのノードオプションの取得

```
pgd node kaboom get-option
__OUTPUT__
Option Name    Option Value
- ------------ --------------------------------------------------
route_dsn      host=kaboom port=5444 dbname=pgddb user=postgres
route_fence    false
route_priority 100
route_reads    true
route_writes   true
```

#### 特定のノードオプションの取得

```
pgd node kaboom get-option route_priority
__OUTPUT__
Option Name    Option Value
- ------------ ------------
route_priority 100
```

#### すべてのノードオプションを json として取得

```
pgd node kaboom get-option -o json
__OUTPUT__
[
  {
    "option_name": "route_dsn",
    "option_value": "host=kaboom port=5444 dbname=pgddb user=postgres"
  },
  {
    "option_name": "route_fence",
    "option_value": "false"
  },
  {
    "option_name": "route_priority",
    "option_value": "100"
  },
  {
    "option_name": "route_reads",
    "option_value": "true"
  },
  {
    "option_name": "route_writes",
    "option_value": "true"
  }
]
```

### pgd node part

#### 概要

pgd node part コマンドは、アクティブなクラスターから PGD ノードを削除し、分離操作を開始します。

このコマンドの動作は、クラスターの Raft プロトコルバージョンに基づいて条件付きです。

1. **Raft プロトコルバージョン 6003 以降**

このシナリオでは、クラスター内のすべての PGD ノードがバージョン製品バージョン 6.4 以上を実行しています。

システムはすぐにノードの名前をその UUID に変更し、バックグラウンドで複雑な分離操作を続行します。このアクションにより、操作全体が完了する前にノードの元の名前を変更できます。最初の part コマンドが返されたら、ノードの UUID ベースの名前を使用して、分割ノードの進行状況を追跡します。

--no-wait オプションは、このシナリオでは非推奨です。

1. **Raft プロトコルバージョン 6003 より前のバージョン**

このシナリオでは、クラスター内のすべての PGD ノードが PGD バージョン製品バージョン 6.4 以上を実行しているわけではありません。

デフォルトでは、コマンドはパーツ操作全体が完了するまで待機します。--no-wait を使用する場合、システムはバックグラウンドでパートオペレーションをスケジュールし、完了を待たずにすぐにアドバイスメッセージを返します。

#### 構文

```
pgd node <NODE_NAME> part [OPTIONS]
```

<NODE_NAME> は、分離するノードの名前です。

#### オプション

pgd node part コマンドでは、次のオプションが使用できます。

| Option | Description | Note |
|---|---|---|
| --no-wait | バックグラウンドでパート操作をスケジュールし、操作が完了するのを待たずに、すぐに戻ります。 | Used for Raft version below 6003 |

Global Options を参照してください。

#### 例

##### ノードの一部 Raft プロトコルバージョン 6003 以降

```
pgd node pgd-a1 part
__OUTPUT__
Starting a part node operation for node: pgd-a1
This may take some time, please wait...
NOTICE:  node pgd-a1 renamed to a867d475-7177-533a-a5b7-44badb2dfe4b and removal␣
↪started in the background
DETAIL:  node is parting nodegroup topgroup (51455156). node name pgd-a1 can be␣
↪reused if required.
```

#### --no-wait オプションを使用したノードの一部 6003 より前の Raft プロトコルバージョン

```
pgd node pgd-a1 part --no-wait
__OUTPUT__
Starting a part node operation for node: pgd-a1
This may take some time, please wait...
NOTICE:  node pgd-a1 removal started in the background
DETAIL:  node is parting nodegroup topgroup (51455156)
```

#### --no-wait オプションのないノードの一部 6003 より前の Raft プロトコルバージョン

```
pgd node pgd-a1 part
__OUTPUT__
Starting a part node operation for node: pgd-a1
This may take some time, please wait...
NOTICE: Node pgd-a1 has been removed from the BDR group
```

### pgd node set-config

#### 概要

pgd node set-config コマンドは、EDB Postgres 分散クラスターでノードレベルの Postgres 設定 GUC を設定するために使用されます。このコマンドには、superuser つまり Postgres プロセス所有者ユーザーのノード DSN が必要です。

コマンドは、構成を設定した後、ノードに構成をリロードし、GUC はサーバーの再起動が必要であるという警告を表示します。

#### 構文

```
pgd node <NODE_NAME> set-config [OPTIONS] <CONFIG> <VALUE>
```

<NODE_NAME> は、オプションを取得するノードの名前です。

また、 <CONFIG> は、取得する特定のノード構成の名前であり、<VALUE> は、それに設定する値です。

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

#### 特定のノード構成を設定する

```
pgd node pgd-a1 set-config bdr.batch_inserts 20 --dsn "host=pgd-a1 port=5432␣
↪user=postgres "
__OUTPUT__
Command executed successfully
```

#### 値にスペースを使用して特定のノード構成を設定します

```
pgd node pgd-a2 set-config unix_socket_directories "/var/run/edb-pge, /tmp, /var/lib/
↪postgresql" --dsn "host=pgd-a2 port=5432 user=postgres "
__OUTPUT__
WARN: Config change requires a Postgres restart to take effect.
Command executed successfully
```

WARN メッセージは、Postgres サーバーが再起動されるまで構成の変更が有効にならないことを示していることに注意してください。

### pgd node set-option

#### 概要

pgd node set-option コマンドは、 EDB Postgres 分散クラスターでノードレベルのオプションを設定するために使用されます。

#### 構文

```
pgd node <NODE_NAME> set-option [OPTIONS] <OPTION> <VALUE>
```

<NODE_NAME> は、オプションを取得するノードの名前です。

また、<OPTION> は、取得する特定のノードオプションの名前であり、<VALUE> は、それに設定する値です。

次のオプションが使用できます。

#### ノードオプション

| Option | Description | Type |
|---|---|---|
| route_priority | 同じノードグループ内の他のノードに対するノードの相対ルーティング優先度。書き込みリーダーを選択する場合にのみ使用されます。 | integer |

次のページに続く

表 208 – 前のページからの続き

| Option | Description | Type |
|---|---|---|
| route_fence | ノードをフェンスするように設定します。フェンスされると、ノードは接続マネージャーから接続を受信できません。したがって、書き込みリーダーになることも、読み取り専用ノードプールで使用することもできません。レプリケーションは影響を受けません。 | bool |
| route_writes | 書き込みをこのノードにルーティングできるかどうか、つまりノードが書き込みリーダーになることができるかどうかを決定します。 | bool |
| route_reads | 読み取り専用接続をこのノードにルーティングできるかどうかを決定します PGD 5.5.0 以降。 | bool |
| route_dsn | 接続マネージャーがこのノードに接続するために使用する dsn。 | string |

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

#### 特定のノードオプションを設定する

```
pgd node kaboom set-option route_priority 100
__OUTPUT__
Command executed successfully
```

#### 値にスペースを使用して特定のノードオプションを設定します

```
pgd node kaboom set-option route_dsn "host=kaboom port=5444 dbname=pgddb user=postgres
↪"
__OUTPUT__
Command executed successfully
```

### pgd node setup

#### 概要

pgd node setup コマンドは、クラスター内の PGD データノードを構成するために使用されます。新しいノードのセットアップ、既存のノードのクラスターへの参加、またはクラスターへのノードの論理参加を実行するために使用できます。

!!!重要「物理的結合のバージョン要件」

pgd node setup が物理的結合ローカルノードが起動して実行されていないときにリモートノードからデータをコピーする場合、ソースノードと参加ノードの両方がまったく同じ PGD バージョンである必要があ

ります。物理的結合を使用して、別の PGD バージョンのノードを既存のクラスターに参加させることはできません。ローリングアップグレードの場合、代わりに論理結合を使用してください。詳細については、ノード参加を使用したローリングアップグレード を参照してください。

コマンドの動作は、コマンドで指定されたローカルノードとリモートノードの状態によって異なります。

これがクラスター内の最初のノードの場合、 pgd node setup は initdb を実行し、 PGD ノードをセットアップします。

これが最初のノードではなく、ローカルノードが起動して実行されていない場合、pgd node setup はクラスターへのノードの物理的参加を実行します。これにより、初期化プロセスの一部としてリモートノードからローカルノードにデータがコピーされ、ローカルノードがクラスターに参加します。これは、新しいノードにデータをロードする最も速い方法です。

ローカルノードが起動して実行されており、リモートノードにも到達可能である場合、pgd node setup はクラスターへのノードの論理参加を実行します。これにより、クラスター内に新しいノードが作成され、リモートノードからストリーミングレプリケーションが開始されます。これは、新しいノードを既存のクラスターに追加するための推奨される方法です。

ローカルノードがアップして実行されており、リモートノード dsn が提供されていない場合、pgd node setup は、ノードが指定されたグループの一部でない場合、ノードグループのスイッチを実行します。

### ユーザーとロール

pgd node setup コマンドを実行するには、スーパーユーザーロールが必要です。スーパーユーザーロールは、データディレクトリを作成し、データベースを初期化するために使用されます。データベースを作成するには、スーパーユーザーロールに CREATEDB 特権が必要です。

--dsn オプションで指定されたユーザーが存在しない場合は作成されます。PGD 機能を管理できる bdr_superuser ロールのみが付与されます。ただし、データベースに対する他の特権はありません。

### 透過的データ暗号化 TDE のサポート

透過的データ暗号化 TDE は、EDB Postgres Advanced Server および EDB Postgres Extended Server バージョン 15 以降で利用可能なオプション機能です。

pgd node setup コマンドは、新しいデータベースサーバーを初期化するとき、または既存のノードをクラスターに参加させるときにこの機能をサポートします。pgd node setup によって公開される TDE オプションは、次の例外を除き initdb TDE options に似ています。

- オプション--data-encryption は、ブール値のみのフラグです AES キー長さの値は受け入れません。
- 代わりに、オプション --data-encryption-keylen を使用して AES キー長を指定します。

### 構文

```
pgd node <NODE_NAME> setup [OPTIONS] -D <PG_DATA>
```

### 引数

- <NODE_NAME\> 作成するノードの名前。これは、クラスター内のノードを識別するために使用される名前です。クラスター内で一意である必要があります。

### オプション

| Option | Description |
|---|---|
| *--listen-addr <LISTEN_ADDR>* | 構成されたノードが着信接続をリッスンするアドレス、および他のノードがこのノードに接続するために使用するアドレス。これは通常、少なくとも`localhost`に設定されますが、任意の有効なアドレスに設定できます。デフォルトは`localhost`です。 `--dsn`からの`host`値もこのリストに追加されます。 |
| *--initial-node-count <INITIAL_NODE_COUNT>* | クラスター内のノードまたはクラスター内に含まれる予定のノードの数。ノードのさまざまなリソース設定を計算するために使用されます。デフォルト 3 です。 |
| *--bindir <BINDIR>* | <BINDIR> バイナリが存在するディレクトリを指定します。デフォルトは、実行中の pgd バイナリが置かれているディレクトリです。 |
| *--log-file <LOG_FILE>* | ログファイルへのパス。postgres 起動ログに使用されます。デフォルトでは、*postgres-<port>.log* `という名前の現在のディレクトリにあるファイルに書き込みます。ここで、ポート値は`--dsn`オプションの`port`属性から取得されます。 |
| *-D, --pgdata <PG_DATA>* | ノードのデータディレクトリとして<PG_DATA>を使用します。環境変数`PGDATA`でも設定されます。これは有効なディレクトリであり、コマンドを実行するユーザーが書き込み可能である必要があります。 |
| *--superuser <SUPERUSER>* | `initdb`のスーパーユーザー名。デフォルトは`postgres`です。 |
| *--node-kind <NODE_KIND>* | 作成するノードの種類を指定します。デフォルトは`data`です。サポートされている値は`data`、*witness*、`subscriber-only`です。 |
| *--group-name <GROUP_NAME>* | ノードグループ名。指定しない場合、ノードはアクティブノードのグループに追加されます。これは、グループの最初のノードの必須引数です。 |
| *--create-group* | 指定されたグループがまだ存在しない場合、このフラグを設定して作成します。これは、最初のノードのデフォルトで true になります。 |
| *--cluster-name <CLUSTER_NAME>* | ノードに参加するクラスターの名前。クラスターを初めてセットアップするときに、これは`parent node group`を作成するために使用されます。指定しない場合、デフォルトの`pgd`。 |
| *--cluster-dsn <CLUSTER_DSN>* | アクティブな PGD クラスターに属する DSN。これは、クラスターの最初のノードを構成する場合には必要ありませんが、後続のノードには必須です。既存のアクティブノードの DSN を指す必要があります。 |
| *--postgresql-conf <POSTGRESQL_CONF>* | ノードに使用される`postgresql.conf`ファイルのオプションのパス。 |
| *--postgresql-auto-conf <POSTGRESQL_AUTO_CONF>* | ノードに使用される`postgresql.auto.conf`ファイルのオプションのパス。 |

次のページに続く

表 209 – 前のページからの続き

| Option | Description |
|---|---|
| *--hba-conf <HBA_CONF>* | ノードに使用される `pg_hba.conf`ファイルのオプションのパス。 |
| *--update-pgpass* | 設定されている場合、新しいノードパスワードの pgpass ファイルは、現在のユーザーの `.pgpass`ファイルに保存されます。 |
| *--verbose* | 詳細なメッセージを出力します。 |
| *-y, --data-encryption* | データベースサーバーの初期化時に TDE を追加します。 |
| *--data-encryption-keylen <AES_KEYLEN>* | TDE の AES キーの長さ。デフォルトは`128`です。サポートされている値は`128`および`256`です。 |
| *--key-wrap-command <KEY_WRAP_COMMANI* | データ暗号化キーを保護するためのラッピング/暗号化コマンド。`<KEY_WRAP_COMMAND>`はカスタマイズ可能ですが、プレースホルダー`%p`を含める必要があります。このオプションを使用しない場合、`pgd node setup`は環境変数`PGDATAKEYWRAPCMD`にフォールバックします。 |
| *--key-unwrap-command <KEY_UNWRAP_COMM.* | データ暗号化キーにアクセスするための解凍/復号コマンド。`<KEY_UNWRAP_COMMAND>`はカスタマイズ可能ですが、プレースホルダー`%p`を含める必要があります。このオプションを使用しない場合、`pgd node setup`は環境変数`PGDATAKEYUNWRAPCMD`にフォールバックします。 |
| *--no-key-wrap* | キーのラッピングを無効にします。このオプションは、実稼働環境での使用はお勧めできません。 |
| *--copy-key-from <COPY_KEY_FROM>* | 指定された場所から既存のデータ暗号化キーをコピーします。通常、暗号化キーは`pg_encryption/key.bin`に保存されます |

Global Options も参照してください。

`pgd node setup` コマンドは、次の取得方法を優先順に確認することにより、非対話型でデータベースパスワードを見つけようとします。

- `--dsn` グローバルオプションで提供される `password` パラメーターで設定されたパスワード。
- `PGPASSWORD` 環境変数で設定されたパスワード。
- `--dsn` グローバルオプションで提供される `passfile` パラメーターで設定されたパスワードファイル。
- `PGPASSFILE` 環境変数で設定されたパスワードファイル
- コマンドを呼び出すユーザーの `HOME` ディレクトリにあるパスワードファイル `.pgpass`

システムがこれらの方法でパスワードを見つけられない場合、コマンドはパスワードを手動で入力するように求めます。

`--update-pgpass` フラグは、システムがパスワードを正常に取得した後に発生する動作を制御します。設定すると、パスワードが将来の非対話型使用に備えて保存されます。更新するパスワードファイルは、次の優先順位に従います。

- `--dsn option` に提供された接続文字列内の `passfile` パラメーターで指定されたパス。
- `PGPASSFILE` 環境変数で指定されたファイル

- .pgpass ファイルは、コマンドを呼び出したユーザーの HOME ディレクトリにあります。

#### 例

これらの例では、3 つのホスト host-1 、host-2 、および host-3 でクラスターをセットアップして、3 つのノード node-1 、node-2 、および node-3 を作成します。3 つのノードはデータノードであり、グループ名が group-1 である pgd という名前のクラスターの一部です。

PGPASSWORD 環境変数をエクスポートして、コマンドを実行するたびに pgdadmin ユーザーのパスワードを入力する必要を回避することをお勧めします。これは、次のコマンドで行うことができます。

```
export PGPASSWORD=pgdsecret
```

#### 最初のノードの構成

```
pgd node node-1 setup --dsn "host=host-1 port=5432 user=pgdadmin dbname=pgddb" \
- -listen-addr "localhost,host-1" \
- -group-name group-1 --cluster-name pgd \
- D /var/lib/edb-pge/17/main
```

コマンドをステップ実行して、node-1 を設定しています。最初のオプションは--dsn オプションで、これはノードの接続文字列です。これは通常、host=hostname port=5432 user=pgdadmin dbname=pgd に設定されます。これは、ローカル Postgres インスタンスの一般的な接続文字列です。

--listen-address オプションは、ノードが着信接続をリッスンするアドレスを指定するために使用されます。この場合、localhost,host-1 に設定しています。これは、ノードがローカルホストと host-1 アドレスの両方でリッスンすることを意味します。

これはクラスター内の最初のノードであるため、グループ名を group-1 に設定し、クラスター名を pgd 実際にデフォルトに設定します。これはクラスター内の最初のノードであるため、--create-group オプションが自動的に設定されます。

最後に、 -D オプションを使用してノードのデータディレクトリを設定します。これは、Postgres データファイルが保存される場所です。この例では、データディレクトリとして/var/lib/edb-pge/17/main を使用しています。

このコマンドは、データディレクトリを作成し、データベースを PGD 用に正しく初期化します。次に、ノードを起動し、クラスターに参加する他のノードを含む新しい接続で利用できるようにします。

#### 2 番目のノードの構成

```
pgd node node-2 setup --dsn "host=host-2 port=5432 user=pgdadmin dbname=pgddb" \
- -listen-addr "localhost,host-2" \
- D /var/lib/edb-pge/17/main
- -cluster-dsn "host=host-1 port=5432 user=pgdadmin dbname=pgddb"
```

このコマンドは最初のノードに似ていますが、node-2 を設定しています。 --dsn オプションは、ノードの接続文字列であり、通常は host=hostname port=5432 user=pgdadmin dbname=pgd に設定されます。cluster-dsn は、アクティブなノードを指す必要があります。接続マネージャー、またはプロキシエンドポ

イントなどを指すことができます。CLI は、その背後にあるノードの実際の DSN を取得します。この場合、クラスター内の最初のノードの接続文字列である host=host-1 port=5432 user=pgdadmin dbname=pgd に設定しています。

### 3 番目のノードの構成

```
pgd node node-3 setup --dsn "host=host-3 port=5432 user=pgdadmin dbname=pgddb" \
- -listen-addr "localhost,host-3" \
- -cluster-dsn "host=host-1 port=5432 user=pgdadmin dbname=pgddb" \
- D /var/lib/edb-pge/17/main
```

このコマンドは 2 番目のノードに似ていますが、node-3 を設定しています。 --dsn オプションは、ノードの接続文字列であり、通常は host=hostname port=5432 user=pgdadmin dbname=pgd に設定されます。cluster-dsn は、アクティブなノードを指す必要があります。接続マネージャー、またはプロキシエンドポイントなどを指すことができます。CLI は、その背後にあるノードの実際の DSN を取得します。この場合、クラスター内の最初のノードの接続文字列である host=host-1 port=5432 user=pgdadmin dbname=pgd に設定しています。

### 分割およびドロップされたノードのクラスターへの参加

```
pgd node node-2 setup --dsn "host=host-2 port=5432 user=pgdadmin dbname=pgddb" \
- -listen-addr "localhost,host-2" \
- -cluster-dsn "host=host-1 port=5432 user=pgdadmin dbname=pgddb" \
- D /var/lib/edb-pge/17/main
```

このコマンドは、後続のノードの設定に似ていますが、 node-2 を再度設定します。 --dsn オプションは、ノードの接続文字列であり、通常は host=hostname port=5432 user=pgdadmin dbname=pgd に設定されます。 cluster-dsn は、アクティブなノードを指す必要があります。接続マネージャー、またはプロキシエンドポイントなどを指すことができます。CLI は、その背後にあるノードの実際の DSN を取得します。この場合、クラスター内の最初のノードの接続文字列である host=host-1 port=5432 user=pgdadmin dbname=pgd に設定しています。

これは、メンテナンスなどのアクティビティのためにノードがクラスターから parted および dropped され、クラスターに再参加する必要がある場合に役立ちます。コマンドは、クラスターへのノードの論理参加を実行します。これにより、クラスター内に新しいノードが作成され、リモートノードからストリーミングレプリケーションが開始されます。

### TDE オプションを使用した最初のノードの構成

```
pgd node node-1 setup \
- -dsn "host=host-1 port=5432 user=pgdadmin dbname=pgddb" \
- -listen-addr "localhost,host-1" \
- -group-name group-1 --cluster-name pgd \
- D /var/lib/edb-pge/17/main \
- -data-encryption \
- -key-wrap-command "openssl enc -e -aes-128-cbc -pbkdf2 -pass pass:secret -out %p" \
- -key-unwrap-command "openssl enc -d -aes-128-cbc -pbkdf2 -pass pass:secret -in %p"
```

TDE 対応クラスターの pgd node setup コマンドは標準のコマンドに似ていますが、3 つの追加 TDE 固有オプション--data-encryption 、--key-wrap-command 、および--key-unwrap-command が必要です。

簡単にするために、この例の--key-wrap-command および--key-unwrap-command の値は単純なパスフレーズを使用しています。ただし、外部のキー管理サービス KMS などのより安全なメカニズムを構成できます。

この例では、 --data-encryption-keylen オプションを明示的に設定せず、キー長をデフォルトの 128 ビットに設定しています。

#### TDE オプションを使用した 2 番目のノードの構成

```
pgd node node-2 setup \
- -dsn "host=host-2 port=5432 user=pgdadmin dbname=pgddb" \
- -listen-addr "localhost,host-2" \
- D /var/lib/edb-pge/17/main \
- -cluster-dsn "host=host-1 port=5432 user=pgdadmin dbname=pgddb" \
- -key-wrap-command "openssl enc -e -aes-128-cbc -pbkdf2 -pass pass:secret -out %p" \
- -key-unwrap-command "openssl enc -d -aes-128-cbc -pbkdf2 -pass pass:secret -in %p"
```

このコマンドは、最初のノードで使用されるセットアップに似ていますが、必要な TDE オプションを保持します。残りの引数は、クラスター内の後続のノードをセットアップするときに使用される標準のコマンドと一貫しています。TDE オプションの値は、初期ノードに提供されたものと同じである必要があります。

### pgd node show

#### 概要

pgd node show コマンドは、 EDB Postgres 分散クラスターにノードレベルの情報を表示するために使用されます。

#### 構文

```
pgd node <NODE_NAME> show [OPTIONS]
```

<NODE_NAME> は、情報を表示するノードの名前です。

#### オプション

コマンド固有のオプションはありません。 Global Options を参照してください。

#### 例

#### ノード情報の表示

```
pgd node kaboom show
__OUTPUT__

#  Summary
```

(次のページに続く)

(前のページからの続き)

```
Node Property   Value
- -------------- ------------
Node Name       kaboom
Group Name      dc1_subgroup
Node Kind       data
Join State      ACTIVE
Node Status     Up
Node ID         2710197610
Snowflake SeqID 2
Database        pgddb

#  Options

Option Name    Option Value
- ------------- ------------------------------------------------
route_dsn      host=kaboom port=5444 dbname=pgddb user=postgres
route_fence    false
route_priority 100
route_reads    true
route_writes   true
```

### ノード情報を JSON として表示する

```
pgd node kaboom show -o json
__OUTPUT__
[
  {
    "Summary": [
      {
        "info": "Node Name",
        "value": "kaboom"
      },
      {
        "info": "Group Name",
        "value": "dc1_subgroup"
      },
      {
        "info": "Node Kind",
        "value": "data"
      },
      {
        "info": "Join State",
        "value": "ACTIVE"
```

(次のページに続く)

(前のページからの続き)


```
      },
      {
        "info": "Node Status",
        "value": "Up"
      },
      {
        "info": "Node ID",
        "value": "2710197610"
      },
      {
        "info": "Snowflake SeqID",
        "value": "2"
      },
      {
        "info": "Database",
        "value": "pgddb"
      }
    ]
  },
  {
    "Options": [
      {
        "option_name": "route_dsn",
        "option_value": "host=kaboom port=5444 dbname=pgddb user=postgres "
      },
      {
        "option_name": "route_fence",
        "option_value": "false"
      },
      {
        "option_name": "route_priority",
        "option_value": "100"
      },
      {
        "option_name": "route_reads",
        "option_value": "true"
      },
      {
        "option_name": "route_writes",
        "option_value": "true"
      }
    ]
  }
]
```

### pgd node upgrade

#### 概要

pgd node upgrade コマンドは、 EDB Postgres 分散クラスター内のノードの Postgres バージョンをアップグレードするために使用されます。

!!!note PGD 6.4 以降、pgd node setup は、シングルユーザーモードで BDR オブジェクトを作成する前に NextOID カウンターを進め、これらのオブジェクトがアップグレードの失敗の原因となるシステムレンジ OID<16384 を受信しないようにします。アップグレードを実行する前に、コマンドは、既存の BDR オブジェクトがシステム範囲 OID で作成されたかどうかを確認します。検出された場合、コマンドはすぐに失敗し、影響を受けるオブジェクトを識別する明確なエラーが表示されます。復旧する手順については、システムレンジの *OID* エラーの解決 を参照してください。

#### 構文

```
pgd node <NODE_NAME> upgrade [OPTIONS] --old-bindir <OLD_BINDIR> --new-bindir <NEW_
↪BINDIR> --old-datadir <OLD_DATADIR> --new-datadir <NEW_DATADIR> --database
↪<DATABASE> --username <USER_NAME>
```

<NODE_NAME> はアップグレードするノードの名前であり、<OLD_BINDIR> 、<NEW_BINDIR> 、<OLD_DATADIR> 、<NEW_DATADIR> 、<DATABASE> 、および<USER_NAME> は、新旧の Postgres インスタンスの bin ディレクトリ、新旧の Postgres インスタンスデータディレクトリ、データベース名、およびクラスターのインストールです。それぞれユーザー名。

#### オプション

次の表は、pgd node upgrade コマンドで使用可能なオプションをリストしています。

| Short | Long | Default | Env | Description |
|---|---|---|---|---|
| -b | --old-bindir | | PGBINOLD | 古い Postgres インスタンスの bin ディレクトリ |
| -B | --new-bindir | | PGBINNEW | 新しい Postgres インスタンスの bin ディレクトリ |
| -d | --old-datadir | | PGDATAOLD | 古い Postgres インスタンスデータディレクトリ |
| -D | --new-datadir | | PGDATANEW | 新しい Postgres インスタンスデータディレクトリ |
| | --database | | PGDATABASE | PGD データベース名 |
| -p | --old-port | 5432 | PGPORTOLD | 古い Postgres インスタンスポート |
| | --socketdir | /var/run/postgresc | PGSOCKET-DIR | アップグレード中にポストマスターソケットに使用するディレクトリ |

次のページに続く

表 210 – 前のページからの続き

| Short | Long | Default | Env | Description |
|---|---|---|---|---|
| | --new-socketdir | /var/run/postgresc | PGSOCKET-DIRNEW | 新しいクラスターのポストマスターソケットに使用するディレクトリ |
| | --check | | | チェックのみを実行し、クラスターを変更しないように指定します |
| -j | --jobs | 1 | | 同時に使用するプロセスまたはスレッドの数 |
| -k | --link | | | 新しいクラスターにファイルをコピーする代わりにハードリンクを使用する |
| | --old-options | | | 古い postgres コマンドに渡すオプション、複数の呼び出しが追加されます |
| | --new-options | | | 新しい postgres コマンドに渡すオプション、複数の呼び出しが追加されます |
| -N | --no-sync | | | アップグレードされたクラスター内のすべてのファイルがディスクに書き込まれるのを待たない |
| -P | --new-port | 5432 | PGPORTNEW | 新しい Postgres インスタンスのポート番号 |
| -r | --retain | | | 正常に完了した後でも SQL とログファイルを保持する |
| -U | --username | | PGUSER | クラスターのインストールユーザー名 |
| | --clone | | | 効率的なファイルのクローン作成を使用する |
| | --copy-by-block | | | 異なる暗号化設定を使用するクラスター間でデータを移行するために使用されます。このオプションは、透過的データ暗号化 TDE を使用するデータベースでサポートされています |
| | --key-unwrap-command | | PG-DATAKEYUN-WRAPCMD | データ暗号化キーを解凍し復号化し、コピーするファイルにアクセスするコマンド。コマンドは、`pgd node setup`を使用したサーバーの初期化中に指定されたものと同じである必要があります |

Global Options も参照してください。

### システムレンジの OID エラーの解決

コマンドがシステムレンジ OID エラーで失敗した場合、クラスターは、シングルユーザーモードで BDR オブジェクトを作成する前に NextOID カウンターを進めなかった 6.4 より前のバージョンの pgd node setup でセットアップされています。影響を受けるオブジェクトに、新しい OID を再割り当てすることはできません。

エラーを解決するには、ノードを分割し、標準の psql セッションシングルユーザーモードではなくで BDR 拡張機能をドロップして再作成し、構造を同期せずにクラスターに再参加して既存のデータを保持します。

1. ノードをクラスターから分離します。

```
pgd node <node-name> part
```

2. BDR データベースに接続した psql セッションで、拡張機能をドロップします。

```
DROP EXTENSION bdr CASCADE;
```

3. 標準の psql セッションシングルユーザーモードではなく、拡張機能を再作成します。

```
CREATE EXTENSION bdr;
```

4. *bdr.create_node()* および *bdr.join_node_group()* を使用してノードを再作成し、グループに再参加します。synchronize_structure を'none' に設定します。
5. pgd node upgrade を再度実行します。

### 例

次の例では、“kaolin”は、クイックスタートデモクラスターからアップグレードするノードの名前です。

### ノードの Postgres バージョンをアップグレードする

```
pgd node kaolin upgrade --old-bindir /usr/pgsql-16/bin --new-bindir /usr/pgsql-17/bin␣
↪--old-datadir /var/lib/pgsql/16/data --new-datadir /var/lib/pgsql/17/data --
↪database pgddb --username enterprisedb
```

### ハードリンクを備えたノードの Postgres バージョンをアップグレードする

```
pgd node kaolin upgrade --old-bindir /usr/pgsql-16/bin --new-bindir /usr/pgsql-17/bin␣
↪--old-datadir /var/lib/pgsql/16/data --new-datadir /var/lib/pgsql/17/data --
↪database pgddb --username enterprisedb --link
```

### 効率的なファイルクローン作成を使用してノードの Postgres バージョンをアップグレードする

```
pgd node kaolin upgrade --old-bindir /usr/pgsql-16/bin --new-bindir /usr/pgsql-17/bin␣
↪--old-datadir /var/lib/pgsql/16/data --new-datadir /var/lib/pgsql/17/data --
↪database pgddb --username enterprisedb --clone
```

#### 別のポート番号を持つノードの Postgres バージョンをアップグレードする

```
pgd node kaolin upgrade --old-bindir /usr/pgsql-16/bin --new-bindir /usr/pgsql-17/bin␣
↪--old-datadir /var/lib/pgsql/16/data --new-datadir /var/lib/pgsql/17/data --
↪database pgddb --username enterprisedb --old-port 5433 --new-port 5434
```

#### 透過的データ暗号化 TDE を使用してノードの Postgres 拡張バージョンをアップグレードする

```
pgd node kaolin upgrade --database pgddb -B /usr/lib/edb-pge/16/bin --socketdir /var/
↪run/edb-pge/ --old-bindir /usr/lib/edb-pge/15/bin  --old-datadir /var/lib/edb-pge/
↪15/main --new-datadir /var/lib/edb-pge/16/main --username postgres --key-unwrap-
↪command "openssl enc -d -aes-128-cbc -pbkdf2 -pass pass:secret -in %p" --copy-by-
↪block
```

### pgd nodes

pgd nodes コマンドは、 EDB Postgres 分散クラスター内のノードを表示するために使用されます。

#### サブコマンド

- *pgd commit-scopes list* ノードをリストします。

### pgd nodes list

#### 概要

pgd nodes list コマンドは、 EDB Postgres 分散クラスター内のノードを表示するために使用されます。デフォルトでは、これはノード名、グループ名、ノードの種類、ノードの参加状態、およびアップまたはダウンを示します。

#### 構文

```
pgd nodes list [OPTIONS]
```

#### オプション

pgd nodes list コマンドでは、次のオプションが使用できます。

| Short | Long | Description |
|---|---|---|
| | --versions | ノードに関するバージョン情報のみを表示します。ノードごとに、 BDR バージョンと Postgres バージョンが表示されます。 |
| -v | --verbose | ノードに関する詳細情報を表示します。このオプションは、ノードごとに、ノード ID、Snowflake シーケンス ID、およびデータベース名を追加します。 |

次のページに続く

表 211 – 前のページからの続き

| Short | Long | Description |
|---|---|---|
| | --group | グループ名でフィルタリングします。 |

共通のグローバルオプションについては、 Global Options を参照してください。

### 例

#### すべてのノードをリストする

```
pgd nodes list
__OUTPUT__
Node Name Group Name   Node Kind Join State Node Status
- -------- ------------ --------- ---------- -----------
kaftan    dc1_subgroup data      ACTIVE     Up
kaboom    dc1_subgroup data      ACTIVE     Up
kaolin    dc1_subgroup data      ACTIVE     Up
kapok     dc2_subgroup data      ACTIVE     Up
kaput     dc2_subgroup data      ACTIVE     Up
karakul   dc2_subgroup data      ACTIVE     Up
```

#### 特定のグループ内のすべてのノードをリストする

```
pgd nodes list --group dc1_subgroup
__OUTPUT__
Node Name Group Name   Node Kind Join State Node Status
- -------- ------------ --------- ---------- -----------
kaftan    dc1_subgroup data      ACTIVE     Up
kaboom    dc1_subgroup data      ACTIVE     Up
kaolin    dc1_subgroup data      ACTIVE     Up
```

#### すべてのノードと詳細情報をリストする

```
pgd nodes list --verbose
__OUTPUT__
Node Name Group Name   Node Kind Join State Node Status Node ID    Snowflake SeqID␣
↪Database
- -------- ------------ --------- ---------- ----------- ---------- --------------- --
↪------
kaftan    dc1_subgroup data      ACTIVE     Up          3490219809 1               ␣
↪pgddb
kaboom    dc1_subgroup data      ACTIVE     Up          2710197610 2               ␣
↪pgddb
kaolin    dc1_subgroup data      ACTIVE     Up          2111777360 3               ␣
↪pgddb
```

(次のページに続く)

(前のページからの続き)

```
kapok     dc2_subgroup data      ACTIVE     Up          1599108790 4              ␣
↪pgddb
kaput     dc2_subgroup data      ACTIVE     Up          1321108721 5              ␣
↪pgddb
karakul   dc2_subgroup data      ACTIVE     Up          1012686451 6              ␣
↪pgddb
```

#### すべてのノードのバージョン情報をリストする

```
pgd nodes list --versions
__OUTPUT__
Node Name BDR Version Postgres Version
- -------- ----------- ----------------------------------
kaboom    5.7.0       15.12.0 (Debian 15.12.0-1.bullseye)
kaftan    5.7.0       15.12.0 (Debian 15.12.0-1.bullseye)
kaolin    5.7.0       15.12.0 (Debian 15.12.0-1.bullseye)
kapok     5.7.0       15.12.0 (Debian 15.12.0-1.bullseye)
kaput     5.7.0       15.12.0 (Debian 15.12.0-1.bullseye)
karakul   5.7.0       15.12.0 (Debian 15.12.0-1.bullseye)
```

### pgd raft

`pgd raft` コマンドは、 EDB Postgres 分散クラスター内の Raft コンセンサス状態を監視および管理するために使用されます。

#### サブコマンド

- *pgd raft disable* 指定された PGD ノードまたはグループで Raft コンセンサスを無効にします。
- *pgd raft enable* 指定された PGD ノードまたはグループで Raft コンセンサスを有効にします。
- *pgd group set-leader* Raft グループ内のリーダーシップの移転を手動でトリガーします。
- *pgd cluster show* クラスターの Raft ステータスを表示します。
- *pgd raft sync-snapshot* リーダーからエクスポートし、すべてのノードでインポートすることにより、すべてのノードで Raft 状態を同期します。

### pgd raft disable

#### 概要

`pgd raft disable` コマンドは、EDB Postgres 分散クラスター内の 1 つ以上のノードで Raft コンセンサスワーカーを停止します。ワーカーが無効になると、ノードは Raft 選挙への参加を停止し、コンセンサスログエントリの送信または受信を停止し、書き込みリーダーとして機能しません。

ワーカーを有効にすると、ノードは Raft 選択、ログレプリケーション、およびクラスター全体の状態変更に参加できるようになります。

> **注釈**
>
> コンセンサスワーカーは、 *pgd raft enable* を使用して明示的に有効になるまで、または Postgres ノードが再起動されるまで、最初に発生するまで無効な状態のままです。

### 構文

```
pgd raft disable [OPTIONS]
```

### オプション

`pgd raft disable` コマンドでは、次のオプションが使用できます。

| Short | Long | Description |
|---|---|---|
| | --node | 特定のノードを名前でターゲットにして、その Raft ワーカーを無効にします。 |
| | --group | 特定のグループ内のすべてのノードをターゲットとします。 |
| | --all | クラスター内のすべてのノードをターゲットとします。 |

共通のグローバルオプションについては、 Global Options を参照してください。

### 例

#### 特定のノードで Raft を無効にする

```
pgd raft disable --node kaftan
__OUTPUT__
Raft disabled successfully on node kaftan
```

#### グループ内のすべてのノードの Raft を無効にする

```
pgd raft disable --group dc1_subgroup
__OUTPUT__
Raft disabled successfully on 3 nodes in group dc1_subgroup
```

#### クラスター全体で Raft を無効にする

```
pgd raft disable --all
__OUTPUT__
Raft disabled successfully on 6 nodes in the cluster
```

### pgd raft enable

#### 概要

`pgd raft enable` コマンドは、EDB Postgres 分散クラスター内の 1 つ以上のノードで Raft コンセンサスワーカーを有効にします。

ワーカーを有効にすると、ノードは Raft 選択、ログレプリケーション、およびクラスター全体の状態変更に参加できるようになります。

#### 構文

```
pgd raft enable [OPTIONS]
```

#### オプション

`pgd raft enable` コマンドでは、次のオプションが使用できます。

| Short | Long | Description |
|---|---|---|
| | --node | 特定のノードを名前でターゲットにして、その Raft ワーカーを有効にします。 |
| | --group | 特定のグループ内のすべてのノードをターゲットとします。 |
| | --all | クラスター内のすべてのノードをターゲットとします。 |

共通のグローバルオプションについては、 Global Options を参照してください。

#### 例

##### 特定のノードで Raft を有効にする

```
pgd raft enable --node kaftan
__OUTPUT__
Raft enabled successfully on node kaftan
```

##### グループ内のすべてのノードで Raft を有効にする

```
pgd raft enable --group dc1_subgroup
__OUTPUT__
Raft enabled successfully on 3 nodes in group dc1_subgroup
```

##### クラスター全体で Raft を有効にする

```
pgd raft enable --all
__OUTPUT__
Raft enabled successfully on 6 nodes in the cluster
```

### pgd raft set-leader

#### 概要

`pgd raft set-leader` コマンドは、Raft グループ内のリーダーシップの移転を手動でトリガーします。現在のリーダーにステップダウンするように指示し、指定されたターゲットノードの新しい Raft リーダーとしての選択を促進します。

#### 構文

```
pgd raft set-leader [OPTIONS] <LEADER>
```

#### 引数

| Argument | Description |
|---|---|
| *<LEADER>* | 新しい Raft リーダーになるノードの名前。 |

#### オプション

`pgd raft set-leader` コマンドでは、次のオプションが使用できます。

| Short | Long | Description |
|---|---|---|
| | --group | リーダーを変更するグループの名前。省略すると、グローバルグループが使用されます。 |
| | --wait-for-completion | 設定されている場合、CLI は、ターゲットノードがリーダーシップを正常に引き継ぐまで、または操作がタイムアウトになるまで、クラスター状態をブロックおよび監視します。 |

共通のグローバルオプションについては、 Global Options を参照してください。

#### 例

##### グローバルグループのリーダーを変更する

```
pgd raft set-leader kaboom
__OUTPUT__
Command submitted successfully
```

#### すでにリーダーであるノードを設定しようとします

```
pgd raft set-leader kaftan --group dc1_subgroup
__OUTPUT__
Node kaftan is already the raft leader
```

#### リーダーを変更して、完了を待ちます

--wait-for-completion フラグを使用すると、60 秒のステートメントタイムアウトが設定されます。CLI は、既存の接続を使用して新しいリーダーを検証しようとするか、プロキシ接続マネージャーのようなトラフィックが再ルーティングされる場合は、新しい接続で再試行します。

```
pgd raft set-leader kaolin --group dc1_subgroup --wait-for-completion
__OUTPUT__
Command executed successfully
```

### pgd raft show

#### 概要

pgd raft show コマンドは、EDB Postgres 分散クラスターの Raft コンセンサス情報を表示します。デフォルトでは、3 つのセクションが表示されます。

- **State** すべてのノードのコンセンサス状態、bdr.stat_raft_state から取得。
- **フォロワー** リーダーノードから見たフォロワーの状態、bdr.stat_raft_followers_state からソース。
- **ジャーナル**グローバルコンセンサスジャーナルの詳細、bdr.global_consensus_journal_details から出典。

3 つすべてを表示するにはセクションフラグを指定せずにコマンドを実行するか、--state 、--followers 、または--journal を使用してそのセクションのみを表示します。これらのフラグは相互排他的です。

#### 構文

```
pgd raft show [OPTIONS]
```

#### オプション

pgd raft show コマンドでは、次のオプションが使用できます。

#### セクションフラグ

| Short | Long | Description |
|---|---|---|
| | --state | State セクションのみを表示します `bdr.stat_raft_state`からのコンセンサス状態 |

次のページに続く

表 216 – 前のページからの続き

| Short | Long | Description |
|---|---|---|
| | --followers | Followers セクション`bdr.stat_raft_followers_state`からのフォロワー状態のみを表示します。 |
| | --journal | Journal セクションのみを表示します`bdr.global_consensus_journal_details`からのグローバルジャーナルの詳細。 |

これらのフラグは相互排他的です。

### セクションオプション

| Short | Long | Applies to | Default | Description |
|---|---|---|---|---|
| | --group | All states | | グループ名で出力をフィルタリングします。 |
| -n | --limit | --journal | 20 | 表示されるジャーナルエントリの数を制限します。 |

表形式出力には、使用可能なフィールドのサブセットが表示されます。完全なフィールドセットには`--output json`または`--output markdown`を使用します。

共通のグローバルオプションについては、 Global Options を参照してください。

### 例

#### すべてのノードの Raft の状態を表示する

```
pgd raft show --state
__OUTPUT__
Group Name    Node Name State         Leader Name Term Commit Index Apply Index Last␣
↪Log Index Is Voting Nodes Voting Nodes Protocol Version
- ------------ --------- ------------- ----------- ---- ------------ ----------- -----
↪--------- --------- ----- ------------ -----------------
dc1_subgroup  kaftan    RAFT_LEADER   kaftan      1    4            4           4
         true       3     3            0
dc1_subgroup  kaboom    RAFT_FOLLOWER kaftan      1    4            4           4
         true       3     3            0
dc1_subgroup  kaolin    RAFT_FOLLOWER kaftan      1    4            4           4
         true       3     3            0
democluster   kaftan    RAFT_LEADER   kaftan      0    335          335         335
         true       3     3            5007
democluster   kaboom    RAFT_FOLLOWER kaftan      0    335          335         335
         true       3     3            5007
democluster   kaolin    RAFT_FOLLOWER kaftan      0    335          335         335
         true       3     3            5007
```

この出力では、 dc1_subgroup はローカルルーティングを使用するデータグループであり、 democluster はグローバルルーティングを使用するトップレベルグループです。

Term 列には、現在の Raft 用語番号が表示されます。 Commit Index は、最後にコミットされたログエントリのインデックスを示します。Apply Index および Last Log Index は、ノードがそれぞれログエントリを適用および受信した範囲を示します。Is Voting は、ノードが Raft 選挙に参加するかどうかを示します。Nodes は、グループ内のノードの合計数を示します。 Voting Nodes は、Raft コンセンサスに参加するノードの数を示します。Protocol Version は、使用中の Raft プロトコルのバージョンを示します。

### 特定のグループのフォロワーの状態を表示する

```
pgd raft show --followers --group dc1_subgroup
__OUTPUT__
Group Name    Node Name Sent Commit Index Match Index Clock Drift (ms)
- ------------ --------- ---------------- ----------- ----------------
dc1_subgroup  kaboom    4                4           -2
dc1_subgroup  kaolin    4                3           5
```

Sent Commit Index は、リーダーが各フォロワーに最後に送信したコミットインデックスです。Match Index は、そのフォロワーでレプリケートされたことが確認された最大のログインデックスです。Clock Drift (ms) は、ミリ秒単位のリーダーとフォロワー間のおおよそのクロック差です。

### カスタム制限を使用して仕訳項目を表示する

```
pgd raft show --journal --limit 5
__OUTPUT__
Group Name    Origin Log Index Term Message Type  Response Type
- ------------ ------ --------- ---- ------------ --------------------------
democluster   kaftan 335       0    CONFIG_CHANGE CONFIG_CHANGE_RESPONSE
democluster   kaftan 334       0    CONFIG_CHANGE CONFIG_CHANGE_RESPONSE
democluster   kaftan 333       0    CONFIG_CHANGE CONFIG_CHANGE_RESPONSE
democluster   kaftan 332       0    CONFIG_CHANGE CONFIG_CHANGE_RESPONSE
democluster   kaftan 331       0    CONFIG_CHANGE CONFIG_CHANGE_RESPONSE
```

### 特定のグループのすべてのセクションを表示する

```
pgd raft show --group dc1_subgroup
__OUTPUT__

#  State

Group Name    Node Name State         Leader Name Term Commit Index Apply Index Last␣
↪Log Index Is Voting Nodes Voting Nodes Protocol Version
- ------------ --------- ------------- ----------- ---- ------------ ----------- -----
↪--------- --------- ----- ------------ ----------------
dc1_subgroup  kaftan    RAFT_LEADER   kaftan      1    4            4           4
         true       3      3            0
dc1_subgroup  kaboom    RAFT_FOLLOWER kaftan      1    4            4           4
         true       3      3            0
```

(次のページに続く)

(前のページからの続き)

```
dc1_subgroup  kaolin    RAFT_FOLLOWER kaftan      1    4            4           4
         true       3     3            0

#  Followers

Group Name    Node Name Sent Commit Index Match Index Clock Drift (ms)
- ------------ --------- ---------------- ----------- ----------------
dc1_subgroup  kaboom    4                4           -2
dc1_subgroup  kaolin    4                3           5

#  Journal

Group Name    Origin Log Index Term Message Type  Response Type
- ------------ ------ --------- ---- ------------ --------------------------
dc1_subgroup  kaftan 4         1    CONFIG_CHANGE CONFIG_CHANGE_RESPONSE
dc1_subgroup  kaftan 3         1    CONFIG_CHANGE CONFIG_CHANGE_RESPONSE
dc1_subgroup  kaftan 2         1    CONFIG_CHANGE CONFIG_CHANGE_RESPONSE
```

### pgd raft sync-snapshot

#### 概要

`pgd raft sync-snapshot` コマンドは、 EDB Postgres 分散クラスター内のすべてのノードで Raft 状態を同期させます、その状態が分岐した場合。

コマンドは、変更を行う前にプリフライトチェックを実行します。BDR バージョンを検証し、Raft リーダーを識別し、すべてのノードに到達可能であることを確認し、状態が既に同期しているかどうかを確認します。状態が既に一貫している場合、`--ignore-sync-status` が指定されていない限り、コマンドは変更せずに終了します。

変更が必要な場合、次のコマンド

1. すべてのノードで Raft を無効にします。
2. リーダーノードから現在の Raft 状態をエクスポートします。
3. リーダーの Raft 状態をすべてのフォロワーノードにインポートします。
4. すべてのノードで Raft を再度有効にし、リーダーが選択されるのを待ちます。

サブスクライバー専用グループは独自の Raft インスタンスを持たず、このコマンドの有効なターゲットではありません。サブスクライバ専用ノードを含む Raft 状態を同期するには、`--group` なしでコマンドを実行します。

#### 構文

```
pgd raft sync-snapshot [OPTIONS]
```

### オプション

`pgd raft sync-snapshot` コマンドでは、次のオプションが使用できます。

| Option | Default | Description |
|---|---|---|
| --group | | 名前で特定のグループをターゲットにします。指定しない場合、デフォルトのトップレベルグループ。 |
| --check | false | クラスターに変更を加えずに、プリフライトチェックのみを実行します。 |
| --ignore-sync-status | false | Raft の状態がすべてのノードで既に一貫している場合でも、同期を続行します。 |

共通のグローバルオプションについては、 Global Options を参照してください。

### 例

#### プリフライトチェックのみを実行する

```
pgd raft sync-snapshot --check
__OUTPUT__
Performing Raft Sync-Snapshot Pre-flight Checks
- ------------------------------------------------
Checking BDR version                                    ok (6.4.0)
Resolving target group                                  ok (dc-1)
Identifying raft leader                                 ok (leader: pgd-a1)
Discovering cluster nodes                               ok (3 nodes)
Connecting to pgd-a1 (leader)                           ok
Connecting to pgd-a2                                    ok
Connecting to pgd-a3                                    ok
Validating node states                                  ok
Checking raft sync status                               ok (out of sync)

Raft Sync-Snapshot Pre-flight Report
- -----------------------------------
Target group:    dc-1
Raft leader:     pgd-a1
Nodes in scope:  3 (pgd-a1, pgd-a2, pgd-a3)

All pre-flight checks passed. Ready for sync-snapshot.
```

### グローバルグループを同期する

```
pgd raft sync-snapshot
__OUTPUT__
Performing Raft Sync-Snapshot Pre-flight Checks
- ------------------------------------------------
Checking BDR version                                        ok (6.4.0)
Resolving target group                                      ok (dc-1)
Identifying raft leader                                     ok (leader: pgd-a1)
Discovering cluster nodes                                   ok (3 nodes)
Connecting to pgd-a1 (leader)                               ok
Connecting to pgd-a2                                        ok
Connecting to pgd-a3                                        ok
Validating node states                                      ok
Checking raft sync status                                   ok (out of sync)

Performing Raft Sync-Snapshot
- ----------------------------
Disabling raft on pgd-a1                                    WARNING:  disabling the␣
↪consensus
ok
Disabling raft on pgd-a2                                    ok (already disabled)
Disabling raft on pgd-a3                                    WARNING:  disabling the␣
↪consensus
ok
Exporting snapshot from leader (pgd-a1)                     ok
Importing snapshot on pgd-a2                                ok
Importing snapshot on pgd-a3                                ok
Re-enabling raft on pgd-a1                                  ok
Re-enabling raft on pgd-a2                                  ok
Re-enabling raft on pgd-a3                                  ok
Waiting for raft recovery                                   ok (leader: pgd-a3)

Raft sync-snapshot completed successfully.
```

## pgd replication

pgd replication コマンドは、 EDB Postgres 分散クラスターのレプリケーションステータスのさまざまな側面を表示するために使用されます。

#### サブコマンド

- *pgd cluster show* クラスターのレプリケーションステータスを表示します。

### pgd replication show

#### 概要

`pgd replication show` コマンドは、EDB Postgres 分散クラスターでのレプリケーションステータスを表示するために使用されます。

オプションなしのデフォルトでは、次のレポートを生成します。

- ノードレプリケーションの進行状況ノード間のレプリケーションステータスのマトリックス。
- Replication Slots 各ノードのスロットのレプリケーションスロットステータス。
- サブスクリプションノード間の各サブスクリプションのサブスクリプションステータス。
- 分析レプリケーション 各ノードの分析レプリケーションステータス。

オプションを使用して、出力を上記のレポートのいずれかに制限できます。`--verbose` オプションを使用して、デフォルトレポートの詳細を増やし、各ノードの他のノードへの接続の LSN とレプリケーションラグを表示できます。

#### 構文

```
pgd replication show [OPTIONS]
```

#### オプション

`pgd replication show` コマンドでは、次のオプションが使用できます。

| Short | Long | Description |
|---|---|---|
| | --nodes | ノード間のレプリケーション状態のみをマトリックス形式で表示します。 |
| | --slots | 各ノードのレプリケーションスロットを表示します。 |
| | --subscriptions | ノード間の各サブスクリプションのサブスクリプションステータスを表示します。 |
| | --analytics | 各ノードの分析レプリケーションステータスを表示します。 |
| -v | --verbose | レプリケーションステータスに関する詳細情報を表示します。 |

共通のグローバルオプションについては、 グローバルオプション を参照してください。

### –slots

これは、 BDR レプリケーションスロットのステータスを示します。verbose フラグを使用した出力は、スロットがアクティブか、レプリケーション状態切断、ストリーミング、キャッチアップ、おおよその遅延などの詳細を提供します。

マトリックスビューでは、バイトラグが括弧内に表示される場合があります。maxOf(WriteLag, FlushLag, ReplayLag, SentLag) です。

#### 例

#### EDB Postgres 分散クラスターでのレプリケーションステータスの表示

```
pgd replication show
__OUTPUT__

#  Node Replication Progress

Node   kaboom kaftan kaolin
- ----- ------ ------ ------
kaboom -       **
kaftan  *-*
kaolin  **       -

#  Replication Slots

Group Name   Origin Node Target Node Slot Name                      Active State    ␣
↪Write Lag Replay Lag Sent Lag Bytes Write Lag Bytes Replay Lag Bytes
- ----------- ----------- ----------- ------------------------------ ------ ---------␣
↪--------- ---------- -------------- --------------- ----------------
dc1_subgroup kaboom      kaftan      bdr_pgddb_democluster15_kaftan t      streaming␣
↪00:00:00  00:00:00   0               0               0
dc1_subgroup kaboom      kaolin      bdr_pgddb_democluster15_kaolin t      streaming␣
↪00:00:00  00:00:00   0               0               0
dc1_subgroup kaftan      kaboom      bdr_pgddb_democluster15_kaboom t      streaming␣
↪00:00:00  00:00:00   0               0               0
dc1_subgroup kaftan      kaolin      bdr_pgddb_democluster15_kaolin t      streaming␣
↪00:00:00  00:00:00   0               0               0
dc1_subgroup kaolin      kaboom      bdr_pgddb_democluster15_kaboom t      streaming␣
↪00:00:00  00:00:00   0               0               0
dc1_subgroup kaolin      kaftan      bdr_pgddb_democluster15_kaftan t      streaming␣
↪00:00:00  00:00:00   0               0               0

#  Subscriptions

Origin Node Target Node Last Applied Tx Timestamp      Last Applied Tx Age␣
↪Subscription Status
```

(次のページに続く)

(前のページからの続き)

```
- ---------- ---------- ---------------------------- ------------------- ----------
↪---------
kaboom      kaftan      2025-02-21 19:18:12.661520 UTC 00:00:18.616        replicating
kaboom      kaolin      2025-02-21 19:18:12.661520 UTC 00:00:18.939        replicating
kaftan      kaboom      2025-02-21 19:18:12.658069 UTC 00:00:18.787        replicating
kaftan      kaolin      2025-02-21 19:18:12.658069 UTC 00:00:18.943        replicating
kaolin      kaboom      2025-02-21 19:18:12.663201 UTC 00:00:18.782        replicating
kaolin      kaftan      2025-02-21 19:18:12.663201 UTC 00:00:18.614        replicating

#  Analytics Replication Progress

Origin Node Replicating Node Replicated LSN Last Updated
- ---------- ---------------- -------------- ------------
```

### ノード間のレプリケーション状態のみをマトリックス形式で表示する

```
pgd replication show --nodes
__OUTPUT__
Node   kaboom kaftan kaolin
- ----- ------ ------ ------
kaboom -       **
kaftan  *-*
kaolin  **       -
```