

「IAなくしてAIなし」、DX時代のデータ基盤構築の成功法則とは？

デジタルトランスフォーメーション(DX)実現の鍵は、いかにデータを有効活用できるかにかかっています。しかし従来の情報系システムに見られた、DWHを中心としたデータの可視化、分析の仕組みだけでは、残念ながらDXのニーズに応えることができません。情報系システムに加えて、新たに「分析系システム」の考え方を取り入れ、新たなデータプラットフォームを確立する必要があります。

従来のDWHアプローチではDXのデータ活用ニーズに応えられない

「デジタルトランスフォーメーション(DX)」という言葉がビジネスシーンにおいて大きく取り沙汰されるようになりました。特に近年ではコロナ禍に伴い、多くの企業がビジネスモデルやワークスタイルの大幅な転換を余儀なくされています。DXへの取り組みはまさに「待ったなし」の状況ですが、実際にDXで効果を上げた企業の例はまだ極めて少ないのが実状です。

DXを代表するテクノロジーといえば、何と言っても「AIを使ったビッグデータ解析」です。これを実践するために機械学習やディープラーニングの環境はかなり整備されてきたものの、その処理対象となる肝心のデータの整備が遅々として進んでいません。このことが、AI活用とDX推進を阻害する最大の要因となっているのです。

「AI向けのデータ整備」で必要なのが「Information Architecture(IA)」,つまりデータ活用のためのアーキテクチャを再考することです。しかし、実はAI向けのデータ整備やIAは、これまで企業がデータ活用を推進する上で解決のアプローチ、つまり「サイロ化克服」のために社内に散在しているデータをデータウェアハウス(DWH)構築して統合しようとする方法とは相性が悪いという課題があります。

その背景としては、DWHは「変更に弱い」という現状があります。しかしながら、DXの文脈でのビッグデータ分析では、決められたデータ集計・分析処理をくり返すことでは不十分です。新しい洞察を得るために、あるいはビジネス環境の変化に対応するために、AI分析には新たなデータを投入しながら、知見や

精度の高い予測モデルの発見していく必要があるからです。つまり、当初からデータモデルを頻繁に変更することが前提となります。

しかしDWHのアプローチでは、データ分析への過程において「ステージング環境へのデータ移行」→「ETL(Extract/Transform/Load)でのデータ加工」→「ヒストリカルデータのDWHへの格納」→「個別の分析ニーズに適したデータを抽出・加工しデータマートへ変換する」など、多くのデータレイヤーが生成されます。データ項目を追加する場合には各データレイヤーの構築や変換処理に多くの時間とコストがかかってしまうのです。

このように、DWHのアプローチでは、AI時代に適したIAを実現するのは難しいと言わざるを得ません。従来のDWHアプローチが抱える課題を洗い出し、DX時代にふさわしいデータ活用を実現するには、一体どのようなIAが適しているのかを解説します。

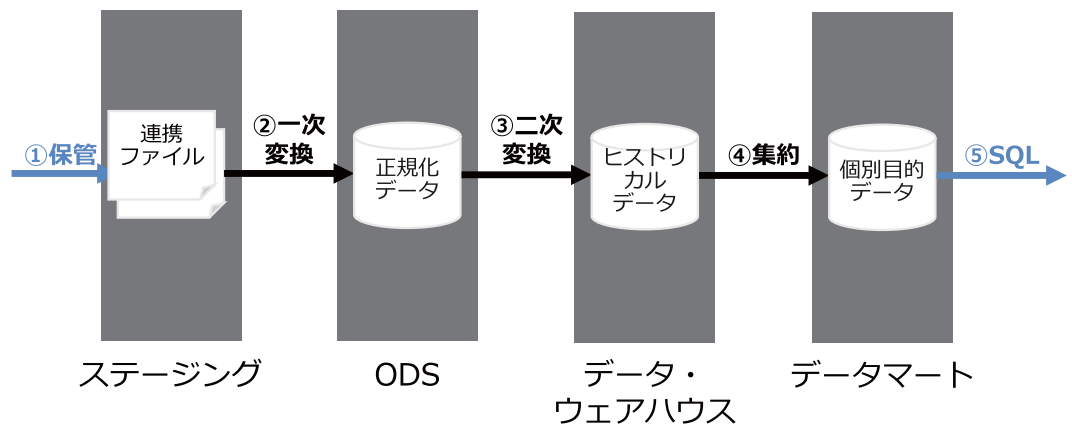
まず、DXを実現するために、「DWHアプローチが抱えるIAの課題」をまとめます。

DWHではデータの品質や信頼性を担保するために、いわゆる「第三正規形」に至るまでデータの正規化処理しますが、上述の課題に加え、これもデータ活用のスピードや柔軟性を損なう要因となっています。

ここまで正規化せずとも、その手前の「第一正規形」のままでも、データ分析には必要十分なデータかもしれません。第三正規化までしなければ、データベースの構築負荷はかなり軽減され、鮮度の高いデータを用い、データ活用をスピードアップできるはずなのです。

実際、ある企業では1年間かけて大規模なDWHを構築したものの、「ユーザー受け入れテスト」で業務側から欲しいレポートとは異なるといった問題が多数、指摘されました。要件定義フェーズで念入りにデータモデルを設計したとしても、机上の検討だけで実際の利用シーンを十分に想定したデータモデルを作り上げるのは難しいのです。データ品質についても慎重に検討したものの、実際の業務を想定したテストでは適切なデータが入っていないなどと次々と問題が露呈しました。

さらに言えば、このような大規模DWH構築プロジェクトの最大の課題のひとつは、システム構築に時間がかかりすぎる点にあります。長い時間をかけて設計、構築し、ようやく完成した時には既にビジネス要件が変わってしまっており、当初の設計では実践的な分析ができないというケースも決して珍しくないのです。



これまでのデータウェアハウス・アプローチは変更しにくい

情報系システムとは別に「分析系システム」を新たに設ける

では、従来のDWHアプローチが抱えるこれらの課題を克服する「DX時代のIA」とは何でしょうか。

これを考える上では、まずシステムを「SoR (System of Record) / SoE (System of Engagement)」と「Sol (System of Insight)」の2種類に分ける必要があります。前者は社内業務や顧客サービスを支えるオンライン系の各種業務システムが該当し、後者はデータ可視化・分析のための情報

系システムが該当します。

情報系システムのインフラは通常、DWHを中心に構成されます。前項でも述べたように、DWHは構築に時間がかかります。また、日次夜間バッチでデータの移行や変換を完了させるなど一定レベル以上の処理性能や可用性が求められます。一方IBMでは、この情報系システムに加えて、新たに「分析系」システムをSol領域に加えることを提唱しています。

分析系システムとは、情報系システムとは異なり、あらかじめ決められたデータだけでなく、あらゆるデータを柔軟に追加、蓄積して自由に分析できる仕組みを提供するものです。情報系システムに相当する性能や可用性は必要なく、より柔軟にデータを扱えるようになります。

このような分析系システムを作ることが、DXに向けたデータ活用の第一歩と言えます。この分析系システムに継続してデータを蓄積していき、データや分析モデルの有効性を検証します。その過程で使いそうなデータやモデルが見つかったら、それを改めて情報系システムや業務系システムに反映させて、業務の中に本格的に組み込み価値の創出を図ります。

IBMでは、業務系システムと情報系システム、そして分析系システムから成るこのような構成を、DXに向けたデータ活用のためのリファレンスアーキテクチャとして提唱しています。近年注目される「データ

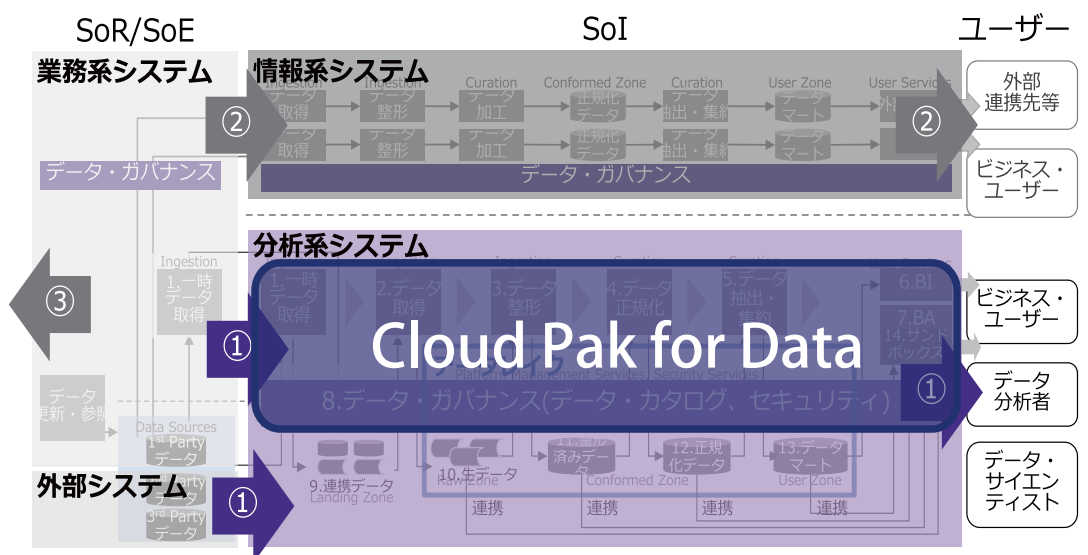
レイク」の考え方も、この分析系システムのコンセプトは包含しています。

IBMでは、分析系システムのためのプラットフォーム製品「Cloud Pak for Data」を提供しています。この製品は、さまざまなデータレポジトリと接続してデータを収集し、それを整理・加工、分析する機能を備えた統合データプラットフォームです。OpenShift上に構築され、データの増加に合わせてスケール可能です。増えたデータを管理するためのデータ・カタログの機能も含まれています。

リファレンスアーキテクチャにおいて重視されているのが、データについての統制を適切に行うための「データ・ガバナンス」です。適切なデータ・ガバナンスを実現するには、上述のような「データ・カタログ」と呼ばれる仕組みも必須となります。データカタログの仕組みによって、どこにどのようなデータがあり、それぞれがどんな性質を持つのかを把握できるようになります。



IBMが考える成功法則



Cloud Pak for Dataが実現する領域

データ・ガバナンスの仕組みに加え、組織・プロセスの確立が不可欠

データ・カタログとは、データに関するメタデータを収集、蓄積することにより、各種データの所在を明確化し、データの品質を可視化するとともに、データの適切な利用を管理するための仕組みです。

管理すべきメタデータには、ビジネス視点でデータ項目の名称や意味を定義した「ビジネス・メタデータ」、物理テーブル・物理ファイルの名称や定義情報、データ間の連携などを示す「テクニカル・メタデータ」、データ参照時の制約や更新履歴など、参照時に留意すべき観点を示した「オペレーショナル・メタデータ」の3種類が存在します。

これらデータ・カタログを実際に構築するにあたっ

ては、すべてのデータを対象に一度に作るというのはあまり現実的ではありません。実際には、業務やデータ分析のユースケースごとに必要とされるデータも異なるため、ユースケースごとに、そこで扱われるデータについてデータ・カタログを作成し、段階的に整備していくやり方が推奨されます。

こうしてデータ・カタログを整備し、データの所在や特性をシステムで把握できるようにする一方、その仕組みを基に、社内でデータが適正に扱われるよう統制を掛けるデータ・ガバナンスの体制を整える必要があります。

データ・ガバナンスが有効に機能していないと、

データ・カタログの概念

メタデータの収集・蓄積により、データの所在の明確化、データ品質の可視化、データ利用の適切な管理を行うこと

メタデータの種類

ビジネス・メタデータ

ビジネスの視点でデータ項目の名称と意味を定義した

テクニカル・メタデータ

物理テーブルや物理ファイルの名称と定義情報、データ間の関連を示したものの

オペレーショナル・メタデータ

データ参照時の制約や更新履歴など参照時に留意すべき観点を示したものの



データの所在の明確化

データ・レイヤーごとにデータの所在がわかるよう整備し、ユーザーが参照できるようにする



データ品質の可視化

データの欠損値や整形加工・来歴を可視化し、分析時の参考情報とする



データ利用の適切な管理

使用許諾と公開範囲、オーナーシップの管理をし、データ管理の統制をはかる

データ・カタログとは？

各ユーザー部門



自ら欲しいデータをアクセス、編集を行うため、各種データがサイロ化し、横断的なデータ活用にはできていない

現状の課題



IT部門

ユーザー部が作成するデータを管理・監視することや、要望を聞きつつ、必要なデータを全て用意する事は困難

新たな価値の創出

- ✓必要な部門から小さく、**クイックにスタート**
- ✓これまで利用が困難だった**他部門のデータ**などを有効活用できる
- ✓IT部門やパワーユーザーに頼らず、**セルフサービス**でデータ活用できる

データ・ガバナンス

新たな組織の組閣



データ・リスク管理とコンプライアンス

- ✓事前にデータ構造の設計が不要で、**アジャイル的に開発**できる
- ✓新規のデータ要件に**迅速に対応**できる
- ✓データ・ガバナンスの**効果的な展開**が可能になる

ビジネスとITをつなぐデータ・ガバナンス

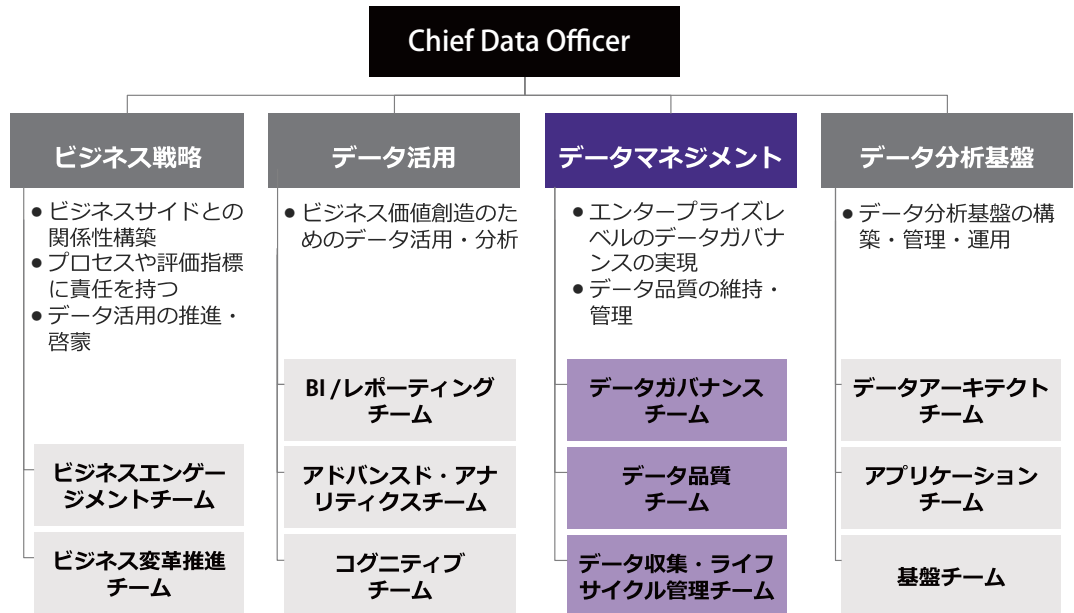
業務部門はそれぞれの目的に沿って独自にデータ活用を進めてしまうため、データのサイロ化に歯止めが掛からなくなってしまいます。またそうした状況においては、IT部門の管理業務の負荷も高まる一方で、データ漏えいのリスクも増大してしまうでしょう。

そうした事態を防ぐには、やはり全社的なデータ・ガバナンスの取り組みが不可欠であり、その際にはITツールの導入をはじめとするテクノロジー面での施策だけでなく、業務プロセスと組織体制面での施策も欠かせません。

とくに組織体制は極めて重要です。IBMではこれまでさまざまな企業のデータ・ガバナンスを支援してきた経験から、取り組みを一過性のもので終わらせず継続的なDXを実現するには、CDO(Chief Data Officer)の配下にビジネス戦略チーム、データ活用チーム、データマネジメントチーム、そしてテ

クノロジーチームの4つの組織を置くのが最も効果的との結論に至っています。

データプラットフォーム構築を成功させる最大のポイントは、とにかく欲張らずに、できる範囲からスタートし、段階的に規模を拡大していくことです。上述のCloud Pak for Dataのように、データの増大に合わせて柔軟に規模をスケールアップしていくことができる製品を利用し、スケラビリティに優れたIAの仕組みを導入することが、DX実現に向けた第一歩として極めて重要だと言えるでしょう。IBMでは、企業のデータ活用とDXを促進すべく、このような製品に加え、IA (Information Architecture)の構築、データ・ガバナンスを含む組織体制の構築をどのように進めるかについてもご支援しています。



継続的なDXを実現するための組織体制例

日本アイ・ビー・エム株式会社

●お問い合わせ

日本アイ・ビー・エム株式会社

お問い合わせフォーム：<https://ibm.biz/otoiawase>

※お問い合わせをいただく際には、「SB クリエイティブ掲載『データ基盤』の件」と、お申し付けください。



