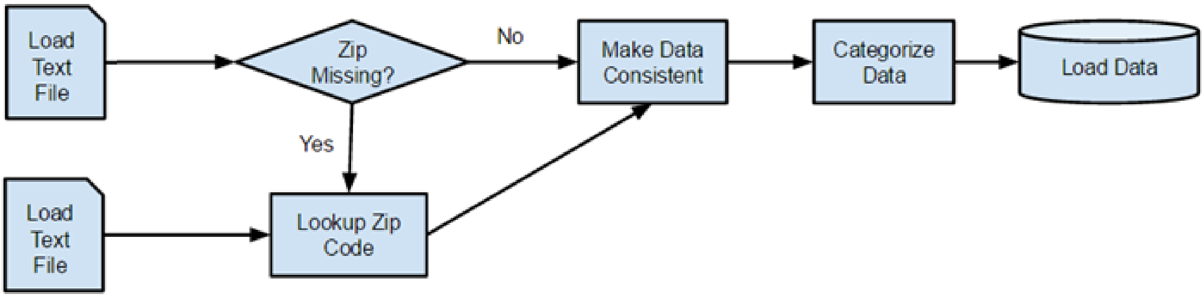
PDI Transformation Basics (Use Case #1, #2, #3 and Future use Case #1)

Demonstrate Exercise #1

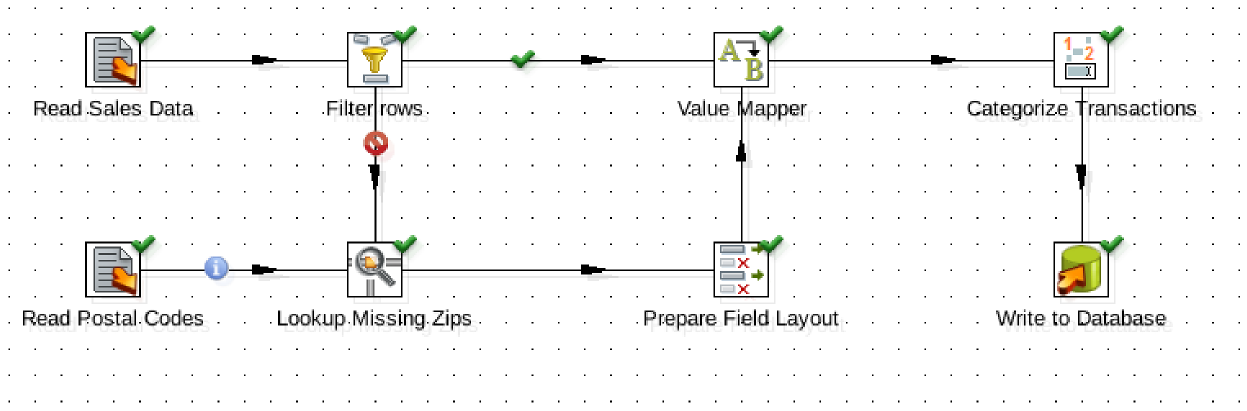
This exercise will step you through building your first transformation with Pentaho Data Integration introducing common concepts along the way. The exercise scenario includes a CSV file of sales data and a flat file (.csv) of postal code data that you will load into a database so that mailing lists can be generated. There are several issues with the data that we want to correct or enrich:

- **Incomplete** – Some of the records have missing zip codes. We need to perform a lookup on these records to get the appropriate zip code.
- **Inconsistent** – Some records have the country as United States and some records has it as USA. We want to make our data consistent.
- **Uncategorized** – We will want to categorize our data based on each transaction size.

The logic looks like this:



The workflow above, once implemented in Pentaho Data Integration, will result in a transformation that looks like the transformation below:



Getting Started with PDI

Ingest Data

CSV Input

1. If not already running, open the Spoon design tool by pressing the button that appears on the bottom of the tool bar that appears on the desktop.

2. Open the “getting_started_with_pdi_final.ktr” from the ~/Desktop/WorkshopTraining/04_getting_started_with_pdi folder for reference purposes.

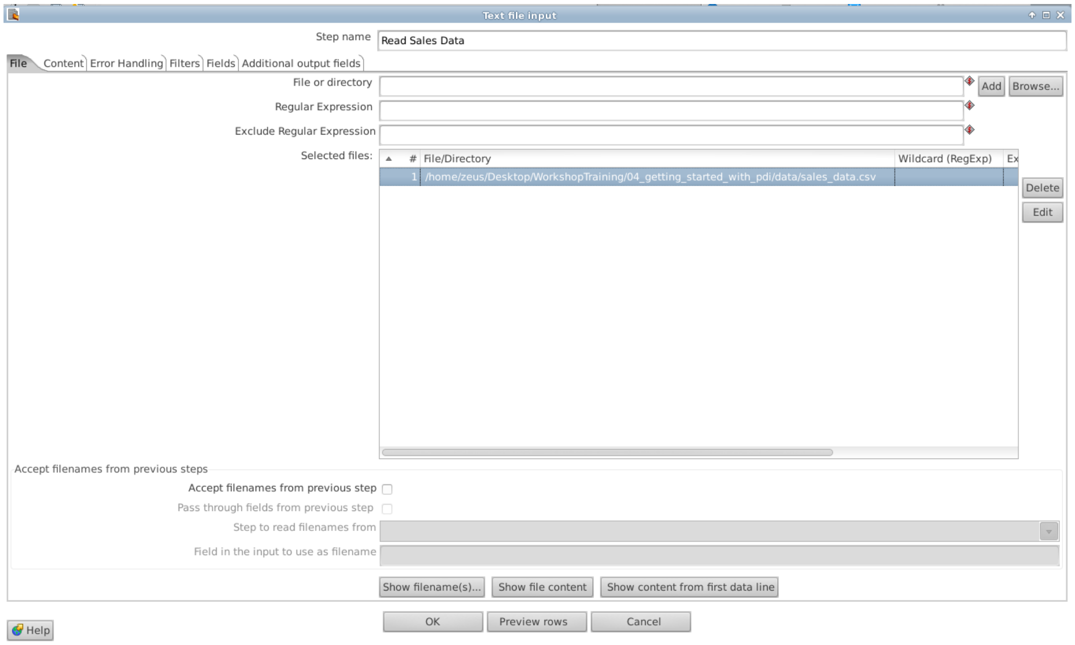
3. Create a New Transformation by going to File -> New -> Transformation

4. Make sure the Design tab is selected, expand the Input Folder, and drag the Text file input step to your canvas.

5. Open the Text file input step and edit the following:

   a. Change the step name to Read Sales Data

   b. Click on Browse and select the “sales_data.csv” file from Desktop/WorkshopTraining/04_getting_started_with_pdi/data/sales_data.csv. Press Add. See dialog below.



Text file input

Step name: Read Sales Data

Selected files:

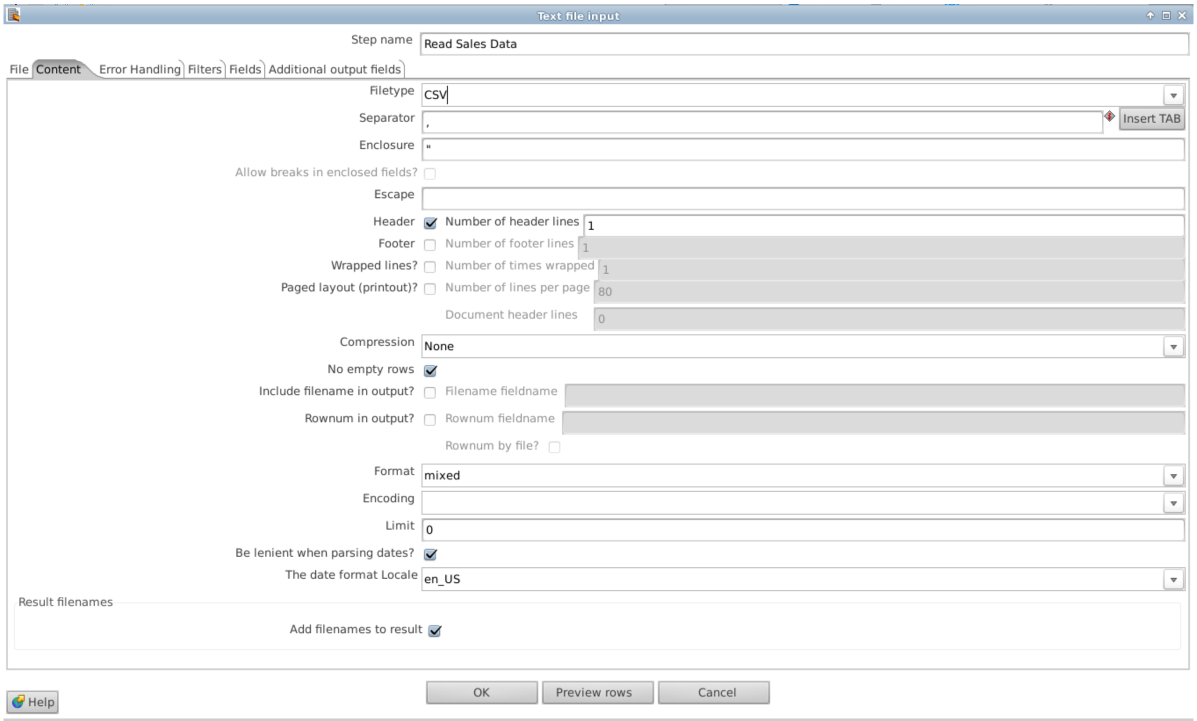
#	File/Directory	Wildcard (RegExp)	Ex
1	/home/zeus/Desktop/WorkshopTraining/04_getting_started_with_pdi/data/sales_data.csv		

Accept filenames from previous steps

Accept filenames from previous step □

Pass through fields from previous step □

c. Switch to the Content tab, and change Separate to a "," (comma). Also, change Format in the dropdown to "mixed".



Step name: Read Sales Data

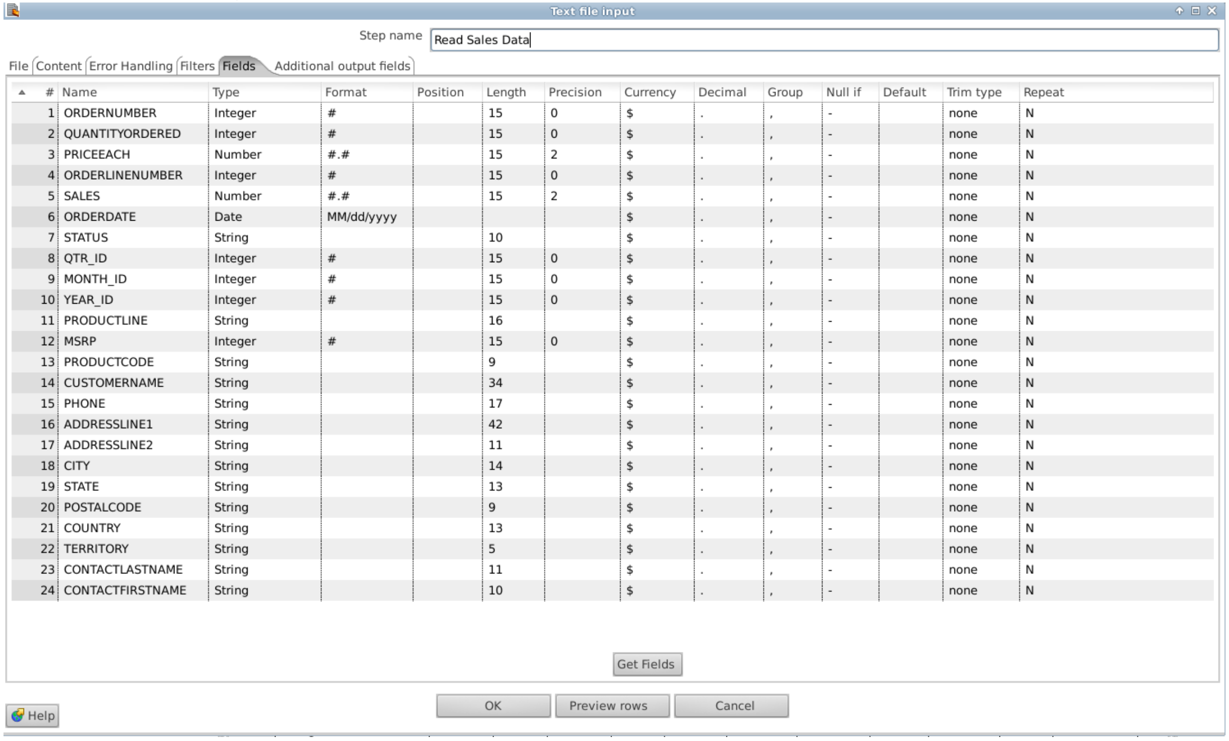
File | Content | Error Handling | Filters | Fields | Additional output fields

Filetype: CSV
Separator: ,

Enclosure: "
Allow breaks in enclosed fields? □
Escape:
Header ☑ Number of header lines: 1
Footer □ Number of footer lines: 1
Wrapped lines? □ Number of times wrapped: 1
Paged layout (printout)? □ Number of lines per page: 80
Document header lines: 0
Compression: None
No empty rows ☑
Include filename in output? □ Filename fieldname:
Rownum in output? □ Rownum fieldname:
Rownum by file? □
Format: mixed
Encoding:
Limit: 0
Be lenient when parsing dates? ☑
The date format Locale: en_US

Result filenames
Add filenames to result ☑

d. Switch to the Fields tab and press the Get Fields button. In the dialog that comes up, enter a 0. This will profile all available rows.

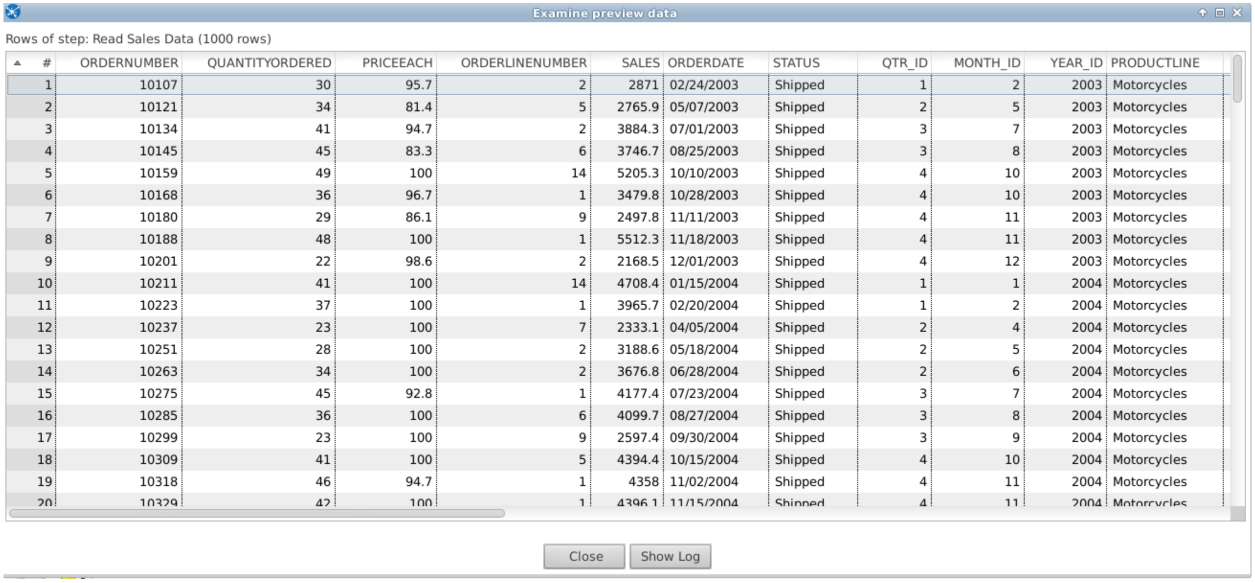


Step name: Read Sales Data

File | Content | Error Handling | Filters | Fields | Additional output fields

#	Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default	Trim type	Repeat
1	ORDERNUMBER	Integer	#		15	0	$	.	,	-		none	N
2	QUANTITYORDERED	Integer	#		15	0	$	.	,	-		none	N
3	PRICEEACH	Number	#.#		15	2	$	.	,	-		none	N
4	ORDERLINENUMBER	Integer	#		15	0	$	.	,	-		none	N
5	SALES	Number	#.#		15	2	$	.	,	-		none	N
6	ORDERDATE	Date	MM/dd/yyyy				$	.	,	-		none	N
7	STATUS	String			10		$	.	,	-		none	N
8	QTR_ID	Integer	#		15	0	$	.	,	-		none	N
9	MONTH_ID	Integer	#		15	0	$	.	,	-		none	N
10	YEAR_ID	Integer	#		15	0	$	.	,	-		none	N
11	PRODUCTLINE	String			16		$	.	,	-		none	N
12	MSRP	Integer	#		15	0	$	.	,	-		none	N
13	PRODUCTCODE	String			9		$	.	,	-		none	N
14	CUSTOMERNAME	String			34		$	.	,	-		none	N
15	PHONE	String			17		$	.	,	-		none	N
16	ADDRESSLINE1	String			42		$	.	,	-		none	N
17	ADDRESSLINE2	String			11		$	.	,	-		none	N
18	CITY	String			14		$	.	,	-		none	N
19	STATE	String			13		$	.	,	-		none	N
20	POSTALCODE	String			9		$	.	,	-		none	N
21	COUNTRY	String			13		$	.	,	-		none	N
22	TERRITORY	String			5		$	.	,	-		none	N
23	CONTACTLASTNAME	String			11		$	.	,	-		none	N
24	CONTACTFIRSTNAME	String			10		$	.	,	-		none	N

6. Press the Preview rows button and make sure data is displaying from the CSV file as below.



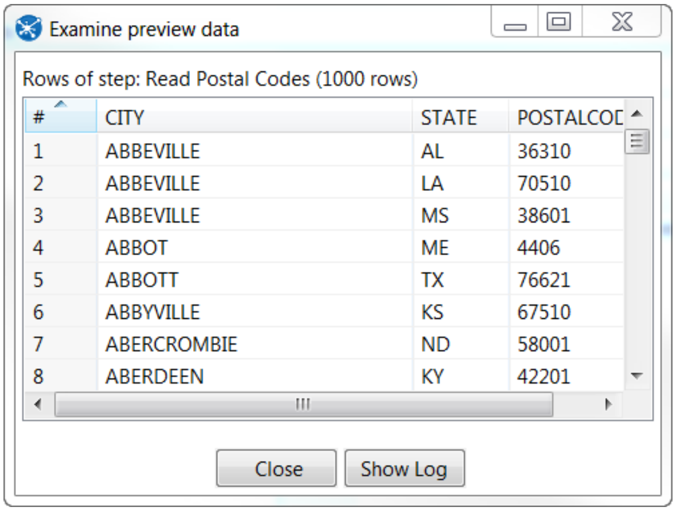
Rows of step: Read Sales Data (1000 rows)

#	ORDERNUMBER	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	ORDERDATE	STATUS	QTR_ID	MONTH_ID	YEAR_ID	PRODUCTLINE
1	10107	30	95.7	2	2871	02/24/2003	Shipped	1	2	2003	Motorcycles
2	10121	34	81.4	5	2765.9	05/07/2003	Shipped	2	5	2003	Motorcycles
3	10134	41	94.7	2	3884.3	07/01/2003	Shipped	3	7	2003	Motorcycles
4	10145	45	83.3	6	3746.7	08/25/2003	Shipped	3	8	2003	Motorcycles
5	10159	49	100	14	5205.3	10/10/2003	Shipped	4	10	2003	Motorcycles
6	10168	36	96.7	1	3479.8	10/28/2003	Shipped	4	10	2003	Motorcycles
7	10180	29	86.1	9	2497.8	11/11/2003	Shipped	4	11	2003	Motorcycles
8	10188	48	100	1	5512.3	11/18/2003	Shipped	4	11	2003	Motorcycles
9	10201	22	98.6	2	2168.5	12/01/2003	Shipped	4	12	2003	Motorcycles
10	10211	41	100	14	4708.4	01/15/2004	Shipped	1	1	2004	Motorcycles
11	10223	37	100	1	3965.7	02/20/2004	Shipped	1	2	2004	Motorcycles
12	10237	23	100	7	2333.1	04/05/2004	Shipped	2	4	2004	Motorcycles
13	10251	28	100	2	3188.6	05/18/2004	Shipped	2	5	2004	Motorcycles
14	10263	34	100	2	3676.8	06/28/2004	Shipped	2	6	2004	Motorcycles
15	10275	45	92.8	1	4177.4	07/23/2004	Shipped	3	7	2004	Motorcycles
16	10285	36	100	6	4099.7	08/27/2004	Shipped	3	8	2004	Motorcycles
17	10299	23	100	9	2597.4	09/30/2004	Shipped	3	9	2004	Motorcycles
18	10309	41	100	5	4394.4	10/15/2004	Shipped	4	10	2004	Motorcycles
19	10318	46	94.7	1	4358	11/02/2004	Shipped	4	11	2004	Motorcycles
20	10329	42	100	1	4396.1	11/15/2004	Shipped	4	11	2004	Motorcycles

7. Click on the Save button and save this transformation as getting_started_with_pdi_student in the ~/Desktop/WorkshopTraining/student_files/04_getting_started_with_pdi folder.

Text File Input

1. Expand the Input folder on the Design Tab and drag Text file Input underneath the Read Sales Data Step
2. Rename the step to Read Postal Codes
3. Click on browse Desktop/WorkshopTraining/04_getting_started_with_pdi/Zipssortedbycitystate.csv then click on the add button
4. Click on the Content tab and change the Separator to ,
5. Click on the fields tab and select Get Fields, enter 0 then Ok then click Ok
6. Click on Preview to validate the data



Rows of step: Read Postal Codes (1000 rows)

#	CITY	STATE	POSTALCOD
1	ABBEVILLE	AL	36310
2	ABBEVILLE	LA	70510
3	ABBEVILLE	MS	38601
4	ABBOT	ME	4406
5	ABBOTT	TX	76621
6	ABBYVILLE	KS	67510
7	ABERCROMBIE	ND	58001
8	ABERDEEN	KY	42201

7. Click Close and Ok

Transform Data

Filter Missing Zip Code

1. Expand the Flow folder in the Design tab

2. Drag the Filter Rows step to the right of the Read Sales Data Step

3. Connect the Read Sales Data step to the Filter rows step by holding down the Shift key, press and hold the left mouse button and drag the pointer to the Filter rows step and let the button go. Select Main Output of Step
4. Open the Filter row step and rename it to Filter Missing Zip and click ok

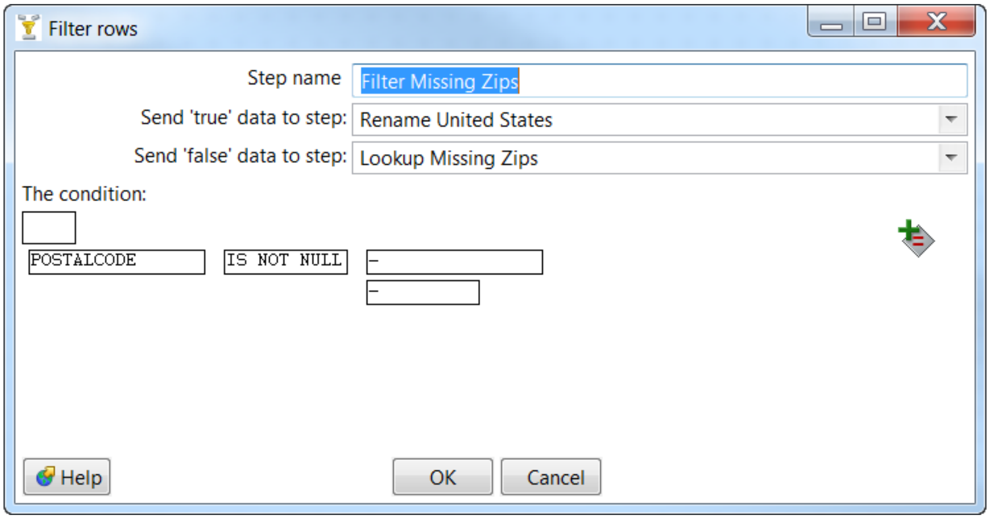
5. Expand the Transform folder under the Design tab and drag over the Value Mapper step to the right of the Filter Missing Zip step, Open the step and rename it to Rename United States

6. Expand the Lookup folder under the Design tab and drag over the Stream lookup step underneath the Filter Missing Zip Step and rename it to Lookup Missing Zips

7. Connect the Filter Missing Zip step to the Value Mapper step and select Result is TRUE

8. Connect the Filter Missing Zip step to the Stream Lookup step and select Result is FALSE

9. Open the Filter Missing Zips step and change the values to reflect the following



Step name: Filter Missing Zips

Send 'true' data to step: Rename United States

Send 'false' data to step: Lookup Missing Zips

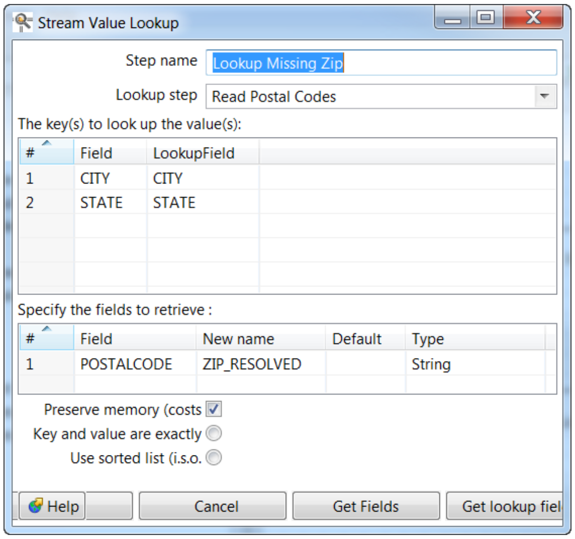
The condition:

POSTALCODE IS NOT NULL

Click Ok

Lookup Missing Data

1. Connect the Read Postal Codes step to the Lookup Missing Zips step
2. Open the Lookup Missing Zips step
3. Select the Read Postal Codes step as your Lookup Step
4. Enter the following values



Step name: Lookup Missing Zip

Lookup step: Read Postal Codes

The key(s) to look up the value(s):

#	Field	LookupField
1	CITY	CITY
2	STATE	STATE

Specify the fields to retrieve :

#	Field	New name	Default	Type
1	POSTALCODE	ZIP_RESOLVED		String

Preserve memory (costs ☑

Key and value are exactly ○

Use sorted list (i.s.o. ○

5. Click Ok

Prepare Field Layout

After you resolve the missing zip code information, the last task is to clean up the field layout on your lookup stream. Cleaning up makes it so that it matches the format and layout of your other streams going to the final write to database step that we will do at the end of this exercise.

1. Expand the Transform folder in the Design Tab and drag over the Select Values step to the right of the Lookup Missing Zips step

2. Connect the Lookup Missing Zip step to the Select values rows step

3. Open the Select values step and rename to Prepare Field Layout

4. Click on the Get fields to select button

5. Highlight the ZIP_RESOLVED row (row #25) and press the CTRL and Up Arrow until the ZIP_RESOLVED row is underneath the POSTALCODE ROW. It should now be row #19.

6. Delete the POSTALCODE row

7. Click on the Meta-data, select ZIP_RESOLVED, enter POSTALCODE in the Rename to field, give it a type of String and a length of 8

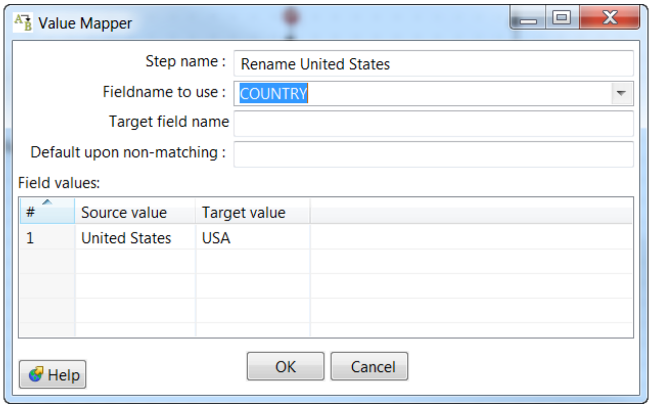
8. Click Ok

9. Connect the Prepare Field Layout step to the Rename United States step

Value Mapping

1. Open the Rename United States step

2. Select Country as the Fieldname to use, Source Value = United States and Target value = USA



Step name : Rename United States

Fieldname to use : COUNTRY

Target field name

Default upon non-matching :

Field values:

#	Source value	Target value
1	United States	USA

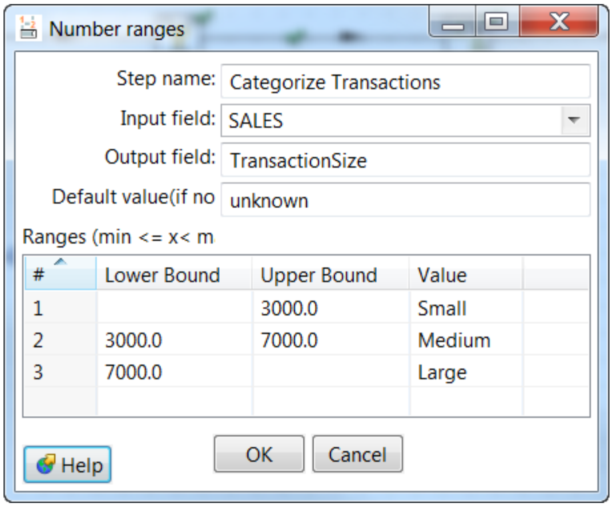
3. Click Ok

Categorizing Measures into Dimensions

1. Expand the Transform folder in the Design tab and drag the Number range step to the right of Rename United States.

2. Connect the Rename United States to the Number range step.

3. Open the step and rename it to Categorize Transactions, then enter the following values:



Step name: Categorize Transactions

Input field: SALES

Output field: TransactionSize

Default value(if no unknown

Ranges (min <= x< m

#	Lower Bound	Upper Bound	Value
1		3000.0	Small
2	3000.0	7000.0	Medium
3	7000.0		Large

Load Data

1. Expand the Output folder in the Design tab and drag the Table output step below Categorize Transactions
2. Connect the Categorize Transactions step to the Table output step
3. Edit the Table Output step and rename it to Load Data
4. Select the workshop_postgres connection
5. Enter workshop1_yourinitials
6. Click the SQL button the click on execute, then Ok, then Close
7. Select the Truncate table checkbox
8. Click on the Ok button
9. Save your transformation and then run your transformation using the default settings.

Visualizing the Data

1. Right click on the Load Data Step and select Visualize and then Analyzer. This will open up an analysis perspective that gives you the ability to slice and dice your data
2. From the Level grouping on the left, drag TERRITORY to the Rows Layout Section, drag YEAR ID to the COLUMNS layout section, then from the Measure section on the left, drag SALES to the measures section on the Layout Panel
3. Click on the graph icon on the upper right corner 📊

Review Exercise #1

What We Covered…

- Basic understanding of creating a transformation
- Ingesting data from a JSON format file
- Ingesting data from a text format profiling the data within the file
- Filtering data
- Categorizing data
- Loading data into a database table
- Visualizing data