

Pentaho 8.0 Overview

Pedro Alves

The forward-looking statements contained in this document represent an outline of our current intended product direction. It is provided for information purposes only and is not a commitment to deliver any new or enhanced product or functionality, or that we will pursue the product direction described. Facts and circumstances may occur which may impact current plans, resulting in changes to the information in this presentation. This information is current only as of the date it is made and should not be relied upon in making purchasing decisions. The development, release (if at all), and timing of any features or functionality described for the Pentaho products remains at the sole discretion of Pentaho.

Pentaho Business Analytics Platform



Data Engineer



Data Analyst / Data Scientist



Business Analyst



Consumer

Production Reporting

Interactive Query and
Analysis

Custom and Self-Service
Dashboards

Pentaho Data Integration

Data Preparation | Integrated Machine Learning

OPEN AND EMBEDDABLE



Operational Data



Big Data

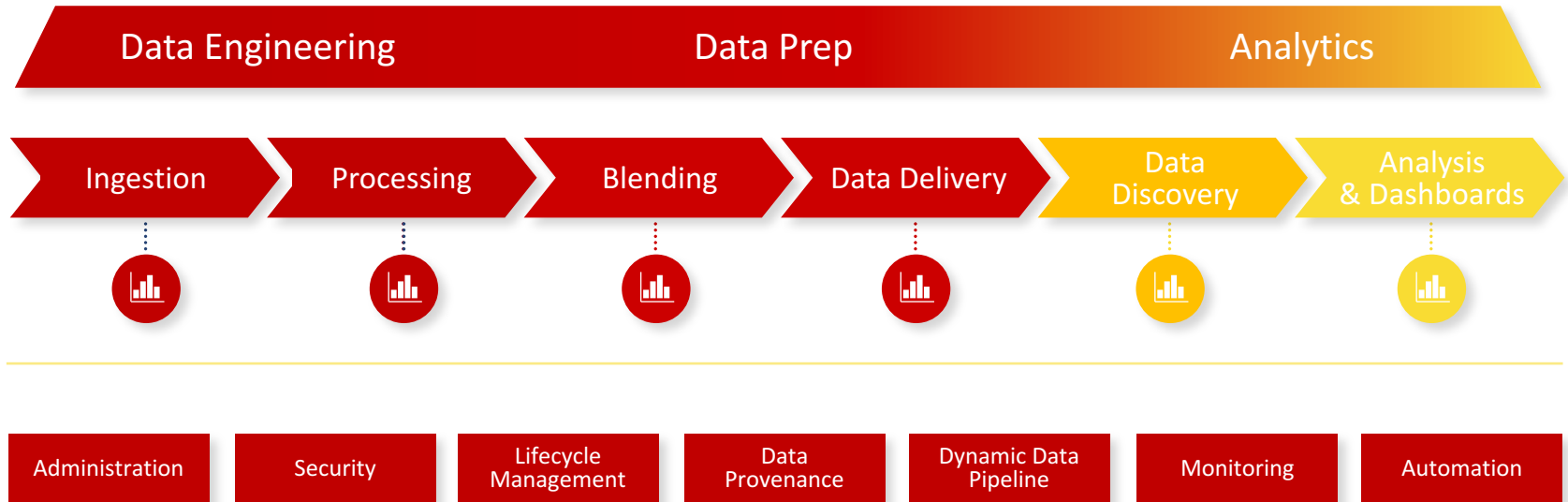


Data Stream



Public/Private Clouds

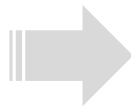
Future Vision: A Single Consistent Experience



Introducing Pentaho 8.0

Challenge #1

Data volumes and velocity are growing exponentially



Pentaho 8.0

Broadens connectivity to streaming data sources

- Connect to Kafka streams
- Stream processing with Spark
- Big data security with Knox

Challenge #2

Processing and storage resources are constrained



Pentaho 8.0

Optimizes processing resources

- Enhanced Adaptive Execution (AEL)
- Native Avro and Parquet handling
- Worker nodes for “Scale-out”

Challenge #3

Shortage of Big Data talent and lack of productivity



Pentaho 8.0

Boosts team productivity across the pipeline

- Data explorer filters
- Improved repository UX
- Extended operations mart

Pentaho 8

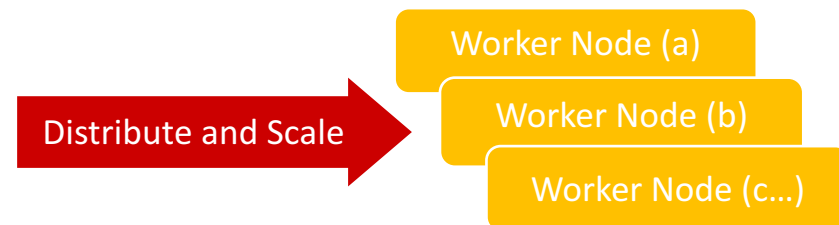
Platform and Scalability

Worker Nodes

New theme

Scale work items across multiple nodes (containers)

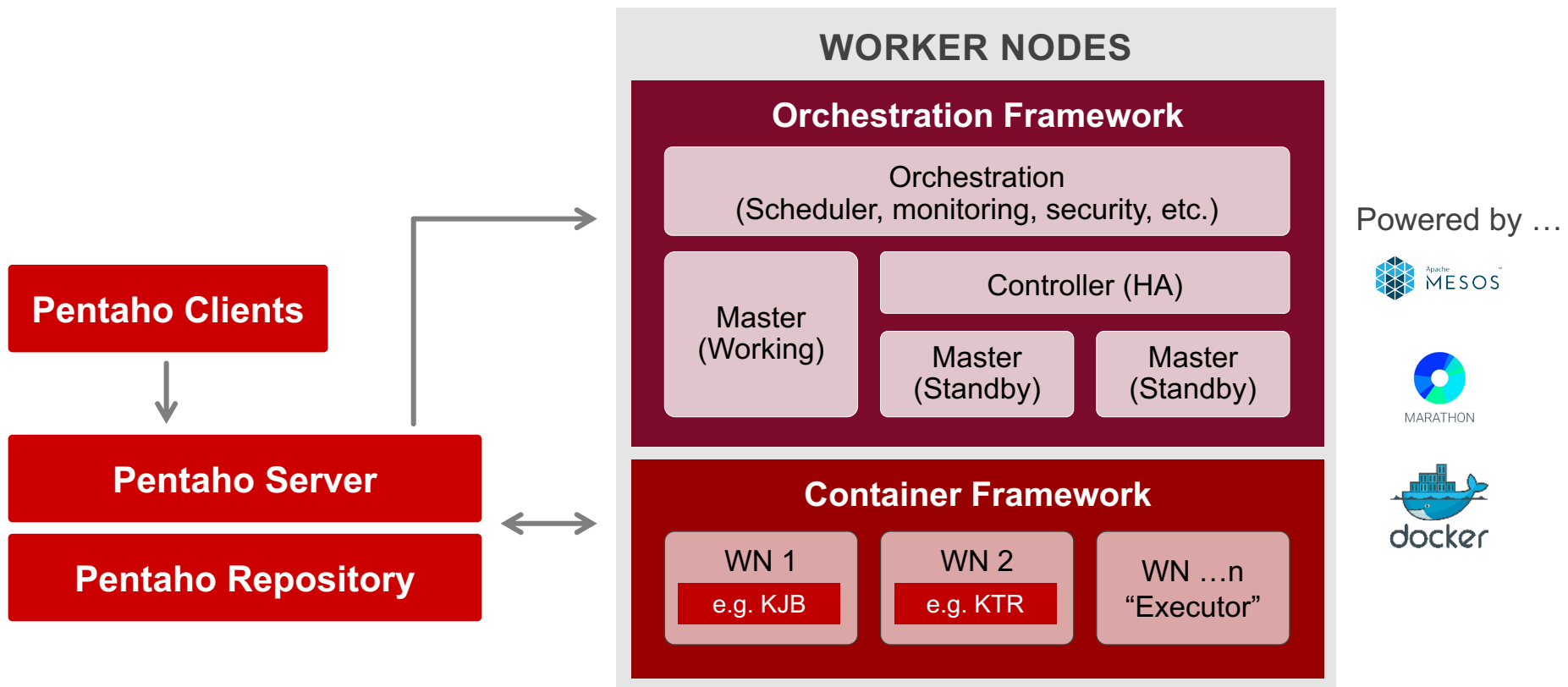
- Easily add and remove resources as required
- Monitor and balance changing workloads
- Deploy on premise, cloud and hybrid



NEW in Pentaho 8.0

- ✓ Container framework
- ✓ Orchestration framework
- ✓ Node monitoring
- ✓ Enhanced HA implementation

Worker Nodes Architecture



← → ↻ 127.0.0.1:8080/pentaho/Home?locale=en_US

File View Tools Help

Home ▾ admin ▾

Browse Files

Create New

Manage Data Sources

Documentation

Getting Started

Welcome Samples Tutorials

Learn More About Pentaho Business Analytics

This video will provide you a high level overview of:

- How Pentaho works
- What it can do
- How to get started
- Where to go for help

Data Integration

Streaming support!

Run configurations for Jobs

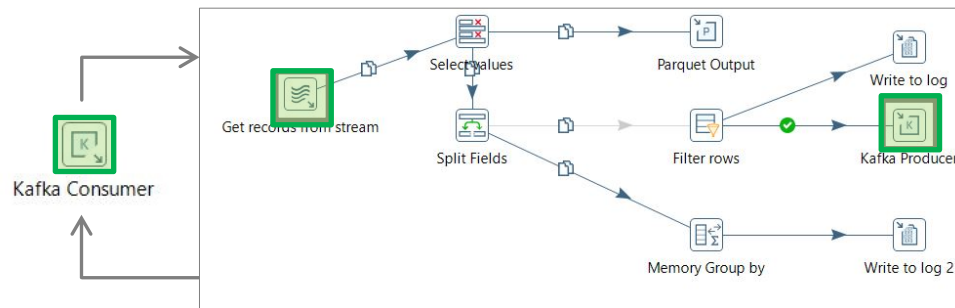
Filters in Data Explorer

New Open / Save experience

Streaming for Time Sensitive Insight

Enable use cases that require real-time processing, monitoring and aggregation

- Real-time device monitoring
- Log-file aggregation
- Notifications
- And more...

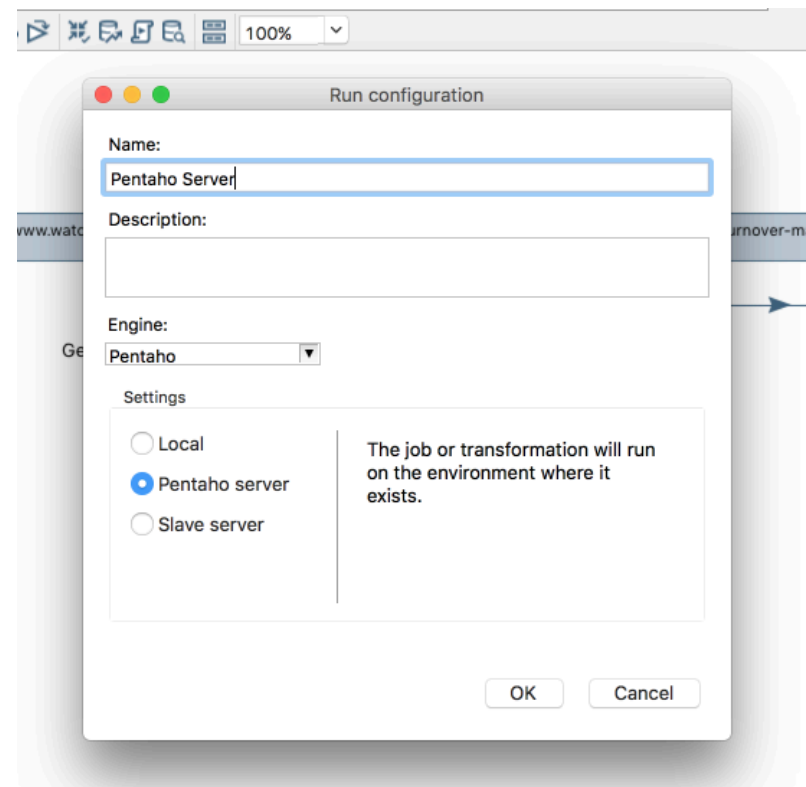


NEW in Pentaho 8.0

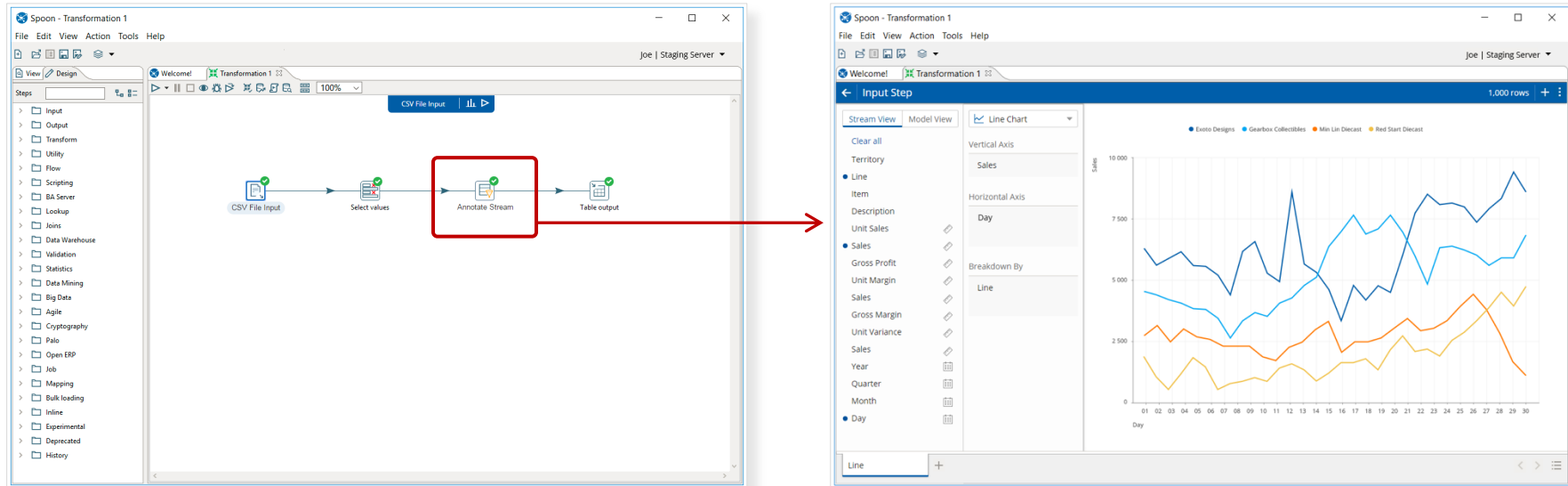
- ✓ Kafka Producer Step
- ✓ Kafka Consumer Step
- ✓ Get records from stream Step
- ✓ Spark streaming via AEL

Run Configuration

- Execute on server
- Leverage worker nodes



Pentaho 7.0 – Data Explorer



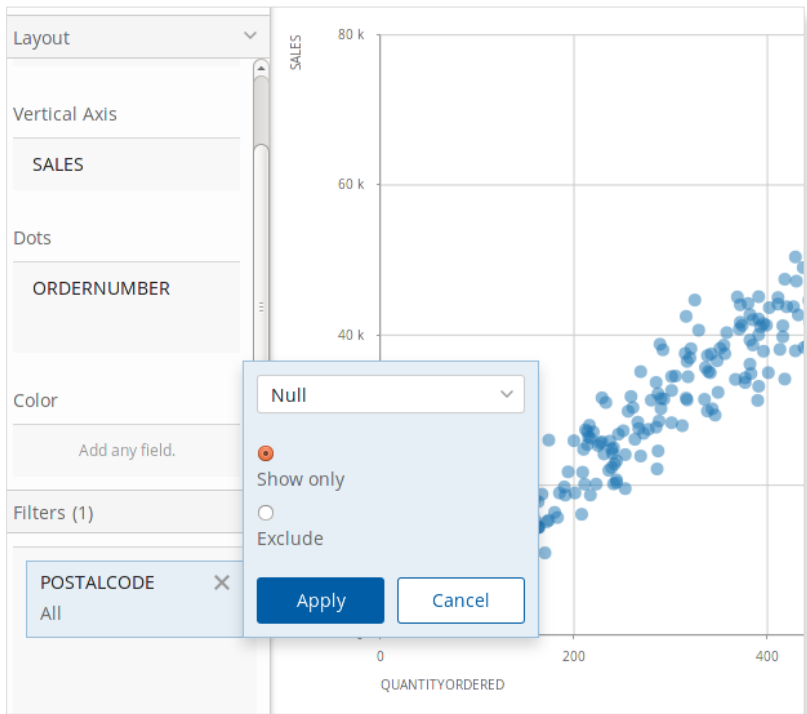
Access visualizations during data prep for inspection and prototyping

Enhanced data inspection in PDI

- Identify data to be cleaned or removed
- Deliver data to the business more quickly

New in Pentaho 8.0

- ✓ Numeric filters
- ✓ String filters
- ✓ Include/Exclude data points

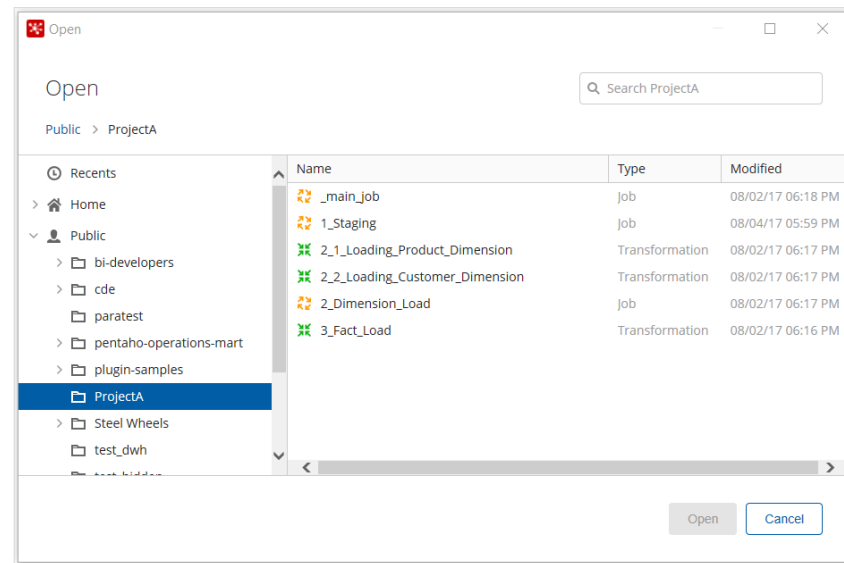


Enhanced user experience for content repository

- Consolidated open and save dialogs
- Enhanced search
- “Recently used” files option
- “Sticky” directories

Business Benefits

- Increased productivity



Big Data

Improvements on AEL

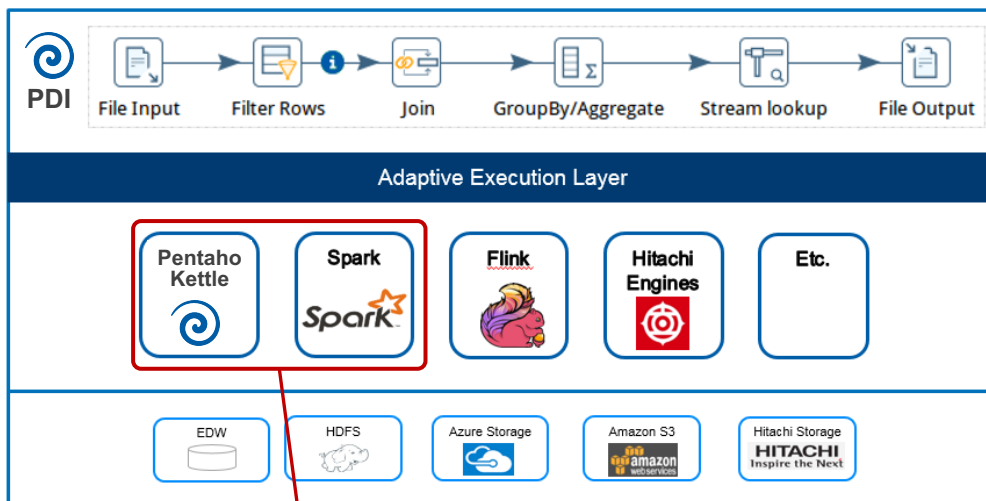
Big Data File Formats - Avro and Parquet

Big Data Security - Support for Knox

VFS improvements for Hadoop Clusters

Pentaho 7.1 – Adaptive Execution for Spark

Transformation
Logic



Processing
Engines

Storage
Layer

- ✓ No Coding
- ✓ Build Once
- ✓ Execute on Any*
Engine

*Currently Available Engines

Enhanced Adaptive Execution

Simplified setup

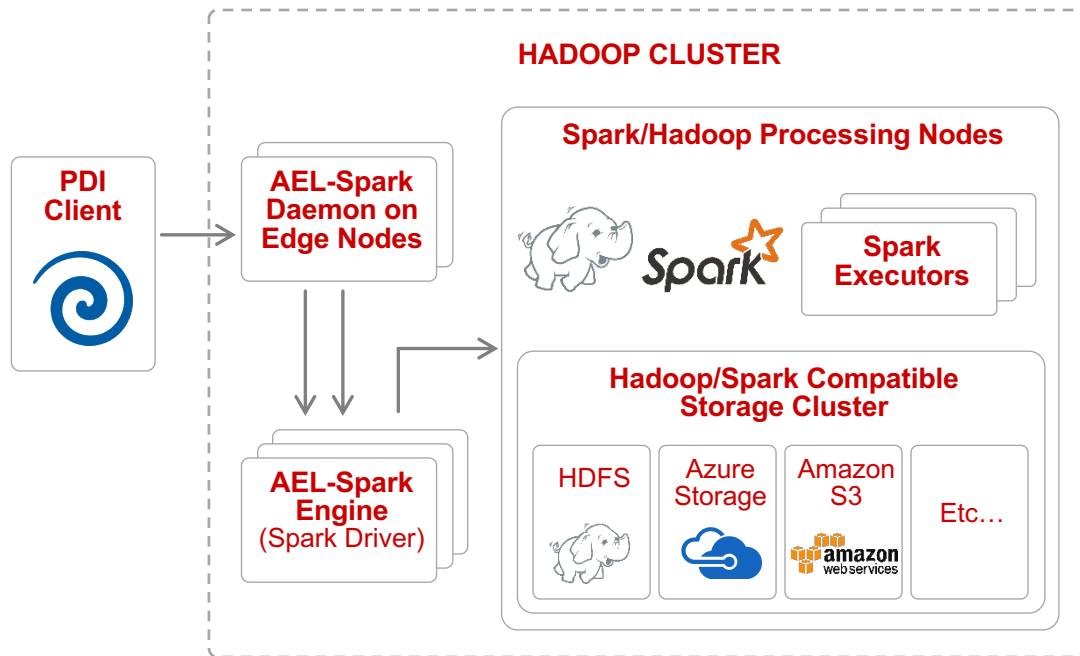
- Eliminated “Zookeeper” component
- Reduced number of setup steps

Hardened deployment

- Fail-over at the edge
- Kerberos impersonation for client

More flexible

- Support multiple run configurations
- Customize cluster settings per job type



Visual handling of data files with common big data file formats

- Reading and writing files with specific steps
- Natively executes in Spark via AEL

NEW in Pentaho 8.0

- ✓ Avro Input and Output Steps
- ✓ Parquet Input and Output Steps

The screenshot shows the 'Parquet Output' configuration window. The 'Step name' is 'HDFS - RetailWeblog_Cleansed File'. The 'Location' is set to 'HDFS'. The 'Folder/File name' is 'hdfs://users/rsaha/cleansed_weblog.parquet'. There are checkboxes for 'Overwrite existing output file' and 'Dictionary encoding'. The 'Fields' tab is active, showing 'Compression' as 'None', 'Version' as 'parquet 2.0', 'Row group size (MB)', 'Data page size (KB)', and 'Page size (KB)' as '1024'. The 'Options' tab shows 'Extension' as 'parquet', 'Include date in file name', 'Include time in file name', and 'Specify date time format'.

Knox Gateway for More Secure Clusters

Pentaho can interact with data services in Knox gateway protected Hadoop clusters*

Benefits

- Highly secured Knox-protected Hadoop clusters can now be integrated in ETL pipeline

*Knox is only available for Hortonworks HDP distro.

Hadoop Cluster

Cluster name:
Knox HDP Cluster

Use a gateway to connect to the cluster

Gateway

URL:
https://hdpcluster8443/gateway

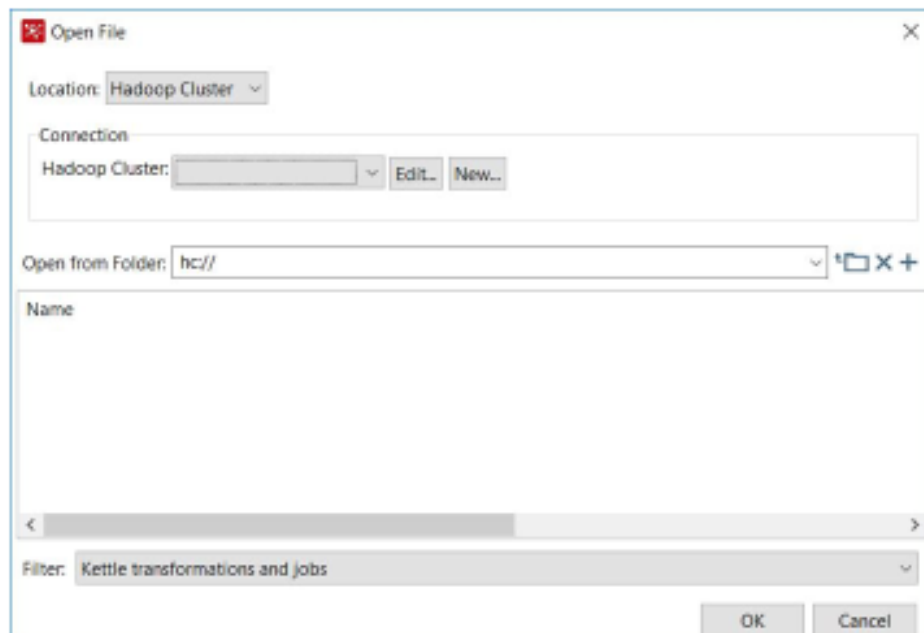
Username: admin Password: ●●●●

Help Test OK Cancel

VFS improvements for Hadoop Clusters

In order to simplify the overall lifecycle of jobs and transformations we made the hadoop clusters available through VFS, on the format

hc://hadoop_cluster/.



Others

Ops Mart for Oracle, MySQL, SQL Server
Platform password security improvements

PDI mavenization

Documentation changes on help.pentaho.com

Feature Removals:

Analyzer on MongoDB

Mobile Plug-in (Deprecated in 7.1)

MindTouch Documentation Improvements (continued)

- Streamlining of site content pages:
 - Simpler site structure and click paths to accelerate user content access
 - Eliminates pages with little content value
 - Fun fact: 7.1 contained 1200+ pages; 8.0 contains 900+ pages with no elimination of valuable content
- Ongoing migration of big data steps from the wiki pages to the MindTouch Doc site

Business Benefits

- Improved access to on-line documentation

Data Integration

- Drag and drop numeric and non-numeric filters in Data Explorer
- Chart actions in Data Explorer to include or exclude data
- Cleaner and easier design/UX for PDI Repository dialogs
- Consolidated PDI Repository dialogs, replacing multiple pop-ups
- PDI Repository Dialogs remember last opened directory
- Jobs can leverage Run Configurations for seamless user experience
- Run Configuration option to run jobs on Pentaho Server

Streaming

- Kafka Producer step for outputting streaming data
- Kafka Consumer step for streaming data input
- Get Records from Stream step for stream processing
- Ability to run AEL on Spark Streaming for transformations that use Kafka streaming ingest

Enterprise Platform

- Worker Nodes Scale-Out solution to drive superior agility
- Ops Mart for Oracle, MySQL, SQL Server
- Ruby Theme – new platform branding for browser and client tools
- Platform password security improvements
- PDI Mavenization for infra alignment

Big Data

- Simplified Spark AEL setup
- Spark AEL fail-over and load balancing
- Support for Spark AEL on Hortonworks
- Enhanced security for Spark AEL including Kerberos impersonation and secure connectivity
- Customization of Spark AEL processing (i.e. memory and other settings)
- Parquet input and output steps
- Avro input and output steps
- Big Data Security with HDP Knox Gateway
- VFS Improvements for named Hadoop clusters

Additional Items

- Big Data Sandbox VM updates
- Cumulative service pack framework
- Documentation improvements on help.pentaho.com

Q&A