



What's New in Pentaho Data Integration 3.0?

Copyright © 2007 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 November 12, 2007

Contents

- Contents..... 2
- Purpose of This Document..... 3
- Performance and Scalability..... 3
 - Separation of Data and Metadata 3
 - Optimized Flat-file Handling 3
- ETL Developer Productivity 3
 - Integrated Debugger 3
 - Updated User Interface..... 4
- Community-Fueled Evolution 5
 - Univariate Statistics Plugin 5
 - New Data Sources and Transformation Steps 6
 - New Job Entries..... 6

Purpose of This Document

This document is intended for users who have a working familiarity with the capabilities of Pentaho Data Integration. This document is focused on introducing the new capabilities delivered in Pentaho Data Integration 3.0. It is not intended to be a complete review of Pentaho Data Integration's functional capabilities.

Performance and Scalability

Separation of Data and Metadata

Pentaho Data Integration 3.0 delivers architectural enhancements providing a clean separation of data and metadata. This redesign of the internal data engine results in significant performance improvements and reduced memory consumption by reducing internal creation of java objects during transformation execution. The performance boost resulting from this change ranges by step type from 15% in the Table Output Step to over 900% in the Select Values Step. For more details on the performance improvements of individual steps, please see the Kettle 3.0 Change Log document found in the docs directory of your Pentaho Data Integration installation.

Optimized Flat-file Handling

Pentaho Data Integration 3.0 includes numerous optimizations designed to improve performance when working with large flat files (.txt or .csv files). Text files are a popular option when organizations need to integrate data from non-relational sources like mainframe systems. To accelerate flat-file handling, Pentaho Data Integration 3.0 includes the following enhancements:

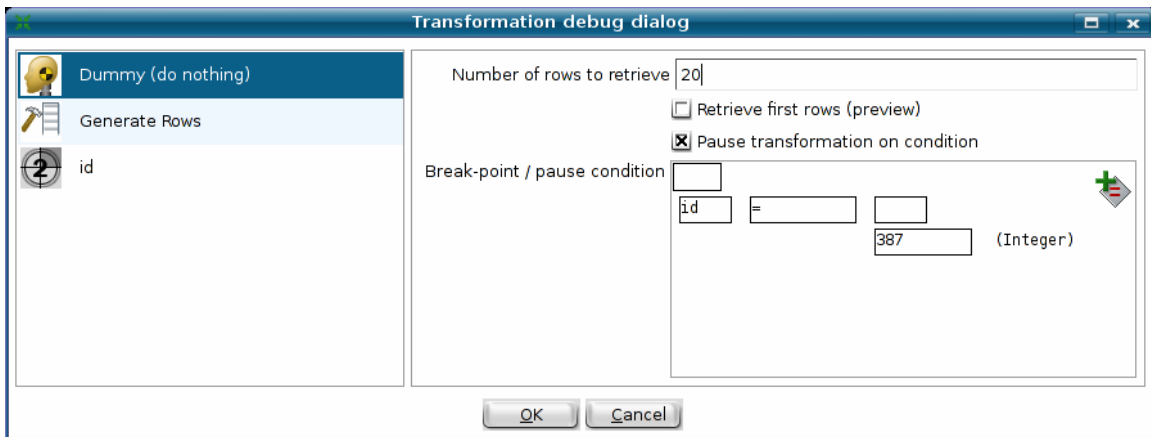
- Lazy Type Conversion – delays data type conversions as long as possible with the goal of not performing any conversions where the source and target data types are the same
- Parallel file reading – provides the ability for multiple step copies to read from the same file in parallel, thereby allowing you to maximize the I/O capacity of your ETL environment
- Non-blocking I/O (NIO)– improves performance when reading large blocks of data

These enhancements, along with numerous changes to streamline the flat file algorithms, allow Pentaho Data Integration 3.0 to process massive flat files over 5 times faster than previous releases.

ETL Developer Productivity

Integrated Debugger

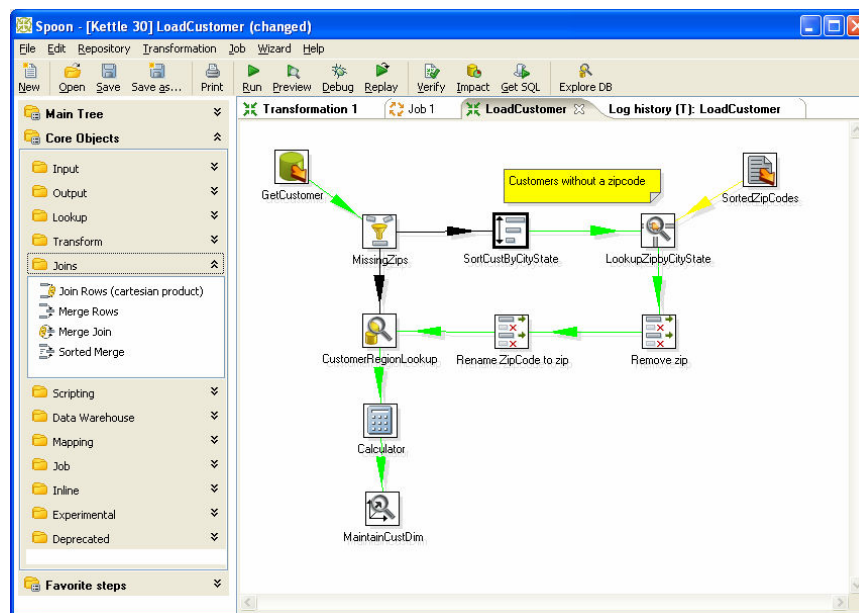
Pentaho Data Integration 3.0 adds an integrated debugger designed to improve ETL developer productivity by providing conditional breakpoints in transformation execution, the ability to pause and resume transformation execution, and the ability to specify the number of rows to be used in test executions.



Pentaho Data integration 3.0 provides an integrated debugger to quickly identify and correct issues.

Updated User Interface

Pentaho Data Integration 3.0 includes an updated user interface to provide more intuitive selection and use of the product's rich library of transformation steps and job entries. It also includes a new toolbar to make it easy for ETL developers to quickly access frequently-used features.



Pentaho Data integration 3.0 provides updated user interface as well as a new toolbar to make it easy to access frequently-used features.

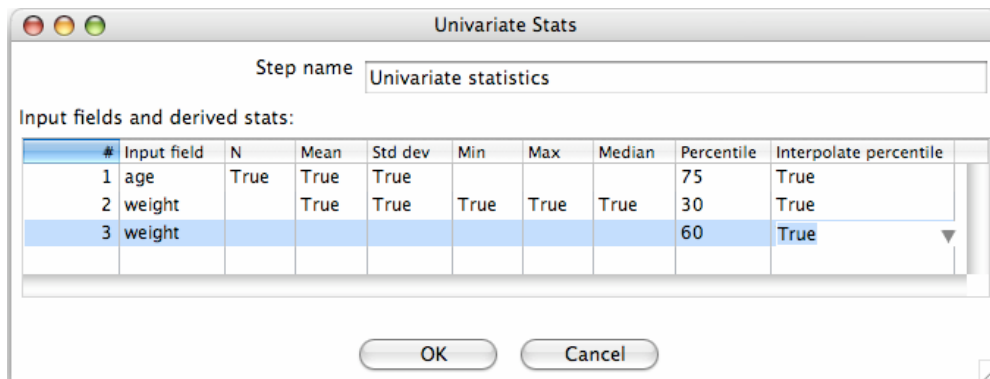
Community-Fueled Evolution

Pentaho Engineering and the Pentaho Data Integration Community have driven significant investment and evolution in the functional capabilities of Pentaho Data Integration 3.0. This release is the result of the hard work from over 20 core committers from 5 different countries and constitutes over 2000 commits and 56,000+ net new lines of code. The following sections describe a sample of the new plugins, transformation steps, and job entries introduced in Pentaho Data Integration 3.0.

Univariate Statistics Plugin

Pentaho Data Integration 3.0 provides a new plug-in to allow univariate statistics calculations during transformation processes. These calculations include mean, median, standard deviation, as well as percentile calculations. These calculations allow organizations to enrich their data during ETL processes. To download the Univariate Statistics Plugin, please visit the Pentaho Data Integration Plugins page:








<http://wiki.pentaho.org/display/EAI/List+of+Available+Pentaho+Data+Integration+Plug-Ins>



Pentaho Data integration 3.0 includes new univariate statistics calculations that can enrich data during the transformation and loading process.

New Data Sources and Transformation Steps








Pentaho Data Integration 3.0 adds new data sources including Sybase IQ, BMC Remedy AR System, LDAP directories, and Microsoft Access, and more.

Icon	Step Name	Description
	LDAP Input	Extract data from an LDAP server
	Regular Expression Evaluator	evaluates a Regex expression against a specified field on the stream
	MS Access Input	extract data from MS Access
	Mondrian Input	extract data from Pentaho Mondrian OLAP server
	Mondrian Closure Generator	creates a closure table by calculating all possible parent-child relationships and attaching the distance (in levels) from the parent to the child
	Append Streams	append the data from one stream to the tail of another stream
	Get File Row Counts	return a list of row counts from one or more files

New Transformation Steps provide integration with LDAP data, Microsoft Access, Mondrian, and more.

New Job Entries

Pentaho Data Integration 3.0 adds new job entries including file copy and delete, FTP to remote destination, Unzip packed files, XML document verification, and customizable logging.

Icon	Step Name	Description
	XSD Validator	validates an XML document against an XML Schema Definition (XSD)
	DTD Validator	validates an XML document against a Document Type Definition (DTD)
	Write to Log	write a specific entry to the execution log
	Unzip	unzip a file or group of files
	FTP Put	upload files using FTP
	Delete files	delete a set of files and/or directories
	Copy Files	easily move files and/or directories from one location to another

New Job Entries provide the ability to easily copy, delete, and ftp files.