

ビッグデータ時代のビジネス・インテリジェンス ～次世代ビジネス・インテリジェンス～

桑田 修平 ((株) NTT データ) 中川 慶一郎 ((株) 数理システム)

概要 大量のデータと Hadoop を用意さえすれば、利益につながる有用な情報が抽出できるようになると思われている風潮が少なからず存在する。“ビッグデータ”というバズワードに惑わされ、具体的な活用シーンをイメージできていない場合に最初に陥りやすい落とし穴の 1 つである。その一方で、早い段階から戦略的にビッグデータ活用に取り組んだ企業は、的確な状況把握と迅速な意思決定が可能となり、経営メリットや競争優位性を顕在化させ始めている。本論文では、これから本格化するビッグデータ時代を迎えるにあたって、鍵となるアプローチや基盤技術を提案し、その有効性を示す 2 つの先進的事例を紹介する。

1. ビッグデータ時代の到来

企業内外に蓄積されたデータを分析・活用する取り組みは、“ビッグデータ時代”以前から数多く行われてきており、ビジネスの世界では、“ビジネス・インテリジェンス (BI)”というキーワードで認知されてきた。ここで、BI とは、「企業内外に散在する膨大なデータを分析して、経営意思決定に活用する IT システム、取り組み、方法論、管理手法を総称するコンセプト」と我々は定義している[1]。これまでの BI が対象としてきたデータは、例えば、商品の在庫状況や POS データといった、比較的小規模な数値データであり、主に、経営判断やマーケティングなどに活用されてきた。

これに対して、“ビッグデータ時代”をむかえた現在では、これまで扱ってこなかったようなデータが活用の対象に含まれるようになった(図 1 参照)[2]:

- ライログ・データ

ブログや twitter での発言や、EC サイトなどの購買履歴・検索履歴などの、個人の行動履歴情報

- センシング・データ

RFID や各種センサ等から送られてくる、人やモノの位置情報や加速度、変位などの物理量

これらのデータが、いわゆる“ビッグデータ”である。センサやネットワーク、データベースなどの基盤技術の発展によって、様々なデータが容易に取得可能となり、大量に蓄積したデータの活用が模索され始めた。現在、これらのビッグデータを活用しようという機

運がビジネス・技術両面で急速に高まっている。

ここで、ビッグデータの特徴を以下に整理しておく:

- データのサイズ:

計算機 1 台のメモリには載りきらないサイズのデータを扱うことになる。そのため、スタンドアロン型の既存の分析ツールは利用できない。現状では、テラバイト級のデータを対象とするケースが多い。

- データの種類:

数値データに加えて、テキストデータも活用の対象となる。つまり、構造化データと非構造化データの両方を上手に活用できる仕組みや方法論が必要となる。現状は、数値データのみを対象とするケースが多い。

- データの到着間隔

データごとにそれが獲得される間隔は様々であり、

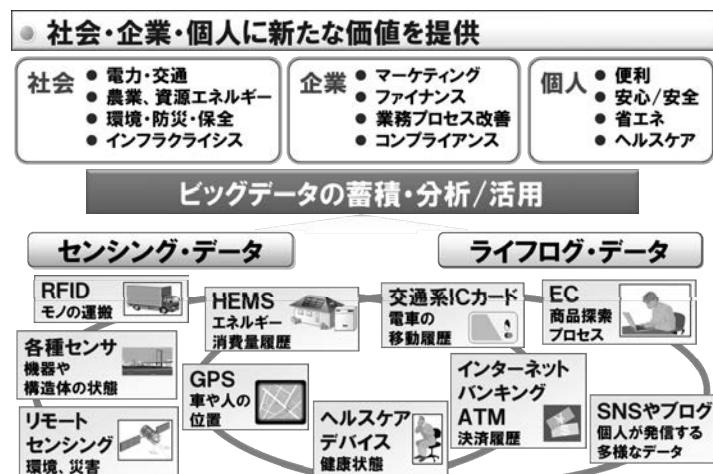


図 1 ビッグデータ活用への期待

情報の“鮮度”に合わせた処理が必要となる。例えば株価情報などのようにミリ秒単位で収集されるデータもあれば、顧客情報などのように月に1回届くデータもある。

これまでにないボリュームと多様性を持つビッグデータを上手に活用することができれば、限られたリソースを有効利用したり、変化の兆候を察知して機先を制したり、より魅力的なサービスを提供したりと、社会・企業・個人それぞれにメリットをもたらすことができる：

- ・社会…限られたリソースを最適に配分することで、社会全体を効率的に運営可能となる。
- ・企業…経営状況の見通しが難しい現状において、あらゆる環境変化に対し即時対応可能となる。
- ・個人…ユーザの行動に沿った、気の利いたサービスが適切なタイミングで提供可能となる。

しかし、「従来のデータベースでは蓄積しきれない」、「データの発生に処理が追いつかない」といったことなどから、膨大な情報量を前に手をこまねいているのが多くの企業の現状である。

2. ビッグデータ活用の切り口

「ビッグデータを活用したいと考えているが、具体的な活用シーンがイメージできていない」という課題に対する我々の解が、“次世代ビジネス・インテリジェンス”である。次世代BIとは、ビッグデータを分析・活用するための、我々が提唱する、ビッグデータ時代に対応したビジネス・インテリジェンスである。

次世代BIによって、ビッグデータを新たな競争優位性の源泉とすることが可能となる。そこでは、データそのものやHadoopなどの基盤技術は、次世代BIを実現するための手段として利用される。あくまで、目的を達成するためにビッグデータを活用するのであり、ビッグデータがあるから何かをしようと考えると、途端に検討が進まなくなる。当たり前であるが重要なポイントである。

ここで、我々は、これまでに数多くの分析コンサルティングやDWH（データウェアハウス）の開発・運用を手がけてきた。そこでの実績やノウハウをもとに、BIを独自に4つのタイプ（集計分析型BI、発見型BI、WHAT-IF型BI、プロアクティブ型BI）に分類している[1]。データ活用の仕方をタイプ分けしておくことで、現状の課題や目的に合致した分析シナリオを速やかに実施することが可能となる。4つのBIの概要を以下に紹介する：

・集計分析型BI、発見型BI

現在広く普及している代表的なタイプのBIである。集

計分析型BIでは、Excelのピボットテーブルのような多次元分析ツール（OLAP: On-Line Analytical Processing）を使ってデータを集計し、その結果に基づいて「見える化」を行う。また、発見型BIでは各種データマイニング手法を駆使して、大量データの中から隠れた関係性や規則性を発見する。

・WHAT-IF型BI、プロアクティブ型BI

WHAT-IF型BIとは、業務の新しい“やり方”をデザインすると同時に、その効果を事前に試算するタイプのBIである。プロアクティブ型BIは、ユーザ行動の理解にもとづき知的サービスを提供するタイプのBIである。

上記の分類は、ビッグデータ時代以前から整理・活用してきたデータ活用の切り口である。BIが注目されるようになったきっかけは、「見える化」の普及である。「見える化」は変革の第一歩と位置づけられ、これまでにも様々な企業で多くの実績を挙げてきた。しかし、ただ単に見えるようになれば良いというわけではなく、これを変革へと発展させていく必要がある。情報分析活用を核にして業務改革やサービス革新を実現し、企業に変革をもたらす新しいタイプのBI、それがWHAT-IF型BIとプロアクティブ型BIである。特に、状態が時々刻々と変化する状況下でのデータ活用を想定するプロアクティブ型BIが、ビッグデータ時代のデータ活用において核となるBI、すなわち、次世代BIである。

もう少し詳しく説明すると、プロアクティブ型BIとは、ユーザ行動の一歩先をリードする“プロアクティブな”サービスや機能を提供するためのBIである。例えば、ECサイトにおいて、顧客が商品を検索したときに関連商品を併せて表示するレコメンドサービスや、ユーザの不審な決済行動を検出して即座に決済を凍結するマネーロンダリング防止といった先進的なサービスがこれにあたる。そのため、ユーザの次の行動を高精度に予測するモデルや、ユーザの行動や状況をリアルタイムに把握できる仕組みが必要となり、センシング・データやログ・データなどのビッグデータや、各種IT基盤の出番となる。

明確な目的（切り口）のもとで、ビッグデータを適切に活用することによって、初めて、社会・企業・個人に革新をもたらすサービスが生まれる。

3. 基盤技術モデル

前述のとおり、ビッグデータには、これまでに扱ってきたデータとは異なる特有の特徴があるため、それに対

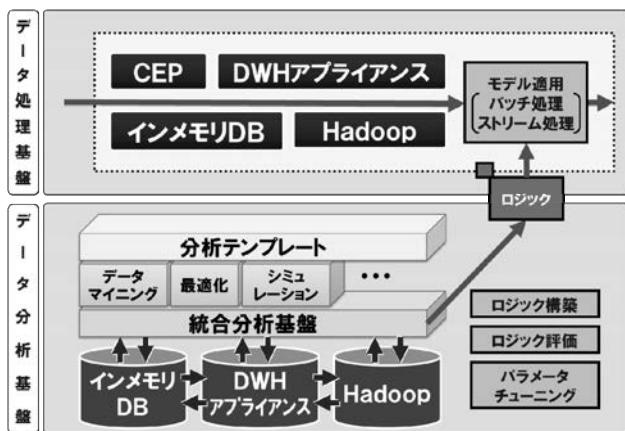


図2 ビッグデータ分析基盤

応した処理基盤を構成し、適切に運用する必要がある。つまり、プロアクティブ型BIを実現可能とするデータ分析基盤は、その都度検討する必要が生じる。

これに対して、我々は、独自の基盤技術モデルである、「ビッグデータ分析基盤」を提唱している(図2参照)。ここで、ビッグデータ分析基盤は、以下の2つの基盤によって構成される：

①データ処理基盤

ヒト・カネ・モノの動きなどの膨大なストリーミング・データを要件に合わせて迅速に処理する基盤

②データ分析基盤

バックヤードでデータを蓄積した後、大規模データマイニングをバッチ処理などで行い、統計モデルを導き出してフロントエンド処理への適用を図る基盤

ビッグデータ分析基盤におけるポイントは、様々な機能や特性を持つ複数の基盤技術を、必要に応じて、効果的に連携させることにある(図3参照)。

・データベースの構成

ビッグデータを活用する前提として、様々なデータ(膨大な量・粒度・鮮度)に対応可能なデータベース構成が必要となる。従来の情報システムはRDBMS(Relational Data Base Management System)を中心に構築されてきたが、最近ではNoSQL(Not only SQL)の利用も増えている。RDBMSはアドホックな集計、NoSQLはバッチをベースとした集計、とそれぞれ異なる特性を持つため、目的や用途に応じて両者を組み合わせたデータベース構成をとる必要がある。

・データ処理方式

ビッグデータの処理方式は、リアルタイムに発生するデータと蓄積されたデータで大きく分けられる。さら

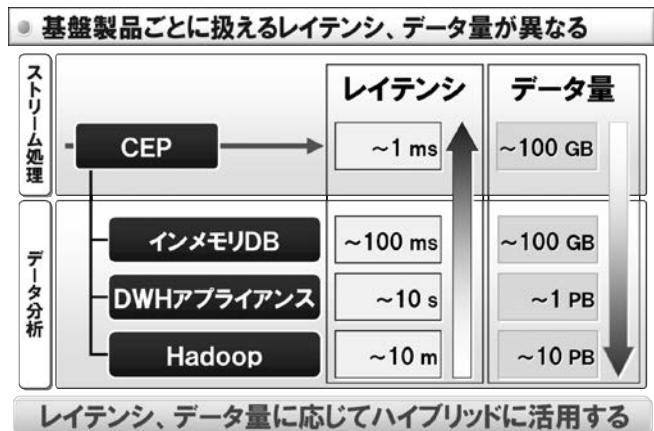


図3 基盤の構成ポイント

に、蓄積されたデータでは蓄積状況に応じて最適な処理方式が存在する。例えば、リアルタイムのデータ処理はCEP(Complex Event Processing)[3]で、数時間前までの時系列データ処理はインメモリデータベースやKVS(Key Value Store)の層で、数日～数カ月前のデータ処理はDWHアプライアンスの層で、それ以上の時間をさかのぼる場合は、Hadoop[4]上のMap/Reduce層において並列分散処理を行う。

情報システムで取り扱うデータの構造や処理の内容に応じて、各層のアーキテクチャを柔軟に組み合わせることで、ビッグデータ活用に最適なパフォーマンスを発揮できるようになる。このときに重要な観点が、図3に挙げたレイテンシとデータ量である。

ビッグデータ活用にあたっては、対象とするデータの性質や、利用する分析手法に適した基盤の構築が必要不可欠である。従って、分析技術と基盤技術のそれぞれの専門家が、互いの技術を持ち寄って、ビッグデータ活用の最適な仕組み作りを行う動きが必須となる。どちらか一方の専門家だけでは、次世代BIの実現は困難である。

4. ビッグデータ活用の先進的事例

早い段階から戦略的にビッグデータ活用に取り組んだ一部の企業は、経営メリットや競争優位性を顕在化させ始めている。本論文では、我々が取り組んだ2つの先進的事例を紹介する。いずれもプロアクティブ型BIとして分類される事例である。

1つ目は、2010年に取り組んだ事例で、橋梁に取り付けた数十のセンサから逐次送信される、変位や傾斜などの物理量を利用して、リアルタイムに橋梁の異常検知を行った検証実験である。2つ目は、2012年に取り組んだ事例で、複数のECサイト上に蓄積された、総計テラバイト級の購買履歴を合わせて分析することで、ECサイ

トをまたぐレコメンデーションを可能とするシステムの開発事例である。

なお、業務上公開できない情報があるため、詳細が分かり難い表現になっている部分がある。ご容赦頂きたい。

4.1 リアルタイム橋梁異常検知

4.1.1 課題

高度経済成長期に建設された橋梁の老朽化が進み、現在、橋梁の予防保全や長寿命化が重要な課題の1つに挙げられている [5]。主に定期的な目視による点�査が行われており、橋梁によっては10年に1度しか点検されていない場合もある。また、人手による点検では、様々な箇所の劣化状況を正確に把握することは困難な作業となる。

全ての橋梁を、人手で常に点検・監視することは非現実的である。そこで、一部の点検・監視作業を機械に任せ、人手で見るべき箇所を自動で割り出すことで、リソースを有効活用することが考えられている。

4.1.2 ビッグデータ活用のポイント

橋梁の各所にセンサを取り付け、センサを取り付けた部分の変位や傾斜、加速度などをリアルタイムに収集・把握することで、橋梁に異常が起きていないかモニタリングする（図4参照）。

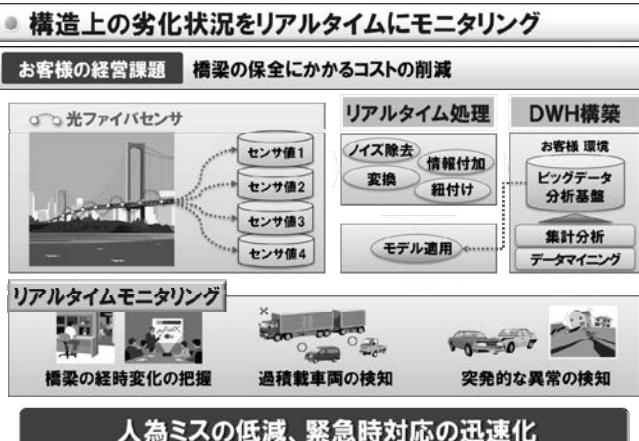


図4 橋梁センサデータのログ分析

この事例でのビッグデータとは、前述の「センシング・データ」に該当し、1つのセンサから1秒間に最大で125回データが送信されてくる。実証実験では数十個のセンサから送られてくるストリーム・データを利用して異常検知を行った。データの概要は以下のとおりである：

- ・データの種類：変位、傾斜、ひずみ、加速度
- ・データのサイズ：各々数キロバイト程度
- ・データの取得間隔：毎秒125回（変位は毎秒1回）

なお、センサから送られてくるデータはそのままでは物理的な意味を持たないため、その都度、変位などの物理量に変換を行う処理が必要となる。

ここで、複数の拠点から逐次送信されるストリーム・データを個別に分析するのではなく、複数のストリーム・データをまとめて分析することが、本実証実験におけるビッグデータ活用のポイントである。具体的には、橋梁の各地点で生成される物理情報を、その位置関係を考慮した上で一括して分析（リアルタイムで異常検知）を行う。そうすることで、数値的に外れ値かどうかをリアルタイムに判定するだけでなく、異常が生じたと判断された場合に、どのような種類の異常が生じたかをルーブベースで判定するロジックを構築・実装した。

実証実験では、異常の種類は次の3つを想定した：

- ①外力（風や地震など）による異常
- ②構造の劣化による異常
- ③センサ自体の故障による異常

例えば、複数の拠点のセンサ値どうしに位置関係を加味した上の相関関係が現れている場合（“遅延相関” [6] が現れている場合）、風や地震などの外力によって橋梁全体が揺れている、というような判断を行う。また、遅延相関が全く見られない場合には、センサそのものが故障をした、というような判断を行う。

4.1.3 基盤構成

ビッグデータ分析基盤として、本実証実験で求められた要件は以下のとおりである。

・以下の処理を1秒以内で処理できること

- A) センサ値を物理量に逐次変換する
- B) 変換した複数の物理量をまとめて処理する

そこで、ミリ秒レベルでのデータ処理を可能とするCEPをデータ処理基盤として採用した。異常検知を行うロジックの構築と評価に関しては、数キロバイト程度のデータであったため、データ分析基盤を特別に用意することはせず、通常の開発機1台のみを利用した。

データ処理の流れを以下に示す：

1. センサから送信される値を受け取る
2. 受け取った値を、変位などの物理量に変換する
3. 複数の物理量をまとめ、異常検知を行う
4. 異常であると判断された場合、種類を判定する
5. 判定結果を出力する

データの受け取り部分はC言語で実装し、2と3の処理はCEP上で実装した。また、4の異常の種類を判定する際には、C言語で作成した分析ライブラリをCEPから呼び出す形で実装を行った。

4.1.4 実証結果

実データを用いた実証実験を行い、1秒以内で処理が終わることを確認した。また、異常検知については、地震が実際に発生した時間を含むデータを用いて実験を行い、異常として検知可能であること、かつ、その異常が外力による異常であると判定されることを確認した。なお、本事例は、米国IDG(International Data Group)主催の世界的なプログラム「The Computerworld Honors Program」Security & Safetyカテゴリにおいて、CEPを用いた先進的なビッグデータ活用事例として21st Century Achievement Award Finalistに選出されている[7]。

4.2 クロスレコメンデーション

4.2.1 課題

例えば、事業部ごとにECサイトを運営しているような場合、新規に立ち上げたばかりのサイトと古くからあるサイトでは、その利用者数に大きな開きがあることが多い。また、年齢層ごとにターゲットを絞って、複数のサイトを運営している場合には、ある年齢層から急激に離反者が増え、利用者数が激減するようなこともある。

このような状況に対して、多くの利用者がいるサイトから、利用者が少ないサイトへ、利用者を“誘導する”ことが考えられる。ここで、誘導する1つの手段として、別のサイトの商品をレコメンド(推薦)することで、サイトを移動して貰うことが考えられている。複数のサイト(ドメイン)をまたぐところから、そのようなレコメンドは、クロスレコメンドと呼ばれることがある[8]。

4.2.2 ビッグデータ活用のポイント

サイトごとの一定期間分の購買履歴を、それぞれ収集し、まとめて分析することで、ユーザごとに、異なるサイトのレコメンド商品を事前に抽出しておく。そして、あるECサイトの商品を閲覧しているユーザに対して、別のECサイトの商品をレコメンドする(図5参照)。

ここで、サイトを意識せずに1つの大きな履歴データとして、既存のレコメンドアルゴリズムを適用することも考えられる。しかし、そのようにすると、特定のサイトの商品だけがレコメンドされる可能性があるため、その商品が属すサイトの区別を踏まえた上でレコメンド商品を抽出できるアルゴリズムを利用する方が好ましい。また、ECサイトによっては、対象ユーザの履歴が存在しない場合もあるため、デフォルトのレコメンド商品リストを用意しておくなどの対応方法も事前に検討してお

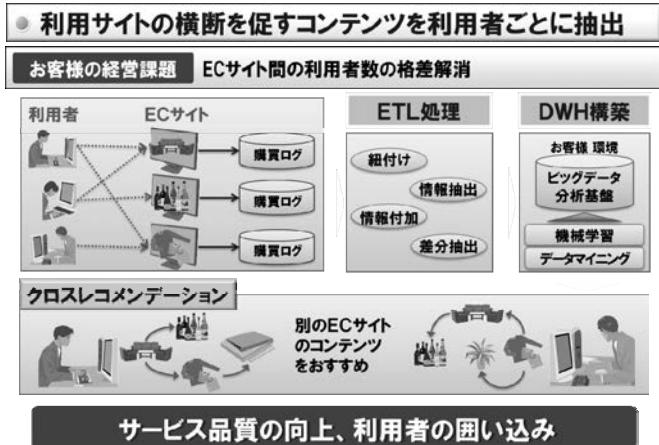


図5 複数のECサイトの購買ログ分析

く必要がある。

この事例でのビッグデータとは、第1節で述べた「ライログ・データ」に該当し、本取組みにおいては、複数のECサイトの購買履歴は毎日定時に送信される。そして、日々蓄積される購買履歴に基づいて、利用者ごと/ECサイトごとに、レコメンドする商品を定期的に洗い替えする。データの概要は以下のとおりである：

- ・データの種類：購買履歴、利用者情報、商品情報
- ・データのサイズ：合計で数テラバイト
- ・データの取得間隔：1日1回定時に送信される

4.2.3 基盤構成

ビッグデータ分析基盤として求められた要件は以下のとおりである。

・以下の処理を1日以内で処理できること
A) 利用者ごとにレコメンド商品リストを作成する
ここで、対象とするデータには非構造化データが含まれており、さらに、大量のデータを一定時間内に処理する必要があったため、データ分析基盤としてHadoopを採用した。また、レコメンド商品リストを抽出した後は、利用者ごとにレコメンド商品を瞬時に提示する必要があるため、データ処理基盤として、インメモリDBを採用した。データ処理の流れを以下に示す：

1. 購買履歴や利用者情報、商品情報を受け取る
2. データの前処理を行う
3. 利用者ごとにレコメンド商品リストを作成する
4. レコメンド商品リストをメモリ上に保持する
5. 要求ごとに、求められたリストを返す

本システムは全てJavaで実装した。現在、数台のラックで構成されたレコメンドシステムとして実稼働中であり、機能追加等のさらなる拡張が検討されている。

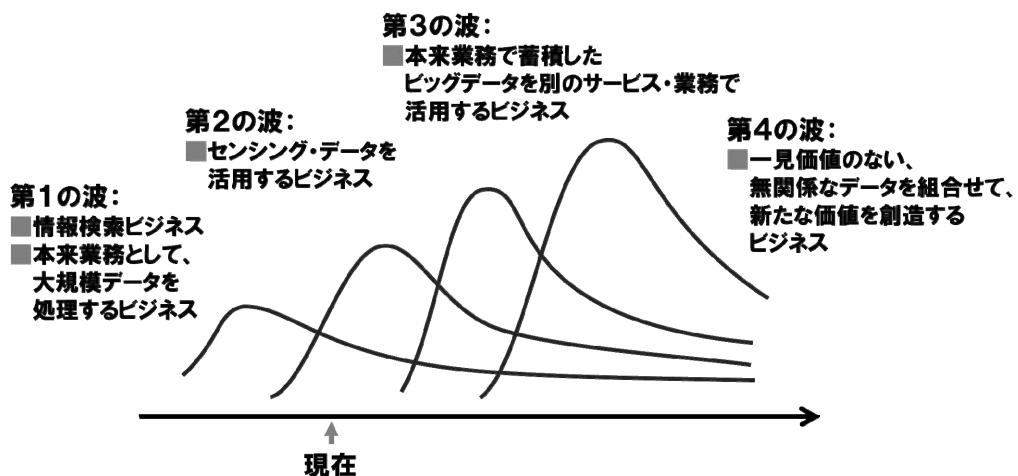


図6 ビッグデータの発展モデル

5. ビッグデータの今後

本論文では、ビッグデータの活用法を具体的にデザインするための切り口、次世代BIを紹介した。今回紹介した先進的事例を通して、ビッグデータに対する漠然としたイメージを、少しでも具体化して頂くことができたら幸いである。

ビッグデータというと、データの収集・分析基盤の導入に目を向けがちであるが、先に述べたとおり、そのような新たなIT基盤と、分析の定石となる分析デザイン(4つのBI)が、まさに両輪で推進されるべきである。

我々は、分析デザインとして、これまでの分析実績を独自に体系化した4つのBI、特にプロアクティブ型BIを核とし、ビッグデータの活用実績を積んできている。また、それと合わせて、IT基盤として、CEPやHadoopなどの基盤アーキテクチャを構成要素を持つビッグデータ分析基盤を提唱し、実証実験等を通して、データの特性や分析手法に合致した基盤の構成や開発に関するノウハウを蓄積中である。

ただし、紹介した先進的事例を見ても分かるとおり、ビッグデータ時代は始まったばかりである(図6参照)。複数の基盤を巧妙に連携させ、より高度なデータ処理を行う段階にはまだ至っていない。しかし、数年のうちに、本格的なビッグデータ時代を迎える、様々なデータ活用やサービスを目や耳にすることは確かであろう。

来たるべき“第3の波”や、“第4の波”を上手にのりこなすためにも、周りに惑わされることなく、ビッグデータを活用する本来の目的を見極め、目的達成に必要な十分な基盤のもとで、ビッグデータを活用していくことが重要である。

参考文献

- 1) NTTデータ技術開発本部: BI革命, NTT出版 (2009).
- 2) NTTデータ技術開発本部: 1冊でわかるビッグデータ「ビッグデータ活用の勘所」(日経BPムック), 日経BP社 (2012).
- 3) 桑田修平, 中川慶一郎: CEPを用いたストリーム・データ分析, オペレーションズ・リサーチ, Vol.56, No.9, pp.511-517 (2011).
- 4) 太田一樹, 下垣徹, 他: Hadoop徹底入門, 翔泳社 (2011).
- 5) 二羽淳一郎, 他: 特集「橋梁の長寿命化」, 橋梁基礎, Vol.44, No.8 (2010).
- 6) Y. Sakurai, S. Papadimitriou and C. Faloutsos: BRAID: Stream Mining through Group Lag Correlations, ACM SIGMOD Conference, pp.599-610 (2005).
- 7) <http://www.nttdata.com/jp/ja/news/release/2012/060500.html>
- 8) J. Tang, S. Wu, J. Sun and H. Su: Cross-domain Collaboration Recommendation, ACM SIGKDD Conference, pp.1285-1293 (2012).

桑田 修平 (正会員)

E-mail: kuwatas@nttdata.co.jp

2003年4月 (株)NTTデータ入社。2005年4月～2007年9月 日本電信電話(株)NTTコミュニケーション科学基礎研究所。現在、(株)NTTデータ技術開発本部サービスイノベーションセンター主任。

中川 慶一郎 (非会員)

E-mail: nakagawaki@msi.co.jp

1992年4月 NTTデータ通信(株)入社。2002年4月 同社技術開発本部。2012年4月 (株)数理システム出向。現在、(株)数理システム取締役。

投稿受付：2012年09月17日

採録決定：2012年12月03日

編集担当：武田浩一（日本IBM）