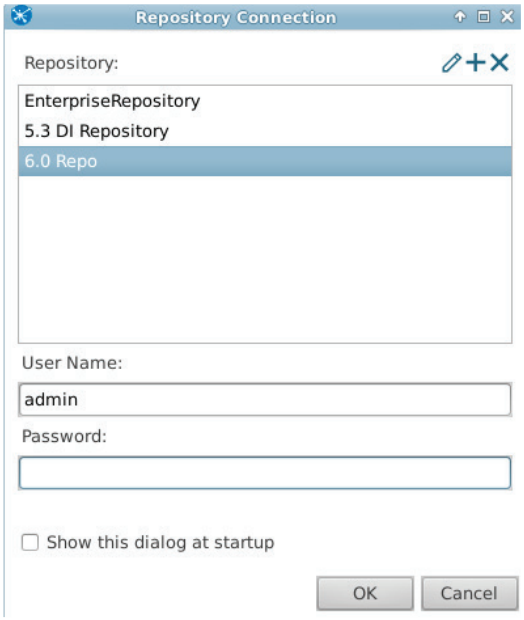
HITACHI
Inspire the Next

Pentaho Data Integration Build Transformations

Pentaho Data Integration, or PDI, is a comprehensive Data Integration platform allowing you to access, prepare, analyze and immediately derive value from both traditional and big data sources. During this lesson, you will be introduced to PDI's graphical design environment, Spoon. You will learn how to create your first transformation to load sales transaction data from a CSV file into an H2 database.

If this is your first time launching PDI you will be prompted to login to the Enterprise Repository. The repository allows you to store and share data integration jobs and transformations and provides access to Enterprise Edition features that will not be covered in this video such as document revision history and scheduling. You can click 'Cancel' to continue on to the design environment.

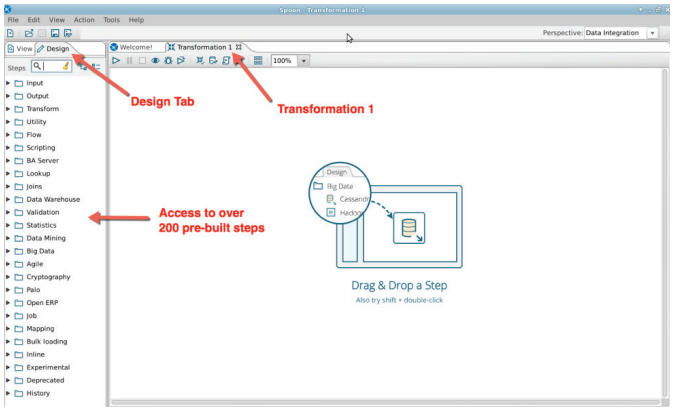
Step by step with Pentaho:
1. On the 'Repository Login' dialog, click Cancel.



Begin creating your first transformation by selecting File | New | Transformation from the menubar. A new tab titled 'Transformation 1' appears for you to begin designing your data flow using the design palate on the left side of the screen.

Pentaho Data Integration provides over 200 pre-built steps including input and output steps for reading and writing data, transformation steps for manipulating data, look up steps for enriching data, and a number of purpose-built steps for working with Big Data platforms such as Hadoop and NoSQL databases.

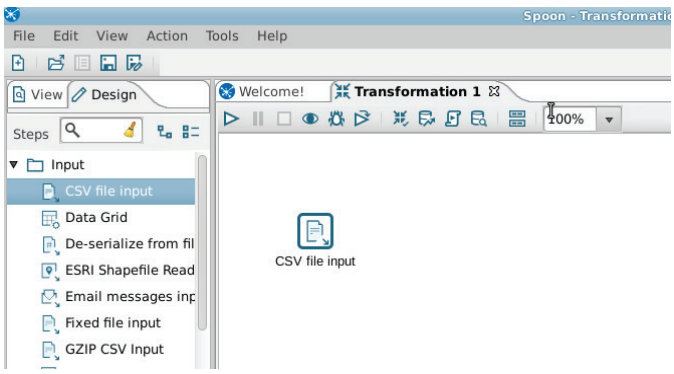
Step by step with Pentaho Corporation:
1. Click to expand the Input folder and then click again to collapse.
2. Click to expand the Transform folder and then click again to collapse.
3. Click to expand the Big Data folder and then click again to collapse.



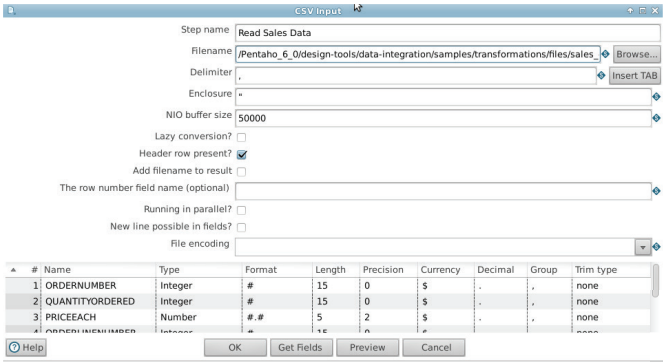
4. Click on the Input folder to open it and leave open for the next section of the demonstration

We'll add our first step in the data flow by dragging the 'CSV file input' step from the Design Palate onto the canvas to the right. Double-click on the step to begin editing the configuration. We'll provide a friendly name, and then click browse to select the CSV file containing the sales transactions we want to read.

Step by step with Pentaho:
1. Drag the CSV file input step onto the canvas
2. Double-click on the step to open the edit dialog

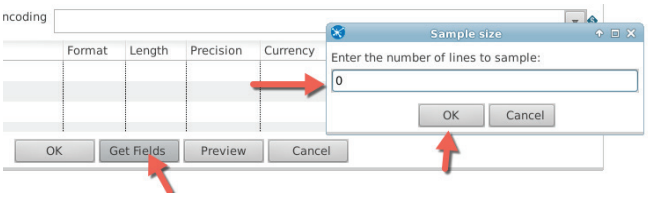


3.In the Step name field, enter 'Read Sales Data'
4. Click on 'Browse…', select the sales_data.csv file (C:\ProgramFiles\pentaho\design-tools\data-integration\samples\transformations\files) and click 'Open'
5.Uncheck the 'Lazy conversion' option

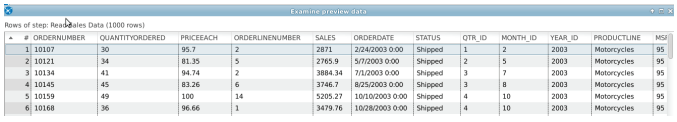


Next we'll use the 'Get Fields' button to bring back a list of fields to be read from the CSV. PDI will analyze a sample of the data to suggest metadata about the fields including field names from the header row if present and the data type and when finished, you are presented with a summary of the results of the scan. Upon closing the scan results, you can see that each of the fields to be read is listed in the Field list. You can now preview the data to ensure our step is properly configured. Everything looks correct, so we'll click 'OK' to continue building our transformation.

Step by step with Pentaho:
1. Click 'Get Fields'
2. In the Sample Size dialog, enter '0' to sample all fields, then click OK



3. In the Scan results dialog, scroll slowly to display some of the results (maybedown to Field nr 5., and then click 'Close'.
4. Click the Preview button, then click OK on the Preview size dialog. Slowly scrollthrough some of the preview data before clicking the 'Close' button.

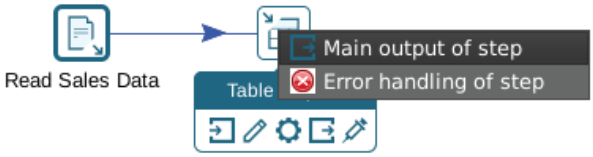


5. Click 'OK' to exit the step configuration.

Now we have our CSV input step for reading the sales data. Let's add a Table output step to write that data into a relational database. You describe the flow of data in your transformations by adding hops between steps on the canvas. Using the hover menu, we'll draw a hop from our 'Read Sales Data' step to the newly added 'Table output' step.

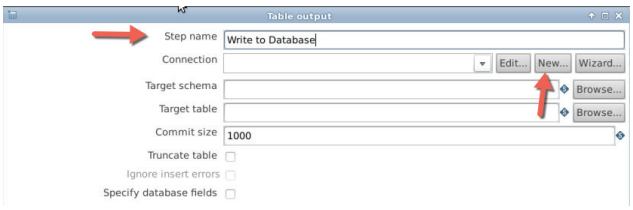
Step by step with Pentaho:
1. Close the Input folder in the Design palate, and expand the Output folder.
2. Drag a Table output step onto the canvas, allow the hover tip on 'How to create anew hop' remain on screen for several seconds.

3. Click to select the 'Read Sales Data' step, and then use the hover menu to draw a hop from 'Read Sales Data' step to the 'Table output' step.
4. When prompted, select this as the 'Main output of the step'.

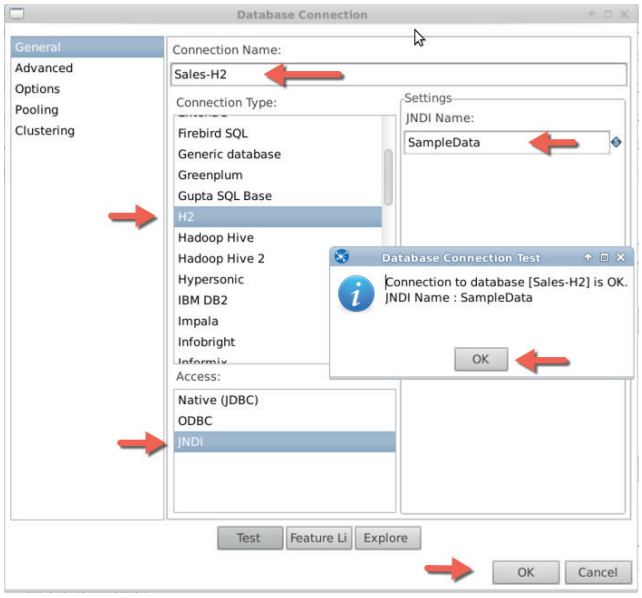


We'll configure the 'Table output' step.. After providing a connection name and enteringthe connection details, we can use the 'Test' button to ensure our connection is properly configured. We want to write our data to a table named Sales, so we'll enter 'Sales' in the Target table field. By clicking on the 'SQL' button, PDI will suggest any SQL necessary to ensure the step works correctly. Since the target table does not exist, you will see a CREATE TABLE DDL statement to execute to prepare for executing our transformation. After executing the DDL, we're now ready to save and run the transformation.

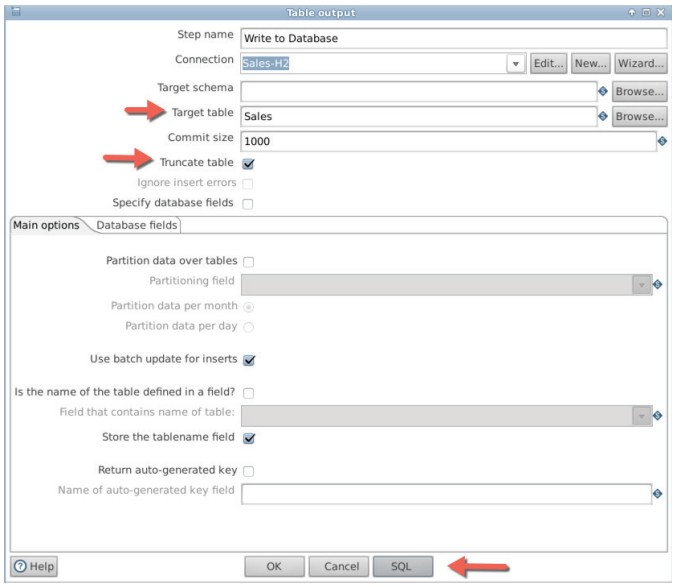
Step by step with Pentaho:
1. Double-click on the Table output step to open the edit dialog.
2. Enter a Step name of 'Write to Database'



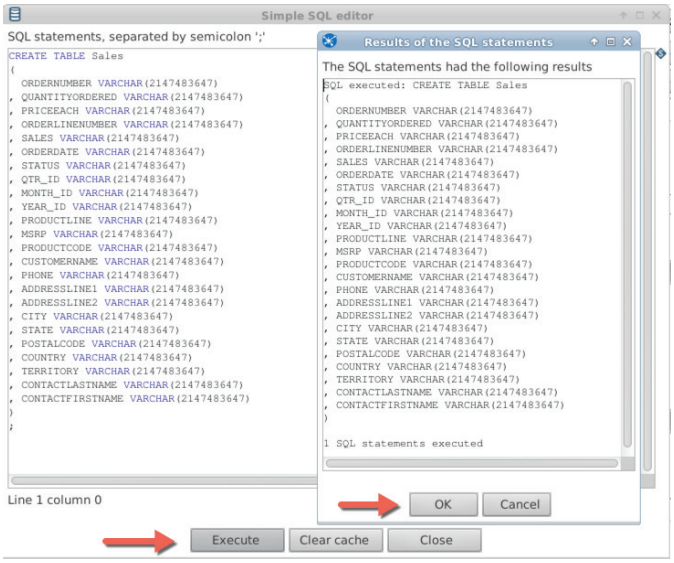
3. *Click the 'New…' button to begin creating a connection*
4. *Enter 'Sales – H2' as the connection name*
5. *Select 'H2' under Connection Type*
6. *Select JNDI under Access:*
7. *Enter 'SampleData' as the JNDI Name*
8. *Click 'Test'*
9. *On the Connection Test dialog, then click OK to close, then OK again to close to exit the Database Connection dialog and complete the connection creation.*



10. *Enter 'Sales' in the Target table field*
11. *Check the 'Truncate table' option (so that multiple runs of the transformation won't keep adding duplicate rows)*
12. *Click the 'SQL' button, pause/scroll slowly to show the generated DDL*



13. *Click Execute, pause for 2 seconds to show the results dialog, then click OK to close it, then click close on the Simple SQL editor dialog.*

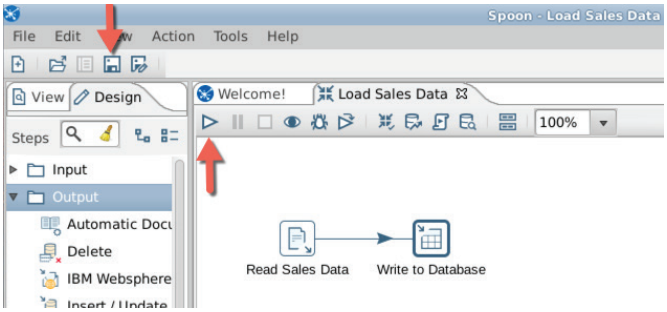


14. *Click OK on the Table output edit dialog (returning to the canvas)*

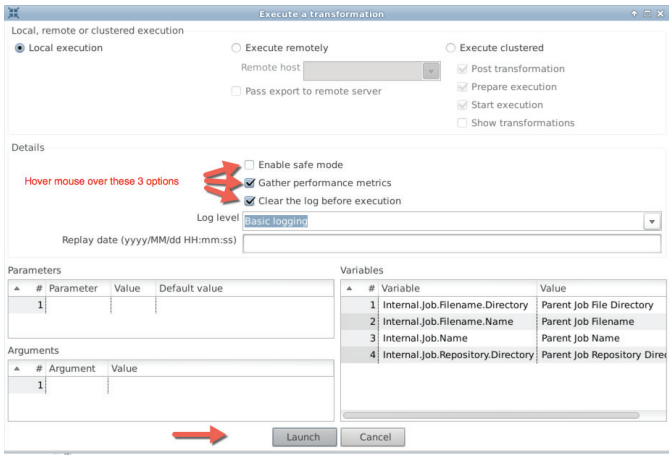
We'll save the transformation as 'Load Sales Data'. We are ready to run the transformation. As you can see, Pentaho Data Integration provides options for running transformations locally, remotely on a dedicated PDI server, or against a cluster of PDI servers for processing large volumes of data or reducing execution times. For this example, we'll simply run the transformation locally.

Step by step with Pentaho:

1. *Click the 'Save' button on the toolbar*
2. *Enter the name 'Load Sales Data' and click the 'Save' button*
3. *Click the Play button on the sub-toolbar to run the transformation*



4. *As you describe the run options, hover the mouse over the three options*
5. *Click Launch to run the transformation*



Congratulations on building and running your first PDI transformation! You can see inthe 'Step Metrics' tab that we have successfully loaded 2823 records into our targetdatabase. To find out more information about the powerful Pentaho platform try another lesson or contact Pentaho and start your free proof of concept with the expertise of a Pentaho sales engineer.

Contact Pentaho at http://www.pentaho.com/contact/

Hitachi Vantara

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 95050-2639 USA
www.HitachiVantara.com | community.HitachiVantara.com

Regional Contact Information
Americas: +1 866 374 5822 or info@hitachivantara.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hitachivantara.com
Asia Pacific: +852 3189 7900 or info.marketing.apac@hitachivantara.com

P-041-A DS February 2018