HITACHI
Inspire the Next

ビッグデータ時代のデータ分析で、成功・失敗をわけるものは何か。

株式会社 日立製作所　サービスプラットフォーム事業本部

データの下準備を容易にする
データパイプラインのススメ

急速なデジタル化が進み、競合他社との競争が激化する現代では、いかにビッグデータを利活用するかがビジネス成功の成否を握る。これまでの既存データをはじめ、SNSやセンサーデータなど多種多様なデータの集約が可能となったいま、データ分析の結果を新ビジネスの創生や業務の改善にいかしていくことが企業競争力の根源となる。

ところが、いざデータ分析を始めてみたものの、うまくいかなかったという声を多く耳にする。一体、何が原因なのか。どうすれば、データ分析は成功するのだろうか。

その答えが、**データパイプライン**という構想である。

この記事では、データ分析がうまくいかない原因と、それを解消しデータ分析を有用なものにするデータパイプラインについて解説していく。「データ分析　成功事例」という検索ワードではたどり着かなかった「データ分析を成功へと導くヒント」が見つかるはずである。

1 データ分析の前提となるデータの下準備とは

データ分析がうまくいかないという人たちの話を俯瞰（ふかん）してみると、データ分析の前の準備段階でつまずいているケースが多い。

「いろいろな軸で分析したいけど、社内のデータが整理されていないので、分析に必要なデータがうまく集められない。」「分析したいデータの形式がバラバラなので、変換して共通の形式にそろえるのに非常に手間がかかる。」という声が多いからだ。

データ分析の前提として、非常に重要なファクターとなるのが**データの下準備**である。

下準備とは、「**データを収集する**」「**データを変換・統合する**」「**いろいろなデータをブレンディング（掛け合わせ）する**」という作業のことで、これを経て分析に有用なデータが作られる。こうした分析用のデータが準備できていなければ、いくら分析ツールや機械学習、人工知能といった仕組みが整備されていても、有益な分析をするのは難しい。しかし、データの下準備が十分にできていない企業が非常に多いという現実がある。

なぜ失敗するのか

分析がうまくいかない原因は、データの下準備ができていないことにある

つまり、データの下準備ができていない企業は、「データを収集する」「データを変換・統合する」「いろいろなデータをブレンディングする」のどこかでつまずいてしまっているのである。

まずは、データが収集できないケース。

これは、主にビッグデータを管理する仕組みが整っていない場合が多い。音声や映像などの非構造化データや、センサーデータ、ログデータといった半構造化データを格納するのは簡単ではないため、その環境自体が整備されていないことが多いのだ。企業内にはあらゆるビッグデータが存在するのに、有効活用されずにデータが破棄されてしまっているのである。

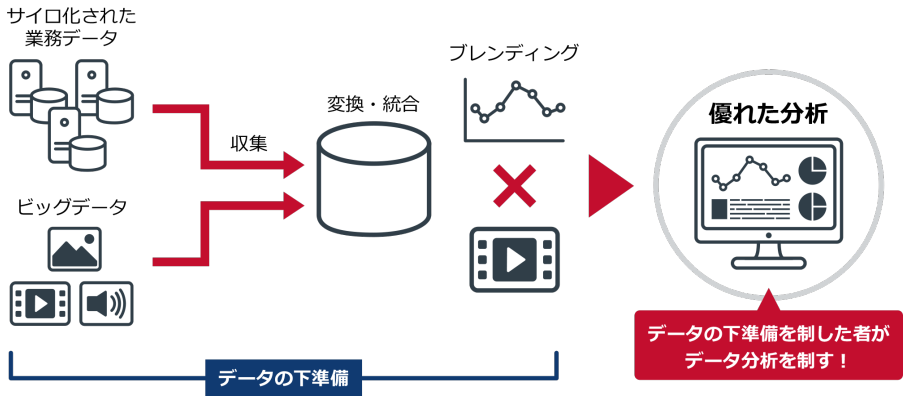
次に、データを変換・統合できないケース。

これは、データのサイロ化が関係している。長い年月、企業のシステムは部門ごとに個別最適で構築されてきた。それぞれのシステムは分断・孤立していて、扱っているデータの形式や定義も異なるため、統合する前の変換作業が困難になっている。また、企業に散在するデータソースに対して、どこに何のデータがあるのか把握できていないという問題もあり、統合するための設計も非常に困難になっている現実がある。

最後に、データのブレンディングができないケース。

これは、構造化データと非構造化データ・半構造化データをブレンドする仕組みがわからない、手作業でブレンドするので時間がかかり過ぎる、そもそもビッグデータが格納できていないのでブレンドできない、といったことが理由になっている。

こうした課題を抱えているため、データの下準備がうまくできないのだ。しかし、データ分析を成功させている企業とは、データの下準備を制している企業である。ここで立ち止まっていては、優れた分析はできない。つまり、市場競争で優位性を得るチャンスを失う、ということを忘れてはいけない。



図：データの下準備を制した者がデータ分析を制す

2 データ分析とは反復プロセスである

データの下準備ができたら、ようやく分析が始まる。 ここで、ひとつ重要なポイントがある。データ分析とは、一度やったら終了ではなく、繰り返し行うものだということである。「別の観点で分析したいから新しいデータを提供して」といったことが、**繰り返し起こるのがデータ分析であり、そのたびに新たなデータに対する下準備が必要となる。**つまり、「データ分析」と「データの下準備」は反復プロセスになっている。しかし、分析軸を変えるたびに下準備を手作業でやっていては、あまりに手間がかかるし、下準備用のプログラムをそのつど開発するのは多大な工数がかかるであろう。こういった背景から反復プロセスが継続できず、データ分析がうまくいかない原因のひとつとなっている。

なぜ失敗するのか

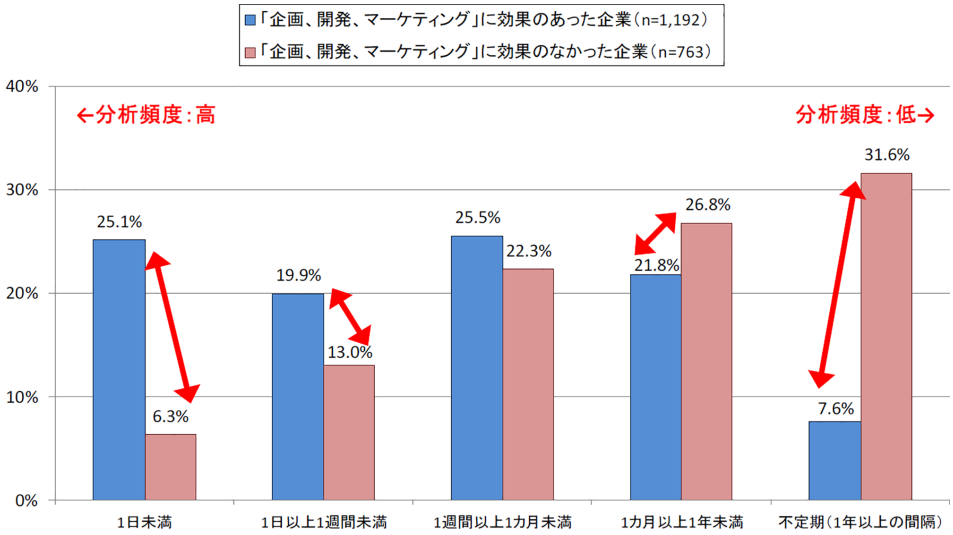
分析がうまくできない原因は、反復プロセスが継続できないことにある

データの下準備や分析をデータサイエンティストに頼んでいる企業の場合、欲しい分析結果を現場がもらうまでに時間がかかることがよくある。データサイエンティストは他の業務と掛け持ちで作業しているケースが多いため、タイムリーに分析結果を手に入れるのは難しいといった実態があるのだ。そうなると、情報の鮮度が落ちてしまうという弊害も発生する。

更新が頻繁に繰り返されるビッグデータは、すぐに情報が古くなってしまうため分析に時間をかけていられない。従来のデータ活用と比べて、情報に求められる鮮度が異なるのだ。

つまり、**高い頻度で分析してこそ価値があるのが、ビッグデータの利活用**なのである。

総務省「ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究」（平成27年）によると、ビッグデータの分析の頻度が1日未満（1日に1回以上分析をしている）という企業は、高い割合で分析の効果を得ている。全体を見ても、効果を得ている企業の過半数は、1か月未満の頻度で分析を行っている。



図の出典：総務省「ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究」図表4-21

この結果からもわかるように、高い頻度で分析を行うことがビッグデータ利活用においては重要である。近年では、セルフサービスBIの普及に見られるように、現場や各部門で目的に沿ってデータ分析をして、ビジネスの変化に柔軟に対応したいというニーズが増えている。こうした動きに乗り遅れないよう、**現場に迅速に分析用のデータを供給できる環境を整備し、データの下準備とデータ分析を柔軟に、かつ高い頻度で反復できる仕組みを整える**ことが、非常に重要になってきている。

3 データ分析を成功へと導くデータパイプライン

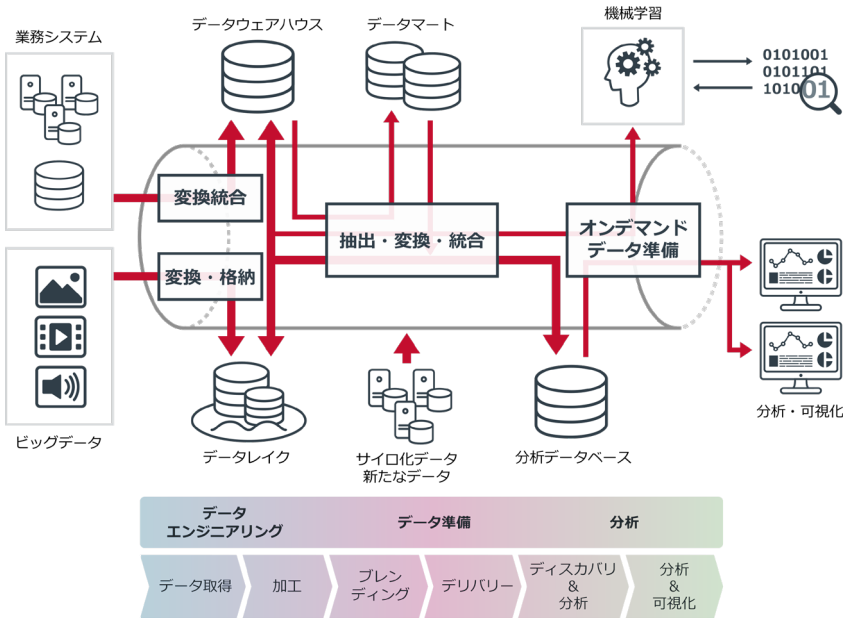
データ分析がうまくいかない原因は、「データの下準備ができていない」ことと、「反復プロセスを継続できない」ことだとわかった。では、これらを解決してくれる仕組みはないのだろうか。

それが、**データパイプラインという構想**である。

企業内に散在するシステムは、扱うデータこそ異なるものの、データの下準備では同じような作業プロセスを経ている。

そこに注目し、データの収集、変換・統合、ブレンドといった共通作業を一本化した構想がデータパイプラインである。この最大の特長は、**データ分析における「手作業」を減らせること。最も手間がかかっているデータの下準備を効率化することで、データ利活用が促進できる**のだ。

図にすると、次のようになる。



図：データパイプライン

データパイプラインで、各工程は次のように改善される。

データエンジニアリング

サイロ化されていた業務システムのデータは、容易にデータ形式を変換できるようになる。すべてのデータが簡単に統合できるようになるため、手作業の手間は劇的に軽減。また、これまでのようにデータソースが把握できない問題も解消され、必要なデータはすべて取り出せるようになる。ビッグデータを管理する環境も用意されているため、さまざまな非構造化データ・半構造化データも簡単に格納できるようになる。

データ準備

既存の業務データとビッグデータを抽出して、これらのデータをブレンディングすることが容易になる。これまでのように、限られたデータソースから分析用データを準備するのではなく、複数のデータソースによるブレンディングが可能となるので、分析の幅が拡大。より有用な洞察へとつながり、ビジネスにおける競争優位の獲得にいかすことができるようになる。構築後に、新たにサイロ化されたデータが発生しても、簡単に変換・統合できるようになるため、増え続けていくデータの管理を心配する必要もなくなる。

分析

分析用のデータが容易に作成できるようになることで、データサイエンティストからのデータ要求や、多様化する分析ニーズにも迅速に対応できるようになる。その結果、分析現場へ迅速に分析用データを供給することが可能となる。

大量データの交通整理をしてくれるデータパイプライン

データパイプラインの優れた点は、これらの工程をあたかもパイプラインのように**シームレスにつなぎ、データ発生源から分析現場までのデータフローを全社で共有**できることにある。これまでの運用のように、工程ごとに別々のツールを導入したり、手作業でのデータ共有に苦労したりすることがなくなり、全工程をスムーズに効率的に実施できるようになるのだ。

POINT

データ収集から分析までの全プロセスをシームレスにつなぎ、
増え続ける多様なデータを分析現場へ迅速に供給できる「データパイプライン」

データパイプラインとは、例えるなら大量データの交通整理である。このデータパイプラインにデータを要求すれば、大量かつ絶え間なく流れているデータを交通整理してくれ、タイムリーに欲しいデータを取得でき、「データ分析」と「データの下準備」という反復プロセスが可能になるのである。まさに、データ分析に必要不可欠な基盤ではなかろうか。

優れた分析を可能にしてくれる、データパイプライン。これを部門最適にとどまらず、全社横断で構築することで、データ分析はより価値あるものとなり、その分析結果を新ビジネスの創生や業務の改善にいかすことで、競争優位性の獲得が期待できるであろう。

データパイプラインを構築するには？

データパイプラインは、優れたデータ分析を可能とする理想的な構想である。しかし、見よう見まねで構築しても決して成功はしない。次の3つの条件を満たしたデータパイプラインを構築することが、データ分析の成功へとつながる。

1. ビジネスのニーズに応じて進化できるような、柔軟性が高いものであること
2. データエンジニアリング、データ準備、分析の3工程が人やプロセス、ツールなどの仕組みによって分断されずに、密に連係していること
3. 「反復」が同じことの繰り返しではなく、新たなニーズに対応しながら拡張・進化を繰り返せること

ホワイトペーパー「8項目の必須チェックリスト」は、データパイプラインの本格的な構築にあたって、データ接続やデータ準備、そのほかデータパイプラインの効率的な管理などに関する注意点や、考慮すべき点を詳細にまとめた情報となっている。

ホワイトペーパーのダウンロードはこちら
https://www.hitachi.co.jp/products/it/bigdata/platform/pentaho/download/index.html#wp3274

製品に関する詳細・お問い合わせは、営業担当員または下記へ

■製品情報サイト
https://www.hitachi.co.jp/pentaho/index.html

■インターネットでのお問い合わせ
https://www.hitachi.co.jp/products/it/bigdata/platform/pentaho/inquiry/index.html

株式会社 日立製作所　サービスプラットフォーム事業本部

2021.2