ビジネス・アナリティクスを支える分析ソリューション

鍋谷　鼎一

ビッグデータの登場以来，大量データをビジネスに活用する取り組みに注目が集まっている．近年では，大量データを「どう蓄積し処理するか」から「ビッグデータで何をするのか」に焦点が移るにつれ，データ分析を通じてビジネスに資する洞察を引き出す「ビジネス・アナリティクス」にも大きな関心が集まっている．ビジネス・アナリティクスへの関心の高まりは，分析ソフトウェアにも大きな変化をもたらしており，本稿ではビジネス・アナリティクスの動向を紹介した後，分析ソフトウェアの観点からその変化を整理する．また，「分析技術の組合せ」，「大規模データ分析技術」にフォーカスし，具体的な事例を交えて紹介する．

キーワード：ビッグデータ，ビジネス・アナリティクス，In-Database Analytics

1. はじめに

2008 年 6 月，当時新入社員だった筆者のノートを見返すと

「全データはメモリ不足で中断．データを分割して

これから回しても報告に間に合わない…！」

という悲鳴にも似たメモがある．

もちろん，筆者が未熟だったことが大きな原因の一つだが，当時，分析ツールで提供されているアルゴリズムは

・分析対象データ全体がメモリ上に存在

・シングルスレッド動作

が前提となっているものが大半であり，一度に処理できるデータ量に限界があった．その背景には，搭載メモリ量を超えるデータを分析するためにはアルゴリズムを根本から見直す必要があること，また理論的に並列処理が可能なアルゴリズムであっても，並列処理を行うには分析処理のほかに「データ分割」，「スレッド／プロセス制御」，「共有データ管理」，「結果の集約」などを実装する必要があり，実装コストがビジネスメリットに見合わなかったことがある．

しかし，2010 年頃からビッグデータが大きな注目を集め，その状況に変化が訪れる．ビッグデータというキーワードは今では広く認知されるようになったが，そもそもビッグデータという概念には厳格な定義はなく「情報量が増えすぎて，分析に使用するデータがメモリに収まり切らなくなり，分析用ツールの改良が必要となった」というのが，ビッグデータと呼ばれるようになった背景である [1]．

「必要は発明の母」の諺のとおり，ビッグデータへの関心の高まりは Hadoop などの分散並列処理基盤の普及を始め多くの変化をもたらした．また，ビッグデータへの関心が「どう蓄積し処理するか」ということから「ビッグデータで何をするのか」ということに焦点が移るにつれ，データ分析を通じてビジネスに資する洞察を引き出す「ビジネス・アナリティクス」にも関心が向けられるようになり，分析ソフトウェアにも変化をもたらしている．

本稿では，筆者の経験談を踏まえつつビッグデータがもたらした変化を分析ソフトウェアの観点で整理し紹介したい．まずビジネス・アナリティクスの動向を紹介し，そこで必要とされる分析技術について整理する．分析ソフトウェアの大きな変化として「大規模データ分析技術」，「分析プラットフォーム」にフォーカスし，具体的な取り組みを紹介する．

2. ビジネス・アナリティクスの動向

企業内外に蓄積されているデータをビジネスに活用するという考えは，古くは 1970 年代の意思決定支援システム [2] を始め，さまざまな形で議論されてきた．近年，ビッグデータやビジネス・アナリティクスが注目を浴びている背景には Google，Amazon といった先進的な企業の取り組みにより情報のもつ価値が再認識されたことや分析技術による競争優位性への期待がある [3]．

しかし，ビッグデータをどのように活用するかにつ

なべたに　こういち
株式会社 NTT データ
〒 135–8671 東京都江東区 3–3–9 豊洲センタービルアネックス

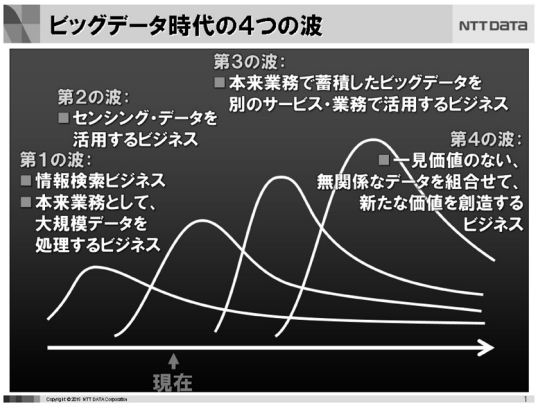

図 1　ビッグデータ時代の四つの波

いては，企業によっても，また同一企業内でも部署によって大きく異なる．ここでは，ビッグデータの活用段階とビジネス領域を四つの大きな波として整理し [4]，各段階で必要とされる分析技術についてまとめる．

第 1 の波は，本来業務として大規模データを処理するビジネス領域である．ここでの対象データは，たとえばコンビニエンス・ストアの POS データや，通信会社の利用履歴であり，これらを活用した販売促進や解約防止を目的とした分析はビッグデータという語が登場する以前からビジネスに組み込まれてきていた．

第 2 の波は，センシング・データを活用するビジネス領域である．この領域でのビッグデータ活用は，IoT (Internet of Things) を抜きに考えることはできない．IoT とは，従来主にインターネットに接続されていたパソコン，サーバ，モバイル機器だけではなく，自動車や自動販売機といった市中の機器や橋やトンネルといった社会インフラに至るまですべての“モノ”(Things) をつなごうとするものである．“モノ”に取り付けられたセンサーからヒト・モノ・カネの動きに関する膨大で詳細なデータをリアルタイムに収集できるようになってきており，現在はこれらのデータを業務課題解決にいかに活用するかが問われる時代といえる．

第 3 の波は，これまでに本来業務で蓄積したデータを二次活用して別のサービス・業務で活用する領域である．たとえば，飲料業界ではネットワークを介し各自動販売機の商品の実売本数，在庫本数，滞留時間などの販売情報をリアルタイムに把握することができるようになってきている．ここで，今まで現地で確認する以外知りようがなかった販売状況を遠隔からリアルタイムに把握することは業務レベルの課題であり，こうしたデータ活用が一次利用，すなわち第 2 の波に対応する．一方で，配送担当者はこのエリア，この季節にはこれくらい売れるという肌感覚をもっており，それを基に補充商品のバランスを調整していた．このような勘と経験に基づく暗黙知を継続して継承していくことはきわめて困難である．暗黙知を形式知化し，組織的に知識を継承していくことは経営レベルの課題であり，たとえば収集した販売状況データから販売量予測をモデル化を通じ，販売量予測・補充計画を形式知化する取り組みなどは二次利用，すなわち第 3 の波に対応する．

最後に第 4 の波は，一見価値のない無関係なデータを組み合わせて，新たな価値を創造するビジネス領域である．たとえば，「Google クライシスレスポンス」 [5] では，カーナビシステムから収集される通行実績情報を Google マップに重ねることで通行可能なルート情報を提供している．災害発生時に車で被災地を目指す人々が通行可能なルートを見つけることができ，高い評価を受けた．Google など先進的な企業以外ではまだ事例が少ない状況だが，政府によるオープンデータの整備も進んできており今後，この領域での活用が普及していくと考えられる．

3.　ビッグデータの活用段階と必要とされる分析技術

ビッグデータの活用段階とビジネス領域について四つの波という形で概説したが，次に段階ごとに分析の観点から必要とされている技術を整理したい．

まず，第 1 の波の特徴はビッグデータというキーワードが出現する前から大規模データをビジネスに活用することが行われており，分析技術としてはアソシエーション分析などデータベース上で比較的実装しやすい集計ベースの手法が用いられることが多い．

第 2 の波における分析の特徴は，間断なく届く大量のセンサーデータに対する高速かつ高度な分析が必要な点である．たとえば，橋梁などにセンサーを取り付け損傷状況や震災時などの損傷状況の把握を行うシステムでは各センサーから毎秒数十〜数千サンプル程度の測定値が届く．これらのデータを収集，加工したうえで異常検知や損傷状態の推定をリアルタイムに行う必要がある．

まずビッグデータ登場以前のデータ活用では，一般的に図 2 で示されるシステム構成のようにデータを蓄積する DB，分析用の環境，分析した結果（予測モデルなど）を実行する AP サーバがそれぞれ別の環境として構築されることが多い．

この場合，データを活用して分析するフローとしては

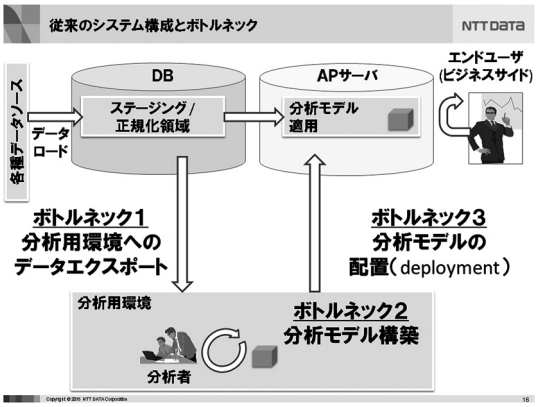

図 2　ビッグデータ登場以前のシステム構成

1. DB から分析に必要なデータを分析用の環境にエクスポート
2. 分析ツールを使って分析作業を実施
3. 分析結果を情報系システムに反映

という手順を踏むことが多いが，それぞれ

1. 大量のディスク I/O，ネットワーク I/O が発生し，データ転送に時間を要する
2. 分析対象データがメモリ上に載りきらない場合は特別な処理が必要
3. 分析結果を人手で情報系システムに配置するための作業 (deployment) が必要

などそれぞれ時間を要する作業が発生する．第 2 の波の特徴であるセンサーなどから届く大量データを分析し，結果を反映するためには上述の各作業の時間を縮める必要がある．

それぞれの作業を効率化するために各種取り組みが行われているが，一例を挙げると

1. データをエクスポートせず，DB 上で分析処理を行う In-Database Analytics
2. 逐次学習することでメモリ上にデータが載りきらないような大量データでも分析を可能にするオンライン機械学習アルゴリズム
3. PMML (Predictive Model Markup Language) などモデル記述言語による export/import を活用した分析モデルの自動配置

といった技術が普及してきている．具体的には In-Database Analytics については各ベンダから具体的な製品としてリリースされ始めており，また SI ベンダにおいても研究開発，技術検証が進められている．In-Database Analytics については 5 節で詳述したい．オンライン機械学習については，NTT ソフトウェアイノベーションセンタと株式会社 Preferred Networks が共同開発したオンライン機械学習向け分散処理基盤 Jubatus [6] が知られている．回帰，分類，外れ値検知，クラスタリングといったアルゴリズムが実装されておりオープンソースとして公開されていることもあり今後のさらなる発展が期待される．分析モデルの自動配置についても IBM 社の SPSS Modeler など商用製品で PMML による export/import がサポートされるようになってきており，この面でも効率化が進むことが期待される．

第 3, 4 の波の特徴は高度な分析もさることながら，マッシュアップという多種多様なデータ，技術を組み合わせる考え方が使われている．ここで，第 1, 2 の波，すなわち「今までの分析の仕方」と第 3, 4 の波，すなわち「これからの分析の仕方」の違いについて筆者の考えを述べたい．第 2 の波までは，分析要件が比較的明確なことが多く，個々の技術者の得意技だけでカバーできることから個人で完結して作業することができた．特に，数値系のデータに対する技術とテキストに対する技術は要求される勘所が違うこともありそれぞれが独立して活動することが多かった[1]．一方，第 3, 4 の波ではデータの種類，目的が多様化することに伴い，必要となる技術も多様化している．特に最近では Twitter を始めとする SNS データの入手が容易になったこともあり，SNS データと他のデータを組み合わせて分析することで，ネット上のユーザの反応やセンチメントを即座にとらえることが可能になってきた．たとえば，2013 年の参議院選挙時における候補者／政党のツイート，リツイート関係から Twitter ユーザの関心政党の変遷をグラフマイニング技術を活用して可視化した事例 [7] や株式に関連するツイートを抽出し，そのセンチメント（ポジティブ表現，ネガティブ表現の有無でセンチメントを定量化）と株式市場との関連性を分析した事例 [8] などテキスト，数値のみならずグラフ構造なども組み合わせた分析事例が出てきている．

ここで重要になってくるのは，必要となる技術が多様化するにつれ複数の技術者が協力してプロジェクトを進める必要がある点である．もちろん，チームを束ねるリーダには複数の技術領域に関する知見と組織を超えて技術者をコントロールする必要がありその役割が大きいのは言うまでもない．分析ソフトウェアの観点からは，各技術者が個別の分析ツールを持ち寄って

[1] 筆者は NTT データ技術開発本部，NTT データ数理システムと分析を生業とする二つの組織に所属したが，いずれの組織でも数値分析系とテキスト分析系は別チーム／別組織であった．

図 3　NTT データのビッグデータ戦略

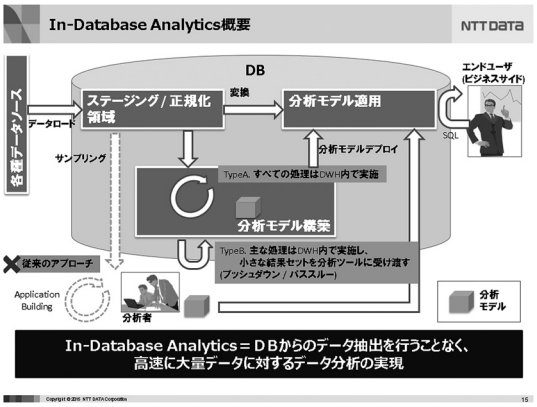

図 4　In-Database Analytics の概要

分析を進めると分析ツール間の連携や全体管理ができなくなり，分析効率の低下はもちろん引き継ぎやメンテナンスの効率低下につながってしまう．そのため，ビジネスの現場では統一的に分析フローを整理，管理できるソリューションの利用が進んでおり商用パッケージでは IBM 社の SPSS Modeler や NTT データ数理システム社の Visual Analytics Platform，オープンソースでは Rapid Miner や KNIME などが知られており，その活用例を 6 節で触れたい．

4.　NTT データのビッグデータ戦略

2, 3 節の動向を踏まえ，NTT データでは以下の戦略でビッグデータビジネスに取り組んでいる．

図 3 に示したとおり，

・データ分析方法論

・Business Analytics 基盤

・Big Data 基盤

の強みを中心に活動を行っている．具体的には，数百を超える分析事例を体系化した方法論 BICLAVIS，分析プラットフォーム Visual Analytics Platform を中心とした NTT データ数理システム社のパッケージ群および In-Database Analytics，DWH アプライアンス，CEP などを中心とした大規模リアルタイム処理基盤技術を核にビジネス展開を進めている．

次節以降では，大規模データ処理技術の一つである In-Database Analytics と分析プラットフォーム Visual Analytics Platform について述べたい．

5.　In-Database Analytics

3 節でも触れたが，従来ではデータを蓄積する環境と分析する環境が分かれており大量データの転送がボトルネックとなっていた．In-Database Analytics 技術とは，図 4 に示すようにデータを蓄積する基盤である DB 側で分析処理を行うものであり，データを抽出し，移動する時間が不要になるというメリットがある．さらに，DB エンジン自体の並列処理度が向上し続けていることもあり，使用する分析アルゴリズムが並列処理に対応していれば大幅な速度向上を実現することができることもメリットである．

筆者が所属するグループでは，クラスタリング手法の一つである K-means 法を DB 上で実行できるようにし，1,000 万件から 100 億件データでその性能を調査した [9]．調査では，100 億件のデータを 56 コアで実行した場合で 4 時間半で処理を完了しており，さらにノードを 560 コア，1,120 コアで実行した場合にそれぞれ約 10 倍，約 20 倍と linear に性能が伸びることを確認した．DB・分析ツール間でのデータ移動が不要になるとともに，高い並列処理性能を示しており大規模データを分析する際の有効な手段になることを示している．現状，サポートされている分析技術が

・回帰

・分類 (Random Forest)

・クラスタリング (K-means)

と限定的であり今後の拡充が望まれるが，大規模データに対する分析技術として活用場面は増えていくと考えられる．

6.　分析プラットフォームとその活用例

3 節で述べたとおり，今後は多種多様なデータに種々の分析手法を組み合わせて活用するシーンが増えてくると考えられる．NTT データ数理システム社が提供する各種分析パッケージ

・Visual Mining Studio（汎用データマイニングシステム）

・Text Mining Studio（テキストマイニングツール）
・Numerical Optimizer（汎用最適化パッケージ）
・S^4（汎用シミュレーションツール）

は，Visual Analytics Platform という分析プラットフォーム上でシームレスに連携可能である．この Visual Analytics Platform 上で統計分析，データマイニング，数理最適化といった複数の分析技術を組み合わせて金融機関の事務センタの事務量予測，要員のスケジューリングを実現した事例を紹介したい．

各金融機関は，営業店で受け付けた口座開設や税金の収納業務を集約して行う事務センタを運営している．ある大手金融機関様から 2010 年頃に団塊の世代の退職や労働力人口の減少を見据え事務センタの運営をより効率化したいとの相談があり要員配置効率化プロジェクトがスタートした．

効率化のキーポイントは，生産管理の分野で発展した考え方である

・山崩し
　時限性の低い業務を後ろ倒しすることで業務量を平準化
・多能化
　一部要員に複数業務を習得してもらい，ピークを渡り歩いてもらうことでより少ない要員で業務遂行を実現

を組み込んだことである．紙面の都合で詳細は割愛するが，いずれも業務要件などの制約を充足しつつ，ピーク業務量やコスト（人件費，研修コストなど）を最小化する数理計画問題として定式化することができる．

他の事例 [10, 11] でも述べられているが，現実的に取り扱えるレベルの数理計画問題として定式化できるかどうかももちろん重要だが，現場利用者の肌感覚にあった要件，基礎数値を揃える[2]ことも現場展開につなげるためには非常に重要である．

この事例では，基礎数値や入力データとして

・標準時間（事務別 1 件当たりの処理時間）
・事務別予測事務件数

が必要で，それぞれ集計分析[3] や統計分析による予測モデルの構築が必要になる．

[2] 本事例では現場の実態と整合性の取れた数値を得るまで半年を要した．

[3] プロジェクト開始当時はお客様側に標準時間算出に必要な情報が保管されていなかった．そこで，事務センタを訪問しストップウォッチ片手に処理時間を計測したが，現在はシステムログから処理時間を取得している．エラー解析用に蓄積しているシステムログを事務センタ運営の基礎値算出に活用しており，2 節で挙げた第 3 の波の好例といえる．

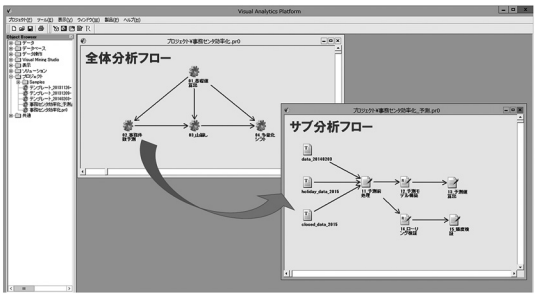

図 5　Visual Analytics Platform

このように統計分析，数理最適化技術を組み合わせて事務件数予測，要員シフトを継続的にお客様に出すためには，各分析コンポーネントを統一的に維持運用管理していくことが必要になる．

分析プラットフォーム Visual Analytics Platform 上では，データや分析処理を表すアイコンをつなげていくことで処理を実装することが可能である．さらに複数の処理アイコン群を一つのモジュールとして実装・管理できるため，複数の技術者が分担して各モジュールを実装していき，モジュールをつなぎ合わせることで一つの大きな分析フローとして管理することが可能になっている．本事例では，図 5 に示すように基礎数値の算出，事務件数予測，山崩し，シフト表作成をそれぞれ Visual Analytics Platform 上で分析コンポーネントとして実装し統一的に運用管理している．

以前は，要素技術ごとに分析ツールが異なりどうしても作業が分断したり，引き継ぎ時の抜け漏れにつながっていたが，本事例では足かけ 5 年以上，筆者を含め 7 人の担当者が引き継いで運用を回してきた．

今後，必要となる技術が多様化するにつれ複数の技術者が関わるプロジェクトは増えてくるものと想定され，分析プラットフォームの重要性はますます高まるものと考えられる．

7. おわりに

本稿では，ビジネス・アナリティクスの動向を紹介し，分析ソフトウェアの観点から最近の変化について紹介した．

読者の中には，Deep learning など先進的な分析技術の成功例が喧伝されるのに対し In-Database Analytics や分析プラットフォームに登場する技術に目新しさがないと感じた方もいるかもしれない．もちろん，Deep learning を始めとした技術の発展は今後も続くと考えられるがビジネス・アナリティクスの観点から見るとあくまで要素技術の一つに過ぎず，むしろいか

に柔軟に既存の技術と組み合わせて価値を生み出していくことが必要となると考えている．

また，筆者が本稿冒頭で触れた悲鳴を上げてから 7 年が経つが，若手社員が分析プラットフォーム上で分析アルゴリズムを並列実行させているのを見ると分析ソフトウェアの進化を改めて実感する．しかし，分析ツール連携や分析アルゴリズムの並列処理は大きな進展を遂げた一方，実ビジネスの現場では 3 節で述べたように幅広い分野に知見をもち，技術者をまとめられる人財が不足しておりボトルネックとなっている．私見だが，幅広い分野の知見をもち，ビジネス・アナリティクスの領域で活躍している諸先輩を見ていると OR 分野出身の方が多く，OR とビジネス・アナリティクスで活用される分析技術との相性のよさを示しているのではないかと感じている．ビジネスニーズの拡大とそれを支える分析ソフトウェアの進化によりビジネス・アナリティクスはこれからますます魅力的なビジネス領域になると期待している．本稿が，OR を学んでいる学生，OR 分野の出身者の方にビジネス・アナリティクスに関心をもっていただく一助になれば幸いである．

謝辞　本稿の執筆にあたり貴重なご意見，アドバイスをいただいた株式会社 NTT データ数理システム 中川慶一郎氏，株式会社 NTT データ 横川雅聡氏に感謝の意を表します．

参考文献

[1] ビクター・マイヤー＝ショーンベルガー，ケネス・クキエ（斎藤栄一郎訳），『ビッグデータの正体』，講談社，2013.
[2] M. S. Scott-Morton, *Management Decision Systems: Computer-Based Support for Decision Making*, Harvard University Press, 1971.
[3] 中川慶一郎，“ビッグデータ時代におけるビジネス・アナリティクス，”情報未来，**40**, pp. 18–21, 2013.
[4] 野村総合研究所，『ビッグデータ革命』，アスキー・メディアワークス，2012.
[5] Google クライシスレスポンス, http://www.google.org/intl/ja/crisisresponse/
[6] オンライン機械学習向け分散処理フレームワーク, http://jubat.us/ja/
[7] 飯田恭弘，岸本康成，藤原靖宏，塩川浩昭，鬼塚真，“大規模グラフ構造データからのコミュニティ抽出と重要度計算—高速化への取組みと応用—，”人工知能，**29**, pp. 472–479, 2014.
[8] NTT データ，Twitter データを用いた金融マーケット向け「Twitter センチメント指標」を開発, http://www.nttdata.com/jp/ja/news/services_info/2014/2014030701.html
[9] NTT データ，独自の In-Database Analytics 技術により従来比 1,000 倍以上の件数の高速データ分析に成功, http://www.nttdata.com/jp/ja/news/release/2015/042700.html
[10] 池上敦子，“ナース・スケジューリング，”オペレーションズ・リサーチ：経営の科学，**54**, pp. 401–407, 2009.
[11] 鈴木敦夫，藤原祥裕，“手術室のスケジューリング支援システムについて，”オペレーションズ・リサーチ：経営の科学，**58**, pp. 515–523, 2013.

