HITACHI
Inspire the Next

Pentaho Data Integration

Big Data and More: 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.

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. The Pentaho adaptive big data layer includes plug-ins for Hadoop distributions from Cloudera, Hortonworks, MapR, Amazon EMR, and Microsoft HDInsights, as well as popular NoSQL databases like MongoDB and Cassandra — for consumable and actionable analytics.

Integrate and Blend Big Data With Existing Enterprise Data

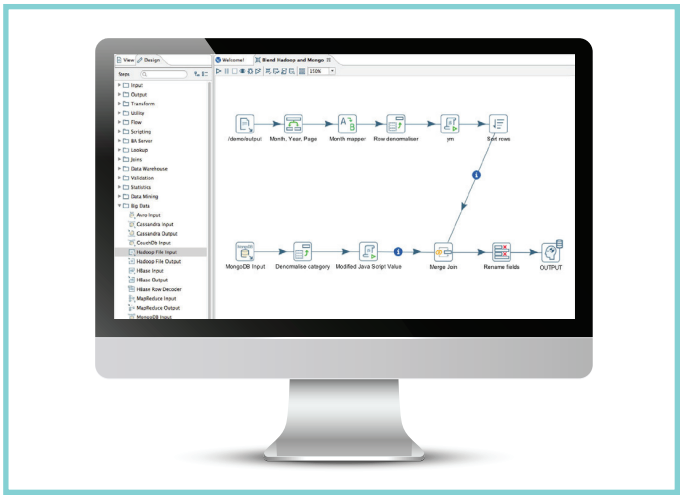
With broad connectivity to any data type and high performance Spark and MapReduce execution, Pentaho simplifies and speeds the process of integrating existing databases with new sources of data. Pentaho Data Integration's graphical designer includes:

- Intuitive, drag-and-drop designer to simplify the creation of analytic data pipelines.
- Rich library of pre-built components to access, prepare, and blend data from relational sources, big data stores, 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 time and reduces the complexity of integrating big data sources. Pentaho provides:

- Code-free Hadoop data transformation design that empowers 15x faster productivity vs. hand-coding and executes in-cluster for high performance.
- Template-based approach to rapidly onboard data sources into Hadoop via metadata injection feature set.
- Seamlessly switch between execution engines such as Spark and Pentaho's native engine to fit data volume and transformation complexity.



Drag and drop data transformation in Pentaho Data Integration

- Integration with advanced analytic models from R and Weka to operationalize predictive intelligence while reducing data prep time.

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:

- RDBMS: Oracle, DB2, MySQL, Microsoft SQL Server.
- Spark and Hadoop: Cloudera, Hortonworks, Amazon EMR, MapR, MS Azure HDInsights.
- NoSQL databases: MongoDB, Cassandra, HBase.
- Analytic databases: Vertica, Greenplum, Teradata, SAP HANA, Redshift.
- Business applications: Salesforce, Google Analytics.
- Files: XML, JSON, Excel, CSV, txt, and more.

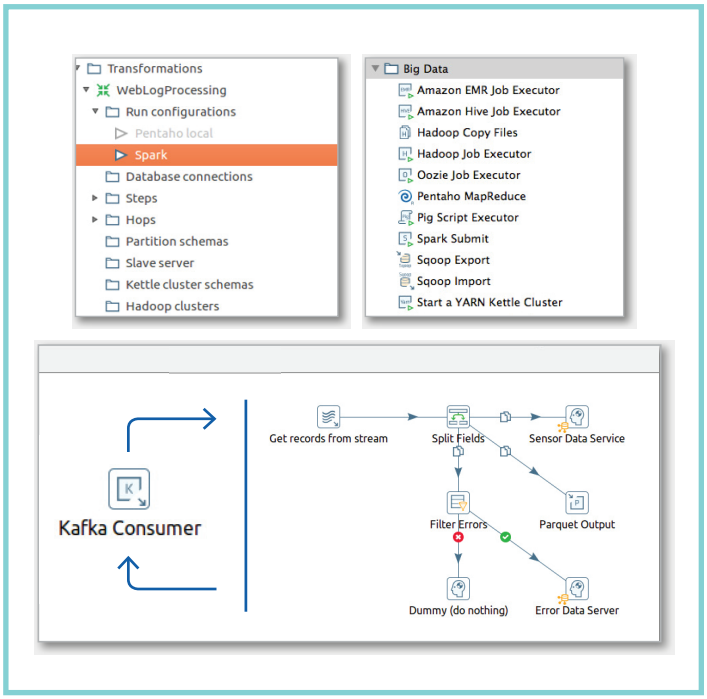
To increase the performance of data extraction, loading and delivery processes, Pentaho offers the following capabilities:

- 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 basic 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.
- De-duplicate and cleanse inconsistent and redundant data.
- Validate, standardize and correct name, address, email and telephone data.



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
2845 Lafayette Street
Santa Clara, CA 95050-2639 USA
www.HitachiVantara.com | community.HitachiVantara.com

Regional Contact Information
Americas: +1 866 374 5822 or info@hitachivantara.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hitachivantara.com
Asia Pacific: +852 3189 7900 or info.marketing.apac@hitachivantara.com

P-016-A DG February 2018