ビジネス・アナリティクスの最新動向（3）

横川雅聡（よこがわ　まさとし）
株式会社 NTT データ
中川慶一郎（なかがわ　けいいちろう）
株式会社 NTT データ数理システム
生田目　崇（なまため　たかし）
中央大学理工学部

1. はじめに

前回は分析シーンや目的を整理したうえで，分析技術のトレンドという視点からビジネス・アナリティクス（BA: Business Analytics）の動向を論じた．第 3 回目の今回は BA を支える IT 基盤について解説する．

はじめに，BA の側面から見た IT 基盤の変遷について説明する．次にビッグデータ時代の BA を支える IT 基盤についてまとめ，今後の方向性を試論する．なお，ここでの IT 基盤とは，データの蓄積・検索・処理を行うデータ基盤と，分析手法や分析フロー管理などを IT 上で実装した分析基盤に限定し，ハードウェアなどについての記述は必要最小限度にとどめた．

2. BA から見た IT 基盤の変遷

分析技術と IT 基盤の関係を振り返ると，提唱された新しい分析技術に IT 基盤の性能が追いつく姿が見られる一方で，IT 基盤の進化が分析技術を牽引してきたというように，二重らせん的な発展を遂げてきたといえよう．

ここでは，コンピュータのダウンサイジングが始まった 1980 年代後半からビッグデータ時代の今日に至るまでの情報分析活用と，そこでの IT 基盤の変遷を説明する（図 1）．また，前回説明した BA における 4 つの分析シーンや分析技術との対応も併せて整理する．

2.1　黎明期

現在の BA の源流は 1980 年代後半まで遡ること

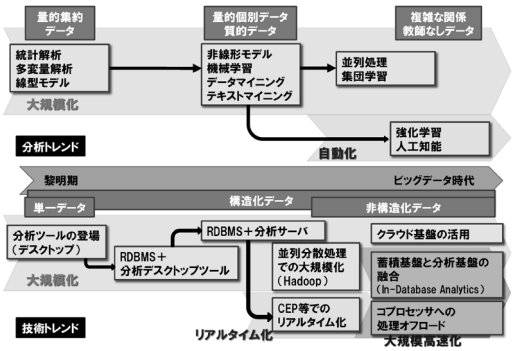

図 1　データ分析および基盤技術の変遷

ができる．当時の情報分析活用は，高度な分析というよりも業務システムの延長としてデータをいかにまとめるかということに主眼が置かれていた．多変量解析をはじめとする分析技術が PC 上で実行できるようになったのもこの時期であるが，ビジネスの分野で活用していこうとする機運はそれほど高いものではなかった．

この時期の情報分析活用を 4 つの分析シーンに当てはめると，集計分析型の BA を導入し始めた段階といえる．また，それを支える IT 基盤も未成熟であった．したがって，分析といっても PC（デスクトップ）上で動作する分析ツールに対して，単一データを用いた処理を行うことが中心であり，PC のリソース制約に基づく限られた分析にとどまっていたといってよい．

2.2　ビジネス・インテリジェンス時代

1990 年代中盤に入り，情報分析活用の新しい潮流としてビジネス・インテリジェンス（BI: Business Intelligence）がビジネスの世界を賑わし始めた．データ

分析の重要性がさまざまなところで問われるに従い，RDBMS（Relational Database Management System）をベースとした分析専用のDBであるデータウェアハウス（DWH: Data Warehouse）の構築が進んだ．

例えば，小売業ではPOSシステムの普及が一巡するとともに，粒度の細かいデータが日々取得できることになった．その結果，販売実績，プロモーション，購入者といった多次元のデータを扱う時代となり，それらの間の相関関係，因果関係への分析に関心が集まるようになり，DWHの導入が加速していった．

また，コンピュータの性能向上に伴い，多変量解析などと比較して，よりマシンパワーを必要とするデータマイニングなどの分析技術が注目されるようになったのもこの時期であり，分析シーンにおいても発見型のBAが加わった．

その後，BAで扱うデータ量も徐々に増加の一途をたどり，大規模な分析はデスクトップ上での処理から，サーバ上に移行するようになった．しかし，依然として分析の対象は構造化データが中心であり，データ基盤と分析基盤はそれぞれ別々に存在していた．

2.3 RDBMSの全盛とビッグデータの萌芽期

2000年代に入ると，「見える化」の追い風に乗ってBIが普及期を迎えることになり，情報分析活用の領域においてもRDBMSは，「それ以外にデータ基盤はない」と考えられるほど，絶対的な存在となった．

また，CRMやSCMなどIT基盤と分析技術を融合した業務改革の流れに伴い，データマイニングや機械学習といった高度な分析技術も業務の中に使われるようになった．この時期になると高度分析を業務に組み込み，その効果を事前に試算しようとするWHAT-IF型の分析シーンも徐々に現れてくるようになる．

一方，RDBMS全盛の裏で，データ基盤において，新しい変化が産声をあげていた［1］．1つは，これまでの大規模化の流れをさらに加速させ，超大量データを並列分散処理機構で高速にバッチ処理する流れであり，具体的にはHadoopに代表されるMapReduce，さらにはNoSQL（Not only SQL）と呼ばれる一連の技術である．NoSQLは文字どおりRDBMS以外のDB全般を指す用語であるが，その多くが非構造データも扱うので，BAの対象も非構造データまで広がることとなった．

もう1つは，時系列的に発生するストリーム・データをリアルタイムに処理する流れであり，CEP（Complex Event Processing：複合イベント処理）と呼ばれる技術に代表される．しかし，両者はそれぞれ異なる分野で適用が進んだため，2つの流れはいったん分かれることになる．

2.4 ビッグデータ時代

2010年代に入っても大規模化，リアルタイム化の流れは加速し続け，2012年はビッグデータ元年といわれるようになった．また，この頃からスマート・グリッドやスマート・コミュニティといった，いわゆるスマート・ビジネスが注目されるようになり，情報分析活用によって知的なサービスを実現するプロアクティブ型の分析シーンが本格的に始動することになる．

このような流れを受け，IT基盤ではRDBMSをベースに，ハードウェア，ソフトウェアが一体になったDWH専用機であるDWHアプライアンスが本格普及する一方で，RDBMS一辺倒であった情報系システムも，内部処理の並列化および後述するMapReduce機構の搭載など，進化をし続けている．

一方，メインプロセッサ以外にもコプロセッサを活用して処理を高速化するようなHPC（High Performance Computing）で培われた技術がエンタープライズ・ビジネスに適用され始めており，コモディティ・ハードウェアを用いた大規模な処理システムは遠い存在ではなくなってきている［2］．

今後は，IT基盤自体も無尽蔵のコンピュータ・リソースをオンデマンドで使い倒す，クラウドの活用に移っていくであろう．

また，深層学習や集団学習といった機械学習をベースとした最新の分析技術をビジネスに利用しようとする試みも徐々にではあるが始まりつつあり，ここでのIT基盤としてもHPCの実用化がさらに進んでいくであろう．

3. ビッグデータ時代のBAを支えるIT基盤

ここでは，ビッグデータ時代を代表する2つの

IT 基盤である Hadoop と CEP について詳細に説明する．

3.1 Hadoop

大規模並列処理技術の代表である Hadoop は，google の技術者のレポートをもとに，OSS として開発されたシステムであり，大規模データ分析，高速バッチ処理の要素技術である．特にバッチを高速化することは，運用時間中に分析結果を入手することで意思決定を迅速化したり，システム運用要員のコストを削減したりする効果が見込めるので，ビジネスへの導入インパクトは大きい．

Hadoop は主に 2 つの構成要素からなる．1 つが分散ファイル・システムの HDFS（Hadoop Distributed File System）であり，もう 1 つが分散処理フレームワークの MapReduce である（図 2）．

HDFS は複数のサーバ上に 1 つのファイル・システムを構築する，いわゆる「分散ファイル・システム」である．利用者は物理的に複数のサーバが存在することを意識することなく，1 つのファイル・システム空間を利用できる．

一方，MapReduce は複数のサーバを利用して並列に処理を行うためのフレームワークである．また，HDFS と連携してデータ配置のローカリティを活用し，処理を効率的に実行する仕組みを持つ．

MapReduce は，必ずしもすべての分散処理に対して有効にあてはまる処理のタイプではないため，Map 処理と Reduce 処理のフレームワークに上手く当てはまる分散処理とはどのようなタイプの処理なのかを見極めて活用する必要がある．

例えば，集計分析型の BA の場合，Hadoop を用いてデータと計算処理を分割したうえで一度集計し，その結果のみを取り出して最後にマージすることで，大規模なデータを効率的に集計することができる．

3.2 CEP

データ基盤に蓄積されたデータではなく，発生したデータに対して分析を行う場合は，リアルタイムに発生するイベントやストリーム・データに対する処理技術が有効となる．CEP とは，このような処理を行う IT 基盤である．

CEP はデータ処理をすべてメモリ上で実行するので，用途によっては毎秒数十万から 100 万イベントものスループットを実現し，入力から出力までのレイテンシを 100 万分の 1 秒のレベルにまで抑え込むことができる．RDBMS の SQL が要求の都度，蓄積された対象データを読み取って検索・集計するのに対し，CEP ではストリーム・データから必要な情報だけを差分計算することで，イベントが発生したタイミングとほぼ同時に結果が得られる．

CEP では，絶えず発生するビジネス・イベントを単一あるいは複数のストリーム・データとして入力アダプタで受信し，特定のイベントをモニタリングして実行する「継続的クエリ」と呼ばれるデータ処理を行い，続くアクションを出力アダプタ経由で起動する（図 3）．

また，多くの CEP 製品には「キャプチャ」と「プレイバック」機能がある．実際に流れるストリーム・データをキャプチャし，任意の速度で再生することで，CEP 上のモデルのシミュレーションを行い，改善につなげることができる．これは瞬時の WHAT-IF 分析を行うことに相当する．

CEP の具体的な利用イメージとしては，（1）小売業での POS システムや製造業での MES（製造実行システム）が出力する現時点の情報をリアルタイムにモニタリングする，（2）Web サイト上のクリック・ストリームに対してレコメンドを行う，（3）

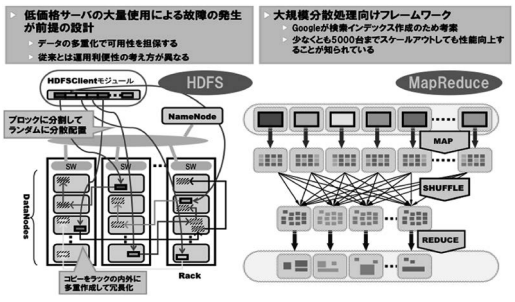

図 2　HDFS と MapReduce

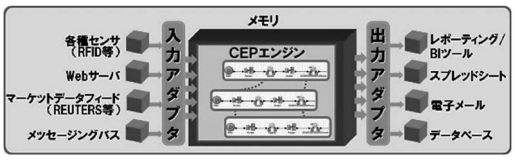

図 3　CEP の概要

IT 機器の一連の操作から不正利用を検出する，といったことが挙げられる．

CEP を活用する際のポイントは，発生するイベントに対するリアルタイムの「モニタリング」と，それに続く「コントロール＝アクション」の一連の流れにある．

4. IT 基盤から見たビッグデータ活用のポイント

本節では，IT 基盤の面からビッグデータ時代の BA を実践するための技術的なポイントを述べる．

4.1 バッチはなくならない

まずは蓄積された大量データに対する処理に対して見ていく．集計分析型 BA を中心にこれまで行われていたバッチ処理は，一見古い方式のようにも思われるが，全件データに対する操作が必要となる分野にはこれからも必須となる．例えば，Web 上のログ分析や，売上データに対する分析などが挙げられる．バッチ処理の大きな欠点は，全データを検索もしくは照会しなければならないため，データ量が増えれば増えるほど，処理に膨大な時間がかかる点にある．

多くの分析者を悩ませているこのバッチ処理時間をまずは短縮しないことには，ビッグデータ活用の現実味は帯びてこない．前述の Hadoop はバッチ高速化の代表的な技術といえる．

4.2 Raw データの蓄積・活用

分析の実施に先立って必ず意識しなければならないのは Raw データの蓄積である．データを活用した分析を行うにしても，分析対象のデータが蓄積されていなければ分析を始めることすらできない．あるいは分析に必要なデータの量が足りていないために分析に耐えられないといったことも起こりうる．

分析の検討をするにしても一定量の Raw データは必要であり，その蓄積には時間を要することから，例えば Web ログ分析を行うと決定したら，分析の詳細を詰めるよりも先に Web システムから得られる Raw データの蓄積を開始するといったことが望ましい．

この点が RDBMS での DWH と Hadoop との大きな違いとなる．DWH はあらかじめ定めた活用の計画に従ってデータ蓄積するのに対して，Hadoop は Raw データのまま格納することが一般的であり，活用の柔軟性を持たせることができる．

また，アドホックな分析を行うことを考えると，その都度どこかからデータを取り寄せてくるようでは，タイムリーな分析は行えない．分析を行うために必要なデータが Hadoop クラスタに蓄積されており，任意のタイミングで利用可能な状態にあるということも重要である．

4.3 流れるデータをリアルタイムに分析

バッチ処理時間の短縮には，処理を高速化させる以外にも方法はある．バッチ処理自体をなくしてしまうことである．

データを大量に蓄積してから集計などのデータ処理を行うから大変なのであり，蓄積する前にデータ処理を施し，必要な結果だけを集めることができれば，バッチ処理時間の短縮になる．もちろん，先に述べたとおり，すべてのバッチ処理がなくなるというわけでは決してない．

また，バッチ処理をいくら高速化しても，いったんデータを蓄積してから都度処理を起動するようなタイプのアプリケーションでは，どうしてもタイムラグが発生する．場合によってはビジネスのスピードに対応できないこともあり，その場合は別のアーキテクチャが求められる．前回説明した BA の特徴と照らし合わせてみると，例えば，金融アルゴリズム取引，不正取引の検出，テレコムなどでのサービスレベル低下検出，橋りょうなどの構造物（もしくは製造ラインの品質）のモニタリングなどでは，その時々で複雑な状況下でリアルタイムな判断が求められる分野である［3］．

こうした分野で従来のようにデータを蓄積してからそれを取り出して分析をする場合，データが発生してからアクションを起こすまでに，ある程度のタイムラグが生じることが避けられないが，これが大きな機会損失につながってしまう場合も少なくない．

4.4 分析モデルのデプロイメント

第 1 回目にも書いたが，今後はセンサ・ネットワークの普及に伴い，交通，環境，防災などの社会インフラ，「ヒト」「モノ」「カネ」の動向追跡，スマート・グリッドの中核となる HEMS（Home Energy

Management Systems），医療，FA（Factory Automation），BAS（Building Automation System：ビル設備総合管理システム）などといった多くの分野で BA の活用が想定される．

多くのセンサから絶えず発生し続けるデータをリアルタイムに処理していくために，CEP は必要不可欠な技術であることは間違いない．

CEP の技術が登場した当初は，先にも述べた，極限のリアルタイム性が求められる金融アルゴリズム取引など一部の分野と見られてきた．それがこのビッグデータ時代においては，適用分野の広がりを見せ，重要な技術となっている．

BA における CEP の活用ポイントとしては，バッチ処理で構築された分析モデルを，CEP に組み込み，発生するストリーム・データに対してリアルタイムに分析モデルを適用して分析結果を得て，アクションにつなげていくところにある．

5．新しい潮流：データ基盤と分析基盤の融合

Hadoop や CEP といった新技術を使う BA だけではなく，これまでの DWH と分析ツールを組み合わせる BA もまだまだ現役である．しかし，大量データに対するデータ分析では，データ基盤（DWH）と分析基盤（分析ツール）との間のデータ転送がボトルネックになるため，データ基盤内での分析処理を実現する In-Database Analytics のような技術が求められている．

これまでの分析ツールは，「DWH からのデータ抽出」「分析モデルの構築」「予測などの分析処理の実施」を DWH などと連携しつつ，主に分析ツール側で行ってきた．しかし，今後はその役割が見直され，大部分の処理は大量データを蓄積する DWH 側にオフロードされ，分析ツール側では最小のデータに対しての必要最低限の処理を担当していくようなケースが増えてこよう（図 4）．

In-Database Analytics は，分析処理（＝Analytics）をデータベースの中で（＝In-Database）実施する技術を指す．DWH 内に格納されたデータを読み込み，各種加工処理（JOIN，GROUPING，SORT など）を並列で行ったうえで，DWH が備える MapReduce 機構上に実装した分析アルゴリズムに中間的な結果セットを受け渡す．DWH から分析

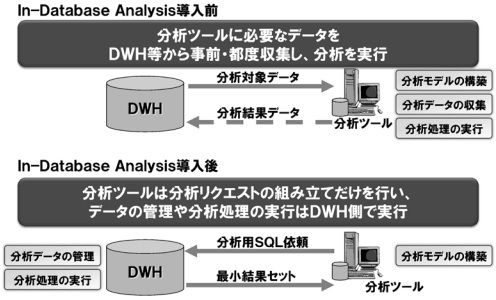

図 4　BA における役割分担の見直し

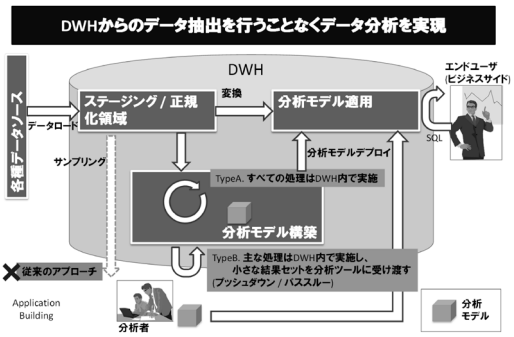

図 5　In-Database Analytics の流れ

アルゴリズムへのデータ受け渡しには UDF（User Defined Function）の I/F を利用した拡張関数を用いる．DWH と分析アルゴリズムは同じメモリ空間を利用し，並列分散処理を実現している．

In-Database Analytics が実現できた背景としては，データ量の増加とともに，データ活用のニーズも増加しており，各 RDBMS ベンダがそのニーズに合わせて DWH としての並列処理度を高めてきている流れがある（図 5）．

大量データがすでに蓄積されている DWH 内で BA を行うことで，これまではあきらめてきた大量データを用いた処理が可能となり，処理の試行回数を増やしてより効果的に解を導出することができる．

In-Database Analytics における分析モデルの構築には，2 つのタイプが存在する．

1 つは分析モデルの構築に必要となる処理をすべて DWH 内で実施するタイプである．もう 1 つは処理の一部を DWH にオフロードしてモデルの構築自体は分析ツール側で実施するタイプである．こちらは生成される分析モデルが二次元の表形式で表

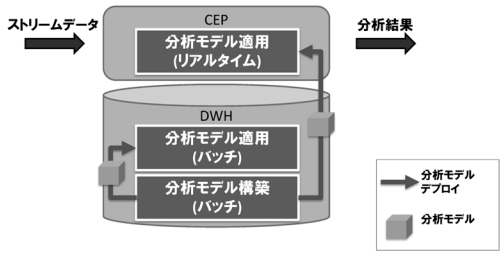

図6　各基盤技術の連動

しにくい場合などに使われるが，両タイプともにほとんどの処理をDWHの処理能力で解決するので，大量データをDWHの外に出す必要がない［4］.

6. 今後の発展

これまではバッチ処理中心で，データ基盤と分析基盤が融合したHadoop，流れるデータからリアルタイムに結果を出していくCEP，構造化データを中心としたデータ基盤であるRDBMSに分析機能を盛り込んだIn-Database Analyticsについてそれぞれ説明してきたが，現状はまだおのおのがそれぞればらばらに存在している状態になっている.

今後のBAを支えるIT基盤として必要なポイントは，それぞれの基盤技術がシームレスに連動し，分析者などからのアクセスを容易にする，もしくは，分析した結果をスムーズに日々のオペレーションに組み込んでいくことにある（図6）.

そのための動きは実はすでに始まっており，分析ツール側から各基盤への処理のオフロードが実装されてきている．これにより，分析者はこれまで慣れ親しんだツールを使いつつ，これまでできなかった大量データに対する分析を，データ量を意識せずに実施可能となっている.

今後は，HadoopやDWHで構築された分析モデルをCEPに動的に組み込むことで，リアルタイムに分析モデルを適用する動きも加速することが想定される.

BAで扱われるデータ量や，求められる処理速度，解くべき課題の複雑性は今後も増加していくことが想定され，BAを支えるIT基盤技術もビッグデータの3つのVに合わせて進化を続けるであろう.

以前のRDBMSのように絶対的な存在となるIT基盤技術が将来的には登場することも想定されるが，しばらくはいろいろな基盤技術を組み合わせ，処理をそれぞれ最適な場所で行っていく工夫が必要となる.

参考文献

［1］桑田修平，海老沢和則，中川慶一郎「M2Mデータの活用法と処理基盤」『電子情報通信学会誌』Vol. 96，No. 5，pp. 347–353（2013）

［2］中川慶一郎，小林佑輔（編著）『データサイエンティストの基礎知識　挑戦するITエンジニアのために』リックテレコム（2014）

［3］横川雅聡，黒田寿男「メニーコア時代のHPC（High Performance Computing）」（2013）
http://www.nttdata.com/jp/ja/insights/trend_keyword/20131114 01.html

［4］横川雅聡「In-Database Analyticsとリアルタイム処理」（2014）
http://www.nttdata.com/jp/ja/insights/trend_keyword/20141009 01.html

略歴

横川　雅聡（よこがわ　まさとし）
2000年東京工業大学工学部制御システム工学科卒業．2009年株式会社NTTデータ技術開発本部．データウェアハウス，ビジネスインテリジェンスおよびビッグデータに関する研究開発，コンサルティングに従事.

中川　慶一郎（なかがわ　けいいちろう）
2000年早稲田大学大学院理工研究科博士課程満期退学．博士（工学）．2012年株式会社NTTデータ数理システム取締役．マーケティング・エンジニアリングおよびビジネス・アナリティクスに関する研究開発，コンサルティングに従事.

生田目　崇（なまため　たかし）
1999年東京理科大学大学院工学研究科博士課程修了．博士（工学）．2013年中央大学理工学部経営システム工学科教授．マーケティング・サイエンスおよび経営科学に関する研究に従事.