HITACHI
Inspire the Next

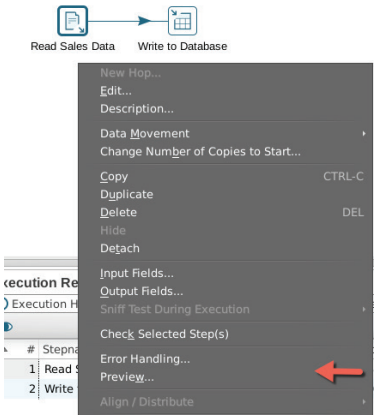
Enriching Data

Pentaho Data Integration is a comprehensive data inegration platform allowing you to access, prepare, analyze and derive value from both traditional and big data sources. This lesson is a continuation of the lesson on building your first transformation. Here we will introduce the preview feature of PDI and use a combination of steps to cleanse our data by looking up missing postal code information.

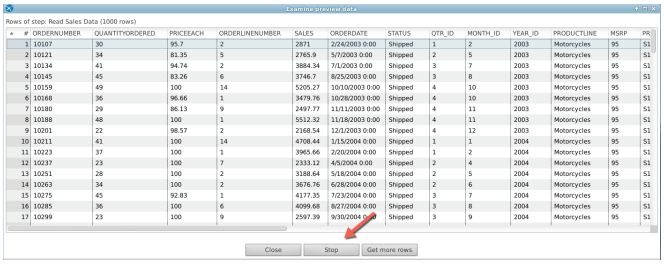
We begin by previewing the rows read by our CSV input step by right-clicking on the step and selecting 'Preview'. Here you specify the number of rows to preview and can optionally configure break-points which pause execution based on a defined condition, such as a field having a specific value or exceeding a threshold. After clicking the 'Quick Launch' button, you can preview data and see that several of our input rows are missing values for the POSTALCODE field.

Step by step with Pentaho:

1. *Right-click on the Read Sales Data step and choose Preview.*



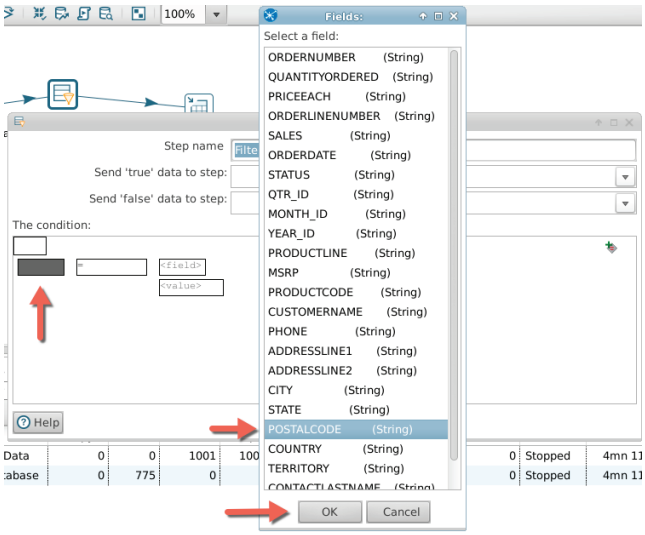
2. *Hit the 'Quick Launch' button*
3. *Click STOP button on the preview window to end the preview.*



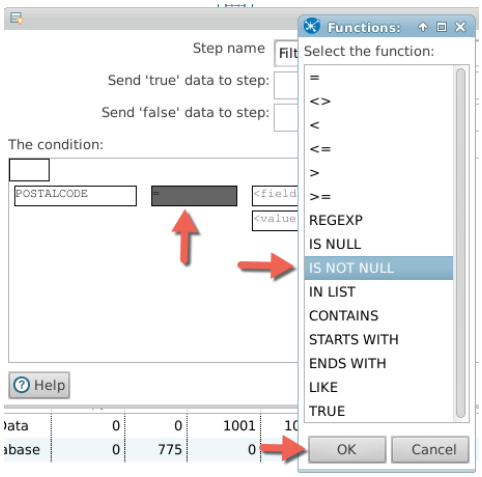
To cleanse our input records, we will resolve the missing postal codes using a lookup file containing City, State and postal code data. First, we use the Filter Rows step to separate rows the rows with missing postal codes. We will define the condition as any time where the POSTALCODE field IS NOT NULL. When this condition is true, we'll send the rows directly to the Write to Database step.

Step by step with Pentaho:

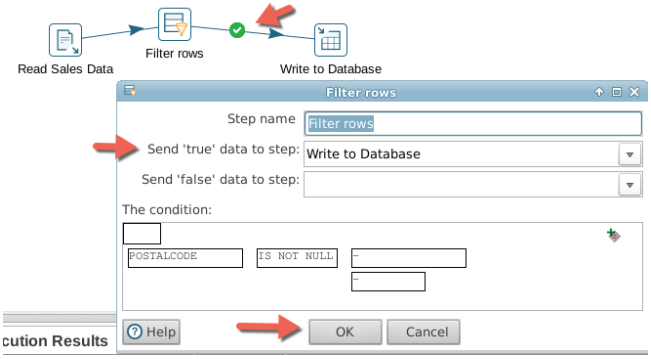
1. *Expand the Flow folder in the Design Palate and Drag a Filter Rows step onto the canvas, then drag it onto the hop between Read Sale Data and Write to Database steps until it makes that hop bold then release it. You should see that it has now become part of the hop.*
2. *Double-click on the Filter Rows to open the edit dialog*
3. *Click on the left <field> operand, select POSTALCODE, then click 'OK'*



4. *Click on the operator (=), select IS NOT NULL, then click 'OK'*



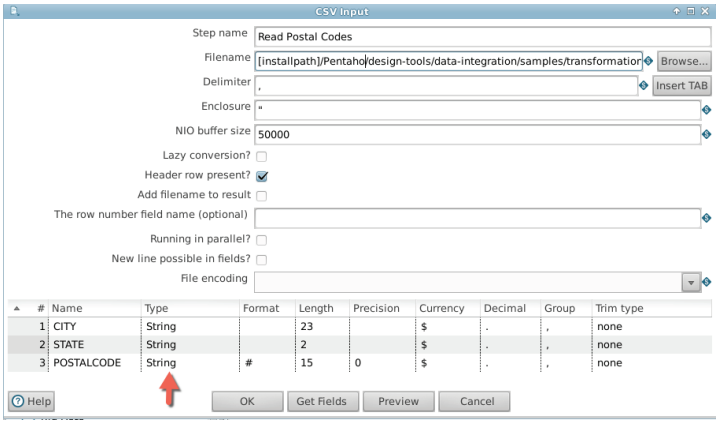
5. *Select the Write to Database step for the "Send 'true' data to step" option*
6. *Click OK to exit the Filter rows edit dialog*



Next, we will create the lookup branch of our transformation. We drag another CSV file input step onto the canvas and edit the configuration to point to our file containing the City, State and Postal Code data.

Step by step with Pentaho:

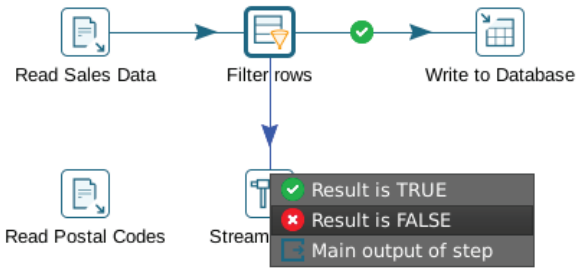
1. *Expand the Input folder in the Design palate and drag a CSV file input onto the canvas, double-click on it to open the step editor.*
2. *Enter 'Read postal codes' as the Step Name*
3. *Click 'Browse" and select the Zipssortedbyci-tystate.csv ([Install path]\pentaho\design-tools-\data-integration\samples\transformations\files), then click 'Open'*
4. *Uncheck the 'Lazy conversion' option*
5. *Click the 'Get Fields' button, click 'OK' on the Sample Size dialog, notice that the POSTALCODE is reporting as Integer, we will correct that next. , click 'Close' on the Scan results dialog*
6. *Next to line 3 POSTALCODE choose the drop down under Type and select String*
7. *then click 'OK' to exit the CSV Input step editor dialog*



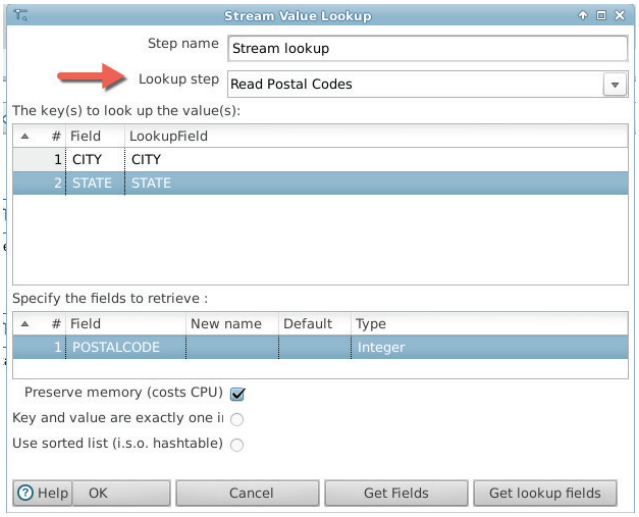
We'll use the Stream Lookup step to add resolved postal code values to the rows where they are missing. After dragging the step onto the canvas, we'll connect it into our data flow by drawing a hop from our Filter rows step and defining is as where to send rows where our condition is FALSE, meaning the postal code is missing. Then we'll create a hop from our Read postal codes step. To configure the Stream lookup, we begin by selecting the Read postal codes as our Lookup step. For our keys, we will use the CITY and STATE fields from our sales data and lookup file. Finally, we will select the POSTALCODE as our field to retrieve by clicking the 'Get lookup fields' button, then removing the CITY and STATE fields that we don't need.

Step by step with Pentaho:

1. *Expand the Lookup folder in the Design palate and drag a Stream lookup step onto the canvas*
2. *Create a hop from the Filter rows step to the Stream lookup step, choosing 'Result is FALSE' as the hop type 3.*



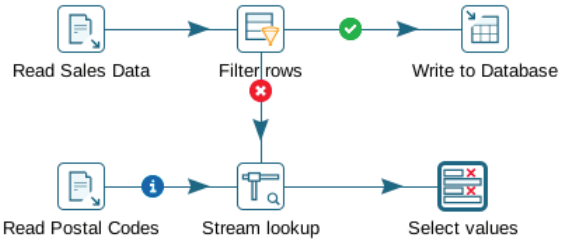
3. *Create a hop from the Read postal codes step to the Stream lookup step, selecting 'Main output of the step' as the hop type*
4. *Double-click on the Steam lookup step to open the step editor*
5. *Click the drop down arrow next to Lookup step and select Read Postal Codes*
6. *In row 1 of the keys table, select CITY as the field and CITY as the LookupField*
7. *In row 2 of the keys table, select STATE as the field and STATE as the LookupField*
8. *Click the 'Get lookup fields' button*
9. *Select row 1 in the Specify the fields to retrieve table (CITY), and hit delete*
10. *Select row 1 in the Specify the fields to retrieve table (STATE), and hit delete*
11. *You should now only have the POSTALCODE field in the table with a type of String*
12. *Click 'OK' to exit the step edit dialog*



Finally, we'll clean up some of the metadata so that the fields flowing through our transformation align and can be loaded correctly into our target database. If we look at the output fields for the Filter Rows step, we see that the POSTALCODE is defined as a String with a character length of 9. Looking at the output fields for our Stream lookup, we see that the new lookup field POSTALCODE_1 is at the bottom of the list and is an Integer type. For this, the Select Values step can be used to perform tasks such as renaming fields, reordering fields or changing a field's data type.

Step by step with Pentaho:

1. *Select (left-click), then right-click on the Filter rows step and choose 'output fields'*
2. *Highlight the POSTALCODE (line 20), then click 'Cancel' to exit*
3. *Select (left-click), then right-click on the Stream lookup step and choose 'output fields'*
4. *Highlight the POSTALCODE_1 (line 26), then click 'Cancel' to exit*
5. *Enter 'select' in the search field at the top of the Design palate to filter the list, then drag a 'Select values' step onto the canvas (under the Transform folder)*
6. *Draw a hop from Stream lookup to Select Values*

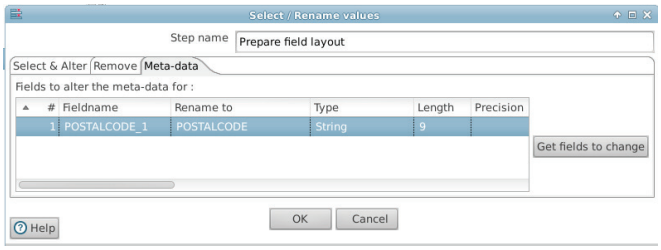


In configuring the Select values step, we'll begin by making sure our postal code lookup field, POSTALCODE_1, is in the proper order by retrieving the list of fields to select, and moving it to the proper location next to the original postal code field. Next, we want to remove the original POSTALCODE field containing the missing values. Finally, we want to rename the POSTALCODE_1 field to POSTALCODE and alter its data type to match the raw type being read from our sales data CSV file. With our field layout prepared, we can now connect the Select values step up to our Write to Database step to send the cleansed rows to the database.

Step by step with Pentaho:

1. *Double-click on the Select values step to open the step editor*
2. *Enter a Step name of 'Prepare field layout'*
3. *Click on the 'Get fields to select' button*
4. *Highlight the POSTALCODE_1 field and move it up to line 21 (CTRL-Up) or (right click - Move UP option)*
5. *Click on the 'Remove' tab*
6. *In line 1 of Fields to remove, click the drop down in the Fieldname column and select POSTALCODE*
7. *Click on the 'Meta-data' tab*
8. *In line 1 of the Fields to alter table:*
    a. *Select POSTALCODE_1 as the Fieldname*
    b. *Enter 'POSTALCODE' in the Rename to column*
    c. *Select 'String' in the Type column*
    d. *Enter '9' in the Length column*

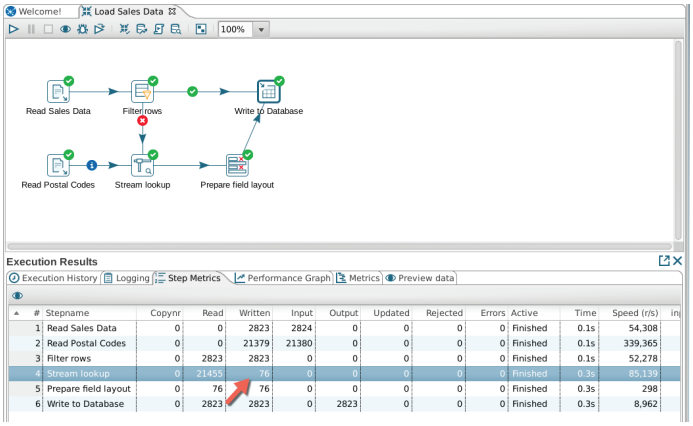
9. Click 'OK' to exit the step editor dialog
10. Draw a hop from the Prepare field layout step to the Write to Database step



Our transformation is now complete and we can preview the cleansed rows that will be written to the database. You can now see that each row contains a valid postal code and we're now ready to re-run the full transformation. After the transformation has completed, we can review the Step Metrics table and see that 76 records were corrected using the Stream Lookup. You have seen how Pentaho Data Integration provides a simple path to enriching your data and creating "analysis ready" data.

Step by step with Pentaho:

1. *Select (left-click) the Write to Database step, then right-click on it and select Preview*
2. *Click the 'Quick Launch' button*
3. *Scroll over to show the viewed the POSTALCODE column with all missing values corrected, pause 1-2 seconds, click 'Stop' to exit the preview window*
4. *Click the PLAY button to open the Execute dialog, then click Launch to execute the transformation*
5. *In the Step Metrics tab, select the Stream Lookup row and mouse over the Written column to highlight the 76 updated rows that contained missing values.*



Congratulations on enriching your data! 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-043-A DS February 2018