

## シスコと Informatica の共同ソリューションによるデータウェアハウスの最適化



今日の情報を基礎とするビジネス志向において、組織として、顧客の獲得および維持、業務効率の向上、製品およびサービス提供の強化、新規市場への参入といった目標の達成に向け、さまざまなソースのデータを統合する必要に迫られています。

今日の情報を基礎とするビジネス志向において、組織として、顧客の獲得および維持、業務効率の向上、製品およびサービス提供の強化、新規市場への参入といった目標の達成に向け、さまざまなソースのデータを統合する必要に迫られています。

企業がこうした目標を達成するために、クリーンかつ安全でアクセスしやすく、タイムリーに利用できる実用的なデータに加え、データ量、種類、速度の驚異的な増大に適応できる拡張性を備え、多様なエンタープライズ データ処理ワークロードにも対応できる分析環境が必要となります。データウェアハウスを補完し、拡張性への新たな要求に対応するものとして、Apache Hadoop などのビッグデータ プラットフォームが登場しました。従来の「スケール アップとスケール アウト」アプローチがハイエンド サーバまたはアプライアンスと高価な共有ストレージ (SAN やネットワーク接続ストレージ (NAS) など) を使用してインフラストラクチャを拡張するのに対し、最新のデータ アーキテクチャでは、より大量のデータをコスト効率よく取り込み、処理できます。

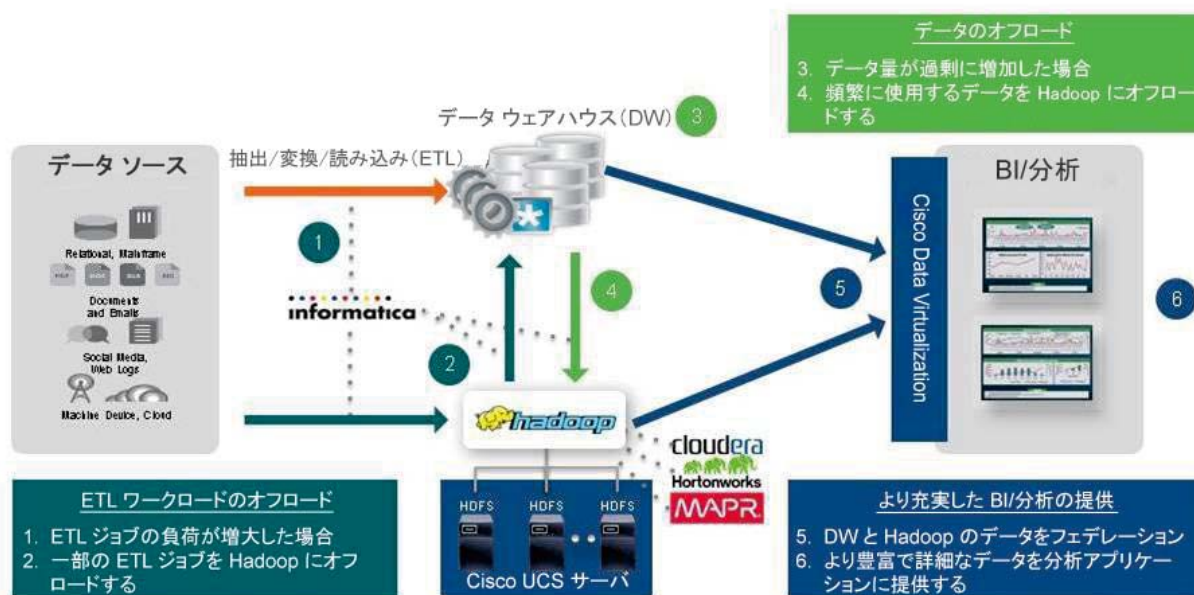
Hadoop および関連ツールを使用してデータ ウェアハウスを最適化することにより、組織には以下のメリットがもたらされます。

- 頻繁に使用するデータをコスト効率の高いストレージにオフロードできるため、データ ウェアハウスのストレージが最適化されます。
- より広範かつ大量のデータを取り込み、保存できるため、ビジネス上の新たな洞察を導き出すことができます。集約されたデータを提示する従来のアプローチとは異なり、データを未加工のまま保存できるため、より高精度の洞察が得られます。
- データ変換 (抽出/読み込み/変換 (ELT)) およびデータ品質ワークロードを移行によって分離できるため、データ ウェアハウス CPU のキャパシティが最適化されます。
- 処理をデータにプッシュする分散環境によって、ネットワーク パフォーマンスがボトルネックとなる問題が解消されます。
- 非構造化データや半構造化データの保存および操作が容易です。

## ソリューション概要

シスコと Informatica は、お客様が Hadoop の価値を迅速に獲得できるように、共同データ ウェアハウス最適化ソリューションを提供しています(図 1)。このソリューションは処理およびストレージをデータ ウェアハウスから Hadoop にオフロードを単一のプラットフォームで実現します。このソリューションは、Cisco Unified Computing System™ (Cisco UCS®)、Informatica Big Data Edition、お客様が選択した Hadoop ディストリビューション、および Cisco® Data Virtualization で構成されます。このソリューションでは、現在のデータ ウェアハウス使用率の分析やオフロードの可能性評価のサービスを提供しています。削減の可能性を見いだした場合は、Cisco UCS サーバにインストールされた Hadoop クラスタ上のシンプルなビジュアル開発環境を使用し、Informatica Big Data Edition によってデータ変換プロセスおよびデータ品質プロセスを実行できます。Cisco Data Virtualization を使用して分散データ環境のフェデレーションを行うことにより、すべてのデータへの一元的なアクセスによるビジネス インテリジェンスおよび分析が可能となります。

図 1 データ ウェアハウス最適化アーキテクチャ



## 製品

### Cisco UCS Intergrated Infrastructure for Big Data

Cisco UCS は、IT プロセスのポリシーベースでの自動化および標準化により、組織がビジネスを行う方法を一変させます。Cisco UCS は、業界標準の x86 アーキテクチャ サーバ、ネットワーキング、およびエンタープライズクラスの管理を 1 つのシステムに統合します。このシステム構成は、一元化されたモデルベース管理によって全面的にプログラム可能であるため、導入の簡素化および迅速化につながります。統合 I/O インフラストラクチャでは、高帯域幅で低遅延のユニファイド ファブリックを使用して、ネットワーキング、ストレージ I/O、および管理トラフィックをサポートします。システムに手動でコンポーネントを組み込むと時間がかかり、エラーも発生しやすくなりますが、このシステムはそれを排除し、自己認識および自己統合型の優れたインフラストラクチャを備えた Wire-Once (配線は初回のみ) アーキテクチャとなっています。

Cisco UCS Integrated Infrastructure for Big Data はビッグデータ向け Cisco Common Platform Architecture (CPA) の第 3 世代であり、スケールアウト アプリケーションに対するさまざまな要求に対応することを目的とした拡張性の高いアーキテクチャです。このソリューションは、農業、教育、金融、医療、サービス プロバイダー、娯楽、保険、公共部門といった分野で広く採用されています。Cisco UCS Integrated Infrastructure ソリューションは高い処理性能と容量の向上をもたらします。さらに、業界トップクラスのパートナーとともにより付加価値の高い包括的なソリューションも提供します。

Cisco UCS Integrated Infrastructure for Big Data では、コンピューティング、ストレージ、接続、および統合管理の機能を備えた包括的なパッケージを簡単に構成確定、手配にいたるため、迅速な導入が可能であり、予測可能なパフォーマンスと総所有コスト (TCO) の削減を実現します。

### Informatica Big Data Edition

Informatica Big Data Edition と Hadoop により、組織は従来より多くの各種データを迅速かつ経済的に統合し、準備を整えることができます。Informatica では、数百ものビルド済みコネクタを使用して、顧客との取引や対話、メインフレーム データ、サーバログ、センサー データなど、あらゆる種類のデータを一括してリアルタイムで簡単に取り込むことができます。

Informatica は、アプリケーション、データベース、マシン、ソーシャル メディア、および業界標準の形式 (FIX、SWIFT、HL7、HIPAA、EDI、ASN.1 など) 用ビルド済みパーサーとのネイティブ接続を使用した Hadoop への高速データ取り込みをサポートします。Informatica は、バッチ処理、ログベースのレプリケーション、変更データ キャプチャ、リアルタイム ストリーミングなど、複数のデータ取り込み方法もサポートしています。

これにより、Informatica 開発者は、シンプルなビジュアル開発環境、ビルド済み変換、および再利用可能なビジネス ルールを使用して、データのプロファイリング、解析、統合、クレンジング、精製を最大 5 倍の速度で実行できます。データパイプラインが最適化され、基盤となる処理テクノロジーから切り離されるため、開発者は Hadoop の継続的な変更や進化に応じて再作業が必要となるリスクを負うことなく新たな革新的な技術を取り入れることができます。Informatica テクノロジーに関するスキルを持つ開発者の数は全世界で 100,000 人を超えているため、組織は既存のスキルを活用でき、プロジェクトの要員を迅速に確保できます。

ビジュアル開発環境: 現在でも、ほとんどの Hadoop 開発は、Informatica PowerCenter などの ETL ツールが存在していなかった 10 年前の ETL コード開発と同じような方法を手動で行なっています。グラフィカルなコードレス開発では、開発時間が最大で 500 % 短縮され、手動での Hadoop コーディングでは見逃されるデータ エラーを検出できることが証明されています。

- ユニバーサル データ アクセス: Hadoop を使用して多様なデータ ソースのデータを保存し処理する組織では、従来のデータ ソースから取得したすべての関連データを新しい種類のデータと統合した処理が、困難なチャレンジとなることが多くなります。Informatica Big Data Edition では、Hadoop に出し入れするデータの前処理および後処理を簡単かつ確実に実行できます。
- 高速なデータ取り込み: アプリケーション、データベース、マシン、ソーシャル メディア、および業界標準の形式用ビルド済みパーサーとのネイティブ接続を使用することで、ソース システムおよびターゲット システムとの間において、または直接 Hadoop やデータ ウェアハウスに対して、ビッグデータのアクセス、読み込み、変換、抽出を行うことができます。ログ ファイルやマシン データおよびセンサー データをリアルタイムで収集し、そのデータを一括して直接 Hadoop に流し込むことができます。
- 包括的なデータ変換: Hadoop は多様なデータの保存には優れていますが、すべての関連するデータ タイプ間において意味を見だし理解する能力は大きな課題となっています。Informatica Big Data Edition は、Hadoop でのデータ統合、データ解析、およびデータ品質に使用できる広範なビルド済み変換機能を備えています。あらゆる種類の構造化データおよび非構造化データを Hadoop 内でネイティブに統合、クレンジング、および利用可能な状態にできます。

- 包括的データ ガバナンス: Hadoop 上のデータをプロファイリングすることにより、データを理解し、データ品質の問題を検出し、データ パイプラインでコラボレーションします。保護が必要な機密データなど、Hadoop 上のデータドメインおよび関係の検出を自動化できます。ソース データから Hadoop を経由してターゲット アプリケーションに至るデータの移動過程の全体を示すエンドツーエンドのデータシステムによって、データ パイプラインの完全な透明性を確保できます。

### Cisco Data Virtualization

Cisco Data Virtualization は、異種データのフェデレーションと複雑なデータの抽象化および簡素化を行い、その結果をデータ サービスまたはリレーショナル ビュー（論理ビジネス ビュー）として、ビジネス インテリジェンスおよび分析ツールや他の情報ダッシュボードなどの利用側アプリケーションに送信します。Cisco Data Virtualization は、高度なクエリー最適化技術により、きわめて高い性能を発揮します。

選択したウェアハウス データを Hadoop にオフロードした後で、Cisco Data Virtualization を使用して、両方のデータソースのフェデレーションを行い、データの単一ビューを作成します。ウェアハウスだけでなく Hadoop ビッグデータ ストアにもアクセスでき、より多くのデータを利用できるため、分析およびビジネス インテリジェンス レポートがさらに充実したものとなります。

Cisco Data Virtualization の主な機能は次のとおりです。

- データ アクセス: 多様なソースのデータに接続し、開示します。
- データ フェデレーション: データ ウェアハウスや Hadoop ビッグデータ ストアなどに対するクエリーを処理し、最適化します。最適化されたアルゴリズムにより、さまざまな種類のデータ ソースに対するクエリーが高速化されます。
- データ配信: データ サービスを介して、分析およびビジネス インテリジェンス ツールなどの多様な利用側アプリケーションにデータを配信します。

### サービス

シスコと Informatica は、このソリューションの一部として、お客様のプロジェクトを成功させることを目的としたオンサイト プロフェッショナル サービスを提供しています。両社は、作業を適切かつ期限どおりに完了するための文書化された実証済みのタスクが定義された導入方法に準拠しています。それらのサービス概要を以下に示します。（日本でのサービス提供状況については別途ご相談ください）

- 評価: お客様が製品を導入する前に、当社のチームがお客様のデータ ウェアハウスや IT 環境を調査します。次に、お客様のデータおよび変換ワークロードの全体像を明らかにします。そのうえで、オフロードの対象と最善のオフロード方法を示したアクション プランを策定します。
- 仮想化: お客様のデータ ウェアハウスおよびビッグデータ クラスタの上にデータ抽象化層を配置します。この層によって、複雑さが物理データ層に隠されます。データおよび ETL は、そのデータを使用するビジネス インテリジェンスおよび分析アプリケーションを中断させることなく移行できます。
- 移行: 移行するデータおよび ELT と ELT ワークロードをマッピングします。データ移動アプローチを決定し、特定されたデータおよびワークロードをウェアハウスからビッグデータ ターゲットに移行します。
- 運用: 当社は、お客様が実稼働への移行後にソリューションを実行できるように運用計画の作成をサポートします。管理者が必要に応じて参照できるように標準運用手順を文書化します。

さらに、物理インフラストラクチャの計画および設計サービスを実施し、予測可能で拡張可能な環境の導入を可能にするソフトウェア構成も提供します。当社は以下の点についてお客様をサポートします。

- データウェアハウス、Informatica サーバ、および Hadoop ノード間の複数のデータトラフィック ストリームを処理するネットワークのプロビジョニング
- Cisco Data Virtualization 層を通してデータウェアハウスと Hadoop のデータのフェデレーションを行うビジネスインテリジェンスおよび分析アプリケーションによる利用を処理するネットワークインフラストラクチャの導入
- 初期データ読み込みの実行中に発生する大量の I/O トラフィックや Hadoop 処理中のノード間トラフィックの緩和

## ソリューションのメリット

### ビジネス上のメリット

Informatica とシスコのデータウェアハウス最適化ソリューションは、高い投資回収率により、純利益に大きな影響をもたらします。次のようなビジネス上のメリットがあります。

- コスト管理: データおよび計算処理をオフロードすることにより、テクノロジーへの投資効果を高めることができます。
- 分析の強化: 現在および最近の履歴だけでなく、通常はアーカイブされアクセスできない長期履歴データにもアクセスできます。
- 競争優位性: 企業の膨大なデータ資産に有効な分析を加えることにより、生産性を高め、ビジネス状況の変化に対応できます。
- リスクの軽減: ビッグデータおよび論理データウェアハウスの導入には、実証済みのソフトウェア、ネットワーキング、およびコンピューティングインフラストラクチャを使用します。

### 技術的なメリット

このソリューションを構成するシスコと Informatica のクラス最高レベルの製品およびサービスに特有の技術的なメリットも数多くあります。主な技術的なメリットは次のとおりです。

- 直線的な拡張性: このプラットフォームは、単純にデータの増加に応じて拡張できるように設計されています。スイッチングコンポーネントの追加や、システム接続を何らかの形で再設計する必要性にせまられることなく、Hadoop ノードを最大 160 個まで増設できます。さらに、サーバを最大 10,000 台まで追加でき、すべてのサーバを Cisco UCS Central ソフトウェアによって 1 つの管理ドメイン内で管理できます。
- データ移行の簡素化: Informatica Big Data Edition によって、データ変換およびデータ品質ジョブの処理が簡素化されます。使いやすいビジュアル開発環境に加え、数百ものビルド済みコネクタおよび変換機能が用意されているため、開発者の生産性が手動コーディングを行う場合と比較して最大 5 倍に向上します。
- クエリーの最適化: Cisco Data Virtualization には、異種データのフェデレーションに使用できる包括的な手法やアルゴリズムが用意されています。この知的財産は、数百年もの研究開発の成果です。
- 高性能: Cisco UCS は、ETL オフロードジョブおよび通常の Hadoop ジョブのさまざまなワークロードを処理するために必要な適応性の高性能を実現します。このソリューションの低遅延で無損失の 10 Gbps ユニファイドファブリックは、十分な冗長性を備えており、アクティブ-アクティブ構成により、他社のソリューションよりも高性能を発揮します。クライアントからの I/O 負荷のピーク時にも速やかな応答を維持したまま対応できます。

- 導入および管理の容易さ: Cisco UCS プラットフォームでは、ビッグデータ アプリケーションとエンタープライズ アプリケーションを同じ管理および接続ドメイン内でホストできるため、データセンターの管理がさらに簡素化します。Informatica Big Data Edition は Hadoop フレームワークの能力を最大限に活用するため、構築したすべてのデータ パイプラインが実稼働に対応でき、パフォーマンス、拡張性、可用性、保守性の向上をもたらします。

## まとめ

Cisco UCS と Informatica は、Hadoop ベンダーと連携し、ビッグデータの能力を取り込むことで、応答時間の向上と総所有コスト(TCO)を削減し、既存のエンタープライズ データ ウェアハウス環境に根本的な変化と、データ量の増加や新しいデータタイプの登場という課題への対応を可能にします。このアプローチでは、Cisco UCS CPA for Big Data と Cisco Data Virtualization に Informatica Big Data Edition を組み合わせることで、これらの製品がもたらす生産性向上と導入の簡素化によって運用コストの削減にもつながり、迅速な導入と、最高性能または大容量に向けたカスタマイズが可能なソリューションが実現します。

## 詳細情報

Cisco UCS ビッグデータ ソリューションの詳細については、[http://www.cisco.com/web/JP/solution/datacenter/big\\_data/index.html](http://www.cisco.com/web/JP/solution/datacenter/big_data/index.html) を参照してください。

Cisco UCS Integrated Infrastructure for Big Data の詳細については、<http://blogs.cisco.com/datacenter/cpav3> [英語] を参照してください。

Cisco Data Virtualization の詳細については、<http://www.cisco.com/web/JP/services/enterprise-it-services/data-virtualization/index.html> を参照してください。

ビッグデータ用 Cisco Validated Designs の詳細については、<http://www.cisco.com/c/en/us/solutions/enterprise/data-center-designs-cloud-computing/bigdata.html> [英語] を参照してください。

Informatica Big Data Edition の詳細については、<http://www.informatica.com/us/products/big-data/informatica-big-data-edition> [英語] を参照してください。

©2015 Cisco Systems, Inc. All rights reserved.

Cisco、Cisco Systems、およびCisco Systemsロゴは、Cisco Systems, Inc.またはその関連会社の米国およびその他の一定の国における登録商標または商標です。本書類またはウェブサイトに掲載されているその他の商標はそれぞれの権利者の財産です。

「パートナー」または「partner」という用語の使用は Cisco と他社との間のパートナーシップ関係を意味するものではありません。(1502R)

この資料の記載内容は2015年2月現在のものです。

この資料に記載された仕様は予告なく変更する場合があります。



シスコシステムズ合同会社

〒107 - 6227 東京都港区赤坂9-7-1 ミッドタウン・タワー

<http://www.cisco.com/jp>

お問い合わせ先