Pentaho | HITACHI Inspire the Next

DATASHEET

Pentaho Data Integration

Ingest, Blend, Cleanse and Prepare Diverse Data From Any Source, in Any Environment — With No Code.

With Pentaho Data Integration, managing the enormous volumes, variety, and velocity of data is simplified.

By allowing data preparation from any source and automating your data pipeline, Pentaho Data Integration allows you to curate data better for your business user. This software delivers business analytics 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 various data sources in the cloud or on premises, including files, relational databases, big data sets and more.

Turn Data Into Actionable Insights

More than just ETL (Extract, Transform, Load), Pentaho Data Integration is a codeless data orchestration tool that blends diverse data sets into a single source of truth as a basis for analysis and reporting. Effortlessly managed in a drag-and-drop graphical interface, so you can easily track where it's coming from, where it's going and how it's transforming.

Data Processing Performance and Productivity

Pentaho Data Integration speeds performance time, reduces the complexity of integrating big data sources, and 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.

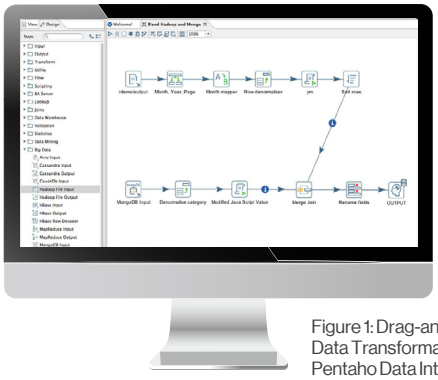

Figure 1: Drag-and-Drop Data Transformation in Pentaho Data Integration

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

Warren Chang, VP of Engineering, Borderfree

Scalability, Simplicity, and Self-Service

With broad connectivity to any data type and high-performance Spark and MapReduce execution, Pentaho technology simplifies and speeds the process of integrating existing databases with new sources of data.

- 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.

Deployment Options

- On-Premises: Install on your own infrastructure for full control.
- Cloud Deployment: Deploy on cloud platforms like AWS, Google Cloud, or Azure.
- Hybrid Deployment: Flexibility to use a combination of on-premises and cloud solutions.

Integration and Extensibility

- API Integration: Comprehensive REST and SOAP APIs for seamless integration with other systems.
- Plugin Architecture: Extend Pentaho's capabilities with a rich plugin ecosystem.
- Third-Party Tool Integration: Integrates with popular BI tools, databases, and other enterprise applications.

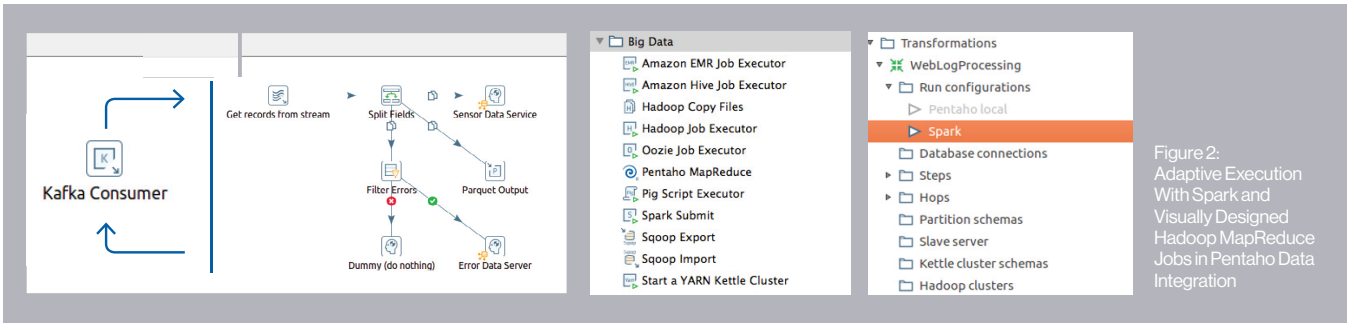

Figure 2: Adaptive Execution With Spark and Visually Designed Hadoop MapReduce Jobs in Pentaho Data Integration

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, Postgres, IBM MQ
- Spark and Hadoop: Cloudera, Hortonworks, Amazon EMR, MapR (HPE Ezmeral Data Fabric), Microsoft Azure HDInsights, and Elastic Search
- NoSQL databases and object stores: MongoDB, Cassandra, HBase, Hitachi Content Platform, AWS S3, Google Cloud Storage, Microsoft Azure ADLS Gen 2
- Analytic databases: Redshift, Snowflake, Vertica, Greenplum, Teradata, SAP HANA, Amazon Redshift, Google Big Query
- Business applications: SAP, Salesforce,Google Analytics
- Files: XML, JSON, Microsoft Excel, CSV, txt, Avro, Parquet, ORC, EBCDIC (mainframe), 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, including Amazon Redshift and Snowkflake.
- 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 technology 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 technology 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.
- Replace file names and locations with simple business names by integrating with Pentaho Data Catalog.

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.

Discover Pentaho Data Integration →

Pentaho Data Integration Technical Specifications and Compatible Components
Learn More

Pentaho

pentaho.com