（執筆 / 安田 央奈）

データマネジメントの高度化に対応するための DataOps の導入
俊敏で柔軟なデータ処理を可能にする新しいデータマネジメント手法

データに基づく意思決定は、企業の競争優位性の獲得に重要な役割を果たすと期待され、必要なデータを必要なタイミングで取り出すことのできるデータマネジメントの構築が重要となってきている。しかし、外部環境が激しく変化し、データの急増、ビジネス目標の高度化に伴って、高品質なデータを得るためのプロセスは複雑化している。企業は数多くの課題に対応しながら、多種多様な大量のデータを俊敏かつ柔軟に処理できる体制を構築する必要がある。

本稿では、高品質なデータを俊敏にかつ継続的に処理できるようにするための新たなデータマネジメント手法である DataOps について詳述する。

1. データマネジメントの高度化にいかにして対応すべきか

現代の市場を取り巻く外部環境の変化はスピードも速く不確実であるため、企業は迅速で正確な意思決定を求められるようになっている。精度の高いデータ分析による客観的な洞察は正確な意思決定を支援し、企業の競争優位性の獲得に重要な役割を果たすことが期待される意思決定に必要なデータを必要なタイミングで取り出すことのできるデータマネジメントを計画し実践することは、昨今の企業活動において重要である。

ビジネスの現場では、製品開発や顧客提供サービスなど様々なユースケースにおいてデータ分析が活用されている。しかし、そのビジネスの現場からのデータに対する要求は従来よりも高度化しており、必要なデータを必要なタイミングで入手することが難しくなっている。

まず、分析に必要とされるデータの量が急増している。売上データや顧客の購買履歴データは、特定の店舗に限らず全国・全世界の店舗のデータ、自社の Web ページ、その他の EC サイトなど様々なソースからデータを収集して、より多くのデータから分析を行っている。

データの種類も多様化している。企業はより深い洞察を得るために Excel や CSV ファイル、リレーショナルデータベースで管理され構造化されたデータの他にも、XML や JSON ファイルといった半構造化データや、画像やテキスト、音声などの非構造化データを活用している。構造や品質の異なるデータに対しては、それぞれ適切な取り扱いが求められている。

データの品質に対しての要求も高まっている。誤記や欠落のない正確で完全なデータであること、最新のデータであることなど、ビジネスの現場は精度の高い分析を行うためにより高品質なデータを求めている。また、様々なソースからデータを収集して一つのデータに統合することも求められているため、異なるソース間での用語やフォーマットの違いを整えて一貫したデータにしたり、重複を排除した一意なデータに整形するなど、高品質なデータを得るためのプロセスは複雑化している。

そして、外部環境の激しい変化に伴い、ビジネスの現場の目的に適うデータの変化が急速であることは一つの大きな課題である。例えば、データの取り扱いに関する各種法規制の施行と改正がある。日本における個人情報保護法は 3 年ごとの見直しとなっている。企業によっては GDPR や、2023 年に CCPA から CPRA（カリフォルニア州プライバシー権法）に改正された欧米のデータ規制にも対応していく必要がある。Safari や Chrome などの主要ブラウザに

よるサードパーティクッキーの廃止によって、これまで活用していたデータが取得できず、ファーストパーティクッキーなどの代わりとなる別のデータへの対応が必要となる。それだけでなく、ターゲットとなる顧客層の変化に伴う新たな分析軸の追加や、競合他社の新たな動向によって、必要なデータがどのようなものであるか、どのような管理が適切であるかは絶え間なく変化している。外部環境が変化しても必要とするデータの仕様や条件を変化させず使い続ければ、陳腐化し低品質化したデータによって誤った意思決定を行い、市場における競争の優位性や顧客からの信頼を損なう恐れがある。

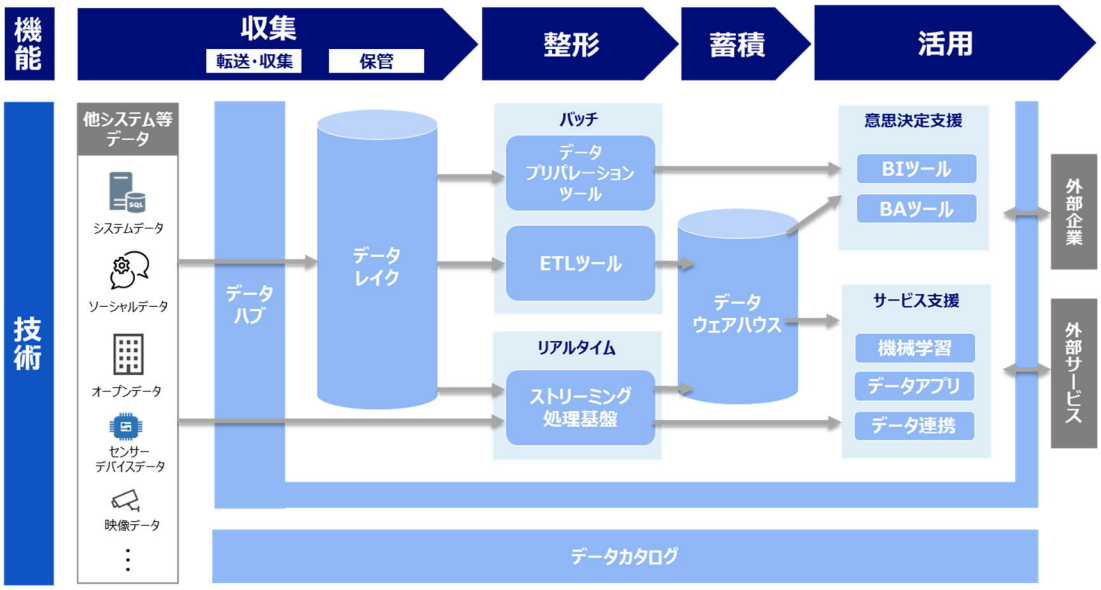
つまりデータに基づいた意思決定に取組む企業においては、データの利活用を取り巻く課題に対応した、多種多様な大量のデータを俊敏かつ柔軟に処理できるデータマネジメントの構築が必要になる。それが構築できなければ、必要なデータを必要なタイミングで取り出すことは難しくなる。データに基づいて意思決定を行うデータドリブン型の先進的な企業では、従来のデータマネジメント手法では限界があることが認識され、より俊敏で柔軟なデータ処理を可能とするデータマネジメント手法の活用が進み始めている。

2. DataOps によるデータマネジメントの俊敏性と柔軟性の向上

データマネジメントには、データの「転送・収集、保管、整形、蓄積、活用」といった各処理のプロセスが含まれている（図 1）。例えばフォーマットが統一されていない複数のソースから収集したデータを一つに統合して分析する場合には収集したデータをそのままの状態で分析に使うことはできないため、一旦データレイクに保管し、データフォーマットの変換やクレンジング、統合などデータの前処理を行ったうえで、データを蓄積する領域となるデータウェアハウスやデータマートへ活用・分析用のデータとして移動させるといった処理が行われる。

これらの処理プロセスを効率的に行うためには、データの分析環境を設計・実装・運用する役割を担うデータエンジニア、データの処理・分析・結果の評価を行い、ビジネスにつながる知見を生み出すデータサイエンスの専門家、データや分析の結果を受けてビジネスの現場においてデータを扱うユーザなど、様々な専門知識やスキルを持つ人材が必要とされている。

図 1　データ活用基盤の全体像[1]



データエンジニア、データサイエンティスト、データユーザがデータマネジメントのどのプロセスを担うかは企業によって状況は異なるが、「転送・収集、保管、整形、蓄積」を担当するデータ作成側と、「活用」を担当するデータ消費側にチームとプロセスが分断されることが多い。チームごとに担当範囲を特化することによってそれぞれのチームが高い専門性を持つことができるようになる。一方でデータ消費側からデータ作成側へ必要なデータを要求したり、データ作成側からデータ消費側へ処理が完了したデータを渡すなど、チームを横断してデータの配信をスムーズに行う必要がある。チーム間の連携プロセスが適切に調整されていなければ、データの品質、適時性、透明性、機動性を妨げることになる。その結果、データの受け渡しのタイミングが遅れたり、データの要求が正しく伝わらず不完全なデータが作成され作業に手戻りが生じたり、データデリバリーの遅延を招く恐れがある。

高いビジネス目標と複雑化するデータマネジメントプロセスに対応しながらチームやプロセス間のギャップを解消するためには、データマネジメントに先進的に取組む企業によって様々なデータマネジメントの手法が実践され、その有効性が検証されてきた。例えば、「顧客との協調」や「変化への対応」を原則[2]とするアジャイル開発や、データ処理のパイプラインを自動化し常に新しいデータを利活用できる継続的デリバリーなどがある。それらの様々な手法から収束するかたちで、DataOps という新しい手法が実践され始めている。

DataOps とは、技術のみならず人やプロセスの変革に取組み、膨大な数のデータソースから作成される高品質なデータを、データの消費者に向けて俊敏にかつ継続的に配信するためのデータマネジメント手法である。DataOps はクラウドや AI・機械学習などの新たな技術と、アジャイルや継続的デリバリーなどの手法を複合的に組み合わせて従来のデータマネジメントプロセスを変革することで、データの急増や要求の急速な変化、チームやプロセスの分断によって生じる様々な課題の解消が期待できる。

企業の状況やユースケースによって組み合わせのベストプラクティスは異なるが、以下の 3 点は基本的要素として

1　「DX 白書 2023」図表 5-34「データ活用基盤の全体像」

2　https://agilemanifesto.org/iso/ja/manifesto.html

備えておくことが望まれる。

(1) データマネジメントプロセスの自動化

人力での操作では時間がかかっているデータ処理を効率化するため、各種データマネジメントの処理やワークフローの自動化を行う。自動化によって、各処理にかかる時間を短縮し、分離されている処理タスク同士を連携させタスク同士の間にあるリードタイムを短縮し、データデリバリーを迅速に行うことが期待できる。また、手動操作によるミスやエラーを減らし生産性を向上させ、データの品質を高く保つうえでも効果的である。

(2) データマネジメントプロセスに対する可観測性

可観測性とは、データマネジメントを行うシステムにおけるデータの状態や、データ処理プロセスの状態が可視化され、理解可能な状態であることを示す。データマネジメントプロセス全体を可視化し、データの品質に影響を与えている欠陥やデータデリバリーにおけるボトルネックをすばやく特定して解決するために、データマネジメントプロセスの観測は必要不可欠な要素である。

(3) 共通の目標に向けて協力できる部門横断的な組織構成

データの作成側とデータの消費側の双方が、高品質なデータの俊敏かつ継続的な配信を共通の目標とし、その実現に向けて協力する必要がある。高度なデータマネジメントプロセスの実現と、変化に対する俊敏で柔軟な対応のためには、異なる部署・役割の人々による互いのノウハウや専門知識を生かしたシームレスな連携が必要である。データの作成と消費を融合する中間的な役割など、これまでとは違った新しい役割やその役割を組み込んだ組織構成が求められるようになってくる。

3. DataOps のケイパビリティ構築を支援する技術や取組

DataOps は、データの処理を行うデータ作成とデータ分析をビジネスに活かすデータ消費側の双方のプロセスが連続的に循環し、継続的に改善されていくことが重要である。特に技術と人とプロセスの変革が相互に補完しあうことが重要である。そのためには、個別のデータ処理にかかるコストを削減し効率化する自動化と、欠陥やボトルネックを特定し解決するためのデータマネジメントプロセス全体に対する可観測性、そして互いのノウハウや知識、役割をシームレスに連携するための部門横断的な組織構成のいずれも欠かすことのできない要素となる。新しい技術を活用したツールの導入など技術への投資は、コミュニケーションやプロセスの変革などを強力に支援することが期待できる。

DataOps のケイパビリティ構築を支援する二つの取組について紹介する。

(1) DataOps のケイパビリティ構築を支援する技術的な取組

① データの前処理の自動化

データの処理プロセスのなかでも、前処理にあたる整形のプロセスは他のプロセスと比較して多大な時間と労力が費やされるため、多くの企業に共通する課題となっている。2022 年に Anaconda 社が実施した調査[3]でデータマネジメントにおける一連の処理プロセスに対してそれぞれどれくらいの時間を費

[3] https://www.anaconda.com/state-of-data-science-report-2022

やしているのかを尋ねたところ[4]、「データ準備」に 22%、「データ整形」に 16%が費やされているとの回答が示された。データマネジメントにかかる業務時間の約 40%がデータの前処理に費やされており、データの前処理のプロセスの自動化はデータ作成側の業務を効率化し、負担を軽減する重要な取組である。

データプリパレーションツールは、データの整形処理をするためプログラミング言語を使用せずに GUI 操作で実行し、AI・機械学習がデータの欠損値や異常値を検知し補完するなど、簡易な操作あるいは自動でデータを整形することができるツールである。従来はプログラミングによるルールベースのデータの整形処理でも迅速的かつ効率的にデータを処理できているが、昨今のデータの増加、多様化、分散化に伴って複雑な処理が必要となるデータに対しては AI・機械学習によるデータの整形処理の自動化は必要不可欠となってきている。例えば、ウエスタンユニオン社、トムソン・ロイター社、欧州トヨタ社といった世界に複数の事業拠点を持ち、数億件分の顧客データを分析している企業は、AI・機械学習によって自らデータソースを分析してデータの整形及びデータの作成を自動で行う Tamr などのデータマネジメントの自動化のツールを活用することで高品質な顧客データを迅速に獲得できる体制を構築している[5]。データプリパレーションツールは、ソリューションベンダーや導入事例が多く確認され、ツールとして生産性の安定期に入っている[6]。各ベンダーのデータプリパレーションツールは得意とする処理や機能があり、導入に際して目的に合わせて適切に選定する必要がある。それぞれのツールは差別化が成されているため、PoC などを通じて事前の評価を行うことで、導入効果を高めることが可能となる。

仮想統合ツールは、データ自体の意味・構造などのデータに関する付随情報であるメタデータを活用することで、大容量のデータ自体を実際に抽出、ロード、変換せずとも、データ自体の構造や状況を仮想的に捉えてビューを作成し、その上で統合データを作成することが可能となる。メタデータを活用し仮想統合する場合のデータソースから活用までの流れを図 2 に示す。仮想統合では収集したメタデータに基づいてデータを仮想的に再現するだけでなく、GUI 操作で整形処理のシミュレーションを行う。整形処理のシミュレーションをもとに物理的な統合データを作成する機能を有するツールもあるため、データソースの追加や新たな統合データの作成を俊敏かつ柔軟に行うことが可能となる。仮想統合はデータ作成の俊敏性と柔軟性のメリットがある反面、それらのメリットはベンダーがツールにあらかじめ搭載しているコネクタやデータ処理機能に依存しており、自社向けにスクラッチ開発されたシステムや一部のアプリに対応していない場合がある。その場合には、仮想統合ツールの導入に際して自社のシステムに対応するカスタムコネクタの開発や、独自システムからデータを抽出する API の設定を行う必要がある。

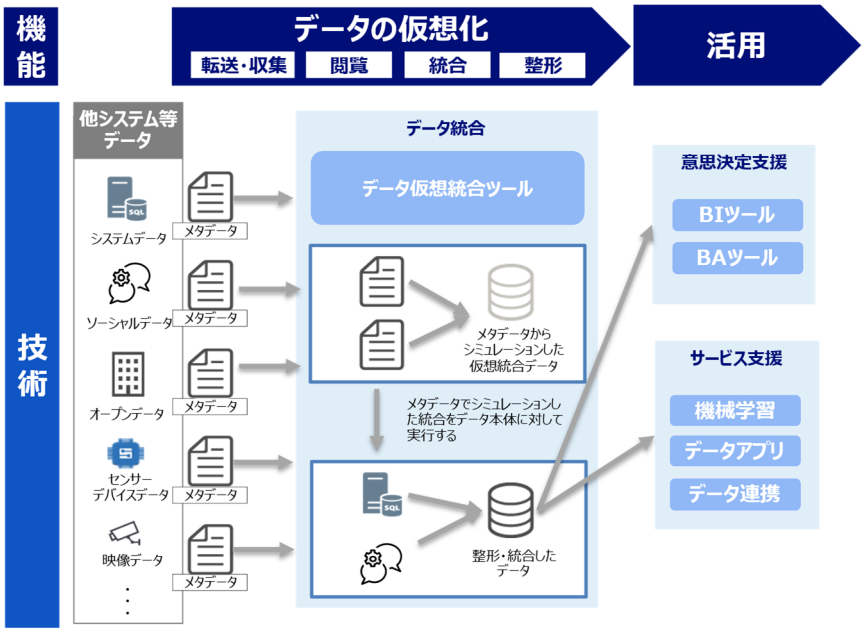
---

[4] アンケート原文は「How much time they spend on the above tasks?」であり、各工程に対する回答時間から各割合が算出されている。

[5] https://www.tamr.com/customer-stories/

[6] https://www.databricks.com/resources/ebook/hype-cycle-for-data-management

図 2　仮想統合のデータ活用基盤の全体像



従来のデータプリパレーション、仮想統合は一部機能の自動化や可視性の向上によってデータ前処理の俊敏性と柔軟性を高めているが、より高度な自動化によって人力による操作を限りなくゼロに近づける ETL 自動化ツールが提供され始めている。ETL 自動化ツールとは、データの抽出（Extract）、変換（Transform）、ロード（Load）の処理を行う ETL プロセスを自動で行う技術である。ETL 自動化ツールは、プログラムや AI・機械学習によって抽出元のデータの構造を自動分析し、ロード先でのデータの蓄積に適した構造に自動的に変換処理を行う。このツールの導入は、データの保管から蓄積までに消費していた労力やデリバリー時間をゼロへと近づけられる可能性がある。例えば、2022 年に AWS が年次イベント「re:Invent 2022」で発表した「ゼロ ETL」という ETL 自動化ツールは、RDBMS である Amazon Aurora のデータを抽出し、データウェアハウスである Amazon Redshift へと自動的にかつニアリアルタイムでロードすることが可能である。

ETL 自動化ツールはデータマネジメントの自動化の潮流の中でも比較的新しい技術であり、現時点で可能なのは特定のデータソースやロード先に対する限られた自動処理にとどまる。AWS のゼロ ETL は Amazon Aurora から Amazon Redshift 間の限定的な機能となっている[7]。他の ETL 自動化ツールも、主要な業務アプリケーションのデータをデータウェアハウスへロードするコネクタを備えており主要な前処理のプリセットが備えられているが、プリセットにない前処理に関してはユーザの手動による設定が必要となる。ETL の自動化については、抽出とロードの高度な自動化が進んでいるが、変換については人力による操作を完全にゼロにすることは難しいのが現状である。

[7] https://aws.amazon.com/jp/events/reinvent-recap

② データマネジメントシステムの統合管理

データレイクやデータプリパレーションツール、データウェアハウスなどのデータマネジメントの処理機能がそれぞれ独立したシステムとして構築されていると、システムを統合管理する仕組みが必要となる。従来は個別にシステムを設計し、個別にデータパイプラインやアクセスコントロールを設定する必要があったが、昨今はデータマネジメント一連のプロセスを統合管理するツールが活用され始めている。プロセスを統合管理することで以下のような効果が期待できる。

一つ目に、データマネジメントの一連のプロセス間でのデータの受け渡しを自動的に処理できるようにプロセス全体を運用管理して、データデリバリーにかかる時間を短縮できる。

二つ目は、データ処理フローの管理運用をシームレスに行うことが可能になる。新しいデータソースからデータを抽出する設定や、保存先のデータレイクやデータウェアハウスの変更のほか、新しいシステムの追加などとのシステム同士の連携操作を比較的容易に行うことができる。

三つ目に、データデリバリーの一連のプロセスを俯瞰的に観測することができる。データの「転送・収集、保管、整形、蓄積、活用」の一連のフローからログなどのメタデータを収集し、エラーやボトルネックの原因をデータマネジメントプロセス全体から分析し、特定することが可能となる。

データマネジメントシステムの統合管理ツールのプロバイダーはAWSやGoogle、Databricksなどの、データの収集から分析まで広範囲な機能をサポートしている規模の大きなデータマネジメントクラウドサービスプロバイダーにとどまっているため、選択肢はあまり多くはない。組織のデータの利活用のユースケースや要件に合わせた適切な統合管理ツールの選択が重要であり、ベンダーロックインによるリスクにどのように対処するかの注意が必要である。

(2) DataOps実践に向けた人とプロセスの変革

DataOpsの特徴 は、膨大な数のデータソースから作成される高品質なデータの俊敏かつ継続的な配信である。その実践のために、これまでのデータマネジメントに必要とされてきたプロセスや新しいプロセスを組み合わせてプロセスの間で密接に連携をとる必要がある。データの収集のプロセスでは、常に変化するデータソースに対応しながら、必要なデータの発見とアクセスを俊敏に行い、次の段階であるデータの保管や整形などのプロセスと連携することが重要である。整形のプロセスでは、蓄積や活用のプロセスが常に高品質なデータに迅速にアクセスできるように、ビジネスの現場のデータの用途や要求を理解するといった連携が必要になる。

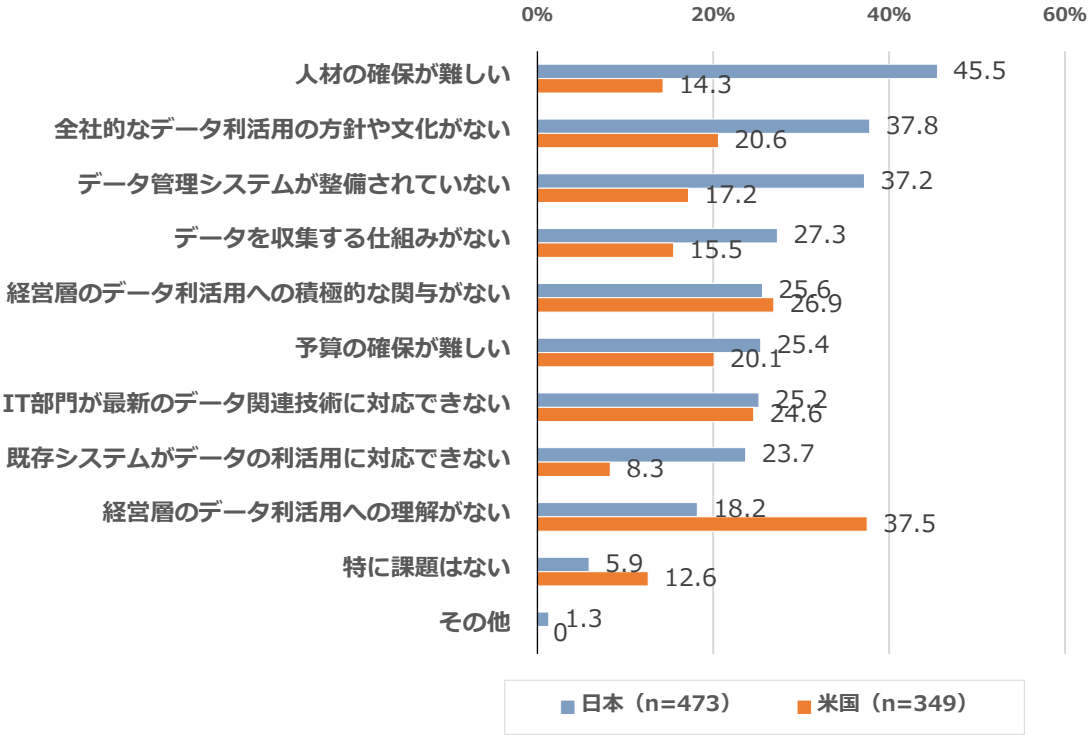
また、DataOpsの実践においては、作成されたデータに見つかった問題点などのフィードバックを他のプロセスと共有し、解決のために連携をとることが推奨される。例えば、データの消費者がデータの作成者から受け取った整形済みデータに問題を発見した場合。自身で対応できる範囲内で修正して対処しようとするのではなく、データの収集や整形などの他のプロセスにフィードバックを共有して、原因を放置させないことが大切である。根本の原因を明らかにしていくことで問題の再発を防ぎ、データマネジメントプロセスの全体を改善していくことができる。

このように DataOps には、これまでとは違ったプロセス同士の密接な連携など、変革的なデータマネジメントを実践するチームが必要不可欠である。DataOps の実践を目指す企業においてはチーム構築の手段を人材獲得に依存せず、プロセスの変革に目を向けることが重要となってくる。

データ利活用に取組む多くの企業は、データエンジニアやデータサイエンスプロフェッショナル[8]のように専門スキルを持つ人材の獲得によってデータマネジメントチームを組織してビジネスの推進に大きな成果を上げてきた。このようなデータエンジニア、データサイエンスプロフェッショナルなどの人材主導によるデータマネジメントチーム組織運営の成功例にならおうとして、人材の確保がデータの利活用推進の重要課題であると意識している日本企業は少なくはない。「DX 白書 2023」にも掲載されている「企業を中心とした DX 推進に関する調査」において、日本と米国それぞれの企業にデータ整備・管理・流通の課題を尋ねたところ、日本企業は「人材の確保が難しい」と回答したのは 45.5%に対し、米国は 14.3%であった（図 3）。

需要に対する供給の不足から人材獲得の競争は激しさを増しており、データエンジニアなどの人材を新たに雇い入れるというデータマネジメントチームの構築方法は多くの企業では困難になってきている。

図 3　データ整備・管理・流通の課題[9]



しかし、昨今はデータエンジニアリングを専門としないビジネスの現場の人材でも自動化ツールの支援によってデータ処理操作が可能である。つまり、これまでデータ消費側とされてきた人々をデータ作成の人材として活かしながら、多種多様な大量のデータを俊敏かつ柔軟に処理する DataOps 型のデータマネジメントチームを構築することができるようになっている。この手法は、データ消費側のプロセスにデータ作成側のプロセスを一部融合し、両者が分離することで生じていたデータデリバリーの遅延や適時性、透明性、機動性の課題

[8] 「デジタルスキル標準(https://www.ipa.go.jp/jinzai/skill-standard/dss/index.html)」の DX 推進スキル標準における「DX の推進において、データを活用した業務変革や新規ビジネスの実現に向けて、データを収集・解析する仕組みの設計・実装・運用を担う人材」

[9] 「DX 白書 2023」図表 5-67「データ整備・管理・流通の課題（複数回答）」を日本企業の回答が多い項目順に編集

が解消されることが期待できる。

データの前処理の役割をデータエンジニアに限定しないデータマネジメント手法によって DataOps に必要な各処理プロセスの連携は密接になるが、データエンジニアリングの専門家ではないためデータの構造の理解を深めることに時間を費やし、知識不足からくるミスがデータの品質を低下させるリスクが生じることも考えられる。このようなリスクに対処するためには、データを利用する従業員に対するデータマネジメントの基礎知識やツールの操作方法のトレーニングの実施が重要である。

4. DataOps を実践することで組織はどのように変わるか

データマネジメントの自動化ツールや、統合されたシステム、プロセス間の密接な連携など DataOps に基づいたデータマネジメントを継続することによって、多種多様な大量のデータを俊敏かつ柔軟に処理できる体制の構築と維持に繋げることができる。この DataOps の実践は企業のデータドリブンな組織運営を支援し、組織を以下のように変えていくことが期待される。

(1) データプロダクトを実践し、データの更なる高品質化を推進

高品質なデータを迅速に得ることによって組織外でもデータ対して高い価値を認められるようになり、新たなビジネスチャンスや新たな価値を創出することが可能となる。高品質なデータの活用によってビジネス上の利益を追及するデータプロダクトという手法を実践していけるようになる。

データプロダクトを実践している企業はデータの使用頻度や、使いやすさ、財務的な影響などを測定し評価することで、データプロダクトによる価値創出を最大化するべく更なる品質向上に努めている。例えば、35 名のデータプロダクトマネージャーからなるチームを組んでデータプロダクトを推進している米国のマーケティング企業 Vista 社は同社のデータプロダクトが毎年 9000 万ドル近い増分利益をもたらしていることを測定しており[10]、データの品質が事業にいかに大きな影響を及ぼしているのか明確化している。データとビジネス上の利益の密接な関係を数値化し、観測することで、組織にとってのデータマネジメントの意義はさらに重要なものとなってくる。

(2) データの民主化の促進

ツールの導入や組織構成の見直しによって、特定の技能の有無によらずあらゆる従業員がデータに基づいて意思決定を行えるようになることで、データの民主化の促進が期待される。データとビジネスニーズの密接な結びつきは組織全体でのデータの利活用促進に必要な条件である。例えば、スイスの製薬大手ノバルティス社はデータプラットフォームの統合や人工知能の導入、技術担当チームの増員などデータの利活用に様々な投資を行ったが、データサイエンティストがビジネス現場の業務を理解していなかったため当初は多くの取組が不調に終わった。その後、ノバルティス社は効率性やパフォーマンスの改善に取組んだ。ビジネスニーズを理解している現場側の従業員をデータサイエンティストと共に働かせ、現場の従業員が自分自身でデータを処理できるように研修に力を入れるようになった。ビジネスニーズに紐づいたデータ利活用が推進され、データに基

[10] https://hbr.org/2022/10/why-your-company-needs-data-product-managers

づく売上予測や、注文システムの再構築などのイノベーションが起きるようになってきた[11]。

データの民主化はビジネスニーズに即したイノベーションを生み、組織全体のデータの利活用を飛躍させることが可能となる。

【お問合せ先】
独立行政法人情報処理推進機構
社会基盤センター　イノベーション推進部　先端リサーチグループ
E-mail：ikc-ar-info@ipa.go.jp
電話：03-5978-7522

[11] https://dhbr.diamond.jp/articles/-/8857