8 ビッグデータ活用におけるガバナンス

後藤 厚宏（情報セキュリティ大学院大学）

社会に浸透するビッグデータ

宇宙空間の観測データから我々人間のライフログに至る実世界のさまざまなデータを活用して，より効率が良く高度な社会を実現する「ビッグデータ」が注目を集めている．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．これを，農業・漁業，運輸などの幅広い分野に適用することで，生産性の向上が期待されている．企業組織においては，その情報システム（IT）の利活用の最大化を検討するITガバナンスの対象が大きく膨らむことに相当する．

このようなビッグデータが，社会基盤の一部として，我々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，ビッグデータを取り巻くシステムや制度の情報セキュリティガバナンスが重要となる．

本稿では，ビッグデータ活用の安心・安全を議論する土台として，ビッグデータ処理を**図-1**のように5つのフェーズ：センシング，トランスポート，クレンジング，分析，応用の流れとして捉える．次章ではビッグデータのソースの特徴について，続いて5つの処理フェーズにおけるリスクの変化について示す．データ品質の確保と処理プラットフォームの技術課題を検討し，エコシステムと適応的セキュリティの観点から，ビッグデータを我々の社会生活の安心・安全に活かすには何を為すべきか，ビッグデータ活用におけるガバナンスについて考えたい．

ビッグデータのソース

ビッグデータのソースの特性を，自然環境系，組込み機器系，社会活動系の3つに分けて見てみよう．

■ 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・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）で分散し，データ量が爆発的に増加することである．このため，地域ごとのデータ管理主体が連携できる仕組みが必要である．

■ 組込み機器系のデータ

電力計や携帯端末のGPSなどの組込み機器系のデータは，自然環境系のデータと同様に，緯度経度，Watt/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異なる．これに伴い，プライバシーに配慮する必要性が生じる．

■ 社会活動系のデータ

Webやクレジットカードの利用履歴，医療分野における薬剤・臨床データなど，我々の社会活動に直結するデータである．すでに，産業としてWebアクセス履歴などを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をケアする安全なエコシステムの構築である．

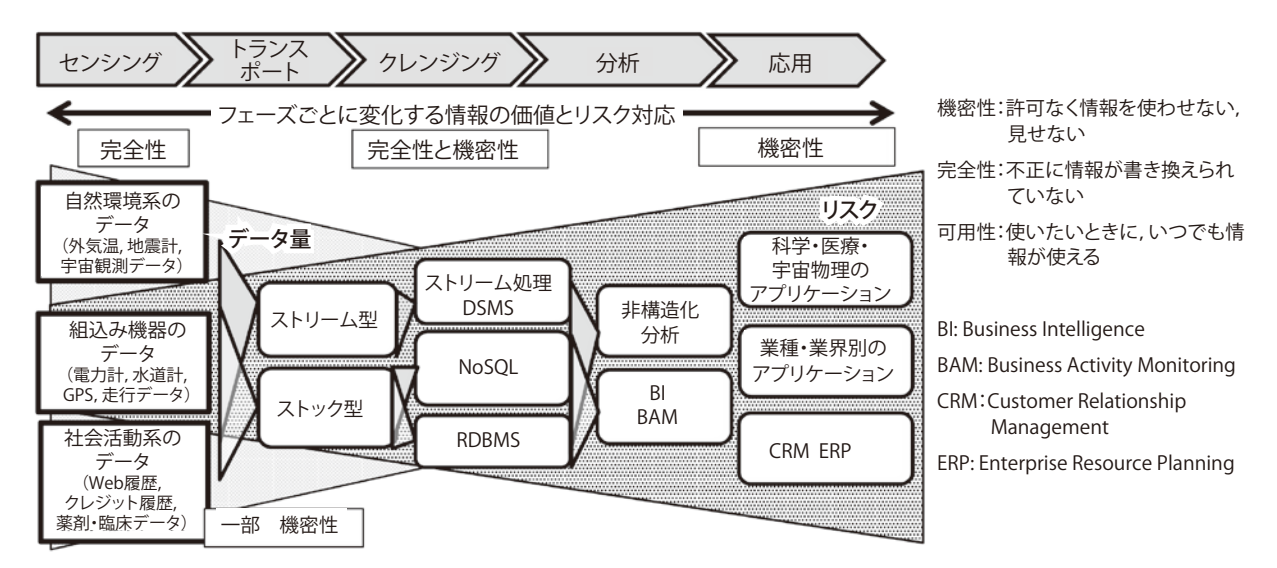

図 -1　ビッグデータの処理フェーズとセキュリティリスク

処理フェーズとリスクの変化

図 -1 のビッグデータ処理の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．このため，情報セキュリティガバナンスにおける情報・脅威の特定とリスク対応をフェーズごとに行うことがセキュリティ対策投資の効率化・最適化につながる．

■ 初期段階（センシングとトランスポート）

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としての価値は，センサネットワークなどによる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，完全性と可用性は求められる．たとえば，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の場合，ソースが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．

社会活動系のデータは，パーソナルデータである場合が多いため当初から機密性が高い．さらに，医療分野のストリーム型の臨床データの場合は，環境センサと同様に，完全性，可用性も求められる．

■ 情報が整理された状態（クレンジングと分析）

大量のデータが整理され意味と価値を持ちはじめるフェーズである．たとえば，断片化していた個人に関する履歴情報やほかの情報が関連づけられるため，個人のプロファイルデータとしてのデータ管理が求められる．

■ 新たな価値の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，企業活動などでは経営層による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．このため，情報の価値判断と管理方法の変更を即時に実施できる手順と体制が重要になる．

データ品質の確保

ビッグデータ処理の前半部分では，情報の品質の確保が重要である．品質を構成する要素として，デ

ータの「由来・系譜（Data provenance）」と「ノイズ」について考える．

■ データの由来と系譜

データの由来・系譜の信頼性は，センサそのものの真正性や，センサからネットワークを介して収集されるデータが途中で改竄されていないこと（完全性）を証明することで担保する． 大量のデータの完全性を証明する技術の１つにストレージ証明技術[1]がある．この技術は，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．

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ことが重要となる．

■ ノイズの除去とノイズの付加

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や，センサの故障などによる情報の欠損，誤情報の混入を想定する必要がある．このようなノイズの除去は，呼び名の通り，クレンジングフェーズの役割である．

一方，プライバシー保護のためにノイズを付加（匿名化）することもクレンジングフェーズの重要な役割である．街頭カメラや店舗内カメラの映像等においては，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[2]．匿名化という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が低減できるため，価値創造の自由度を高める効果が期待できる．

処理プラットフォームの課題

分析フェーズでは，データ分析がストック型かストリーム型か，対象データが構造化データか非構造化データかによって利用されるシステムが異なる．

ストック型構造化データの分析システムは，BAM（Business Activity Monitoring）やBI等の企業情報システムの延長線にある．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される場合が多い．

一方，大量の非構造化データにはRDBMSではないNoSQLシステムが広く活用されている．すでに多数のオープンソースソフトウェアが，正にビッグデータ向けのプラットフォームとして，実用に供されている．ただし，それらのNoSQLシステムは歴史が浅いため，実環境での「セキュリティ面の鍛錬」が不足している懸念もある．

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し可用性を高められる秘密分散技術，機密保持したままデータ分析を可能とする秘匿計算技術[3]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．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，匿名化と異なり，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ストリーム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[4]．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．

エコシステムと適応的セキュリティ

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，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（高度道路交通システム）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なされている．加えて，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，ブレーキの緩急等）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サービスも現れている（**図-2**）．この例で注目すべきは，自動車本体の保守のために設置したセンサのデータが，地域社会の（交通）安

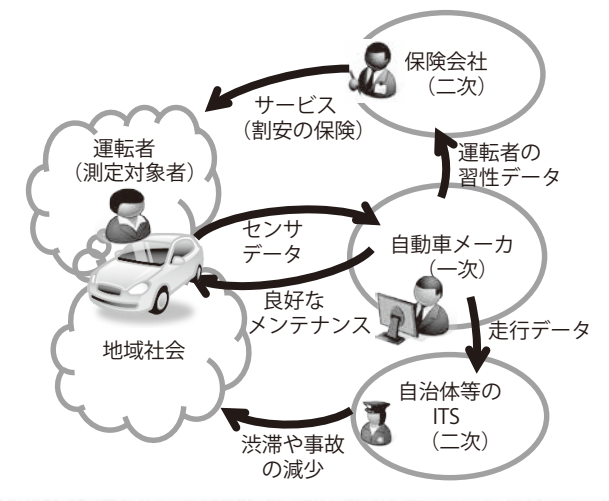

図 -2　ビッグデータのエコシステム例

全と割安な自動車保険という異なる目的（と受益者）に活用されている点である[5]．このようにビッグデータのエコシステムの拡大によって多様なステークホルダ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．

■ 適応的なセキュリティの重要性

セキュリティ対策の基本は，情報と脅威の特定，リスクの分析，そのリスクに見合った対処策（リスク対応）の先を見越した（Proactive な）作り込みである．一方，ビッグデータ処理においては，収集した巨大データに対して，分析アプリケーションを色々と試しながら，探索的にゴールを求める場合が多い．このため，システムにおける「情報の価値」を事前に特定することが容易ではなく，セキュリティ設計にとって難しい環境となる．

異なる業種の事業者にまたがるエコシステム（図 -2）の場合はさらに難しくなる．一次事業者が二次事業者以降における情報の価値を事前に見積もれないことに相当するためである．多様なステークホルダからなる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，事前の作り込みに加え，適応的なセキュリティシステムの実現が重要

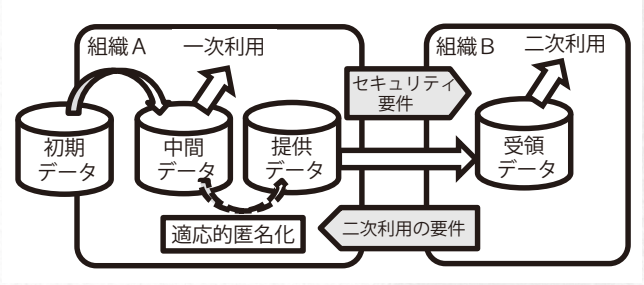

図 -3　適応的セキュリティ

になる．たとえば**図 -3** のように，組織 B が事業者のデータの二次利用の仕方を「要件」としてフィードバックし，組織 A は，それに基づいて適応的にデータを匿名化し，組織 B に求める「セキュリティ要件」を伝達する仕組みである．

ビッグデータに求められる取り組み

ビッグデータを活用した自由度の高い価値創造を進めるためには，それを支えるシステムのガバナンスにおける新たな取り組みとして，①センシングから応用に至る間で情報の価値と特性がダイナミックに変化することを考慮すること，②多様なステークホルダによる価値創造に柔軟に対応できる適応的なセキュリティ対策を実現すること，の 2 つが鍵になる．

参考文献

1) 有田正剛：ストレージ証明，http://www28.atwiki.jp/cryptospace/pages/19.html
2) 新井裕之 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 (Aug. 2007).
3) 千田浩司 他：高機密データも安全に二次利用可能な「秘密計算技術」，NTT 技術ジャーナル (Mar. 2014).
4) 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 (15 July 2006).
5) 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，信学技報，Vol.112, No.463, NS2012-284, pp.693-698 (Mar. 2013).

(2015 年 5 月 14 日受付)

後藤 厚宏（正会員）goto@iisec.ac.jp
並列・分散処理，インターネットセキュリティ技術の研究開発等に従事．2011 年より情報セキュリティ大学院大学教授．本会フェロー．現在，本会理事．