

企業情報システムとデータの活用範囲の拡大

Corporate Information System and Expanded Practical Use of Data

山 崎 慎 一

要 約 企業のあらゆる業務にコンピュータが活用されている。企業が自社の業務プロセスから生まれるデータを蓄積して、ビジネス課題の解決やビジネス戦略に活用するためのデータウェアハウスを構築するのは普通になり、経営を支援するためのビジネス・インテリジェンスという用語は一般的になっている。データ活用を企業の重要戦略活動とするイノベーターは、従来からのデータ分析に加えて、これまで対象としてこなかったビッグデータを活用し、新たなビジネスに利用するようになってきている。多くの企業がこれに続くようにビッグデータの活用に注目し、利用しようとする大きな動きがある。ビッグデータの活用がイノベーターからマジョリティに移ろうとしている。

本稿では、企業のデータ分析の必要性、分析対象のデータ、分析の技術などを明らかにする。さらに、分析を必要とする企業に対してITサービス企業が提供すべきサービスを明らかにする。

Abstract Almost every enterprise uses computer systems for their business processes. It has been commonly accepted that enterprises build the data warehouses to store data derived from own business information systems for solving business matters and using the business strategies. The term “business intelligence” has been a generality applicable. In addition to conventional data analytics, innovators defining data analytics for their important strategic activities have begun to use “big data” left out of analytical subject of enterprises for a long time, and applied it in new business fields. There are big waves where many enterprises take notice to use big data and apply big data in their business. The waves of big data are moving from innovators to majorities of enterprises.

In this paper, we show the necessity of data analytics, objects of data analytics, and technologies of analytics in the enterprise. Furthermore we show several services that IT service companies have to offer to enterprises that need data analytics.

1. はじめに

企業のあらゆる業務にコンピュータが活用されており、業務プロセスからは多くのデータが発生し、血液のように社内の各業務の間を流れ、必要な情報のやりとりでビジネスが成り立っている。自社の業務プロセスから生まれるデータを様々な課題の解決に利用しようとするれば、必然的に日々のデータを収集、蓄積し利用に備えなければならない。1990年代初めに提唱されたデータウェアハウスの概念^[1]は、広く企業に理解され、今や多くの企業では、データウェアハウスあるいは目的別に特化したデータマート^[2]を構築し、ビジネスの意思決定に活用している。経営戦略を支える知識を生み出す仕組みや活動は、ビジネス・インテリジェンスと呼ばれ、スピード経営^[3]が求められる現代の企業活動に必須になっている。データ活用を企業の最重要活動に位置付けているイノベーターは、データ分析に力をいれてきたが、さらに対象デー

タを拡大して、これまでその取り扱いがやかいかだったビッグデータの活用を始めている。国内では、2012年はビッグデータ活用元年とも言われており、多くの企業が注目している。ビッグデータの活用という大きな動きが企業を取り巻きつつある。

本稿では、2章で企業のデータ活用の現状と、新たな対象データについて概観を述べる。3章では、データウェアハウスの改善とビッグデータの処理手法について、4章では、分析データの収集とマーケティングについて、5章では、分析の深さについて、6章では、ITサービス企業が顧客企業に提供すべきものについて、それぞれ説明する。最後に、日本ユニシスの立場を明らかにする。

2. データ活用

データ分析の時代である^{[4][5]}。企業の基幹系システムと呼ばれるトランザクション処理が、業務プロセスのコスト削減を主体とするのに対して、情報系システムであるデータウェアハウス中心の処理は、業務分析や新たなビジネス戦略へのデータ活用である。企業で発生するデータはさまざまだが、大きく次の2種に分類できる。

① 構造化データ

構造化データは、業務プロセスで発生するデータで、業務システム内にある。商品売上の個数や金額などの企業決算に連なるデータがその典型で、主体となるデータ型は数値型である。これらのデータは、ビジネスの要求に従い事前にその属性やデータ型が決められ正規化される。業務システムでは、ビジネス要求を満たすためにデータモデルが設計され実装される。実装にはリレーショナル・データベースが使用され、データのあいまい性は排除される。分析に使われるデータは通常、構造化データである。

② 非構造化データ

非構造化データは、構造化データと異なり、メールや社内文書のテキストデータのような非定型データである。システムログやWebアクセス・ログも含まれる。これらのデータは正規化されておらず、あいまいな表現のテキストであったり、時にはバイナリ・データであったりする。そのためデータベースでの取り扱いは面倒で効率も悪く、分析にはほとんど利用されてこなかった。

これまでのデータウェアハウスには、構造化データが蓄積されてきた。業務プロセスで発生するデータは正規化されているが故に、データウェアハウスのデータも、その目的に従って統合され、時系列に蓄えられて使用される。構造化データをデータウェアハウスに蓄積し利用するためのデータ・モデリングは確立し、使用されるソフトウェアは一般化してきた。データウェアハウスを構築するためのETL（Extract：抽出，Transform：変換，Load：ロード）ツールや、利用するための分析ツールは数多く市販されている。データウェアハウスの概観を図1に示す。

2011年に話題となったビッグデータは、構造化データと非構造化データが組み合わされたデータで、これまでのデータに比べ、Volume（大量）、Variety（多様）、Velocity（高頻度）という3Vの特徴があるとされている（図2）。ビッグデータの分析は、いままでの構造化データ中心に行われてきた分析とは異なる三つの特徴を持つ未開拓のデータを対象とするため、これまで処理するのが困難であったが、次章以降で述べる新しい技術で分析可能となってきた。

ビッグデータの解釈はさまざま存在するが、本稿では「企業のビジネス戦略に活用するためのインターネット関連の非構造化データ」とする。

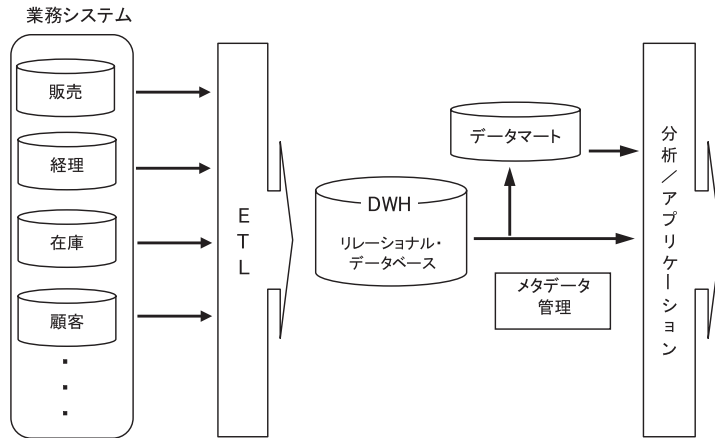


図1 データウェアハウス概要

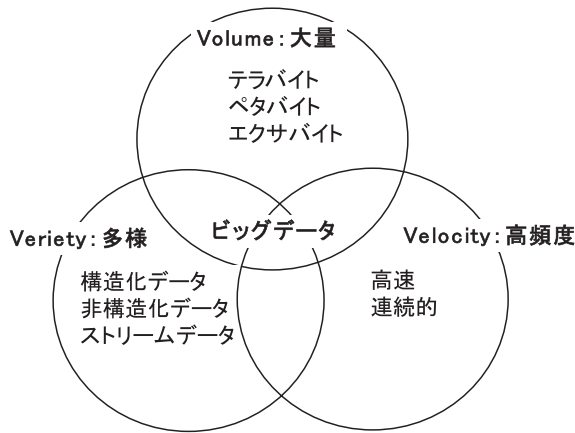


図2 ビッグデータ概念

3. データウェアハウスの改善とビッグデータの処理手法

これまでの分析対象だった構造化データも企業ビジネスの拡大あるいは企業合併などで大量になっており、データウェアハウスの効率向上は必須になっている。増大する構造化データを量に応じてスケーラブルに処理可能にするために、新しいデータベースやデータ処理技術が出現した^[6]。これは、次の五つの要因によって実現されている。

- ① ハードウェアの性能向上： CPU 性能，メモリ容量，ネットワーク性能
- ② 並列処理： 超並列処理 (MPP: Massively Parallel Processing)，分散処理
- ③ インメモリ処理： メモリ容量拡大に伴う処理
- ④ 最適化処理： 新アルゴリズム (メモリ有効利用，データベース機能拡大)
- ⑤ I/O 削減： データ圧縮による I/O 容量削減，カラム処理による I/O 容量削減

2002年に米国のベンチャー企業から超並列処理をソフトウェア、ハードウェアに採用したDWHアライアンスが登場した。2012年現在では、主要なデータベース・ベンダーの全てが、DWHアライアンスを提供するようになった(表1)。

表1 DWHアライアンス

名称	提供ベンダー
IBM Netezza High Capacity Appliance	IBM
Oracle Exadata Database Machine	Oracle
Greenplum Data Computing Appliance	EMC
SQL Server Parallel Data Warehouse	Microsoft
HP Enterprise Data Warehouse Appliance	HP
Teradata Active Enterprise Data Warehouse	Teradata

従来からの技術を応用した新しいカラム指向データベースも参入してきた。インメモリ・データベースも進化し、ハードウェアでは大量メモリの搭載が可能になり、大量データの処理には選択肢が多くなった。非構造データの格納にキー・バリュー・ストアを使うこともある。既存データベースもデータ圧縮などの更なる機能改善が図られ、データベース市場は活発な状況にある。ただひとつのデータベースで全てのデータ処理に対応するのではなく、データの種類や処理の目的に適したデータベースを選択することが重要になった^{[7][8]}。

一方ビッグデータに対応したスケーラブルな分散処理技術としてオープンソースのHadoopが利用されるようになった。Hadoopは、MPPと同様シェアード・ナッシングな環境での分散処理である。複数のサーバにまたがるHDFS(Hadoop Distributed File System)にロードしたデータを複数のサーバにあるMapReduceエンジンが並列に処理する。簡略化すると、次のような処理である。

- ① 大量のデータを分解し、複数サーバにロードする
- ② それぞれのサーバにロードされたデータについて、個々に単純な数え上げなどのMap処理をする
- ③ サーバでの処理結果を合わせ、集計するReduce処理をする。

シンプルな並列処理で、これらを繰り返して実行することで、これまでリレーショナル・データベースでは困難であった大量の非構造化データの処理が可能になった。

Hadoopの処理は、機能を拡張して複雑な問い合わせ処理に対応してきたリレーショナル・データベースに対するアンチテーゼである。Hadoopは、単機能ゆえに、ペタバイト級のデータの集計処理が可能となった。Hadoopは、ビッグデータの特徴である不要な部分や重複部分などを排除するクレンジング処理に適しているため、ビッグデータをHadoopで目的に沿って処理し、その結果をリレーショナル・データベースのデータウェアハウスに格納して分析することで、高度な分析に使えるようになった。例えばWebアクセス・ログは、分解して不要部分を取り去り集計した後に、目的別にデータウェアハウスやデータマートに格納する。ソーシャル・データや文書は、言語解析や単語の分類、キーワード索引の作成後、分類してデータウェアハウスあるいはデータマートに格納する。

Hadoop が出現した 2006 年当時、既存リレーショナル・データベース処理と Hadoop は、相反するもののように捉えられていたが、2012 年現在は、データの特徴に応じて Hadoop とリレーショナル・データベースを使い分け、互いに不足する部分を補完するものと理解されている。データに応じた処理とデータウェアハウスの関係を図 3 に示し、処理の流れを①から④で解説する。

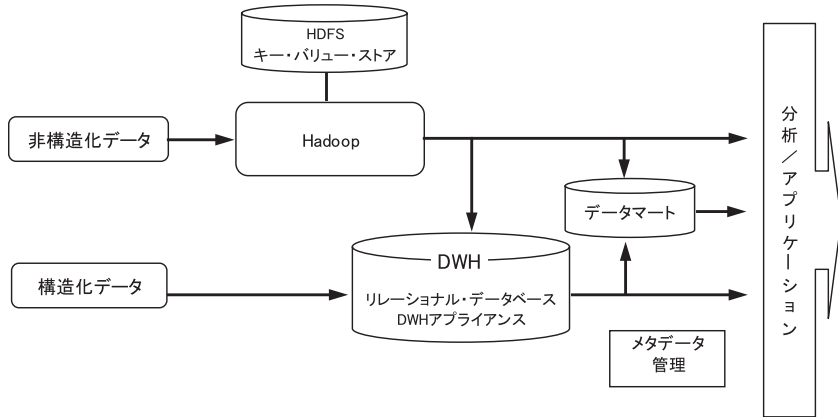


図 3 データウェアハウスと Hadoop

- ① 構造化データは、目的別にデータウェアハウスやデータマートに格納する。
- ② 非構造化データは、HDFS などの分散ストアに格納し Hadoop 処理をする。処理結果は、目的別にデータウェアハウスやデータマートに格納する。あるいは直接処理をする。
- ③ データウェアハウスのデータ処理、Hadoop 処理の過程で、互いのデータを利用する場合がある。このとき、データウェアハウスと Hadoop の間で、相互にデータのやり取りを可能にする。
- ④ メタデータ管理で、データウェアハウスあるいはデータマートに格納されているデータの源泉、データの関連、特性などを管理して、分析するデータの関連を分析者に提供する。

4. データ収集とマーケティング

これまで自社の商品販売データのように、自社内の業務プロセスで発生したデータと外部にある各種の公開データ、業界団体のデータ、あるいは調査会社から購入可能なデータを使って分析が行われてきた。

インターネットの拡大とともに、さらに多くの外部データが効率的に収集可能になり、分析に利用できるようになった。ソーシャル・ネットワークで個人から発信されるデータを収集し分析することで、個人のライフスタイルの変化や、企業の販売する商品に対する意見や批判などを製品に活かすことが可能になった。企業外データと企業内データの双方を分析し関連させて、企業活動に利用することが始まっている。企業内部および外部のデータを、それぞれ構造化データ、非構造化データに分類したものが表 2 である。

表2 企業の内部、外部データの種類

データ種類	企業内部	企業外部
構造化データ	業務オペレーション・データ 販売、預金、経理、財務、調達、 顧客、製造、品質	各種公開データ 市場調査(業種、業態)、政府発表、 経済指数、企業決算、株式市場
非構造化データ	メール、文書、画像、映像 システム・ログ、Webアクセス・ログ 医療カルテ、RFIDデータ、 株式トレーディング・データ	ソーシャル・ネットワーク・データ BLOG、交通トラフィック、カー・セン サー、 GPSデータ、農地センサー、 インターネット・公開データ

損害保険会社は、被害を及ぼしそうな大型台風の予測進路上に自社の保険顧客がいれば、家屋などの損害を予測でき、台風が過ぎ去った後の、被害者への連絡や損害対策に即時に対処できるようにする。気象データと顧客データを利用することで、すばやい行動が可能になるのである。

顧客に販売した機器から定期的に位置データと機器状態データを発信させ、機器が安全に使用されているかチェックし、保守の時期には顧客に連絡をするようなことが行われつつある。将来の自動車保険では、契約車に車の状態と運転状況データを発信する機器を設置できれば、保険会社はデータを収集し分析できる。事故率などの統計値と実際の運転データに基づいて、契約者個別に契約額を算定することが可能になる。流行の「駅ナカ」(駅構内で展開される一定規模の商業スペース)では、構内店舗で決済機能付き携帯電話によって支払った販売データと、付随して得られる位置データを組み合わせて客の動線を分析し、店舗の配置や通路計画などに活用している。

損害保険会社の例では、気象データという公開されるデータを自ら取得する。自動的に収集可能な販売データと異なり、能動的にデータを取りに行かなければデータが集まらない。機器データと駅ナカの例では、データの発生の仕組みを自ら仕掛けて、データを集めている。これまでのデータ収集と販売推進の手法とはまったく異なり、データを発生させることから考慮しなければならない。またソーシャル・ネットワークのデータは、分類し整理して利用しなければならない。これまでの正規化されたデータに比べ、利用可能になるまでに、あいまいな部分を排除し分析の目的に合うように加工する前処理が必須である。

このように、企業の外部で発生するデータを利用するには、これまでの企業の内部で発生するデータ利用と比べ、データ収集方法からマーケティング手法が異なる。

5. 分析の深さ

データ分析には、分析の深さに応じて段階があり、事前定義されたレポートから始めて、いくつかのレベルがある。分析の深さのレベルを表3に示す。深さのレベルと定義についての解釈は、いくつも存在している。過去データのレポートと統計処理の適用でレベルを分ける定義や、予測レベル以降は過去の分析とは異なることから、予測、現在事実、最適化を高度分析とする分類もある^[4]。本稿では、分析に使用するデータの種類を基に、分析のレベルを6段階に分けている。

表3 分析のレベル

分析データ	分析のレベル	内容	処理
過去データ	事実	何が起きたのか	事前定義型 定型検索
	詳細	何処で, 何が	非定型検索 OLAP
	理由分析	何故か, 原因は	要因分析
	予測	何が起きそうか	モデル作成 事実と理由から予測
過去データ + リアルタイム データ	現在事実	今何が起きているか	業務プロセスと連携
	対応と最適化	ベストな対応	自動アクション

分析の深さ ↓

表3のはじめの四つのレベルは、過去データの分析である。目的別に整理され時系列に蓄積された過去の事象からパターンを見つけ、予測モデルまで達する。下の二つのレベルは、高度なリアルタイム分析とも言われており、今現在起きている事象、すなわちリアルタイム・データと過去データを合わせて分析する。

クレジットカード会社は、顧客のクレジットカードが使用されるごとに、カード盗難などによる不正使用かどうかチェックしている。カード決済処理と同時に使用箇所、購買金額などが過去のデータと比較される。非常に短時間に、狭い範囲での高額な同種類の製品の決済などはチェックの対象になり得る。海外での決済なども過去データと比較される。このような要件は、リアルタイムあるいはニア・リアルタイム処理される。道路のトラフィック・データから、道路状況を判断し、最適な信号制御をすることや、刻々と変化する企業の株価と売買状況に応じて、事前定義されたルールと過去のデータを参照し、株式売買をリクエストする証券会社のアルゴリズム・トレードも典型である。

6. IT サービス企業の役割

データ分析について、対象データ、データウェアハウスの技術および Hadoop による分散処理技術を明らかにし、データ分析の深さのレベルを示した。先進的な企業は、高度なデータ分析を強化し、さらにビッグデータの活用を始めているが、すべての企業が分析技術を持ち、分析技術者を抱えているわけではない。むしろ大半の企業は、高度な分析もまだ行っておらず、データ分析の理解も足りず、レポートに留まっていることが多い。データ分析を自社で推進することができなければ、IT サービス企業にその役割を求めることができる。IT サービス企業は、分析機能を企業に提供するために、必要な機能を持つ企業を吸収するなどして準備を進めている^[6]。

企業で必要な、分析基盤の調達、分析の実施、分析人材育成について、社内で準備可能な場合と社外に求める場合の比較を表4に示す。IT サービス企業にとっては、社外調達に応えるサービスを提供することが必須である。

表4 分析サービスの調達

	社内調達	社外調達
分析基盤	社内構築	クラウドで調達 PaaS(Platform as a Service) IaaS(Infrastructure as a Service)
分析の実施	社内組織で実施	フルアウトソース Database as a Services
分析人材育成	社内	教育 コンサルティング

7. おわりに

これまで、企業の意味決定について、その行動や判断に役立てるためのデータ分析の重要性と、台頭しているビッグデータの分析をするための技術について紹介してきた。さらに企業の業務プロセスで発生するデータと、Web空間で発生し取得したデータを関連させて分析することで、ビジネス・インテリジェンスに利用できることを明らかにした。また、ITサービス企業が、データ分析を必要とする顧客に、何を提供すべきかを示した。

日本ユニシスは、2005年に国内で最初に、最新技術を採用した分析基盤であるDWHアプリケーションに取り組んだ⁸⁾。ICT (Information and Communication Technology) 事業にも早くから取り組み、既に多くの顧客にサービスを提供している。さらに、メインフレームの時代から長くオペレーションズ・リサーチや人工知能に取り組んできており、それらの研究の積み重ねが、データの高度分析やデータ・マイニングの先端技術に生かされている。顧客分析や売上分析で定評のあるソフトウェア MiningPro21[®] や、テキストマイニングなどの技術を提供しており、顧客スコアリング、需要予測、自動発注をはじめ、多くのシステムで実証してきた。最近では、Web空間のソーシャルメディアのビッグデータを活用する感性分析 (Sentiment Analysis) を行うソフトウェア TopicExplorer[®] も提供している。将来的には大量となる広域地域の医療データ分析にHadoopを適用したシステムの構築実績もある。ITサービス企業として顧客に様々な分析サービスを提供可能である。

日本ユニシスでは、今後とも、これら自社保有技術を活用し、企業の内部および外部で発生するデータの分析、それらを関連させた企業のデータ分析に関するあらゆるサービスを提供し、顧客の期待に応える所存である。

-
- 参考文献** [1] William H. Inmon, Building the Data Warehouse, John Wiley & Sons, Inc, 1992.
邦訳: W. H. インモン著, 藤本 康秀訳, はじめてのデータウェアハウス構築, オーム社, 1995.
- [2] Ralph Kimball: The Data Warehouse Toolkit, John Wiley & Sons, Inc, 1996.
邦訳: ラルフ・キンボール著, 藤本 康秀監訳, データウェアハウス・ツール・キット, 日経BP社, 1998.
- [3] ビル・ゲイツ著, 大原 進訳, 思考スピードの経営, 日本経済新聞社, 1999.
- [4] トーマス・H・ダベンポート, ジェーン・G・ハリス著, 村井 章子訳, 分析力を武器とする企業, 日経BP社, 2008.
- [5] イアン・エアーズ著, 山形 浩生訳, その数学が戦略を決める, 文芸春秋, 2007.

- [6] 羽生貴史, データベース技術の今後の動向, ユニシス技報, 日本ユニシス, Vol.29 No.2, 通巻 101 号, 2009 年 8 月.
- [7] Michael Stonebraker, Ugur Cetintemel, "One Size Fits All": An Idea Whose Time Has Come and Gone, ICDE 2005.
- [8] 行成 敦, データベースの動向と課題—日本ユニシスの取り組みと今後の動向, ユニシス技報, 日本ユニシス, Vol.26 No.1 通巻 89 号, 2006 年 5 月

執筆者紹介 山崎 慎一 (Shinichi Yamazaki)

1972 年東京都立大学工学部卒業。同年日本ユニバック(株)(現日本ユニシス)入社。オンラインリアルタイム・システム, ホットスタンバイ・システムなど各種システム開発, データベース管理システム保守業務に従事, 1995 年からデータウェアハウス構築, 2005 年から, DWH アプライアンスを担当, データ・モデリングが研究分野。現在, 総合技術研究所インキュベーションラボに所属。情報処理学会会員。

