

Pentaho Data Integration

The Power To Access, Prepare and Blend Multiple Data Sources Faster

DATASHEET

With Pentaho from Hitachi Vantara, managing the enormous volumes and increased variety and velocity of data entering organizations is simplified. Pentaho Data Integration (PDI) delivers analytics-ready data to end users faster with visual tools that reduce time and complexity. Without writing SQL or coding in Java or Python, organizations immediately gain real value from their data, from sources like files, relational databases, Hadoop and more, which are in the cloud or on premises.

Turn Big Data Into Actionable Analytics

Pentaho's adaptive big data layer allows you to plug into popular big data stores with flexibility and insulation from change. Data can be accessed once, then processed, combined and consumed anywhere. For consumable and actionable analytics, Pentaho's adaptive big data layer includes plug-ins for Hadoop distributions from Cloudera, Hortonworks, MapR, Amazon Web Services Elastic MapReduce (AWS EMR), Google Cloud and Microsoft Azure HDInsight, object stores such as Hitachi Content Platform, as well as popular NoSQL databases like MongoDB and Cassandra.

Integrate and Blend Big Data With Existing Enterprise Data

With broad connectivity to any data type and high-performance Spark and MapReduce execution, Pentaho simplifies and speeds the process of integrating existing databases with new sources of data. Pentaho Data Integration's graphical designer includes:

- Intuitive, drag-and-drop designer to simplify the creation of analytics data pipelines (see Figure 1).
- Rich library of prebuilt components to access, prepare and blend data from relational sources, big data stores on premises or in the cloud, enterprise applications and more.
- Ability to spot check data in flight with immediate access to analytics, including charts, visualizations and reporting, from any data prep step.
- Powerful orchestration capabilities to coordinate and combine transformations, including notifications and alerts.
- Integrated enterprise scheduler for coordinating workflows and debugger for testing and tuning job execution.

Big Data Processing Performance and Productivity

Pentaho speeds performance time and reduces the complexity of integrating big data sources. Pentaho provides:

- Code-free data transformation design that empowers 15 times faster productivity versus hand-coding and executes in-cluster for high performance.
- Template-based approach to rapidly onboard data sources into Hadoop via metadata injection feature set.
- Ability to seamlessly switch between execution engines, such as Spark and Pentaho's native engine, to fit data volume and transformation complexity (see Figure 2).
- Integration with advanced analytics models from R, Python, Scala and Weka to operationalize predictive intelligence while reducing data prep time.



Figure 1. Drag-and-Drop Data Transformation in Pentaho Data Integration

Broad Connectivity and Data Delivery

Pentaho Data Integration offers broad connectivity to a variety of diverse data, including all popular structured, unstructured and semi-structured data sources. Some examples include:

- Relational database management system (RDBMS): Oracle, IBM® DB2®, MySQL, Microsoft SQL Server.
- Spark and Hadoop: Cloudera, Hortonworks, Amazon EMR, MapR, Microsoft Azure HDInsights.
- NoSQL databases and object stores: MongoDB, Cassandra, HBase, Hitachi Content Platform.
- Analytic databases: Redshift, Snowflake, Vertica, Greenplum, Teradata, SAP HANA, Amazon Redshift, Google Big Query, Microsoft Azure SQL Data Warehouse (DW).
- Business applications: SAP, Salesforce, Google Analytics.
- Files: XML, JSON, Microsoft Excel, CSV, txt, Avro, Parquet, ORC, unstructured files with metadata, including audio, video and visual files.

To increase the performance of data extraction, loading and delivery processes, Pentaho offers the following capabilities:

- Native connectivity and bulk-loading to most common data sources.
- Data services to virtualize transformations without staging, making data sets immediately available to reports and applications.
- Automatic creation and publishing of metadata models to drive faster analytic results.
- Process streaming data in real time.

Data Profiling and Data Quality

Pentaho provides data profiling capabilities, such as row counts, mathematical functions and identification of null values, as well as data quality operators, such as string manipulators, mapping functions, filtering and sorting. For name and address verification capabilities, Pentaho integrates with leading data quality vendors, such as Human Inference and Melissa Data. Pentaho data profiling and data quality capabilities help:

- Identify data that fails to comply with business rules and standards.
- Deduplicate and cleanse inconsistent and redundant data.
- Validate, standardize and correct name, address, email and telephone data.

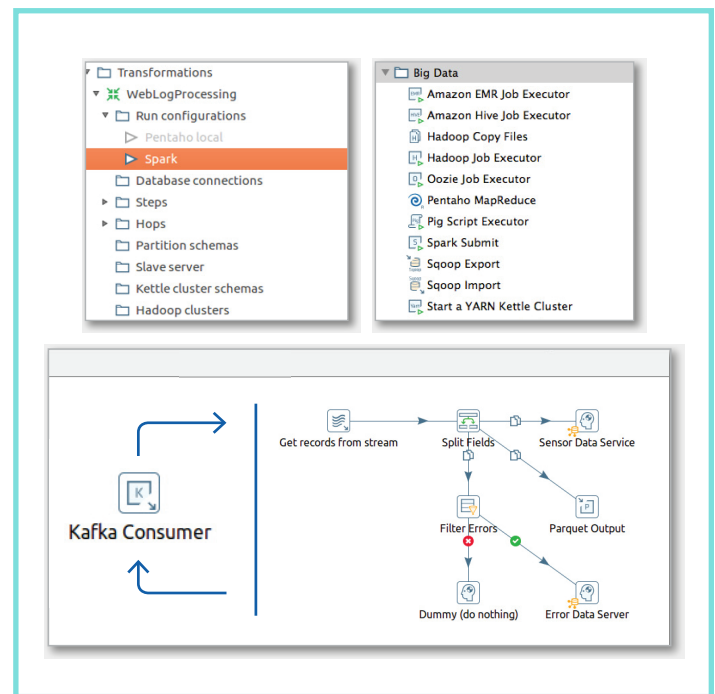


Figure 2. Adaptive Execution With Spark and Visually Designed Hadoop MapReduce Jobs in PDI

Powerful Administration and Management

Pentaho Data Integration provides out-of-the box capabilities for managing operations for data integration projects. These capabilities include:

- Shared repository for collaboration among data analysts, developers and data stewards.
- Content management, versioning and locking to easily version jobs for roll-back to prior versions.
- Control over security privileges for users and roles and integration with third-party security systems; ability to set permissions for creating, reading or executing jobs and transformations.

borderfree
from pitney bowes

“Moving data across a business is an art. Pentaho transforms art into better business value.”

– Warren Chang, VP of Engineering, Borderfree

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
HitachiVantara.com | community.HitachiVantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
HitachiVantara.com/contact

HITACHI is a registered trademark of Hitachi, Ltd. Pentaho is a trademark or registered trademark of Hitachi Vantara Corporation. Microsoft, HDInsights, Azure and SQL Server are trademarks or registered trademarks of Microsoft Corporation. IBM and DB2 are trademarks or registered trademarks of International Business Machines Corporation. All other trademarks, service marks, and company names are properties of their respective owners.

P-016-D BTD July 2019

