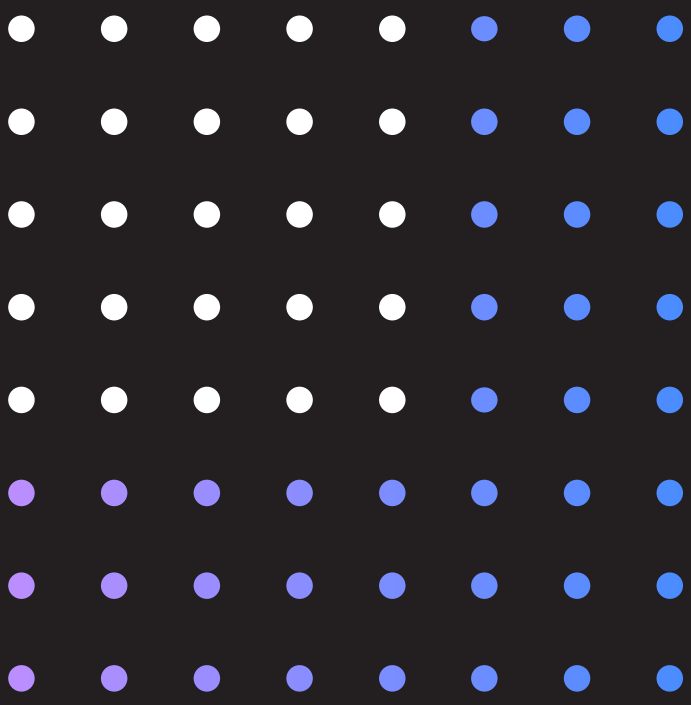
インテリジェント・データ・カタログおよびデータレイク・ガバナンスによるビジネス対応データの提供

IBM Watson Knowledge Catalog は、機械学習を利用したデータ・ガバナンス・プラットフォームによってデータレイクの課題解決を支援します。



IBM

目次

要点

- 信頼できる洞察を得ようとデータを保存、分析するための
  データレイクを構築しても、そこから期待どおりの価値を
  実現している組織はほとんどありません。
- 企業および規制上のポリシーを遵守しつつ、データ・アクセス、
  準備、統合、および利用者への提供を行うのは非効率的です。
  DataOps は、そうした非効率性に伴って組織が直面する
  問題を解決します。
- 一般的なデータレイクの課題として、データレイクへの新規
  データ・ソースのインポートが複雑でコストがかかる、内部と
  外部のデータ・セットを理解できない、データ・ガバナンスの
  信頼性が低い、セルフサービス・データ準備ツールを使用
  できない、データレイク内のデータを検索および理解できない、
  などが挙げられます。
- カタログ化、データ品質、およびデータ検出機能を備えた
  エンタープライズ・データ・ガバナンス・プラットフォームは、
  不完全なデータレイク・プロジェクトを刷新してビジネス
  価値の真の発信源にすることができます。
- IBM Watson Knowledge Catalog は IBM Cloud Pak™ for
  Data 上で利用でき、データの検出、カタログ化、品質確保、
  およびガバナンスのための機械学習 (ML) カタログを提供
  します。データ利用者はデータ資産、データ・セット、分析
  モデルを素早く検出、キュレート、分類、共有できます。
- 組織が自社のデータを深く理解できていないと、ML やディープ・
  ラーニングを初めとする様々なタイプの人工知能 (AI) を
  用いて情報を信頼しつつ利用することがますます難しくなります。

DataOps アプローチによるデータレイクの課題解決

10 年前、すべてのエンタープライズ・データを格納できる中央データ・ストアを構築するための、柔軟で汎用的なアプローチを求める声が高まり、その解決策がデータレイクでした。データレイクとは、事実上どのタイプのデータも保存できる、多目的のデータ・ストレージ環境です。データレイクを使用することで、ビジネス・アナリストやデータ・サイエンティストはデータ・セットを移動することなく、それぞれのデータ・セットに最も適した分析エンジンやツールを適用することもできます。

通常、これらのデータレイクは Apache Hadoop と Hadoop Distributed File System (HDFS) に、Apache Hive や Apache Spark などのエンジンを組み合わせて構築されていました。データレイクが拡大するにつれて、一連の問題が顕在化してきました。このテクノロジーでは、構造化および非構造化データの膨大で多様なコレクションを収集、保存、分析するために物理的な拡張を行うことはできましたが、そうした機能をどのように業務ワークフローに組み込むかという現実的な問題にはほとんど注意が向けられていませんでした。

データレイク・プロジェクトの 80% 超は、アナリティクスやデータ・サイエンスの成功にとって必要なデータの検出、インベントリーおよびキュレーションが最大の阻害要因となり、2022 年までには価値を提供できなくなるでしょう。1 その結果、「どのデータをデータレイクに保存すればよいか?」、「そのデータを誰が使用するのか?」、「どうすればデータが検出しやすくなるか?」、「このデータの発信元はどこか?」、「データの誤使用を防止するにはどうすればよいか?」といった疑問が生じ、その多くは答えがありませんでした。人、プロセス、およびテクノロジーの問題への対処においてこうした深刻な制限があっては、データレイクの実装に失敗するのは当然です。

現在、多くの組織は自社の失敗を認識し、データレイク実装の先導チームを刷新して、データレイクを適切に実装するために 2 回目、3 回目、さらには 4 回目の取り組みを始めています。今度の取り組みではデータ運用、すなわち DataOps に重点が置かれています。

このホワイト・ペーパーでは、データレイクで発生する共通の課題を評価し、DataOps などの新しいアプローチを紹介します。これらのアプローチによって、データスワンプなどの問題点を、組織のビジネスに対応したデータ・パイプラインの中心的存在へと転換することができます。

DataOps は、協働的なデータ管理のプラクティスであり、組織内のデータ管理者とデータ利用者間における通信の改善、統合、データ・フローの自動化などに重点を置いています。

DataOps の概要

DataOps は、DevOps やデータ管理、およびデータ・ガバナンスのベスト・プラクティスで構成される共通のフレームワークであり、複数のステークホルダーが関与するデータ・フローの開発および保守を共同で行うための手段となります。企業および規制上のポリシーを遵守しつつ、データ・アクセス、準備、統合、および利用者への提供を行うのは非効率的であり、それに伴って組織が直面する問題を解決するのが DataOps の目的です。このような非効率性は事業部門、分析チーム、さらには運用プロセスでも見受けられます。

DataOps の方法論を実行するには、人、プロセス、テクノロジーの問題に対処する必要があり、それによってデータレイク実装の成否が決まります。テクノロジーの側面から見ると、DataOps はデータの取り込みと統合、データ品質、データ・ガバナンス、およびデータ消費に対応するために緊密に統合されたエンドツーエンドのプラットフォームを使用し、ガバナンスを備えたデータレイクを作成することを重視しています。継続的なデータ・パイプラインを企業全体で維持するために、取り込みプロセスの一環として、データ品質の検証ルールが自動的に実行されます。取り込みプロセスはパイプラインの中心部分を構成するデータ・カタログと完全に統合する必要があります。データ利用者は、データ・カタログからデータ品質スコアやデータ・プロファイルの結果にアクセスでき、組織が関連データ内で同じデータを扱っていることを確信できます。

データ量の増加ペースは、組織がデータから価値を創出する能力を上回りつつあります。'SoI (Systems of Insight; 洞察システム) を使用するうえで最大の課題は何かという質問に対し、1) 40% の組織は、既存のビジネス・プロセスとソース・データを統合し、それを分析すること、2) 39% の組織は、データ量が増加する中でそれらを入手、収集、管理し、ガバナンスを行使すること、と答えています。2 今日では、データレイク・テクノロジーに既に投資した多大な時間と資源を保護するだけではなく、他に代替策がないという事実が重要です。AI の実装から、さらには総合的な分析の実行まで、できる限り多くのデータをあらゆる角度から完全に把握できる機能が不可欠です。つまり、すべてのデータを 1 カ所で保存、分析、管理できるアーキテクチャーが必要です。多くの場合、これらの要件を満たす唯一の現実的な選択肢は、ガバナンスを備えたデータレイクです。

現代の企業は、DataOps 用のビジネス対応データ・パイプラインをサポートすることで、データから価値を創出する方法を見出すことができます。またそうすべきです。

エンタープライズ・データレイクの使用における課題

データの共有
企業内のチームが新しいデータ・セットを入手または作成する場合、大抵はデータの価値やそれに関連する機密性を強く意識します。例えば、そのデータに商業上の機密情報や個人情報 (PII) または顧客データが含まれている場合、チームはその情報の取り扱い方法を理解し、チームの誰もデータを誤使用しないように注意を払います。

さらにはチームの外部、つまりチーム以外のデータ利用者が、データの価値やデータの誤使用によるリスクについて同じ認識を共有していないことも承知しています。当然、こうしたリスクがあることで、チームはデータの共有や管理下にない場所でのデータの保存に対して非常に慎重になります。

これはデータレイクにとっては望ましいことではありません。データレイクを無秩序なデータの集積場としか見ていない企業は、自社にとって貴重なデータをデータレイクに委ねることに極めて消極的です。その結果、企業は別の領域でそのデータのメリットを享受できず、データレイクをエンタープライズ・データを共有するためのセルフサービス・リポジトリーとして使用するというコンセプト全体が破綻してしまいます。

データの統合
チームがデータレイクへのデータの統合に賛同したとしても、そのプロセスはきわめて複雑です。データレイクの本来のコンセプトは、データをそのままの形式でインポートすることであり、従来のデータウェアハウスの複雑な抽出、変換およびロード (ETL; Extract, Transform and Load) プロセスは必要ありません。しかし実際には、ほぼ全てのデータ・ソースは、何らかの意味ある分析に役立てるためにある程度の前処理を必要とします。

そのため、新しいデータ・ソースをデータレイクに統合するのに数カ月を要することもよくあります。さらに、これまでデータの大半はエンタープライズ・システムではなく、小規模な運用上のサイロに保存されていたため、統合するソースは全体で数十、数百にもなる可能性があります。

これは、ビジネス・アナリストやデータ・サイエンティストが必要としている情報が、多くの場合はデータレイクに追加されていない、あるいは数カ月または数年間も追加されない可能性があることを意味します。これもデータレイク導入の大きな障害になります。

データの保存
コモディティー・ストレージや計算リソースのコストはここ数年間で大幅に低下したとはいえ、Hadoop クラスターは無償ではありません。高性能なデータウェアハウス・アプライアンスに大量のデータを保存するよりも、それをデータレイクに保存する方が費用対効果は格段に向上しますが、そのコストは依然として高額です。

さらに、従来どおりデータウェアハウスに保存されているデータと異なり、データレイクに保存されているビッグデータの価値/量の比率は比較的低くなります。ほんのわずかな量の高価値なデータを検出するために、膨大な量のデータを保存しなければならない場合もあります。

データ・サイエンティストにとってどのデータ・セットが有益で価値があるかがわからなければ、データレイクの底に沈んで決して使用されることのないデータを統合して保存するために、多額の資金を投入することにもなりかねません。

データの検出
最も貴重な保存すべきデータ・セットを特定し、それらを共有するようステークホルダーを説得し、データレイクにそれらのデータを適切に統合できたとしても、他のユーザーがそれを正しく

エンタープライズ・データレイクの使用における課題

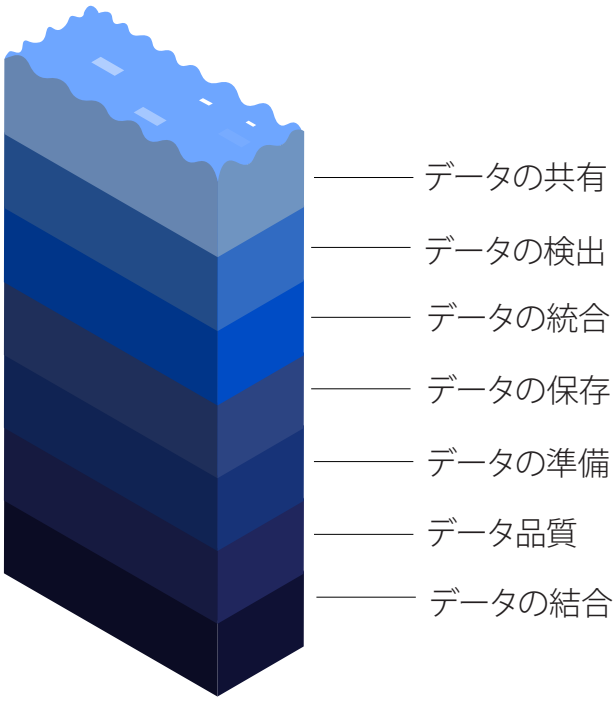


図 1. データレイク・テクノロジーを導入した企業で発生する共通の問題

検出、理解、使用できるようにする必要があります。データレイク内のデータの品質は、さらにもう 1 つの課題です。データ品質の程度が不明であっても、そのデータはデータレイクに追加されます。

残念ながら、ほとんどのデータレイクにおいて、この問題に対処するのは容易ではありません。多くの場合、データは関連データを伴わずに保存されるため、新規ユーザーが元のデータ所有者の助言なしにそのデータから意味を読み取ることは困難であるか、不可能です。用語はそれぞれの分野に特化されていることが多いため、ある事業分野で使用されているメートル法が、他の事業分野ではまったく別の名称で呼ばれていたり、微妙に異なる方法で定義されていたりします。混乱や誤解が生じる可能性が大きいため、多くのデータ・セットはそれをよく知らないアナリストにとってほとんど役に立たないか、危険ですらあります。

内部データと外部データの結合
最後に、最大規模のデータレイクであっても、社内のデータ・サイエンティストが使いたがる全てのデータ・セットを保存しようとしてはいけません。例えば、データ・サイエンティストの 1 人が地理空間を分析したい、あるいはアルゴリズムに天候データや株価を組み込みたいという理由だけで、データレイクに Google マップ、Weather.com®、Bloomberg などの完全なコピーをインポートするのは合理的ではありません。

データレイクには、ビジネス・アナリストが分析に必要なデータがすべて保存されているわけではないため、アナリストは複数のアプリケーションで時間をかけてデータを探さなければなりません。役立つ分析を行うには、ほとんどの場合内部と外部のデータ・

セットを組み合わせる必要があり、これもデータレイク導入の障害となります。また、ユーザーの立場から見た場合は、データレイクに対して認められた価値が下がることになります。

データの準備
データの準備を難しくしている要因は、データの在処の把握からそのフォーマットまでさまざまです。アナリティクスで活用できるようにデータを準備する作業は、データ利用者にとって非常に効率が悪く時間がかかります。データ利用者はデータ分析やモデル化、あるいはビジネスへの影響に関する洞察の取得に注力する代わりに、情報を検出、整理、フォーマットすることに時間の大半を費やします。

ガバナンスを備えたデータ・セットへのアクセスが制限されていることも、準備フェーズで IT への依存度が過度に高まった原因です。このようなアクセスの制限は、セルフサービス機能やデータ・リテラシーのスキルを全社的に強化してこの障害を緩和する必要があることを示しています。

データ品質
データを無造作にデータレイクに追加すると、そのデータは役に立たなくなる可能性があります。データに対してデータ品質ルールまたは検証ルールを適用した後データレイクに追加しないと、データレイクは信頼できる有用なデータを提供することができません。高品質なデータは、意思決定に用いるデータの信頼性を決定づける重要な特性です。データは貴重な資産であり、データが組織から組織に渡る際には十分管理する必要があります。情報源がますます多様化し、規制遵守への取り組みが活発化する中では、これらの多様なデータ・ソースから一貫性があって信頼でき再利用可能な方法で情報を統合しアクセスすることが重要です。

ガバナンスを備えたデータレイク構築のための総合的アプローチ

ほとんどのデータレイクは、オープン・ソース・プロジェクトの Apache Hadoop とその広範囲なエコシステムを活用して、データ・ストレージ層や分析エンジンに対応しています。当然のことながら、Hadoop に関与するオープン・ソース・コミュニティーは現在のデータレイク実装における問題点を認識しており、最近では多様な問題に個別に対応することを目的とした多数のプロジェクトが開始されています。同じく市場には、同様の問題の解決を謳ったさまざまな独自開発のツールが登場しています。

このようなツールを使用して、データレイクの問題を 1 つ 1 つ解決していきたくなります。データ・セットの数が増えすぎて管理できなくなれば、カタログ・ツールを追加します。必要なデータを検出できないという苦情をユーザーから受ければ、検索機能でフロントエンドを強化します。さらに、データ・スチュワードがデータの発生源や使用者を追跡できなくなれば、データ・リネージュ・ツールとデータ・ガバナンス・フレームワークを導入します。

この方法は単純なようですが、実際には、こうした段階的アプローチは複雑さが大幅に増大して保守性が低下するという代償を伴います。データレイクの規模と範囲が拡大するにつれ、特にそれが顕著になります。データレイクに新しいデータ・ソースを追加すると ETL の要件が複雑化するように、新しいツールを追加するとデータレイクの機能的ではない要件が複雑化しがちです。

ビジネス・アナリストが有効利用できるようにデータの統合、品質管理、およびカタログ化を行う統合型のエンドツーエンド・プラットフォームとは異なり、各ツールには多くの場合、それぞれ独自の障害管理やロギングの手法が組み込まれています。その結果、トラブルシューティングや問題解決に多大な時間を要することになりかねません。

データレイクでよく発生する問題を技術面ではなく概念面から見てみると、段階的アプローチにおけるさらに重大なもう 1 つの欠点が明らかになります。ここで重要な点は、拡張性、検出の容易さ、統合、データ品質、およびデータ・ガバナンスは個別の問題ではなく、相互に密接に関連していると認識することです。これらの問題を解決するには、より包括的なアプローチが必要です。

拡張性、検出の容易さ、統合、データ品質、およびデータ・ガバナンスは個別の問題ではなく、相互に密接に関連しています。これらの問題を解決するには、情報管理に対する包括的なアプローチが必要です。

IBM Watson Knowledge Catalog のデータ検出、データ・カタログ、データ品質

IBM Watson Knowledge Catalog は IBM Cloud Pak for Data 上で使用でき、データ資産、データ・セット、分析モデルや、組織の他のメンバーとの関係性を素早く検出、キュレート、分類、および共有するのに役立ちます。IBM Watson Knowledge Catalog を使用して、データ・ガバナンス・チームはビジネス・グロッサリー、ポリシー、およびルールを定義し、ガバナンスに対応した高度なワークフローを構成できます。このカタログは、データ・エンジニア、データ・スチュワード、データ・サイエンティスト、ビジネス・アナリストが信頼できる唯一の情報源としての機能を果たし、彼らが信用し確信を持って使用できるデータへのセルフサービス・アクセスを提供します。

IBM Cloud Pak for Data 上で動作する IBM Watson Knowledge Catalog などのソリューションは、今日のデータレイクにおける主要問題の解決に必要なすべての機能を、単一の総合プラットフォームで提供することができます。メタデータを収集、保存、管理し、データ系列を追跡できる効果的なツールとしては、データレイクには様々な欠点があります。このカタログは、相互に関連する問題の根本原因を突き止めるのに役立ちます。

データレイクの価値は様々な点において、データそのものに依存するのと同じくらい、保存されるメタデータに依存します。データの発生源、作成者、内容、使用を許可されたユーザー、使用方法などを説明するメタデータがなければ、そのデータ自体にはほとんど価値がありません。ユーザーはそのデータを検出できず、もし検出できたとしてもその意味が理解できない、確実に信頼することができない、あるいは使い方がわからないといった状況が生じます。

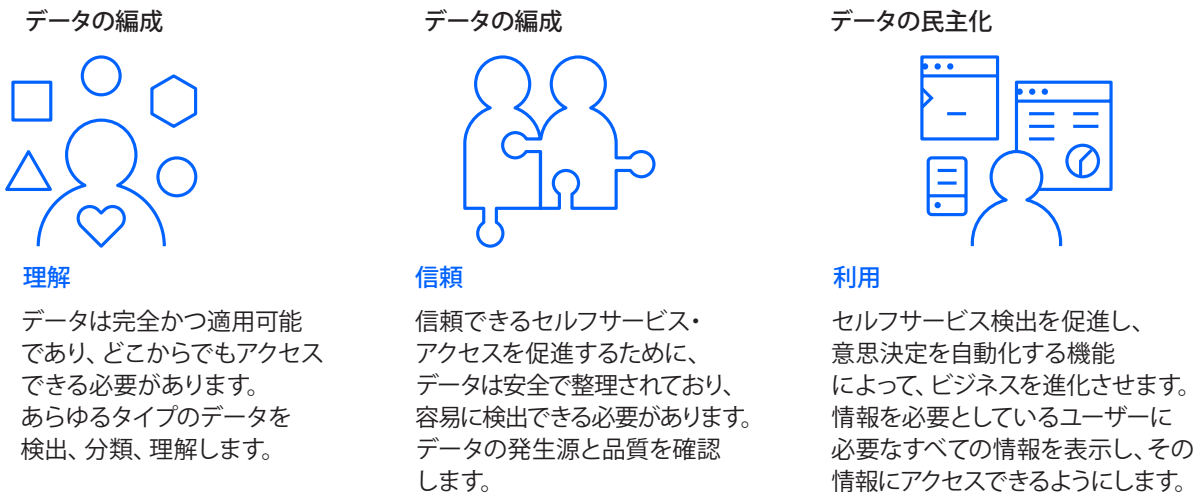


図 2. データ検出、データ・カタログ、データ・ガバナンスに用いる IBM Watson Knowledge Catalog のさまざまな機能

信頼できる唯一の情報源と単一のアクセス・ポイント

IBM Cloud Pak for Data 上で動作する IBM Watson Knowledge Catalog は、メタデータを主要優先事項とすることで前述の問題に対処します。その中心的役割を果たすのが、強力なカタログ・エンジンです。このエンジンはデータレイク、データウェアハウス、トランザクション・システム、あるいはスプレッドシートなど、データがどこに配置されていようと、企業がアクセスできるすべてのデータ・セットや分析資産の索引を作成します。それが構造化、非構造化データであるか、オンプレミスかクラウド上にあるかなどは関係ありません。さらにこのカタログには、企業が利用登録している専用データ・サービスやオープン・データの API など、外部のデータ・セットやデータ・ソースも追加できます。

このデータ・カタログは、すべてのデータ・セットに関する信頼できる唯一の情報源であるだけでなく、単一のアクセス・ポイントにもなります。AI を活用した検索およびサジェスト機能によって、ビジネス・アナリスト、データ・サイエンティスト、データ品質エンジニア、データ・ガバナンス・チームは資産の検出が容易になります。また、使用可能なメタデータが表示されるため、ユーザーは検出したデータの内容を理解し、それが有用かどうかを判断することができます。

組み込み済セルフサービス・データ準備機能を使用すると、アナリティクスおよび AI アプリケーションでの本番使用に向けてデータを変換する時間を短縮できます。ビジネス・アナリストやデータ・サイエンティストがデータの準備や分析に時間を費やす必要はありません。エンタープライズ規模のデータ準備ソリューション (IBM InfoSphere Advanced Data Preparation など) と統合すると、カタログを通じて作成された、ガバナンスを備えたデータ・セットが最も適したコンテキストとともに表示されるので、ビジネスに関する洞察やビジネス・ユーザー向けのアクションを強化することができます。こうした統合により、データ・パイプライン全体でコラボレーションが促進されます。

拡張性、検出の容易さ、統合、データ品質、およびデータ・ガバナンスは個別の問題ではなく、相互に密接に関連しています。これらの問題を解決するには、情報管理に対する包括的なアプローチが必要です。

このカタログはチーフ・データ・オフィサー (CDO) オフィスのデータ・スチュワードも支援します。例えば、タグを付けてデータ・セットを分類し、そのデータ系列や使用状況を自動的に追跡したり、組み込み型のビジネス・グロッサリーを活用してデータのビジネス用語を標準化したりします。これにより、データ・スチュワードは各データ・セットの内容、機密情報や PII の保存場所、アクセスを許可するユーザーなどをより簡単に把握できるようになります。

組織内外の複数のデータ・ソースに対応する単一カタログ

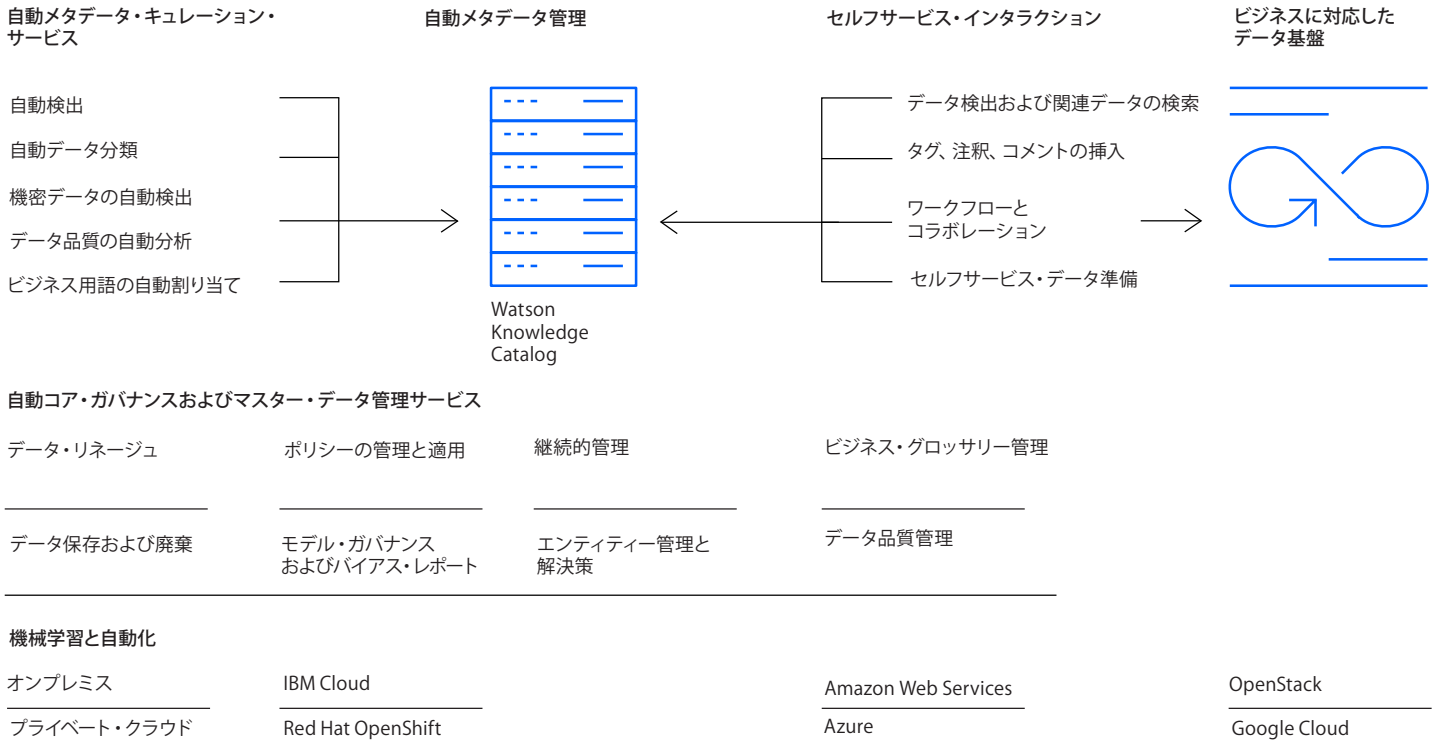


図 3. IBM Watson Knowledge Catalog のインテリジェントなメタデータ・インデックスとデータ (構造化および非構造化) は、元のシステムに常駐できますが、ユーザーは素早くデータを検出して高度なアナリティクスに利用することができます。

メタデータを主要優先事項とする IBM Watson Knowledge Catalog は信頼できる唯一の情報源であり、企業がアクセスできるすべてのデータ・セットへの単一のアクセス・ポイントです。

組み込み型のインテリジェント・データ検出

検出をさらに容易にするために、カタログではデータ・セットや分析資産にタグとコメントを挿入できます。これによってメタデータの補強や関連データの補足ができるため、同僚が必要なデータを見つけやすくなります。このソリューションには、ML を使用して各データ・セットの内容を自動的に分類するデータ検出アルゴリズムも組み込まれています。名前、アドレス、郵便番号、社会保障番号など共通のフィールド・タイプを指定することで、作成者がデータに手動で注釈を付ける必要性が軽減されます。このソリューションでは、自動化機能と ML を導入してデータ・キュレーションとメタデータ管理を自動化します。組み込み型のデータ品質機能を使用すると、詳細なデータ・プロファイル、データの品質管理、検証ルールを実行できます。

データ運用の自動化によって、データ品質およびガバナンスを維持しつつ、データ・パイプラインがキュレートされます。また、データレイクには高い品質とガバナンスを備えたデータが継続的に蓄積されます。

同様に、自社の資産のインテリジェント・メタデータ・モデルを追加することで、EU 一般データ保護規則 (GDPR; General Data Protection Regulation) やカリフォルニア州消費者プライバシー法 (CCPA; California Consumer Privacy Act) などを独自の方法で自動的に適用できます。

IBM Cloud Pak for Data で動作する IBM Watson Knowledge Catalog は、基本的には全てのデータ利用者に対し、信頼できる高品質かつビジネスに対応したデータを提供します。

このソリューションの全てのコンポーネントはマイクロサービスとして開発されており、拡張性、エラー管理、セキュリティー、ロギングなどの非機能要件に対し、統一された設計方針と共通のアプローチが採用されています。

IBM Watson Knowledge Catalog では ML を搭載したエンタープライズ・ガバナンス・プラットフォームを利用できるため、規模にかかわらず AI に対応できます。

IBM Watson Knowledge Catalog は、段階的かつユーザーが自力で行うアプローチのようにエラーとパフォーマンス・ボトルネックで混乱することはありません。IBM Watson Knowledge Catalog は ML を搭載したエンタープライズ・ガバナンス・プラットフォームを利用できるため、規模にかかわらず AI に対応できます。

IBM Watson Knowledge Catalog は以下の 3 種類の形式でご利用いただけます。

- IBM Cloud™ の Software as a Service (SaaS) として
- IBM Cloud Pak for Data に組み込み
- IBM Watson Studio に統合

IBM Watson Knowledge Catalog などのソリューションは、データレイクの取り組みで本来期待されていた価値を引き出すことができます。インテリジェントなカタログ化およびガバナンス機能を備えた Watson Knowledge Catalog を使用すると、信頼性が高くガバナンスが有効で、AI に対応したデータレイクを構築することができます。

AI 用にガバナンスを備えたデータレイクを構築する 4 つの利点

1. 高い品質とガバナンスによるデータの信頼性および信用性の確保

- データ品質機能を使用すると、データの品質を改善し、高品質なデータをデータレイクで使用できるようになります。
- ガバナンス・ポリシーは自動的に設定、適用されます。したがって、データを検出すると、その使用可否や使用方法もわかります。
- ユーザーが評価、コメント、その他の情報を追加するときにそのデータをキュレートできます。これにより、他のユーザーはそのデータ・セットが自分にとって有効かどうかを判断できます。

2. データ利用者への支援

- 業務部門 (LOB; Line-Of-Business) は積極的にデータを共有します。これは、ガバナンスが正しく機能し、データが誤使用から保護されていることを確信しているからです。
- 動的なデータ・ポリシーや制約を使用することでコラボレーションを強化し、信頼できる企業資産へとデータを変換させることができます。
- 他のユーザーが価値を引き出せるようにユーザーが関連タグやメタデータを追加することで、時間の経過と共にデータの検出や再使用がしやすくなります。
- データの保存場所を問わず、組織が所有するすべてのデータ・セットに単一のインターフェースでアクセスできます。

3. 時間の節約

- 自動データ検出によって、新規データ・セットのメタデータを追加するための時間と作業が低減されます。
- 自動データ・キュレーションおよびメタデータ管理によって、メタデータの検出や用語の割り当てにかかる時間、さらにはビジネス・グロッサリーの作成時間も短縮されます。
- シンプルで直観的なセルフサービス・データ準備ツールを使用すると、データ利用者はデータ準備の時間を短縮し、洞察の検出により多くの時間を割くことができます。
- データ・サイエンティストやビジネス・アナリストは、より短い期間でより良い分析を行えるようになります。
- AI を搭載したインテリジェントな検索機能を使用すると、必要なデータを数秒で検出できます。他の部門からデータが提供されるのを数週間も待つ必要はありません。

4. 増大するデータとコストの管理

- 価値の低いデータ・セットをデータレイクに追加する費用をなくすことで、ストレージ・コストを最適化できます。
- 組織が利用登録している外部データ・セットはすべて表示可能です。これにより、必要以上の登録料金を支払うリスクが減少します。
- データに対するユーザーの需要に基づき、データレイクに追加する新規データ・ソースに優先順位を付けることができます。これにより、最も価値の高いデータ・ソースを最初に取り込むことができます。

データの価値を引き出す

あなたの職場が CDO オフィスまたは IT 部門であっても、職務が LOB のデータ・サイエンティストまたはアナリストであっても、あなたと同僚は同じ目標を共有しています。確実に期待に応えるデータレイクを構築できれば、あなたは自分の業務をより簡素化し、かつ生産性を高めることができます。ただしそれだけではありません。現時点では対抗できる組織がほとんどない競争上の優位を企業が獲得するうえで、あなたは重要な役割を果たすこともできるのです。

競合他社がいまだにデータスワンプで四苦八苦している間に、データレイクのデータを正しく整理することができれば、他社にとっては夢でしかない可能性を切り開くことができます。真の先行者利益は、これまで利用されていなかったデータの価値を最初に引き出した者に与えられます。

まとめ

すべてのデータの保存場所、使用者、アナリティクス業務における
データの価値を理解してください。

DataOps イニシアチブにはデータ・カタログが不可欠です。
なぜならデータ・カタログは、データ・ガバナンス、データ品質、
積極的なポリシー管理を統合し、オープンなメタデータ管理を
自動化できるからです。

インテリジェントなカタログ化およびガバナンス機能を備えた
IBM Watson Knowledge Catalog を使用すると、信頼性が高く
ガバナンスが有効で、AI に対応したデータレイクを構築する
ことができます。このカタログはデータ統合、データ品質、
およびガバナンスをデータレイク環境に組み込んで、ビジネスに
対応した DataOps 用のデータを配信することができる、唯一の
信頼できる情報源です。

詳細情報

詳しくは、次の Web サイトをご覧ください。

ibm.com/cloud/watson-knowledge-catalog

日本アイ・ビー・エム株式会社
〒103-8510
東京都中央区日本橋箱崎町19-21

IBM のホーム・ページ:
ibm.com

1.Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders–
Gartner, Sept 2019

2. The Forrester Wave: Machine Learning Data Catalogs, Q2 2018

ASW12449-JPJA-03