



What's New in Pentaho Data Integration Enterprise Edition 4.1

Copyright © 2010 Pentaho Corporation. Redistribution permitted. All trademarks are the property of their respective owners.

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

Last Modified on October 28, 2010

Contents

- Contents 2
- Purpose of This Document 3
- Pentaho Data Integration Enterprise Edition 4.1 3
- Pentaho Data Integration for Hadoop 3
- Enhancements to Hops 4
- Metadata Injection 4
- General Steps and Job Entries..... 4
- New Transformation Steps..... 4
- New Job Entries 5

Purpose of This Document

This document introduces new capabilities delivered in Pentaho Data Integration (PDI) 4.1. It is intended to address people who have a working familiarity with the capabilities of Pentaho Data Integration (PDI), but is not a complete review of Pentaho Data Integration's functional capabilities.

Pentaho Data Integration Enterprise Edition 4.1

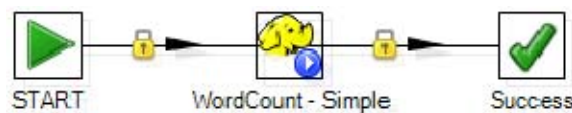
This PDI release includes: integration with Apache Hadoop, making it easy to leverage Hadoop for storing and processing very large data sets; usability improvements for working with hops; the first ever support for the concept of Metadata Injection; and a number of new general purpose transformation steps and job entries.

Pentaho Data Integration for Hadoop

More and more enterprises are turning to Hadoop to reduce costs and improve their ability to extract actionable business insight from the vast amount of data being collected throughout the enterprise. Hadoop's massive parallel processing capabilities, along with the ability to store extremely large amounts of data in a low cost and reliable manner, make it an attractive option for building Business Intelligence solutions for Big Data. However, Hadoop presents many challenges to traditional BI Data Integration users, including a steep technical learning curve, a lack of qualified technical staff, and the lack of appropriate tools for performing data integration and business intelligence tasks with Hadoop.

Pentaho Data Integration Enterprise Edition 4.1 delivers comprehensive integration with Hadoop, which lowers the technical barriers to adopting Hadoop for Big Data projects. By using Pentaho Data Integration's easy-to-use, graphical design environment, ETL Designers can now harness the power of Hadoop with zero Java development to address common Data Integration use cases including:

- Moving data files into and out of the Hadoop Distributed File System (HDFS)
- Input/Output data to and from Hadoop using standard SQL statements
- Coordination and execution of Hadoop tasks as part of larger Data Integration and Business Intelligence workflows
- Graphical Design of new MapReduce jobs taking advantage of Pentaho Data Integration's vast library of pre-built mapping and data transformation steps



Pentaho for Hadoop simplifies the use of Hadoop for analytics including file input and output steps as well as managing Hadoop jobs

Pentaho Data Integration Enterprise Edition 4.1 supports the latest releases of Apache Hadoop as well as popular commercial distributions such as Cloudera Distribution for Hadoop and Amazon Elastic MapReduce. For information and best practices on how to incorporate Hadoop into your Big Data architecture, visit <http://www.pentaho.com/hadoop/resources.php>.

Enhancements to Hops

Pentaho Data Integration 4.1 enhances the handling of hops between steps and job entries by allowing all hops downstream from a certain point or among all selected steps or job entries to be enabled or disabled. This allows for easier debugging of a faulty step at the end of the transformation and you can now disable and enable hops simply by clicking on them once. In addition, when hops are split, target and error handling info is retained.

Metadata Injection






Pentaho Data Integration 4.1 supports for the first time in data integration history the concept of Metadata Injection. Metadata Injection offers increased flexibility for developers who want to treat their ETL metadata as data. Last-minute injection of file layout and field selection into a transformation template makes this possible. It can drastically reduce the number of data transformations in situations where patterns can be discovered in the data integration workload. Implemented as a metadata injection step, this feature allows developers to dynamically set step properties in transformations. The step exposes all the available properties of the step and enables injection of file names, the removal or renaming of fields, and other metadata properties.









General Steps and Job Entries

In addition to the Pentaho for Hadoop functionality, Pentaho Data Integration 4.1 includes a number of new steps and job entries designed to increase developer productivity. These include a conditional blocking step, JSON and YAML input steps, a string operations step, and a write to file job entry step. Below is a complete list of new steps and transformations.

New Transformation Steps






Pentaho Data Integration 4.1 adds the following new transformation steps:

Icon	Step Name	Description
	Hadoop File Input	Processes files from an HDFS or Amazon S3 location.
	Hadoop File Output	Creates files in an HDFS location.
	Conditional Blocking Step	Block this step until steps finish, allows building step logic depending on some others steps execution
	JSON Input Step	Enables JSON step to execute even if defined path does not exist
	JSON Output Step	Create JSON block and output in a field of a file.

	LDAP Output Step	Perform Insert, Upsert, Update, Add and Delete operations on records based on their DN.
	YAML Input Step	Enables reading information from a YAML file.
	Email Messages Input	Read POP3/IMAP server and retrieve messages.
	Generate Random Credit Card Number	Generates random valid Credit Card numbers.
	String Operations Step	Enables string operations including trimming, padding, lowercase/uppercase, InitCap, Escape (XML, SQL, CDATA, HTML), extract only digits, remove special characters (CR, LF, Espace, Tab)
	S3 File Output	Creates files in an S3 file location.
	Run SSH Commands	Runs SSH commands and returns results.
	Output steps metrics	Returns metrics for one or more steps within a transformation.

New Job Entries

Pentaho Data Integration 4.1 adds the following new job steps/entries:

Icon	Step Name	Description
	Amazon EMR Job Executor	Executes Map/Reduce jobs in Amazon EMR
	Hadoop Copy Files	Copies files to and from HDFS or Amazon S3
	Hadoop Job Executor	Executes Map/Reduce jobs in Hadoop
	Hadoop Transformation Job Executor	Executes PDI transformation-based Map/Reduce jobs in Hadoop
	Write to File Job Entry	At job level, directly write some data (static or in variables)