Google Cloud

統一され、柔軟で
アクセスしやすい未来のデータ

テクノロジー企業とスタートアップ企業は、成功のために以下のことが必要であると学んでいます。

- データは社内全体で統合されるべきものですが、社内のみならず、サプライヤーやパートナー間においても同様です。そのためには、非構造化データを解放し、組織的サイロと技術的サイロを取り除くことが必要です。
- テクノロジー スタックには、オフラインのデータ分析から、リアルタイムの機械学習まで、さまざまなユースケースをサポートするだけの柔軟性が必要です。
- テクノロジー スタックはまた、どこからでもアクセス可能でなければなりません。さまざまなプラットフォーム、プログラミング言語、ツール、オープン スタンダードにも対応する必要があります。

データを最大限に活用することがなぜ競争上の優位性になるのか

ページ 02

イノベーションに注力できるよう、データをどう働かせるか

ページ 04

総合的なデータ ウェアハウス オプションを選ぶことが重要な理由

ページ 11

自信をもってデータ移行の道すじをつけていくには

ページ 12

Google Cloud

1章

データを最大限に活用することがなぜ競争上の優位性になるのか

データの重要性については誰もが理解していますが、自社のデータから革新的なビジネスと顧客の分析情報を抽出できている企業はほとんどありません。データを最大限活用するとはどういうことでしょうか？また、なぜそれは難しいのでしょうか。

データを最大限活用できるということは、データを使って製品やオペレーションに関する意思決定ができるということです。では、自問してみてください。自社の顧客の期待がどう変化してきているかをご存じですか？また、カスタマーエクスペリエンスの向上のために、データを活用できているでしょうか。課題という観点から、データエンジニアやサイエンティスト達が現在どのようなことに力を注いでいるかを自問してみてください。

幅広い市場開拓の意思決定とともに、革新的な製品の方向性とユーザーエクスペリエンスを推進するうえで、データは非常に重要です。データをうまく使いこなすことで、大きな競争上の優位性を手に入れることができます。これが、モダナイゼーションとオペレーションの規模を拡大し続け、現在および将来のデータコストを正当化し、組織の成熟度と意思決定能力を向上させるために、多くのテクノロジー企業やスタートアップが、大きな重圧のもと取り組んでいる理由です。

一方で、アクセス、ストレージ、一貫性のないツール、コンプライアンス、セキュリティといった課題もあり、深堀りしてデータの真の価値を引き出すことは困難です。

レガシー システムを引き継ぎ、そこに新しいシステムを統合しなくてはならない場合もあります。すべてのデータを 1 つのクラウドに置くべきでしょうか。それとも、複数のクラウドに分散させるべきでしょうか。従来は垂直統合されていた分析スタックをモダナイズし、水平スケーリングが可能なプラットフォームと連携させるにはどうすればよいでしょうか。

あるいは、データをリアルタイムで処理する代わりに、バッチ処理またはマイクロバッチ処理をしているかもしれません。結果として、アーキテクチャがオーケストレーション システムとスケジューリングにより複雑化され、競合や耐障害性に対するメンテナンスが必要になります。バッチ アーキテクチャの管理と保守にかかる運用のオーバーヘッドは高額で、データ レイテンシの点でも妥協しなくてはなりません。

全データへのアクセシビリティが低いことや、データが届くと同時に処理と分析を行う能力を持たないことは、大きなデメリットです。最新の技術スタックは、ストリーミング スタックである必要があります。つまり、データの規模拡大に対応し、利用可能な最新のデータを使用して、非構造化データを組み込んで理解しなければなりません。最先端の分析チームは、AI / ML でプロセスをテストおよび運用化して焦点をオペレーションからアクションに移しています。

Google Cloud

2 章

イノベーションに注力できるよう、データをどう働かせるか

データを働かせるとはどういう意味でしょうか。カスタマー エクスペリエンスの向上、新規顧客へのリーチ、収益の増加などがありますが、重要なのは、革新を可能にするということです。こういった成果を達成するためのデータ プラットフォームの選択に際し、原則を 2 つおすすめしたいと思います。

原則 1: 簡潔さとスケーラビリティ

おそらく、使用できるデータは現在大量にあることでしょう。また、それが急激に増大しているため、その量に対応しつつ ROI を維持または上昇させたいとお考えでしょうか。将来的にデータ量がどれくらい（たとえば 1 テラバイトなど）になるかを予測し、その量を処理できるようにシステムを設計するとします。もし成長がその予測を上回るようであれば、システムの全面的な移行を検討することになるでしょう。または、予測される成長に合わせてスケーリングできるデータ ウェアハウスを選択しても、増大する処理ニーズにより管理が複雑化することも考えられます。

小型のシステムの方が、一般的にシンプルです。しかし、昔のように、使いやすいシステムと優れたスケーラビリティのあるシステムのどちらかを選ばなければいけないということはないのです。サーバーレス アーキテクチャは、クラスタ管理の必要性を排除します。そして、コンピューティングとストレージの両方が大規模でも処理可能であるため、データサイズが技術的容量を超えてしまうという心配は不要です。

簡潔さとスケーラビリティ両方の観点から、サーバーレスのデータ プラットフォームをご提案しています。また、ソフトウェアのインストール、クラスタ管理、クエリの微調整などが必要となるオプションは避けることをおすすめします。

原則 2: アジリティと継続的なコスト削減

コンピューティングとストレージを組み合わせたデータ管理システムでは、たとえ不要であっても、増大するデータ量に対応するためにコンピューティングをスケールアップしなくてはなりません。これには費用がかかるため、妥協をして、分析ウェアハウスには最新 12 か月分のデータのみ保存することになるかもしれません。また、データの即時のユースケースがないために、そのデータを含めないことを選択する場合も考えられます。その結果、データがないために後で仮説をテストできなくなり、開始するには新しいパイプラインが必要となるかもしれません。

その他のシステムでは、コンピューティングとストレージをそれぞれ別にスケーリングし、料金を支払うことができても、手動でのクラスタの設定、スケーリング、最適化が必要になる可能性があります。インフラストラクチャ管理をできる限り減少させるために、信頼性、パフォーマンス、組み込みデータ保護を強化したサーバーレスでマルチクラウドのデータ ウェアハウス（たとえば BigQuery など）を検討してみてください。

コストと管理のほかに、アジリティも重要です。データが変更された場合、それに気づいてから対応するまでどのくらい時間がかかるでしょうか。お使いのソフトウェアやツールの新しいバージョンが出た場合、新しい機能を使いこなすまでどれくらいの時間がかかるでしょうか。アジリティを高めるには、より少ない操作で、さまざまなワークロードに適用できる柔軟なツールを選択することが重要です。

Redshift といったシステムのクエリは、効率を上げることを目的とした最適化が必要です。これは可能なテストの量を制限するため、問題があると疑われる場合にのみデータを抽出して pull することになるかもしれません。コンピューティングとストレージの分離の欠如による妥協と、一方でのデータ ウェアハウスを最適化する必要性により、制限が大きくなります。

BigQuery などでは、クエリを事前にプランニングする必要も、データセットをインデックスに登録する必要もありません。ストレージとコンピューティングを分離すれば、クエリの費用を心配することなく、データの取り込みを行えます。また、データサイエンティストは、クラスタやデータ ウェアハウスのサイズを気にすることなく、アドホック クエリで新しいアイデアをテストすることができます。

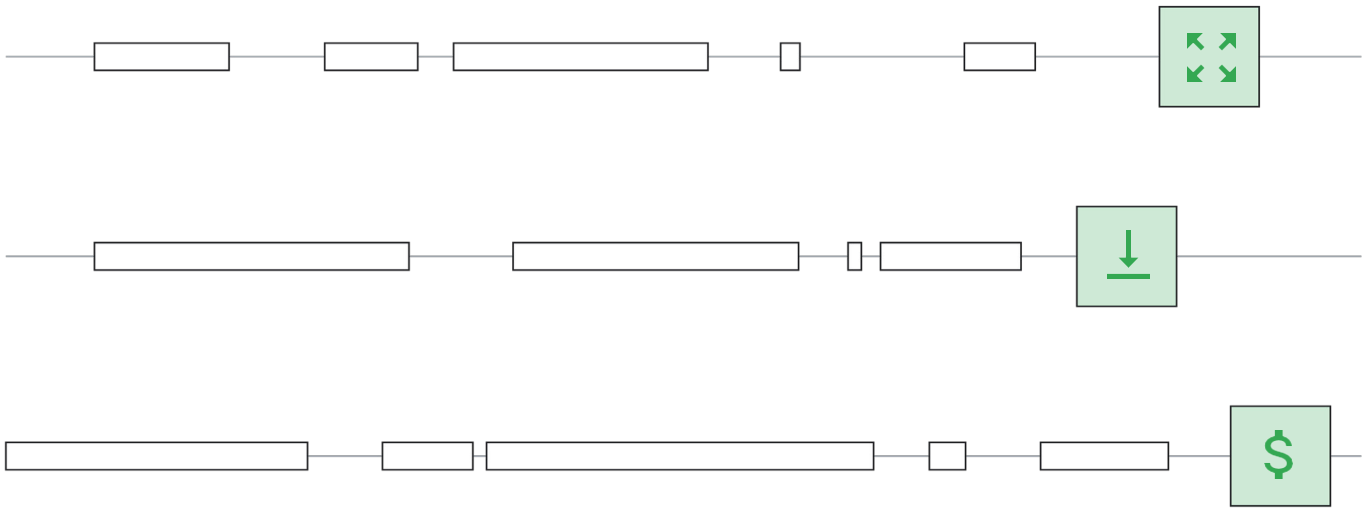
簡潔でスケーラブル、かつ柔軟で費用対効果の高いプラットフォームが、どのようにイノベーションの実現に役立つかを見てきました。次は、それを実現するために、データがどう役に立つかを考えてみましょう。

Google Cloud

リアルタイムでデータドリブンの意思決定をする

ビジネス オペレーションのスピードはますます加速しています。顧客の期待もまた変化しています。かつては取引の調整や返品の承認をするのは3 日以内でよかったのに、今や即座の回答が求められます。より高速でタイムリーな意思決定によりストリーミングの需要はさらに高まっています。

リアルタイムでデータをキャプチャし、そのデータをビジネスチームが低レイテンシでクエリできるようにすることが求められます。また、ストリーミング パイプラインは、スケーラブルで復元性があり、管理の負担も少ないものにしたいところです。これこそが、チームがビジネスのスピードに合わせリアルタイムに反応できる唯一の方法なのです。BigQuery は当然、ネイティブにストリーミング データの取り込みをサポートし、そのデータを SQL ですぐに分析できるようにします。BigQuery の使いやすい Streaming API に加え、Dataflow を使えば、季節性があったり突発的であったりするワークロードを、コストを過剰に費やすことなく管理できます。



データサイロの解消

多くの組織がサイロを作ってしまう原因は、チームがそれぞれ独自のデータを抱え込み、部署や事業部門ごとに別々にデータを保管していることにあります。このため、部門をまたぐ分析を行う場合、このようなサイロを解消する方法を探さなくてはなりません。たとえば、抽出（ETL）パイプラインを実行してデータを取得し、使用しているデータ ウェアハウスに組み込むなどです。しかし、データを所有している部門にとって、多くの場合、パイプラインを維持するメリットはほとんどありません。結果、パイプラインは次第に古くなり、データは陳腐化して有効性も失われます。

今日では、多くの企業が組織的サイロ以外に、部門のニーズ、能力の整合、規制圧力に応じて、マルチクラウド戦略を採用しています。こうした企業はしばしば、オンプレミスにある旧来のデータレイクとデータ ウェアハウスへの投資という現実も抱えていることが多いのです。現在のマルチクラウド、ハイブリッド クラウドの現実は、サイロ化されたデータの管理とアクセスについてもう一段階上の改善を必要とします。

データ ファブリックやデータメッシュと呼ばれる、共通のコントロール ペインを持つ分散型ウェアハウスに移行することで、部門、クラウド、オンプレミス システムをまたいで、高品質なデータにアクセスする能力が高まります。これにより、商品の販売状況や顧客行動などのビジネスの課題を解決でき、その場でのデータクエリが可能になります。

BigQuery は、こうしたデータメッシュの技術的基盤を提供します。組織内の誰がデータを所有していても、組織全体のユーザーがデータアセットとインサイトにアクセスし、これを管理、保護、共有できます。たとえば、BigQuery にすべてのデータを取り込み、再利用可能な関数とマテリアライズド ビューを提供できます。さらに、データを移動せずに ML モデルをトレーニングすることも可能です。これにより、非技術分野のエキスパート（権限のあるパートナーやサプライヤーを含む）がスプレッドシートやダッシュボードのような慣れ親しんだツールを使用してデータに容易にアクセスして、SQL を使ったクエリを実行できます。

これは「ハブ アンド スポーク」に例えることができます。BigQuery はデータを含むハブです。スポークは報告ツール、ダッシュボード、ML モデル、ウェブ アプリケーション、レコメンデーション システムなどです。これらは、データをコピーすることなく、BigQuery からデータをライブで読み取ります。たとえば、Looker では、データを可視化してユーザーの日々のワークフローに統合できます。この方法により、データのユーザビリティ、セキュリティ、そして品質を同時に向上できます。

すべてのデータへのアクセスを簡略化

従来は、非構造化データと半構造化データにはデータレイクが最適で、構造化データにはデータ ウェアハウスが最適でした。この分離によって技術的なサイロが生まれ、形式の相違をまたぐことが難しくなりました。管理が容易でコストが低いため、すべてのデータがデータレイクに保存されるようになります。分析情報を抽出するために分析ツールを使用する場合は、データはウェアハウスに移動されます。

最近よく使用されるようになった「レイクハウス」は、これら 2 つを、すべてのデータタイプに適合する統合環境に結合します。BigQuery をデータ ウェアハウスとデータレイクの両方として使用できるのです。BigQuery の Storage API を使えば、ストレージに直接アクセスして、通常データレイクに関連づけられているワークロードを強化できます。データは BigQuery に信頼できる単一の情報源として保存できるため、作成して維持するコピー数は少なくなります。代わって、ダウンストリーム処理を SQL 変換で行うことができます。変換は論理ビューで保存され、データを移動させる必要はありません。

使い勝手の良さはやはり重要です。30 分や 3 時間ではなく、30 秒でクエリから結果を取得できるのであれば、データを使用した意思決定の可能性が広がります。

AI / ML でワークロードを迅速にテストし運用化する

データ サイエンティストはどのくらいの速さでテスト実行できるでしょうか。実際のユーザーでテストを評価するために、開発を停止してモデルを運用化する必要があるかもしれません。データ サイエンティストは、過去のデータを使用してモデルを開発し、反復処理を行い、そのモデルをエンジニアに引き継ぎます。エンジニアは、多くの場合、モデルを本番環境システムに組み込むために完全に書き換え、A/B テストを行います。そして待機してからモデルに反復処理を施し、再度運用化します。このサイクルは停止と再開を繰り返し少しずつ進められます。数多くのコードが書き直され、チーム間で必要なすべての調整が行われる中でエラーも発生します。この方法はアジリティに欠けるため、データ サイエンティストは想定されるほどのテストを行うことができません。定型業務になるまでにかかる時間どころか、プロジェクトがどのくらいの長さになるか、それが成功するのかという予測が困難になります。これを解決するには、データ サイエンティストに強力かつ使い慣れたツールを提供する必要があります。Vertex AI Workbench を使うと、データ サイエンティストは Jupyter Notebook で効果的に仕事をしつつ、集中トレーニング、迅速なテスト、高速なデプロイを行うことができます。

データに基づく差別化を真剣に検討しているのであれば、収集しているデータから最高の価値を引き出したいと考えることでしょう。そのためには、データ サイエンス チームの生産性をできる限り高める必要があります。簡単なことでも時間がかかったり、難しすぎたりすることを理由にモデル構築の機会を逸してはなりません。

事前構築およびローコード モデルの品質は非常に重要です。Vertex AI の AutoMLは、ノーコード環境で利用可能な最高水準の AI モデルを提供します。これにより、高速なベンチマークと優先順位付けが可能になります。自社データでエンティティの抽出やVertex AI Matching Engine といった事前構築モデルを使用することで、分類や回帰にとどまらず、データからの価値の創造を大いに促進できます。

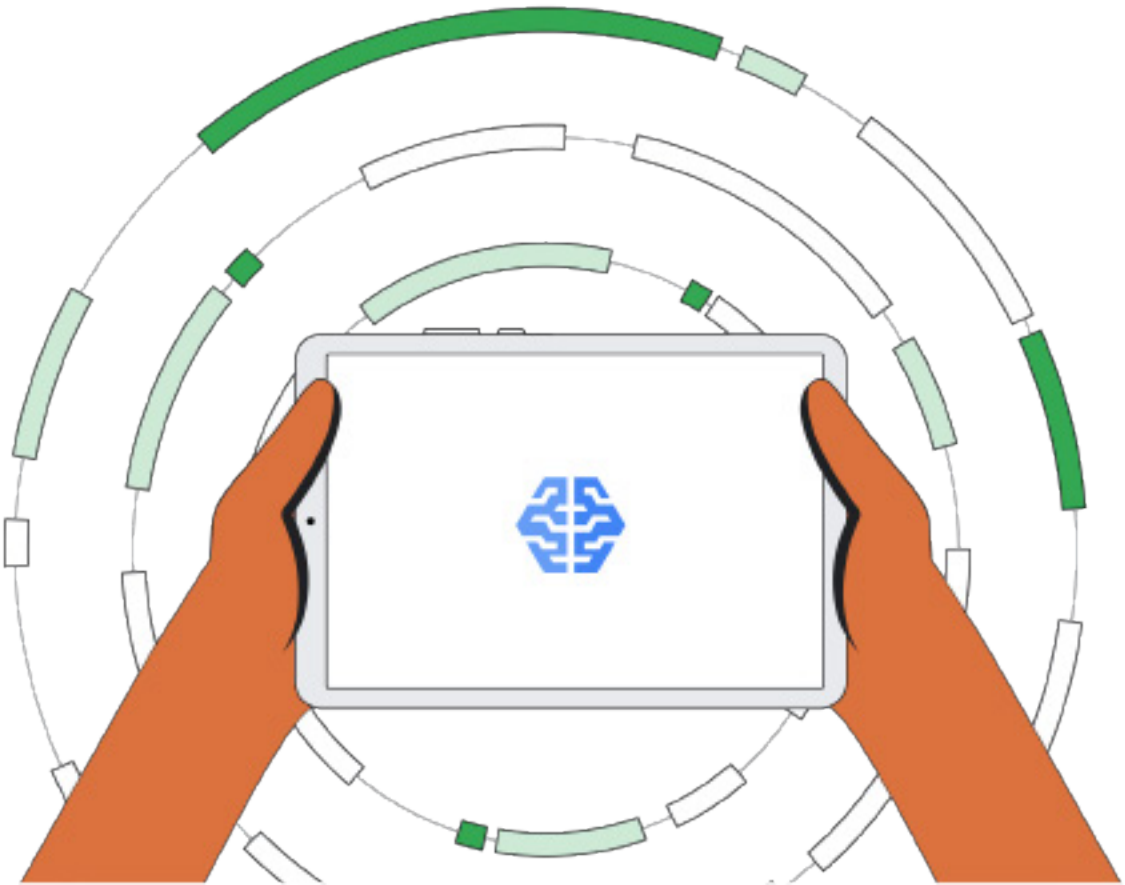
データ アジリティを保つためには、早い段階で頻繁にエンドツーエンドのテストを行うことが重要です。Vertex AI Pipelines を使えば、テストの履歴を振り返り、ベンチマークやエンドポイントの比較、シャドウモデルによる A/B テストを行うことが可能です。 コードはコンテナ化されているため、同じコードを開発システムと本番環境システム間で使用できます。データ サイエンティストは Python で作業し、プロダクション エンジニアは完全にカプセル化されたコンテナを取得します。両チームは、Vertex AI Prediction を使用してモデルの運用化について標準化できるため、迅速に動くことができます。

BigQuery ML により、ドメイン エキスパートは、カスタムモデルをトレーニングしてアイデアの実現可能性を頻繁にテストできます。このときに使用するのは SQL のみで、従来のデータ サイエンス ツールの知識は不要です。 この結果、本番環境に近いシステムでテストを行うことができ、実現可能性の調査にかかる時間も数か月から数日に短縮されます。BigQuery の ML モデルを Vertex AI にデプロイすることで、先ほど説明したようなメリットが得られます。Looker を使用することで、整合性のあるデータモデルをすべてのデータ上に作成し、LookML を使用してデータのクエリを実行できます。これにより、組織の全ユーザーが、読みやすいレポートとダッシュボードを作成して、データのパターンを調べることができます。

本番環境で真の価値を推進するには、システムがデータを取り込み、処理し、提供できる必要があります。また、機械学習は顧客のコンテキストに応じてカスタマイズされたサービスをリアルタイムで推進する必要があります。一方で、継続的に稼働している本番環境のアプリケーションは、モデルについて、常時再トレーニング、デプロイ、セキュリティを確認する必要があります。受信データの場合、データの前処理と検証を行い、品質問題がないことを確認する必要があります。その後、特徴量エンジニアリングと、ハイパーパラメータ調整によるモデルのトレーニングを行います。

これらの多重フェーズの ML ワークフローを容易にオーケストレーションして管理し、確実に繰り返し実行するためには、データ サイエンスと機械学習のインテグレーションが不可欠です。MLOps ツールと自動化ワークフローにより、迅速な継続的デリバリーが可能になり、本番環境までのモデルの管理が簡素化されます。 Google のすべての AI プロダクトには、抽象化レイヤにかかわらず単一のワークフローと語彙があります。そして、カスタムモデルと AutoML モデルは同じ形式と技術基盤を使用しているため、簡単に交換できます。

たとえば、ライブの膨大な量のデータ ストリームに異常検出を適用し、不正行為に対処するにはどうすればよいでしょう。正しいアプローチとしては、サンプルデータ ストリームを生成して一般的なネットワーク トラフィックをシミュレートします。そして、そのストリームを Pub/Sub に取り込んでから、DLP を使用して個人情報（PII）をマスクしたうえで、BigQuery ML の k 平均法クラスタリングを使用する異常検出モデルを BigQuery で作成およびトレーニングします。その後、ライブデータにモデルを適用し、Dataflow を使用してリアルタイム検出を行います。また、Looker を使用してダッシュボード、アラート、アクションを作成することで、確認されたイベントに対応します。



Google Cloud

3 章

総合的なデータ ウェアハウス オプションを選ぶことが重要な理由

BigQuery と Redshift についてお話してきましたが、利用できるデータハウスのオプションはこれだけではありません。Snowflake や Databricks など、主要な 3 つのクラウド上で動作するデータ分析プロダクトは他にもあります。では、BigQuery を選んだ場合、クラウド ロックインは問題になりえるでしょうか。

まず念頭に置くべきは、BigQuery は、Google Cloud に保存したデータだけを分析するものではないということです。BigQuery Omni は、Amazon S3 や Azure Blob Storage にあるデータを Google Cloud Console からシームレスにクエリする機能を提供します。

けれども実際には、Snowflake や Databricks を使用している場合、AWS から Google Cloud（あるいはその逆）に移行する切り替え費用の方が低いのです。しかし、別のデータ ウェアハウスに移行する場合の費用はどうでしょう。Snowflake から BigQuery へ、または Databricks から EMR に移行する場合を考えてみましょう。シナリオは異なりますが、切り替え費用はやはりかかります。

どのシナリオにおいても切り替え費用は発生するので、最終的には長い目で見て一番役に立つツールやプラットフォームを選ぶとよいでしょう。プラットフォームの差別化機能、現在のコスト、将来的なイノベーションの速度などを考慮して選択します。Snowflake を選んだ場合、データ ウェアハウスに特化した会社が、その分野でより迅速なイノベーションを起こすと確信しているということです。BigQuery を選んだ場合、多くのデータや AI 技術を発明したことで知られる会社が、プラットフォーム全体でイノベーションを起こし続けることを期待していると言えます。

私たちは、革新的で統合されたプラットフォームが、イノベーションの優れたフライホイール効果を生み出すと確信しています。Google Kubernetes Engine（GKE）のようなマネージド サービスを提供することで、コンテナ イメージの読み込みが速くなります。これにより、サーバーレス Spark の動作も向上します。サーバーレス Spark は BigQuery 上のデータを操作できるので、BigQuery の価値も高まります。 個別のプロダクトよりもプラットフォームに重きを置く方が、より大きなフライホイール効果を生み出すでしょう。

Google Cloud

4 章

自信をもってデータ移行の道すじをつけていくには

データ移行にはどのくらいの期間が必要でしょうか。6 か月？2 年？どのくらいの労力が必要となるでしょうか。またそれだけの価値があるでしょうか？

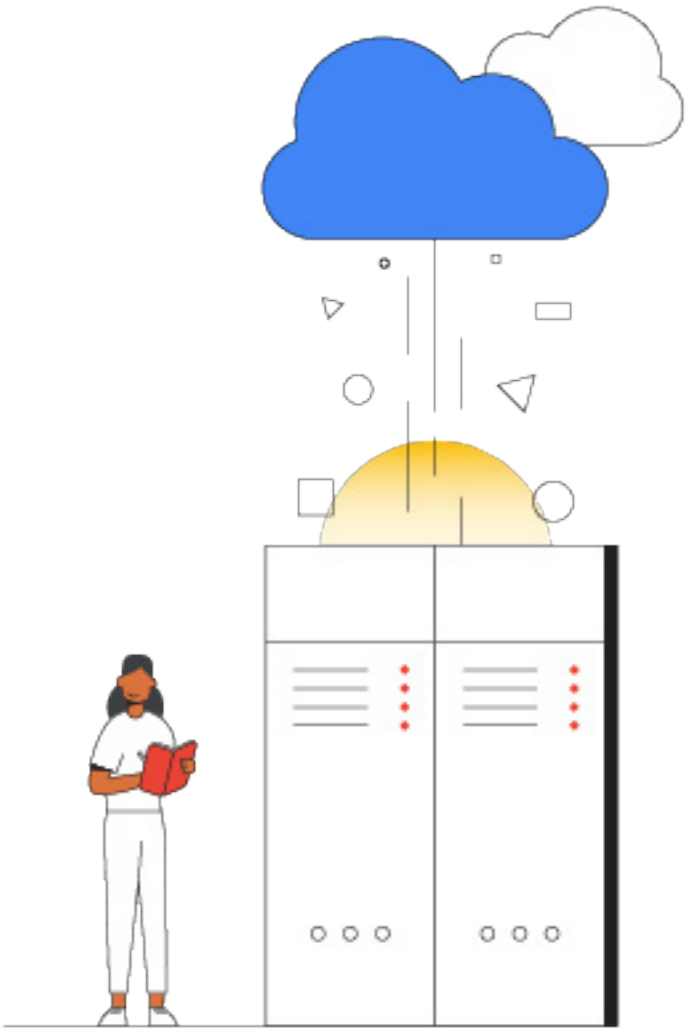
クラウド間の移行の方が、オンプレミスからクラウドへの移行より簡単と言えます。通常、オンプレミスの方が技術的奥行きがあるからです。とにかく、「いかに早くイノベーションを起こせるか」というような目標に集中することが重要です。

現在まだ行えていないけれど、ぜひやってみたい革新的なことを思い浮かべてください。そして、それらを実行するために必要な新しいプロジェクトやデータ転送などをセットアップしてみましょう。新しいユースケースの構築や、必要なデータソースのミラーリングのお手伝いをぜひさせてください。しばらくの間は、多くのユースケースがオンプレミスで動作しながら、オンプレミス環境や他のクラウド プロバイダからリアルタイムまたはバッチでミラーリングされたデータを活用するという、ハイブリッド環境になることと思います。

2 つめに考慮すべきはコスト面でしょう。実行中の高価な Teradata インスタンスを見てみましょう。お客様が BigQuery に切り替えることで、コストを半分に削減されるのを見てきました。自動評価ツールと自動 SQL トランスパイラがスクリプトの大部分を変換してくれるため、移行も以前に比べてずっと簡単です。クライアントは Teradata と会話していると思っていても、実は BigQuery と会話している、というような仮想化の方法も用意しています。すべてをシャットダウンすることなく、移行を支援できる方法はたくさんあります。こうした移行ツールを使って、高価な Teradata や Hadoop のワークロードからの移行が可能です。

3 つめの考慮項目は、SAP、Salesforceシステム、Oracle などの ERP システムを検討することです。サプライチェーンを最適化し、見込み顧客のスコアリングを行い、不正行為の検出をしたいと考える場合、分析ワークロードを ERP システムに接続できるようにしておくことが大事です。サード パーティーのコネクタを使って、これらのシステムからデータを取得し、そのデータを使ってクラウド上で最新の AI 駆動型ユースケースを構築できます。

これらを行う順序は、ご自身のユースケースに応じて決まります。スタートアップ企業であれば、イノベーションから始めて、次に費用の最適化、最後に既存のパイプラインやコネクタを活用できます。事業がサプライチェーンに大きく依存している場合、ERP コネクタから始めることになるでしょう。この 3 つをどのような順番で行うにせよ、貴重なデータのかなりの部分をクラウドに移行していることがおわかりになると思います。残っているものを見て、移行する価値があるかを検討しましょう。多くの場合、答えはノーです。本当に必要な 70〜80% のワークロードを移行させた後は、厳しい決断を下さなければならないのです。残りの 20〜30% は移行する価値があるのか、それとも書き換えや別のタスクを検討すべきかという決断です。今あるものをすべてそのまま移行するとなると、オンプレミスで抱えていた技術的負債を、新しいクラウド環境においても再現することとなり、データの価値に注力することが難しくなります。



Google Cloud

次のステップに進む準備はできていますか？

今日は、データを使いこなす方法やその意義、またクラウドのデータ ウェアハウスへ移行する際に直面しうる検討事項についてお話ししました。

著しい優位性を得るためにインサイトを使用し、会社のコストを削減し、データと AI の利用を最適化して生産性を向上させるために Google Cloud がどう役立つかに興味をお持ちの場合は、ぜひご連絡ください。

お問い合わせ方法

関連情報

- ご自身の組織がどの種類のデータ処理ユニットかを知る
- 組織の種類に応じたアナリティクス データ プラットフォームの構築方法については、こちらのホワイトペーパーをお読みください。