



Create DI Solutions

Copyright Page

This document supports Pentaho Business Analytics Suite 5.2 GA and Pentaho Data Integration 5.2 GA, documentation revision October 7, 2014, copyright © 2014 Pentaho Corporation. No part may be reprinted without written permission from Pentaho Corporation. All trademarks are the property of their respective owners.

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 460

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

Introduction

Pentaho Data Integration (PDI) is a flexible tool that allows you to collect data from disparate sources such as databases, files, and applications, and turn the data into a unified format that is accessible and relevant to end users. PDI provides the Extraction, Transformation, and Loading (ETL) engine that facilitates the process of capturing the right data, cleansing the data, and storing the data using a uniform and consistent format.

PDI provides support for slowly changing [dimensions](#), and surrogate key for data warehousing, allows data migration between databases and application, is flexible enough to load giant datasets, and can take full advantage of cloud, clustered, and massively parallel processing environments. You can cleanse your data using transformation steps that range from very simple to very complex. Finally, you can leverage ETL as the data source for Pentaho Reporting.

Note: **Dimension** is a data warehousing term that refers to logical groupings of data such as product, customer, or geographical information. **Slowly Changing Dimensions** (SCD) are dimensions that contain data that changes slowly over time. For example, in most instances, employee job titles change slowly over time.

Common Uses of Pentaho Data Integration Include:

- Data migration between different databases and applications
- Loading huge data sets into databases taking full advantage of cloud, clustered and massively parallel processing environments
- Data Cleansing with steps ranging from very simple to very complex transformations
- Data Integration including the ability to leverage real-time ETL as a data source for Pentaho Reporting
- Data warehouse population with built-in support for slowly changing dimensions and surrogate key creation (as described above)

Audience and Assumptions

This section is written for IT managers, database administrators, and Business Intelligence solution architects who have intermediate to advanced knowledge of ETL and Pentaho Data Integration Enterprise Edition features and functions.

You must have installed Pentaho Data Integration to examine some of the step-related information included in this document.

If you are novice user, Pentaho recommends that you start by following the exercises in *Getting Started with Pentaho Data Integration* available in the Pentaho InfoCenter. You can return to this document when you have mastered some of the basic skills required to work with Pentaho Data Integration.

What this Section Covers

This document provides you with information about the most *commonly used* steps. For more information about steps, see [Matt Caster's blog](#) and the [Pentaho Data Integration wiki](#).

Refer to [Administer DI Server](#) for information about administering PDI and configuring security.

Pentaho Data Integration Architecture

PDI Components

Spoon is the design interface for building ETL jobs and transformations. Spoon provides a drag-and-drop interface that allows you to graphically describe what you want to take place in your transformations. Transformations can then be executed locally within Spoon, on a dedicated Data Integration Server, or a cluster of servers.

The **Data Integration Server** is a dedicated ETL server whose primary functions are:

Execution	Executes ETL jobs and transformations using the Pentaho Data Integration engine
Security	Allows you to manage users and roles (default security) or integrate security to your existing security provider such as LDAP or Active Directory
Content Management	Provides a centralized repository that allows you to manage your ETL jobs and transformations. This includes full revision history on content and features such as sharing and locking for collaborative development environments.
Scheduling	Provides the services allowing you to schedule and monitor activities on the Data Integration Server from within the Spoon design environment.

Pentaho Data Integration is composed of the following primary components:

- **Spoon.** Introduced earlier, Spoon is a desktop application that uses a graphical interface and editor for transformations and jobs. Spoon provides a way for you to create complex ETL jobs without having to read or write code. When you think of Pentaho Data Integration as a product, Spoon is what comes to mind because, as a database developer, this is the application on which you will spend most of your time. Any time you author, edit, run or debug a transformation or job, you will be using Spoon.
- **Pan.** A standalone command line process that can be used to execute transformations and jobs you created in Spoon. The data transformation engine Pan reads data from and writes data to various data sources. Pan also allows you to manipulate data.
- **Kitchen.** A standalone command line process that can be used to execute jobs. The program that executes the jobs designed in the Spoon graphical interface, either in XML or in a database repository. Jobs are usually scheduled to run in batch mode at regular intervals.
- **Carte.** Carte is a lightweight Web container that allows you to set up a dedicated, remote ETL server. This provides similar remote execution capabilities as the Data Integration Server, but does not provide scheduling, security integration, and a content management system.

What's with all the Culinary Terms?

If you are new to Pentaho, you may sometimes see or hear Pentaho Data Integration referred to as, "Kettle." To avoid confusion, all you must know is that Pentaho Data Integration began as an open source project called, "Kettle." The term, K.E.T.T.L.E is a recursive that stands for **Kettle Extraction Transformation Transport Load Environment**. When Pentaho acquired Kettle, the name was changed to **Pentaho Data Integration**. Other PDI components such as Spoon, Pan, and Kitchen, have names that were originally meant to support a "restaurant" metaphor of ETL offerings.

Create DI Repository Connections

There are several tasks that must be done first before following these tutorials. These are the tasks that must be done first.

- Your administrator must have [installed Pentaho Data Integration](#) and configured the DI server and its client tools as described in [Configure the DI Server](#) and [Configure the PDI Tools and Utilities](#).
- You must also [start the DI server](#) and [login to Spoon](#).

Create a Connection to the DI Repository

You need a place to store your work. We call this place the DI Repository. Your administrator may have created a connection to the DI repository during the configuration process. If you need to make another repository connection or if your administrator did not create a connection to the DI repository, you can create the connection.

1. Click on **Tools > Repository > Connect**.
2. If you have unsaved files open and you've made modifications, you are prompted to save them. Click **OK** to dismiss the message, then save the files and try to connect to the repository again.
3. The **Repository Connection** dialog box appears.
4. In the **Repository Connection** dialog box, click the add button (+).
5. Select **DI Repository: DI Repository** and click **OK**. The **Repository Configuration** dialog box appears.
6. Enter the URL associated with your repository. Enter an ID and name for your repository.
7. Click **Test** to ensure your connection is properly configured. If you see an error message, make sure you started your [DI server is started](#) and that the **Repository URL** is correct.
8. Click **OK** to exit the **Success** dialog box.
9. Click **OK** to exit the Repository Configuration dialog box. Your new connection appears in the list of available repositories.
10. Select the repository, type your user name and password, and click **OK**.
11. If you do not have files open, the process for connecting to the repository completes. If you have files open, a message appears that varies depending on your permissions.
 - *Would you like to close all open files?* This message appears if you have Create, Read, and Execute permissions. You can choose to close open transformation and job files or to leave them open. Click **Yes** or **No**.
 - *You have limited permissions. Some functionality may not be available to you. Would you like to close all open files now?* This message appears if you have Read and Create permissions. You can choose to close open transformation and job files or to leave them open. Click **Yes** or **No**.
 - *You have limited permissions. Some functionality may not be available to you. All open files will be closed.* This message appears if you have Read and Execute, or only Read permissions. You must close all open transformation and job files to continue. Click **OK**.

NOTE:

For more information on permissions, including what functionality is available for each, see [Use Pentaho Security on the DI Server](#).

Disconnect from the DI Repository

To disconnect from the DI Repository, complete these steps.

1. Save all open files.
2. Select **Tools > Repository > Disconnect Repository**.
3. If you have any unsaved, modified files open, you are prompted to save them.
4. A message appears. The content of the message depends on the permissions that you have.
 - *Would you like to close all open files?* This message appears if you have Create, Read, *and* Execute permissions. You can choose to close open transformation or jobs files, or to leave them open. Click **Yes** or **No**.
 - *All open files will be closed.* This message appears if you do *not* have all of the permissions (Create, Read, *and* Execute). Clicking the **OK** button closes all open files.

Interface Perspectives

The **Welcome** page contains useful links to documentation, community links for getting involved in the Pentaho Data Integration project, and links to blogs from some of the top contributors to the Pentaho Data Integration project.

Close the Welcome Page to proceed to Spoon.

- [Use Perspectives Within Spoon](#)
- [Tour Spoon](#)
- [Model Perspective](#)
- [Visualization Perspective](#)
- [Instaview Perspective](#)
- [Customize the Spoon Interface](#)

Use Perspectives Within Spoon

Pentaho Data Integration (PDI) provides you with tools that include ETL, modeling, and visualization in one unified environment — the Spoon interface. This integrated environment allows you to work in close cooperation with business users to build business intelligence solutions more quickly and efficiently.

File:/pdi_etl_perspectives.png

When you are working in Spoon you can *change perspectives*, or switch from designing ETL jobs and transformations to modeling your data, and visualizing it. As users provide you with feedback about how the data is presented to them, you can quickly make iterative changes to your data directly in Spoon by changing perspectives. The ability to quickly respond to feedback and to collaborate with business users is part of the Pentaho Agile BI initiative.

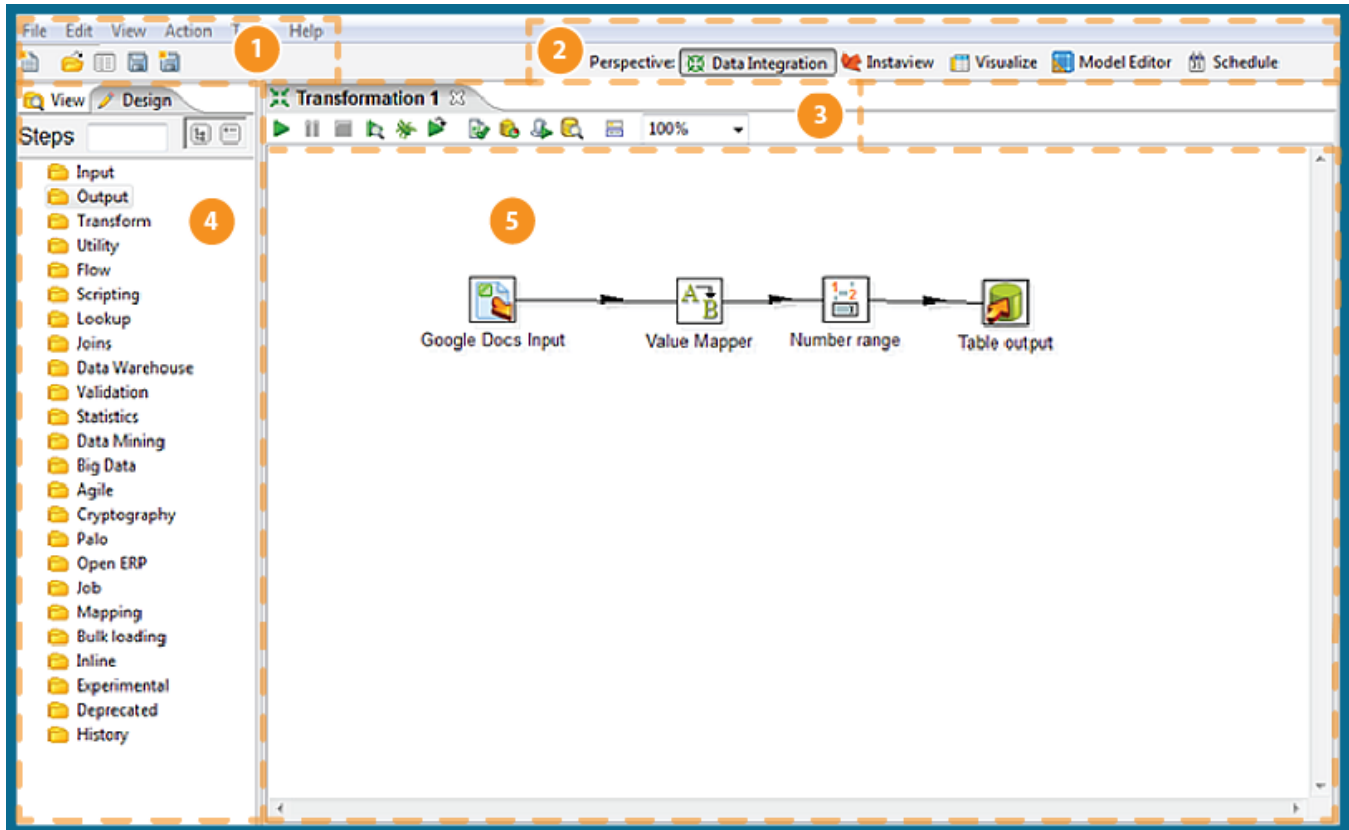
From within Spoon you can change perspectives using the **Perspective** toolbar located in the upper-right corner.

File:/pdi_perspective_toolbar_full.png

The perspectives in PDI enable you to focus how you work with different aspects of data.

- [Data Integration perspective](#)—Connect to data sources and extract, transform, and load your data
- [Model perspective](#)—Create a metadata model to identify the relationships within your data structure
- [Visualize perspective](#)—Create charts, maps, and diagrams based on your data
- [Instaview perspective](#)—Create a data connection, a metadata model, and analysis reports all at once with a dialog-guided, template-based reporting tool
- **Schedule perspective**—Plan when to run data integration jobs and set timed intervals to automatically send the output to your preferred destinations


















Tour Spoon



Component Name	Name	Function
1	Toolbar	Single-click access to common actions such as create a new file, opening existing documents, save and save as.
2	Perspectives Toolbar	<p>Switch between the different perspectives.</p> <ul style="list-style-type: none"> • Data Integration — Create ETL transformations and jobs • Instaview — Use pre-made templates to create visualizations from PDI transformations • Visualize — Test reporting and OLAP metadata models created in the Model perspective using the Report

Component Name	Name	Function
		<p>Design Wizard and Analyzer clients</p> <ul style="list-style-type: none"> • Model Editor — Design reporting and OLAP metadata models which can be tested right from within the Visualization perspective or published to the Pentaho BA Server • Schedule — Manage scheduled ETL activities on the Data Integration Server
3	Sub-toolbar	Provides buttons for quick access to common actions specific to the transformation or job such as Run , Preview , and Debug .
4	Design and View Tabs	<p>The Design tab of the Explore pane provides an organized list of transformation steps or job entries used to build transformations and jobs. Transformations are created by simply dragging transformation steps from the Design tab onto the canvas and connecting them with hops to describe the flow of data.</p> <p>The View tab of the Explore pane shows information for each job or transformation. This includes information such as available database connections and which steps and hops are used.</p> <p>In the image, the Design tab is selected.</p>
5	Canvas	Main design area for building transformations and jobs describing the ETL activities you want to perform

Table 1. Spoon Icon Descriptions

Icon	Description
	Create a new job or transformation
	Open transformation/job from file if you are not connected to a repository or from the repository if you are connected to one
	Explore the repository
	Save the transformation/job to a file or to the repository
	Save the transformation/job under a different name or file name (Save as)
	Run transformation/job; runs the current transformation from XML file or repository
	Pause transformation
	Stop transformation
	Preview transformation: runs the current transformation from memory. You can preview the rows that are produced by selected steps.
	Run the transformation in debug mode; allows you to troubleshoot execution errors
	Replay the processing of a transformation
	Verify transformation
	Run an impact analysis on the database
	Generate the SQL that is needed to run the loaded transformation.
	Launch the database explorer allowing you to preview data, run SQL queries, generate DDL and more
	Hide execution results pane
	Lock transformation

- [VFS File Dialogues in Spoon](#)

VFS File Dialogues in Spoon

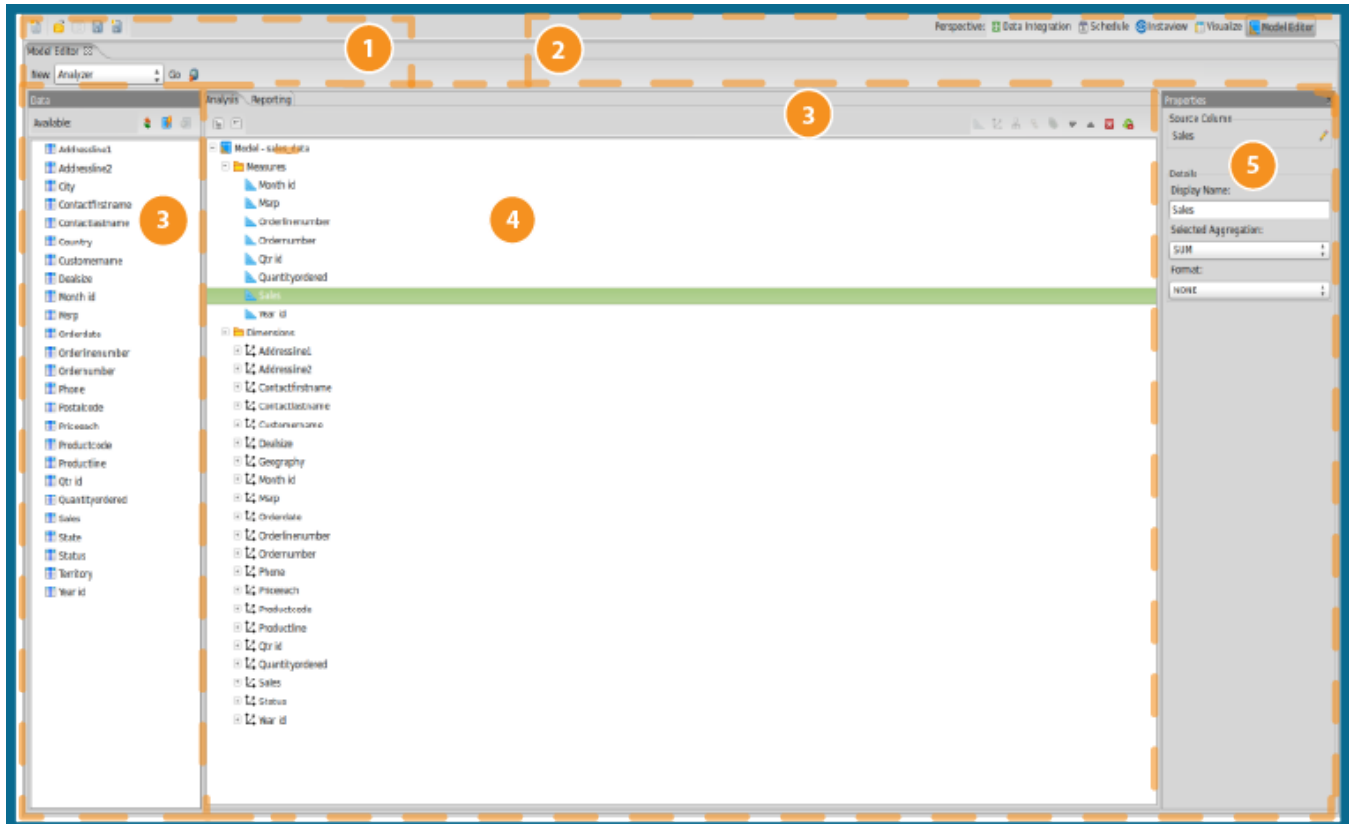
Some job and transformation steps have virtual filesystem (VFS) dialogues in place of the traditional local filesystem windows. VFS file dialogues enable you to specify a VFS URL in lieu of a typical local path. The following PDI and PDI plugin steps have such dialogues:

- File Exists
- Mapping (sub-transformation)
- ETL Meta Injection
- Hadoop Copy Files
- Hadoop File Input
- Hadoop File Output

Note: VFS dialogues are configured through certain transformation parameters. Refer to [Configure SFTP VFS](#) for more information on configuring options for SFTP.

Model Perspective

The **Model** perspective is used for designing reporting and OLAP metadata models that can be tested from within the **Visualize** perspective or published to the Pentaho BA Server.

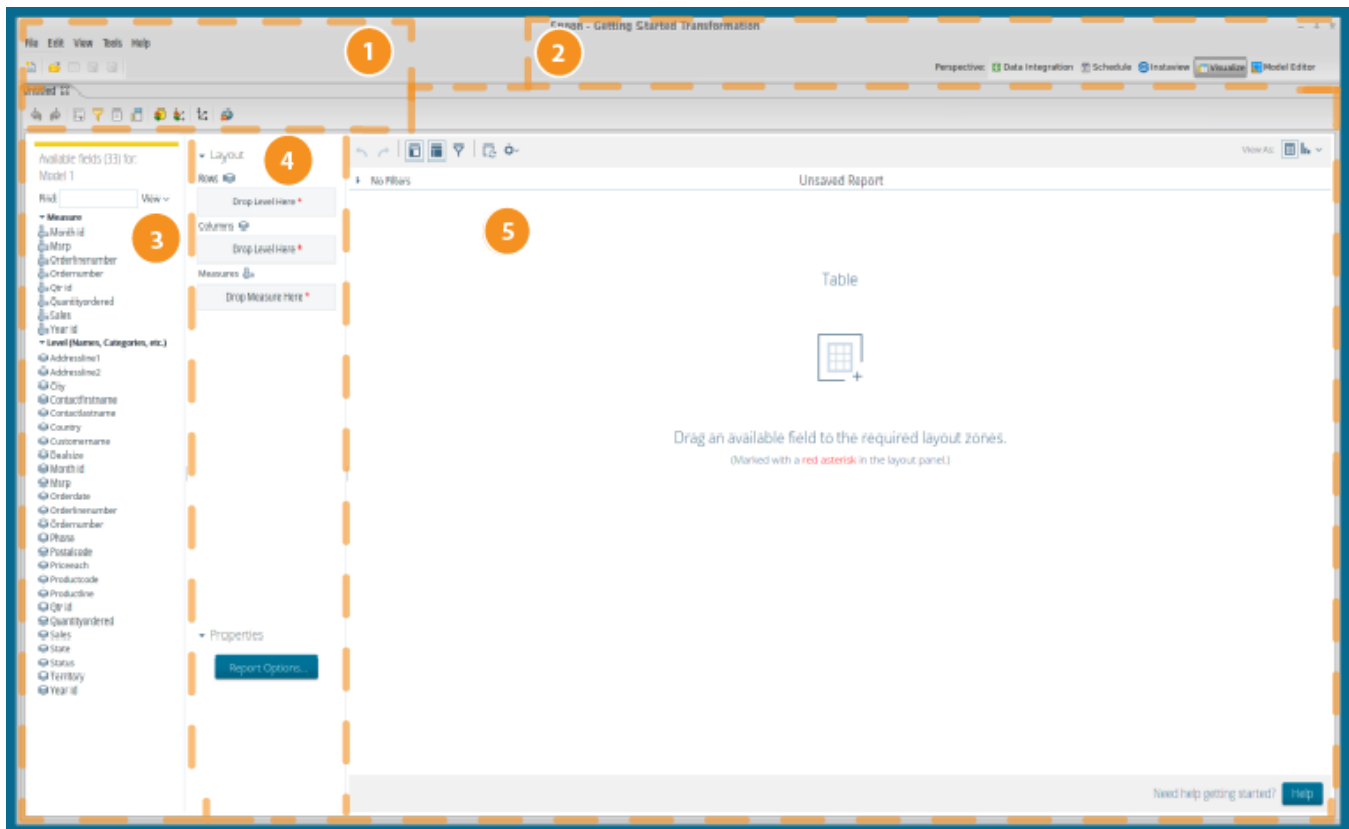


Component Name	Description
1-Menubar	The Menubar provides access to common features such as properties, actions and tools.
2-Main Toolbar	The Main Toolbar provides single-click access to common actions such as create a new file, opening existing documents, save and save as. The right side of the main toolbar is also where you can switch between perspectives.
3-Data Panel	Contains a list of available fields from your data source that can be used either as measure or dimension levels (attributes) within your OLAP dimensional model.

Component Name	Description
4- Model Panel	Used to create measures and dimensions of your Analysis Cubes from the fields in the data panel. Create a new measure or dimension by dragging a field from the data panel over onto the Measures or Dimension folder in the Model tree.
5-Properties Panel	Used to modify the properties associated with the selection in the Model Panel tree.

Visualization Perspective

The **Visualize** perspective allows you to test reporting and OLAP metadata models created in the **Model** perspective using the Report Design Wizard and Analyzer clients respectively.



Component Name	Description
1-Menubar	The Menubar provides access to common features such as properties, actions, and tools.
2-Main Toolbar	The Main Toolbar provides single-click access to common actions such as create a new file, opening existing documents, save and save as. The right side of the main toolbar is also where you can switch between perspectives.
3-Field List	Contains the list of measures and attributes as defined in your model. These fields can be dragged into the Report Area to build your query.

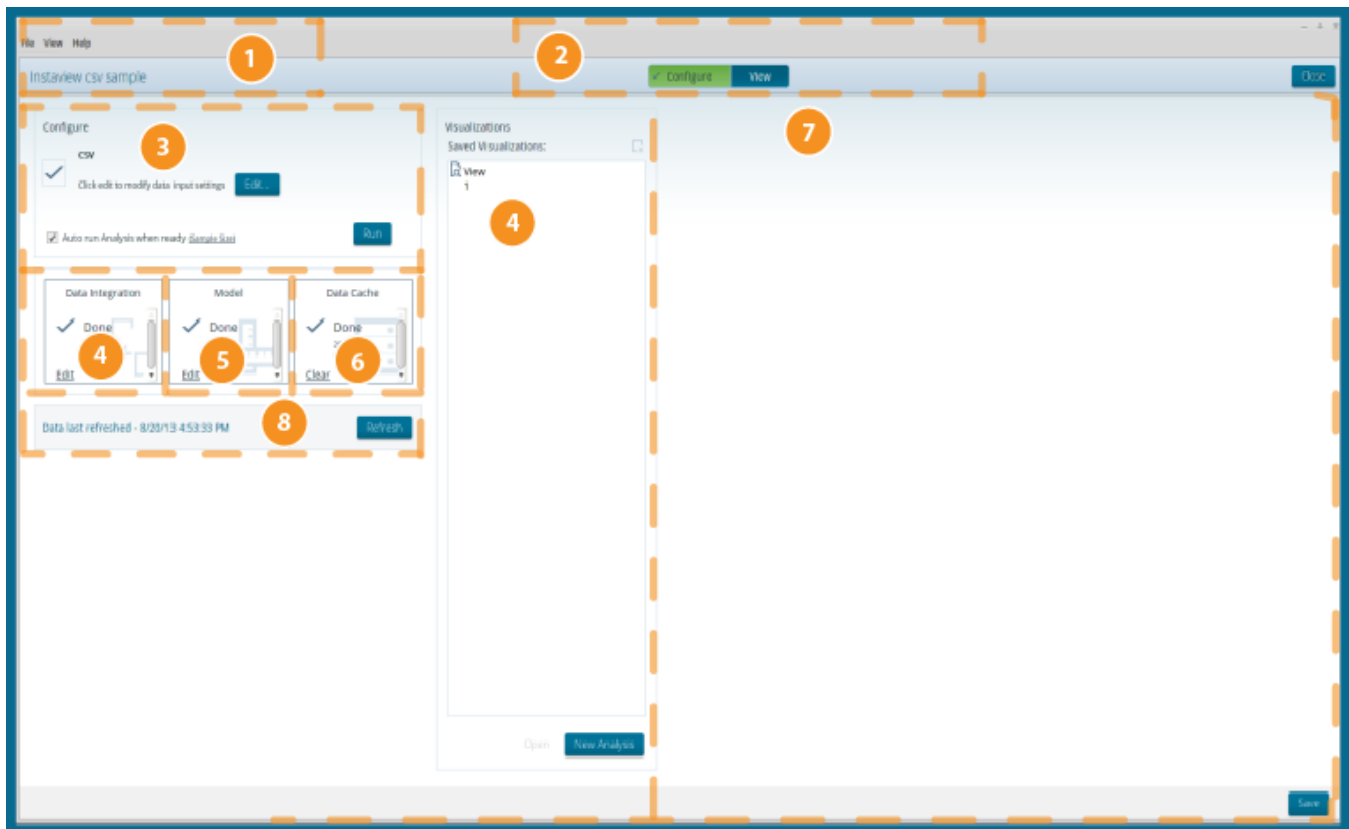
Component Name	Description
4-Layout	Allows you to drag Measures and Levels into the Row , Column , and Measures area so you can control how it appears in the workspace.
5-Canvas	Drag fields from the field list into the Report Area to build your query. Right click on a measure or level to further customize your report with sub-totals, formatting, and more.

Instaview Perspective

With Instaview, you can access, transform, analyze, and visualize data without having extensive experience designing business analytics solutions or staging databases. Instaview gives you immediate access to your data so you can quickly explore different ways to structure and present your data as a complete business analytics solution. In addition to extracting and loading the data, Instaview gives you the ability to manipulate the data to make it fit your specific needs from within one simple tool.

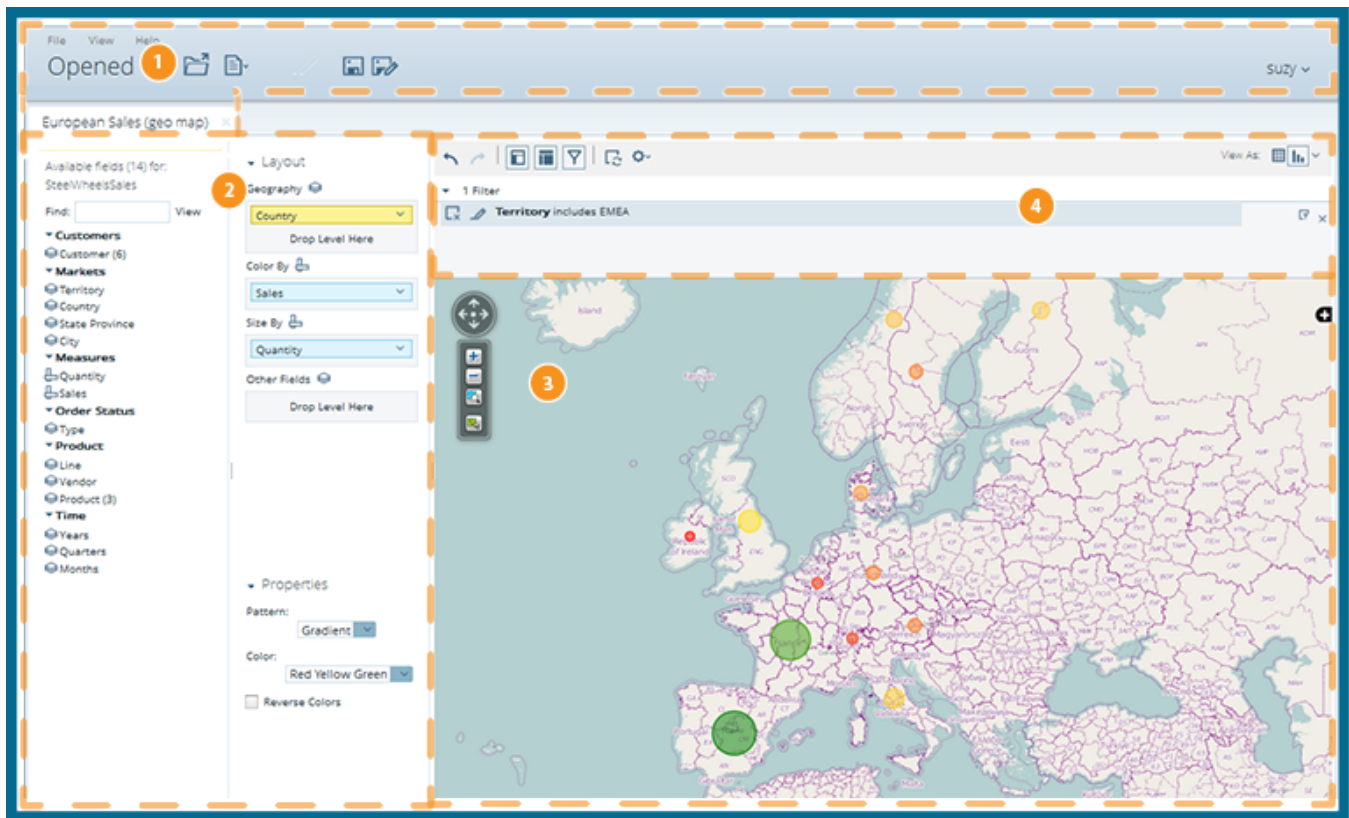
When you create an Instaview you

- Create a new data source from which to extract and transform your data.
- Create a data model to define how columns and fields relate to one-another.
- Create an Analyzer Report with tables and charts from your transformed data.



Component	Description
1 - Instaview	A combination of a valid data connection, a data integration transformation, a metadata data source template, and one or more Analyzer reports. You can only have one Instaview at a time.

Component	Description
2 - Configure View	<p>The Configure/View mode toggle allows you to switch between Cofigure mode and View mode.</p> <ul style="list-style-type: none"> • Configure mode enables you edit a data connection, data integration transformation, metadata data source template, and Analyzer report. It also provides the means to clear the Data Cache. • View mode enables you to create reports and visualizations from a valid Instaview data source. From within this view you can drag and drop fields from (measurements or dimensions) your data onto the Reporting canvas.
3 - Configure data source panel	<ul style="list-style-type: none"> • The Edit button takes you to the data connection dialog and allows you to edit the data connection settings for the current Instaview. • The Auto run Analysis when ready option, if checked, will automatically create a new Analyzer report after pressing Run. • The Run button lets you manually start the Instaview data transformation. Pressing Run will modify the data integration transformation or metadata model if changes were made within the Configure panel, if necessary.
4 - Data Integration panel	<p>Provides the means to access and edit the data integration transformation for the current Instaview. Editing will open the Data Integration perspective in PDI.</p>
5 - Model panel	<p>Enables you to edit the metadata model for the current Instaview. Editing will open the Model perspective in PDI.</p>
6 - Data Cache panel	<p>Provides the means to clear the data cache.</p>
7 - Visualizations panel	<p>Displays existing Views and provides the means to open existing, create new, and delete Instaviews. You can also rename an existing visualization by right-clicking an item within this panel.</p>
8 - Refresh display	<p>Displays when the current Instaview was last run. If your data is connected to a live data source this displays the last time the data was accessed by Instaview.</p> <p>The Refresh button provides the means to manually refresh the current Instaview.</p>



Item	Name	Function
1	Opened view	Displays quick access buttons across the top to create and save new Analysis reports, Interactive reports, and Dashboards. Opened reports and files show as a series of tabs across the page.
2	Available Fields and Layout panels	<p>Use the Available Fields and Layout panels to drag levels and measures into a report.</p> <p>Your report displays changes in the Report Canvas as you drag items onto the Layout panel.</p> <p>Delete a level or measure from your report by dragging it from the Layout panel to the trashcan that appears in the lower right corner of the Report Canvas.</p>
3	Report Canvas	<p>Shows a dynamic view of your report as you work to build it. The look of your report changes constantly as you work with Available Fields and Layout panels to refine it.</p> <p>The Report Canvas shows different fields based on the chart type selected.</p>
4	Analyzer Toolbar and Filters	Use the Analyzer Toolbar functions to undo or redo actions, hide lists of fields, add or hide filters, disable the auto-refresh function, adjust settings, and change the view of your report.

Item	Name	Function
		Use the Filters panel to display a list of filters applied to the active report, or edit or delete filters.

Customizing the Spoon Interface

Kettle Options allow you to customize properties associated with the behavior and look and feel of the Spoon interface. Examples include startup options such as whether or not to display tips and the Welcome page, and user interface options such as fonts and colors. To access the options, in the menu bar, go to **Tools > Options...**

The tables below contain descriptions for options under the **General** and **Look & Feel** tabs, respectively. You may want to keep the default options enabled initially. As you become more comfortable using Pentaho Data Integration, you can set the options to better suit your needs.

General

Option	Description
Default number of lines in preview dialog	Sets the default number of lines that are displayed in the preview dialog box in Spoon
Maximum nr of lines in the logging windows	Specifies the maximum limit of rows to display in the logging window
Central log line store timeout in minutes	no def given
Max number of lines in the log history views	Specifies the maximum limit of line to display in the log history views
Show tips at startup?	Sets the display of tips at startup
Show welcome page at startup?	Controls whether or not to display the Welcome page when launching Spoon
Use database cache?	Spoon caches information that is stored on the source and target databases. In some instances, caching causes incorrect results when you are making database changes. To prevent errors you can disable the cache altogether instead of clearing the cache every time.
Open last file at startup?	Loads the last transformation you used (opened or saved) from XML or repository automatically
Auto save changed files?	Automatically saves a changed transformation before running
Only show the active file in the main tree?	Reduces the number of transformation and job items in the main tree on the left by only showing the currently active file
Only save used connections to XML?	Limits the XML export of a transformation to the used connections in that transformation. This is helpful while

Option	Description
	exchanging sample transformations to avoid having all defined connections to be included.
Ask about replacing existing connections on open/import?	Requests permission before replacing existing database connections during import
Replace existing connections on open/import?	This is the action that takes place when there is no dialog box shown, (see previous option)
Show Save dialog?	Allows you to turn off the confirmation dialogs you receive when a transformation has been changed
Automatically split hops?	Disables the confirmation messages that launch when you want to split a hop
Show copy or distribute dialog?	Disables the warning message that appears when you link a step to multiple outputs. This warning message describes the two options for handling multiple outputs: 1. Distribute rows - destination steps receive the rows in turns (round robin) 2. Copy rows - all rows are sent to all destinations
Show repository dialog at startup?	Controls whether or not the Repository dialog box appears at startup
Ask user when exiting?	Controls whether or not to display the confirmation dialog when a user chooses to exit the application
Clear custom parameters (steps/plugin-ins)	Clears all parameters and flags that were set in the plug-in or step dialog boxes.
Display tool tips?	Controls whether or not to display tool tips for the buttons on the main tool bar.

Look & Feel

Option	Description
Fixed width font	This option customizes the font that is used in the dialog boxes, trees, input fields, and more; click Edit to edit the font or Delete to return the font to its default value.
Font on workspace	This option customizes font that is used in the Spoon interface; click Edit to edit the font or Delete to return the font to its default value.
Font for notes	This option customizes the font used in notes that are displayed in Spoon; click Edit to edit the font or Delete to return the font to its default value.
Background color	This option sets the background color in Spoon and affects all dialog boxes; click Edit to edit the color or Delete to return the background color to its default value.

Option	Description
Workspace background color	This option sets the background color in the graphical view of Spoon; click Edit to edit the background color or Delete to return the background color to its default value.
Tab color	This option customizes the color that is being used to indicate tabs that are active/selected; click Edit to edit the tab color or Delete to return the color to its default value.
Icon size in workspace	Affects the size of the icons in the graph window. The original size of an icon is 32x32 pixels. The best results (graphically) are probably at sizes 16,24,32,48,64 and other multiples of 32.
Line width on workspace	Affects the line width of the hops in the Spoon graphical view and the border around the step.
Shadow size on workspace	If this size is larger than 0, a shadow of the steps, hops, and notes is drawn on the canvas, making it look like the transformation floats above the canvas.
Dialog middle percentage	By default, a parameter is drawn at 35% of the width of the dialog box, counted from the left. You can change using this option in instances where you use unusually large fonts.
Canvas anti-aliasing?	Some platforms like Windows, OSX and Linux support anti-aliasing through GDI, Carbon or Cairo. Enable this option for smoother lines and icons in your graph view. If you enable the option and your environment does not work, change the value for option "EnableAntiAliasing" to "N" in file \$HOME/.kettle/.spoonrc (C:\Documents and Settings\<user>\.kettle\.spoonrc on Windows)
Use look of OS?	Enabling this option on Windows allows you to use the default system settings for fonts and colors in Spoon. On other platforms, the default is always enabled.
Show branding graphics	Enabling this option will draw Pentaho Data Integration branding graphics on the canvas and in the left hand side "expand bar."
Preferred Language	Specifies the preferred language setting.
Alternative Language	Specifies the alternative language setting. Because the original language in which Pentaho Data Integration was written is English, it is best to set this locale to English.

Terminology and Basic Concepts

Before you can start designing transformations and jobs, you must have a basic understanding of the terminology associated with Pentaho Data Integration.

- [Transformations, Steps, and Hops](#)
- [Jobs](#)
- [More About Hops](#)

Transformations, Steps, and Hops

A **transformation** is a network of logical tasks called *steps*. Transformations are essentially *data flows*. In the example below, the database developer has created a transformation that reads a flat file, filters it, sorts it, and loads it to a relational database table. Suppose the database developer detects an error condition and instead of sending the data to a Dummy step, (which does nothing), the data is logged back to a table. The transformation is, in essence, a directed graph of a logical set of data transformation configurations. Transformation file names have a .ktr extension.



The two main components associated with transformations are **steps** and **hops**:

Steps are the building blocks of a transformation, for example a text file input or a table output. There are over 140 steps available in Pentaho Data Integration and they are grouped according to function; for example, input, output, scripting, and so on. Each step in a transformation is designed to perform a specific task, such as reading data from a flat file, filtering rows, and logging to a database as shown in the example above. Steps can be configured to perform the tasks you require.

Hops are data pathways that connect steps together and allow schema metadata to pass from one step to another. In the image above, it seems like there is a sequential execution occurring; however, that is not true. Hops determine the flow of data *through* the steps not necessarily the sequence in which they run. When you run a transformation, each step starts up in its own thread and pushes and passes data.

Note: All steps are started and run in parallel so the initialization sequence is not predictable. That is why you cannot, for example, set a variable in a first step and attempt to use that variable in a subsequent step.

You can connect steps together, edit steps, and open the step contextual menu by clicking



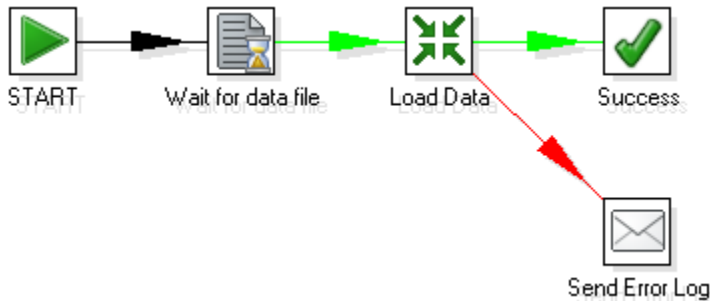
to edit a step. Click the down arrow to open the contextual menu. For information about connecting steps with hop, see [More About Hops](#).



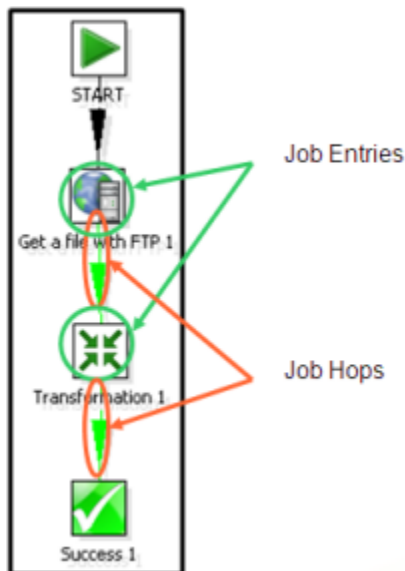
A step can have many connections — some join two steps together, some only serve as an input or output for a step. The data stream flows through steps to the various steps in a transformation. Hops are represented in Spoon as arrows. Hops allow data to be passed from step to step, and also determine the direction and flow of data through the steps. If a step sends outputs to more than one step, the data can either be copied to each step or distributed among them.

Jobs

Jobs are workflow-like models for coordinating resources, execution, and dependencies of ETL activities.



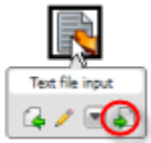
Jobs aggregate up individual pieces of functionality to implement an entire process. Examples of common tasks performed in a job include getting FTP files, checking conditions such as existence of a necessary target database table, running a transformation that populates that table, and e-mailing an error log if a transformation fails. The final job outcome might be a nightly warehouse update, for example.



Jobs are composed of **job hops**, **job entries**, and **job settings**. Hops behave differently when used in a job, see [More About Hops](https://help.pentaho.com/Documentation/5.2/OL0/OY0/040/000). Job entries are the individual configured pieces as shown in the example above; they are the primary building blocks of a job. In data transformations these individual pieces are called steps. Job entries can provide you with a wide range of functionality ranging from executing transformations to getting files from a Web server. A single job entry can be placed multiple times on the canvas; for example, you can take a single job entry such as a transformation run and place it on the canvas multiple times using different configurations. Job settings are the options that control the behavior of a job and the method of logging a job's actions. Job file names have a .kjb extension.

More About Hops

A hop connects one transformation step or job entry with another. The direction of the data flow is indicated by an arrow. To create the hop, click the source step, then press the <SHIFT> key down and draw a line to the target step. Alternatively, you can draw hops by hovering over a step until the hover menu appears. Drag the hop painter icon from the source step to your target step.



Additional methods for creating hops include:

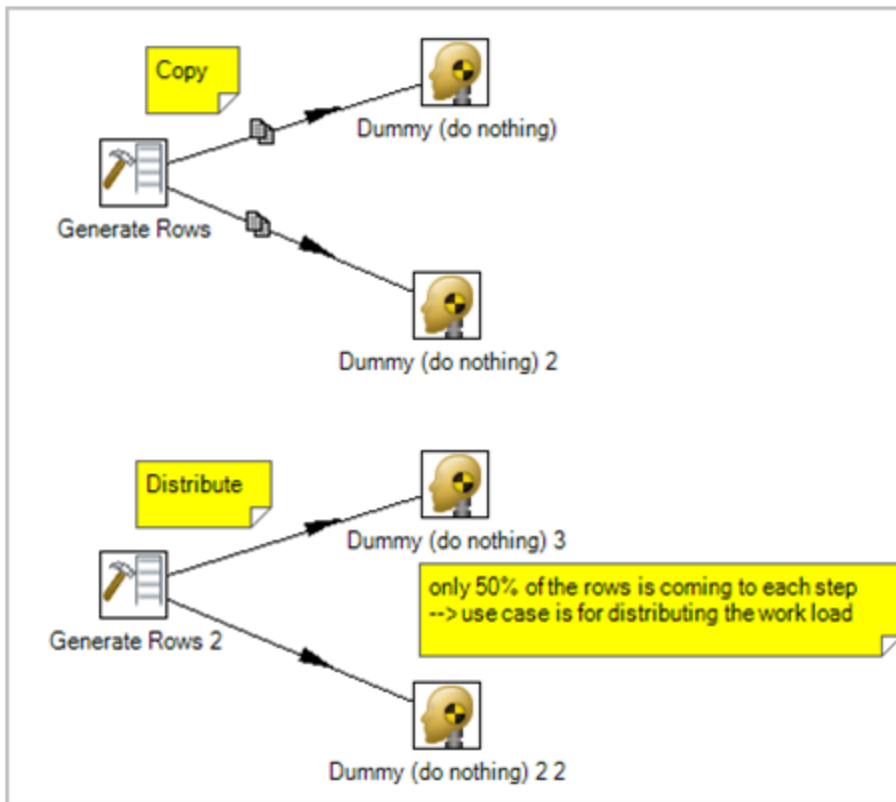
- Click on the source step, hold down the middle mouse button, and drag the hop to the target step.
- Select two steps, then choose New Hop from the right-click menu.
- Use <CTRL + left-click> to select two steps the right-click on the step and choose **New Hop**.

To **split a hop**, insert a new step into the hop between two steps by dragging the step over a hop. Confirm that you want to split the hop. This feature works with steps that have not yet been connected to another step only.

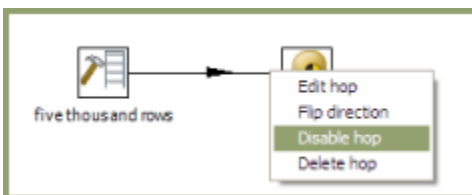
Loops are not allowed in transformations because Spoon depends heavily on the previous steps to determine the field values that are passed from one step to another. Allowing loops in transformations may result in endless loops and other problems. Loops are allowed in jobs because Spoon executes job entries sequentially; however, make sure you do not create endless loops.

Mixing rows that have a different layout is not allowed in a transformation; for example, if you have two table input steps that use a varying number of fields. Mixing row layouts causes steps to fail because fields cannot be found where expected or the data type changes unexpectedly. The trap detector displays warnings at design time if a step is receiving mixed layouts.

You can specify if data can either be **copied**, **distributed**, or **load balanced** between multiple hops leaving a step. Select the step, right-click and choose **Data Movement**.

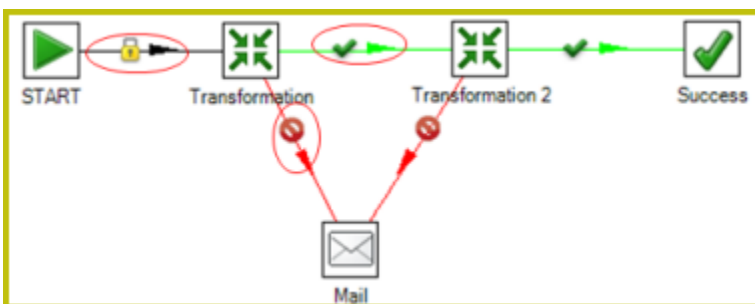


A hop can be enabled or disabled (for testing purposes for example). Right-click on the hop to display the options menu.



Job Hops

Besides the execution order, a hop also specifies the condition on which the next job entry will be executed. You can specify the **Evaluation** mode by right clicking on the job hop. A job hop is just a flow of control. Hops link to job entries and, based on the results of the previous job entry, determine what happens next.



Option	Description
Unconditional	Specifies that the next job entry will be executed regardless of the result of the originating job entry
Follow when result is true	Specifies that the next job entry will be executed only when the result of the originating job entry is true; this means a successful execution such as, file found, table found, without error, and so on
Follow when result is false	Specifies that the next job entry will only be executed when the result of the originating job entry was false, meaning unsuccessful execution, file not found, table not found, error(s) occurred, and so on

Create Transformations

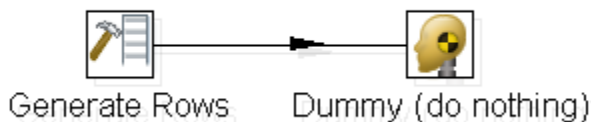
This exercise is designed to help you learn basic skills associated with handling steps and hops, running and previewing transformations. See [Get Started with DI](#) for a comprehensive, "real world" exercise for creating, running, and scheduling transformations.

- [Get Started](#)
- [Save Your Transformation](#)
- [Run Your Transformation Locally](#)
- [Build a Job](#)

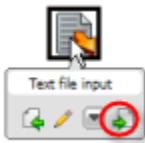
Get Started

Follow these instructions to begin creating your transformation.

1. Click **File > New > Transformation**.
2. Under [the Design tab](#), expand the **Input** node, then select and drag a **Generate Rows** step onto the canvas.
Note: If you don't know where to find a step, there is a search function in the left corner of Spoon. Type the name of the step in the search box. Possible matches appear under their associated nodes. Clear your search criteria when you are done searching.
3. Expand the **Flow** node; click and drag a **Dummy (do nothing)** step onto the canvas.
4. To connect the steps to each other, you must add a hop. Hops describe the flow of data between steps in your transformation. To create the hop, click the **Generate Rows** step, then press and hold the **<SHIFT>** key and draw a line to the **Dummy (do nothing)** step.



Note: Alternatively, you can draw hops by hovering over a step until the hover menu appears. Drag the hop painter icon from the source step to your target step.



5. Double click the **Generate Rows** step to open its edit properties dialog box.
6. In the **Limit** field, type 100000. This limits the number of generated rows to 100,000.
7. Under **Name**, type **FirstCol** in the **Name** field.
8. Under **Type**, type **String**.
9. Under **Length**, type **150**.
10. Under **Value**, type **My First Step**. Your entries should look like the image below. Click **OK** to exit the Generate Rows edit properties dialog box.

Generate Rows

Step name:

Limit:

Never stop generating rows: ☐

Interval in ms (delay):

Current row time field name:

Previous row time field name:

Fields :

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Value
1	FirstCol	String		150					My First Step

Help OK Preview Cancel

11. Now, save your transformation. See [Save Your Transformation](#).

Save Your Transformation

Follow the instructions below to save your transformation.

1. In Spoon, click **File > Save As**. The **Transformation Properties** dialog box appears.
2. In the **Transformation Name** field, type **First Transformation**.
3. In the **Directory** field, click the **Folder Icon** to select a repository folder where you will save your transformation.
4. Expand the **Home** directory and double-click the **admin** folder. Your transformation will be saved in the **admin** folder in the DI Repository.
5. Click **OK** to exit the **Transformation Properties** dialog box. The **Enter Comment** dialog box appears.
6. Click in the **Enter Comment** dialog box and press **<Delete>** to remove the default text string. Type a meaningful comment about your transformation. The comment and your transformation are tracked for version control purposes in the DI Repository.
7. Click **OK** to exit the **Enter Comment** dialog box and save your transformation.




Run Your Transformation Locally

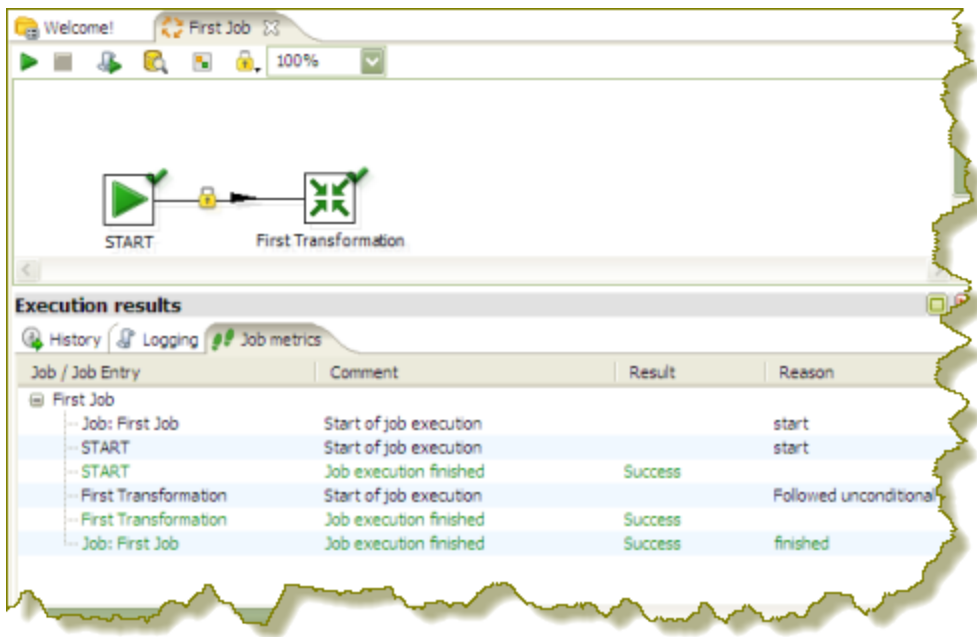
In [Get Started](#), you created a simple transformation. Now, you are going to run your transformation locally, which is a local execution. Local execution allows you to execute a transformation or job from within the Spoon on your local device. This is ideal for designing and testing transformations or lightweight ETL activities.

1. In Spoon, go to **File > Open**. The contents of the repository appear.
2. Navigate to the folder that contains your transformation. If you are a user with administrative rights, you may see the folders of other users.
3. Double-click on your transformation to open it in the Spoon workspace.
Note: If you followed the exercise instructions, the name of the transformation is **First Transformation**.
4. In the upper left corner of the workspace, click **Run**. The **Execute a Transformation** dialog box appears. Notice that **Local Execution** is enabled by default.
5. Click **Launch**. The **Execution Results** appear in the lower pane.
6. Examine the contents under **Step Metrics**. The Step Metrics tab provides statistics for each step in your transformation such as how many records were read, written, caused an error, processing speed (rows per second) and more. If any of the steps caused the transformation to fail, they are highlighted in red.
Note: Other tabs associated with Execution Results require additional set up. See [Performance Monitoring and Logging](#).

Build a Job

You created, saved, and ran your first transformation. Now, you will build a simple job. Use jobs to execute one or more transformations, retrieve files from a Web server, place files in a target directory, and more. Additionally, you can schedule jobs to run on specified dates and times. The section called [Get Started with DI](#) contains a "real world" exercise for building jobs.

1. In the Spoon menubar, go to **File > New > Job**. Alternatively click  (New) in the toolbar.
2. Click the **Design** tab. The nodes that contain job entries appear.
3. Expand the **General** node and select the **Start** job entry.
4. Drag the Start job entry to the workspace (canvas) on the right.
The Start job entry defines where the execution will begin.
5. Expand the **General** node, select and drag a **Transformation** job entry on to the workspace.
6. Use a hop to connect the Start job entry to the Transformation job entry.
7. Double-click on the **Transformation** job entry to open its properties dialog box.
8. Under **Transformation specification**, click **Specify by name and directory**.
9. Click  (Browse) to locate your transformation in the solution repository.
10. In the **Select repository object** view, expand the directories. Locate **First Transformation** and click **OK**. The name of the transformation and its location appear next to the **Specify by name and directory** option.
11. Under **Transformation specification**, click **OK**.
12. Save your job; call it **First Job**. Steps used to save a job are nearly identical to saving a transformation. Provide a meaningful comment when saving your job. See [Saving Your Transformation](#).
13. Click  (Run Job) in the toolbar. When the **Execute a Job** dialog box appears, choose **Local Execution** and click **Launch**.



The **Execution Results** panel opens displaying the job metrics and log information for the job execution.

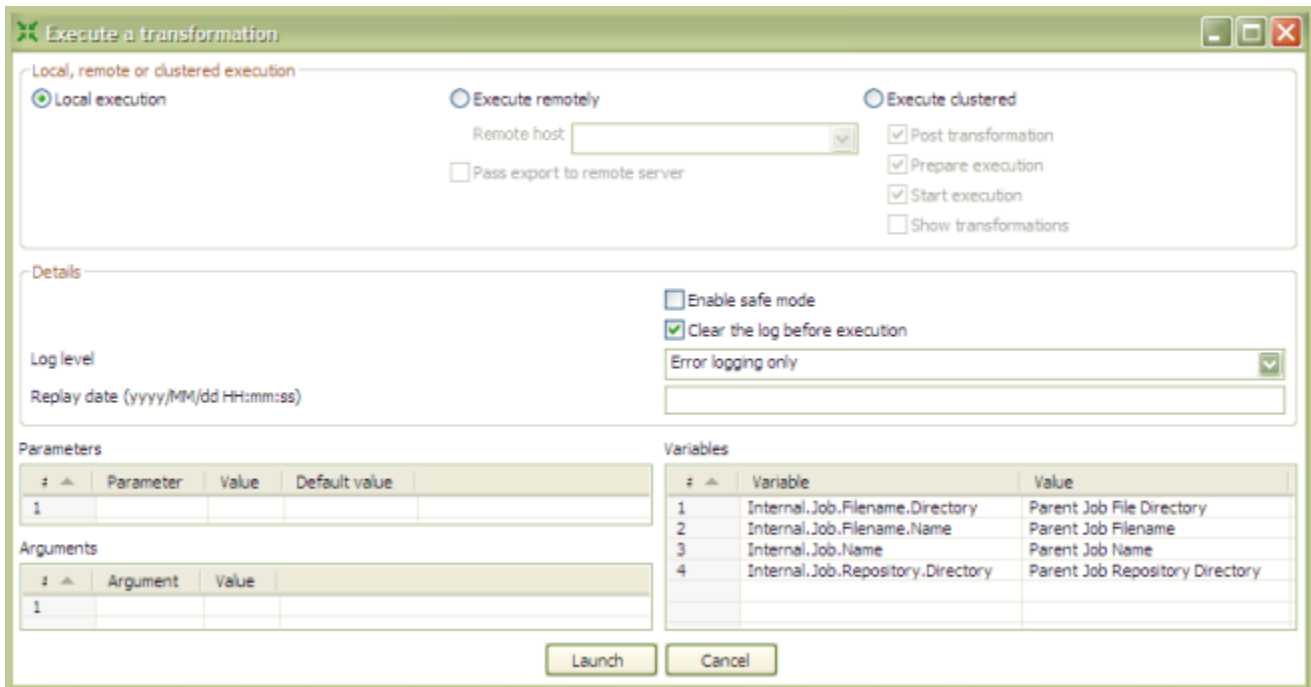
Executing Transformations

When you are done modifying a transformation or job, you can run it by clicking

../pdi_admin_guide/images/run.png

(Run) from the main menu toolbar, or by pressing F9. There are three options that allow you to decide where you want your transformation to be executed:

- **Local Execution** — The transformation or job executes on the machine you are currently using.
- **Execute remotely** — Allows you to specify a remote server where you want the execution to take place. This feature requires that you have the Data Integration Server running or Data Integration installed on a remote machine and running the Carte service. To use remote execution you first must set up a slave server (see [Setting Up a Slave Server](#)) .
- **Execute clustered** — Allows you to execute a transformation in a clustered environment.



The dialog box titled "Execute a transformation" contains the following sections:

- Local, remote or clustered execution:** Three radio buttons: "Local execution" (selected), "Execute remotely", and "Execute clustered".
 - Under "Execute remotely": "Remote host" (dropdown), "Pass export to remote server" (checkbox).
 - Under "Execute clustered": "Post transformation" (checked), "Prepare execution" (checked), "Start execution" (checked), "Show transformations" (checkbox).
- Details:**
 - "Enable safe mode" (checkbox).
 - "Clear the log before execution" (checked).
 - "Log level" dropdown: "Error logging only" (selected).
 - "Replay date (yyyy/MM/dd HH:mm:ss)" text field.
- Parameters:** Table with 5 columns: #, Parameter, Value, Default value, and an empty column.
- Arguments:** Table with 4 columns: #, Argument, Value, and an empty column.
- Variables:** Table with 3 columns: #, Variable, and Value.

Buttons: "Launch" and "Cancel".

#	Parameter	Value	Default value	
1				

#	Argument	Value	
1			

#	Variable	Value
1	Internal.Job.Filename.Directory	Parent Job File Directory
2	Internal.Job.Filename.Name	Parent Job Filename
3	Internal.Job.Name	Parent Job Name
4	Internal.Job.Repository.Directory	Parent Job Repository Directory

- [Initialize Slave Servers in Spoon](#)
- [Executing Jobs and Transformations from the Repository on the Carte Server](#)
- [Impact Analysis](#)

Initialize Slave Servers in Spoon

Follow the instructions below to configure PDI to work with Carte slave servers.

1. Open a transformation.
2. In the **Explorer View** in Spoon, select **Slave Server**.
3. Right-click and select **New**. The **Slave Server** dialog box appears.
4. In the Slave Server dialog box, enter the appropriate connection information for the Data Integration (or Carte) slave server. The image below displays a connection to the Data Integration slave server.

Option	Description
Server name	The name of the slave server
Hostname or IP address	The address of the device to be used as a slave
Port	Defines the port you are for communicating with the remote server
Web App Name	Used for connecting to the DI server and set to pentaho-di by default
User name	Enter the user name for accessing the remote server
Password	Enter the password for accessing the remote server
Is the master	Enables this server as the master server in any clustered executions of the transformation

Note: When executing a transformation or job in a clustered environment, you should have one server set up as the master and all remaining servers in the cluster as slaves.

Below are the proxy tab options:

Option	Description
Proxy server hostname	Sets the host name for the Proxy server you are using
The proxy server port	Sets the port number used for communicating with the proxy
Ignore proxy for hosts: regexp separated	Specify the server(s) for which the proxy should not be active. This option supports specifying multiple servers using regular expressions. You can also add multiple servers and expressions separated by the ' ' character.

5. Click **OK** to exit the dialog box. Notice that a plus sign (+) appears next to **Slave Server** in the Explorer View.

Executing Jobs and Transformations from the Repository on the Carte Server

To execute a job or transformation remotely on a Carte server, you first need to copy the local `repositories.xml` from the user's `.kettle` directory to the Carte server's `$HOME/.kettle` directory. The Carte service also looks for the `repositories.xml` file in the directory from which Carte was started.

For more information about locating or changing the `.kettle` home directory, see [Changing the Pentaho Data Integration Home Directory Location \(.kettle folder\)](#).

Impact Analysis

To see what effect your transformation will have on the data sources it includes, go to the **Action** menu and click on **Impact**. PDI will perform an impact analysis to determine how your data sources will be affected by the transformation if it is completed successfully.

Working with the DI Repository

In addition to storing and managing your jobs and transformations, the DI repository provides full revision history for documents allowing you to track changes, compare revisions and revert to previous versions when necessary. This, in combination with other features such as enterprise security and content locking make the DI repository an ideal platform for providing a collaborative ETL environment.

Note: If you prefer to manage your documents as loose files on the file system, click **Cancel** in the **Repository Connection** dialog box. You can also stop the Repository Connection dialog box from appearing at startup by disabling the **Show this dialog at startup** option.

- [Deleting a Repository](#)
- [Managing Content in the DI Repository](#)
- [Working with Version Control](#)

Deleting a Repository

When necessary, you can delete a DI repository or Kettle Database repository. Follow these instructions

1. In the **Repository Connection** dialog box, select the repository you want to delete from the list of available repositories.
2. Click **Delete**. A confirmation dialog appears.
3. Click **Yes** to delete the repository.

Managing Content in the DI Repository Explorer

Content is managed in the **Browse** tab in the **Repository Explorer** window. In the **Repository Explorer** window, you can perform common tasks such as those listed below:

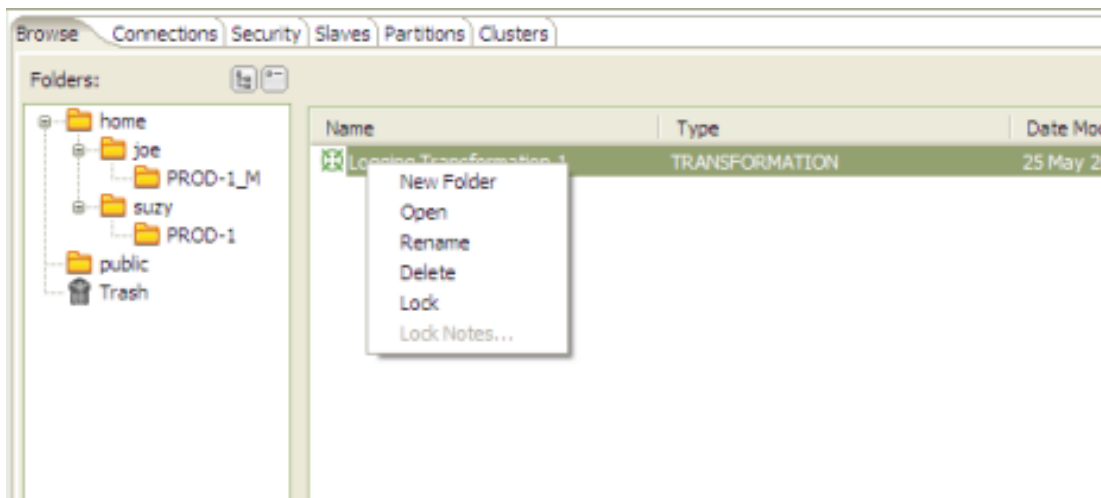
- Create a new folder in the repository
- Open a folder, job, or transformation
- Rename a folder, job or transformation
- Delete a folder, job, or transformation
- Lock a job or transformation

Access Repository Explorer Window

To access the **Repository Explorer** window, do these things.

1. Connect to a repository. To learn how to do this, see [Create DI Repository Connections](#).
2. Select **Tools > Repository > Explore**.
3. The **Repository Explorer** window appears.

Note: Permissions set by your administrator determine what you are able to view and tasks you are able to perform in the repository.



Create a New Folder in the Repository

To create a new folder in the repository, perform these steps.

1. In the **Browse** tab in the **Repository Explorer** window, right-click on the folder that you want to create the new folder under. For example, if you want to create a new folder under `public`, right-click the `public` folder.
2. Select **New Folder**.
3. Enter the name for the new folder in the **Name New Folder** window.

4. Click **OK**.

Open a Folder, Job, or Transformation

To open a folder, job, or transformation, right-click on the folder, job, or transformation and select **Open**.

NOTE:

To select more than one file, hold down the CTRL or SHIFT keys as you select the folders, jobs, or transformations, then right-click and select **Open**.

Rename a Folder, Job or Transformation

To rename a folder, job, or transformation, in the Repository Explorer window do these things.

1. In the **Browse** tab in the **Repository Explorer** window, right-click the folder, job, or transformation and select **Rename**.
2. Enter the new name in the **Name** window that appears.
3. Click **OK**.
4. Enter a comment when prompted.
5. Click the **OK** button.

In the **Version History** tab, a comment appears that indicates that the file has been renamed.

Move Objects

To **move** objects, such as folders, jobs, or transformations, in the repository, select the object, then click-and-drag it to the desired location in the navigation pane on the left. You can move an object in your folder to the folder of another repository user.

NOTE:

To select more than one file, hold down the CTRL or SHIFT keys as you select the folders, jobs, or transformations, then right-click and select **Move**.

Restore Objects

To **restore** an object you deleted, double-click the **Trash** icon. The object(s) you deleted appear in the right pane. Right-click on the object you want restored, and select **Restore** from the menu.

Delete a Folder, Job, or Transformation

To delete a folder, job, or transformation, in the Repository Explorer window do these things.

1. In the **Browse** tab in the **Repository Explorer** window, right-click the folder, job, or transformation and select **Delete**.
2. Click **Yes** when the warning message appears.
3. Click **OK**.

NOTE:

To select more than one file, hold down the CTRL or SHIFT keys as you select the folders, jobs, or transformations, then right-click and select **Delete**.

Lock and Unlock Jobs and Transformations

You can lock or unlock jobs and transformations. Locking and unlocking jobs and transformations protect them from being edited by other users.

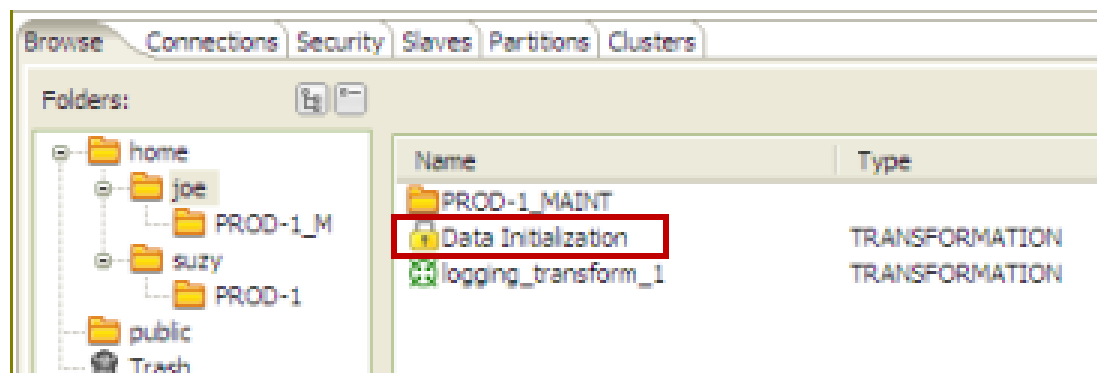
Lock a Job or Transformation

To lock a job or transformation, complete these steps.

1. In the **Browse** tab in the **Repository Explorer** window, right-click the job, or transformation and select **Lock**.
2. Enter the notes in the **Lock Notes** window that appears.
3. Click **OK**. The job or transformation icon changes to show a padlock.

NOTE:

The lock status icons are updated on each PDI client only when the **Repository Explorer** is accessed. If you want to refresh lock status in the **Repository Explorer**, exit and access it again. Also, select more than one file, hold down the CTRL or SHIFT keys as you select the folders, jobs, or transformations.



View Lock Notes

To view notes that were entered when the job or transformation was locked, do these things.

1. In the **Browse** tab in the **Repository Explorer** window, right-click the job, or transformation and select **Lock Notes**.
2. The lock note appears in a pop up window.
3. Click **OK** to dismiss the note.

Unlock a Job or Transformation

To unlock a job or transformation, complete these steps.

1. In the **Browse** tab in the **Repository Explorer** window, right-click the job, or transformation and select **Lock**.
2. The icon for the job or transformation returns to normal; the padlock icon disappears.

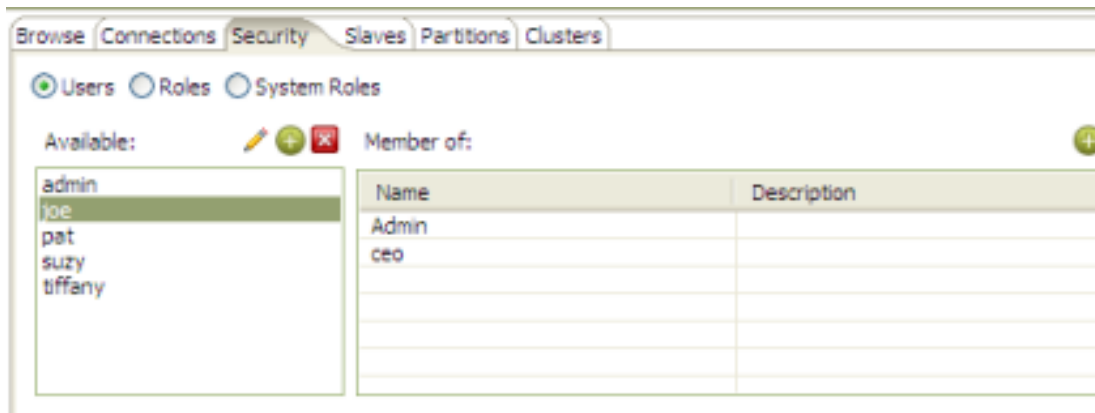
NOTE:

To select more than one file, hold down the CTRL or SHIFT keys as you select the folders, jobs, or transformations.

Access Connection, Security, and Cluster Information

In addition to managing content such as jobs and transformations, click the **Connections** tab to manage (create, edit, and delete) your database connections in the DI Repository. See [Managing Connections](#) for more information about connecting to a database.

Click the **Security** tab to manage users and roles. Pentaho Data Integration comes with a default security provider. If you do not have an existing security such as LDAP or MSAD, you can use Pentaho Security to define users and roles. You must have administrative privileges to manage security. For more information, see the section called [Administer the DI Server](#).



You can manage your slave servers (Data Integration and Carte instances) by clicking the **Slaves** tab. See [Setting Up a Slave Server](#) for instructions.

Click the **Partitions** and **Cluster** tabs to manage partitions and clusters. See [Creating a Cluster Schema](#) for more information.

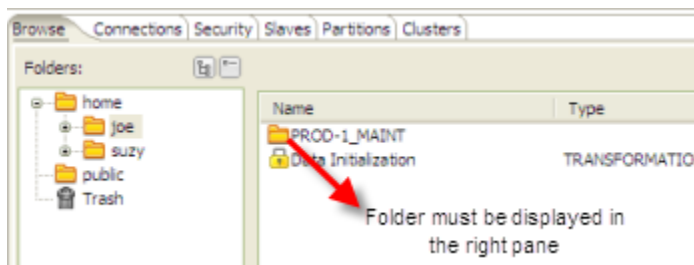
- [Setting Folder-Level Permissions](#)
- [Exporting Content from Solutions Repositories with Command-Line Tools](#)

Setting Folder-Level Permissions

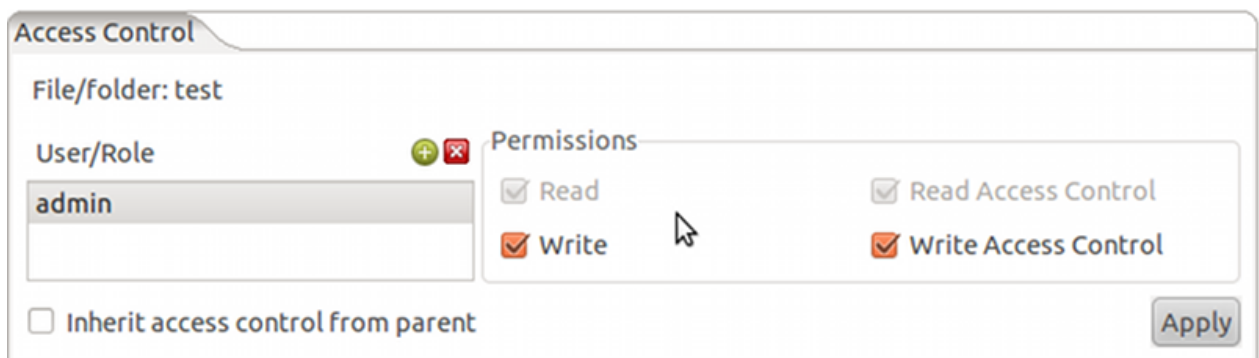
You can assign specific permissions to content files and folders stored in the DI Repository. Setting permissions manually will override inherited permissions if the access control flags allow. Follow the instructions below to set folder-level permissions.

1. Open the Repository Explorer (**Tools > Repository > Explore**).
2. Navigate to the folder to which you want permissions set and click to select it.

The folder must appear in the right pane before you can set permissions.



3. In the lower pane, under the **Permissions** tab, disable **Inherit security settings from parent**.
4. Click **Add** to open the **Select User or Role** dialog box.
5. Select a user or role to add to the permission list. Use the yellow arrows to move the user or role in or out of the permissions list. Click **OK** when you are done.
6. In the lower pane, under the **Access Control** tab, enable the appropriate **Permissions** granted to your selected user or role.



If you change your mind, use **Delete** to remove users or roles from the list.

7. Click **Apply** to apply permissions.
- [Access Control List \(ACL\) Permissions](#)

Access Control List (ACL) Permissions

These are the permissions settings for DI Repository content and folders.

Note: You must assign both **Write** and **Manage Access Control** to a directory in order to enable the selected user to create subfolders and save files within the folder.

Type	Value
Read	If set, the content of the file or contents of the directory will be accessible. Allows execution.
Manage Access Control	If set, access controls can be changed for this object.
Write	If set, enables read and write access to the selected content.
Delete	If set, the content of the file or directory can be deleted.

Exporting Content from Solutions Repositories with Command-Line Tools

To export repository objects into XML format, using command-line tools instead of exporting repository configurations from within Spoon, use named parameters and command-line options when calling Kitchen or Pan from a command-line prompt.

The following is an example command-line entry to execute an export job using Kitchen:

```
call kitchen.bat /file:C:\Pentaho_samples\repository\repository_export.kjb  
"/param:rep_name=PDI2000" "/param:rep_user=admin" "/param:rep_  
password=password"  
"/param:rep_folder=/public/dev"  
"/param:target_filename=C:\Pentaho_samples\repository\export\dev.xml"
```

Parameter	Description
rep_folder	Repository Folder
rep_name	Repository Name
rep_password	Repository Password
rep_user	Repository Username
target_filename	Target Filename

It is also possible to use obfuscated passwords with Encr, the command line tool for encrypting strings for storage/use by PDI. The following is an example command-line entry to execute a complete command-line call for the export in addition to checking for errors:

```
@echo off  
ECHO This an example of a batch file calling the repository_export.kjb
```

```

cd C:\Pentaho\pdi-ee-<filepath>--check--</filepath>{{contentVars.
PDIvernum3}}>\data-integration

call kitchen.bat /file:C:\Pentaho_samples\repository\repository_export.kjb
"/param:rep_name=PDI2000"
"/param:rep_user=admin" "/param:rep_password=password" "/param:rep_
folder=/public/dev"
"/param:target_filename=C:\Pentaho_samples\repository\export\dev.xml"

if errorlevel 1 goto error
echo Export finished successful.
goto finished

:error
echo ERROR: An error occurred during repository export.
:finished
REM Allow the user to read the message when testing, so having a pause
pause

```


Working with Version Control

Whenever you save a job or transformation in the DI Repository, you are prompted to provide a comment. Your comments are saved along with your job or transformation so that you can keep track of changes you make. If you have made a change to a transformation or job that you do not like, you can choose to restore a specific version of that job or transformation. It is important to provide descriptive version control comments, so that you can make good decisions when reverting to a version of a job or transformation.

- [Examining Revision History](#)
- [Restoring a Previously Saved Version of a Job or Transformation](#)

Examining Version History

To examine revision history for a job or transformation...

1. In Spoon menubar, go to **Tools > Repository > Explore**. The **Repository Explorer** window opens.
2. In the navigation pane on the left, locate and double-click the folder that contains your job or transformation.
3. Click on a transformation or job from the list to select it. The **Version History** associated with transformation or job appears in the lower pane.

Administrative users see the **home** folders of all users on the system. If you are not logged in as an administrator, you see your **home** and **public** folders. Your **home** folder is where you manage private content, such as transformations and jobs that are in progress. The **public** folder is where you store content that you want to share with others.

Right-click on the line under Version History that contains the transformation or job you want to examine. Choose **Open** to open the transformation or job in Spoon.

Restoring a Previously Saved Version of a Job or Transformation

To restore a version of a job or transformation...

1. In Spoon menubar, go to **Tools > Repository > Explore**. The **Repository Explorer** window opens.
2. Browse through the folders to locate the transformation or job that has multiple versions associated with it.
3. Right-click on a transformation or job from the list to select it.
4. Select **Restore**.
5. Write a meaningful comment in the **Commit Comment** dialog box and click **OK**. The version is restored. Next time you open the transformation or job, the restored version is what you will see.

Reusing Transformation Flows with Mapping Steps

When you want to reuse a specific sequence of steps, you can turn the repetitive part into a *mapping*. A mapping is a standard transformation except that you can define mapping input and output steps as placeholders.

- Mapping Input Specification — the placeholder used for input from the parent transformation
- Mapping Output Specification — the placeholder from which the parent transformation reads data

Note: Pentaho Data Integration samples that demonstrate the use of mapping steps are located at `...samples\mapping\Mapping`.

Below is the reference for the **Mapping (sub-transformation)** step:

Option	Description
Step name	Optionally, you can change the name of this step to fit your needs.
Mapping transformation	Specify the name of the mapping transformation file to execute at runtime. You can specify either a filename (XML/.ktr) or a transformation from the repository. The Edit button opens the specified transformation under a separate step in the Spoon Designer.
Parameters	Options under the Parameters tab allow you to define or pass PDI variables down to the mapping. This provides you with a high degree of customization. Note: It is possible to include variable expressions in the string values for the variable names. Note: Important! Only those variables/values that are specified are passed down to the sub-transformation.
Input Tabs	Each of the Input tabs (may be missing) correspond to one Mapping Input Specification step in the mapping or sub-transformation. This means you can have multiple Input tabs in a single Mapping step. To add an Input tab, click Add Input . <ul style="list-style-type: none">• Input source step name— The name of the step in the parent transformation (not the mapping) from which to read• Mapping target step name — The name of the step in the mapping (sub-transformation) to send the rows of data from the input source step• Is this the main data path? — Enable if you only have one input mapping ; you can leave the Mapping source step name and Output target step name fields blank

Option	Description
	<ul style="list-style-type: none"> • Ask these values to be renamed back on output? — Fields get renamed before they are transferred to the mapping transformation Note: Enabling this option renames the values back to their original names once they move to the Mapping output step. This option makes your sub-transformations more transparent and reusable. • Step mapping description — Add a description of the mapping step • Source - mapping transformation mapping Enter the required field name changes
Output Tabs	<p>Each of the Output tabs (may be missing) correspond to one Mapping Output Specification step in the mapping or sub-transformation. This means you can have multiple Output tabs in a single Mapping step. To add an Output tab, click Add Output.</p> <ul style="list-style-type: none"> • Mapping source step — the name of the step in the mapping transformation (sub-transformation) where that will be read • Output target step name — the name of the step in the current transformation (parent) to send the data from the mapping transformation step to. • Is this the main data path? — Enable if you only have one output mapping and you can leave the Mapping source step and Output target step name fields above blank. • Step mapping description — Add a description to the output step mapping • Mapping transformation - target step field mapping — Enter the required field name changes
Add input / Add output	Add an input or output mapping for the specified sub-transformation

Arguments, Parameters, and Variables

PDI has three paradigms for storing user input: arguments, parameters, and variables. Each is defined below, along with specific tips and configuration information.

- [Arguments](#)
- [Parameters](#)
- [Variables](#)

Arguments

A PDI argument is a named, user-supplied, single-value input given as a command line argument (running a transformation or job manually from Pan or Kitchen, or as part of a script). Each transformation or job can have a maximum of 10 arguments. Each argument is declared as space-separated values given after the rest of the Pan or Kitchen line:

```
sh pan.sh -file:/example_transformations/example.ktr argOne argTwo argThree
```

In the above example, the values **argOne**, **argTwo**, and **argThree** are passed into the transformation, where they will be handled according to the way the transformation is designed. If it was not designed to handle arguments, nothing will happen. Typically these values would be numbers, words (strings), or variables (system or script variables, not PDI variables).

In Spoon, you can test argument handling by defining a set of arguments when you run a transformation or job. This is accomplished by typing in values in the **Arguments** fields in the **Execute a Job** or **Execute a Transformation** dialogue.

Parameters

Parameters are like local variables; they are reusable inputs that apply only to the specific transformation that they are defined in. When defining a parameter, you can assign it a default value to use in the event that one is not fetched for it. This feature makes it unique among dynamic input types in PDI.

Note: If there is a name collision between a parameter and a variable, the parameter will take precedence.

To define a parameter, right-click on the transformation workspace and select **Transformation settings** from the context menu (or just press **Ctrl-T**), then click on the **Parameters** tab.

- [VFS Properties](#)

VFS Properties

`vfs . scheme . property . host`

The **vfs** subpart is required to identify this as a virtual filesystem configuration property. The **scheme** subpart represents the VFS driver's scheme (or VFS type), such as `http`, `sftp`, or `zip`. The **property** subpart is the name of a VFS driver's ConfigBuilder's setter (the specific VFS element that you want to set). The **host** optionally defines a specific IP address or hostname that this setting applies to.

You must consult each scheme's API reference to determine which properties you can create variables for. Apache provides VFS scheme documentation at <http://commons.apache.org/vfs/apidocs/index.html>. The `org.apache.commons.vfs.provider` package lists each of the configurable VFS providers (`ftp`, `http`, `sftp`, etc.). Each provider has a `FileSystemConfigBuilder` class that in turn has `set*(FileSystemOptions, Object)` methods. If a method's second parameter is a **String** or a **number** (Integer, Long, etc.) then you can create a PDI variable to set the value for VFS dialogues.

The table below explains VFS properties for the SFTP scheme. Each property must be declared as a PDI variable and preceded by the `vfs.sftp` prefix as defined above.

Note: All of these properties are optional.

SFTP VFS Property	Purpose
<code>compression</code>	Specifies whether zlib compression is used for the destination files. Possible values are zlib and none .
<code>identity</code>	The private key file (fully qualified local or remote path and filename) to use for host authentication.
<code>authkeypassphrase</code>	The passphrase for the private key specified by the identity property.
<code>StrictHostKeyChecking</code>	If this is set to no , the certificate of any remote host will be accepted. If set to yes , the remote host must exist in the known hosts file (<code>~/.ssh/known_hosts</code>).

Transformation properties

Transformation | **Parameters** | Logging | Dates | Dependencies | Miscellaneous | Monitoring

Parameters :

#	Parameter	Default Value	Description
1	vfs.http.proxyHost	192.168.1.250	A parameter that exists but is not used by any steps in this transformation
2	vfs.sftp.StrictHostKeyChecking	no	Accept the encryption key of any SFTP server
3	vfs.sftp.authkeyphrase.192.168.56.101	password	The passphrase for the private key used for authentication to the specified SFTP host
4	vfs.sftp.identity.192.168.56.101	C:\temp\auth\private_key	The private key used for authentication to the specified SFTP host

OK SQL Cancel

- [Configure SFTP VFS](#)

Configure SFTP VFS

To configure the connection settings for SFTP dialogues in PDI, you must create either variables or parameters for each relevant value. Possible values are determined by the VFS driver you are using.

You can also use parameters to substitute VFS connection details, then use them in the VFS dialogue where appropriate. For instance, these would be relevant credentials, assuming the parameters have been set:

```
sftp://${username}@${host}/${path}
```

This technique enables you to hide sensitive connection details, such as usernames and passwords.

See [VFS Properties](#) for more information on VFS options. You can also see all of these techniques in practice in the **VFS Configuration Sample** sample transformation in the `/data-integration/samples/transformations/` directory.

Variables

A variable in PDI is a piece of user-supplied information that can be used dynamically and programmatically in a variety of different scopes. A variable can be local to a single step, or be available to the entire JVM that PDI is running in.

PDI variables can be used in steps in both jobs and transformations. You define variables with the **Set Variable** step in a transformation, by hand through the **kettle.properties** file, or through the **Set Environment Variables** dialogue in the **Edit** menu.

The **Get Variable** step can explicitly retrieve a value from a variable, or you can use it in any PDI text field that has the diamond dollar sign



icon next to it by using a metadata string in either the Unix or Windows formats:

- `${VARIABLE}`
- `%%VARIABLE%%`

Both formats can be used and even mixed. In fact, you can create variable recursion by alternating between the Unix and Windows syntaxes. For example, if you wanted to resolve a variable that depends on another variable, then you could use this example: `${%%inner_var%%}`.

Note: If there is a name collision with a parameter or argument, variables will defer.

You can also use ASCII or hexadecimal character codes in place of variables, using the same format: `$(hex value)`. This makes it possible to escape the variable syntax in instances where you need to put variable-like text into a variable. For instance if you wanted to use `$(foobar)` in your data stream, then you can escape it like this: `$(24){foobar}`. PDI will replace `$(24)` with a `$` without resolving it as a variable.

- [Variable Scope](#)
- [Internal Variables](#)

Variable Scope

The scope of a variable is defined by the location of its definition. There are two types of variables: global environment variables, and Kettle variables. Both are defined below.

- [Environment Variables](#)
- [Kettle Variables](#)

Environment Variables

This is the traditional variable type in PDI. You define an environment variable through the **Set Environment Variables** dialogue in the **Edit** menu, or by hand by passing it as an option to the Java Virtual Machine (JVM) with the -D flag.

Environment variables are an easy way to specify the location of temporary files in a platform-independent way; for example, the `${java.io.tmpdir}` variable points to the `/tmp/` directory on Unix/Linux/OS X and to the `C:\Documents and Settings\<username>\Local Settings\Temp\` directory on Windows.

The only problem with using environment variables is that they cannot be used dynamically. For example, if you run two or more transformations or jobs at the same time on the same application server, you may get conflicts. Changes to the environment variables are visible to all software running on the virtual machine.

Kettle Variables

Kettle variables provide a way to store small pieces of information dynamically in a narrower scope than environment variables. A Kettle variable is local to Kettle, and can be scoped down to the job or transformation in which it is set, or up to a related job. The **Set Variable** step in a transformation allows you to specify the related job that you want to limit the scope to; for example, the parent job, grandparent job, or the root job.

Set Kettle Variables

Kettle variables configure various PDI-specific options such as the location of the shared object file for transformations and jobs or the log size limit. You can set Kettle variables using two methods.

- Set Kettle Variables in Spoon
- Set Kettle Variables Manually

If you are running a Pentaho MapReduce job, you can also set Kettle and environment variables in the Pentaho MapReduce job entry.

Set Kettle Variables in Spoon

To set Kettle variables in Spoon, complete these steps.

1. In Spoon, select **Edit > Edit the kettle.properties** file.
2. In the **Kettle Properties** window, modify the variable value.
3. If you want to add a variable, complete these steps.
 - a. Right-click on a line number, then select **Insert before this row** or **Insert after this row**.
 - b. Enter the variable name and value.
 - c. If you want to reposition the variable, right-click on the row number again, then select **Move Up** or **Move Down**.
4. Click the **OK** button.

Set Kettle Variables Manually

To edit Kettle variables manually, complete these steps.

1. Open the `kettle.properties` file in a text editor. By default, the `kettle.properties` file is typically stored in your home directory or the `.pentaho` directory.
2. Edit the file.
3. When complete, close and save the file.

Set Kettle or Java Environment Variables in the Pentaho MapReduce Job Entry

Pentaho MapReduce jobs are typically run in distributed fashion, with the mapper, combiner, and reducer run on different nodes. If you need to set a Java or Kettle environment variable for the different nodes, such as the `KETTLE_MAX_JOB_TRACKER_SIZE`, set them in the Pentaho MapReduce job entry window.

NOTE:

Values for Kettle environment variables set in the Pentaho MapReduce window override the Kettle environment variable values in the kettle.properties file. To view the list of variable names, descriptions, see the [Set Kettle Variables in Spoon](#) section of this document.

To set kettle or java environment variables, complete these steps.

1. In Spoon, double-click the **Pentaho MapReduce** job entry, then click the **User Defined** tab.

2. In the **Name** field, set the environment or Kettle variable you need.

For Kettle environment variables, type the name of the variable in the **Name** field, like this:

KETTLE_SAMPLE_VAR.

For Java environment variables, preface the value with the `java.system.` prefix, like this:

java.system.SAMPLE_PATH_VAR.

3. Enter the value of the variable in the **Value** field.

4. Click the **OK** button.

Internal Variables

The following variables are always defined:

Variable Name	Sample Value
Internal.Kettle.Build.Date	2010/05/22 18:01:39
Internal.Kettle.Build.Version	2045
Internal.Kettle.Version	4.3

These variables are defined in a transformation:

Variable Name	Sample Value
Internal.Transformation.Filename.Directory	D:\Kettle\samples
Internal.Transformation.Filename.Name	Denormaliser - 2 series of key-value pairs.ktr
Internal.Transformation.Name	Denormaliser - 2 series of key-value pairs sample
Internal.Transformation.Repository.Directory	/

These are the internal variables that are defined in a job:

Variable Name	Sample Value
Internal.Job.Filename.Directory	file:///home/matt/jobs
Internal.Job.Filename.Name	Nested jobs.kjb
Internal.Job.Name	Nested job test case
Internal.Job.Repository.Directory	/

These variables are defined in a transformation running on a slave server, executed in clustered mode:

Variable Name	Sample Value
Internal.Slave.Transformation.Number	0..<cluster size-1> (0,1,2,3 or 4)

Variable Name	Sample Value
Internal.Cluster.Size	<cluster size> (5)

Note: In addition to the above, there are also **System** parameters, including command line arguments. These can be accessed using the [Get System Info](#) step in a transformation.

Note: Additionally, you can specify values for variables in the **Execute a transformation** dialog box. If you include the variable names in your transformation they will appear in this dialog box.

Rapid Analysis Schema Prototyping

Data Integration offers rapid prototyping of analysis schemas through a mix of processes and tools known as **Agile BI**. The Agile BI functions of Pentaho Data Integration are explained in this section, but there is no further instruction here regarding PDI installation, configuration, or use beyond OLAP schema creation. If you need information related to PDI in general, consult the section on [installing PDI](#) and/or the section on [working with PDI](#) in the Pentaho InfoCenter.

Note: Agile BI is for **prototyping only**. It is extremely useful as an aid in developing OLAP schemas that meet the needs of BI developers, business users, and database administrators. However, **it should not be used for production**. Once your Agile BI schema has been refined, you will have to either hand-edit it in Schema Workbench to optimize it for performance, or completely re-implement the entire model with Schema Workbench.

- [Creating a Prototype Schema With a Non-PDI Data Source](#)
- [Creating a Prototype Schema With a PDI Data Source](#)
- [Prototypes in Production](#)

Creating a Prototype Schema With a Non-PDI Data Source

Your data sources must be configured, running, and available before you can proceed with this step.

Follow the below procedure to create a OLAP schema prototype from an existing database, file, or data warehouse.

Note: If you are already using PDI to create your data source, skip these instructions and refer to [Creating a Prototype Schema With a PDI Data Source](#) instead.

1. Start Spoon and connect to your repository, if you are using one.

```
cd ~/pentaho/design-tools/data-integration/ && ./spoon.sh
```

2. Go to the **File** menu, then select the **New** sub-menu, then click on **Model**. The interface will switch over to the **Model** perspective.
3. In the **Properties** pane on the right, click **Select**. A data source selection window will appear.
4. Click the round green + icon in the upper right corner of the window. The **Database Connection** dialogue will appear.
5. Enter in and select the connection details for your data source, then click **Test** to ensure that everything is correct. Click **OK** when you're done.
6. Select your newly-added data source, then click **OK**. The **Database Explorer** will appear.
7. Traverse the database hierarchy until you get to the table you want to create a model for. Right-click the table, then select **Model** from the context menu. The Database Explorer will close and bring you back to the Model perspective.
8. Drag items from the **Data** pane on the left and drop them into either the **Measures** or **Dimensions** groups in the **Model** pane in the center. The Measures and Dimensions groups will expand to include the items you drag into them.
9. Select each new measure and dimension item, and modify its details accordingly in the **Properties** pane on the right.
10. Save your model through the **File** menu, or publish it to the BA Server using the **Publish** icon above the Model pane.

You now have a basic OLAP schema. You should test it yourself before putting it into production. To do this, continue on to [Testing With Pentaho Analyzer and Report Wizard](#).

Creating a Prototype Schema With a PDI Data Source

1. Start Spoon and connect to your repository, if you are using one.

```
cd ~/pentaho/design-tools/data-integration/ && ./spoon.sh
```

2. Open the transformation that produces the data source you want to create a OLAP schema for.
3. Right-click your output step, then select **Model** from the context menu.
4. Drag items from the **Data** pane on the left and drop them into either the **Measures** or **Dimensions** groups in the **Model** pane in the center. The Measures and Dimensions groups will expand to include the items you drag into them.
5. Select each new measure and dimension item, and modify its details accordingly in the **Properties** pane on the right.
6. Save your model through the **File** menu, or publish it to the BA Server using the **Publish** icon above the Model pane.

You now have a basic OLAP schema. You should test it yourself before putting it into production. To do this, continue on to [Testing With Pentaho Analyzer and Report Wizard](#).

Testing With Pentaho Analyzer and Report Wizard

You must have an analysis schema with at least one measure and one dimension, and it must be currently open and focused on the Model perspective in Spoon.

This section explains how to use the embedded Analyzer and Report Design Wizard to test a prototype analysis schema.

1. While in the Model perspective, select your visualization method from the drop-down box above the Data pane (it has a **New:** to its left), then click **Go**. The two possible choices are: **Pentaho Analyzer** and **Report Wizard**. You do not need to have license keys for Pentaho Analysis or Pentaho Reporting in order to use these preview tools.
2. Either the Report Design Wizard will launch in a new sub-window, or Pentaho Analyzer will launch in a new tab. Use it as you would in Report Designer or the Pentaho User Console.
3. When you have explored your new schema, return to the Model perspective by clicking **Model** in the upper right corner of the Spoon toolbar, where all of the perspective buttons are. Do not close the tab; this will close the file, and you will have to reopen it in order to adjust your schema.
4. If you continue to refine your schema in the Model perspective, you must click the **Go** button again each time you want to view it in Analyzer or Report Wizard; the Visualize perspective does not automatically update according to the changes you make within the Model perspective.

You now have a preview of what your model will look like in production. Continue to refine it through the Model perspective, and test it through the Visualize perspective, until you meet your initial requirements.

Prototypes in Production

Once you're ready to test your OLAP schema on a wider scale, use the **Publish** button above the Model pane in the Model perspective, and use it to connect to your test or development BA Server.

You can continue to refine your schema if you like, but it must be republished each time you want to redeploy it.

Note: Agile BI is for **prototyping only**. It is extremely useful for developing OLAP schemas that meet the needs of business analytics developers, business users, and database administrators. However, **it should not be used for production**. Rather, once your Agile BI schema has been refined, you will have to either hand-edit it in Schema Workbench to optimize it for performance, or completely re-implement the entire model with Schema Workbench.

Using the SQL Editor

The **SQL Editor** is good tool to use when you must execute standard SQL commands for tasks such as creating tables, dropping indexes and modifying fields. The SQL Editor is used to preview and execute DDL (Data Definition Language) generated by Spoon such as "create/alter table, "create index," and "create sequence" SQL commands. For example, if you add a Table Output step to a transformation and click the SQL button at the bottom of the Table Input dialog box, Spoon automatically generates the necessary DDL for the output step to function properly and presents it to the end user through the SQL Editor.

Below are some points to consider:

- Multiple SQL Statements must be separated by semi-colons.
- Before SQL Statements are sent to the database to be executed, Spoon removes returns, line-feeds, and separating semi-colons.
- Pentaho Data Integration clears the database cache for the database connection on which you launch DDL statements.

The SQL Editor does not recognize the dialects of all supported databases. That means that creating stored procedures, triggers, and other database-specific objects may pose problems. Consider using the tools that came with the database in these instances.

Using the Database Explorer

The **Database Explorer** allow you to explore configured database connections. The Database Explorer also supports tables, views, and synonyms along with the catalog, schema, or both to which the table belongs.

A right-click on the selected table provides quick access to the following features:

Feature	Description
Preview first 100	Returns the first 100 rows from the selected table
Preview x Rows	Prompts you for the number of rows to return from the selected table
Row Count	Specifies the total number of rows in the selected table
Show Layout	Displays a list of column names, data types, and so on from the selected table
DDL	Generates the DDL to create the selected table based on the current connection type; the drop-down
View SQL	Launches the Simple SQL Editor for the selected table
Truncate Table	Generates a TRUNCATE table statement for the current table Note: The statement is commented out by default to prevent users from accidentally deleting the table data
Model	Switches to the Model perspective for the selected table
Visualize	Switches to the Visualize perspective for the selected table

Unsupported Databases

It may be possible to read from unsupported databases by using the generic database driver through an ODBC or JDBC connection. Contact Pentaho if you want to access a database type that is not yet in our list of [supported components](#).

You can add or replace a database driver files in the `lib` directory located under `...\design-tools\data-integration`.

Performance Monitoring and Logging

Pentaho Data Integration provides you with several methods in which to monitor the performance of jobs and transformations. Logging offers you summarized information regarding a job or transformation such as the number of records inserted and the total elapsed time spent in a transformation. In addition, logging provides detailed information about exceptions, errors, and debugging details.

Reasons you may want to enable logging and step performance monitoring include: determining if a job completed with errors or to review errors that were encountered during processing. In headless environments, most ETL in production is not run from the graphical user interface and you need a place to watch initiated job results. Finally, performance monitoring provides you with useful information for both current performance problems and capacity planning.

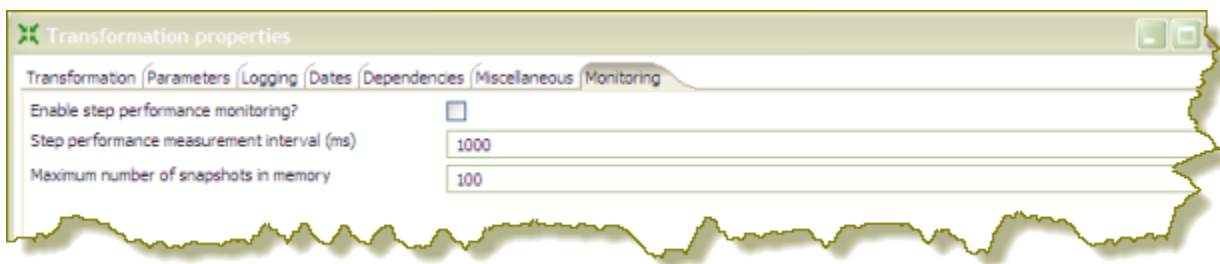
If you are an administrative user and want to monitor jobs and transformations, you must first set up logging and performance monitoring in Spoon. For more information about monitoring jobs and transformations, see the section [Administer the DI Server](#).

- [Monitoring Step Performance](#)
- [Logging Steps](#)
- [Logging Transformations](#)
- [Pentaho Data Integration Performance Tuning Tips](#)

Monitoring Step Performance

Pentaho Data Integration provides you with a tool for tracking the performance of individual steps in a transformation. By helping you identify the slowest step in the transformation, you can fine-tune and enhance the performance of your transformations.

You enable the step performance monitoring in the **Transformation Properties** dialog box. To access the dialog box right-click in the workspace that is displaying your transformation and choose, **Transformation Settings**. You can also access this dialog box, by pressing <CTRL + T>.

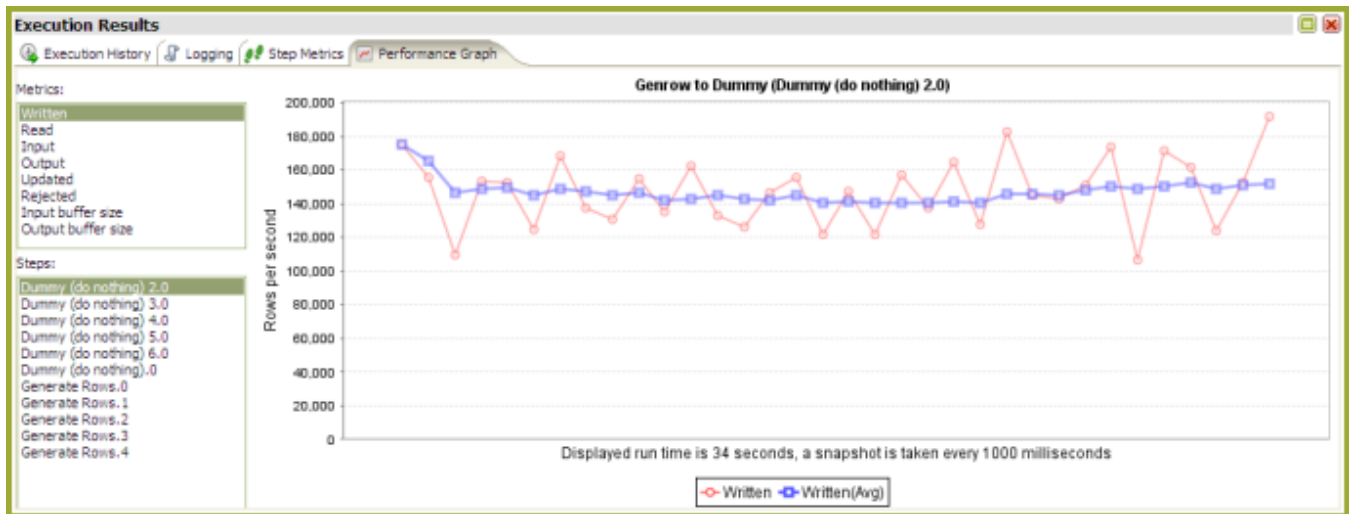


As shown in the sample screen capture above, the option to track performance (**Enable step performance monitoring?**) is not selected by default. Step performance monitoring may cause memory consumption problems in long-running transformations. By default, a performance snapshot is taken for all the running steps every second. This is not a CPU-intensive operation and, in most instances, does not negatively impact performance unless you have many steps in a transformation or you take a lot of snapshots (several per second, for example). You can control the number of snapshots in memory by changing the default value next to **Maximum number of snapshots in memory**. In addition, if you run in Spoon locally you may consume a fair amount of CPU power when you update the JFreeChart graphics under the Performance tab. Running in "headless" mode (Kitchen, Pan, DI Server (slave server), Carte, Pentaho BI platform, and so on) does not have this drawback and should provide you with accurate performance statistics.

- [Using Performance Graphs](#)

Using Performance Graphs

If you configured step performance monitoring, with database logging (optional), you can view the performance evolution graphs. Performance graphs provide you with a visual interpretation of how your transformation is processing. To enable database logging, enable the option **Enable step performance monitoring** within the **Transformation Properties / Monitoring** dialog box.



Follow the instructions below to set up a performance graph history for your transformation.

1. Right-click in the workspace (canvas) where you have an open transformation. Alternatively, press <CTRL +T>. To enable the logging, you also need to enable the option **Enable step performance monitoring** in the **Transformation Properties/Monitoring** in the dialog. The **Transformation Properties** dialog box appears.
2. In the Transformation Properties dialog box, click the **Logging** tab. Make sure **Performance** is selected in the navigation pane on the left.
3. Under **Logging** enter the following information:

Option	Description
Log Connection	Specifies the database connection you are using for logging; you can configure a new connection by clicking New .
Log Table Schema	Specifies the schema name, if supported by your database
Log Table Name	Specifies the name of the log table (for example L_ETL)
Logging interval (seconds)	Specifies the interval in which logs are written to the table

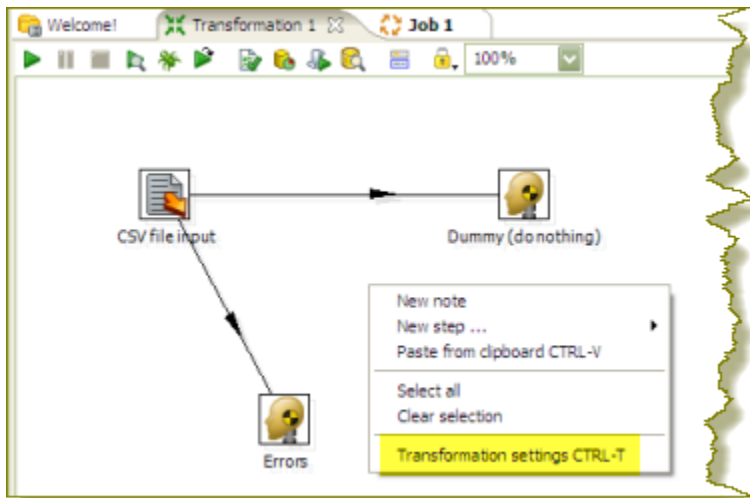
Option	Description
Log record timeout (in days)	Specifies the number of days old log entries in the table will be kept before they are deleted

4. Enable the fields you want to log or keep the defaults.
5. Click **SQL** to create your log table. The Simple SQL Editor appears.
6. Click **Execute** to execute the SQL code for your log table, then click **OK** to exit the **Results** dialog box.
Note: You *must* execute the SQL code to create the log table.
7. Click **Close** to exit the Simple SQL Editor.
8. Click **OK** to exit the Transformation Properties dialog box.

Logging Steps

Follow the instructions below to create a log table that keeps history of step-related information associated with your transformation.

1. Right-click in the workspace (canvas) where you have an open transformation. Alternatively, press <CTRL +T>.



The **Transformation Properties** dialog box appears.

2. In the Transformation Properties dialog box, click the **Logging** tab. Make sure **Step** is selected in the navigation pane on the left.



3. Under **Logging** enter the following information:

Option	Description
Log Connection	Specifies the database connection you are using for logging; you can configure a new connection by clicking New .
Log Table Schema	Specifies the schema name, if supported by your database
Log Table Name	Specifies the name of the log table (for example L_STEP)

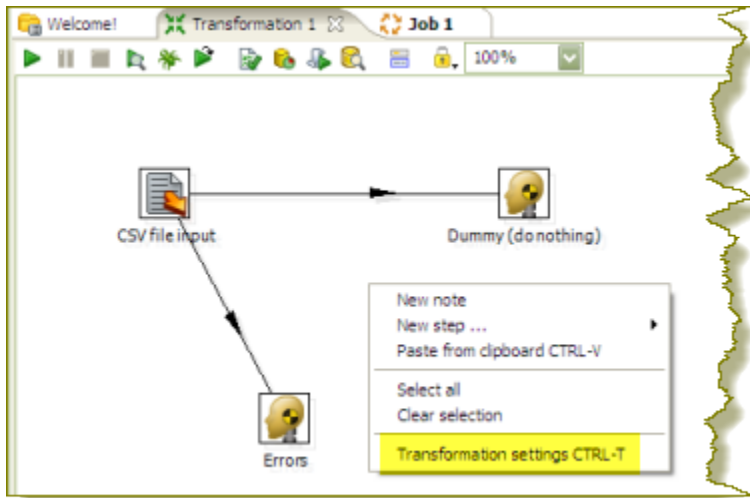
Option	Description
Logging interval (seconds)	Specifies the interval in which logs are written to the table
Log record timeout (in days)	Specifies the number of days old log entries in the table will be kept before they are deleted

4. Enable the fields you want to log or keep the defaults.
5. Click **SQL** to create your log table. The Simple SQL Editor appears.
6. Click **Execute** to execute the SQL code for your log table, then click **OK** to exit the **Results** dialog box.
Note: You *must* execute the SQL code to create the log table.
7. Click **Close** to exit the Simple SQL Editor.
8. Click **OK** to exit the Transformation Properties dialog box.

Logging Transformations

Follow the instructions below to create a log table for transformation-related processes:

1. Right-click in the workspace (canvas) where you have an open transformation. Alternatively, press <CTRL+T>.



The **Transformation Properties** dialog box appears.

2. In the Transformation Properties dialog box, click the **Logging** tab. Make sure **Transformation** is selected in the navigation pane on the left.



3. Under **Logging** enter the following information:

Option	Description
Log Connection	Specifies the database connection you are using for logging; you can configure a new connection by clicking New .
Log Table Schema	Specifies the schema name, if supported by your database
Log Table Name	Specifies the name of the log table (for example L_ETL)

Option	Description
Logging interval (seconds)	Specifies the interval in which logs are written to the table
Log record timeout (in days)	Specifies the number of days old log entries in the table will be kept before they are deleted
Log size limit in lines	Limits the number of lines that are stored in the LOG_FIELD (when selected under Fields to Log); when the LOG_FIELD is enabled Pentaho Data Integration will store logging associated with the transformation in a long text field (CLOB)

4. Enable the fields you want to log or keep the defaults.
5. Click **SQL** to create your log table. The Simple SQL Editor appears.
6. Click **Execute** to execute the SQL code for your log table, then click **OK** to exit the **Results** dialog box.
Note: You *must* execute the SQL code to create the log table.
7. Click **Close** to exit the Simple SQL Editor.
8. Click **OK** to exit the **Transformation Properties** dialog box.

The next time you run your transformation, logging information will be displayed under the **Execution History** tab.

Pentaho Data Integration Performance Tuning Tips

The tips described here may help you to identify and correct performance-related issues associated with PDI transformations.

Step	Tip	Description
JS	Turn off compatibility mode	<p>Rewriting JavaScript to use a format that is not compatible with previous versions is, in most instances, easy to do and makes scripts easier to work with and to read. By default, old JavaScript programs run in compatibility mode. That means that the step will process like it did in a previous version. You may see a small performance drop because of the overload associated with forcing compatibility. If you want make use of the new architecture, disable compatibility mode and change the code as shown below:</p> <ul style="list-style-type: none">• <code>intField.getInteger() > intField</code>• <code>numberField.getNumber() > numberField</code>• <code>dateField.getDate() > dateField</code>• <code>bigNumberField.getBigNumber() > bigNumberField</code>• and so on... <p>Instead of Java methods, use the built-in library. Notice that the resulting program code is more intuitive. For example :</p> <ul style="list-style-type: none">• checking for null is now: <code>field.isNull() > field==null</code>• Converting string to date: <code>field.Clone().str2dat() > str2date(field)</code>• and so on... <p>If you convert your code as shown above, you may get significant performance benefits.</p> <p>Note: It is no longer possible to modify data in-place using the value methods. This was a design decision to ensure that no data with the wrong type would end up in the output rows of the step. Instead of modifying fields in-place, create new fields using the table at the bottom of the Modified JavaScript transformation.</p>
JS	Combine steps	<p>One large JavaScript step runs faster than three consecutive smaller steps. Combining processes in one larger step helps to reduce overhead.</p>
JS	Avoid the JavaScript step or write a custom plug in	<p>Remember that while JavaScript is the fastest scripting language for Java, it is still a scripting language. If you do the same amount of work in a native step or plugin, you avoid the overhead of the JS scripting engine. This has been known to result in significant performance gains. It is also the primary reason why the Calculator step was created — to avoid the use of JavaScript for simple calculations.</p>
JS	Create a copy of a field	<p>No JavaScript is required for this; a "Select Values" step does the trick. You can specify the same field twice. Once without a rename, once (or more) with a rename. Another trick is to use <code>B=NVL(A,A)</code> in a Calculator step where B is forced</p>

Step	Tip	Description
		to be a copy of A. In version 3.1, an explicit "create copy of field A" function was added to the Calculator.
JS	Data conversion	Consider performing conversions between data types (dates, numeric data, and so on) in a "Select Values" step (version 3.0.2 or higher). You can do this in the Metadata tab of the step.
JS	Variable creation	If you have variables that can be declared once at the beginning of the transformation, make sure you put them in a separate script and mark that script as a startup script (right click on the script name in the tab). JavaScript object creation is time consuming so if you can avoid creating a new object for every row you are transforming, this will translate to a performance boost for the step.
N/A	Launch several copies of a step	There are two important reasons why launching multiple copies of a step may result in better performance: <ol style="list-style-type: none"> 1. The step uses a lot of CPU resources and you have multiple processor cores in your computer. Example: a JavaScript step 2. Network latencies and launching multiple copies of a step can reduce average latency. If you have a low network latency of say 5ms and you need to do a round trip to the database, the maximum performance you get is 200 (x5) rows per second, even if the database is running smoothly. You can try to reduce the round trips with caching, but if not, you can try to run multiple copies. Example: a database lookup or table output
N/A	Manage thread priorities	In versions 3.0.2 and higher, this feature that is found in the "Transformation Settings" dialog box under the (Misc tab) improves performance by reducing the locking overhead in certain situations. This feature is enabled by default for new transformations that are created in recent versions, but for older transformations this can be different.
Select Value	If possible, don't remove fields in Select Value	Don't remove fields in Select Value unless you must. It's a CPU-intensive task as the engine needs to reconstruct the complete row. It is almost always faster to add fields to a row rather than delete fields from a row.
Get Variables	Watch your use of Get Variables	May cause bottlenecks if you use it in a high-volume stream (accepting input). To solve the problem, take the "Get Variables" step out of the transformation (right click, detach) then insert it in with a "Join Rows (cart prod)" step. Make sure to specify the main step from which to read in the "Join Rows" step. Set it to the step that originally provided the "Get Variables" step with data.
N/A	Use new text file input	The new "CSV Input" or "Fixed Input" steps provide optimal performance. If you have a fixed width (field/row) input file, you can even read data in parallel. (multiple copies) These new steps have been rewritten using Non-blocking I/O (NIO) features. Typically, the larger the NIO buffer you specify in the step, the better your read performance will be.
N/A	When appropriate, use lazy conversion	In instances in which you are reading data from a text file and you write the data back to a text file, use Lazy conversion to speed up the process. The principle behind lazy conversion is that it delays data conversion in hopes that it isn't necessary (reading from a file and writing it back comes to mind). Beyond helping with data conversion, lazy conversion also helps to keep the data in "binary" storage form. This, in turn, helps the internal Kettle engine to perform faster data serialization (sort, clustering, and so on). The Lazy Conversion option is available in the "CSV Input" and "Fixed input" text file reading steps.

Step	Tip	Description
Join Rows	Use Join Rows	You need to specify the main step from which to read. This prevents the step from performing any unnecessary spooling to disk. If you are joining with a set of data that can fit into memory, make sure that the cache size (in rows of data) is large enough. This prevents (slow) spooling to disk.
N/A	Review the big picture: database, commit size, row set size and other factors	Consider how the whole environment influences performance. There can be limiting factors in the transformation itself and limiting factors that result from other applications and PDI. Performance depends on your database, your tables, indexes, the JDBC driver, your hardware, speed of the LAN connection to the database, the row size of data and your transformation itself. Test performance using different commit sizes and changing the number of rows in row sets in your transformation settings. Change buffer sizes in your JDBC drivers or database.
N/A	Step Performance Monitoring	Step Performance Monitoring is an important tool that allows you identify the slowest step in your transformation.

Working with Big Data and Hadoop in PDI

Pentaho Data Integration (PDI) can operate in two distinct modes, job orchestration and data transformation. Within PDI they are referred to as jobs and transformations.

PDI jobs sequence a set of entries that encapsulate actions. An example of a PDI big data job would be to check for existence of new log files, copy the new files to HDFS, execute a MapReduce task to aggregate the weblog into a click stream and stage that clickstream data in an analytic database.

PDI transformations consist of a set of steps that execute in parallel and operate on a stream of data columns. The columns usually flow from one system, through the PDI engine, where new columns can be calculated or values can be looked up and added to the stream. The data stream is then sent to a receiving system like a Hadoop cluster, a database, or even the Pentaho Reporting Engine.

The tutorials within this section illustrate how to use PDI jobs and transforms in typical big data scenarios. PDI job entries and transformation steps are described in the [Transformation Step Reference](#) and [Job Entry Reference](#) sections of Administer the DI Server.

PDI's Big Data Plugin

The Pentaho Big Data plugin contains all of the job entries and transformation steps required for working with Hadoop, Cassandra, and MongoDB.

By default, PDI is pre-configured to work with Apache Hadoop 0.20.X. But PDI can be configured to communicate with most popular Hadoop distributions. Instructions for changing Hadoop configurations are covered in the [Configure Your Big Data Environment](#) section.

For a list of supported big data technology, including which configurations of Hadoop are currently supported, see the section on [Supported Components](#).

Using PDI Outside and Inside the Hadoop Cluster

PDI is unique in that it can execute both outside of a Hadoop cluster and within the nodes of a hadoop cluster. From outside a Hadoop cluster, PDI can extract data from or load data into Hadoop HDFS, Hive and HBase. When executed within the Hadoop cluster, PDI transformations can be used as Mapper and/or Reducer tasks, allowing PDI with Pentaho MapReduce to be used as visual programming tool for MapReduce.

These videos demonstrate using PDI to work with Hadoop from both inside and outside a Hadoop cluster.

- Loading Data into Hadoop from outside the Hadoop cluster is a 5-minute video that demonstrates moving data using a PDI job and transformation: <http://www.youtube.com/watch?v=Ylekzmd6TAc>

- Use [Pentaho MapReduce](#) to interactively design a data flow for a MapReduce job without writing scripts or code. Here is a 12 minute video that provides an overview of the process: <http://www.youtube.com/watch?v=KZe1UugxXcs>.
- [Pentaho MapReduce Workflow](#)
- [PDI Hadoop Job Workflow](#)
- [Hadoop to PDI Data Type Conversion](#)
- [Hadoop Hive-Specific SQL Limitations](#)
- [Big Data Tutorials](#)

Pentaho MapReduce Workflow

PDI and Pentaho MapReduce enables you to pull data from a Hadoop cluster, transform it, and pass it back to the cluster. Here is how you would approach doing this.

PDI Transformation

Start by deciding what you want to do with your data, open a PDI transformation, and drag the appropriate steps onto the canvas, configuring the steps to meet your data requirements. Drag the specifically-designed Hadoop **MapReduce Input** and Hadoop **MapReduce Output** steps onto the canvas. PDI provides these steps to completely avoid the need to write Java classes for this functionality. Configure both of these steps as needed. Once you have configured all the steps, add hops to sequence the steps as a transformation. Follow the workflow as shown in this sample transformation in order to properly communicate with Hadoop. Name this transformation Mapper.

File:/hadoop_transformation_workflow.jpg

Hadoop communicates in key/value pairs. PDI uses the **MapReduce Input** step to define how key/value pairs from Hadoop are interpreted by PDI. The **MapReduce Input** dialog box enables you to configure the **MapReduce Input** step.

File:/hadoop_mr_input.jpg

PDI uses a **MapReduce Output** step to pass the output back to Hadoop. The **MapReduce Output** dialog box enables you to configure the **MapReduce Output** step.

File:/hadoop_mr_output.jpg

What happens in the middle is entirely up to you. Pentaho provides many sample steps you can alter to create the functionality you need.

PDI Job

Once you have created the Mapper transformation, you are ready to include it in a **Pentaho MapReduce** job entry and build a MapReduce job. Open a PDI job and drag the specifically-designed **Pentaho MapReduce** job entry onto the canvas. In addition to ordinary transformation work, this entry is designed to execute mapper/reducer functions within PDI. Again, no need to provide a Java class to achieve this.

Configure the **Pentaho MapReduce** entry to use the transformation as a mapper. Drag and drop a Start job entry, other job entries as needed, and result jobentries to handle the output onto the canvas. Add hops to sequence the entries into a job that you execute in PDI.

The workflow for the job should look something like this.

File:/hadoop_transformation_job_workflow.jpg

The Pentaho **MapReduce** dialog box enables you to configure the Pentaho MapReduce entry.

File:/hadoop_transformation_job_config.jpg

PDI Hadoop Job Workflow

PDI enables you to execute a Java class from within a PDI/Spoon job to perform operations on Hadoop data. The way you approach doing this is similar to the way you would for any other PDI job. The specifically-designed job entry that handles the Java class is **Hadoop Job Executor**. In this illustration it is used in the **WordCount - Advanced** entry.

File:/hadoop_job_workflow.jpg

The **Hadoop Job Executor** dialog box enables you to configure the entry with a `jar` file that contains the Java class.

File:/hadoop_job_config.jpg

If you are using the Amazon Elastic MapReduce (EMR) service, you can **Amazon EMR Job Executor**. job entry to execute the Java class. This differs from the standard Hadoop Job Executor in that it contains connection information for Amazon S3 and configuration options for EMR.

File:/hadoop_emr_job.jpg

Hadoop to PDI Data Type Conversion

The **Hadoop Job Executor** and **Pentaho MapReduce** steps have an advanced configuration mode that enables you to specify data types for the job's input and output. PDI is unable to detect foreign data types on its own; therefore you must specify the input and output data types in the **Job Setup** tab. This table explains the relationship between Hadoop data types and their PDI equivalents.

PDI (Kettle) Data Type	Apache Hadoop Data Type
<code>java.lang.Integer</code>	<code>org.apache.hadoop.io.IntWritable</code>
<code>java.lang.Long</code>	<code>org.apache.hadoop.io.IntWritable</code>
<code>java.lang.Long</code>	<code>org.apache.hadoop.io.LongWritable</code>
<code>org.apache.hadoop.io.IntWritable</code>	<code>java.lang.Long</code>
<code>java.lang.String</code>	<code>org.apache.hadoop.io.Text</code>
<code>java.lang.String</code>	<code>org.apache.hadoop.io.IntWritable</code>
<code>org.apache.hadoop.io.LongWritable</code>	<code>org.apache.hadoop.io.Text</code>
<code>org.apache.hadoop.io.LongWritable</code>	<code>java.lang.Long</code>

For more information on configuring **Pentaho MapReduce** to convert to additional data types, see <http://wiki.pentaho.com/display/BAD/Pentaho+MapReduce>.

Hadoop Hive-Specific SQL Limitations

There are a few key limitations in Hive that prevent some regular Metadata Editor features from working as intended, and limit the structure of your SQL queries in Report Designer:

- **Outer joins are not supported.**
- **Each column can only be used once in a SELECT clause.** Duplicate columns in SELECT statements cause errors.
- **Conditional joins can only use the = conditional unless you use a WHERE clause.** Any non-equal conditional in a FROM statement forces the Metadata Editor to use a cartesian join and a WHERE clause conditional to limit it. This is not much of a limitation, but it may seem unusual to experienced Metadata Editor users who are accustomed to working with SQL databases.

Big Data Tutorials

These sections contain guidance and instructions about using Pentaho technology as part of your overall big data strategy. Each section is a series of scenario-based tutorials that demonstrate the integration between Pentaho and Hadoop using a sample data set.

- [Hadoop Tutorials](#)
- [MapR Tutorials](#)
- [Cassandra Tutorials](#)
- [MongoDB Tutorials](#)

Hadoop Tutorials

These tutorials are organized by topic and each set explains various techniques for loading, transforming, extracting and reporting on data within a Hadoop cluster. You are encouraged to perform the tutorials in order as the output of one is sometimes used as the input of another. However, if you would like to jump to a tutorial in the middle of the flow, instructions for preparing input data are provided.

- [Loading Data into a Hadoop Cluster](#)
- [Transforming Data within a Hadoop Cluster](#)
- [Extracting Data from a Hadoop Cluster](#)
- [Reporting on Data within a Hadoop Cluster](#)

Loading Data into a Hadoop Cluster

These scenario-based tutorials contain guidance and instructions on loading data into HDFS (Hadoop's Distributed File System), Hive and HBase using Pentaho Data Integration (PDI)

- [Prerequisites](#)
- [Using a Job Entry to Load Data into Hadoop's Distributed File System \(HDFS\)](#)
- [Using a Job Entry to Load Data into Hive](#)
- [Using a Transformation Step to Load Data into HBase](#)

Prerequisites

To perform the tutorials in this section you must have these components installed.

PDI—The primary development environment for the tutorials. See the [Data Integration Installation Options](#) if you have not already installed PDI.

Apache Hadoop 0.20.X—A single-node local cluster is sufficient for these exercises, but a larger and/or remote configuration also works. If you are using a different distribution of Hadoop see [Configure Your Big Data Environment](#). You need to know the addresses and ports for your Hadoop installation.

***Hive**—A supported version of Hive. Hive is a Map/Reduce abstraction layer that provides SQL-like access to Hadoop data. For instructions on installing or using Hive, see the [Hive Getting Started Guide](#).

***HBase**—A supported version of HBase. HBase is an open source, non-relational, distributed database that runs on top of HDFS. For instructions on installing or using HBase, see the [Getting Started section of the Apache HBase Reference Guide](#).

**Component only required for corresponding tutorial.*

- [Sample Data](#)

Sample Data

The tutorials in this section were created with this sample weblog data.

Tutorial	File Name	Content
Using a Job Entry to Load Data into Hadoop's Distributed File System (HDFS)	weblogs_rebuild.txt.zip	Unparsed, raw weblog data
Using a Job Entry to Load Data into Hive	weblogs_parse.txt.zip	Tab-delimited, parsed weblog data
Using a Transformation Step to Load Data into HBase	weblogs_hbase.txt.zip	Prepared data for HBase load





Using a Job Entry to Load Data into Hadoop's Distributed File System (HDFS)

In order to follow along with this tutorial, you will need

- Hadoop
- Pentaho Data Integration

You can use PDI jobs to put files into HDFS from many different sources. This tutorial describes how to create a PDI job to move a sample file into HDFS.

If not already running, start Hadoop and PDI. Unzip the sample data files and put them in a convenient location: [weblogs_rebuild.txt.zip](#).

1. Create a new Job by selecting **File > New > Job**.
2. Add a Start job entry to the canvas. From the **Design** palette on the left, under the **General** folder, drag a **Start** job entry onto the canvas.
File:/loading_data_into_hdfs_step2.png
3. Add a Hadoop Copy Files job entry to the canvas. From the **Design** palette, under the **Big Data** folder, drag a **Hadoop Copy Files** job entry onto the canvas.
File:/loading_data_into_hdfs_step3.png
4. Connect the two job entries by hovering over the **Start** entry and selecting the output connector
File:/loading_data_into_hdfs_step4a.png
, then drag the connector arrow to the **Hadoop Copy Files** entry.
File:/loading_data_into_hdfs_step4.png
5. Enter the source and destination information within the properties of the **Hadoop Copy Files** entry by double-clicking it.
 - a. For **File/Folder source(s)**, click **Browse** and navigate to the folder containing the downloaded sample file `weblogs_rebuild.txt`.
 - b. For **File/Folder destination(s)**, enter `hdfs://<NAMENODE>:<PORT>/user/pdi/weblogs/raw`, where `NAMENODE` and `PORT` reflect your Hadoop destination.
 - c. For **Wildcard (RegExp)**, enter `^.*\.txt`.
 - d. Click **Add** to include the entries to the list of files to copy.
 - e. Check the **Create destination folder** option to ensure that the `weblogs` folder is created in HDFS the first time this job is executed.

When you are done your window should look like this (your file paths may be different).

File:/loading_data_into_hdfs_step5.png

Click **OK** to close the window.

6. Save the job by selecting **Save as** from the **File** menu. Enter `load_hdfs.kjb` as the file name within a folder of your choice.

7. Run the job by clicking the green Run button on the job toolbar

File:/loading_data_into_hdfs_result_run.png

, or by selecting **Action > Run** from the menu. The **Execute a job** window opens. Click **Launch**.

An **Execution Results** panel opens at the bottom of the Spoon interface and displays the progress of the job as it runs. After a few seconds the job finishes successfully.

File:/loading_data_into_hdfs_step7.PNG

If any errors occurred the job entry that failed will be highlighted in red and you can use the **Logging** tab to view error messages.

8. Verify the data was loaded by querying Hadoop.
 - a. From the command line, query Hadoop by entering this command.

```
hadoop fs -ls /user/pdi/weblogs/raw
```

This statement is returned

```
-rwxrwxrwx 3 demo demo 77908174 2011-12-28 07:16 /user/pdi/weblogs/raw/weblog_raw.txt
```

Using a Job Entry to Load Data into Hive

In order to follow along with this tutorial, you will need

- Hadoop
- Pentaho Data Integration
- Hive

PDI jobs can be used to put files into Hive from many different sources. This tutorial instructs you how to use a PDI job to load a sample data file into a Hive table.

Note: Hive could be defined with external data. Using the external option, you could define a Hive table that uses the HDFS directory that contains the parsed file. For this tutorial, we chose not to use the external option to demonstrate the ease with which files can be added to non-external Hive tables.

If not already running, start Hadoop, PDI, and the Hive server. Unzip the sample data files and put them in a convenient location: [weblogs_parse.txt.zip](#).

This file should be placed in the `/user/pdi/weblogs/parse` directory of HDFS using these three commands.

```
hadoop fs -mkdir /user/pdi/weblogs
hadoop fs -mkdir /user/pdi/weblogs/parse
hadoop fs -put weblogs_parse.txt /user/pdi/weblogs/parse/part-00000
```

If you previously completed the [Using Pentaho MapReduce to Parse Weblog Data](#) tutorial, the necessary files will already be in the proper directory.

1. Create a Hive Table.
 - a. Open the Hive shell by entering `'hive'` at the command line.
 - b. Create a table in Hive for the sample data by entering

```
create table weblogs (
  client_ip      string,
  full_request_date string,
  day           string,
  month         string,
  month_num     int,
  year          string,
  hour          string,
  minute        string,
  second        string,
  timezone      string,
  http_verb     string,
```

```
uri      string,  
http_status_code  string,  
bytes_returned    string,  
referrer         string,  
user_agent       string)  
row format delimited  
fields terminated by '\t';
```

- c. Close the Hive shell by entering 'quit'.
2. Create a new Job to load the sample data into a Hive table by selecting **File > New > Job**.
3. Add a Start job entry to the canvas. From the **Design** palette on the left, under the **General** folder, drag a **Start** job entry onto the canvas.
File:/loading_data_into_hdfs_step2.png
4. Add a Hadoop Copy Files job entry to the canvas. From the **Design** palette, under the **Big Data** folder, drag a **Hadoop Copy Files** job entry onto the canvas.
File:/loading_data_into_hdfs_step3.png
5. Connect the two job entries by hovering over the **Start** entry and selecting the output connector
File:/loading_data_into_hdfs_step4a.png
, then drag the connector arrow to the **Hadoop Copy Files** entry.
File:/loading_data_into_hdfs_step4.png
6. Enter the source and destination information within the properties of the **Hadoop Copy Files** entry by double-clicking it.
 - a. For **File/Folder source(s)**, enter `hdfs://<NAMENODE>:<PORT>/user/pdi/weblogs/parse`, where **NAMENODE** and **PORT** reflect your Hadoop destination.
 - b. For **File/Folder destination(s)**, enter `hdfs://<NAMENODE>:<PORT>/user/hive/warehouse/weblogs`.
 - c. For **Wildcard (RegExp)**, enter `part-.*`.
 - d. Click the **Add** button to add the entries to the list of files to copy.

When you are done your window should look like this (your file paths may be different)

File:/loading_data_into_hive_step6_result.png

Click **OK** to close the window.

7. Save the job by selecting **Save as** from the **File** menu. Enter `load_hive.kjb` as the file name within a folder of your choice.
8. Run the job by clicking the green Run button on the job toolbar
File:/loading_data_into_hive_result_run.png
, or by selecting **Action > Run** from the menu. The **Execute a job** window opens. Click **Launch**.

An **Execution Results** panel opens at the bottom of the Spoon interface and displays the progress of the job as it runs. After a few seconds the job finishes successfully.

File:/loading_data_into_hive_result.png

If any errors occurred the job entry that failed will be highlighted in red and you can use the **Logging** tab to view error messages.

9. Verify the data was loaded by querying Hive.
 - a. Open the Hive shell from the command line by entering `hive`.
 - b. Enter this query to verify the data was loaded correctly into Hive.

```
select * from weblogs limit 10;
```

Ten rows of data are returned.

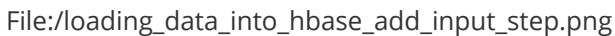
Using a Transformation Step to Load Data into HBase

In order to follow along with this tutorial, you will need

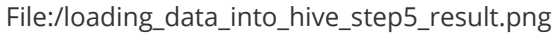
- Hadoop
- Pentaho Data Integration
- HBase

This tutorial describes how to use data from a sample flat file to create a HBase table using a PDI transformation. For the sake of brevity, you will use a prepared sample dataset and a simple transformation to prepare and transform your data for HBase loads.

If not already running, start Hadoop, PDI, and HBase. Unzip the sample data files and put them in a convenient location: [weblogs_hbase.txt.zip](#)

1. Create a HBase Table.
 - a. Open the HBase shell by entering `hbase shell` at the command line.
 - b. Create the table in HBase by entering `create 'weblogs', 'pageviews'` in the HBase shell. This creates a table named `weblogs` with a single column family named `pageviews`.
 - c. Close the HBase shell by entering `quit`.
2. From within the Spoon, create a new transformation by selecting **File > New > Transformation**.
3. Identify the source where the transformation will get data from. For this tutorial your source is a text file (`.txt`). From the **Input** folder of the **Design** palette on the left, add a **Text File Input** step to the transformation by dragging it onto the canvas.

4. Edit the properties of the **Text file input** step by double-clicking the icon. The **Text file input** dialog box appears.
5. From the **File** tab, in the **File or Directory** field, click **Browse** and navigate to the `weblog_hbase.txt` file. Click **Add**.

The file appears in the **Selected files** pane.



6. Configure the contents of the file by switching to the **Content** tab.
 - a. For **Separator**, clear the contents and click **Insert TAB**.
 - b. Check the **Header** checkbox.
 - c. For **Format**, Select **Unix** from the drop-down menu.



7. Configure the input fields.
 - a. From the **Fields** tab, select **Get Fields** to populate the list the available fields.
 - b. A dialog box appears asking for **Number of sample lines**. Enter **100** and click **OK**.

- c. Change the **Type** of the field named **key** to **String** and set the **Length** to 20.

File:/loading_data_into_hbase_text_file_input_fields_tab.png

Click **OK** to close the window.

8. On the **Design** palette, under **Big Data**, drag the **HBase Output** to the canvas. Create a hop to connect your input and **HBase Output** step by hovering over the input step and clicking the output connector

File:/loading_data_into_hbase_step7a.png

, then drag the connector arrow to the **HBase Output** step.

File:/loading_data_into_hbase_step7.png

9. Edit the **HBase Output** step by double-clicking it. You must now enter your Zookeeper host(s) and port number.
 - a. For the **Zookeeper hosts(s)** field, enter a comma separated list of your HBase Zookeeper Hosts. For local single node clusters use `localhost`.
 - b. For **Zookeeper port**, enter the port for your Zookeeper hosts. By default this is `2181`.
10. Create a HBase mapping to tell Pentaho how to store the data in HBase by switching to the **Create/Edit mappings** tab and changing these options.
 - a. For **HBase table name**, select **weblogs**.
 - b. For **Mapping name**, enter `pageviews`.
 - c. Click **Get incoming fields**.
 - d. For the alias **key** change the **Key** column to **Y**, clear the **Column family** and **Column name** fields, and set the **Type** field to **String**. Click **Save mapping**.

File:/loading_data_into_hbase_step9.png

11. Configure the HBase out to use the mapping you just created.
 - a. Go back to the **Configure connection** tab and click **Get table names**.
 - b. For **HBase table name**, enter `weblogs`.
 - c. Click **Get mappings for the specified table**.
 - d. For **Mapping name**, select `pageviews`. Click **OK** to close the window.

Save the transformation by selecting **Save as** from the **File** menu. Enter `load_hbase.ktr` as the file name within a folder of your choice.

12. Run the transformation by clicking the green **Run** button on the transformation toolbar

File:/loading_data_into_hbase_result_run.png

, or by choosing **Action > Run** from the menu. The **Execute a transformation** window opens. Click **Launch**.

An **Execution Results** panel opens at the bottom of the Spoon interface and displays the progress of the transformation as it runs. After a few seconds the transformation finishes successfully.

File:/loading_data_into_hbase_result.png

If any errors occurred the transformation step that failed will be highlighted in red and you can use the **Logging** tab to view error messages.

13. Verify the data was loaded by querying HBase.

- a. From the command line, open the HBase shell by entering this command.

```
hbase shell
```

b. Query HBase by entering this command.

```
scan 'weblogs', {LIMIT => 10}
```

Ten rows of data are returned.

Transforming Data within a Hadoop Cluster

These tutorials contain guidance and instructions on transforming data within the Hadoop cluster using Pentaho MapReduce, Hive, and Pig.

- [Using Pentaho MapReduce to Parse Weblog Data](#)—How to use Pentaho MapReduce to convert raw weblog data into parsed, delimited records.
- [Using Pentaho MapReduce to Generate an Aggregate Dataset](#)—How to use Pentaho MapReduce to transform and summarize detailed data into an aggregate dataset.
- [Transforming Data within Hive](#)—How to read data from a Hive table, transform it, and write it to a Hive table within the workflow of a PDI job.
- [Transforming Data with Pig](#)—How to invoke a Pig script from a PDI job.

Extracting Data from a Hadoop Cluster

These tutorials contain guidance and instructions on extracting data from Hadoop using HDFS, Hive, and HBase.

- [Extracting Data from HDFS to Load an RDBMS](#)—How to use a PDI transformation to extract data from HDFS and load it into a RDBMS table.
- [Extracting Data from Hive to Load an RDBMS](#)—How to use a PDI transformation to extract data from Hive and load it into a RDBMS table.
- [Extracting Data from HBase to Load an RDBMS](#)—How to use a PDI transformation to extract data from HBase and load it into a RDBMS table.
- [Extracting Data from Snappy Compressed Files](#)—How to configure client-side PDI so that files compressed using the Snappy codec can be decompressed using the Hadoop file input or Text file input step.

Reporting on Data within a Hadoop Cluster

These tutorials contain guidance and instructions about reporting on data within a Hadoop cluster.

- [Reporting on HDFS File Data](#)—How to create a report that sources data from a HDFS file.
- [Reporting on HBase Data](#)—How to create a report that sources data from HBase.
- [Reporting on Hive Data](#)—How to create a report that sources data from Hive.

MapR Tutorials

These tutorials are organized by topic and each set explains various techniques for loading, transforming, extracting and reporting on data within a MapR cluster. You are encouraged to perform the tutorials in order as the output of one is sometimes used as the input of another. However, if you would like to jump to a tutorial in the middle of the flow, instructions for preparing input data are provided.

- [Loading Data into a MapR Cluster](#)
- [Transforming Data within a MapR Cluster](#)
- [Extracting Data from a MapR Cluster](#)
- [Reporting on Data within a MapR Cluster](#)

Loading Data into a MapR Cluster

These tutorials contain guidance and instructions on loading data into CLDB (MapR's distributed file system), Hive, and HBase.

- [Loading Data into CLDB](#)—How to use a PDI job to move a file into CLDB.
- [Loading Data into MapR Hive](#)—How to use a PDI job to load a data file into a Hive table.
- [Loading Data into MapR HBase](#)—How to use a PDI transformation that sources data from a flat file and writes to an HBase table.

Transforming Data within a MapR Cluster

These tutorials contain guidance and instructions on leveraging the massively parallel, fault tolerant MapR processing engine to transform resident cluster data.

- [Using Pentaho MapReduce to Parse Weblog Data in MapR](#)—How to use Pentaho MapReduce to convert raw weblog data into parsed, delimited records.
- [Using Pentaho MapReduce to Generate an Aggregate Dataset in MapR](#)—How to use Pentaho MapReduce to transform and summarize detailed data into an aggregate dataset.
- [Transforming Data within Hive in MapR](#)—How to read data from a Hive table, transform it, and write it to a Hive table within the workflow of a PDI job.
- [Transforming Data with Pig in MapR](#)—How to invoke a Pig script from a PDI job.

Extracting Data from a MapR Cluster

These tutorials contain guidance and instructions on extracting data from a MapR cluster and loading it into an RDBMS table.

- [Extracting Data from CLDB to Load an RDBMS](#)—How to use a PDI transformation to extract data from MapR CLDB and load it into a RDBMS table.
- [Extracting Data from Hive to Load an RDBMS in MapR](#)—How to use a PDI transformation to extract data from Hive and load it into a RDBMS table.
- [Extracting Data from HBase to Load an RDBMS in MapR](#)—How to use a PDI transformation to extract data from HBase and load it into a RDBMS table.

Reporting on Data within a MapR Cluster

These tutorials contain guidance and instructions about reporting on data within a MapR cluster.

- [Reporting on CLDB File Data](#) —How to create a report that sources data from a MapR CLDB file.
- [Reporting on HBase Data in MapR](#)—How to create a report that sources data from HBase.
- [Reporting on Hive Data in MapR](#)—How to create a report that sources data from Hive.

Cassandra Tutorials

These tutorials demonstrate the integration between Pentaho and the Cassandra NoSQL Database, specifically techniques about writing data to and reading data from Cassandra using graphical tools. These tutorials also include instructions on how to sort and group data, create reports, and combine data from Cassandra with data from other sources.

- [Write Data To Cassandra](#)—How to read data from a data source (flat file) and write it to a column family in Cassandra using a graphic tool.
- [How To Read Data From Cassandra](#)—How to read data from a column family in Cassandra using a graphic tool.
- [How To Create a Report with Cassandra](#)—How to create a report that uses data from a column family in Cassandra using graphic tools.

MongoDB Tutorials

These tutorials demonstrate the integration between Pentaho and the MongoDB NoSQL Database, specifically how to write data to, read data from, MongoDB using graphical tools. These tutorials also include instructions on sorting and grouping data, creating reports, and combining data from Mongo with data from other sources.

- [Write Data To MongoDB](#)—How to read data from a data source (flat file) and write it to a collection in MongoDB
- [Read Data From MongoDB](#)—How to read data from a collection in MongoDB.
- [Create a Report with MongoDB](#)—How to create a report that uses data from a collection in MongoDB.
- [Create a Parameterized Report with MongoDB](#)—How to create a parameterize report that uses data from a collection in MongoDB.

Implement Data Services with the Thin Kettle JDBC Driver

The Thin Kettle JDBC Driver provides a means for a Java-based client to query the results of a transformation. Any Java-based, JDBC-compliant tool, including third-party reporting systems, can use this driver to query a Kettle transformation by using a SQL string via JDBC. With the Thin Kettle JDBC Driver, you can blend, enrich, clean, and transform data from multiple sources to create a single data federation source. You can also seamlessly integrate with Enterprise Service Buses (ESB).

Details on how to use the Thin Kettle JDBC Driver [appear on the wiki](#).

- [Configuration of the Kettle JDBC Driver](#)
- [Example of How to Use the Kettle JDBC Driver](#)
- [JDBC Driver and SQL Reference](#)

Transactional Databases and Job Rollback

By default, when you run a job or transformation that makes changes to a database table, changes are committed as the transformation or job executes. Sometimes, this can cause an issue if a job or transformation fails. For example, if you run a job that updates then syncs two tables, but the job fails before you can write to the second table, the first table might be updated and the other might not, rendering them both out of sync. If this is a concern, consider implementing job rollback by making the transformation or job databases (or both) transactional. When you do this, changes to a data source occur only if a transformation or job completes successfully. Otherwise, the information in both data sources remain unchanged.

The following links provide general information on how to make databases transactional. The wiki provides [more detail](#).

- [Make a Transformation Database Transactional](#)
- [Make a Job Database Transactional](#)

Make a Transformation Database Transactional

To make a transformation database transactional, complete these steps.

1. In Spoon, open a transformation.
2. Right-click an empty space in the transformation's tab and select **Transformation Settings** from the menu that appears.
3. Click the **Miscellaneous** tab.
4. Enable the **Make the transformation database transactional** checkbox.
5. Click **OK** to close the window.

Make a Job Database Transactional

To make a job database transactional, complete these steps.

1. In Spoon, open a job.
2. Right-click in an empty space in the job's tab. Select **Job Settings** from the menu that appears.
3. Click the **Transactions** tab.
4. Enable the **Make the job database transactional** checkbox.
5. Click **OK** to close the window.

Interacting With Web Services

PDI jobs and transformations can interact with a variety of Web services through specialized steps. How you use these steps, and which ones you use, is largely determined by your definition of "Web services." The most commonly used Web services steps are:

- [Web Service Lookup](#)
- [Modified Java Script Value](#)
- [RSS Input](#)
- [HTTP Post](#)

The Web Service Lookup Step is useful for selecting and setting input and output parameters via WSDL, but only if you do not need to modify the SOAP request. You can see this step in action in the **Web Services - NOAA Latitude and Longitude.ktr** sample transformation included with PDI in the `/data-integration/samples/transformations/` directory.

There are times when the SOAP message generated by the Web Services Lookup step is insufficient. Many Web services require the security credentials be placed in the SOAP request headers. There may also be a need to parse the response XML to get more information than the response values provide (such as namespaces). In cases like these, you can use the Modified Java Script Value step to create whatever SOAP envelope you need. You would then hop to an HTTP Post step to accept the SOAP request through the input stream and post it to the Web service, then hop to another Modified Java Script Value to parse the response. The **General - Annotated SOAP Web Service call.ktr** sample in the `/data-integration/samples/transformations/` directory shows this theory in practice.

Scheduling and Scripting PDI Content

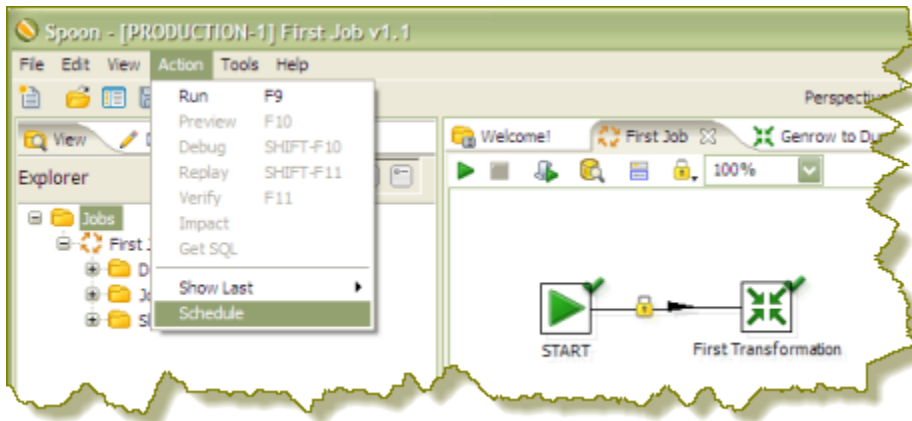
Once you're finished designing your PDI jobs and transformations, you can arrange to run them at certain time intervals through the DI Server, or through your own scheduling mechanism (such as **cron** on Linux, and the **Task Scheduler** or the **at** command on Windows). The methods of operation for scheduling and scripting are different; scheduling through the DI Server is done through the Spoon graphical interface, whereas scripting using your own scheduler or executor is done by calling the **pan** or **kitchen** commands. This section explains all of the details for scripting and scheduling PDI content.

- [Scheduling Transformations and Jobs From Spoon](#)
- [Command-Line Scripting Through Pan and Kitchen](#)

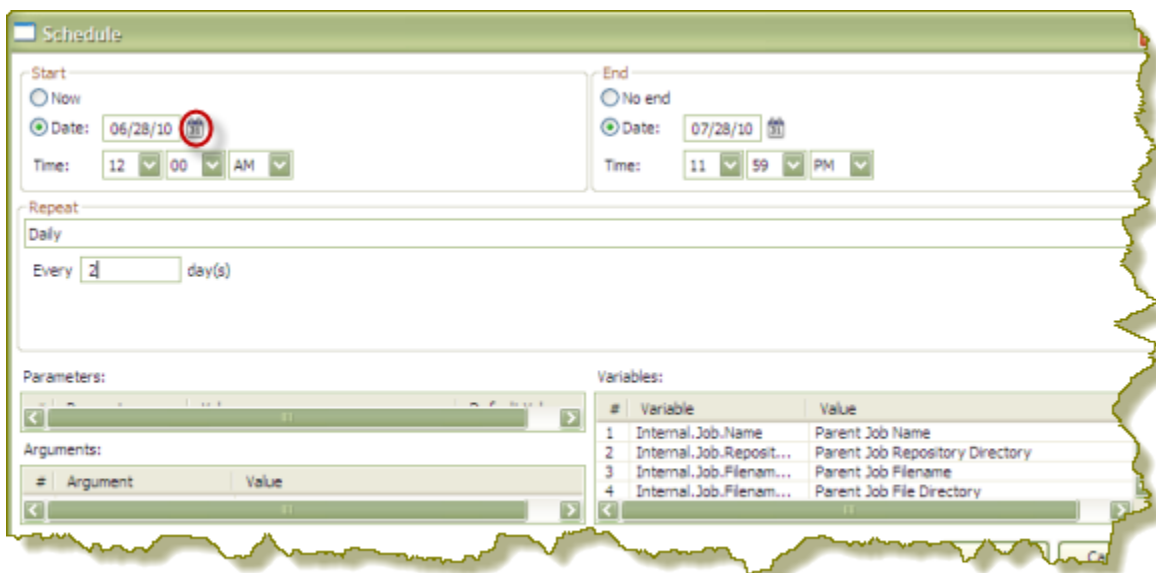
Scheduling Transformations and Jobs From Spoon

You can schedule jobs and transformations to execute automatically on a recurring basis by following the directions below.

1. Open a job or transformation, then go to the **Action** menu and select **Schedule**.



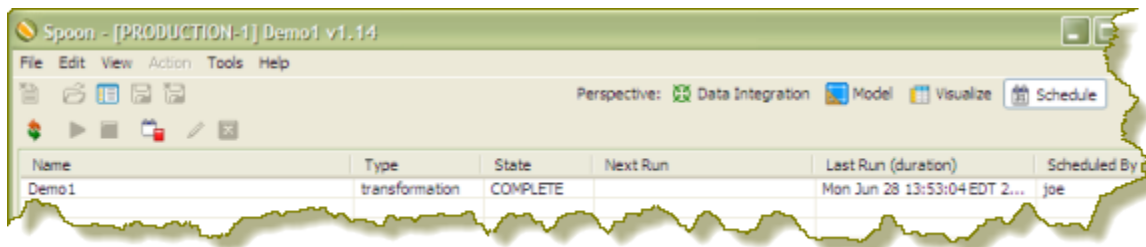
2. In the **Schedule a Transformation** dialog box, enter the date and time that you want the schedule to begin in the **Start** area, or click the calendar icon (circled in red) to display the calendar. To run the transformation immediately, enable the **Now** radio button.



3. Set up the **End** date and time. If applicable, enable the **No end** radio button or click on the calendar and input the date and time to end the transformation.
4. If applicable, set up a recurrence under **Repeat**.

End date and time are disabled unless you select a recurrence. From the list of schedule options select the choice that is most appropriate: **Run Once, Seconds, Minutes, Hourly, Daily, Weekly, Monthly, Yearly.**

5. Make sure you set parameters, arguments and variables, if available. Click **OK**.
6. In the Spoon button bar, click the **Schedule** perspective.



From the Schedule perspective, you can refresh, start, pause, stop and delete a transformation or job using the buttons on the upper left corner of the page.



Command-Line Scripting Through Pan and Kitchen

You can use PDI's command line tools to execute PDI content from outside of Spoon. Typically you would use these tools in the context of creating a script or a cron job to run the job or transformation based on some condition outside of the realm of Pentaho software.

Pan is the PDI command line tool for executing transformations.

Kitchen is the PDI command line tool for executing jobs.

Both of these programs are explained in detail below.

- [Pan Options and Syntax](#)
- [Kitchen Options and Syntax](#)
- [Importing KJB or KTR Files From a Zip Archive](#)
- [Connecting to a DI Solution Repositories with Command-Line Tools](#)
- [Exporting Content from Solutions Repositories with Command-Line Tools](#)

Pan Options and Syntax

Pan runs transformations, either from a PDI repository (database or enterprise), or from a local file. The syntax for the batch file and shell script are shown below. All Pan options are the same for both.

```
pan.sh - option = value arg1 arg2
```

```
pan.bat / option : value arg1 arg2
```

Switch	Purpose
rep	Enterprise or database repository name, if you are using one
user	Repository username
pass	Repository password
trans	The name of the transformation (as it appears in the repository) to launch
dir	The repository directory that contains the transformation, including the leading slash
file	If you are calling a local KTR file, this is the filename, including the path if it is not in the local directory
level	The logging level (Basic, Detailed, Debug, Rowlevel, Error, Nothing)
logfile	A local filename to write log output to
listdir	Lists the directories in the specified repository
listtrans	Lists the transformations in the specified repository directory
listrep	Lists the available repositories
exprep	Exports all repository objects to one XML file

norep	Prevents Pan from logging into a repository. If you have set the KETTLE_REPOSITORY, KETTLE_USER, and KETTLE_PASSWORD environment variables, then this option will enable you to prevent Pan from logging into the specified repository, assuming you would like to execute a local KTR file instead.
safemode	Runs in safe mode, which enables extra checking
version	Shows the version, revision, and build date
param	Set a named parameter in a name=value format. For example: -param:FOO=bar
listparam	List information about the defined named parameters in the specified transformation.
maxloglines	The maximum number of log lines that are kept internally by PDI. Set to 0 to keep all rows (default)
maxlogtimeout	The maximum age (in minutes) of a log line while being kept internally by PDI. Set to 0 to keep all rows indefinitely (default)

```
sh pan.sh -rep=initech_pdi_repo -user=pgibbons -pass=lumburghsux -trans=TPS_
reports_2011
```

```
pan.bat /rep:initech_pdi_repo /user:pgibbons /pass:lumburghsux /trans:TPS_reports_
2011
```

- [Pan Status Codes](#)

Pan Status Codes

When you run Pan, there are seven possible return codes that indicate the result of the operation. All of them are defined below.

Status code	Definition
0	The transformation ran without a problem.
1	Errors occurred during processing
2	An unexpected error occurred during loading / running of the transformation
3	Unable to prepare and initialize this transformation
7	The transformation couldn't be loaded from XML or the Repository
8	Error loading steps or plugins (error in loading one of the plugins mostly)
9	Command line usage printing

Kitchen Options and Syntax

Kitchen runs jobs, either from a PDI repository (database or enterprise), or from a local file. The syntax for the batch file and shell script are shown below. All Kitchen options are the same for both.

```
kitchen.sh - option = value arg1 arg2
```

```
kitchen.bat / option : value arg1 arg2
```

Switch	Purpose
rep	Enterprise or database repository name, if you are using one
user	Repository username
pass	Repository password
job	The name of the job (as it appears in the repository) to launch
dir	The repository directory that contains the job, including the leading slash
file	If you are calling a local KJB file, this is the filename, including the path if it is not in the local directory
level	The logging level (Basic, Detailed, Debug, Rowlevel, Error, Nothing)
logfile	A local filename to write log output to
listdir	Lists the directories in the specified repository
listjob	Lists the jobs in the specified repository directory
listrep	Lists the available repositories
export	Exports all linked resources of the specified job. The argument is the name of a ZIP file.
norep	Prevents Kitchen from logging into a repository. If you have set the KETTLE_REPOSITORY, KETTLE_USER, and KETTLE_PASSWORD environment variables, then this option will enable you to prevent

Switch	Purpose
	Kitchen from logging into the specified repository, assuming you would like to execute a local KTR file instead.
version	Shows the version, revision, and build date
param	Set a named parameter in a name=value format. For example: -param:FOO=bar
listparam	List information about the defined named parameters in the specified job.
maxloglines	The maximum number of log lines that are kept internally by PDI. Set to 0 to keep all rows (default)
maxlogtimeout	The maximum age (in minutes) of a log line while being kept internally by PDI. Set to 0 to keep all rows indefinitely (default)

```
sh kitchen.sh -rep=initech_pdi_repo -user=pgibbons -pass=lumburghsux -job=TPS_
reports_2011
```

```
kitchen.bat /rep:initech_pdi_repo /user:pgibbons /pass:lumburghsux /job:TPS_
reports_2011
```

- [Kitchen Status Codes](#)

Kitchen Status Codes

When you run Kitchen, there are seven possible return codes that indicate the result of the operation. All of them are defined below.

Status code	Definition
0	The job ran without a problem.
1	Errors occurred during processing
2	An unexpected error occurred during loading or running of the job
7	The job couldn't be loaded from XML or the Repository
8	Error loading steps or plugins (error in loading one of the plugins mostly)
9	Command line usage printing

Importing KJB or KTR Files From a Zip Archive

Both Pan and Kitchen can pull PDI content files from out of Zip files. To do this, use the ! switch, as in this example:

```
Kitchen.bat /file:"zip:file:///C:/Pentaho/PDI_Examples/Sandbox/linked_executable_
job_and_transform.zip!Hourly_Stats_Job_Unix.kjb"
```

If you are using Linux or Solaris, the ! must be escaped:

```
./kitchen.sh -file:"zip:file:///home/user/pentaho/pdi-ee/my_package/linked_
executable_job_and_transform.zip\\!Hourly_Stats_Job_Unix.kjb"
```

Connecting to a DI Solution Repository with Command-Line Tools

To export repository objects into XML format using command-line tools instead of exporting repository configurations from within Spoon, use named parameters and command-line options when calling Kitchen or Pan from a command-line prompt.

The following is an example command-line entry to execute an export job using Kitchen:

```
call kitchen.bat /file:C:\Pentaho_samples\repository\repository_export.kjb
"/param:rep_name=PDI2000" "/param:rep_user=admin" "/param:rep_
password=password"
"/param:rep_folder=/public/dev"
"/param:target_filename=C:\Pentaho_samples\repository\export\dev.xml"
```

Parameter	Description
rep_folder	Repository Folder
rep_name	Repository Name
rep_password	Repository Password
rep_user	Repository Username
target_filename	Target Filename

Note: It is also possible to use obfuscated passwords with Encr a command line tool for encrypting strings for storage or use by PDI.

The following is an example command-line entry to execute a complete command-line call for the export in addition to checking for errors:

```
@echo off
ECHO This an example of a batch file calling the repository_export.kjb

cd C:\Pentaho\pdi-ee-<filepath>--check--</filepath>{{contentVars.
PDIvernum3}}>\data-integration
```

```
call kitchen.bat /file:C:\Pentaho_samples\repository\repository_export.kjb
"/param:rep_name=PDI2000"
"/param:rep_user=admin" "/param:rep_password=password" "/param:rep_
folder=/public/dev"
"/param:target_filename=C:\Pentaho_samples\repository\export\dev.xml"

if errorlevel 1 goto error
echo Export finished successfull.
goto finished

:error
echo ERROR: An error occured during repository export.
:finished
REM Allow the user to read the message when testing, so having a pause
pause
```

Exporting Content from Solutions Repositories with Command-Line Tools

To export repository objects into XML format, using command-line tools instead of exporting repository configurations from within Spoon, use named parameters and command-line options when calling Kitchen or Pan from a command-line prompt.

The following is an example command-line entry to execute an export job using Kitchen:

```
call kitchen.bat /file:C:\Pentaho_samples\repository\repository_export.kjb
"/param:rep_name=PD12000" "/param:rep_user=admin" "/param:rep_
password=password"
"/param:rep_folder=/public/dev"
"/param:target_filename=C:\Pentaho_samples\repository\export\dev.xml"
```

Parameter	Description
rep_folder	Repository Folder
rep_name	Repository Name
rep_password	Repository Password
rep_user	Repository Username
target_filename	Target Filename

It is also possible to use obfuscated passwords with Encr, the command line tool for encrypting strings for storage/use by PDI. The following is an example command-line entry to execute a complete command-line call for the export in addition to checking for errors:

```
@echo off
ECHO This an example of a batch file calling the repository_export.kjb
```

```
cd C:\Pentaho\pdi-ee-<filepath>--check--</filepath>{{contentVars.  
PDIvernum3}}>\data-integration  
  
call kitchen.bat /file:C:\Pentaho_samples\repository\repository_export.kjb  
"/param:rep_name=PDI2000"  
"/param:rep_user=admin" "/param:rep_password=password" "/param:rep_  
folder=/public/dev"  
"/param:target_filename=C:\Pentaho_samples\repository\export\dev.xml"  
  
if errorlevel 1 goto error  
echo Export finished successful.  
goto finished  
  
:error  
echo ERROR: An error occurred during repository export.  
:finished  
REM Allow the user to read the message when testing, so having a pause  
pause
```

Transformation Step Reference

This section contains reference documentation for transformation steps.

Note: Many steps are not completely documented in this section, but have rough definitions in the Pentaho Wiki: <http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps>.

- [Agile](#)
- [Big Data](#)
- [Input](#)
- [Output](#)
- [Transform](#)
- [Utility](#)
- [Flow](#)
- [Scripting](#)
- [Lookup](#)
- [Joins](#)
- [Data Warehouse](#)
- [Validation](#)
- [Statistics](#)
- [Palo](#)
- [Job](#)
- [Mapping](#)
- [Bulk Loading](#)
- [Inline](#)
- [Data Mining Steps](#)
- [Experimental](#)
- [Deprecated](#)

Agile

The PDI transformation steps in this section pertain to Agile Mart.

- [MonetDB Agile Mart](#)
- [Table Agile Mart](#)

Big Data

The PDI transformation steps in this section pertain to Big Data operations.

Note: PDI is configured by default to use the Apache Hadoop distribution. If you are working with a Cloudera or MapR distribution instead, you must install the appropriate patch before using any Hadoop functions in PDI. Patch installation is covered in [Select DI Installation Options](#) and [Getting Started with PDI and Hadoop](#).

- [Avro Input](#)
- [Cassandra Input](#)
- [Cassandra Output](#)
- [CouchDB](#)
- [Hadoop File Input](#)
- [Hadoop File Output](#)
- [HBase Input](#)
- [HBase Output](#)
- [HBase Row Decoder](#)
- [MapReduce Input](#)
- [MapReduce Output](#)
- [MongoDB Input](#)
- [MongoDB Output](#)
- [Splunk Input](#)
- [Splunk Output](#)
- [SSTable Output](#)

Input

The PDI transformation steps in this section pertain to various methods of data input.

- [CSV file input](#)
- [Data Grid](#)
- [De-serialize from file](#)
- [Email messages input](#)
- [ESRI Shapefile Reader](#)
- [Fixed file input](#)
- [Generate random credit card numbers](#)
- [Generate random value](#)
- [Generate Rows](#)
- [Get data from XML](#)
- [Get File Names](#)
- [Get File Rows Count](#)
- [Get repository names](#)
- [Get SubFolder names](#)
- [Get System Info](#)
- [Get tables names](#)
- [Google Analytics](#)
- [Google Docs Input](#)
- [GZIP CSV Input](#)
- [HL7 Input](#)
- [IBM Websphere MQ Consumer](#)
- [JMS Consumer](#)
- [JSON Input](#)
- [LDAP Input](#)
- [LDIF Input](#)

- [Load file content in memory](#)
- [Microsoft Access Input](#)
- [Microsoft Excel Input](#)
- [Mondrian Input](#)
- [OLAP Input](#)
- [Property Input](#)
- [RSS Input](#)
- [S3 CSV Input](#)
- [Salesforce Input](#)
- [SAP Input](#)
- [SAS Input](#)
- [Table input](#)
- [Text file input](#)
- [XBase Input](#)
- [XML Input Stream \(StAX\)](#)
- [Yaml Input](#)

Output

The PDI transformation steps in this section pertain to various methods of data output.

- [Automatic Documentation Output](#)
- [Delete](#)
- [IBM Websphere MQ Producer](#)
- [Insert - Update](#)
- [JMS Producer](#)
- [Json Output](#)
- [LDAP Output](#)
- [Microsoft Access Output](#)
- [Microsoft Excel Output](#)
- [Microsoft Excel Writer](#)
- [RSS Output](#)
- [OpenERP Object Output](#)
- [Properties Output](#)
- [RSS Output](#)
- [S3 File Output](#)
- [Salesforce Delete](#)
- [Salesforce Insert](#)
- [Salesforce Update](#)
- [Salesforce Upsert](#)
- [Serialize to file](#)
- [SQL File Output](#)
- [Synchronize after merge](#)
- [Table Output](#)
- [Text File Output](#)
- [Update](#)

- [XML Output](#)

Transform

The PDI transformation steps in this section pertain to various data modification tasks.

- [Add a checksum](#)
- [Add constants](#)
- [Add sequence](#)
- [Add value fields changing sequence](#)
- [Add XML](#)
- [Calculator](#)
- [Closure Generator](#)
- [Concat Fields](#)
- [Example Plugin](#)
- [Get ID from slave server](#)
- [Number range](#)
- [Replace in string](#)
- [Row denormaliser](#)
- [Row flattener](#)
- [Row Normaliser](#)
- [Select values](#)
- [Set field value](#)
- [Set field value to a constant](#)
- [Sort rows](#)
- [Split field to rows](#)
- [Split Fields](#)
- [String operations](#)
- [Strings cut](#)
- [Unique rows](#)
- [Unique rows \(HashSet\)](#)

- [Value Mapper](#)
- [XSL Transformation](#)

Utility

The PDI transformation steps in this section pertain to various conditional and data processing tasks.

- [Change file encoding](#)
- [Clone row](#)
- [Delay row](#)
- [Edi to XML](#)
- [Execute a process](#)
- [If field value is null](#)
- [Mail](#)
- [Metadata structure of stream](#)
- [Null if](#)
- [Process files](#)
- [Run SSH commands](#)
- [Send message to Syslog](#)
- [Table Compare](#)
- [Write to log](#)
- [Zip file](#)

Flow

The PDI transformation steps in this section pertain to various process control tasks.

- [Abort](#)
- [Append Streams](#)
- [Block this step until steps finish](#)
- [Blocking Step](#)
- [Detect empty stream](#)
- [Dummy \(do nothing\)](#)
- [ETL Metadata Injection](#)
- [Filter rows](#)
- [Identify last row in a stream](#)
- [Java Filter](#)
- [Job Executor](#)
- [Prioritize streams](#)
- [Single Threader](#)
- [Switch / Case](#)
- [Transformation Executor](#)

Scripting

The PDI transformation steps in this section pertain to formula and script execution.

- [Execute row SQL script](#)
- [Execute SQL script](#)
- [Formula](#)
- [Modified Java Script Value](#)
- [R Script Executor](#)
- [Regex Evaluation](#)
- [Rule Accumulator](#)
- [Rule Executor](#)
- [User Defined Java Class](#)
- [User Defined Java Expression](#)

Lookup

The PDI transformation steps in this section pertain to status checking and remote service interaction.

- [Call DB Procedure](#)
- [Check if the a column exists](#)
- [Check if file is locked](#)
- [Check if webservice is available](#)
- [Database join](#)
- [Database lookup](#)
- [Dynamic SQL row](#)
- [File exists](#)
- [Fuzzy match](#)
- [HTTP client](#)
- [HTTP Post](#)
- [MaxMind GeoIP Lookup](#)
- [REST Client](#)
- [Stream lookup](#)
- [Table exists](#)
- [Web services lookup](#)

Joins

The PDI transformation steps in this section pertain to database and file join operations.

- [Join Rows \(Cartesian product\)](#)
- [Merge Join](#)
- [Merge Rows \(diff\)](#)
- [Multiway Merge Join](#)
- [Sorted](#)
- [XML Join](#)

Data Warehouse

The PDI transformation steps in this section pertain to data warehouse functions.

- [Combination Lookup/Update](#)
- [Dimension Lookup/Update](#)

Validation

The PDI transformation steps in this section pertain to data validation.

- [Credit Card Validator](#)
- [Data Validator](#)
- [Mail Validator](#)
- [XSD Validator](#)

Statistics

The PDI transformation steps in this section pertain to statistics and analysis.

- [Analytic Query](#)
- [Group By](#)
- [Memory Group by](#)
- [Output Steps Metrics](#)
- [R script executor](#)
- [Reservoir Sampling](#)
- [Sample Rows](#)
- [Univariate Statistics](#)

Palo

The PDI transformation steps in this section pertain to interactivity with Palo business intelligence software.

- [Palo Cell Input](#)
- [Palo Cell Output](#)
- [Palo Dim Input](#)
- [Palo Dim Output](#)

Job

The PDI transformation steps in this section pertain to interactivity with a PDI job that is calling this transformation (a parent job).

- [Copy rows to result](#)
- [Get files from result](#)
- [Get rows from result](#)
- [Get Variables](#)
- [Set files in result](#)
- [Set Variables](#)

Mapping

The PDI transformation steps in this section pertain to value mapping.

- [Mapping Input Specification](#)
- [Mapping Output Specification](#)
- [Mapping \(sub-transformation\)](#)
- [Simple Mapping \(sub-transformation\)](#)

Bulk Loading

The PDI transformation steps in this section pertain to bulk loading of data.

- [Infobright Loader](#)
- [ElasticSearch Bulk Insert](#)
- [Greenplum Bulk Loader](#)
- [Greenplum Load](#)
- [Ingres VectorWise Bulk Loader](#)
- [LucidDB Streaming Loader](#)
- [MonetDB Bulk Loader](#)
- [MySQL Bulk Loader](#)
- [Oracle Bulk Loader](#)
- [PostgreSQL Bulk Loader](#)
- [Teradata Fastload Bulk Loader](#)
- [Vertica Bulk Loader](#)

Inline

The PDI transformation steps in this section pertain to inline data modification.

- [Injector](#)
- [Socket reader](#)
- [Socket writer](#)

Data Mining Steps

The PDI transformation steps in this section pertain to using Data Mining (Weka) plugins.

- [ARFF Output](#)
- [Knowledge Flow](#)
- [WEKA Forecasting](#)
- [WEKA Scoring](#)

Experimental

The PDI transformation steps in this section are experimental steps.

- [Script](#)
- [SFTP Put](#)

Deprecated

The PDI transformation steps in this section pertain to steps which have been deprecated.

- [Aggregate Rows](#)
- [Get previous row fields](#)
- [Google Analytics Input](#)
- [LucidDB Bulk Loader](#)
- [Streaming XML Input](#)
- [XML Input](#)

Job Entry Reference

This section contains reference documentation for job entries.

Note: Many entries are not completely documented in this section, but have rough definitions in the Pentaho Wiki: <http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Job+Entries>.

- [File Encryption](#)
- [Big Data](#)
- [General](#)
- [Mail](#)
- [File Management](#)
- [Conditions](#)
- [Scripting](#)
- [Bulk Loading](#)
- [XML](#)
- [Utility](#)
- [Repository](#)
- [File Transfer](#)
- [Palo](#)
- [Deprecated](#)

File Encryption

The PDI job entries in this section pertain to file encryption operations.

- [Decrypt files with PGP](#)
- [Encrypt files with PGP](#)
- [Verify file signature with PGP](#)

Big Data

The PDI job entries in this section pertain to Hadoop functions.

Note: PDI is configured by default to use the Apache Hadoop distribution. If you are working with a Cloudera or MapR distribution instead, you must install the appropriate patch before using any Hadoop functions in PDI. Patch installation is covered in [Data Integration Installation](#) and [Work with Big Data](#).

- [Amazon EMR Job Executor](#)
- [Amazon Hive Job Executor](#)
- [Hadoop Copy Files](#)
- [Hadoop Job Executor](#)
- [Oozie Job Executor](#)
- [Pentaho MapReduce](#)
- [Pig Script Executor](#)
- [Sqoop Export](#)
- [Sqoop Import](#)

General

The PDI job entries in this section pertain to general data integration functions.

- [DUMMY](#)
- [Example Plugin](#)
- [Job](#)
- [Set variables](#)
- [START](#)
- [Success](#)
- [Transformation](#)

Mail

The PDI job entries in this section pertain to email operations.

- [Get mails \(POP3/IMAP\)](#)
- [Mail](#)
- [Mail Validator](#)

File Management

The PDI job entries in this section pertain to file input/output operations.

- [Add filenames to result](#)
- [Compare folders](#)
- [Convert file between Windows and UNIX](#)
- [Copy Files](#)
- [Create a folder](#)
- [Create a file](#)
- [Delete a file](#)
- [Delete filenames from result](#)
- [Delete files](#)
- [Delete folders](#)
- [File Compare](#)
- [HTTP](#)
- [Move Files](#)
- [Process result filenames](#)
- [Unzip file](#)
- [Wait for file](#)
- [Write to file](#)
- [Zip file](#)

Conditions

The PDI job entries in this section pertain to conditional functions.

- [Check DB Connections](#)
- [Check Files Locked](#)
- [Check if a folder is empty](#)
- [Check webservice availability](#)
- [Checks if files exist](#)
- [Columns exist in a table](#)
- [Evaluate files metrics](#)
- [Evaluate rows number in a table](#)
- [File exists](#)
- [Simple evaluation](#)
- [Table exists](#)
- [Wait for](#)

Scripting

The PDI job entries in this section pertain to script execution.

- [JavaScript](#)
- [Shell](#)
- [SQL](#)

Bulk Loading

The PDI job entries in this section pertain to bulk loading of data.

- [BulkLoad from MySQL into File](#)
- [BulkLoad into MSSQL](#)
- [BulkLoad into MySQL](#)
- [MS Access Bulk Load](#)

XML

The PDI job entries in this section pertain to XML validation and XSL execution.

- [Check if XML file is well formed](#)
- [DTD Validator](#)
- [XSD Validator](#)
- [XSL Transformation](#)

Utility

The PDI job entries in this section pertain to a variety of special-case job operations.

- [Abort job](#)
- [Display MsgBox Info](#)
- [HL7 MLLP Acknowledge](#)
- [HL7 MLLP Input](#)
- [Ping a host](#)
- [Send information using Syslog](#)
- [Send SNMP trap](#)
- [Talend Job Execution](#)
- [Truncate tables](#)
- [Wait for SQL](#)
- [Write to Log](#)

Repository

The PDI job entries in this section pertain to PDI database or solution repository functions.

- [Check if connected to repository](#)
- [Export repository to XML file](#)

File Transfer

The PDI job entries in this section pertain to file transfer operations.

- [FTP Delete](#)
- [Get a file with FTP](#)
- [Get a file with FTPS](#)
- [Get a file with SFTP](#)
- [Put a file with FTP](#)
- [Put a file with SFTP](#)
- [Upload files to FTPS](#)

Palo

The PDI job entries in this section pertain to Palo databases.

- [Palo Cube Create](#)
- [Palo Cube Delete](#)

Deprecated

The PDI job entries in this section pertain to job entries which have been deprecated.

- [SSH2 Get](#)
- [SSH2 Put](#)

About PDI Marketplace

Use Marketplace to share, download, and install plugins developed by members of the user community or Pentaho. The Marketplace presents a list of plugins, indicates whether they have been installed, and displays other information about the plugin, such as where to obtain technical support. Access Marketplace from the **Help** menu in Spoon. To learn how to use Marketplace to install a plugin, see the [Install Only DI Tools](#) article. To learn more about Marketplace, [visit our wiki](#).

Troubleshooting

This section contains information about changing the Kettle Home directory. More troubleshooting tips will be added to this document in the future.

- [Changing the Pentaho Data Integration Home Directory Location \(.kettle folder\)](#)
- [Kitchen can't read KJBs from a Zip export](#)
- [Generating a DI Repository Configuration Without Running Spoon](#)
- [Unable to Get List of Repositories Exception](#)
- [Database Locks When Reading and Updating From A Single Table](#)
- [Force PDI to use DATE instead of TIMESTAMP in Parameterized SQL Queries](#)
- [PDI Does Not Recognize Changes Made To a Table](#)
- [Using ODBC](#)
- [Sqoop Import into Hive Fails](#)

Changing the Pentaho Data Integration Home Directory Location (.kettle folder)

The default Pentaho Data Integration (PDI) HOME directory is the user's home directory (for example, in Windows `C:\Documents and Settings\{user}\.kettle` or for all other *nix based operating systems `($HOME/.kettle)`). The directory may change depending on the user who is logged on. As a result, the configuration files that control the behavior of PDI jobs and transformations are different from user to user. This also applies when running PDI from the Pentaho BI Platform.

When you set the `KETTLE_HOME` variable, the PDI jobs and transformations can be run without being affected by the user who is logged on. `KETTLE_HOME` is used to change the location of the files normally in `[user home].kettle`.

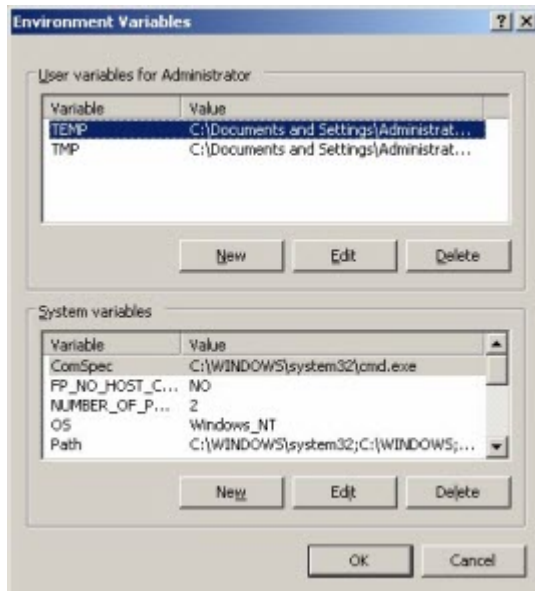
Note: The PDI home directory is independent of the PDI application directory.

Below is a short description of each item in the HOME directory:

Item	Description
<code>kettle.properties</code>	Default properties file for variables
<code>shared.xml</code>	Default shared objects file
<code>db.cache</code>	The database cache for metadata
<code>repositories.xml</code>	The local repositories file
<code>.spoonrc</code>	User interface settings, last opened transformation/job
<code>.languageChoice</code>	User language (delete to revert language)

To change set the `KETTLE_HOME` variable...

Step	Description
Step 1	<p>Set the <code>KETTLE_HOME</code> variable according to your needs. This can be performed system wide by the operating system or just before the start of PDI using a shell script or batch (for example, use the <code>SET</code> command).</p> <p>The <code>KETTLE_HOME</code> variable can be set system wide on Windows systems using the environment variables settings (see below):</p>



Step 2

Point the KETTLE_HOME to the directory that contains the .kettle directory. The .kettle gets appended by PDI. For example, when you have stored the common files in C:\Pentaho\Kettle\common\, .kettle you need to set the KETTLE_HOME variable to C:\Pentaho\Kettle\common).

The method above can also be used for configuration management and deployment to switch between the test, development, and production environments when other variables like the database server of the connection is defined in `kettle.properties`.

For testing purposes set a variable in the `kettle.properties` file of your defined .kettle home directory. Set the KETTLE_HOME directory accordingly by using the operating system SET command. Start Spoon and go to **Edit > Show Environment Variables**. You should see the variables defined in `kettle.properties`.

- [Changing the Kettle Home Directory within the Pentaho BI Platform](#)

Changing the Kettle Home Directory within the Pentaho BI Platform

You can set the KETTLE_HOME directory in the BA Server:

1. When started as a service, edit the registry: HKEY_LOCAL_MACHINE\SOFTWARE\Apache Software Foundation\Procrun 2.0\pentahobiserver\Parameters\Java
Note: On 64-bit systems, the Apache Software Foundation is under **Wow6432Node**.
2. Add a new line (**not a space!**) to the **Options** associated with the KETTLE_HOME variable, for example, –
Dcatalina.base=C:\Pentaho\3.0.1\Installed\server\biserver-ee/
tomcat

```
[...]  
-XX:MaxPermSize=256m  
-DKETTLE_HOME=C:\Pentaho\Kettle\KETTLE_HOME
```

3. Reboot the server.
4. When you start the BA Server from the command line, you must edit the ...server\biserver-ee\start-pentaho.bat (see below):

```
[...]  
set CATALINA_HOME=%PENTAHO_PATH%tomcat  
set CATALINA_OPTS=-Xms256m -Xmx768m -XX:MaxPermSize=256m -Dsun.rmi.dgc.  
client.gcInterval=3600000 -Dsun.rmi.dgc.server.gcInterval=3600000 -DKETTLE_  
HOME=C:\Pentaho\Kettle\KETTLE_HOME  
call startup  
endlocal  
:quit
```

Kitchen can't read KJBs from a Zip export

Note: This also applies to Pan and KTR files in Zip archives.

If you are trying to read a KJB file from a Zip export but are getting errors, you may have a syntax error in your Kitchen command. Zip files must be prefaced by a ! (exclamation mark) character. On Linux and other Unix-like operating systems, you must escape the exclamation mark with a backslash: \!

Windows:

```
Kitchen.bat /file:"zip:file:///C:/Pentaho/PDI_Examples/Sandbox/linked_executable_
job_and_transform.zip!Hourly_Stats_Job_Unix.kjb"
```

Linux:

```
./kitchen.sh -file:"zip:file:///home/user/pentaho/pdi-ee/my_package/linked_
executable_job_and_transform.zip\!Hourly_Stats_Job_Unix.kjb"
```

Generating a DI Repository Configuration Without Running Spoon

Because it is not always convenient to launch a repository export from Spoon, jobs and transformations can be launched using the command-line tools Kitchen and Pan.

To deploy a job from Kitchen or a transformation from Pan that will export a `repositories.xml` file, follow these instructions.

- [Connecting to a DI Solution Repositories with Command-Line Tools](#)

Connecting to a DI Solution Repository with Command-Line Tools

To export repository objects into XML format using command-line tools instead of exporting repository configurations from within Spoon, use named parameters and command-line options when calling Kitchen or Pan from a command-line prompt.

The following is an example command-line entry to execute an export job using Kitchen:

```
call kitchen.bat /file:C:\Pentaho_samples\repository\repository_export.kjb
"/param:rep_name=PDII2000" "/param:rep_user=admin" "/param:rep_
password=password"
"/param:rep_folder=/public/dev"
"/param:target_filename=C:\Pentaho_samples\repository\export\dev.xml"
```

Parameter	Description
rep_folder	Repository Folder
rep_name	Repository Name
rep_password	Repository Password
rep_user	Repository Username
target_filename	Target Filename

Note: It is also possible to use obfuscated passwords with Encr a command line tool for encrypting strings for storage or use by PDI.

The following is an example command-line entry to execute a complete command-line call for the export in addition to checking for errors:

```
@echo off
ECHO This an example of a batch file calling the repository_export.kjb

cd C:\Pentaho\pdi-ee-<filepath>--check--</filepath>{{contentVars.
PDIvernum3}}>\data-integration

call kitchen.bat /file:C:\Pentaho_samples\repository\repository_export.kjb
```

```
"/param:rep_name=PDI2000"
  "/param:rep_user=admin" "/param:rep_password=password" "/param:rep_
folder=/public/dev"
  "/param:target_filename=C:\Pentaho_samples\repository\export\dev.xml"

if errorlevel 1 goto error
echo Export finished successfull.
goto finished

:error
echo ERROR: An error occured during repository export.
:finished
REM Allow the user to read the message when testing, so having a pause
pause
```

Unable to Get List of Repositories Exception

When you are working with a repository and trying to execute a job or transformation remotely on a Carte server, the following error message often appears:

There was an error posting the transformation on the remote server:

org.pentaho.di.core.exception.KettleException:

Unable to get a list of repositories to locate repository 'repo'

- [Executing Jobs and Transformations from the Repository on the Carte Server](#)

Executing Jobs and Transformations from the Repository on the Carte Server

To execute a job or transformation remotely on a Carte server, you first need to copy the local `repositories.xml` from the user's `.kettle` directory to the Carte server's `$HOME/.kettle` directory. The Carte service also looks for the `repositories.xml` file in the directory from which Carte was started.

For more information about locating or changing the `.kettle` home directory, see [Changing the Pentaho Data Integration Home Directory Location \(.kettle folder\)](#).

Database Locks When Reading and Updating From A Single Table

If you create read and updated steps to or from a single table within a transformation you will likely experience database locking or slowed processing speeds.

For example, if you have a step which reads from a row within a table--a *Table Input* step--and need to update the step (with the *Update* step) this could cause locking issues.

Note: This is known to often cause difficulty with MS SQL databases in particular.

- [Reading and Updating Table Rows Within a Transformation](#)

Reading and Updating Table Rows Within a Transformation

Reading rows and updating rows on a table, within a single transformation, can cause the database to stop updating, referred to as locking, or slow down processing speeds.

Reading rows and updating rows in the same transformation on the same table should be avoided when possible as it is often causes these issues.

A general solution compatible with all databases is to duplicate the table to be read/updated, then create separate read/update steps. Arrange the steps to be executed sequentially within the transformation, each on a different, yet identical, version of the same table.

Adjusting database row locking parameters or mechanisms will also address this issue.

Force PDI to use DATE instead of TIMESTAMP in Parameterized SQL Queries

If your query optimizer is incorrectly using the predicate `TIMESTAMP`, it is likely because the JDBC driver/database normally converts the data type from a `TIMESTAMP` to a `DATE`. In special circumstances this casting prevents the query optimizer of the database not to use the correct index.

Use a `Select Values` step and set the `Precision` to 1 and `Value` to `DATE`. This forces the parameter to set as a `DATE` instead of a `TIMESTAMP`.

For example, if Oracle says it cannot use the index, and generates an error message that states:

```
The predicate DATE used at line ID 1 of the execution plan contains an implicit
data type conversion on indexed column DATE. This implicit data type
conversion prevents
the optimizer from selecting indices on table A.
```

After changing the `Precision` to 1, setting the `Value` as a `DATE`, the index can be used correctly.

PDI Does Not Recognize Changes Made To a Table

If PDI does not recognize changes you made to a table, you need to clear the cache of database-related meta information (field names and their types in each used database table). PDI has this cache to increase processing speed.

If you edit the table layout outside of Spoon, field changes, deletions or additions are not known to PDI. Clearing the cache addresses this issue.

To clear the database-related meta information cache from within Spoon:

Select the connection and then `Tools > Database > Clear Cache`. Or, `Database connections > Clear complete DB cache`.

Using ODBC

Although ODBC can be used to connect to a JDBC compliant database, Pentaho does not recommend using it and it is not supported. For details, this article explains "Why you should avoid ODBC."

<http://wiki.pentaho.com/pages/viewpage.action?pageId=14850644>.

Sqoop Import into Hive Fails

If executing a Sqoop import into Hive fails to execute on a remote installation, the local Hive installation configuration does not match the Hadoop cluster connection information used to perform the Sqoop job.

Verify the Hadoop connection information used by the local Hive installation is configured the same as the Sqoop job entry.