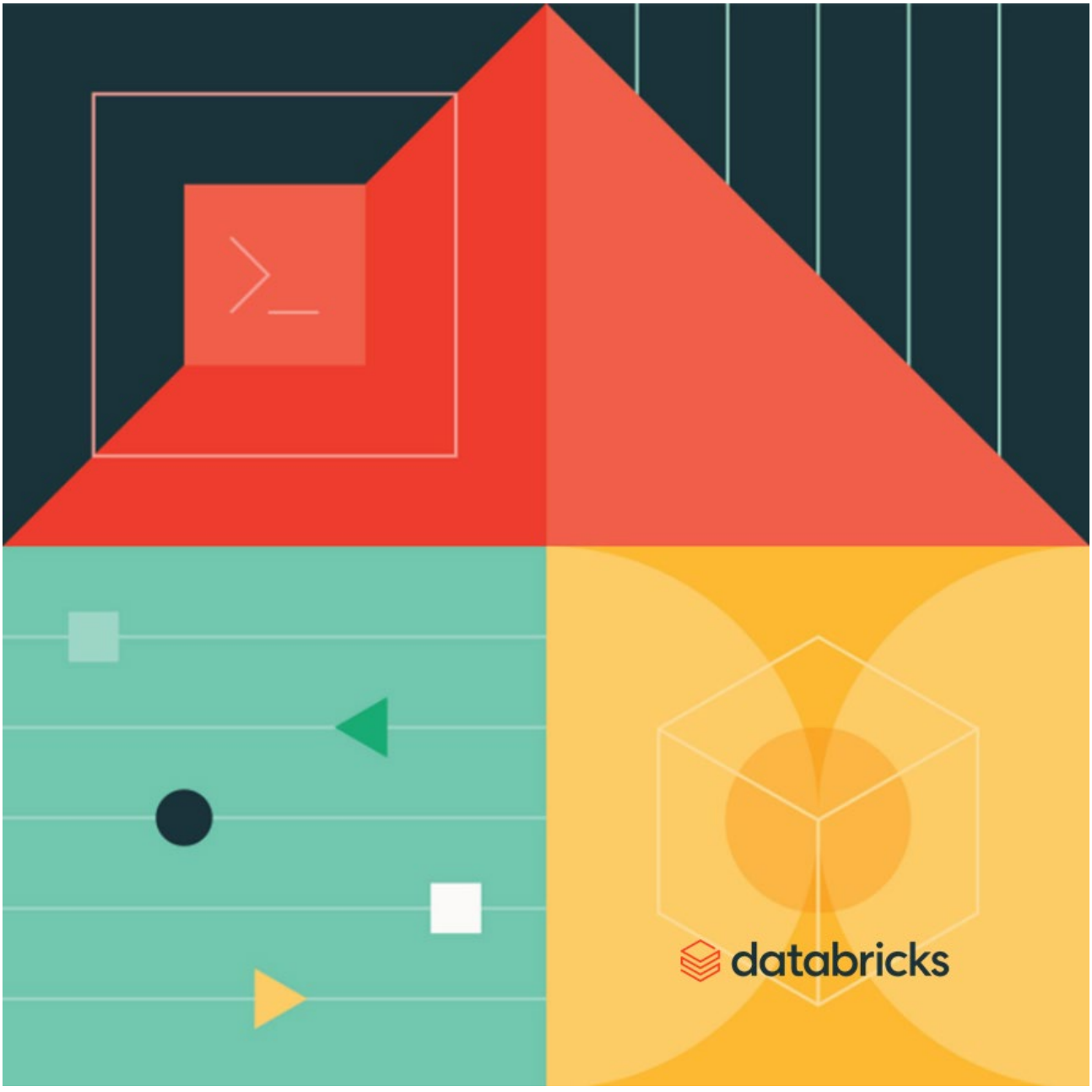
データレイクハウスの構築

英語版原題：Building the Data Lakehouse

ビル・インモン（Bill Inmon）― データウェアハウスの父

メアリー・レビンズ（Mary Levins）、ランジート・スリバスタバ（Ranjeet Srivastava）



本書について

原版の出版者

Technics Publications
2 Lindsley Road, Basking Ridge, NJ 07920 USA
ホームページ：https://www.TechnicsPub.com

著作権および書籍管理番号

初版：2021 年

ISBN、印刷版：　9781634629669
ISBN、Kindle 版：　9781634629676
ISBN、ePub 版：　9781634629683
ISBN、PDF 版：　9781634629690
米国議会図書館管理番号（LCCN）：　2021945389

執筆者謝辞

本書の執筆にあたり、Databricks 社の Bharath Gowda 氏より多大なご支援をいただきました。また、同社の Sean Owen、Jason Pohl 両氏のご協力にも感謝します。

日本語版について

本資料は、出版者である Technics Publications の許諾を得て、Databricks が翻訳したものです。なお、本資料には、第 1 章から第 3 章までの内容が含まれています。本資料についてのお問い合わせは、データブリックス・ジャパンまでお寄せください。

databricks

データブリックス・ジャパン株式会社
お問い合わせ：03- 6821-1670（代表）
https://databricks.com/jp/company/contact

目次

はじめに

かつてアプリケーションはシンプルであった。しかし現在では、あらゆる種類のデータ、テクノロジー、ハードウェア、その他のガジェットが存在する。データは無数の場所から多様な形式で生成され、その量は膨大である。

組織が分析目的で使用するデータは 3 種類存在する。第一のデータは、従来の構造化データである。主にトランザクションにより生成され、最も長く存在しているデータである。第二のデータは、電子メール、コールセンターの会話、契約書、医療記録などから生成されるテキストデータである。かつてテキストは、保存のみが可能で、コンピュータが分析できない「ブラックボックス」であった。

しかし今では、テキスト ETL（抽出・変換・ロード）技術によって、テキストの標準的な分析が可能となった。第三のデータは、アナログ／IoT データである。ドローン、電子眼カメラ、温度計、腕時計など、あらゆるマシンからこの種類のデータが生成される。アナログ／IoT データは、構造化データやテキストデータとは異なり、多様性・不規則性が高い。さらに、膨大な量のデータが自動生成されている。アナログ／IoT データは、データサイエンティストの領域である。

以前は、これらのデータは全て「データレイク 」と呼ばれる場所に取り込まれていた。しかし組織は、データをデータレイクに取り込むだけでは無意味であることに気づく。データを有用で分析可能なものにするには、1) データの関連付け、2) データを分析インフラに綿密に配置してエンドユーザーがデータを利用できるようにすること、この 2 つの条件を満たす必要があった。

この 2 つの条件を満たさない限り、データレイクはデータスワンプとなり、しだいに無用のものと化す。

分析の要件を満たさないデータレイクは、時間とコストの無駄となる。

そこで誕生したのがデータレイクハウスである。データレイクハウスは、データレイクに不足していた要素を補い、有用性と生産性をもたらした。言い換えれば、データレイクハウスを利用しないデータレイクの構築は、単に高コストで無用である。データレイクは、時間の経過とともに、高額な負債へと変化する。

分析と機械学習に必要な第一の要素は、分析インフラである。分析インフラには、成熟度の異なるものが混在する。例えば、データレイクハウスの分析インフラには、次のような要素がある。

- メタデータ
- データリネージ
- 容積測定
- 作成履歴
- トランスフォーメーションの記述

分析や機械学習に必要なデータレイクハウスの第二の要素は、共通コネクタの認識と活用である。共通コネクタは、あらゆる種類のデータの組み合わせや比較を可能にする。共通コネクタがなければ、データレイクハウスに格納されたさまざまな種類のデータの関連付けは極めて困難、もしくは不可能である。しかし、共通コネクタを活用することで、あらゆる種類のデータの関連付けができるようになる。

データレイクハウスにより、他の方法では困難、もしくは不可能なレベルの分析や機械学習が実現する。しかし、他のアーキテクチャ構造と同様に、データレイクハウスを有効活用するには、アーキテクチャに対する理解と詳細な計画が必要となる。

第 1 章：データレイクハウスへの進化

進化には長い時間が必要である。その進行は極めてゆっくりであり、各ステップを日々観察することはできない。絵の具が乾いていく様子を見守るようなものである。しかし、1960 年代から始まったコンピュータ技術の進化は、極めて高速に進んでいる。

テクノロジーの進化

その昔、コンピュータの世界はシンプルであった。データを取り込み、処理し、出力する。最初の記録媒体は紙テープである。紙テープへの記録は自動化されていたが、記録できる量はごくわずかで、フォーマットは固定されていた。次に登場したのがパンチカードである。パンチカードの問題点は、やはりフォーマットが固定されていることであった。膨大な枚数のパンチカードは、多くの紙を消費し、カードの束を崩してしまうと、順番を元に戻すための煩雑な作業が必要になることも問題であった。

その後、磁気テープによる近代的なデータ処理が始まった。これにより、膨大な量のデータを、固定されていないフォーマットに保存して利用できるようになった。しかし、磁気テープには、特定のレコードを見つけるのにファイル全体を検索しなければならないという問題があった。すなわち、磁気テープのファイルでは、データを順番に検索しなければならなかったのである。

また、磁気テープが壊れやすいことはよく知られており、データの長期保存には不向きであった。

そして、ディスクストレージの登場である。ディスクストレージは、データへの直接アクセスを可能にし、現代の IT 処理の可能性をさらに拡げた。ディスクストレージでは、順番に検索するのではなく、レコードに直接アクセスすることが可能である。初期には高いコストや入手困難などの課題があったが、時間の経過とともに、ディスクストレージは大幅に安価、大容量になり、利用が拡大された。

オンライントランザクション処理（OLTP）

データの直接アクセスが可能になったことが、高性能なアプリケーションの実現につながった。

高性能なストレージでデータへの直接アクセスができるようになったことで、オンライントランザクション処理（OLTP）システムが可能になった。OLTP システムが利用可能になると、企業は、コンピュータがビジネスの中核になったことに気づく。現在 OLTP は、オンライン予約システムや、銀行の窓口システム、ATM システムなどに利用されている。今やコンピュータは、顧客と直接対話できるようになっている。

初期のコンピュータは、反復的な処理に有用であった。しかし、オンライントランザクション処理システムでは、コンピュータは顧客との直接的な対話に活用できる。これにより、コンピュータのビジネス価値は飛躍的に高まることとなる。

コンピュータアプリケーション

アプリケーションの数は急増し、あらゆる目的のアプリケーションがいたるところに出現するようになった。

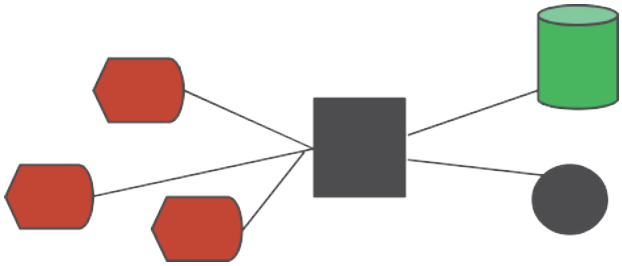


図 1-1：さまざまな目的の多様なアプリケーション

データ整合性の問題

アプリケーションの増加に伴い、予期せぬ問題が発生した。コンピュータ初期のエンドユーザーの不満は、自分のデータが見つけられないことであったが、アプリケーションの増加につれて、適切なデータを見つけられないことに変化した。

エンドユーザーの不満は、「データが見つからないこと」から「適切なデータが見つからないこと」に変化した。この変化は些細なことのように聞こえるが、実際にはそうではなかった。

アプリケーションの普及に伴い、データの整合性という問題が発生する。さまざまな場所に同一のデータが存在し、しばしば異なる値で現れるのである。エンドユーザーは、適切なデータか否かを判断するために、複数のアプリケーションの中からどのバージョンのデータを使うかを見極める必要があった。適切なバージョンのデータを使用できなければ、ビジネス上の判断を誤ってしまうおそれがある。

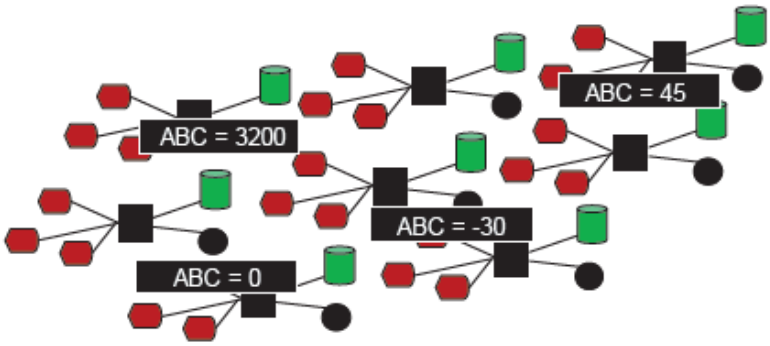


図 1-2：意思決定の基礎となる適切なデータを見つけるのは困難であった

適切なデータを見つけるという課題の重要さは、当初はあまり認識されていなかった。しかし、時間の経過とともに、意思決定に適切なデータを見つけることの複雑さが認識され、アプリケーションの構築だけではなく、アーキテクチャへの新たなアプローチが必要とされた。マシンやテクノロジー、コンサルタントが増えることにより、データの整合性の問題は、改善するどころか悪化する結果となっていたのである。

テクノロジーが増えることで、データの整合性が取りにくくなるという問題が拡大した。

データウェアハウス

そこで誕生したのがデータウェアハウスである。データウェアハウスでは、分散したアプリケーションデータが、別の物理的な 1 つの場所に集約してコピーされる。このようにしてデータウェアハウスは、アーキテクチャ上の問題に対応するアーキテクチャソリューションとなった。

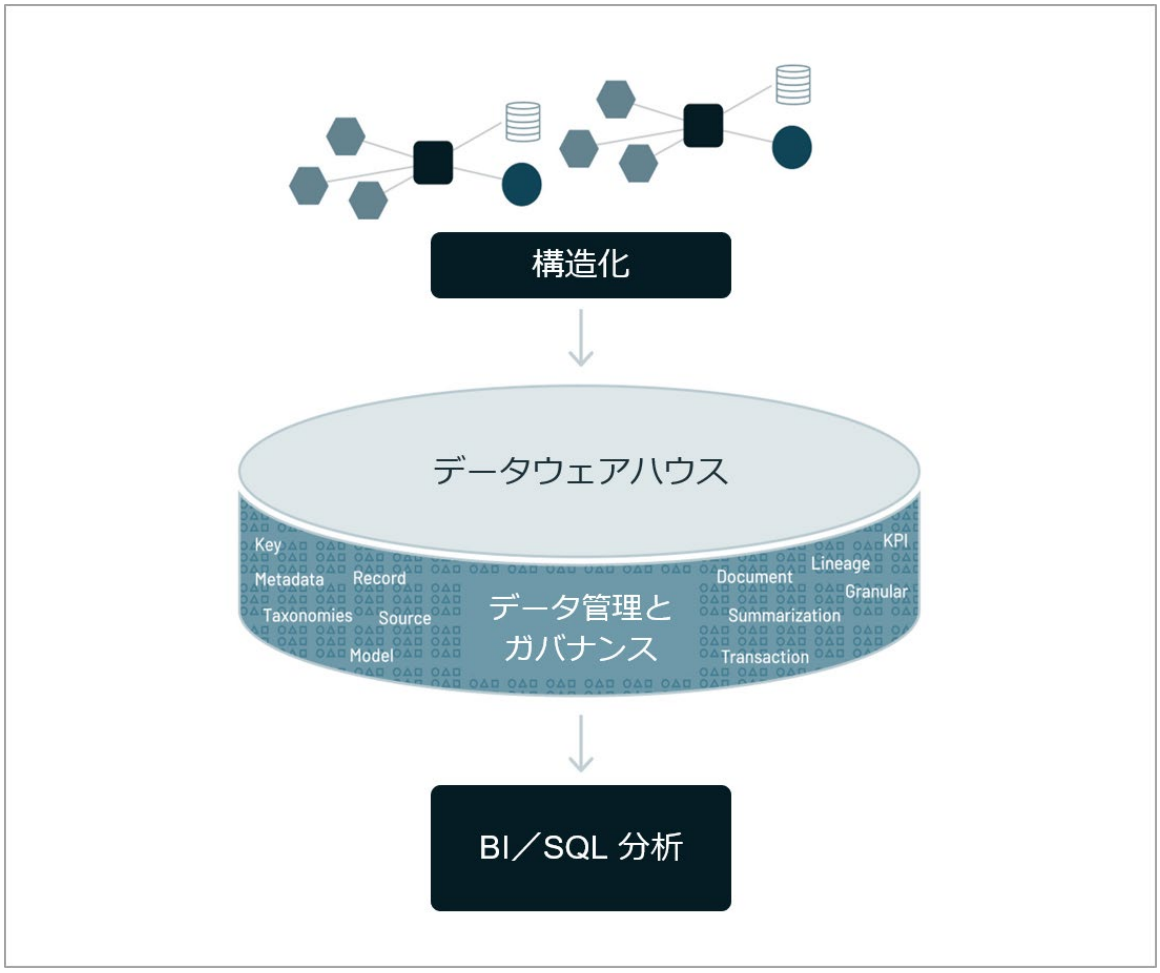


図 1-3：データウェアハウスを中心とする新たなインフラが必要となった

しかし、データを集約して物理的に別の場所に配置するだけのアーキテクチャでは不十分である。そこで、データウェアハウスを中心とした新たなインフラが設計された。データウェアハウスを中心とした新たなインフラにより、データウェアハウスにあるデータの利用が可能になり、分析が容易になった。データウェアハウスは重要だが、データウェアハウスの周辺の分析インフラなしでは、エンドユーザーにとってデータウェアハウスは価値を持たない。分析インフラには次のようなものがある。

- **メタデータ：** どのデータがどこに配置されたかを示すガイド
- **データモデル：** データウェアハウス内で見つかったデータの抽象化
- **データリネージ：** データウェアハウス内のデータの起源と変換の履歴
- **サマライズ：** データウェアハウス内のデータを生成するアルゴリズムに関する説明
- **KPI：** 主要な性能指標
- **ETL：** アプリケーションからのデータを自動的に企業データに変換するテクノロジー

履歴データの問題

データウェアハウスによって、分析処理に新たな扉が開かれた。これまでは、古いデータやアーカイブデータを容易かつ効率的に格納する便利な場所がなく、組織におけるシステムへのデータ保存期間は、1 週間、1 か月、あるいは四半期分が一般的であった。1 年分あるいは 5 年分のデータを保管する組織はほとんど存在しなかった。しかし、データウェアハウスにより、10 年またはそれ以上の期間のデータが格納できるようになった。

また、時系列データの長期的な記録が可能になったことが大きな意味をもたらした。例えば、顧客の購買傾向を知りたい組織にとって、顧客の過去の購買パターンを理解することは、現在および将来の購買パターンの予測につながる。

過去は、未来を予測するための重要な判断材料となった。

データウェアハウスによって、分析の世界に、データ保存期間の長さという新たな要素が加わることとなる。これで履歴データが負担になることはなくなった。

データウェアハウスは、重要な役割を果たし便利なものではあるが、ほとんどの場合において、処理できるのは構造化データおよびトランザクションベースのデータのみである。構造化環境やデータウェアハウスでは、利用できるデータの種類が限定されることに留意すべきである。

テクノロジーの進化により、構造化データだけではなく、多様なソースからさまざまな種類のデータが出現するようになった。コールセンターやインターネット、データを生成するマシンなど、データはあらゆる場所のあらゆるデバイスから生成される。テクノロジーの進化が生み出したものは、構造化、トランザクションベースデータにとどまらなかった。

しかし、テキスト、IoT、画像、音声、動画、ドローンなど、エンタープライズが扱うデータの多様化に対応するには、データウェアハウスでは限界があった。さらに、機械学習（ML）や人工知能（AI）の台頭により、SQL を使用しないデータへの直接アクセスが必要な反復的アルゴリズムが導入され、データウェアハウスでは扱えないデータが増大した。

あらゆる種類のデータへの対応

データウェアハウスは重要かつ有用である反面、ほとんどのケースで、処理できるのは構造化データが中心である。しかし、組織にはさまざまな種類のデータが存在する。下の図は、組織が扱うデータを大まかに分類したものである。

図 1-4：データの種類

構造化データとは、一般的には、組織の日常業務の中で生成されるトランザクションベースのデータである。テキストデータとは、組織内での書簡、電子メール、会話により生成されるデータである。非構造化データには、IoT、画像、動画、他のアナログベースのデータなどが含まれる。

構造化データ

最初に出現したデータの種類は構造化データである。構造化データは、ほとんどケースでトランザクション処理によって生成される。トランザクションが実行されるとレコードが書き込まれる。トランザクションの種類には、販売、支払い、通話、銀行取引などがある。それぞれの新たなレコードの構造は、過去のものに類似する。

処理の類似性は、銀行での預金を例に挙げてみるとわかりやすい。ある顧客が銀行の窓口で預金し、次の顧客も窓口で預金する。口座番号や入金額は異なっていても、どちらのレコードも同じ構造を持つ。

同じ構造を持つデータが繰り返し書き込まれることから「構造化データ」と呼ばれる。

構造化データは、一般的に、トランザクションごとに生成されるレコード、すなわち膨大な数のレコードから成るものであり、ビジネスの根幹に関わるトランザクションから生成される構造化データのビジネス価値は高いものとなる。

テキストデータ

未加工のテキストがそれほど有用でない主な理由は、コンテキストの理解あってこそ意味を持つからである。単なる未加工のテキストの読み取りや分析だけでは不十分である。

テキストの分析には、テキストとそのコンテキストの両方を理解する必要がある。

テキストには、考慮すべき点が他にもある。テキストは、英語、スペイン語、ドイツ語など、言語の世界の中に存在することを考慮しなければならない。また、テキストには、予測可能なものと予測不可能なものがある。予測可能なテキストと、予測不可能なテキストでは、分析の方法が大きく異なる。同じワードが複数の意味を持つことも、的確な分析結果を得る障壁となる。例えば、「レコード」というワードは、曲を録音したレコード盤を意味することもあれば、レースの記録を意味することもある。また、別のものを意味することもある。未加工のテキストの読み取りや分析には、他にもさまざまな課題がある。

テキストデータ ETL

構造化されたフォーマットでテキストを作成することは現実的に可能である。これは「テキストデータ ETL」というテクノロジーとして知られている。未加工のテキストを読み取り、テキストとコンテキストの両方を識別しながら、標準的なデータベースのフォーマットに変換する。これにより、構造化データとテキストの融合が可能になる。あるいは、テキスト単体で独立した分析を行うこともできる。

アナログデータ／IoT データ

自動車や時計、製造機械などのマシンが動作すると、アナログデータが生成される。マシンが動作している間は、マシンから測定値が生成される。温度、化学成分、速度、時刻など、測定対象は多岐にわたり、アナログデータは、同時に測定・取得されるさまざまな不特定データで構成される。

電子眼カメラ、温度モニター、映像機器、テレメトリ、タイマーなど、アナログデータはさまざまなソースから生成される。

アナログデータは通常、頻繁に生成される。マシンの種類や処理内容によるが、1 秒、10 秒などの秒単位、あるいは分単位の周期で測定されることが多い。

ほとんどの測定値は正常範囲内であり、有用でないかもしれない。しかし、ときには極めて興味深い正常範囲外の測定値を取得することがある。

アナログや IoT のデータを取得し、利用・管理するうえでの課題は、次の点をどう決定するかにある。

- 取得、測定するデータの種類
- データ取得の頻度
- 正常範囲

他にも、収集したデータ量、必要な場合のデータ変換、異常値の発見と除去、アナログデータと他のデータとの関連付けなど、さまざまな課題がある。原則として、正常範囲内のデータはバルクストレージに、正常範囲外のデータは別のストアに格納する。

問題解決との関連性でデータを保存する方法もある。従来から、ある種のデータは他の種類のデータよりも問題解決への関連性が高いとされている。アナログデータの分析にあたっては、分析者は一般的には次のポイントに着目する。

- データの値
- 多数回にわたるデータの傾向
- 相関パターン

その他の非構造化データ

エンタープライズにおいて生成されているデータの大半は、画像や音声、動画などの非構造化データである。

非構造化データは、通常ではテーブル形式の構造を持たないため、一般的なデータベースのテーブルには格納できない。アナログデータや IoT データの量は膨大であり、これらのデータセットの格納や管理にはかなりのコストがかかる。

また、非構造化データを SQL のみのインタフェースで分析するのは容易なことではない。しかし、クラウドに安価な Blob ストレージが登場したことで、柔軟性のあるクラウドコンピューティングや機械学習アルゴリズムによる非構造化データへの直接アクセスが可能になった。エンタープライズは、これらのデータセットの可能性を理解し始めている。

非構造化データの新たなユースケースの例は次のとおりである。

画像データ

- X 線、CT、MRI スキャンの医療画像解析により、放射線科医を支援
- 画像解析により、ホテルやレストランの所有地や料理の写真を分類

- ビジュアルサーチによる商品発見により、e コマース企業の顧客エクスペリエンスを改善
- ソーシャルメディア上の画像からブランドを識別し、マーケティングキャンペーンのターゲットとなる購買層を特定

音声データ

- コールセンターの音声データの自動文字起こしにより、顧客サービスを改善
- 会話型 AI により、音声を認識し、人間の会話に近いコミュニケーションを実現
- 製造工場のさまざまな機械音を音声 AI で識別し、設備の不具合をプロアクティブに検知

動画データ

- 動画による店内行動解析で人数のカウント、キュー分析、ヒートマップなどの情報を提供し、顧客と商品の関わり方を把握
- 動画解析によるインベントリ自動追跡および、製造工程での製品の不具合の検知
- 政府機関・自治体において、動画解析によって公共インフラの状態や使用状況を把握し、メンテナンスの時期などに関わる意思決定を支援
- 顔認証によって認知症患者が施設を離れたことを検知し、医療従事者に対して対応を促すためのアラートを発行

ビジネス価値

ビジネス価値は、データの種類によって異なる。第一に、日常業務におけるビジネス価値。第二に、長期戦略的なビジネス価値。第三に、機械の管理や運用におけるビジネス価値である。

構造化データとビジネス価値には、当然強い関係性がある。組織における日々の業務では、トランザクションと構造化データが発生する。また、テキストデータとビジネス価値の間にも、強い関係性がある。テキストは、ビジネスには不可欠なものである。

しかし、現在のビジネスにおけるアナログ／IoT データとの関係性は異なる。大規模なクラウドコンピューティングリソースや機械学習のフレームワークの利用によってもたらされるアナログ／IoT データの可能性については、やっと最近理解され始めたところである。例えば、組織では、画像データの活用による製造段階での品質問題の検知、コールセンターの音声データによる顧客感情の分析、石油やガスのパイプラインなどにおける遠隔操作の動画データによる予知保全を実施している。

データレイク

データレイクは、組織におけるあらゆる種類のデータを集約したリポジトリである。

データレイクに格納されているデータは、構造化データ、テキストデータ、アナログ／IoT データの 3 種類である。データレイクに存在するデータには、さまざまな課題があるが、最大の課題は、アナログ／IoT データの形式や構造が、データウェアハウスにある従来の構造化データとは大きく異なることである。さらに問題を複雑にしているのは、データレイクにある多様な種類のデータのそれぞれの量の違いである。他の種類のデータと比較して、はるかに膨大な量のアナログ／IoT データがデータレイクに格納されている。

企業は、あらゆるデータをオフロードする場所としてデータレイクを利用している。データレイクが、Apache Parquet や ORC などの汎用的でオープンなファイル形式でデータを保持するファイル API を備えた、低コストのストレージシステムであることが主な理由である。また、オープンフォーマットにより、データレイクのデータは、機械学習システムをはじめとするさまざまな分析エンジンから直接アクセスできるようになっている。

当初は、データを抽出してデータレイクに置くだけでよいと考えられていた。データがデータレイクに格納すれば、エンドユーザーは、そこからデータを検索し、分析を行えると考えたのである。しかし、組織は、データをデータレイクに格納することと、データレイク内のデータを利活用できることは全く別の話であることに気づく。言い換えると、エンドユーザーのニーズとデータサイエンティストのニーズは、全く異なるものであった。

図 1-5：オープンフォーマットを基盤とするデータレイク

エンドユーザーは、次のような障壁に直面した。

- 必要なデータはどこにあるのか？
- データユニットの相互関係は？
- データは最新か？
- データの正確度は？

データレイクに期待されていた事柄の多くは、トランザクションのサポート、データ品質やガバナンスの確保、性能の最適化といった重要なインフラ機能の欠如により、実現されていない。その結果、エンタープライズの持つデータレイクのほとんどが、データスワンプになっているのが実情である。

データスワンプのデータは放置され、時間の経過とともに陳腐化する。

現在のデータアーキテクチャにおける課題

各システムの限界を補うためによく使用されるのが、複数のシステム（データレイク、複数のデータウェアハウス、その他の特殊目的のシステム）を同時に使用するという方法であるが、次のような問題が発生する。

1. **デュアルアーキテクチャに起因する高コストなデータ移動：**アナログ／IoT データの 90%以上は、ファイルへのオープンな直接アクセスによる柔軟性と、安価なストレージの利用による低コストであることを理由に、データレイクに格納されている。データレイクの性能不足や品質問題を解決するために、エンタープライズでは、ETL（抽出・変換・ロード）を実行し、データレイクのデータの小規模なサブセットをダウンストリームのデータウェアハウスにコピーし、重要な意思決定の支援や BI アプリケーションに利用している。このデュアルシステムアーキテクチャでは、データレイクとデータウェアハウスの間のデータ ETL に、継続的なエンジニアリングを必要とする。ETL の各ステップには、処理の失敗や、データ品質の低下を誘発するバグの発生などのリスクがあり、データレイクとデータウェアハウスの一貫性を維持することは困難でコストもかかる。その間も ETL はデータを取り込み続ける。
2. **限定的な機械学習のサポート：**機械学習とデータ管理の融合に関して多くの研究がなされている一方で、TensorFlow、PyTorch、XGBoost などの主要な機械学習システムはいずれも、データウェアハウス上でうまく機能しないのが実情である。比較的小規模なデータを扱うビジネスインテリジェンス（BI）とは異なり、機械学習システムは、複雑な非 SQL コードを使用して大規模なデータセットを処理する。
3. **オープン性の欠如：**データウェアハウスは、データを独自のフォーマットに固定する。そのため、データやワークロードを他のシステムに移行するにはコストがかかる。また、アクセスは主に SQL を介してのみ行われるため、機械学習システムをはじめとする分析エンジンをデータウェアハウスで動作させることは困難である。

データレイクハウスの登場

データスワンプの問題を解決すべく、「データレイクハウス」と呼ばれる新たなクラスのデータアーキテクチャが出現した。データレイクハウスは、次のような要素を持つ。

- 構造化環境からのデータ
- テキスト環境からのデータ
- アナログ／IoT 環境からのデータ
- レイクハウスのデータを読み解く分析インフラ

データレイクハウスでは、データウェアハウスと類似のデータ構造とデータ管理機能を、データレイクに使用される低コストのストレージに直接実装する。このオープンで標準化された新たなシステムデザインにより、アナログ／IoT データ分析が可能になる。

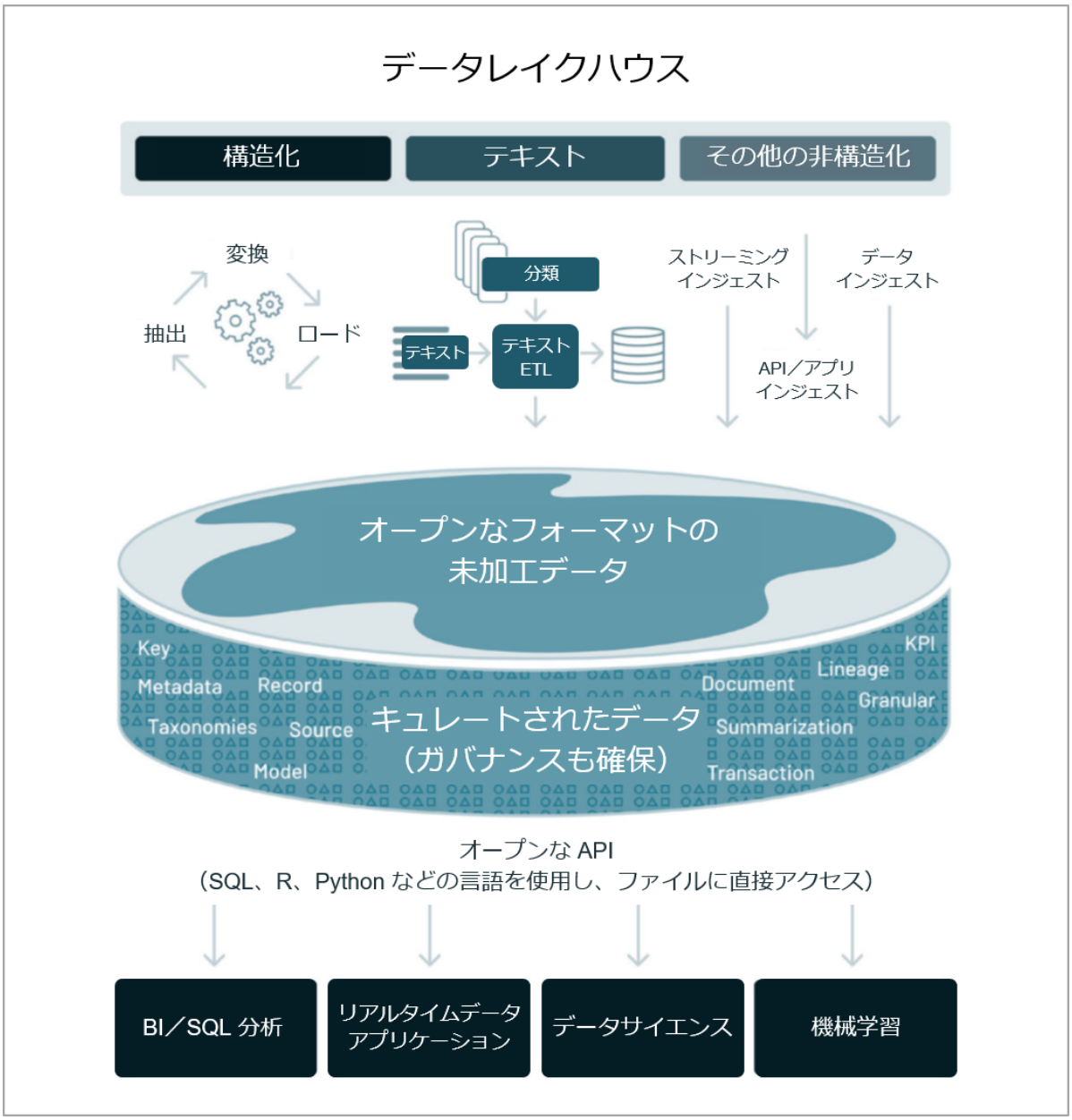


図 1-6：データレイクハウスのアーキテクチャ

データレイクハウスのアーキテクチャは、既存のデータレイク上に構築することで、前述の現在のデータアーキテクチャの主要な課題を解決する。

データレイクハウスアーキテクチャによるアナログ／IoT コンポーネントの構築では、主に以下の機能が中核となっている。

レイクファースト

Amazon S3、Azure Blob Storage、Google Cloud などの低コストのストレージ上のデータレイクに既に格納されている、多くの構造化データおよび、テキストデータその他の非構造化データを利用できる。

データレイクの信頼性と品質の確保

- **トランザクションのサポート：**SQL などのツールを利用して、複数のユーザーが同時にデータの読み取り／書き込みを行う場合でも、ACID トランザクションによって一貫性が保たれる。
- **スキーマのサポート：**スター／スノーフレークスキーマなどの DW スキーマアーキテクチャをサポートし、堅牢なガバナンスと監査メカニズムを提供。
- **スキーマの適用：**任意のスキーマを指定して適用し、不良データによるデータ破損を防止。
- **スキーマの進化：**常に変化するデータに対応し、エンドユーザーは、DDL の煩雑な作業を必要とせずにテーブルのスキーマを変更して自動適用できる。

ガバナンスとセキュリティ制御

- **Scala、Java、Python、SQL API による DML のサポート：**データセットのマージ、更新、削除を可能にし、GDPR／CCPA コンプライアンスを確保し、さらに、変更データの取得などのユースケースを簡素化する。
- **履歴機能：**データに対する変更の詳細を全て記録し、変更に関する完全な監査証跡を提供。
- **データのスナップショット：**監査、ロールバック、実験の再現などを目的とした、以前のバージョンのデータへのアクセスおよび、元の状態への復元を可能にするスナップショットを提供。
- **ロールベースのアクセス制御：**テーブルの行／列レベルでの高粒度のセキュリティとガバナンスを提供。

性能の最適化

ファイル統計とデータ圧縮によってファイルサイズを適切に調整し、キャッシュ、多次元クラスタリング、Z オーダー、データスキップなどをはじめとする最適化手法の利用を可能にする。

機械学習のサポート

- **多様なデータの種類に対応：**画像、動画、音声、半構造化データ、テキストなど、多くの新しいアプリケーションのデータの格納、精製、分析、アクセスをサポート。
- **非 SQL による効率的な直接読み込み：**R や Python のライブラリを用いた機械学習実験における、大量のデータの直接読み込みをサポート。
- **データフレーム API のサポート：**機械学習ワークロードにおけるデータアクセスのクエリ最適化機能を備えた宣言型のデータフレーム（DataFrame）API をビルトインで提供（TensorFlow、PyTorch、XGBoost などのシステムは、データ操作の主要な抽象化手法としてデータフレームを採用している）。
- **機械学習実験におけるデータのバージョニングとスナップショット：**監査やロールバック、機械学習実験の再現が必要な場合に、以前のバージョンのデータの復元やアクセスを可能にする、データのバージョニングおよびスナップショットを提供。

オープン性の提供

- **オープンファイルフォーマット：**Apache Parquet や ORC などのオープンなファイルフォーマットをサポート。
- **オープン API：**カスタムエンジンやベンダーロックインを回避し、データへの直接アクセスが可能なオープン API を提供。
- **多様なツール、言語のサポート：**SQL によるアクセスだけでなく、他の機械学習、Python／R ライブラリなどのツールやエンジンに対応する多言語サポート。

データウェアハウス、データレイク、データレイクハウスの比較

	データウェアハウス	データレイク	データレイクハウス
データフォーマット	クローズド、独自仕様のフォーマット	オープンフォーマット	オープンフォーマット
データの種類	構造化データ、一部の半構造化データ	全ての種類：構造化、半構造化、テキスト、非構造化（未加工）データ	全ての種類：構造化、半構造化、テキスト、非構造化（未加工）データ
データアクセス	SQL のみ	SQL、R、Python その他の言語によるファイルへの直接アクセスを可能にするオープン API	SQL、R、Python その他の言語によるファイルへの直接アクセスを可能にするオープン API
信頼性	ACID トランザクションによる高品質、高信頼性データ	低品質、データスワンプ	ACID トランザクションによる高品質、高信頼性データ
ガバナンスとセキュリティ	テーブルの行／列レベルでの高粒度のセキュリティとガバナンス	ファイルレベルでの粗いセキュリティ（不十分なガバナンス）	テーブルの行／列レベルでの高粒度のセキュリティとガバナンス
性能	高	低	高
スケーラビリティ	スケーリングにより、コストが爆発的に増大	あらゆる種類／量のデータを低コストでスケーリング	あらゆる種類／量のデータを低コストでスケーリング
ユースケースのサポート	BI、SQL アプリケーション、意思決定支援に限定	機械学習に限定	単一のデータアーキテクチャで BI、SQL、機械学習に対応

データレイクハウスのアーキテクチャは、データウェアハウス（DWH）市場の黎明期に見られたものに匹敵するインパクトをもたらすものである。オープンな環境でのデータ管理、エンタープライズの各部門から得られる多様なデータの集約、データレイクのデータサイエンスとデータウェアハウスのエンドユーザー分析を組み合わせた機能が相乗効果を発揮し、データレイクハウスを利用する企業におけるデータ活用の可能性を引き出す。

第 2 章：データサイエンティストとエンドユーザー

最初にアプリケーション、次にデータウェアハウスが登場し、その後さまざまな種類のデータが出現した。データの量と多様性が増大し、あらゆる種類のデータがデータレイクに格納されるようになった。

データレイク

データレイクは当初、未加工データのリポジトリとして考えられていた。さまざまなソースから収集されたデータは、誰もがアクセスできるよう、そのままデータレイクに蓄積されていた。

図 2-1：データレイク

分析インフラ

時は流れ、データレイクには分析のためのインフラが必要であることがわかってきた。分析インフラは、データレイクの未加工データをもとにして構築され、次のような役割を持っていた。

- データの関連付け
- データの適時性の識別
- データ品質の検証
- データリネージの把握

これらの役割（コンポーネント）については、第 3 章で解説する。

図 2-2：多様なコンポーネントを有する分析インフラ

異なる対象ユーザー

分析インフラとデータレイクは、対象ユーザーが異なっていた。

データレイクの主なユーザーはデータサイエンティストであった。データサイエンティストは、組織にとって意味のある新たなデータの傾向を発見する。

図 2-3：新たなデータの傾向を発見するデータサイエンティスト

データレイクや分析インフラを利用する他のコミュニティは、エンドユーザーであった。エンドユーザーの役割は、ビジネスの継続的な生産性と収益性の向上を図ることである。

図 2-4：ビジネスの成長を図るエンドユーザー

分析ツール

エンドユーザーとデータサイエンティストの大きな違いは、データ分析に利用するツールが異なることである。データサイエンティストは、主に統計解析ツールを利用する。探索的なツールを使用するデータサイエンティストもいるが、多くは統計解析ツールを利用する。

一方、エンドユーザーは全く異なる方法でデータ分析を行う。エンドユーザーは、シンプルな計算や視覚化を目的としたツールを利用し、チャートやダイアグラムの作成、データの視覚化などを重視する。

データサイエンティストのツールは、大まかに集積されたデータを扱う。一方、エンドユーザーのツールは、一貫性があり明確に定義されたデータを扱う。

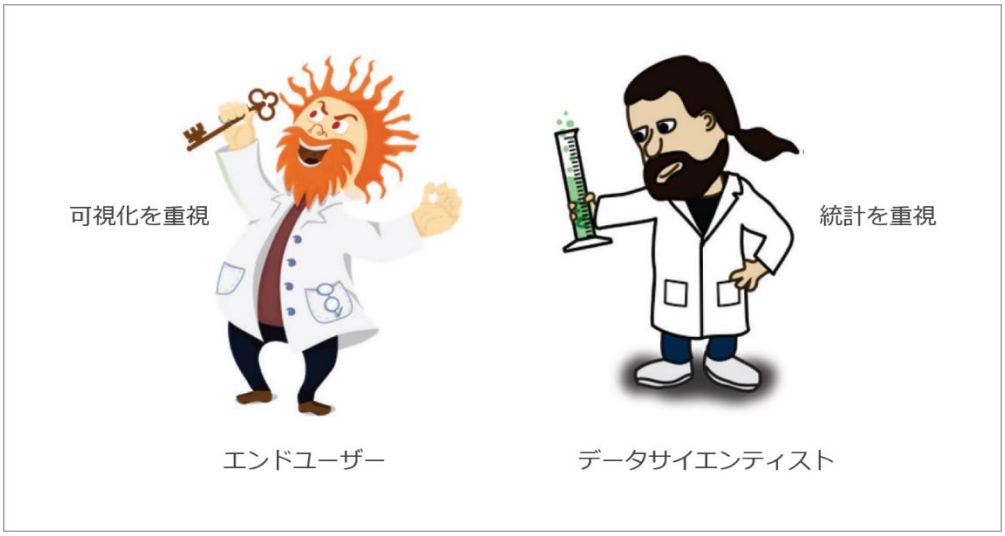


図 2-5：データサイエンティストとエンドユーザーは異なる視点を持つ

分析内容

分析ニーズにも違いがある。データサイエンティストは、データから、核心的なパターンや傾向を発見しようとする。新たなパターンや傾向の発見は、組織の存続性や収益性の改善につながる。

エンドユーザーの関心は、データの新たなパターンの発見ではなく、むしろ、既存のデータのパターンの再計算や再検討を行うことである。例えば、エンドユーザーは、収益性、新規顧客数、新規の販売形態など、月次・四半期ごとの KPI に関心がある。

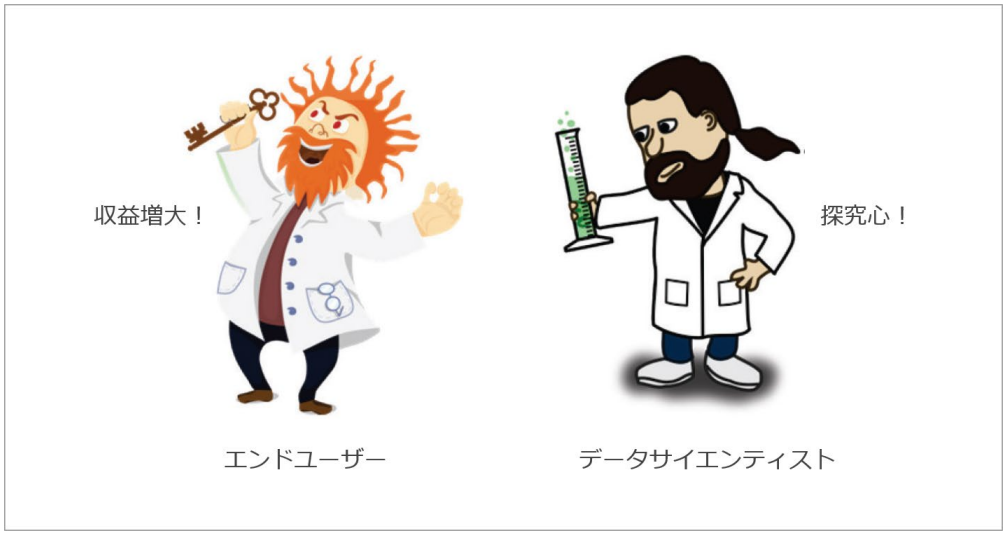


図 2-6：データサイエンティストの探求心が、差別化要素の創出力となる

分析へのアプローチも、データサイエンティストとエンドユーザーでは大きく異なる。エンドユーザーは、定期的に生成されるデータのパターンをもとに、シンプルな計算方法を用いて同じ分析を何度も繰り返す。データサイエンティストは、発見を目的とする分析を行う。

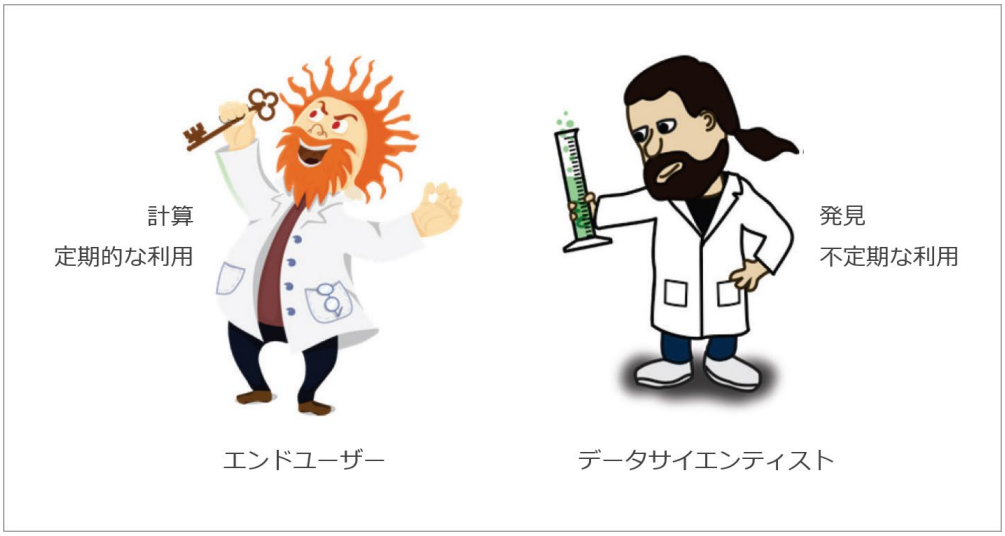


図 2-7：エンドユーザーとデータサイエンティストの分析へのアプローチの違い

データの種類

データサイエンティストは、粒度の低い、多くの場合はマシンによって生成される、多様なデータを扱う。データ探索では、さまざまな種類のデータを広範に検証することが必要となる。

一方、エンドユーザーは、高度に整理され、定期的に生成され、集約された（あるいは大まかに集約された）データを扱う。月次、週次、日次で同種類のデータを調査して再計算する。

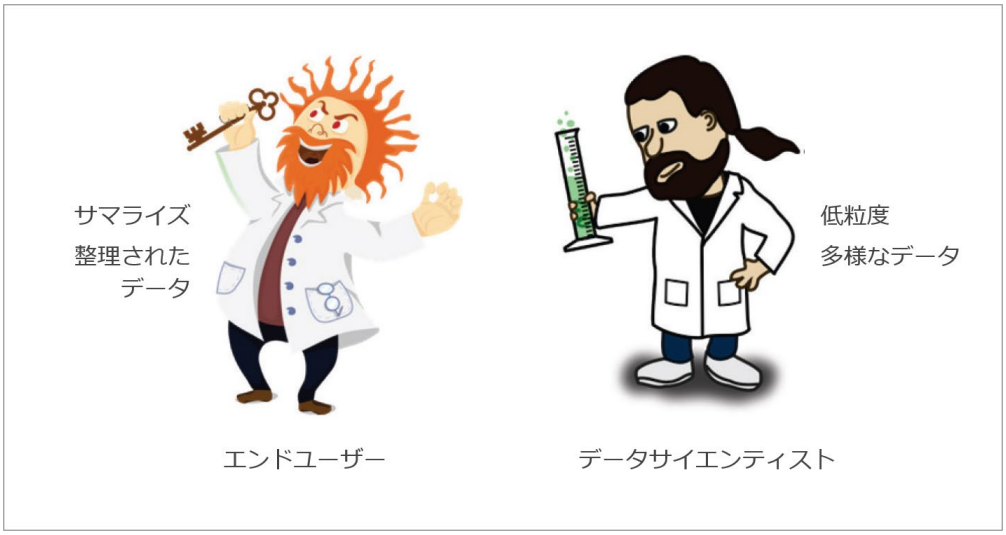


図 2-8：エンドユーザーとデータサイエンティストが扱うデータの種類の違い

ニーズが異なるため、それぞれがデータレイクの異なる部分に関心を持つ。データサイエンティストは未加工データ、エンドユーザーは分析インフラのデータに関心を持つ。

ニーズの違いが、他方のデータへの一切の関わりを遮断するか？答えは No である。エンドユーザーがデータレイクにある未加工データを探索・利用することも、データサイエンティストが分析インフラを利用することも、可能であるべきである。

図 2-9：エンドユーザーとデータサイエンティストが関心を持つデータの違い

データサイエンティストは分析インフラの有用性を認めるであろう。しかし、データサイエンティストがデータ分析の技術を習得したとしても、実環境では、データのゴミ処理にほとんどの時間を費やすことになる。データのクリーニングに時間の 95%を費やすことになり、分析に費やせる時間は 5%となる。

異なるタイプのユーザーが、それぞれの目的でデータレイクハウスを利用する。データレイクハウスの目的は、あらゆるユーザーにサービスを提供することである。

第 3 章：データレイクハウスにおけるデータの種類

データレイクハウスはさまざまな種類のデータの融合体であり、データの種類にはそれぞれ特徴がある。

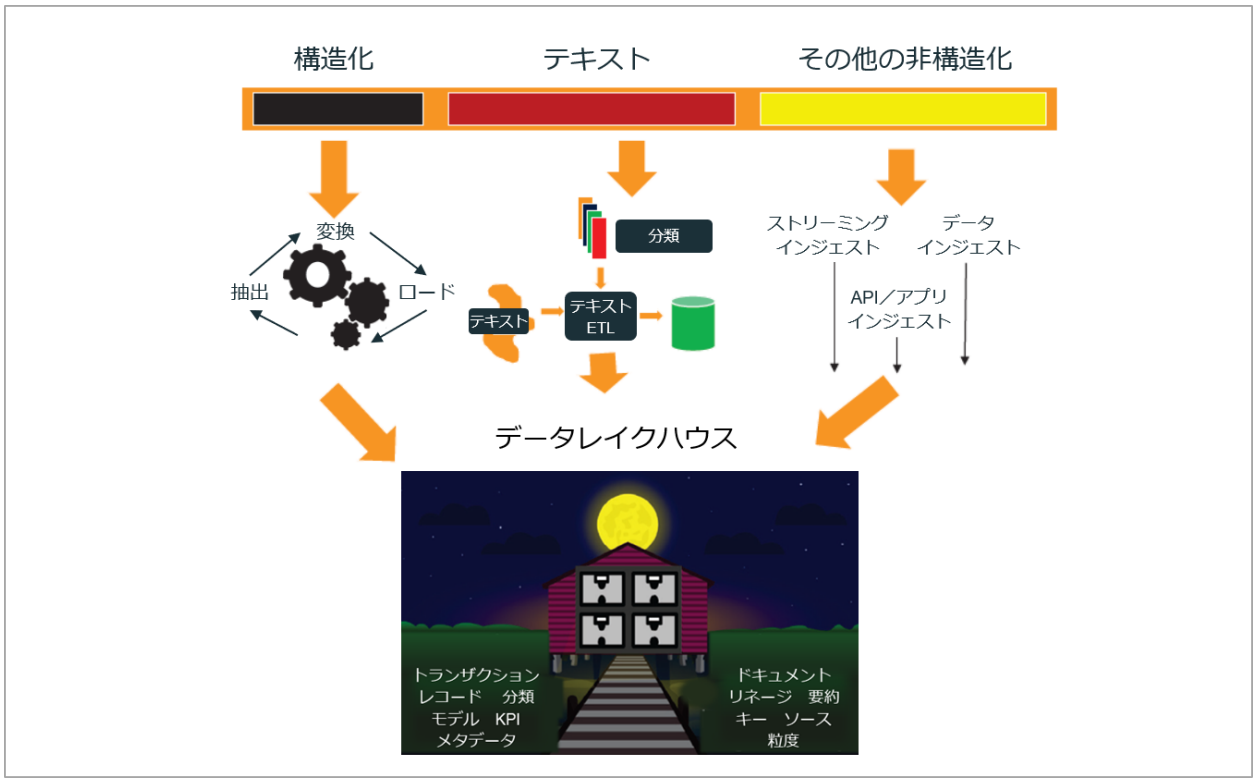


図 3-1：データレイクハウスとインフラ

データレイクハウスは次の要素で構成されている。

- データレイク：未加工のテキストデータを格納
- 分析インフラ：記述的情報をエンドユーザーに提供
- さまざまな種類のデータ：構造化データ、テキストデータ、その他の非構造化データ

データレイクハウスで探索できるデータはオープンである。

次に、各コンポーネントについて詳しく見ていく。

データの種類

データレイクハウスは、次の 3 種類のデータを扱う。

- 構造化データ：トランザクションベースのデータ
- テキストデータ：音声会話や文書からのデータ
- その他の非構造化データ：アナログ／IoT データ（一般的にはマシンから生成されるデータ）

図 3-2：データレイクハウスの 3 種類のデータ

構造化データ

コンピュータによって最初に処理されたデータは構造化データである。構造化データは主にトランザクションの実行から生成される。トランザクションから流れるデータから、1 つ以上の構造化されたレコードが書き込まれる。

構造化された環境にはレコードが含まれる。レコードの構造は統一されており、レコードには、キーや属性など、さまざまな種類の情報が含まれる。さらに、特定の構造化レコードの場所の検索に役立つインデックスも含まれる。

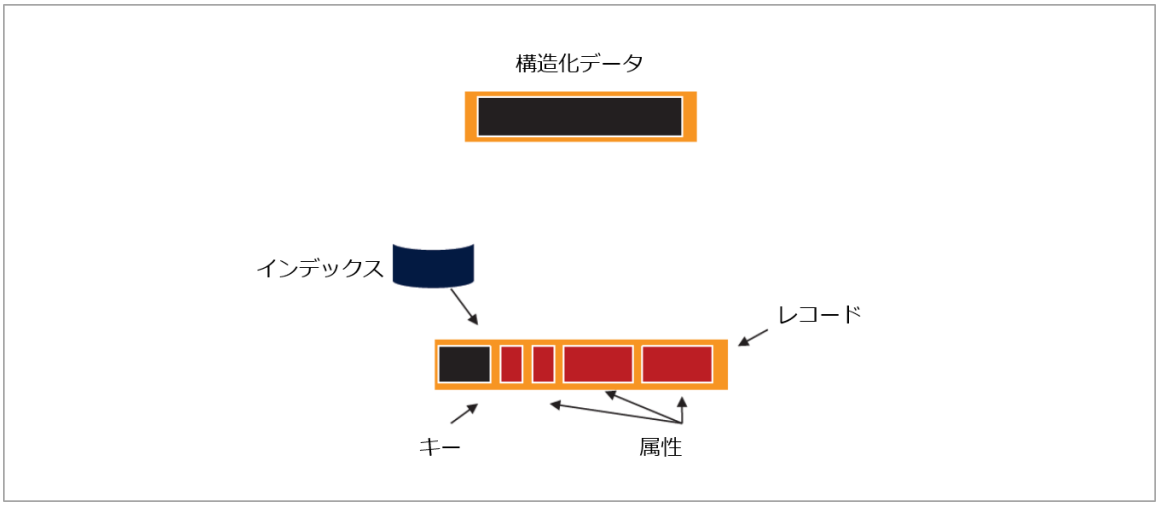


図 3-3：構造化された環境のコンポーネント

レコードは、データベースによって構造化された環境で作成され、個別にまたはまとめてアクセスできる。データベース内のレコードは、削除や修正が可能である。

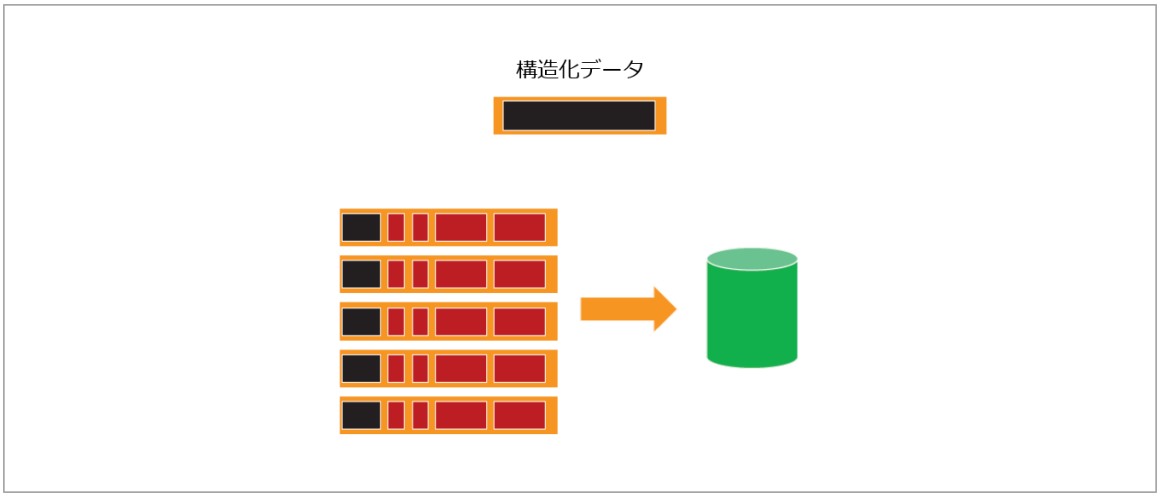


図 3-4：構造化されたレコードをデータベースに取り込む

テキストデータ

第二の種類はテキストデータである。通話、電子メール、インターネットなど、テキストデータはあらゆる場所に存在する。未加工のテキストデータは、データレイクハウスに格納してもあまり意味を持たないため、通常はデータベース形式で格納される。データベース形式にすることで、テキストに対して分析ツールを使用できるようになる。データレイクハウスに未加工データを格納することも可能である。

データレイクハウスに格納するテキストデータには次のものがある。

- データのソース
- 目的のワード
- ワードのコンテキスト
- ドキュメント内のワードのバイトアドレス

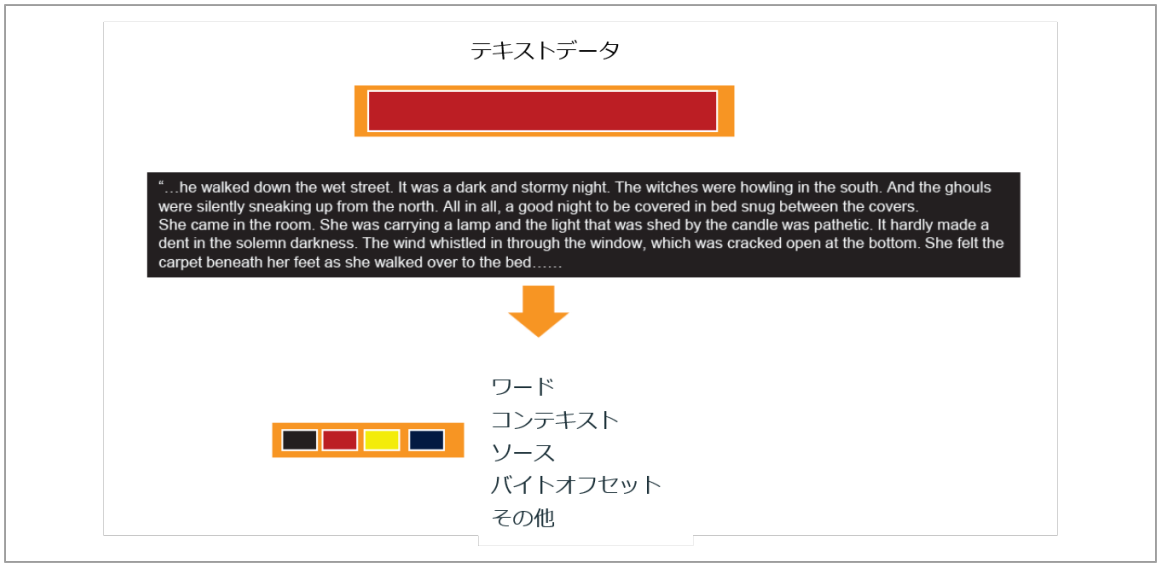


図 3-5：データレイクハウスに格納するテキストデータの種類

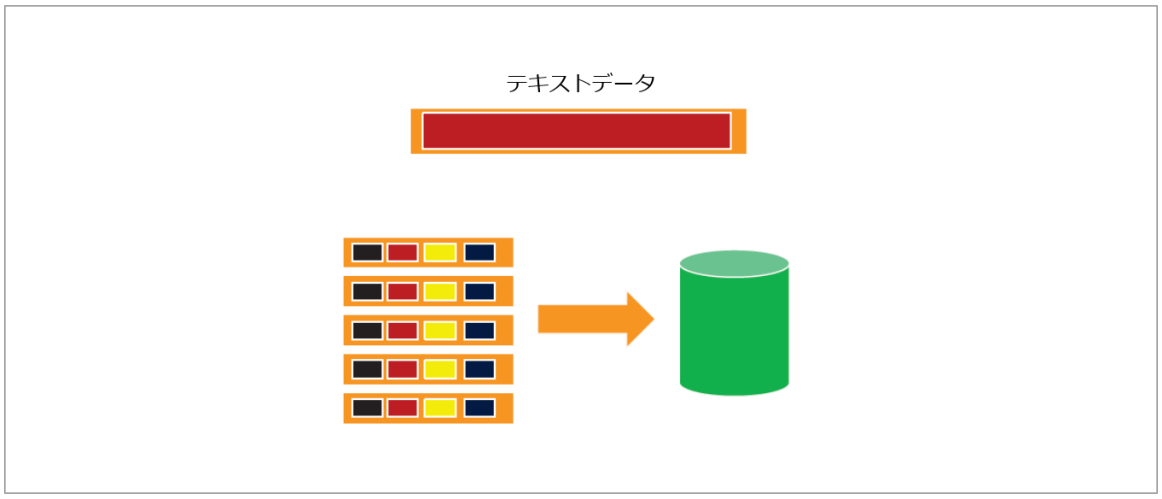


図 3-6：テキストをデータベース構造に変換

その他の非構造化データ

第三の種類は、その他のカテゴリの非構造化データである。非構造化データは、一般的には、マシンによって生成、収集されるアナログデータや IoT データのことを意味する。マシンが収集する測定値は、取得するデータの種類やマシンの用途によって異なる。測定値にはさまざまな種類がある。

- 時間
- 温度
- 処理速度
- 処理を行うマシン
- 処理のシーケンス番号

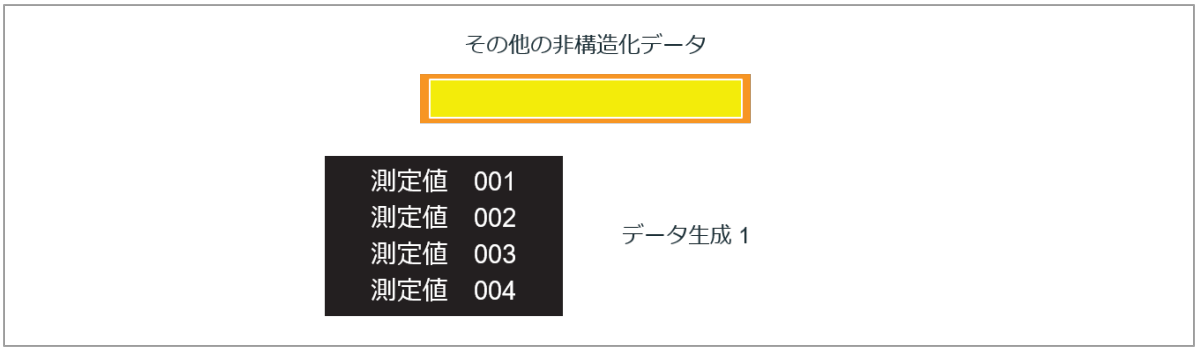


図 3-7：マシンから収集される測定値

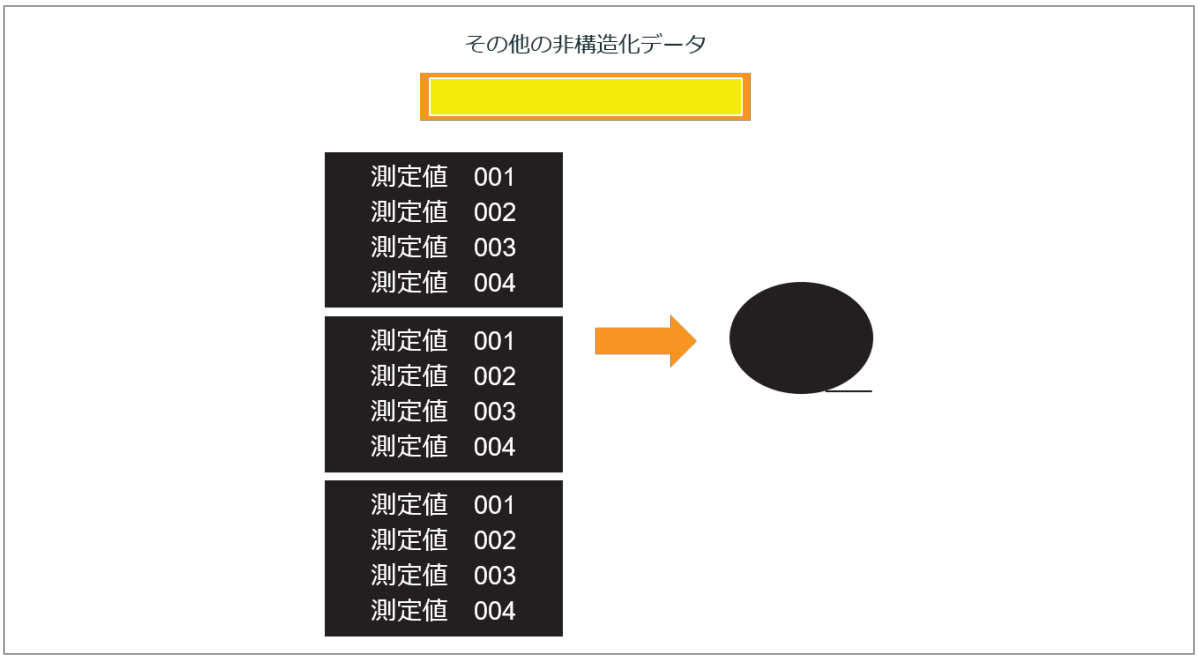


図 3-8：測定値は、測定値は秒や分単位などの周期で連続測定・取得され、データベースに格納される

データの種類によるデータ量の比較

データレイクハウスに格納される膨大なデータの量はデータの環境（種類）によって大きく異なる。通常では構造化データの量が最も少なく、テキストデータは構造化データよりも多いが、その他の非構造化データ（アナログ／IoT データ）の量には及ばない。その他の非構造化環境は圧倒的な量のデータを生成する。

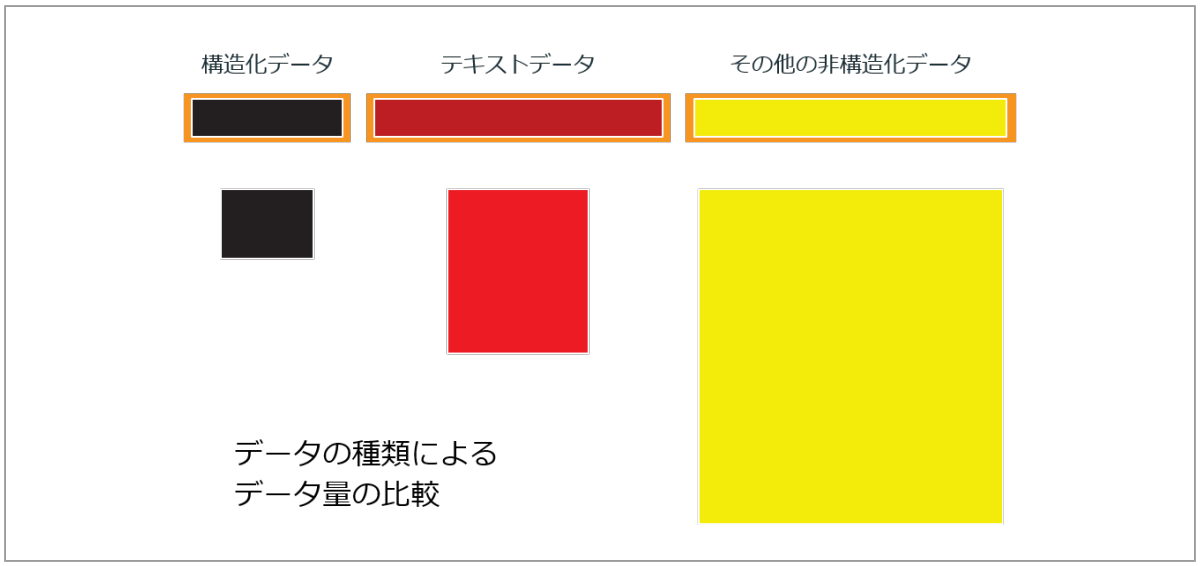


図 3-9：一般的な企業におけるデータ量の比較

多様なデータの種類の関連付け

データレイクハウスには、異なるデータ環境のデータを関連付けるという重要な機能を備えている。データレイクハウスで分析を行う場合に、異なる環境のデータの関連付けは極めて有用である。

分析を行うには環境内で共通キーを持つ必要があり、環境によって使用できる共通キーが異なる。データによっては、共通キーを拒むものや、共通キーが存在しないものがある。

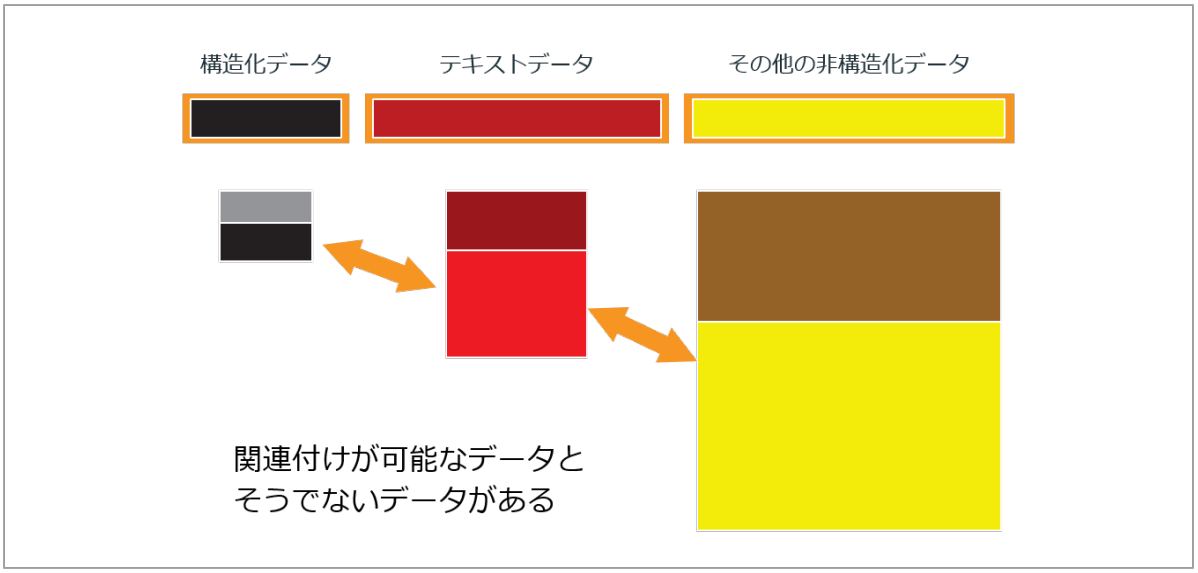


図 3-10：分析に必要な共通キーの有無

アクセス確率にもとづくデータの分類

アナログや IoT 環境からのデータは、バルクストレージに格納されることが多い。バルクストレージは低コストであり、データを格納する時点で都合がよいからである。しかし、バルクストレージは分析には適していない。そこで、アナログ／IoT データの一部をバルクストレージに格納し、残りを標準的なディスクストレージに格納するという手法が用いられる。アクセス確率が低いデータをバルクストレージに格納し、アクセス確率が高いデータをディスクストレージに格納する。この手法により、一部のデータを分析に利用するというニーズと、膨大な量のデータを格納するというニーズの両方を満たす。

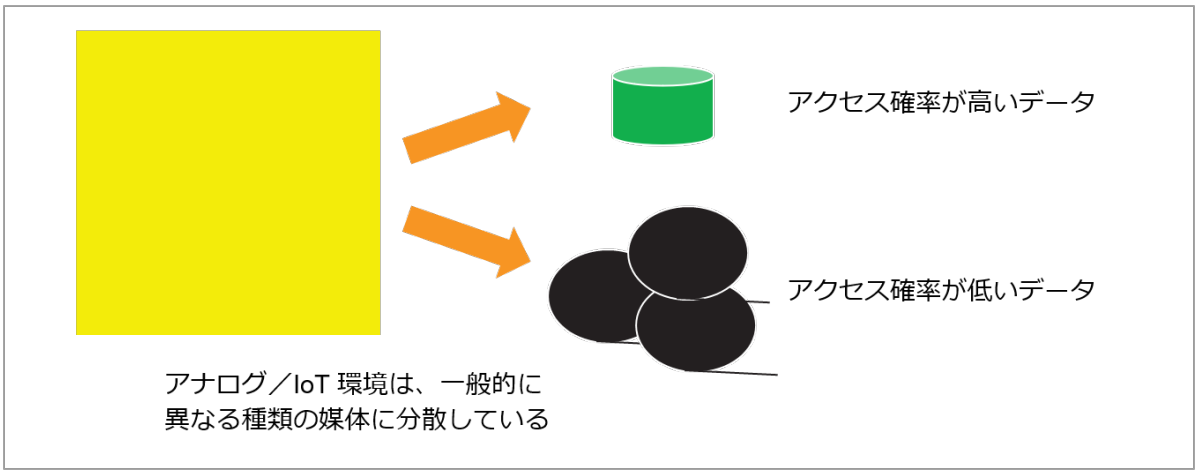


図 3-11：データの種類によって異なる媒体に格納

IoT／アナログデータの関連付け

アナログや IoT 環境からのデータは、関連性がある場合とそうでない場合がある。容易かつ自然に関連付けられるデータと、他のデータと容易に関連付けられない独立したデータがある。

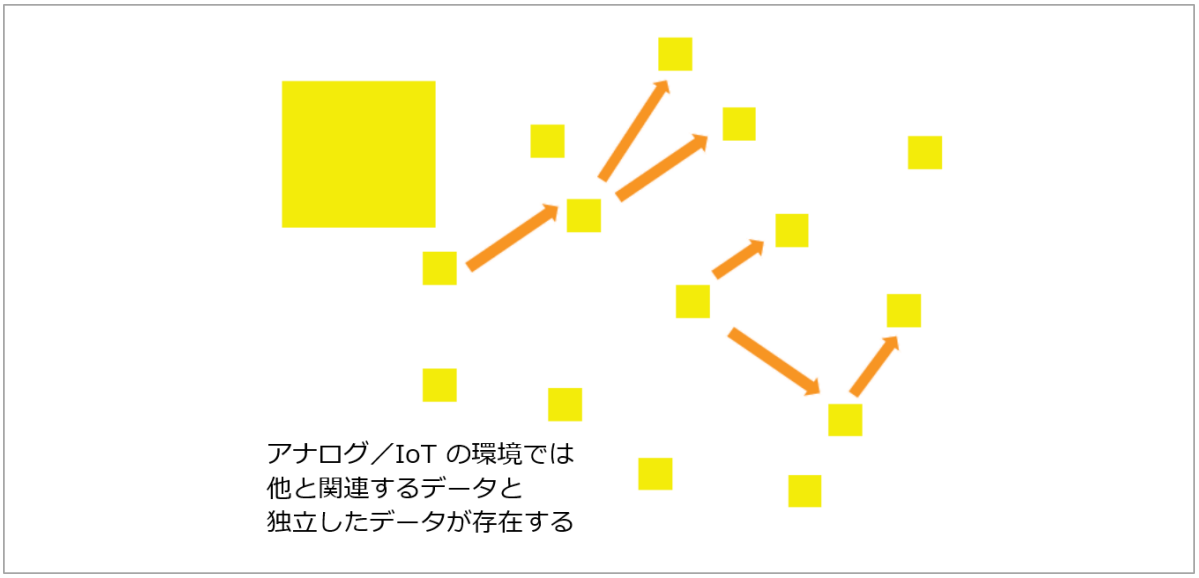


図 3-12：アナログ／IoT 環境には、他と関連するデータと独立したデータがある

データレイクハウスには多様な種類のデータが存在し、容量、構造、関連性をはじめとする属性はデータの種類によって異なる。異なる種類のデータに共通キーがない場合の関連付けには共通コネクタを使用する。

共通コネクタは、定型の手段がない場合にデータを関連付ける手段である。

共通コネクタは多数あり、以下はその例である。

- 地理／位置情報
- 時間
- 金額
- 名前

共通コネクタのシンプルな使用例として、地理データを考えてみる。さまざまな州における骨密度調査による大量のＸ線データがあるとしよう。

カリフォルニア州とメリーランド州のデータを選択し、それぞれの州のＸ線データを集約する。エンドユーザーは、各州のデータを使用して分析を行う。

一方、構造化された環境では、骨密度の低下に関連する医薬品の購入データを使用する。この場合も、カリフォルニア州とメリーランド州における購入データを選択する。エンドユーザーは、州ごとの医薬品の購入状況の差異を特定する。同様に、Ｘ線データから、州ごとの骨密度状況の差異を特定する。これにより、それぞれの州における医薬品の消費量と骨密度の差異分析をあわせた分析が可能となる。

骨のＸ線データと医薬品の購入データの間にはキー構造は存在しない。データを結びつけているのは、分析のために選択した地域と、骨密度の情報が収集された地域だけである。

同様に、金額や時間などを共通コネクタとして使用した分析も可能である。前述の例以外にも、多くの共通コネクタがある。例えば、人に関するデータでは、次のような共通コネクタが考えられる。

- 性別
- 年齢
- 人種
- 体重
- その他の身体の測定値

これらの共通コネクタを使うことで、異なる種類のデータの関連付けが可能になる。

分析インフラ

分析インフラはデータレイクにある未加工のデータをもとに構築される。分析インフラは、大規模な図書館のカードカタログに類似している。大きな図書館を訪れる際の状況を考えてみよう。まず、どのように本を探すだろうか？本棚を１つ１つ順番に見て歩きながら目当ての本を探していたら、図書館に長時間滞在することになる。一方、カードカタログを使用すれば、素早く効率的に検索できる。探している本をカードカタログで見つけたら、図書館のどこに行くべきかがわかる。このカードカタログと同様の役割を持つのが分析インフラである。

分析インフラは、分析対象のデータがデータレイクのどこにあるかを迅速かつ容易に見つけるための参照手段である。

図 3-13：分析インフラは、カードカタログのような役割を持つ

データが見つかれば、アクセスや分析が可能になる。データレイクと分析インフラによって見つかったデータの読み取りと分析にはさまざまな方法がある。

- SQL、R、Python
- Tableau、Qlik、Excel、PowerBI などの BI ツール
- リアルタイムアプリケーション
- データサイエンス、統計分析
- 機械学習

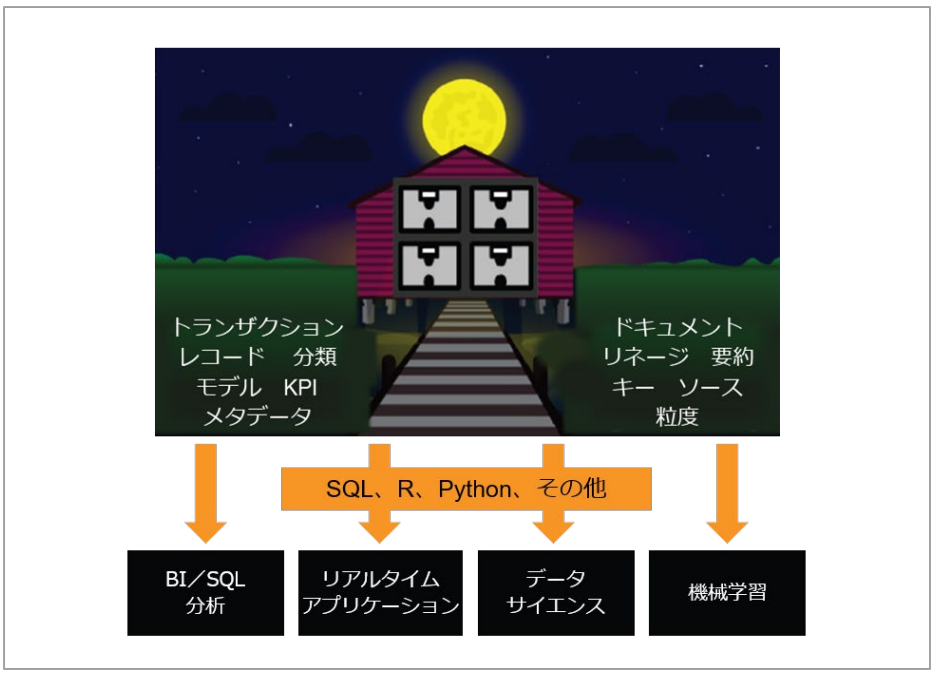


図 3-14：データレイクと分析インフラによって見つかったデータの読み取りと分析