

Pentahoデータ統合

ビッグデータをより強力に：
様々なデータソースに対する迅速なアクセス、準備、ブレンディング

データシート

Pentahoなら、企業に押し寄せる膨大で多様、高頻度での処理が求められるデータも簡単に管理することができます。

Pentahoデータ統合 (PDI) は、簡単に操作可能なビジュアルツールによって、分析に必要なデータを短時間でエンドユーザーに提供できます。SQLやJava、Pythonなどのコーディングなしで、企業は自社のデータやファイル、リレーショナルデータベース、Hadoopなどのデータソースのデータを、それがクラウド上にあってもオンプレミスでも、価値ある情報に変えることができます。

ビッグデータを有益な分析データに

Pentahoのアダプティブビッグデータレイヤにより、主要なビッグデータストアに柔軟に接続できます。接続先システムに変更があっても影響を受けることはありません。一度データにアクセスすると、その後どこからでも処理、結合、および利用ができます。分析をさらに便利にするPentahoのアダプティブビッグデータレイヤには、Cloudera、Hortonworks、MapR、AmazonなどのHadoopディストリビューション用プラグインだけでなく、NoSQLデータベースであるMongoDBやCassandra用のプラグインが含まれています。

ビッグデータと既存のエンタープライズデータの統合およびブレンディング

あらゆる種類のデータ形式に対応する幅広い接続性と高パフォーマンスなHadoop内処理により、Pentahoは既存のデータベースと新しいデータソースの統合プロセスを簡略化、高速化します。Pentahoデータ統合のグラフィカルデザイナーの特長は以下のとおりです。

- 直感的なドラッグ&ドロップ操作で、分析データパイプラインの作成を簡素化 (図1)
- リレーショナルソース、ビッグデータストア、エンタープライズアプリケーションなどへアクセスし、データを抽出・準備・ブレンドするためのプレビルドコンポーネントを提供するライブラリ
- データ準備プロセスの任意のポイントで、チャート、可視化、レポートを含むアナリティクスを実行でき、データのスポットチェックが可能
- 通知やアラートなど、変換済みデータの調整や統合をサポートする強力なオーケストレーション機能
- ワークフロー調整のためのエンタープライズスケジューラーと、ジョブ実行のテストやチューニング用のデバッガーを統合

ビッグデータ統合および大規模データ処理

Pentahoは、ビッグデータソースの統合にかかる時間を短縮し、複雑さを軽減します。Pentahoには以下の特長があります。

- 手作業のコーディングに比べ何倍も速い生産性を達成するコーディング不要のHadoopデータ変換設計、およびクラスター内での実行による高パフォーマンスを実現
- テンプレートによるアプローチで、メタデータ・インジェクション機能を使用して迅速にデータのオンボーディング作業を実施
- データ量や複雑さに応じて、SparkやPentahoなどの実行エンジンをシームレスに切り替え可能 (図2)
- 予測分析を可能とするR、Python、ScalaおよびWekaの高度な分析モデルを統合すると同時に、データの準備時間を削減



図1. Pentahoデータ統合でのドラッグ&ドロップによるデータ変換

幅広い接続性とデータ配信

Pentahoデータ統合は、構造化、非構造化、および半構造化のデータソースすべてを含む多種多様なデータへの接続に対応しています。以下はその一部の例です。

- RDBMS: Oracle、DB2、MySQL、Microsoft SQL Server
- Spark、Hadoop: Cloudera、Hortonworks、Amazon EMR、MapR、Microsoft Azure HDInsights
- NoSQLデータベース: MongoDB、Cassandra、HBase
- 分析用データベース: Vertica、Greenplum、Teradata、SAP HANA、Google Big Query、Microsoft Azure SQL Data Warehouse(DW)
- ビジネスアプリケーション: Salesforce、Google Analytics
- ファイル: XML、JSON、Excel、CSV、txtなど

データの抽出、ロード、配信プロセスのパフォーマンスを向上させるため、Pentahoでは次の機能を用意しています。

- 一般的なデータソースへのネイティブ接続およびバルクロード
- データを一時退避することなく変換データを仮想化するデータサービスにより、レポートとアプリケーションに必要なデータセットの即時利用が可能
- より速く分析結果を得るためにメタデータモデルを自動作成して発行
- ストリームデータをリアルタイムに処理可能

borderfree
from pitney bowes

「企業をまたがるデータの移動はアートと言えます。
Pentahoはアートをより良いビジネス価値に変えるのです。」

– Warren Chang氏、
Borderfree社 エンジニアリング担当副社長

Hitachi Vantara

本データシートは、Hitachi Vantara Corporationが提供するデータシートの抄訳です。

HITACHIは、株式会社日立製作所の商標または登録商標です。Pentahoは、Hitachi Vantara Corporationの日本及びその他の国における商標または登録商標です。Amazonは、米国その他の諸国における、Amazon.com, Inc.またはその関連会社の商標です。OracleとJavaは、Oracle Corporation及びその子会社、関連会社の米国及びその他の国における登録商標です。IBM、DB2は、世界の多くの国で登録されたInternational Business Machines Corporationの商標です。Microsoft、SQL Server、Excel、およびAzureは、米国Microsoft Corporationの米国およびその他の国における登録商標または商標です。SAP、SAP HANA、および本文書に記載されたその他のSAP製品、サービス、ならびにそれぞれのロゴは、ドイツおよびその他の国々におけるSAP SEの商標または登録商標です。その他記載の会社名、製品名などは、それぞれの会社の商標もしくは登録商標です。

P-016-C BTD November 2018

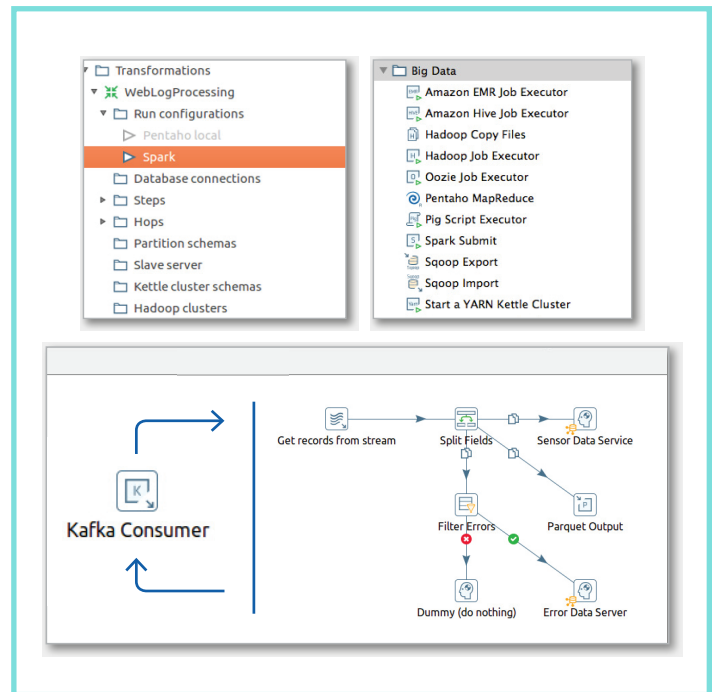


図2. Sparkの実行適用と、PDIでビジュアルに設計できるHadoop MapReduceのジョブ

データプロファイリングとデータ品質

Pentahoは、行数カウント、数学関数、null値の識別などの基本的なデータプロファイリング機能と、文字列操作、マッピング関数、フィルタリングや並べ替えなどのデータ品質に関する機能を提供します。名前および住所の検証機能に関しては、Human Inference社やMelissa Data社などの代表的なデータ品質ベンダーと連携しています。Pentahoのデータプロファイリングおよびデータ品質機能は以下のような用途に便利です。

- ビジネスルールや標準に準拠していないデータの特定
- 統一されていないデータや重複データの非重複化とクレンジング
- 名前、住所、メール、および電話番号データの検証、標準化、訂正

強力な管理機能

Pentahoデータ統合により、データ統合プロジェクトの運用管理が容易になります。次の機能が用意されています。

- データアナリスト、開発者、データ管理者間の連携に役立つ共有リポジトリ
- 旧バージョンへのロールバックを容易に行うためのコンテンツ管理、バージョン管理、ロック機能
- ユーザーやロールベースのセキュリティ権限の管理、およびデータの作成、読み込み、ジョブ実行、変換に対する操作許可の設定

