Pentaho

DATASHEET

Pentaho Data Integration

Data Integration That Delivers Clarity—Not Complexity.

More than just ETL (Extract, Transform, Load), Pentaho Data Integration is a no-code, end-to-end data orchestration platform that blends diverse datasets into a **single source of truth** as a basis for analysis, reporting, and AI initiatives.

With a drag-and-drop graphical interface, users can track data from ingestion through transformation to delivery—without writing code.

- Design and optimize pipelines with ease
- Built for scalability, simplicity, and self-service
- Enhanced data observability, quality, and governance with built-in lineage and metadata

Pentaho helps you navigate the growing volume and complexity of data, empowering business and technical teams to deliver trusted insights faster.

> "Moving data across a business is an art. Pentaho transforms art into better business value."
>
> **Warren Chang**, VP of Engineering, Borderfree

Drive measurable business value through faster time-to-insight, operational efficiency, and smarter data collaboration

Pentaho prepares data from any source, structured or unstructured, and orchestrates pipelines for delivery to any destination. By simplifying and automating the process—from ingestion to output—it unlocks self-service analytics and accelerates time to value.

Turn Data into Actionable Insights

Pentaho transforms how organizations access, prepare, and deliver data by seamlessly integrating data from disparate systems and different formats. With visual tools and drag-and-drop functionality, teams can trace every step of the data lifecycle—from origin to output—with clarity and control.

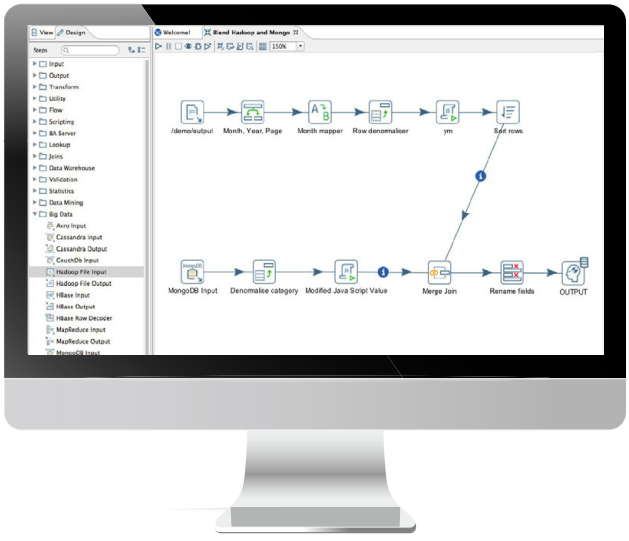


Figure 1: Drag-and-Drop Data Transformation in Pentaho Data Integration

Pentaho Data Integration

Data Processing Performance and Productivity

Pentaho reduces complexity in managing large-scale data pipelines while improving performance across the board.

- Code-free transformation design for accelerated development
- Metadata injection for efficient onboarding of new sources
- Cluster-executable pipelines for optimized throughput

Scalability, Simplicity, and Self-Service

Pentaho provides robust, visual pipeline design and broad connectivity for self-service access across data types and environments.

- Intuitive drag-and-drop designer
- Library of prebuilt connectors for databases, big data, cloud storage, and apps
- Visual profiling, charting, and in-flight analytics
- Orchestration with alerts and coordination of transformations
- Built-in enterprise scheduling and debugging tools

Deployment Options

- On-Premises: Maintain full control on your infrastructure
- Cloud Deployment: Deploy to AWS, Google Cloud, Azure, and more
- Hybrid Deployment: Flexibility to combine on-prem and cloud environments
- Containerized Deployment: Support for Docker and Kubernetes for scalable, portable operations

Integration and Extensibility

- API Integration: REST and SOAP APIs for seamless connectivity
- Plugin Ecosystem: Extend Pentaho with integrations for SAP, Salesforce, Kafka, ElasticSearch, Google Analytics, and more
- GenAI-Ready: Connect to LLMs and GenAI-powered parsing tools via purpose-built plugins
- AI/ML Operationalization: Operationalize models in R, Python, Spark, and Weka as part of your pipeline

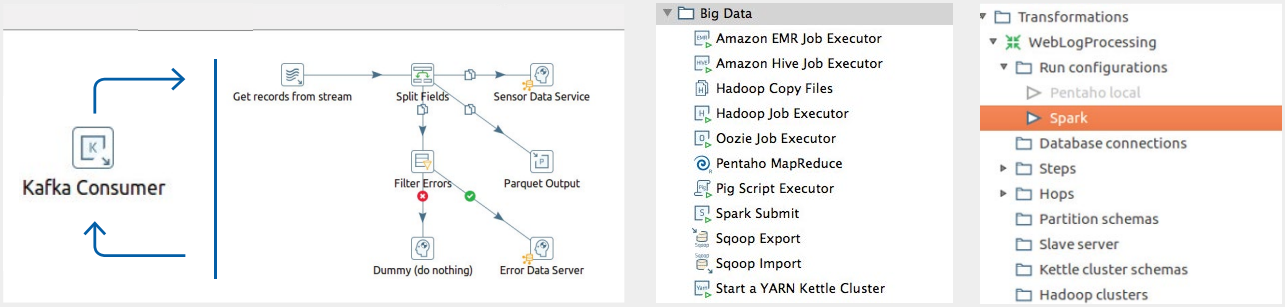


Figure 2: Adaptive Execution With Spark and Visually Designed Hadoop MapReduce Jobs in Pentaho Data Integration

Pentaho Data Integration

Broad Connectivity and Data Delivery

Pentaho connects to virtually any data source or format—from relational databases and Hadoop to NoSQL stores, cloud storage, and files. It enables high-performance ingestion and delivery pipelines through:

- Native connectivity and optimized bulk loading
- Virtual data services for instant access
- Automated publishing of metadata models
- Real-time streaming and delivery capabilities

Data Profiling and Data Quality

With built-in profiling and cleansing features, Pentaho ensures trusted, validated data pipelines.

- Detect rule violations and missing or malformed data
- Standardize and correct common fields like names, emails, and locations
- Integrate with leading data quality vendors
- Simplify naming and access with data catalog integration
- **Drive data observability with integrated lineage tracking, metadata management, and pipeline visibility via OpenLineage standards**

Pentaho also supports adaptive pipelines that respond to metadata changes to ensure smarter orchestration across evolving data environments.

DataOps Capabilities

Pentaho enables collaboration and governance through enterprise-grade operational tools.

- Centralized repository for team collaboration
- Versioning, rollback, and content locking
- Role-based access control and third-party security integration

Discover Pentaho Data Integration →

Pentaho Data Integration Technical Specifications and Compatible Components

Learn More

About Pentaho

Trusted by over 70% of the Fortune 100, Pentaho's end-to-end data integration and intelligence platform enables the foundational and AI-grade data fitness that drives faster and more confident decision-making.

Pentaho

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
pentaho.com

© Hitachi Vantara LLC 2025. All Rights Reserved. HITACHI and Pentaho are registered trademarks of Hitachi, Ltd. in the U.S. and other countries. All other trademarks, service marks and company names are properties of their respective owners.

P-BTD-DS-Pentaho-Data-Integration-17June25-B