■特別寄稿論文

ビッグデータとその解析手法

Big Data and Analysis Methods for It

日本経済大学 福島 綾一*
Japan University of Economics Ryoichi FUKUSHIMA
電気通信大学 植野 真臣**
University of Electro Communication Maomi UENO

Abstract : *This article introduces and reviews recent analyses methods on big data. We first introduce several kinds of definitions of big data based on Volume, Variety, Velocity, and Value extracted from data. Next, we describe that big data can be classified into three types – 1. various kinds of data with very large volume or producing very fast, 2. sparse data and 3. universal data –. Then we derive three important factors for utilizing big data – big data technologies, visualization and techniques for analyzing big data – and introduce the details of each factor corresponding to that three types of big data.*

Keywords : *big data, data analysis, data structure, machine learning*

1. はじめに

近年，様々な分野のデータがこれまでにない規模で増加している．米国 IDC 社の調査によると，世界中で生成またはコピーされたデータの総量は 2013 年に 4.4 ゼタバイト（ZB，1ZB は 10 億テラバイト〈TB〉）に上るという．そして 2 年毎に 2 倍以上のペースで増え続け，2020 年には 44ZB になると予測している（IDC/EMC, 2014a）．日本国内に限ると 2014 年に 495 エクサバイト（EB，1EB は 100 万 TB），2020 年には 2.2ZB になると予測している（IDC/EMC, 2014c）．

一方で 2013 年の全世界のデータのうち，解析対象となりえるものは 22% と見込まれたものの，実際に解析されたのは全体の 5% に留まったという（IDC/EMC, 2014b）．

質，量共に大きく変化しているデータを解析するには，従来とは異なる基盤技術や解析手法が必要となる．

米国 Google 社は 2003 年，2004 年，2006 年と分散ファイルシステムの Google File System，分散アプリケーション処理モデルの MapReduce，分散データベースシステムの Bigtable を相次い

＊日本経済大学 経営学部 経営学科 専任講師
＊＊電気通信大学大学院 情報システム学研究科 教授

で発表した（Chang et al., 2008; Dean & Ghemawat, 2008; Ghemawat et al., 2003）．これは自社の検索エンジンを支える巨大データ群の管理基盤の公開でもあった．

オープンソース・ソフトウェアプロジェクト支援団体の Apache ソフトウェア財団は，これらの発表に触発され，分散ファイルシステムと Map Reduce を実装した Hadoop[1]の最初のバージョンを 2007 年に，Bigtable をモデルとした HBase[2]を 2008 年にリリースした．これによって膨大なデータであってもフリーソフトウェアで処理できるようになり，国内外で導入が進められた．

そして 2010 年頃よりこの膨大なデータに対して「ビッグデータ」という語句が当てられ，技術革新と競争，生産性の次のフロンティアとして注目されるようになった．

本稿ではまず「ビッグデータ」の定義を再確認し，データサイズが巨大なもの＝ビッグデータではないことを示した上で，「ビッグデータ」と呼ばれるものに 3 体様あることを示す．次にこの 3 体様とビッグデータ利活用で重要となる 3 要素の関係に基づき，ビッグデータ工学，データの可視化，ビッグデータ解析手法について述べる．

2.「ビッグデータ」の定義

ビッグデータは 3 つの V，すなわち Volume（データ量），Variety（データの多様性），Velocity（データの発生速度，処理速度）もしくは，これに Value（経済的価値）を加えた 4 つの V を伴うものと受け止められている．

前者の 3V の考え方は，これらを三次元的に管理することで情報資産の価値が増大するとした米国 META Group 社（現 米国 Gartner 社）のレポート（Laney, 2001）を援用したものである．

後者の 4V は，2011 年の米国 IDC 社の定義によるもので，*Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.* と，3V に経済的価値の抽出を加えてビッグデータ技術として捉えている（Gantz & Reinsel, 2011）．

以下の項でそれぞれについて述べる．

2.1　Volume（データ量）

Apache Hadoop Project がビッグデータを「一般的なコンピュータでは収集や管理，処理がしきれないデータセット」としたことを受け，米国 McKinsey 社は *"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.* と 2011 年の自社レポートで定義した（Chen et al., 2014; Manyika et al., 2011）．このレポートでは，データサイズについて *big data in many sectors today will range from a few dozen terabytes to multiple petabytes* と述べている．

ビッグデータのサイズに言及する時，この一文を引用して数十 TB から数ペタバイト（PB，1PB は 1,000TB）とすることが多い．しかし，McKinsey 社は自ら意図的な主観に基づいて定義していると述べており（*This definition is intentionally subjective*），何 TB 以上だからビッグデータであるとは定義しない（*we don't define big data in terms of being larger than a certain number of terabytes*）としている．

Web 2.0 の提唱者の一人である O'Reilly は，企業において既存のデータ管理システムや解析システムでは対応しきれず，システム刷新の決断を迫られた時のデータをビッグデータだと定義した（Reilly et al., 2009）．このため数百 GB のデータがビッグデータになる企業もあれば，数百 TB に至ってビッグデータとなる企業もあるとしている．

また，朝野は多変量データとの比較からビッグデータは欠測値だらけのスパース（sparse）な行列（疎行列）であると述べている（朝野，2014）．

このビッグデータのスパース性については，樋口が大手総合スーパーマーケットの ID 付き POS

データを例に挙げている．［商品，消費者，時刻］で構成される 3 次元空間に［値段，個数］のデータが埋め込まれると考えると，時刻を日単位としてもそのデータは 50,000 商品×2,000 万人×365 日といった規模になる．反面，個々の消費者に注目すると 1 年間で 50,000 商品中 500 商品も購買することはなく，毎日スーパーに行く顧客も限られることから，この 3 次元空間はほとんど欠測になると指摘している（樋口，2013）．この場合，従来の統計解析のようにサンプル数を増やしてもスパースなデータが増えるだけで，個々のデータの欠測値が埋まるわけではない．このため，新たな解析手法が必要となる．

Mayer-Schönberger らは，*ビッグデータは，必ずしも絶対数で「ビッグ」である必要はない．結果としてビッグになりやすいだけだ*と述べている．またデータ集合の規模が大きいだけでは，ビッグデータとは言えないとし，無作為抽出標本のような簡便法を使うのではなく，データ全体を解析対象にすることがビッグデータの条件だとしている（Mayer-Schönberger & Cukier, 2013）．この点は情報通信白書〈平成 25 年版〉においても「悉皆に近い大規模性」と表している（総務省，2013）．

では，実際にデータ解析者が携わっているデータサイズはどうだろうか．ビッグデータやデータ解析に関するポータルサイトの KDNuggets では 2015 年 8 月，データ解析者を対象にこれまで解析した最大のデータサイズのアンケート調査を実施した．459 人の回答のうち，11TB 以上の解析経験がある者は 11.5%に留まる一方，最頻値は 1.1～10GB（19.6%）となっている．2013 年，2014 年の調査と比較してもこうした傾向は変わらず，半数以上が 1.1GB～1TB の範囲の解析を行っていることがうかがえる（Piatetsky, 2015）（図 1）．

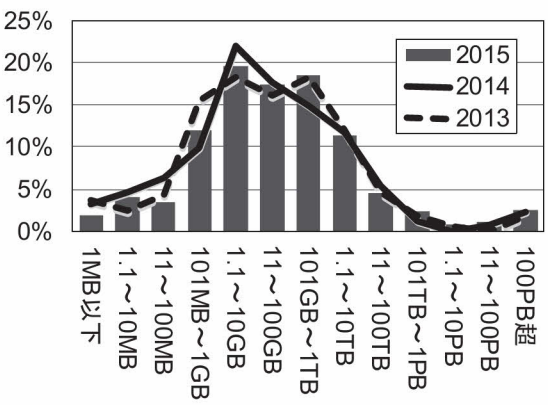

図 1　KDnuggets 2015 Poll: Largest Dataset Analyzed
（（Piatetsky, 2015）より筆者作成）

2.2　Variety（データの多様性）

ビッグデータはデータの量のみでなく，データの多様性，すなわち，データの種類の多さもその特徴となる．データの種類は，以下の 3 種に分けることができる．

1. 構造化データ：リレーショナルデータベースのように，各項目の型やサイズが厳密に定義され，行列で表現できるデータ．経理データや POS データ，アクセスログなど．
2. 半構造化データ：厳密な定義がない，もしくは行列では表現できないものの XML で表現できるデータ．IoT（Internet of Things）による GPS データや RFID データ，センサーデータなど．
3. 非構造化データ：映像，画像，音声や，ブログ記事や Twitter のつぶやきのようなテキストデータ．

従来のデータ処理では，構造化データのみを扱うことが多かったが，構造化データのみでなく，半構造化データ，非構造化データを含んだ多様なデータであることもビッグデータの定義に含まれている．

特に近年のデータ量の爆発的な増加は，半構造化データ及び非構造化データの伸びによるものである．国内の 2014 年の構造化データの流通量は 2010 年比で 1.9 倍に留まるのに対し，半構造化データは 2.8 倍，非構造化データは 3.3 倍となっている（株式会社情報通信総合研究所，2015）（図 2）．

2.3　Velocity（データの発生速度，処理速度）

対象とするデータが多様化する中で，データの発生速度もまた急激に上昇している．わずか60秒の間にTwitterでは423,000ツイートが投稿され，Googleは312万回検索されている．また電子メールは1億4千7百万通送信されている[3)]．

当初は従来のコンピュータ技術では処理しきれなくなった規模のデータ解析としてMapReduceによるバッチ処理が行われてきたが，センサー情報を用いた異常検知を行う場合にはリアルタイム性が求められる（Chen et al., 2014; Philip Chen & Zhang, 2014）．

2.4　Value（解析結果の価値）

Brynjolfssonらはデータを重視した意思決定をしている企業が，そうでない企業と比べて生産性が5〜6%高いことを示した（Brynjolfsson et al., 2011）．

総務省（2013）はビッグデータ分析による成果として分析結果入手の時間短縮，モデルの精度向上，リアルタイムでの状況把握を挙げている．また，その結果として企業や社会において意思決定の高度化，意思決定の迅速化，業務実行の精度向上，業務改革の迅速化，新規商品・サービスの投入といった効果が現れるとしている．

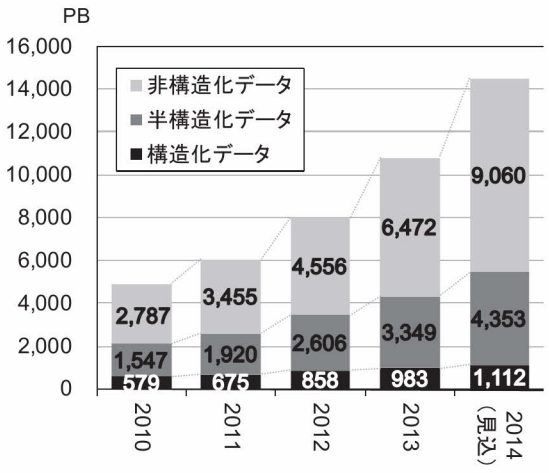

図2　国内データ流通量の推移
（株式会社情報通信総合研究所（2015）より筆者作成）

2.5　ビッグデータの3体様と利活用のための3要素

これらのことから，以下の3つのいずれかに該当するものが「ビッグデータ」であると言える．そしてその内容は構造化データから非構造化データまで多岐にわたる．

① 従来の技術や設備では処理しきれないほど膨大，または高速に生成される多様なデータ
② 従来の統計解析では対応できないスパースなデータ
③ 標本ではなく，母集団そのものであるデータ

McKinsey社はビッグデータの利活用で重要な手法・技術としてビッグデータ工学，データの可視化，ビッグデータ解析手法の3つを挙げている（Manyika et al., 2011）．先の3点をこの3要素に当てはめると図3となる．これらについて次章以降にて述べる．

3. ビッグデータ工学

「ビッグデータ工学」とは膨大なデータを集約，操作，管理，分析するための一連の技術を指す．本章では，その中核をなすオープンソース・ソフトウェア（OSS：Open Source Software）であ

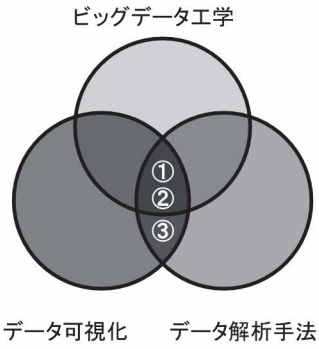

図3　ビッグデータの3体様と利活用のための3要素との関係
（Manyika et al.（2011）より筆者作成）

るHadoopと，Hadoopを取り巻く一連のソフトウェアやツール群について紹介する．

3.1 Hadoop及びHDFS, MapReduce, YARN

Hadoop（Hadoop 1.0）は2007年にリリースされた並列分散処理のミドルウェアであり，Hadoop 2.0は2012年に開発が始まった後継バージョンである（Vavilapalli et al., 2013）．

Hadoop 1.0はHDFS（Hadoop Distributed File System）と，MapReduceが強固に結び付いた形で実装されていた．

HDFSは大規模なコンピュータ群を一つの巨大なストレージとして扱う分散ファイルシステムであり，各ノードに重複してファイルの断片を保存することでフォールトトレラントかつスケールアウトを可能にしている．

MapReduceはクラスタ資源の管理と，分散並列にバッチ処理を行うデータ処理機構の両方を担う．

MapReduceのデータ処理機構は，入力データを変換するMapフェーズと変換データを基に結果を出力するReduceフェーズに大別され，それぞれのフェーズを多段に組み合わせていくことで最終的な結果を得ることができる．例えば，Googleは検索エンジンのインデックス更新処理に100段のMapReduce処理を行ってきた（Peng & Dabek, 2010）．

その後，MapReduceが担っていたクラスタ資源の管理機構をYARN（Yet Another Resource Negotiator）へと切り離したHadoop 2.0がリリースされた（Vavilapalli et al., 2013）．

これにより1万ノードを超えるクラスタや，バッチ処理以外の処理にも柔軟に対応できるようになり，それまでMapReduce上で動作していたアプリケーションやミドルウェアは，YARN上の別の分散並列処理エンジン上でより効果的に動作するようになった（図4）．

3.2 Hadoopスタック

前述の通りHadoopはミドルウェアのため，実際のMapReduce処理に当たってはJava言語によるプログラミングが必要であった．その生産性向上のため，独自のスクリプト言語（Olston et al., 2008）によってデータフローを指示したり，SQLライクな言語（Thusoo et al., 2009）をMapReduce処理に変換したりする仕組みが提案された．

Hadoop 2.0では前述の通りMapReduceに代わる分散並列処理エンジンが開発され，そのエンジン上で動作するコンポーネントも開発されている．

こうした各コンポーネントの重なりをHadoopスタックもしくはHadoop技術スタックと呼ぶ（図5）．

Hadoopスタックの個々の要素はほぼOSSで構成されており，それぞれ頻繁に更新されている．各コンポーネントのバージョンや依存関係に留意しながらセットアップするのは煩雑なため，主要なコンポーネント及びHadoopと連携するツールをまとめたパッケージ（ディストリビューション）が複数の企業から有償，無償で配布されている．

前述のHDFS，YARN，MapReduce以外の主なコンポーネントは以下の通りである．

3.2.1 分散並列処理エンジン

MapReduce以外の処理を，分散並列で行うためのミドルウェアである．バッチ処理や対話処理に対応するTez[4]，インメモリ処理基盤のSpark[5]，ストリーム並列処理を行うStorm[6]などがある．

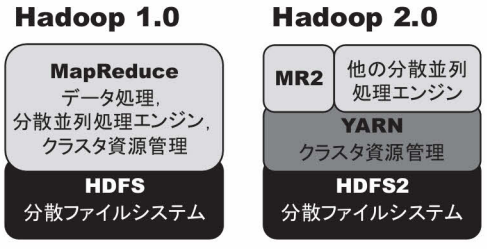

図4 Hadoop 1.0と2.0とのシステム構成の違い
（Murthy（2013）より筆者作成）

3.2.2　スクリプト処理

独自のスクリプト言語でデータフローを記述すると，MapReduce 用のプログラムコードに変換して実行する Pig[7]，MapReduce 処理を独自の API で隠蔽し，Java 以外のプログラミング言語で実行できるようにする Cascading[8]がある．

3.2.3　SQL，リアルタイム SQL

SQL に似た言語を用いて指示をすると，内部で Java プログラムに変換して実行する Hive（Hive on Tez)[9]，HDFS や HBase に対してリアルタイムに SQL を実行できる Impala[10]，Presto[11]がある．また，Spark と Drill[12] は CSV ファイルや JSON ファイルのように，構造化データと準構造化データの混在環境下でもリアルタイムに SQL が実行できるようになっている．

3.2.4　NoSQL

NoSQL（Not Only SQL）とはリレーショナルデータベース以外のデータベースの総称である．NoSQL のデータモデルは大きく 4 つに分かれ，それぞれに対応したシステムがある．

1. キー・バリュー型：プログラミング言語における連想配列のように Key と Value のペアを単位としてデータを格納する．代表的なシステムに redis[13]，Cloud Datastore[14]がある．
2. 列指向型：リレーショナルデータベースが行を単位とした表構造のデータモデルなのに対し，列指向型は文字通り列を単位とした表構造となっている．列単位の集計や更新が素早く行え，データ圧縮効率も高い．Cassandra[15]，HBase がこれに当たる．
3. ドキュメント指向型：JSON や XML といったファイルがデータの単位になり，リレーショナルデータベースのようなスキーマ定義をせずに利用できる．主なシステムとして MongoDB[16]，CouchDB[17]が挙げられる．
4. グラフ指向型：最短経路問題などグラフ問題に特化したデータベースである．Neo4j[18] はグラフ指向型データベースの中で圧倒的な人気を誇っている[19]．

3.2.5　機械学習

機械学習コンポーネントは，Hadoop 1.0 の時から MapReduce 上で動作してきた Mahout[20]，Spark 上で動作する MLlib[21]，NTT ソフトウェアイノベーションセンタが開発に携わっていた Jabatus[22] などがある．

3.2.6　ストリーム処理

5.1 節にて後述する．

図 5　Hadoop スタックの概念図
（(Page, 2014；日経コンピュータ，2015）より筆者作成）

3.2.7 検索

Solr[23]は，Hadoopクラスタ上でも稼働する全文検索エンジンである．また，HDFSでは一般的なファイルシステムのようなファイル検索が行えないため，Solrで作成されるインデックスが用いられる．

3.2.8 グラフ処理

ソーシャルネットワークなど要素同士の関係がグラフ表現されるデータに対してグラフのまま処理をするコンポーネントとしてGiraph[24]，Spark上で動作するGraphX[25]がある．

4. データの可視化

データの可視化は，解析結果を理解したりさらに改善したりするための図表や画像，動画を作成する技術である（Manyika et al., 2011）．日本学術会議の提言では，より具体的に次元圧縮，特徴抽出や画像処理などを含むとしている（日本学術会議，2014）．

また実務の面から，データの性質や特徴が明らかではない場合，その把握のために可視化に重点を置いて探索的にデータを解析していくことも重要であるとの指摘がある（あんちべ，2015）．

データの可視化に当たってはMicrosoft社のExcelの利用の他，BI（Business Intelligence）ツールが用いられる．

5. ビッグデータ解析手法

ビッグデータの解析に当たっては，従来のデータマイニングやマーケティング手法の応用に加え，ビッグデータ特有の問題に対応した手法が提案されている．

本章では2.5節で定義したビッグデータの3体様（膨大または高速に生成される多様なデータ，スパースなデータ，母集団そのものであるデータ）に対する解析手法について述べる．なお「膨大なデータ」についてはビッグデータ工学の範疇であるため本章では触れない．

5.1 高速に生成されるデータへの対応

MapReduceは大規模データの並列分散処理を目的としていることから，一連の処理の途中で一時ファイルの読み書きが発生する仕組みになっている．このため高頻度に生成されるデータを処理するのには適していない．

Sparkのストリーム処理用フレームワークであるSpark Streamingは，MapReduceでの一時ファイルを効率的に各ノードのメモリに格納するResilient Distributed Dataset（RDD）（Zaharia et al., 2012）と，それらの処理をコントロールするDiscretized stream（D-stream）（Zaharia et al., 2012）アルゴリズムにより0.5〜2秒程度のバッチ処理に細かく分解することで高頻度データに対応している．

生成されたデータを1件ずつ処理する本来の意味でのストリーム処理ではないが，複数個所で同時に障害が発生しても並列に復旧作業が行えたり，全体のレイテンシを低下させないようにスループットの遅いノードを検知したりする仕組みを持っている．

一方，Twitter社がOSSとして公開し，後にApacheソフトウェア財団のプロジェクトになったStorm（Toshniwal et al., 2014）はストリーム処理に特化した実行環境である．

Stormでは入力データを1件ずつ高速に処理することができるようになっている．しかし多段階に組み合わせた際に下流の処理速度が上流に比べて遅いと処理待ちのプロセスであふれてしまうため設計段階から十分注意しなければならなかった．

Twitter社ではこうした問題を解決するために後継システムのHeronを開発し，社内のStormクラスタを全てHeronに置換えたと発表した（Kulkarni et al., 2015）．

5.2 スパースなデータへの対応

スパースなデータでは，そのデータの特徴（p）の数が得られるサンプル数（n）に比べ非常に大きく，$p<n$であることを前提とする従来の統計

解析では対応できない.

またデータの次元数は，計算にかかる時間に指数的に影響を及ぼすため，仮に分散並列処理にて解析するにしてもある程度次元を落とす必要がある. そのための手法として特徴選択と次元削減(次元圧縮）がある.

5.2.1　特徴選択

特徴選択では，特徴量の中から有用なものだけを選び，その他のものは削除する．具体的な手法として凸最適化問題である L1 正則化法（Lasso）(Tibshirani, 1996) 及び L2 正則化法 (Ridge 回帰) (Hoerl & Kennard, 2000)，ランダムに複数の決定木を生成し，多数決によって有効な特徴量を決定するランダムフォレスト（Svetnik et al., 2003）などがある.

5.2.2　次元削減（次元圧縮）

次元削減はできるだけ元の情報量を損なわないように特徴量を合成し，低次元のデータに変換する．主成分分析で知られる Karhunen - Loève 展開や，線形判別分析（Fisher, 1936）などが用いられる.

5.3　母集団そのものであるデータへの対応

従来の統計解析では，得られたデータは何らかの分布に従った集団（母集団）の部分集合として捉え，部分集合から母集団を推定してきた.

一方，母集団もしくは母集団とみなせるだけの大量のデータを解析できるのであれば，その結果から規則性や判断基準を見出し，未知の情報に対する予測が可能となる.

5.3.1　教師あり学習

教師あり学習は機械学習のうち，入力データ x に対応した出力（正答）y の組み合わせを事前に学習しておき，未知の x を与えた際に対応する y を予測する手法である.

人間の神経細胞を模したプログラム（ニューロン)を複数組み合わせたニューラルネットワーク，ニューラルネットワークをさらに階層的に積み重ねたディープラーニング，ベイズ分類器などが知られている.

ディープラーニングは Google が猫の画像の自動判別ができるようになったと発表したことでよく知られるようになった（Le, 2013）.

この研究では 1,000 万枚の画像を 3 日間かけて 1,000 台のコンピュータに学習させている．まさにビッグデータ時代だからこそできた解析手法と言える.

5.3.2　教師なし学習

教師なし学習は，（大量の）入力データからコンピュータ自身が規則性を見出す手法である.

ベイジアンネットワーク，Latent Dirichlet Allocation（LDA）などのクラスタリングやグラフィカルモデリング手法が典型である.

ベイジアンネットワークは個々の事象を矢印で結び, それぞれの関係を条件付確率で表す（植野, 2013）．矢印の向きが因果を表しており，確率的に因果関係を説明できる.

ベイジアンネットワークでの解析結果を用いることで，期待効果の最大化や情報量最大化を達成できる．これにより不確定な状況から最善の行動を選択することができると期待されている.

6.　おわりに

本稿ではビッグデータについて複数の定義があることを示し，それらを 3 体様として整理した. そしてその 3 体様とビッグデータ利活用に重要な 3 要素との関係に基づいてビッグデータの解析手法について解説した.

データ解析に当たってまず重要なのは目的の設定である（例えば，あんちべ（2015）；日経コンピュータ（2015））．その上で目的に沿ったデータ収集と解析を行うことになる.

またデータ解析には広範な知識が不可欠である．データサイエンティスト協会は 2015 年 11 月にデータ解析者（データサイエンティスト）が備えるべきスキルセットを公開した（一般社団法人

データサイエンティスト協会，2015）．これによるとデータ解析手法だけでなく，プログラミングやシステム構築，企業経営についての知識も求められていることが分かる．

一方，日本情報経営学会では，2011 年に河本らがデータから情報を形成するプロセスを体系化し，そのプロセスに対する課題解決ミッションとその推進体のあり方，必要な人材等について提案している（河本・細川ほか, 2011）．この提案はビッグデータ時代にも適用できる考え方というよりも，むしろこのような時代にこそ必要なものであると考える．

データ解析ソリューション提供事業者やデータ解析ソフトウェア販売会社の方に話を聞くと，異口同音に経営者の中には「データを用意してマウスをクリックするだけで素晴らしい結果が得られる」と考えている人が少なからずいるという．

本稿がそうした誤解を解く一助となり，ビッグデータ時代の企業，個人の取り組みの参考となることを期待して本稿の結びとする．

注

1）Apache Hadoop: http://hadoop.apache.org/
2）Apache HBase: http://hbase.apache.org/
3）Internet Live Stats（http://www.internetlivestats.com/）より
4）Apache Tez: https://tez.apache.org/
5）Apache Spark: http://spark.apache.org/
6）Storm: http://storm.apache.org/
7）Apache Pig: http://pig.apache.org/
8）Cascading: http://www.cascading.org/
9）Apache Hive: https://hive.apache.org/
10）Apache Impala（旧 Cloudera Impala）: http://www.cloudera.com/content/www/en-us/products/apache-hadoop/impala.html
11）Presto: https://prestodb.io/
12）Apache Drill: https://drill.apache.org/
13）redis: http://redis.io/
14）Cloud Datastore: https://cloud.google.com/datastore/
15）Apache Cassandra: http://cassandra.apache.org/
16）MongoDB: https://www.mongodb.org/
17）Apache CouchDB: http://couchdb.apache.org/
18）Neo4j: http://neo4j.com/
19）DB-Engines Ranking of Graph DBMS. http://db-engines.com/en/ranking/graph+dbms
20）Apache Mahout: http://mahout.apache.org/
21）MLlib: http://spark.apache.org/mllib/
22）Jabatus: http://jubat.us/ja/
23）Apache Solr: http://lucene.apache.org/solr/
24）Apache Giraph: http://giraph.apache.org/
25）GraphX: http://spark.apache.org/graphx/

参考文献

IDC/EMC（2014c）「デジタルユニバースの機会」IDC/EMC Report
https://japan.emc.com/collateral/analyst-reports/idc-digital-universe-2014-japan.pdf（2015 年 12 月 1 日）．

朝野熙彦（2014）『ビッグデータの使い方・活かし方―マーケティングにおける活用事例』東京図書．

あんちべ（2015）『データ解析の実務プロセス入門』森北出版．

植野真臣（2013）『ベイジアンネットワーク』コロナ社．

河本薫・細川嘉則・河村真一・野波成・岡村智仁・大西道隆・津崎賢治・小林宏樹・三上彩（2011）「企業においてデータからの情報形成力を強化するのに必要なミッションと推進体のあり方」『日本情報経営学会誌』Vol. 31，No. 3，pp. 32-40.

株式会社情報通信総合研究所（2015）『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 報告書』
http://www.soumu.go.jp/johotsusintokei/linkdata/h27_03_houkoku.pdf（2015 年 12 月 1 日）．

一般社団法人データサイエンティスト協会（2015）『データ社会に求められる新しい才能とスキル』
http://www.slideshare.net/DataScientist_JP/ss-55326920（2015 年 12 月 1 日）．

総務省（編）（2013）『情報通信白書〈平成 25 年版〉』日経印刷．

日経コンピュータ（編）（2015）『すべてわかるビッグデータ大全 2015－2016』日経 BP 社．

日本学術会議（2014）『ビッグデータ時代に対応する人材の育成』
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t198-2.pdf（2015 年 12 月 1 日）．

樋口知之（2013）「データ・サイエンティストがビッグデータで私たちの未来を創る」『情報管理』Vol. 56，No. 1，pp. 2–11.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H.（2011）“Strength in Numbers: How Does Data-Driven Deci-

sionmaking Affect Firm Performance?," *SSRN Electronic Journal*.
http://doi.org/10.2139/ssrn.1819486

Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., & Gruber, R. E. (2008) "Bigtable," *ACM Transactions on Computer Systems*, Vol. 26, No. 2, pp. 1-26.

Chen, M., Mao, S., & Liu, Y. (2014) "Big Data: A Survey," *Mobile Networks and Applications*, Vol. 19, No. 2, pp. 171-209.

Dean, J., & Ghemawat, S. (2008) "MapReduce: Simplified Data Processing on Large Clusters," *Communications of the ACM*, Vol. 51, No. 1, pp. 107-113.

Fisher, R. A. (1936) "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, Vol. 7, No. 2, pp. 179-188.

Gantz, B. J., & Reinsel, D. (2011) "Extracting Value from Chaos," *IDC iView*, Vol. 1142, pp. 9-10.

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003) "The Google File System," *ACM SIGOPS Operating Systems Review*, Vol. 37, No. 5, pp. 29.

Hoerl, A. E., & Kennard, R. W. (2000) "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, Vol. 42, No. 1, pp. 80.

IDC/EMC (2014a) "The Digital Universe of Opportunities—Report," IDC/EMC Report, http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf.

IDC/EMC (2014b) "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things," http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Kulkarni, S., Bhagat, N., Fu, M., Kedigehalli, V., Kellogg, C., Mittal, S., Patel, J. M., Ramasamy, K., & Taneja, S. (2015) "Twitter Heron: Stream Processing at Scale," *In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 239-250.

Laney, D. (2001) "3D Data Management: Controlling Data Volume, Velocity and Variety," *META Group Research Note*, Vol. 6, No. 70.

Le, Q. V. (2013) "Building High-level Features Using Large Scale Unsupervised Learning," *In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8595-8598). IEEE.
http://doi.org/10.1109/ICASSP.2013.6639343

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011) "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute.

Murthy, A. (2013) "Apache Hadoop 2 is now GA!," http://hortonworks.com/blog/apache-hadoop-2-is-ga/ (2015年12月1日).

Mayer-Schönberger, V., & Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Eamon Dolan/Houghton Mifflin Harcourt (斎藤栄一郎訳 (2013)『ビッグデータの正体 情報の産業革命が世界のすべてを変える』講談社).

Olston, C., Reed, B., Srivastava, U., Kumar, R., & Tomkins, A. (2008) "Pig Latin," *In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data - SIGMOD '08, p. 1099*.

Page, B. (2014) ""YARN Ready" —Accelerating the Adoption of Enterprise Hadoop." http://hortonworks.com/blog/yarn-ready-accelerating-adoption-enterprise-hadoop/ (2015年12月1日).

Peng, D., & Dabek, F. (2010) "Large-scale Incremental Processing Using Distributed Transactions and Notifications," *In Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation, Vol. 10, pp. 1-15*.

Philip Chen, C. L., & Zhang, C.-Y. (2014) "Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data," *Information Sciences*, Vol. 275, pp. 314-347.

Piatetsky, G. (2015) "Poll Results: Where is Big Data? For Most, Largest Dataset Analyzed is in Laptop-size GB Range."
http://www.kdnuggets.com/2015/08/largest-dataset-analyzed-more-gigabytes-petabytes.html (2015年12月1日).

Reilly, T. O., Winge, S., & Schneider, S. (2009) *Big Data: Release 2.0 Issue 2.0.11* O'Reilly Media, Inc.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003) "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.," *Journal of Chemical Information and Computer Sciences*, Vol. 43, No. 6, pp. 1947-58.

Thusoo, A., Sarma, J. Sen, Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., & Murthy, R. (2009) "Hive," *Proceedings of the VLDB Endowment*, Vol. 2, No. 2, pp. 1626-1629.

Tibshirani, R. (1996) "Regression Shrinkage and Selec-

tion via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, pp. 267-288.

Toshniwal, A., Donham, J., Bhagat, N., Mittal, S., Ryaboy, D., Taneja, S., Shukla, A., Ramasamy, K., Patel, J. M., Kulkarni, S., Jackson, J., Gade, K., & Fu, M. (2014) "Storm@twitter," *In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data - SIGMOD '14, pp. 147-156*.

Vavilapalli, V. K., Seth, S., Saha, B., Curino, C., O'Malley, O., Radia, S., Reed, B., Baldeschwieler, E., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., & Shah, H. (2013) "Apache Hadoop YARN," *In Proceedings of the 4th annual Symposium on Cloud Computing - SOCC '13, pp. 1-16*.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S., & Stoica, I. (2012) "Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing," *In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation*.

Zaharia, M., Das, T., Li, H., Shenker, S., & Stoica, I. (2012) "Discretized Streams: An Efficient and Fault-tolerant Model for Stream Processing on Large Clusters," *In Proceedings of the 4th USENIX Conference on Hot Topics in Cloud Computing, pp. 10*. USENIX Association.