データ分析の精度と柔軟性を確保する
―クラウド型データ分析基盤のすすめ―

機械学習やディープラーニングなど、大量のデータに基づいた分析に注目が集まっているが、このような分析を業務に取り入れる場合に欠かせないのがデータの運用という視点である。そこで本稿では、データ分析の精度と柔軟性を確保するためのデータ運用をシステムとしてどう考えるべきか解説する。

NRIネットコム デジタルインテグレーション事業本部
クラウド事業推進部 アーキテクチャデザイン課長代理
喜早（きそう）　彬（あきら）
専門はクラウドサービスを生かしたシステムの設計・開発

欠かせないデータ運用の視点

世はまさにデータ分析の時代である。生活者は商品・サービスについてインターネットで調べ、気に入ればそこから購入するし、スマートフォンにダウンロードしたアプリを使って購入することも普通になった。また、オンライン上のデータにとどまらず、機械の稼働状況や人の活動データを収集することも容易となった。これを背景に、企業は、生活者に自社の商品・サービスを認知させ、購入への動機づけを行うために、さまざまなデータから価値ある知見を導き出す取り組みをこぞって進めている。

データ分析の取り組みが進むとともに、分析手法への注目も高まっている。統計学的手法や、機械学習・ディープラーニングといったアルゴリズムを使った手法である。こうした手法の進化がデータ分析の価値を高めることは言うまでもないが、忘れてならないのは、データ分析は1回実行すればよいものではなく、継続的に日々の業務へ価値をもたらさなければならないということだ。そのためには、データ分析の精度を保ち続けることが重要である。

データ分析の精度に影響を及ぼすポイントは、「データの鮮度」と「データを使う際の柔軟性」である。データは“生き物”であり、収集されるデータは日々変わっていく。データが新しいほど、分析対象の今の姿を的確に捉えることができる。データの新しさに加えて、分析の前処理となるデータの加工が重要となる。加工とは、収集したデータを分析に適した形にしたり、データを集計したりするなど、扱いやすい形に変形させることである。このように、データを新しく保ったり、データを柔軟に加工して価値のあるものにしたりすることを、ここではデータ運用と呼ぶ。データ運用は、データ分析作業の8割ほどを占めると言う人もいるくらい、非常に重要である。

鍵となるデータ分析基盤

的確なデータ運用を可能にする仕組みを、ここではデータ分析基盤（以下、分析基盤）

と呼び、必要な機能や構築方法について解説する（図１参照）。

(1) 分析基盤の構成

分析基盤は大きく分けて３つの層で構成される。

①ETL層

ETLとはExtract（抽出）、Transform（変換）、Load（書き出し）の略であり、分析したいデータをファイル形式などに抽出し、データが欠けている部分に対する穴埋め処理や、データ構造の変更など、データに対する編集処理を必要に応じて行い、次のデータレイク層に書き出すことを意味する。

②データレイク層

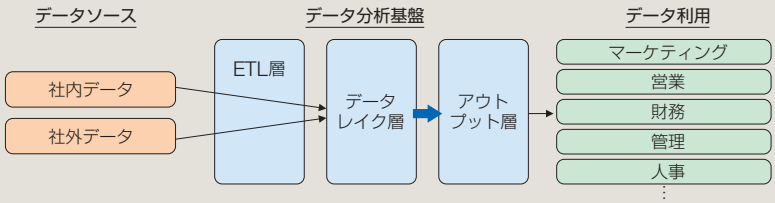
データレイクとは文字通りデータの湖を意味し、ETL層で処理されたデータが集積される場所となる。対象となるすべてのデータを集約することで、1つのシステムにとどまらない横断的なデータ分析を行うための拠点とすることができる。

③アウトプット層

データレイク層に集約されたデータを、要件に応じて使いやすい形に整形するのがこのアウトプット層である。例えば、データレイク層に保存されているデータから、分析に使用したいデータだけをデータウェアハウスのデータベースへ投入する、といった使い方ができる。

分析基盤を以上の３層構成とすることにより、データを使用する際の柔軟性を確保できる。また、ETLの処理の速さと、アウトプット層での適切なサービスの利用により、素早い運用を行うことができるため、データの鮮度も保たれる。

図１　データ分析の流れとデータ分析基盤の構成



(2) クラウドサービスの活用

分析基盤はクラウドサービスとの相性が良い。理由は以下の３点である。

①データ容量に制限がない

分析基盤は、他の一般的なシステムに比べて桁違いのデータ量を扱うが、クラウドサービスは事実上、容量の上限がないため、将来的にデータサイズが増大しても対応できる。ただし、クラウドサービスのストレージは従量課金制であることが多い。これは、使用したデータ量のみのコストで済むことを意味するが、逆に言えばデータ量が増えた分だけコストが増えていくことになるので、そのバランスを考える必要はある。

②データ利用サービスが豊富

クラウドサービスでは、システム構築に必要なものがパーツのような形で多数提供されている。例えば、データウェアハウスに適した高速な検索を行えるデータベースや、機械学習を簡単に利用できるAPI（プログラムやデータを他のプログラムから利用するためのインターフェース）などがある。また、大量のデータをバッチ処理で分析する用途には、分散処理のサービスが提供されている。これらは、いずれもWebブラウザーから簡単に行えるようになっている。これらのサービスを組み合わせて使うことで、基盤構築に必要だった時間を大幅に削減でき、データを活用

した価値創出という本来の目的にリソースを集中できるようになる。

③柔軟なリソース確保が可能

分析基盤に求められる性能は、基幹系や情報系といった通常のシステムとは異なる。通常のシステムであれば、常時ある程度の負荷がかかる想定でシステムの構成を検討するが、分析基盤の場合は、データ分析を行う短い時間に大量の負荷がかかる（リソースが必要になる）。しかも、この負荷はユーザーが分析を行うタイミングで発生するため、いつリソースが必要になるかを想定することが難しい。このような不明確な要件に対しても、クラウドサービスを使ってシステムを構築しておけば、負荷に応じて自動的にリソースの増減が行える。いつ使うかもわからない大量のリソースを、前もってずっと保有し続けておく必要はない。

分析基盤構築の重要ポイント

分析基盤を構築する際には、以下の３点が重要なポイントになる。

①データの活用方法を具体化・明確化する

前述の、分析基盤のアウトプット層に当たる部分の話になるが、「誰が」「どのような目的で」「何を求めるか」によって、分析基盤に必要なものが変わる。例えば、「発注経験の浅い店員が」「店舗での精度の高い発注作業のために」「機械学習を使った発注量の提案をしてほしい」というような具合だ。このような明確化によって、利用すべきデータを具体的に検討することができるようになり、そのデータの量や種別に応じて最適なシステム構成が決まる。このように、業務にどう生かしたいかを踏まえながら、これらの項目を明確化することが大切である。

②データの選定を慎重に行う

分析基盤では、個人情報のように取り扱いに注意を要するデータを対象とすることもある。この場合、各事業者で定めたセキュリティポリシーに応じた検討が必要になるが、基本的にはデータをそのまま分析基盤に投入することは避け、何らかの加工を行うことが望ましい。例えば、生の個人情報の代わりに、個人を特定するIDを使う。もし個人情報の復元が必要であれば、取り込んだIDをキーとして、データ元へ問い合わせて個人情報を取得するというのも１つの方法である。この問題はアウトプット層の構成にも大きく影響するので、設計段階で慎重に方針を検討する必要がある。

③データの利用環境を限定する

これは、主にBI（ビジネスインテリジェンス）ツールや自前の分析プログラムを使う場合に留意する点だ。これらのツールやプログラムは、各個人のPC上で操作することが可能である。すなわち、データを各個人のPCに持ち出して作業できるということである。個人の端末上に保存されたデータは、一般にセキュリティの統制が及ばない。そのため、セキュリティが確保されたサーバー上にBIツールや分析プログラムを実行できる環境を用意し、各個人のPCからそのサーバーに接続して使用するようにすることが望ましい。この場合、分析環境には、ユーザーが十分に快適と感じられるくらいの性能が必要である。自分のPCの方が高性能だと、ルール

違反と知りながらもそちらへデータを保存して分析してしまう可能性があるからだ。

以上の3点を検討する際には、将来起こり得ることを想定したり、分析手法の専門知識が必要となったりするため、多くのプロジェクトを経験しているベンダーに相談しながら進めるのがよいだろう。

分析基盤がもたらす効果

分析基盤は、全社横断的なデータのハブでもある。各組織が持つデータを分析して結果を共有することで、以下のような効果が生まれるだろう。

(1) 既存業務の効率化

分析基盤の上に保持された大量のデータを利用することで、経験を積んだベテランのナレッジを全社の資産とすることができるようになる。

例えば、NTTドコモが2017年2月に発表した「AIタクシー」がある。これは、携帯電話の位置情報から割り出したエリアごとの人の数や、気象情報、地理情報などを基に、乗車を希望している人が何人いるかを区域ごとに予測して、ドライバーに配信するシステムである。予測には、過去の乗車実績を学習させた人工知能（AI）が用いられる。空車走行を減らせる効果はもちろん、経験の浅いドライバーが乗客を効率よく獲得できるようになるといった効果もある。

このように、大量のデータを分析する仕組みは、人の仕事を助け、一部の仕事を代替する。その結果、1人が生み出す価値を増大させたり、人にしかできない仕事にリソースを重点的に投入したりすることが容易になる。

(2) 新しいビジネスチャンスの発見

複数のブランドを持つ企業が、従来はブランドごとにユーザーの行動分析を行っていたとする。これらの情報を1カ所に集積し分析することで、思いもよらない発見があるかもしれない。分析基盤に保存すべきデータは、システム上のデータに限らない。例えば、対面販売のようなオフライン環境での売上高と、Webページ上のユーザーの行動履歴を結び付けることも可能である。これにより、ユーザーがどういう経路で商品を認知し、どんな関心を持ち、購入に至ったのかということを、オフラインとオンラインをまたいで追跡できるようになる。それは、新しいビジネスチャンスの発見につながるはずだ。

データを基に別の業態へ進出した企業もある。眼鏡の製造販売で知られるジンズが開発・販売している製品に、眼鏡型ウェアラブルデバイス「JINS MEME」がある。これは、眼球の動きから使用者の集中の度合いを計測するものである。同社は、この製品で取得したデータを基に、「世界一集中できる環境」をうたった会員制のワークスペース「Think Lab」を2017年12月にオープンさせた。データ分析を新しいビジネスにつなげた例といえる。

デジタルトランスフォーメーションという大きな変革の時代、さまざまな活動を記録したデジタルデータを分析し、そこから有用な知見を導き出す分析基盤は、企業にとって変革の時代を生き抜くための必須の備えとなるだろう。■