



最新のデータ ウェア ハウスを構築する ための基本ガイド



Google Cloud

最新のデータ ウェアハウスを構築しましょう

エグゼクティブ サマリー	1
最新のデータ ウェアハウスで考慮すべきこと	2
Google Cloud のデータ プラットフォームでできること	4
使ってみる	7

エグゼクティブ サマリー

データ量の増加や、顧客が求めるリアルタイムのエクスペリエンスに対応するには、強力なデータ基盤が不可欠です。かつてない高速で大規模なデータ ストリーム、グローバルなビジネスニーズ、IT に精通したユーザーなどの要因により、IT チームにはアジリティが厳しく求められるようになりました。このような変化の中であっても、従来のデータ ウェアハウスで新たな需要に対応しようとしていることが多いのです。しかし、このような従来型データ ウェアハウスでは、もはや太刀打ちできなくなっています。新しい需要に対して十分な準備ができていないのです。現在データ関連の業務に携わる人々によると、最新のデータ戦略を構築しようとする際に、従来型のテクノロジーに伴う制約が足かせになっているようです。



最新のデータ ウェアハウスで 考慮すべきこと

このガイドでは、最新のクラウド データ ウェアハウスをデプロイする際に考慮する点を説明します。

- ・ ストレージとコンピューティングの分離
- ・ 大規模なスケールに対応したシステム
- ・ データ ガバナンス機能とセキュリティの組み込み
- ・ AI や ML を組み込んだ高度な分析
- ・ 使いやすいリアルタイム ストリーミング分析

企業のモダナイゼーションを妨げているものは何でしょうか。Google のお客様の多くが、大半の時間をシステム エンジニアリングに費やし、データの分析には 15% 程度しかかけていないことがわかっています¹。メンテナンス作業に多大な時間がかかっているということです。システムが時代遅れであれば、データ戦略やアジリティは促進されません。しかしそうした複雑な従来型インフラストラクチャを維持するためだけに、企業が人材を雇用し続けているのが現実です。

そういったチームに最新のシステムを導入することで、他のチームが必要なときに迅速かつ安全にデータを把握できるようになり、作業効率が向上します。最近では、どのチームも高度な分析へのアクセスを必要としています。データは貴重な通貨であり、コラボレーションを改善し、業務を高速化する決め手となっています。データの増加とデータ オペレーションの増加を分離することが、ユーザーのアクセスや分析情報への道を開きます。

多くのお客様が、システム エンジニアリングに大半の時間を費やしています。

¹ Google 内部データ、2020 年 1 月。

大量のデータの増加、お客様の需要、リアルタイムのストリーミングデータ機能の出現も、ビジネスデータ分析の状況に変化をもたらしました。以前は夜通しのデータオペレーションや既製のレポートが一般的でしたが、ビジネスチャンスがグローバル化したことで、データウェアハウスはストリーミングデータやバッチデータを読み込みながら、同時クエリもサポートしなければならなくなりました。従来のシステムでは、ハードウェアと費用の高さが大きな制約となっており、その維持に苦慮しています。

ストリーミングデータがうまく機能すれば、企業は、オンラインカスタマーエクスペリエンスのカスタマイズ、複雑な製造システムにおけるリアルタイムなメンテナンス要件の予測、リアルタイムPOSトランザクションデータに基づく一貫した視点での在庫状況の把握、財務リスクの計算、不正行為の検出、何十万人ものプレイヤーのモバイルゲームエクスペリエンスの向上が実現できます。

今日のデータの増加と、お客様のニーズを先取りするために必要なスピードに対処するには、新しいツールが必要です。人間のスピードでは不十分です。しかしAIとMLが成熟したことで、容易にデプロイして分析のニーズに対応できるようになりました。AIやMLはすでに小売業のような業種のありかたを変えつつあります。予測分析は、予測やその他のタスクを提供することで、ビジネスが的確な意思決定を下せるようにします。

最近のGartnerの調査²によると、回答者がAIを採用するうえでの主な課題は以下のとおりです。

- スタッフのスキル (56%)
- AIのメリットと使い方の理解度 (42%)
- データのスコープまたは品質 (34%)

AIを導入する際の 主な課題

56%

スタッフのスキル

42%

AIのメリットと使い方の
理解度

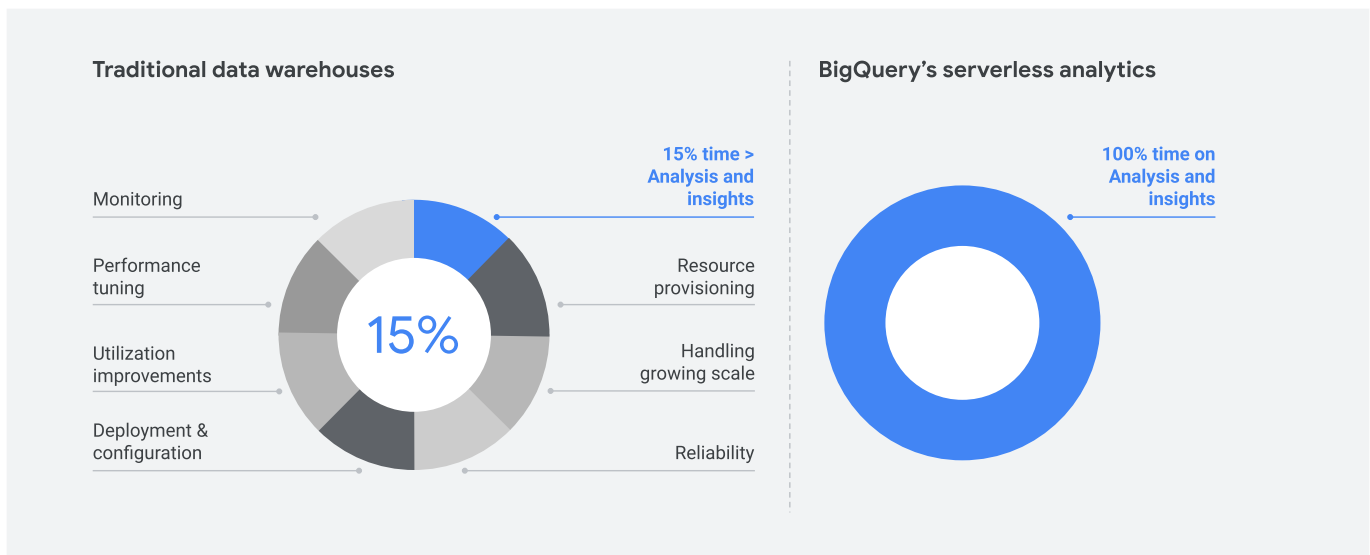
34%

データのスコープまたは品質

² Gartner, 「[Survey Analysis: AI and ML Development Strategies, Motivators and Adoption Challenges](#)」、Jim Hare および Whit Andrews、2019年6月。

Google Cloud のデータ プラットフォームでできること

新しいクラウド インフラストラクチャのオプションにより、最新のアナリティクス戦略を簡単にサポートできます。他のクラウドのデプロイと同様に、従来の非効率なアーキテクチャをパブリック クラウドにそのまま移植してしまう可能性があります。それを避けるには、データ ウェアハウスの総所有コスト (TCO) を考慮するとよいでしょう。従来のテクノロジーに多大な費用がかかっているが、それに見合うビジネスのアジリティが得られていない様子が把握できるからです。



BigQuery のサーバーレス モデルによる費用と時間の節約

ESG によると、BigQuery はオンプレミス ソリューションと比較して、全体的な費用を 3 年間で 52% 削減でき、AWS と比較しても 41% 削減できることがわかりました³。

Google Cloud は、データ インフラストラクチャの観点から、最新のクラウド データ ウェアハウスの要件に対応しています。[コンピューティングとストレージのレイヤ分離](#)は、ビジネスのアジリティを実現するために不可欠です。ストレージとコンピューティングが分離され、独立してオンデマンドでスケールできるようになれば、高価なコンピューティング リソースを常に稼働させておく必要はありません。

³ ESG Publication、「[The Economic Advantages of Migrating Enterprise Data Warehouse Workloads to Google BigQuery](#)」、2019 年 3 月。

Google は、最新のビジネスニーズに対応するデータ ウェアハウスを設計しました。このサーバーレスのコンセプトはシンプルですが、ギガバイト、ペタバイト規模のデータ分析に適用できる強力なものです。これは以下を意味します。

- BigQuery のオンデマンド分析エンジンは、特定のクエリのコンピュティング要件に基づいて、その場でプロビジョニングされます。お客様がノードやクラスタを定義する必要はありません。
- BigQuery は、オンプレミスや他のクラウド インフラストラクチャとは異なり、処理に必要なデータ量に応じてクエリのパフォーマンスを管理します。そのため、たとえば、2 週間あるテーブルに対するクエリを実行していない場合でも、パフォーマンスは常にクエリを実行しているテーブルの場合と変わりはありません。クエリの実行の最適化や、複数のデータコピーとウェアハウスの作成にかかる時間を減らすことができます。
- BigQuery は標準的な SQL インターフェースを使用しており、[Looker](#) との統合により可視化を実現します。
- BigQuery のようなクラウド データ ウェアハウスでは、これまでのように常にメンテナンスを行う必要がないため、データベース管理者 (DBA) は、計画とイノベーションに費やす時間を増やし、維持に要する時間を減少できます。
- ストリーミング データ分析は、移植性、言語選択、再利用可能なコードを提供するオープンソースの SDK を介して、バッチ処理とストリーム処理を統合する処理ツールである [Dataflow](#) で実現します。
- BigQuery を使用すると、データの移動やサードパーティ ツールを必要とせずに高度な機械学習タスクに取り組むことができます。売上予測や顧客セグメント作成などの予測分析をデータソースで直接実行できます。

「より多くのデータを収集することで、イノベーションが加速し、意思決定もスマートに行えるようになりました。Google Cloud Platform 上でリアルタイム分析を使用することで、わずか数週間でセールス コンバージョンを大幅に増やすことができました。Google Cloud Platform がなければ、クリック ストリーム データの件数を 100 倍に増やすことはできなかったでしょう。」

Zulily 製品管理担当ディレクター
Bindu Thota 氏

ガバナンスとセキュリティを主役に

Google Cloud の BigQuery は、効果的なデータ ガバナンス戦略と運用モデルを簡単に構築できるように設計されています。大規模運用にも対応できる組み込みのデータ保護機能、データを管理するためのツールとテクノロジー、業界標準への準拠、透明性に基づく信頼を提供します。ここでは、最新の Cloud Platform におけるデータ ガバナンスについて、重要なヒントを紹介します。

1. **検出と評価。**クラウドにおけるデータ ガバナンスのベスト プラクティスでは、どのようなデータ資産を持っているかを把握するために、初期のデータ検出と評価フェーズを設けることが求められます。
2. **分類と整理。**データ資産を評価し、その属性をスキャンして、後続の組織のために分類します。このプロセスでは、オブジェクトに機密データが含まれていて、特別な分類が必要かどうかも推測できます。
3. **データのカタログ化。**構造的なメタデータ、データ オブジェクトのメタデータ、ガバナンス指令に関連した機密性レベルの評価を含むデータカタログを維持します。
4. **品質の検証と評価。**データ コンシューマによってデータ品質の要件が異なるため、データ品質の期待値を文書化する手段と、データの検証とモニタリング プロセスをサポートする技術やツールを提供することが重要です。
5. **データアクセスの管理。**利用可能な資産へのアクセスを提供し、不正なアクセスを防止します。
6. **監査、追跡、モニタリング。**モニタリング、監査、追跡（誰が、いつ、何を、どのような情報で行ったかなど）を行うことで、セキュリティ チームはデータを収集し、脅威を特定して、ビジネスに損害や損失が生じる前に対処できます。
7. **データの保護。**漏洩したデータの内容が読み取られないようにするために、保存時の暗号化、転送時のデータの暗号化、データのマスキング、完全削除を含むデータ保護の手段を追加で導入します。



始めましょう

デジタルトランスフォーメーションの良いところは、技術的な変化とビジネスおよび文化的なニーズが並行して起きるところです。あるお客様によれば、BigQuery では使い慣れた SQL インターフェースが使われているため、少人数で作業量が過大になっているデータサイエンティストのグループからデータサイエンスの業務を移行し、他の多くの従業員に任せることができたということです。それにより、データサイエンティストが機械学習モデルのトレーニングのためにサイロ化したデータレイクを作成して、プロジェクトごとにデータを抽出してさまざまなリポジトリに配置する必要がなくなりました。従業員は時間を節約し、必要なフォーマットに直接分析情報を活用でき、データサイエンティストはトレーニングモデルの構築作業に集中できるようになりました。

Google は、AI や高度な分析をより身近なものにするツールを設計することで、データとユーザーの距離を縮めています。Google Cloud の BigQuery データウェアハウスを使用することで、ユーザーは大規模なデータセットをインタラクティブに分析できるため、分析情報を共有し、顧客分析を使用して、より良いプロダクトエクスペリエンスを構築できるようになります。BigQuery に組み込まれた ML 機能により、データサイエンティストやデータアナリストは、簡単な SQL ステートメントだけで巨大な構造化データセットまたは半構造化データセットに対する機械学習モデルを BigQuery 内に直接構築、デプロイできるようになります。さあ、どこから始めましょうか？

貴社のデータウェアハウスは、ビジネスのイノベーションをサポートする準備ができていますか？次への一歩を踏み出しましょう。Google Cloud のウェブサイトにてぜひアクセスしてください(google.com/data-warehouse-modernization)。





Google Cloud