簡単に学べる

Snowflake特集号第2版

クラウドデータウェアハウジング

for dummies®

クラウドデータウェアハウスについて

データウェアハウスソリューションの比較

クラウドデータウェアハウスの選び方

提供:

snowflake®

Joe Kraynak
David Baum

Snowflakeについて

Snowflakeは、モダンなデータウェアハウジングを、すべてのデータユーザーにとって効率が良く、コストが手軽でアクセスが容易なものにするという明確なビジョンを持ってスタートしました。Snowflakeがあれば、複数のクラウドを利用した瞬間的な弾力性、セキュアなデータ共有、秒単位の課金体系を利用するデータ主導型企業が可能になります。従来のオンプレミスとクラウドのソリューションはこの点に苦心してきましたが、Snowflakeは、データウェアハウジングの威力、ビッグデータプラットフォームの柔軟性、クラウドの弾力性を、従来のソリューションに比べるとわずかなコストで統合可能にする、新しいクラウド向けに構築されたアーキテクチャを備えた新製品を開発しました。Snowflake: あなたのデータに制限はありません。

詳細については、**Snowflake (snowflake.com)** をアクセスしてください。

クラウドデータウェアハウジング

Snowflake特集号第2版

著者：Joe Kraynak、David Baum

for dummies®

Cloud Data Warehousing For Dummies®、Snowflake特集号第2版

出版:
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-71462-0 (pbk); 978-1-119-71468-2 (ebk)

アメリカ合衆国にて製作

10 9 8 7 6 5 4 3 2 1

謝辞

この本と製作にかかわった全ての人を誇りに思います。読者の皆様の事業や組織向けにカスタマイズした「*For Dummies*」シリーズの作成については、メール（info@dummies.biz）にてお問い合わせいただくか、www.wiley.com/go/custompubをご覧ください。製品またはサービス向けの「*For Dummies*」ブランドのライセンスについては、BrandedRights&Licenses@Wiley.comまでお問い合わせください。

本書の出版にあたりご協力いただきました皆様に心より御礼申し上げます。

デベロップメント ディレクター：
Nicole Sholly

プロジェクト エディター：
Martin V. Minner

エグゼクティブ エディター：Steve Hayes

エディトリアル マネージャー：
Rev Mengle

事業開発担当：Karen Hattan

プロダクション エディター：
Mohammed Zafar Ali

Snowflake 寄稿者チーム：
Vincent Morello、*Clarke Patterson*、*Leslie Steere*、*Kent Graziano*

目次

はじめに

エグゼクティブ、マネージャー、あるいはアナリストであるあなたは、知識が力であり、ふさわしいタイミングで分析されたデータは、十分な情報を集めた上で決断を下し、競争優位を獲得するのに必要なインサイトをもたらすことを十分理解しています。現在、組織はかつてないほど多くの重要なデータを収集しています。これには、データマートやクラウドベースのアプリケーション、自動生成されたデータなど、内外の幅広いソースが含まれます。

不幸なことに、過去30年のデータハウスのアーキテクチャは、極度に大規模で多様化したデータセットによる負荷を受けて酷使され続けています。データがデータウェアハウスに入力され、分析を開始できるまでアナリストが24時間以上待つことも珍しくありません。複雑なクエリをデータに対して実行すると、さらに待つ場合もあります。多くの場合、データの処理と分析に必要なストレージとコンピュートリソースは不十分です。これは、システムのハングやクラッシュにつながります。これを避けるために、ユーザーとワークロードは待ち行列に入る必要があり、その結果さらに遅れが出ます。さらに最近では、さまざまな形のデータレイクなど代替的なアプローチが現れました。しかし、それらのソリューションは独特な制約をもたらしました。

効率性と競争力を保つために、組織は絶え間なく生成される大量のデータが持つパワーを活かし、複雑なデータ分析を実行できなければなりません。幸いなことに、10年以上前にクラウドコンピューティングの商業化が起こり、この問題への対処に役立ち、期待を上回るコンピューターハードウェア、アーキテクチャ、ソフトウェアの進歩を促しています。

本書の概要

*Cloud Data Warehousing For Dummies*第2版へようこそ。本書では、組織がどのようにすれば大量のデータが持つ威力を手頃なコストで便利に使いこなし、効率を高めて未加工データを価値あるビジネスインテリジェンスに変えることができるのかを解説します。

データが多いほど、大きなチャンスへの扉が多く開きます。しかし同時に、ほとんどの場合、大きな問題をともないます。大きなチャンスを活かすためには、多様な形式でデータを格納および整理できるデータウェアハウスソリューションを導入して、データへの便利なアクセスを提供し、データを分析できるスピードを上げる必要があります。これは、で

きる限りコスト効率を上げて実行されければなりません。本書はその方法を解説します。

本書で使用するアイコン

本書では、以下のアイコンを使用してヒントや覚えておくべきポイントなどを強調します。

ヒント

ヒントは、タスクを実行する簡単な方法や、組織でのクラウドデータウェアハウジングの上手な使用方法を紹介します。

ポイント

このアイコンは、クラウドデータウェアハウジングの理解と応用に没頭するさいに、覚える価値がある概念を強調します。

ケーススタディ

本書のケーススタディは、組織がどのようにクラウドデータウェアハウジングを応用して費用を削減し、データアナリティクスのスピードとパフォーマンスを大幅に向上させているかについて明らかにします。

本書を超えて

本書の内容にご興味を持たれ、さらに詳しく知りたい場合は、www.snowflake.comにアクセスしてください。同社の詳細と商品を閲覧できるほか、無料の試用Snowflakeが提供されています。さまざまなプランと価格の詳細を入手でき、ウェビナーの閲覧、ニュースレターへのアクセス、今後のイベント情報の入手、ドキュメントとその他のサポートへのアクセスが可能で、同社と連絡を取ることができます。Snowflakeは皆様からのご連絡をお待ちしております。

本章の内容

- » データウェアハウジングの過去から現在までをたどる
- » クラウドデータウェアハウスのメリットを理解する
- » クラウドデータウェアハウジングは今の経済環境でどのような場面に適しているか

第1章 クラウドデータウェアハウジングでスピードアップ

どのような形であれ、クラウドコンピューティングとサービス型ソフトウェア(SaaS)には数十年の歴史があります。しかし、サービス型クラウドデータウェアハウス(DWaaS)は、従来のオンプレミスデータウェアハウジングとそれに類似するソリューションに変わる選択肢として、ごく最近に出現しました。なぜ今なのでしょうか？何が変わったのでしょうか？本章では、これらの質問に回答します。

データウェアハウスの定義から始め、このテクノロジーがクラウドに向かう様子を示すデータウェアハウジングの進化をたどります。そして、どのようにすれば組織がクラウドDWaaSからメリットを受けられるかを考察し、今日のデータ主導型経済で競うために、多くの企業がクラウドデータウェアハウジングに頼る理由について説明します。

データウェアハウスの定義

データウェアハウスは、情報とインサイトをもたらすトレンド、パターン、相関関係を明らかにするために、データの格納と分析に特化したコンピューターシステムです。従来、組織は、マーケティング、販売、生産、財務を含む内部ソース(通常、トランザクションのデータベース)から収集されたデータを格納および統合するためにデータウェアハウスを

使用してきました。データウェアハウスは、企業がこうしたトランザクションのデータベースから直接データを分析すると、通常のトランザクション活動の負荷とそのデータの分析に必要なワークロードの下で遅延が発生する(さらにクラッシュすることもある)ことが理解されたために出現しました。そのため、データウェアハウス内にすべてのデータが分析用に複製され、データベースはトランザクションに集中させたのです。

長い年月を経て、データソースは内部の事業活動と外部のトランザクション以外にも拡大しました。現在、データソースには、急激に容量、多様性、速度を増しつつある、ウェブサイト、携帯電話とアプリ、オンラインゲーム、オンラインバンキングアプリ、さらに機械からのデータが含まれます。最近では、組織はモノのインターネット(IoT)デバイスから大量のデータも取得しています。

データウェアハウジングの進化

歴史的に見ても、企業は、明確に定義され、高度に構造化された形式のデータを合理的に予測できる比率と容量で収集しました。従来のテクノロジーの処理速度は上がりましたが、オンプレミスの演算能力とストレージ不足にくわえてリソースを増加させる困難さにより、データのアクセスと使用はすべてのユーザーが受け入れられるパフォーマンスになるよう慎重に管理され、制限されました。このため、組織は分析サイクルが非常に長くなったとも言えます。

時間の変化(図1-1を参照).テクノロジーが進歩すると、組織は大容量のデータに基づいて重大な事業判断を決定できるようになります。

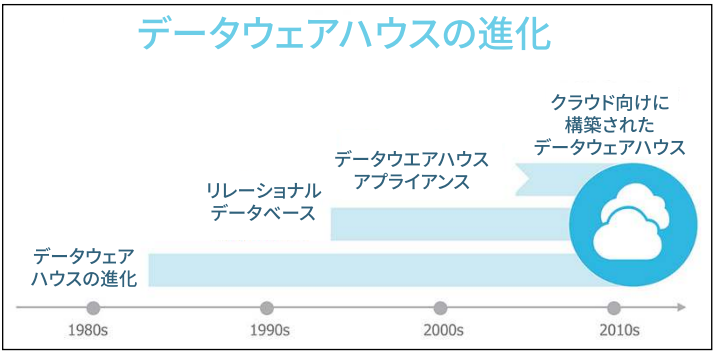

図 1-1：従来のシステムがクラウドデータウェアハウスの出現をもたらしました。

それは、マーケットリーダーや成熟企業だけではありません。小規模で機敏な市場参入者が、ほんの数か月や数年という期間で既存の業界を次々と変えて行っているのです。彼らはデータを使用して実行することで、チャンスを見出し、小売りやビジネスベンダーと顧客との関わりを変える製品やサービスを開発しています。

従来のデータウェアハウジングの限界に気付く

従来のデータウェアハウスは、今必要とされているデータの容量、多様性、速度を処理するように設計されていませんでした。新しいシステムは、これらの欠点に対処できるよう設計され、組織が現在必要とするデータのアクセスと分析に対応しようとしています。現在、明らかになっている問題は次のとおりです:

- データソースの数と種類が増えた結果、綿密でありながら多大なコストがかからない分析を可能にするためには、1か所にさらに多様なデータ構造が共存する必要があります。
- 従来のアーキテクチャはユーザーとデータ統合活動の間に競合を引き起こしており、新しいデータをデータウェアハウスにロードすると同時にユーザーに適切なパフォーマンスを提供することは困難です。
- 特定の間隔のバッチでデータをロードすることは現在も一般的ですが、多くの組織は、継続的なデータロード(*マイクロバッチ*)とストリーミングデータ(*インスタントロード*)を必要としています。
- 従来のデータウェアハウスを、今日の増大するストレージとワークロードの需要を満たすために拡張すると、可能な場合でも高価になり、手間がかかり、処理が遅くなります。
- さらに最近では、代替的なデータプラットフォームが複雑であり、特別なスキルや、多くのチューニングと構成を必要とすることが珍しくありません。これらは、データソース、ユーザー、クエリの数と種類が増えることで悪化します。

解決策となるテクノロジーと設計

うれしいことに、テクノロジーとデータウェアハウジング アーキテクチャ(モダンデータウェアハウスの設計と構築ブロック)は、次に挙げる

イノベーションにより、データ主導型経済の需要に対応するよう進化してきました。

- **クラウド：**モダンデータウェアハウスの進化を推し進める主な要因は、クラウドです。これにより、ほとんど制限のないアクセス、低コストなストレージ、改善された拡張性、クラウドベンダーへのデータウェアハウジングの管理とセキュリティのアウトソーシング、実際に使用したストレージとコンピューティングリソースに応じた従量制の支払いが可能になります。
- **超並列処理(MPP)：**1つのコンピューティング操作を分割し、多数の独立したコンピュータープロセッサーで同時に実行するMPPは、2000年代初期に出現しました。この作業分割は、ソフトウェアがこのアプローチを活用するように設計されていれば、データのストレージと分析が高速化されます。
- **列指向型ストレージ：**従来、データベースでは表計算シートで見られるような行に格納されていました。たとえば、これに顧客や小売のトランザクションに関するすべての情報を含めることができました。従来の方法でデータを抽出するには、1つの要素を取得するためにシステムでその行全体を読み取る必要があります。これには労力と時間がかかります。列指向型ストレージであれば、1つのレコードの各データ要素が1列に格納されます。このアプローチにより、ユーザーは、会費が支払い済みであるジムの会員などの1つのデータ要素を照会でき、レコードの全体(この場合、各会員のID番号、名前、年齢、住所、都道府県、市町村、支払い情報など)を読み取る必要はありません。このアプローチにより、この種の分析クエリの応答を大幅に高速化できます。
- **ベクトル処理：**この形式の*データアナリティクス*(結論を出すためにデータを検査するサイエンス)向けデータ処理は、最新の革新的な半導体デザインを活用します。このアプローチは、数十年前に、旧型のハードウェア技術をベースとして構築された古いデータウェアハウスソリューションに比べると、はるかに高速なパフォーマンスを発揮します。
- **ソリッドステートドライブ(SSD)：**ハードディスクドライブ(HDD)と異なり、SSDはフラッシュメモリ上にデータを格納し、データのストレージ、抽出、分析が高速になります。SSDを活用するソリューションは、パフォーマンスの大幅な向上を実現できます。

データウェアハウジングの進化を推し進めるテクノロジーの進歩とその他のトレンドについては、第2章を参照してください。

クラウドデータウェアハウスの紹介

クラウドデータウェアハウジングは、必要なハードウェア、ソフトウェア、インフラストラクチャの購入、設置、インストール、構成に多大な初期費用をかけずに最新のテクノロジーとアーキテクチャを活用できる、企業にとって費用対効果が優れた方法です。クラウドデータウェアハウジングのさまざまな構成・選択肢は、一般に、以下の3つのカテゴリに分かれます。

- **クラウドインフラストラクチャに配置される従来型データウェアハウスソフトウェア：**この選択肢は、オリジナルのコードベースを再利用するため、従来のオンプレミスデータウェアハウスに似ています。依然として、データウェアハウスを構築および管理するITの専門知識が必要です。ハードウェアとソフトウェアの購入、設置、インストールの必要はありませんが、引き続き重要な構成とチューニングを実行し、定期的なバックアップなどの作業を行う必要があるでしょう。
- **マネージドサービスとしてサードパーティによりクラウドでホストおよび管理される従来型データウェアハウス：**この選択肢では、サードパーティのプロバイダーがITの専門知識を提供しますが、引き続き、従来のデータウェアハウスと同じ制約の多くを経験する可能性があります。データウェアハウスは、ベンダーが管理するデータセンターに設置されたハードウェアをホストします。これは、*アプリケーションサービスプロバイダー(ASP)*と呼ばれる業界に似ています。顧客は、使用を予想するディスク容量とコンピュートリソース(CPUとメモリ)を事前に指定する必要が依然としてあります。
- **真のSaaS型データウェアハウス：**この選択肢は、しばしば*DWaaS*と呼ばれます。ベンダーが、すべてのハードウェアとソフトウェアを備えた、データウェアハウスに必要なパフォーマンス、ガバナンス、セキュリティの確立と管理に関連するすべてのタスクのほとんどを取り除いた完全なクラウドデータウェアハウスソリューションを提供します。通常、クライアントはストレージとコンピューティングリソースの使用量と使用時間のみに応じて支払います。また、この選択肢では、パフォーマンスに影響を与えずに無制限のワークロード数を同時に運用しながら、各ワークロードに特化した無制限の演算能力を追加して、需要に応じてスケールアップ・ダウンを行うことができます。

クラウドデータウェアハウジングソリューションの詳細な比較については、第5章を参照してください。

クラウドデータウェアハウスが必要な理由

データを信頼して顧客へのサービスを向上させ、業務を合理化し、業界をリードしている組織は、クラウドデータウェアハウスのメリットを享受できるでしょう。大規模な従来のデータウェアハウスとは異なり、クラウドでは、大小のビジネスがニーズと予算に合わせてデータウェアハウスのサイズを決め、物事の日々そして年々の変化に応じてシステムを動的に拡大または縮小できるようになります。

ここでは、最先端のクラウドデータウェアハウステクノロジーが、企業の業務を大幅に向上させることができるいくつかの分野を挙げます。

- **カスタマーエクスペリエンス（顧客体験）：**エンドユーザーの行動をリアルタイムでモニターすることにより、組織は個々の顧客のニーズに合わせて、製品、サービス、特別なオファーを調整することができます。顧客感情を分析すること、つまり、大量のソーシャルメディアの投稿、ツィート、その他のオンライン活動を分析することにより、企業は顧客を深く理解できます。
- **品質保証：**また、組織はストリーミングデータを使用して、顧客サービスの問題や製品の欠点の初期警告信号をモニターすることもできます。数日や数週間ではなく、数分または数時間で対応策を取ることができます。これは、データソースがコールセンターの苦情記録しかない時代には不可能でした。
- **業務の効率性：***オペレーショナルインテリジェンス*(OI)は、組織にとって、コストの節減、利益の増大、プロセスの合理化、市場原理への迅速な対応が可能になる場面を特定するための、ビジネスのモニターとイベントの分析で構成されます。組織からデータウェアハウスを管理する手間を省くことで、データの分析に集中できます。
- **イノベーション：**市場動向を把握するためバックミラーを見るよう、過去のデータに頼る、旧来の方法を取るのではなく、企業は最新データソースとデータアナリティクス(予測的、規範的、機械学習)を使用して、トレンドに気付き、活用できるため、未知の、あるいは予想外の競合よりも先に業界のディスラプションを実行できます。


ポイント

企業データのほとんどは、多数の異なるデータベースに格納されています。質問すべき問題：どのようにすれば、そのようなデータにアクセスできるのでしょうか？データのすべてを抽出、格納、分析するコストはいくらかかるでしょうか？何もしなければ、どうなるでしょうか？ここが、クラウドデータウェアハウジングの出番となります。

本章の内容

- » データアクセスと分析の増大する需要に対応する
- » 現在のデータの作成方法および使用方法に対応する
- » 新しく改良されたテクノロジーで問題に取り組む

第2章 モダンデータウェアハウスが生まれた理由

クラウドデータウェアハウジングは、データソース、容量、種類の変化、データアクセスとアナリティクスの重要性増大、そしてデータのストレージ、アクセス、アナリティクスの効率を大幅に上げるテクノロジーの改良、という3つの重要なトレンドがあいまって出現しました。本章では、これらのトレンドをさらに詳しく説明し、どのようにすればデータウェアハウスでクラウドのメリットを活用してこれらのトレンドに対応できるのかを明らかにします。

データの容量、種類、速度のトレンドの考察

本書でデータについて述べる場合の単位はペタバイトです。1ペタバイトは100万ギガバイトです。これは、約5,000億ページの標準的な印刷されたテキスト、または58,333本の約2時間の高解像度動画に相当します。データは、ビジネスの日常業務から、モバイルデバイスでウェブサイトとソフトウェアアプリケーションを使用する人々、そしてデジタルデバイスや機械装置の毎日の活動から大量に発生します。

このセクションでは、クラウドデータウェアハウジングの需要の原因となったデータの変化とデータの使用に注目します。

爆増するデータの管理

それほど遠くない昔、ビジネスでは、一般に人間がシステムに手入力したデータを管理していました。顧客、クライアント、パートナーなどの外部ソースからのデータがある場合もありました。データ量は比較的小さく、予測可能でした。データは企業のデータセンターで格納および管理され、セキュリティが保証されました。現在、*オンプレミスメソッド*として知られる方法です。

今、ビジネスの世界は、本書で既に述べたさまざまなソースや、一覧に挙げるには数と種類が多すぎるその他のソースで利用されるデータの爆増を経験しています。このデータの容量と多様性により、従来のオンプレミスデータウェアハウスは簡単に対応できなくなる可能性があります。ユーザーとユーザーが任意の時間に処理するワークロードによる過負荷のため、データの処理と分析がしばしばシステム停止の原因となり、クラッシュさえ引き起こします。

データの激増に対応するためには、新しい視点が必要になります(図2-1を参照)。論点を、「組織のデータウェアハウスがどの程度の規模にする必要があるか」という観点から、「大容量のデータ処理に必要な規模にあつれきを起こさずに低コストでデータウェアハウスを拡張できるかどうか」に移す必要があります。

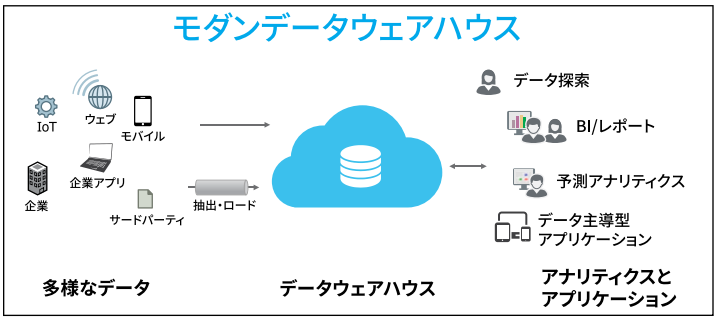

図 2-1：モダンデータウェアハウスは、すべてのユーザーに対応するすべてのデータを可能にします。


ポイント

クラウドデータウェアハウジングがきっかけとなったユースケースは、次々と発生しています。たとえば、クラウドを使用して自社のデータを格納するSaaS生まれの企業と大企業が、そのデータをマネタイズ(販売)しています。これをサービスとしてパッケージ化し、可能な限り優れたインサイトでさらに優れたビジネスの決断を下したい他の組織に販売します。

クラウドで発生したデータの活用

ほんの数例を挙げると、顧客関係管理(CRM)ソフトウェア、基幹業務ソフト(ERP)のソフトウェアスィート、広告購入プラットフォーム、オンラインマーケティングツールを含むSaaSの急速な採用を組織はこれまで行ってきました。クラウドのおかげで、新しいSaaS企業は、1〜2台のノートパソコンの価格で事業を開始することができます。これらのSaaS製品は、クラウドに格納される大量の貴重なデータを生み出します。さらに、SaaSベンダーが自社のオンプレミスデータセンターよりも優れた安全性を提供できることに気付いています。

SaaS/クラウドアプリケーションの需要も高まっています。その展開の容易さは、オンプレミスアプリケーションの起動と実行に必要な事項とは比較にもなりません。過去には、企業がデータを生成するために運用する重要なエンタプライズアプリケーションの数はわずか5〜10であったこともありました。現在、１つのシステムにマーケティングデータ、別のシステムに財務、さらに別のベンダーに製品情報など、それぞれ独自のデータサイロを作ってしまう可能性があり、データが包括的で最適な分析のために統合されていない数百、数千のアプリケーションを中規模の組織が所有していても珍しくありません。

現在、クラウドに組織のデータの大部分があるので、それらのデータを統合する場所もクラウド内になるのが自然です。クラウドデータウェアハウジングを使用すれば、高価で時間がかかり、クラウドネイティブデータの量が増大するにつれて意味を失うデータセンター内での作業をしなくてすむようになります。

自動生成データの使用

機械によって自動生成されたデータは、スマートフォン、センサー情報、空調管理、石油掘削装置、ホームセキュリティシステム、スマートメーターなどを含む、インターネットを通じてデータを通信するデバイスを無限に収集する*モノのインターネット(IoT)*に関連する重要なトピックです。IoTデバイスから収集および分析されたデータにより、製品とプロセスの向上、装置のモニター、不具合の回避に必要な保守の予測が可能になります。

しかし、自動生成されたデータの多くは、信号対雑音比が良くありません。貴重なデータが含まれますが、「雑音」も多いのです。そのため、価値のあるビットを見つけるためにデータのすべてを格納しなければならないことが多いのです。さらに、自社のデータセンター以外で発生するデータの割合が増えています。これにより、クラウドがその無制限に近い拡張性とあいまって、データを格納および統合するのが当然の場所になります。

データ探索の検証

データの分析はデータ探索から始まります。興味深く価値があるつながりを特定し、それらをレポートとアナリティクスの形でデータユーザーに表示します。データ探索は新しいコンセプトではありませんが、データ量が増大すると資源集約的な作業になります。

データ探索には、しばしば大規模なデータセットが関わります。また、これは度々検証の性質を帯び、従来のオンプレミスデータウェアハウスを配置するための相当額の初期費用を正当化するのに必要なROI評価が複雑になります。これに対して、クラウドは必要に応じてデータウェアハウスをスケールアップおよびスケールダウンすることができ、組織が高額な先行投資を確約しなければならない問題を回避できる、使用量に応じた料金を提供します。

データレイクの紹介

さまざまな形式で大量の未加工データのすべてを1か所で所有する必要が高まり、現在、レガシーデータレイクと考えられているものが発生しています。組織は、データの変換と変換されたデータから価値あるインサイトを引き出すすることがほぼ不可能だったため、これらのソリューションに法外なコストがかかることにすぐに気付きました。

しかし、データレイクへの当初の関心は、企業がすべてのデータを1か所に合理的なコストで格納することを望んでいる点を明らかにしました。モダンクラウドデータウェアハウスを既存のデータレイクに追加するか、データウェアハウス内にデータレイクを構築すれば、ほぼ無制限に近いストレージとコンピュートリソースを使用して、無制限の構造化データと半構造化データを優れた費用対効果でローディング、変換、そして分析するというデータレイクの本来の構想を簡単に実現できます。

レポートとアナリティクスでトレンドを探る

データ主導型の意思決定を経営陣やデータサイエンティストだけに任せることはなくなりました。今では、企業のほとんどすべての業務状況を改善するために使用されています。しかし、組織内でデータのアクセスとアナリティクスの需要が増加し、ワークロードが従来型のデータウェアハウスからストレージとコンピュートリソースを奪うために、システムのスピードが低下したり、クラッシュする可能性があります。効率性が下がり、企業はシステムを維持するためにインフラストラクチャの追加に時間と資金を投資する必要に迫られます。

このセクションでは、人々がデータにアクセスしたり使用したりする方法に変化が起きているトレンドと、それらのトレンドがどのようにモダンなクラウド用に構築されたデータウェアハウスソリューションのニーズを促進しているのかを解説します。

アナリティクスを可能にする弾力性の使用

ここでは、クラウドに構築された弾力性のあるデータウェアハウジングが、次に挙げるデータを使用して実行可能性を広げるシナリオをいくつか紹介します。

- データ探索には多くのメリットがあります。しかし、この種の分析に最適とされるようなオンデマンドかつ弾力性に富む拡張性を備え、大規模なデータセットの分析に必要となるコンピュートリソースを理解するのはほぼ不可能です。
- 常時発生するアドホックデータ分析で、単一で具体的なビジネスの質問に回答します。動的な弾力性と各ワークロードの専用リソースにより、他のワークロードのスピードを下げずにクエリが可能になります。
- イベント駆動型アナリティクスは、常にデータが必要です。アナリティクスは、ビジネスをリアルタイムまたはほぼリアルタイムでモニターできるように、新しいデータを取り込み、レポートとダッシュボードを継続的に更新します。ストリーミングデータを取り込み、処理するには、データフローのバリエーションと急激な増加に対応する弾力的なデータウェアハウスが必要です。

綿密な事前計画をラピッドイテレーションに置き換える

企業は新しいアイディアの市場性を確かめる際に、通常、綿密な事前計画またはラピッドイテレーションの2つの経路を経ます。最初の選択肢は、チャンスや新製品のアイディアをじっくり考え、アイディアを検討し、そのアイディアが消費者の需要を作り出すよう期待するといった、時間のかかるプロセスです。ラピッドイテレーションには、市場でアイディアを迅速にテストし、商品の有望なバージョンが成果を見せるまで何度も繰り返す作業が含まれます。そこから、プロセスが再び始まります。

ラピッドイテレーションは、定評のある競合を排除し、業界全体のビジネスの方法を変えるために最も効果的なプロセスとして出現しました。しかし、これを成功させるためには大量の正確なデータを高速で収集および分析する必要があります。クラウドデータウェアハウジングとアナ

リティクスの進歩により、データの正確性を維持しながらのラピッドイテレーションが実用化されました。

データアナリティクスの増大する需要を満たす

ケーススタディ

Janaは15か国以上の新興国市場で3,000万人以上のスマートフォンユーザーに、無料で無制限のインターネットアクセスを提供しています。同社のmCent Androidアプリによって、Janaは顧客のモバイルインターネットのコストを、スポンサー付きのコンテンツを利用する4,000ブランドに負担させています。

新しいブランドのコンテンツやmCentの機能を取り入れる際に、Janaはユーザーの注目、ライフタイムのユーザー価値、重要業績評価指標(KPI)を含む主要指標を分析および測定します。

Janaと同社のデータが成長するにつれて、同社の当初のアナリティクスアーキテクチャは同社のビジネスにとって効率的に機能できなくなりました。クエリは遅くなり、テーブルのスキャンは実行できなくなりました。容量とバックアップシステムの追加と、Janaのオープンソースデータリポジトリの管理のための時間が、ますますかかるようになりました。

図でわかるように、Janaは、同社のデータプラットフォームコンポーネントのほとんどをアップグレードし、これらの障害を克服し、以下に挙げるメリットを享受するために、クラウドに構築されたデータウェアハウスを使用してシステムを合理化しました。

- 多様なデータの急増するストリーミングを処理および分析するビジネスの需要に遅れを取らずに対応できます。
- 企業のあらゆる場所でアナリティクスの利用をさらに促すことができます。Janaの従業員８割が、データウェアハウスにアクセスしています。
- 管理の諸経費を大幅に節減できます。

Janaの、より高速で、よりコストが安く、効率が向上したデータウェアハウスへの変換。

アナリティクスの組み込み

多くの企業にとって、アナリティクスの運用は切り離された独特のビジネスプロセスです。しかし、クラウド内に構築されることが多くなったビジネスアプリケーションにアナリティクスを組み込むトレンドが強くなっています。これらのアプリケーションは、アプリケーションを照会するユーザー数と、ユーザーがデータを分析するために実行するクエリ数(ワークロード)の著しい変動に対応します。クラウドは、クラウドベースのアプリケーションから組織のクラウドデータウェアハウスへのデータ転送を容易にします。クラウドでは、その拡張性と弾力性によりユーザーとワークロードの変動を適切にサポートできます。

モダンデータウェアハウスに必須なテクノロジー

テクノロジーのイノベーションにより、可用性、簡素性、コスト、パフォーマンスの点でデータウェアハウジングとアナリティクスの改善が可能になります。このセクションでは、モダンデータウェアハウスに欠かせない主要なテクノロジーに注目します。

クラウド

クラウドの性質は、データウェアハウジングに特に適しています。他の文脈でも述べましたが、クラウドによって実現できるものを知ることは大切です。

- **無制限のリソース：**クラウドインフラストラクチャは、ほとんど制限のないリソース、オンデマンド、数分または数秒以内というスピードを実現します。組織は、使用状況に応じて秒単位で支払います。パフォーマンスを損なうことなく、ユーザーとワークロードの規模が何であれ、動的にサポートすることが可能になります。
- **資金を節約、データに集中：**クラウドで構築されるソリューションを選択する企業は、ハードウェア、ソフトウェア、その他のインフラストラクチャの高価な初期投資と、オンプレミスシステムの保守、更新、安全性のコストを回避します。企業は、その代わりにデータの分析に集中します。
- **当然の統合ポイント：**いくつかの推定によると、分析したいデータの80%は、企業のデータセンター以外のアプリケーションからのものです。データをクラウドにまとめると、何百万ドルにもなるハードウェアとソフトウェアを初期に購入して、リ

ソースを保守する技術スタッフに給与を支払う必要がないため、社内のデータセンターを構築するよりも容易かつ低コストになります。

列指向型ストレージ、処理

前述のとおり、列指向型ストレージは、データのストレージ、抽出、分析の効率性とパフォーマンスを大幅に向上させ、システムユーザーは迅速に結果にアクセスできるようになります。

ソリッドステートドライブ(SSD)

ハードディスクドライブ(HDD)と異なり、SSDはフラッシュメモリ上にデータを格納するため、データのストレージ、抽出、分析が高速になります。これの改良は、SSDを効率的に使用するよう設計されたデータウェアハウスの演算能力を高めます。

NoSQL

*NoSQL(「構造化問い合わせ言語(SQL)だけではない」の省略語)*は、IOTやソーシャルメディアによって生成されたデータなどの新しい形式のデータの格納と分析を可能にし、組織のデータアナリティクスを強化および拡張するテクノロジーを指します。従来のデータウェアハウスは、これらのデータの種類の対応には適しません。そのため、JSON、Avro、XMLなどの新しいアプローチが、これらの「半構造化」データ形式を処理するために近年出現しました。

これらのNoSQLシステムのいくつかは、従来のデータウェアハウスを置き換える目的で設計されましたが、最終的には補完のみにとどまりました。半構造化データから価値を得るために、多くの場合、組織はNoSQLシステムからデータを抽出および変換し、ビジネスユーザーが簡単にアクセスできるようにデータを従来のデータウェアハウスにロードする必要があります。その結果、両方の種類のシステムのメリットを活用したい企業(前述のケーススタディのJanaなど)にとって、この技術は複雑性とコストをさらに増大させます。

そのため、モダンなクラウドで構築されたデータウェアハウスは、組織が2つのシステムに対して金銭を支払い、管理しなくても済むように、構造化(従来型)データ形式と半構造化データ形式の取り込みとクエリを組み込み、それらに最適化されている必要があります。

本章の内容

- 適切なデータウェアハウスソリューションを選ぶ
- 高いコストパフォーマンスを追求する
- データのセキュリティ、保護、ガバナンスを優先する

第3章 モダンデータウェアハウスの選択基準

第2章で解説した傾向は新しい種類のデータウェアハウスの必要性と機会をもたらしました。つまり、今のデータ容量、多様性、速度や、組織によるデータの新たな使用方法に対応した技術です。こうしたソリューションでは、クラウドを含む重要なテクノロジーイノベーションを活用する必要があります。

データウェアハウスの市場においては、基準のチェックリストは、どの選択肢が自分のニーズに最適かを判断するのに役立ちます。組織に最適なデータウェアハウスソリューションを見つけるために、本章のチェックリストを検討してください。

現在、そして将来のニーズを満たす

システムの真の弾力性はビジネスに利点をもたらしますが、それよりも大切なことがあります。コンピュートリソースとストレージの両方に独立した拡張の可能性を持たせるべきです。そうすれば、コンピュートの性能のみを増やす必要がある場合、ストレージを増やさなくても済みます。その逆の場合も同様です。弾力性があるデータウェアハウスには重要な能力があります。

1か所ですべてのデータを格納および統合する

従来とは異なる半構造化データには、前章で検討されたように、従来のデータの制約を超えたデータアナリティクスのインサイトを強化できる可能性があります。しかし、これには、組織がデータを分析できるようにする前に新しいデータの種類をロードおよび変換するための新しいアプローチが必要です。ほとんどの従来のデータウェアハウスは、このようなデータタイプを処理するために、パフォーマンスか柔軟性を犠牲にしています。モダンデータウェアハウスは、読み込む前に半構造化データを変換することが必要となる、厳密な従来の構造を事前に設計およびモデル化する必要性を取り除く必要があります。また、そのようなデータタイプに対して、元々の形式のままでクエリのパフォーマンスを最適化する必要もあります。全体的に見て、データウェアハウスは柔軟に多様なデータをサポートし、パフォーマンスの問題を回避する必要があります。

すべてのデータを1か所に効率よくロードすることは重要です。しかし、より精密なアナリティクスのために、そのような多様なデータの種類のすべてを統合することは別問題です。モダンデータウェアハウスは、一度NoSQLシステムに収容された半構造化データと、従来の企業のリレーショナルデータベースに固有な構造化データを自動的に統合する必要があります。インストールや設定の必要性がまったくないことに加えて、チューニングとパフォーマンスは組み込まれているべきです。最も重要なのは、すべてのデータを管理するために、2つの別々のシステムを保守し、それらに料金を支払う必要があってはならないことです。

既存のスキル、ツール、専門知識をサポートする

従来のデータウェアハウスは、そのテクノロジーに40年の歴史があり、クラウドに対応するための再設計が容易でないという理由だけでも時代遅れです。これは、従来のデータウェアハウスが依存するSQLという言語が業界の主流であり続けていることも意味しています。このため、SQLデータウェアハウスと通信する、新旧のデータ管理、データ転換、統合、視覚化、ビジネスインテリジェンス、アナリスティクスツールが幅広く存在します。標準SQLの役割は確立しており、SQLスキルを持つ人々が大勢います。

ケーススタディ

異なるデータの分析

Chimeは、モバイル世代に対応したスマートバンクです。Chimeは、ビジネスに価値をもたらしながら会員の体験を充実させるために、モバイル、ウェブ、バックエンドサーバープラットフォームのあらゆる場所でデータを収集および分析しています。

Chimeでの重要業績指標の分析は多くの時間と労力を必要とし、FacebookとGoolgeの広告サービスを含む多数のサービスからのデータの収集と分析も行います。また、Chimeは他のサードパーティのアナリティクスツールからもイベントを抽出しました。ツールが提供する内容は殆どJSONなど、半構造化データでした。

Chimeは、新しいクラウドデータウェアハウスにより次に挙げる必要条件を満たしました。

- 構造化データと半構造化データを効率よく提供し、標準的なSQLデータベーステーブルを使用しながら、ほぼリアルタイムでクエリが使用できる。
- 同社のデータウェアハウスにロードされる新しいデータタイプすべてに対応する新しいモデルを設計する必要をなくし、データパイプラインを簡素化する。
- ワークロードの需要と管理コストに応じてスケールアップ・ダウンができる。
- サードパーティのデータアナリティクスツールと迅速かつスムーズに統合できる。
- データの抽出および分析に複雑なプログラミング言語を必要とする他の選択肢の代わりに、SQLを使用できる。

Chimeのアナリストは、現在、会員サービスを充実させ、クエリ結果の待ち時間を短縮し、データ分析に費やす時間を増やすために、さらに多くのシナリオを作っています。

従来のデータウェアハウスはSQLをサポートしますが、半構造化データを効率よく格納および処理するために必要な能力はサポートしません。そのため、多くの組織はNoSQLソリューションなどの代替的なアプローチに目を向けました。これらのシステムには制約があり、別の問題をもたらします。一般には利用されず、SQLに対応できないかもしれない特別な知識とスキルが必要になるのです。モダンデータウェアハウスは、先進のテクノロジーで構築されながらも、包括的で確立された標準(SQLなど)で作られ、Spark、Python、Rのコンピューティング言語などの業界で一般的に利用されている他のスキルとツールとの互換性を持つ必要があります。

組織の費用を節減する

従来のデータウェアハウスは、ライセンス料金、ハードウェア、サービス、データウェアハウスのセットアップ、管理、配置、チューニングに必要な時間と専門知識、データのセキュリティを保ち、バックアップを作成するコストに何百万ドルもかかる場合があります。さらに、ビジネス要件を満たし、今のデータの容量と多様性を最大限に活用するデータウェアハウスの構築は、どんな組織にとっても法外な費用がかかることが少なくありません。

モダンデータウェアハウスは、これらの問題に対してはるかに低い価格で対応するはずです。たとえば、ユーザーがリソースの必要分のみに支払うように、ストレージとコンピュートを別々に拡張するか？ワークロードと並行処理も同様に拡張するか？多様なデータ構造をサポートし、1か所で多様なデータを統合するか？ダウンタイムは最小限またはゼロになり、アップグレードを自動的、または段階的な進め方で実現する選択肢が提供されるか？そして最終的に、複雑でコストが高く、悩みの種となる手作業でのシステムの調整とチューニングが不要になり、これらのすべてが自動的に実行され、最大のパフォーマンスを得ることができるか？(クラウドデータウェアハウスの比較については、第5章を参照してください。)


ポイント

クラウドデータウェアハウジングにより、従来のオンレプミスソリューションのコストに比べわずかのサービス料金ですべてをカバーします。クラウドベースのソリューションは多様で、すべてが同じではありません。こうした違いは、価値あるデータインサイトを得るために顧客が支払う必要がある金額も変わってきます。

データの復元力と回復を実現する

さまざまな種類のデータウェアハウスの故障が、データの損失や不整合の原因となる可能性があります。そのため、データウェアハウスはデータを安全、最新、そして使用可能な状態に保つ必要があります。従来のデータウェアハウスは、通常、定期的にバックアップを実行してデータを保護します。これは、貴重なコンピュートリソースを消費し、進行中のワークロードを妨げます。また、定期的なバックアップは追加ストレージも必要であり、最新のデータを保存できず、データが不整合になることが少なくありません。

モダンデータウェアハウスは、システムの耐久性、復元力、可用性を確保する際にそれ自体を管理する必要があります。バックアップ処理がバックグラウンドで実行されているために進行中のワークロードが妨げられたり、パフォーマンスが悪化したり、サービスが実行できな

い状態になったりしてはなりません。データをコピーしてどこか他に移動させなくても保護できる賢明な方法を使用すべき、そしてそれは安価であるべきです。最終的にマルチクラウドアーキテクチャを持つことで、ビジネスの拡大に応じてデータとワークロードを物理的な場所と、Amazon、Microsoft、Googleなどの主要なクラウドベンダーの両方の間で移動させる携帯性が手に入ります。

保存中のデータと転送中のデータのセキュリティを保つ

データセキュリティは、次に挙げる2つの主な分野をカバーします。

- **機密性：**データへの不正アクセスを防ぐ
- **完全性：**データが修正されず、破損もなく、適切に管理され、品質が維持されていることを保証する

また、モダンデータウェアハウスは、マルチレベルの*ロールベースアクセス制御(RBAC)*もサポートする必要があります。これにより、ユーザーは閲覧が許可されているデータのみに確実にアクセスできます。さらにセキュリティを高めるには、*多要素認証(MFA)*が必要です。MFAにより、ユーザーがログオンすると、システムが第2の認証要求を多くの場合携帯電話に送信します。そして、電話に送信されたパスコードの入力が求められます。これにより、権限のない人物が盗まれたユーザー名とパスワードを使用しても、システムにアクセスできなくなります。

データガバナンスは、企業のデータが適切にアクセスおよび使用され、すべてのデータ漏洩を防ぎ、あらゆる法令を順守するように管理および保護されることを保証します。また、企業がその従業員と共有するデータの品質を維持するために厳格な監視も必要です。不良データには、ビジネスの決断が手遅れになったり不適切になったり、売上が失われたり、コストが上昇したりする結果につながる恐れがあります。データの品質監視を担当するデータ管理者は、データが破損した、または不正確になった場合、更新の頻度が不十分で無意味になった場合、あるいは状況に合わないデータが分析される場合を特定できます。

データの暗号化は平文を暗号文に置き換える暗号化アルゴリズムの適用を意味し、もう1つの必要なセキュリティ機能です。ソリューションの大部分は「キーの管理」です。暗号化されたデータを復号するには暗号化キーが使用されます。データの保護に加えて、データを解読するキーも保護する必要があります。どのぐらいの期間、同じキーを使用するか？キーが不正にアクセスされた場合、何が起きるか？これらのすべてに管理が必要です。データウェアハウスは、暗号化キーを暗号化する階層的

なキーラッピングアプローチと、1つのキーの使用回数を制限する強固なキーローテーションプロセスを採用する必要があります。

さらに、モダンクラウドデータウェアハウスのソリューションプロバイダーは、脆弱性を先回りしてチェックするために*侵入テスト*として知られているセキュリティテストを定期的に実行しなければなりません。ベンダーは、パフォーマンスに影響を与えずにこれらの措置を着実かつ自動的に講じなければなりません。

クラウドデータウェアハウスのセキュリティとガバナンスの本格的な議論については、第8章を参照してください

業界標準のエンドツーエンドセキュリティを使用するデータウェアハウスを選びます。SOC 1/SOC 2 Type IIやISO/IEC 27001などのセキュリティ監査に合格したソリューションを見つけましょう。

データパイプラインを合理化する

*データパイプライン*は、クエリをサポートする形式でデータウェアハウスにデータをインポートする*抽出・加工・ロード(ETL)*プロセスを主に指します。データパイプラインが遅いと、ユーザーが分析などでデータにアクセスするのに非常に長い待機時間が必要になります。複数のソースからの非リレーショナルデータのストリーミングが、多様性、数、サイズの点で急増すると、問題が悪化します。

モダンデータウェアハウスは、プロセスの全体的な複雑性を抑えて、データパイプラインを通じたデータの移動を高速化します。モダンなソリューションでは、データを変換するためにNoSQLなどの複雑なシステムを追加しなくても半構造化データを元々の形式で効率良くロードし、クエリをただちに使用可能にします。これにより、ユーザーはSQLデータベースのクエリと同じ方法でデータにただちにアクセスできます。このようなソリューションは、新しいデータへのアクセスを爆速化し、取り込みと変換のプロセスを1日から1時間未満にまで短縮することができます。

価値を生み出す時間を最適化する

ソリューションの展開は大仕事になってはいけません。昔は手作業だった重要な手続きは自動化されるべきです。何より、すべてのユーザーがいつでも使用でき、従来のシステムにかかる費用のほんの一部ですべてのデータの種類を網羅できるソリューションを選択すべきです。そのようなシステムが迅速に提供するデータインサイトは組織の合理化を進め、顧客の役に立ち、業界をリードする能力を向上させるのに役立つはずです。

本章の内容

- 価値を生み出すまでの時間のギャップを縮める
- ストレージとコンピューティングのコストを大幅に削減する
- 動的な弾力性を活用する
- 管理とセキュリティをアウトソーシングする

第4章 オンプレミス対クラウドデータウェアハウジング

新しいデータウェアハウスの市場で最初に考慮すべき選択肢は、データウェアハウスをどこに置くか、つまり組織のデータセンターに置くか、クラウドでサービス型ソフトウェア(SaaS)として利用されるかです。従来のオンプレミスデータウェアハウジングは、クラウドが実行可能なプラットフォームになるずっと以前に設計された、成熟し確立されたテクノロジーです。クラウドが急速に採用され、クラウドが提供するメリットを最大限に活用できるデータウェアハウスソリューションのニーズがあります。本章では、クラウドデータウェアハウジングを従来のオンプレミスシステムと比較しながら、重要な検討事項を提示します。

価値を生み出す時間の評価

従来のデータウェアハウスの配置(第3章を参照)には1年以上かかり、データからインサイトを抽出するまでにプロジェクトが複数年に及ぶ場合もあります。今日、ビジネスの機敏性は、プロジェクトをサポートする主要な関係者にとって重要です。プロジェクトの成功に責任を負い、ビジネスとテクノロジーを実現する人たちは、プロジェクトが本稼働する前にチームや企業を去る可能性があります。また、そのような長いサイクルでは、プロジェクトが経済の悪化、業績、仕様の予定外の変更によりプロジェクトを実行に移せないリスクの影響も受けます。

さらに、オンプレミスソリューションは現代の半構造化データの処理に適しません。これにはオープンソースのNoSQLソリューションを追加する必要があり、複雑性がさらに増し、新しいデータウェアハウスの導入期間が長くなってしまいます。

適切に実行すれば、クラウドウェアハウスは数週間またはほんの数か月で軌道に乗せることが可能です。そのため、立ち上げと稼働に必要な時間のほとんどは、他のデータソースからのデータ抽出と、データウェアハウスからインサイトを抽出するフロントエンドアナリティクスツールの構成に費やされるはずです。

ストレージとコンピューティングのコスト構成

オンプレミスデータウェアハウスは、ハードウェア、ソフトウェア、管理の点でコストが高くつきます。更に、サーバー、追加されるストレージデバイス、システムを収納するデータセンターの間取り、データにアクセスするための高速なネットワーク、システムの稼働を維持するのに必要な電源と冗長電源などが含まれます。ウェアハウスが極めて重要な場合は、障害復旧サイトを設定するコストを追加します。また、組織がデータウェアハウスソフトウェアとアドオンパッケージのソフトウェアライセンス料金に数十万ドルを支払うことは珍しくありません。データウェアハウスへのアクセス権が与えられる顧客とサプライヤーを含めてエンドユーザーを追加すると、それらのコストは大幅に増加する場合があります。そして、多くの場合、もとのラインセスコストの２割にもなる年間サポート契約を継続的に支払います。さらに、オンプレミスデータウェアハウスは、システムを配置および保守するために特別なITスタッフを必要とします。これは、問題が発生した際にボトルネックになる可能性があり、ベンダーではなく顧客とのシステムの責任が残ります。

クラウドデータウェアハウスでは、オンプレミスシステムの先行投資と継続的な費用が、シンプルな運営コストベースの価格設定に取って代わります。実際に使用したストレージとコンピューティングリソースに基づいた料金を毎月支払います。控えめに言っても、クラウドデータウェアハウスソリューションの年間コストを、類似のオンプレミスシステムのコストの10分の1にすることが可能です。

サイジング、バランシング、チューニング

最善のパフォーマンスのために、オンプレミスデータウェアハウスはモデル作成、サイジング、バランシング、チューニングが必要になり、それらにはかなりの額の初期投資と継続的なモニタリングと管理コストが必要です。多くの場合、こうした構成に含まれる内容は次のとおりです。

- 中央演算装置(CPU)の数とスピード
- メモリ容量
- 必要なストレージ容量に対応するディスクの数とサイズ
- 入出力(I/O)*帯域幅*(所定時間に転送可能なデータ量の指標)
- データの種類を含めデータウェアハウス構造を定義するカスタムデータモデルと、更新頻度

オンプレミスデータウェアハウスの場合、組織は多くの場合、システム構成を1年のわずかな期間に過ぎないかもしれないピーク時の使用に合わせて決定します。たとえば、データウェアハウスの全機能が必要なのは、事業年度の各四半期末または各年度末のみという場合があるのです。しかし、その場合でも、システムのスケールアップやダウンを容易にできないため、1日24時間、毎日、ピーク能力に合わせて費用を支払わなければなりません。

弾力的なクラウドデータウェアハウジングは、次に挙げる2つの重要な優位性をもたらします。

- 能力計画と管理(システムのサイジング、バランシング、チューニング)の複雑性とコストは、システムに組み込まれ、自動化され、サブスクリプションのコストに含まれます。
- ピーク時と安定した使用時期の間で変動するワークロードの需要を満たすための動的なプロビジョニングストレージとオンザフライのコンピュートリソースも同様です。容量は、必要な時に必要なものを持つことです。しかし、すべてのワークロードが同じように作られているわけではありません。弾力的なクラウドデータウェアハウスを使用すれば、どのリソースがどのユーザーとワークロードに割り当てられているのか、非常に細かく知ることができます。

データの準備とETLのコストの検討

オンプレミスデータウェアハウスは、データソースのすべてからデータを抽出する必要があります。その後、データをデータウェアハウスにロードする*前*に、システム内で、多くの場合、厳格なデータ構造に合わせてデータを変換する必要があります。主な課題は、高額で有限な処理能力とストレージに制約を受けることです。その結果、データ変換が他のデータ処理ジョブと競合するのを回避するため、通常の業務時間外にデータ変換を行わなければならないのです。これは様々な観点で高価です。さらに、半構造化データは、従来のデータ構造特有の整合性のある行列になりません。高容量、高速で出力されるデータでもあります。

最も優れた、クラウド構築ソリューションは、半構造化データを変換せずに直接ロードできます。このようなソリューションでは、従来のデータウェアハウスよりも最大50倍高速で新しいデータにアクセスできます。さらに、無制限のクラウドストレージのコストは低いため、データを定期的に集計して限定するのではなく、データアナリストはデータのすべてにアクセスできます。


ケーススタディ

データパイプラインの最適化

オンラインゲームスタジオのDoubleDownは、データウェアハウスにロードするデータを準備するために、データパイプラインにNoSQLシステムを追加しました。しかし、このアプローチにより、DoubleDownの毎日のイベントログ(ユーザーのクリックと、ゲーマーのアクティビティによって生成されるその他のデータ)に長い処理時間がかかるようになりました。この企業では、ある1日のデータに翌日の午後3時までアクセスすることができなくなりました。さらに悪いことには、データの計算クラスターのどれかがダウンすると、データが喪失してしまいました。

DoubleDownは、半構造化データを最初に変換せずに直接ロードし、データをただちにクエリで使用できるシステムを選びました。これによりデータパイプラインの品質とパフォーマンスが向上しました。アナリストは100倍近く速くデータを得ることができ(24時間が15分になりました)、同社の以前のパイプラインでは頻繁に起きたエラーのほとんどすべてがなくなり、定期的な集計ではなく完全なデータ粒度がアナリストに提供され、DoubleDownのデータパイプラインのコストが80パーセント削減されました。

DoubleDownのアナリストは、現在、迅速なデータ主導型決定を目指して、新製品のリリースによるデータにただちにアクセスできます。

特別なビジネスアナリティクスツールのコストの追加

第3章で述べたように、従来のオンプレミスデータウェアハウスは、現代のデータ容量、多様性、速度の処理には適していません。その結果、組織は、従来のリレーショナルデータを格納するオンプレミスのエンタープライズSQLデータウェアハウスと、非リレーショナルデータを格納し、オンプレミスかクラウドで稼動できるNoSQLビッグデータプラットフォームの2つのデータプラットフォームを運用します。

残念ながら、これらの新しいシステムは管理が大幅に複雑になり、SQLのツールと専門知識ほどには普及していない特別なツールと専門知識が必要です。結局のところ、SQLには数十年間の歴史があり、その一方でNoSQLシステムが登場したのは比較的最近です。

理想的なクラウドデータハウジングソリューションは両方の技術のベストを提供できます。つまり、リレーショナルデータと非リレーショナルデータを統合する柔軟性を提供しつつ、データクエリのためすぐに利用可能なSQLツールとスキルをサポートします。


ヒント

最新のデータウェアハウスを選択する際には、データウェアハウスの管理に必要なスキルと専門知識の費用および可用性だけでなく、データウェアハウスと組み合わせて使用される多くのアナリティクスやその他のツールの費用と可用性も検討してください。

拡張性と弾力性の考慮

従来のデータウェアハウスは、ユーザーがプロセスと限定されたリソースを奪い合うため、システムの速度低下とクラッシュを起こす傾向があります。これらのシステムはストレージとコンピュートが1つのコンピュータークラスタ(コンピューターのグループ)上で密接につながり、片方のみを増やすコストが高くなります。

最新の、クラウドに構築されるデータウェアハウスソリューションは、事実上無制限のストレージとコンピュートを提供しますが、ストレージとコンピュートの拡張性が独立しているデータウェアハウスを検討してください(図4-1を参照)。理想的なクラウドデータウェアハウスは、次に挙げる3つの方法で拡張できます。

- **ストレージ：**クラウドストレージは本質的に拡張性があり、変化するニーズに合わせてストレージ容量を簡単に調節できます。

- **コンピュート：**データのロードとクエリの処理に使用されるリソースは、いつでも、ワークロードの数と強度の変化に応じて、簡単にスケールアップまたはスケールダウンできます。
- **ユーザーとワークロード(並行処理)：**固定された演算リソースを使用するソリューションは、ユーザーとワークロードが増えるにつれて遅くなります。組織は、しばしばデータを別々のデータマートに複製し、一部のワークロードを通常の業務時間外に移し、パフォーマンスを維持するためにユーザーを待たせなければなりません。クラウドのみが、互いのパフォーマンスに影響を与えずに、1つのデータにすべてのユーザーがアクセスできる、無制限に近い数のユーザーまたはワークロードに対応する任意サイズの専用コンピュートクラスタを追加することで、データウェアハウスを「スケールアウト」できます。

ストレージとコンピュートのコストを互いに低く抑えるために、2つを簡単かつ別々に拡張できるよう切り離すクラウドソリューションを選んでください。また、このソリューションでは、パフォーマンスを損なわずにより多くのユーザーとワークロードをサポートするため、並行実行するスケールアウトも可能であるべきです。

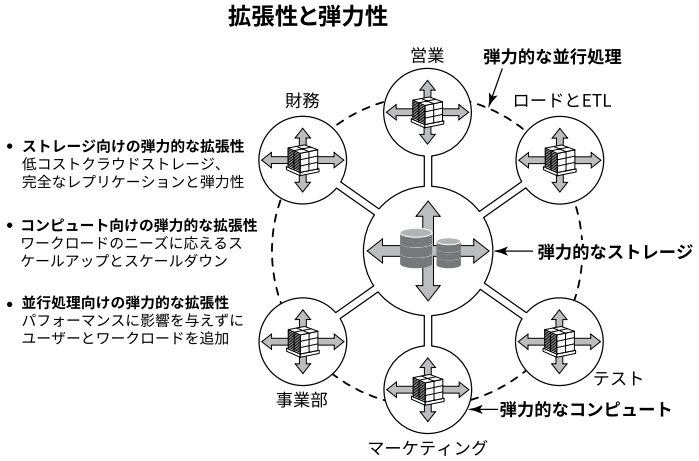

図 4-1：理想的なデータウェアハウスは、次に挙げる3つの方法で拡張できます。

遅延とダウンタイムの削減

オンプレミスソリューションを使用する多くの企業は、大きく分けて2つの不満を抱えています。前日に収集されたデータがウェアハウスで使用できるまでに、数時間または1日以上待つ必要があります。大規模なデータセットで複雑なクエリを実行するために、同様に長い時間を待つ必要があります。いくつかのケースでは、複数の並行処理がフリーズしたり、システムがクラッシュしたりする場合があり、遅延とダウンタイムがさらに延びます。

実質的に無限のストレージとコンピュートリソースがあるクラウドデータウェアハウスソリューションは、動的な弾力性を持つよう構築され、需要の増大に合わせて、スケールアップ、スケールダウン、スケールアウトを行う態勢が十分に整っています。ただし、遅延を減らし、予定外のダウンタイムをなくすには、システムリソースの単純な強化では不十分です。優れたソリューションは、手作業のチューニングをしなくても効率よくクエリを実行できるように、データパイプラインを合理化し、データを格納します。

これらのすべての種類のパフォーマンス問題に対処し、ダウンタイムを最小化するソリューションを探しましょう。どの程度データとアナリティクスに素早くアクセスできるかは、業務と競争力を保つ能力に大きな影響を及ぼす可能性があります。

セキュリティ対策コストの検討

1つの漏洩が瞬く間に広報の悪夢に変わり、ビジネスで損失を出し、規制当局から法外な罰金を命じられる結果になることがあります。クラウドはセキュリティリスクの危険性を招きますが、データセンターよりも高いセキュリティを保つことが可能です。

オンプレミスデータウェアハウスに決めた場合、自分だけが機密データの安全性の責任を負います。これには、ファイアウォールの保護、セキュリティプロトコル、保存中と転送中のデータの暗号化、ユーザーのロールと権限、出現するセキュリティの脅威のモニターと対応に、絶えず慎重に配慮することが含まれます。

効果的なデータセキュリティは複雑で、特に人材面で導入にコストがかかります。セキュリティ対策の導入が不適切な場合、突破されるとさらなる費用がかかります。

多くの顧客に対応するクラウドデータウェアハウジングプロバイダーは、業務仕様のエンドツーエンドデータウェアハウスセキュリティを実現するための専門知識とリソースを提供できます。保存中と転送中の両方のデータのセキュリティを保つために、業界標準のエンドツーエンドの暗号化を保証するプロバイダーを選定してください。

データ保護と復元の費用

オンプレミスデータウェアハウスは、装置の故障、停電や電圧の急激な変化、盗難や破壊行為、災害(火災、洪水、地震など)によるデータの損失に対して脆弱です。データを保護するために、定期的にバックアップを作成し、それを遠隔地に保存する必要があります。データの損失を防ぎ、データウェアハウスが入力されるデータとクエリを常に処理できるようにするには、バックアップ電源も必要です。災害に襲われた場合、最新のバックアップを使用してデータを回復させるために、スキルを持つスタッフが所定の場所にいる必要があります。データウェアハウスが極めて重要な場合は、サービスが途切れないように自動フェイルオーバーを保証するためのソフトウェア、ライセンス、プロセスとともに、地理的に離れた障害復旧サイト(もう1つのデータセンター)も必要になります。

クラウドプロバイダーは、データの保護と回復に対応する最適なソリューションを提供します。その性質により、データはオフプレミスに格納されます。一部のクラウドベースのソリューションは、2か所以上の物理的に離れた場所で自動的にデータのバックアップを作成します。地理的に孤立しているデータセンターも、障害復旧機能を提供します。クラウドデータセンターには冗長電源があり、長時間の停電の際でも稼動し続けます。クラウドプロバイダーは、コストを数千以上のクライアントに分散させることで、これらの保護を従来よりも低コストで提供できます。


ヒント

独自のデータバックアップを管理しない場合は、検討中のクラウドデータウェアハウスプロバイダーに、サービスの構成について必ず質問してください。同様に、障害復旧保護が必要な場合は、プロバイダーのアーキテクチャが地理的に離れた複数のセンターを使用していることを確認してください。また、プロバイダーが複数のクラウドプロバイダーを横断するソリューションを提供していて、災害が起きた場合に別のクラウドにデータウェアハウスのインスタンスを切り替える必要があるのかもご確認ください。

本章の内容

- パフォーマンスに影響を与える要因を検討する
- データの保護とセキュリティを確保するソリューションを選ぶ
- 管理コストの節約を評価する

第5章 クラウドデータウェアハウスソリューションの比較

クラウドの導入増加にともない、レガシーのオンプレミスベンダーと最近の市場参入者がそれらのデータウェアハウス製品のクラウド版を提案するようになりました。もちろん、2つのソリューションに違いは存在します。本章では、クラウドデータウェアハウスの違いのいくつかと、期待すべき事項について説明します。

クラウドでのデータウェアハウジングに取り組むアプローチを理解する

次に挙げる複数のクラウドアプローチが提供するデータウェアハウス能力には相当な違いがあります。

- **サービス型インフラストラクチャ(IaaS)：**顧客が、クラウドプラットフォームプロバイダーが提供するコンピューターに従来のデータウェアハウスソフトウェアをインストールする必要があります。顧客がクラウドハードウェアとデータウェアハウスソフトウェアのあらゆる側面を管理します。データウェアハウスの能力は、オンプレミスハードウェアを使用して配置された同じソフトウェアと同一です。

- **サービス型プラットフォーム(PaaS)：**このハイブリッドアプローチを使用するデータウェアハウスベンダーは、クラウドサービスとしてハードウェアとソフトウェアを提供します。ベンダーは、ハードウェアの配置、ソフトウェアのインストール、ソフトウェアの構成を管理します。顧客はソフトウェアを管理、チューニング、最適化します。
- **サービス型ソフトウェア(SaaS)：**データウェアハウスベンダーが、ハードウェアとソフトウエアの管理のすべての側面を含めて、すべてのハードウェアとソフトウェアを提供します。通常、サービスには、ソフトウェアとハードウェアのアップグレード、セキュリティ、可用性、データ保護、最適化が含まれます。

これらのシナリオのすべてにおいて、データセンターそのものとデータウェアハウスをサポートするハードウェアの購入、配置、構成の作業が顧客からベンダーに移ります。こうしたメリット以外にも、利用するサービス内容によって、そのメリットとデメリットは使いやすさからセキュリティや可用性までさまざまな優位性があります。


ポイント

データウェアハウスプロバイダーが、クラウドを使用した従来のデータウェアハウスへのアクセスのみを提供する場合、そのソリューションは元のオンプレミスのアーキテクチャと機能に類似している可能性が高いと言えます。

アーキテクチャの比較

多くのベンダーが、元々オンプレミス環境用に設計および展開したクラウドデータウェアハウスを提供しています。これらの従来型アーキテクチャはクラウドよりはるか以前に作られ、馴染みのある選択肢であることがメリットです。また、クラウド向けに構築されるデータウェアソリューションは、クラウドのメリットを活用すべきです(図5-1を参照)。クラウドに最適化されたアーキテクチャに構築されるソリューションを見つけるには、以下の特徴に注視して探してください。

- すべてのデータが一元化されたストレージ
- コンピュートとストレージのリソースが独立した拡張性を持つ
- リソースを奪い合わず無制限に近い並行処理
- パフォーマンスを悪化させずにデータのロードとクエリを同時に実行できる
- ビジネスの継続性を強化し、拡張を簡素化するために複数の地域とクラウドを横断してデータを複製できる

- APIをセットアップしたり、面倒なETL手順を確立したりしなくてもデータを共有できる
- システム全体に適用される強固なメタデータサービス。(メタデータは、ファイルのサイズ、作成者、作成日時など、データ自体に関するデータを指します。)クラウドに最適化されたアーキテクチャは、ユーザーに対して自動的かつ透過的にデータストレージが拡張および縮小される、サービスとしてのストレージも活用します。古いアーキテクチャ向けに設計されたデータストレージは高価であり、拡張性が限定されます。

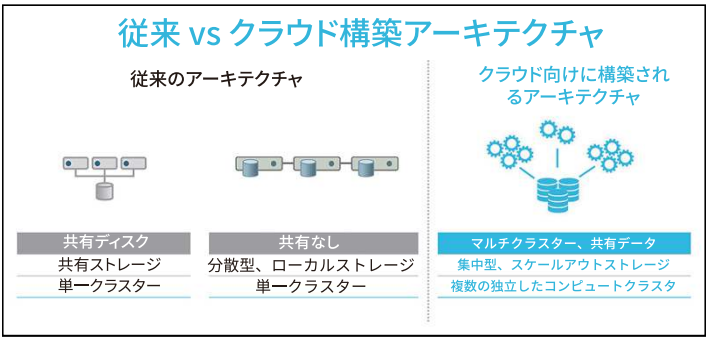

図 5-1：クラウドに最適化されたアーキテクチャがパフォーマンスを合理化する方法。

データの多様化管理の評価

クラウドデータウェアハウジングの採用を推し進める重要要因は、企業のデータセンターの外にあるクラウドで生じるデータ量の増大に端を発しています。ほとんどの場合、この非リレーショナルデータは、オンプレミスまたはクラウドにある従来のデータウェアハウスにロードされる前に変換される必要があります。このアプローチはかなり複雑になり、新しいデータへのアクセスが遅れます。

このようにデータの量と多様性が増大したことで、クラウドが統合ポイントとして当然のように機能しはじめました。この問題の理想的な対処方法は、リレーショナルデータと非リレーショナルデータの両方を処理でき、データのロードやクエリの処理中に非リレーショナルデータの変換やパフォーマンスの妥協を必要としないクラウドデータウェアハウスを使用することです。


ポイント

データは、従来型のクラウドベースウェアハウスにロードされる前に変換される必要があります。または、組織が非リレーショナルデータを処理する追加システムを購入し、保守する必要があります。

拡張性と弾力性の評価

すべてのクラウドデータウェアハウスが同じ種類の弾力性を備えているわけではありません。高度なソリューションは、スケールアップ、スケールダウン、オンザフライが可能で、システムをオフラインにしたり、読み取り専用モードにする必要もありません。


ヒント

拡張性が劣るソリューションの欠点を考察します。

- 手作業での再構成が必要なクラウドデータウェアハウスは、リソースを拡張するために注意深い計画立案とベンダーとの協力が欠かせません。
- 拡張には、データの再分配とシステムの再構成のために、ダウンタイムまたは読み取り専用モードへの切り替えが必要になる場合があります。
- ほとんどのクラウドデータウェアハウスの提供内容では、同じノード上にコンピュートとストレージをセットし、顧客はコンピュートとストレージのうち片方のみを増やせばいい場合でも、両方を拡張しなければならなくなります。
- ほとんどがオンプレミスソリューションの「クラウドに流れ着いた」バージョンであり、ピーク時に使用する場合に備えて、大き過ぎてめったに使わない構成を買うはめになります。そのうちに、利用できるリソースを上回り、高価なアップグレードの必要性に直面するでしょう。

並行処理能力の比較

*並行処理*は、2つ以上のタスクを同時に実行したり、2人以上のユーザーが計算ソリューションにアクセスできたりする能力です。従来のデータウェアハウスでは、コンピュートとストレージのリソースが固定され、並行処理は制限されます。しかし、クラウドではコンピュートとストレージは固定されません。クラウドに最適化されたアーキテクチャは、次の2つの方法で並行処理をサポートします。

- 複数のユーザーが、パフォーマンスを悪化させずに同じデータにクエリを同時に実行できます。
- ロードとクエリが並行して発生できるため、リソースの競合なしに複数の同時のワークロードが可能です。

SQLとその他のツールのサポートの確保

ほとんどのビジネスインテリジェンス(BI)、抽出・加工・書き込み(ETL)、データアナリティクスのツールは、標準SQLをサポートするデータウェアハウスと通信できます。ただし、すべてのクラウドデータウェアハウスソリューションが、標準SQLを完全にサポートしているわけではありません。たとえば「クラウドデータウェアハウス」として位置づけられるビッグデータソリューションは、殆どがNoSQLソリューションであり、SQLのサポートは不完全または非標準です。これらの新しいアナリティクスツールのサポートは重要ですが、データのクエリに関してはSQLはいまだに業界標準です。データウェアハウスは、データ管理、データ変換、データ統合、視覚化、BI、そしてその他の種類のアナリティクスのためにSQLツールをサポートすべきです。

バックアップ/回復サポートの確認

オンプレミスと多くのクラウドデータウェアハウジングのソリューションでは、顧客は自社のデータをバックアップとデータレプリケーションのツールを使用して保護する必要があります。しかし、一部のクラウドデータウェアハウスソリューションには、サービスの一部としてデータ保護が提供されます。


ポイント

最適な保護のために、データの過去のバージョンを自動的に保存できるか、使用するデータをオンラインバックアップで自動的に複製できるるソリューションを選定ください。ソリューションは、事業の完全な継続性を目指して、同じクラウドプロバイダーが複数の地域、または、複数のクラウドプロバイダーによるレプリケーションによって損失または破損したデータのセルフサービスリカバリーも可能なことが望ましいです。

復元力と可用性の確認

*復元力*は、コンポーネント、ネットワーク、あるいはデータセンターの障害の最中でも自動的に機能し続けるためのデータウェアハウスの能力です。*可用性*は、ユーザーがシステムに常時アクセスできる能力です(「アップタイム」として知られています)。クラウドデータウェアハウスサービスが、顧客が可用性と弾力性にどの程度責任を負うかは場合によります。最も基本的なレベルのクラウドデータウェアハウスサービスでは、顧客がシステムモニターを担当し、障害を検出し、ときによっては防ぐことが必要があります。また、データウェアハウスの複製コピーが障害の際に使用できるように、顧客がデータレプリケーションも管理しなければならないケースもあります。これの対極に位置するベンダーは、サービスの一部としてモニター、レプリケーション、自動フェイルオーバーを提供します。

可用性はソフトウェアアップグレードの要因でもあります。アップグレードの間、ベンダーによってアプローチの方法が異なります。

- **基本：**顧客がアップグレードと関連するダウンタイムを管理します。
- **より望ましい：**ベンダーがアップグレードを管理し、ユーザーに今後のアップグレードを通知します。そのため、ユーザーはダウンタイムに備えて計画を立てることができます。
- **最善：**ベンダーが、ユーザーに影響を与えない、またはユーザーがダウンタイムの対象にならない透過的なアップグレードを提供します。また、ベンダーが顧客に自動アップグレードのオプトインまたはオプトアウトを許可することもできます。そうすれば、顧客は希望するタイミングでアップグレードを受けることができます。


ヒント

クラウドデータウェアハウスソリューションがサポートする可用性の「9」が何桁あるかに(99.9XXパーセントのアップタイム)注視して選定しましょう。

パフォーマンスの最適化

クラウドの有利な点の1つは、必要な時のみに料金を支払えばよい、膨大な使用可能リソースを持てることです。需要に応じてパフォーマンスを最適化でき、新しいリソースを取り込む管理の手間を省くクラウドデータウェアハウスを選びましょう。


ポイント

リソースを加減するためにアクティビティを中断または遅らせるデータウェアハウスは避けましょう。一部のソリューションでは、データの再分配とメタデータの再計算を含む管理作業も必要です。

クラウドデータセキュリティの評価

クラウドは、一般的にオンプレミスデータストレージよりセキュリティが低いと見なされますが、「セキュアな」オンプレミスデータセンターへの侵入が懸念されるため、クラウドソリューションは次第に受け入れられてきました。これらのインシデントにより、企業のデータのセキュリティを保つ能力には限界があることが明らかになりました。クラウドデータウェアハウジングのオファーは、物理的なデータセンターのセキュリティに対する責任をソリューションベンダーに移しますが、以下の点に注意してください。セキュリティ機能はベンダーによって異なります。

- 基本的なクラウドデータウェアハウスのオファーは、いくつかのセキュリティ能力のみを提供し、暗号化、アクセス制御、セキュリティのモニターなどは顧客任せです。
- その他のソリューションは暗号化やアクセス制御などの機能を提供し、顧客はそれらの使用を選択できますが、選択しなければシステムは脆弱なままになります。
- サービスを重視するクラウドデータウェアハウスのサービスは、セキュリティのための機能を組み込み、暗号化、暗号キー管理、キーのローテーション、侵入検出などをサービスの一部として提供します。

管理コストの構成

従来のデータウェアハウスは、顧客の時間、労力、専門知識をかなり必要とします。1人以上のデータベース管理者(DBA)が、ソフトウェアのパッチとアップグレード、データのパーティショニングと再パーティショニング、インデックス管理、ワークロード管理、統計の更新、セキュリティの管理とモニター、バックアップとレプリケーション、クエリのチューニングと再書き込みなどを実行する必要があります。

古いオンプレミステクノロジーで構築される基本レベルのクラウドデータウェアハウスソリューションでは、現在でも顧客がこれらの全作業を管理する必要があります。新しいデータウェアハウジングサービスでは、最新の設計と自動化により、この管理費の多くを削減するか取り除くことができます。

セキュアなデータ共有を可能にする

多くのビジネスは、サードパーティのデータリポジトリー、サービス、ストリームを活用して業務を強化できます。FTP、API、電子メールなどの従来のデータ共有方法では、データをコピーして顧客に送信する必要があります。これらの面倒でコストが高く、リスクがある方法はすぐに陳腐化し、絶えず最新バージョンで更新する必要がある静的データの共有がベースになります。第6章では、クラウドで構築されるデータウェアハウスでは、最新の管理されたセキュアなデータ共有がどのように可能になるかについて詳しく説明します。

ヒント

今日の強固なデータ共有方法では、最新データを所在場所を変えずに交換できます。

グローバルデータレプリケーションの許可

*データレプリケーション*により、クラウドにデータのコピーが複数作成されます。この種のグローバルなフットプリントは、障害復旧と事業の継続性に欠かせないだけではありません。複数地域にETLパイプラインをセットアップせずに、グローバルな顧客基盤とデータを共有する場合にも便利です。一流のデータウェアハウスベンダーでは、複数の地理的な地域と、アマゾンウェブサービス(AWS)、Microsoft Azure、Google Cloud Platform (GCP) などのクラウド間で、簡単にデータを共有できます。これらのグローバルレプリケーション能力により、市場が拡大し、パートナーとの関係構築が容易になり、アナリティクスとデータ共有に対応する完成度の高いエコシステムが可能になります。

ワークロードの確実な隔離

データウェアハウスのスピードとパフォーマンスにとって重要な要因は、ワークロードを隔離する能力です。効率を上げるため、クラウドデータウェアハウスでは、並行して実行する必要があるユーザーとプロセスのワークロードを隔離するために、(さまざまなサイズの)コンピュートリソースの複数のプールを容易に構成できなければなりません。これにより、競合が排除され、各ワークロードに合わせてサイズが決定されたリソースが提供されます。理想的なのは、こうした別々のワークロードが同じデータに同時にアクセスでき、必要性によって容易にオンとオフを切り替えられることです。

すべてのユースケースを可能にする

従来の環境ではさまざまなデータシステムが多種多様のユースケースを扱います。ユースケースには、運用報告のためのデータウェアハウス、部門のレポートとアナリティクスのためのデータマート、データ探索のためのデータレイク、予測アナリティクスなどのアクティビティのための特別なツールなどがあります。これらのそれぞれに、ハードウェア、データのコピー、個別管理などが必要です。

多様なユースケースを一緒にクラウドで運用できるようにするには、複製するデータの従来の形式を保ちながら、ストレージにトラブルやコストが発生することなく、データウェアハウスがテーブル、スキーマ、データベースの複数のコピーをクローンできる高速で効率的な方法をサポートする必要があります。また、クラウドデータウェアハウスは、以前のバージョンのデータに簡単にアクセスおよびロールバックできるタイムトラベルなどの機能を使用するデータ変換ジョブによって起こるエラーや問題からも、容易に回復できなければなりません。

本章の内容

- » データ共有の重要性を認識する
- » 効率的なデータ共有アーキテクチャを確立する
- » データ共有の機会を活用する

第6章 データ共有を可能にする

データ共有は、共有すべき貴重な資産を持つ企業の決定によって、企業内と企業間の両方でデータへのアクセスを提供する行動です。データを使用可能にする、あるいはデータを共有する組織が、*データプロバイダー*です。共有されたデータを使用する組織が、*データコンシューマー*です。組織はデータプロバイダー、データコンシューマー、またその両方であることもあります。

組織はその内部であらゆるデータを生成および共有していることに加えて、多くがサードパーティのデータリポジトリー、サービス、ストリームを活用して業務を強化しています。たとえば、金融サービス組織はさまざまな市場、金融、経済の指標を活用して、顧客に新しい製品を提案するのに役立つ優れたデータモデルを作成することがあります。

社内そして外部の市場との交換を通じて、世界の急増するデータソースからもたらされる潜在的な価値は豊富にあります。しかし最近まで、相当なリスク、コスト、トラブル、そして遅延なしにデータ共有をするためのテクノロジーは存在しませんでした。データ共有の商業利用には1世紀近い歴史がありますが、これまでのすべての方法には制限がありました。すべての組織が、すぐに使用できる最新データに必要に応じてアクセスし、ただちにそれを利用できたか、その可能性は想像がつくことでしょう。データは、データプロバイダーによる分析を待つ必要はなくなりました。データコンシューマーの手に渡り、データコンシューマーが再構築するのです。セキュアな制御された環境内ですぐにアクセスでき、いつでも使用できるのです。

技術上の問題に取り組む

ファイル転送プロトコル(FTP)、クラウドストレージ(Amazon S3、Box、Dropboxなど)、アプリケーションプログラミングインターフェイス(API)、電子メールなどの従来のデータ共有方法は、共有されたデータのコピーを作成し、データコンシューマーに送付する必要があります。これらの面倒でコストが高く、リスクがある方法は静的データを作り出しますが、このデータはすぐに古くなり、最新バージョンで更新する必要性や、継続的なデータの移動と管理を必要とします。

新しいデータ共有テクノロジーにより、組織はデータスライスを簡単に共有し、セキュア・制御された方法で共有されたデータを受け取ることができます。これらのテクノロジーには、データの移動、抽出・変換・ロード(ETL)テクノロジーも、データを最新に保つための継続的なアップデートも必要ありません。FTPを使用してデータを転送する必要はなく、アプリケーションをリンクするためにAPIを設定する必要もありません。データはコピーされるのではなく共有されるため、クラウドストレージを追加する必要はありません。この新しいアーキテクチャにより、図6-1に示されているように、データプロバイダーは、データコンシューマーが瞬間的に発見、クエリ、改善できるよう、簡単かつセキュアにデータを公開できます。

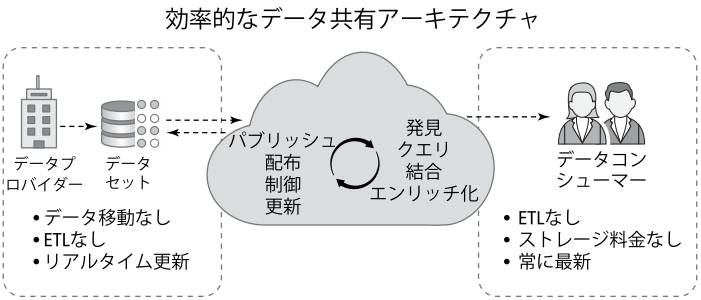

図 6-1：リアルタイムなデータ共有に対応する効率的なアーキテクチャ。

マルチテナントのクラウドに構築されるデータウェアハウスは、クラウドエコシステムの権限を与えられたメンバーが最新の読み取り専用バージョンのデータを活用できるため、データ共有サービスに最適なプラットフォームを提供します。データプロバイダーは、ベンダー、サプライチェーンパートナー、ロジスティクスパートナー、顧客、多くのその他の関係者とデータを共有できます。これらのクラウドに構築されるソリューションは、クラウドコンピューティングとデータウェアハウジングの最先端技術を活用します。データを物理的に内外のコンシューマーに転送するのではなく、ウェアハウスでは、SQLを使用して最新データセットの制御された部分に読み取り専用アクセスを許可します。

データ共有を成功させる

組織がデータ共有に着手すると、ほとんどは似たような経緯をたどります。

1. **内部のコラボレーション：**コラボレーションを強化し、データサイロを破壊しながら事業部と関連会社とともに企業内でデータが共有されます。
2. **ビジネスインサイト：**所有するデータの完成度が上がるほどコラボレーションが強化され、データ共有が規範になるにつれて優れたビジネスインサイトが得られます。
3. **顧客アナリティクス：**企業は、製品またはサービスの価値を向上させるために顧客志向のアナリティクスを構築します。これは、データの収益化の第一歩です。
4. **高度なアナリティクス：**顧客がさらに多くのデータを求めるため、企業は顧客に豊富な情報をデータから提供するためにカスタムアナリティクスサービスを開発します。
5. **データサービス：**企業は、顧客にデータモデリング、データエンリッチメント、データアナリティクスなどのデータ補強サービスも提供するために、内部データセットを活用します。
6. **データエクスチェンジ：**企業は、通常はデータマーケットプレースやデータエクスチェンジを通じて外部データを調達し、より多くの対象ユーザーにデータ製品を提供して、データ製品を向上させる方法を模索します。

データの収益化

ほとんどの組織は既にデータを共有しているか、その計画を持っていますが、データを収益化する方法については見逃している可能性があります。データの収益化を追求するマーケットプレースは限りなく急速に拡大しています。調査会社のIDCは「2019 Predictions for Digital Transformation（2019年のデジタルトランスフォーメーションの予測）」において、企業の80%が2020年までにデータの管理と収益化機能を開発し、2023年までに事業体の95%が、新しいデジタル重要業績評価指標(KPI)のセットを取り入れると予測しています。

ヒント

適切なデータ共有アーキテクチャがあれば、新しい製品、サービス、市場機会を見つけるために多くのデータを簡単に分析できます。

収益機会の最大化


ケーススタディ

Environics Analyticsは、北米有数のデータアナリティクス企業です。データ主導によるインサイトを3,000以上のクライアントに配信するために、Environicsは大量の人口統計、ロケーション、消費者のデータを取り込み、分析しています。

Environicsは最近、アナリティクスアクティビティをデータとワークロードをいくらでも処理できるクラウドに構築されたデータウェアハウスに移動しました。それに組み込まれたデータエクスチェンジサービスによって、顧客は新しいデータを見つけ、瞬時に入手することができます。Environicsの製品開発担当シニアバイスプレジデントのSean Howard氏によると、セキュアなデータ共有サービスを所有して便利なデータ配信メカニズムを提供すると、収益の拡大につながる機会が大量に生まれます。クラウドプラットホームは、ITチームの助けを借りなくても、各ユーザーの分析ニーズに合わせてスケールアップとスケールダウンを素早く行うことができます。

以前は、Environicsのデータサイエンティストが自分のコンピューターにデータセットを保存し、FTPを使用して完成品をクライアントと共有しました。これは内部で混乱を引き起こし、成長の足かせになっていました。数十億行のイベントを含む大規模なデータセットを調べるには、ハードウェアの設置、SQLサーバー環境の構築、クエリパフォーマンスの最適化、ストレージとコンピュートリソースの使用状況のモニターを行うためにITチームの継続的なサポートが必要でした。

現在は、需要に応じて拡張する分析環境があり、データサイエンティストはあらゆる業界、ソース、またはファイルタイプから大きなデータセットのプロトタイプを安心して作ることができます。彼らは、数十億の未加工データポイントを有望なデータ製品に変換できます。セキュアなデータ共有サービスはカスタマーロイヤルティを高め、達成までコストを抑え、バージョン管理を劇的に簡素化しながら不要なファイル転送を削減します。

小売り、銀行、信用組合、不動産会社、非営利団体、政府機関が、消費者と市場について十分に情報を検討して決断を下すことを可能にするデータエクスチェンジを使用しています。Environicsは、現在、データロードを迅速に処理し、ほぼリアルタイムのアナリティクスを可能にする継続的なデータ取り込みサービスを利用して、モノのインターネット(IoT)データとその他のビッグデータのソースを実証しています。「データエクスチェンジへの参加は、ビジネスの真の成長を促し、より多くの有望なクライアントに私たちのデータを理解していただくのに役立っています」とHoward氏は語りました。

本章の内容

- 障害復旧力と事業継続性を強化する
- ベンダーに縛られないクラウド間のポータビリティを可能にする
- グローバルな拡張構想を採用する
- マルチクラウド環境でセキュリティと管理を簡素化する

第7章 マルチクラウド戦略で選択肢を最大限に増やす

複数の地域やクラウドに広く展開することができるデータウェアハウスを保有すると、データ共有、事業の継続性、地域への浸透に大きなメリットをもたらします。Flexaの「2019 State of the Cloud(2019年のクラウド状況)」レポートによると、組織の84%に市場の現実を織り込んだマルチクラウド戦略があります。それがアマゾンウェブサービス(AWS)、Microsoft Azure、Google Cloud Platformのどれであれ、各クラウドサービスが対応するサービスには多少の違いがあります。

データウェアハウスを使用して世界への展開を目指す組織にとって、クロスクラウド戦略には意味があります。ニーズに最も合うクラウドストレージベンダーを選ぶことができる一方で、世界中のどこでも自由でセキュアなデータの移動が可能になるのです。たとえば、組織内の各部門にそれぞれ独自のクラウド要件があるかもしれません。すべての事業部が同じプロバイダーを使用する需要を満たすよりも、むしろマルチクラウド戦略の方が各部門にとって最適なクラウドの使用を実現できます。こうした柔軟性を求める場合は、複数のクラウド環境をサポートし、クロスクラウドのサポートを提供するSaaSプロバイダーを探しましょう。

クロスクラウドを理解する

*マルチクラウド*は、複数の異なるクラウドにデータを格納できることを意味します。クロスクラウドは、それらのすべてのクラウドからデータに同時にアクセスして、1つのクラウドから別のクラウドに分析操作をシームレスに移行し、クラウド間でデータを共有できることを意味します。ユーザーは1つのクラウドベンダーに縛られないため、これがクラウドデータウェアハウジングの究極の目標になります。これが重要な理由を以下に挙げます。

- すべてのクラウドプロバイダーが全地域で運営しているわけではないため、グローバル企業にとって戦略的優位になります。
- 使用してるクラウドとは別のクラウドで標準化している企業を買収する場合に有用です。
- データの共有または収益化を計画している場合、複数の地域とクラウドに広がる一元的データ管理プラットフォームがあると、対応可能な市場が広がります。

次のセクションでは、クロスクラウドデータウェアハウスを可能にするテクノロジーを評価します。

ヒント

クラウド構成の違いを解決するという大変な作業をやり遂げたデータウェアハウスベンダーと協力して、すべてのクラウドに及ぶ共通コードベースの上にソリューションを構築しましょう。

グローバルレプリケーションの活用

*データレプリケーション*は、局所的な停電中でもデータの可用性を維持するために、複数の場所にデータを格納するプロセスです。また、これは複数の地域とクラウドを横断してデータの共有を可能にする基本的なテクノロジーです。データウェアハウスには、地域展開の選択肢を最大限に広げ、事業の継続性を可能にし、業務を世界中に拡大するために高度なデータレプリケーションテクノロジーが必要です。

データウェアハウスプラットフォームでは、主要なデータに対する操作のパフォーマンスを低下させずに、複数の地域とクラウドにレプリケーションを作成することができます。

サービスの中断を最小限に抑える

クロスクラウドデータウェアハウスレプリケーションは、ビジネスに欠かせない障害復旧シナリオにとって重要です。停電が発生した場合に、ダウンタイムを発生させずに瞬間的にデータ処理アクティビティを再開できることを保証します(図7-1を参照)。ただし、適切なデータレプリケーションテクノロジーがなければ、大規模なデータウェアハウスの地理的なバックアップの回復には、数時間または数日かかる場合があります。その場合、回復時間の目標を達成できますか？

データウェアハウスベンダーに、あらゆるサイズの、あらゆるクラウドや地域に存在するデータベースで、瞬間的なアクセスと回復に対応しているかを問い合わせましょう。世界の特定の地域で災害が起きた場合、別の地域またはクラウドサービスで複製されたデータにだだちにアクセスできなければなりません。データウェアハウスプロバイダーがデータベースを複製し、複数のクラウドプラットフォームと地域にわたって同期状態を維持できるかどうかを調べましょう。

複数地域とクロスクラウドのレプリケーション

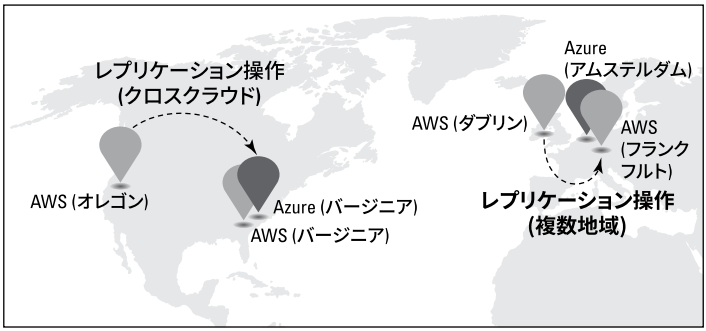

図 7-1：グローバルなデータレプリケーションが、停電中の事業の継続性を保証します。

マルチ・クロスクラウドのサポート

データポータビリティは、大量のデータを所有するすべての組織に関わる問題です。各パブリッククラウドプロバイダーの地域の普及率は異なります。地理的な領域とクラウドの間でのデータとワークロードの移動は、クロスクラウドアーキテクチャであれば容易になります。

データポータビリティによって、業界でデータを特定の国または地域内に残すことが必要であれば、法規制の順守が簡単になります。異なるクラウドベンダーを使用する別の企業と合併あるいは買収しても、この場合は容易になります。

データ主権を満たす

企業が成長すると、業務地域内でデータ処理業務を行うことが期待される場合があります。マルチクラウド戦略によって、各地域で最強のクラウドを選択する柔軟性が得られるため、レイテンシーを最低に抑え、地域の居住要件を守り、データ主権指令に遵守するアーキテクチャをセットアップできます。データへのアクセスを犠牲にせずに遠隔地に業務運営を拡大できるようになり、企業全体にとって単一のソースを維持できることの真の価値が分かることでしょう。

また、データレプリケーションはデータの共有化と収益化、共同事業者との情報交換を容易にし、その一方で「データソースは単一の場所に存在し、移動しなくてもアクセスが可能である」というデータ共有の基本原則をすべて順守できます。

セキュリティの簡素化

複数のクラウドで作業する場合、クラウドプロバイダーのすべてに同じセキュリティの構成やテクニックをどうすれば適用できるでしょうか？監査の証跡とイベントログの違いを解決する必要がありますか？貴社のサイバーセキュリティエキスパートは異なるルールセットに対応する必要がありますか？あるいは、データを暗号化する複数のキー管理システムに手を加えますか？すべてのクラウドプラットフォームに展開する一元化されたコードベースは、それらの操作のすべてを簡素化します。特別なスキルセットを持つ人々を雇用する必要も、複数のクラウドの微妙な違いを熟知した状態を維持する必要もありません。


ポイント

高度なレプリケーションテクノロジーがあれば、パイプラインのセットアップやデータのコピー、セキュリティの相違を解決せずに、多くの地域と異なるベンダークラウドの間でデータを容易に共有できます。これによって市場が広がり、パートナーの参加が容易になり、データの分析と共有のための強固なエコシステムがもたらされます。

本章の内容

- 包括的なデータセキュリティを確立する
- プライバシー規制を順守する
- 認証と認定を検証する
- データの維持、保護、可用性を向上させる

第8章 データセキュリティを高める

クラウドセキュリティには真実があります。それはほとんどの場合、クラウドにあるデータはデータセンターにあるデータよりも安全という事実です。Tom Davenport、Ashish Verma、David Linthicumを含むチームが執筆した、2019年のDeloitteによるITエグゼクティブに対する調査で、組織の9割以上が主にクラウドプラットフォームにデータを保存していることがわかりました。データのセキュリティとガバナンスが、組織にとってデータをクラウドに移行する一番の推進力であったことが調査で確認されました。

SaaSクラウドプロバイダーは、数千、さらには数百万の顧客に対応しています。プロバイダーには、業務仕様のエンドツーエンドのデータセキュリティを提供できるリソースの余裕があります。しかしながら、すべてのクラウドプロバイダーがデータを守る努力をしているわけではありません。詳しく調べれば、セキュリティ能力のばらつきが非常に大きいことがわかるでしょう。

データを保護し、関連規定を遵守することが、クラウドデータウェアハウスサービスのアーキテクチャ、導入、運用の基本になるはずです。サービスのすべての側面で、現在のセキュリティ脅威と進化中のセキュリティ脅威の両方を対象とする多層的なセキュリティ戦略の一環として、データの保護が中核に位置づけられなければなりません。この戦略では、包括的なモニター、警告、検証可能なサイバーセキュリティ対策と組み合わせて、外部インターフェイス、アクセス制御、データストレージ、物理的なインフラストラクチャ構築に取り組む必要があります。

デフォルトでデータを暗号化

データの暗号化は、平文を暗号文に置き換える暗号化アルゴリズムの適用を意味しています。これはセキュリティの基本です。データが自社設備を出てからインターネットを通じてデータウェアハウスに到達するまで、ディスクに保存されるとき、中継所に移動するとき、データベースオブジェクト内に配置されるとき、仮想データウェアハウス内にキャッシュされるときにデータを暗号化します。クエリ結果も暗号化しなければなりません。これらのすべてが組み込まれる必要があります。これらはオプションではなく必須です。

ベンダーは、データを復号する暗号解読キーも保護する必要があります。優れたサービスプロバイダーは、キーの階層モデルを採用したAES 256ビット暗号化を提供します。この方法では暗号化キーを暗号化し、キーのローテーションを構成することで1つのキーが使用できる期間を限定します。


ポイント

データは多くの場所に存在する可能性があります。各ポイントでデータフローを保護および制御する必要があります。すべてのデータは、転送中も保存中も自動的にエンドツーエンドで暗号化される必要があります。

アクセス制御の適用

データの安全性を確保することは、包括的なセキュリティの１つの側面にすぎません。データの漏洩は、多くの場合、ユーザーが簡単に破られるパスワードを選び、不十分な認証手続きがそれに重なった結果です。クラウドデータウェアハウスサービスは、常にユーザーと認証情報を検証し、ユーザーにアクセスが許可されたデータのみにアクセス権を付与する必要があります。

出発点は、ユーザーが閲覧を許可されたデータのみにアクセスできる*ロールベースのアクセス制御*です。アクセス制御は、テーブル、スキー

マ、データウェアハウスへの仮想的拡張を含むすべてのデータベースオブジェクトに適用される必要があります。利便性とセキュリティを最大限に高めるには、ユーザーの携帯電話に送信される1回限りのセキュリティコードなどの2段階認証が必要な*多要素認証（MFA）*についてもクラウドデータウェアハウスが提供する必要がありあります。

シングルサインオン手順と*フェデレーション認証*を使用すると、ユーザーが他の許可されたアプリケーションからデータウェアハウスサービスに直接ログインすることが容易になります。フェデレーション認証はID管理とアクセス制御の手順を一元化し、チームによるユーザーアクセス権限の管理を容易にします。


ヒント

管理者からアクセス権を付与しない限り、クラウドデータウェアハウスベンダーは顧客の暗号化されていないデータにアクセスすることはできません。

パッチ適用、更新、ネットワークのモニター

ソフトウェアのパッチとセキュリティの更新は、利用可能になり次第、すべての関連するソフトウェアコンポーネントにインストールされなければなりません。また、ベンダーは定期的に独立したセキュリティ会社によるセキュリティテスト(侵入テストとしても知られています)を実施し、脆弱性を事前にチェックする必要があります。

データセンターの物理的なセキュリティ対策には、誰も不正にアクセスできないよう、生体認証によるアクセス制御、武装警備員、ビデオ監視が含まれているべきです。すべての物理的および仮想的機械は、監査、モニター、警告に関する厳密なソフトウェア手順でさらに管理される必要があります。さらなるセキュリティ手段として、ファイル整合性監視(FIM)ツールが重要なシステムファイルが改ざんされていないことを保証し、IPアドレスのホワイトリストはデータウェアハウスへのアクセスを信頼されているネットワークのみに限定します。(ホワイトリストとは、電子メールのブロックプログラムがメッセージの受信を許可する電子メールアドレスまたはドメイン名のリストです。)

ネットワークを監視するサイバーセキュリティモニターシステムによって生成されるセキュリティ「イベント」は、不正開封を防止するセキュリティ情報およびイベント管理(SIEM)システムで自動的にログインされなければなりません。疑わしいアクティビティが検出されると、自動警告がセキュリティスタッフに送信される仕組みが必要でし。

データの保護、維持、冗長性の確保

災害時には、クラウドデータウェアハウスプロバイダーとのサービスレベル契約(SLA)により、指定された保持期間内にテーブルまたはデータ

ベースにあるデータの以前のバージョンを瞬間的に回復または照会できるようにしなければなりません。完全なるデータ保持戦略においては、同じクラウドの地域またはゾーン内でのデータレプリケーション以上のことが必要となり、地理的な冗長性のために、複数の使用可能なゾーンにデータを複製する必要があります。オプションで、他のゾーンへの自動フェイルオーバーによる継続的なビジネス業務を保証することもできます。

テナントを隔離する必要性

データウェアハウスベンダーがマルチテナントのクラウド環境を使用し、そこで多くの顧客が同じ物理インフラストラクチャを共有する場合は、各顧客の仮想データウェアハウスがすべての他のデータウェアハウスから隔離されていることを確認します。ストレージについては、この隔離が仮想マシン層まで伸びていて、各顧客のデータストレージ環境が、独立したディレクトリと一意の暗号化キーによって他のすべての顧客の環境から隔離されている必要があります。一部のベンダーは専用の仮想プライベートネットワーク(VPN)も提供し、顧客のシステムからクラウドデータウェアハウスまでをつなぎます。これらの専用サービスによって、データウェアハウスの最も機密性の高いコンポーネントは他の顧客のコンポーネントから完全に隔離されます。

ガバナンスとコンプライアンスの維持

データガバナンスの整備によって、企業のデータは適切にアクセスおよび使用され、毎日のデータ管理業務はすべての関連法規の要件を遵守して行われるようになります。ガバナンスポリシーにより、データの所有権とアクセス可能性を管理する規則と手順を確立します。これらのガイドラインに共通して該当する情報の種類には、クレジットカード情報、社会保険番号、誕生日、IPネットワーク情報、地理位置情報の座標が含まれます。

認証と認定の需要

コンプライアンスとは、強固なサイバーセキュリティ対策の実施だけではありません。データウェアハウスプロバイダーが、必要なセキュリティ手順を実施していることを証明できることも必要です。データの漏えいは、修復に多額なコストがかかり、顧客との関係を永久に損なう可能性があります。

業界標準の認証レポートは、クラウドベンダーが適切なセキュリティ制御を使用していることを検証します。たとえば、クラウドデータウェアハウスベンダーは、脅威とセキュリティのインシデントを適切にモニ

ターおよび対応し、十分なインシデント対応手順を実施していることを実際に示す必要があります(図8-1を参照)。

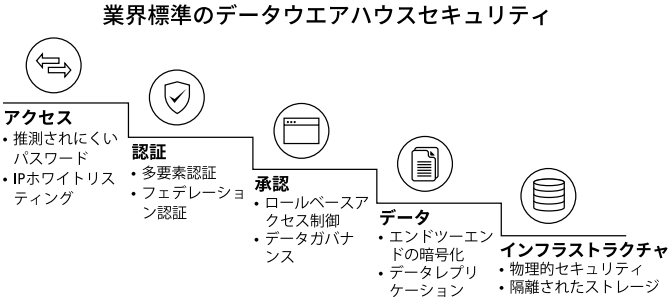

図 8-1：すべてのデータトラフィックが暗号化されセキュアであり、クラウドプロバイダーがすべての関連する認定証を保有していることを確認してください。

ISO/IEC 27001およびSOC 1/SOC 2 Type IIなどの業界標準の技術証明の他に、クラウドプロバイダーがすべての適用される政府と業界の規制に準拠していることも確認します。ビジネスによっては、これにPCI、HIPAA/正常性情報信頼アライアンス(HITRUST)、FedRAMPの認定も含まれる場合があります。

証拠を要求し、ベンダーにカバーレターだけでなく、各関連標準のレポート全文のコピーを必ず提出させます。たとえばSOC 2 Type IIレポートは、過去12か月間、適切な技術上と管理上の制御が実施されていたことを検証します。コンプライアンスのPCI-DSS認証は、ベンダーがクレジットカード情報を適切に保存および処理しているかどうかを明らかにします。保護された正常性情報を扱う場合は、ベンダーがHIPAAガイドラインを順守することが必要です。


ヒント

コンプライアンスと認証は、データウェアハウスベンダーがセキュリティに本格的に取り組み、透明性が高い企業であることを証明します。

クラウドベンダーは、協業するサードパーティのソフトウェアベンダーが法令を順守し、定期的なセキュリティ監査を実行しているというエビデンスも提出しなければなりません。データのセキュリティは、テクノロジーチェーンの中の最も脆弱なリンクと同程度しかありません。そのため、すべての関係者が強固なセキュリティ制御を実施し、業界標準のセキュリティ対策を順守していることを確認します。コンプライアンスの証拠がない場合は、補足書類を入手します。

ポイント

独立した監査人によって、業界で認められるエンドツーエンドのセキュリティ対策を整備していることを実際に確認されたクラウドプロバイダーのみと協業します。こうしたコンプライアンスの検討事項は、この重要なデータリポジトリーに対して満たすべき最低要件とするべきです。

包括的なセキュリティ体制を強く要求する

セキュリティを問題なく維持するのはコストがかかり、専門知識が必要です。装置の故障、ネットワーク違反、保守のミスは、データの損失と不整合につながる可能性があります。包括的なセキュリティ対策は多くの側面を含みます。クラウドデータウェアハウスベンダーには、偶発的または意図的な破壊を防ぐ手順が必要です。セキュリティの能力が不十分で、暗号化やアクセス制御、セキュリティの監視を顧客に任せるベンダーも存在します。セキュリティはデータウェアハウスサービスの基本であり、あなたがデータセキュリティのために何か特別なことをする必要はないはずです。

ヒント

セキュリティ認定証を明示しているベンダーは、信頼できるセキュリティ対策プログラムを提供している可能性がはるかに高くなります。

本章の内容

» コストパフォーマンスが高いストレージ環境を作る

» アーキテクチャと価格設定を通じて最高の価値とパフォーマンスを手に入れる

第9章 データウェアハウスコストの最小化

本章では、クラウドに構築されるデータウェアハウスの運用方法と、データウェアハウスベンダーが長期的なコストの最小化を支援できる方法について考察します。

ストレージコストの最小化

格納できるデータが多いほど、優れたインサイトを得ることができます。幸い、Amazon、Microsoft、Googleのクラウドストレージは比較的低コストであり、格納するデータの量と種類が制限されることはないはずです。クラウドデータウェアハウスベンダーが、ストレージコストを過剰に請求していないかを確認するため、契約条件を確認します。ベンダーは、定価料金表をあなたに直接渡すべきです。ウェアハウスベンダーはデータを3〜5倍の情報量まで*圧縮*して、付加価値を提供できます。3倍の圧縮は、格納するデータ量とコストが3分の1になることを意味します。

使用に関する合意書の条件を調べてください。支払う必要があるのは使用するストレージに対してのみです。超過分や「予約済み」のストレージ容量に対して支払う必要はありません。また、開発やテストのアクティビティ用にデータウェアハウス内で作成するデータベースのクローンに対しても支払う必要もありません。データはコピーされるのではなく、複数回参照できるようになるはずです。そのため、ストレージに対して余分に支払う必要はありません。

クラウドデータウェアハウスでは、構造化データと、JSONなどの半構造化データも格納および照会できる必要があります。最後に、*マルチクラウド機能*を提供するベンダーを探しましょう。データウェアハウスを別のクラウドストレージ環境に移行する場合、将来のコストを節約できるためです。

コンピュート効率の最大化

コンピュートリソースはストレージリソースより高価なため、データウェアハウスサービスは各リソースを別々に拡張でき、利用ベースの価格決定モデルに沿って正確に必要なコンピュートリソースだけを簡単に割り当てることができるようにする必要があります。ベンダーはあなたが使用するリソースのみに秒単位で課金し、コストの上昇を避けるため、使用しない時はコンピュートリソースを自動的に停止すべきです。サブスクリプションベースとは異なり、利用ベースの価格設定ではリソースの消費方法を選ぶことができます。

また、柔軟な契約条件で各ワークロードに対するコンピュートクラスターの「適切なサイズ」も決定できるようにすべきです。低いコンピュート要件で抽出・加工・ロード(ETL)ジョブを実行している場合、過大なプロビジョニングクラスターのコストを発生させるのではなく、小さなクラスターをワークロードにマッチさせることができます。新しい機械学習モジュールのテストが必要な場合は、より大きなクラスターを利用できます。これにより、利用コストを最小限に抑えながら各ワークロードに応じてきめ細かく拡張することができます。ウェアハウスの実行コストはオンプレミスウェアハウスよりも安価で、クラウドバーションはネット上をクロールしながら大きなリソースを使用し、限定された結果を出します。ワークロードは各ワークロード専用のコンピュートクラスターのおかげで、遅くなることはなく、止まることもありません。

本章の内容

- データウェアハウスのニーズと成功基準をリスト化する
- 総所有コストのすべての要因を検討する
- 購入前に試験稼働用のデータウェアハウスを検証する

第10章 クラウドデータウェアハウジングを始める6つのステップ

本章では、組織のクラウドデータウェアハウスを選ぶにあたり、6つの重要なステップを案内します。プロセスはデータウェアハウスのニーズの評価から始まり、最善の選択肢の検証で終わります。最後には、自信を持ってソリューションの選択に役立つ計画を立案できるでしょう。

ステップ1：ニーズの評価

適切なデータウェアハウスは、現在のニーズを満たし、将来のニーズにも対応できる必要があります。そのため、データの性質、既にあるスキルとツール、利用ニーズ、ビジネスの将来計画、どのようにすればデータウェアハウスが想像以上にビジネスを発展させることができるかについて検討してください。

- **データ：**データウェアハウスの対象にすべきデータタイプは何ですか？新しいデータが作られる割合は？データをウェアハウ

スに移動する頻度は？現在、アクセスできない重要なデータは何ですか？

- **既存のスキル、ツール、プロセスに適合させる：**チームのどのツールとスキルを、さまざまなクラウドデータウェアハウスの選択肢に適用しますか？どのプロセスにクラウドデータウェアハウスの影響が及びますか？
- **利用：**どのユーザーとアプリケーションがデータウェアハウスにアクセスしますか？どんな種類のクエリを実行しますか？ユーザーがアクセスする必要があるデータの量は？そのスピードは？時間の経過とともにワークロードはどのように変動しますか？どんなパフォーマンスをユーザーとアプリケーションが必要としますか？データウェアハウスへのアクセスが必要なユーザーの数と、リソースの制約のために現在はアクセスしないユーザーの数は？
- **データ共有：**組織内で、顧客またはパートナーあるいはその両方とデータをセキュアに共有する計画はありますか？計画がある場合、どんな種類のデータを共有しますか？データマーケットプレースまたはエクスチェンジを作ってデータの収益化もしますか？データコンシューマーに未加工データへのアクセスを許可しますか？また、アナリティクスなどのデータサービスも提供してデータを充実させますか？
- **グローバルアクセス：**Amazon S3やMicrosoft Azure、Google Cloud Platformなどのパブリックオブジェクトストアにデータを格納する計画はありますか？特定の機能、地域、またはデータ主権との関係の維持を必要とする要件がありますか？障害復旧を強化、あるいはグローバルな事業の継続性を保証するために、地域の配置の選択肢を最大化するクロスクラウドアーキテクチャが必要ですか？
- **リソース：**データウェアハウスを管理する人材はいますか？可用性、パフォーマンス、セキュリティのモニターと管理に対してどのくらいの投資額を見込んでいますか？データウェアハウスの配置とテストに絞った専門知識がありますか？また、DevOpsチームはこれを合理化しますか？

ステップ2：移行、あるいは最初からやり直し

すべてのクラウドデータウェアハウスプロジェクトは、既存環境をどの程度新しいシステムに移行すべきか、クラウドデータウェアハウスのために何を新たに構築すべきかの評価から始める必要があります。抽出・加工・ロード(ETL)プロセスの設計から、データモデルとソフトウェア

配置のライフサイクルメソッドまで、これらの決定ですべてに対処する場合があります。検討事項:

- **これはまったく新しいプロジェクトですか？**その場合、制約の下で既存の実装を続けるのではなく、クラウドデータウェアハウスの能力を最大限に活用するプロジェクトを設計するほうが、合理的な選択肢であることが多いです。
- **現在のシステムのどの部分が最も多くのトラブルを引き起こしていますか？** 巧みに計画された移行では、最も問題の大きいワークロードをまずクラウドデータウェアハウスに移動させることを念頭に取り組む場合があります。または、素早く成果を上げるために、単純なワークロードを移動する場合もあります。
- **現在のシステムのなかで、クラウドデータウェアハウスへの移行後には解消される制約に対応しているのはどの側面ですか？**適切なクラウドソリューションでは、リソースの制約を回避するため、容量の追加に必要な破壊的行為を回避するため、あるいはコストを最適化するために設計されたツールとプロセスが不要な場合があります。
- **現在のユーザーとアプリケーションは、どのような方法でデータウェアハウスにアクセスしていますか？**SQLなどの業界標準のインターフェイスに依存し、標準的なETLとビジネスインテリジェンスのツールを使用するユーザーやアプリケーションは、新しい手法に適応するのに少しの変更しか必要としない場合があります。
- **データとアナリティクスの要件は、将来変わる可能性がありますか？**進化することを前提に作られたソリューションは予想よりも長く存続し、セキュアなデータ共有とグローバルなデータアクセスなどの高度な機能を活用する新しい機会が訪れる可能性があります。


ポイント

大規模で複雑な従来型のデータウェアハウスがある場合、クラウドデータウェアハウスを快適に使用するために、システムのごく一部を移行してください。その後、クラウドの範囲を何度も拡大できます。

ステップ3：成功基準の策定

新しいクラウドデータウェアハウスへの移行が成功したかどうかを、どのようにして測定しますか？重要なビジネス上および技術上の要件を設定してください。その基準は、パフォーマンス、並行処理、簡素性、総所有コスト(TCO)に焦点を当てるべきです。

ポイント

新しいクラウドデータウェアハウスに以前のシステムでは使用できなかった機能があり、その機能が新しいソリューションのビジネスおよび技術上の成功の評価に関連する場合は、必ず対象に含めてください。

新しいソリューションの成功基準を設定する際に、定量化できる基準、定性的な基準、定量化できる基準の測定方法、定性的な基準の評価方法を決めて、成功をどのように測定するかを決定します。

ケーススタディ

レイテンシ問題の解決

White Opsはサイバーセキュリティサービスの先進プロバイダーです。統計分析を採用する従来のアプローチと異なり、White Opsはロボットと人間との交流活動を区別し、新しい詐欺のパターンを暴いて特徴を明らかにする作業に取り組みながら犯罪活動と戦っています。この絶え間ないプロセスは、大量のデータの格納と処理が必要です。

White Opsは、データを格納および処理するために以前はNoSQLシステムに依存していました。しかしながら、結果のレイテンシがワークロードによっては24時間以上ありました。リクエストが多いほど、遅延は長時間に及びました。

生産性とパフォーマンスを上げるために、White Opsは中核言語としてSQLを使用するクラウドデータウェアハウスを導入し、サービスとして提供しました。このデータウェアハウスによって、White Opsはすべてのデータを1か所に置き、弾力的に拡張し、標準SQLを使用して多様なデータを照会し、詐欺を防止する提案の開発を加速させました。

現在、White Opsは大規模なデータを集約および拡張して、豊富なプログラミングスキルを持つスペシャリストに頼らずにデータにアクセスできるようになり、顧客がオンライン詐欺の壊滅的な影響を回避する手助けをしています。

ステップ4：ソリューションの評価

データウェアハウスのニーズと成功基準を決定すると、ソリューションの評価を開始する準備が整います。本書全体で、利用できる選択肢の違いを詳しく説明しています(第3章、第　4章、第5章を参照)。比較する際には、次に挙げる基準を満たしているか確認してください。

- 現在と将来のニーズに対応

- 構造化データと半構造化データの統合、すべてを1か所に格納、データサイロの作成を避ける
- 既存のスキル、ツール、専門知識をサポート
- データの損失から保護し、容易にデータの回復が可能
- 業界標準のパスワード保護と暗号化を使用するデータ保護
- データとアナリティクスが常時使用できることを保証
- データパイプラインを合理化し、可能な限り短時間で新しいデータが分析に使用できるようにする
- できる限り早く新しいデータウェアハウスのメリットを得られるように、価値を生み出す時間を最適化
- ワークロード隔離に専用リソースをあてる
- 最新データをコピーや移動することなくデータを共有でき、データプロバイダーとコンシューマーを簡単につなげることができる
- データベースを複製し、複数のアカウント、クラウドプラットフォーム、地域でそれらの同期状態を保ち、事業の継続性を強化し、拡張を合理化する
- 開発やテスト用としてデータベースのコピーをせずにクローニングを提供し、レポート、データ探索、予測アナリティクスなど複数のユースケースをサポートする
- 以前のバージョンのデータをロールバックして、エラーや攻撃によって失われたデータを簡単に回復する
- コンピュートとストレージを別々に、そして自動的に拡張し、パフォーマンスの速度を下げずに並行処理を拡張する

ステップ5：総所有コストの計算

価格に基づいてクラウドデータウェアハウスを選ぶ場合は、ライセンス(通常、ユーザー数ベース)、ハードウェア(サーバー、ストレージ、ネットワーク)、データセンター(オフィスのスペース、電力、管理、保守、進行中の管理)、データセキュリティ(パスワード保護と暗号化)、可用性と弾力性を保証するソリューション、拡張性と並行処理のサポート、配置と中継の環境を作成するコストを含めて、従来のデータウェアハウスのTCOを検討します。

一部のソリューションでは、複数のデータマートの構築と管理、複数のデータマートでデータの複数コピーを保有、スタッフのトレーニング、

多様なデータを処理するための複数のシステム(SQLとNoSQLなど)の保有などの追加コストを考慮する必要があります。

クラウドデータウェアハウスの選択肢のコスト計算は、通常は容易ですが、ベンダーのサービスによって異なります。サービス型データウェアハウス(DWaaS)を選択して、すべてをベンダーにアウトソーシングすることを想定すれば、毎月のサブスクリプション料金を基にTCOを計算できます。サービス型インフラストラクチャ(IaaS)またはサービス型プラットフォーム(PaaS)のソリューションを選ぶ場合は(第5章を参照)、ソリューションに含まれないソフトウェア、管理、サービスのコストを追加する必要があります。


ヒント

通常、組織はデータウェアハウスの予想寿命全体でのTCOを計算します。これは一般的に1～3年とされています。主な注意事項：多くの場合、人々はクラウドシステムは高い稼働率で24時間休まず稼動すると思っており、クラウドソリューションが需要の変化に対応してスケールアップとスケールダウンを大胆に実行し、しかも秒単位の課金ができるため節約につながることを見落としています。

テップ6：コンセプト実証を行う

さまざまなクラウドデータウェアハウスの選択肢を調べ、デモを見て、質問をし、各ベンダーの担当者に会ったら、選択する前にコンセプト実証(PoC)を必ず実行してください。PoCではソリューションを検証し、ソリューションがどの程度ニーズに応え、成功基準を満たすかを見極めます。これはテスト稼働と考えてください。これには通常1日から2日間かかりますが、数週間かけて実施される場合もあります。ソリューションが満足のいくパフォーマンスを上げる場合にはその製品を購入するという一般的な了解をしたうえで、有望なベンダーにPoCを依頼します。あるいは、クラウドデータウェアハウジングの場合は、このサービスをサブスクリプション方式で支払います。


ヒント

PoCをセットアップする際には、解決したい問題だけでなく、クラウドソリューションを導入する場合に可能性のあるすべての要件と成功基準のリストを作成します。

データウェアハウジングのニーズと成功基準を出発点として、包括的なリストを作ります。新しいデータウェアハスは、現在のデータウェアハウスで実行できるすべての機能を上回る機能を持ちながらも、現在のシステムの欠点を克服するものであることを確認してください。複数のベンダーとPoCを行う場合は、それぞれに同じチェックリストを使用します。

クラウドデータウェアハウジングの威力を利用して競争優位を獲得する

現代の企業は、急速に増大しつつあるデータにアクセスして、可能な限り優れたインサイトを獲得しています。さらに組織は、事業部門全体として、またビジネスのエコシステム内で、データを収益化するためのデータの交換を行い、セキュアなデータの共有と共有データの入手を望んでいます。しかし、このようなデータへのアクセスは、従来のデータアナリティクスプラットフォームにとって悩みが尽きない、さらに大きな問題をもたらします。最先端を走っている企業は、全ビジネスユーザーにとって、すべてのデータを格納および分析するために最も効率的かつコストパフォーマンスが高い方法はクラウドデータウェアハウジングであることに気付いています。本書は、この新しくエキサイティングなテクノロジーから何が得られるのか、どうすれば組織はその恩恵を受けることができるのかについて明らかにします。

本書の内容

- クラウドデータウェアハウスが出現した理由
- クラウドデータウェアハウスの違い
- 異なるデータウェアハウスの評価方法
- セキュリティとガバナンスが重要な理由
- クロスクラウドソリューションのメリット
- モダンなデータ共有で優れたインサイトを生み出す方法
- 実際のケーススタディ

snowflake®

Joe Kraynak はベテランのDummiesのライターであり、さまざまなトピックの著書と共著書があります。**David Baum** は、化学とテクノロジーを専門とするフリーランスのビジネスライターです。

ビデオ、ステップバイステップの写真、ハウツー記事、またはお買い物をするには、**Dummies.com**にアクセスしてください！

ISBN: 978-1-119-71462-0

for dummies®

Also available as an e-book

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.