招 待 論 文

社会の安心・安全に向けたビッグデータ処理ネットワークの課題

情報セキュリティ大学院大学
教 授 後藤 厚宏

1. はじめに

「大量の情報から新たな価値を創出するビッグデータ時代の到来」が昨今の話題である．宇宙空間の観測データからわれわれ人間の活動ログ（ライフログ）に至る実世界のセンサ情報を活用して，より効率が良く高度な社会を実現する潮流である[1]．従来から，マーケティングや企業戦略システムとして，さまざまなデータ分析手法が開発され活用されてきた．そのような分析手法が，農業・漁業，運輸などの分野に適用され，生産性向上の鍵として期待されている．同様に，ビッグデータは，社会の安心・安全のためにも活用されつつある．米国では，Twitter (注1) や検索キーワードのログ分析によって，医療機関における診療状況よりも早く，インフルエンザの流行の予兆を見つけることができたと報告がある[2].

このようなビッグデータが，社会基盤の一部として，われわれの社会生活や産業を支える「新たな価値」を生み出すのであれば，当然ながら，それを支えるシステムは信頼性が高く，セキュアである必要がある．本稿では，ビッグデータ処理ネットワーク全体の安心・安全を議論する土台として，ビッグデータ処理を**第1図**のように5つのフェーズ：センシング（Sensing），トランスポート（Transport），クレンジング（Cleansing），分析（Analyze），応用（Apply）の流れとして捉え，それぞれの特徴と技術課題を検討しながら，ビッグデータをわれわれの社会生活の安心・安全に生かすには何をなすべきかについて考えたい．

2. センシングから価値創造へ

ビッグデータのソースは，外気温計や地震計などの自然環境のセンサ，電力計や携帯端末のGPS（Global Positioning System）などの組込み機器，Web履歴，クレジット履歴など人の社会活動のログなど，さまざまなものがある．

- 自然環境系のデータ

外気温や地震計のデータに代表されるストリームデータであり，センサネットワークによって収集・集約される．

ビッグデータ処理のゴールは，大規模サイエンスから，農業・水産業・エネルギー産業など多岐にわたる．共通する課題は，データソースが地球規模（将来的には宇宙規模）で分散され，さらにデータ量が爆発的に増加することである．見込まれるデータ量は，ネットワークを介して送受しきれる規模を容易に超えるといわれており，広域センサネットワークのなかで「処理（の一部）をどのように分散配置するか」が将来課題である．

- 社会活動系のデータ

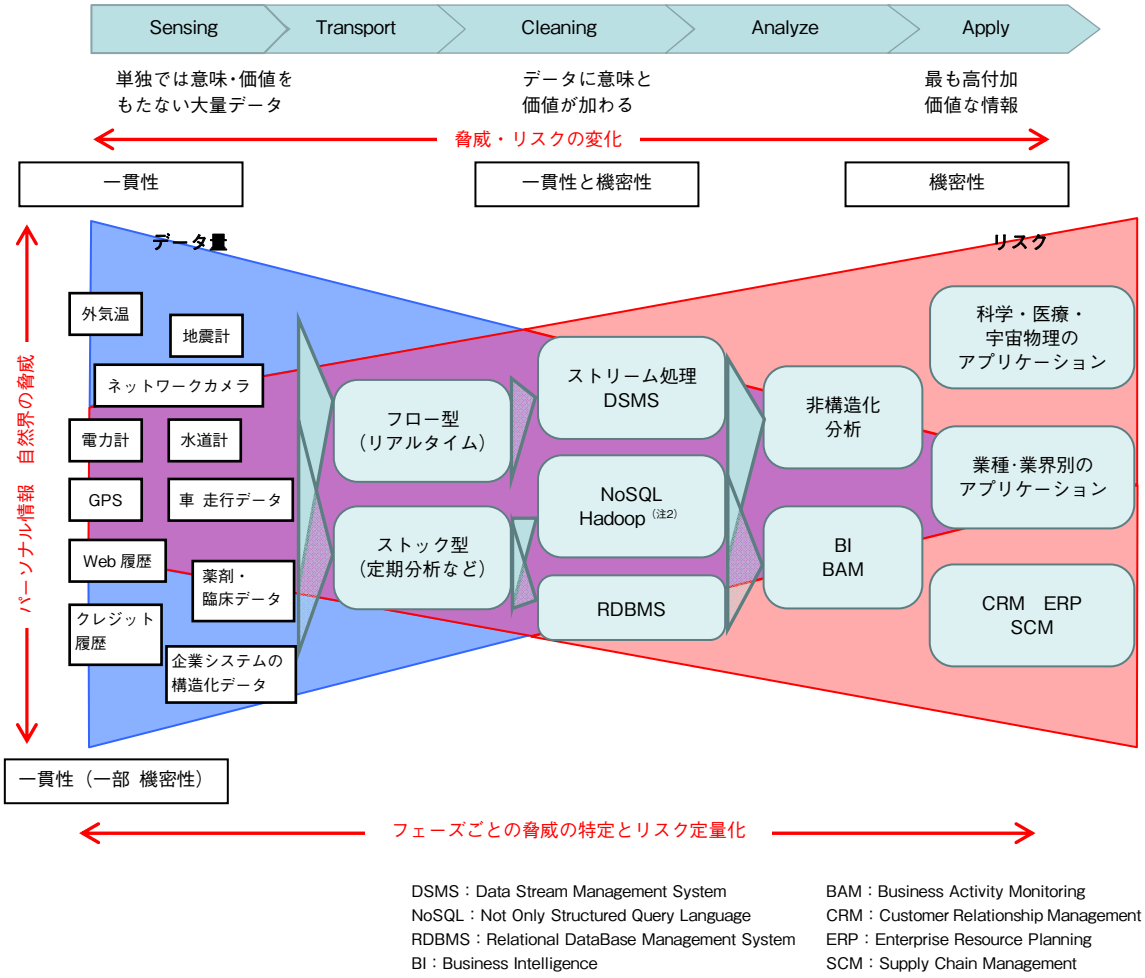
Webアクセス履歴，TwitterやBlogの発信ログ，クレジットカードの利用履歴や，医療分野における薬剤・臨床データなど，われわれの社会活動に直結するデータである．

既に，産業としてWeb履歴などを多用なマーケティングに活用する動きが盛んである．このような社会活動系データの利活用促進における課題は，多様なステークホルダーが介在するエコシステム（4.3節）を，安心・安全なネットワークとしていかにして構築するかである．

- 組込み機器のデータ

自然環境系と社会活動系の中間的な性質をもつ組込み機器のデータは利活用の期待が高い．家庭やオフィスビルでの省エネが重視されてから，電力計，水道計など，社会のなかに埋め込まれたセンサをソースとするシステムが拡大しつつある．広域ネットワークを介して広範囲のデータを収集できる環境を作ることにより，スマート

（注1）Twitter, Inc.の米国またはその他の国々における登録商標または商標



第 1 図　ビッグデータの処理プロセス

シティとして地域社会全体の最適化が期待できる．

組込み機器のデータは，自然環境系のセンサと同様に，緯度経度，W/h，m^3/hなどの物理的情報であるが，それらは人間行動・社会活動に対応したものであることが違いである．例えば，センサ機器としてのライブカメラは，火山の火口を常時監視することにより安心・安全に寄与し，街中やイベント会場では防犯カメラとして地域の安心・安全や犯罪捜査に役立つ．両者の違いは，プライバシーへの配慮の必要性の有無である．

3. セキュリティリスクの特徴と変化

ビッグデータ処理ネットワークのセキュリティリスクは，基本的に取り扱う情報の質と量に依存する．センシング段階では，一般的に単位量当たりの情報の価値は相対的に低く，クレンジング，分析，応用とビッグデータの処理プロセスの下流に行くほど，生成されるデータ量は絞り込まれ，情報としての価値が高まると考えるのが妥当であろう．このため，フェーズごとに情報資産としての価値（リスク）を判断し，脅威を特定するモデルを構築することがシステム全体のセキュリティ対策投資の効率化・最適化につながる．例えば，暗号化やアクセスコントロール，監査といった各技術対策要素をどの程度実装すべきか，というガイドは，それぞれのフェーズにおける情報の質・量・構造化の度合いから見積もる被害コストとバランスがとれる指標として作成する必要がある．

- 初期段階（センシング，トランスポートのフェーズ）

Web上から収集したテキストデータ，自然環境や機器に組み込まれたセンサが収集したデータの情報資産としての価値は，センサネットワークなどの収集コストとデータ量が主要素となる．単独の情報価値が低いため機密性は重視されないが，4.1節で議論するように，一貫性と

（注2）Apache Software Foundationの商標または登録商標

可用性は求められる．例えば，気象の急激な変化を捉えて避難誘導するようなシステムを想定した場合，ソースとなるセンサデータに関して，正しいセンサからの正しいデータであるかどうかは，システムとしてのアウトプットである「避難誘導の是非」を左右するためである．

社会活動系のデータは，個人に直結するパーソナルデータである場合が多いため，自然環境系と異なり，当初から機密性が高い．さらに，医療分野での臨床データの場合などは，フロー型のデータの場合もあり，環境センサと同様に，一貫性，可用性も求められる．

- 情報が整理された状態（クレンジング・分析フェーズ）

Hadoopプラットフォームで処理された出力結果など，大量のデータが整理され意味と価値をもち始めるフェーズである．例えば，断片化していた個人に関する履歴情報や他の情報が関連づけられて，プロファイルデータとなりうるため，データ管理が不十分であるとプライバシー侵害の懸念につながる．

- 新しい情報が生まれる応用フェーズ

BI（Business Intelligence）や応用領域ごとの分析システムの出力など，アプリケーションシステムを通じてビジネスに活用される状態であり，企業活動などではトップレベルの意思決定の材料にもなる．

企業・組織が独自の手法・ノウハウなどで生み出す最も高付加価値な情報として，企業としての競争力の源泉になる．初期段階では機密性がなかった情報も，この段階になると機密情報となる．ここで生まれる価値ある情報はアナリストによる試行錯誤の結果である場合も多く，その情報価値の大小が事前に見えにくい．

4. セキュリティ課題と対策技術

4.1 情報の品質の確保

ビッグデータ処理においては，最終的な分析結果の価値を左右する情報の品質（Information quality）の確保が最も重要である．品質を構成する要素はデータの「由来・系譜（Data provenance）」と「ノイズ」である．

- 由来と系譜

データの由来・系譜の信頼性は，センサそのものの認証や，データの一貫性（改竄（かいざん）されていないこと）の証明によって確保する．後述するビッグデータのエコシステムにおいては，一次事業者が確保したデータの由来・系譜の信頼性を，適切な契約行為によって二次事業者に引き継ぐプロセスが重要となる．

自然環境向けセンサネットワークのためのセキュア通信プロトコルは，これまでに多くの研究がなされている[3]．その初期に提案されたSNEP（Sensor Network Encryption Protocol）では，ブロードキャストによるセンサネットワーク通信において，データ機密性（盗聴対策）や一貫性の証明（偽造データ，古いデータ，繰り返しデータでないこと）を小さいオーバヘッドで提供する方式が提案されている．

スマートフォンなどのGPSによる位置情報取得やライブカメラなどの組込み系および社会活動系のセンサネットワークでは，インターネットプロトコルが主流である．音声・映像データの安全な送受のために，SSL（Secure Socket Layer）通信を高効率・低負荷で実現できる技術（本特集[4]）も広まっている．

センサから収集される大量のデータの一貫性を証明するストレージ証明技術[5]が期待されている．本技術は，クラウド事業者 が，クライアントから預かった“巨大データ”を（捨てたり壊したりしないで）完全に保持していることを“小さい証明書”で保証するプロトコルである．

- ノイズの除去と付加

もう1つの品質の要素であるノイズは「除去」と「付加」の両方がある．自然環境系の計測データなどでは，センサ特性としてのノイズは避けられない．加えて，センサの故障などによる情報の欠損，誤情報の混入も想定する必要がある．このようなノイズの除去は，クレンジングフェーズの役割の1つである．

一方，プライバシー保護のためにノイズを付加（匿名化[6]）することもクレンジングフェーズの重要な役割の1つである．街頭カメラや店舗内カメラの映像などにおいては，自動認識技術を活用して人物にぼかしを入れることも可能になっている[7]．ノイズの付加によってデータの品質は下がるが，プライバシー侵害のリスクも下げられるため，応用面で価値創造の自由度を高める効果が期待できる．

4.2 分析プラットフォームのセキュリティ

分析フェーズでは，データ分析がストック型かフロー型か，および対象となるデータの構造化の度合い（構造化データ，非構造化データ）によってシステムアーキテクチャが異なる（**第1図**）．

構造化データを対象とするストック型の分析システムの代表は，BAM（Business Activity Monitoring）やBIなどの企業情報システムである．これらのプラットフォームには，RDBMS（Relational DataBase Management System）などの既存システムが活用できる機会は多い．

一方，同じストック型でも，大量の非構造化データにはHadoopに代表されるNo-SQL（Not Only Structured Query Language）システムが広く活用されている．正にビッグデータ向けのプラットフォームとして，オープンソースソフトウェアも多数実用に供されているが，セキュリティ面では，実環境での「鍛錬」が不足している懸念が指摘されている[8]．これは，RDBMSが企業システムや金融システムとして，数十年にわたり実環境にて鍛えられてきたのに対し，No-SQLシステムの歴史が浅いことに起因する．

ストック型の分析プラットフォームでは，ストレージからの情報漏えいリスクを軽減する秘密分散技術[9]，機密保持したまま分析を可能とする秘匿計算[10]の研究が盛んである．両者とも正にビッグデータに向けた性能向上が必要である．GPSの位置情報などを，提供者のプライバシーを保護したまま効率よく分析できるPAT（Privacy Aware Transformation）技術も提案されている[11]．これらの技術は，変換時に分析アルゴリズムが固定されるが，PPDM（Privacy Preserving Data Mining）[6]の摂動法と異なり，データの精度を維持したまま元データの秘匿が可能である．

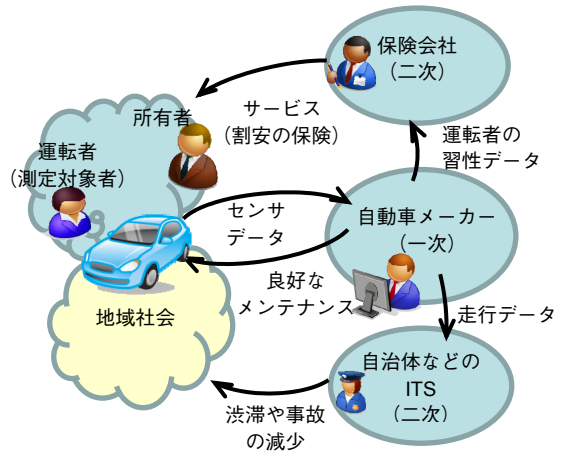
非構造化データを対象とするフロー型の分析プラットフォームDSMS（Data Stream Management System）は，主に on-memory でストリームデータを扱う[12][13]．将来の主役として期待されるが，技術面ではコスト性能比の向上が重要である．

4.3 価値創造を高めるエコシステム

ビッグデータを社会の安心・安全に生かすうえで，そのエコシステムを注視する必要がある．収集されたセンサデータを活用して利用者にサービスを提供する一次事業者に加え，その解析結果（統計データ）や一次処理後のデータを受けて付加価値サービスを提供する二次事業者など，さまざまなステークホルダーが，ビッグデータのエコシステムに登場する．

第2図は，自動車の制御用センサを例に，センサの保有者（=自動車の所有者），設置者（=自動車メーカー），測定対象者（=運転者）とサービス提供者（一次，二次）とその受益者の関係を示している[14]．

近年の自動車では，エンジンやメカ部分に多数のセンサが埋め込まれており，これまでは走行時や保守点検時に閉じた環境で利用されてきた．これらのセンサ情報を，ITS（Intelligent Transport Systems）の入力情報として実時間に集約し，道路交通の安全や渋滞回避に活用する試みがある．この場合は，地域の自治体がセンサデータの二次事業者であり，運転者を含む地域住民が受益者となる．



第2図　自動車制御用センサデータの活用事例

走行距離だけでなく，運転者の運転習性データ（加速の仕方，ブレーキの緩急など）を制御用センサから取り出して保険会社に提供することより，自動車の保険料の割引や加算に活用する，本格的な“pay-as-you-drive（PAYD）”のサービスも現れている．

この例は，センサの設置とデータ収集の当初目的が自動車本体のメンテナンスにあるものが，地域社会の（交通）安全と割安や自動車保険という異なる目的（と受益者）に活用するという価値創造の例である．

このようにビッグデータのエコシステムを拡大することによって多様なステークホルダーによる価値創造の自由度が高まる一方，システム全体のセキュリティ設計の難しさも増すことになる．

4.4 適応的なセキュリティの重要性

ITシステムでは，最初にシステムのゴールがあり，そのゴールを達成するためのアプリケーションを開発し，そのアプリケーションにデータを与えてゴールを達成するという手順が基本である．一方，ビッグデータにおいては，収集した巨大データに対して，分析アプリケーションをいろいろと試しながら，探索的にゴールを求める場合もある．

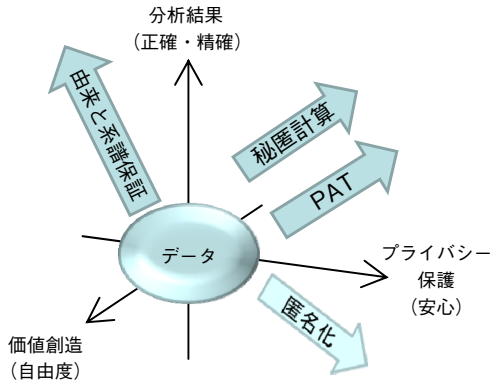
セキュリティ対策の基本は，情報資産と脅威の特定，リスクの見積もり，そのリスクに見合った対処策のProactiveな作り込みである．このため，上述のような探索的な状況では，システムにおける「情報資産の価値」を事前に特定することが難しいため，セキュリティ設計にとって厳しい環境と言える．

前節のエコシステムのように異なる業種の事業者にまたがってデータが活用される場合，一次事業者が二次事業者以降における情報資産の価値が事前に見積もれない

ことに相当する．多様なステークホルダーからなるビッグデータのエコシステムにおいて自由度の高い価値創造を可能とするためには，柔軟性のあるセキュリティシステムをいかにして実現するかが重要になる．例えば，後段の事業者から前段の事業者へデータ活用の仕方についてフィードバックし，それに基づいてデータ匿名化やセキュリティ対策のパラメータを適応的に変更するような仕組みが求められる．

4.5 セキュリティ技術の特徴づけ

ビッグデータ処理ネットワークのセキュリティ対策技術を，分析結果（正確・精度），プライバシー保護（安心），価値創造（自由度）の3軸で特徴づけたものを**第3図**に示す．データの由来・系譜の保証は，分析結果の精度と価値創造の必須要件である．秘匿計算やPAT技術は，分析精度とプライバシー保護の両立を狙うが，分析の自由度は制約を受ける．プライバシー保護のための匿名化は精度とのトレードオフとなる．一方，適切な匿名化によってエコシステムでのデータの流通が可能になれば，さらなる価値創造が期待できる．



第3図　ビッグデータの技術課題の特徴づけ

5. おわりに

ビッグデータを社会の安心・安全に生かすために，それを支えるシステム全体のセキュリティ確保が重要である．センシングから応用に至る間で情報としての価値と特性が変化することを考慮して，フェーズごとに脅威を特定するモデルを構築し，適切なセキュリティ対策を導入することが必要である．また，対策技術のさらなる高度化が期待される．

参考文献

[1] ソフトバンクビジネス＋IT, “サイバーフィジカルシステムとは何か：東京大学 教授 喜連川 優氏に聞く,” http://www.sbbit.jp/article/cont1/22970, 参照 Oct. 21.2013
[2] Son Doan, et al., “Enhancing twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses,” IEEE 2nd International Conference on Healthcare Informatics, Imaging and Systems Biology, Sep. 2012.
[3] Adrian Perrig 他，ワイヤード／ワイヤレスネットワークにおけるブロードキャスト通信のセキュリティ， 溝口文雄（訳），共立出版，2004.
[4] 田中裕之　他, “音声・映像機器の暗号技術,” パナソニック技報，vol.59, no.2, pp. 23-28, 2013.
[5] 有田正剛, “暗号理論教室 - ストレージ証明,” http://www28.atwiki.jp/cryptospace/pages/155.html, 参照 Oct. 21.2013.
[6] Charu C. Aggarwal et al., Privacy-preserving data mining: models and algorithms, Springer, 2008.
[7] 新井啓之 他, “インテリジェントな映像モニタリングを目指して,” NTT技術ジャーナル，vol.19, no.8, pp.8-12, 2007.
[8] Jeremy Glesner et al., “Security considerations for federal Hadoop deployments”, Hadoop World 2011.
[9] 松尾正克　他, “排他的論理和を用いた(k,n)しきい値秘密分散法”, パナソニック技報，vol.59, no.2, pp. 29-34, 2013.
[10] NTT報道発表資料, “医療統計処理における秘密計算技術を世界で初めて実証”, 2012年2月，http://www.ntt.co.jp/news2012/1202/120214a.html, 参照 Oct. 21. 2013.
[11] Ikuo Nakagawa, et al., “PAT: Privacy aware transformation method and applications for location services”, 電子情報通信学会技術研究報告, vol. 113, no. 94, IA2013-6, pp. 31-36, 2013.
[12] 櫻井 保志, “時系列データのためのストリームマイニング技術,” 情報処理，vol. 47, no. 7, pp. 755-761, 2006.
[13] 白石 陽, “センサネットワークのためのデータベース技術,” 情報処理，vol. 47, no. 4, pp. 387-393, 2006.
[14] 麻生享路 他, “センサデータを活用する社会に向けたプライバシーに係る課題の多角的考察,” 電子情報通信学会技術研究報告, vol. 112, no. 463, NS2012-284, pp. 693-698, 2013.

《プロフィール》

後藤　厚宏（ごとう　あつひろ）
1979　東京大学 工学部　電子工学科卒業
1981　東京大学大学院 情報工学専攻 修士課程修了
1984　東京大学大学院 情報工学専攻 博士課程修了
　　工学博士
1984　日本電信電話公社 武蔵野電気通信研究所 入社
1985-2011　日本電信電話（株）にて研究開発に従事
2007-2010　NTT 情報流通プラットフォーム研究所長
2010-2011　NTT サイバースペース研究所長
2011-現在　情報セキュリティ大学院大学　教授

専門技術分野：
　クラウドセキュリティ，ID 管理，並列分散処理