Eon モードでのデータベース管理者のオペレーション関連の考慮事項

原文はこちら

本ドキュメントの目的は、Vertica Enterprise モードをすでにお使いのユーザー様が、Vertica Enterprise モードと Vertica Eon モードの動作上の類似点と相違点のいくつかについて習得することです。このドキュメントでは、Eon モードのアーキテクチャーについて基本的な知識があることを前提としています。Eon モードのアーキテクチャーの詳細については、Vertica documentation を参照してください。

これは 2 部構成のシリーズの第 1 部です。

プロジェクション設計

Eon モードは、Enterprise モードのように、分散プロジェクション（Segmented Projection）、複製プロジェクション（Unsegmented Projection）、アグリゲートプロジェクション（Aggregate Projection）をサポートします。分散プロジェクションのバディープロジェクションを作成する必要はありません。

プロジェクションの定義、プロジェクションの使用方法、およびプロジェクションの作成と最適化のプロセスは、Enterprise モードに似ています。アンカーテーブルには、少なくとも 1 つのスーパープロジェクションが必要で、クエリスペシフィックプロジェクションを 0 個以上持つことができます。

Vertica には、最適化されたプロジェクション設計を作成するのに役立つデータベースデザイナーが用意されています。Enterprise モードのデータベースからエクスポートされたデザインファイルは、ファイルを変更せずに Eon モードデータベースにインポートできます。

Eon モードでの分散プロジェクションの例を次に示します。

```
=> CREATE PROJECTION values_p (a, b, c)
AS SELECT a, b, c
FROM values
SEGMENTED BY HASH(a) ALL NODES;   --KSAFE keyword is optional
```

K-Safety と高可用性

Eon モードでは、分散プロジェクションがシャードに分割されます。データベース内のシャードの数は、クラスター内のパーマネントノードの数に関係なく、Eon モードデータベースを作成したときに指定したものに関連付けられます。

K-Safety が 1 であるデータベースでは、セグメントシャード（Segment shard）には少なくとも 2 つの異なるノードが登録され、すべてのノードがサブスクライブするレプリカシャード（Replica shard）があります。各シャードには 1 つのプライマリノードサブスクリプションがあります。

例えば、データベースに 3 つのノードと 3 つのシャードがある場合、次のシステムテーブルのクエリを使用してノードのサブスクリプションを表示します。

```
testeon=> SELECT node_name , shard_name ,subscription_state,is_primary FROM
node_subscriptions ORDER BY 1,2;
node_name         | shard_name  | subscription_state | is_primary
--------------------+-------------+--------------------+------------
 v_testeon_node0001 | replica     | ACTIVE             | f
 v_testeon_node0001 | segment0001 | ACTIVE             | f
 v_testeon_node0001 | segment0002 | ACTIVE             | t
 v_testeon_node0002 | replica     | ACTIVE             | t
 v_testeon_node0002 | segment0001 | ACTIVE             | t
 v_testeon_node0002 | segment0003 | ACTIVE             | f
 v_testeon_node0003 | replica     | ACTIVE             | f
 v_testeon_node0003 | segment0002 | ACTIVE             | f
 v_testeon_node0003 | segment0003 | ACTIVE             | t
(9 rows)
```

セッションを開始すると、Vertica はクエリに対応する一連のノードとシャードサブスクリプションを割り当てます。セッションサブスクリプションの一部であったノードがダウンすると、セッションのノードおよび／またはシャードサブスクリプションが自動的に調整されます。この処理はユーザーには見えない形で実行されますが、100%のデータを照会することが可能です。

テストデータベースには、Eon モードで動作する 3 つのノードと、3 つのシャードがあります。すべてのノードが UP になると、次のセッションサブスクリプションが作成されます。

```
testeon=> SELECT node_name , shard_name
FROM session_subscriptions WHERE is_participating;
     node_name      | shard_name
--------------------+-------------
 v_testeon_node0002 | replica
 v_testeon_node0001 | replica
 v_testeon_node0003 | replica
 v_testeon_node0001 | segment0001
```

```
 v_testeon_node0003 | segment0002
 v_testeon_node0002 | segment0003
(6 rows)
```

v_testeon_node0003 が DOWN になった後、セッションのサブスクリプションが変更されました。シャード segment0002 は、ノード v_testeon_node0001（ハイライト表示されているもの）を介して使用できるようになりました。v_testeon_node0002 上のシャード segment0003 のサブスクリプションは自動的にプライマリに切り替わります。

```
testeon=> SELECT node_name , shard_name FROM session_subscriptions WHERE
is_participating;
     node_name      | shard_name
--------------------+-------------
 v_testeon_node0002 | replica
 v_testeon_node0001 | replica
 v_testeon_node0001 | segment0001
 v_testeon_node0001 | segment0002
 v_testeon_node0002 | segment0003
(5 rows)
```

v_testeon_node0002 と v_testeon_node0003 が DOWN になると、シャード segment0003 が UP ノードによってサブスクライブされていないため、データベースは UNSAFE モードでシャットダウンします。

ノードの再起動とリカバリ

Eon モードでノードを再起動すると、ノードはクラスターに参加した後 UP 状態に移行します。Eon モードでは、カタログイベントのリプレイ、履歴データのリカバリ、あるいは Replay Delete は発生しません。

ノードが UP 状態に移行すると、ノードがサブスクライブするシャードのシャード subscription_state が INACTIVE から PENDING に変更されます。PENDING 状態では、ノードは同じシャードにサブスクライブしている別のノードからシャードメタデータを受信します。ノードが別のノードからシャードメタデータを受信した後、シャードサブスクリプションの状態は、PASSIVE 状態に遷移します。PASSIVE 状態では、ノードはウォームデポのシャードサブスクリプションの欠落したファイルをキューに入れてフェッチします。キューに入れられたすべてのファイルがフェッチされると、シャードサブスクリプションの状態は ACTIVE とマークされます。あるノードのすべてのシャードサブスクリプションが ACTIVE 状態になると、そのノードはクエリを実行できるようになります。

構成パラメーターEnableDepotWarmingFromPeers を使用して、ノードのリカバリ中にデポのウォーミングをスキップすることができます。このパラメーターが 0 に設定されていると、ノードはデポをウォーミングせずに PENDING 状態から ACTIVE 状態に遷移します。この場合、デポに見つからないデータにヒットしたクエリは、最初のヒット時にパフォーマンスの低下を示します。デポに見つからないデータファイルは、今後のアクセスのために共有ストレージから自動的にフェッチされます。

ノードダウン時のパフォーマンス

Enterprise モードでは、Vertica オプティマイザによって生成されたクエリプランが異なる可能性があるため、すべてのノードが UP でない場合、一部のクエリのパフォーマンスが大幅に低下する可能性があります。Eon モードでは、Enterprise モードで起こりうるノードダウン時のパフォーマンス劣化と同じ問題には直面しません。Eon モードでは、バディープロジェクションがないため、すべてのノードが UP しているかどうかにかかわらず、クエリプランは同じです。

ノードの追加と削除（リバランシング）

Enterprise モードでは、新しいノードがクラスターに追加または削除されると、セグメント数とセグメント境界が変更されます。既存のノード上の ROS コンテナは、セグメントレイアウトの変更に合わせて分離して転送する必要があります。このプロセスは、リバランスと呼ばれ、各ノードのデータのサイズに応じて、かなりの時間がかかる I/O インテンシブな処理です。クラスター内のすべてのノードでリバランス処理を実行するには、すべてのノードが UP 状態である必要があります。したがって、DOWN のノードはクラスターから削除できません。

Eon モードでは、ノードがクラスターに追加または削除されるときに、rebalance_shards 関数を実行して、ノード間でシャードサブスクリプションを再配布できます。ノードが新しいシャードにサブスクライブすると、クエリを実行できるようになる前にデポがウォームアップします。このプロセスでは ROS ファイルを分割する必要はなく、rebalance_cluster よりもはるかに高速です。シャードサブスクリプションへの変更は、すべてのノードが UP していない場合に許可されるため、すべてのノードが UP でない場合に、ノードをクラスターに追加または削除できます。

次の内容は、4 つのシャードを持つ 3 ノードクラスターのノードサブスクリプションを示しています。

```
testeon=> SELECT node_name , shard_name ,subscription_state,is_primary FROM
node_subscriptions ORDER BY 1,2;
     node_name      | shard_name  | subscription_state | is_primary
--------------------+-------------+--------------------+------------
 v_testeon_node0001 | replica     | ACTIVE             | f
 v_testeon_node0001 | segment0001 | ACTIVE             | t
 v_testeon_node0001 | segment0002 | ACTIVE             | t
 v_testeon_node0001 | segment0004 | ACTIVE             | f
 v_testeon_node0002 | replica     | ACTIVE             | t
 v_testeon_node0002 | segment0001 | ACTIVE             | f
 v_testeon_node0002 | segment0003 | ACTIVE             | f
 v_testeon_node0002 | segment0004 | ACTIVE             | t
 v_testeon_node0003 | replica     | ACTIVE             | f
 v_testeon_node0003 | segment0002 | ACTIVE             | f
 v_testeon_node0003 | segment0003 | ACTIVE             | t
(11 rows)
```

admintools を使用して新しいノードをクラスターに追加し、rebalance_shards を実行して 4 ノードのクラスターにすることができます。

```
testeon=> SELECT rebalance_shards();
 rebalance_shards
-------------------
 REBALANCED SHARDS
(1 row)
```

次の内容では、新しいノードが追加され、シャードがリバランスされた後のノードのサブスクリプションを示しています。

```
testeon=> SELECT node_name , shard_name ,subscription_state,is_primary FROM
node_subscriptions ORDER BY 1,2;
     node_name      | shard_name  | subscription_state | is_primary
--------------------+-------------+--------------------+------------
 v_testeon_node0001 | replica     | ACTIVE             | f
 v_testeon_node0001 | segment0001 | ACTIVE             | f
 v_testeon_node0001 | segment0002 | ACTIVE             | t
 v_testeon_node0002 | replica     | ACTIVE             | t
 v_testeon_node0002 | segment0001 | ACTIVE             | t
 v_testeon_node0002 | segment0004 | ACTIVE             | f
 v_testeon_node0003 | replica     | ACTIVE             | f
 v_testeon_node0003 | segment0002 | ACTIVE             | f
 v_testeon_node0003 | segment0003 | ACTIVE             | t
 v_testeon_node0004 | replica     | ACTIVE             | f
 v_testeon_node0004 | segment0003 | ACTIVE             | f
 v_testeon_node0004 | segment0004 | ACTIVE             | t
(12 rows)
```

K-Safety を 0 に減らし、1 つのノードを除くすべてのノードを削除することによって、クラスターを単一のノードに縮小できます。

```
testeon=> SELECT mark_design_ksafe(0);
WARNING 6022:  Setting K-safety to 0 could result in catastrophic data loss
in the event of a failure.  Do not use k=0 in a production environment. For
test, dev or other non-production environments, K=0 may be acceptable however
Vertica still recommends a minimum value of K=1
  mark_design_ksafe
----------------------
 Marked design 0-safe
(1 row)
```

次に、新しいシャードサブスクリプションを表示します。

```
testeon=> SELECT node_name , shard_name ,subscription_state,is_primary FROM
node_subscriptions ORDER BY 1,2;
     node_name      | shard_name  | subscription_state | is_primary
--------------------+-------------+--------------------+------------
 v_testeon_node0001 | replica     | ACTIVE             | t
 v_testeon_node0001 | segment0001 | ACTIVE             | t
 v_testeon_node0001 | segment0002 | ACTIVE             | t
 v_testeon_node0001 | segment0003 | ACTIVE             | t
 v_testeon_node0001 | segment0004 | ACTIVE             | t
(5 rows)
```

フォールトグループとエラスティッククラススケーリング

Enterprise モードでは、フォールトグループを作成して Vertica のラックを認識させ、バディーノードを同じラックに配置することができます。大規模クラスターで実行されるクエリのバッファー要件を軽減するテラスルーティングを利用することもできます。

Eon モードでは、クラスターにノードを追加するとクエリスループットに効果があります。フォールトグループを作成して、エラスティックスループットのスケーリングを利用することができます。フォールトグループ内のノードの数がシャードの総数以上である場合、フォールトグループ内のノードに接続する Vertica セッションは、そのフォールトグループ内のノードのみを使用してクエリを実行します。フォールトグループ内のノードに接続する Vertica セッションは、UP 状態のノードがクエリに応答するのに十分なノードがない場合にのみ、他のフォールトグループ内のノードを使用できます。

```
testeon=> CREATE FAULT GROUP group1;
CREATE FAULT GROUP
testeon=> CREATE FAULT GROUP group2;
CREATE FAULT GROUP
testeon=> ALTER FAULT GROUP group1 ADD node  v_testeon_node0001;
ALTER FAULT GROUP
testeon=> ALTER FAULT GROUP group1 ADD node  v_testeon_node0002;
ALTER FAULT GROUP
testeon=> ALTER FAULT GROUP group1 ADD node  v_testeon_node0003;
ALTER FAULT GROUP
testeon=> ALTER FAULT GROUP group2 ADD node  v_testeon_node0004;
ALTER FAULT GROUP
testeon=> ALTER FAULT GROUP group2 ADD node  v_testeon_node0005;
ALTER FAULT GROUP
testeon=> ALTER FAULT GROUP group2 ADD node  v_testeon_node0006;
ALTER FAULT GROUP
```

更新されたノードのサブスクリプションを次に示します。

```
testeon=> SELECT node_name , shard_name ,subscription_state,is_primary FROM
node_subscriptions ORDER BY 1,2;
     node_name      | shard_name  | subscription_state | is_primary
--------------------+-------------+--------------------+------------
 v_testeon_node0001 | replica     | ACTIVE             | f
 v_testeon_node0001 | segment0001 | ACTIVE             | f
 v_testeon_node0001 | segment0002 | ACTIVE             | f
 v_testeon_node0002 | replica     | ACTIVE             | t
 v_testeon_node0002 | segment0001 | ACTIVE             | f
 v_testeon_node0002 | segment0003 | ACTIVE             | f
 v_testeon_node0003 | replica     | ACTIVE             | f
 v_testeon_node0003 | segment0002 | ACTIVE             | f
 v_testeon_node0003 | segment0003 | ACTIVE             | f
 v_testeon_node0004 | replica     | ACTIVE             | f
 v_testeon_node0004 | segment0001 | ACTIVE             | t
 v_testeon_node0004 | segment0003 | ACTIVE             | f
 v_testeon_node0005 | replica     | ACTIVE             | f
 v_testeon_node0005 | segment0001 | ACTIVE             | f
 v_testeon_node0005 | segment0002 | ACTIVE             | t
 v_testeon_node0006 | replica     | ACTIVE             | f
 v_testeon_node0006 | segment0002 | ACTIVE             | f
 v_testeon_node0006 | segment0003 | ACTIVE             | t
(18 rows)
```

ノード v_testeon_node0004 を介して vsql を使用してデータベースに接続すると、フォールトグループ group2 からのノードだけが取得されます。

```
testeon=> SELECT node_name , shard_name FROM session_subscriptions WHERE
is_participating;
     node_name      | shard_name
--------------------+-------------
 v_testeon_node0004 | replica
 v_testeon_node0005 | replica
 v_testeon_node0006 | replica
 v_testeon_node0005 | segment0001
 v_testeon_node0006 | segment0002
 v_testeon_node0004 | segment0003
(6 rows)
```

WOS とロードストラテジー

Eon モードでは、WOS は使用できません。Eon モードでは、バルクロードが推奨されています。

Tuple Mover

Eon モードでは、Tuple Mover 処理はグローバルレベルの処理であり、シャードへのプライマリーサブスクリプションを持つノードが mergeout 処理の実行を担当します。mergeout 処理の出力は、共用ストレージおよびシャードに加入している他のノードに出力されます。Eon モードでは、mergeout のためにプロジェクションを確認して選択する mergeout アルゴリズムに変更はありません。Eon モードでは、WOS がないため、moveout 処理が無効になります。

エポック

Eon モードでは、WOS が存在しないため、last_good_epoch 値は常に最新（-1）です。AdvanceAHMInterval 構成パラメーターの値を変更しない限り、AHM エポックは 3 分ごとに last_good_epoch に進められます。

Delete、Update と Replay Delete

Eon モードでは、delete および update のプランは Enterprise モードと同じです。Replay delete のプランは、moveout 処理、リカバリ、リバランス、またはバディープロジェクションの追加がないため、mergeout 処理にのみ適用されます。WOS がないため、すべての update と delete のプランは、ROS または DVROS ファイルに書き込みます。

リソースプールとワークロード管理

リソースプール、リソースプールに関連するシステムテーブル、およびワークロードを管理するためのリソースプールの設定は、Eon モードと Enterprise モードの場合で同じです。Eon モードでは、複数のフォールトグループを設定することで、エラスティックスループットのスケーリングを利用できます。

DML クエリは、セッションサブスクリプションの一部ではないノード上でごく少量のメモリを予約しますが、ピアツーピアのキャッシュリフィルの一部としてファイルを受け取ります。

S3 から削除された ROS ファイルをクリア

DROP TABLE、TRUNCATE TABLE あるいは mergeout 処理の結果として ROS ファイルが削除されると、ファイルはデポから直ちに削除され、2 時間後に共有ストレージ上から削除されます。

システムテーブル、Vertica カタログ、バックアップとリストア、2 つの Vertica ノード間の移行、アップグレードなどのトピックをカバーする本書の第 2 部は近日中に発行予定です。