（執筆 / 安田 央奈）

「データの民主化」
従業員によるデータ利活用の拡大

データマネジメントの一連のプロセスのうち、データの準備に費やされている時間は分析に費やされている時間より多く、データ準備工程の効率化は重要な課題となっている。データプリパレーションツールはデータの整形や統合といった準備処理を簡単な操作、あるいはAI・機械学習で自動実行できる機能を備えており、データ準備を効率化するにとどまらず、非技術者がデータ準備を実行することを可能にする。

本稿では、データ準備工程の効率化におけるデータ準備処理を行うツールの簡易化・自動化の潮流と、それによる「データの民主化」について詳述する。

1. データマネジメントにおける低付加価値業務

顧客データの分析によるパーソナライズサービスの提供や、気候や人の移動といったあらゆるデータからのリスク予測など、データ分析に基づいた意思決定を行うことは、ビジネスにおいて必要不可欠になりつつある。先進的な企業はデータからの更なる価値創造を追求し新しい技術を取り入れており、データマネジメントで活用されている技術の潮流はデータ利活用を戦略に組み込んでいくうえで注目し対応していかなくてはならない動向となっている。

データによる価値創造やイノベーションにおいて、データエンジニアやデータサイエンティストは中心的役割となることが期待されている人材である。データサイエンティストとは統計学や情報科学理論に基づいてデータから洞察を得るプロフェッショナルであり、データエンジニアとはデータ利活用の基盤を構築・運用する技能を持ったプロフェッショナルであり、企業によっては非常に高額な報酬を設けている専門職である。

ところが、データサイエンティストやデータエンジニアに集まる期待や脚光とは裏腹に、データを収集、整形、分析するデータマネジメントの一連のプロセスの大半は創造的で変革的な価値創造ではなく、付加価値の低い地味な作業が占めている。

2016年にCrowdFlower社（現Figure Eight社）がデータサイエンティストに対して、データマネジメントのプロセスのうち何に最も時間を使っているのか調査[1]を実施した。そして、79%がデータ準備工程に最も時間を使っているとの結果が示された（Cleaning and organizing data60%、Collecting data sets19%の合計である）。2020年にAnaconda社が実施した別の調査[2]ではデータマネジメントの一連の各工程において、データサイエンティストがそれぞれ何割の時間を費やしているのかが調べられた。最も高い割合となったのがデータ準備（Data preparation）の22%で、次いでデータ整形（Data cleansing）17%となった。データマネジメントにおいて、業務時間の約40%がデータ準備工程に費やされているのである。

新たな洞察を見出す分析の工程よりも、付加価値の低いデータ準備工程に長い時間が費やされているとの結果は衝撃的であり、データ準備工程の効率化は重要な課題として度々議論されてきた。

---

[1] https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

[2] https://know.anaconda.com/rs/387-XNW-688/images/Anaconda-2021-SODS-Report-Final.pdf

本稿では、データ準備工程の効率化におけるデータ準備処理を行うツールの簡易化・自動化の潮流と、それによる「データの民主化」について詳述する。

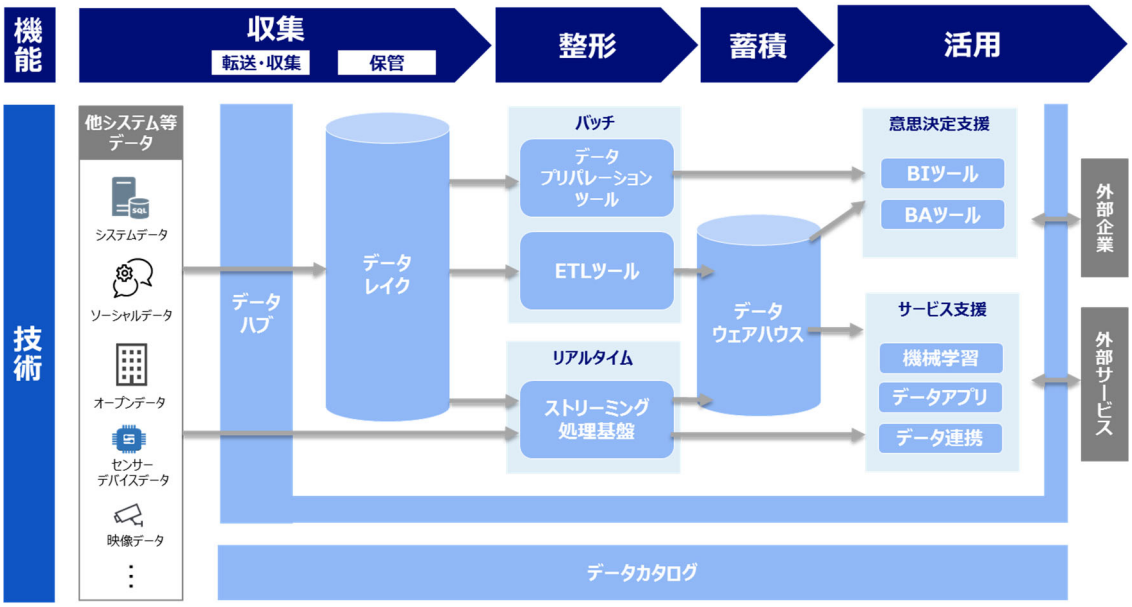
2. 非技術者によるデータ準備が可能となる簡易化・自動化の技術潮流

(1) データ利活用におけるデータ整形の必要性

データ利活用をビジネス戦略上高いプライオリティに位置づけている企業は、財務データ以外にもシステムデータやソーシャルデータ、IoT など各種デバイスから生成されるデータも利活用の対象としており、このように多様なデータから価値ある洞察を得るためには分析に応じてデータを準備することが必要である。

データは発生してから分析できるようになるまでに図表 1 に示す「転送・収集、保管、整形、蓄積、活用」の各段階に応じて適切な形状で取り扱われ処理される。

図表 1　データ活用基盤の全体像[3]



「転送・収集」「保管」の段階では様々なソースからデータを集めてくる。データレイクを使えば、データの構造やファイルフォーマットを問わず読み込んだデータをそのままの構造で保管することが可能となる。保管時のデータに対する加工処理を最小限に留め、比較的発生時の生データに近い状態で保管できるデータレイクの特性は、将来的に必要となる可能性があるデータを保管する点においても重要であり、必要性は高まっている。

一方、データウェアハウスには、AI・機械学習やグラフ作成の可視化ツールで分析できる状態に整形し、構造化したデータを蓄積する。多様なソースから収集したため生じる表記ゆれの修正や、不要データの削除、関連データの統合などの整形は、データから洞察を得るうえで重要な準備工程である。

例えば、COVID-19 やインフルエンザなどの感染症対策のため、感染者数の予測分析をしようとする場合。公的機関が公開している感染者数の変移、病院の患者収容状況、交通機関が公開する人の移動増減など各ソース

[3] 「DX 白書 2021」図表 42-6「データ活用基盤の全体像」

から関連データを収集するが、感染症表記が「COVID-19」と「コロナ」で統一されていなければ「コロナ」表記になっているデータを「COVID-19」に統一するよう修正し、位置に関する情報を番地まで含んでいるデータと含んでいないデータが混じっている場合には番地を削除して市区町村までの表記に統一したり、番地まで含んでいる別のデータを作成または入手したりして、データを整形する必要がある。

データ利活用のユースケースが増えれば、それに合わせたデータを作成しなくてはならないため、データマネジメントが効率よく実施できていなければ、使いたい時に使いたいデータや分析結果を得られず、ビジネス上の成果を得られなくなってしまう。従来は、データの変換や統合を実行できるプログラミング言語の Python や Ruby、データベース言語の SQL の技能を習得しているデータサイエンティストやデータエンジニアなどの技術者が整形処理の主な担い手とならざるをえなかったが、データプリパレーションツールがその状況を変え始めている。

(2) データプリパレーションツールによる整形処理の簡易化・自動化

昨今のデータマネジメントにおいて、データへの接続や整形処理をノーコード/ローコード[4]による操作で実行したり、整形が必要なデータを AI・機械学習が検知したり、簡易あるいは自動でデータ準備を行えるデータプリパレーションツールの導入が広まりつつある。当機構が 2021 年に日米企業に実施したアンケート調査ではデータプリパレーションツール（データ整備ツール）を活用していると回答した日本企業は 21.1%（「全社的に活用している」が 6.8%、「事業部で活用している」が 14.3%）だが、米国企業は 70.4%（「全社的に活用している」が 50.1%、「事業部で活用している」が 20.3%）と高い比率を示した。

データプリパレーションツールの導入により、従来データ準備を担ってきた技術者側のプログラムを組む手間や時間を軽減するにとどまらず、非技術者である事業部側でもデータの準備を容易に実行できるようになる。

データプリパレーションツールは主に以下の機能を有している[5]。

① データ接続

データベースやファイルに接続し、目的のデータを取得する機能である。

多様な外部ソースとの接続をサポートするコネクタを備えており、新たなソースを追加する場合のパイプラインを容易に構築することができる。

② データ確認

取得したデータの分布や、整形処理前後のデータを確認する機能である。

非構造化データファイルのデータからメタデータ[6]を分析して、構造化させてテーブル状に表示したり、簡易なグラフを自動作成したりできる。一部のデータプリパレーションツールでは、クレジットカードや電話番号など個人に関する情報を自動検知して、そのデータが含まれる列や箇所をハイライト表示し取扱い注意を促す機能が組み込まれている。

③ データ整形

---

[4] プログラミング言語なしで処理を実行できる（ノーコード）、あるいは簡単なプログラミングで処理を実行できる（ローコード）機能・特性

[5] 「DX 白書 2021」図表 42-11「データプリパレーションツールの主な機能」をもとに作成

[6] ここでのメタデータとは、データの意味や構造、特性などといったデータに関する付随情報を示す

欠損データの補完や、表記ゆれの修正、データの分割、重複データや外れ値データの削除などといったデータを整形する機能。ツールにより整形パターンは異なる。

例えば、データの作成元の違いから日付データの書き順に「dd-mm-yy」や「yy-mm-dd」、「mm-dd-yy」のようなバラつきがあった場合に、指定した日付の書き方に自動的に修正するといった機能を使うことができる。ツールによっては、AI・機械学習により整形の必要なデータを検知して表示する高度な機能を有する。

④ データ結合

指定する条件のもと、複数のデータを結合する機能。

複数のソースから関連するデータを収集し統合することで分析の精度を高めたり、より深い洞察を得たりすることが期待できる。

これらの機能を使用すれば、上述の感染者数予測分析のようなデータも事業部内で準備して分析できる。データの専門家でない者がデータプリパレーションツールを使って準備する場合、一度の操作では高品質のデータ作成は困難であるが、操作性の容易さから試行錯誤の繰り返しを通じてデータを洗練させていくような使い方ができる。

データプリパレーションツールの利活用からは、データ準備時間の短縮の成果も期待できる。データ準備の時間短縮は、データを使ったビジネスプロセス全体の時間短縮にも繋がる。例えば、GlaxoSmithKline 社（以下 GSK 社）では、以前は科学者や研究者は IT 部門から必要なデータが届くまで数週間から数カ月待たなくてはならなかった。GSK 社はビックデータ分析の環境を整えるためにデータ利活用に関する技術や知見を集約した CoE（Center of Excellence）を構築し、データ統合やストリーミングと合わせてデータプリパレーションツールの Trifacta[7]を導入して科学者や研究者自身でもデータにアクセスできる環境を整えた。科学者や研究者はデータプリパレーションツールを使って IT 部門に頼らずに自身で必要なデータを入手できるようになったことから、通常 6 ヶ月かかる新薬開発のプロセスを 2 週間に短縮することができた。

データプリパレーションツールを活用すると非技術者でもデータ整形ができるようになるが、データプリパレーションツールのみでは必要なデータへのアクセスを実現することはできない。非技術者がデータ準備に加わるにはデータプリパレーションツールと合わせて、必要となるデータが纏まって保管されるデータレイクを取り入れたデータマネジメントも今まで以上に重要となってくる。ビジネスの知見を有する事業部側がデータを観察することで、今まで使っていなかったデータを分析し、新たな洞察を得られる可能性もある。

データレイクは、もともとはデータの多様化・大容量化に応じた拡張性の高いデータストアであるが、データプリパレーションツールのユーザにとって、必要となるデータの保管場所が集約されてアクセス先が明確化しているという効果も発揮している。GSK 社のように、データプリパレーションツールの導入だけでなく、データやそれに関する技術と知見の集約を合わせることが重要となる。

---

[7] https://www.trifacta.com/customers/gsk/

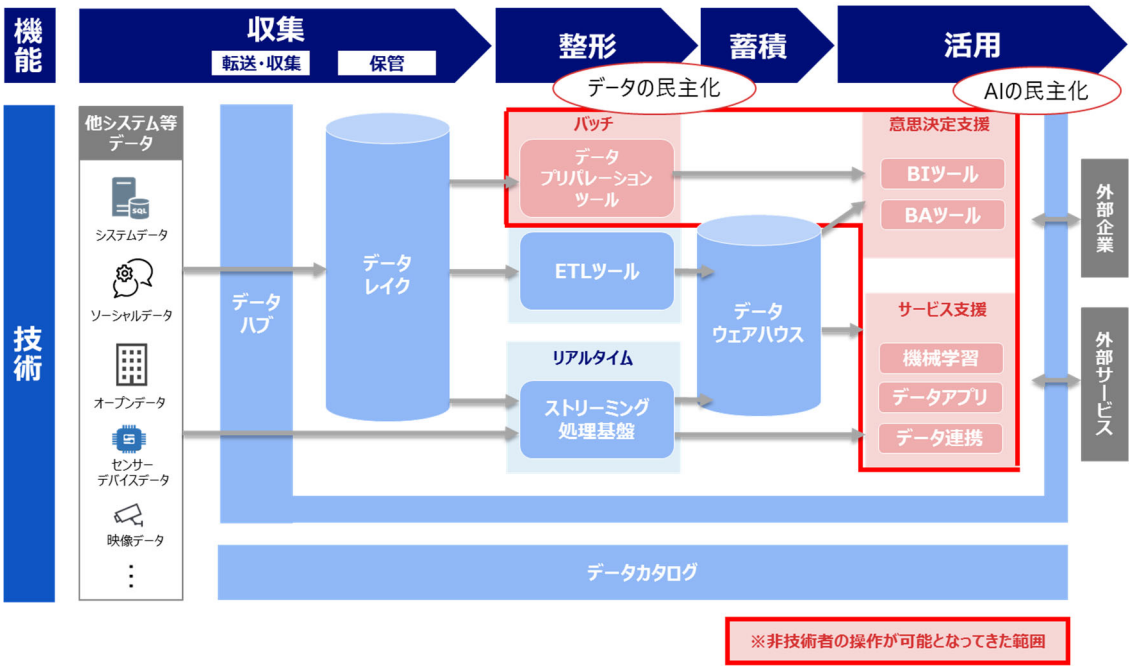
3. 「AI の民主化」から「データの民主化」

データ分析に基づいた意思決定がビジネスにおいて必要不可欠となっている昨今、非技術者がユーザとして使えるツールはデータプリパレーションツールだけではない。企業に存在するデータを活用して、特定の技能の有無によらずあらゆる従業員が意思決定し問題解決する手段に加え、効率化と生産性向上により企業全体の活動を大きく底上する「AI の民主化」には、もっと早くから様々な企業が対応している。例えば米国の Ford 社は 2018 年時点でマーケティング部門や生産部門などの事業部門に AI などの予測分析ツールを使いこなす市民データサイエンティストを全世界で 3,200 人以上擁している[8]。

データが発生してから分析できるようになるまでに「転送・収集、保管、整形、蓄積、活用」の各段階を経るが、その中で最もビジネスの現場に近い活用の段階は早くから非技術者向けの支援ツールが導入され「AI の民主化」が進んでいた。

現時点ではまだデータマネジメントの一連のプロセスのうち、非技術系事業部の従業員が処理を行える領域は限られているが、データプリパレーションツールの簡易化・自動化の潮流を受けて、非技術者が自身で処理できる領域は「活用」だけに留まらず「整形」へと広がり、「AI の民主化」から「データの民主化」へと拡大している（図表 2）。分析ツールやデータプリパレーションツールの提供会社は導入しやすく使いやすいソリューションを開発し、データサイエンティスト以外にビジネスユーザや非技術者を彼らのツールのユーザとして視野に入れはじめているのである。

図表 2　データマネジメントプロセスにおける「データの民主化」の拡大[9]



[8] The AI Summit New York 2018 でのフォード講演より

[9] 「DX 白書 2021」図表 42-6「データ活用基盤の全体像」をもとに作成

「データの民主化」は次第に企業全体に影響を及ぼしていく。そうなると、ビジネスとデータを今まで以上に密接に結びつけることができる。データの利活用に加わる従業員を広く増やし、あらゆる事業においてデータに基づいた意思決定を行って効率化や生産性向上を促進するデータドリブンな組織へと変革していくことができる。

例えば、英国の文房具グッズの e コマース企業 Papier 社は Facebook や Google など様々な媒体に広告を掲載しており、広告閲覧やクリックストリーム、トランザクションなど様々なデータを収集して顧客の 360 度分析を行っていた。Papier 社[10]はデータの抽出・変換・ロードを行う ETL 処理を自動化する Fivetran や BI ツールの Looker の導入により、CTO が一週間のうち丸一日にあたる時間をかけていた ETL の修正やデータベースの更新を効率化し、あらゆる従業員が常に最新のデータを使えるよう環境を整えた。その結果、今までは専門チームからの回答を待たなくてはならなかった分析を、従業員のうちの三分の二が自らツールを使って分析を行うようになり、データドリブンな組織へと変わっていった。

4. 「データの民主化」に日本企業はどのように対応していくべきか

データプリパレーションツールを筆頭に、データマネジメントにおけるツールの潮流は多様化・大容量化から、簡易化・自動化へと発展していっている。各種ツールにおいては非技術者をユーザと想定した設計が進んでいる。「データの民主化」はデータ利活用を推進していく企業が注視しておくべき潮流である。

「データの民主化」は新しい技術の導入や、それを使いこなす人材の育成だけでなく、データドリブンな文化や組織風土の醸成が重要となってくる。「データの民主化」はあらゆる従業員を対象にして企業全体に影響が及んでいくが、ボトムアップによるものではなく、企業全体を見通し組織的に作り上げるものである。最高データ責任者を任命したり、企業におけるデータ利活用の知見や技術を集約して CoE（Center of Excellence）などの推進チームを構成したりして、組織的に「データの民主化」を推進していく必要がある。

データ利活用のますますの効率化や、データドリブンな組織への変革を目指し「データの民主化」を推進する日本企業への推奨事項を以下のとおり三点にまとめる。

(1) データ利活用の技術や知見を集約する CoE を構築する

CoE は特定の領域やテーマに関する知見や技術を集約し、卓越（エクセレンス）した専門家によるチームの構成である。CoE はその高い専門性をもって、企業の課題解決における中心的役割を果たす。課題解決に取り組む各事業部にそのノウハウでもって支援したり、事業部同士の協調を調整したり、あるいは課題解決の実行部隊となる場合もある。

「データの民主化」という組織変革的な状態を目指していくうえで、CoE のような専門性の集約組織の構成

[10] https://www.fivetran.com/case-studies/case-study-papier

は重要である。データの CoE を構成することで、全社におけるデータとそのユースケースを把握し、各事業部における課題や要望を吸い上げ、どのように解決していくべきか組織横断的な方法を策定していくことができる。

(2) 事業部側の非技術者でも利用できるデータプリパレーションツールや AI・機械学習ツールを取り入れていく

「データの民主化」は、従来中核を担ってきたデータサイエンティストやデータエンジニア以外でもデータ利活用に積極的に関与していける点が新しい。貴重なデータに関する専門人材を創造的で変革的な価値創造に集中させ、データの整形や分析は各事業部がセルフサービスで実施できるように、非技術者でも利用できる処理が簡単で自動化されたデータプリパレーションツールや AI・機械学習ツールの導入が求められる。Python など整形に関するプログラミング技術を習得している人材でも、データプリパレーションツールを活用することで作業効率を上げることが期待できる。

データプリパレーションツールは製品によって簡易化・自動化できる機能は異なるため、自社のデータ利活用に合ったものを検討し、ツールの機能以上の複雑な整形や分析が必要となるときのためデータサイエンティストやデータエンジニアとの連携は継続しておくことが望ましい。

(3) データサイエンスに基づいた洞察ができる人材を増やすこと

データの利活用においてはプログラミングのスキルと、データサイエンスのスキルの両方が重要である。プログラミングのスキルが必要な作業をデータプリパレーションツールで支援するだけでなく、多くの従業員が統計学や数学モデリングの基礎を理解してデータサイエンスに基づいた洞察ができるようになると、データ利活用の効果を飛躍させることができる。

そのためには、統計学や数学モデリングに関して学べる研修などの機会を設けたり、そのようなスキルを既に身につけている専門性の高い人材と交流したり、スキル向上の施策が必要である。少なくとも、非技術者である従業員が、データ利活用でどのようなことができるのか、なぜ自分たちのビジネスにとってそれが重要であるのか理解と納得できる状態となれば、データドリブンな組織風土を全社的に醸成していけるだろう。

【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