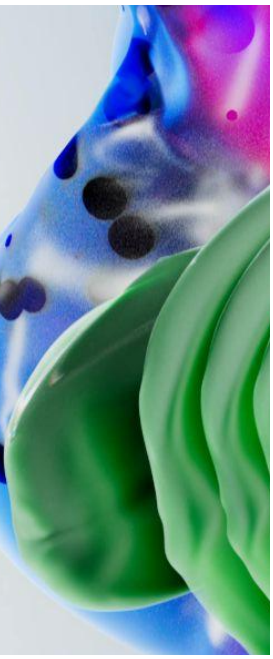
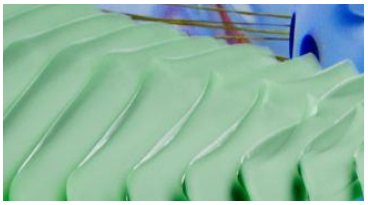
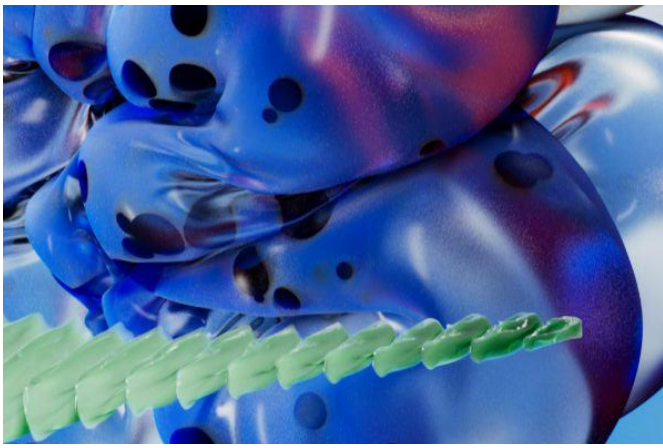
Google Cloud

データレイク移行の実践ガイド







データの可能性を引き出して AI トランスフォーメーションを加速させる

作成者: Google Cloud、グループ プロダクト マネージャー
Dana Soltani

目次

エグゼクティブサマリー

データレイクをクラウドに移行することを計画している場合、このガイドを活用して成功に導くことができます。過去 5 年間にわたり数え切れないほどのエンタープライズデータレイク移行を観察して得た見識を生かし、**移行プロセスを最適化する方法についてまとめました**。

「なぜ」を理解する

データレイクの進化と、クラウドベースのアプローチでモダナイズすべき説得力のある理由について探ります。従来のシステムが抱える課題と、スケーラブルで費用対効果に優れた革新的なクラウド環境に移行する利点について学びます。

情報に基づいた経済的意思決定

コスト最適化戦略や、大幅なコスト削減を達成した企業の実例など、クラウドの移行の経済的側面について詳しく調べます。移行戦略をビジネスの SLA に適合させ、ニーズに合った適切なアプローチを選択する方法を学習します。

移行の計画

検出と評価から実行と最適化まで、このガイドでは 5 つのフェーズからなる明確な移行プロセスを紹介します。各段階における詳細なチェックリストと考慮すべきポイントにより、構造化された効率的な移行プロセスが確保されます。

万全な移行体制

データ、メタデータ、ワークロード、ガバナンス、ワークフローなど、データレイクのさまざまなコンポーネントを移行するためのベストプラクティスを採用します。クラウドでデータの完全性を確保して、ダウンタイムを最小限に抑え、パフォーマンスを最適化する方法を学びます。

Google Cloud のサポートを活用

Google Cloud により提供されるサービス、ツール、専門知識の包括的なスイートについて調べます。これには、アカウントチームによる個別のサポート、自動移行ツール、コンサルティング サービス、技術的な支援、広範なパートナーエコシステムへのアクセスが含まれます。

このガイドのガイダンスとベストプラクティスに従うことにより、データレイクを Google Cloud に自信を持って移行し、スケーラビリティ、高い費用対効果、最先端のデータ分析を実現できるだけでなく、AI / ML 機能へのアクセスが可能になります。

1 章

データレイクの進化

組織がデータを収集、保存、分析する方法は、長年にわたって進化してきました。データベースからデータ ウェアハウス、データレイクに至るまで、この進化はインターネット、ビッグデータ分析、そして今では AI など、他のイノベーションと歩調を合わせて進んできました。現在の状況を理解するため、データレイクがどこまで進化してきたかを振り返ってみましょう。

すべてはビッグデータから始まった

ビッグデータの初期、従来のデータウェアハウスは、生成される情報の量と種類の増加に対応するのに苦労していました。構造化データ向けに設計されているため、非構造化データや半構造化データはほとんど活用されていませんでした。2010 年代初頭に登場したデータレイクは、このギャップを埋めることを目指していました。データレイクは、構造や使用目的に関係なく、元データをネイティブ形式で保存する巨大なリポジトリであり、あらゆる種類のデータを収集して分析する新たな機会を組織に提供します。

Hadoop は、この分野で先駆的な役割を果たしました。Google MapReduce の論文[1]によれば、Hadoop 分散ファイルシステム（HDFS）は大量のデータを保存でき、処理フレームワーク（MapReduce）はこのデータを並行して分析できます。その後、より高速で効率的な処理エンジンとして Apache Spark が登場しました。組織が所有および運営するデータセンター内でオンプレミスのデータレイクを構築し、維持し始めるまで、それほど時間はかかりませんでした。そこでは、Hadoop や Spark などのアーキテクチャを活用して、安全かつ費用対効果の高い方法で専有データを保存および処理することができ、目的を果たすことができました。

従来のデータレイクには課題が伴っていた

オンプレミスのデータレイクではコントロール性とセキュリティが高まりますが、組織がデータアセットを最大限に活用することは困難です。次のような課題があります。

スケーラビリティの制限

ハードウェア インフラストラクチャの物理的な容量による制約があるため、増大するデータ量と処理需要に対応するためにスケールアップするには費用と時間がかかりました。これによりパフォーマンスのボトルネックが発生し、データやワークロードの急増に対応する能力に限りがありました。

高い初期費用とメンテナンス費用

オンプレミスのデータレイクを構築および維持するには、ハードウェア、ソフトウェアライセンス、IT インフラストラクチャへの多額の先行投資が必要でした。そこにハードウェアのアップグレード、ソフトウェアアップデート、セキュリティパッチの適用などの継続的なメンテナンスが加わり、総所有コストが増加していました。

管理オーバーヘッド

ハードウェアのプロビジョニング、ソフトウェアのインストール、構成、パフォーマンス調整、セキュリティ管理などのタスクには、専門知識と専用のIT リソースが必要でした。これにより、データ分析や分析情報の生成など、より高い価値を生み出すアクティビティに費やす時間が減っていました。

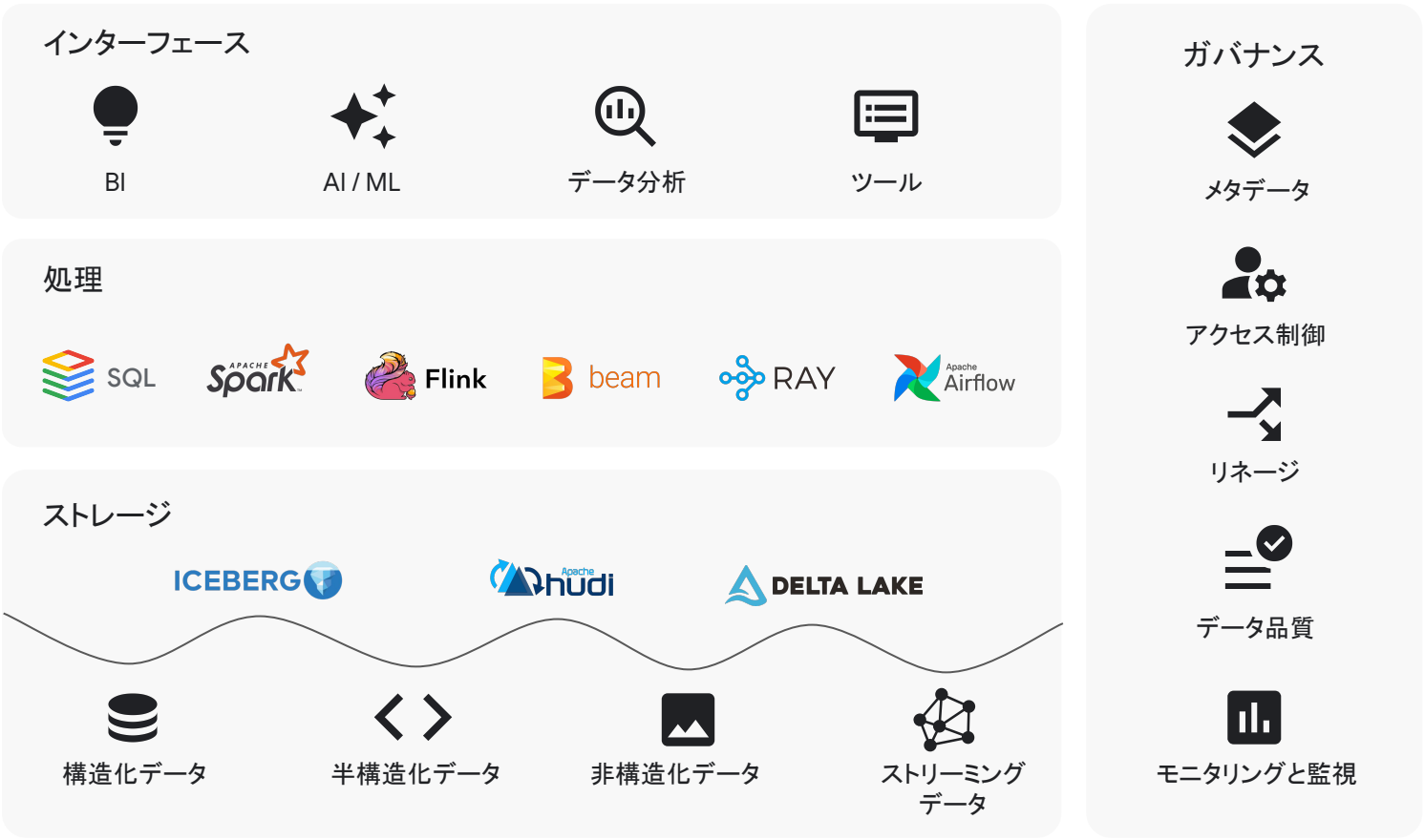
イノベーションへのアクセスが限られる

サーバーレスコンピューティング、AI / ML プラットフォーム、高度な分析ツールなど、分析、AI、クラウドネイティブサービスの最新の進歩は、オンプレミスのデータレイクと互換性がありませんでした。これによりイノベーションが妨げられ、データからビジネス上の優位性を引き出すことが困難になっていました。

1. Dean, J および Ghemawat, S、2004 年 MapReduce: Simplified Data Processing on Large Clusters

クラウドがデータレイクにもたらした変化

クラウド コンピューティングにより、データレイクが新たなレベルに引き上げられました。現在のクラウドベースの最新のデータレイクは、従来のデータレイクのデプロイでしばしば直面する大きな課題を解決できるよう設計されています。オンプレミスのストレージとコンピューティングの静的でサイロ化された環境と比較すると、これらは弾力性と短命性が高く、組織がデータから新たな機会を引き出せるように設計された 4 つのレイヤで構築されています。



ストレージ

ベースとなるレイヤでは、構造化データ、半構造化データ、非構造化データがファイルとして汎用クラウド ストレージに保存されます。Apache Parquet、Avro、ORC が最も一般的なファイル形式です。SQL でデータをクエリするには、Hive テーブル形式がよく使用されていますが、Apache Iceberg などのより新しい形式に取って代わられつつあります。Iceberg には、Atomicity（原子性）、Consistency（一貫性）、Isolation（独立性）、Durability（永続性）（ACID）トランザクション サポートに加えて、ペタバイト規模のテーブルの効率性と、スキーマおよびパーティションの進化、タイムトラベル、マテリアライズドビューなどの高度な機能が備わっています。

処理

ストレージとコンピューティングの分離はオンプレミスのデータレイクから引き継がれ、クラウドで強化されます。クラウドでは、処理はエフェメラルであり、コンピューティング リソースは費用を最小限に抑えるために必要な場合にのみプロビジョニングされます。ユースケースに応じて、Apache Spark、Flink、Ray などの複数の処理エンジンを使用して、データはバッチまたはリアルタイムで処理されます。

ツールとインターフェース

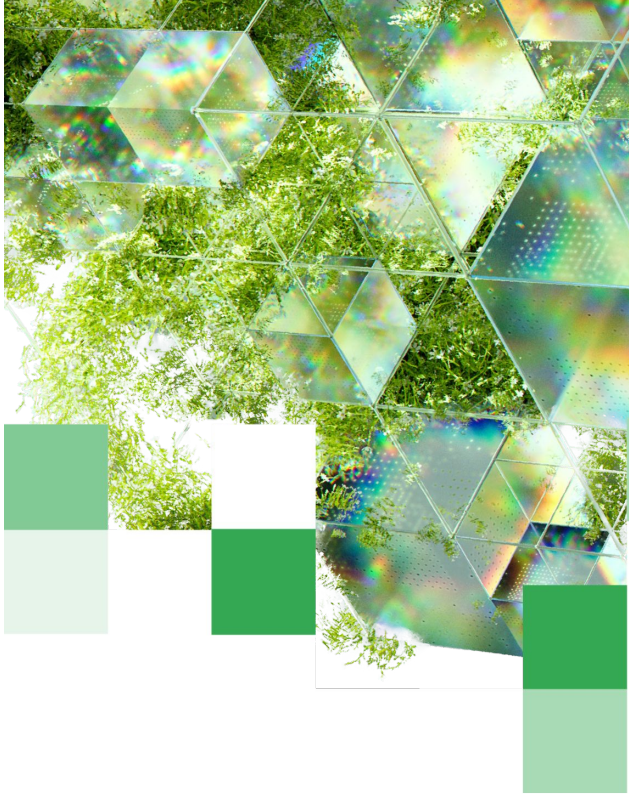
これは、さまざまなユーザーがデータレイクと接触するレイヤであり、ユースケースはアドホック分析から、ビジネス アプリケーションにおける継続的な予測のためのML パイプラインの構築まで多岐にわたります。ユーザーは、BI ツール、データサイエンス ノートブック、SQL ワークベンチ、API などを通じてデータレイクを操作します。

ガバナンス

データレイクには、厳格な管理が必要な機密データやミッション クリティカルなプロセスが含まれることがよくあります。メタデータ管理、データ品質管理、リネージのトラッキング、アクセス制御などの多様な機能により、運用コストを下げられるだけでなく、大きな被害をもたらすデータ漏洩のリスクを軽減できます。

この構造により、分析が価値を生み出すまでの時間と、分析情報が組織全体に行き渡るまでの時間が短縮されます。これは、現在でも多くの組織が使用しているオンプレミス データレイクの費用対効果を大幅に上回ります。

クラウド データレイクはデータと AI の可能性を引き出す



多くの組織は、現在のデータインフラストラクチャがAI 向けに構築されているわけではないことに気づき始めています。従来のデータレイクでは、 AI の基盤となる大量の非構造化データを効率的に保存および処理することができません。多くの場合、ウェアハウス、レイク、クラウドにわたりデータサイロは残ります。そして多くの組織には、AI モデルの構築とサービングに必要なコンピューティング リソースが不足しています。

これらの制限を克服するため、従来のデータレイクの進化における自然な次のステップはモダナイゼーション、つまりデータレイクをクラウドに移行し、アーキテクチャをモダナイズしてすべてのデータを1 つの統合プラットフォームにまとめることです。

クラウド データレイクにより、組織は次のことが可能になります。

データドリブンな意思決定の迅速化

クラウドでは、データマネジメントとデータ処理が統合されているため、データを活用して競争上の優位性を高め、エンドカスタマーにとっての価値を高めることが容易になります。

AI の導入の加速

構造化データと非構造化データ全体にわたるテスト、モニタリング、ガバナンスのためのツールに簡単にアクセスでき、AI によるイノベーションを加速させるための ML モデルも利用可能です。

複雑さの軽減

クラウド データレイクでは通常、多層ストレージ、処理エンジン、ML および AI モデルのさまざまなオプションが利用可能であり、データとAI の活用を簡素化し、費用を削減するのに役立ちます。

ガバナンスの強化

クラウドでは、データと分析情報の高い品質、信頼性、規制の遵守を確保しやすくなります。

Definity は費用削減とパフォーマンス向上のために BigQuery に移行

150 年を超える歴史を持つカナダの大手P&C 保険会社であるDefinity は、急速に進化する業界の需要に対応するには、データ インフラストラクチャをモダナイズする必要があることを認識していました。以前のオンプレミスの Cloudera プラットフォームでは、十分にスケーリングとイノベーションを行い、データとAI の力を最大限に活用することができませんでした。Definity はわずか 10 か月で Google Cloud に移行し、BigQuery 上にデータストアを構築し、Vertex AI を分析および AI プラットフォームとしてセットアップしました。1 年間のインフラストラクチャおよび運用コストを 30% 以上削減して、デプロイ時間を63% 短縮し、インフラストラクチャのセットアップを10 倍高速化しました。

> 当社は、スケーラビリティ、費用対効果、Vertex AI への接続性を考慮に入れて、BigQuery を選択しました。これは、当社の統合データ分析プラットフォームのニーズに最適なソリューションです。BigQuery では、あらゆるデータタイプがサポートされているため、データ形式に関係なく、エンタープライズ データから最大限の価値を引き出すのに最適なプラットフォームです。この結果、データ インフラストラクチャの管理ではなく、 AI / ML によるイノベーションに集中できるようになりました。"
>
> Definity、
> データ プラットフォームおよびクラウド
> エンジニアリング、アソシエート VP、
> **Nitin Mathur 氏**

2 章

データレイク移行の経済性

データレイクをクラウドに移行すると、さまざまな点でメリットが生まれます。クラウドモデルへの移行によって大幅な費用削減が実現するだけでなく、効率性の向上、分析情報のすばやい入手、ビジネス上の意思決定の改善、最新の AI イノベーションの活用が可能になります。そして、これらすべては収益機会の増加と競争上の優位性につながります。

20～62%
の費用削減

組織がセルフマネージドまたはオンプレミスの代わりにマネージドクラウド データレイクを使用した場合に実現[2]

80～90%
の費用削減

レイクハウス アーキテクチャに移行し、マネージド クラウド サービス上でデータレイクとウェアハウスを組み合わせることによって実現[2]

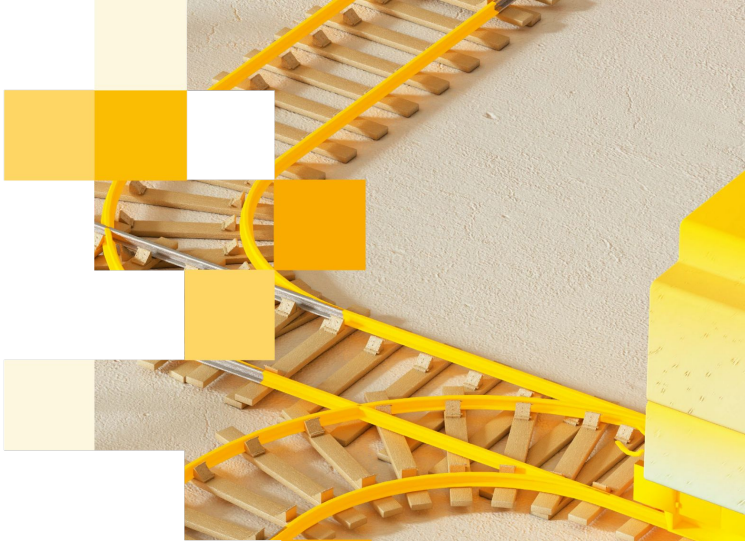
費用は変化の主な原動力

オンプレミスのデータレイクが経済性がますます低下していることは間違いありません。まず、インフラストラクチャの構築と維持にかかるコストのため、CapEx と OpEx が高くなります。さらに、増え続けるデータの量と多様性に対応するため、オンプレミスのデータレイクでは費用のかかるハードウェアの追加が必要になることがよくあります。最終的に、インフラストラクチャの管理と保護にはIT の専門知識が必要となり、人件費が増加します。

それから、機会損失があります。オンプレミスソリューションには、クラウドベースのデータレイクのような柔軟性とアジリティが欠けているため、迅速なイノベーションとAI の導入が妨げられ、変化するビジネスニーズへの対応が難しくなります。その結果、多くの企業は、オンプレミスのデータレイクにとどまることで持続可能性を犠牲にしています。

2. ESG、2022 年、Google Cloud Dataproc の経済的メリットの分析

クラウドにより費用構造が変わる



クラウド データレイクの費用構造は、従量課金制のユーティリティモデルに似ており、使用量に応じて費用が増減します。

主な構成要素は次のとおりです。

ストレージ

これは基本的な費用であり、保存容量に対してギガバイト単位で課金されます。また、データ量、アクセス頻度、選択したストレージ階層（ホット、コールド、アーカイブ）などの要素の影響を受けます。

コンピューティング

これらの費用はデータの処理と分析に対して発生し、抽出、変換、読み込み（ETL）、クエリ、ML などのタスクに使用されるコンピューティングインスタンスのタイプと期間に基づいて課金されます。

マネージド サービス

分析エンジン、データウェアハウス エンジン、BI ツール、ストリーミング サービスなどのマネージド サービスの使用にかかる費用。

データ転送

これらの費用は、クラウド環境に、クラウド環境から、またはクラウド環境内でデータを移動したときに発生します。

オーケストレーションと管理

これらは、データと AI のパイプラインのスケジュール設定とモニタリングに関連するコストです。

データ ガバナンス、セキュリティ、コンプライアンス

アクセス制御、暗号化、監査などの機能は、全体的な費用に影響を与える可能性があります。

これらの費用の構成要素をそれぞれ理解して最適化することは、クラウド データレイクの総所有コストを管理するために不可欠です。

Eureka は Google Cloud でダウンタイムを減らしてアプリのデプロイを加速

テクノロジー企業のEureka は、大規模なデータセットにAI と ML を適用することによって、組織がインテリジェンスを大規模に引き出すのを支援しています。同社は、データ分析サービスをより効率的に運用するためにGoogle Cloud に移行し、BigQuery と Dataproc を使用することにより、サーバーのダウンタイムを削減して、新しいアプリケーションのデプロイを加速させています。さらに、費用の削減も実現しました。BigQuery ストレージによって18% 程度もの費用削減を実現できたためです。

> データ分析機能の向上には、目を見張るものがありました。さらに、Google Cloud に移行してからは、オンプレミス ソリューションを使用していたときと比べて、ダウンタイムとサービスの停止を大幅に削減できました。"

Eureka、
CMO
Michael Hawkins 氏

全文を読む →

移行の初期費用の計画

オンプレミスのデータレイクをクラウドに移行する場合、さまざまな初期費用がかかります。それらの費用について最初から把握しておくことで、予期しない驚きや予算超過を回避できます。

移行の初期費用には通常、以下が含まれます。

クラウド インフラストラクチャ

移行プロセス中のオンプレミスのデータレイクがまだ使用中で廃止できない期間、仮想マシン、ストレージ、ネットワークコンポーネントをプロビジョニングするための費用です。

プロフェッショナル サービス

ほとんどの企業組織は、クラウド移行を支援するため、コンサルティング、実装サポート、トレーニングを利用しています。

データ転送

これには、現在のデータセンターからの下り（外向き）料金に加え、帯域幅費用がかかる場合があります。

データの変換とクレンジング

データをクラウド環境に適合させて、互換性を確保するため、これには多大な費用がかかる可能性があります。

ライセンス料と利用料金

データレイクの移行に必要なクラウドベースのソフトウェアとツールに費用がかかる可能性があります。

費用はクラウドで最適化できる

クラウド データレイクの費用を最適化するには、予防的かつ継続的な取り組みが必要です。ここでは、パフォーマンスとスケーラビリティを維持しながら費用を大幅に削減するため、移行前、移行中、移行後に実行できる具体的なアクションをいくつか紹介します。

コンピューティングおよびストレージ リソースのサイズ適正化

ワークロード パターンを分析し、実際のニーズに基づいてコンピューティング容量を調整します。自動スケーリングを使用し、需要に応じてリソースを動的に調整します。データアクセス頻度に基づいて適切なストレージ階層（ホット、コールド、アーカイブ）を選択します。

費用対効果の高い料金モデルの利用

予測可能なワークロードやフォールトトレラントなワークロードには、予約済みインスタンスまたはスポットインスタンスを活用します。クラウドプロバイダが提供する継続利用割引や他の料金オプションを調べます。

データ ストレージの最適化

データを圧縮してストレージのフットプリントを削減し、データライフサイクル管理ポリシーを利用して、古くなったデータをより安価なストレージ階層に自動的に移動します。使用されていないデータを定期的に削除またはアーカイブします。

データ パイプラインの最適化

効率的なデータの取り込みと処理のパイプラインを設計します。リージョンやアベイラビリティゾーンをまたがるデータ移動を最小限に抑え、ネットワーク費用を削減します。データ変換タスクにサーバーレスコンピューティングを使用すると、実際の使用量に対してのみ課金されます。

クラウド支出のモニタリングと分析

クラウド プロバイダのコスト管理ツールを活用することにより、支出を追跡して、費用発生源を特定し、予算アラートを設定します。費用レポートを定期的に確認し、支出パターンを分析して最適化可能な分野を特定します。

定期的な見直しと最適化

データレイク環境を継続的にモニタリングし、ニーズの変化に応じて戦略を調整します。新しいクラウド サービスとコスト最適化のベストプラクティスに関する最新情報を常に入手します。

データ ガバナンス ポリシーの実装

明確なデータの保持ポリシーを確立して、データ品質基準を適用することにより、ストレージ費用を最小限に抑え、データの使いやすさを高めます。

クラウドネイティブのツールとサービスの使用

ELT / ETL プロセスには、BigQuery や Dataproc 上のサーバーレス Spark などのマネージドサービスを利用します。これは、セルフマネージドソリューションよりも費用対効果が優れています。

LiveRamp は Hadoop を Google Cloud に移行

LiveRamp は、ビジネスの拡大に伴い、オンプレミス環境のデータセンターにおけるスペースと電力に関連する制約に直面し、ビジネスの目標を達成する能力が制限されていました。そこで、弾力性の高い環境を活用するためHadoop から Dataproc に移行するという戦略的な決定が下され、それが功を奏しました。

> 当社は、オンデマンド VM と Spot VM のバランスを適切に取ったことで、特定のクラスタで約 30% のコスト削減を実現しました。これは、エンジニアが有効な A/B テスト フレームワークを構築したことで、複数の構成でクラスタやジョブを実行して最も信頼性が高く、管理可能で費用対効果の高い構成を見つけることができた結果です。また、あるアプリケーションは 10 倍以上の速さで実行できるようになりました。"
>
> LiveRamp
> シニア エンジニアリング マネージャー
> Mithun Bondugula 氏

全文を読む →

移行のリスクを慎重に管理する必要がある

データレイクをクラウドに移行する際、組織にリスクをもたらす可能性がある大きな分野は 3 つあります。注意すべき点とリスクを軽減する方法は次のとおりです。

01 データ中心のリスク

データ損失

ネットワークの問題、抽出/ 読み込み時のエラー、あるいは新しい環境での誤った削除などにより、転送時にデータが失われる可能性があります。緩和策として、堅牢なデータ検証、バックアップ/ リカバリーメカニズム、完全性を確保するためのデータ品質ツールの使用などが挙げられます。

データの破損

データは、転送時または変換プロセス中に破損する可能性があります。これは、互換性のない形式、エンコードの問題、または移行スクリプトのバグが原因となる可能性があります。移行前と移行後の徹底的なデータプロファイリングと、スキーマ マッピングの検証が不可欠です。

データ セキュリティ

あらゆる移行フェーズでデータセキュリティを維持することが最も重要です。リスクには、不正アクセス、転送時のデータ侵害、またはクラウド環境の構成ミスによるデータ漏洩などがあります。プロジェクト全体を通じて暗号化、強力なアクセス制御、セキュリティ監査が重要となります。

02 プロジェクトの管理および実行のリスク

非現実的なタイムライン

移行の複雑さと必要な時間を過小評価すると、早計な決定を下したり、ミスが増えたり、最終的にはプロジェクトが失敗したりする可能性があります。予期しない問題が生じた場合に備えて余裕のある現実的なタイムラインにするなど、綿密な計画を立てることが重要です。

スキルギャップ

クラウドの専門知識（クラウドプラットフォーム、データ移行ツール、セキュリティのベストプラクティスに関するスキルを含む）の不足は、プロジェクトの進行にとって大きな妨げとなる可能性があります。多くの場合、既存のスタッフをトレーニングするか、経験豊富なクラウドの専門家を採用することが必要となります。

コミュニケーション不足

移行プロジェクトにはさまざまな関係者が関与します。コミュニケーション不足は、誤解、期限超過、対立につながる可能性があります。明確なコミュニケーションチャネルを確立し、すべての関係者に情報が伝わり、連携がとれるようにします。

不十分なテスト

本番環境に影響を与える前に潜在的な問題を特定して対処するには、テストが不可欠です。急いだり不十分だったりすると、費用のかさむ問題が移行後に発生する可能性があります。パフォーマンスやセキュリティのテストを含む包括的なテスト戦略がどうしても必要です。

03 クラウド特有のリスク

ベンダー ロックイン

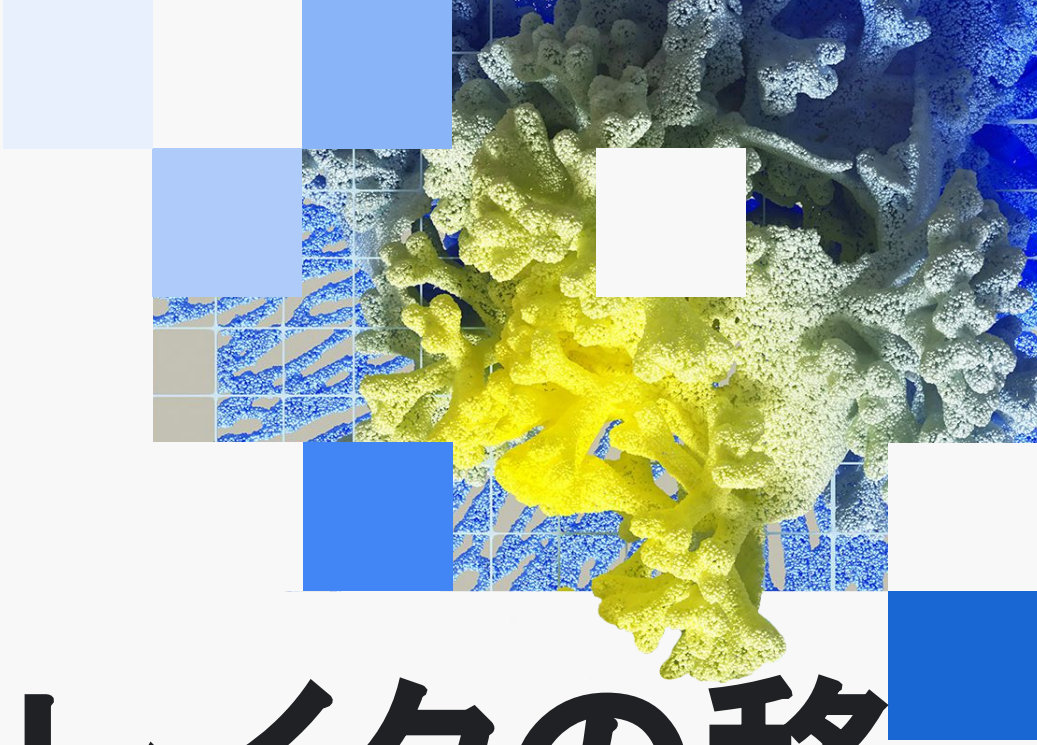
単一のテクノロジーやプロバイダに過度に依存すると、将来の柔軟性が制限され、費用が上昇する可能性があります。このリスクを軽減するには、複数の技術オプションとプロバイダを評価し、相互運用性を重視した戦略を検討してください。

予期しないクラウド費用

クラウドの料金計算は複雑になる場合があります。適切に計画していない場合、データ転送、ストレージ、コンピューティングに関連する予期しない費用が発生する可能性があります。徹底した費用の見積もり、費用の継続的なモニタリング、最適化戦略が不可欠です。

統合に関する課題

クラウド データレイクを既存のオンプレミスシステム、アプリケーション、またはデータソースに統合するのは複雑になる場合があります。ハイブリッド環境全体でシームレスなデータフローと互換性を確保するには、慎重な計画と統合テストが必要です。



3 章

データレイクの移行を成功させるための計画と実行

データレイクをクラウドに移行するプロセスは複雑になる可能性があり、綿密な計画と完璧な実行が必要です。この章では、現在の環境の評価、クラウド戦略の定義から、移行のオーケストレーション、継続的な成功のための最適化に至るまで、重要なステップについて詳しく説明します。

レイク移行の5つの段階

フェーズ 01

調査

すべてのデータアセット、ワークロード、ユーザー、権限、ワークフローのインベントリを作成することにより、データレイクの現在の状態を把握します。この情報は、移行の範囲を決定し、潜在的な課題を特定するために非常に重要です。

フェーズ 02

評価

最適なクラウドアーキテクチャを定義します。これには、適切なクラウドサービスの選択、ガバナンスポリシーの確立、費用管理の決定のほか、データマネジメントおよびセキュリティのプロセスとプロトコルを大筋を決めることが含まれます。

フェーズ 03

計画

タスクとサブタスクの定義、役割と責任の割り当て、タイムラインの確立、予算設定など、移行の詳細なプロジェクト計画を作成します。

フェーズ 04

実行

テーブルとデータ、ユーザーと権限、ワークロードとワークフローをクラウドに移行します。移行したデータとアプリケーションをテストおよび検証し、新しい環境で期待どおりに動作することを確認します。

フェーズ 05

最適化

移行が完了したら、パフォーマンスの微調整、費用対効果の向上、監視およびガバナンスプロセスの強化を継続的に行います。このフェーズでは、クラウドネイティブな機能とサービスを使用することで、スケーラビリティ、信頼性、メンテナンス性を向上させることができます。

Squarespace が Google Cloud の分析レイクハウスでエスカレーション件数を 87% 削減

Squarespace は、スケーラビリティを強化して、メンテナンスのオーバーヘッドを削減し、データオペレーションにおけるイノベーションを促進するため、オンプレミスの Hadoop インフラストラクチャを Google Cloud に移行し、主に BigQuery、Dataflow、Cloud Composer を活用して統合データプラットフォームを構築しました。

> Hadoop エコシステムを Google Cloud に無事に移行できたことで、インフラストラクチャのメンテナンスにまつわる大きな負担が消失しました。移行前の数か月と移行直後の数か月を比較すると、エスカレーション件数が 87% 減少しています。データ プラットフォーム チームとデータ インフラストラクチャ チームはさまざまなサービスやファイル システムの状態のモニタリングに集中する必要がなくなり、現在では社内ユーザーがビジネスを推進するために必要とする新機能やより優れたソフトウェアの提供に注力しています。"

Squarespace、
シニア ソフトウェア エンジニアリング マネージャー
Douglas O'Connor 氏

Squarespace、
シニア スタッフ ソフトウェア エンジニア
Constantinos Sevdinoglou 氏

全文を読む →

始める前に組織計画を立てることが不可欠

データレイクの移行を確実に成功させるには、まず経営幹部からの支援を確保し、IT、データ エンジニアリング、データ サイエンス、BI、ビジネス ユニットなど、さまざまな部門の代表者で構成された部門横断型チームを編成する必要があります。基盤となるクラウドインフラストラクチャを担当する専任のプラットフォームチームを含め、各チームメンバーの役割と責任を明確に定義します。

次に、全員に情報を提供して、コラボレーションを促進するため、明確なコミュニケーションチャネルと報告経路を確立します。

定期的なミーティング、進捗レポート、一元的なコミュニケーション プラットフォームは、透明性を維持し、障害にすばやく対処するのに役立ちます。移行プロセス全体を通して、チームが協力して、専門知識を共有し、互いから学ぶことを奨励します。

スキルギャップを解消するため、クラウドテクノロジーとベストプラクティスに関するトレーニングやスキルアップの機会を提供します。これにより、チームは新しいクラウド環境を効果的に操作し、その機能を最大限に活用できるようになります。チームのスキルを補完し、移行プロセスを加速させるため、経験豊富なパートナーと協力することを検討します。

そして最後に、組織に残る潜在的な懸念や変化への抵抗に前もって対処します。そのとき、クラウド移行のメリットを伝え、チームメンバーが新しい環境に適応できるようサポートを提供することが役立ちます。

フェーズ 01

調査

調査フェーズでは、現在のデータレイク環境について包括的に把握することが目標です。これにより、移行の範囲を決定して、考えられる課題とリスクを特定し、移行を成功させるのに必要なリソースとタイムラインを見積もることができます。

同時に、移行が技術的だけでなく戦術的にも成功するよう、ビジネスのサービスレベル契約（SLA）を含む組織のニーズをカタログ化する必要があります。SLA は、データとワークロードの期待されるパフォーマンス、可用性、セキュリティレベルを定義するものであり、移行戦略の中心に据える必要があります。

現在の状態からのマッピングが必要な要素として以下をインベントリに含めます。

データアセット

データソース、データの種類と量、その場所、さらにデータに依存するダウンストリームの利用者（システムやユーザー）をすべて識別し、カタログ化します。データ検出ツールは、このプロセスを自動化するのに役立ちます。中核となるビジネス機能をサポートし、最も厳しい SLA が適用される重要なデータには特に注意してください。こうすることで、そのデータをまず移行することを優先し、重要な業務の中断を最小限に抑えることができます。さらに、メタデータ（つまり、スキーマ、データリネージ、データ品質メトリック、ビジネス コンテキストなどのデータを説明する情報）もカタログ化します。

ガバナンス

データの分類、データの保持、アクセス制御ルール、コンプライアンス要件など、現在のデータガバナンス ポリシーを徹底的に文書化します。データレイクにアクセスできるすべてのユーザーとグループ、およびそれぞれのロールと権限を識別します。最後に、既存のデータガバナンス ツールとプロセスが、選択したクラウドプラットフォームと互換性があるかどうかを確認します。場合によっては、クラウド プロバイダのセキュリティおよびガバナンスフレームワークに合わせて、それらを調整する必要があります。

ワークロード

ETL パイプライン、データ変換ジョブ、分析クエリ、ML モデルなど、データレイクとやり取りするすべてのプロセスとアプリケーションを特定します。他のシステムやデータソースへの依存関係を評価し、選択したクラウドプラットフォームとの互換性を確認することにより、必要なコード変更やアーキテクチャの変更を計画できるようにします。データアセットの場合と同様、ビジネスに不可欠なワークロードを特定して優先順位を付けます。

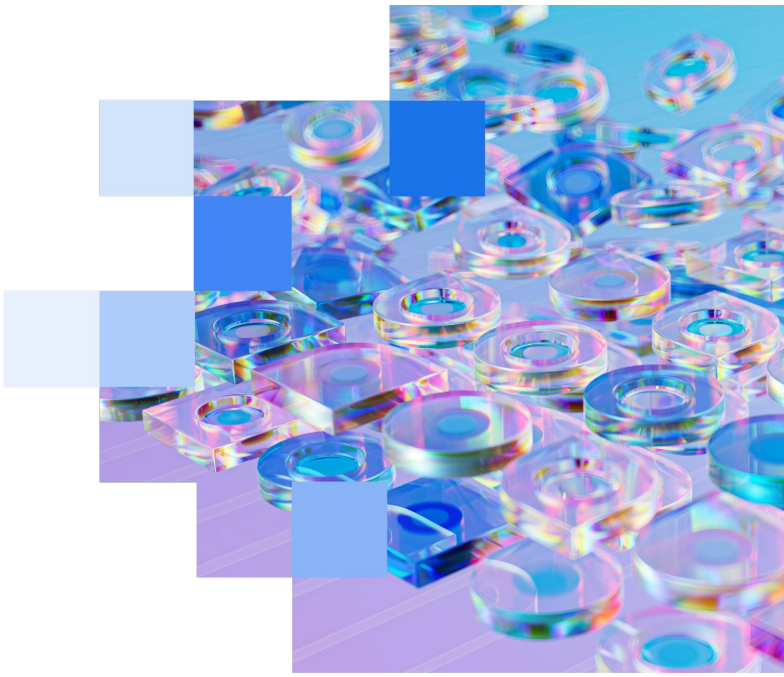
ワークフロー

ETL パイプライン、データ変換ジョブ、ML ワークフローなど、データレイクに関連付けられているすべてのワークフローと有向非巡回グラフ（DAG）を識別します。それらの依存関係、スケジュール、リソース要件を分析します。ワークロードやガバナンスと同様、現在のワークフローオーケストレーションツールと選択したクラウドプラットフォームの互換性を確認し、必要な変更をはっきりと把握します。

フェーズ 02

評価

ここまでで、データレイクについて理解が深まったので、最適なクラウド アーキテクチャを定義できるようになりました。これには、適切なクラウドサービスの選択、ガバナンスポリシーの確立、費用管理の決定のほか、データマネジメントおよびセキュリティのプロセスとプロトコルの大筋を決めることが含まれます。



移行戦略の定義

適切な移行戦略の選択は、データレイクの複雑さ、予算とタイムラインの制約、希望するクラウド最適化レベル、リスク許容度などの要因によって異なります。明確に定義された戦略があれば、スムーズで効率的な移行プロセスが確保されて、中断が最小限に抑えられ、クラウド投資の価値が最大限に高められます。

データレイクの移行に関して、組織は通常 3 つの戦略のいずれかを選択します。

リフト＆シフト

最小限の変更で、既存のデータレイクを「現状のまま」クラウドに移行します。このアプローチは、スケジュールがタイトな場合や初期段階でのサービス中断を最小限に抑えたい場合に適していますが、スケーラビリティ、費用対効果、高度な分析などのクラウドのメリットを活用できなくなる可能性があります。

リフトして最適化

クラウドに移行した後、マネージドクラウド サービスを使用してデータレイクアーキテクチャを段階的に改善し、効率を高めます。速度と最適化を両立させるこの戦略では、コストとパフォーマンスを反復的に最適化しながら、クラウドのメリットを早い段階で手に入れることができます。

モダナイゼーション

データレイクを再構築し、クラウドネイティブのサービスと機能を最大限に活用します。この包括的なアプローチでは、長期的なメリットが最も大きくなりますが、多額の先行投資とより長期の実装タイムラインを伴います。

アーキテクチャの設計

新しい環境でパフォーマンス、スケーラビリティ、費用対効果を最適化するには、クラウド アーキテクチャの設計に細心の注意を払います。その一環として、可能であればSLA 保証が組み込まれたクラウドネイティブサービス（マネージドデータ ウェアハウスやサーバーレスデータ処理エンジンなど）を活用して、ビジネスのSLA を満たすか、それを超えることができるという保証を得ます。

クラウドベースのデータレイクのアーキテクチャには、6 つの主要な要素が含まれます。

01 ストレージ

オブジェクト ストレージ

クラウド オブジェクト ストアは、元データと処理済みデータ（特に画像、動画、ログファイルなどの非構造化データ）に対して、スケーラブルで費用対効果の高いストレージを提供します。

テーブル形式

Apache Iceberg のような最新のテーブル形式は、Apache Hive などの従来のテーブル形式に比べて大きなメリットがあります。これらの形式には、ACID プロパティ、スキーマの進化、およびタイムトラベル機能が備わっており、これらの機能は、アップデートが頻繁に実施されてスキーマが進化するデータレイクにとって重要です。

02 コンピューティング

サーバーレス ストリーミング

クラウド プロバイダは、Pub/Sub や Dataflow などのサーバーレス ストリーミング プラットフォームや、Apache Flink および Apache Kafka のマネージド サービスを提供しています。これらのプロダクトは、イベントドリブンなリアルタイムのデータ処理とデータ変換に最適です。

マネージド エンジン

クラウド プロバイダは、大規模なデータ処理、分析、ML 向けに、管理されたサーバーレスのSpark および Ray サービスを提供しています。これらのサービスによってインフラストラクチャ管理が簡素化され、最適化されたパフォーマンスが実現します。

データ ウェアハウジング

構造化データと分析ワークロードには、複雑なクエリに対して高いパフォーマンスとスケーラビリティを提供するクラウド データ ウェアハウス サービスを検討します。さらに、Apache Iceberg などのオープン形式を使用して、構造化データと非構造化データに対する分析および ML ワークロードを強化するため、BigQuery などの統合データプラットフォームを検討することもできます。

03 ネットワーキング

VPC

Virtual Private Cloud（VPC）を使用して、データレイク環境を分離し、ネットワークアクセスを制御します。

データ転送

クラウドプロバイダが提供する専用の接続サービスを使用して、オンプレミス環境とクラウド間のデータ転送を最適化します。

04 セキュリティ

アクセス制御

Identity and Access Management（IAM）のロールとポリシーを使用してきめ細かなアクセス制御を実装し、機密データへのアクセスを制限します。

暗号化

クラウドによって提供される暗号化サービスを使用して、保存データおよび転送中のデータを暗号化します。

脅威の検出

脅威の検出とモニタリングにはクラウドネイティブのセキュリティツールを使用します。

05 データ ガバナンスとメタデータの統合

データ ガバナンス

明確なデータガバナンスポリシーを確立し、クラウドベースのツールを活用してきめ細かなアクセス制御とコンプライアンスを実現します。モダナイゼーション戦略を長期的に考える際は、クラウドプロバイダのデータからAI へのフルガバナンス機能を検討してください。

クラウドベースのメタストア

クラウドプロバイダは、データレイク用の一元化されたメタデータリポジトリを提供することにより、統一された統合コンポーネントでデータ検出、ガバナンス、リネージのトラッキング、品質モニタリングを可能にしている場合があります。

06 移行ツール

データ移行サービス

クラウドプロバイダは、データ検証、変換、スキーマ マッピングなどの機能を備えたさまざまなデータ移行サービスを提供しています。移行する必要があるデータの量と種類、所要時間の要件、許容できるダウンタイム レベルを検討しながら、他の移行シナリオ向けのオープンソースツールや商用 ETL ソリューションについて検討します。

一括データ転送

大規模なデータセットの場合、クラウド プロバイダが提供する一括データ転送ツールを検討します。変更部分のみが転送される増分データ移行は、ダウンタイムを最小限に抑え、データの整合性を維持するのに役立ちます。

フェーズ 03

計画

移行をスムーズに実施して成功させるには、徹底した計画が不可欠です。ここで、詳細なプロジェクト計画を作成します。これは、実行フェーズが進むにつれて更新されていき、実際に何が起こったかを文書化する生きたドキュメントとして機能します。

プロジェクト計画の主な要素は次のとおりです。

タイムラインとマイルストーン

移行の主要なマイルストーンと、それぞれの予想されるタイムラインを定義します。段階的なアプローチにより、サービスの停止を最小限に抑えることができます。優先して移行する必要がある重要なデータとワークロードを文書化し、緊急性と重要度に応じて移行のその後のフェーズを計画します。

タスクと依存関係

移行プロセスを管理可能なステップに分割し、各段階で依存関係を文書化します。

役割と責任

移行における主要な関係者とその役割を特定します。関係者がつながるためのコミュニケーションチャネルを文書化し、必要に応じてエスカレーション手順を定義します。

予算

クラウドリソース、移行ツール、プロフェッショナル サービスなど、移行に関連する推定費用をおおまかに決めます。

ロールバック プラン

なんらかの問題が発生してビジネスクリティカルなプロセスに影響が及ぶ場合に実行するステップを文書化します。

フェーズ 04

実行

データレイクコンポーネントをクラウドに転送する準備が整いました。このフェーズで考慮すべき重要なポイントは次のとおりです。

- 移行時にデータの完全性を確保する方法
- 重要なビジネスプロセスのダウンタイムとサービス中断を最小限に抑える方法
- 移行したデータとワークロードを検証する方法
- 発生する可能性のある問題に対処する方法

移行プロセス全体を通じて、パフォーマンスと可用性を継続的にモニタリングし、SLA に準拠していることを確認します。計画フェーズ中に確立されたコミュニケーションチャネルを使用して、問題やサービス中断の可能性について関係者に注意喚起します。ビジネスの SLA に沿って移行作業を進めることにより、移行をスムーズに行って、ダウンタイムを最小限に抑え、クラウド内の重要なデータとアプリケーションの完全性を維持することができます。

移行の主な構成要素について詳しく見ていきましょう。

データ

精度と完全性を確保するには、移行前、移行中、移行後にデータの完全性を検証します。これには、チェックサムの比較、スキーマに対するデータの検証、レコード数や他の統計データの調整が含まれる可能性があります。クラウドプロバイダは、データ検証、変換、スキーママッピングなどの機能を備えたさまざまなデータ移行サービスを提供しているため、適切なデータ移行ツールを選択することも重要です。最後に、データ移行とは、ビットやバイトの移動だけではない点に留意してください。新しいクラウド環境に合わせてデータを変換し、クラウドネイティブのツールやサービス向けに最適化することも重要です。これには、新しいデータレイクでデータを効率的に使用できるよう、データ クレンジング、スキーマの変更、形式の変換を行うことが含まれる可能性があります。

ワークフローの移行

ワークフローとDAG を移行して適応させると、データ処理と ML のパイプラインで継続性が確保され、クラウドベースのオーケストレーション サービスのスケーラビリティと効率の高さを生かせるようになります。複雑なデータパイプラインの定義、スケジュール設定、モニタリングのためのマネージドサービスを提供するワークフロー オーケストレーション サービス（Cloud Composer など）の使用を検討します。これらのサービスでは、多くの場合、視覚的な DAG エディター、バージョン管理、モニタリングなどの機能が提供されます。同時に、ワークフローをリファクタリングしてクラウド向けに最適化することも検討できます。移行したワークフローと DAG をクラウド環境で徹底的にテストすることで、期待どおりに機能し、パフォーマンス要件を満たしていることを確認します。

メタデータ

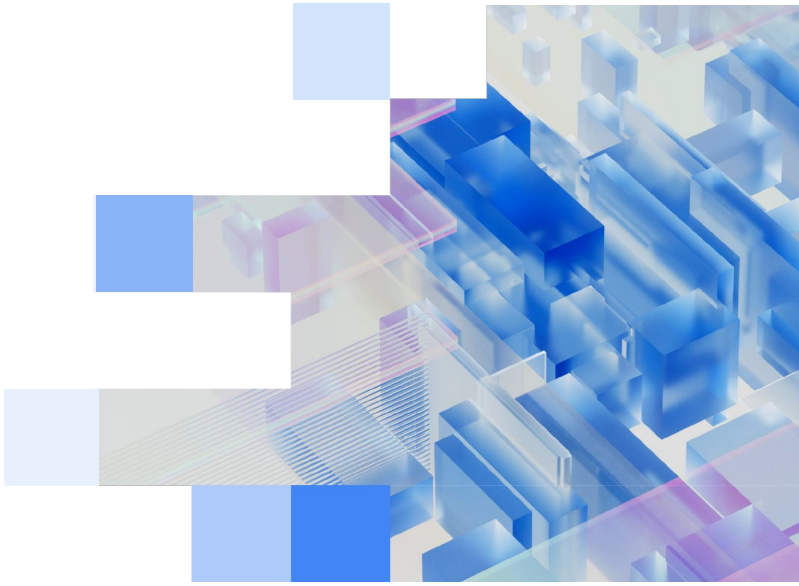
メタデータの移行は、継続的なデータ検出とデータ ガバナンスにとって、また移行先でもデータの価値と目的を維持するために重要です。クラウドプロバイダが提供するクラウドネイティブのメタデータ管理ツールを使用します。これらのツールには、多くの場合、メタデータの自動検出、分類、リネージのトラッキングの機能が備わっています。移行時は、ソースメタデータとターゲットメタデータ間の明確なマッピングを維持して、データリネージを保持し、クラウドにおいてデータが正確に解釈され、使用されるようにします。移行したメタデータの完全性と精度を検証することを忘れないでください。

ワークロード

ワークロードをリファクタリングすると、クラウドネイティブ サービスを活用し、パフォーマンスを最適化することができます。大規模なデータ処理には、Spark、Flink、Ray のマネージド サービスを使用したサーバーレスコンピューティング、またはサーバーレスデータウェアハウスを検討してください。ビジネス上の重要性と依存関係に基づいてワークロードの移行に優先順位を付け、段階的なアプローチを採用することでリスクを軽減し、ダウンタイムを最小限に抑えます。移行されたワークロードをクラウド環境で徹底的にテストして、機能性、パフォーマンス、データの完全性を確保し、クラウド モニタリング ツールを使用してボトルネックや最適化の機会を特定します。

ガバナンス

移行をスムーズに行ってサービスの中断を最小限に抑えながら、クラウドベースのデータレイクでロールの定義、ユーザー グループの作成、アクセス制御ポリシーの適用を行うには、ユーザーのアクセスと権限をきめ細かく制御できる IAM サービスを使用すると役立ちます。移行プロセス全体を通じてデータアクセスポリシーが一貫して適用され、機密データがガバナンス ポリシーに従って保護されていることを確認します。また、ユーザーのアクセスと権限を定期的に監査し、あらゆる潜在的なセキュリティリスクを特定して、それに対処します。移行時にこれらのステップを実行することにより、より安全でコンプライアンスを確保した環境を構築できると同時に、承認されたユーザーが、作業を効果的に行うために必要なデータにアクセスできるようにします。



フェーズ 05

最適化

データレイクの移行に成功した後も、仕事は終わりません。クラウドに移行した後、費用とパフォーマンスの両面で環境を継続的に最適化する機会が生じます。そのための主な方法をいくつか紹介します。

パフォーマンスの調整

クラウドネイティブツールを使用してボトルネックを特定し、クエリを最適化することにより、データレイクのパフォーマンスを継続的にモニタリングおよび微調整します。

費用の最適化

サーバーレスプロダクトの採用、適切なインスタンス タイプの使用、スポットインスタンスの活用、データ ストレージの最適化などの戦術は、費用の削減に役立ちます。

監視とガバナンスの強化

データ品質のモニタリング、リネージのトラッキング、コンプライアンスの自動化にクラウドネイティブ ツールを使用して、データ ガバナンスを強化します。

クラウドネイティブ サービスの使用

データレイクのスケーラビリティ、信頼性、セキュリティ、機能をさらに強化するため、新しいクラウドサービスと機能について継続的に調査し、採用します。

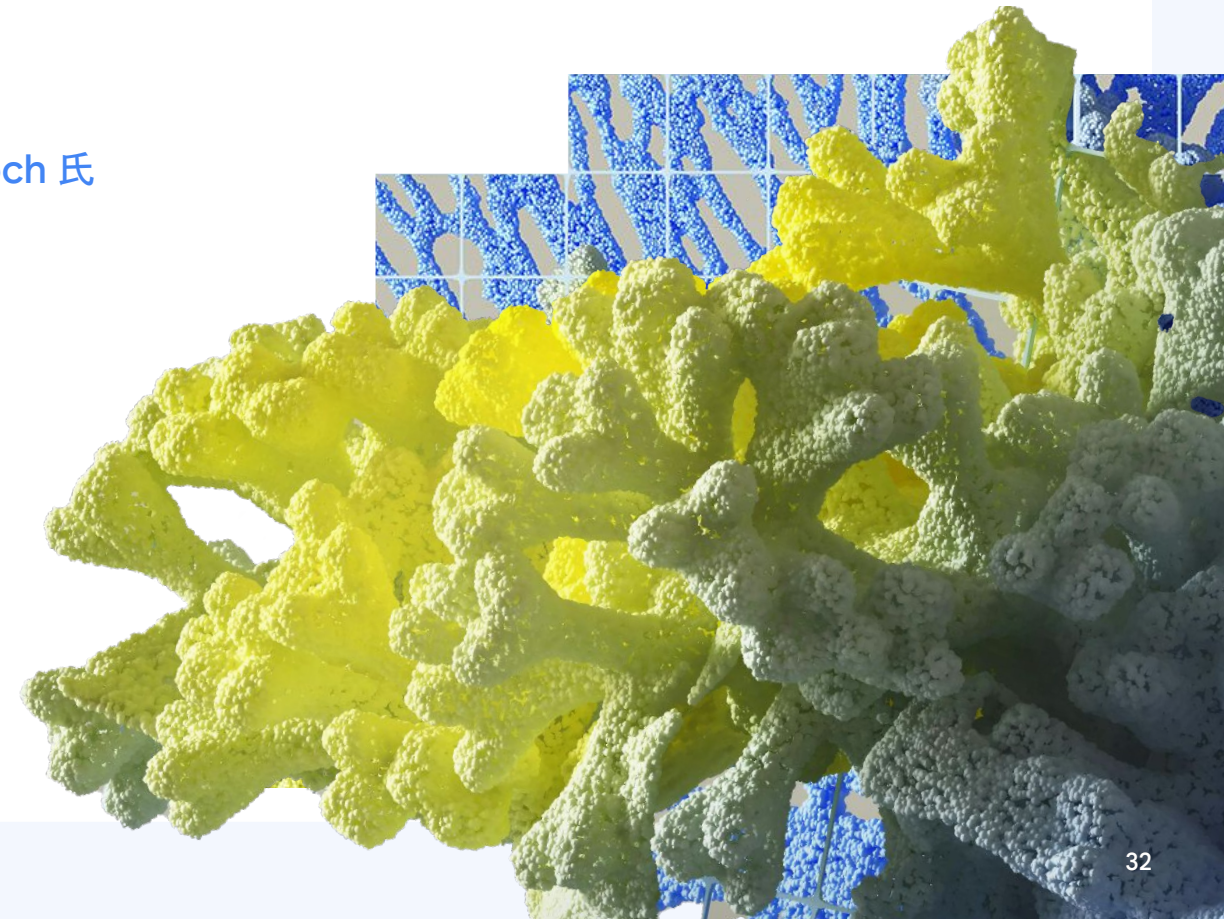
General Mills は、パートナーのサポートにより、データレイクの Google Cloud への移行期間を 30% 短縮

General Mills は、オンプレミスのデータと分析のエコシステムでは対応できなくなり、AI イノベーションに備えたいと考えたため、BigQuery へのデータレイク移行プロセスを開始し、そのビジョンの実現に向けて Accenture に支援を求めました。

同社の変革プログラムは元々3 年近くかかる予定でしたが、Accenture のサポートを受けて30% 以上短縮し、わずか21 か月で計画を遂行できました。

> Accenture は、当社の移行戦略の実行を支援するうえで重要な役割を果たしました。同社はすでに Google Cloud での経験があり、当社が目指していた、より広範なアーキテクチャに関する経験もありました。その種のインサイトは大いに役立ちました。"
>
> General Mills、
> デジタルコア VP
> Jason Staloch 氏

全文を読む →



計画に役立つチェックリスト

このチェックリストでは、データレイクをクラウドに移行する際に考慮すべきポイントを包括的に確認することができます。必ず、このチェックリストを実際のニーズや要件に合わせて調整してください。

チェックリストをダウンロード →

4 章

Google ができること

Google Cloud には、データレイク移行プロジェクトをサポートし、組織全体でデータドリブンなイノベーションを加速させるためのサービスおよびリソースの包括的なスイートが用意されています。

Google Cloud アカウント チーム

Google Cloud アカウントチームは、移行プロセス全体にわたってパーソナライズされたガイダンスとサポートを提供します。現在の環境を評価して、クラウド戦略を定義し、カスタマイズされた移行計画を策定できるよう支援します。さらに、一本化された連絡窓口として機能し、さまざまなGoogle Cloud リソースを調整して、プロジェクトの順調な進行をサポートします。

Google の 自動移行 ツール

Google Cloud には、データとワークロードをクラウドに移動するプロセスを効率化するための自動移行ツールが用意されています。たとえば、BigQuery 移行サービスは、オンプレミスのデータレイクやデータウェアハウスなどのさまざまなソースの検出と、そこからGoogle Cloud へのデータの転送を簡素化します。

Google コンサルティング サービス

複雑な移行シナリオの場合、Google コンサルティング サービスが専門家のガイダンスと実践的なサポートを提供できます。実績のあるGoogle Cloud コンサルタントは、さまざまな業界のお客様と協力し、複雑で大規模なデータレイク移行プロジェクトを数多く成功させてきました。

ベストプラクティスに加えて、移行に伴う課題や細かな点に関する深い見識を示すことで、お客様が自社のニーズの評価、クラウド アーキテクチャの設計、データとワークロードの移行を行い、データレイクを最適化してパフォーマンスと費用対効果を高められるようにします。同時に、複雑な移行に伴うリスクを軽減して、タイムラインを短縮し、全体的な移行費用を最小限に抑えることができるよう支援します。Google コンサルティング サービスを利用し、ビジネスの目標に沿って移行を確実に成功させることで、最終的にビジネス成果の向上を実現できます。

Migration Black Belt プログラム

移行時の技術的な課題に対処し、プロセスを加速させるため、Google Cloud では、パフォーマンス、スケーラビリティ、費用対効果を考慮に入れてクラウドデータレイクの設計を最適化するアーキテクチャレビューなど、経験豊富なエンジニアによる技術的支援を提供しています。Google Cloud の Migration Black Belt エンジニアリング チームは、クラウドにおけるデータとワークロードの移行、セキュリティ、ガバナンスに関するベスト プラクティスのガイダンスも提供できます。

金銭的インセンティブ

移行の初期費用を相殺できるよう、Google Cloud はさまざまな金銭的インセンティブプログラムを用意しています。これらのプログラムでは、2 つのシステムを並行して運用する費用を相殺できるようクラウドの使用に対してクレジットが付与されたり、移行の初期費用の一部がカバーされたりします。

Google Cloud パートナープログラム

Google Cloud パートナープログラムでは、データレイクの移行とクラウドテクノロジーの専門知識を持つ認定パートナーを擁する広大なネットワークを利用することができます。これらのパートナーは、移行の計画と実装から継続的なサポートやマネージド サービスまで、さまざまなサービスを提供しています。

移行のインセンティブの対象となるお客様

詳しくはこちら →

データレイクを Google Cloud に移行する理由

データレイクをGoogle Cloud に移行すると、他のクラウドプロバイダと比べて魅力的なメリットがいくつも得られます。

優れたデータ分析機能

Google Cloud はデータ分析とAI の分野をリードしていることで知られており、AI 対応の統合データプラットフォームである BigQuery や、費用対効果の高いフルマネージド Spark サービスであるDataproc などの強力なツールを提供しています。これらのツールにより、あらゆる種類のデータに対してより高速かつ効率的なデータ処理と分析が可能になり、データから貴重な分析情報を引き出すことができます。

オープンで柔軟なエコシステム

Google Cloud は、オープンソーステクノロジーの採用に積極的であり、さまざまなデータレイク アーキテクチャとツールをサポートする柔軟なエコシステムを備えています。これは、既存の投資を最大限に活用し、お客様独自のニーズに最も適したソリューションを選択できることを意味します。

イノベーションと AI / ML

Google Cloud は AI と ML の分野におけるイノベーションの最前線に位置しており、インテリジェントなアプリケーションの構築に使用できる事前トレーニング済みモデル、カスタムモデル開発ツール、特殊なAI / ML インフラストラクチャを提供しています。

優れた費用対効果

Google Cloud には、競争力のある価格設定とさまざまなコスト最適化ツールおよび戦略が用意されているため、クラウド費用の管理と投資価値の最大化に役立ちます。

セキュリティとコンプライアンス

Google Cloud は、セキュリティとコンプライアンスに精力的に取り組んでおり、データを保護し、規制要件を満たすためのセキュリティ機能および認証の包括的なセットを備えています。

グローバル インフラストラクチャ

Google Cloud のグローバル インフラストラクチャにより、データレイクで高可用性、低レイテンシ、スケーラビリティが実現し、信頼性の高いパフォーマンスが確保されるだけでなく、世界中のどこからでもデータにアクセスできるようになります。

次のステップに進む準備はできていますか？

データレイクの移行を開始するには、Google Cloud チームにお問い合わせください。移行プロセスを加速させることを目的とした、期間限定のインセンティブプログラムをぜひご利用ください。

今すぐこちらのフォームに記入し、Google Cloud でデータの可能性を最大限に引き出しましょう。

協力者:

Google Cloud、プロダクト マーケティング マネージャー
Angela Soares

Google Cloud、AI 向け分析データ プラットフォーム、GTM ストラテジスト
Adnan Hasan

Google Cloud、シニア アウトバウンド プロダクト マネージャー
Sajal Agarwal

Google Cloud、プロダクト マーケティング マネージャー
Jill Hardy

概要チェックリスト

データ移行チェックリスト

フェーズ 01

調査

データアセットのインベントリ作成

- ▢ すべてのデータソース（データベース、テーブル、ファイルなど）をカタログ化する
- ▢ データの量と種類を把握する
- ▢ データ品質を評価し、潜在的な問題を特定する
- ▢ データリネージと依存関係を文書化する

ワークロードの分析

- ▢ データレイクにアクセスするワークロードとプロセスを特定する
- ▢ データアクセスのパターンと頻度を把握する
- ▢ 各ワークロードのパフォーマンス要件を分析する

ユーザーと権限の文書化

- ▢ データレイクにアクセスできるすべてのユーザーとグループを特定する
- ▢ 各ユーザー / グループのアクセスレベルと権限を文書化する

ワークフローのマッピング

- ▢ データ パイプラインとETL プロセスを文書化する
- ▢ 一般的なデータ変換および拡充のステップを特定する
- ▢ システム間のデータリネージと依存関係を分析する

フェーズ 02

評価

クラウド プロバイダの選択

以下の条件に基づいてクラウドプロバイダを評価する。

- ▢ 提供サービス（ストレージ、コンピューティング、分析、AI）
- ▢ 料金モデルとコスト最適化オプション
- ▢ セキュリティ機能とコンプライアンス認証
- ▢ スケーラビリティとパフォーマンス能力
- ▢ 地理的可用性とデータ所在地要件

移行戦略の定義

- ▢ リフト＆シフト
- ▢ リフトして最適化
- ▢ モダナイゼーション

クラウド アーキテクチャの定義

- ▢ クラウドの移行先データレイクアーキテクチャを設計する
- ▢ 適切なクラウドストレージソリューション（オブジェクトストレージ、データウェアハウスなど）を選択する
- ▢ データ処理のためのコンピューティングリソースを決定する（例 サーバーレスプロダクト、仮想マシン）

ガバナンス ポリシーの確立

- ▢ データ セキュリティポリシーとアクセス制御を定義する
- ▢ データ品質基準と検証手順を確立する
- ▢ 関連する規制（PCI、GDPR、HIPAA など）を遵守する
- ▢ データの保持および削除ポリシーを実装する

費用管理の決定

- ▢ ストレージ、コンピューティング、他のサービスのクラウド費用を見積もる
- ▢ 費用最適化戦略を実装する（リソースのサイズ適正化、予約済みインスタンスなど）
- ▢ 予算のモニタリングとアラートメカニズムを確立する

フェーズ 03

計画

移行計画の策定

- 移行するデータとワークロードの優先順位を決める
- 移行タスクと依存関係の大筋を決める
- タイムラインとマイルストーンを確立する

役割と責任の割り当て

- 移行における主な関係者とその役割を特定する
- コミュニケーションチャネルとエスカレーション手順を定義する

予算の設定

- クラウドリソース、移行ツール、プロフェッショナルサービスに予算を割り当てる
- 予算に対する移行コストを追跡およびモニタリングする

フェーズ 04

実行

データの移行

- データ移行に適切なツールとプロセスを使用する
- 移行時にデータの完全性と整合性を確保する
- 移行したデータをソースデータと照合して検証する

ユーザーと権限の移行

- クラウドにおけるユーザーの認証と承認を構成する
- ユーザーのロールと権限を新しい環境に移行する

ワークロードの移行

- データレイクにアクセスするワークロードとプロセスを移行する
- クラウドでデータパイプラインとETL プロセスを構成する
- 移行したワークロードをテストして検証する

監視とトラブルシューティング

- 移行プロセスを継続的にモニタリングする
- 移行時に発生した問題やエラーに対処する
- 移行の進行状況と生じた課題を文書化する

フェーズ 05

最適化

パフォーマンスの最適化

- データのストレージと処理を微調整して最適なパフォーマンスを得る
- クラウドネイティブ サービスを使用してパフォーマンスを高める
- パフォーマンス指標をモニタリングおよび分析してボトルネックを特定する

費用の管理

- クラウド費用を継続的にモニタリングし、最適化の機会を特定する
- 需要に応じてリソースをサイズ適正化し、使用量を調整する
- クラウド プロバイダが提供するコスト管理ツールとサービスを使用する

セキュリティの強化

- データ保護のためのセキュリティのベストプラクティスを実装する
- クラウドネイティブのセキュリティ機能（暗号化、アクセス制御、脅威検出）を使用する
- セキュリティポリシーと手順を定期的に見直し、更新する