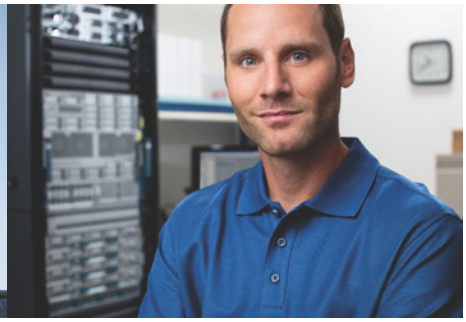
データの準備を効率化

Cisco UCS 統合インフラストラクチャ上の Cisco Data Preparation

ソリューション概要
2015 年 9 月





概要

統合インフラストラクチャ

- 企業のビッグデータ環境を支える、業界トップクラスのプラットフォームを構築できます。

セルフサービス型のビッグデータ準備

- ガバナンス、コラボレーション、および大規模環境に配慮して構築されたプラットフォーム上で、運用分析、予測モデリング、パッケージ化された分析ツールなどを使用して、データ統合およびデータ品質の最適化を自動実行できます。
- セルフサービス型のアプリケーションを導入することで、技術者ではないビジネス アナリストでも、分析に役立つデータを容易に収集、調査、クリーニング、結合、および強化できるようになります。

データの即時検証

- データ準備要求は迅速に処理され、ユーザは処理状況をリアルタイムで検証できます。

データ品質の向上

- フルテキスト検索、インタラクティブなテキスト/数値フィルタ、ヒストグラム、データ品質ヒート マップなどを使用することで、データのパターン、エラー、重複、および脱落を特定できます。

自動化されたデータ シェイピング プラットフォーム

- ピボットと分割を使用して、データの取り込み、クリーニング、結合を行うことで、分析用データ セットを迅速に作成できます。

ビックデータ ワークロードに対応した拡張性

- 複雑なスイッチング インフラストラクチャ層を追加することなく、単一ラックおよび複数ラック展開を拡張できます。

適切なビッグデータ ソリューションを活用すれば、分析のためのデータ準備に多くの時間を費やすことなく、必要とする回答を迅速に入手できます。

組織が多くのセルフサービス型ビジネス インテリジェンスや分析アプリケーションを使用するようになり、分析用のデータの準備に多くの時間が費やされる傾向にあります。データを準備するには、さまざまなソースからのデータ セットの集約、重複データや空白フィールドの特定、スペルチェック、カラムの分割や再編、コンテキスト用のデータ追加、といった作業がともないます。通常データの準備作業には多くの時間を要しますが、強力なビジネス インテリジェンス ツールを使用すれば、質問をして判断を下すというプロセスなしに、ごく短時間でこの作業を完了できます。Cisco Unified Computing System™（Cisco UCS®）上で実行される Cisco® Data Preparation は、データ量が増え続ける中で、アナリストの生産性を向上し、データ カオスのリスクを軽減し、データで得られた知見からより大きな価値を引き出すことを可能にします。

Cisco Data Preparation

Cisco Data Preparation は、技術者ではないビジネス アナリストが、分析に役立つ raw データを簡単に収集、クリーニング、結合、強化することを可能にするエンドユーザ向けアプリケーションです（図 1）。このソリューションは、アナリストに次のような機能を提供します。

- **追加**：Hadoop 分散ファイル システム（HDFS）ファイル、リレーショナル データベース、スプレッドシート、フラット ファイルなどのデータを、存在している場所に関係なく取り込めます。
- **調査**：インタラクティブな検索により、データ品質に関する問題を特定できます。この作業には、フルテキスト検索、インタラクティブなテキスト/数値フィルタとヒストグラム、さらにはパターン、エラー、重複、およびスパース データや欠損データなどを強調表示するビジュアルなデータ品質ヒート マップなどが使用されます。
- **クリーニング**：コードやスクリプトを作成することなく、データセットの一部または全体にわたり、高度なアルゴリズムを適用できます。このソリューションにより、データの不

© 2015 Cisco Systems, Inc. All rights reserved. This document is Cisco Public Information.

整合、ギャップ、重複などが強調表示されるため、アナリストはブランクを埋めたり、重複データを削除または名前変更したり、一貫性のない大文字使用を修正するなど、データ品質を向上するための各種タスクを実行できます。

- シェイピング：データのピボット化/ピボット化の解除、カラムの分割、集約などの作業をワンクリックで実行して、分析処理により適したデータ セットを迅速に作成できます。
- 強化：オリジナルのデータ セット内に、分析に必要なコンテキストが含まれていない場合に、必要なデータを追加できます。たとえば、標準的な 5 桁の米国郵便番号に 4 桁の拡張コードを追加するなどです。
- 結合：複数のデータセットを迅速に取り込んで、データのマージに最適なフィールドを特定できます。Cisco Data Preparation は、複数のデータ セットにわたって共通属性を自動的に検出し、ベストマッチ オプションを提示することが可能です。そのためアナリストは、分析に最適な組み合わせを簡単に選択できます。次にこれらのデータ セットから、単一の回答セットが作成されます。重複するリファレンスは、スクリプト、SQL、またはピボット テーブルやマクロのような複雑なスプレッドシート機能を使用する必要なしに、重複除去された信頼性の高いエンティティにマージされます。
- 公開：回答セットは、公開されると、QlikView、Tableau、Microsoft Excel、およびその他の ODBC 準拠の分析ツールやアプリケーションから、オープン データベース コネクティビティ（ODBC）ライブ クエリを介して、直接使用できるようになります。

拡張性に配慮した設計

このソリューションは、インタラクティブなセルフサービス型のデータ準備を大規模に実行できるように設計された 4 層アーキテクチャを使用します（図 2）。Apache Spark をベースとする Cisco Data Preparation と Cisco UCS の組み合わせによりセルフサービス型ソリューションが可能になり、データ量の増加に応じて簡単に拡張可能なインフラストラクチャ上で、自動化されたデータ統合機能が提供されます。

- ユーザインターフェイス層：アプリケーションのフロントエンドとして、HTML5 および Web ソケット テクノロジーを使用して設計された、ビジュアルでダイナミックなマルチユーザ インターフェイスが用意されており、一般的なコンシューマ アプリケーションと同等のインタラクティブで直感的な操作性を実現します。
- Web サービス：Lightweight Java 層により、ユーザ インターフェイスから渡されたアクションが適切に変換されて、基盤となるプラットフォームに渡されます。この層により、テナント、ユーザ、プロジェ

図 1：Cisco Data Preparation は、分析の向上に必要なデータの収集、処理、および強化を支援

© 2015 Cisco Systems, Inc. All rights reserved. This document is Cisco Public Information.

クト、およびセルレベルの修正用ルールに関する重要な処理が実行されて、ガバナンス機能のための包括的なバックボーンが提供されます。

- **並列インメモリ パイプライン型エンジン:**このエンジンは、独自の機械学習、潜在的セマンティック インデックス作成、統計的パターン認識、およびテキスト分析技法を使用します。データの処理には、ベクトル クエリ プロセッサによってパフォーマンスを加速し、大規模かつ多様な構造化データと非構造化データをリアルタイムで処理できる、モデルフリー環境が使用されます。
- **ファイル管理とストレージ:**すべてのデータ セットは、HDFS 上で動作するライブラリに保存され、このライブラリを介して処理されます。

さまざまな使用例に最適

Cisco Data Preparation は操作が容易で、高速なインメモリ処理アーキテクチャを採用しているため、さまざまなタイプのビジネス課題に対応可能です。このソリューションは、特定の業種や職務だけを対象とするものではなく、あらゆる組織に次のようなメリットをもたらします。

- データから得られる知見を収益に変換
- 顧客レコードをマージおよび調整
- オンラインの Web 小売情報にアクセス（ユーザ トランザクションや製品移動時間など）
- サプライチェーン内の停滞ポイントを特定
- 在庫を管理
- セールス組織のために製品価値バンドルを確立
- データ品質を評価
- コンプライアンスを評価
- アプリケーション間でデータを移行

ビッグデータ向け Cisco UCS 統合インフラストラクチャ上に構築

Cisco Data Preparation は、ビッグデータ向け Cisco UCS 統合インフラストラクチャを基盤としており、スケールアウト アプリケーションのニーズに適合するように設計されたスケーラブルなアーキテクチャです。包括的かつ発注が容易なパッケージとして提供されるこのインフラストラクチャは、コンピューティング、ストレージ、接続、および統合管理機能などを備えています。

Cisco UCS 6200 シリーズ ファブリック インターコネクト
ファブリック インターコネクトは、システム全体の接続と管理を一元化します。通常、ビッグデータ アプリケーションに対応しているクラスタには大量のノードが存在します。冗長ペアで展開される Cisco UCS ファブリック インターコネクトは、高帯域幅で低遅延の接続、アクティブ-アクティブの冗長性、高パフォーマンス、および卓越した拡張性で、これらのノードを支えることが可能です。

このシステムは、接続されたすべてのインフラストラクチャ コンポーネントの管理を統合および一元化します。Cisco UCS Manager は、サービス プロファイルによる迅速で一貫性のあるサーバ構成をサポートしており、クラスタ全体のファームウェア アップデートの操作を一元化するなどして、継続的なシステム メンテナンス作業を自動化します。さらに、クラスタ全体の状態に関するアラームや通知などのオプションを備えた、高度なモニタリング機能を提供します。

Cisco UCS C シリーズ ラック サーバ
Cisco UCS C240 および C220 M4 ラック サーバは、コンパクトな設計でありながら、コンピューティング、I/O、およびストレージ容量に関する広範なニーズに対応可能です。これらのラック サーバには Intel® Xeon® プロセッサ E5-2600 v3 ファミリが搭載され、12 Gbps シリアル接続 SCSI（SAS）のスループットにより、旧世代のサーバよりはるかに優れたパフォーマンスと効率性が提供されます。

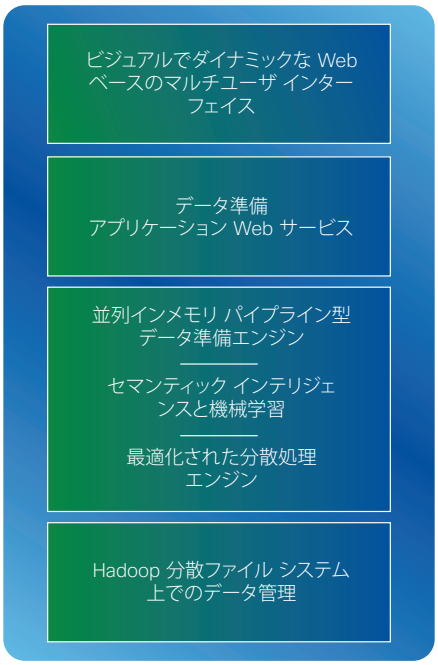


図 2: Cisco Data Preparation のアーキテクチャ

© 2015 Cisco Systems, Inc. All rights reserved. This document is Cisco Public Information.

表 1：Cisco Data Preparation のリファレンス アーキテクチャ

コンポーネント	説明
接続性	・ Cisco UCS 6296UP 96 ポート ファブリック インターコネクト X 2 ・ スイッチング インフラストラクチャの追加なしに、最大 80 サーバまで拡張可能
Spark パイプライン サーバ	・ 以下を搭載した Cisco UCS C240 M4 ラック サーバ X 8： ・ Intel Xeon プロセッサ E5-2680 v3 ファミリ CPU X 2 ・ 512 GB のメモリ ・ 120 GB 内蔵 SSD ブート ドライブ X 2 ・ 1.6 TB SSD ドライブ（データ ストレージ用）X 12
アプリケーション サーバ	・ 以下を搭載した小型フォーム ファクタ（SFF）Cisco UCS C220 M4 ラック サーバ X 4： ・ Intel Xeon プロセッサ E5-2680 v3 ファミリ CPU X 2 ・ 256 GB のメモリ ・ 1.2 TB SAS SFF ハードディスク ドライブ（HDD）X 2
メタデータ管理サーバ	・ 以下を搭載した Cisco UCS C220 M4 ラック サーバ X 3： ・ Intel Xeon プロセッサ E5-2680 v3 ファミリ CPU X 2 ・ 256 GB のメモリ ・ 1.2 TB SAS SFF HDD X 2 ・ 1.6 TB SSD X 2
データ ライブラリ サーバ	・ ビッグデータ向け Cisco UCS 統合インフラストラクチャ ・ Cloudera 5.4 以降

このサーバは CPU をデュアル構成で使用し、最大 768 GB のメイン メモリ（ビッグデータ アプリケーションの場合は、通常 128 または 256 GB）と、さまざまなディスクおよびソリッドステートディスク（SSD）オプションをサポートしています。Cisco UCS C220 M4 サーバは、業界トップクラスのコンピューティング密度を提供し、Cisco UCS C240 M4 サーバは、バランスの取れたコンピューティング リソースとストレージ リソースを提供します。

表 1 は、データ準備用リファレンス アーキテクチャの推奨コンポーネントを示したものです。

まとめ

緊急性の高いニーズに対してタイムリーに回答を得るために、データ準備の迅速化が必要とされている場合には、Cisco Data Preparation の利用をご検討ください。この革新的なソリューションは、データ準備の柔軟性と応答性を向上して、価値実現を迅速化し、データ アナリストの生産性を高めることが可能です。インテリジェントなセルフサービス型のデータ準備機能は、関連データを特定して、分析に必要な準備作業を行ううえで、大きな効果を発揮します。

詳細情報

Cisco Data Preparation 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。

Cisco Smart Play プログラムの詳細については、http://www.cisco.com/jp/go/smartplay/ を参照してください。

ビッグデータ向け Cisco UCS ソリューションの詳細については、http://www.cisco.com/jp/go/bigdata/ を参照してください。

ビッグデータ向けシスコ共通プラットフォーム アーキテクチャ（CPA）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。

©2016 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。(1602R)
この資料の記載内容は2016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。

お問い合せ先

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

LE-53502-00JA 16.02