ビッグデータ解析の現状

本レポートでは、既にエクサバイト級のデータが流通していると推定されるビッグデータの現状ならびに解析基盤の技術動向とそのリアルタイム化に伴うビッグデータ解析の変化について解説します。

3.1 ビッグデータの現状

近頃では、至る所で見聞きするようになった「ビッグデータ」ですが、その実態は相変わらず非常に捉え辛いままです。語り手の立場や意見によって「ビッグデータ」の意味するところが大きく異なることや、有効だとされる活用事例が様々な業種をまたいで多岐にわたることが、その主たる原因でしょう。ビッグデータに関わる現状の網羅的な把握を試みた総務省の平成25年度の調査では、国内でのビッグデータの流通量は年々拡大していることが報告されています。

メディア別の流通量では、POS、RFID、GPSから得られるデータの総量が多く、経年推移では医療系データ（電子カルテ、画像診断）とM2M系データ（GPS、RFID）が大きく伸びていると報告されています。この調査では、メディアについて「データは出生が多様で様々な種類に及んでる」とし、

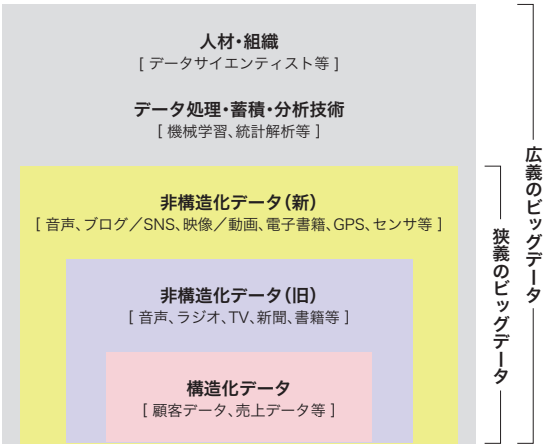
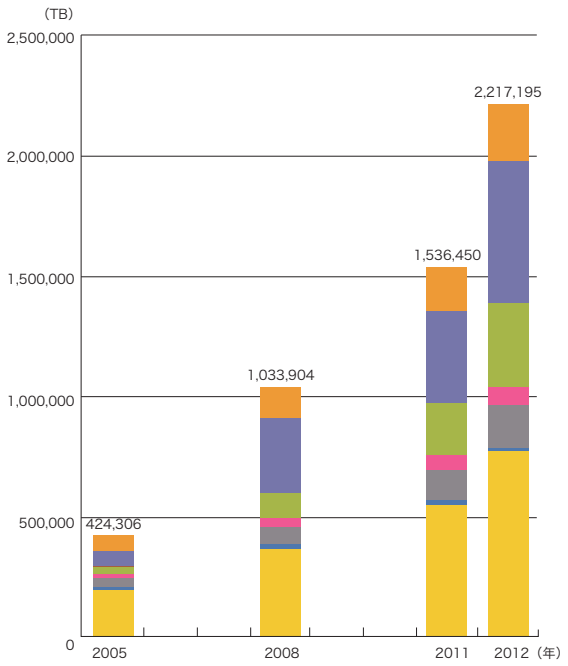


図-1 データ形式の異なる3種類のデータ



（単位：TB）

	2005	2008	2011	2012(年)
アクセスログ	521	1,008	1,566	2,147
Blog、SNS 等記事	155	379	721	1,129
電子メール	64,566	121,085	178,501	238,048
気象データ	877	1,958	4,176	6,229
RFID データ	64,519	309,559	370,440	583,942
GPS データ	33,116	104,928	228,801	348,143
E コマースにおける販売ログ	12	28	50	69
携帯電話	17,765	36,068	56,544	75,881
固定 IP 電話	38,035	80,958	132,017	178,474
CTI 音声ログデータ	2,509	5,662	10,878	15,019
画像診断	71	191	445	630
電子カルテ	227	636	1,032	1,635
業務日誌	1	2	4	6
電子レセプト	2	4	5	7
POS データ	201,823	371,272	551,026	765,424
経理データ	100	161	234	397
顧客 DB	4	7	11	16
合計	424,306	1,033,904	1,536,450	2,217,195

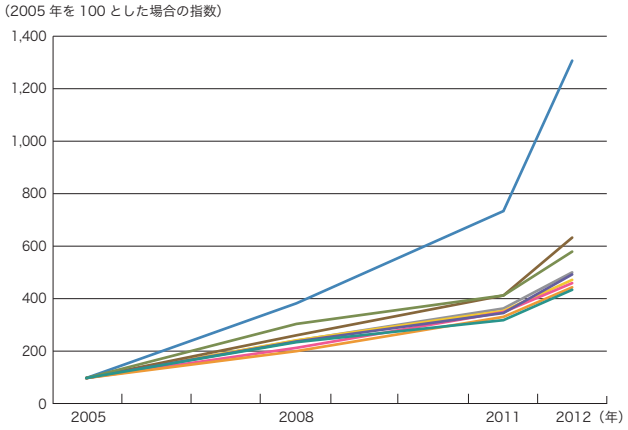
図-2 ビッグデータ流通量の推計 *1

＊1 http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc113220.html

「データ形式の異なる3種類のデータに分類される(図-1)」と説明されていますが、データ分析の視点から見ると非構造データ(新)に分類されるデータ群の中には、POS、RFID、GPSなどのように標準フォーマットが規定されている(すなわち構造化されている)事例もありますし、映像や音声、テキストなどのストリームデータでも、タイトルや著作者などのメタ情報を内蔵する事例があるので、実際には、構造型データと非構造型データを内包する複合型のデータと理解するべきでしょう。図-2は、国内で流通するビッグデータは非構造化データよりも構造化データが圧倒的に多いことを示しています。

次に、産業別でのビッグデータの流通量(図-3)と蓄積量(図-4)の推定値について、総務省の調査では次のように報告しています。

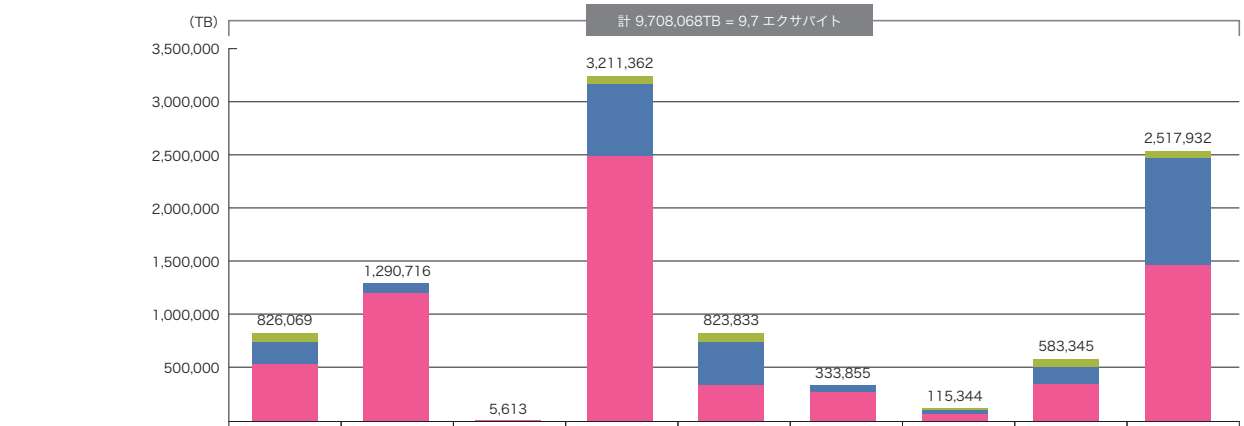
この推定の大変興味深いところは、流通量に関しては(不動産業を除くと)概ね同じ特性で増加しているのに対し、蓄積



(2005 年を 100 とした場合の指数)

	2005	2008	2011	2012(年)
製造業	100	233	320	431
建設	100	303	411	576
電力・ガス・水道	100	200	327	436
商業	100	239	345	496
金融・保険	100	211	349	459
不動産	100	378	732	1,310
運輸	100	260	410	632
情報通信	100	242	351	462
サービス	100	236	363	499

図-3 ビッグデータ流通量の推移(産業別) *2



	製造業	建設	電力・ガス・水道	商業	金融・保険	不動産	運輸	情報通信	サービス
蓄積量 _ 社内	543,156	1,194,956	3,722	2,468,219	334,843	270,270	76,887	350,485	1,455,067
蓄積量 _ 社外国内	203,562	90,900	1,376	678,545	412,397	63,237	30,258	167,949	990,394
蓄積量 _ 社外国外	79,350	4,860	515	64,598	76,594	348	8,199	64,911	72,472
蓄積量 _ 合計	826,069	1,290,716	5,613	3,211,362	823,833	333,855	115,344	583,345	2,517,932

図-4 ビッグデータ蓄積量(産業別、2012年) *3

*2 http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc113220.html
*3 http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/html/nc113230.html

量は業種ごとに大きく異なることです。ビッグデータの蓄積量は、その利活用の度合いを表しているとすれば、これは業種ごとのビッグデータ利活用に対する取り組みの温度差を示していると理解することもできますが、業界ごとの慣行の違いが反映されている可能性もあります。産業別の蓄積量では、BtoBの業種よりもBtoCの業種の方が蓄積量は多いこと、更に、ビッグデータの多くが所有者の内部に留まっている非公開データであることにも注目されます。

総務省の調査結果に基づけば、既にエクサバイト級のデータが流通していると推定されることから、「ビッグデータの利活用は始まっている」と主張しても良いでしょう。しかしながら、「大量のデータソースから収集・蓄積したビッグデータを選別・分析することにより、新たな知見を得る」というビッグデータ本来の利活用は、端緒についたところという現状が浮かび上がってきます。

さて、今後のビッグデータの利活用を促進・加速する「ビッグデータに関わる今日的な課題」という議論では、「ビッグデータの共有化をどのよう、促進するのか？」、また、「ビッグデータの解析により、どのような新たな知見を得られるのか？」の2つの課題が掲げられることが多いのですが、その文脈で頻繁に語られるのが、M2M、IoT、CPSの3つのキーワードです。各々の定義を確認すると、Machine-to-Machine（M2M）は、有線や無線の通信システムを用いて同じ種類のデバイスの間でのコミュニケーションを可能にする技術、Internet of Things（IoT）は、インターネットのような構造の中で仮想的に表現されるユニークに識別可能なオブジェクト、Cyber-Physical System（CPS）は、物理的な実体を制御する協調型計算エレメントによるシステムということで、各々は独立した概念というよりは、センサーなどのデバイスから構成されるネットワークに対して、3つの視点を提示していると理解するべきなのかもしれません。例えば、RFIDタグを用いた荷物の集配管理システムでは、個々の荷物にRFIDタグが付与されますが、その荷物がRFIDリーダーの近傍を通過すると、その位置情報を付加されたデータが生成されます。このデータを蓄積しておいて任意の荷物について位置情報をトレースすると、その荷物がどういう経路で移動して、現在どこで保管されているのかを調べることができます。更に、集配管理システムが取り扱うすべての荷物についてトレースすれば、荷物が集中している配送センターを正確に把握することができます。スマートフォンが普及した今日では、同じ方法で人間の動態も把握できるわけですが、各デバイスが生成するデータを集めるとビッグデータが形成され、業務の最適化やマーケティングに有益な知見が得られると考えられています。

3.2 ビッグデータ解析基盤の技術動向：リアルタイム化

前述のようなM2M、IoT、CPSのコンセプトに基づくデバイスネットワークから生成されるビッグデータの解析には、即時性が求められることは自明ですので、必然的にビッグデータ解析基盤は、リアルタイム解析の要請に応えなければならなくなります。

ビッグデータ解析の基盤技術と言えば、特にMapReduceに基づく分散処理基盤の技術を思い浮かべる方々が多いかと思いますが、バッチ処理を前提として設計されたMapReduceのアーキテクチュアをリアルタイム解析に適用するには困難が伴います。バッチ処理ではジョブが終了するまでに処理結果が確定しないので、処理時間分の遅延が発生してしまいます。この時間を縮小する対策としては、まず処理そのものの最適化・高速化を図ることになりますが、自ずと限界があります。次なる対策としてはバッチ処理のサイズを小さくしていくことになりますが、あまり小さくしてしまうとバッチ処理として意味をなさなくなってしまいます。通常、MapReduceジョブは数分から数時間という処理時間を要することが一般的ですが、これをウェブサービスで許容される応答時間の数秒程度まで時間短縮することは事実上不可能だと言えるでしょう。

ビッグデータ解析基盤のリアルタイム化の研究では、2つのアプローチが検討されてきました。その1つは、「ユーザのリクエストにリアルタイムで応答する」アプローチで、この場合「MapReduceの基盤そのものを機能強化」して応答性能を向上させるケースと、「MapReduceを内包する基盤システム」として考えて、内部的にはMapReduceを利用しつつ、それ以外のところで応答性能の向上を図るケースがあります。もう1つは、「実際に実時間でデータ処理を実行する」、いわゆる「リアルタイムストリーム処理」を実現する基

盤技術を用いるアプローチで、この場合はMapReduceに置き換わる形になります。

「MapReduceの基盤そのものを機能強化」の事例としては、MapReduce Online[*4]があります。この事例ではHadoopを大幅に改造してMap処理とReduce処理の間のデータの受け渡しのパイプライン化を図っています。この機能拡張によりユーザは処理中のジョブの状態を詳細に把握する、すなわちイベントモニタリングが可能になります。また、MapReduceアプリケーションでストリーム処理を記述することが可能です。

「MapReduceを内包する基盤システム」の事例としては、GoogleのDremel[*5]のアイデアを踏襲する2つのオープンソース・クローン「Apache Drill[*6]」と「Cloudera Impala[*7]」が上げられます。いずれも、リアルタイムでのデータ処理を行っているわけではありませんが、それに匹敵する低遅延でのクエリー応答性能を発揮します。

「リアルタイムストリーム処理」の事例としては、Apache Storm (Twitter Storm)[*8]があります。Stormは元々Twitterの分析を行っていたBackType社が開発したシステムですが、Twitter社によるBackType社の買収にともなってApache Project経由でオープンソース化された経緯がありますが、それ自体は汎用のビッグデータ処理基盤です。

Stormは、Complex Event-Processing (CEP) を実現するストリームエンジンが搭載されており、Spout/Boltと呼ばれるビッグデータ処理を行うエンティティに欠損のないデータストリームの供給を保証します。Stormでのストリームは、Tupleと呼ばれるデータ単位のフローで表現され、SpoutとBoltを繋ぎ合わせるTopolozyを定義することにより、ストリーム処理全体を記述します。Spoutはデータソースを表現するエンティティであるのに対し、Boltはデータの変換・加工を司るエンティティです。定義や機能は全く異なりますが、MapReduceにおけるMap及びReduceを想像すると理解しやすいでしょう。

StormもHadoopと同じようにクラスタを構成することが可能で、Nimbus、Zookeeper、Supervisor、Workerの4種類のソフトウェアが動作します。Nimbusは、Workerのスケジューリングやモニタリングを司るマスター、Zookeeperは、Hadoopでも使用されている分散ロックマネージャ、SupervisorはNimbusからのリクエストを受け付けWorkerの起動・停止を制御し、Workerは、実際の処理を実行するプロセスとして機能します。これらのソフトウェアをクラスタ内のノードに適切に配置することにより、高いスケーラビリティや耐障害性を実現しています。Storm自身は、Clojureと呼ばれるLISPライクな言語で記述されていますが、JavaVMの上で動作します。従って、Javaをはじめとする JavaVMの上で動作する各種の開発言語を使ってSpoutやBoltを記述することができます。

以上、ビッグデータ解析基盤のリアルタイム化に関する課題と幾つかの事例を紹介してきましたが、Apache Stormの登場をリアルタイムビッグデータ解析基盤のデファクトスタンダード化のトレンドと捉える見方が増えているように思えます。やはり、Hadoop (MapReduce)と共通するシンプルなプログラミングモデルとTwitter分析での実績は、汎用的なプラットフォームに期待される要件でしょう。

では、StormはHadoopに置き換わるオープンソースのビッグデータ解析基盤になるのでしょうか？この質問に対する答えは、「両者による棲み分けが進む」ということになりそうです。というのも、すべてのビッグデータ解析が即時性を求められる訳ではないからです。前述のような、即時応答を求められるビッグデータ解析の需要は、Stormに置き換わって行くでしょうが、従来のバッチ処理で十分(あるいはそれが必要)な需要にはHadoopが使われ続けることになるでしょう。またHadoopでの解析の前処理(データの整形、フィルタリング、マッチング)にStormを活用したり、ラムダアーキテクチュアと呼ばれるバッチ処理とストリー

*4 http://db.cs.berkeley.edu/papers/nsdi10-hop.pdf
*5 http://research.google.com/pubs/pub36632.html
*6 http://incubator.apache.org/drill/
*7 http://blog.cloudera.com/blog/2012/10/cloudera-impala-real-time-queries-in-apache-hadoop-for-real/
*8 http://storm.incubator.apache.org/

ム処理を組み合わせたハイブリッドなシステム構築法も提案されています。現時点では、両者は相互補完の関係にあるという理解が一般的なようです。

3.3 リアルタイム化に伴うビッグデータ解析の変化

ビッグデータ解析基盤のリアルタイム化は、ビッグデータの定義でよく語られる3Vs（Volume、Variety、Velocity）のうちのVelocity方向への能力拡大への対応と理解されています。3Vsの定義を世に広めたGartnerによると、Velocityでは「データの生成と処理の高速性」について説明していますが、例えば、センサーデータやログデータの解析や、GPS情報を用いた時空間データ解析、あるいはソーシャル・メディアから入手できるストリームデータの解析などが、Velocityの要件が問われる具体的な事例として上げられます。このような既存の解析では、異常検知や空間移動の履歴、センチメント分析など、蓄積している過去データから新たな知見を得る事例が良く知られていますが、リアルタイムビッグデータ解析基盤の即時性の向上により、今後は、より時間軸に重点をおいた時系列分析や、それに基づく予測などへと解析方法の多様化が進むと考えられています。

3.3.1 ソーシャルビッグデータとしてのWikipedia

我々は、ビッグデータの時系列に着目した分析の1例として、Wikipedia Pageview Count（Wikipedia PVC：http://www.gryfon.iij-ii.co.jp/ranking/）を用いたトレンド分析を試みています。周知のように、Wikipediaは最も成功しているインターネット百科事典です。非常に開かれた運営方針が採用されており、そのデータベース等は無償で入手できることから、研究など様々な用途に活用されています。その一部として公開されているWikipedia PVCは、2013年1月頃から公開されるようになりましたが、これは各Wikipediaページの直近の1時間あたりのページビュー数を1 ～2時間の更新頻度で公開しています。Wikipedia PVCとWikipediaデータベースを組み合わせると、社会的トレンドを示す時系列データとして利用することができますが、これはインターネット経由で入手できるソーシャルビッグデータの1例と見なせます。百科事典としての特性を持つため、一般的なソーシャルメディア（SNSやブログ）と比較すると次のような特徴があります。

- 利用者自身の改変が認められていますが、ガイドライン等により記述内容の発散を防ぐ運用が行われているため社会全般に関する最大公約数的情報が得られます。
- 百科事典であるが故に網羅性に優れており、閉じた空間内でLinked Dataを形成しています。
- 一般利用者のサービス認知度が高く、利用者は未知のトピックの詳細や関連情報を知るために利用しています。
- 多言語に対応しており、各言語間でのページごとの対応付けが明確である事例が多いです。

トレンド分析などにおいて、一般的なソーシャルメディアから取得したメッセージ等を用いてテキスト分析を行う場合、用語法の統一などが図られていないため、データ解析を行う上での障害となる場合が多いですが、意味的な発散が抑制されているWikipediaデータでは、そのような障害は起こりづらく、人間に理解しやすい分析結果が得られると考えています。

3.3.2 Wikipedia PVCの時系列変動の分析

百科事典として広く認知されているWikipediaでは、メインページ、あるいは外部のサーチエンジンから、知りたいトピックについて検索して該当するページにたどり着く利用パターンが一般的でしょう。

あるWikipediaページに着目して、そのPVCの時系列的変動を観察すると、いずれかのタイミングでピークが発生した後、徐々に減衰をすることが確認できます。特に、顕著なピークが現れる幾つかのページについて更に調査を進めたところ、テレビ放送やネットニュースで報道されたトピックを扱うページが反応していることが分かりました。すなわち、テレビ番組の視聴者やネットニュースの読者は、未知のトピックが現れた場合にWikipediaでその内容を調べる行動を取っている仮説が成り立ちます。

そこで、この仮説を検証するために、連続ドラマに着目しドラマの各回の放映時間と、Wikipedia PVCの時系列的変動との関係を調査しました。連続ドラマに着目した理由は、連続ドラマが放映されている時間には対応するWikipediaページの、PVCにピークが発生する確率が極めて高いこ

とが確認できていたことによります。また、テレビドラマの場合は「視聴率」という広く認知されている指標があり、Wikipedia PVCによるピークとの関係を調べることができることも重要でした。

Wikipedia PVCは、2008年以降に民放で放映された連続ドラマ334件すべてのデータを入手できます。その中から、Wikipedia PVCにデータ欠損のないドラマ244件について、次のような方法で分析を行いました。

1. ドラマ1話分の放映時間を1時間と仮定し、放映時間から次回放映1時間前までの168時間を1つの時系列データとして抽出する。
2. 抽出した時系列データに対し回帰分析を行い、得られた係数をその回の社会的関心度とする。
3. 各回の社会的関心度とその回の平均視聴率の相関を調べ、その有意性を確認する。

[2]の回帰分析では、計量経済学で用いられる「社会的イベントは指数関数に基づいて変動する」との知見に基づき、回帰式 $pvc = \alpha * \exp(\beta * t) + \gamma$ を用いて分析を行いました。更に、放送時間の遅延や拡大を考慮してピーク値の補正を行いました。

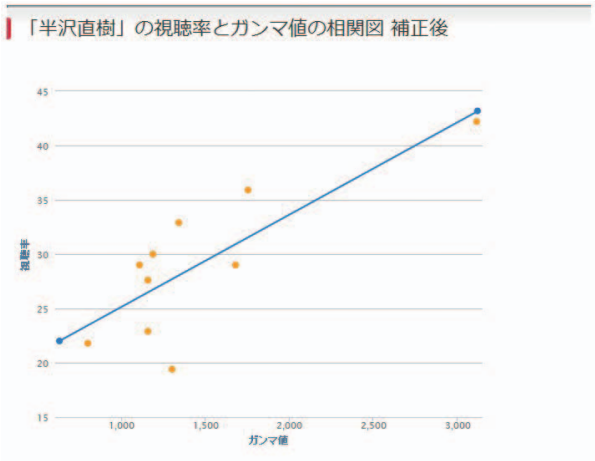


図-5「半沢直樹」の各回のγ値と平均視聴率との相関分析結果

非線形回帰分析の結果はいずれも放映時に高いピークを示し、その後γ値に収束します。分析により得られた係数α,β,γについて平均視聴率との相関を調べたところ、γ値との間に相関が見られました。テレビドラマ「半沢直樹」の各回のγ値と平均視聴率との相関分析結果を図-5に示します。

γ値と平均視聴率の相関は、本稿執筆時には、有効データ244件のうち40件で分析を行い「有意性あり」の結果を得ています。残る204件についても相関分析を順次実施して行きますが、特に、視聴率の低いドラマの場合には、Wikipedia PVC変動と視聴行動が同期していない事例も見つかっているため、有効データにおいて「有意性あり」が確認できる件数と、この解析方法が適用できるWikipedia PVC値の範囲を明らかにする計画です。

3.4 まとめ

本稿では、M2M、IoT、CPSのキーワードで語られる現在のビッグデータに関わる現状と、それに伴ってリアルタイム化を指向する解析基盤に関わる技術動向、更に時系列データからの知見を得るビッグデータ解析手法の多様化について概観しました。

「ライブで精緻なデータを取得」して、「即時性の高い処理基盤」を活用し、「今起きている何かを把握する」試みからは、従来の「静的でマクロな知見」を得るビッグデータ解析とは大きく異なる「動的でミクロな知見」が得られるものと想像しています。この新たな知見の中には「これから起こる何かの予兆」が含まれている可能性があります。誰よりも早くこのような予兆を見つけることが、今後のビッグデータに関わる課題になるかもしれません。

執筆者：

藤田 昭人(ふじた あきと)
株式会社IIJイノベーションインスティテュート(IIJ-II)企画開発センター チーフアーキテクト。2008年IIJ入社。構造化オーバーレイ研究の知見を活用したクラウドコンピューティング技術の研究開発に従事している。