Pentaho Data Integration Installation Guide

Help and Support Resources

If you have questions that are not covered in this guide, or if you would like to report errors in the documentation, 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 on the topics covered in this guide, visit http://www.pentaho.com/training.

Limits of Liability and Disclaimer of Warranty

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

Pentaho (TM) and the Pentaho logo are registered trademarks of Pentaho Corporation. All other trademarks are the property of their respective owners. Trademarked names may appear throughout this document. Rather than list the names and entities that own the trademarks or insert a trademark symbol with each mention of the trademarked name, Pentaho 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.

Company Information

Pentaho Corporation
Citadel International, Suite 340
5950 Hazeltine National Drive
Orlando, FL 32822
Phone: +1 407 812-OPEN (6736)
Fax: +1 407 517-4575
http://www.pentaho.com

E-mail: communityconnection@pentaho.com

Sales Inquiries: sales@pentaho.com

Documentation Suggestions: documentation@pentaho.com

Sign-up for our newsletter: http://community.pentaho.com/newsletter/

Contents

Installation Overview

This guide explains how to install Pentaho Data Integration Enterprise Edition version 4.2 on both servers and workstations using either the standalone PDI graphical installer, or equivalent archive packages.

The PDI Installation Utility or Packages Will Provide	You Must Supply On Your Own
Data Integration server **Data Integration tools:** • Spoon (graphical interface) • Kitchen (CLI job interface) • Pan (CLI transformation interface) • Carte (CLI execution engine for PDI content) **Pentaho Enterprise Console** **A Java Runtime Environment 6.0** (only if you use the PDI graphical installer)	**An operating system:** • Linux (Red Hat Enterprise Linux 5, SUSE Linux Enterprise 10) • Windows (XP, 7) • Solaris 10 • Apple OS X (10.5 or newer) **A Java Runtime Environment 6.0** (only if you use the PDI archive packages) **One or more data sources:** • Any JDBC-compliant database • A spreadsheet • A flat file containing comma-separated values **A Hadoop cluster:** (for Hadoop deployments only) • Apache Hadoop 0.20.2 and Hive 0.7.0 • Cloudera CH3 Hadoop-0.20.2-cdh3u0 and hive-0.7.0-cdh3u0

There are two ways to install. There are 32-bit and 64-bit graphical installers available for both Windows and Linux, and archive packages that will deploy to Windows, Linux, OS X, and Solaris.

There are three installation paths. Different pieces of PDI will be installed on servers and workstations, unless you choose to run everything on one development machine. Typically, the Data Integration Server and Pentaho Enterprise Console will go on the server, and each workstation will have its own copy of the data integration tools. There are special installation instructions in this guide for installing PDI to a Hadoop node.

This is for Enterprise Edition only. You will need Enterprise Edition license keys for the BI Platform and Pentaho Data Integration in order to complete installation. If you are installing to a Hadoop node, you must have a Pentaho BI Suite For Hadoop license as well.

Installation Methods

There are two ways to install Pentaho Data Integration: through the graphical installer, and through archive packages. The installer provides a 32- or 64-bit Java Runtime Environment (depending on which version you download), a master control script for starting and stopping the PDI servers, a Start menu entry in Windows, automatic service configuration in Windows, and it auto-checks for available port numbers and reassigns them if there are port collisions.

The archive packages are more useful for **headless**, **unattended** (if you have your own deployment scripts), and **remote** installations. Installing from archives is the easiest and quickest way to deploy to one server and many workstations. If you perform an archive-based installation, you will have to manually configure your operating system to start the PDI servers as boot services. This guide provides basic instructions for these processes, but you will need to modify them to accommodate your operating environment.

Installation Checklist: Server

The Installation Checklist is a concise list of instructions intended to show a high-level overview of the installation and configuration process. It also serves as a quick reference for administrators and developers who have performed several installations in the past and only need a brief rehash of the necessary steps. If you need more details than are provided in this checklist, consult the appropriate section in the verbose instruction set that comprises the rest of this guide.

Step	Procedure	Done
Step 1	Download the DI Server archive package from the Enterprise Edition FTP site.	
Step 2	Create a `/pentaho/server/` directory in an appropriate location on your filesystem.	
Step 3	Unpack the DI Server and Pentaho Enterprise Console archive packages to `/pentaho/server/`.	
Step 4	Start the DI Server and Pentaho Enterprise Console.	
Step 5	Log into the Pentaho Enterprise Console, which by default is located at `http://localhost:8088`.	
Step 6	Install an Enterprise Edition key for Pentaho Data Integration.	
Step 7	Continue on to the workstation installation procedures.	

Installation Checklist: Workstation

The Installation Checklist is a concise list of instructions intended to show a high-level overview of the installation and configuration process. It also serves as a quick reference for administrators and developers who have performed several installations in the past and only need a brief rehash of the necessary steps. If you need more details than are provided in this checklist, consult the appropriate section in the verbose instruction set that comprises the rest of this guide.

Step	Procedure	Done
Step 1	Download either the PDI "all in one" installer or the client tool archive package from the Enterprise Edition FTP site.	
Step 2	Create a `/pentaho/design-tools/` directory in an appropriate location on your filesystem.	
Step 3	Unpack the PDI client archive package to `/pentaho/design-tools/`.	
Step 4	Install a PDI Enterprise Edition license key using the command line tool in the `/pentaho/design-tools/data-integration/license/` directory.	
Step 5	Start Spoon.	
Step 6	Create or connect to an enterprise repository.	
Step 7	Test the new PDI installation by creating, sharing, and scheduling new content.	
Step 8	Remove any temporary files, such as the original archive packages you installed from.	

Installation Checklist: Hadoop

The Installation Checklist is a concise list of instructions intended to show a high-level overview of the installation and configuration process. It also serves as a quick reference for administrators and developers who have performed several installations in the past and only need a brief rehash of the necessary steps. If you need more details than are provided in this checklist, consult the appropriate section in the verbose instruction set that comprises the rest of this guide.

Step	Procedure	Done
Step 1	Download the Pentaho Hadoop Distribution (PHD) archive package from the Enterprise Edition FTP site.	
Step 2	Stop the Hadoop service.	
Step 3	Navigate to your Hadoop root directory.	
Step 4	Unpack the Pentaho Hadoop Distribution archive package to the Hadoop root.	
Step 5	On your Hadoop nodes, install Enterprise Edition keys for Pentaho Data Integration and Pentaho Hadoop.	
Step 6	Start the Hadoop service.	
Step 7	Apache deployments must unpack the Hadoop patch archive to each Pentaho client tool and server. (Cloudera deployments do not require this step).	

Alternative: The "All-In-One" Graphical Installation Utility

Pentaho provides self-contained 32-bit and 64-bit graphical installation utilities for both the Windows and Linux platforms. These utilities will install the DI Server, Pentaho Enterprise Console, Java Runtime Environment, and Pentaho Data Integration client tools on a single machine. This makes it useful for evaluation or limited development purposes, but impractical for larger production deployments.

Consult the Welcome Kit email that was sent to you after completing the sales process. This email contains user credentials for the Enterprise Edition FTP site, and download instructions that tell you where to find the installer in the directory hierarchy. Here are the file names:

- **Windows 32-bit installer:** `PDI-4.2.0-GA-i386.exe`
- **Windows 64-bit installer:** `PDI-4.2.0-GA-x64.exe`
- **Linux 32-bit installer:** `PDI-4.2.0-GA-i386.bin`
- **Linux 64-bit installer:** `PDI-4.2.0-GA-x64.bin`

You can also find direct FTP links in the Pentaho Knowledge base in the **Enterprise Software** section. These links are particularly helpful in situations where you do not have access to an FTP client to navigate a directory structure; a direct link to a file on an FTP server can easily be downloaded with a Web browser, or by using a command line program such as wget.

Note: It is also possible to use the much larger BI Suite installation utility to install PDI as a server or client. However, the download is considerably larger, and the process takes a little longer.

Prerequisites

In order to install the Pentaho Data Integration server, you should be familiar with system administration operations pertaining to network services, including modifying your firewall to open specific ports, and adding services to the system startup and shutdown scripts. In most installation scenarios, this will require you to be comfortable using your operating system's command line interface and/or graphical system administration tools. You must have the ability to install software, open firewall ports, and start and stop system services on the machine you are installing on. If you don't have the requisite access or permissions, having the cooperation of someone who does (the sysadmin or DBA) is probably good enough.

Installing the Pentaho Data Integration client tools onto user workstations is a much simpler process that requires little more than being able to unpack a file archive or run the graphical installation utility as a regular user.

Software Requirements

Supported operating systems:

- **Windows:** XP SP2, 7
- **Linux:** SUSE Linux Enterprise Desktop and Server 10, Ubuntu 9.04 and 9.10, and Red Hat Enterprise Linux 5 are officially supported, but most others should work
- **Solaris 10**
- **OS X:** 10.5

Java

If you are installing from archive packages, you must have a Sun Java Runtime Environment (JRE) version 1.6 (6.0) installed. **The graphical installer will provide you with a JRE; the archive packages will not.**

Note: The GNU Compiler for Java, or GCJ for short, interferes with the way many native Java programs work on Linux, including some of the components of the Pentaho BI Suite. If you are using a Linux distribution that installs GCJ by default (which includes all of the most popular distros), then before you begin installation you must remove, disable, or circumvent GCJ. If you cannot remove it, you can simply ensure that your PENTAHO_JAVA_HOME variable is properly set (instructions for this are below), and add the Java Runtime Environment's /bin/ directory to the beginning of your PATH variable in ~/.bashrc or /etc/environment, then relog before continuing.

Supported Web browsers

Linux, Solaris, and other Unix-like operating systems must have either Mozilla Firefox 3.8 or newer, or the XulRunner library version 1.9 or higher in order to use the embedded Analyzer in Pentaho Data Integration.

64-bit support

Your environment can be either 32-bit or 64-bit as long as it meets the above requirements.

The aforementioned configurations are officially supported by Pentaho. Other operating systems such as FreeBSD and OpenBSD; other Java virtual machines like Blackdown; other application servers such as Liferay and Websphere; and other Web browsers like Opera may work without any problems. However, the Pentaho support team may not be able to help you if you have trouble installing or using the BI Suite under these conditions.

How to Check Your Java Version

The Pentaho BI Suite requires a Java Runtime Environment (JRE) or Java Development Kit (JDK) version 1.6 (sometimes referred to as 6.0). Follow the procedure below to see which version of Java is installed on your system and configured to be the default Java executable.

Note: There may be multiple JREs or JDKs on your system, but only one can be set as the global default. Any of these Java instances can be explicitly used for any Java program; if no specific Java executable is called, the default is used. Pentaho establishes a specific system variable named **PENTAHO_JAVA_HOME** to declare which Java instance it will use.

1. Open a terminal or command prompt window.
2. Type this command in: **java -version** and press **Enter**.

   Along with the Java version, the bit-ness (32-bit or 64-bit) and patch level will also show in the output.

```
java version "1.6.0_21"
Java(TM) SE Runtime Environment (build 1.6.0_21-b06)
Java HotSpot(TM) 64-Bit Server VM (build 17.0-b16, mixed mode)
```

Hardware Requirements

The Pentaho BI Suite does not have strict limits on computer or network hardware. As long as you meet the minimum software requirements (note that your operating system will have its own minimum hardware requirements), Pentaho is hardware agnostic. There is, however, a recommended set of minimum system specifications:

Server:

- **RAM:** at least 2GB
- **Hard drive space:** at least 2GB for the software, and more for solution and content files
- **Processor:** dual-core AMD64 or Intel EM64T

Workstation:

- **RAM:** at least 1GB
- **Hard drive space:** at least 1GB for the software, and more for solution and content files
- **Processor:** dual-core AMD64 or Intel EM64T

It's possible to use less capable machines, but in most realistic scenarios, the too-limited system resources will result in an undesirable level of performance.

Your environment does not have to be 64-bit, even if your processor architecture supports it; while all modern desktop, workstation, and server machines have 64-bit processors, they often ship by default with 32-bit operating systems. If you want to run the Pentaho BI Suite in a pure 64-bit environment, you will have to install a 64-bit operating system, ensure that your solution database and Java Runtime Environment are 64-bit, and install the BI Suite via the 64-bit graphical installer, or through the archive-based or manual deployment methods.

Note: A 32-bit JRE has a hard memory limit of 2GB (1.5GB on Windows), so if you have 2GB or more of RAM, you must use a 64-bit JRE on a 64-bit operating system to take full advantage of it. This means that, through architectural limitations, a 32-bit environment will likely be under-powered for a production BI Server deployment.

Obtaining the Archive Packages

Consult the Welcome Kit email that was sent to you after completing the sales process. This email contains user credentials for the Enterprise Edition FTP site, where you can download individual archive packages for the DI Server and Data Integration client tools. Here are the packages you need for each platform and distribution:

- **DI Server for Windows:** `pdi-ee-server-4.2.0-GA.zip`
- **DI Server for Linux/Solaris/OS X:** `pdi-ee-server-4.2.0-GA.tar.gz`
- **Data Integration client tool Windows package:** `pdi-ee-client-4.2.0-GA.zip`
- **Data Integration client tool Linux/Solaris/OS X package:** `pdi-ee-client-4.2.0-GA.tar.gz`
- **Data Integration for Hadoop:** `phd-ee-4.2.0-GA.zip`

If you will be using PDI (or any other Pentaho client tools) with an Apache Hadoop distribution, you will have to get a patch to replace Cloudera-specific Hadoop JARs:

- **Pentaho client tool patches for Apache Hadoop deployments:** `pentaho-apache-hadoop-4.2.0.zip`

If you download the **pdi-ee-server** package, you must also download the Pentaho Enterprise Console package:

- **Pentaho Enterprise Console for Linux/Solaris/OS X:** `pec-3.9.0-GA.tar.gz`
- **Pentaho Enterprise Console for Windows:** `pec-3.9.0-GA.zip`

Server Installation Procedure

To install and configure the Data Integration Server and Pentaho Enterprise Console from archive packages, follow the below procedures in the order they are presented. This process should take less than 20 minutes.

Server Archive Package Deployment

Follow the below instructions to install the Data Integration Server and Pentaho Enterprise Console on a dedicated server. If you intend to deploy the client tools and servers on one machine, you can combine this with *Workstation Installation Procedures* on page 13, or use the PDI graphical installer.

Note: The example commands in this and other sections are specific to Linux. You will have to adjust or ignore them on other operating systems.

1. Create a **/pentaho/server/** directory in an appropriate place in your hierarchy.

   This directory should be accessible to the system users who will be controlling services. Typically only root or the users in the wheel or administrator group will need to do this.

   Note: If you are using the graphical installer, it will create this directory structure for you, so you can skip this step.

   ```
   mkdir -p /home/pentaho/pentaho/server/
   ```
2. Unpack the **pdi-ee-server-4.2.0-GA** archive to `/pentaho/server/`.

   ```
   tar zxvf pdi-ee-server-4.2.0-GA.tar.gz -C /home/pentaho/pentaho/server/
   ```
3. Unpack the **pec-3.9.0-GA** archive to `/pentaho/server/`.

   ```
   tar zxvf pec-3.9.0-GA.tar.gz -C /home/pentaho/pentaho/server/
   ```
4. Switch to the `/pentaho/server/data-integration-server/` directory and run the **start-pentaho** script to start the DI Server.

   ```
   cd /home/pentaho/pentaho/server/data-integration-server/ && ./start-pentaho.sh
   ```
5. Switch to the `/pentaho/server/enterprise-console-server/` directory and run the **start-pec** script to start the Pentaho Enterprise Console.

   ```
   cd /home/pentaho/pentaho/server/enterprise-console-server/ && ./start-pec.sh
   ```

The DI Server and Enterprise Console are now installed, and should be operational. The DI Server will not be accessible from workstations until a license key is installed.

Post-Install Configuration

After you've installed PDI software to your server and workstations, you must perform some extra tasks to register license keys, connect your workstations to the server, and configure the server to start at boot time. Follow the sections below that apply to your situation.

Installing or Updating an Enterprise Edition Key

You must install Pentaho Enterprise Edition keys associated with products for which you have purchased support entitlements. The keys you install determine the layout and capabilities of the Pentaho Enterprise Console, and the functionality of the BI Server and DI Server. Follow the instructions below to install an Enterprise Edition key through the Pentaho Enterprise Console for the first time, or to update an expired or expiring key. If you would prefer to use a command line tool instead, see *Appendix: Working From the Command Line Interface* on page 11.

Note: If your Pentaho Enterprise Console server is running on a different machine than your BI or DI Server, you must use the command line tool to install and update license files; you will not be able to use the Pentaho Enterprise Console for this task.

Note: **License installation is a user-specific operation.** You must install licenses from the user accounts that will start all affected Pentaho software. If your BI or DI Server starts automatically at boot time, you must install licenses under the user account that is responsible for system services. If you have a Pentaho For Hadoop license, it must be installed under the user account that starts the Hadoop service as well as user accounts that run Pentaho client tools that have Hadoop functionality, and the account that starts the DI Server. There is no harm in installing the licenses under multiple local user accounts, if necessary.

1. If you have not done so already, log into the Pentaho Enterprise Console by opening a Web browser and navigating to `http://server-hostname:8088`, changing **server-hostname** to the hostname or IP address of your BI or DI server.
2. Click the **+** (plus) button in the upper right corner of the Subscriptions section.

   An **Install License** dialog box will appear.
3. Click **Browse**, then navigate to the location you saved your LIC files to, then click **Open**.

   LIC files for each of your supported Pentaho products were emailed to you along with your Pentaho Welcome Kit. If you did not receive this email, or if you have lost these files, contact your Pentaho support representative. If you do not yet have a support representative, contact the Pentaho salesperson you were working with.

   Note: Do not open your LIC files with a text editor; they are binary files, and will become corrupt if they are saved as ASCII.
4. Click **OK**.

   The Setup page changes according to the LIC file you installed.

You can now configure your licensed products through the Pentaho Enterprise Console.

Appendix: Working From the Command Line Interface

Though the Pentaho Enterprise Console is the quickest, easiest, and most comprehensive way to manage PDI and/or the BI Server, some Pentaho customers may be in environments where it is difficult or impossible to deploy or use the console. This appendix lists alternative instructions for command line interface (CLI) configuration.

Installing an Enterprise Edition Key on Windows (CLI)

To install a Pentaho Enterprise Edition Key from the command line interface, follow the below instructions.

Note: Do not open your LIC files with a text editor; they are binary files, and will become corrupt if they are saved as ASCII.

1. Navigate to the `\pentaho\server\enterprise-console\license-installer\` directory, or the `\license-installer\` directory that was part of the archive package you downloaded.
2. Run the **install_license.bat** script with the **install** switch and the location and name of your license file as a parameter.

   ```
   install_license.bat install "C:\Users\pgibbons\Downloads\Pentaho BI Platform
    Enterprise Edition.lic"
   ```

Upon completing this task, you should see a message that says, "The license has been successfully processed. Thank you."

Installing an Enterprise Edition Key on Linux (CLI)

To install a Pentaho Enterprise Edition Key from the command line interface, follow the below instructions.

Note: Do not open your LIC files with a text editor; they are binary files, and will become corrupt if they are saved as ASCII.

1. Navigate to the `/pentaho/server/enterprise-console/license-installer/` directory, or the `/license-installer/` directory that was part of the archive package you downloaded.
2. Run the **install_license.sh** script with the **install** switch and the location and name of your license file as a parameter. You can specify multiple files, separated by spaces, if you have more than one license key to install.

   Note: Be sure to use backslashes to escape any spaces in the path or file name.

   ```
   install_license.sh install /home/pgibbons/downloads/Pentaho\ BI\ Platform\ Enterprise\
    Edition.lic
   ```

Upon completing this task, you should see a message that says, "The license has been successfully processed. Thank you."

Starting the DI Server At Boot Time On Linux

This procedure assumes that you will be running your DI Server and Pentaho Enterprise Console server under the **pentaho** local user account. If you are using a different account to start these services, substitute it in the script below.

You can start and stop the DI Server manually at any time by running the **start-pentaho.sh** and **stop-pentaho.sh** scripts. To start the Tomcat server automatically at boot time, and stop automatically during shutdown, follow the below procedure.

1. With root permissions, create a file in `/etc/init.d/` called **pdi**.
2. Using a text editor, copy the following content into the new pentaho script, changing **mysql** to the name of the init script for your database if it is running on the remote machine, or remove **mysql** entirely if you are using a remote database. Secondly, you must adjust the paths to the DI Server and Pentaho Enterprise Console scripts to match your situation.

   ```
   #!/bin/sh -e
   ### BEGIN INIT INFO
   # Provides: pdi
   # Required-Start: networking
   # Required-Stop:
   # Default-Start: 2 3 4 5
   # Default-Stop: 0 1 6
   # Description: Pentaho DI Server
   ### END INIT INFO

   case "$1" in
   "start")
   su - pentaho -c "/home/pentaho/pentaho/server/data-integration-server/start-
   pentaho.sh"
   su - pentaho -c "cd /home/pentaho/pentaho/server/enterprise-console && ./start-pec.sh"
   ;;
   "stop")
   su - pentaho -c "/home/pentaho/pentaho/server/data-integration-server/stop-pentaho.sh"
   su - pentaho -c "cd /home/pentaho/pentaho/server/enterprise-console && ./stop-pec.sh"
   ;;
   *)
   echo "Usage: $0 { start | stop }"
   ;;
   esac
   exit 0
   ```
3. Save the file and close the text editor.
4. Make the init script executable.

   ```
   chmod +x /etc/init.d/pdi
   ```
5. Add the pdi init script to the standard runlevels so that it will run when the system starts, and stop when the system is shut down or rebooted, by using the update-rc.d command.

   This command may not exist on your computer if it is not Debian-based. If that is the case, consult your distribution documentation or contact your distribution's support department to determine how to add init scripts to the default runlevels.

   ```
   update-rc.d pdi defaults
   ```

The Pentaho DI Server will now start at boot time, and shut down when the system stops or restarts.

Workstation Installation Procedures

To install and configure the Data Integration client tools from the archive package, follow the below procedures in the order they are presented. This process should take less than 10 minutes. If you intend to deploy the client tools and servers on one machine, you can combine this with *Server Installation Procedure* on page 10, or use the PDI graphical installer.

Workstation Archive Package Deployment

Follow the below instructions to install the Data Integration client tools on your workstations. If you intend to deploy the client tools and servers on one machine, you can combine this with *Server Installation Procedure* on page 10, or use the PDI graphical installer.

Note: The example commands in this and other sections are specific to Linux. You will have to adjust or ignore them on other operating systems. If you need instructions for installing a license from a Windows command line, see *Installing an Enterprise Edition Key on Windows (CLI)* on page 11.

1. Create a **/pentaho/design-tools/** directory in an appropriate place in your hierarchy.

   If you are using the graphical installer, it will create this directory structure for you, so you can skip this step.

   ```
   mkdir -p /home/pgibbons/pentaho/design-tools/
   ```
2. Unpack the **pdi-ee-client-4.2.0-GA** archive to `/pentaho/design-tools/`.

   ```
   tar zxvf pdi-ee-client-4.2.0-GA.tar.gz -C /home/pgibbons/pentaho/design-tools/
   ```
3. Navigate to the `/pentaho/design-tools/license-installer/` directory.
4. Run the **install_license.sh** script with the **install** switch and as a parameter, the location and name of your license file.

   Note: You must also install the Pentaho Hadoop Enterprise Edition license if you are a Pentaho BI Suite For Hadoop customer and want to use the full Hadoop and Hive functionality in PDI.

   ```
   ./install_license.sh install /home/rwilco/downloads/Pentaho\ PDI\ Enterprise\
    Edition.lic
   ```

The Data Integration client tools are now installed.

Connecting to an Enterprise Repository

Follow the below instructions to create a new enterprise repository connection from a PDI workstation. By default, the DI Server comes with a preconfigured enterprise repository, so there is no need to create one -- only to connect to it.

1. Start Spoon by running the `/pentaho/design-tools/data-integration/spoon` script.

   The **Repository Connection** dialogue will appear.
2. Click the round green **+** icon in the upper right corner of the window.

   The **Repository Type** dialogue will appear.
3. Select **Enterprise Repository** in the list, then click **OK**.

   The **Repository Configuration** dialogue will appear.
4. Ensure that the **URL** field corresponds to your DI Server address and port number. Type in a system-identifiable value (a unique internal name for this repository instance) in the **ID** field, and a friendly name or description in the **Name** field.

   If you only intend to have one repository for all users, you can un-check the **Show this dialogue at startup** option before clicking OK. This will prevent the dialogue from appearing every time you start Spoon. If you need to make repository connection changes later, you can still get to this screen through the **Tools** menu.
5. Use the default credentials of **admin** and **secret** for this repository, and click **OK** to complete repository configuration.

   This account is part of the default PDI configuration. Refer to the *PDI Administrator's Guide* to learn more about setting up users and roles in PDI.

You are now connected to an enterprise repository, and are enabled to begin creating users and roles for your organization.

Hadoop Node Installation Procedure

These instructions assume that your Hadoop cluster is already properly configured and tested.

Note: You must be logged in as the user account that Hadoop is installed to. This user account must have a home directory.

Follow the directions below to install Pentaho Data Integration on a Hadoop node.

Note: In the examples below, your Hadoop root directory is represented as /hadoop. Adjust this and any other paths to match your configuration.

1. Stop the Hadoop service.

   ```
   ~/hadoop/bin/stop-all.sh
   ```
2. Unpack the **phd-ee-4.2.2-GA** archive to the Hadoop root directory.

   ```
   unzip phd-ee-4.2.0-GA.zip -d /hadoop
   ```
3. Navigate to the **license-installer** directory that was unpacked alongside the other files.

   ```
   cd /hadoop/license-installer/
   ```
4. Run the **install_license.sh** script with the sole parameter of the location and name of your license file.

   ```
   ./install_license.sh install ~/downloads/Pentaho\ PDI\ Enterprise\ Edition.lic ~/
   downloads/Pentaho\ Hadoop\ Enterprise\ Edition.lic
   ```
5. Start the Hadoop services.

   ```
   ~/hadoop/bin/start-all.sh
   ```

The Pentaho Data Integration libraries are now installed on this Hadoop node, and will remotely execute Hadoop jobs launched from a PDI workstation. Repeat this process for each node in the cluster.

Apache Patch Deployment

This procedure is only for Apache Hadoop deployments. Cloudera CDH3 deployments do not need to follow these instructions; doing so will make Pentaho client tools inoperable with CDH3.

Pentaho ships its client tools with default support for Cloudera CDH3 Hadoop deployments. In order to get Pentaho client tools to work properly with other supported Hadoop distributions, you must remove some Cloudera-specific JARs and replace them with Apache-specific equivalents. Follow the instructions below to accomplish this.

Note: This procedure covers patch deployment for all Hadoop-aware Pentaho software. If you don't have or are not using some of these programs, then skip those steps and follow only the ones that apply to you.

1. Exit any Pentaho client tools and stop the BI and DI servers if they are running.
2. Unpack the **pentaho-apache-hadoop-4.2.0** zip file to a temporary location.

   ```
   unzip pentaho-apache-hadoop-4.2.0.zip -d /home/pgibbons/temp/
   ```

   This package contains a set of patch archives for each affected Pentaho program.
3. For **PDI client tool** deployments, delete the `/pentaho/design-tools/data-integration/libext/pentaho/hadoop-core-0.20.2-cdh3u0.jar` file, then unpack the **pdi-client** archive to the **data-integration** directory.

   ```
   rm /pentaho/design-tools/data-integration/libext/pentaho/hadoop-core-0.20.2-cdh3u0.jar
    && unzip /home/pgibbons/temp/pentaho-apache-hadoop-4.2.0/pdi-client.zip -d /pentaho/
   design-tools/data-integration/
   ```
4. For **DI Server** deployments, delete the `/pentaho/server/data-integration-server/tomcat/webapps/pentaho-di/WEB-INF/lib/hadoop-core-0.20.2-cdh3u0.jar` file, then unpack the **pdi-server** archive to the **data-integration-server** directory.

   ```
   rm /pentaho/server/data-integration-server/tomcat/webapps/pentaho-di/WEB-INF/
   lib/hadoop-core-0.20.2-cdh3u0.jar && unzip /home/pgibbons/temp/pentaho-apache-
   hadoop-4.2.0/pdi-server.zip -d /pentaho/server/data-integration-server/
   ```

5. For **Report Designer** deployments, delete the `/pentaho/design-tools/report-designer/lib/hadoop-core-0.20.2-cdh3u0.jar` file, then unpack the **prd** archive to the **report-designer** directory.

   ```
   rm /pentaho/design-tools/report-designer/lib/hadoop-core-0.20.2-cdh3u0.jar && unzip /
   home/pgibbons/temp/pentaho-apache-hadoop-4.2.0/prd.zip -d /pentaho/design-tools/
   report-designer/
   ```

6. For **BI Server** deployments, delete the `/pentaho/server/biserver-ee/tomcat/webapps/pentaho-di/WEB-INF/lib/hadoop-core-0.20.2-cdh3u0.jar` file, then unpack the **bi-server** archive to the **biserver-ee** directory.

   ```
   rm /pentaho/server/biserver-ee/tomcat/webapps/pentaho-di/WEB-INF/lib/hadoop-
   core-0.20.2-cdh3u0.jar && unzip /home/pgibbons/temp/pentaho-apache-hadoop-4.2.0/bi-
   server.zip -d /pentaho/server/biserver-ee/
   ```

7. For **Metadata Editor** deployments, delete the `/pentaho/design-tools/metadata-editor/libext/JDBC/hadoop-core-0.20.2-cdh3u0.jar` file, then unpack the **pme** archive to the **metadata-editor** directory.

   ```
   rm /pentaho/design-tools/metadata-editor/libext/JDBC/hadoop-core-0.20.2-cdh3u0.jar
    && unzip /home/pgibbons/temp/pentaho-apache-hadoop-4.2.0/pme.zip -d /pentaho/design-
   tools/metadata-editor/
   ```

The Cloudera Hadoop JARs have now been replaces with Apache-specific versions.

Adding PDI Enterprise Repository Content Support to the BI Server

If you are using a Pentaho Data Integration (PDI) enterprise repository (through a Data Integration Server) to store PDI jobs and transformations, and you plan on using those jobs and transformations in action sequences that will be run on the BI Server, you must install some BI Server plugins from the PDI client tool package. This is not a typical scenario, but there is no harm in performing it if you aren't sure of the details.

1. Download a PDI Enterprise Edition 4.2.0 client tool archive package from the Pentaho Knowledge Base or Enterprise Edition FTP Site.

   The package name (available in both tar.gz and zip formats) is: **pdi-ee-client-4.2.0-GA**
2. Unpack the archive to a temporary location.
3. Edit the `/pentaho/server/biserver-ee/pentaho-solutions/system/kettle/settings.xml` file.
4. Change the value of the **<repository.type>** node from **files** to **rdbms**.
5. Enter your enterprise repository connection information in the proper nodes.
6. Enter the location of your local **repositories.xml** file in the **<repositories.xml.file>** node.

   Note: This file is created on your PDI client workstation when you establish a connection to an enterprise repository. Once you have made all of your repository connections on a workstation, copy the **repositories.xml** file to the `~/.kettle/` directory on the BI Server and DI Server machines. If the client tool and servers are all on the same machine, you do not have to copy the file. If you have not yet established any repositories, you will have to revisit this procedure later when your PDI environment is fully configured.
7. Copy the contents of `/data-integration/plugins/` to the `/pentaho/server/biserver-ee/pentaho-solutions/system/kettle/plugins/` directory.

   ```
   cp -r /tmp/data-integration/plugins/* /home/pentaho/pentaho/server/biserver-ee/
   pentaho-solutions/system/kettle/plugins/
   ```
8. Remove the unpacked archive.

   ```
   rm -rf /tmp/data-integration/
   ```

Your BI Server is now configured to

Adding a JDBC Driver

Before you can connect to a data source in any Pentaho server or client tool, you must first install the appropriate database driver. Your database administrator, CIO, or IT manager should be able to provide you with the proper driver JAR. If not, you can download a JDBC driver JAR file from your database vendor or driver developer's Web site. Once you have the JAR, follow the instructions below to copy it to the driver directories for all of the BI Suite components that need to connect to this data source.

Note: Microsoft SQL Server users frequently use an alternative, non-vendor-supported driver called JTDS. If you are adding an MSSQL data source, ensure that you are installing the correct driver.

Backing up old drivers

You must also ensure that there are no other versions of the same vendor's JDBC driver installed in these directories. If there are, you may have to back them up and remove them to avoid confusion and potential class loading problems. This is of particular concern when you are installing a driver JAR for a data source that is the same database type as your Pentaho solution repository. If you have any doubts as to how to proceed, contact your Pentaho support representative for guidance.

Installing JDBC drivers

Copy the driver JAR file to the following directories, depending on which servers and client tools you are using (Dashboard Designer, ad hoc reporting, and Analyzer are all part of the BI Server):

Note: For the DI Server: before copying a new JDBC driver, ensure that there is not a different version of the same JAR in the destination directory. If there is, you must remove the old JAR to avoid version conflicts.

- **BI Server:** `/pentaho/server/biserver-ee/tomcat/lib/`
- **Enterprise Console:** `/pentaho/server/enterprise-console/jdbc/`
- **Data Integration Server:** `/pentaho/server/data-integration-server/tomcat/webapps/pentaho-di/WEB-INF/lib/`
- **Data Integration client:** `/pentaho/design-tools/data-integration/libext/JDBC/`
- **Report Designer:** `/pentaho/design-tools/report-designer/lib/jdbc/`
- **Schema Workbench:** `/pentaho/design-tools/schema-workbench/drivers/`
- **Aggregation Designer:** `/pentaho/design-tools/agg-designer/drivers/`
- **Metadata Editor:** `/pentaho/design-tools/metadata-editor/libext/JDBC/`

Note: To establish a data source in the Pentaho Enterprise Console, you must install the driver in both the Enterprise Console and the BI Server or Data Integration Server. If you are just adding a data source through the Pentaho User Console, you do not need to install the driver to Enterprise Console.

Restarting

Once the driver JAR is in place, you must restart the server or client tool that you added it to.

Connecting to a Microsoft SQL Server using Integrated or Windows Authentication

The JDBC driver supports Type 2 integrated authentication on Windows operating systems through the **integratedSecurity** connection string property. To use integrated authentication, copy the **sqljdbc_auth.dll** file to all the directories to which you copied the JDBC files.

The **sqljdbc_auth.dll** files are installed in the following location:

```
<installation directory>\sqljdbc_<version>\<language>\auth\
```

Note: Use the **sqljdbc_auth.dll** file, in the x86 folder, if you are running a 32-bit Java Virtual Machine (JVM) even if the operating system is version x64. Use the **sqljdbc_auth.dll** file in the x64 folder, if you are running a 64-bit JVM on a x64 processor. Use the **sqljdbc_auth.dll** file in the IA64 folder, you are running a 64-bit JVM on an Itanium processor.

Testing and Cleanup

You should now have a complete PDI environment, from the DI Server to individual client workstations. Before you go into production, you should perform the following tests:

- Ensure that each PDI workstation has an enterprise repository connection.
- Create a new job, transformation, and/or Agile BI analysis schema.
- If you installed PDI to Hadoop nodes, ensure that you can run Hadoop jobs on them.
- Schedule a job or transformation and ensure that the schedule executes properly.
- From one PDI user workstation, share a job or transformation with another user and verify that this other user can access it.
- Physically restart the server and ensure that the DI Server and Enterprise Console are automatically started as services.

Once you're certain that your PDI environment is ready for production, you can remove any installation artifacts, such as ZIP or tar.gz archives and installers.

Note: You should now continue on to the *PDI Administrator's Guide* for further guidance on system administration, configuration, and maintenance.