第１部

特集1 2012年度WIDEボード合宿での議論
― ビッグデータの実状把握とWIDEとして取り組むべきこと ―

第１章　概要

WIDEプロジェクトでは、1994年以降、毎年8月上旬に3日間で合宿形式で、ボードメンバーを中心にした検討会（WIDEボード合宿）を開催してきた。

WIDEボード合宿で議論する検討課題は毎回ごとに異なるが、検討領域の研究開発や社会展開の状況を技術を中心把握し、WIDEプロジェクトとして活動すべき項目を整理し、具体的な活動に結び付けることを目的としている。

2012年度は、ビッグデータ（Big Data）をテーマにして、議論を展開した。結論の概観は、以下の通り。

（1）現実には、ビッグデータは存在しない。
（2）広域分散環境で動作するシステムは存在しない。
（3）ビッグデータを収集・保存・移動・提供する基盤の研究が、十分に行われていない。
（4）データのハンドリングに関する管理ポリシーが確立されていない。
（5）「オープンデータ」に関する取り組みが、ほとんど、行われていない。

これら、ビッグデータに関する技術とポリシーの現状把握をもとに、WIDEプロジェクトとして、ビッグデータに関係する研究開発の活動を、WIDEプロジェクトの大方針である“Evidenced-Based Research”に基づき行うべきとの結論となった。

第２章　現状のビッグデータに関する研究

WIDEボード合宿では、参加者にテーマが事前に与えられ、そのテーマに関する調査と考察を行ってくることになっている。その調査・考察の発表内容に関する議論を展開し、ビッグデータに関する研究の現状を以下のように整理・要約した。

（1）セキュリティー（Security）
基本的に、”Walled－Garden”と呼ばれる閉域システムとして、ビッグデータシステムが構築・運用されている。それぞれの、閉域型のビッグデータシステムは、それぞれ、異なる技術仕様・運用ポリシーを用いて設計・構築・運用・管理されており、悪意ユーザの存在や、外部システムからのアクセスや外部システムとの協調・連携は、考慮されていないものが、ほとんどである。
（2）名前（Naming）、番号（Numbering）
基本的に、閉域システムとしての設計・構築であるので、システム内で用いられる各種の識別子は、「グローバルなユニーク性を考慮しない・前提としない」という考え方となっている。すなわち、他システムとの相互接続を前提としない設計となっている。
（3）応用領域（Application Area）
アプリケーションは、特定のデータ、特定のビジネスドメインに紐付いた形で実現されているのが実状である。すなわち、特定ドメインにおける「問い」に対し、個別に設計・最適化され、実装される。そのため、クローズシステムで、制限付き、特化された形での実現が一般的となっている。
（4）標準化（Standardization）
ドメインごとのシステムの設計・実装となっている

上に、技術的な観点とビジネス的な観点から、ベンダーごとに異なる技術を用いているのが多くみられる。要素技術としては、異なるビジネス領域で同じ技術を用いている場合も少なくないが、システムとしては、基本的には、相互接続性を持たないものになっている。

(5)データ(Data)

データの保存と解析を、多様なデータに対して、多様な観点での解析が行われることが要求条件となり、基本的な方向性は、タグ付けされた構造化データ様式を用いたシステムとなるのが一般的である。しかし、タグ付けや構造化データの取り扱いは、処理オーバーヘッドが増加するのが一般的であり、高速化には、さまざまな工夫が行われているが、大規模システムの開発効率化とオープン化のための「モジュール化」をどのように行うべきなのか、研究開発の途上にある。

(6)ストーレッジ(Storage)

ファイルシステム、データ様式を含んだ、分散ストーレッジ化への本質的な対応が必要となる。実際には、非常に大規模なデータをハンドリングするシステムは、さほど多くないのが、実状であるので、多くの場合には、ストーレッジのパフォーマンンスが、システムのボトルネックとなってはいないのが、現状である。

第3章　2012年度WIDEボード合宿での議論の概要

本章では、プレゼンテーションされ、議論された内容について、「データと生成場所」、「解釈・分析技術」、「アプリケーション」、「実現技術」、「アーキテクチャ」、「プライバシ」、「運用」の7項目で整理する。

3.1　ビッグデータで扱われるデータとデータを生み出す主体

ビッグデータで扱われるデータの例として、プレゼンテーションで示されたものを列挙する。扱われるデータは広範囲にわたるため、様々な整理方法が可能である。ここでは、生み出される情報自身と、生み出す主体である場所やモノによる区分で示した。また、データはデータ自身に付随する情報としてメタデータとともに扱われる。これらメタデータについても例を示す。

(1)データの例

議論に出てきた、ビッグデータで扱われうる、あるいは、扱われているデータを以下に示す:

- 医療行為に関連した各種センシング情報
- 自動車に付随するセンサ
- 放射線
- 地震に関連したセンシングデータ
- Twitterによるツイート
- 気象情報
- ネットワーク(構成する要素についての情報、流れている情報についての情報、運用管理に関連する情報)

「ビッグデータ」とはいえ、それぞれのデータ自体は、かならずしも膨大であるとは限らない。ここで、ネットワークに関する情報は、WIDEプロジェクトが自身でもち、活用できる情報であり、今までも有効に活用してきたものと言える。

(2)データに関係するもの

データの生成されるモノあるいは場所として、以下が例として挙げられた:

- データを生み出すもの自身:
  - 人
  - センサ
  - ネットワークノード
  - サーバネットワークノード
  - サーバ
- 特定目的のセンサの集合体として:
  - 病院
  - 自動車
  - HEMS/BEMS
  - インターネットノード
  - ウェブサイト
  - クラウドソーシング

プライバシの視点ではあるが、データに関係する利害関係者の例として、以下の整理があった[1]:

- データ取得対象(Data Collection Target)
- データ収集者(Data Collectors)
- データマーケットとアグリゲータ(Data Markets / Aggregators)
- データユーザ(Data Users)
- データモニタ / プロテクタ(Data Monitors / Protectors)

(3)メタデータ

データは、メタデータとして、以下のような情報を付随して持ちうる:

- データソースにまつわる属性情報の例:
  - 生成者(生成場所)
  - 所有者
- データ自身にまつわる属性情報の例:
  - データ形式
  - 単位
  - タイムスタンプ
  - 精度に関する情報(信憑性/ソース信頼性、精度)

(4)プライバシ

プライバシの扱いは、ビッグデータにおいて重要な点である。

データ収集者(Data Collector)は、必然的にデータ取得対象(Data Collection Target)に対する全ての情報を持ち得る。そのため、データ収集者が情報を配布しようとするとき、プライバシを考慮して情報を適切に処理する、すなわち、プライバシを担保する手法の適用が必要な場合がある。

ここで、プライバシを担保するとは、個人を示すID(Personal Identity)と、誕生日・性別・郵便番号などによる準ID(Quasi Identity)、公開されてはならない情報(Sensitive Attributes)との関係において、公開されてはならない情報と個人を結びつけられないようにする、ということである。

この実現のための手法として、k-Anonimity[2]、l-Diversity[3]、t-Closeness[4]等の手法が提案されている。

3.2 解釈・分析技術

データは、機械学習やデータにタグ付けすることで特徴を抽出し、得られた特徴をもとに分析をおこなう。特徴抽出には、対象となるアプリケーション領域固有なノウハウが必要であり、ノウハウを持つ専門家を巻き込むのが難しいという点が指摘された。ここで、機械学習と解析ツールの細部の議論は行わなかった。一部、実現技術との関連でビッグデータに特化したものについてのプレゼンテーションがあった。

3.3 アプリケーション領域

アプリケーションは、特定のデータに紐付いた形で実現されている。すなわち、特定ドメインにおける「問い」に対し、個別に設計され、実装される。そのため、クローズシステムで、制限付き、特化された形での実現が一般的と考えられ、アプリケーション固有の議論となった。

以下のようなアプリケーションが例として示された:

- 小売り大手におけるシステム(シェルとパイプの組み合わせによる開発手法USP Lab.による)
- 教育における学習記録の活用
- 天文学における情報
- WIDEにおける分析対象のデータとして、ネットワークトラフィック情報

3.4 実現技術

ビッグデータを扱うシステムを構築するのにあたっては、High Performance Computing(HPC)領域や、ストレージにおいても、ネットワークを用いた分散システムによって構築するのが標準的といえる。

そして、アプリケーションアーキテクチャの視点では、大規模分散型のパラダイムが適用されている。それに伴い、様々な分散型システム最適化技術が適用可能と言える。

議論で現れたキーワードは以下の通りである:

- ハードウエア: HPC / ストレージ
- ソフトウエア: 分散パラダイム
- セキュリティ:クローズドシステム主体なので特別でない

- 名前付け: 実装依存で、クローズド
- ストレージ: サイズ / 信頼性 / 遅延 / 保存期間

構築を支援するためのソフトウエアとしては、ビジネスインテリジェンス向けの製品として、データマイニングに向いた形のデータベース製品があり、たとえば、カラム志向のデータベースなどが例としてあげあられる。加えて、昨今では、PFIのJubatus（http://jubat.us/）のように、ビッグデータ向きに最適化された機械学習システムもある。

MapReduce/Hadoopがビッグデータで引き合いに良く引き合いに出されるが、データ量に対する武器であり、データ量に対して対応することが可能になった点が重要だが、用いる資源に対し、効率が良いとは言えない。

3.5 アーキテクチャ

アーキテクチャに関連した議論では、以下のようなキーワードが見られた:

- パラダイム: 大規模分散 / エッジ・ヘビー / データのポータビリティ＆モビリティ
- モデリング: データ / データを生み出す対象自身（例: 空間とセンサの関係）
- ネーミング・識別子: グローバルvsローカル

この中で、特徴的なのは、「エッジ・ヘビー」型についての議論である[5]。

ビッグデータのシステムを、ネットワーク視点でみると、既存の、MapReduce/Hadoopを指す「ビッグデータシステム」は、1ヶ所に情報が集積されて処理されるため、ネットワーク的に見て近傍に集積されたクラスタで用いられる技術が主体であるといえ、広域分散処理とはなっていない。すなわち、各所で収集された情報をデータセンタに集約して処理するスタイルが主である。

これに対し、データのボリュームが増すにしたがって、情報のソースに近い点でデータを集積し処理をする、「エッジ・ヘビー型」へと変化していくという予測がある。

エッジ・ヘビー型のシステムの場合は、必然的に大規模広域分散型のシステムとなるため、より、ネットワークの用い方という視点での検討が重要になると考えられる。

3.6 運用

ビッグデータシステムの運用に関連したディスカッションでは、以下のキーワードが見られた。主には、データの扱いを中心とした視点であり、人材の不足における議論であったと言える。

- 運用ポリシー
  - 責任範囲・義務・権利
  - 業界 / 国
  - グローバル・オープン
  - データ保存
    - » 恒久保存
    - » メタデータの保存: セマンティクス・プロファイル / ポリシー含む
    - » オーナの特定
- 人
  - 統計学者・データサイエンティストの必要性
  - Chief Analytical Officerの価値
- 標準化・国際化
  - 対象領域に特化され、共通基盤となるものが少ない

第４章　WIDEが考えるビッグデータの検討項目

以下で、議論の中で重要と考えられた５つの点、「システム設計の見直し」、「データサイエンティストの必要性」、「エッジ・ヘビー」、「情報永続性とメタデータ」、「オープン性」それぞれについて簡単にまとめる。

4.1 ビッグデータのインパクトによるシステム設計前提の見直し

ビッグデータは、分散システム／ネットワークシステムのデザインに見直しを迫っている。たとえば、それは、処理、ストレージ、ネットワーク能力のバランスの変化に伴うものである。

ここで、何が変わってきているかに着目すると、1)データの生成、2)収集方法、3)データの保存方法、4)データ処

理と解析、ユーザとのインタラクションである。

あらゆる部分が変化しているとも言えるので、これまでのシステムモデルの前提を見直すべきである。

- データの生成と収集においては、machine-to-machine（M2M）、crowd-sourcing、キャリアなどのオンラインサービスプロバイダの持つユーザに関するデータ収集などが例にあげられる。
- データの保存においては、NoSQLデータベース分散ストレージといった、古いが、見直されている手法があげられるだろう。
- データ処理と解析においては、クラウドコンピューティング、分散処理（MapReduce等）などの構築手法にあわせ、データマイニング、機械学習、統計手法といった分野で、手法の変化が見られる
- ユーザとのインタラクションという点では、ユーザにカスタマイズされたサービス（オンラインお勧めシステム、ソーシャルメディア等）、継続的に生成されるデータの解析処理の必要性、オンライントレード、オークション、アドサービスなどのリアルタイムトランザクション、データの可視化等の視点が重要である。

4.2　データサイエンティストの必要性

一連の議論で、特に強調された点は、ビッグデータシステムにおいては、データありきの解析、すなわち、「手元にこのデータがあるので、何か良い情報が得られないか」というアプローチは、意味ある結果を得られないことが多いという点である。

このアプローチを取るためには、大元の「問い」を起点とし、解析対象となるデータの領域知識（Domain Knowledge）を持ち、情報分析能力をもつもの、すなわち、データサイエンティストによって分析プランを立案するのが重要である。

一方、データサイエンティスト自身が不足しているので、育成が急務であるという認識であった。

4.3　エッジ・ヘビー

アーキテクチャの節で取り上げたように、処理するデータセンタに集約するのではなく、データの出元に近い場所で集約処理する「エッジ・ヘビー」アーキテクチャが求められるという予測がある。

この方式には一定の利点があると考えられるので、今後、データの出元と処理する場所の関係については、議論してゆく必要がある。

4.4　情報永続性とメタデータの重要性

多量のデータという視点で国立天文台の情報とメタデータの重要性についての議論があった。

天文台のデータは、ロスレスで情報を保持する必要があるということと、データについての充分な情報をメタデータとして持たないと、情報として役に立たないという意見があった。そのため、データは、記録されるだけでなく、将来有効活用するために必要な情報を合わせて持つことが重要である。

4.5　オープン性の追求

データの有効利用には、多目的・複数参加者によるデータ利用や、業界間を跨いだ水平統合の追求が必要である。様々なデータを広く提供するための、オープンデータ志向が重要である。

そして、これらのデータを活用するためのシステムやアルゴリズムを有効活用するためには、システムのオープン性も重要である。

第５章　WIDEとしての取組み

ビッグデータは、言うならば、データ・ルネッサンスの実現である。データを様々な形で利用可能とし、それらを、広域分散システムによる、ふんだんに用意できる資源を背景に、様々な手段で処理加工し、ユーザで利用できるようにすることにより、様々な事柄に対して、検証可能性を高めることが可能となるだろう。

これにより、現在のWIDEプロジェクトにおけるキーワードの一つである、「検証可能な研究（Evidence Based

Research)」が可能となる。

WIDEとしてのビッグデータに関する研究開発のスコープと方向性は、以下に整理される。

(1)グローバル

地理的なグローバル性、すなわち、遅延とそれに伴う同期の問題にどのように対応、解決するのか。また、論理的な観点でのグローバルの意味を考え、どのように実現すべきなのか。

(2)システム間、領域間での相互接続性

垂直統合型のシステム設計・構築・運用を、どのようにして、水平統合型、水平連携型にすることができるのか。

(3)分散データ処理

グローバルドメインでの分散データ処理の実現。

(4)オープン性

データは、関係するステークホルダによって、「利用」、「加工・処理」、「アクセス」可能にするべきである。

- 利用権:データは、システムに害を与えない限り、すべての人が利用可能な権利を持つ。
- アクセス権:ネットワークを通してデータを獲得可能にする。
- 加工・処理権:獲得したデータの解析・加工を行う権利を持つ。

(5)透明性

(6)標準化

ドメインごとの標準化ではなく、ドメインに共通する技術の標準化を進める。

(7)データの信憑性

科学技術的な根拠に基づいたデータの提示が行われるような仕組み・体制を作る必要がある。また、誰でも、提示されたデータの検証(再計算による再検査など)を行うことが可能な、データ自身とデータ処理方法に関する透明性を担保する必要がある。

(8)システム開発(running codeを第一に)

- 自律的なシステム構築が、全体のシステムの高度化・高機能化に貢献するような構造を目指すべき。
- 最初から、完璧で詳細なシステム設計を行うことは、適切ではない。これまでの、経験と見識をもとに、最小要求条件を満足するシステムからスタートし、運用を行いながら、必要な機能拡張・修正が可能なアーキテクチャの設計を行うべきである。

(9)データ収集の継続性

データの保持と消去は、基本的には、運用者の自律的な判断に任せるべきであり、すべてのシステムにおいて、義務化・ディフォルト化するべきではない。たとえば、「データを消す権利」を一般的に適用することは、システムの経済面での問題を発生させてしまう。一方で、各データのオーナーを把握することは、非常に重要な要求条件となる。データの人格権を定義するよりは、データは誰に属しているかという観点で設計することがよいかもしれない。

(10)ストーレッジの重要性

データの広域(グローバル空間)での移動(Portability, Mobility, Latency)に対応可能なシステムを設計・構築する必要がある。

(11)データの収集・転送・分析

データ自身が、その属性と処理方法の情報を持つなど、データ自身の高度化が必要かもしれない。

(12)名前と識別子

特定ドメインに閉じない、グローバルな空間での名前と識別子が利用される環境の構築が重要である。

(13)データ表現法

データの持つポリシーやプロファイルが表現可能な様式である必要がある。システム側が、データの様式を暗に把握して、動くシステムでなく、データ自身がシステムに、その処理ポリシー・方法を陽に提示するような形式が必要とされる。

(14)運用ポリシー

今後の課題。責任と権利の関係は、組織によって異なるものである。

(15)秘匿性

運用ポリシーとも深く関与する問題であるし、透明性との関係もしっかりと議論・整理されなければならない。

WIDEプロジェクトとしては、上記の整理に基づき、ビッグデータに関する研究開発を推進するが、ビッグデータ

の研究開発に関する技術的な課題やポリシーなどの課題は、これまで、WIDEプロジェクトが行ってきた研究・開発・運用の方向性そのものであり、また、「Evidenced-Based Research」の方向性そのものである。ただ、システムの構成要素の技術的条件は、継続的に変化しており、それに対応した、具体的なシステムの再設計や修正を行う必要がある。さらに、ビッグデータの領域は、インターネットによって構築された、グローバルな空間に透明性を持ったデジタルデータの流通基盤を、フラグメント化しようとしていると捉えることもできるであろう。各特定ドメインに閉じたシステムの構築ではなく、共通の技術を適用した、ビッグデータシステムの構築を行い、さらなる、イノベーションを継続することを可能にするための研究開発を行うべきである。

第６章　Appendix 1：2001年以降のWIDEボード合宿の検討テーマ

- 2001年：インターネットセキュリティー
- 2002年：Peer-to-Peerアーキテクチャ
- 2003年：モバイルリアリティー
- 2004年：ラムダネットワーク
- 2005年：Gridコンピューティングの解剖
- 2006年：インターネット的な無線技術
- 2007年：Future Internetアーキテクチャ
- 2008年：「次」のインターネット
- 2009年：2010年以降の体制
- 2010年：Hand-Madeシステム
- 2011年：震災対応システム
- 2012年：ビッグデータ

第７章　Appendix 2：開催の詳細

- 場所:8月10日～ 8月12日長野県野辺山（八ヶ岳グレイスホテル）
- 参加者23名（ゲスト 4、WIDEメンバ 3、ボードメンバ 16）
- 議論の題目:今から10年後の情報とインターネットのありかたについて幅広い視点で議論
- キーワード：
  情報システムパラダイムの変革、ビッグデータ、データ活用の発想の転換、クラウドコンピューティング、High Performance Computing、目的特化型データベース（NoSQL）、セキュリティ、プライバシー、Evidence Based Research
- ゲストによるプレゼンテーション
  - USP Lab.：シェルスクリプトによるアプリケーションプログラミング
  - 国立情報学研究所：Hadoop動向、インターネットトラフィック分析、ビル環境情報のモデリング
  - Preferred Infrastructure：ビッグデータ分析技術、Jubatasによるリアルタイム分析
  - トヨタITC：ビッグデータ処理に関連したオープンソースについて
- ボードメンバによるプレゼンテーション
  - 対象データ
    - » データルネッサンスとネット中立性
    - » データ量の膨大化と医療
    - » 自動車、Twitter、放射線、地震
  - データ解釈技術とアプリケーション
    - » Data Collection、Analysis and Application
    - » Education area
    - » Internet measurement and (big) data
    - » Privacy and Anonymization
  - 実現技術
    - » High Performance Computing
    - » DSP/RTB、application example for (big) data
    - » Storage
  - アーキテクチャ