IBM Db2 Warehouse on Cloud の高可用性とリカバリー

Aniket Kulkarni, IBM
Littleton, Massachusetts USA
email: aniket.kulkarni@us.ibm.com

Venkatesh Gopal, IBM
Leawood, Kansas USA
email: gopalv@us.ibmcom

要約
クラウド・サービスを測定する主要メトリクスの一つは可用性です。クラウド・サービス・プロバイダーは自社のサービスを、基盤コンポーネントの故障リスクを緩和するように構築しなければなりません。故障リスクはサービスの規模と複雑さが成長するにつれて増加します。

高可用性はこれらのリスクに対処するサービスの機能です。もっとも基本的なレベルでは、高可用性 (HA) は (a) コンポーネントの故障を検出し、(b) 自動的にそこからリカバリーする機能によって達成されます。Kubernetes などのコンテナ・オーケストレーション、管理システムである程度は可能ですが、特定種類のサービスでさらに高いレベルの信頼性を保証するためにその上に構築するさまざまな方法が存在しています。

このホワイトペーパーでは、IBM のクラウド・データウェアハウスである IBM® Db2® Warehouse on Cloud で高可用性と信頼性を達成している方法に焦点をあてます。はっきりと強調するのは、コンテナ化したサービス・モデル、多層信頼性、絶望的な故障からのリカバリー機能です。

多層信頼性
Db2 Warehouse on Cloud は顧客にシェアード・ナッシング、大規模パラレル・クラウド・データ・ウェアハウスをコンテナ化したサービス [1] モデルとして IBM Cloud™ Container サービス上に構築して、提供します。IBM Cloud Container サービス自体は、コンテナ化したアプリケーションの自動展開、スケーリング、管理用のオープンソースのコンテナ・オーケストレーション・システム、Kubernetes[2] 上に構築されています。

Db2 Warehouse on Cloud で高可用性と信頼性をサポートするために、IBM は Kubernetes 基盤の上に以下のような多層信頼性デザインを構築しました :

- 層 1: システムに単一故障点がありません。すべてのコンポーネントに 1 つまたは複数の冗長コピーがあり、複数の冗長コンポーネントが故障したときには自動リカバリーが開始されます
- 層 2: コンテナ内のコンポーネントが故障したら、自動でコンテナ内リカバリー
- 層 3: コンテナ自体が故障したら、自動でコンテナ・リカバリー
- 層 4: コンテナをホスティングしているサーバーがシステム故障を起こしたら、自動でサービス・リカバリー

さらに Db2 Warehouse on Cloud はユーザーが定義し、管理する高速バックアップ、リストア機能を提供しています。サービスには 24x7 DevOps とモニタリング・サポートもあり、複数層信頼性レイヤーでも対応できない複数の、ないしはリカバリーできない故障の際に は故障とシステム・ヘルスを追跡します。

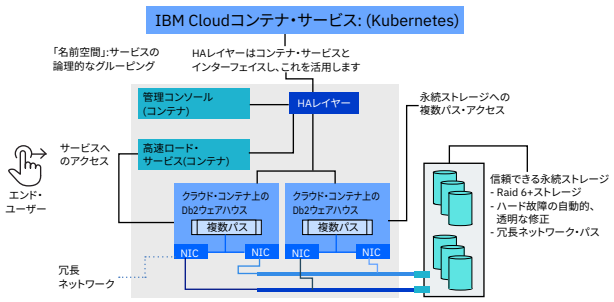

図 1: Db2 Warehouse on Cloud サービスの信頼性の裏にあるアーキテクチャーの高レベル・ビュー

- サービスのそれぞれのコンポーネントはコンテナです
- HA レイヤーは Kubernetes を拡張、強化して、高速の HA を提供します
- 永続ストレージへのネットワーク接続は冗長でマルチパスによって論理的にグループ化されています
- HA レイヤーは コンテナ内部のコンポーネントが故障したときに、コンテナを再起動することなく、コンテナをリカバリーするので、すばやくリカバリーできます
- 永続ストレージにはネットワーク、ストレージ・レベルの複数の冗長性があります

それぞれの層を眺めるとともに、バックアップ、リストア機能を深くまで探ってみましょう。

信頼性層 1: コンポーネント冗長性

Db2 Warehouse on Cloud はエラスティックなデータ・ウェアハウスで、ユーザーは CPU コア (「コンピュート」) とストレージ・キャパシティをオンドマンドで独立してスケールできるので、変化する業務要件に対応できます。このエスティック性を可能にするアーキテクチャーはコンピュートとストレージの疎結合によって構築されています。コンピュート・キャパシティは Kubernetes の上に構築された IBM Cloud コンテナ・サービスによって管理されています。ストレージはネットワーク接続の高パフォーマンス IBM Cloud ブロック・ストレージに配置されています。[3]

ネットワーク自体が信頼性とパフォーマンスのために重要なコンポーネントなので、IBM はネットワークにも単一故障点がないことを保証しています。基盤サーバーのそれぞれには冗長ネットワーク・インターフェイス・コントローラー (NIC)2 枚が接続され、ストレージ・バックプレーンにつながる別々のパスに接続されていて、可能な場合には (ロード・バランスによって) 高パフォーマンスを保証するように束ねられます。さらに、ストレージ・バックプレーンには一対の冗長「ターゲット」、ないしはアクセス・ポイントがあり、これによってストレージ・インフラストラクチャーへのネットワーク・アクセスが可能になります。コンピュート (コンテナ) 側では、アクティブ / パッシブな複数パス構成で、ストレージ・パスの一方が故障しても、システムがもう一方に瞬時に切り換えることで、中断のない「フェイルオーバー」を保証します。

上記に加えて、ストレージ・バックエンドは複数の冗長、ホットスワップ対応の SSD を RAID 6+ 構成で利用しています。これは従来の RAID 6 の改良版で、SSD の管理、問題の修復を中断なしに可能にします。

信頼性層 2: コンテナ内リカバリー

従来の Kubernetes や Web サービス・モデルでは、コンテナ内のコンポーネントの故障にはコンテナ全体の再起動で対応します。単純で効果的ですが、これは基幹系のアプリケーションに高レベルの稼働時間を要求するユーザーをサポートするのには充分なものではありません。

大規模なデータセットを扱うクラウド・データ・ウェアハウスの負荷の性質を考慮すると、コンテナの故障コンポーネント一つをリカバリーするほうがコンテナ全体の再起動よりも高速です。

完全な再起動にはコンテナをストレージと切り離して、コンピュートを捨てて、新しいコンピュートを選択してから、ストレージを再結合する必要があります。これに対して、影響を受けたコンポーネントに直接対応すると、大きく時間が短縮されて、目標復旧時間 (RTO) を最小に保てます。

従来のモデルのこの欠点にはリカバリーを 2 つの層、層 2 と層 3 に分割することで対応しています。Db2 Warehouse on Cloud では「HA レイヤー」はコンテナそれぞれに拡張されていて、「コンテナ HA レイヤー」という適切な名前で呼ばれています。コンテナ HA レイヤーはコンテナ内の各コンポーネントの健康を追跡して、コンポーネントのどれかが故障したら、リカバリーを行います。

たとえば Db2 コンテナでデータ・パーティション・プロセスのひとつで故障が発生したとします。そのコンテナのコンテナ HA レイヤーがこの故障を検出して、それらのデータ・パーティションの再起動を試みます。その後、データベース・リカバリーでデータの一貫性を保証します。さらに、故障が連続して発生したら、たとえば1つのデータ・パーティションの故障が別のパーティションのリカバリー中に発生したら、コンテナ HA レイヤーは進行中のリカバリーを打ち切って、両方のパーティションのクリーン・リカバリーを一緒に行います。

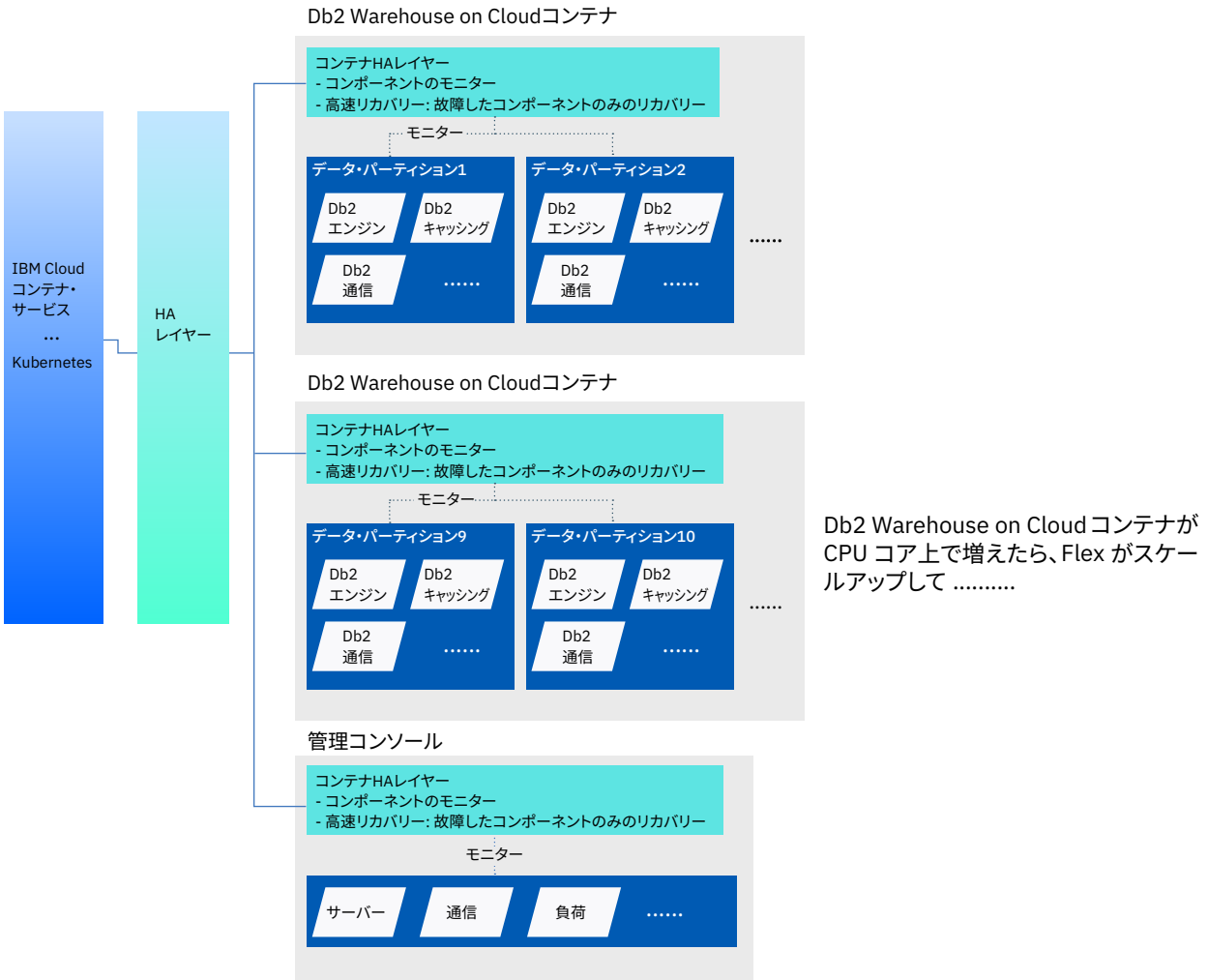

図 2: 層 2 コンテナ内 HA:

- それぞれのコンテナには HA レイヤーが組み込まれているので、そのコンテナ内で有されたコンポーネントのサブセットのみが故障した場合には高速 HA を提供しようとします
- Db2 Warehouse on Cloud コンテナでは HA レイヤーはそれぞれのパーティションとパーティション内のコンポーネントをモニターします。パーティションの一部が故障すると、それらのパーティションのみが自動的に再起動され、リカバリーされます
- Db2 Warehouse on Cloud 以外のコンテナでは、コンテナ内のそれぞれのコンポーネントが同様にモニターされて、可能であればリカバリーされます
- HA レイヤーは Container Service の HA レイヤーとインターフェイスし、それを拡張します。これによって層 2 リカバリーが不可能な場合には層 3 リカバリーにエスカレーションできます

信頼性層 3: コンテナ・リカバリー

コンテナのどれかで複数のエラーが発生すると、ないしはコンテナ自体が終了された場合は、層 3 リカバリーが開始されます。この場合、HA レイヤーはコンテナ再起動の従来の Kubernetes HA モデルに戻ります。そして :

1. 影響を受けたコンテナはクリーンに停止されて、クリーンアップされます。
2. 関連の永続ストレージが切り離されてから、コンテナ・サービスは新たなコンピュート・コア群にコンテナを再スケジュールします。
3. その時点でストレージは自動的に再結合されて、コンテナが起動されます。

複数コンテナで構成される、Db2 Warehouse on Cloud などの複数コンテナ・サービスでは HA レイヤーがすべてのコンテナに渡る同期データベース・リカバリーを調整していることに注意してください。

層 2 信頼性からのフォールバック以外に、このリカバリーはネットワーク・レベルでの複数故障の際にデーター貫性を保つ重要なユース・ケースにも対応しています。とくに、冗長ネットワーク・パスのそれぞれで複数のネットワーク故障が発生するレア・ケースでは、データー貫性を守るために Db2 Warehouse on Cloud はストレージを読み取り専用保護モードにします。

HA レイヤーがこの状況を検出すると、コンテナは強制停止されて、切り離され、再スケジュールされ、結合されて、再起動されます。これによってコンテナが健全なネットワークを持つ別の基盤サーバーに移動されて、ストレージの保護モードが解除されるので、データー貫性を保って、システムのリカバリーに成功します。

信頼性層 4: サービス・リカバリー

コンテナ・リカバリーはもっとも伝統的な Kubernetes HA モデルです。HA レイヤーによって Kubernetes はコンテナにコンピュート・コアを提供している基盤サーバー自体の故障を検出できます。

リカバリー動作は層 3 とよく似ています。すべてのコンテナが停止されて、すべての影響を受けるストレージが切り離されて、すべての影響を受けるコンテナが再スケジュールされて、すべてのストレージが結合されて、すべての影響を受けるコンテナが再起動されます。

故障した、ないしは応答のなくなったサーバーは連続してハートビートが返ってこないことで HA レイヤーに検出されます。猶予期間が経過すると、リカバリー動作（すでに言及したもの) が行われ、サーバーには「スケジュール不可」と印が付けられて、他のコンテナを開始するのには利用されなくなります。これによって運用チームは影響を受けたサーバーの保守を行って、エンド・ユーザーにさらなる中断を与えることなく回復させられます。

バックアップとリストア :

ユーザーがデータベースのバックアップ、リストアを行える機能はクラウド・サービスをユーザーのミスから守るのにきわめて重要です。極端なレア・ケースでは、全体的な故障からリカバリーするための最終手段でもあります。Db2 Warehouse on Cloud ではユーザーが単純なインターフェイスでこの機能を自分で管理できます :

- ユーザーは自社の業務にもっとも都合がいいときにバックアップが実行できるようスケジュールできます。設定したスケジュールに基づいてバックアップは 24 時間ごとに実行されます。
- 最新 7 回のバックアップが維持されて、必要に応じて、ユーザーはボタンを1回クリックするだけでそれらのバックアップのひとつからデータベースをリストアできます。

リダイレクト・オン・ライト・スナップショット : 電光のようにすばやいバックアップとリストアの鍵

信頼できる永続ストレージ層はバックアップとして高速の、ほとんど瞬時のリダイレクト・オン・ライト・スナップショットもサポートしています。データベースのそれぞれのパーティションはネットワーク接続の信頼できる、永続ストレージ・ボリュームに支援されています。図 3 に示したのは 3 TB – 4.2 TB のデータベースで、スナップショットベースのバックアップの所要時間 (分単位) が、従来のデータベース・バックアップ・テクノロジーに対して劇的に短いことです。スナップショット・バックアップはほんの数分 (1-3 分) ですが、従来のバックアップには 2 時間から 3.5 時間かかります。

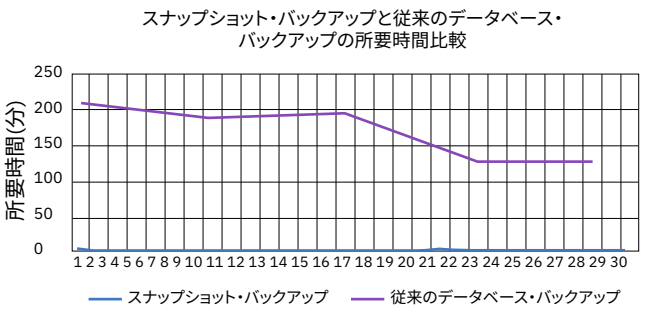

図 3: このグラフはサイズが 3 TB – 4.2 TB のデータベースの、1カ月間のスナップショット・バックアップと従来のデータベース・バックアップのバックアップ所要時間 (分単位) の比較です。このデータは本番負荷を実行している複数のシステムで測定したものです。

ROW はコピー・オン・ライト (COW) スナップショット・テクノロジーを最適化した変種で、信頼性の高い、永続ストレージで用いられます。データ・ブロックは参照によって結びつけられていて、スナップショットはリーダーが正しいブロック群を読めるようにした参照群にすぎません。

たとえば、ユーザーがスナップショットを取得した後でボリュームの 1 ブロックを変更したら、ストレージ・ボリュームは新たな場所に変更を書き込んで、ブロックの現行版として新たな場所を指すように参照を更新します。スナップショットは最初のブロックを参照し続けますが、他のブロックは変更されないままなので、スナップショットも現行ビューもそれらを参照します。

Db2 Warehouse on Cloud はシステム負荷を1分程度「停止」して、その休止状態を利用して、負荷を「再開する」前にボリュームのスナップショットをとります。最新のバックアップ 7 つが保持されます。

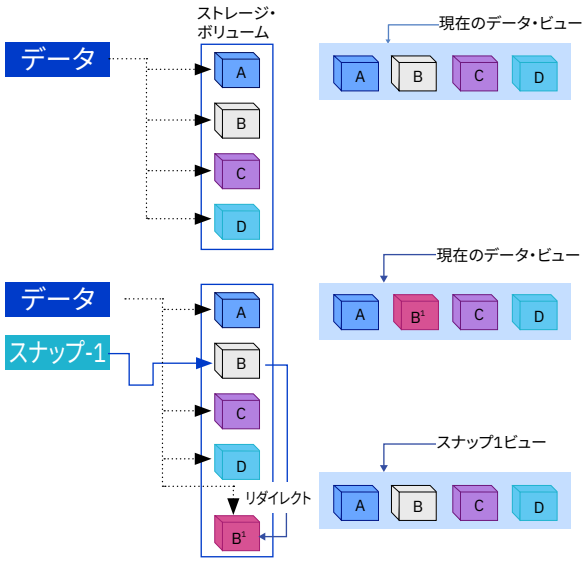

図 4: この図は ROW 動作を示します。ボリュームは参照を利用して、ブロックの最新バージョンを結びつけて、エンド・ユーザーに読ませます。スナップショット (Snap-1) が作成されてから B が B’ に変更されると、元のブロックはそのままで、B’ が新しい場所に書き込まれます。

Snap-1 は元の B ブロックを指していますが、データの現在ビューは B’ を示しますスナップショット (Snap-1) が作成されてから B が B’ に変更されると、元のブロックはそのままで、B’ が新しい場所に書き込まれます。

Snap-1 は元の B ブロックを指していますが、データの現在ビューは B’ を示します

このやり方で、ROW によって、ユーザーは「時間旅行リストア」に対応できます。つまり、ユーザーはスナップショット 1 をリストアできますが、その一方でスナップショット 2 から 7 を保存しておけるので、ユーザーは必要に応じてそのスナップショットのいずれかを個別にリストアできます。

詳細情報

Db2 Warehouse on Cloud のインターフェイスに直接触れたいのであれば、ガイド付き製品ツアー もご利用いただけます。

日本アイ・ビー・エム株式会社
〒 103-8510
東京都中央区日本橋箱崎町 19-21

Produced in Japan
December 2018

1 “IBM Cloud Kubernetes Service.”
https://www.ibm.com/cloud/container-service

2 “Production-Grade Container Orchestration.”
https://kubernetes.io/

3 “Block Storage.”
https://www.ibm.com/cloud/block-storage

Please Recycle

27022627-JPJA-00