データ ウェアハウスとデータ レイクの融合

レイクハウスは多くのユース ケースに対応するように進化する

ホワイト
ペーパー

VENTANA RESEARCH

後援：

databricks

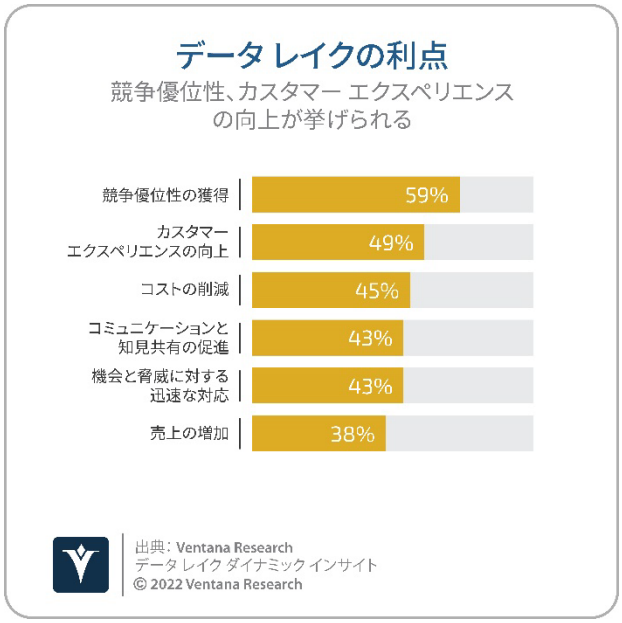
tableau

目次

データ ウェアハウスとデータ レイクの価値

これまで何十年もの間、組織では、組織内のさまざまな部分から情報を引き出して分析を行う必要性を認識してきました。製品の収益性を分析するには、生産コスト、販売コスト、カスタマー サービス コストといった情報が必要です。財務計画には、販売情報、業務情報、マーケティング情報、および従業員情報が必要です。これらのさまざまな情報源をまとめることで、一貫性のある情報のセットに対して、豊富な分析を実行するのが容易になります。当社の調査では、圧倒的多数 (91%) の組織が、分析によって事業活動やプロセスが改善されたと報告しています。データ ウェアハウスの利点は明らかであるため、長い間、企業の情報アーキテクチャの基礎的な構成要素として活用されています。ビッグ データを収集して保管することが多くの企業の標準的なアプローチとなるにつれ、データを中央のリポジトリに統合するというコンセプトの延長線としてデータ レイクが作成されるようになりました。

データ レイクには、データ ウェアハウスと同様、多くの利点があります。今回の調査によると、データ レイクから得られた利点として各組織が挙げた中で最も多かったのは、競争優位性を得られることでした。また、カスタマー エクスペリエンスの向上や、売上増加やコスト削減による収益の改善も報告されています。さらに、データ レイクが市場の機会や脅威への迅速な対応に役立っているとも報告しています。これらの利点の主な理由は、データ レイクでは詳細な情報が得られるので、他の方法では不可能な分析が可能になることです。たとえば、多くの予測分析には詳細なデータが必要であり、一般的にデータ ウェアハウスで利用可能な集計データでは正確に実行することができません。数百万人の顧客を持つあるグローバルなテクノロジー メディア企業は、ビデオ音声アプリケーションから生のテレメトリー データをデータ レイクに収集しています。このデータを人工知能と機械学習 (AI/ML) の技術を使用して分析することで、パーソナライズされた視聴者体験を生み出すことに成功して賞を獲得しました。



データ ソースを 1 つにまとめると管理と統制が容易になりますが、データ ウェアハウスとデータ レイクは、データ品質、データ一貫性、データ アクセスを管理するための一元的な場所として機能します。また、分析に必要な情報をどこで探せばよいかという混乱を解消し、データに素早くアクセスできるように調整や最適化を行うことができます。さらに上記で指摘したように、幅広いデータ ソースからデータ ウェアハウスやデータ レイクにフィードすることで、高度な分析を行うことが可能になります。

アーキテクチャにおける持続的な課題

データ ウェアハウスとデータ レイクは、それぞれ異なる目的で設計されています。データ ウェアハウスは構造化されたリレーショナルなデータ テーブルを扱うために設計され、データ レイクは膨大な量の生の詳細なデータを扱うために設計されています。データ ウェアハウスは一般的に、集計データ (製品別、顧客別、地域別の日次合計売上高など) を扱います。データ レイクは、非構造化データ (テキスト、画像、音声、ビデオ、ログ ファイルなど) を収集して管理します。

データ ウェアハウスは構造化されたリレーショナルなデータ テーブルを扱うために設計され、データ レイクは膨大な量の生の詳細なデータを扱うために設計されています。

データ ウェアハウスは、アドホック クエリ、レポート作成、ダッシュボード、セルフサービスによるインタラクティブな可視化など、特定の種類の分析を行うために設計されています。データ ウェアハウスは通常、定期的にソース システムから情報を取得するバッチ処理に依存しています。データがデータ ウェアハウスにロードされると、他のバッチ処理が実行されて合計が集計され、組織で必要とされる一般的なレポートや可視化の多くが迅速に処理されるようになります。これらのバッチ処理は完了までに数時間かかることが多く、今日多くの組織が必要とするリアルタイム処理とは根本的に合致しません。

さらに、データ ウェアハウスにデータを追加するには、一連の複雑なデータ処理とデータ準備を行うパイプラインが必要です。まず、ソース システムから情報を抽出し、この情報をクレンジングして異なるシステムから取得したデータの不整合を解決する必要があります。次に、データを分析のために変換したり準備したりする必要があります。たとえば、難解なシステム コードを、より理解しやすい値や計算で得られたメトリックに変換します。多くの場合、これらの準備プロセスにはそれぞれ別個のテクノロジーが使用されるため、運用コストと管理コストが増加します。

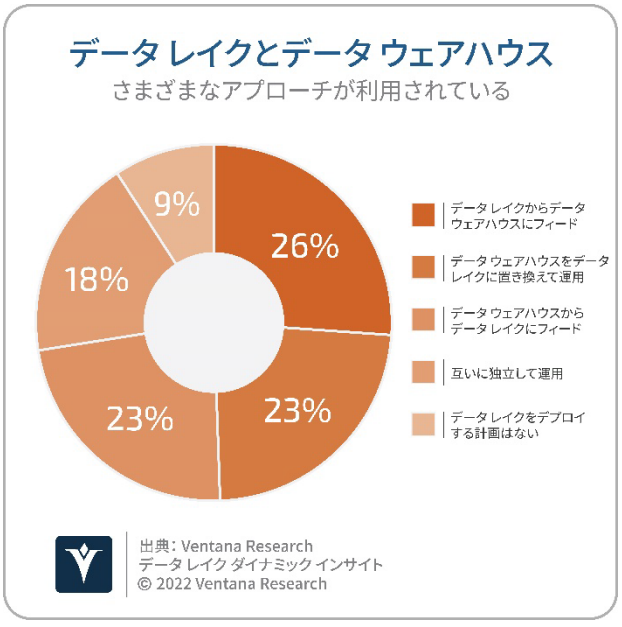
データ レイクは、データ探索、予測モデリング、自動意思決定など、データ ウェアハウスとは異なるタイプの分析を行うために設計されています。また、データ ウェアハウスには適さない種類のデータを扱うことができる柔軟性もあり、人気が高まっています。しかし、必ずしも万能ではありません。データ レイクには生データが入るため、データの質はそのデータを生成するシステムの質によって決まります。複数のソースから集められたデータは、一貫性を保つために合理化する必要があります。また、大量のデータに対するクエリは、インタラクティブな分析を行うには十分な速度で処理されないことが多いため、データ量が課題になります。

レイクハウスのアプローチは両者の長所を兼ね備えている

データ ウェアハウスとデータ レイクを組み合わせることで、両者の長所を活用することができます。これらを結合したアーキテクチャを利用すると、構造化データと非構造化データの両方がサポートされ、生の詳細データだけでなく、変換して集計されたデータも提供されます。このアーキテクチャは、データのリアルタイム処理とストリーミングを必要とするアプリケーションや分析に対応するだけでなく、あらゆるデータを分析する組織のデジタル変革の取り組みを支援するために必要な拡張性も備えています。

たとえば、1 日の訪問者数が 50 万人、商品数が 40 万点のヨーロッパのオンライン小売業者は、レイクハウスのアプローチを使用して 10 億行以上あるダッシュボードを提供し、分析に必要なすべての関連情報を確実に入手できるようにしています。この実装はクラウド上に展開されているので、クラスタ管理が簡素化され、あらゆる規模での運用がサポートされます。機械学習のネイティブ サポートにより、データ サイエンス チームは、ダッシュボードとインタラクティブな可視化をサポートする同じインフラストラクチャ上で、簡単かつ迅速にモデルの開発、展開、および追跡を行うことができます。この小売業者は、顧客エンゲージメントを高め、よりパーソナライズされたコンテンツを提供することにより、最終的に売上を倍増させることができました。

今回の調査によると、この小売業者が行ったように、多くの組織がこの 2 つのアプローチを組み合わせようとしています。ほぼ 4 分の 3 (73%) の組織が、何らかの形でデータ ウェアハウスとデータ レイクを結合させています。4 分の 1 (26%) の組織は、データ レイクからデータ ウェアハウスにデータをフィードしています。さらに 4 分の 1 (23%) の組織は、データ ウェアハウスからデータ レイクにフィードしています。また、4 分の 1 (23%) の組織では、データ ウェアハウスをデータ レイクに置き換えて運用しています。この 3 番目のシナリオを表す「レイクハウス」という用語はここから生まれました。



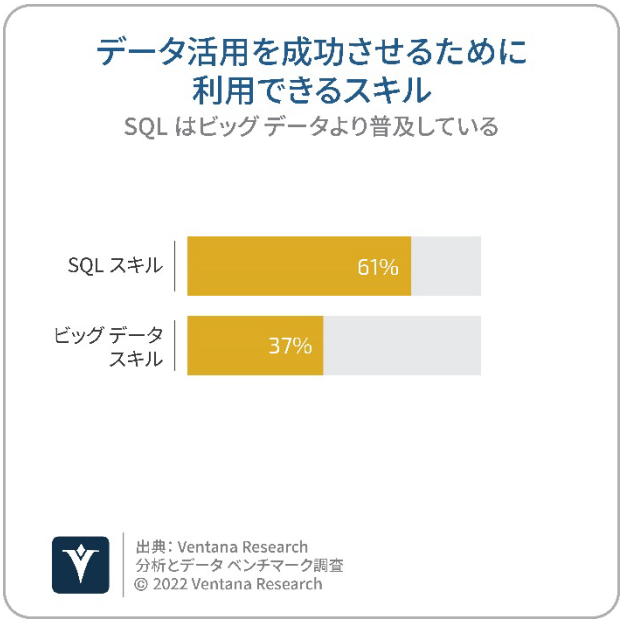
このような結合されたビッグ データの実装を行うときのアーキテクチャとして、クラウド上のオブジェクト ストレージが採用されるようになっています。多くの場合、ビッグ データの実装には、多数のノードからなるクラスタが必要になるため、その取得に時間がかかり、設定も複雑になりがちです。クラウド ベースの実装では、クラウド プロバイダーが複雑な部分の多くを管理し、数分でクラスタが利用可能になる

ため、これらの問題の多くが解決されます。オブジェクト ストレージは、大量のデータを管理するための低コストでスケーラブルなプラットフォームを提供します。その結果、今回の調査では、ビッグ データを扱う組織の 3 分の 2 (65%) が、実稼働環境でオブジェクト ストアを使用していると報告しています。

オブジェクト ストレージを基盤とするレイクハウスのもう 1 つの利点として、オープン性が挙げられます。オブジェクト ストレージでは、Apache Hudi、Apache Iceberg、Delta Lake など、さまざまなオープン ソースのファイル形式を利用することができます。これらのファイル形式は、機械学習やデータ エンジニアリングなどのさまざまなスキル セットをサポートするとともに、R や Python などのさまざまなツールにも対応します。また、これらのファイルは、他のツールから直接アクセスすることができ、組織がデータを管理および統制する方法の選択肢が広がります。

このような利点がある一方で、単一のクラウド プロバイダーとそのオブジェクト ストレージ モデルだけに依存しすぎるのは危険な場合もあります。実際、今回の調査では、約半数 (42%) の参加企業が複数のクラウド プロバイダーを利用しています。基盤となるストレージ モデルに依存しないオープン レイクハウス アーキテクチャを採用することで、柔軟性、移植性、相互運用性を高めることができます。また、分析やビジネス インテリジェンス (BI) を始め AI/ML に至るまで、さまざまなユース ケースをサポートできるようにもなります。

レイクハウスでの最新の分析機能を使ったアクセスを民主化する



各組織は、データと分析を広く利用できるようにするために、まだ苦慮しています。今回の調査では、4 分の 3 (72%) の組織で、分析を利用しているのは従業員の半分以下でした。残念ながら
各組織は、ビッグ データを分析する能力に関しては、他のソースからのデータを分析する場合ほどは自信を持っていません。レイクハウスには SQL アクセスが組み込まれているため、各組織はより簡単にビッグ データにアクセスして分析することができます。レイクハウスの SQL は、いくつかのベンダーが性能と機能の幅を示す TPC ベンチマークを発表するまでに進歩しました。レイクハウスでの SQL アクセスが重要なのは、必要な SQL スキルを有していると報告した組織が 3 分の 2 近く (61%) あるのに対し、必要なビッグ データ スキルを有していると報告した組織は 3 分の 1 強 (37%) しかないためです。

各組織は、一貫性のある単一のデータ セットで運用することで、パフォーマンスを高めることができます。データ レイクの使用に関して報告された最大の利点の 1 つは、組織全体のコミュニケーションと知識の共有が向上することです。残念ながら、分析プロセスで最も時間がかかるのはデータの準備であると、3 分の 2 以上 (69%) の組織が報告しています。また、IT 部門が統合または準備していないデータをビジネス ユーザーが扱えるようにすることに抵抗がないと回答したのは、調査参加企業の半数弱 (46%) でした。レイクハウスは、データ サイエンスなどの分析プロセス用にデータを統合および準備する手段を IT 部門に提供することで、これらの課題に対処します。

ある例では、数千億円規模の金融サービス企業がレイクハウスのアプローチを利用して、自社の標準的なビジネス インテリジェンス ツール (モバイル アクセスを含む) からデータを利用できるようにしました。これにより従業員は、以前なら作業チケットを申請して数日または数週間後に取得していたメトリックに、セルフサービス方式でアクセスできるようになりました。さらに、この組織では、SQL、Python、R、Scala を同じノートブックに混在させることが可能です。また、データ サイエンス モデルの結果をレイクハウスにアンロードして、セルフサービスのダッシュボードで利用できるようにしています。レイクハウスをビジネス インテリジェンス ツールと統合することで、多くの組織が苦労しているセルフサービス アプローチを実現することができています。

> **レイクハウスのアプローチを採用すると、データ ウェアハウスとデータ レイクの両方を管理する必要がなくなります。**

セルフサービスと民主化の妨げとなる問題のいくつかは、データ ガバナンスに関わるものです。データ レイクとデータ ウェアハウスを別々に導入した場合、まったく異なるポリシーを 2 セット維持する必要がありますが、それを実施するには、多くの場合それぞれ異なるスキルが必要とされます。また、このアプローチでは、2 つの環境間で共有されるデータやユーザーの労力が二重に必要となります。レイクハウスのアプローチを採用すると、データ ウェアハウスとデータ レイクの両方を管理する必要がなくなります。単一の一貫したモデルでアクセスやガバナンスを管理することができます。その結果、各組織はさまざまな分析、BI、および AI/ML のユースケースをより簡単に管理できるようになり、セルフサービスとその利点を実現しやすくなります。

ビジネス インテリジェンスの先を見据える

レイクハウスのアプローチには、多くの組織が依存しているビジネス インテリジェンス ツールと分析ツールをサポートする機能が含まれています。レポートとダッシュボードは、これらの分析の多くのバックボーンを形成し、今回の調査に参加した組織の 80% 以上で使用されています。

レイクハウスでこれらの機能やインタラクティブなビジュアライゼーションなどの他の分析機能を提供することには、いくつかの利点があります。データを直接レイクハウスにストリーミングできるため、レイテンシが小さくなり、データの鮮度が高まります。また、データの維持や更新の対象となるコピーの数も少なくて済みます。さらに、BI 用に別個のウェアハウスを用意する必要がないため、ソフトウェア ライセンスのコスト、コンピューティング インフラストラクチャのコスト、メンテナンスのコストを削減することができます。

しかし、多くの組織は BI の先を見据えて動いています。AI/ML ベースの分析は、必須要件となっています。調査参加企業のほぼ 9 割 (87%) が、すでに AI/ML を使用しているか、使用を計画しています。たとえば、米国のある大手ヘルスケア企業は、レイクハウスを利用して、機械学習により顧客のその時のニーズを特定し、よりパーソナライズされた体験を提供しています。予測モデルを使用してパーソナライゼーションを強化した結果、服薬遵守が 1.6% 向上しました。つまり、指示に従って時間通りに服薬する患者が増えました。これは、場合によっては生死を左右することもあるので、大きな改善です。この種の分析の多くは、データ ウェアハウスに保管されているデータでは実行できません。一般にデータ ウェアハウスには必要な詳細データが含まれていないためです。AI/ML が検出する相関関係の多くは、どの商品が一緒に購入されているかなど、詳細なレベルにしか存在しないため、詳細なデータが必要になります。同様に、顧客感情の高度な分析には、カスタマー サービスで行ったやりとりのテキストなどの非構造化データが必要です。また、大量の画像データを収集および処理できるデータ アーキテクチャの進化に伴い、画像処理も一般的になってきています。

データ ウェアハウスでは容易に対応できない新たな分析分野として、ストリーミング データがあります。しかし、リアルタイム分析を行うための十分なテクノロジーを備えていると報告している組織は、わずか半数 (52%) にすぎません。ストリーミング データは、ゆるく構造化されたファイルや情報の流れとして届くことが多いため、リレーショナル データベースの行と列に必ずしもきれいに収まるとは限りません。データのストリームを到着と同時にリアルタイムで処理することで、差し迫った状況に反応し、手遅れにならないうちに対応することができます。

各組織は、さまざまなストリーミングのユース ケースに直面します。前述のように、レイクハウスに供給するデータ パイプラインをストリーミング機能で強化することにより、データ レイテンシが小さくなり、より一貫性のある最新の業務情報を提供することが可能になります。データが生成される際にイベント ストリーム処理を行えば、在庫がなくなった製品を販売してしまうなど、カスタマー エクスペリエンスを損なうような状況を回避できます。また、リアルタイム ストリーム分析を行うと、不正行為の検出や予知保全などのリアルタイムのユース ケースをサポートすることができます。

このソリューションをレイクハウス、データ レイクなどと呼ぶかは別として、データ ウェアハウスとデータ レイクの両方の機能を活用することが重要です。これにより、BI、AI/ML、リアルタイムなど、組織が必要とするあらゆる分析に対応し、現在および将来のワークロードをサポートする

ために必要なコスト効率の高いスケーリングが可能になります。この 2 つの世界を融合させ、組織のデータの価値を最大化しましょう。

次のステップ

- データ レイクの利点を確認し、このアーキテクチャがまだ導入されていない場合は、それを導入する。
- アナリスト、データ エンジニア、データ サイエンティストなど、すべての関係者のニーズを把握する。
- データ レイクとデータ ウェアハウスを統合するための選択肢を評価する。
- データ レイクとデータ ウェアハウスの両方を活用するために必要な労力を最小限に抑えるために、レイクハウスのアプローチを検討する。
- 収集したすべてのデータを、ビジネス インテリジェンスやデータ サイエンス技術で利用できるようにすることでデータの価値を最大化する。

Ventana Research について

Ventana Research は、もっとも権威があり信頼されている、ビジネス技術リサーチ/アドバイス サービス会社です。当社は、ベンチマーク調査や技術調査評価、教育ワークショップ、および当社の調査/アドバイス サービス Ventana On-Demand などを含む、一連の独自調査に基づく情報を通じて、主流となっている技術や革新的な技術に関する洞察と専門的なガイダンスを提供しています。当社の他に類を見ない、ビジネス プロセスとパフォーマンスの最適化に技術が果たす役割についての理解とベスト プラクティスは、各業界のビジネス/IT 機能に渡る人々、プロセス、情報、および技術に対する厳格な調査に基づくベンチマークを基盤にしています。このベンチマーク調査と当社のマーケットのカバー範囲、および数百件もの技術プロバイダーに対する深い専門知識が、技術への投資から得られる価値を増強しながらコストやリスクを低減するための教育と専門知識を当社のお客様に提供できることを証明しています。

Ventana Research は、業界最高の総合的なアナリストと調査範囲を提供しています。世界中のビジネス/IT プロフェッショナルが当社コミュニティのメンバーとなっており、世界各国の高く評価されているメディアや関連パートナーのように、Ventana Research からの恩恵を受けています。当社の認識と分析は、Twitter、Facebook、および LinkedIn などのブログやソーシャル メディア チャネルを通じて毎日配信されています。

Ventana Research が、ベンチマーク調査、教育、およびアドバイザー サービスを介して、どのように企業における情報や技術の使用を進歩させるのかについては、www.ventanaresearch.com をご覧ください。