Special edition paper

ビッグデータ分析用高速サーバの導入

Introduction of High-Speed Analysis Data Warehouse



志小田　雄宇*　真柴　史明*　荒井　浩**

JR-EAST has many types of “Big Data” such as train operational data, equipment management data, passengers'(or customers') opinion data. It is hoped that analysis of these data will apply many services and internal operations, for example, information service for passenger to the individual, statistical prediction based on operations. However, Big Data analysis take an immense amount of time on general-purpose computer because of huge volumes of data. Therefore, we built Big Data analysis platform system, which including High-Speed Analysis DWH (Data WareHouse: DWH), DataMining software and ETL software. We opened the way for efficient Big Data analysis by introduction of this system.

●キーワード：ビッグデータ、データマイニング、データウェアハウス、ETL ツール

1. はじめに

近年のユビキタスネットワーク環境の整備やソーシャルメディアの普及など「ネットワーク・サービスレベル」での進歩と、スマートフォン・タブレットの普及やデータストレージ機器の低価格化など「デバイスレベル」での進展は、多様で膨大なデジタルデータを日々爆発的に創出し、蓄積することを可能にしてきている。[1] 全世界で生み出されるデータ量は、2009年の0.8ゼタバイト（0.8×10^{21}バイト）から、2020年には35ゼタバイト（35×10^{21}バイト）に増加すると予想されている。[2]

上述のような大容量データは、総称して「ビッグデータ」と呼ばれ、現在、このビッグデータの利活用をとおして企業や組織の成長戦略を推し進める社会的気運が高まっている。

当社においても、鉄道輸送や設備管理など、多くのビックデータを所有していることから、これらを利活用した「お客さまサービス」や「社内の業務革新」の実現が期待されている。

一方で、ビッグデータは、データ容量が非常に大きいため汎用コンピュータでは処理に時間がかかりすぎること、使用するデータにはエラーが含まれている場合が多いことなどから、そのままでは利活用することが難しい。そこで、ビッグデータの利活用を効率的に進めるためには、データの処理・分析に特化した専用のハードウェアとソフトウェアを導入することが必要となる。

このような背景から、フロンティアサービス研究所では、JR東日本研究開発センター内にビッグデータ分析プラットフォーム（以後、高速データ分析システム）を構築し、社内ビッグデータの有効活用を推進していくことにした。本論文では、今回構築した高速データ分析システムの概要と当研究所におけるビッグデータ利活用の将来像について紹介する。

2. 高速データ分析システム

2.1 高速データ分析システムの概要

図1にビッグデータ分析のプラットフォームとして導入した「高速データ分析システム」の構成イメージ図を示す。

構築したシステムは、分析端末からデータマイニングツール（データ分析ソフトウェア）用サーバまたはETLツール（データ変換・アップロードソフトウェア）用サーバにアクセスすることで、高速サーバに目的の処理や分析を実行させ、結果を分析端末で確認することができるようになっている。以下の項より、各機器およびソフトウェアの概要と特徴について簡単に紹介する。

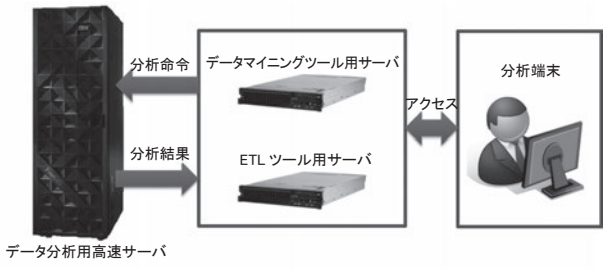

図1　高速データ分析システム

2.2 データ分析用高速サーバ

導入したデータ分析用高速サーバの外観を図2に示す。本機器は、膨大なデータベース上の特定のレコードを高速に検索・処理・分析できるように目的特化したサーバであり、一般的には高速分析データウェアハウス（Data WareHouse: DWH）などと呼ばれている（以後、高速サーバと呼ぶ）。

*JR東日本研究開発センター　フロンティアサービス研究所
**株式会社ビューカード　（元　フロンティアサービス研究所）

図2　高速サーバの外観（手前側が高速サーバ）

高速サーバは、他のシステム機器と比較して下記の特徴を有している。

(1) 簡易性

・1～2日程度でシステム構築を完了することできる

・複雑で時間がかかるデータベースの設計等が不要である

(2) 高パフォーマンス

・従来の大規模システムに比べ、チューニングなしで10倍～100倍のパフォーマンスを実現できる

・大容量データの高速なデータ処理が可能である

(3) 低コスト

・チューニング不要であることから、従来の大規模システムに比べ、導入コストとランニングコストの両面で優位性がある

また、高速サーバには、高速化のために特殊なアーキテクチャが実装されているが、その主要なものを下記にあげる。

(1) 非対称型超並列処理（Asymmetric Massively Parallel Processing：AMPP）

図3にAMPPアーキテクチャのデータ処理イメージを示す。ユーザーやアプリケーションがリクエストした高速サーバへの命令は、まずSMP（Symmetric Multi-Processing：対称型マルチプロセッシング）層で、クエリ解析、実行計画生成、最適化や並列化などを実施する。SMP層で並列化されたジョブがMPP（Massively Parallel Processing：超並列処理）層に渡され、MPP層にある多数のノード（CPU、メモリ、FPGA（Field-Programmable Gate Array）等で構成されるデータ処理ハードウェア）によって並列的に実行される。このように、SMPとMPPの長所を併せ持つアーキテクチャであることから、ユーザーやアプリケーションは、複雑な処理を実行せずに、高速並列処理を実現できる。

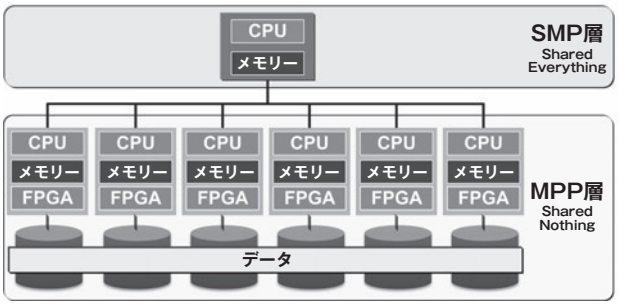

図3　AMPPアーキテクチャ
（出展：日本アイ・ビー・エム(株)；PureSystem資料）

(2) FPGAによるストリーミング処理

HDDから読み出されたデータは、まずMPP層内の各FPGAに入力される。FPGA内では、読み出されたデータのレコード絞り込みやカラム抽出をストリームで実行する。絞り込みされたデータは、メインメモリに転送され、CPUで最終的に処理される。このように専用ハードウェアを使用してリクエストの返答に必要なデータだけを抽出していることから、CPUでの処理を最小限に抑えられる。

2.3　データマイニングツール

高速サーバでの分析（高速サーバへの問い合わせ）は、データベース言語やプログラミング言語で実装することも可能であるが、プログラム作成とその検証に大きな手間と労力を要する。このため、GUI（Graphical User Interface）ベースで直感的にデータを分析することができるソフトウェア（データマイニングツール）を導入した（図4）。データマイニングツールを導入する利点を下記に示す。

①GUIであることから直感的な操作で分析が可能

②データ分析の手順が一目で確認可能

③豊富な統計モデルを実装済み

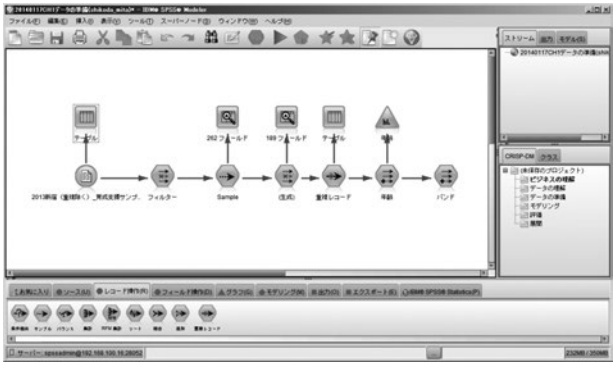

図4　データマイニングツール

2.4 ETLツール

1章でも述べたようにビッグデータは、分析のために蓄積しているデータではないことがあるため（代表例はさまざまな機器のログデータ）、文字化けや欠損値などエラーレコードを含有している場合がある。このため、これらのデータを分析するためには、エラーレコードを一定のルールに従って加工する必要がある。本システムでは、高速サーバへのデータロードと同時に自動的にデータを変換・加工できるETLツールを導入した。導入したETLツールはデータマイニングツールと同様にGUIベースで処理フローを生成可能なことから、操作性に優れている。

2.5 分析端末

分析端末は汎用のパーソナルコンピュータである。現在、分析室には5台の端末が設置済みで、それぞれの端末から高速サーバなどを利用した分析を実施することができる。図5に分析端末を設置している分析室の外観を示す。

図5 データ分析室

2.6 ネットワーク

高速データ分析システムのネットワーク構成を図6に示す。

ネットワークは、外部へのデータ漏洩防止の観点から、社内専用回線や他のインターネット網から独立したスタンドアロン方式を採用した。

分析端末－各サーバの通信回線はLANケーブルによる配線で伝送速度は1Gbpsである。一方、各サーバ間の通信回線は、大容量データの転送を行うことから、FCケーブルで接続されており、伝送速度は10Gbpsとなっている。

尚、高速サーバのホストマシンは冗長化されており、片方が故障しても、処理を継続できるよう設計されている。

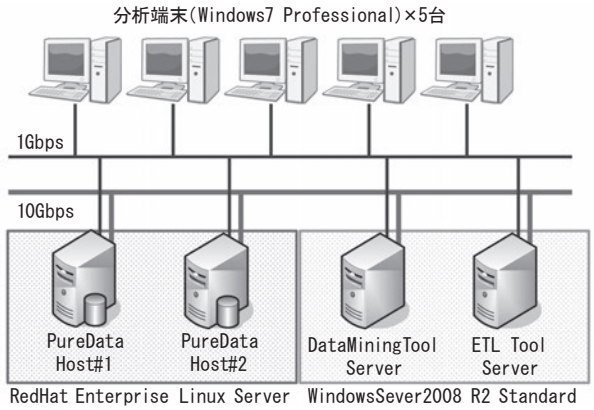

図6 ネットワーク構成図

2.7 セキュリティ対策

今回構築した高速データ分析システムは、2.6でも述べたように、外部ネットワークから独立したスタンドアロンとなっていることから、外部からの攻撃による情報漏洩のリスクは小さい。一方、本システムで最も心配されるのは、USBフラッシュメモリなどによるデータ入出力時のウィルス感染とデータ持ち出しによる情報漏洩である。このため、本システムでは下記のセキュリティ対策を実施している。

(1) 情報漏洩防止ソフトウェアの導入

電子媒体の情報持ち出しによる情報漏洩を防止するために、「情報漏洩防止ソフトウェア」を導入した。

ユーザーは、情報セキュリティ管理者の許可がなければ、データを分析端末から持ち出すことができない。さらにすべてのデータは、移動履歴などのファイル操作がロギングされており、過去履歴から持ち出し人物や流出経路などの特定が可能である。

(2) ウィルス対策ソフトウェアの導入

分析端末には、「ウィルス対策ソフトウェア」を導入した。尚、本システムは外部インターネット網に接続していないため、ウィルス定義ファイルは定期的に手動アップデートを実施している。

3. 高速データ分析システム導入の効果

データ分析は、さまざまな視点からモデル作成を行い、失敗を繰り返して精度や知見を高めていく「トライアンドエラー方式」で作業を進めていく。このため、本システムの導入による処理速度の高速化によって、分析効率が格段に向上（デスクトップサーバマシンで数時間を要していた処理を本システムでは数分、場合によっては数十秒で処理可能）し、今後分析をより深度化させることで、新たな知見が明らかになることが期待される。

また、従来のデータベースシステムは操作が難しく、ユーザーが表やグラフを得るまでの作業が煩雑であることから、誰もが簡単に必要な情報を得られる仕組みについて検討を進めている。4章では、システムの運用開始後にフロンティアサービス研究所で実施したプロトタイプの一例を紹介する。

4. これまでの分析成果の紹介

膨大な情報を実際に現場で活用する場合、迅速かつ簡単にデータを確認できることようにすることが運用上望ましいと考えられる。このため、実運用を考慮した研究の一環として、BI（Business Intelligence）ツールを活用したプロトタイプのアプリケーションに関する研究を実施している。

BIツールとは、高速サーバなどのデータベースに格納された大量の数値データを高度な知識がなくても扱えるようにしたソフトウェアのことである。

図7・8はこれまでの研究で作成したビッグデータ利活用に関するプロトタイプの画面イメージである。このプロトタイプでは、ユーザーは膨大なデータが対象であることを意識せず、マウス操作数クリックの簡易な操作で表やグラフを取得できる。

図7　操作画面イメージ

図8　表示画面イメージ

このほかにも、シミュレーション結果などの可視化を行うシステムにBIツールを活用できないか、実運用上の視点から検証を行っている。

5. 当社のビッグデータ利活用の将来像

これまでは、ビッグデータ分析のために導入したシステムの概要とシステムを利用した分析成果について紹介してきたが、最後に今後実施予定の研究テーマと私たちが考えている当社におけるビッグデータ利活用の将来像について、いくつか紹介する。

(1) 業務革新の実現

私たちは、現場第一線から本社まで、誰でも簡単にデータ分析結果（予測値）に基づいた「意思決定・行動」を取れるようなシステム（アプリケーション）を構築することで社内の業務革新を実現したいと考えている。具体的には、各種施策の精緻な効果予測や駅構内の混雑予測などのシミュレータの実用化をめざしていく。

(2) お客さまサービスの向上

(1)のシミュレータなどの情報を活用して、お客さまに応じたリアルタイム性の高い情報提供を実現したいと考えている。これは、お客さまがいま求めている情報（駅や列車の混雑予測、遅延予測等）を的確に把握し、迅速に配信するサービスの実用化をめざしていく。さらに確度の高い情報を適切なタイミングで配信することは、輸送障害時の混雑緩和などへの効果も期待できる。

上述のような将来像の実現に向けて、フロンティアサービス研究所では今後順次、データ分析やシステム開発・試験を実施していく予定である。

6. まとめ

フロンティアサービス研究所では、社内外のビッグデータ分析を推進するためのプラットフォームとして「高速データ分析システム」を導入した。導入に当たっては、分析効率化、操作性、セキュリティなどを考慮して、機器・ソフトウェアの選定を行った。今後、導入したシステムを活用して「社内業務革新」と「お客さまサービスの向上」を実現するアプリケーションの研究開発を推進していく予定である。

参考文献

1）喜連川優；情報爆発とこれから，電子情報通信学会誌，Vol.94, No.8, pp.662～666,2011.

2）John Gantz, David Reinsel; The Digital Universe Decade - Are You Ready?, IDC-IVIEW, 2010