aws analytics

モダン分析アーキテクチャの構築

データは必要不可欠

今日、データおよび分析は企業にとって必要不可欠な要素です。データからビジネス価値を生み出すことに成功した企業が他社を上回る存在になるでしょう。Aberdeen のアンケートによると、最新のデータレイク分析プラットフォームを導入した企業では、同業他社を 9% 上回る収益の増加が見られました。これら企業のリーダーは、データレイクに保存されたログファイル、クリックストリーム、ソーシャルメディア、インターネット接続デバイスからのデータといったデータソースに対して、機械学習などの新しいタイプの分析を実行することができたからです。

最新のデータレイク分析は、企業が以下のことを実現するうえで役立ちます。

- 機会をすばやく特定し、対応する
- ビジネスを迅速に成長させる
- より多くの顧客を引き寄せ、維持する
- 生産性を向上させる
- 先を見越してデバイスを維持する
- 十分な情報に基づく意思決定を行う

必要なデータにアクセスできない原因は何か?

データからより多くの価値を引き出したいと考えているが、今日の最新のデジタルビジネスによって生み出されるすべてのデータを取得、保存、分析するのに苦労しているという声が企業から常に寄せられています。データは飛躍的に増加し、新しいソースからも来るようになっており、ますます多様化しています。また、不特定多数のアプリケーションやユーザーが、データに安全にアクセスでき、より短時間に分析できなければなりません。現在のデータのサイズ、複雑さ、多様なソースに対して、過去に機能したのと同じテクノロジーやアプローチはもはや通用しません。

お客様はデータを異なるサイロに保存してきたため、データの蓄積量が多くなるにつれて、分析を実行するのが難しくなってきました。お客様は分析を容易にするために、データすべてをひとつのリポジトリ、つまりデータレイクに保存したいと考えています。企業は、選択した標準ベースのデータ形式を使用して、データをあらゆる規模で安全かつ低コストで保存する必要があります。そして、現在と将来の分析ユースケースで確実にニーズを満たせるようにするために、幅広い分析エンジンを使用して、さまざまな方法でデータ分析する柔軟性を求めています。また、インサイトをさらに前に進め、今後の動向を正確に予測するために、履歴データの運用レポートからリアルタイム分析と機械学習の実行へとシフトしていく必要もあります。

既存のデータインフラストラクチャの課題

ほとんどの企業が、報告および分析目的のデータウェアハウスを構築してきました。これらの企業では、独自のトランザクション処理システムやその他のデータベースを含む、さまざまなソースからのデータを使用しています。また多くの企業では、一般的に「ビッグデータ」と呼ばれるデータ、または高度に構造化されたデータウェアハウスには適さないデータの分析用に、Hadoop フレームワークを構築してきました。データウェアハウスおよびビッグデータフレームワークの構築と実行は、複雑で高価なものとなってきました。

従来のデータウェアハウスの課題

従来のデータウェアハウジングシステムでは、以下のようなさまざまな問題や要求事項がありました。

- 初期投資として数百万ドルものソフトウェアとハードウェアのコストがかかる
- 計画と調達に数か月かかる
- セットアップが困難
- 実装とデプロイのプロセスに時間が必要
- 自分でデータモデルを定義し、データを取り込む必要がある
- データウェアハウス管理者のチームを雇う必要がある
- クエリが高速で実行できる状態を保ち、データ損失から保護する必要がある
- ミッションクリティカルな分析では、高度に正規化されたデータのみが必要とされる
- 多くのサイロ化されたデータストア内の大量のデータ (ダークデータ)
- ダークデータが分析用のデータウェアハウスにたどり着くことはない
- スケーリングが困難

データボリュームが増加した場合、またはより多くのユーザーが分析と報告を利用する必要がある場合には、低速なクエリパフォーマンスを受け入れるか、高価な更新プロセスに時間と労力を投資するかを選択する必要があります。実際、一部の IT チームでは既存のサービスレベルアグリーメント (SLA) を順守するために、データの増強やクエリの追加を推奨していません。多くの場合、企業ではこの問題を軽減するために複数のデータマートをセットアップしています。これらのデータマートには、データウェアハウス内にあるデータのサブセットのコピーが含まれています。こうすることで、実行時間の長い独自のクエリが、ミッションクリティカルなビジネスの運営や意思決定のパフォーマンスおよび SLA に影響を与えることはなくなります。これにより、データおよび分析のインフラストラクチャは複雑化し、企業は自社のデータウェアハウスとデータマートに関して選択したベンダーにさらに囲い込まれることになります。

さまざまな新しいデータに対して異なる分析エンジンが必要

また、従来のウェアハウスは、クリックストリーム、ログ、ソーシャルメディア、IoT センサー、その他の最新のデータソースから来るさまざまな半構造化データや非構造化データを処理するのに効果的ではありません。これらのデータタイプは、正規化された、構造化データモデル内には適合しません。これらの新しいデータタイプを処理できる、ビッグデータ用の Hadoop フレームワークといった異なる分析エンジンが必要です。このため、データの保存場所、データにアクセスできるユーザー、各データに最適な分析エンジンといったデータ戦略は、非常に複雑なものとなります。

バッチデータのロードの遅延

バッチデータのロードは課題となります。抽出、変換、ロード (ETL) ジョブは、データウェアハウスに対する分析の負荷が最小限である夜間に実行されるのが一般的です。夜間の ETL ジョブは、24 時間年中無休でデータへのアクセスが必要なグローバル企業には適していません。また、夜間の ETL ジョブは、即時に答えを必要とするユーザーがインサイトを得ることを遅らせる原因ともなります。今日の目まぐるしい速さで進む世界では、明日まで待つということでは遅すぎる場合があります。

最新の分析に対する限られたサポート

従来のデータウェアハウスでは、高度な機械学習や予測ワークロードはいずれもサポートされていないか、限られた形でのサポートのみが提供されています。そのため、従来のデータウェアハウスでは、リアルタイム分析または予測分析といった最新のユースケース、また高度な機械学習を必要とするアプリケーションをサポートすることができません。

データのセキュリティ保護のために回避策となるソリューションが必要

セキュリティとデータプライバシーも問題となります。医療および金融サービスといった機密性の高いデータを扱う業界では、データウェアハウスが ISO、HIPAA、FedRAMP などに準拠している必要があります。EU 一般データ保護規則 (GDPR) の規定により、顧客の機密データを保管時と移動時の全段階で暗号化しなければならないため、IT 部門にはさらなる負担がかかっています。こうした規制のなかには、突然の通知で、迅速に記録を取り出して更新または削除するよう組織に求めるものもあります。従来のデータウェアハウスでは、対応策として組織が高額なソリューションを実装しなければならない場合が多くあり、その影響でアナリストが機密データを分析できないケースが多発していました。

ビッグデータシステムの複雑さ

ビッグデータのプラットフォームでも同様の問題が発生してきました。また、データが大量かつ多様であるため、複雑なアルゴリズムで分析を実行しなければならず、スキルのある人材の不足も問題でした。その上、データウェアハウスとビッグデータシステムにまたがってデータ分析を行うのは複雑なタスクであり、時間もかかります。

最適なデータストレージ

組織がデータを異なるサイロに保存してきたため、データの蓄積量が多くなるにつれて、分析を実行するのが難しくなってきました。組織は分析を容易にするために、データすべてをひとつのリポジトリ、つまりデータレイクに保存したいと考えています。選択した標準ベースのデータ形式を使用して、データをあらゆる規模で安全かつ低コストで保存する必要があります。そして、現在と将来の分析ユースケースで確実にニーズを満たせるようにするために、幅広い分析エンジンを使用して、さまざまな方法でデータ分析する柔軟性を求めています。

データレイクとは

データレイクとは、あらゆる規模ですべての構造化データと非構造化データを保存可能な、一元化されたリポジトリです。データを変換したり構造化したりしなくてもそのまま保存できるほか、ダッシュボードや可視化からビッグデータ処理、リアルタイム分析、機械学習まで、さまざまなタイプの分析を実行できるため、より良い意思決定につなげることができます。ビジネスの意思決定に必要なデータは、今やデータレイクで処理し、修復し、そこからデータウェアハウスへロードできるようになりました。

データウェアハウスとデータレイク両方のニーズ

データウェアハウスとデータレイクはそれぞれ異なるニーズとユースケースに対応しているため、一般的な組織では、要件に応じてこの両方が必要となります。データウェアハウスを使用する組織がデータレイクのメリットに注目しているように、データレイクを使用することでウェアハウスを進化させ、広範なクエリ機能、データサイエンスのユースケース、さらには新しい情報モデルを発掘するうえでの高度な機能をウェアハウスで利用可能にすることができます。ガートナー社は、この技術進化を「Data Management Solution for Analytics」、通称「DMSA」と命名しました。

特性	データウェアハウス	データレイク
データ	トランザクションシステム、業務データベース、基幹業務アプリケーションからのリレーショナルデータ	IoT デバイス、ウェブサイト、モバイルアプリケーション、ソーシャルメディア、企業アプリケーションからの非リレーショナルデータとリレーショナルデータ
スキーマ	データウェアハウスの実装以前に設計 (スキーマオンライト)	分析時に書き込み (スキーマオンリード)
料金/パフォーマンス	高コストのストレージを使用、クエリ結果の取得は最速	低コストのストレージを使用してクエリ結果をより速く取得
データ品質	高度にキュレートされたデータで、事実の情報源として機能	任意のデータで、キュレートできるかどうかは不明 (raw データ)
ユーザー	ビジネスアナリスト	(キュレートされたデータを使用する) データサイエンティスト、データデベロッパー、ビジネスアナリスト
分析	バッチレポート、BI、可視化	機械学習、予測分析、データ検出、プロファイリング

分析パイプラインを作成する

データを分析するには、まずデータを収集し、処理して保存する必要があります。このタスクは、分析パイプラインとしてとらえることができます。ソースシステムからデータを抽出し、データを処理してから、分析を行うデータストアにそのデータをロードします。分析パイプラインは、データベースやアプリケーション、デバイスといった異種のソースから流入する大容量データに対応できるよう設計されています。

1. データを収集する
2. データを処理する
3. データを保存する
4. データを分析し、可視化する
5. 将来の成果を予測する

例として、次の図 1 をご覧ください。

図 1: 分析パイプライン

データを収集する

データには、トランザクションデータ、ログデータ、ストリーミングデータ、Internet of Things (IoT) データなど、さまざまな種類があります。それぞれの種類のデータは、個々の種類とその用途に適したデータストアに保存されているかもしれません。データストアの一部はトランザクションデータやリレーショナルデータに合わせて最適化されており、別のデータストアは非リレーショナルデータや非構造化データに適しているためです。組織の戦略として望ましいのは、データ、さらにはデータを生成または消費するアプリケーションに最も適した専用のデータベースを使用することです。

- **トランザクションデータ:** 一般的に、e コマースの購入取引や財務取引などのデータは、リレーショナルデータベース管理システム (RDBMS) または NoSQL データベースシステムに保存されます。どのデータベースソリューションを選択すべきかは、ユースケースとアプリケーションの特性によって左右されます。RDBMS ソリューションは、トランザクションを記録する場合や、トランザクションの表の行を複数箇所更新しなければならない場合などに適しています。NoSQL データベースは、データが正常に構造化されておらず定義済みのスキーマに適合しない場合や、スキーマが非常に頻繁に変更される場合などに適しています。

- **ログデータ:** システム生成されたログを確実にキャプチャすることで、ログに保存された情報を活用して問題のトラブルシューティングを行い、監査を実施し、分析を実行することができるようになります。データレイクは、分析に使用するログデータのストレージソリューションとして人気です。

- **ストリーミングデータ:** ウェブアプリケーションやモバイルデバイスのほか、多くのソフトウェアアプリケーションおよびサービスでは、膨大な量のストリーミングデータ (ときには 1 時間あたり数テラバイト) が生成されることがあるため、絶え間なく収集、保存、処理する必要があります。このデータにはさまざまなものが含まれるため、多くの場合、半構造化データまたは非構造化データと呼ばれています。

- **IoT データ:** 世界中のデバイスとセンサーが、絶え間なくメッセージを送信しています。現在、多くの組織がこのデータをキャプチャしてインテリジェンスを引き出す必要をますます感じています。

データを処理する

収集プロセスでは、データソースからデータを収集または抽出し、データを変換して、データを別の保存先に保存します。この保存先には、別のデータベースのほか、データレイク、データを処理または分析するデータウェアハウスなどの分析サービスが挙げられます。

バッチデータとリアルタイムデータ

処理ワークフローには、バッチとリアルタイムの 2 種類があります。

これまでと同様、バッチデータのロードは、まだ広く使用されています。毎晩のバッチジョブでは、データがシステムから抽出され、分析可能な形式に変換されて、送信先にロードされます。このため、必要としている人にデータが利用可能になるまで時間がかかります。

リアルタイム処理では、伝送中のデータが保存される前に、インメモリのインラインデータ変換が実行されます。これらのストリーミング技術を使用すると、リアルタイムで大規模なデータの取り込みが可能になるため、ほぼ瞬時に分析が実行できます。

- **抽出、変換、ロード (ETL):** ETL は、複数のソースからデータを取り出し、もしくは抽出し、定義されたターゲットスキーマ (書き込み時のスキーマ) に合わせて変換し、送信先のデータストアにロードするプロセスです。ETL は通常、継続的に進行するプロセスで、明確に定義されたワークフローが夜間などの一定の時間に実行されます。ETL ジョブのセットアップと実行は手間のかかる作業であり、完了するまでに数時間かかる場合があります。

- **抽出、ロード、変換 (ELT):** ELT は ETL のバリアントであり、変換が行われる前に、抽出されたデータがターゲットシステムにロードされます。スキーマは、データの読み取り時、または使用時に定義されます (読み取り時のスキーマ)。ELT は通常、ターゲットシステムに十分な変換処理能力があり、事前定義されたスキーマと一致しない方法でデータを探索する場合に適しています。

- **リアルタイム処理:** リアルタイムのデータストリーミングサービスを使用することで、継続的なストリーミングデータを収集、処理、分析し、リアルタイムアプリケーションおよび分析ソリューションに大規模に配信できます。ここで重要なのは、毎晩のバッチ ETL ジョブの完了を待たずに、データをすぐに分析できることです。デベロッパーは、リアルタイムアプリケーションを簡単に構築し、安全で可用性が高く、耐久性があり、スケーラブルなフルマネージドサービスを利用できます。ストリーミングデータをレコードごとに、またはスライディングタイムウィンドウで連続して増分的に処理し、処理されたデータを、相関関係、集計、フィルタリング、サンプリングなどのさまざまな分析に使用できます。

データを保存する

データは、データレイク、またはデータウェアハウスのような分析ツールに保存できます。

データレイクは、構造化データおよび非構造化データを含む、すべてのデータの一元化されたリポジトリです。データレイクではスキーマが定義されていないため、ビッグデータ分析、全文検索、リアルタイム分析、機械学習など、追加的な分析が可能になります。すべてのデータの中央リポジトリとしてデータレイクを使用する組織が増加しており、ダウンストリームアプリケーションや分析ツールで使用されています。

データウェアハウスでは、分析用に最適化された、事前定義済みのスキーマが利用されます。データは厳選され、複数のデータソースからの信頼できる単一の情報源として機能します。

- **データレイク:** データレイクは、さまざまなタイプのデータおよび分析アプローチを組み合わせるのに必要な規模、俊敏性、柔軟性に対応しており、従来のデータサイロやデータウェアハウスでは不可能だった方法で、より深いインサイトを得ることができます。これにより、幅広い分析および機械学習サービスを柔軟に利用し、セキュリティやガバナンスを侵害することなく、すべての関連するデータに簡単にアクセスすることが可能になります。

- **データウェアハウス:** データウェアハウスは、1 つ以上のデータソース (またはデータレイク) からの情報の中央リポジトリです。データは事前定義済みのデータモデルに合わせて変換、修復、重複排除されます。データウェアハウスは、データ分析用に特別に設計されています。これには、大量のデータの読み取り、およびデータ間の関係や傾向の把握が含まれます。データベースは、トランザクションの詳細の記録など、データをキャプチャして保存するために使用されます。データウェアハウスを使用して、BI ツールを活用することにより、大量のデータに対する分析を高速で実行し、隠れたデータのパターンを発掘できます。データサイエンティストは、データウェアハウスのクエリを実行し、オフライン分析を通じて傾向を特定します。組織全体のユーザーは、アドホック SQL クエリ、定期レポート、およびダッシュボードを使用してデータを利用し、重要なビジネス上の決定を行います。

- **データマート:** データマートは、特定の機能領域または対象分野に重点を置いた単純な形式のデータウェアハウスであり、データウェアハウスのデータのサブセットのコピーが含まれています。例えば、組織の部門ごとに特定のデータマートを設定したり、リージョンに基づいてデータマートをセグメント化したりできます。大規模なデータウェアハウスや運用ストアから、またはこの組み合わせから、データマートを構築できます。データマートの設計、構築、管理は簡単に行えます。ただし、データマートは特定の機能領域に重点を置いているため、機能領域全体のクエリは、データ分散により複雑になる可能性があります。

データの分析

データの本当の価値の引き出し方

最新の分析パイプラインでは、さまざまなツールを活用して、データに隠された価値を引き出すことができます。1 つのツールですべてに対応することはできません。どの分析ツールでも、同じソース、つまりお使いのデータレイクにアクセスしてデータを処理できる必要があります。

データウェアハウスとデータレイクへの簡単なアクセス

データレイクからデータウェアハウスにデータをコピーして、構造化され、正規化された、高性能のクエリエンジンを活用するデータモデルに適合させることもできます。同時に、一部のユースケースでは、データウェアハウスの正規化されたデータと関連した非構造化データの分析が必要です。ここで、データの変換と移動による遅延を生じさせずに、データウェアハウスとデータレイクの両方のデータを含むよう、データウェアハウスのクエリを拡張することが、タイムリーなインサイトを得るうえで不可欠です。

他のビッグデータ分析ツールが、データレイク内の同じデータにアクセスできる必要があります。これにより、ビジネスユーザーからデータサイエンティストまで、組織をまたいですべての人が、データと分析結果の両方を信頼することができます。

- **インタラクティブ分析:** インタラクティブ分析では、通常、標準 SQL クエリツールを使用してデータにアクセスし、分析します。エンドユーザーは、すばやい結果と、変更および再実行がすぐにできるクエリを求めています。

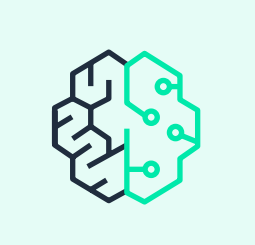
- **データウェアハウジング:** データウェアハウジングでは、分析用に最適化された高性能のクエリエンジンを使用して、ペタバイト規模の大量のデータに対する複雑な分析クエリを実行できます。

- **データレイク分析:** 新しい種類のデータウェアハウスが登場しています。データウェアハウスのクエリをデータレイクに拡張して、データウェアハウスとデータレイクで構造化データや非構造化データを処理し、データを移動せずにエクサバイトにスケールアップすることが可能になります。

- **ビッグデータ分析:** ビッグデータ処理では、Hadoop および Spark のフレームワークを使用して大量のデータを処理します。

- **運用分析:** 運用分析では、既存の運用の改善に重点が置かれ、アプリケーションのモニタリング、ログ、およびクリックストリームデータなどのデータを使用します。

- **ビジネスインテリジェンス (BI):** BI ソフトウェアは、ビジネスの意思決定のためにデータを取得、分析、変換、レポートする、使いやすいアプリケーションです。BI ツールでは通常、データウェアハウスやビッグデータ分析システムなどの分析サービスに保存されているデータを読み取ります。BI ツールでは、レポート、ダッシュボード、可視化を作成して、ユーザーがアドホックで特定のデータを詳しく分析できるようにします。その結果、組織の意思決定が加速、改善され、運用効率が向上します。また、新しい機会、収益の可能性、および市場の傾向を把握して KPI を報告できるようになります。



機械学習の適用

企業がますます大量のデータを生成、保存、分析するのに伴い、業績を上げるためにこれら膨大なデータセットを自動化して利用することが求められています。機械学習への依存度が高まり、タスクの自動化、エンドユーザーやお客様へのパーソナライズされたサービスの提供、データ分析による運用効率の向上などが実現しています。機械学習は、モデルの構築とトレーニング、および本番環境へのデプロイに要するプロセスが複雑で時間がかかるため、しばしば実際より難しいものと思われています。

機械学習プロセス

まず、トレーニングデータを収集、整備して、データセットのどの要素が重要かを見極める必要があります。次に、使用するアルゴリズムとフレームワークを選択する必要があります。アプローチの決定後は、トレーニングによって予測の仕方をモデルにティーチングする必要があり、これには多くのコンピューティングを要します。次に、できるだけ正確な予測を提供できるよう、モデルを調整する必要があります。これにはしばしば、多くの手作業が求められます。完全にトレーニングのできたモデルを開発した後は、モデルをアプリケーションに統合してこのアプリケーションをスケーリングの可能なインフラストラクチャ上にデプロイする必要があります。これらすべてには特化した専門知識と、大量のコンピューティングとストレージが求められ、多くの時間をかけてプロセスの各部を実験し、最適化する必要があります。

機械学習の簡単な設定

機械学習サービス: 機械学習サービスを使用すると、デベロッパーとデータサイエンティストは、あらゆる規模の機械学習モデルを迅速かつ簡単に構築し、トレーニング、デプロイできます。リアルタイムの不正行為検出モデルの実行から、開発中の医薬品における生物学的影響の仮想分析、また野球での盗塁成功の予測などユースケースと業界の全体にわたり、機械学習の適切な実装を妨げる複雑さを低減できます。

- **人工知能 (AI) サービス:** 人工知能サービスでは、あらかじめ用意されたインテリジェンスをアプリケーションとワークフローに提供します。お使いのアプリケーションに簡単に統合して、一般的なユースケース (パーソナライズされたレコメンデーション、コンタクトセンターのモダナイゼーション、安全とセキュリティの強化、カスタマーエンゲージメントの向上など)に対応できます。

- **機械学習 (ML) フレームワーク:** 機械学習フレームワークでは、機械学習アルゴ
リズムを試してカスタマイズできます。機械学習の担当者および研究者に、クラウド上での深層学習をあらゆる規模で加速させるインフラストラクチャとツールを提供します。

データレイク – 未来のイノベーションへの道

データレイクは今後の主流

クラウド内のデータレイクは、多くの企業にとってますます重要な戦略となってきました。より柔軟にデータを処理できるようにし、意思決定者が確実にデータを利用できるようにします。データレイクでは、raw データと処理済みのデータを任意の形式で保存し、アプリケーションおよびエンドユーザーの需要に応じて、後日、転送および変換できます。この概念の背後には、データに適用される分析や質問がまだ特定されていない可能性があるため、未来のイノベーションに備え、すぐにアクセスできる環境にデータを保持しておく、という考え方があります。

エンタープライズレベルでのデータサイロの回避

しかし、他の主要なエンタープライズデータイニシアチブと同様、このコンセプトをエンタープライズに売り込む必要がありますデータレイクでは、さまざまなソースからのデータを取り込みます。そしてデータ統合とセキュリティに必要なすべての要件を備え、すべてを 1 か所に保存します。データレイクは、永遠の課題であったデータサイロに対する答えであり、さまざまに分断された環境を乗り越えて、すべてのデータを最終的に 1 か所で維持します。また、データレイクにより、ビッグデータを構成するさまざまなデータフォーマットを即座に処理したり、統合したりする必要がなくなります。

詳細については、AWS でのデータレイクと分析をご覧ください。

本資料に関する簡単なアンケートにご協力ください。
アンケートはこちら »

AWS について

これまでの 13 年間、アマゾン ウェブ サービスは世界で最も包括的で広く導入されたクラウドプラットフォームとして活躍してきました。AWS では、米国、オーストラリア、ブラジル、カナダ、中国、フランス、ドイツ、インド、アイルランド、日本、韓国、シンガポール、スウェーデン、英国にまたがる 20 の地理的リージョンにある 61 のアベイラビリティーゾーン (AZ) から、コンピューティング、ストレージ、データベース、ネットワーク、分析、ロボティクス、機械学習と人工知能、IoT、モバイル、セキュリティ、ハイブリッド、バーチャルリアリティと拡張現実 (VR と AR)、メディア、アプリケーション開発、デプロイ、マネジメント向けの数多くの機能を備えた、165 以上のサービスを提供しています。急速に成長するスタートアップ企業から最大級のエンタープライズ、および先進的な政府機関に至るまで、何百万ものお客様がインフラストラクチャの実現に AWS を使用して俊敏性を高め、コストの削減を実現しています。AWS の詳細については、**https://aws.amazon.com** を参照してください。