

データ活用文化の醸成に向けて

—企業が意識すべきデータ管理の重要性—

分析自動化ツールの発達により、データ解析のハードルが下がっている。データ分析に関する高度なノウハウやスキルがなくてもデータ分析が行えるようになった一方で、データ管理が障壁となり、データ活用が進まない事例も多い。データ管理は、データ活用文化を企業内に根付かせるための重要な要素である。本稿では、その重要性と取り組み事例について解説する。

野村総合研究所 システムコンサルティング事業本部
システムデザインコンサルティング部 副主任システムコンサルタント

座 吾 実 希

専門はシステム化構想、アナリティクス、アジャイル開発



企業におけるデータ活用活発化の背景と課題

DataRobotやGoogle Analyticsのような分析自動化ツールや、Trifactaに代表されるデータの自動加工ツール（データプレパレーションツール）の発達により、高度な知識・ノウハウを有するデータサイエンティストだけではなく、データ分析を本職としない人がデータ分析に取り組むハードルが、以前と比べて低くなっているといわれている。こうしたことから、事業部門の担当者によるデータ活用への積極的な取り組みが広がっている。野村総合研究所（NRI）で支援するアナリティクスやBI（ビジネスインテリジェンス）ツールの導入検討プロジェクトにおいても、データサイエンティストではない普通の社員が日常的にデータ分析をできる環境を整え、現場でのデータ活用を促進していきたいという声が増えている。

一方、データ分析の専門家とデータ分析初心者の間には、各種分析自動化ツールなどを導入するだけでは補いきれない、データの管

理に関するデータリテラシーの差が存在する。特に、普通の社員が分析対象とするのは、データ形式が標準化しやすいWebマーケティングや在庫管理などの世界とは異なる、各社独自設計の社内システムが少なくない。そうしたシステムでは、一般的なツールでは対応しきれない、データ管理のこつが求められることがある。

一般的にデータリテラシーというと「データから意味を読み取り、分析するスキル」を指すことが多いが、データ管理に関するデータリテラシーは「分析結果の精度を担保するため、分析用データを適切に加工・管理するスキル」を示す。

本稿では、データ分析の初心者が直面しがちなデータ管理に関する3つの課題について述べた後、企業として取り得る対策とその注意点について説明する。

データ分析の初心者が直面する3つの課題

一般的に、現場を知る事業部門の担当者が

データ分析を実施する場合、業務に精通しているため、データサイエンティストに比較すると、データ定義の理解や分析結果の解釈には優れているといわれている。一方で、データ活用に向けて初心者が分析を行ってみると、さまざまな課題に直面する。以下に、データ分析の初心者が直面するデータ管理に関する課題について3つの代表的な例を紹介する。

(1) データ量が膨大でハンドリングできない

データ分析に慣れていないと、そもそもシステムからデータを取り出せなかったり、データが取り出せても、分析に使用する表計算ソフトや各種ツールに投入するデータ容量が大き過ぎるためにアプリケーションの処理限界を超えてしまい、その結果、分析できない状態に陥ったりすることがある。

この場合は、データの処理状況をモニタリングしながら、処理可能なデータ量になるよう加工する、もしくは分析に使用するツールを見直すといったエンジニアリングスキルが必要となる場面もある。しかしながら、データ分析の初心者である現場の担当者には、データ加工技術やツールに関する知見がないため、データ分析に至るまでの過程で挫折してしまうケースも少なくない。

(2) データ定義理解に時間がかかる

一般的に、データ分析の専門家よりも事業部門の担当者や現場スタッフの方が業務に精通しているため、データ理解に要する労力は少なく済むといわれている。しかしながら、データ定義の理解には、どのようなロジックでデータが収集されているのかというシステム設計の観点を要することも多く、システム開発とは無縁の事業部門の担当者が

データを分析することは困難といえる。そうになると、システム担当者と協力してデータの解釈を進める必要があるが、システム担当者からシステムを保守するベンダーへの問い合わせが必要となり、回答までのリードタイムが1、2週間かかることも多々ある。

(3) データクレンジング精度が低い

データサイエンティストがデータ分析に費やす時間の50～80%は、分析対象のデータから異常値や欠損を取り除くデータクレンジング作業が占めると、一般的にいわれている。データ分析初心者はデータクレンジングの重要性に対する認識不足から、クレンジングが中途半端になり、分析結果の質が下がってしまうことも少なくない。

データクレンジングについて顧客情報の分析での一例を挙げると、「株式会社〇〇」や「(株) 〇〇〇」というように、複数の表記で同じ会社が登録されていたり、表記の違いが原因で重複して登録されていたりする場合、分析を行う前にどちらかの表記に統一し、「名寄せ」と呼ばれるデータの重複を取り除く作業を行う必要がある。こうしたデータクレンジングを行わないと、同一の会社を複数の会社として集計してしまったり、表記の違いから別の会社として集計してしまったりして、正確な分析結果が得られないことになる。

効果的なデータ管理の仕組み

上述の課題に適切に対処し、データ活用を促進するには、データ管理のハードルを下げる必要がある。そのためには、組織として分析者の負荷を下げる仕組みを整備することが

重要である。前節で述べた3つの課題を理解し、うまく対処している企業の事例を紹介する。

(1) 分析用中間データの整備

卸売業A社では、分析者が自由にデータを取り出せるよう、週に1回、業務システムに蓄積したデータから分析に使用する一部のデータを別のデータベースに退避して、分析環境を整備している。また、Microsoft Excelのファイル形式やCSV形式、SQL形式など、取り出せるファイル形式を複数用意することで、作業者のスキルレベルに応じたデータ抽出・加工が可能になり、よりスピーディーにデータ分析に着手できる状態になったという。

こうした中間データは、一般的にデータウェアハウス（データを整理・格納しておく倉庫のようなイメージからこう呼ばれる）やデータマート（部門特化型のデータを格納しておく、データウェアハウスより小規模な商店のイメージからこう呼ばれる）と呼ばれるものに当たる。注意すべき点としては、データ活用の目的を明確にせずにデータウェアハウスを構築してしまうと、分析者のニーズを満たすことができず、結局活用されない事態に陥りがちなので、最初は分析目的別などのスモールスタートで、データマートを整備していくことをお勧めする。

(2) データ定義が分散管理されない仕組みの整備

各事業部で管理しているシステムのデータは、システムごとのデータ定義は個別部署で管理している場合が大半である。顧客コードや商品コードなどキーとなる情報の形式が異

なり、お互いにひも付けされていないことも多い。全社でデータ共有されているとしても、項目名やデータ形式を示したテーブル定義書を社内に公開する程度で、その業務的背景や他システムとの関係については、他部門から見るとブラックボックス状態であることが多い。

サービス業B社では、情報共有を事業部門横断のデータ定義のナレッジ共有プラットフォームで標準化している。データの生まれる業務的背景や分析利用時の注意点といったメタデータを管理することで、分析者・データ管理担当者双方の問い合わせ負担を減らし、よりスピーディーなデータ分析を実現している。データ分析者がデータ集計時に気を付けるべき点が共有されており、次に同じ分析をする人が同じミスを繰り返さないといった効果が期待できる。

(3) データクレンジングの効率化およびデータ品質の向上

データクレンジングの経験のない人が、データの重複や欠損を全て手作業で取り除く作業は、非常に非効率的である。分析者が全て手作業で整備しなくても、データプレパレーションツール（データ変換や整形、統合、名寄せといったデータクレンジング作業を自動で行う）を活用するといった方法があるが、コストが高かったりパラメータ設定が複雑だったり、個人レベルでの導入障壁が高い。一方で、分析者のデータクレンジング負荷を減らすため、マスターデータのレベルでデータクレンジングを行う方法を製造業C社が取り入れている。

分析対象となるデータには、業務上で基本

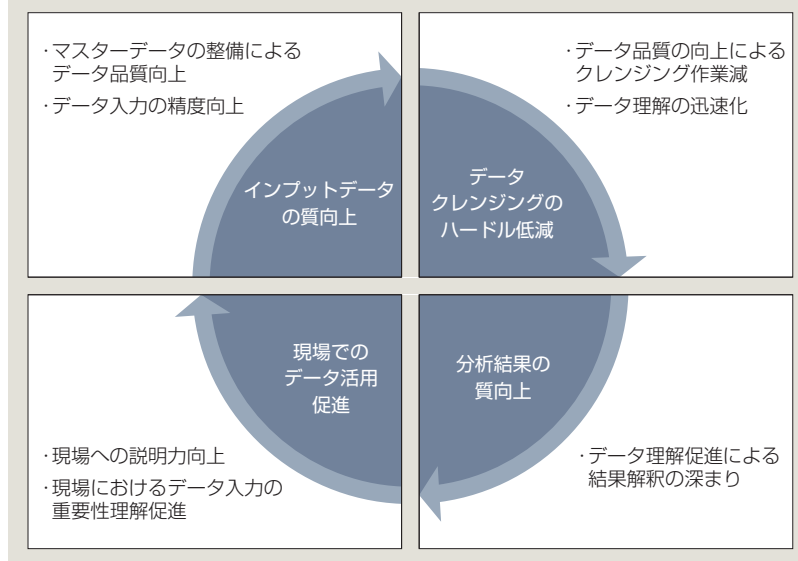
情報となるマスターデータと、業務に伴って発生する出来事の詳細を記録したトランザクションデータの2種類が存在する。トランザクションデータはマスターデータを基点に発生するため、データが発生する前段階で名寄せやクレンジングを行うマスターデータの整備に力を入れる意義は大きい。これはシステム間のデータ不整合を減らし、分析者のデータクレンジングの手間を減らすことで、解析困難なデータの発生を防ぐ観点からも有効である。

製造業などで、材料の仕入れから製品の完成、販売まで、一連のバリューチェーンがシステムで管理されている場合、本来は全てのマスターデータを標準化することが理想的である。しかしコストや期間を考慮した場合、最初は各システムやプロセスごとに見直し、徐々に広げていくというアプローチでもよいだろう。

データ管理の強化で得られるメリット

すでに述べたとおり、データプレパレーションツールやデータ分析自動化ツールの実用化が進み、データ分析に必要なスキル・ノウハウ取得の敷居は低くなっている。しかし、データ分析を現場での活用へとつなげるには、データ管理においてこれまでに述べた3つを行うことでガバナンスを強化し、データ加工やデータ定義のこつこの属人化を排除す

図1 データ管理強化によって得られる好循環サイクル



ることが重要である。そのようにして、専門的な知識を持たない人でもデータ活用に取り組める環境を整えることが、データ活用文化を醸成するための重要成功要因であると考えられる。

また、NRIで支援するアナリティクス関連プロジェクトにおいては、データ管理の強化でデータ活用の文化が現場に根付くと、その結果として、システムに入力するデータの重要性が認知され、入力されるデータの質も向上するという好循環が見られる傾向がある(図1参照)。システム開発面でも、データのメタ情報が共有されたりマスターデータが標準化されることで他システムとのデータ関係が容易になったりするなど、副次的効果も多い。

すぐに対策を打ちにくい課題もあるかもしれない。しかし、データ活用を進めたいのであれば、まずはスモールスタートで、データ管理環境整備に取り組んでみてはどうか。