データレイクの“沼化”を防いで AI 活用を実現する「最強の」データ整理術

データを活用するためには、データをきちんと整理しないといけない

いまやデジタル・トランスフォーメーション（DX）による変革を目指す企業にとって、AI × データの全社的な活用は避けられない。しかし、学習モデルや分析用のデータがすぐに見つからない、探せてもそのままでは使えない、といった課題のせいでプロジェクトを推進できないことも多い。そんな企業のために、いま本当に求められるデータプラットフォームの要件と解決のアプローチ、さらに具体的なソリューションとは何かを解説しよう。

「AI 導入」の前に、対処しなければいけないことがある

IBM の調査によれば、いまや 82%の企業が AI を導入、または検討中という結果が出ており、多くの企業が AI を活用しようとしている。さらに AI に取り組む先進企業のうち、28%が新たなインサイトや価値を創出し、高い業績を示している。（図 1）

もちろん、残りの企業も AI の活用によって結果を出したいと考えているようだが、まだ効果は限定的だ。

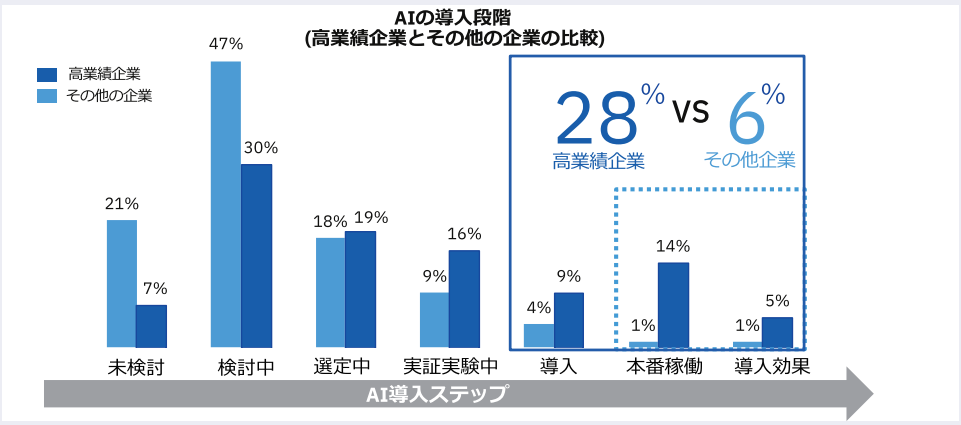
大抵の企業は社内にデータが散在し、どこに何があるかも見えていないという根本的な問題を抱えている。たとえデータを見通せてもデータを取得しづらかったり、データから分析モデルを作れても、何を参照しているのか把握しづらかったり、分析を自由に試行錯誤できる環境がないのだ。

また単に AI を導入し、効果測定だけで終わるのでなく、ライフサイクルを回しながら分析モデルをブラッシュアップしなければ、より良い効果も上がらない。

こういった課題を解決するには、何をすればよいのだろうか。

（図 1）IBM ユーザー調査結果から
今や 82% の企業が AI を導入、または検討中。
高業績企業はより積極的に AI に取り組んでいる

出典：“Shifting toward Enterprise-grade AI”
IBM Institute for Business Value, 2018



8 割以上の企業が AI を導入、または検討しているが、そのうち業績を上げている企業は 3 割弱という結果に留まっている

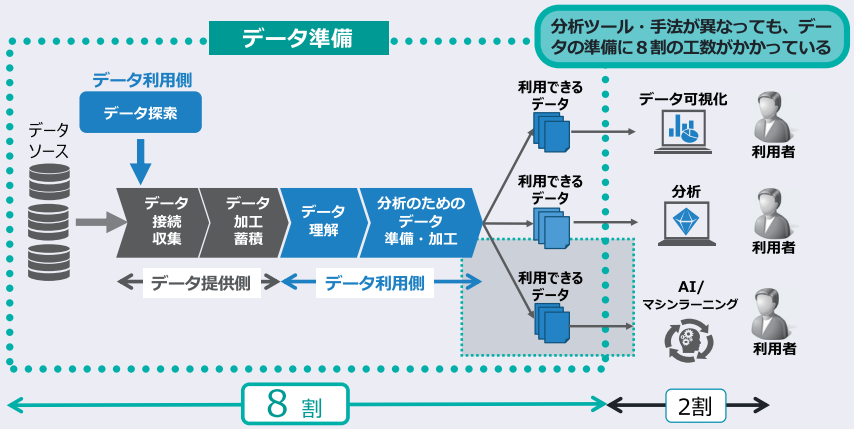
AI 本格導入の具体的な下準備

まずデータをしっかり活用できる IT 環境の整備が肝要だ。とはいえ、そこにも高い障壁が待ち受けている。マシンラーニングや AI などで、ちゃんと使えるデータを準備するのに時間がかかりすぎるのだ。実際に全体工数のうち約 8 割近くが、その前準備で占められているというから驚きだ。(図2)

重要な点は、ユーザーがどんなロケーションでもデータを取れる環境にすること、そして構造化／非構造化データなど、種類にかかわらずデータを取れるようにすることだ。分析に必要なデータを、しっかりと準備・加工しなければ、AI の学習には使えない。そのうえで、学習モデルを構築していくというアプローチになる。

そして精度の高い学習モデルに収斂させるには、結果をデータプラットフォームにフィードバックすることもポイントだ。AI 導入を成功させるには、企業全体でデータと AI の PDCA サイクルを回せる仕組みづくりが必要だ。

(図2) マシンラーニングや AI などデータを活用する前の障壁



マシンラーニングや AI などで、データを活用する前の障壁は高い。全体のうち約 8 割近くが、前準備で占められている

理想的なデータプラットフォームの形

前出のように、データプラットフォームを構築するうえで最も重要なことは、データの準備だ。ほとんどの企業がこの作業に工数を割いており、データを利用する側と提供する側で分断がみられる。

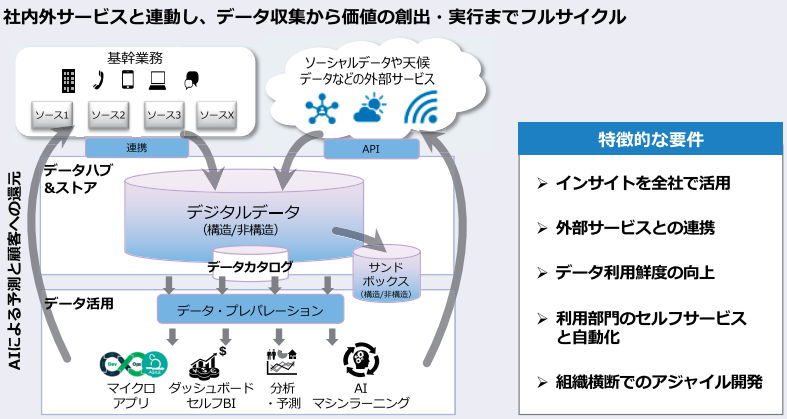
たとえば、データ利用側は、販売促進キャンペーンを張りたいと考えている。データから購買傾向を見て、優良顧客の予測リストをつくれば、より高い効果が得られるはずだ。しかし、その前に本当に必要なデータがどこにあるのかを把握できないのだ。

一方、データ提供側は、データのリクエストを利用側から受けても、どういうデータを渡せばよいのか明確ではないため、データを出すことができない。同じようなデータを他部署から依頼されることも多いが、大量のデータをロードするにも時間がかかる。さらに、それらのデータをガバナンスを効かせて管理するのも大変だ。

ここ数年では、社内外から調達してきたデータを 1 つの大きなデータレイクに溜め込んで、ユーザーに使ってもらうというトレンドがある。

ただし、実際にデータレイクを用意しても、すぐにユーザーがデータを利用できないという実情もある。というのも、データベース管理者ならフィールド名（データの要素の 1 つ）の意味するところが直感的に理解できるが、一般ユーザーはいきなりフィールド名を示されても、それが何のデータか想像できないからだ。そこでデータカタログをつくり、誰でもわかりやすい形で、欲しいデータを見つけられるような工夫が求められるのだ。(次ページ図3)

(図3) 多くの企業が描くAI×データプラットフォーム構想



理想的なデータプラットフォームの形。データ入手からモデル実装まで、継続的な学習を行いながら、PDCA を迅速に回せる

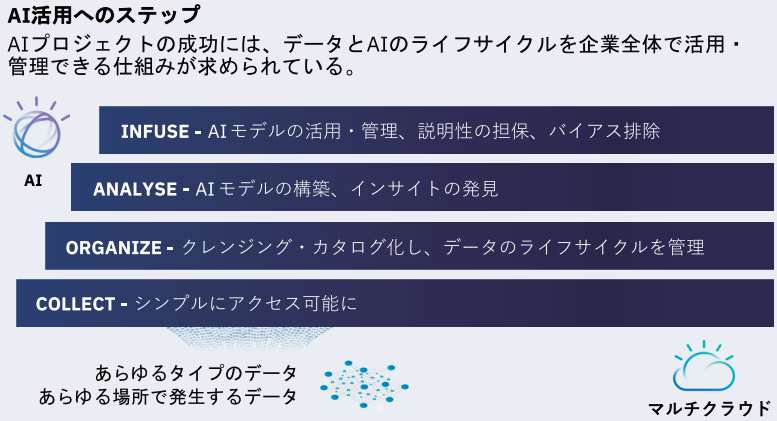
そして、ユーザーが検索したデータをサンドボックスに一時的に保管し、そこから自由に使えるようにすればよい。そのうえで、加工したデータを適用し、AI／学習モデルを作って実装し、ビジネス価値を創出していく。このようにして、データ入手からモデル実装まで、継続的な学習を行いながら、PDCAを迅速に回せるデータ管理プラットフォームが理想的といえるだろう。

「AI Ladder」に基づいた「IBM Cloud Pak for Data」とは?

IBM では AI × データのプロジェクトを加速させるアプローチとして「AI Ladder」というコンセプトを掲げている。これは、「COLLECT」（収集）、「ORGANIZE」（編成）、「ANALYZE」（分析）、「INFUSE」（活用・管理）の 4 つの AI 活用の仕組みをつくるということだ。（図4）

データにシンプルにアクセスできるようにして、そのうえでデータを取得してカタログ化する。さらにビジネス用語とデータベースのフィールド名を紐づけ、ユーザーがデータを活用できる形に整える。それらのデータを、そのまま次の分析環境で使う。もちろん構築した AI モデルの管理まで行えるようにしておく。一見ステップバイステップで取り組む図のように見えるが、これら 4 つを網羅的に行っていく必要がある。

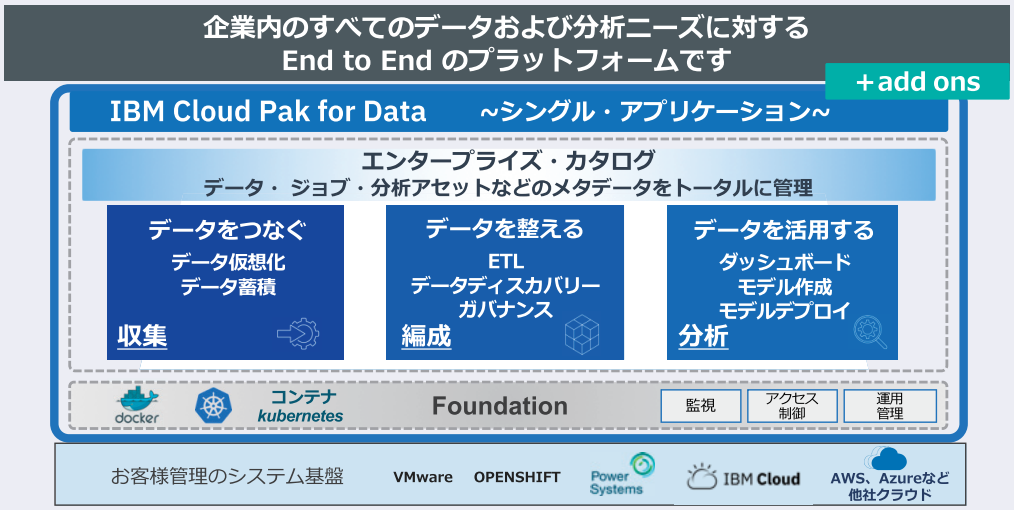
(図4) AI Ladder



AI プロジェクトを成功に導くためのコンセプト「AI Ladder」。4 つの AI 活用の仕組みを着実に作りあげていくことが近道

この AI Ladder のコンセプトに基づいて開発されたのが、次世代データプラットフォーム「IBM Cloud Pak for Data」（以下、Cloud Pak for Data）だ。これは、企業内のすべてのデータや分析ニーズに対応し、あらゆる AI システムの構築や運用に用いることが可能な統合基盤だ。（図5）

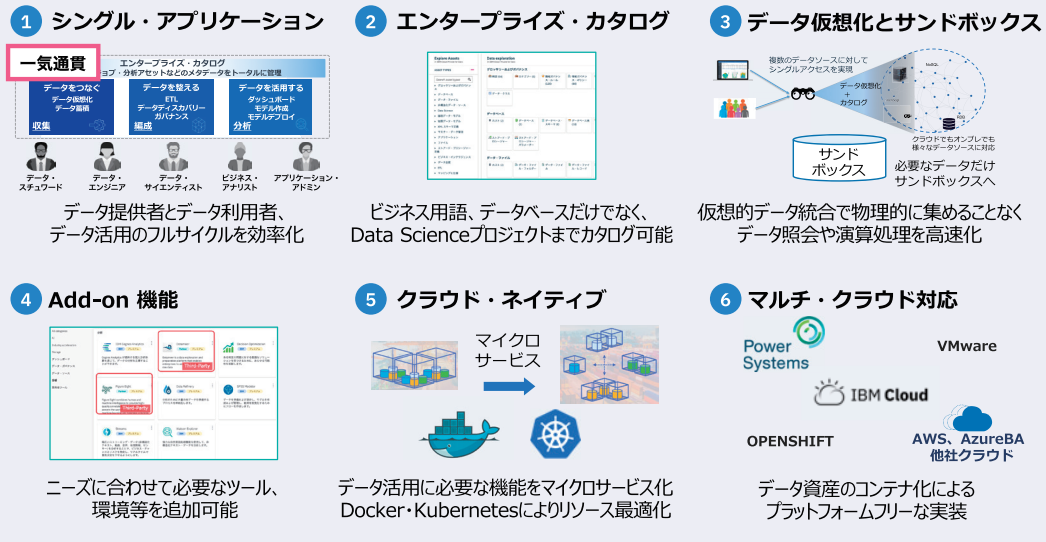
（図5）次世代データプラットフィーム IBM Cloud Pak for Data



次世代データプラットフォーム「IBM Cloud Pak for Data」。あらゆる AI システムの構築や運用に用いることが可能

Cloud Pak for Data の特徴は６つ挙げられる。1つ目は「収集」「編成」「分析」というデータ活用のフルサイクルを一気通貫で回せることだ。２つ目は、誰もがデータを見つけられるエンタープライズ・カタログを用意できること。カタログの範囲が広く、データベースのデータからビジネス用語や構築モデル、分析アセットなども対象とし、横断的に探すことが可能だ。（図6）

（図6）IBM Cloud Pak for Data の特徴



Cloud Pak for Data の特徴。フルサイクルでのデータ活用、広いカタログ範囲、柔軟な Add-on 機能、データ仮想化とサンドボックス、マルチクラウド対応など

3 つ目の特徴は、データ仮想化とサンドボックスだ。仮想的にデータを統合し、欲しいデータだけをサンドボックスに置き、仮想ビューからデータを照会したり、演算処理を高速化するものだ。これにより低コストなスモールスタートで始められ、ユーザー視点でデータレイクを構築できる。

Cloud Pak for Data の 4 つ目の特徴であり、ユニークな機能としては「Add-on 機能」が挙げられる。データプラットフォームに対して、ユーザーが求めるツールや環境を追加して利用できるのだ。これらには、IBM が用意する「SPSS Modeler」「Cognos Analytics」といったツールや新しい Watson サービス群、サードパーティ製のツールやサービスが含まれている。

また IBM では、Cloud Pak for Data の開発にあたり、従来のオンプレのソリューションを分解してマイクロサービス化している。これが 5 つ目の特徴だ。そして 6 つ目の特徴として、データ活用に必要な機能をパッケージングし、コンテナ化した Docker イメージと、Kubernetes で管理している。そのためバージョン管理の負担も軽減される。将来的にオンプレからの移行も意識し、ハイブリッドクラウド環境や、IBM Cloud、AWS、Microsoft Azure などのマルチクラウドにも対応している。

このように Cloud Pak for Data を導入することで、企業全体で AI× データの PDCA サイクルを回せる理想的なデータプラットフォームを構築できるようになるだろう。興味があれば、無料トライアルとして「IBM Cloud Pak Experience」も用意しているので、実際に試してみるとよいだろう。

※本記事は、2019 年 7 月にビジネス +IT にて掲載され、許可を経て転載したものです。

IBM Cloud Pak™ for Data の詳細はこちら https://www.ibm.com/jp-ja/analytics/cloud-pak-for-data

©Copyright IBM Japan, Ltd. 2019
〒103-8510 東京都中央区日本橋箱崎町 19-21
IBM、IBM ロゴ、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、Copyright and trademark information をご覧ください。