Pentaho Data Integration

Big Data & More: The Power to Access, Prepare & Blend Multiple Data Sources Faster

With Pentaho, managing the enormous volumes and increased variety and velocity of data entering organizations, regardless of type of data and number of data sources, is simplified. Pentaho’s complete data integration platform delivers “analytics ready” data to end users 15X faster with visual tools that reduce time and complexity. Instead of coding in SQL or writing MapReduce, organizations immediately gain real value from their data, from data sources like Hadoop, NoSQL and relational data stores, with a graphical designer.

Integrate and Blend Big Data with Existing Enterprise Data

With broad connectivity to any data type and a high performance in-Hadoop execution, Pentaho makes it simplifies and speed the process of integrating existing databases with new sources of data.

Pentaho Data Integration’s graphical designer includes:
- Intuitive, drag and drop designer
- Rich library of pre-built components
- Dynamic transformations, to determine field mappings, validation and enrichment rules using variables
- Integrated debugger for testing and tuning job execution

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 and Intel, as well as popular NoSQL databases Cassandra and MongoDB and Splunk – for consumable and actionable analytics.

Big Data Integration and High-Volume Data Processing

Pentaho speeds time and reduces the complexity of integrating with big data sources. Pentaho’s intuitive graphical design provides:
- Native connectivity to leading Hadoop, NoSQL and analytic databases
- Visual designer for MapReduce jobs to reduce development cycles by as much as 15x
- Data preparation, modeling and exploration of unstructured data sets

Pentaho’s powerful data integration engine provides:
- Multi-threaded engine for fast execution
- Cluster support, enabling distributed processing of jobs across multiple nodes
- Unique in-Hadoop execution for extremely fast performance

Deliver Data to a Wide Variety of Applications

Pentaho’s out-of-the-box data standardization, enrichment and quality capabilities provide information to SaaS providers and ISVs the shape and form most-suited for their applications.
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:

- Standard relational databases, Oracle, DB2, MySQL, SQL Server
- Hadoop, Apache Hadoop, Cloudera, HortonWorks, MapR
- NoSQL databases, MongoDB, Cassandra, HBase
- Analytic databases, Vertica, Greenplum, Teradata
- Packages enterprise applications, SAP
- Cloud-based and SaaS applications, Salesforce, Amazon Web Services
- Files, XML, Excel, flat file and web service APIs

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 delivery in a multi-dimensional format for analytics
- Data delivery through real-time data services for operational 3rd party applications

Team Work and Collaboration for Developers

Pentaho Data Integration is built on a centralized repository where all stakeholders in a data integration project share and collaborate on developing data flows. Pentaho provides:

- Shared repository for collaboration among data analysts, job developers and data stewards
- Content management, versioning and locking to easily version jobs for roll-back to prior versions

Powerful Administration and Management

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

- Managing security privileges for users and roles
- Integrating with existing security definitions in LDAP and Active Directory
- Setting permissions to control user actions, read, execute or create
- Scheduling of data integration flows
- Monitoring and analyzing the performance of data integration processes

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

WHY PENTAHO DATA INTEGRATION?

- **Power of big data orchestration and integration:** Integration of all data - Hadoop, NoSQL and relational - in one platform; In-Hadoop and clustered execution of data processing for maximum scalability
- **Ease of use:** Simple set up; Intuitive graphical designer; No extra code generation; Over 100 out-of-the-box mapping objects, including a visual MapReduce designer for Hadoop
- **Modern and extensible:** 100% Java for cross-platform deployment; Pluggable architecture for adding connectors, transformations and user-defined expressions
- **High value, low cost:** No upfront fees; Subscription license model with no developer/user license fees; No maintenance fees