Prac 1 : Introduction to Pentaho ETL tool

The Data Integration perspective of Spoon allows you to create two basic file types: transformations and jobs.

1. Transformations are used to describe the data flows for ETL such as reading from a source, transforming data and loading it into a target location.
2. Jobs are used to coordinate ETL activities such as defining the flow and dependencies for what order transformations should be run, or prepare for execution by checking conditions such as, "Is my source file available?" or "Does a table exist in my database?"

Launching the PDI graphical designer: Spoon:

1. Start spoon: If your system is Windows, type the following command:

   Spoon.bat

2. As soon as Spoon starts, a dialog window appears asking for the repository connection data. Click the **No Repository** button. The main window appears. You will see a small window with the tip of the day. After reading it, close that window.
3. A **welcome!** window appears with some useful links for you to see.

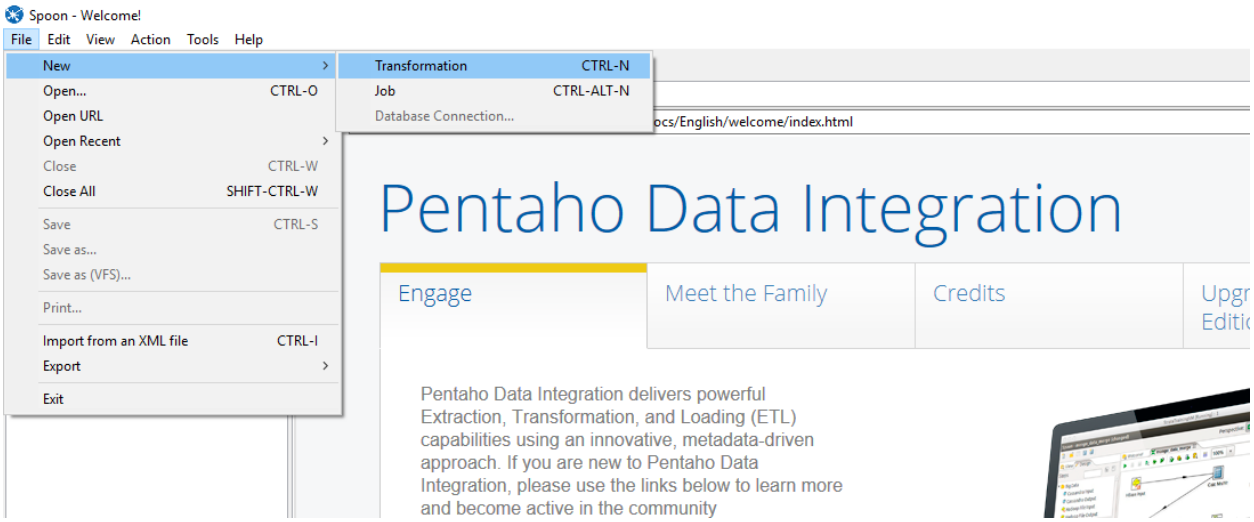
Creating new Transformation

Transformation is a network of logical tasks called *steps*. Transformations are essentially *data flows*. In the example below, the database developer has created a transformation that reads a flat file, filters it, sorts it, and loads it to a relational database table.

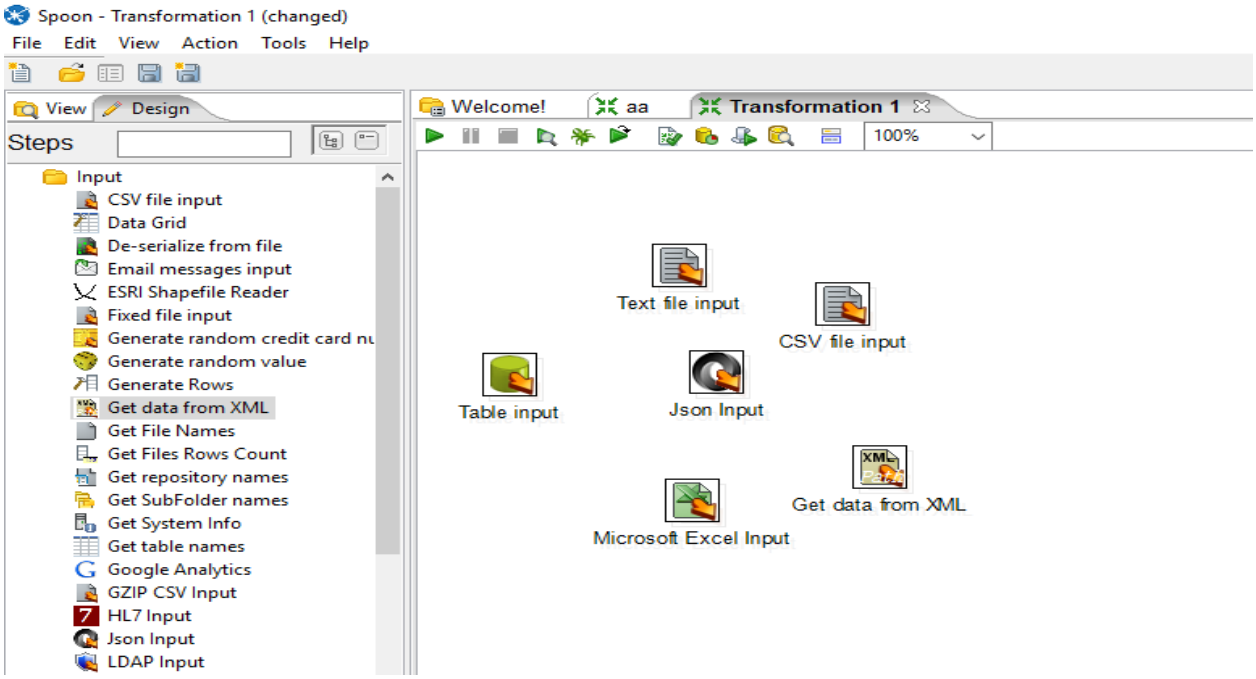
A. INPUT

1. Run the Pentaho Data Integration tools (Spoon).

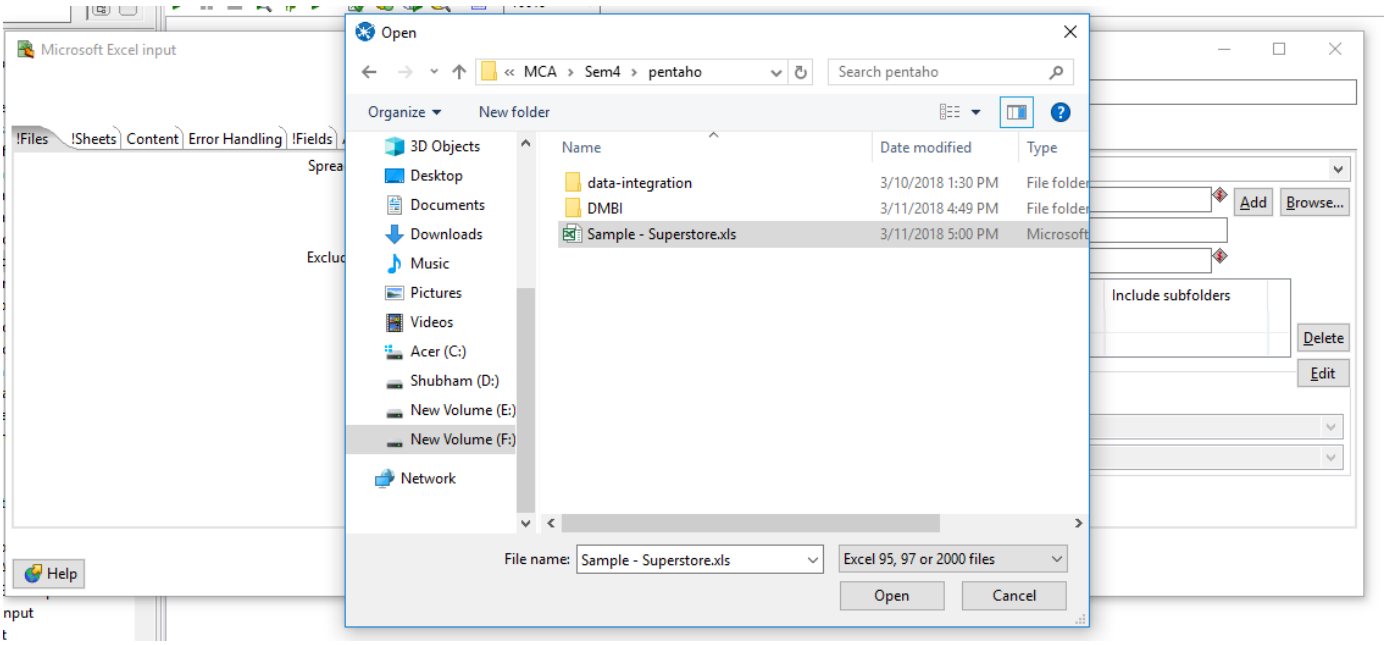
2. Pull down the **File** menu and select the **New** menu item followed by **Transformation**.



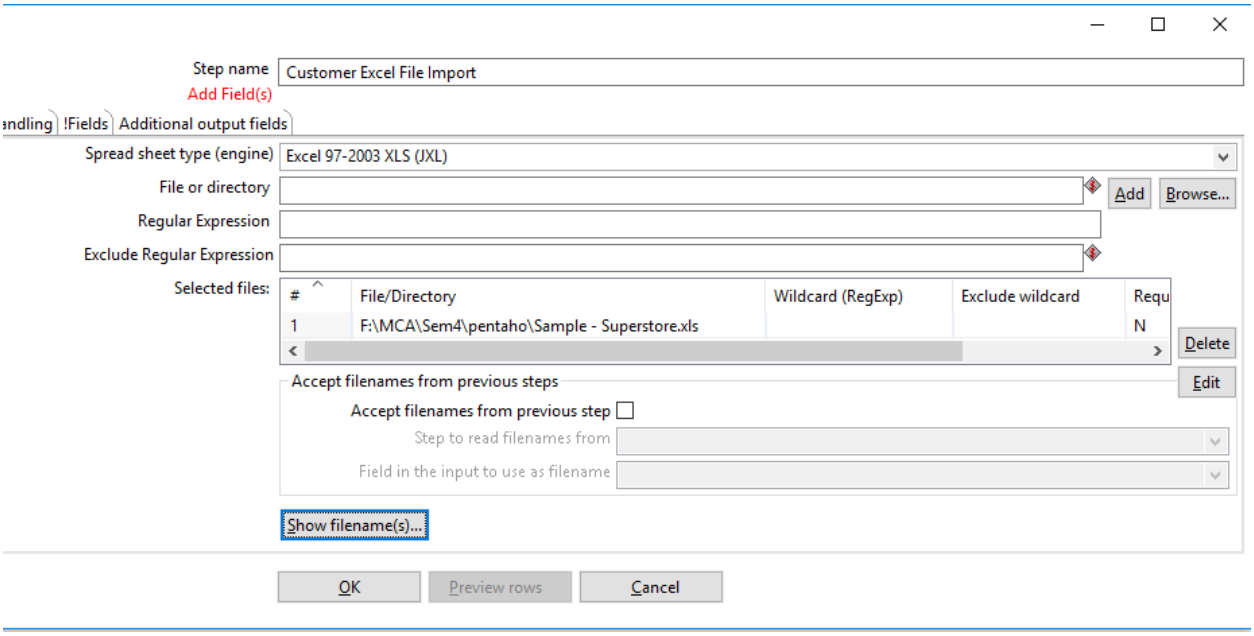
3. Open up the Input folder and drag and drop the **Table Input /JSON Input / Get data From XML / Microsoft excel input /Text File input** step on to the transformation window.



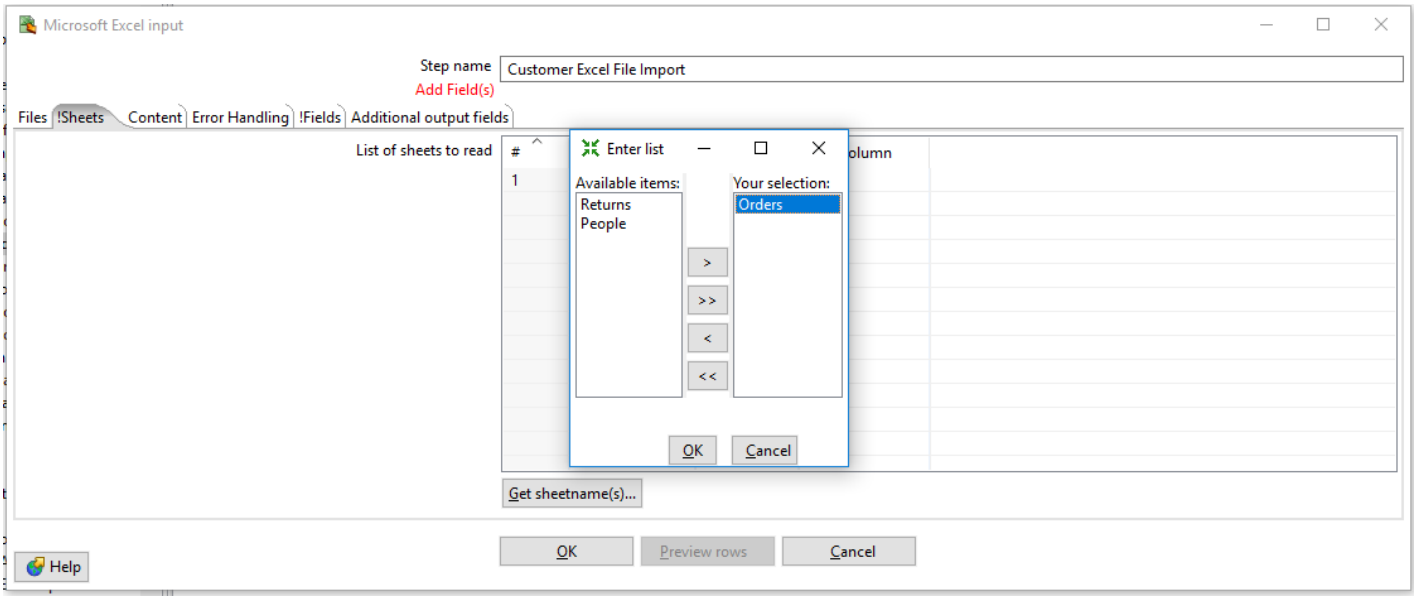
4. Double-click on the Microsoft Excel Input step to view its properties Click on the Browse button next to **Filename** field and navigate to the folder with the Excel files. Select the Excel file as shown below and then click the Open button.:



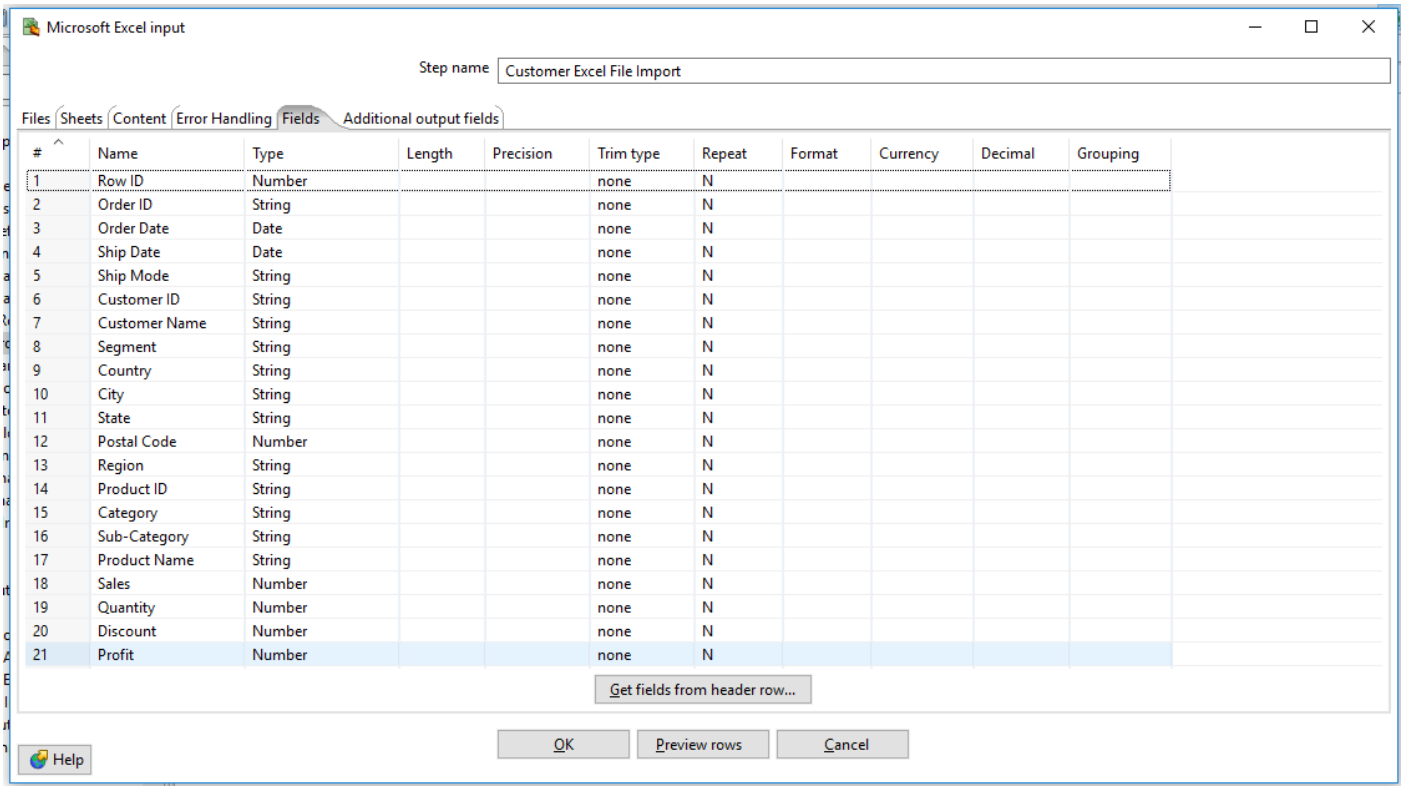
5. Click on ADD button to add excel file .



6. Navigate to Sheet tab to select the sheet which you want to select. And select Sheetname(s) Button. Following form will appear then select sheet and select (>) button to add sheet. Or (<) this button to remove sheet.

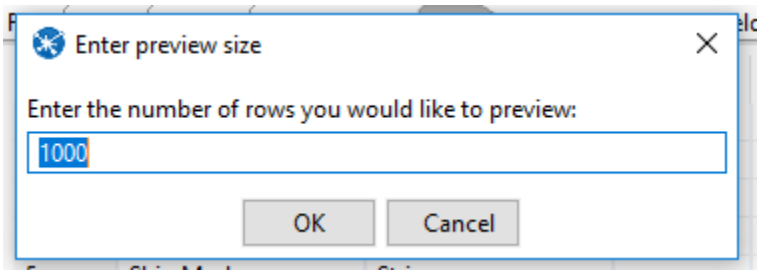


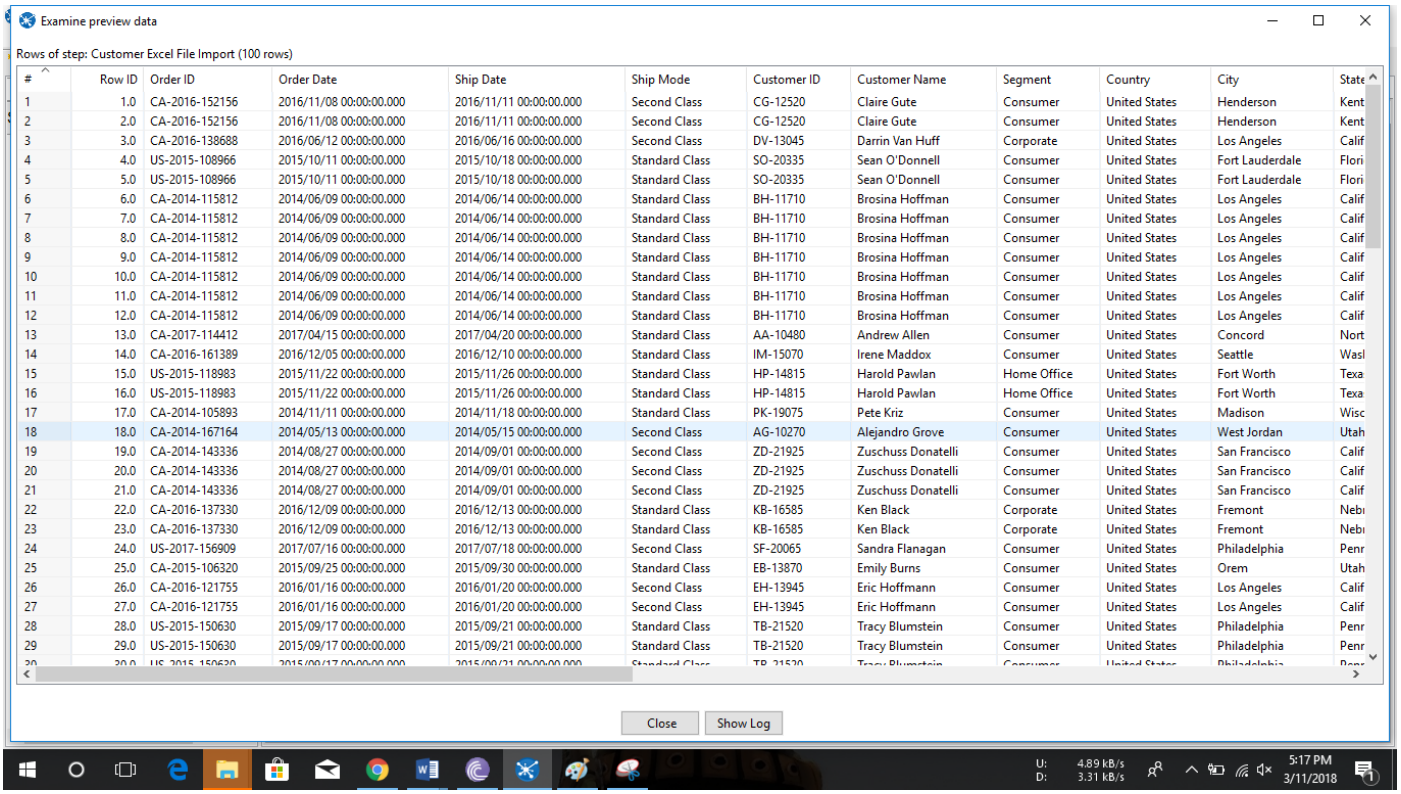
7. Navigate to Fields tab to assert the header name. Click on “Get Fields from header row” button to get column list .



#	Name	Type	Length	Precision	Trim type	Repeat	Format	Currency	Decimal	Grouping
1	Row ID	Number			none	N				
2	Order ID	String			none	N				
3	Order Date	Date			none	N				
4	Ship Date	Date			none	N				
5	Ship Mode	String			none	N				
6	Customer ID	String			none	N				
7	Customer Name	String			none	N				
8	Segment	String			none	N				
9	Country	String			none	N				
10	City	String			none	N				
11	State	String			none	N				
12	Postal Code	Number			none	N				
13	Region	String			none	N				
14	Product ID	String			none	N				
15	Category	String			none	N				
16	Sub-Category	String			none	N				
17	Product Name	String			none	N				
18	Sales	Number			none	N				
19	Quantity	Number			none	N				
20	Discount	Number			none	N				
21	Profit	Number			none	N				

8. Click on “Preview Row” button to select top 1000 row.



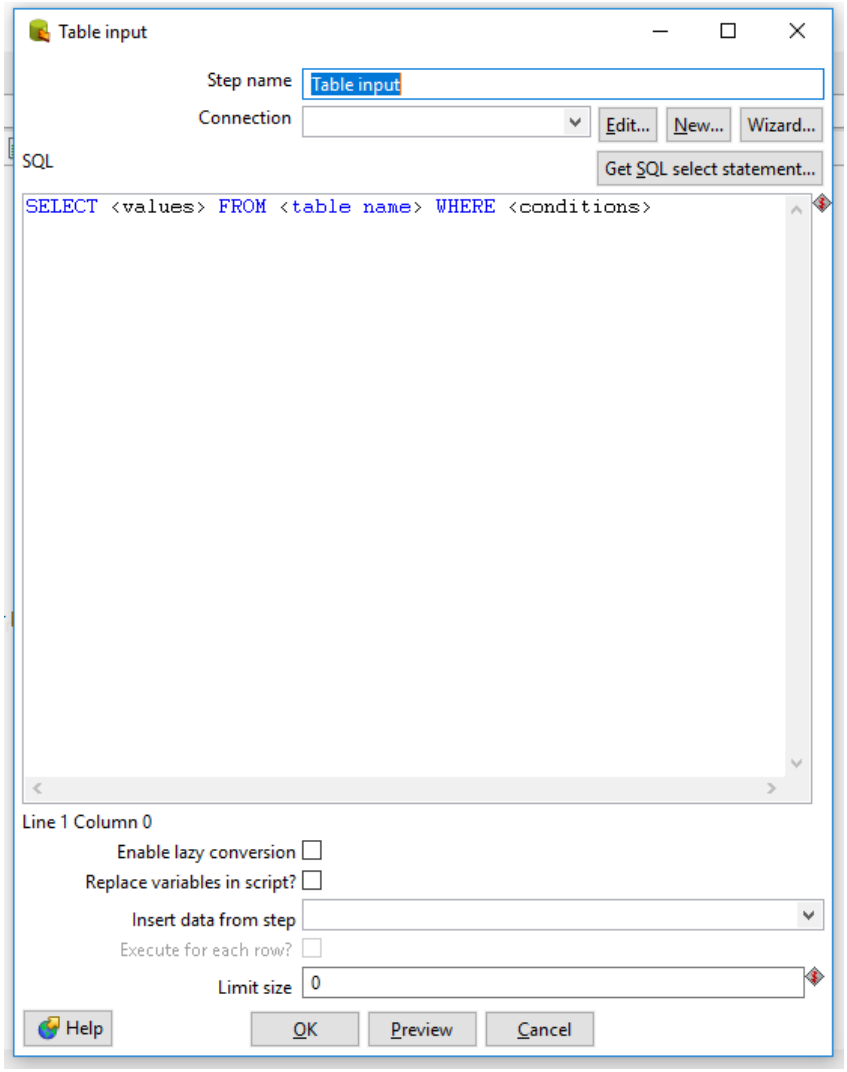


#	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State
1	1.0	CA-2016-152156	2016/11/08 00:00:00.000	2016/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kent
2	2.0	CA-2016-152156	2016/11/08 00:00:00.000	2016/11/11 00:00:00.000	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kent
3	3.0	CA-2016-138688	2016/06/12 00:00:00.000	2016/06/16 00:00:00.000	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	Calif
4	4.0	US-2015-108966	2015/10/11 00:00:00.000	2015/10/18 00:00:00.000	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Flori
5	5.0	US-2015-108966	2015/10/11 00:00:00.000	2015/10/18 00:00:00.000	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Flori
6	6.0	CA-2014-115812	2014/06/09 00:00:00.000	2014/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	Calif
7	7.0	CA-2014-115812	2014/06/09 00:00:00.000	2014/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	Calif
8	8.0	CA-2014-115812	2014/06/09 00:00:00.000	2014/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	Calif
9	9.0	CA-2014-115812	2014/06/09 00:00:00.000	2014/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	Calif
10	10.0	CA-2014-115812	2014/06/09 00:00:00.000	2014/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	Calif
11	11.0	CA-2014-115812	2014/06/09 00:00:00.000	2014/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	Calif
12	12.0	CA-2014-115812	2014/06/09 00:00:00.000	2014/06/14 00:00:00.000	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	Calif
13	13.0	CA-2017-114412	2017/04/15 00:00:00.000	2017/04/20 00:00:00.000	Standard Class	AA-10480	Andrew Allen	Consumer	United States	Concord	Nort
14	14.0	CA-2016-161389	2016/12/05 00:00:00.000	2016/12/10 00:00:00.000	Standard Class	IM-15070	Irene Maddox	Consumer	United States	Seattle	Wasl
15	15.0	US-2015-118983	2015/11/22 00:00:00.000	2015/11/26 00:00:00.000	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth	Texa
16	16.0	US-2015-118983	2015/11/22 00:00:00.000	2015/11/26 00:00:00.000	Standard Class	HP-14815	Harold Pawlan	Home Office	United States	Fort Worth	Texa
17	17.0	CA-2014-105893	2014/11/11 00:00:00.000	2014/11/18 00:00:00.000	Standard Class	PK-19075	Pete Kriz	Consumer	United States	Madison	Wisc
18	18.0	CA-2014-167164	2014/05/13 00:00:00.000	2014/05/15 00:00:00.000	Second Class	AG-10270	Alejandro Grove	Consumer	United States	West Jordan	Utah
19	19.0	CA-2014-143336	2014/08/27 00:00:00.000	2014/09/01 00:00:00.000	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	Calif
20	20.0	CA-2014-143336	2014/08/27 00:00:00.000	2014/09/01 00:00:00.000	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	Calif
21	21.0	CA-2014-143336	2014/08/27 00:00:00.000	2014/09/01 00:00:00.000	Second Class	ZD-21925	Zuschuss Donatelli	Consumer	United States	San Francisco	Calif
22	22.0	CA-2016-137330	2016/12/09 00:00:00.000	2016/12/13 00:00:00.000	Standard Class	KB-16585	Ken Black	Corporate	United States	Fremont	Nebı
23	23.0	CA-2016-137330	2016/12/09 00:00:00.000	2016/12/13 00:00:00.000	Standard Class	KB-16585	Ken Black	Corporate	United States	Fremont	Nebı
24	24.0	US-2017-156909	2017/07/16 00:00:00.000	2017/07/18 00:00:00.000	Second Class	SF-20065	Sandra Flanagan	Consumer	United States	Philadelphia	Penr
25	25.0	CA-2015-106320	2015/09/25 00:00:00.000	2015/09/30 00:00:00.000	Standard Class	EB-13870	Emily Burns	Consumer	United States	Orem	Utah
26	26.0	CA-2016-121755	2016/01/16 00:00:00.000	2016/01/20 00:00:00.000	Second Class	EH-13945	Eric Hoffmann	Consumer	United States	Los Angeles	Calif
27	27.0	CA-2016-121755	2016/01/16 00:00:00.000	2016/01/20 00:00:00.000	Second Class	EH-13945	Eric Hoffmann	Consumer	United States	Los Angeles	Calif
28	28.0	US-2015-150630	2015/09/17 00:00:00.000	2015/09/21 00:00:00.000	Standard Class	TB-21520	Tracy Blumstein	Consumer	United States	Philadelphia	Penr
29	29.0	US-2015-150630	2015/09/17 00:00:00.000	2015/09/21 00:00:00.000	Standard Class	TB-21520	Tracy Blumstein	Consumer	United States	Philadelphia	Penr

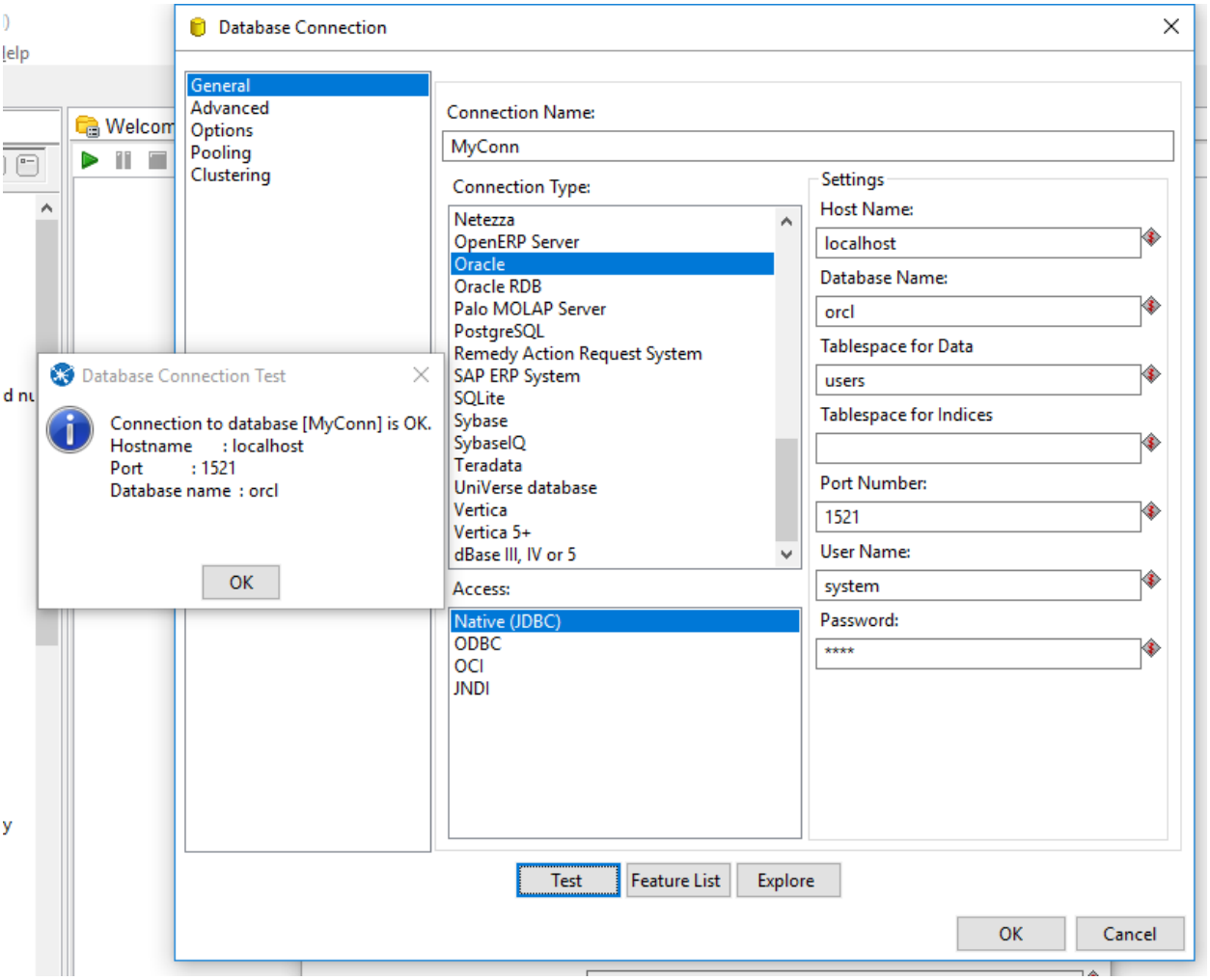
9. Close the preview and make any additional changes required to data types. Once done, click the **OK** button to close up this Excel File input step.

Configuring Table Input :

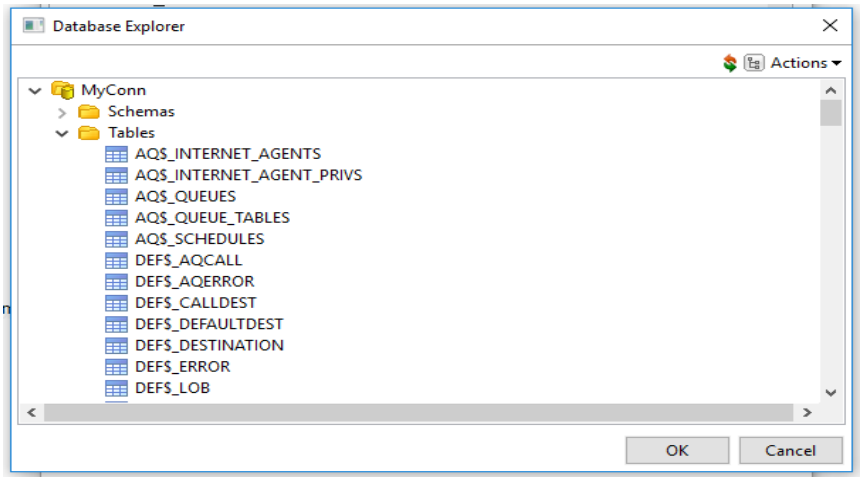
1. Double-click on the Table Input step to view its properties. Select “New” button to configure database connection. Or “Edit” button to edit previously configured connection.



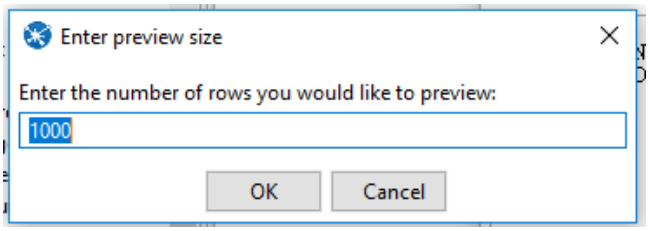
2. Fill the following connection information. And click on “Test” button to verify information.

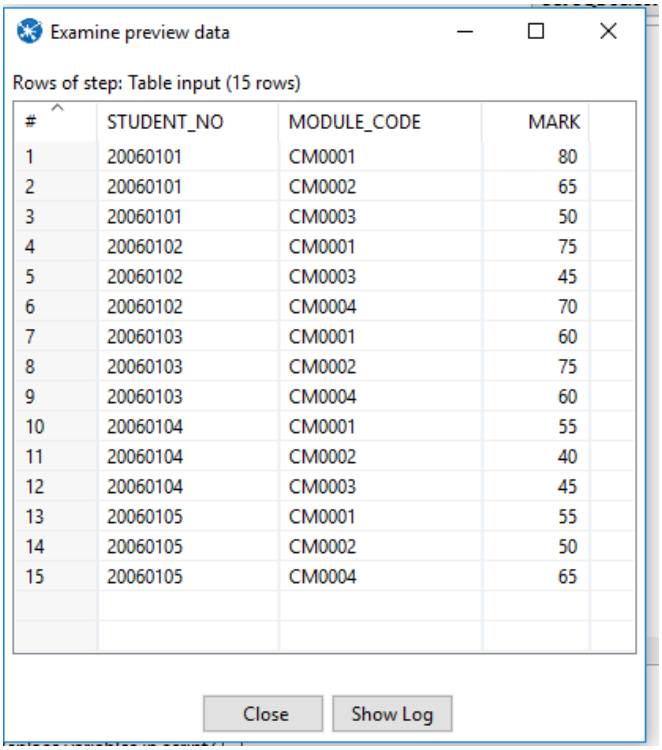


3. Click on “Get SQL select statement” to retrieve table data.



4. Select Table and click on OK. And Click on “Preview” button to preview data of selected table.



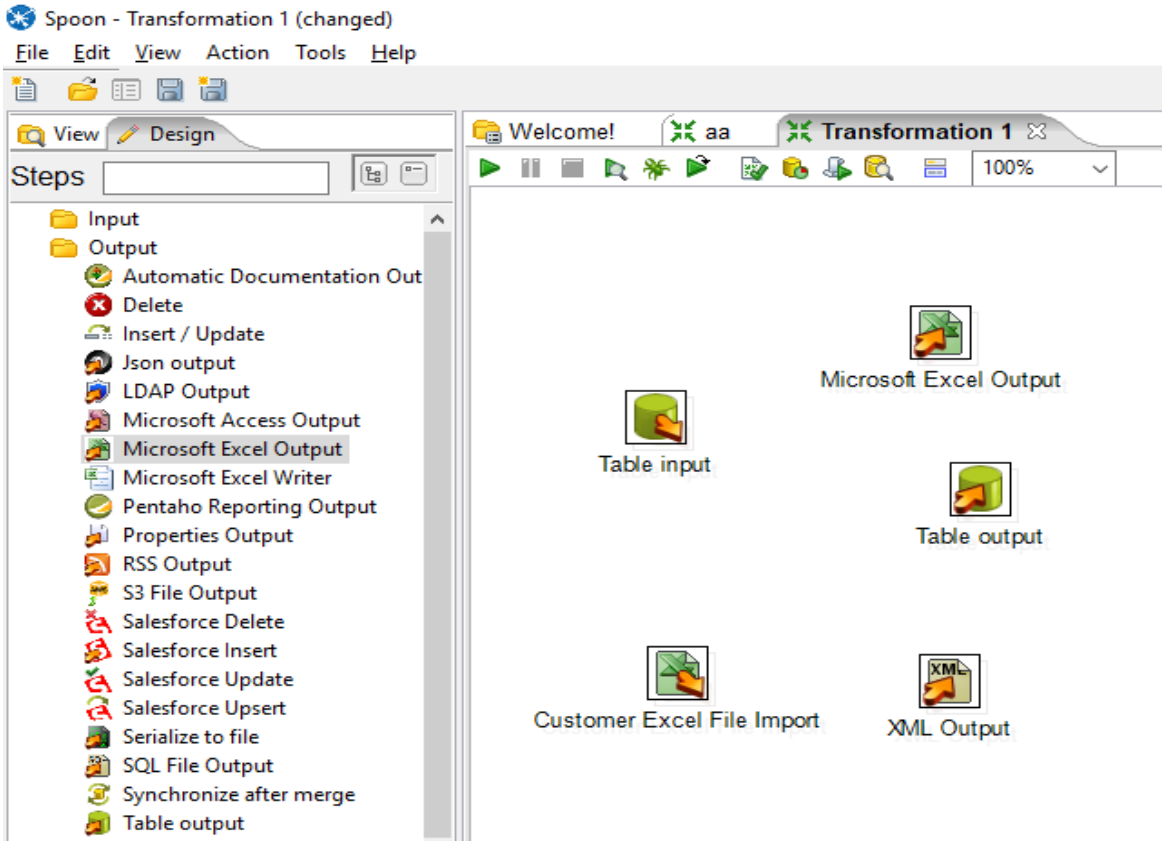


#	STUDENT_NO	MODULE_CODE	MARK
1	20060101	CM0001	80
2	20060101	CM0002	65
3	20060101	CM0003	50
4	20060102	CM0001	75
5	20060102	CM0003	45
6	20060102	CM0004	70
7	20060103	CM0001	60
8	20060103	CM0002	75
9	20060103	CM0004	60
10	20060104	CM0001	55
11	20060104	CM0002	40
12	20060104	CM0003	45
13	20060105	CM0001	55
14	20060105	CM0002	50
15	20060105	CM0004	65

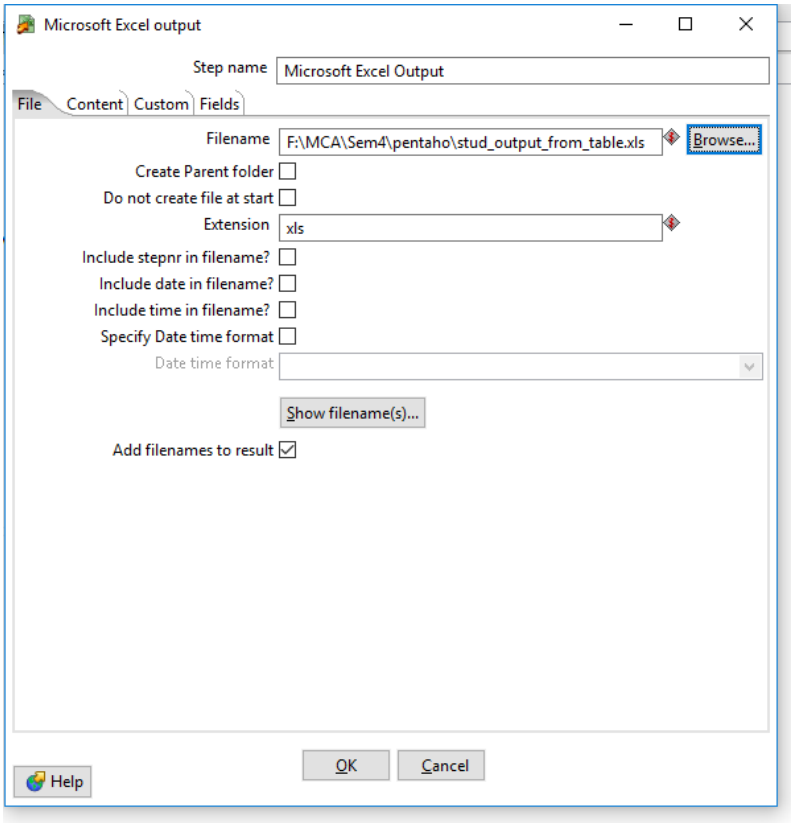
5. Click OK button to close the property window.

B. OUTPUT

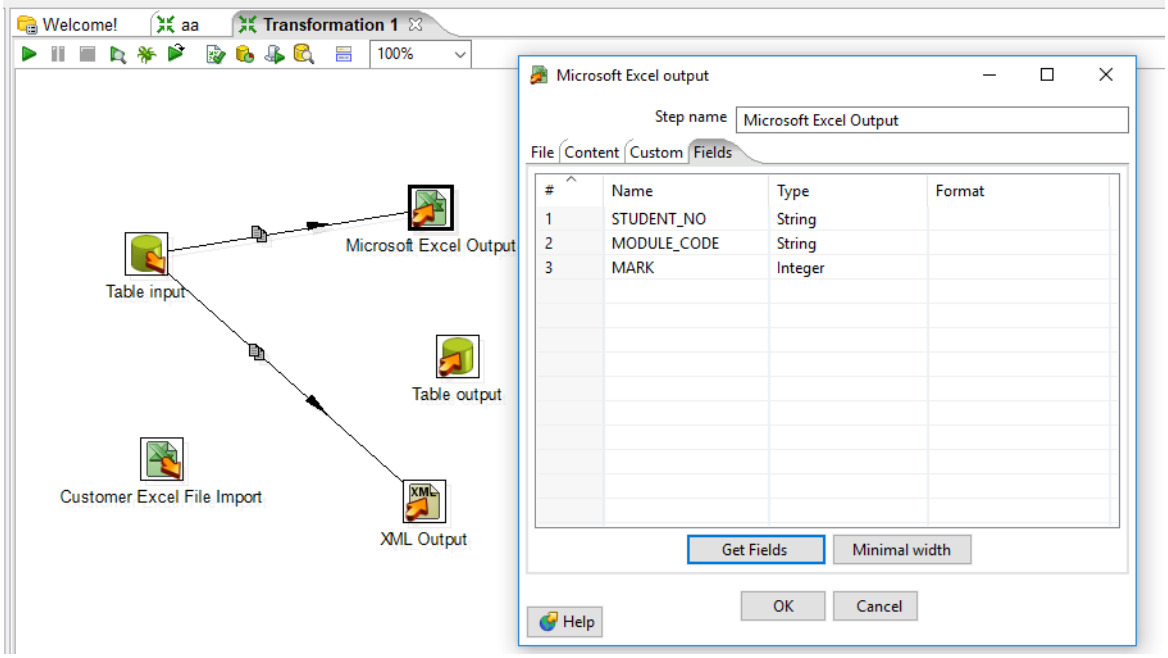
1. Open up the Output folder and drag and drop the Table output /JSON output / XML Output / Microsoft excel output /Text File output step on to the transformation window.



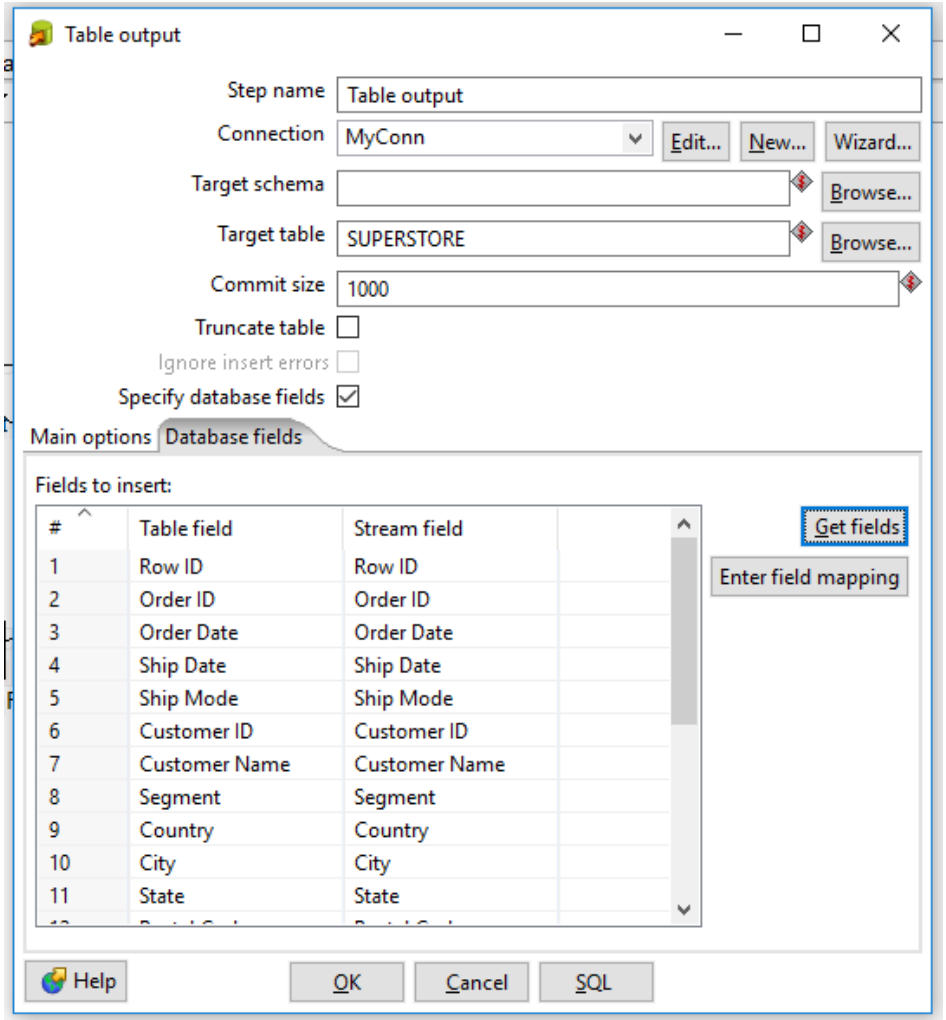
2. Double-click on the Microsoft Excel output step to view its properties Click on the Browse button next to **Filename** field and navigate to the folder to store the output.



3. Make connection to input and output i.e. Create a hop between the **Table Input** and **Microsoft Excel Output** steps Go to field tab of Excel output step and click on “Get fields” button to retrieve column list of table data



4. For Table output, make connection to input to table output. Double click on Table output. Select connection and target table which can be selected from browse button. Check Specify database fields to get Fields and click on “Get fields” button.

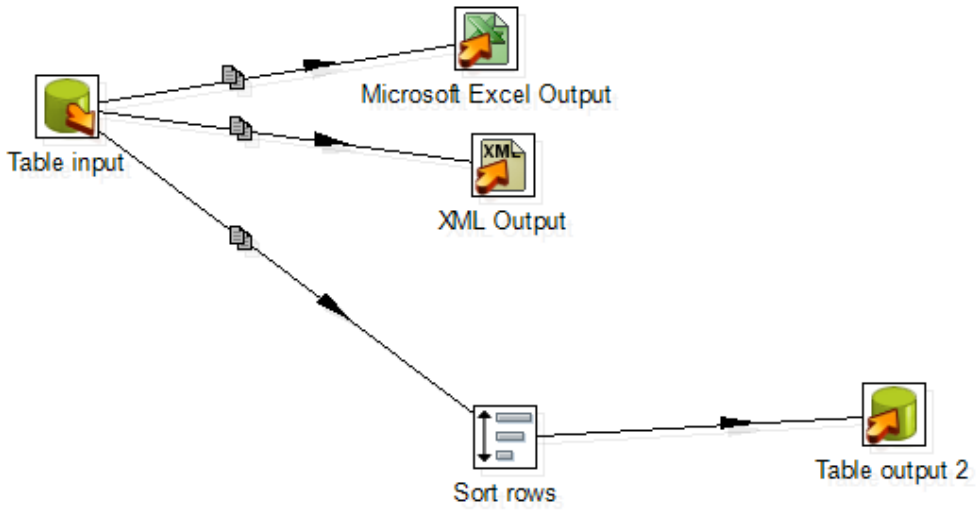


5. Click OK to close the property window.

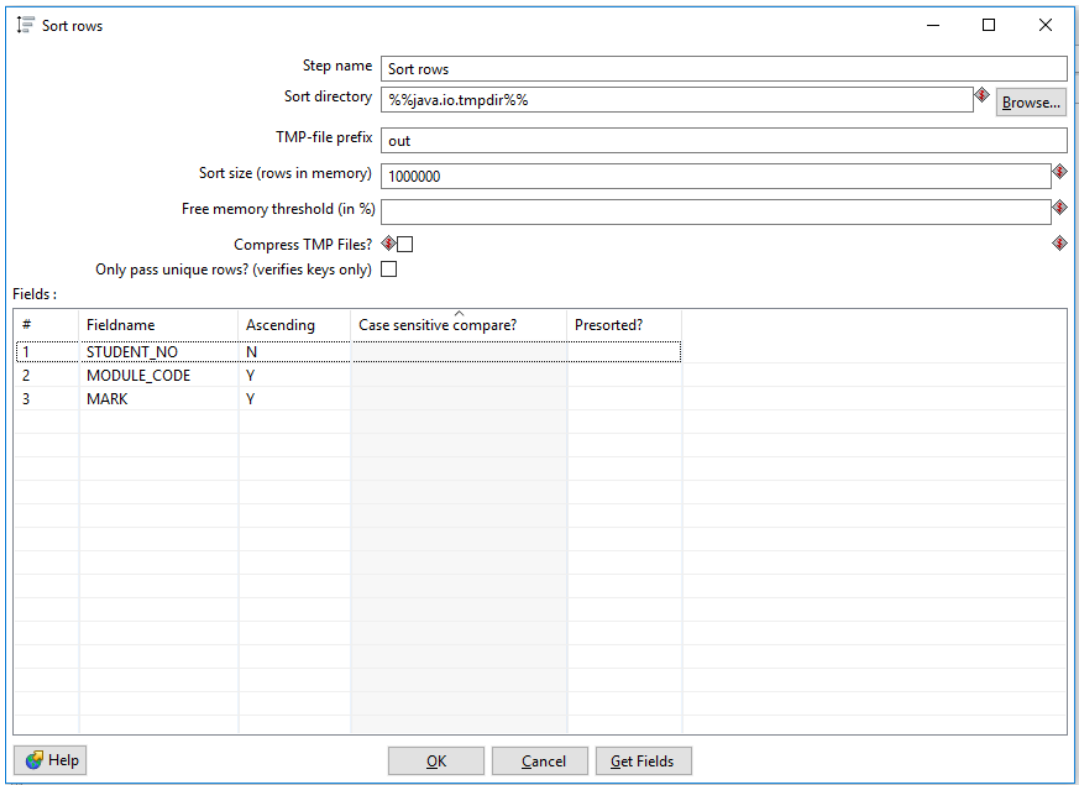
C. Transform

The Pentaho Kettle (PDI) Sort Rows step will sort your data based on field names you specify.

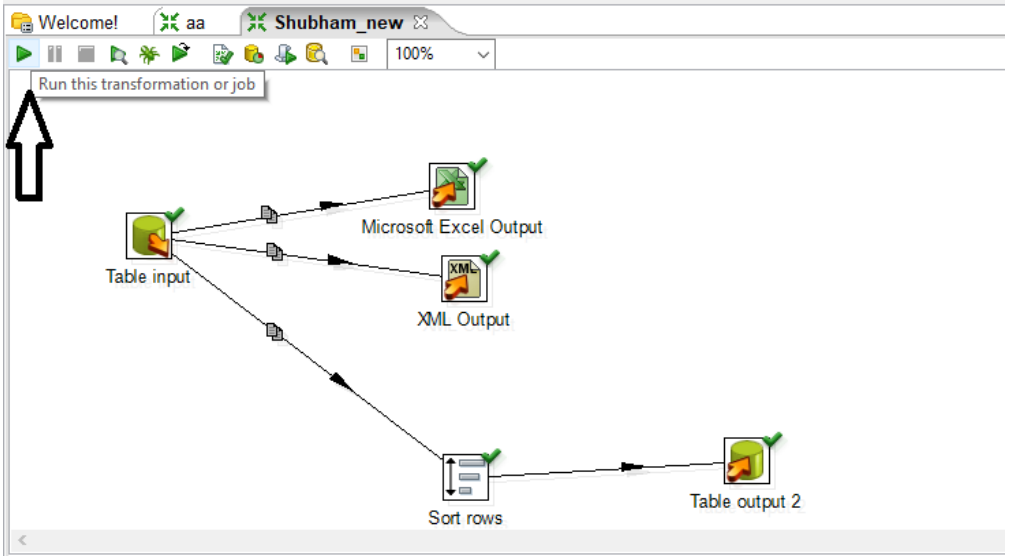
1. Open the Transform folder from the design tab and drag the “sort rows“ to canvas.
2. Create the hop between the Table input and Sort rows to retrieve the information for sorting. And create second hop between “Sort rows” to “ Table output2 “ to get output of transformation in database table.



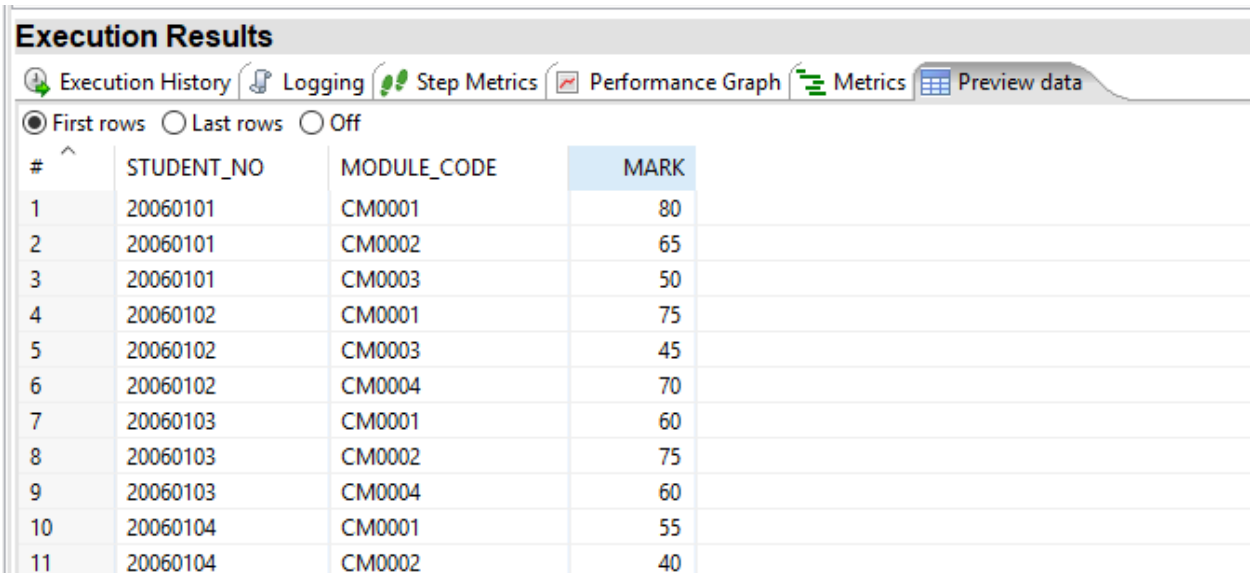
3. Configure the table input to get data from table in which sort row is to be applied as show in above input steps.
4. Double click on the Sort Rows to open property window. Click on the “Get Fields” button to retrieve the table column.
5. In the “fields” all column data is listed. Now selecting “Y” to sort corresponding row in ascending order or “N” to sort in Descending order.



6. Click “OK” to close the window.
7. Open the “Table Output2” property window by double clicking on it. Provide the connection information as shown in above output steps.
8. Select the “Target Table” to store the sorted table data. Click on “Browse” present against the target table.
9. Database explorer will open, then select the output table. And click “OK” button to close window.
10. Click “OK” to close the property window.
11. To run the transformation, click on “Run” button as shown below :



12. A window will open and click on “Launch” button to run. Save the changes if not saved before.
13. In the execution result window, Switch to “Preview data” tab to see the result of Sorted table.



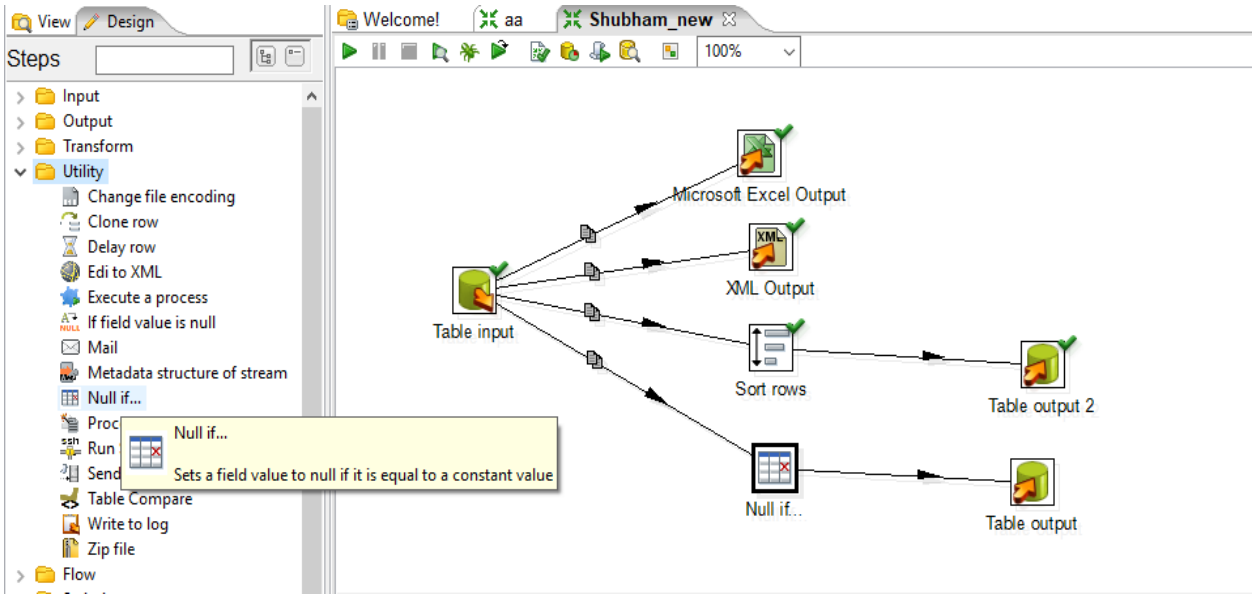
#	STUDENT_NO	MODULE_CODE	MARK
1	20060101	CM0001	80
2	20060101	CM0002	65
3	20060101	CM0003	50
4	20060102	CM0001	75
5	20060102	CM0003	45
6	20060102	CM0004	70
7	20060103	CM0001	60
8	20060103	CM0002	75
9	20060103	CM0004	60
10	20060104	CM0001	55
11	20060104	CM0002	40

D. Utility

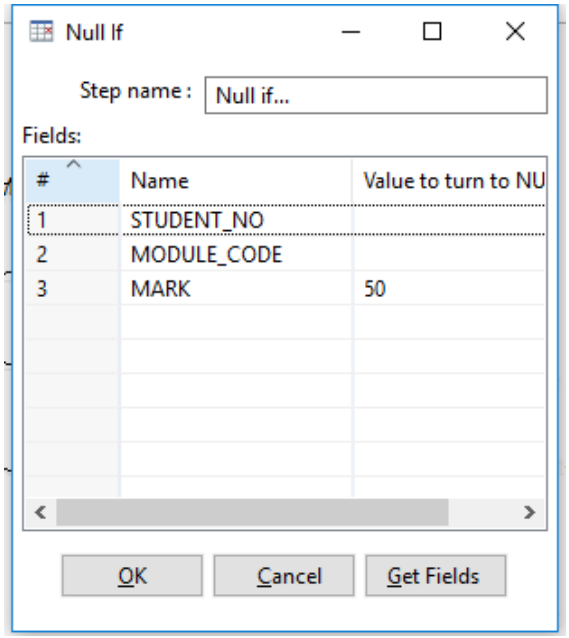
NULL if…

If the string representation of a certain field is equal to the specified value, then the value is set the null (empty).

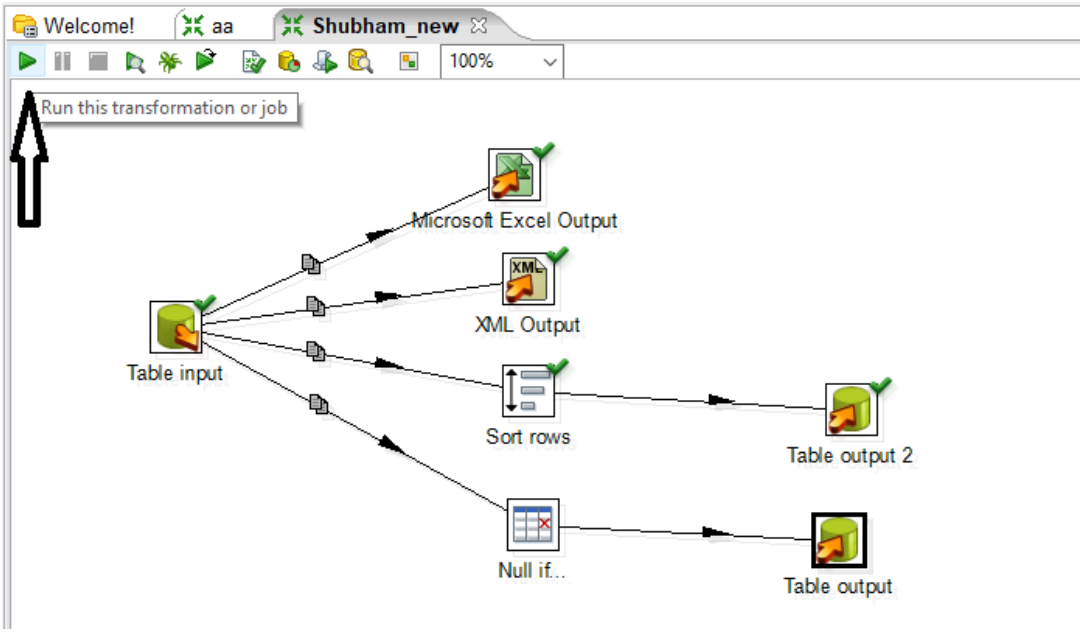
1. Open the utility folder and select the “Null if..” step from the design tab.
2. Create the hop between the “Table input” and “Null if..” . And create second hop between “Null if..” to “ Table output “ to get output of transformation in database table.



3. Configure the table input to get data from table as show in above input steps.
4. Double click on the “Null if..” to open property window. Click on the “Get Fields” button to retrieve the table column.
5. Enter the value which needed to convert to null. For example, in “marks” table those who got 50 marks needed to convert to null. For this write 50 in column mark as shown below:



6. Click “OK” to close the window.
7. Open the “Table Output” property window by double clicking on it. Provide the connection information as shown in above output steps.
8. Select the “Target Table” to store the transformed table data. Click on “Browse” present against the target table.
9. Database explorer will open, then select the output table. And click “OK” button to close window.
10. Click “OK” to close the property window.
11. To run the transformation, click on “Run” button as shown below :



12. A window will open and click on “Launch” button to run. Save the changes if not saved before.
13. In the execution result window, Switch to “Preview data” tab to see the result of Sorted table.

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

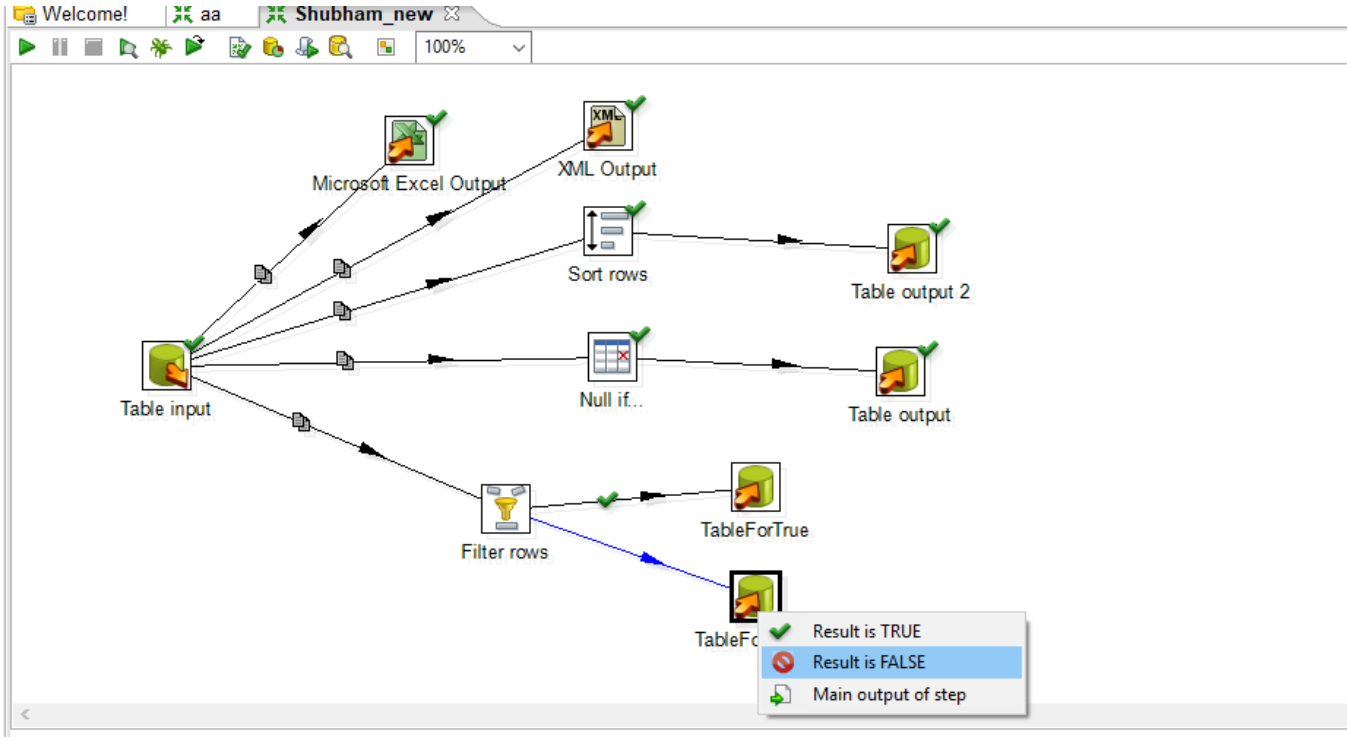
#	STUDENT_NO	MODULE_CODE	MARK
1	20060101	CM0001	80
2	20060101	CM0002	65
3	20060101	CM0003	<null>
4	20060102	CM0001	75
5	20060102	CM0003	45
6	20060102	CM0004	70
7	20060103	CM0001	60
8	20060103	CM0002	75
9	20060103	CM0004	60
10	20060104	CM0001	55
11	20060104	CM0002	40
12	20060104	CM0003	45
13	20060105	CM0001	55
14	20060105	CM0002	<null>
15	20060105	CM0004	65

E. Flow

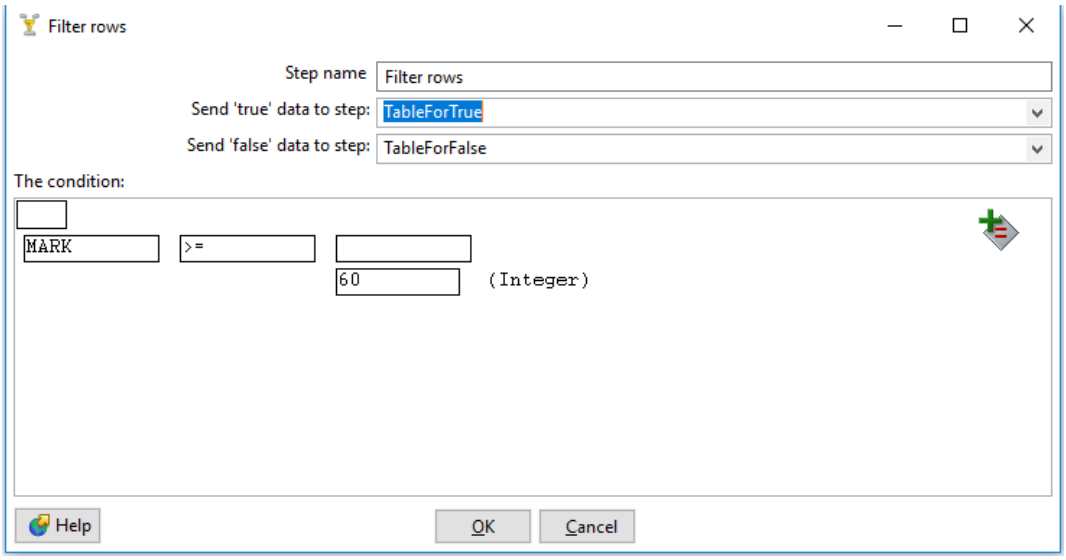
Filter Row

The Filter Rows step allows you to filter rows based on conditions and comparisons. Once this step is connected to a previous step (one or more and receiving input), you can click on the "<field>", "=" and "<value>" areas to construct a condition.

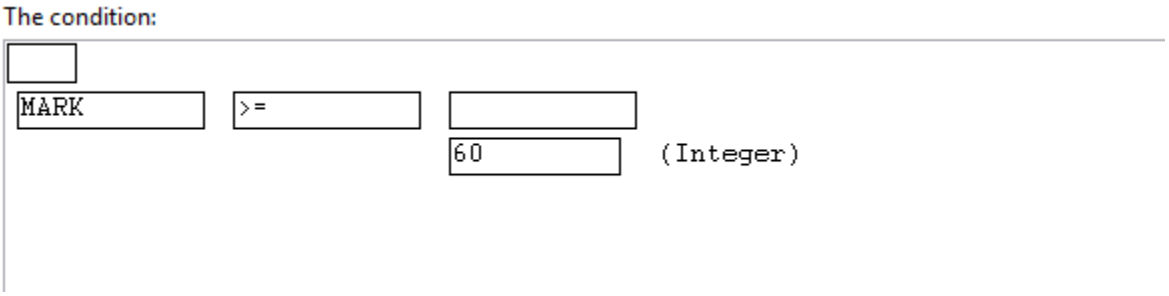
1. Open the Flow folder and select the “Filter row” step from the design tab.
2. Create the hop between the “Table input” and “Filter row” . And create second hop between “Filter Row” to “ Table output3 “ to get output of transformation in database table.



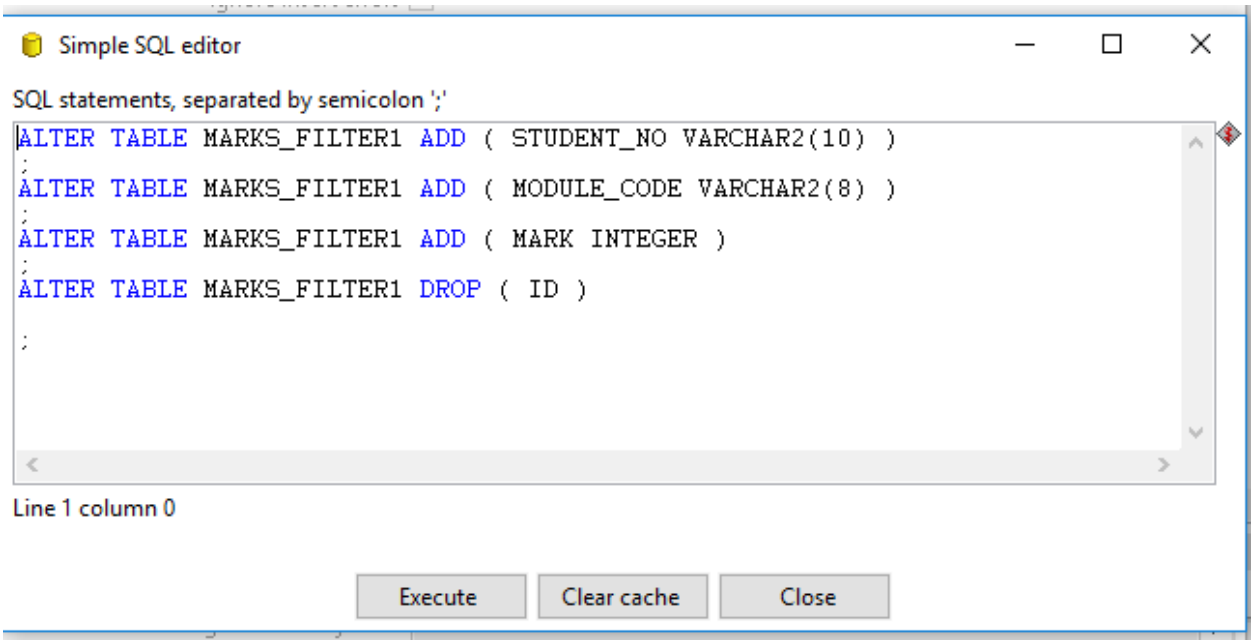
3. Select the Filtering Condition True or False. Then according to condition, the result will filter.
4. Configure the table input to get data from table as show in above input steps.
5. Double click on the “Filter Rows” to open property window. Select the “Table Output” step to send data if true or false as shown below :



6. Now for condition, Select the given boxes to enter condition we want. For example, in this we are going to filter student data according to their marks with following condition.



7. Click “OK” to close this window.
8. Open the “TableForTrue” property window by double clicking on it. Provide the connection information as shown in above output steps.
9. Select the “Target Table” to store the table data which satisfy the condition. Click on “Browse” present against the target table.
10. Database explorer will open, then select the output table. And click “OK” button to close window.
11. Click “OK” to close the property window.
12. Open the “TableForFalse” property window by double clicking on it. Provide the connection information as shown in above output steps.
13. Select the “Target Table” to store the table data which does not satisfy the condition. Click on “Browse” present against the target table.
14. Database explorer will open, then select the output table. And click “OK” button to close window.
15. Click on “SQL” button to verify output table schema following screen will appear.



16. Then click on “Execute” to update table schema.
17. Click “OK” to close the property window.

18. Now run the transformation by clicking on run button.
19. A window will open and click on “Launch” button to run. Save the changes if not saved before.
20. In the execution result window, Switch to “Preview data” tab to see the result of Sorted table.

Result of TableForTrue step :

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	STUDENT_NO	MODULE_CODE	MARK
1	20060101	CM0001	80
2	20060101	CM0002	65
3	20060102	CM0001	75
4	20060102	CM0004	70
5	20060103	CM0001	60
6	20060103	CM0002	75
7	20060103	CM0004	60
8	20060105	CM0004	65

Result of TableForFalse step:

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	STUDENT_NO	MODULE_CODE	MARK
1	20060101	CM0003	50
2	20060102	CM0003	45
3	20060104	CM0001	55
4	20060104	CM0002	40
5	20060104	CM0003	45
6	20060105	CM0001	55
7	20060105	CM0002	50