pentaho

Version 5.1

Explore What's New in DI

Copyright Page

Help and Support Resources

To view the most up-to-date help content, visit https://help.pentaho.com.

If you do not find answers to your questions here, please contact your Pentaho technical support representative.

Support-related questions should be submitted through the Pentaho Customer Support Portal at http://support.pentaho.com.

For information about how to purchase support or enable an additional named support contact, please contact your sales representative, or send an email to sales@pentaho.com.

For information about instructor-led training, visit http://www.pentaho.com/training.

Liability Limits and Warranty Disclaimer

The author(s) of this document have used their best efforts in preparing the content and the programs contained in it. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, express or implied, with regard to these programs or the documentation contained in this book.

The author(s) and Pentaho shall not be liable in the event of incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of the programs, associated instructions, and/or claims.

Trademarks

The trademarks, logos, and service marks ("Marks") displayed on this website are the property of Pentaho Corporation or third party owners of such Marks. You are not permitted to use, copy, or imitate the Mark, in whole or in part, without the prior written consent of Pentaho Corporation or such third party. Trademarks of Pentaho Corporation include, but are not limited, to "Pentaho", its products, services and the Pentaho logo.

Trademarked names may appear throughout this website. Rather than list the names and entities that own the trademarks or inserting a trademark symbol with each mention of the trademarked name, Pentaho Corporation states that it is using the names for editorial purposes only and to the benefit of the trademark owner, with no intention of infringing upon that trademark.

Third-Party Open Source Software

For a listing of open source software used by each Pentaho component, navigate to the folder that contains the Pentaho component. Within that folder, locate a folder named licenses. The licenses folder contains HTML.files that list the names of open source software, their licenses, and required attributions.

Contact Us

Global Headquarters Pentaho Corporation Citadel International, Suite 340

5950 Hazeltine National Drive Orlando, FL 32822

Phone: +1 407 812-OPEN (6736)

Fax: +1 407 517-4575

http://www.pentaho.com

Sales Inquiries: sales@pentaho.com

New Features in Pentaho Data Integration 5.1

Pentaho Data Integration 5.1 delivers many exciting and powerful features that help you quickly and securely access, blend, transform, and explore data.

Data Science Pack with Weka and R

The **R Script Executor**, **Weka Forecasting**, and **Weka Scoring** steps form the core of the Data Science Pack and transforms PDI into a powerful, predictive analytics tool. The **R Script Executor** step, which is new for 5.1, lets you include R scripts in your transformations and jobs. You can customize random seed sampling, limit the batch and reservoir size, adjust logging level messages, and more. You can also choose to load the script from a file at runtime, enabling you to have more flexibility in transformation design.

Related Content:

- R Script Executor Step
- Weka Forecasting and Weka Scoring Steps

YARN Hadoop Distribution Support

PDI includes support for YARN capabilities including enhanced scalability, compatibility with MapReduce, improved cluster use, and greater support for agile development processes. YARN also provides support for non-MapReduce workloads.

Cloudera, Hortonworks, and MapR Hadoop Distribution Support

Use Pentaho's innovative Big Data Adaptive Layer to connect to more Hadoop Distributions, including Cloudera 5, Hortonworks 2.1, and MapR 3.1. These certified and tested YARN-based distributions allow you to use PDI to build scalable solutions that are optimized for performance. Pentaho supports over 20 different Hadoop Distribution versions from vendors such as Apache, Hortonworks, Intel, and MapR. Pentaho also supports Cloudera distributions and is a certified Cloudera partner.

Related Content:

- Configuring Pentaho for Your Hadoop Distribution and Version

YARN for Carte Kettle Clusters

The **Start a YARN Kettle Cluster** and **Stop a YARN Kettle Cluster** entries make it possible to to execute carte transforms in parallel using YARN. Carte clusters are implemented using the resources and data nodes of a Hadoop cluster, which optimizes resources and speeds processing.

Related Content:

- YARN Distribution Configuration
- Start a YARN Kettle Cluster
- Stop a YARN Kettle Cluster

Updated Support for Cassandra and MongoDB

PDI 5.1 provides support for newer versions of Cassandra and MongoDB.

Related Content:

- Pentaho Component Reference
- Cassandra 2.0 Release Notes
- MongoDB 2.6 Release Notes

Security Enhancements

PDI security has been enhanced to include support for more support for standard security protocols and specifications.

AES Password Support

Use Advanced Encryption Standard (AES) to encrypt passwords for databases, slave servers, web service connections, and more. AES uses a symmetric key algorithm to secure your data.

Related Content:

- Apply AES Passwords Encryption

New Execute Permission

You can now choose whether to grant permission to execute transformations and jobs by user role. This provides more finely-tuned access controls for different groups and can be useful for auditing, deployment, or quality assurance purposes.

Related Content:

- Assign Permissions to Use or Manage Database Connections.

Kerberos Security Support

If you are already using Kerberos to authenticate access a data source, with a little extra configuration, you can also use Kerberos to authenticate DI users who need to access your data.

Related Content:

- Implement Kerberos Authentication

Impersonation Support

If your transformation or job must run on a MapR cluster or access its resources, you can use impersonation to specify that another Hadoop user will run transformations or jobs on behalf of the default admin account. Impersonation is useful because it leverages another Hadoop user's existing authentication and authorization settings.

Related Content:

- Use Impersonation to Access a MapR Cluster

Teradata and Vertica Bulkloaders

There are two new bulkloaders steps: Teradata Insert/Upsert TPT Bulkloader and Vertica Bulkloader. Also, newer versions of Teradata and Vertica are now supported.

Related Content:

- Teradata Insert/Upsert TPT Bulkloader
- Vertica Bulkloader

JBoss Platform Support

Deploy PDI on your existing JBoss web application server or a new one. You can also choose whether to store house the DI Repository on a PostgreSQL, MySQL, or Oracle database.

Related Content:

- DI Manual Installation Guide

New Marketplace Plugins

Pentaho Marketplace continues to grow with many more of your contributions. As a testament to the power of community, Pentaho Marketplace is a home for your plugins and a place where you can contribute, learn, benefit from, and connect to others. New contributions include:

- **Vizor:** A realtime monitoring and debugging tool for transforms that run in the Hadoop cluster. Vizor helps you to more easily troubleshoot your transformations and jobs.
- **Riak Consumer and Producer:** Links with Maven to provide dependency management.
- **Load Text From File:** Uses Apache Tika to extract text from files in many different formats, such as PDF and XLS.
- **Top / Bottom / First / Last filter:** Filters rows based on a field's values or row numbers.
- **Map (key/value pair) type:** Provides a ValueMeta plugin for key/value pairs that are backed by a java.util.Map.
- **PDI Tableau Data Extract Output:** Use Pentaho's ETL capabilities to generate a Tableau Data Extract (tde) file.
- **PDI NuoDB:** Provides a PDI database dialect for the NuoDB NewSQL database that works in the cloud.
- **Neo4j JDBC Connector:** Provides a PDI database dialect for the Neo4j graph database.

- **HTML to XML:** Uses JTidy to convert HTML into XML or XHTML.
- **Apache Kafka Consumer and Producer:** Reads and sends binary messages to and from Apache Kafka message queues.
- **LegStar z/OS File Reader:** Reads raw z/OS records from a file and transforms them to PDI rows.
- **Compare Fields:** Compares 2 sets of fields in a row and directs it to the appropriate destination step based on the result. This step detects identical, changed, added, and removed rows.

There are many more new plugins such as **IC AMQP**, **IC Bloom filter**, **JaRE (Java Rule Engine)**, **JDBC Metadata**, **Memcached Input/Output**, **Google Spreadsheet Input/Output**, and **Sentiment Analysis**.

Related Content:

- PDI Marketplace
- Pentaho Marketplace Community Site

Documentation Changes

Our documentation has been moved to a new platform that is easier to use, search, and maintain. There are three new guides. Some documentation has been co-located so that it is easier to find.

- *Define DI Server Advanced Security* explains how to configure LDAP, LDAP/JDBC Hybrid Configuration, Active Directory, and Kerberos for DI use.
- *Troubleshoot DI Server Issues* helps you troubleshoot common DI component issues.
- *DI Manual Installation* explains how to install DI on Tomcat or JBoss with the DI Repository of your choice.
- Transformation Steps and Job Entries have been moved to the Pentaho Data Integration wiki.
- Big Data Configuration information has been moved to the Pentaho Big Data wiki.

Minor Functionality Changes

To learn more about minor functionality changes that might impact your upgrade experience, see the PDI 5.0 to 5.1 Functionality Change article.

PDI Version 5.1 Minor Functionality Changes

The following table describes minor functionality changes for PDI version 5.1. Provisions to preserve backward compatibility and related Jira cases are also listed.

Description	Related Jira Cases
In 5.0, the **Checks if files exist**, **Check if a folder is empty**, and **Evaluate rows number in a table** job entries fail when they are scheduled to run because the number of errors were set internally. In 5.1, the error flag is no longer set, but if you need it to be, set **KETTLE_COMPATIBILITY_SET_ERROR_ON_SPECIFIC_JOB_ENTRIES** to **Y**.	• PDI-10270 • PDI-10644
In 5.0, the strict checking of numbers was enabled by default when numbers were converted from a string. While this was not enforced for integer parsing in 5.0, but is enforced for integer parsing in 5.1. To be backward compatible, set **KETTLE_LENIENT_STRING_TO_NUMBER_CONVERSION** to **Y**.	• PDI-10560 • PDI-6186 • PDI-8974
When using the aggregate functions **MIN** and **SUM** in the **Group by**, **Memory Group by** or **Row Denormaliser** steps, the behavior of including **NULL** values into the aggregate has changed. To support backward compatibility, there are two new variables. • Set **KETTLE_AGGREGATION_MIN_NULL_IS_VALUED** to **Y** to set the minimum to **NULL** if NULL is within an aggregate. Otherwise by default **NULL** is ignored by the **MIN** aggregate and **MIN** is set to the minimum value that is not NULL. • Set **KETTLE_AGGREGATION_ALL_NULLS_ARE_ZERO** to Y to return 0 when all values within an aggregate are NULL. Otherwise by default NULL is returned when all values are NULL.	• PDI-6960 • PDI-9662 • PDI-10250 • PDI-11530
In 5.0, the time zone for the DATE data type was ignored. This has been addressed in 5.1. But, if you want the time zone for the DATE data type to continue to be ignored, set **KETTLE_COMPATIBILITY_DB_IGNORE_TIMEZONE** to **Y.**	• PDI-10749
The default conversion format for Timestamps is **yyyy/MM/dd HH:mm:ss.SSSSSSSSS**. Modify the value of the **KETTLE_DEFAULT_TIMESTAMP_FORMAT** variable if you need to change the default conversion format.	• PDI-11439

New Database Connections set the option **Preserve case of reserved words** to **True** by default. By setting this option by default, reserved words are no longer changed to upper or lower case.	• PDI-12123 • PDI-7893