Keywords you should know. 第100回

知っておきたいキーワード

ビッグデータとは

西山　智†

† KDDI株式会社 技術戦略部
"Big Data" by Satoshi Nishiyama (KDDI Corp., Tokyo)
キーワード：ビッグデータ，データ分析，プラットフォーム，Hadoop，個人情報，通信の秘密

まえがき

「ビッグデータ」がはやり言葉になって，すでに2～3年経ちました．もう「ビッグデータ」はバズワード化したのでしょうか？ 筆者は，これまで着実に行われてきたデータ分析がビッグデータというキーワードで脚光を浴びている場合も多いこと，また実際に役立っている事例も多いことなどから，「人工知能」とか「エキスパートシステム」のように一過性で終わるのではなく，ビッグデータのブームはまだしばらく続くと思います．そこで，今回はビッグデータについてとりあげます．

ビッグデータとは，文字通り，処理するには大きすぎるデータを指していますが，どの程度の大きさが大きすぎるのかは相対的なものであり，特に規定はありません．データの種類についてもどれが該当するのか定義はありません．どれがビッグデータなのかは扱う組織や人により決まります．

例えば，衛星によるセンシングデータやゲノム解析の結果，あるいは小売業におけるPOSデータなどは，ビッグデータの代表例と多くの人が考えるでしょう．通信会社や公共交通機関などのインフラから生じる大量のログも典型的なビッグデータであり，例えば，通信会社はログを処理するために，ペタバイトクラスのデータベースを構築していたりします．一方，個人にとっては，数テラバイトのデータを処理するのも結構大変で，充分ビッグデータと思うかもしれません．

またデータの大きさだけがビッグデータの特徴ではありません．ガートナーは，ビッグデータを表す特徴として3つのV（High Volume：巨大なデータ量，High Velocity：高いデータ入出力速度，High Variety：高いデータの多様性）を定義し，そのいずれかに該当する場合をビッグデータとしています[1]．さらに近年，IBMが4番目のVとして，V（Veracity：正確さ）を追加しました[2]．

ビッグデータの分析

従来のデータマイニングでは，発生頻度が高く，かつ特徴的な事実を抽出することが重要でした．これに対しビッグデータでは，大量のデータが存在することを活用して，従来は無視されていた発生頻度は低い，いわゆるロングテール部分にあるが特徴的な事実を洗い出してビジネスに結び付けるためにも使われます．Amazonは商品販売実績やユーザの商品検索履歴などを活用して，ロングテールを考慮した品ぞろえと個人毎にカスタマイズした商品推薦を行い，ビジネスを成功させました．KDDIでも，大量の通信ログを解析して，非常にまれにしか起こらない障害の原因を探ったりしています．

ビッグデータのもう一つの活用が，統計化して全体傾向をつかむことです．例えば，東日本大震災の際には，自動車会社はカーナビシステムから集まった情報を利用して，東北地域の走行可能な道路を可視化し ☞

映像情報メディア学会誌 Vol. 68, No. 5, pp. 396～398（2014）

☞ 災害復旧に役立てました．携帯電話端末の携帯ネットワークへの位置登録情報や通信時のログは，利用者の位置を推定するために役立つ情報が含まれていますので，人口動態，観光の周回パターンなど，いろいろな分析が行えます[3) 4)]．皆さん，たまに駅や道路で人数を計数している作業員の方を目撃されたことがあるかと思います．通信ログを使えば精度は少し落ちますが，このような交通センサス的な調査を代替する可能性を秘めています．また人の動きは，携帯電話のアプリを使って端末の位置を収集することでも集められますし，駅の自動改札機の情報や監視カメラの映像などからも推定することができます．人以外でも，例えばKDDIでは，基地局の一部に気温や日照などがわかる気象センサを設置しており，気象庁のアメダスよりも細かいエリア毎に降雨の有無がわかります．少し昔には，名古屋市でタクシーのワイパーにセンサを付けて，その作動状況で降雨を推定する実験が行われていました．

ビッグデータを支える処理基盤

ビッグデータを支える処理基盤としてはHadoopが有名です．Hadoopは，Googleの分散KVS（Key-Value Store）であるBigtableとそのプログラミングモデルであるMapReduceの仕様をもとに，オープンソースとしてApacheが実装したもので，大規模なデータを安価かつ安全に格納し，加工するために適しています．名前は開発者の一人の息子が持っていた黄色い象のおもちゃの名前からとっており，象がシンボルとして使われています．分散システムではいかに並列性を稼ぐかが性能に大事ですが，**図1**に示すように，Hadoopがデータの分散や処理の並列化を自動的に行い，利用者は分散を意識することなく，MapとReduceという2種類の処理を記述するだけで，大規模並列処理の性能を享受できます．安価なサーバを多数並べて並列性が得られる上，ソフトはオープンソースで利用できることもあり，研究分野のみならず民間企業でも多数利用されるようになりました．MapReduceはJava言語で記述することから，MapReduceになれていないデータ分析者向けに，HiveやImparaなどのSQL実行系も開発されています．また企業向けには，複数の会社がカスタマイズしたサポート付きHadoopを有償で提供しています．またHadoop向けの専用サーバも販売されており，企業が使いやすい環境が整えられつつあります．

一方で，Hadoopは大規模なデータを効率的に分散処理するために設計されていることから，小規模データの処理を行わせても性能は上がりません．処理を起動するだけで数秒間から下手をすると1分近くかかりますし，ディスクを読み書きするブロックサイズが標準で64 MBと非常に大きく，最低でもこの大きさで並列化を行いますので，小さなデータでは並列度もあがりません．

ビッグデータを支える基盤はHadoopだけとは限りません．**図2**に示すように，DWH（Data Warehouse）向けの分散型RDBMS（Relational Database Management System）のみ，あるいは両者の組合せもよく用いられます．DWH向けの分散RDBMSは，データ格納と検索に特化した設計がなされており，更新はほとんど考慮されていませんが，検索は同じHW規模のHadoopより一般的に高速です．一方，格納するデータ量あたりの単価は，Hadoopより高価な場合が多く，結果的に生に近いデータをHadoopに格納し，DWHには生データを加工した結果を入れる場合が多くなります．この時，Hadoopは生データの格納，検索や集計だけではなく，DWHに格納するデータの前処理（ETL: Extract, Transform, Load）のT（データ加工）としても利用されます．DWH向けの分散型RDBMSはMPP（Massively Parallel Processing）型のものが多く，単純にソフトだけで実現するものから，専用の処理ボードを備えるもの，並列サーバ間のネットワークを高速にしたもの，Flashメモリーを ☞

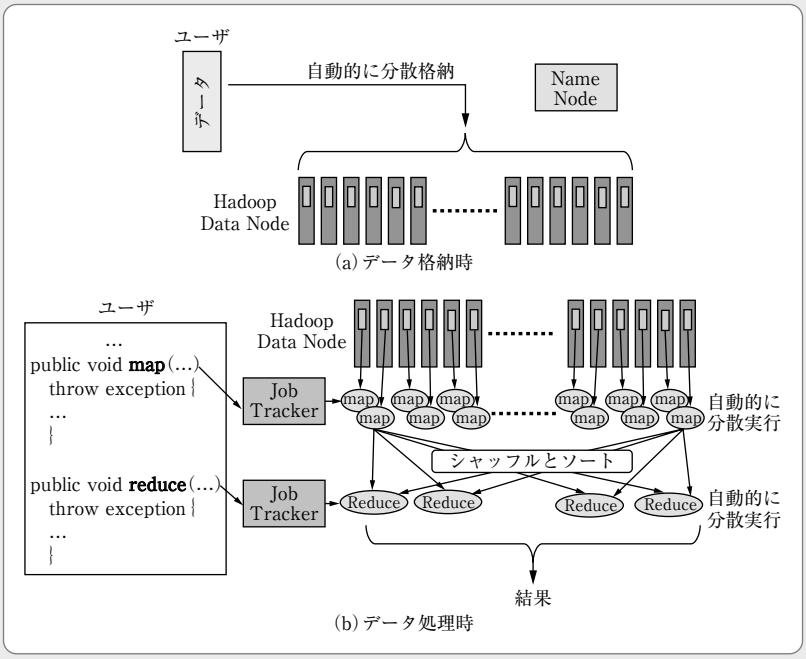

図1 Hadoopによる分散処理

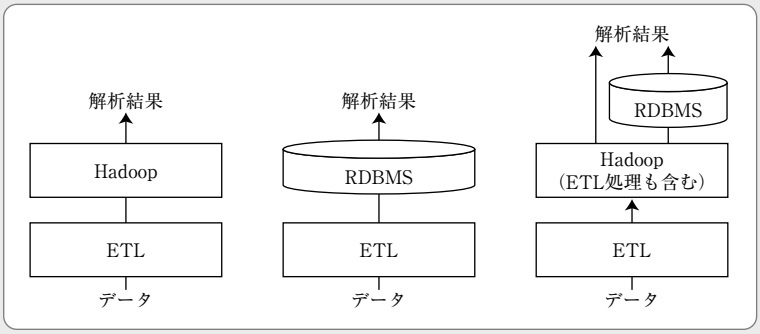

図2　ビッグデータ処理基盤の典型的な構成

☞ 活用するものなど，いろいろなアプローチの製品が競っており，利用する側で自分の用途に合わせて選定する必要があります．また，FacebookやGoogleなどの大規模なサービス事業者は，独自の基盤を構築している場合が多く，例えばFacebookの基盤は，分散処理基盤のソフトウェアからデータ格納用のサーバまで独自開発が行われています．ちなみに，Facebookのビッグデータ基盤はこの1～2年でエクサバイトの規模になると報告されています．

パーソナルデータの取り扱い

前述の通り，いわゆるビッグデータには，通信のログや位置取得履歴，買い物履歴や自動改札の利用履歴など人の活動により生じるものが多くあります．これらのビッグデータには，個人の情報（パーソナルデータ）が含まれることから，プライバシー保護の観点から，その利用には一定の制約が課されています．パーソナルデータを収集し活用する場合には，どのような情報をどのような目的で取得し，第3者への提供などその取り扱いについて，プライバシーポリシー等の形態で事前に利用者に示す必要があります．また利用者に利用を拒否する仕組み（opt-outと呼びます）を提供する必要があります．これらがきちんとできていないと，社会的問題になる可能性があります．最近では，JR東日本のSUICAデータを日立に提供しようとして問題となった事例が記憶に新しいところです．このときは，第3者への提供を利用者に事前に示しておらず，また匿名化処理が甘かったことが問題となり，結局提供中止に追い込まれました．

通信会社の場合，日本国憲法が通信の秘密を保障しており，通信の内容のみならず誰が誰といつ通信したとか，通信した位置など，いわゆる通信の外延情報も通信の秘密の一部とみなされていることから，通信ログの利用にさらに厳しい制約が課されています．通信会社は，通信ログを課金やネットワークの品質改善など，通信会社に必須の目的には利用できますが，マーケティングなどその他の目的には利用できません．利用するためには，利用者の明示的な同意（opt-inと呼びます）が必要となります．諸外国では，匿名化すればopt-inなしに利用したり，大学などの第3者に提供できたりする場合が多いことから，通信ログの活用に関する研究が進んでおり，専門の国際会議[5]も立ち上がっています．日本では研究目的でも，あらためてそれ専用のopt-inを取得する必要があることから，日本からの参加は少なく，研究の立ち遅れが懸念されます．

むすび

本稿では，ビッグデータ解析で得られるものや解析を行うための基盤，それにパーソナル情報を含む場合の取り扱いを述べました．ICT技術の発達に伴い，ますます人やセンサから得られるビッグデータが活用される機会が増えていくと思います．本稿がその活用の一助になれば幸いです．

（2014年1月31日受付）

参　考　文　献

1) D. Laney: "3D Data Management: Controlling Data Volume, Velocity and Variety", Gartner（2001）
2) http://dsnowondb2.blogspot.jp/2012/07/adding-4th-v-to-big-data-veracity.html
3) http://www.dcm-im.com/service/area_marketing/mobile_spatial_statistics/
4) http://www.kddi.com/corporate/news_release/2013/1029a/
5) http://perso.uclouvain.be/vincent.blondel/netmob/2013/

西山　智（にしやま　さとし）
1984年，国際電信電話(株)（現 KDDI(株)）入社．1991年，テキサス大学オースチン校計算機科学科修士課程修了．これまで，データベース，分散通信システム，ユビキタス通信システムの研究に従事．現在は社内でビッグデータの基盤構築と分析の両面に関与．