pentaho®

open source business intelligence™

What’s New in Pentaho Data Integration

Enterprise Edition 4.2

For the latest information, please visit our web site at www.pentaho.com

Contents

Purpose of This Document

This document introduces new capabilities delivered in Pentaho Data Integration (PDI) 4.2. It is intended for users that already have a working familiarity with the capabilities of Pentaho Data Integration, but is not a complete review of Pentaho Data Integration's functional capabilities.

Pentaho Data Integration Enterprise Edition 4.2

Pentaho Data Integration 4.2 is a feature release delivering a broad range of enhancements and new features including enhanced support for Big Data technologies, tighter integration with Pentaho Reporting, improved import/export capabilities, exposure of transformation output data as simple web services, ability to restart jobs and much more. This release also includes large number of enhancements and bug fixes to existing features. For a complete listing of changes in PDI 4.2, please view the release notes in JIRA.

Major Enhancements

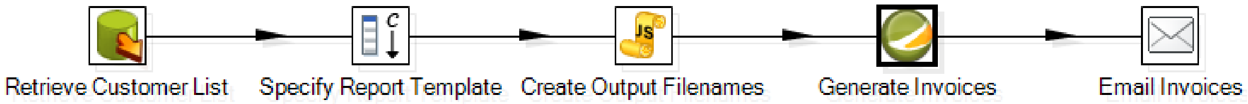
Enhancements for Big Data

PDI 4.2 includes a number of improvements and new features for tackling Big Data use cases including:

- Updated support for the latest Hadoop distributions including Cloudera Distribution for Apache Hadoop 3.0, Apache Hadoop .20.2 and Apache Hive 0.7
- HBase Integration – input and output steps providing the ability to read and write data to HBase with full support for structured and unstructured data including for data type support
- MongoDB – input step for reading structured and unstructured data from MongoDB
- New Bulkloaders – bulk load interfaces provide high speed loading of large volumes of data to relational and analytic databases. This release includes new Bulk Loading connectors for Greenplum (parallel loading through gpload), ElasticSearch, Vectorwise, MySQL, and PostgreSQL.

Pentaho Reporting Integration

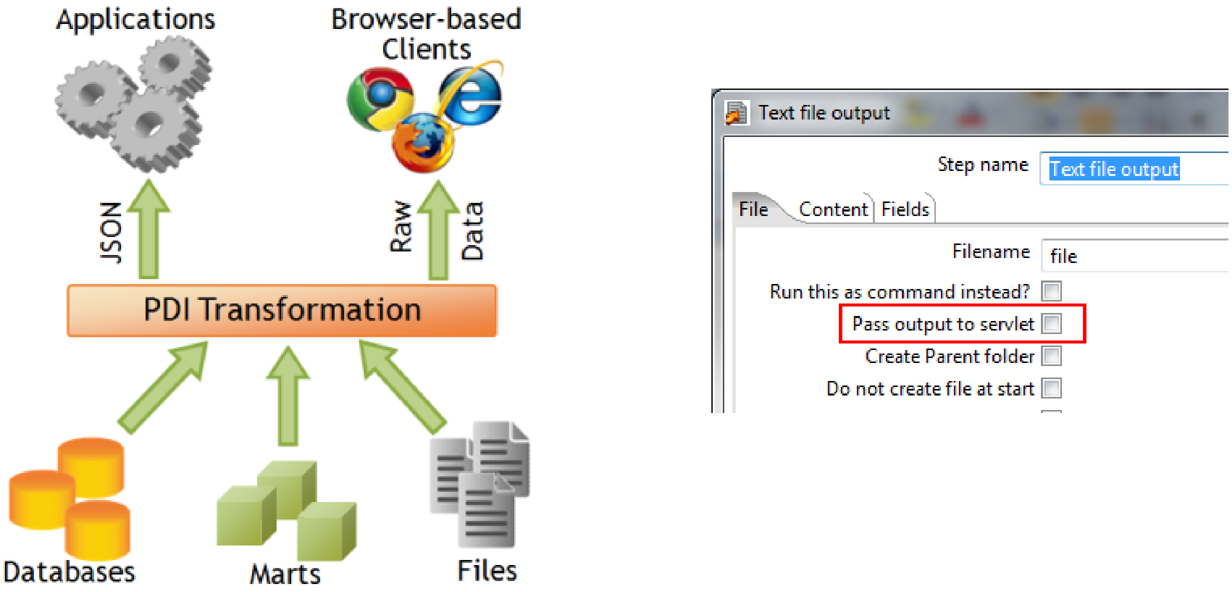
PDI 4.2 includes the Pentaho Reporting engine and two new steps powerful new steps allowing you to take advantage of Pentaho's Reporting capabilities as part of ETL workflows or for automatic generation of simple documentation about your transformations and jobs.



Sample Transformation Illustrating Dynamic Report Generation and Distribution

Exposing Data as a Service

There are many use cases where you need a quick and easy way to share processed data with a remote user or application. PDI 4.2 introduces a simple means to automatically expose either text-based or JSON output as a basic Web Service. To accomplish this, you simply add a Text File Output or JSON Output step to your transformation, enable the ‘Pass output to servlet’ option in the step configuration dialog and then the output can be streamed to a remote location using a simple URL.



Step name: Text file output

Filename: file

Run this as command instead? □

Pass output to servlet □

Create Parent folder □

Do not create file at start □

Transformation Output Data as a Web Service

Additional Enhancements

- New graphical performance and progress feedback when executing transformations in Spoon
- Ability to restart Jobs from a specific Job Entry
- Carte Improvements
    - Ability to perform parallel runs of a clustered transformation
    - Service for listing reserved and free sockets
    - Additional configuration options for setting slave sequences
    - Ability to time-out stale objects using environment variables
- Repository Import/Export improvements
    - Export a specific repository folder
    - Export/Import based on validation rules
    - Command Line Utility for Import/Export

New Transformation Steps

Pentaho Data Integration 4.2 adds the following new transformation steps:

Icon	Step Name	Description
	HBase File Input	Read data from an HBase table.
	HBase File Output	Write data to an HBase table

Icon	Step Name	Description
	Get repository names	List detailed information about Transformations or Jobs stored in a repository. This includes attributes such as name, type and last modified date.
	GZIP CSV Input	Parallel reader for GZIP compressed CSV files.
7	HL7 Input	Reads and parses HL7 messages and outputs a series of values from the message.
	MongoDB Input	Read data from a MongoDB table.
	XML Input Stream	Fast reader for very large XML documents.
	Automatic Documentation Output	Generate basic documentation for transformations and jobs.
	Microsoft Excel Writer	Writes or appends data to an Excel document. Includes support for .XLSX file format.
	Pentaho Reporting Output	Generate Pentaho Reports from and existing PRPT including full support for parameterization.
2	Get ID from slave server	Retrieves unique IDs from a slave server.
	Set field Value	Set the value of a field based on the value of another field.
C	Set field value to a constant	Set the value of a field to a constant.
	ETL Metadata Injection	Inject metadata into an existing transformation prior to execution allowing the creation of highly dynamic ETL solutions.
	Prioritize streams	Set the priority for streams in a transformation flow.
	Single Threader	Executes a transformation snippet in a single thread.
REST	REST Client	Read data from a RESTful service.
	ElasticSearch Bulk Insert	Performs bulk inserts to ElasticSearch
	Greenplum Bulk Loader	Performs parallel (gpdist), bulk inserts into the Greenplum distributed database.
	Ingres VectorWise Bulk Loader	Performs bulk inserts to Ingres Vectorwise
	MySQL Bulk Loader	Performs bulk inserts to MySQL.
	PostgreSQL Bulk Loader	Performs bulk inserts to PostgreSQL.

New Job Entries

Pentaho Data Integration 4.2 adds the following new job entries:

Icon	Job Entry Name	Description
🔒	Decrypt files with PGP	Decrypt files encoded with PGP encryption.
🔒	Encrypt files with PGP	Encrypt a file using PGP encryption.
🔒✓	Verify file signature with PGP	Verify the file signature using PGP.
🐘	Pig Script Executor	Execute a Pig Script in Apache Hadoop.
Dos 010	Convert file between DOS and UNIX	Converts a file from DOS to UNIX format or vice versa.
7 ACK	HL7 MLLP Acknowledge	Creates an HL7 MLLP client server and acknowledges the receipt of an HL7 message.
7 IN	HL7 MLLP Input	Creates an HL7 MLLP client server, accepts a single HL7 message, and sets this as a variable in the parent job.