NTT DATA
Global IT Innovator

NTTデータ
ビッグデータ・リファレンス・アーキテクチャー
第1.0版

NTT DATA Big Data Reference Architecture

ビッグデータ・リファレンス・アーキテクチャは
NTTデータとEVERIS SPAIN, S.L.U.の共同著作物です。

目次

図表

1章 活用が進むビッグデータ

「世の中に大量のデータがあふれ、いかにこれらのデータ（ビッグデータ）を上手く活用するかが今後の企業の競争力を分ける」と言われるにようになって久しい。事実、企業におけるデータ流通量においては、総務省が公表する推計値推移では2005年から2013年までの9年間でデータ流通量は約8.7倍に拡大している。データ活用においては、マーケティング領域ではインターネット広告に適用される「アドテクノロジー」や個別商品ごとの需要予測、そして業務管理・品質管理の領域では製造業における設計等の精度向上や運輸業における運行効率化が挙げられる（図表 1）。

特に重要になると想定されるテーマの1つが、IoT（Internet Of Things）である。
あらゆる製品（モノ）が、ネットワークにつながり、さまざまなモノの状態を把握するためのセンサーが設置され、遠隔地からリアルタイムにモノの状態を把握し、操作することが可能になる。刻々と生成されるこれらの情報を、リアルタイムに活用する新たなサービスが続々と生み出される日は、すぐそこまで来ている。

このような流れを受けて、これまで以上に多くの企業が自社のサービス高度化や新しいビジネス創出のためにデータを活用する仕組みの構築に取り組んでいる。

図表 1　ビッグデータの活用事例

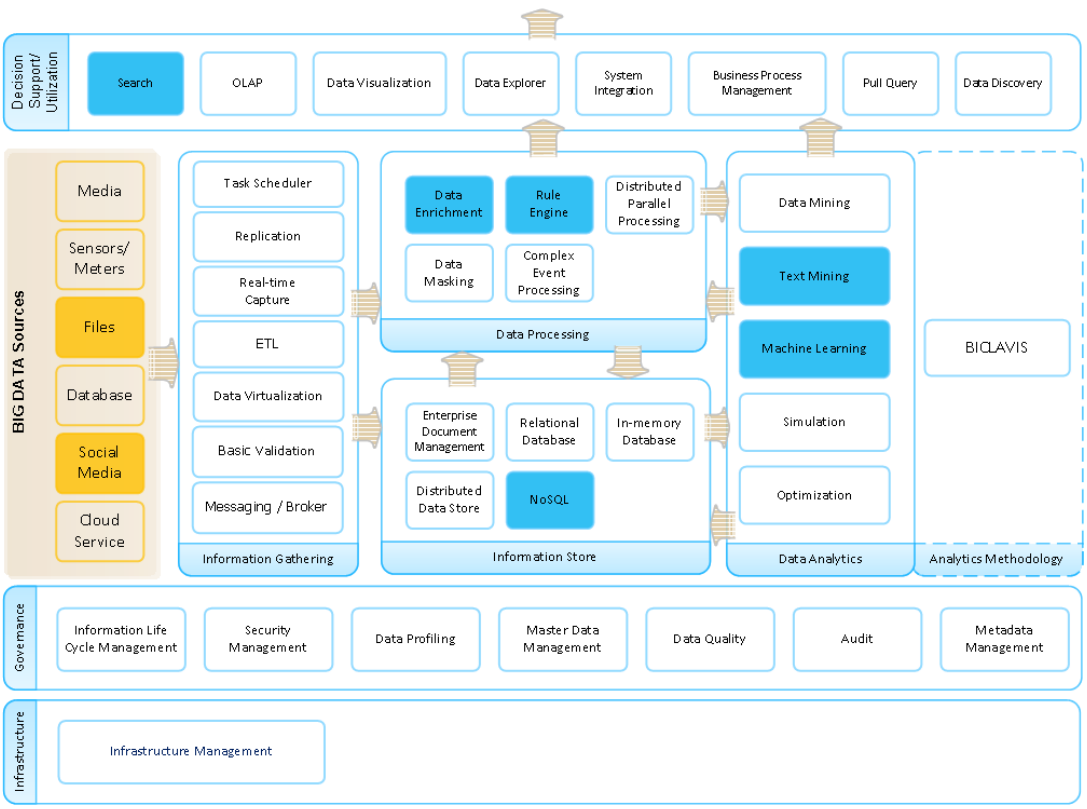
活用分野	概要
マーケティング	・　インターネット広告における「DSP（Demand-Side Platform）」への活用 ・　サービス業における、個別商品の動態情報を活用した需要予測
業務管理・品質管理	・　製造分野における、製造記録情報を活用した設計・加工条件の精度向上 ・　農業における、家畜の動態データを活用した成育状況予測・管理 ・　運輸業における、車載 GPS・乗車人員情報を活用した運行ダイヤの最適化

（『平成 26 年度版　情報通信白書』より作成）

2章 NTTデータ ビッグデータ・リファレンス・アーキテクチャー

データ活用の仕組みについて世の中に目を向けると、大量のデータを分散処理するためのHadoopや、発生したデータを蓄積することなくリアルタイムに分析処理するCEP（Complex Event Processing：複合イベント処理）など個別技術は揃ってきている。更に、これら技術の一部はオープンソースで提供されていることから、誰でも利用できる状況にある。しかし、真にビジネスに活きるビッグデータ活用のためには、単に要素技術を取りそろえるだけではなく、これらの技術を上手く組み合わせ、仕組みを素早く構築し、柔軟に拡張、進化させていくことが重要である。

そのためNTTデータグループでは、グローバルでのビッグデータシステム開発の経験と技術開発を活かし、ビッグデータ・リファレンス・アーキテクチャー（Big Data Reference Architecture 以下BDRA）として体系化している（図表２）。これを基に、各企業の目的や既存ITシステムの状況に合わせ、ビッグデータ活用の在り方を提示することができる。



図表 2 NTTデータ ビッグデータ・リファレンス・アーキテクチャー（BDRA）

ここでは、次章のユースケースに進むにあたり必要となるBDRA構成の紹介にとどめ、特徴は後述することとする。BDRAは大きく3つのプラットフォームと７つのレイヤー（機能の集合体）で構成されている。第一のプラットフォームは多種多様なデータを分析できる状態に加工する役割であり、「収集」、「蓄積」、「処理」の3つのレイヤーで構成される。第二はデータ活用の中核となる分析プラットフォームであり、「分析」、「可視化」の2つのレイヤー、そして第三は全体を管理するためのマネジメントプラットフォームであり、「ガバナンス」、「インフラ管理」の2つのレイヤーで構成される（図表 3）。

図表 3 NTTデータ ビッグデータ・リファレンス・アーキテクチャーのレイヤー

区分	レイヤー	概要
データプラットフォーム	データ収集	● Web 等のメディアや、各種センサー、データベース等、さまざまなデータソースが産み出し、蓄積している多様な情報を集めて、分析可能な形にする機能を集めたレイヤー。ETL を活用した異なる形式のデータの統合や、ソフトウエアやハードウエアなど異なるリソース間での情報を共有し、信頼性や可用性、アクセス容易性を強化するメッセージングやレプリケーション処理などが本レイヤーで実施される。
	データ蓄積	● 大規模データを蓄積し、柔軟に処理する上で必要となるデータベース機能を集めたレイヤー。大規模データ処理を実現する分散データストアや、高速処理を実現するインメモリー・データベース、スケーラビリティが高く柔軟性が特徴である NoSQL などが本レイヤーに含まれる。
	データ処理	● 収集した大規模データを高速に分析するための処理機能やデータマスキングなど分析前に必要となる事前処理機能を集めたレイヤー。大規模データ処理を実現する並列分散処理技術や、高速処理を実現する複合イベント処理技術など、ビッグデータ・ソリューションの中核となる技術が含まれる。

区分	レイヤー	概要
分析プラットフォーム	データ分析	● 蓄積・処理されたデータに対する相関分析や自然言語分析、機械学習などさまざまな分析機能を集めたレイヤー。テキストマイニングやデータマイニングなどの分析手法が本レイヤーに含まれる。また、NTT データが体系化した分析方法論「BICLAVIS」[1]の活用により、さまざまな分析手法の最適かつ複合的活用を実現する。
	可視化	● 分析結果を活用した意思決定を支援するための機能を集めたレイヤー。データ・ビジュアライゼーション（見える化）や OLAP、ビジネス・プロセス・マネジメントなどが本レイヤーに含まれる。
マネジメントプラットフォーム	ガバナンス	● データの品質管理や、データの保護を目的としたデータのマネジメント機能を集めたレイヤー。情報のライフサイクルマネジメントを始めとし、データ・プロファイリングやマスターデータ・マネジメント、メタデータ・マネジメントなどデータのマネジメントを通じ、データの品質管理を実現する。データ保護の視点では、セキュリティ・マネジメントや監査が本レイヤーに含まれる。
	インフラ管理	● 信頼性や可用性、パフォーマンス、スケーラビリティの管理を目的とした、運用面のマネジメントとシステム的なマネジメント双方を実現するレイヤー。

[1] データ分析方法論「BICLAVIS」については後述の「ビッグデータ活用時の課題とBDRAの特徴」において詳細を記す。

3章 BDRAユースケース

前章で紹介したBDRAを用いたユースケースの中から、代表的な４件を紹介する。

3.1.SNSデータを用いた金融マーケット指数の変動予測

ここでは、大規模データのリアルタイム分析技術を活用し、Twitterデータからなる「Twitterセンチメント指標」が株価指数との相関を明らかにしたユースケースを紹介する。

近年、米国の金融市場を中心に、Twitterなどのソーシャル・ネットワーク上の情報活用が進んでいる。日本においても金融分野におけるビッグデータの活用は広まってきており、同じくソーシャル・ネットワーク上の情報活用への関心が高まっている。

このような流れを受けて、NTTデータとNTTデータ数理システムは、Twitterデータを用いて株価指数と相関のある指標「Twitterセンチメント指標」を開発した。これは、株式に関するツイートを抽出し、ポジティブなツイートか、ネガティブなツイートかを自動的に判別し、それぞれの全ツイートに占める割合を指標として、リアルタイムに算出するものである。この「Twitterセンチメント指標」は、「日経平均ボラティリティーインデックス」に対し統計的に有意な関係性があることを35か月分（2011年1月～2013年11月）の数億ツイートにも及ぶデータから検証した。

本ユースケースにおけるビッグデータ活用時のポイントは、分析におけるリアルタイム性の確保、及び、効率的な分析手法の選定である。リアルタイムに分析を行うためには、大規模データから高速にデータを抽出する仕組みが必要であり、また、日本語テキストの処理は英語などの一つ一つの単語が明確に区切られている語に比べ、一つの文から意味を持つ最小単位を見つけ出す処理が必要となるため時間がかかる。そのため、BDRAにおけるデータ処理レイヤーの「並列分散処理」、及びデータ蓄積レイヤーの「分散データストア」（具体的にはHadoop Distributed File System）を活用し、リアルタイムでのツイート分析を実現している。さらに、データ分析レイヤーの「テキストマイニング」や「データマイニング」、「ルール・エンジン」などさまざまな技術を統合的に活用していることも、本ユースケースの特徴のひとつである。

また、効率的な分析手法の選定にあたっては、NTTデータが体系化したデータ分析方法論「BICLAVIS」を活用することで、分析の効率化を図っている。本ケースでは、「Twitterセンチメント指標」と「日経平均ボラティリティーインデックス」の相関を評価するにあたり、「評価・要因分析型」のシナリオ類型を活用している。

3.2. システム開発における設計フェーズの自動化ツール

ここでは、柔軟なデータモデルの構築やメタデータ・マネジメントを活用し、システム開発環境に自動化ツールを導入したユースケースを紹介する。

NTTデータでは、グローバル化の進展などのビジネス環境の変化に伴い、従来の高品質なITシステムを短期に実現する「TERASOLUNA」というオープン系システム開発のための総合ソリューションを提供している。そのソリューションのひとつとして、設計情報を始めとするシステム開発時に作成する情報を集約し、設計書の整合性チェックや設計ノウハウの蓄積などシステム開発の効率化および品質担保に寄与するツール「TERASOLUNA DS」を開発した。これは、通常レビューにおいて人手で行う設計書の整合性や表記揺れチェックの自動化、設計書・ソースコードの高速検索、仕様変更時の影響範囲分析、設計書の入力支援など、システム開発を支援するさまざまな機能を提供するものである。これにより、レビュー作業の削減、仕様変更/バグ発生時の影響範囲特定等、設計作業を大幅に効率化することが可能となっている。

本ユースケースにおけるビッグデータ活用時のポイントは、設計書のフォーマットがプロジェクト毎に異なることに起因するスキーマの複雑化や、新しい文書形式追加時のスキーマ再設計への対応が挙げられる。まずBDRAのデータ収集レイヤーにおける「ETL処理」において、さまざまなフォーマットの設計書をXMLファイルに変換した上で、データ蓄積レイヤーの「NoSQL データベース」を活用することにより、スキーマに依存しない柔軟なデータモデルを構築している。さらに、ガバナンスレイヤーにおける「メタデータ・マネジメント」により設計書の整合性確認を効率的かつ正確に自動化することができている。大規模システム開発における設計書量は4万ファイル、40万ページに及ぶほど膨大である。その設計書に関する構造や属性、記録情報などのメタデータを管理する仕組みを入れることで、人手よるレビューよりも正確な確認、分析を実現している。

3.3.橋梁の状態のリアルタイム解析

ここでは、東京ゲートブリッジやベトナムのメコン川にある東南アジア最長級のカントー橋など、橋梁の状態をモニタリングする既存サービスをベースに、大規模データの高速処理技術を適用した実証実験を紹介する。

社会インフラである橋や道路は人々の生活を支える基盤であり、「いつでも、安全に使えて当たり前」と考えられている。そのため、管理者は迅速に橋梁の異常や損傷を把握し、道路通行再開の判定や使用可能ルートを特定することが求められる。

そこでNTTデータでは、橋梁に設置されている各種センサーを用いて橋桁および橋脚のひずみなどのデータをリアルタイムかつ継続的に収集、解析に取り組んだ。

本ユースケースにおけるビッグデータ活用時のポイントは、短時間で膨大に発生するセンサーデータを遅延なく処理することである。BDRAのデータ処理レイヤーにおける「複合イベント処理」を活用し、大規模災害発生時等に俯瞰監視する橋梁が増加した場合にも、1台のサーバで100橋梁以上のセンサーデータを高速で解析することが可能である。また、異常パターンの抽出にデータ分析レイヤーの「データマイニング」を活用していること、さらに意思決定支援レイヤーの「データ・ビジュアライゼーション」を活用し異常検知結果を視覚的にわかりやすく表示するなど、各レイヤーの技術を組み合わせることで、本システムを実現している。

さらに、NTTデータのデータ分析方法論「BICLAVIS」を活用することにより、異常検知の精度を高めている。センサーデータから検知する異常値には、センサー異常による計測失敗や、強風や地震などの外力が混在している。そこで、低周波成分除去の前処理や、センサー間の位置関係を考慮した遅延相関を利用した判別ロジックを実装し、これらの異常値を橋梁の異常と区別している。これらの異常検出においては、異常パターンの定義が可能なものについては「不正検出型」を、定義が困難なものには「外れ値検出型」のBICLAVISシナリオ類型を活用することで効率化と精度向上を実現している。

3.4. 交通状況の予測システム

ここでは、大規模データのシミュレーション技術や予測・制御分析モデルを活用することで、交通渋滞の緩和を実現したユースケースを紹介する。

交通渋滞は先進国や発展途上国を問わず重要な社会問題となっており、膨大な時間的、経済的な損失が発生しているだけでなく、化石燃料消費やCO2発生などの環境負荷の要因にもなっている。各国で新たな道路建設や道路規制の再設計などに取り組むも、実施コストが高く、渋滞緩和効果の見込みもわからない状況であった。またそのような対策は局所的な効果しかなく、全体最適が実現できないことも問題である。

このような問題を解決するために、NTTデータでは、カーナビやスマートフォンから取得した車両の GPS データをもとに交通状況を予測し、信号制御や交通規制などの交通施策による渋滞緩和の効果を試算できる交通シミュレーションシステムを開発し、中国の吉林市で実証実験を実施した。実証実験では、シミュレーションを利用した渋滞緩和により、バス路線の運行時間を最大27%改善することに成功した。これは、車両、道路、信号など交通環境の数理モデルを構築、計算機上で交通環境を再現し、信号変換シナリオを交通シミュレーションにより評価した上で最適な信号制御を実施するものである。信号制御や交通規制をかけたシナリオでシミュレーションすることで、事前に交通施策の効果を判断できるため、渋滞対策の検討が可能となる。また、交通管理者は渋滞予測結果を見ることで、渋滞が発生しやすい道路・時間帯など現在の交通のボトルネックを把握することもできる。

本ユースケースにおけるビッグデータ活用時のポイントは、大規模な交通量をシミュレーションするための大規模データの処理にある。BDRAのデータ処理レイヤーの「並列分散処理」を活用した高速シミュレーション基盤により、100万台規模の交通量のシミュレーションを実現することができている。「並列分散処理」機能以外に、データ収集レイヤーの「リアルタイム・キャプチャ」機能や意思決定支援レイヤーの「データ・ビジュアライゼーション」機能などBDRAにおける各レイヤーの機能を組み合わせることで効率的かつ適切なシステム構築を実現している。

また、本ユースケースでは、予測や制御に関する分析において、データ分析方法論「BICLAVIS」を活用し、「リスク・シミュレーション型」のシナリオを採用している。

4章 ビッグデータ活用時の課題とBDRAの特徴

前章で紹介したユースケースから見てとれるビッグデータを活用する際に共通する課題と、BDRAの特徴を示す。

・ビッグデータ活用時の課題

（1）複合的なIT基盤技術の組み合わせ

企業が求めるビッグデータ活用シーンを紐解くと、データの「取得、蓄積、処理、分析」のいずれか１つで事足りる訳ではなく、これらが組み合わさっている。特にデータ蓄積においては、これまでのリレーショナル・データベース一辺倒ではなく、活用用途に応じた技術の組み合わせは当たり前になってきており、蓄積したデータをいかに使うか、またはいかにデータを蓄積するかが求められている。

（2）多種多様な業界のデータ分析

ビッグデータ活用に取り組む企業は、金融、IT、社会インフラなど多岐に渡っている。また、企業が求めるデータ分析結果が高度化するにつれ、データ分析手法も複雑化する。企業が求めるビジネススピードを落とすこと無く、これらの要求に答えるためには適切なデータ分析手法の選択が重要である。

（3）データ品質の担保

システム開発の自動化事例に見て取れるよう、企業内に蓄積されたデータは分析することを前提としているものではない。そのため、データをプロファイリングし、分析に有効なものになっているかを早期に検証することが必要となる。データから有効な分析結果を得るため、そのライフサイクルを最適に管理することが重要である。

・BDRAの特徴

上記3つの課題に対し、BDRAは次に示すような特徴を持っている。

（1）迅速・柔軟な技術統合を実現する網羅的なフレーム

BDRAは7つに分けたビッグデータ活用に必要となる技術毎に深い知識を、そして組み合わせの妙を有しており、これらを体系化している。世の中のベンダー製品やオープンソースソフトウェア（OSS）の組み合わせを検証しており、異なるベンダー製品間の組み合わせにも対応できる。また、使用頻度が高い組み合わせに関してはセット化していることに加え、既存ITシステムに合わせた取捨選択も可能である。各ビルディングブロックの概要については、第二弾で説明する。

（2）効果的・効率的な分析アプローチを実現する方法論「BICLAVIS」

NTTデータでは、これまで実施してきた200を超えるデータ分析事例をもとに、「BICLAVIS」という業界横断的なデータ分析方法論を整理している。データ分析業務は得てして属人的になりがちであり、一方で分析を求める業界・業務の裾野は広い。そのため、NTTデータでは、データ分析のノウハウを収集する仕組みを整え、これらの情報を業界横断的に活用できるよう分析の切り口で体系化している。具体的には、分析目的を軸にシナリオとして類型化しデータ分析の目的、手順、手法とそれらを実装した分析テンプレートとして整備している（図表4）。これにより、これまで実績がない業界からのデータ分析依頼に関しても成果をあげることができる。

図表 4　分析シナリオ

分析シナリオ類型		概要
① 予兆発見型		大量のデータから何らかの構造変化や状態変化が起こる兆しを発見する。
② 異常検出型		大量のデータから異常なパターンを自動的かつリアルタイムに検出し、アラーティングにより早期の危機対応を促す。
	不正検出型	事前知識による異常パターンの定義が容易な場合に、定義ファイルに合致する行動や状態を異常と見なして検出する。
	外れ値検出型	事前知識による異常パターンの定義が困難な場合に、標準的な行動や状態から逸脱するものを異常と見なして検出する。
③ 予測・制御型		業務における要因と結果の因果関係を明確化し、要因の操作に伴う結果の変化を予測することによって、要因の適切な水準を把握する。
	収益シミュレーション型	シミュレーションによりあらゆる改善施策の増収効果を試算し、施策の順位付けや選定を行う。
	リスク・シミュレーション型	不確実要素に起因するリスクも考慮したシミュレーションを行い、改善施策の順位付けや選定を行う。
	最適化型	最適化手法を用いて、パフォーマンスを最大化するような改善施策を選定する。
	リスク・ヘッジ型	リスク分散手法を用いて、リスクを最小にするような改善施策を選定する。
④ ターゲティング型		優良顧客や見込み顧客など、重点的に施策を打つべきターゲットを抽出することによって、施策の費用対効果を最大化する。
⑤ 与信管理型		顧客のデフォルト（滞納・倒産）リスクや解約リスクをスコアリング・管理することによって、顧客の選別や乗り換え防止策を打つ。
⑥ 評価・要因分析型		観測対象を比較評価すると同時に、その評価結果を左右する要因を特定し、対象ごとに評価の改善策を提示する。
⑦ コンテクスト・アウェアネス型		個々のユーザの行動履歴や嗜好などのデータから行動の意図を読み取り、一歩先回りした商品やサービスを提供する。
⑧ プロセス・トレース型		成長・発展や悪化のプロセスを抽出し、それを促進または阻害・阻止する要因を特定する。

（3）ビッグデータガバナンスの確立に寄与する「ガバナンス」レイヤー

BDRAは、ビッグデータ活用上必須となるデータの信頼性向上や、セキュリティの向上についてのガバナンス機能を豊富に有している。ビッグデータで活用される情報は、「ガベージイン・ガベージアウト」という言葉があるように、精度の悪いデータからは、意味のない分析結果しか出てこない。

中でもデータの信頼性向上のための機能については、データクレンジングの前にデータのプロファイリングを実施し、データ品質に関するルールを定めるよう定義されており、マスタデータに関しても管理のための体系が確立されている。

さらには、データを保護するセキュリティの観点からも、多様な機能が準備されている。パーソナルデータに関するさまざまな議論がなされる昨今、安心してビッグデータを活用するためには、セキュリティの観点が欠かせないものである。セキュリティ・マネジメントは、一連のメソッドとして整備されており、IT監査や情報セキュリティ監査、データ中心型の監査と保護（DCAP：Data-Centric Audit and Protection）など監査におけるさまざまなポイントも整理されている。

これまでに見たように、BDRAはビッグデータ活用のさまざまな仕組みを、多様な技術を統合して構築した際のノウハウを集約したものである。現在も、NTTデータはこのBDRAを活用してさまざまな仕組みを提供しており、今後も大きく進化させる予定である。

株式会社NTTデータ
技術開発本部 サービスイノベーションセンタ
〒135-6033 東京都江東区豊洲3-3-9 豊洲センタービル アネックス
Tel: 050-5547-7537 Fax: 03-3532-0487
www.nttdata.com/jp/