データレイク導入の検討ポイント

五島 克樹 ：株式会社クロスフィールド

1. はじめに

ビッグデータは、2010 年代前半にいわゆる「バズワード」的な取り上げられ方をして以来、現在に至るまで産業、ビジネスの分野で注目を集めてきました。これまでに多くの企業がビッグデータの活用に取り組み、IoT や AI などとの併用で有効利用に至った事例も出てきています。

最近では、このビッグデータを保管するシステムとして、「データレイク」というワードを耳にすることが多くなってきました。データを保管するシステムといえば、データウェアハウス（DWH）やデータマート（DM）などのシステムが頭に浮かびますが、これらと「データレイク」の違いは何なのか、気になるところです。

本稿では、ビッグデータの解析基盤を構成する際にキーになるといわれている「データレイク」について、その役割を整理し、併せて導入を検討する際のポイントについて考察してみたいと思います。

2. ビッグデータとデータレイクの関係

まず、ビッグデータの定義について確認したいと思います。

> ビッグデータは、（中略）データの量であるとか、サイズを意味していません。ビジネスの判断に必要な「さまざまな形をした、さまざまな性格をもった、さまざまな種類のデータ」を意味し、それらを駆使してビジネスの効率や効果を上げることができる情報と、その情報の使い方のことを表すのです。

（岡村久和監修「IoT 時代のビッグデータビジネス革命」インプレス 2018 年より抜粋）

一例として「IoT 時代のビッグデータビジネス革命」（岡村久和監修、インプレス 2018 年）での定義を掲載しましたが、ビッグデータとデータレイクの関係を把握する上で重要になるポイントは、「**さまざまな形をした、さまざまな性格をもった、さまざまな種類のデータ**」という部分です。

つまり、ビッグデータとして取り扱うデータには、リレーショナルデータベース（RDBMS）や業務システムで扱う構造化データだけでなく、画像や動画、音声、SNS 投稿やメールといった非構造化データも含まれるということになります。

従来のデータ分析では、主に構造化データだけが分析の対象となっており、様々なシステムから収集した構造化データを格納する仕組みとして、データウェアハウスやデータマートが用意されていました。しかしビッグデータ分析では、構造化データに加えて非構造化データも収集・格納する必要が生じます。この必要性を満たすために生まれたのが、データレイクです。

データレイクは、規模にかかわらず全ての構造化データと非構造化データを保存できる仕組みです。データをそのままの形で保存できるため、データを構造化しておく必要がないという特徴を持っています。

3. データウェアハウス、データマートとの違い

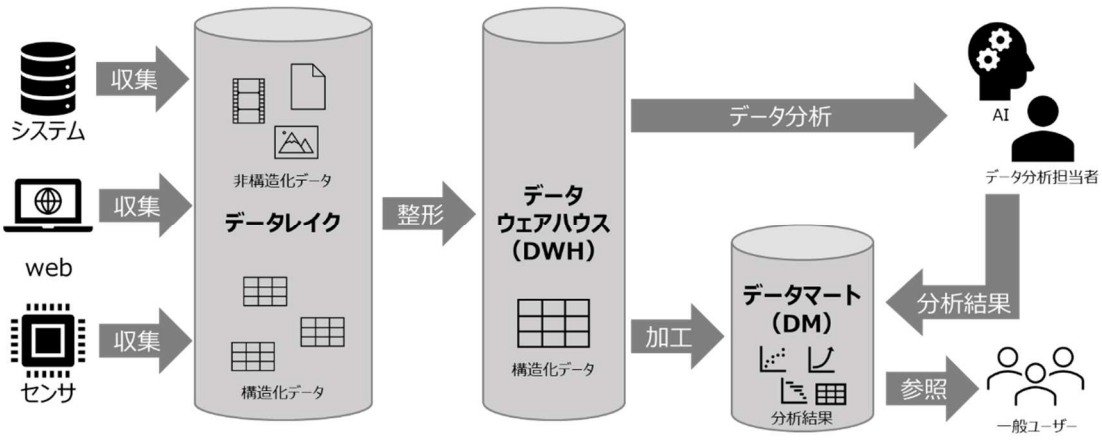
データレイクとデータウェアハウス、データマートの違いを大まかにいうと、前述の通り非構造化データを格納できるか否か、ということになります。

では、構造化データ、非構造化データの両方を格納できるデータレイクさえあればデータウェアハウス、データマートは不要なのか、というとそうではありません。それぞれ、データ分析という大きな仕組みの中で担う役割が異なっているのです。以下、データレイク、データウェアハウス、データマートの役割になります。

データベース	データ分析基盤の中での役割
データレイク	✓ データソース（システムやデバイスなど）から収集したRAW データを、そのままの形で保持する ✓ データレイクがあることで、分析し直す際にデータを再収集する必要がなくなり、データソースに不要な負荷をかけない
データウェアハウス	✓ データを整形（データの補完、修正、削除や、正規化などによって分離されているデータを結合）して、分析しやすい状態で保持する ✓ データ分析担当者は、このデータを利用して統計解析や機械学習を行うことで、データの価値を探索する
データマート	✓ 活用に特化した形式に加工したデータを保持する ✓ データ分析担当者ではない、一般のユーザーがこのデータを利用する

「データソース（システムやデバイスなど）」 → 「データレイク」 → 「データウェアハウス」 → 「データマート」とプロセスでデータを収集・整形・分析することで、はじめてビッグデータ分析ができる、ということになります。データレイクだけを導入してもビッグデータ分析は実現できない点に注意が必要です。

【分析基盤におけるデータレイク、データウェアハウス、データマートの役割】



4. データレイク導入の検討ポイント

最後にデータレイク導入を検討する際のポイントを述べたいと思いますが、具体的なポイントを見る前に、前提として理解しておきたいことがあります。それは「データレイク」はデータ保管の概念（コンセプト）であり、それを実現する手段は様々である、ということです。

“データレイク 製品 比較”などのキーワードで web 検索するとわかることですが、ERP パッケージなどと違い、データレイクの製品比較サイトなどは存在しません。なぜなら現状では、汎用的なデータレイク製品が存在するわけではなく、用途に合わせて Hadoop などのミドルウェアを使って構築するのが主流だからです。そのため、実際にデータレイクを構築する際は、ベンダーに要件を伝えてカスタムメイドで提案を受ける必要があります。

したがって本項では、データレイク導入を検討する際のポイントとして、データレイクの要件定義をする際のポイントを記載したいと思います。

① データレイク導入の目的は何か

まずは何を分析したいのかを明確化することが重要です。

データレイクに限らず、新しい技術を導入する際に、新技術の導入を前提に後付けで利用目的を考えるということが発生しますが、データレイクには相応のコストがかかります。せっかくデータレイクを備えたデータ分析基盤を作ったのに、以前と変わらない分析しか行わないなどということが起こらないように、目的起点でデータレイク=ビッグデータ分析の導入要否を判断しましょう。

なお先行事例を整理すると、ビッグデータ分析の目的は大体以下３つに分類することができます。これらの観点でビッグデータ分析の目的が何なのかを検討してはどうかと考えます。

目的	具体的な利用イメージ
現状を正確に把握する	✓ 現状把握に必要な情報をデータ分析や可視化ツールでわかりやすいレポートに転換すれば、経営層が組織全体の経営状況を、現場職員が特定の業務の進捗状況を把握できるようになる
課題の解決策を導き出す	✓ データ分析やデータマイニングを行うことで事物の法則や異常値を見出すことが可能。その結果、課題の原因を特定し、迅速に適切なアクションを取ることも可能となる。また、施策の効果をデータ分析によって検証し、施策を繰り返して改善することも可能
新たなビジネスチャンスを発見する	✓ 既存製品とサービス、バイヤーとサプライヤ、消費者の好みに関する情報を収集して統合的な分析を行うことで、企業が新たなビジネス機会を発見し、まったく新しいカテゴリの商品とサービスを創出することができる

② データ分析担当者を配置できるか

データレイクに格納されているデータを分析するには、格納されている非構造化データを構造化し、データウェアハウスに格納する必要がありますが、この構造化を行うにはデータの分析を専門とするデータ分析担当者の知見が必要です。一般にはデータ・サイエンティストと呼ばれます。

必ずしも社内の人財を充てる必要はありませんが、陳腐化する前にデータを利用する必要があることを踏まえると、社外の専門家をスポット的に起用するのは現実的ではありません。社外の人財を利用するのであれば、継続的に起用する必要があります。その場合かなりのランニングコストを要するので、社外の専門家を起用しつつ、その間に社内の人財を育成することが必要だと考えます。

③ 分析に利用しうる非構造化データが自社のビジネス上に存在するか

前述の通り、データレイクの特徴は構造化データだけでなく非構造化データも格納できるという点です。自社のビジネスにおいて、分析に利用できるような非構造化データがない場合は、データレイクは不要です（RAW データを直接、データウェアハウス上に保持すればよい）。

一般論になりますが、B2C に比べて B2B のビジネスを行う企業では、ビッグデータ分析に利用できる非構造化データは少ない傾向にあります。

なお、既存のデータをベースに考えるだけでなく、IoT などを利用して分析に役立つ非構造化データを新たに収集することができないかを考えるのも重要なポイントになります。

5. おわりに

非常に簡単ですが、以上がデータレイクの説明となります。データウェアハウスやデータマートなど、既存の仕組みとの違いをご理解いただけたでしょうか。本稿がデータレイク導入を考えていただくきっかけになれば幸いです。

なお、データレイク（湖）の対義語として、データスワンプ（沼）という言葉があります。これは、使えない、陳腐化したデータが溜まってしまった状態を表しています。

導入したデータレイクをスワンプ化させずに有効に利用し続けるには、他のシステムと同じように、データマネジメントやデータガバナンスを確立する必要があることを申し添えて、本稿を終わりにしたいと思います。