

超高速データ分析プラットフォーム 「InfoFrame DWH Appliance」

孝忠 大輔・三国 晋章

要旨

爆発的に増え続けるデータからタイムリーに価値ある情報を抽出するために、ビッグデータをアドホックに分析したいというニーズが高まっています。高速分析プラットフォームとして「InfoFrame DWH Appliance」が注目されている理由と、Webサイトにおけるアクセスログをアドホックに分析し、マーケティングに活用している事例について紹介します。

キーワード

●ビッグデータ ●DWHアプライアンス ●情報活用 ●アドホック分析 ●Webアクセスログ

1. まえがき

ビッグデータから価値ある情報を抽出するためには、定型化された分析アプローチのみではなく、さまざまな角度からデータを眺めるアドホックな分析アプローチが必要となります。本稿では、超高速データ分析プラットフォーム「InfoFrame DWH Appliance（インフォフレーム・データウェアハウス・アプライアンス）」が、ビッグデータをアドホック分析する際に求められる高速性と簡易性を、どのように実現しているかについて説明します。また、InfoFrame DWH Applianceを導入し、Webサイトにおけるアクセスログをアドホックに分析している事例を紹介します。

2. ビッグデータへの分析アプローチ

近年、ネットワークやデバイスの進歩、SNSなどのユーザーコミュニティの活性化により、大量のデータが生み出されるようになりました。従来では扱うことができなかったビッグデータを収集・蓄積・分析することは可能になりましたが、ビジネスへの活用という点では、まだ方法論が確立されていない黎明期といえます。

ビッグデータから価値ある情報を取り出すためには、販売分析や顧客分析に代表される定型化された分析アプローチではなく、さまざまな角度からデータを眺めるアドホックな分析アプローチが必要となります。

従来、多くの企業で実施されている販売分析や顧客分析では、どのようなデータを収集し、どのような指標で分析す

ればよいか、既に体系的に整理されているため、分析の進め方でつまづくことは、ほとんどありません。それに対して、企業戦略の立案や意志決定に役立つ情報をビッグデータから抽出したり、ビッグデータを用いた新たなサービスを作り出したりするための方法論は確立されていないため、手探りで分析を行う必要があります。つまり、ビッグデータに対する分析アプローチは、あらかじめ想定されていない分析パターンを試行錯誤しながら行う「アドホック分析」が中心になるということです。

3. ビッグデータとアドホック分析

「大量、多様、高頻度」という特性を持ったビッグデータをアドホック分析する際に求められる要件は次の3つです。

(1) 高速性：

数十テラバイト（TB）～数ペタバイト（PB）の大量データを高速処理できること

(2) 簡易性：

多種多様なデータを簡単に取り込み、組み合わせて分析できること

(3) 迅速性：

高頻度で発生するデータをタイムリーに分析できること
3つの中で特に考慮すべき要件は「簡易性」です。現在、方法論が確立されておらず手探りで進めなければならないビッグデータの分析では、どのデータ種と、どのデータ種を組み合わせるべきかを事前に想定することは難しいため、あらゆるデータ種を分析対象として分析することになります。

このように、ビッグデータの分析においては数多くのデータ種を取り扱う必要があるため、多種多様なデータを簡単に取り込み、組み合わせて分析できるという「簡易性」が重要になるのです。

4. アドホック分析のための高速分析プラットフォーム

NECでは、ビッグデータを高速かつ簡単にアドホック分析するための分析プラットフォームとして「InfoFrame DWH Appliance」を提供しています。InfoFrame DWH Applianceは、弊社の高性能・高信頼なサーバとストレージに、Netezzaの高度なデータ検索/分析を行うアーキテクチャとソフトウェアを組み合わせて最適化したデータウェアハウス・アプライアンス製品です（図1）。

InfoFrame DWH Applianceが、ビッグデータをアドホック分析する際に求められる高速性と簡易性をどのように実現しているかについて紹介します。

4.1 InfoFrame DWH Applianceの高速性

ビッグデータのような数十TB～数PBの大量データを処理する場合、「できるだけデータを動かさない」ことが大切です。大量データの不要な移動は、I/Oボトルネックの発生や、CPU及びメモリにおける過剰処理につながり、結果としてシステム・パフォーマンスに大きな悪影響をもたらします。InfoFrame DWH Applianceでは、「FPGA（Field Programmable Gate Array）」という汎用デバイスを利用して、データ・ストリーミングのできるだけ早い段階でデータのフィルタリング処

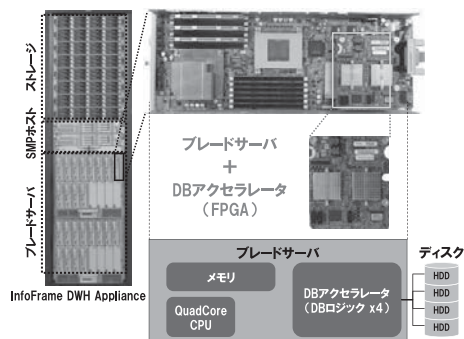


図1 InfoFrame DWH Applianceの構成

理（データがディスクから取り出されるのと同時に不要なデータを取り除くなど）を実施しています。このようにデータの存在するディスクの近くでデータ処理することにより、不必要な大量データの移動を避けています（図2）。

また、InfoFrame DWH Applianceは大量データを高速処理するために、「超並列処理アーキテクチャ（Massively Parallel Processing：MPP）」を採用しています。対称型のマルチプロセッシング・ホストは、超並列処理ノードのグリッドに接続されており、データの検索や分析といった重い負荷に対処するためにMPPグリッドを活用します。InfoFrame DWH Appliance ZA100モデルの場合、1つのクエリを92並列で処理することができます。このようにMPP技術を活用して大量データを並列実行することによって、ビッグデータに対する高パフォーマンスを実現しています（図3）。

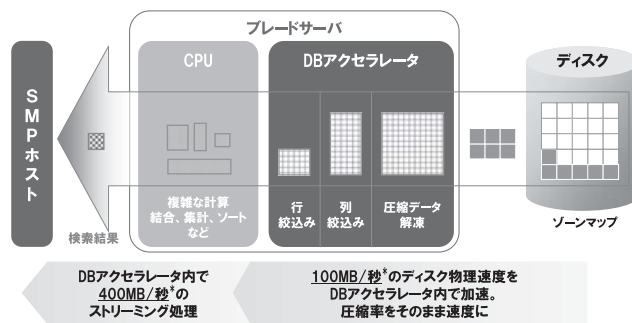


図2 FPGAによるデータ・ストリーミング処理

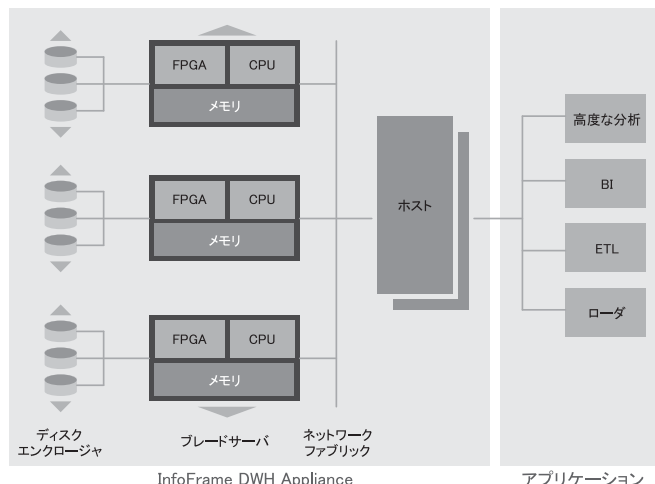


図3 InfoFrame DWH ApplianceのMPPアーキテクチャ

4.2 InfoFrame DWH Applianceの簡易性

InfoFrame DWH Applianceは、「シンプルなアプライアンス構成」を目指して「簡易性」を重視した設計となっています。InfoFrame DWH Applianceにロードされたデータは、データベース管理者が関与しなくても、自動的にデータ分散が行われて格納されます。また、InfoFrame DWH Applianceに投げられたクエリは、超並列グリッド内で自動的に高速処理されます。そのための事前準備として、一般的なDB技術では必要とされるデータベース・チューニング作業（インデックス作成やパーティショニングなど）を行う必要はほとんどありません。

このようなInfoFrame DWH Applianceの持つ「簡易性」は、ビッグデータのアドホック分析において大きな効果を発揮します。場合によっては、数百種類のデータ種を組み合わせることで分析することになるビッグデータの分析において、新たなデータ種が追加されるたびに、データベース・チューニング作業を実施することは現実的ではありません。新たなデータ種を追加しても特にチューニングを必要とせず、ロードすればすぐ分析に使えるという「簡易性」が、ビッグデータのアドホック分析では重要になるのです。

5. 高速分析プラットフォームによるビッグデータ活用事例

ビッグデータをアドホックに分析するという取り組みは、ネット/通信業界の企業において、ビジネスをより大きく成功に導くため、積極的に実施され始めています。本章では、大量に蓄積しているWebアクセスログの分析を円滑・迅速に行うために、弊社のInfoFrame DWH Applianceを導入し、マーケティングやWebサイト改善に活用しているお客様の事例を紹介します¹⁾。

5.1 Webアクセスログの活用

現在発展途上にあるビッグデータの活用例として取り上げられることが多い「Webアクセスログ」は、さまざまな業務領域で活用されています。例えば、コンテンツの企画・開発担当者の場合、アクセスログから読み取れるユーザーの閲覧状況（サイト内での振る舞い・行動）を分析することによって、Webサイトの改善や新サービスの開発につながっています。

また、登録済みユーザーの閲覧履歴を分析し、レコメンド（推奨）につなげたりしています。営業担当者場合は、Webサイトへのアクセス傾向を分析することによって、プロモーション提案や顧客獲得につながっています。

5.2 Webアクセスログのアドホック分析

実際にWebアクセスログ分析を実施されているお客様の声を聞いてみると、Webアクセスログの分析は、定型的な分析というよりは、アドホックな分析になると仰っています。例えば、特定のサービスを利用したことがある人や、特定のプロモーションページを閲覧したことのある人だけに絞り込んだ分析を実施しようとした場合、その都度、分析軸を設定して分析する必要があります。また、ユーザーの詳細な振る舞い・傾向などを把握するためには、Webアクセスログだけではなく、複数データをミックスして分析する必要があります。このように、プロモーションのテーマや企画の目的によって分析内容及び分析データが異なるため、アドホックな分析になるということです。

5.3 InfoFrame DWH ApplianceによるWebアクセスログ分析

InfoFrame DWH Applianceは、大量のWebアクセスログのアドホック分析においても、高パフォーマンスを発揮することができます。弊社では、InfoFrame DWH Applianceを検討されているお客様に対して、POC（Proof of Concept：実機検証プログラム）を実施しています。実際にお客様のデータを使って、InfoFrame DWH Applianceの「高速性」と「簡易性」を確認していただくことが目的です。

お客様のWebアクセスログをお預かりし、POCを実施した結果の例を以下に紹介します。

- (1) 店舗・会員データ、コンテンツのログデータなど約160万件を対象に、従来のシステムでは3時間半程度要していた抽出・集計処理を、InfoFrame DWH Applianceで検証したところ、わずか11秒で完了。
- (2) 3.5億件（20GB）のWebアクセスログに対するあいまい検索をInfoFrame DWH Applianceで検証したところ、3秒で完了。

このようなWebアクセスログ分析を数パターン検証した結果、従来のシステムと比較して、平均で576倍、最高で約1,000

倍のパフォーマンスを実証することができました。また、POCを通じて、InfoFrame DWH Applianceは特別なチューニングが不要であり、運用管理がそれほど負担にならないことを高く評価していただくことができました。その結果、InfoFrame DWH Applianceを導入してWebアクセスログを分析した結果を、コンテンツ担当者や営業担当者へタイムリーに提供することが可能になっています。

6. むすび

Webアクセスログの分析事例でも紹介したように、ビッグデータに対する分析の特徴は、アドホックな分析アプローチになるということです。方法論が確立されておらず手探りで進めなければならないビッグデータの分析においては、何度も試行錯誤できる「高速性」と、多種多様なデータを取り込み、組み合わせて分析できるという「簡易性」が重要となります。弊社はその両方の要件を満たし、大量データをアドホックに分析できるソリューションとして、InfoFrame DWH Appliance (Netezza) を、50社以上のお客様でご活用いただいています。今後も、企業戦略の立案や意志決定に役立つ情報をビッグデータから抽出したり、ビッグデータを用いた新たなサービスを作り出したりするための分析プラットフォームとして貢献できるよう努めてまいります。

*Netezzaは、International Business Machines Corporationの登録商標または商標です。

参考文献

- 1) InfoFrame DWH Appliance導入事例 株式会社ぐるなび様
<http://www.nec.co.jp/library/jirei/gnavi2/contents.html>

執筆者プロフィール

孝忠 大輔
ITソフトウェア事業本部
ITプラットフォームソリューション
事業部
主任

三国 晋章
ITソフトウェア事業本部
ITプラットフォームソリューション
事業部
システム部長

関連URL

InfoFrame DWH Appliance製品情報:
<http://www.nec.co.jp/pfsoft/dwhappliance/>

NEC 技報のご案内

NEC 技報の論文をご覧くださいありがとうございます。
ご興味がありましたら、関連する他の論文もご一読ください。

NEC技報WEBサイトはこちら

NEC技報(日本語)

NEC Technical Journal(英語)

Vol.65 No.2 ビッグデータ活用を支える 基盤技術・ソリューション特集

ビッグデータ活用を支える基盤技術・ソリューション特集よせて
ビッグデータを価値に変えるNECのITインフラ

◇ 特集論文

データ管理/処理基盤

超高速データ分析プラットフォーム [InfoFrame DWH Appliance]
SDN 技術で通信フローを制御する [UNIVERGE PF シリーズ]
大量データをリアルタイムに処理する [InfoFrame Table Access Method]
大量データを高速に処理する [InfoFrame DataBooster]
ビッグデータの活用最適なスケールアウト型新データベース [InfoFrame Relational Store]
高い信頼性と拡張性を実現した Express5800/ スケーラブル HA サーバ
大規模データ処理に対する OSS Hadoop の活用
大容量・高信頼グリッドストレージ iStorage HS シリーズ (HYDRAStor)

データ分析基盤

ファイルサーバのデータ整理・活用を支援する [Information Assessment System]
超大規模バイオメトリック認証システムとその実現
WebSAMの分析技術と応用例～インバリエント分析の特長と適用領域～

データ収集基盤

スマートな社会を実現する M2M とビッグデータ
微小な振動を検知する超高感度振動センサ技術開発とその応用

ビッグデータ処理を支える先進技術

多次元範囲検索を可能とするキーバリューストア [MD-HBase]
高倍率・高精細を実現する事例ベースの学習型超解像方式
ビッグデータ活用のためのテキスト分析技術
ビッグデータ時代の最先端データマイニング
ジオタグ付きデータをクラウドでスケラブルに処理するジオフェンシングシステム
柔軟性と高性能を備えたビッグデータ・ストリーム分析プラットフォーム [Blockmon] とその使用事例

◇ 普通論文

地デジ TV を活用した「まちづくりコミュニティ形成支援システム」

◇ NEC Information

NEWS

スケールアウト型新データベース [InfoFrame Relational Store] が 2 つの賞を受賞



Vol.65 No.2
(2012年9月)

特集TOP