

ビッグデータの利活用が創る スマートな社会とビジネス

Smart Society and Business Creation Using Big Data

三木 良雄

Miki Yoshio

社会に広がるビッグデータの潮流

インターネット上に存在する膨大なデータや、実世界の活動から生まれる、映像、文字、音声など、すべての情報を何らかの形でビジネスに応用しようとする動きから、ビッグデータが大きな潮流になっている。データを分析して業務に活用する考えは古くから存在し、すでに特定の業務では定着しているが、近年のビッグデータ潮流では、ネットサービス事業の成功事例をベースとして、種々の業務においても新たなデータ利活用が模索されている。しかし、このような潮流を解説し、自社ビジネスへの適用を考えたとき、新出の技術を適

用しただけではデータの利活用には至らないことに気がつく。ビッグデータの利活用においては、技術側面だけでなく、データが発生し、最終的にサービスが届けられる実世界、つまり、社会全体を系として捉える必要がある（図1参照）。電力システムや交通システムなどの社会インフラシステムも社会全体に広がりを持つが、システム全体に関与する生活者、事業者がシステム本来の提供物以外の価値を享受できるところがビッグデータにおいて着目すべき点である。

ここでは、ビッグデータ向けに考案、提唱されている新技術と旧来からのIT（Information Technology）との差異や同質性を俯瞰（ふかん）し、その本質を明らかにするとともに、ビッグデータ時代におけるITシステムの提供価値や仕組みについて述べる。併せて、ビッグデータ利活用方法の方向性を示すことで、この特集における個々の論文の位置づけの明確化を図る。

ビッグデータ利活用上の課題

ビッグデータへの期待とともに、利活用に関する不安、あるいは不明確な部分を払拭したいという声を聞くことが多い。この章では、まずビッグデータの利活用に関する課題を明らかにすることで、利活用に関する方向性、可能性の明確化を試みる。

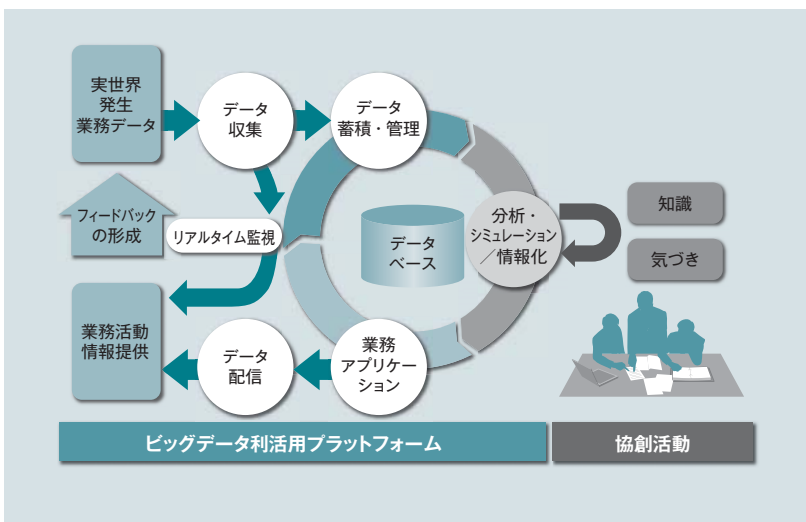


図1 価値創生のデータ流通構造

実世界で発生したデータを収集し分析した結果を基に、業務、サービスとして新たな情報を実世界へ配信する。配信された情報が実世界の行動変化を引き起こすことで、実世界で発生するデータが変化する。この循環がデータの発生から利用に至るまでの関係者全員に価値を生み出す。

ビッグデータ関連技術の利用

ビッグデータにおける代表的な技術には、**KVS**^(a)やHadoop^{※1)}に代表される「NoSQL^(b)」や、**CEP**^(c)に代表される「リアルタイムデータ処理技術」などがある。

(1) No SQL

No SQLは、旧来の帳票や業務書類のように、定型化されRelational Database (リレーショナルデータベース)に格納されたデータを取り扱うのではなく、インターネットや実世界に存在する非定型なデータを取り扱うためのファイル形式を、「SQLではない=No SQL」と表現したものである。技術的にはKVSがKeyとValueから成る簡便なデータ構造を基本としているように、従来のデータベースが行方向のデータに重きを置いていたのに対し、列方向の大量データの読み出しに適した構造を持つ。

同様に、特定のアプリケーションを高速化する目的で考案された技術にHadoopがある¹⁾。Hadoopは多数のノード計算機とストレージへ、**MapReduce**^(d)と呼ばれる方法で、データの分散、処理、収集を行うプログラミングの枠組みである。

(2) リアルタイムデータ処理技術

従来の情報処理システムでは、処理対象となるデータをデータベースなどにいったん保存し、サーバで処理をするというスタイルが一般的であったのに対し、センサーなど時々刻々発信される情報の処理では、リアルタイム性の観点からデータがシステムに取り込まれた瞬間に処理される形態が望ましい。CEPではリアルタイムに入力されるデータ列に対して、単一的な問い合わせ、処理要求を出すことで、リアルタイムに加工された情報を出力として得ることができる。

以上に述べたように、近年のビッグデータ操作における技術はインターネットサービスやオンラインの処理高速化の観点に着目したものが中心であり、処理のデータ操作との整合性に依存した技術に効果を得ている。したがって、処理内容と技術が整合

していない場合には、新技術を採用したのにも関わらず想定した効果が得られないというような状況が課題として見えてくる。

例えば、技術が考案された背景のアプリケーションを無視して、複雑なアプリケーションに適用しようとした場合に、データアクセスのインタフェースやデータ間の更新整合性の保証などで非常に多くの手間を要する。このような、データの同期やアプリケーションとのインタフェースは、旧来のデータベースが長い時間をかけて克服してきた課題であり、効果が最大化される部分に新技術を適用することが重要である。

利活用価値の発見

前述したように、ビッグデータの利活用はインターネットサービスなどで多くの価値を生み出しており、それ以外のビジネスにおいて利活用を考えた場合、データから得られる価値を発見する行為自体が一つの課題となり得る。

ビッグデータの利活用では、(1) 取り扱うデータの全体像を把握、(2) 多変量解析などの数理統計手法、マーケティング手法などによるデータの特徴抽出、(3) 業務指標や他データとの比較、クロス分析、といった流れが一般的である。

このうち(2)の段階で各種分析手法が用いられるが、ビジネスにおける重要性が判明するのは(3)の段階であり、この段階に至って、他のデータの必要性が見えてきたり、そもそものビジネス課題が表面化したりすることも多い。つまり、ビッグデータの価値はビジネス課題の解決と表裏関係となっており、人間の洞察力が重要である一方、ITシステムも価値を増大させる何らかの仕組みを持つことが重要である。この種のシステムにおける課題やその解決方法について、次に述べる。

ビッグデータ利活用システム構築上の課題

ここでは、ビッグデータを利活用するためにシステムを構築する際の課題について述べる。

(a) KVS

Key-value Storeの略。従来のデータベースを補完する、シンプルな構造のデータ保存方式。保存したい任意のデータを「値(Value)」とし、それに対応する「キー(Key)」を設定して、それらをペアで保存する方法。複雑なデータベース設計・運用の負担を軽減でき、拡張性に優れているのが特徴である。

(b) SQL

Structured Query Languageの略。リレーショナルデータベース(1件のデータを複数の項目の集合として表し、データの集合を表という形式で格納するデータ管理方式)の操作を行うために開発された言語。

(c) CEP

Complex Event Processingの略。複合イベント処理。多数のソースから発生するデータをミリ秒レベルで高速処理する技術。データをいったんHDD(Hard Disk Drive)に蓄えるという処理を省き、基本的にメモリ上で一連の処理を行うことで、リアルタイムのデータ分析と出力を可能にしている。近年、金融分野を中心に導入が進んでいる。

(d) MapReduce

多数のコンピュータによる大量データの並列分散処理を実現するフレームワーク。データを抽出・仕分けするMap処理と、抽出・仕分けしたデータを集計するReduce処理の2段階に分けてデータ処理を行う。2004年にGoogle社が開発、発表し、同社の検索エンジンのインデックスデータの生成をはじめ、幅広く利用されている。

※1) Hadoopは、Apache Software Foundationの登録商標または商標である。

(1) データ収集の課題

ビッグデータ利活用において、最初に遭遇するのがデータ収集の課題である。データの所在やアクセス（所有権、通信、データ量、データ種など）の問題により、データを手元に集めること自体に幾つかの課題が存在する。次に、データ全体のプロファイル可視化の課題がある。本格的な利用や詳細な分析に入る前にデータ全体の傾向や平均、分散などの基礎情報を取得する必要がある。しかし、データの全体量が多いこと、形式が多様であること、利活用前段階での大規模設備投資が困難であることなどが実行上の課題となる。

また、この段階では、最終的なビジネス適用の検討、既存ビジネスのデータや課題の整理、さらには業務フローなどのデータが発生し、適用される具体的な箇所の整理なども新たな作業や課題として浮上する。

(2) データ格納

ビッグデータの場合、収集されたデータの多くは、非構造データと呼ばれる形式が固定的ではないデータである。こうした複数の種類のデータを一元的に格納するための課題が存在する。また、全体を俯瞰したり、分析作業を行う際に、目的のデータを取り出すための方法も、データ形式によらない統一的な形で実行できることが望まれるが、そのような一般的な技術体系ははまだ整備されていない。

(3) 既存システムとの連携

ビッグデータの利活用における初期の分析工程は、当該データを利用対象とできるか否かの判定にすぎず、利活用方法を既存の業務システムの中に構築して最終的なゴールとなる。この際に問題になるのは、データ変換と既存データ形式を勘案したデータ管理である。業務に活用される前の素データは非定型であるが、いったん業務利用が決定したデータは、既存の定型業務データに整合した変換を施し、既存DB (Database) に格納するか、非構造のままでも、既存データとの関係が明確な形で管理される必要がある。技術課題はデータ蓄積と同様であるが、既存システムの形式と

の整合も含めた多様なデータ管理が必要となる。

スマート情報システム

ITと価値創生

スマートという単語を接頭語として用いることで機能の高度化を表現する用法を耳にすることが多い。しかし、ビッグデータの利活用の観点から、ITを適用しただけで機能が高度化し、何らかの価値が自動的に生まれるとするのはやや早計である。

ITが基本的に提供できる価値は次の3点に集約される。(1) 情報伝達：人間が日、年という単位で時間を要する情報の伝達を一瞬で可能とする。(2) 情報蓄積：人間の記憶量と精度には限界があるが、メモリやディスクに蓄積された情報は半永続的に保存可能である。(3) 情報処理：人間ならば手順の誤り、時間超過が発生しうる処理を、内容・時間ともに正確に実行可能である。

このように、ITが基本的に有する機能は極めてシンプルであり、価値を生み出すメカニズムがどこに存在するのかを明確化するのは、ITに対する過信、誤解の払拭（ふっしょく）にも有効であると考えられる。

ビッグデータの利活用に成功しているシステムアーキテクチャには一定の傾向がある。例えば、ビッグデータの名称が使われるきっかけを与えたインターネット検索の場合、(1) インターネット上のデータ収集、(2) データの閲覧、検索サービス提供、(3) ユーザーによる検索情報の活用、行動、(4) ユーザー行動によるデータの変化という一連の閉ループが形成されていることが分かる。

ここで重要なのは、ネット上に存在していたデータの提供だけでなく、そのデータを最終的に変化させるだけのユーザーやデータ保有者の行動変化を引き起こし、その変化がこのサービスに関わる全員に対して何らかの価値をもたらしているところにある。価値創生を実現するITシステムをスマート情報システムと定義するならば、

このような価値創生をもたらす閉ループを有するシステムと考えることができる。

店舗における宣伝方法と集客状況のシミュレーションの事例を図2に示す。ここでは、まったく自由に利用者が店舗を選択する場合（ランダムな来店）、一方的に広告を出す場合（一様な広告）、利用者の行動に合わせて離反しそうになったときに積極的な誘導を図る場合（個別広告）の3パターンを示している。

図が示すように、まったく誘導をしない状況に対して、広告（店舗3）への誘導はできているものの、個別の行動に合わせた誘導では圧倒的な集客とともに、ここに掲げた4店舗の中で寡占的状况を生み出している。このように、データの循環が閉じた系では、行動が情報を生み、その情報に基づく行動が新たな行動につながるといった価値の創生がシステムのメカニズムとして引き起こされる。

スマート情報システムにおけるデータ管理

ここでは、ビッグデータ利活用と技術的な課題の中で、スマート情報システムにおける解決法について述べる。

(1) データ格納と処理

スマート情報システムでは、実世界で発生するデータを直接取り扱うためにはデータ量だけでなく、データ多様性、リアルタイム性に対応可能なデータ管理が不可欠となる。データの多様性に関しては、一時的なデータの格納と、利活用という二つの段階に分離される。前者に関しては、検索のような簡易なアプリケーションを前提とするために、大規模ファイルシステムと**インメモリKVS^(e)**、Hadoopの組み合わせといったシステムが適合する。日立は、インメモリのKVS基盤、Hadoop導入サービス、そして、CEPを実現するストリームデータ処理基盤を提供している。

次に、実際の業務に多様性のあるデータを活用する場合、既存の業務システムのデータモデルを大幅に変更したり、アプリケーションやインタフェースを多種準備したりすることは大きな問題である。つま

り、多様なデータの利活用では、形式の異なるデータの統一的な格納と演算、およびアプリケーションとのインタフェースが不可欠となる。また、多様なデータを格納するためには、多次元の複雑なデータどうしの関係を管理する必要がある。この点に関しては、多次元表やグラフ管理により実現することができるが、個別のデータ格納形式を持つことは非効率である。

以上の課題を同時に解決可能なのが、Relational Database技術である²⁾。アプリケーションインタフェースとしてSQLはすでに確立されており、現在も非構造データなどに向け拡張が進んでいる。これに加え、処理速度の問題を原理的に克服することで、高速なJoin演算などにより、複雑なデータ関係を表形式のまま表現できる。日立は高速データ処理システムHitachi Advanced Data Binderプラットフォームなどの革新的技術を提供することに加え、ストレージ仮想化の技術を高度に発展させることで、ビッグデータの格納と処理の容易化を実現する。

(2) データ収集と共有

スマート情報システムでは、先に述べたように、広域に分散したシステムが前提となること、サービスに用いるデータの所有

(e) インメモリKVS

データを外部ストレージではなくサーバの内蔵メモリ上に格納し、直接アクセスすることで、高速データ処理を可能にするインメモリ技術とKVSを組み合わせ、高速で高信頼なデータ処理を実現する技術。

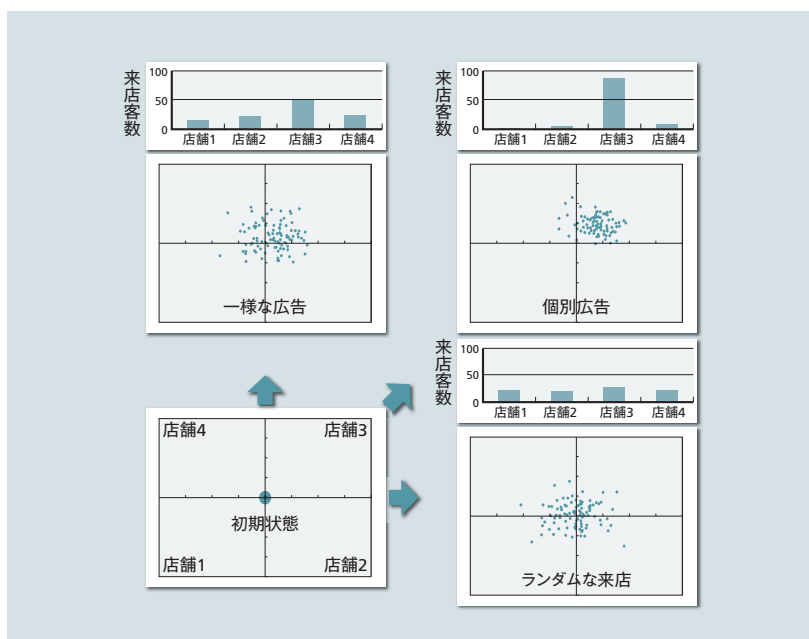


図2 | 集客シミュレーション

閉じた系では、来店客の個別行動に対して広告を出すことが可能となり、一様に広告を出す場合に比べて圧倒的な集客力を持つ。

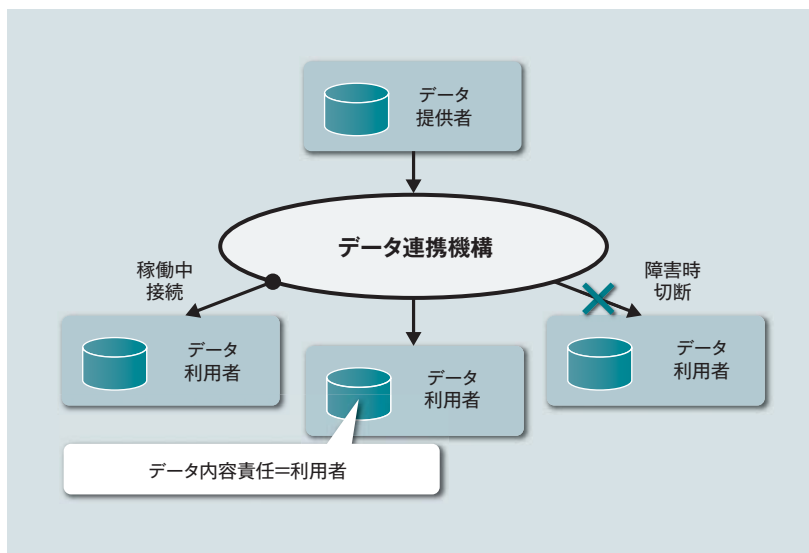


図3 | データ連携機構

スマート情報システムでは社会インフラ事業者、IT (Information Technology) サービス事業者が自由にデータ共有できる環境と障害伝播(ば)の防止、システム拡張性の両立が必要である。

者、関連事業者が複数にまたがることなどが特徴である。しかし、この特徴から次のような課題が新たに生じる。

(a) システム拡張性

スマート情報システムでは、社会インフラ事業者、IT系のサービス事業者、ユーザーとしての一般家庭、企業が、システムやデータを共有する系となる。そのために、新たな事業者やユーザー、サービスが追加される際に共有データ型式の変更やシステムテストのために、すでに稼働しているシステムを停止することは不可能である。このために、稼働中の全体系を停止せずに新たなシステムの追加、削除を可能とする仕組みが必要となる。

(b) 責任分界

複数のシステムが連携して稼働する複合系では、あるシステムが障害で停止しても他のシステムに影響を及ぼさない仕組みが必要である。また、障害復旧の観点からはシステムとしての責任分界線が明確であることが重要であり、上記のシステム動作だけでなく、データ共有におけるデータのマスター管理にも共通する。

(c) データ共有

上記のシステム拡張性や責任分界の明確性を保ちながら、広域、他業種で発生するデータを共有し、かつ保管する必要がある。

以上の課題と機能を実現するために、

日立はデータ収集・蓄積・配信基盤を開発している。この基盤ではデータ連携機構とネットワーク仮想化により、システム間を連携する。データ連携機構に接続される複数のシステムにおいて、データを送出する役割のシステムとデータ配信を受けるシステムが明確に規定されており、データを受けるシステムが障害などにより停止しても、他のシステムに影響を及ぼさないことを原則としている。この種の考えは、自律分散システム³⁾の技術に端を発しており、データ連携機構は制御だけではなく、一般的なITサービスとデータ共有管理に拡張したものである。

日立はこのデータ収集・蓄積・配信基盤を情報制御連携基盤のコア要素として位置づけることで、社会インフラシステムと情報システムが安全・安心に連携したスマート情報システムを提供する(図3参照)。

ビッグデータ利活用事例

ここでは、ビッグデータの利活用事例として、前章で述べたスマート情報システムの適用パターンと有効性について述べる。

ITシステム障害予兆検知

ビッグデータの利活用を、身近なITシステム運用に適用することが可能である。障害予兆をIT機器から出力されるログデータの分析により実施する(図4参照)。まず、ログ情報はストリーム情報処理基盤に入力される。ここでは、時系列的に変化する情報をほぼリアルタイムに処理し、情報の変化傾向からシステムの異常状態を予見し、システム運用オペレータに警報を上げる仕組みとなる。この判断基準や異常が予見される入力パターンの解析は中段、後段の分析システムで実行される。ストリームデータ処理に入力された時系列情報、ならびにシステム稼働情報は中段の時系列格納基盤に格納される。

次に、格納されているシステムの挙動と異常状態の比較検討を行い、機械学習的に傾向を判断し、どのようなリアルタイム入

力があつた場合にシステムの異常予兆と捉えるべきかの情報を、最前段のストリーム情報処理基盤に返す仕組みとなる。

前述したように、スマート情報システムでは情報の閉ループが価値を生み出す構造となっている。この例では、システムが出力する時系列データに基づく予兆検知を学習的に改善していることから、次第に真の異常予兆のみが警報対象となることで、分析精度が向上していくという改善が行われる。

データセンター空調—IT連携

IT機器は動力系機器とは異なり、投入された電力がほぼすべて熱として排出されることになる。そのために、IT機器から発生した熱を空調設備により屋外へ運び出すための電力が、ほぼIT機器と同じだけ必要となる。このことから、ITシステムの消費電力を低減するためにはIT機器を使わないことが、唯一の直接的な手段となるが、一般的には空調効率の改善として、センサーで室内の温度分布と温度変化の傾向を詳細に調べ、空調制御に反映することや、データセンター内の温度分布や冷氣と暖気の直接的な混合を防ぐことなど、熱輸送面での効率改善にとどまっていた。

空調とITの連携制御においては、ITの運用スケジュールや入力データ量の監視など、短期から長期にわたるITの運用スケジュールに基づき、稼働させるIT機器の場所と、それに呼応した空調設備の運転を連動させる（図5参照）。これにより、最小限の空調設備で熱輸送が可能となる。このように、スマート情報システムでは、エネルギーなどの需要側の需要計画や需要実態を把握することで、供給もしくは投入する資源の最適化、最小化を実現することができる。これが、スマートグリッドなどの効用原理である。

ビル環境マネジメント

従来の情報処理システムが、人間に対して処理結果を出力するのに対して、制御シ

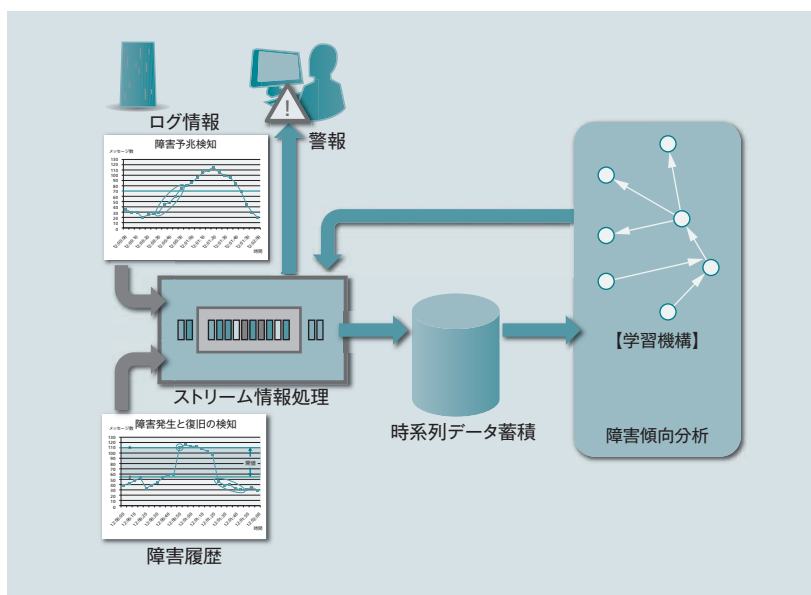


図4 ITシステム障害予兆検知

IT装置から出力されたログ情報をストリーム情報処理で故障傾向を分析し、警報を出す。詳細な分析方法は機械学習機構で生成する。

システムは機器に対して制御信号を出力し、かつ、機器の情報は直接的に制御システムに入力される。ここで、情報処理システムと制御システムを連携させることができれば、スマート情報システムとしてデータの閉ループを形成することが可能となる。ビルの空調や照明の管理に対して、人が実際に体感している情報をフィードバックすることで、既定されている空調や照明の制御以上にエネルギーを節約できる可能性が生

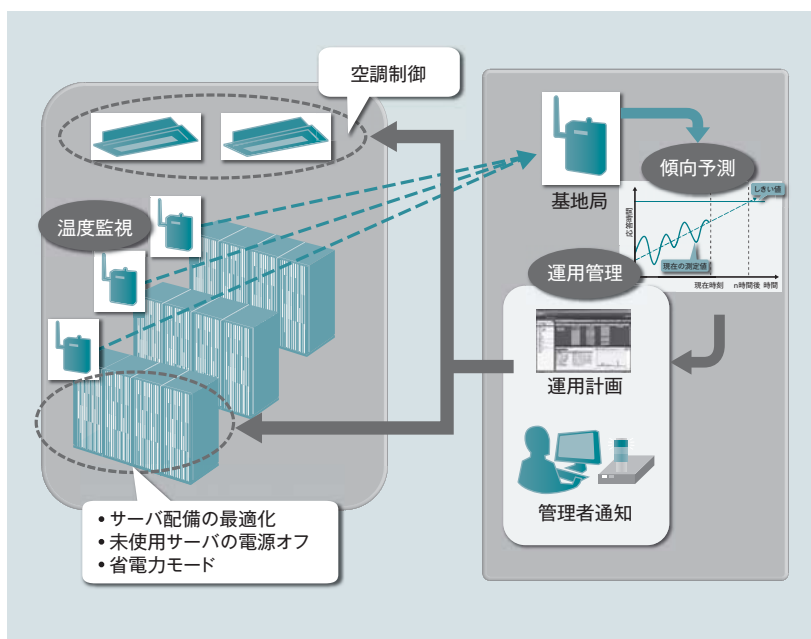


図5 データセンター空調—IT連携

予定されているIT機器の運用計画と、データセンター内の空調状況などから総合的に空調とIT機器の制御を実施して、トータルに消費電力を削減する。

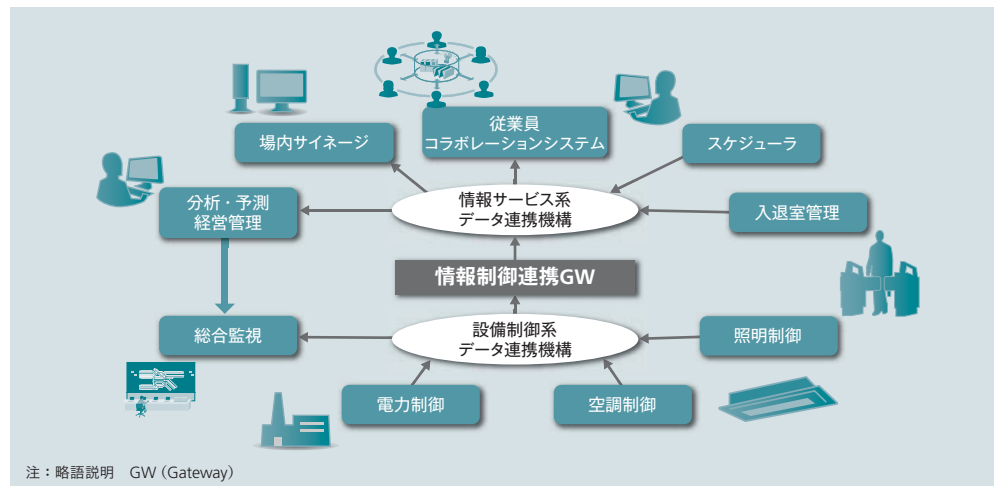


図6 | ビル環境マネジメント

ビルの空調や照明の管理に対して、人間が実際に体感している情報をフィードバックすることで、既定されている空調や照明の制御以上にエネルギーを節約する。

まれる（図6参照）。

将来は業務スケジュールや会議室予約システムを連携することにより、IT機器と空調の連携と同様に、将来の需要予測に基づくむだの最小化を実現することも可能である。

この例のように、インフラと一般サービスとの連携システムでは、あるシステムの障害により、インフラ系システムが停止する事態を回避しなければならない。また、上記のようにシステムは固定的ではなく、柔軟に拡張されることが前提となる。このような場合に、前章で述べた情報制御環境が有効であり、ビッグデータの実世界適用における基本機構であると考えられる。

ビッグデータ活用を通じ、社会イノベーションに寄与

以上、ビッグデータ利活用上の課題を価

値創生、技術の両面からまとめ、その解決方法を述べた。価値創生の観点では、データの収集、分析、配信、利用の流れが一連の閉ループを構成するアーキテクチャが有効であることを示した。技術の観点では、データの格納、管理においては、一次データの格納検索にはNo SQLを用い、複雑なアプリケーションに対してはリアルタイム処理にストリームデータ処理、多様なデータ管理には高速なRelational Databaseが有効であることを示した。

また、データの収集と共有ではシステムの拡張性や信頼性確保に自律分散システムの考え方を一般情報処理に拡張した考え方が有効であることを示した。このように、日立は社会インフラから情報サービスまでの幅広い連携から新たな価値を生み出し、社会全体のイノベーションに寄与していく所存である。

参考文献など

- 1) Hadoop Wiki, <http://wiki.apache.org/hadoop/ProjectDescription>
- 2) 長尾、外：情報の構造とデータベース、岩波講座 情報科学8、岩波書店（1983.6）
- 3) 森：自律分散システム入門、森北出版（2006.9）

執筆者紹介



三木 良雄
1986年日立製作所入社、情報・通信システム社 ITプラットフォーム事業本部 事業統括本部 企画本部 所属
現在、ソフトウェアの製品企画に従事
工学博士
IEEE会員、電子情報通信学会会員、情報処理学会会員