# Table of Contents

1. Introduction to Pentaho.............................................................................................................................................. 0
   1a. High level overview of abilities of Pentaho: ........................................................................................................ 2
   1b. Introduction to tutorial on Pentaho Data Integration (Kettle): .............................................................................. 3
2. Pentaho – Differences between Community and Enterprise Edition .............................................................................. 3
3. Market Share -> Pentaho Vs. Other Competitors: ........................................................................................................ 4
4. Advantages of Pentaho.................................................................................................................................................. 5

5. Pentaho Data Integration Tutorials
   5a Getting Started with Pentaho - Download and Installation......................................................................................... 6
   5b Connecting to MYSQL Database.................................................................................................................................. 8
   5c Transformation: Excel to MYSQL............................................................................................................................... 9
   5d Transformation- XML to RSS 2.0 WXR WordPress....................................................................................................... 12

6. FAQ................................................................................................................................................................................. 15

7. References..................................................................................................................................................................... 18
1. Introduction to Pentaho

Pentaho is powerful business intelligence (BI) tool that brings together IT and business users to access, integrate, blend, visualize, and analyze all data that impacts business results. The suite offers variety of features such as reporting, OLAP pivot tables, dash boarding, analysis, data mining, workflow capabilities and more. The BI suite helps to create solution for real world problems. Pentaho is available to its users in two editions

- Community Edition
- Enterprise Edition

The enterprise edition mainly serves the corporate world with finished product and is a paid edition. The BI Suite we are dealing with is community edition, open source software where we are free to use and distribute its programs. Pentaho is programmed in Java programming language and needs at least Java 1.5 versions installed in the workstation for the software to process. The software programmed in JAVA makes it platform independent and is compatible with Windows, MAC, Linux, Ubuntu, AIX, and Solaris compatible.

1a. High level overview of abilities of Pentaho:

Data Integration: It helps to access manage and blend data from different data sources. It also offers the ability to obtain data in different format and convert data from one format to other.

Business Analytics: Turn the data into insights, process the information and analyze the data that helps to make information driven decisions

Shrey Hatle  Ashiqa Sayeed  Deepti Gupta  Neela G P
Pentaho Data Integration Tool

Reporting: Helps to access and format data from various sources in various formats. We can create, manage and distribute through rich graphical report designer and deliver reports through web or e-mail.

Process Management: Helps to integrate processes from different sources, identify each of the processes individually and execute the processes

1b. Introduction to tutorial on Pentaho Data Integration (Kettle):

Our tutorial mainly concentrates on the abilities of Pentaho in Data Integration section referred as Kettle by Pentaho developers. Pentaho Data Integration is a part of Pentaho Studio that delivers powerful Extraction, Transformation, and Loading (ETL) capabilities using meta-data driven approach. Kettle is an acronym for "Kettle E.T.T.L. Environment". It provides intuitive, graphical, drag and drop design environment that has scalable standard based architecture.

Data Integration consists of these packets:

- Spoon: Graphical interface to design transformations and jobs that can be run using kettle tools
- Kitchen: Program that executes jobs designed by spoon in XML or in database repository
- Pan is a data transformation engine that performs functions of reading, manipulating and writing data to and from various data sources
- Carte is a web server to execute the transformations and jobs remotely. It is also referred as slave server.

Transformations and jobs can describe themselves using a XML file or can be put in Kettle database repository. Pan or Kitchen can then read the data to execute the transformation or to run the job. On the whole, PDI makes data warehouses easier to build, maintain and update.

2. Pentaho – Differences between Community and Enterprise Edition

Pentaho Enterprise Edition differs in 4 main additional features compared to Pentaho community version

- Interactive Reporting: The tool offers ability to perform reporting of the information on the fly. If we have data set or information available, we can create visualization of the information or data. This visualization can be shared across with report summary/notes across team or to higher authorities. It offers the benefit of on the fly visualization of the information and sharing.
- Analyzer: The visualized and analyzed data can be obtained from the analyzer. The tool offers the features of drill through, drill down, and filtered analyzing capability of the data through the analyzer.
Pentaho Data Integration Tool

- The Pentaho data integration commercial tool offers lot more powerful features compared to the open source. It provides option for scheduling, management, timing of the reports created.

- The mobile version of the tool is also available for enterprise edition which is compatible with mobile and tablets which can be downloaded and complete functionality can be available.

- Dashboard Designer: It helps to pick data from various platforms, visualization created, and reports generated drag to a common dashboard and perform design operation and comprehensive analysis on these reports and data visualizations

- Big Data Application Support: There are numerous applications used to connect and integrate to the data sources such as map reduce. Pentaho offers support to all the big data application which connect to these sources and intermediate applications.

- Open source version is pool of good innovative ideas. The enterprise edition has a professional feel to turn it into product like to meet the requirements of business standards.

3. Market Share -> Pentaho Vs. Other Competitors:

One of the established competitors of Pentaho is SAS. Pentaho poses a competitive challenge to the well-established product SAS over the past few years. Below is the diagram showing Google trends of SAS Vs. Pentaho over the last decade.
The open source product is presenting challenges in both established and emerging market. We can expect the growth of old guard companies to be rather flat. Below is the list of products, market share, and their expected growth

- Business Objects (SAP): -0.2% market share growth last year
- Hyperion (Oracle): 2.3% market share growth last year
- SAS: 2.7% market share growth last year
- MicroStrategy: -6.4% product license revenue growth Q3 2010 from Q3 2009
- Pentaho is on target for 150% growth this year

4. **Advantages of Pentaho:**
- Pentaho is an open source software with open licensing
- Easy upgrades
- Software is easy to use and is completely transparent
- Has a reputation of 100% customer satisfaction in providing support, training, consultation, quality assurance, product management and maintenance.
- Extensive QA provides higher development productivity and quality products; High quality Enterprise Development Methodology
- 24x7 Enterprise Support through a subscribed service
- 20% of developers time are allocated to provide customer service and support
- Continuous testing in diverse environments by a huge global community
- Supports embed ability and service oriented architectures
- Highly flexible and is easy to customize
- Supports AJAX and web services
- Supports J2EE,JDBC,MDX,SQL architectures and frameworks
- Platform free and does not involve legacy architecture and migration issues
- Enables to execution of multiple jobs and transformations as part of the same database transaction
5. Pentaho Data Integration Tutorials

5a. Getting started with Pentaho – Downloading and Installation

In our tutorial, we will explain you to download and install the Pentaho data integration server (community edition) on Mac OS X and MS Windows. Please follow the below steps for instructions:

1. To get the Pentaho 5.0.1 stable edition, you can go to community.pentaho.com and click on Download Pentaho CE. This will direct you to the download section under Data Integration. Click on Download.

2. We can also download other versions of Pentaho data integration server. To do so follow the below steps:
   a. Click on Other artifacts
   b. This will redirect you to this site
   c. A click on the Parent folder, will give you options to download various versions of the DI (Data Integration) server.
   d. Click on 4.4.0-stable and then download the pdi-ce-4.4.0-stable.zip
   e. Once the download is complete, extract this folder.
   f. The installation is different for Mac and Windows users:

<table>
<thead>
<tr>
<th>Microsoft Windows</th>
<th>MAC OS X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open the folder once it’s extracted and run the Spoon.bat file to start Pentaho</td>
<td>1. Open Item</td>
</tr>
<tr>
<td>2. Go to the folder where you just downloaded Pentaho. (../Downloads/pentaho/data-integration)</td>
<td>2. Go to the folder where you just downloaded Pentaho. (../Downloads/pentaho/data-integration)</td>
</tr>
<tr>
<td>3. Start the Pentaho Spoon using the following</td>
<td>3. Start the Pentaho Spoon using the following</td>
</tr>
</tbody>
</table>
command - "./spoon.sh"

4. This will open a Repository connection screen. Just click on Cancel

5. This will direct you to the Pentaho DI (Spoon)

To see a full length video on downloading and installing please visit: https://www.youtube.com/watch?v=kSSxCJp2J84

After Logging in

Once you start the server you can start your Internet browser to connect to the server. All major browsers like Microsoft Internet Explorer, Google Chrome, Firefox or Apple Safari can be used.

You can access Pentaho by http://localhost:8080/pentaho/ and this will direct you to http://localhost:8080/pentaho/Login

5b. Connecting to a MySQL Database

Pentaho metadata architecture supports a wide range of data sources. To define a business model, we must describe the database or that data source that we would like to model. We can do this by defining one or more connections in the editor.

To connect to a database, we need to understand the below information first in the connection information dialog box:

Shrey Hatle    Ashiq Sayeed    Deepti Gupta    Neela G P
Connection Types – is a list of database connections that the Pentaho Editor can support.

Method of Access - Under Connection Type, you will see a list of Method of Access. You can either select JDBC or OBDC. JNDI method is used when one would be abstracting the information on the general tab from the metadata domain.

Example Connection: JDBC Connection to MySQL Database

1. In the Pentaho DI console, under View, you can see Database connections under Transformations.
2. Click on Database connection and a Connection dialog will open.
3. You can type in Connection Name, let’s name it Mysql_test.
4. From the list of connection type, select MySQL.
5. Select Native (JDBC) as the access method.
6. Enter localhost as the server host name assuming you have a local Pentaho BI Server running. If the server (or the sample database) is hosted elsewhere, enter the name of that host here.
7. Enter the name of the database, in our example, hatlesh_reciepes.
8. Add username and password for your database.
9. You should now click the Test button at the bottom of the dialog. If there is no error in the connection, you should receive a “Test OK” message. If not, there would be a prompt with a proper error message. Make sure your MySQL service through the XAMPP.

For a detailed video, please visit: https://www.youtube.com/watch?v=8fCBpc-w2H8

5c. Transformation – EXCEL to MySQL

To do the transformation, we need:

An excel sheet containing data to be inserted into MySQL and a table in the database into which we plan to insert the data

Follow these steps to load data into MySQL from an excel sheet using Pentaho spoon. We are using the 4.8.0 stable version to demonstrate the transformation.

1. Create a table in the MySQL database. You can use any tool to do this. We are using MySQL-Workbench to do the same. In our example, we are using a table ‘players’ with columns ‘id’, ‘First_Name’, ‘Last_Name’ and ‘Country’.
2. Then, open an Excel sheet and write the data you want to push into the database. The first row in Excel can be used as the headers in Pentaho. For a better understanding, we will match two headers in Excel to the column name in the players table. This is to demonstrate a ‘guessing’ and a ‘sql modification’ feature in Pentaho.

3. Once you have inserted the data into the Excel sheet, save the file (.xlsx or .xls format).

4. Next, open Pentaho. Once you are in, start a new transformation. In the input folder, drag and drop the Microsoft Excel input icon onto the canvas. Double click on it and give it a step name. It is extremely important to give names to steps. It helps to understand a complex transformation once we are done and also helps someone who will work on it the next time.

5. Go to the file and directory and add the file.

6. Note that, we have created an xlsx file and not an xls file. Make sure you select the right spread type engine in content. By default it is the 97-2003 engine but we want it for 2007 xlsx. So select excel 2007 xlsx (apache). (In 5.0.1, this is present in the add file tab itself).

7. Next, go to sheets and here you can check the sheets from which you want to read the data. If we are reading data from one single Excel sheet, this step is not required. This step should be followed when you are working with multiple sheets. At any point, Pentaho will keep warning you what you need to add so keep a lookout for that.
8. Now go to the **Fields** tab. Here we name the fields. The ‘get fields from header rows’ is self-explanatory and it gets the fields from the first row of the excel file. You may preview rows to see the output of this step.

9. Go to the output folder and select **Table Output**. For the table output, you would need the mysql connection to the table we just created (players in our example).

10. Once you have that set up. Double click on the **Table Output** step and click on **Database fields**. Click on enter **field mapping**. Here you will map your excel columns to your database table. Now you see that the source columns are missing, that is because we have not yet connected the two steps. Let’s connect the two steps. Hover the mouse over the Microsoft excels input step, click here and drag to the other step. Now go to database field again. Now you can see that you have source fields.

11. You can manually add mappings or you can just click ‘**guess**’ and pentaho will map them for you. (This is a really cool feature because you don’t need to keep adding all the transformations in). After we click guess we see that the ‘First_Name’ and ‘Last_Name’ is automatically mapped but ‘Country’ and the ‘Player_ID’ was not mapped. You can manually map this. Just click on ‘Nationality’ in source field, click on ‘Country’ in the target field and then click **ADD**. Click **OK**, save your transformation.
12. Now run the transformation.

So that’s how you load data from an excel file into MySQL database using pentaho. You can also check out the video at https://www.youtube.com/watch?v=RDxdy_8mOa0.

5d. Transformation- XML to RSS 2.0 WXR WordPress

We will show to read data from an xml file and convert it into WordPress xml. We are using the output as WordPress WXR xml for a reason. It is actually an RSS 2.0 type xml but with some modifications. This would help us to explain multiple things.

Select the XML you want to transform (In our example we are using nodes_inputXML.xml file). This has root element nodes and all the child elements are node. This is an old school blog in xml which we got from the University of Maryland Libraries.

1. Open pentaho spoon and search for XML using the search tab on the left. This will highlight everything that is related to xml. For our transformation we need the ‘Get data from XML’ as our input. For the output search for ‘rss output’. Just link them from input to output. You may name them as you want.

2. Now there are two ways you can give in the path for the input files. You can double click on the ‘get data from xml’ icon and then add the xml file there (/Users/ShreyHatle/Box Documents/Lib-ssdr/pentaho/inputxml/univarch_exhibits/nodes_inputXML.xml) or you can take the input from a previous step. This may happen when this step is part of a big transformation. In our example, we will just add the
xml but if you want to take this as a previous step you can use this tab over here and select the XML source is defined in a field check box.

3. Next, click on content. This section uses xpath to iterate over nodes and get data from it. You can click on the ‘Get Xpath node’ button and this will give you a list of all the nodes it finds. For our input xml, we need the xpath to loop on my node element.

4. Next go to the fields tab and click on get fields. This will pull in all the fields that pentaho finds in the xml. Now you can map the fields you want. For example, the creation date needs to be mapped to pubDate, the title would be the title and the data/body would be mapped to description. Put in all the mappings.

5. Once you have all the mappings, you can see that if we want attribute, we select Element type as attribute OR you can also put @created in the xpath and select type as Element. You can click on preview rows to see how it pulls out the data.

6. Next, for an RSS feed the type may be a ‘blog_post’ or an image, we want to change them to ‘post’ or an ‘attachment’. You can do this by adding a value mapper. This will just check the source value and replace the field with the target value. Just mention the field type you need.
7. For WordPress xml, like we said earlier, it has some custom tags. We need some constants like the wp:wxr_version number, wp: base_site_url etc. You can add them in. For this step, we can use the add constants step and put in all the tags that you would need.

8. Next pull in the RSS 2.0 output and open it. On the main Channel tab keep the encoding as utf-8, the version as 2.0. You can also select the channel title field, description and the link field and it will generate an rss 2.0 xml for you. And then in the item tab you can select individual item title field, description, link, pubdate (Marc Batchelor, 2013) etc.

9. But in this example, we are going to convert the input xml into a WordPress xml which is a custom modified xml form of rss 2.0. For this, first check the ‘create custom RSS’ box. This will activate the custom output tab. Once you are in the Custom Output Tab, we need to just add the tags and the field values (which we get from the previous steps). Do this for both the item fields and the channel fields. 

**REMEMBER** - That it will only be activated once you check the custom RSS output.
10. One thing that you will notice in this is that you cannot add any attributes to a field or you cannot map inner tags. Like for e.g. in the WordPress xml they have a `<wp: comment>` tag and it contains `<wp: content: encoded>`, `<wp: comment_author>` tags. We cannot replicate using this.

11. Next, the custom namespace tags allow you to provide the necessary namespaces. Ok, so once we add these namespaces, click on the output file tab and give in the filename. You can also click on filename defined in a field and get the path from a previous step.

12. Click **OK** now and you are done.

13. **Run** this transformation and check the output.

14. You can now see that the output xml has all the custom tags that we wanted.

For a detailed video, please visit: [http://www.youtube.com/watch?v=ZcQqtqf04aE](http://www.youtube.com/watch?v=ZcQqtqf04aE)
6. FAQs

Q1: How to start spoon?
A: Traverse to the integration folder of your installed pdi-ce-4.4.0-stable and double click on spoon.bat to start the application.

Q2: How to troubleshoot if the application doesn't start.
A: Edit the Spoon.bat file and follow the steps below.
   1. Replace in the last line "start javaw" with only "java"
   2. Add a "pause" in the next line
   3. Save and try it again.

Q3: How to overcome the following error:
"Could not find the main class. Program will exit".
A: Add the kettle jars to the classpath.

Q3: How to overcome the following error:
"Exception in thread "main" java.lang.NoSuchMethodError: method java.lang.Class.asSubclass with signature (Ljava.lang.Class;)Ljava.lang.Class; was not found".
A: Download the Java 5 version and recheck the path if the correct version is added to it. Alternatively set the path (set PATH) in the Spoon.bat file

Q4: In a transformation, how to use database connections from repository?
A: Create a new transformation or close and re-open the ones you have loaded in Spoon.

Q5: What are value, row, input stream, hop and note in a transformation?

Value – They are part of a row and can contain any type of data: strings, floating point numbers, unlimited precision BigNumbers, integers, dates or boolean values

Row - A row consists of 0 or more values that are processed together as a single entry.

Input stream - A stack of rows that enters a step
**Hop** - A graphical representation of one or more data streams between two steps. It always represents the output stream for one step and the input stream for another.

**Note** - Descriptive text that can be added to a transformation

**Q6: What are Job Entry, hop and note in a transformation?**

**A:** *Job Entry* – A part of a job that performs a specific task

**Hop** - A graphical representation of one or more data streams between two steps. It always represents the output stream for one step and the input stream for another.

**Note** - Descriptive text that can be added to a job

**Q7: How to copy fields in a row in a job?**

**A:** Use a "Select Values" step renaming a field while selecting also the original one or Use a calculator step and use NLV operation or Use a JavaScript step to copy the field

**Q8: Where to find safe mode option?**

**A:** The safe mode option is available in the Spoon logging window or in the Execute a transformation/Job window.

**Q9: How to integrate database with PDI?**

**A:** 2 tables in same DB: Use a "Table Input" step and do the join in SQL itself

2 tables in different DB: Use the "Database Join" step.

**Q10: Using JNDI in PDI?**

**A:** Edit "jdbc.properties" file in your parent pdi-ce-4.4.0-stable installed folder. Make appropriate changes so the JNDI information matches the application server information.

**Q11: How to check for the return code of a shell script/batch file?**

**A:** The Shell script considers a return code of 0 to mean success. Use hops to control the resulting flow.

**Q12: How to enable display tool tip feature?**

Shrey Hatle          Ashiqa Sayeed          Deepti Gupta          Neela G P
A: To enable the option, open options dialog box by selecting Edit > Options from the menu bar and check the ‘display tool tip’

Q13: How to resolve the following error?

“Kettle.ERROR_0019 - RDBMS access to transformation not allowed when repository type is 'files’”

A: open the settings.xml file located in the pentaho-solutions/system/kettle directory

1. Set the content of the repository.type element to rdbms. If you want to use a PDI repository.
2. Put the name of the repository you want to use in the repository.name element. Same as existing repository defined in PDI’s repositories.xml file.
3. Set the repository.name element to the name of the repository user that is to connect to the repository.
4. Set the repository.password element to the password of the repository user.

Q14: How to switch from file-based repository storage to a database-based repository?

A: Make a backup of the repository.properties file located in the jdbc directory.

1. Overwrite the original repository.properties file with a copy of the RDBMS-specific.properties file of choice.
2. Open the modified repository.properties file and edit it to point it to your database. Provide values for a number of properties. The names of these properties all start with MDRStorageProperty.org.netbeans.mdr.persistence.jdbcimpl. This prefix is followed by a dot and a name that configures a property of a JDBC connection. Typical property names (without the prefix) are:
   - driverClassName: theJavaclassnameofthedriver
   - url: the JDBC connection string
   - userName: thenameofthedatabaseuser
   - password: thepasswordofthedatabaseuser
7. References


