

HA8000-bd/BD10を活用した Hadoopビジネスへの取り組み

Use of High-density Server for Hadoop Enterprise Business

藤田 あずさ

Fujita Azusa

人見 洋一

Hitomi Yoichi

中村 賢

Nakamura Satoshi

高橋 初美

Takahashi Hatsumi

ビッグデータアナリティクスを実現する一つの技術であるOSSの分散処理ソフトウェアHadoop^{※1)}が注目されている。Hadoopはノード数に応じて処理性能が向上する特性があり、多数のサーバノードを並べることでビッグデータアナリティクスに対応する。

日立グループは、Hadoopに適したプラットフォームである高集積・低消費電力を特長とするHA8000-bd/BD10と、Hadoopを使いやすくするための技術支援や関連製品を組み合わせ、製品化している。

1. はじめに

増え続けるビッグデータを活用するためのビッグデータアナリティクスを実現する一つの技術として、Apache Software Foundation (Apacheソフトウェア財団) が開発しOSS (Open-source Software) として公開している分散処理ソフトウェアHadoopが注目を集めている。Hadoopは、大量のデータを多数のコンピュータ上で分散して処理するためのOSSであり、大量の非構造化データを処理し、新しい価値を見つけることが可能と言われている。しかし、新しい技術であるHadoopのプログラム実装、サーバ構成

の決定、性能チューニングなど、使いこなすにはテクニックが必要である。日立グループは、さまざまなビッグデータの利活用に取り組んでおり、Hadoopに関連する取り組みもその一つである。

ここでは、HadoopとエンタリーブレードサーバHA8000-bd/BD10を組み合わせた構築事例と、簡易に導入できるセットとして商品化している「かんたんHadoopソリューション for ログ解析 (QlikView^{※2)} & JP1)」について述べる。

2. HadoopにHA8000-bd/BD10を使うメリット

高集積・低消費電力を特長とするHA8000-bd/BD10は、Hadoopでの活用に適したプラットフォームである (図1 参照)。

2.1 HA8000-bd/BD10の特長

HA8000-bd/BD10は5Uのベースユニットに最大40ブレード搭載可能な高集積プラットフォームである。ブレードは1スロット幅と2スロット幅の2種類あり、2スロット幅のブレードの場合、ベースユニットに20ブレードまで搭載できる。Hadoopを使ってビッグデータを扱う場合でも、導入当初は小規模に10ブレードから始め、データ量の増加に合わせてブレードを追加するというような投資の平準化ができる。

サーバブレードは、運用時31 W^{※3)}の低消費電力で動作可能な省電力設計となっている。1スロットモデル40ブレードフル搭載のHadoopクラスターでも、消費電力1,240 W

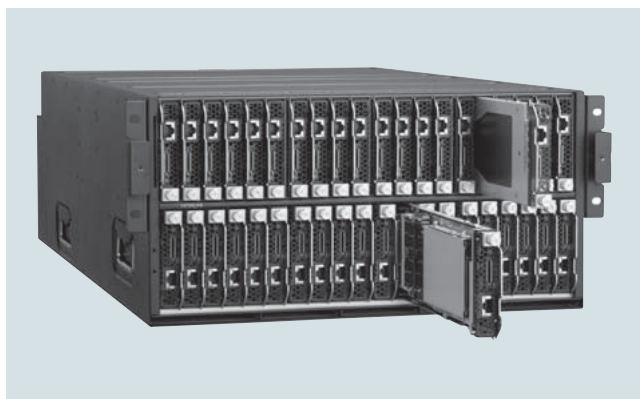


図1 | 高集積エンタリーブレードサーバHA8000-bd/BD10
5Uのベースユニットに40枚のサーバブレードが入る。

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

※2) QlikViewは、QlikTech International ABの商標または登録商標である。

※3) サーバブレード40台をHA8000-bd/BD10用5Uベースユニットに搭載した場合の、通常運転時の消費電力の目安が1,240 Wで、1台当たりが31 W。実際の消費電力は搭載するサーバブレード、モジュールの構成によって変わる。

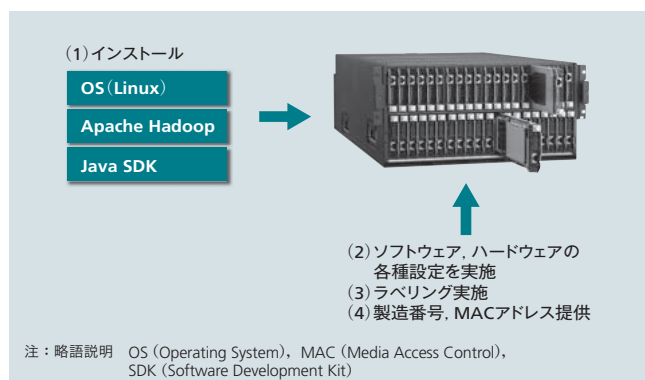


図2 | プラットフォーム設定サービス

Hadoopを実行するために必要なソフトウェアのインストールや設定、ハードウェア管理のための製造番号、MACアドレスなどの管理情報を提供する。

程度での運用が可能であり、電力コストを抑えることができる。

Hadoopは処理を分散させることによって大量データを扱うことができ、ノード数に比例してデータ処理性能が上がる。HA8000-bd/BD10は少ないスペース・消費電力で多くのノードを並べられることから、Hadoopのスケールアウト性能を出す構成を組むことに適している。

2.2 プラットフォーム設定サービスとチューニング

ユーザー企業でHadoopを導入する際の最初のハードルとなるのが、多数のノードへのOS (Operating System) およびHadoopインストールとHadoopの初期設定である。そこで、「インストールサービス for Hadoop」と「プラットフォーム設定サービス for Hadoop」を提供している(図2参照)。

これらのサービスにより、Hadoopシステムに必要なOS (Linux^{※4)}、Apache^{※1)} Hadoop、Java^{※5)} SDK (Software Development Kit) をインストールする。さらに、日立グ

ループにおけるHadoopシステム構築のノウハウを活用し、Hadoopマスタ(親)ノードとスレーブ(子)ノードが動作するための各種パラメータ設定も実施し、提供している。

このサービスを利用すると、サーバ台数が多くなるHadoop環境の構築に必要な作業を事前に実施し、導入の手間を軽減できる。また、導入後、業務運用開始までの期間も短縮可能となる。

Hadoopクラスタの性能を最大限に引き出すためには、アプリケーションの性質とサーバの構成から最適な設定値にチューニングする必要がある。ジョブの実行時間のチューニングでは、Hadoopの基本的なプロパティ、HDD (Hard Disk Drive) 台数、データ圧縮の有無など、そのつど調整が必要になる。日立グループはHadoopに適したサーバとして、HA8000-bd/BD10を使って性能検証を実施している。一例として、Hadoopに付属されているサンプルプログラムのTeraSortをHA8000-bd/BD10で実行し、HDDの台数に応じてタスクの実行に関するパラメータなどをチューニングした結果、チューニング前と比較して約1.8倍のスループットを得ることができた。

3. 大規模構築事例

3.1 Hadoopシステム事例の概要

日立グループは、Hadoopを使った大量ログ解析システムを顧客へ納入しており、一部が稼働済みである(図3参照)。

このシステムの導入における顧客の課題は、大量に発生するログの蓄積、ログの一元管理とデータ抽出・結合の高

※4) Linuxは、Linus Torvalds氏の日本およびその他の国における登録商標または商標である。

※5) Javaは、Oracle Corporation およびその子会社、関連会社の米国およびその他の国における登録商標である。

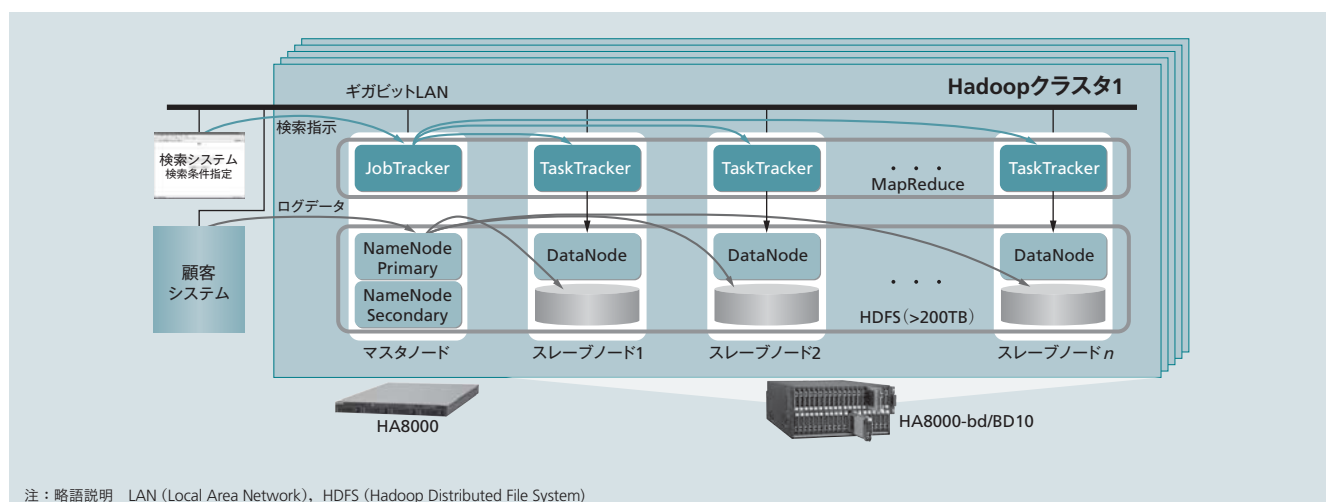


図3 | 大規模構築事例のシステムイメージ

顧客システムからログデータを転送し、HDFSに保存する。検索システムから保存されているログデータの検索を実施する。HadoopマスタノードはHA8000、スレーブノードはHA8000-bd/BD10で構成されている。

速実施、および加速化しているログの発生スピードであった。そこで、Hadoopを大規模ストレージ基盤として導入し、大量のログを蓄積、分散処理することで、これらの課題解決をめざした。現在は、稼働時のデータ需要に合わせてスレーブノード240台でログ検索処理を分散し、高速検索を実現している。

将来的には、データ需要増加に対応して、ノード数を増設し、合計1,000ノード超、総容量1ペタバイトとなる予定になっている。多数のノードの環境構築が必要になるため、環境構築を自動化し、ノード増設作業を省力化できるようにしている。なお、高集積・省電力なHA8000-bd/BD10の適用により、全ノード設置完了時でも、サーバラック5台に収容が可能である。

3.2 Hadoopシステム構築のポイント

Hadoopシステムの構築にあたっては、Hadoopの特性への深い知識と理解が必要であり、従来のシステム構築に比べて、設計が複雑化しやすい。特にリソース設計、運用設計、パラメータ設計、アプリケーション設計などの設計フェーズにおいては、設計と検証を繰り返し、設計を最適化することが必要である。

Hadoopシステムのアプリケーション性能については、業務処理、データの特性による影響が大きいため、本番相当のデータ、アプリケーションを用いて検証することが必要である。この事例においては、100ノードの検証環境を構築し、事前に性能検証を実施することで顧客要求の性能を満たすシステムを実現した。

なお、信頼性を上げるために1,000ノードをマルチクラスタ構成とし、Hadoopマスタノードの障害が発生しても

24時間365日連続的に出力される大量ログを確実に蓄積するように設計している。また、マルチクラスタ構成によってHadoopクラスタ単位でのメンテナンスが可能になり、システム無停止での運用を実現している。

さらに、統合システム運用管理ソフトウェア(JP1/AJS3, JP1/PPM)とHadoopを連携させることで、Hadoopクラスタのリソース管理とジョブ管理を可能にし、加えて各クラスタのHDFS(Hadoop Distributed File System)使用率を監視することでクラスタごとの使用量の均等化を実現している。

4. かんたんHadoopソリューション for ログ解析

4.1 かんたんHadoopソリューションのねらい

Hadoopというキーワードは注目されているが、ユーザー企業からは、Hadoopを使って何ができるのか分からないという声を多く聞いていた。そこでHadoopを活用する用途の一つとして、ログ解析を取り上げ、ユーザー企業での活用イメージを想起できるように「かんたんHadoopソリューション for ログ解析(QlikView & JP1)」を販売している。

このソリューションにより、ユーザー企業のシステム設計負担を軽減し、システム構築立ち上げを迅速化することが可能と考えている。また、Hadoopは分散処理の実行基盤にすぎないため、ユーザー企業においてHadoopシステムを活用するにはBI(Business Intelligence)ツールや運用管理ツールなど他のソフトウェアとの連携が必須である。このソリューションのセットでは、Hadoop自体の環境設定に加えて、これらのソフトウェアとの連携の環境設定などをあらかじめ日立製作所で実施して顧客納入することに

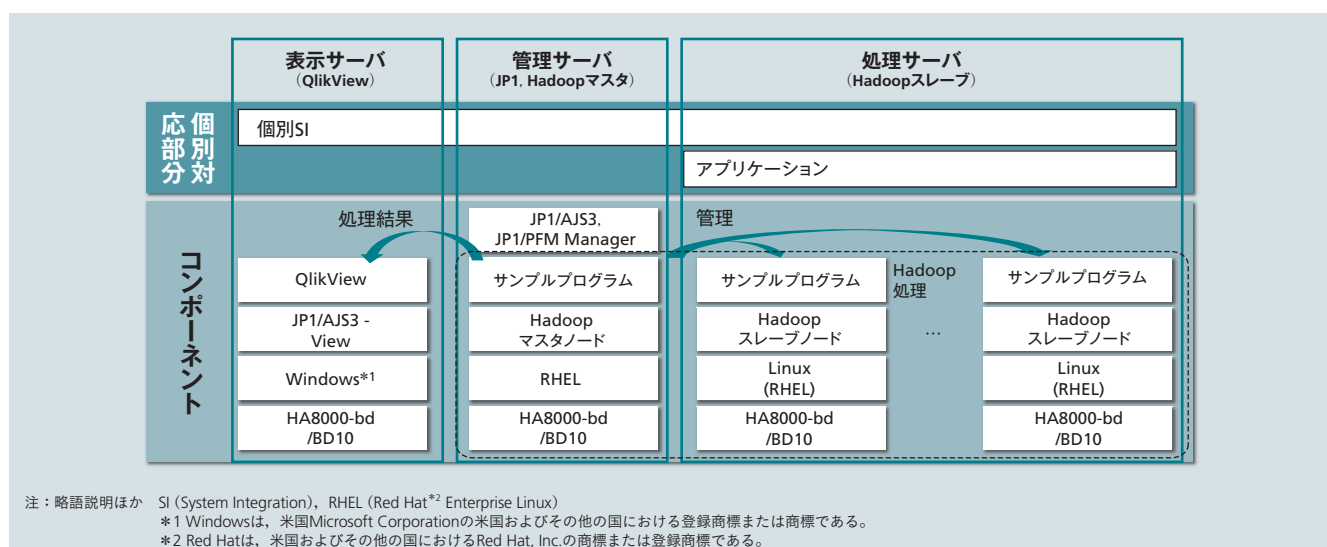


図4 「かんたんHadoopソリューション for ログ解析」のシステム構成概要

QlikViewをインストールした表示サーバ、Hadoopマスタノードである管理サーバ、Hadoopスレーブノードである処理サーバから成る。JP1は管理サーバにインストールされ、全サーバの管理をする。

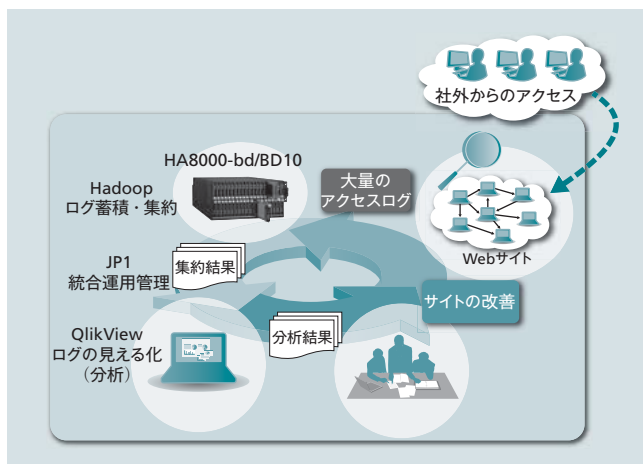


図5 「かんたんHadoopソリューション for ログ解析」の使用イメージ
自社WebサイトへのアクセスログをHadoopで集約し、QlikViewで集約データを分析して、分析結果をWebサイトの改善に適用する。

より、顧客負担の軽減を実現している。

4.2 かんたんHadoopソリューションの概要

かんたんHadoopソリューション for ログ解析（QlikView & JP1）は、Webサイトのアクセス記録（ログ）解析向けに、高集積のエントリーブレードサーバHA8000-bd/BD10と、インメモリ型BIプラットフォーム「QlikView」、統合システム運用管理「JP1」といったソフトウェア群を組み合わせ、効率的なWebログ解析が行えるシステムを容易かつ迅速に導入できる形で提供している（図4参照）。

QlikViewは、データの解析・可視化に適したインメモリ型BIプラットフォームである。QlikViewを使うとシンプルかつスピーディにグラフの作成が可能で、集約処理結果と顧客マスターデータなどを組み合わせ、ドリルダウン分析など自由度の高い分析が可能となる。

また、JP1により、HA8000-bd/BD10とHadoop環境のジョブ管理、稼働監視などを可能にし、サーバ台数が多くなるHadoopシステムの管理を効率化できる。従来システムと連携させる際にも、従来の業務とHadoopの業務を一元的に管理することが可能となる。

さらに、Apache Webサーバのサンプルログと、ログのデータ集約処理を行うHadoopのサンプルプログラムを添付しており、Hadoopに関する詳しい専門知識のないユーザーでも機器導入後すぐにHadoopを利用した試用分析が可能となっている。

このソリューションで提供されるシステムにより、単なるWebへのアクセスログである非構造化データを蓄積し分析することで「誰がどのようなタイミングでWebを参照しているのかを知ること」や「製品の社外発表やWebの更新による効果の確認」などが可能になり、次の一手への対

表1 Hadoopのビジネス活用例

HadoopとHA8000-bd/BD10を組み合わせた「かんたんHadoopソリューション」の適用に有望と考えている業務の例と概要を示す。

適用業務	適用概要
コールセンター	対応音声データを蓄積、分析することによってサービスや製品を改善
商品企画	SNSなどユーザーの声を分析し、トレンド予測することによってヒット商品開発
仕入れ管理	気象データを小売業での仕入れ管理に活用
原価管理	時間がかかりすぎる業務バッチ処理の高速化

注：略語説明 SNS（Social Networking Service）

応が素早くできるようになるという新しい価値を生み出すことが期待できる（図5参照）。

5. おわりに

ここでは、HA8000-bd/BD10を活用したHadoopビジネスと「かんたんHadoopソリューション for ログ解析（QlikView & JP1）」について述べた。

Hadoopの国内市場はまだ黎明（れいめい）期にあり、一部の先進的ユーザーが活用しているにすぎない。

HadoopもLinuxと同様に、これから多くのユーザーに処理基盤として普及し、裾野が広がっていく可能性があると言われている。日立グループは、これからもユーザー企業に採用され、新しい価値を創出することができるソリューションを形にしていきたいと考えている（表1参照）。

執筆者紹介



藤田 あずさ

1997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 システム基盤ビジネス本部 クライアント部 所属
現在、HA8000-bd/BD10と関連ソリューションの製品計画に従事



中村 賢

1992年日立中部ソフトウェア株式会社入社、株式会社日立ソリューションズ プロダクトソリューション事業本部 システムプロダクト事業部 システム基盤本部 第5部 所属
現在、Hadoopを活用したソリューション事業、プロダクト開発に従事



人見 洋一

1989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 開発統括本部 ソフトウェア本部 第2プラットフォームソフトウェア設計部 所属
現在、Hadoop MapReduceサポートサービスのサポート業務に従事



高橋 初美

2006年日立アイ・エヌ・エス・ソフトウェア株式会社入社、パッケージソリューション事業本部 パッケージソリューション部 所属
現在、新規商材のソリューション企画に従事