解説

ビッグデータ斜め読み

応用一般

―流行に惑わされないための要点と将来展望―

吉田圭吾　松崎和賢（(株) 三菱総合研究所）

ビッグデータとは何か?

ビッグデータという言葉を文字通り捉えるならば「大規模なデータ」ということになろう．しかし，単純に「何バイト以上のデータをビッグデータと呼ぶ」といった定義は存在しない．驚異的なデータ量の増加と，データ形態の複雑化により，従来のストレージやデータベースシステムによる管理が困難になり，そのような状況を表現するために，ビッグデータという言葉が使われるようになった背景がある．ビッグデータというと，膨大なデータ量について注目されることが多いが，しばしば 3 つの V，すなわち，Volume（量），Velocity（速度），Variety（多様性）という要素で特徴付けられる．

Volume

2009 年の時点で，米国における従業員 1,000 人以上の企業では，平均 200TB のデータ量を抱えていたと言われている[1]．このように，数百テラ，あるいはペタバイト級の膨大な規模のデータを扱うとなると，格納，転送，検索，分析，可視化などあらゆる面で困難さを伴うことになる．とりわけ分析においては，バッチ処理が時間的な制約に収まるよう，限られたデータ以外は分析の対象外とすることが一般的に行われてきた．データを無駄にすることなくスケーラビリティを確保しようとした場合，システムを高性能化するスケールアップの考え方は，コスト面ですぐに限界に達してしまう．ビッグデータ時代では，必然的に分散システムにより性能を増強できる，スケールアウトの考え方が求められることになる．

データ量の問題は，システム構成だけでなく，アルゴリズムに対しても大きな制約となる．たとえば，映像配信・オンライン DVD レンタル事業を展開する米 Netflix 社は，推薦エンジンの性能を向上させるために，100 万ドルの賞金を懸けたアルゴリズムの開発コンテストを開催していたが，開発アルゴリズムの多くは実際に利用されることがなかった．これは，1 億件のデータに対しては有効であったとしても，数十億件以上のデータに対してはスケーラビリティを確保できず，同社のサービスへの適用に耐えられなかったことが，理由の 1 つとも言われている．

Velocity

Velocity という言葉は，データの発生・更新頻度の高さと，即時的な処理の要求という，ビッグデータの 2 つの特性を表現している．たとえば，商品の購買推薦において，これまでバッチ処理で分析され，一定期間固定化されていた推薦ロジックを，メディア等の取り上げによる購買傾向の急激な変化に応じてリアルタイムに変更することや，建物の空調機器や電源装置のモニタリングデータ，あるいは，大規模プラントに張り巡らされたセンサネットワークが数百ミリ秒ごとに生成する多変量データから，即時に異常検知を行う，といったビッグデータの活用方法が挙げられる．

Variety

Variety は，関係モデルでは扱うことが難しいテキスト，画像，音声のような非構造化データ，部分的に構造化された XML のような半構造化データなど，形態の多様性に富んでいることを指す．これらのデータに対しては，たとえばテキストマイニングのように，これまでにも個別にデータマイニングやデータ処理が適

用されてきたが，形態の異なるデータソースを組み合わせ，包括的な分析を行うことがビッグデータ活用における特徴である．必然的に，NoSQL データベースと呼ばれる，従来的な関係データベースとは異なる設計思想を持つデータベースや，データの内容に対して臨機応変な分析手法の選択が求められることになる．

ビッグデータが注目される理由とその効果

なぜ今，「ビッグデータ」なのか?

以前から情報爆発の問題が叫ばれてきたが，なぜ今，これほどまでにビッグデータという言葉が盛り上がりを見せているのだろうか．それは，これまでビッグデータ処理の問題を克服する技術的側面が注目されてきたが，ここ 3, 4 年でビッグデータのビジネスへの活用が技術的，コスト的に可能となり，ベンダ各社が中核事業として一斉に注力し始めたことが背景にある．そして，その最大の牽引役が Hadoop[☆1] である．

Hadoop は Google が 2003 年と 2004 年にそれぞれ論文発表した，分散ファイルシステム Google File System と，並列処理 MapReduce[☆2] のオープンソースクローンである．データの分割方法や分割後の処理内容を指定する Mapper と，分割データに対する処理結果の集約を行う Reducer という 2 つの関数を書くだけで，並列分散処理を効率的に実装できること，安価なコモディティサーバを複数連携させるスケールアウト構成に向いていることから，ビッグデータに対する有効な対策として注目を集めた．SQL と類似した構文で Hadoop 環境を操作できる Hive[☆3] や，Hadoop 上で稼働する機械学習ライブラリである Mahout[☆4] などの，周辺技術の開発も進んでいる．

その一方で，過熱気味に注目を集めたこともあり，Hadoop のような新技術を採用することこそがビッグデータへの対応と誤解されていることが少なくない．

上述の通り，Hadoop の基本概念はビッグデータを多数の小さな塊に分割して各々を処理し，あとで途中結果を集約する MapReduce に基づいている．このため，別ノードの処理結果を利用する必要がある分散性が低い処理や，オーバヘッド時間がかかるためリアルタイム処理には不向きとされる．とはいえ，集計のような単純なバッチ処理を短時間で処理する点で Hadoop が強力な手段であることに違いはない．Hadoop があらゆる問題を解決する魔法の杖でないという認識を共有し，従来的な関係データベースなどが不得手とする分野を Hadoop などで適切に補完していくことが重要である．

ビッグデータ活用への期待

ビッグデータは 2012 年 1 月に開催された世界経済フォーラム年次総会（ダボス会議）でも議題となり，研究者や技術者の枠を越えて世界的に注目を集めている．ビッグデータは一過性のブームとして消え去るのではなく，産業界をはじめとして，実際に世の中に大きな革新をもたらすと期待されている．テキサス大学オースティン校の研究によれば，データの有用性を 10% 高めることができれば，平均的な Fortune 1000 企業の場合，年間収益は 20 億ドル向上すると推計されている[2]．

最近ではビッグデータを活用した成功事例が各種媒体で紹介されることも増えてきた．その多くは，**図 -1** に示すように，利用者数の多い Web サービス企業や，POS データやアクセスログを大量に保有する大手小売業など，ビッグデータが収集される状況がすでに存在している事例である．そこで本節では，ビッグデータとのかかわりが今後増していく，先の展開について述べたい．

(1) M2M がもたらす次なるビッグデータの波

まず，M2M（Machine to Machine）の拡大により，これまで大規模データとのかかわりが深くなかった業界にビッグデータの波がやってくる．M2M とは，人間を介さずに機器間で相互に情報交換して動作制御を行うシステムを指す．たとえば，建築物に巡らされた歪センサのデータから安全性を判定する構造ヘル

☆1 http://hadoop.apache.org/
☆2 データを分解し，必要な情報を整形して出力する処理（Map）と，その出力を集約する処理（Reduce）からなるプログラミングモデル．
☆3 http://hive.apache.org/
☆4 http://mahout.apache.org/

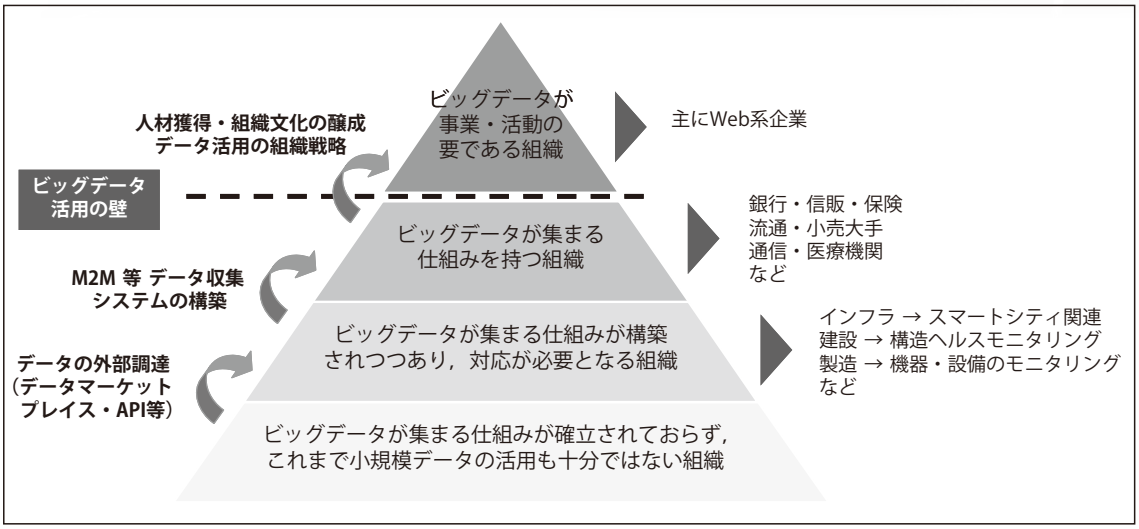

図 -1
ビッグデータの活用段階

スモニタリングや，スマートグリッド（次世代送電網）に接続された需要家の消費状況，気象情報による再生可能エネルギーの出力予測，分散電源の供給能力などに基づいたアンシラリ（系統需給調整）サービスなど，M2M により生成されるデータを活用する事業は今後多く登場していくと言われている．

機械により自動的に生み出すデータ量は，人間の活動により生成されるソーシャルメディアデータやアクセスログ等と比べて遥かに高速に増加する．また，M2M デバイスは 2011 年末の 20 億台から，2020 年には 120 億台にまで急増するという予測もあり[3]，M2M 由来のデータ量は加速的に増加していく．

一方で，例として挙げた建設やインフラ等の業界では，データ解析を積極的に事業の中核として位置付けてきた企業は，これまでのところ多くはない．それゆえ，今後ビッグデータの波と向きあうことになる業界に対して，システムインテグレーションのみならず，データ利活用のコンサルティングや分析代行など，さまざまな事業機会が創出されるに違いない．

(2) データマーケットプレイス

また，これまで小規模なデータさえ十分に活用することがなかった企業では，大量かつ多様なデータを活用可能な状態に整形すること自体が，すでに敷居が高く，複雑な非構造化データに対応するための新規システムに十分な投資ができるとも限らない．これらの企業にとって，既存システムで利用できるように構造化された，あるいは容易に扱えるよう前処理が施された，関心のあるデータだけを外部から調達できれば，余計なコストをかける必要がなく魅力的である．

名　称	概　要
Bluekai Exchange	ターゲティング広告用のクッキー情報売買を仲介
DataMarket	各国の統計や経済指標データの流通に特に注力
Datasift	Twitter のツイートを集め，位置，属性，感情，Klout スコア（影響力）等で選別して提供
Factual	病院，レストランや POI（Point Of Interest）の位置情報を API で提供
Infochimps	CSV 等に整形されたデータの売買を仲介
Kasabi	Linked Data を専門に取扱い
Windows Azure DataMarket	Microsoft のクラウド Azure を介したデータ流通プラットフォーム

表 -1　データマーケットプレイスの一例

今後，このような要求が増えてデータの流通が発展することを見据え，すでに米国を中心に，データマーケットプレイス（データ流通市場）を展開する事業者が登場し始めている．**表 -1** にデータマーケットプレイス事業の一例を示す．データの活用が企業の競争力強化につながることがより明示的になれば，データの流通は活性化していくことだろう．データが流通するようになれば，たとえば，スポーツ用品メーカが，自社で保有するランニングシューズの販売実績データと，データマーケットプレイスから調達した，ユーザ属性が紐付いたランニング記録のデータを掛け合わせ，訴求力が高い新商品の開発や，地域・店舗ごとの緻密な販売戦略につなげるなど，単独では難しかったことが実現できるようになる．

ビッグデータをめぐる事例・動向

民間企業の先行事例

英国 OVUM 社の調査によると，民間企業がビッグデータを活用する目的としては，運用上の意思決定，戦略意思決定，顧客サービス，ビジネス予測が上位に位置付けられている[4]．ビッグデータは，分析手段を変えたとしても，目的まで変えるわけではない．MIT の研究者らによる調査によれば，データに基づく意思決定が企業の業績を平均して 5 ～ 6% 向上させると報告されており[5]，意思決定は民間企業が最も強く関心を寄せている目的である．

たとえば，図 -1 で示した，ビッグデータが集まる仕組みを持つ民間企業の事例としては，Walt Disney 社が挙げられる．同社はテーマパーク入場者，ホテル利用者，ディズニー・チャンネル視聴者の情報などを収集して，サイトの改善や動線分析を行っている．

また，ビッグデータが事業・活動の要である事例としては，小型衛星を利用する米国 Skybox Imaging 社が，多額の投資資金を集め注目されている．同社は，12 個の超小型衛星から高解像度の地球観測画像と映像を取得し，分析するサービスを提供する．年間 11PB に及ぶ地球観測データを用いて，常時情報が更新されるオンラインマップ，農地のモニタリング，豪雨後の河川監視など，複数分野にわたる実践的な事業展開が考えられている．こうしたサービスをできるだけリアルタイムに近い形で実現するために，同社は Hadoop を活用している． C 言語によるライブラリを利用して高速な数値演算を実現するために，Hadoop のタスクとして呼び出せる BusBoy という独自の機構を用意するなど，技術的側面にも注力している．

米国に見る政府系機関の動向

2012 年 3 月に米科学技術政策局 OSTP が公表した Fact Sheet : Big Data Across the Federal Government には，90 件近いビッグデータ関連事業が記載されている．このうち 25 件程度（約 3 割）は，明確な目的を持ちビッグデータを扱う事業であるが，残りについてはビッグデータ処理の支援環境を構築する，基盤側の事業が大半を占める．

明確な目的を持つ事業には The Center of Excellence on Visualization and Data Analytics（CVADA），The Cyber-Insider Threat（CINDER）program などがある．前者は，自然災害やテロ攻撃，サイバー脅威などに対して，ビッグデータを活用して初動の迅速化を図る計画である．後者ではビッグデータを用いて，軍のコンピュータネットワークにおけるサイバースパイ行為を発見する手法の開発を行う．

他の事業については，ビッグデータの高速処理技術の研究に関する委託プロジェクトや，政府系機関が収集しているデータを，研究者が共有・利用しやすくするプラットフォームを構築するプロジェクトが多く，ビッグデータの分析に関しては，第三者に委ねられている場合がほとんどである．

学術（非情報処理）分野の動向

ビッグデータがバズワードになる前から，素粒子物理学，地球科学，天文学，生命科学，気象学，海洋学などの科学分野では，大量の観測データやシミュレーションデータを扱ってきたが，そのデータ量は近年，爆発的に増大している．たとえば，Science 誌に掲載された論文によれば，気象データは 2020 年に 100PB を超え，2030 年には約 350PB に達するとの推計もある[6]．

先取権が重視される科学分野において，ビッグデータの効率的な処理は，いち早く研究成果を上げるために有効である．たとえば，国立天文台は 200 億レコードからなる全天球の観測データの処理を，領域ごとに分割して MapReduce により 70 並列で実行することで，所要時間を 180 日から 3 日まで減らし，研究効率を大幅に向上できることを示した．

一方で，ビッグデータを問題点ではなく，解決策として捉える動きも見られる．チューリング賞受賞者である James Nicholas Gray 氏は，実験科学，理論科学，計算科学（シミュレーション）に次ぐ第 4 の科学的探求のパラダイムとしてデータ集約型科学を提唱した．すなわち，既知の規則やモデルに基づいて演繹的に現象を予測するのではなく，機械学習を駆使し，データから帰納的に潜在的な規則や構造に関する洞

察の獲得を試みるアプローチである.

前述したOSTPのファクトシートでは，ビッグデータを科学的な研究開発に活用するプログラムに約2億ドルを投入することが公表されているが，我が国においても，今後，科学分野でのビッグデータ活用の機運が高まることだろう．しかし，ビッグデータを取り扱うとなると，各専門分野の知識と異なる能力が要求されるため，データ操作や分析処理に長けたデータサイエンティストの貢献が期待される.

ビッグデータ活用の課題

人材の育成と確保

ビッグデータの盛り上がりに対し，適切なデータに対して適切な手法を適用し，解析結果を意思決定へ結びつける人材が不足している．GigaOM社の調査では，50%近くの意思決定責任者は，分析結果を戦略的な意思決定に反映できる担当者が社内に存在しないために，ビジネスインテリジェンス・プロジェクトが失敗に終わったと感じているとの報告がある.

大量データから知見を掘り出せる人材は絶対数が少ない上に，そのような解析者はビッグデータを事業の中核と位置付けている，ごく一部の先行企業に集中しやすい．内部での人材育成・確保が難しい場合，外部からの調達が必要であるが，部外者がデータと接点を持つことへの抵抗感を減らすなど，データ解析に関する組織文化の醸成が必要である．組織文化については，次節でも触れる.

データ解析を意思決定へと結びつけるスペシャリストの育成に関する取り組み事例を挙げると，ノースウェスタン大学マコーミック工学院では，大学院の修士課程としてデータ解析コース（Master of Science Degree in Analytics）を2012年9月から開講する．この課程に対しては民間企業が出資しており，カリキュラムにはSPSSやSASなどの分析ツールを使った実習が含まれる.

また，先に述べた米国政府のビッグデータ関連事業の中には，人材育成に関するプロジェクトも存在する．たとえば，新世代の研究者がビッグデータの技術的課題を解決すべく支援・教育を進めるNSF（米国立科学財団）のCyberinfrastructure Framework for 21st Century Science and Engineering（CIF21）には，約1.2億ドルの予算が配分されている.

Googleのチーフエコノミストである Hal Varian氏が今から3年ほど前に，「今後10年間で最もセクシーな職業は統計家である」と述べている．この発言を信じるならば，データサイエンティストは，あと7年間はセクシーな存在である．データの分析能力を高めておくことで，損をすることはないだろう．ただしその際には，数理科学・統計手法の知識と，それらを実装してビッグデータを解析する技術に加え，仮説を立て，適切な解析手法の選択と分析工程の組み立てを行い，結果から導出された知見を意思決定へ結びつける能力が必要とされていることを，十分に意識すべきである.

データ解析文化の醸成

組織内部でデータ解析に対する深い理解が共有されるための組織文化の醸成も，ビッグデータ活用を進める上で大きな課題である.

一例として，現場への裁量権の付与が挙げられる．仮に先進的なデータ分析基盤を導入し，能力の高いデータサイエンティストを確保したとしても，現場に裁量権がなく，施策の実施や意思決定までに時間がかかっていたのでは，高速処理のメリットをまったく活かすことができない．また，ある仮説に基づき実施されたデータ解析の結果が，最初から有用な知見や高いパフォーマンスの改善をもたらすことは稀であり，通常は仮説構築，データの収集・抽出・解析，業務へのフィードバック，成果の検証，仮説の修正，というサイクルを再帰的に回すことが必要となる．このサイクルをいかに迅速に回せるかが，ビッグデータ活用を成功させる鍵となるが，そのためには，解析担当者と施策担当者が，密に情報を共有できる体制が構築されていなくてはならない.

ビッグデータ時代のプライバシー

DoD（米国国防総省）やDHS（米国安全保障省）に対して，データマイニングの禁止を訴えたデータマ

イニングモラトリアム法案が米国で提出されたのは，2003 年のことであった．監視カメラ映像の解析，個人のインターネット上における行動履歴や位置情報の解析，医療データの解析などの進展により，便利さが提供される反面，プライバシー軽視への警戒や，個人からデータを収集する際にオプトインとすべきかオプトアウトとすべきか☆5 など，高度なデータ解析によるプライバシー侵害への懸念はビッグデータ以前から継承されている．

こうした状況下で期待されるのは，先述のモラトリアム法案の頃から盛り上がりを見せている，プライバシー保護データマイニングの応用である．たとえば，医療の質の向上を目的とした医療データの解析では，カナダの Privacy Analytics 社のように，法令に準拠して利用者別に応じたリスク査定を行うために，匿名化やプライバシー保護データマイニングを実施する企業もある．また，Explorys 社のように，Hadoop を使い，匿名化された医療データの解析を専門とする企業もある．

先述したダボス会議においても，パーソナルデータ・エコシステムと呼ばれる枠組みが議論されている．これは，個人の行動で生み出すデータを構造化して価値を持たせることで，データ流通の活性化や，個人のデータを利活用する新産業の創出が期待される，という議論である．データ開示範囲を自身で制御できれば問題ないという考えもある一方で，巧妙にデータ開示の同意を取られる危険性も孕んでいる．さまざまなデータに価値を見いだそうとするビッグデータの背後には，プライバシーの問題がつきまとうのである．

個人のプライバシーに関するデータを扱いサービスを提供する場合は，サービスの設計段階からプライバシー対策の検討が望まれる．その上で，法令の順守，適切なリスク分析と安全管理措置，およびこれらの方針についての透明性を担保することが必須となる．

☆5 オプトイン（opt in）では個人からデータを取る場合事前に許可を取る．オプトアウト（opt out）では明示的に拒否されない限り許可を取る必要がない．

ビッグデータ時代を勝ち抜くには

先述した 3 つの課題については，ビッグデータに限らず，データ分析を競争力としていく上で，あらゆる場合に当てはまる．特に，これまで積極的にデータ解析に取り組んでこなかった組織においては，手始めに小規模なデータの活用から取り組みをはじめ，組織内におけるデータ解析文化の醸成や，データ分析を意思決定へと結びつけることのできるデータサイエンティスト，プロジェクトマネージャ等の人材育成・獲得に注力することが望ましい．決して世の中を覆うビッグデータの喧伝に判断力を奪われ，目的に適さない IT 投資を行わぬよう注意が必要である．小規模なデータであってもデータ解析文化が醸成された組織は，高機能なツールを保有するが手持ちのデータを活用しきれていない組織と比べて，強い競争力を手にすることだろう．そういった組織がビッグデータを扱うようになったとき，その組織は業界におけるリーダー・ポジションを獲得できるに違いない．人材と組織が整備されてこそ，初めてデータ解析が活かされ，組織の競争力へと繋がるのである．

参考文献

1）McKinsey Global Institute : Big Data : The Next Frontier for Innovation, Competition, and Productivity (2011).
2）Barua, A., Mani, D. and Mukherjee, R. : Measuring the Business Impacts of Effective Data, The University of Texas at Austin (2010).
3）Machina Research : M2M Global Forecast and Analysis 2010-20 (2011).
4）Ovum : Big Data Interest Bubbling under the Surface (Oct. 2011).
5）Brynjolfsson, E., Hitt, L. and Kim, H. : Strength in Numbers : How Does Data-driven Decision-making Affect Firm Performance? in Proc. ICIS 2011.
6）Overpeck, J. T. et al. : Climate Data Challenges in the 21st Century, Science, Vol.331, No.700 (2011).

（2012 年 6 月 5 日受付）

吉田圭吾　keigo@mri.co.jp

2009 年東京大学大学院工学系研究科修士課程修了．同年，（株）三菱総合研究所入社．現在，同社未来情報解析センター研究員．人工知能学会，日本リモートセンシング学会，各会員．

松崎和賢（正会員）kazutaka@mri.co.jp

2007 年東京大学大学院情報理工学系研究科博士課程修了．同年，（株）三菱総合研究所入社．2012 年 6 月より技術研究組合制御システムセキュリティセンターに出向中．博士（情報理工学）．