

ガバナンスの効いたデータ・レイクとは

セルフサービスによるデータ・アクセス機能を拡張すると、データ分析をスピーディーに実行し、アクションにつながる洞察を獲得できます



特長

- 社内ユーザーがセルフサービス形式で幅広いデータに迅速にアクセスできる環境を実現します
- 分析を迅速に実行し、洞察の精度を高めることができます
- 大量データをデータ形式に問わられることなく、コスト効率と拡張性の高い環境に蓄積できます
- データの品質、セキュリティ、ガバナンスを維持できます

現代の企業は大量のデータを収集し、分析を強化することで、拡大する情報源からデータに基づく新たな洞察を生み出しています。増大するデータを最大限に活用するには、全社的にデータへのスピーディーなアクセスを提供する必要があります。同時に、長期的なデータの保存と管理を行うための効率的で使いやすい方法も必要です。

ガバナンスの効いたデータ・レイクを活用すると、このような課題に対応することができます。データ・レイクとは複数のリポジトリから成る共有データ環境のことで、ビッグデータ・テクノロジーを活用するためのものです。俊敏性を持ちながらセキュアで制御されたデータ環境（構造化データと非構造化データの両方に対応）のニーズを満たすために、企業はますますデータ・レイクの活用を模索するようになっていきます。

データウェアハウスとは異なり、必要になるまでデータをそのままの形式で保存することができます。データを迅速に配置・保存することができ、セルフサービスでデータにアクセスし分析できる柔軟で使いやすい環境を提供します。包括的なガバナンス機能があるため、データを重複させることなくデータを簡単に発見・理解・保存することができます。



ガバナンスの効いたデータ・レイクの活用

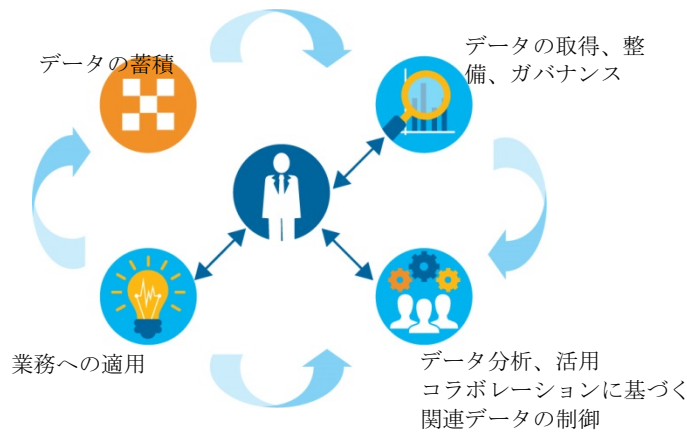


図 1. ガバナンスの効いたデータ・レイクのコンポーネントは効率的にデータの処理と発見を行うため、ユーザーはアナリティクス、レポートニング、意思決定の基盤としてデータ・レイクをより簡単に活用できます。

データ・レイクを通じて主要な課題に対応する

データ・レイクは、企業がビジネス戦略と IT 戦略に関するいくつかの課題に対応するために不可欠な役割を果たします。

- ビジネス・ユーザー、アナリスト、データ・サイエンティスト、開発者が、正確で意味のある洞察を生み出すことができるよう、タイムリーに包括的なデータにアクセスできる環境を実現する。
- スピーディーにデータの準備と分析を行うことで、ビジネスが求めるスピードに対応する。
- データ・サイエンティスト (より深いデータ分析機能が必要な場合)、およびデータ・エンジニア (データ・レイク・ベースのアプリケーションをビジネス部門のユーザーに提供する必要がある場合) の間でコラボレーションを促進する。
- データの品質、セキュリティ、ガバナンスを保証することにより、プライバシーの保護と法令遵守を実現しながら、信頼性が高くわかりやすいデータをユーザーに提供する。
- スケーラブルなデータ蓄積基盤で、急速に増大する一連のデータに対応することで、既存のリソースを最大限に活用しながら、コスト効率良くデータを処理する。

データ・スワンプ (“未整備の沼地”) を回避する

最も効果的なデータ・レイクは、積極的にデータの取り込み、ガバナンス、保護、管理を行う一連の堅牢なコンポーネントを備えています。データ・レイク (湖) を最適化し「データ・スワンプ (“未整備の沼地”)」を回避するには、データ・レイク構築へのプランニングが不可欠です。その名が示すとおり、データ・スワンプとは、すべての利用可能なデータをフル活用できるよう整理、可視化、管理されていない環境を指します。

データ・スワンプでは、ユーザーはデータの生成元やデータの信頼性に自信が持てません。必要なデータが見つかるかどうかかわからず、リポジトリに必要なデータが存在しているかどうかすらわかりません。さらに、データが適切に保護されているかどうか確認を持つことができません。また、通常データに関するビジネス上のコンテキストを確認することもできません。

データ・レイクが提供する機能とは何か

データ・レイクは、以下の機能や価値を提供する環境を指します。

- ユーザーが大量の生データにアクセスできる環境
- 分析モデルを開発・検証後、本番環境に移動するための環境
- データの探索を行い、洞察を獲得するための分析用サンドボックス (試行錯誤できる環境)
- ユーザーがデータを発見し、ビジネス用語と技術的なメタデータを関連付けるために役立つ全社的なデータ・カタログ
- データの加工とクエリーの再利用が行える環境

以下は、データ・レイクだからこそ提供可能な機能、価値ではありません。

- 社内のすべてのデータを保存するデータウェアハウスまたはデータ・マート
- オペレーショナル・データ・ストア (ODS) の代替環境
- 高パフォーマンスな本番環境
- 稼働状況のレポートニング・アプリケーション
- 特定の問題を解決できるよう特定の目的のために構築されたシステム (ただし、データ・レイクから特定目的のために構築されたデータ・マートにデータをフィードすることは可能)

ベスト・プラクティスの活用により、データ・レイクを最適化する

以下のステップに従いベスト・プラクティスを活用すると、企業はデータ・レイクを最適化し、データ・スワンプによる問題を回避できます。

データ・リポジトリを構築する

リポジトリは企業に構造化データと非構造化データを提供します。各リポジトリは独自のワークロード機能を提供するか、一連のデータを独自の観点で処理します。複数のリポジトリはさまざまな種類のデータを保存することができます。企業は必要に応じて、新規のリポジトリを追加し、古いリポジトリを削除する必要があります。

データ・レイク・サービスを実装する

データ・レイク・サービスは、アナリスト、データ・サイエンティスト、開発者、ビジネス・ユーザーによるデータ・レイク・リポジトリへのアクセスを制御し、サポートします。また、これらのサービスによりデータを常に同期することができます。データ・レイク・サービスがデータ・カタログを保有するため、ユーザーは必要なデータを見つけ出し、作業に適しているか検証することができます。また、データ・カタログはデータ・ソースを検証するためのデータ・リネージュ（来歴）を提供するため、ユーザーがデータに対する信頼性を高めることができます。

包括的な情報管理と情報ガバナンスのモデルを構築する

データ・レイク・サービスには、情報管理機能と情報ガバナンス機能を提供することに特化したミドルウェアのサポートが必要です。このミドルウェアには以下の機能が必要です。

- データの移動と加工を行うプロビジョニング・エンジン
- データを処理するユーザー間のコラボレーションを促進するワークフロー・エンジン
- モニタリング機能、アクセス制御機能、監査機能

データ・レイクの活用により、幅広いメリットを享受する

データ・レイクを最適な形に実装すると、企業はデータ環境から実効性のあるビジネス・バリューを効率的に導き出すことができます。データ・レイクを構築することで得られるメリットとしては、以下が挙げられます。

- **社内の幅広い種類のデータに簡単にアクセスできるようになる:** ユーザーはオンプレミスとクラウドの環境に存在する構造化データと非構造化データにアクセスできます。時間のかかる処理を IT 部門に依頼することなく、必要なタイミングで必要なデータに自らアクセスできるようになります。
- **スピーディーにデータを準備する:** データ・レイクは、いくつかの手法を通じてスピーディーにデータの準備を行います。例えば、データ・カタログを実装するとデータに関する知識と理解を深めることができ、その結果、スピーディーにデータの準備を行うことができるようになります。さらに、ハイブリッド・クラウド・インフラを持つデータ・レイクを構築すると、使用目的に最も合致するプラットフォームでデータを保存できるようになります。データが最適なロケーションに存在すると、より短時間でデータへのアクセスと探索を行うことができるため、データの準備と再利用の作業をスピードアップできます。
- **俊敏性を高める:** データの準備がスピードアップすると、ユーザーはさらに多くのデータを探索できます。ユーザーがより俊敏に分析モデルの構築・テストを行うために、サンドボックスを活用することができます。さまざまな方法で分析を試みることができ、場合によっては「試行錯誤」をすばやく実施することで、より迅速に最も生産性の高い分析方法に移行できます。
- **より正確な洞察を提供し、より効果的な意思決定を行う:** データ・レイクを通じてユーザーがより多くのデータにアクセスし、データの準備をスピードアップし、さまざまなデータの処理を試みることができるため、企業はより精度の高い洞察を生み出すことができます。適切に構築したデータ・レイクはデータ・リネージュ（来歴）を管理するため、データの信頼性を担保することができます。このような機能のすべてが、企業によるビジネス上の意思決定の改善につながります。



© Copyright IBM Corporation 2016

日本アイ・ビー・エム株式会社
IBM Analytics

〒103-8510
東京都中央区日本橋箱崎町 19 番 21 号

Produced in Japan
2016 年 7 月

IBM、IBM ロゴ、および ibm.com は、世界の多くの国で登録された International Business Machines Corporation の商標です。他の製品名およびサービス名等は、それぞれ IBM または各社の商標である場合があります。現時点での IBM の商標リストについては、<http://www.ibm.com/legal/copytrade.shtml> をご覧ください。

本書の情報は最初の発行日の時点で得られるものであり、予告なしに変更される場合があります。すべての製品が、IBM が営業を行っているすべての国において利用可能なものではありません。

本書に掲載されている情報は特定物として現存するままの状態を提供され、第三者の権利の不侵害の保証、商品性の保証、特定目的適合性の保証および法律上の瑕疵担保責任を含むすべての明示もしくは黙示の保証責任なしで提供されています。IBM 製品は、IBM 所定の契約書の条項に基づき保証されます。

お客様は自己の責任で関連法規を遵守しなければならないものとします。IBM は法律上の助言を提供することはいたしません。また、IBM のサービスまたは製品が、お客様がいかなる法規も遵守されていることの裏付けとなると表明するものでも、保証するものでもありません。



Please Recycle

同時に、データ・レイクは IT 部門にもいくつかの重要なメリットを提供します。例えば、データ・レイクを活用することで IT 部門は継続的なデータの増大に備えることができます。データ・レイク・アーキテクチャーに Hadoop を組み込むと、IT 部門は大規模な分析が実行できるスケーラビリティとコストメリットの高い環境を実現できます (ただし、処理のスピードアップは実現しない場合があります)。緊急のビジネス上の課題に応えるためのリソースを提供するだけでなく、データ・レイクは手ごろなコストでアクセスの頻度の低いデータに関する検索可能なアーカイブを提供することもできます。ハイブリッド・クラウド環境でデータ・レイクを構築すると、企業は投資コストを削減したうえで迅速にリソースを追加できます。

また多くの場合、データ・レイクを活用すると、大きなコストを伴うエンタープライズ・データウェアハウス (EDW) のリソースが不要になるため、EDW はその他の役割 (パフォーマンスの履歴データの監視と分析を行うビジネスアナリストのサポートなど) をより効率的に果たすことができます。データ・レイクは EDW のサービス・レベル・アグリーメントに影響を及ぼすことなくデータ・リクエストを処理することで、セルフサービスによるアナリティクスを実現します。

迅速なアナリティクスと俊敏な意思決定を実現する

データ・レイクは、現在利用可能な大量のデータを活用するための非常に有効な手法を提供します。データ・レイクを実現する主なベスト・プラクティスを活用することで、企業はデータの信頼性とセキュリティを確保したうえで、全社的に幅広い種類のデータへの柔軟なアクセスを提供することができます。データ・レイクを企業の特定のニーズに合わせて最適化すると、ビジネスの俊敏性を高め、意思決定を改善するために大いに役立ちます。

詳細情報

効果的なデータ・レイクの実装によるメリットとビジネス・チャンスを確認するには、ibm.biz/data_lake_j の Web サイトをご参照ください。