PRACTICAL – 4

Aim: Implementation of ETL transformation using Pentaho.

Theory:

Pentaho Reporting is a suite (collection of tools) for creating relational and analytical reporting. Using Pentaho, we can transform complex data into meaningful reports and draw information out of them. Pentaho supports creating reports in various formats such as HTML, Excel, PDF, Text, CSV, and xml.

Pentaho can accept data from different data sources including SQL databases, OLAP data sources, and even the Pentaho Data Integration ETL tool.

Input File: CSV file input

Transformation: Sort rows, Unique, Concat fields, Number range, Calculator.

Output File: Microsoft excel output

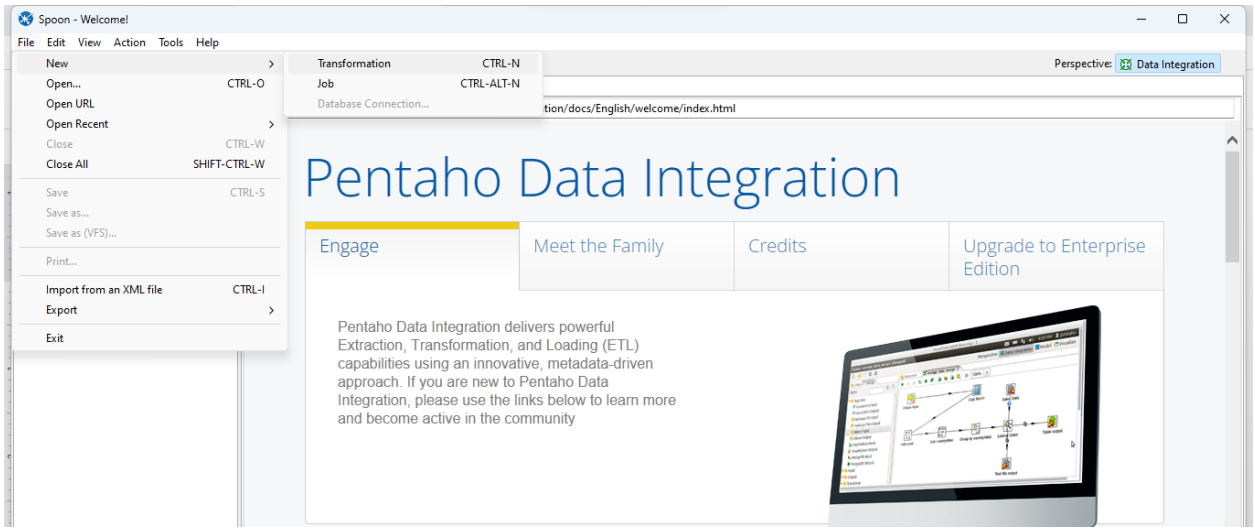
1. Sort Rows Transformation

Steps to implement Data integration using Pentaho.

Step:1 Install the Pentaho set up, and run **spoon batch** file.

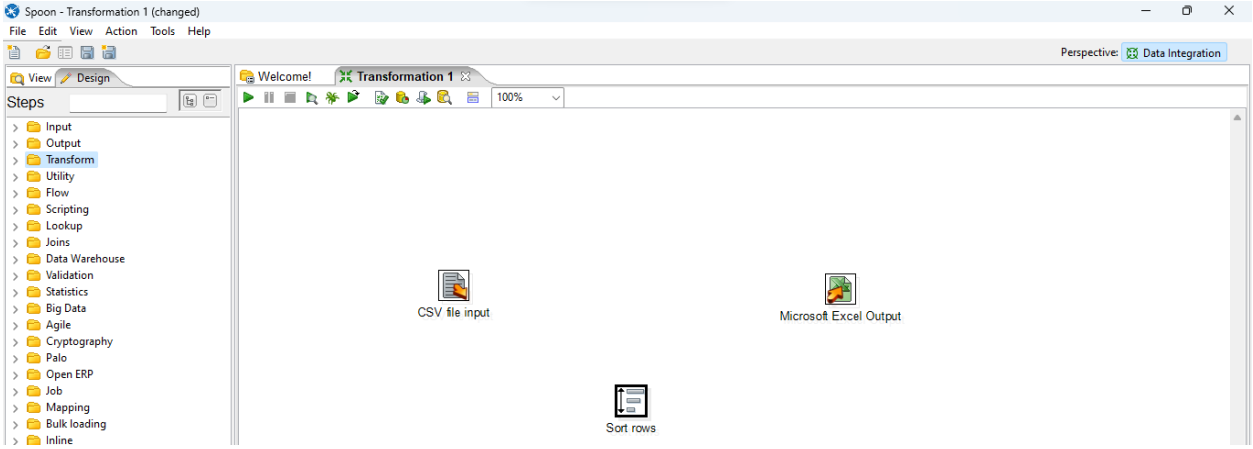
[Pentaho -> Software -> Data-integration -> Spoon batch]

Step:2 Create Transformation. To create click on File-> New -> Transformation



Step:3 Now we have to add input, output and transformation. To add click on input menu and select required input, similarly for output and for transformation.

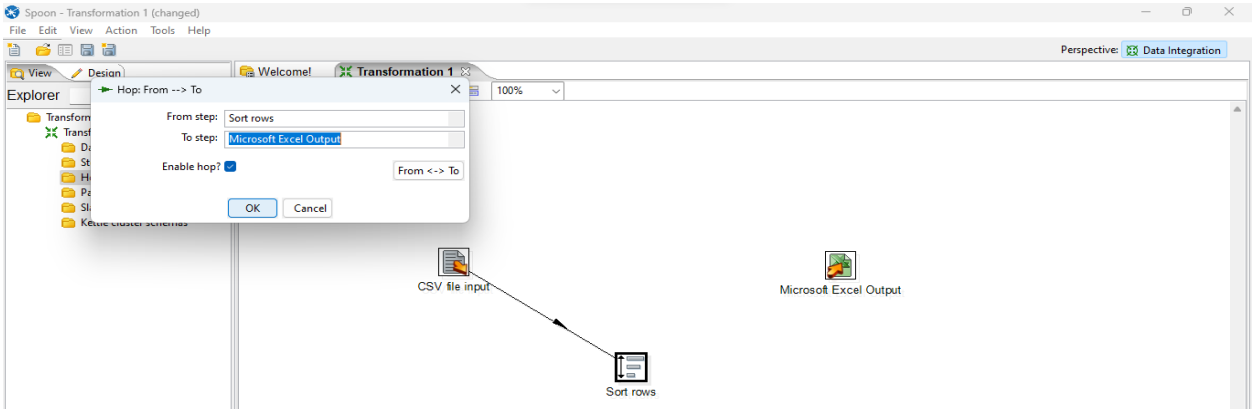
In this case I have selected input as **CSV file input**, output as **Microsoft excel output**, and transformation as **Sort rows.**

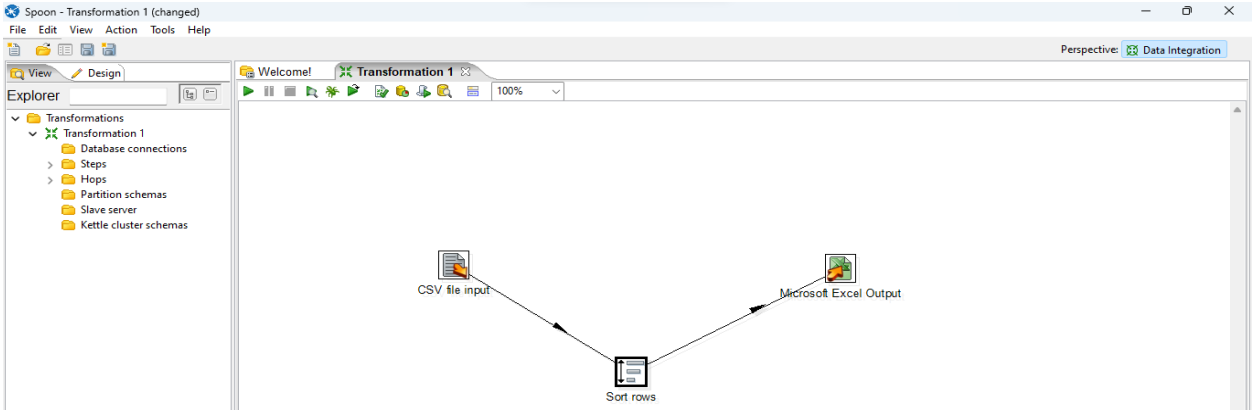


Step:4 To build connection between input, transformation and output.

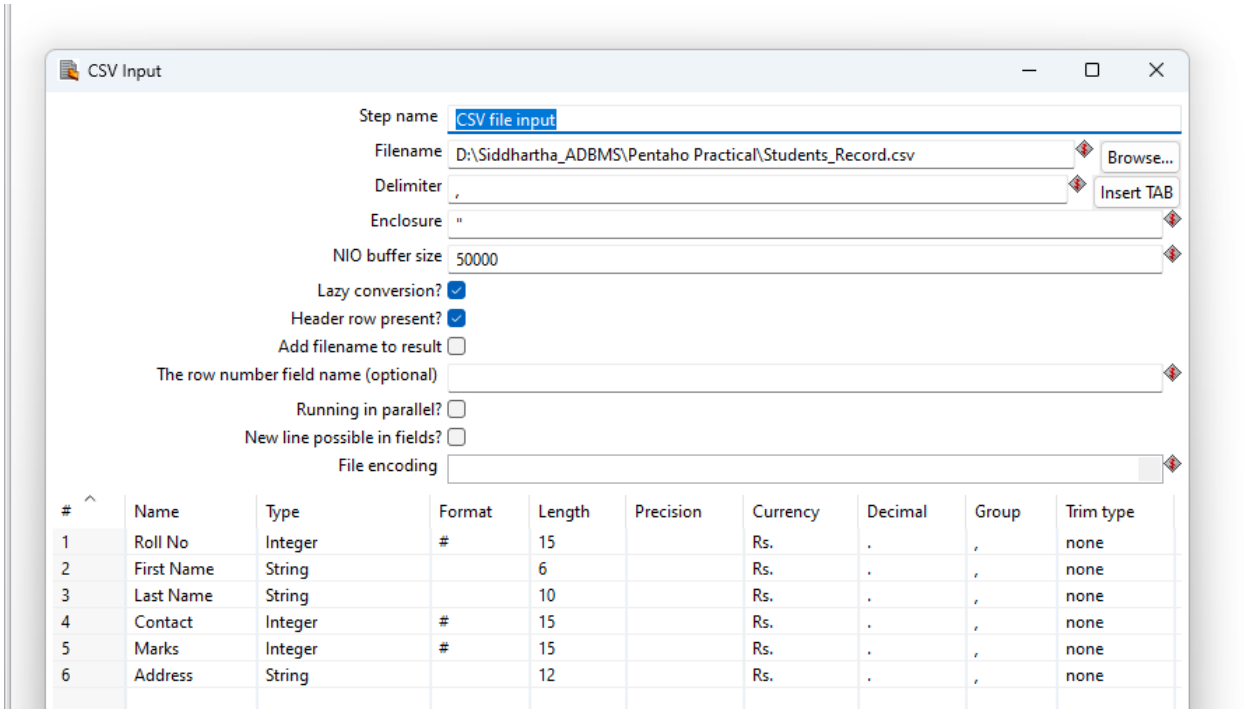
Select view, go to **hop** -> add source and destination and connection will be shown.

Here first ill connect input file to sort rows and then sort rows to excel output.

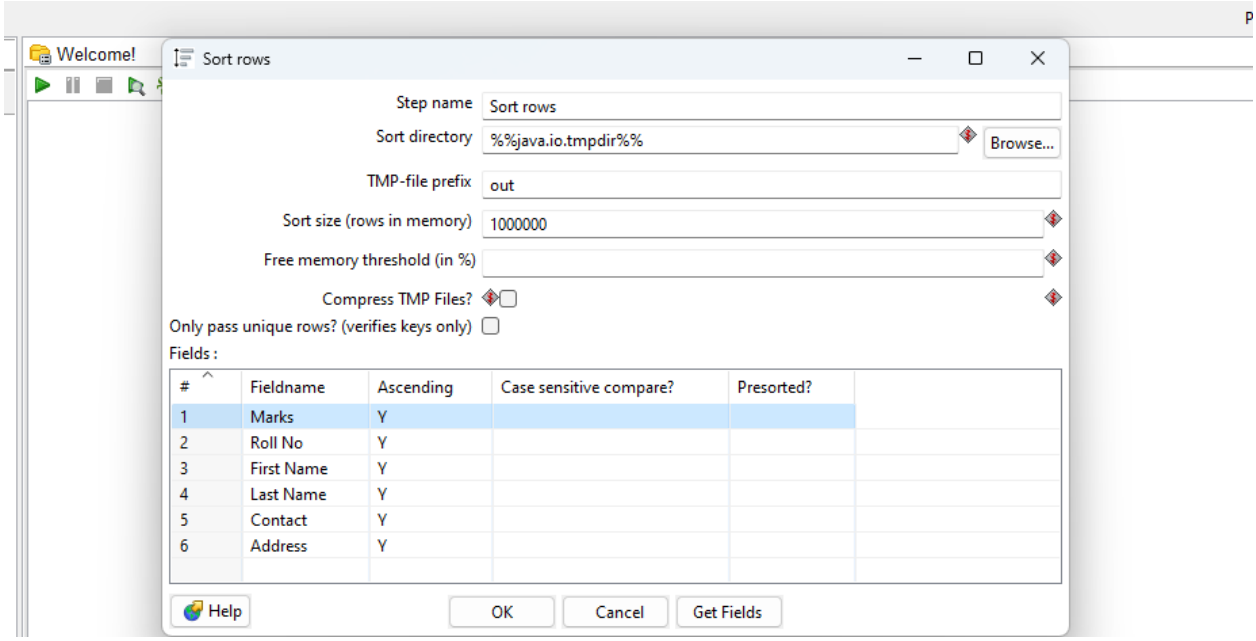




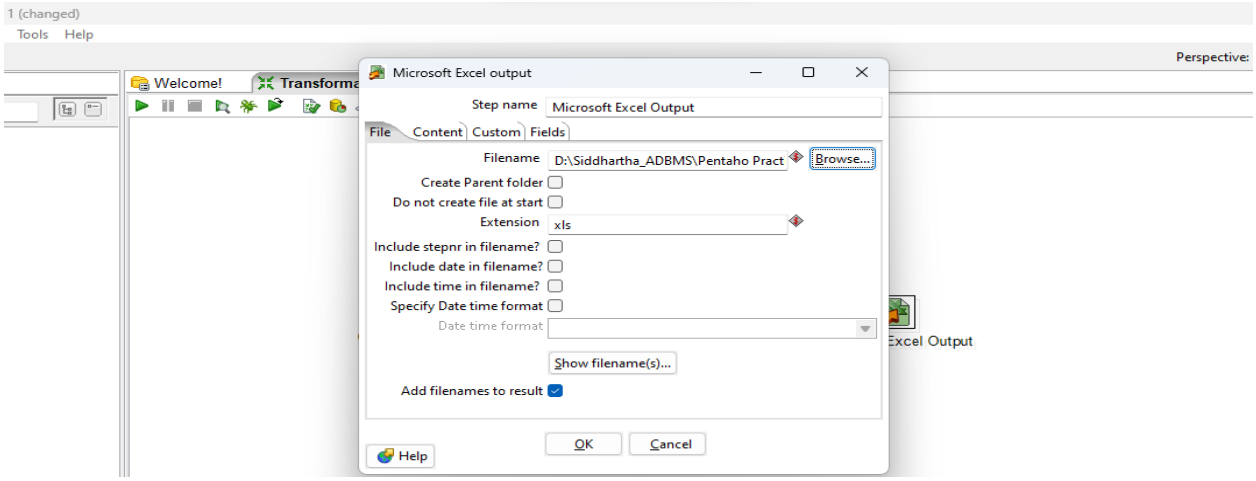
Step:5 After successful hop connection, click on CSV input file and add destination of your input file.



Step:6 Now click on the transformation (Sort Rows) and add the required fields by clicking get fields. Here we can sort the file by mentioned field name.

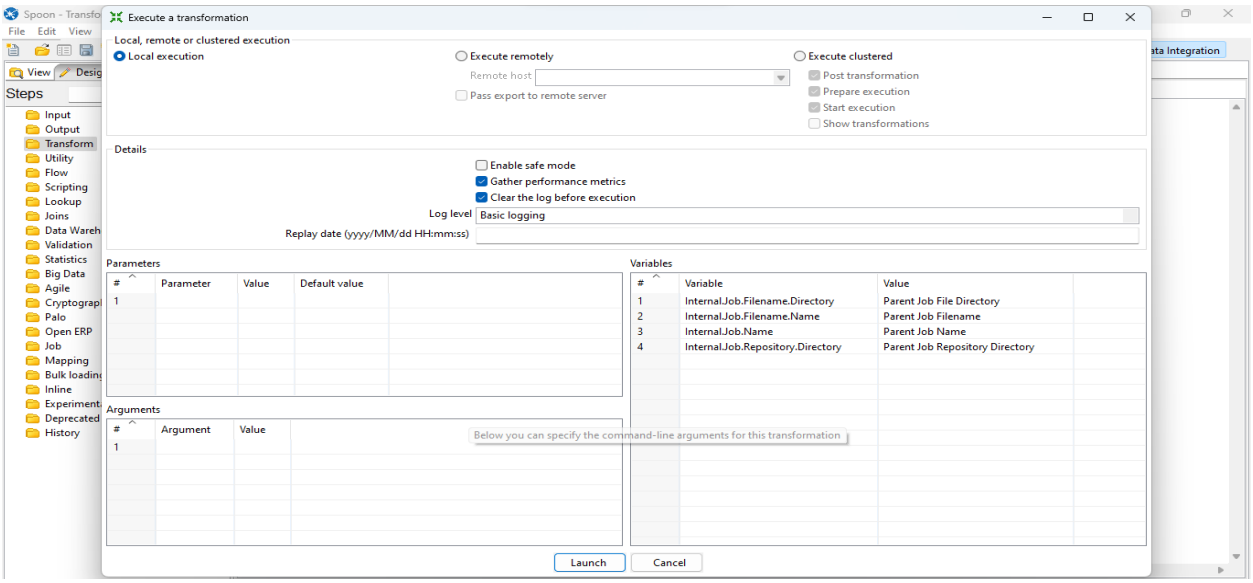


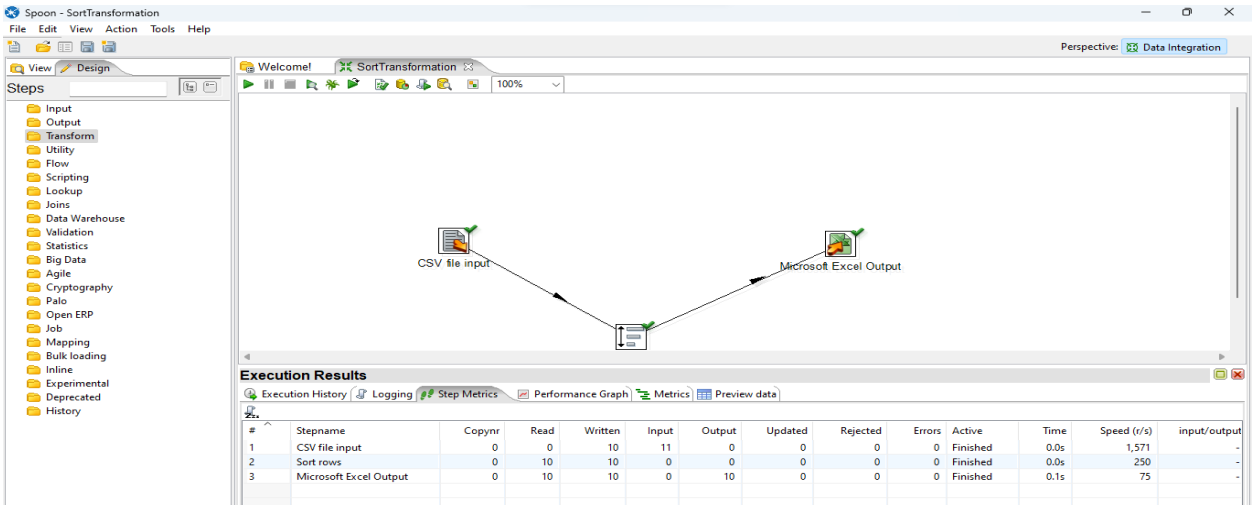
Step:7 Click on Microsoft Excel output to give destination for output file.



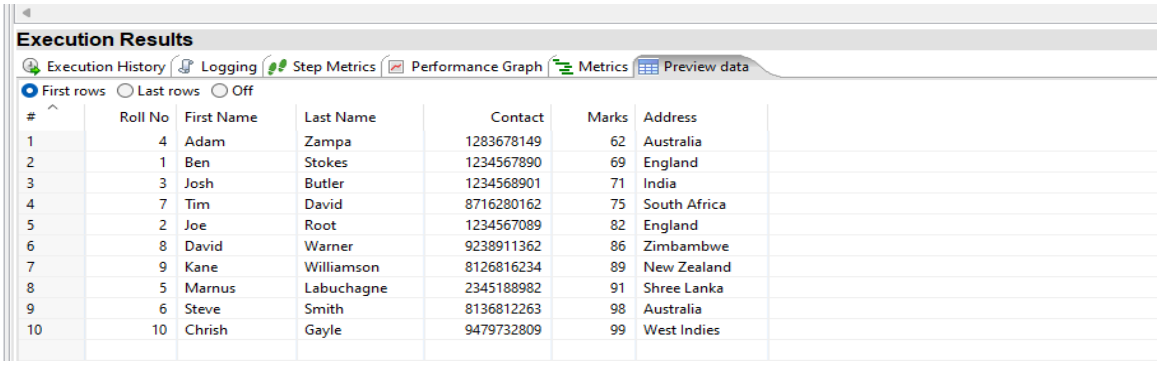
Step:8 Now click on Run and launch the transformation after saving the transformation.

In my case I have saved my transformation named as **SortTransformation.ktr**





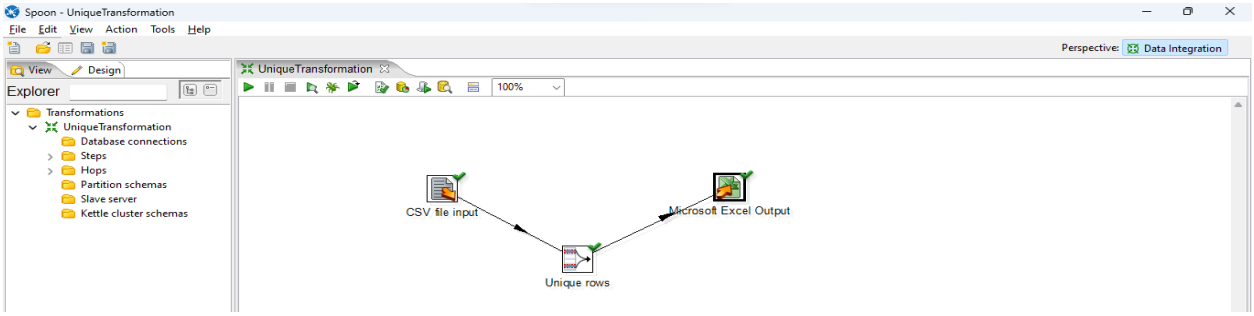
Step:9 After successful run we can see the status of execution and to see the changes inside file, click on Preview Data.

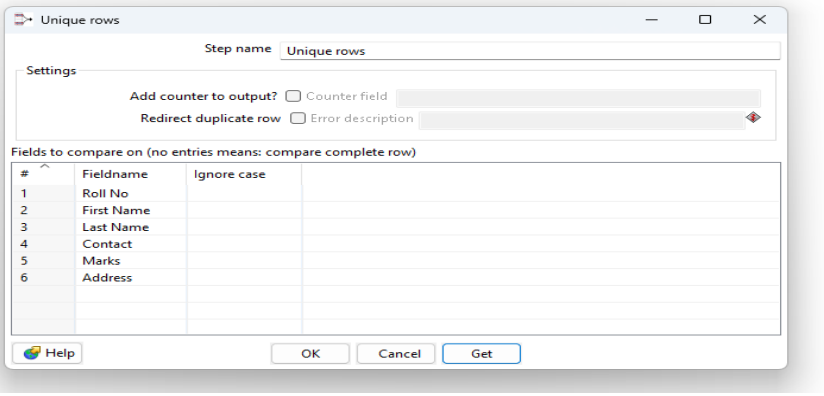


#	Roll No	First Name	Last Name	Contact	Marks	Address
1	4	Adam	Zampa	1283678149	62	Australia
2	1	Ben	Stokes	1234567890	69	England
3	3	Josh	Butler	1234568901	71	India
4	7	Tim	David	8716280162	75	South Africa
5	2	Joe	Root	1234567089	82	England
6	8	David	Warner	9238911362	86	Zimbambwe
7	9	Kane	Williamson	8126816234	89	New Zealand
8	5	Marnus	Labuchagne	2345188982	91	Shree Lanka
9	6	Steve	Smith	8136812263	98	Australia
10	10	Chrish	Gayle	9479732809	99	West Indies

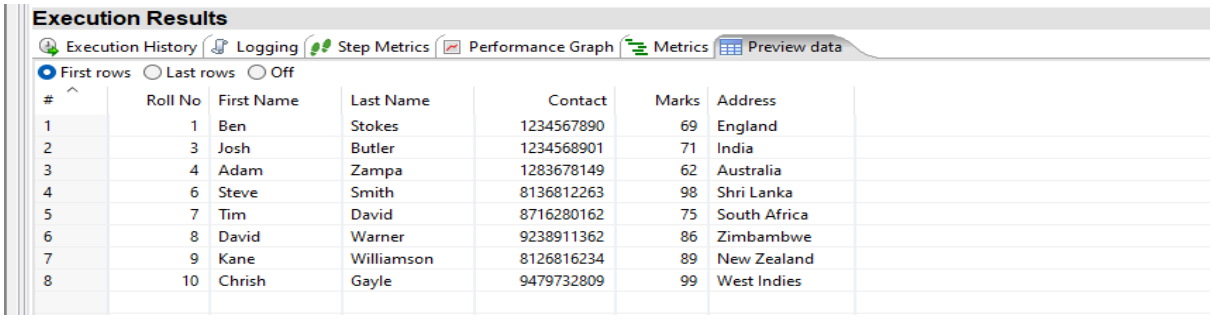
2. Unique Transformation

Returns the unique rows only and delete immediately repeated row.





#	Fieldname	Ignore case
1	Roll No	
2	First Name	
3	Last Name	
4	Contact	
5	Marks	
6	Address	

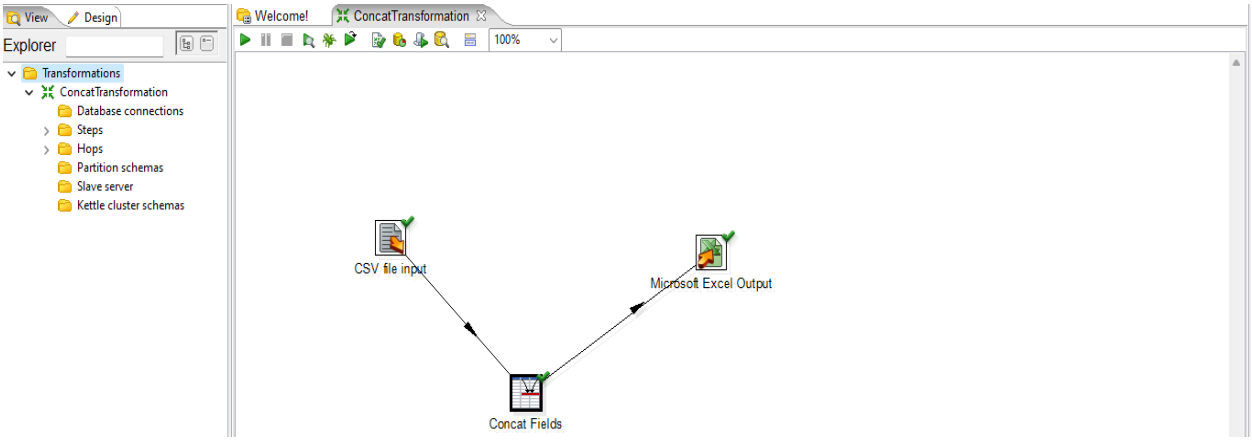


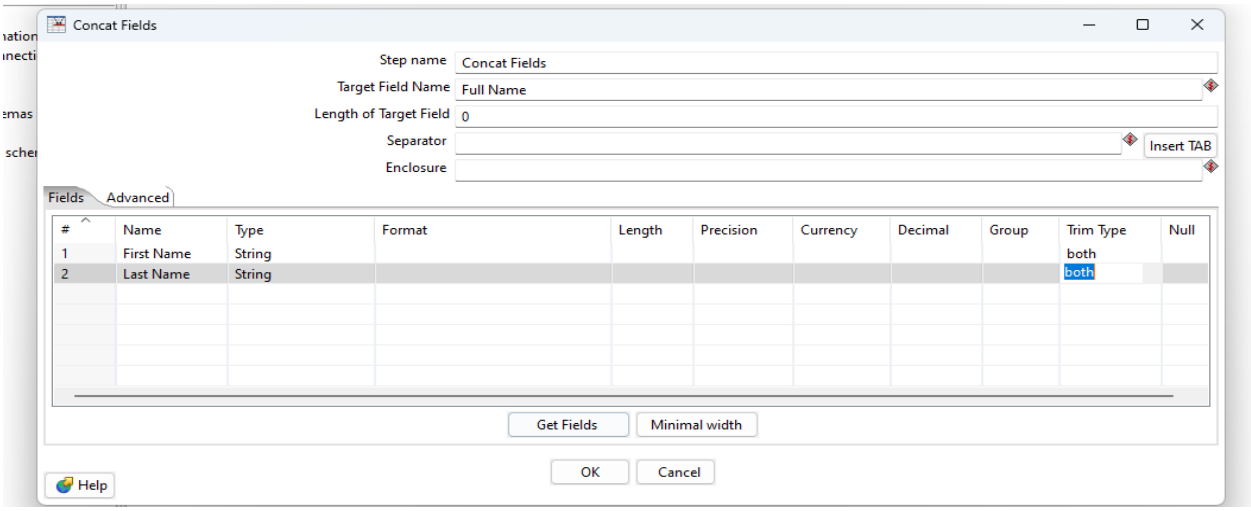
#	Roll No	First Name	Last Name	Contact	Marks	Address
1	1	Ben	Stokes	1234567890	69	England
2	3	Josh	Butler	1234568901	71	India
3	4	Adam	Zampa	1283678149	62	Australia
4	6	Steve	Smith	8136812263	98	Shri Lanka
5	7	Tim	David	8716280162	75	South Africa
6	8	David	Warner	9238911362	86	Zimbambwe
7	9	Kane	Williamson	8126816234	89	New Zealand
8	10	Chrish	Gayle	9479732809	99	West Indies

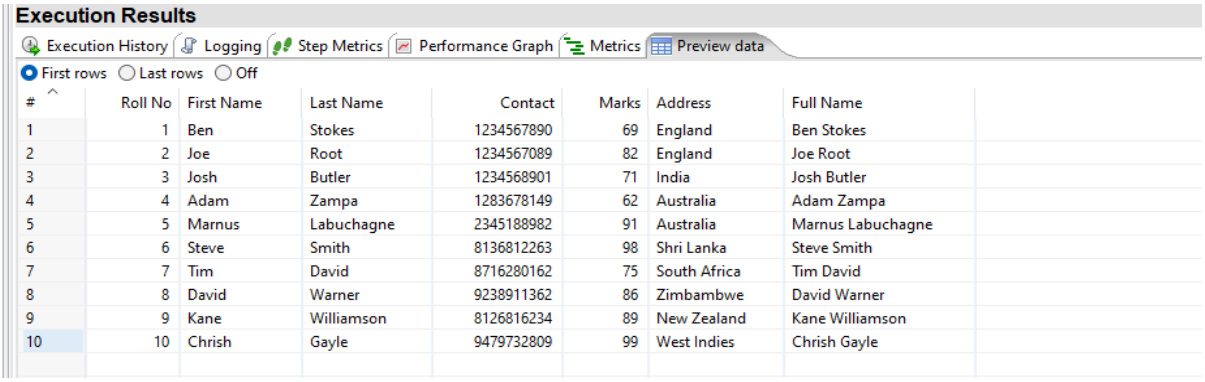
3. Concat Fields Transformation

Used to concat/merge two different fields.

In my case I am concating First Name and Last Name, storing it into new field Full Name

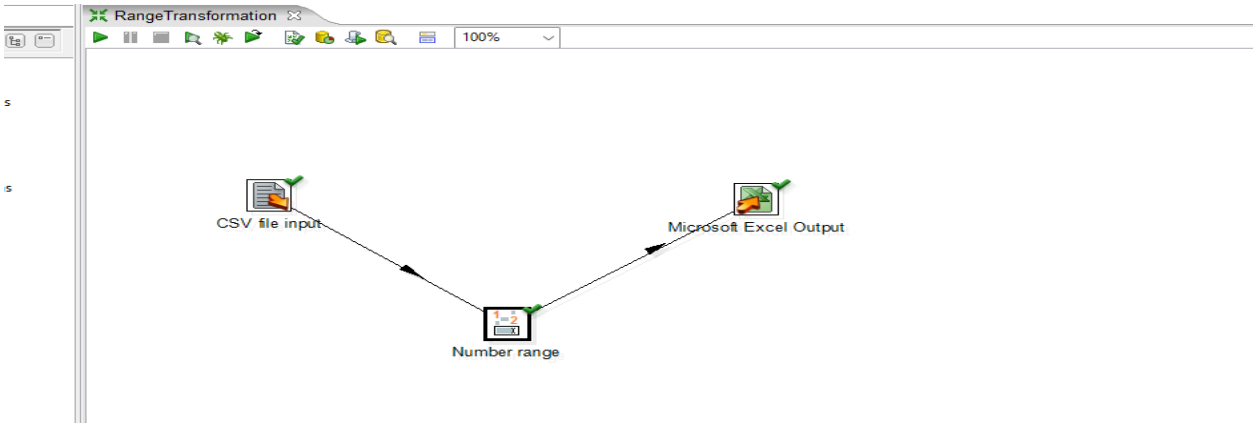


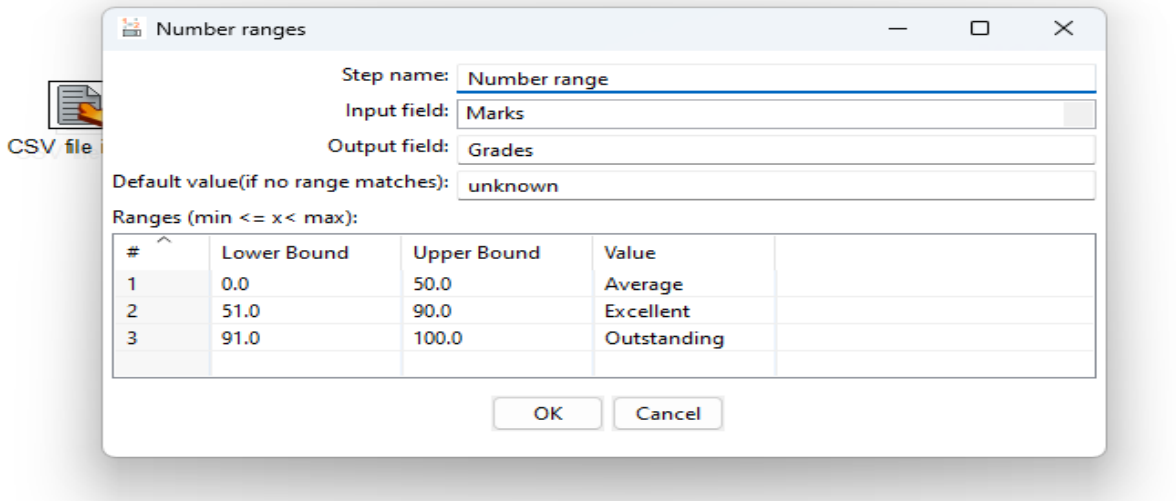


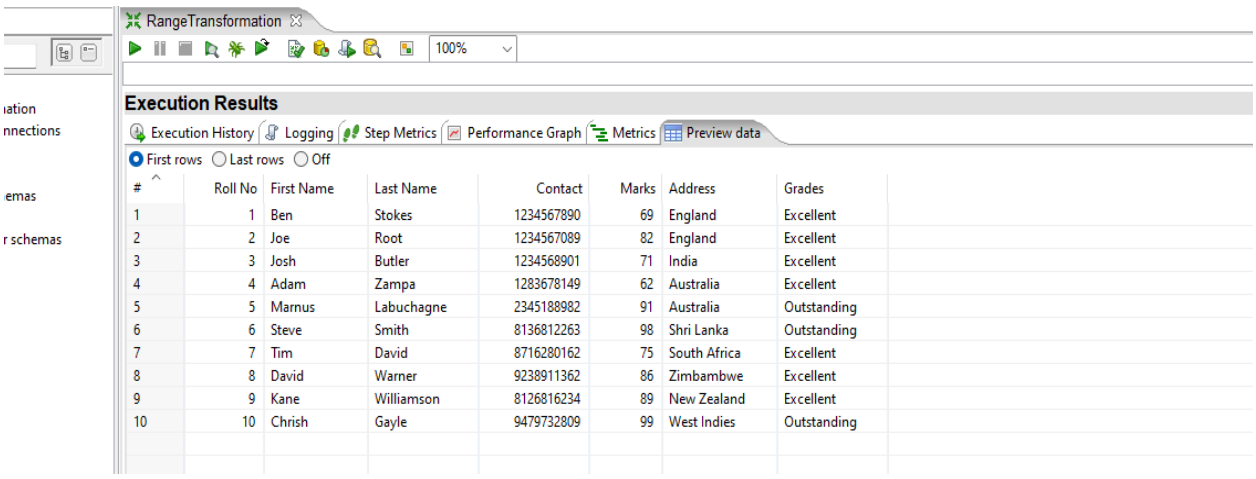


#	Roll No	First Name	Last Name	Contact	Marks	Address	Full Name
1	1	Ben	Stokes	1234567890	69	England	Ben Stokes
2	2	Joe	Root	1234567089	82	England	Joe Root
3	3	Josh	Butler	1234568901	71	India	Josh Butler
4	4	Adam	Zampa	1283678149	62	Australia	Adam Zampa
5	5	Marnus	Labuchagne	2345188982	91	Australia	Marnus Labuchagne
6	6	Steve	Smith	8136812263	98	Shri Lanka	Steve Smith
7	7	Tim	David	8716280162	75	South Africa	Tim David
8	8	David	Warner	9238911362	86	Zimbambwe	David Warner
9	9	Kane	Williamson	8126816234	89	New Zealand	Kane Williamson
10	10	Chrish	Gayle	9479732809	99	West Indies	Chrish Gayle

4. Number Range Transformation

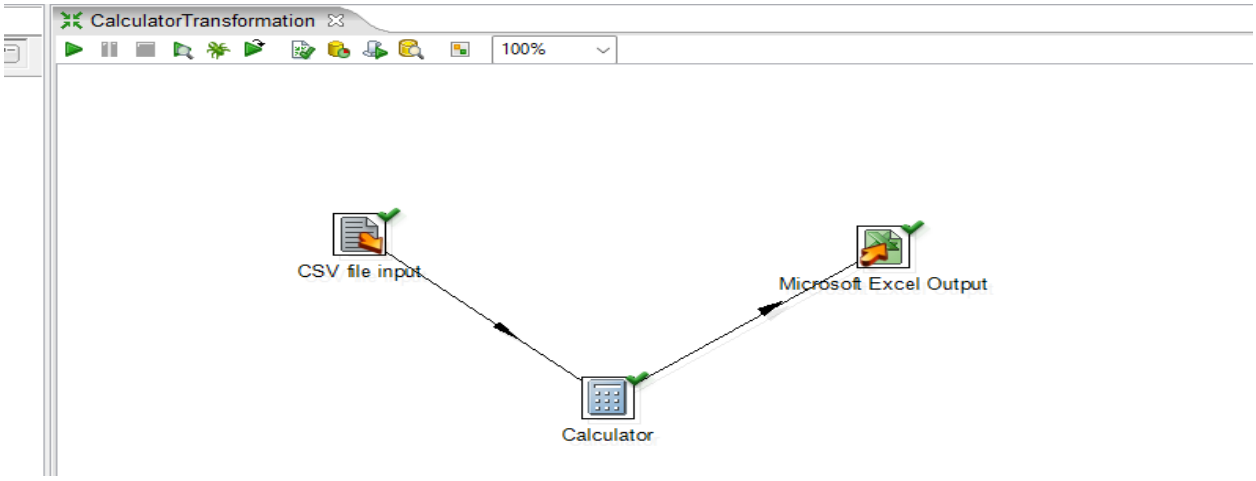


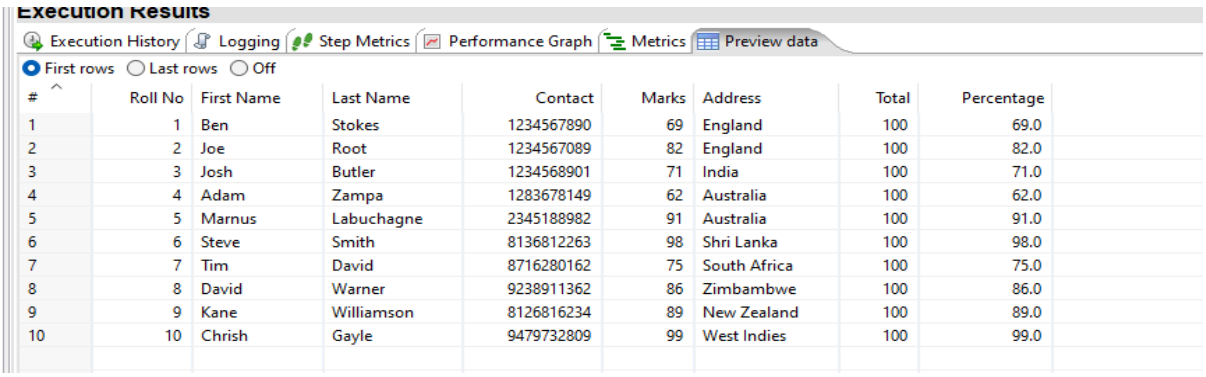




#	Roll No	First Name	Last Name	Contact	Marks	Address	Grades
1	1	Ben	Stokes	1234567890	69	England	Excellent
2	2	Joe	Root	1234567089	82	England	Excellent
3	3	Josh	Butler	1234568901	71	India	Excellent
4	4	Adam	Zampa	1283678149	62	Australia	Excellent
5	5	Marnus	Labuchagne	2345188982	91	Australia	Outstanding
6	6	Steve	Smith	8136812263	98	Shri Lanka	Outstanding
7	7	Tim	David	8716280162	75	South Africa	Excellent
8	8	David	Warner	9238911362	86	Zimbambwe	Excellent
9	9	Kane	Williamson	8126816234	89	New Zealand	Excellent
10	10	Chrish	Gayle	9479732809	99	West Indies	Outstanding

5. Calculator Transformation





#	Roll No	First Name	Last Name	Contact	Marks	Address	Total	Percentage
1	1	Ben	Stokes	1234567890	69	England	100	69.0
2	2	Joe	Root	1234567089	82	England	100	82.0
3	3	Josh	Butler	1234568901	71	India	100	71.0
4	4	Adam	Zampa	1283678149	62	Australia	100	62.0
5	5	Marnus	Labuchagne	2345188982	91	Australia	100	91.0
6	6	Steve	Smith	8136812263	98	Shri Lanka	100	98.0
7	7	Tim	David	8716280162	75	South Africa	100	75.0
8	8	David	Warner	9238911362	86	Zimbambwe	100	86.0
9	9	Kane	Williamson	8126816234	89	New Zealand	100	89.0
10	10	Chrish	Gayle	9479732809	99	West Indies	100	99.0

Input file: Table input

Transformation: Select values, String Operation, Row Normalizer, Sort rows, Concat fields

Output file: Microsoft excel file

For Database firstly we have to connect our input file to database, and from there it will fetch all the records of different table, and accordingly we can select any one table on which we are going to perform transformation.

Step: 1 After selecting input file as table input, click on table input and enter

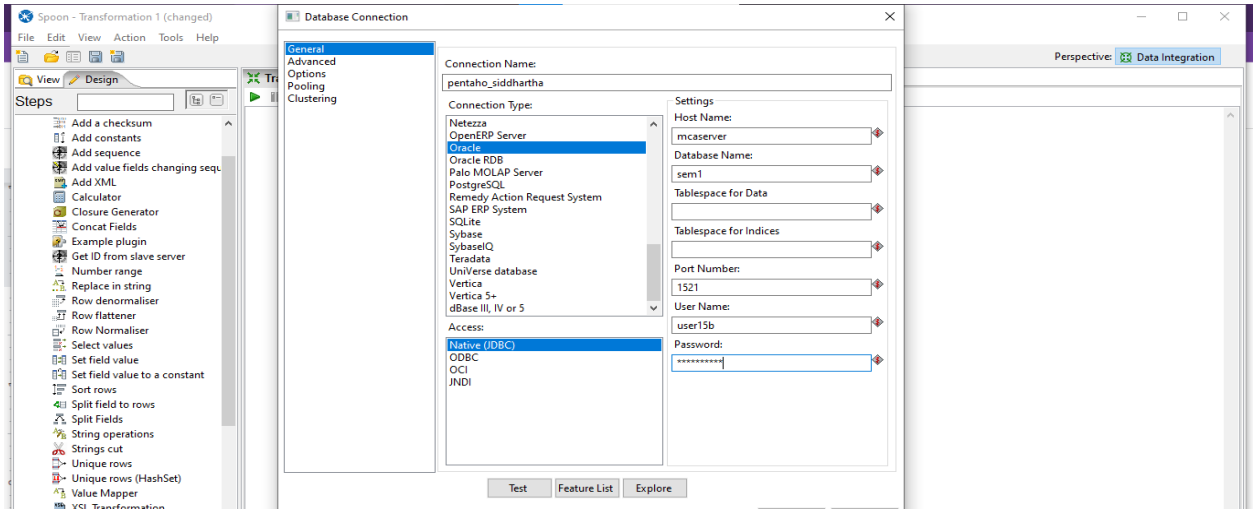
Host Name: mcaserver

Database Name: sem1

User Name: user15b

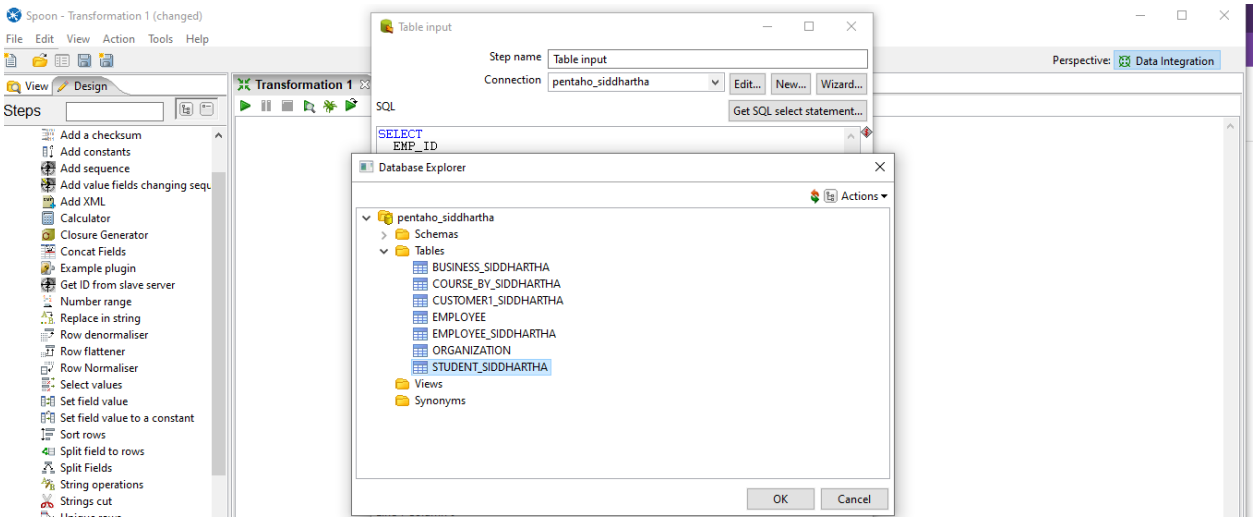
Password: _____

Click on test and you will receive a pop up of successful connection.

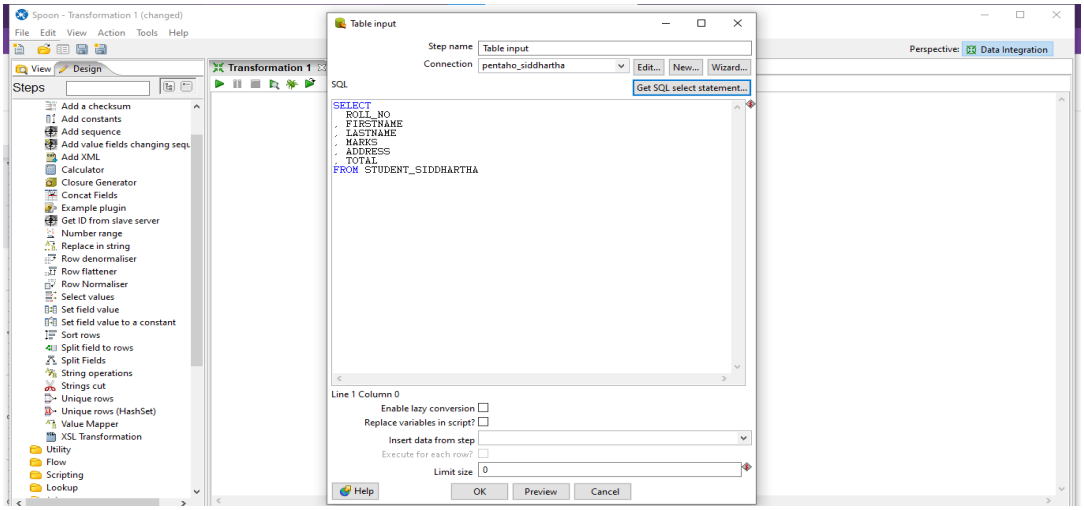


Step:2 Click on get SQL select statement, it will run select query in and fetch all the tables created inside your database.

Here I am selecting table student_siddhartha.



Step:3 after selecting required table we can see the fields/attributes of table and we can also preview all the data of table student_siddhartha.



Examine preview data

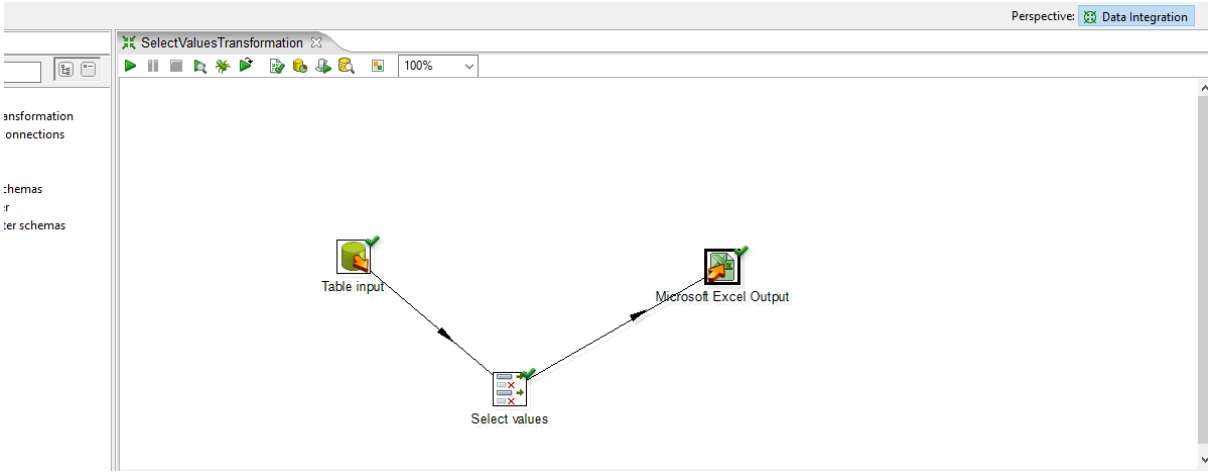
Rows of step: Table input (10 rows)

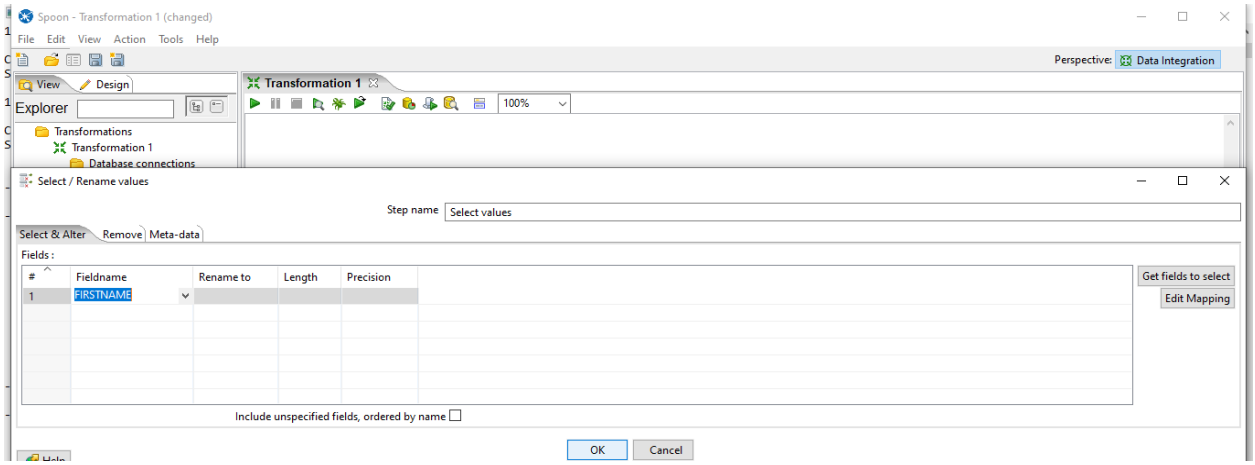
#	ROLL_NO	FIRSTNAME	LASTNAME	MARKS	ADDRESS	TOTAL
1	1	Ben	Stokes	69	England	100
2	2	Joe	Root	82	England	100
3	3	Josh	Butler	71	India	100
4	4	Adam	Zampa	62	Australia	100
5	5	Marnus	Labuchagne	91	Australia	100
6	6	Steve	Smith	98	Shri Lanka	100
7	7	Tim	David	75	South Africa	100
8	8	David	Warner	86	Zimbambwe	100
9	9	Kane	Williamson	89	New Zealand	100
10	10	Chrish	Gayle	99	West Indies	100

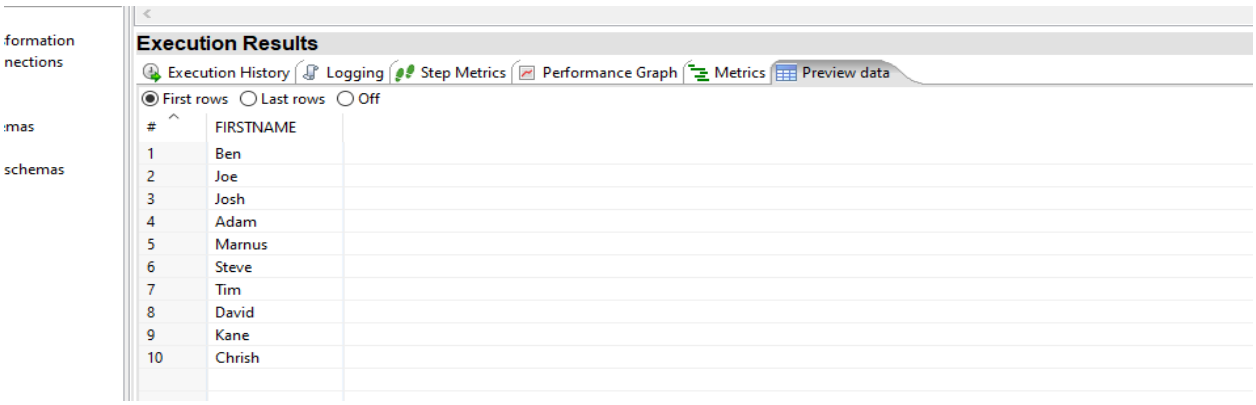
1. Select Values Transformation

It will display the data of given field.

In my case I have given field name as First Name so it will display all the data of



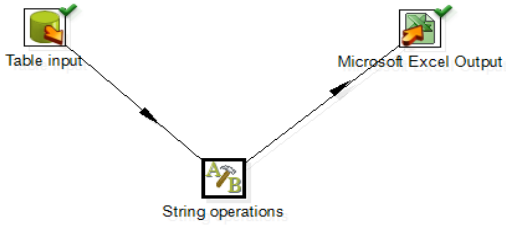


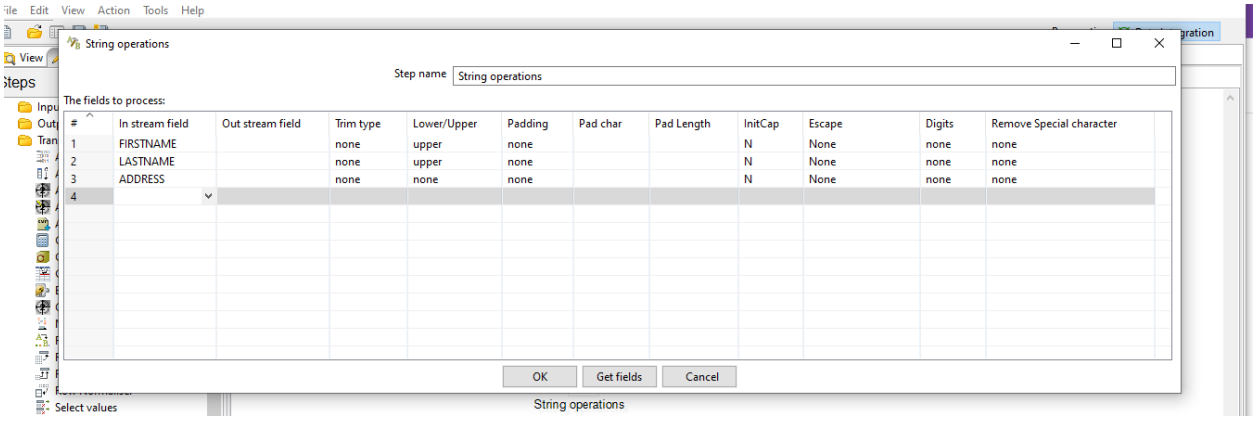


2. String Operation Transformation

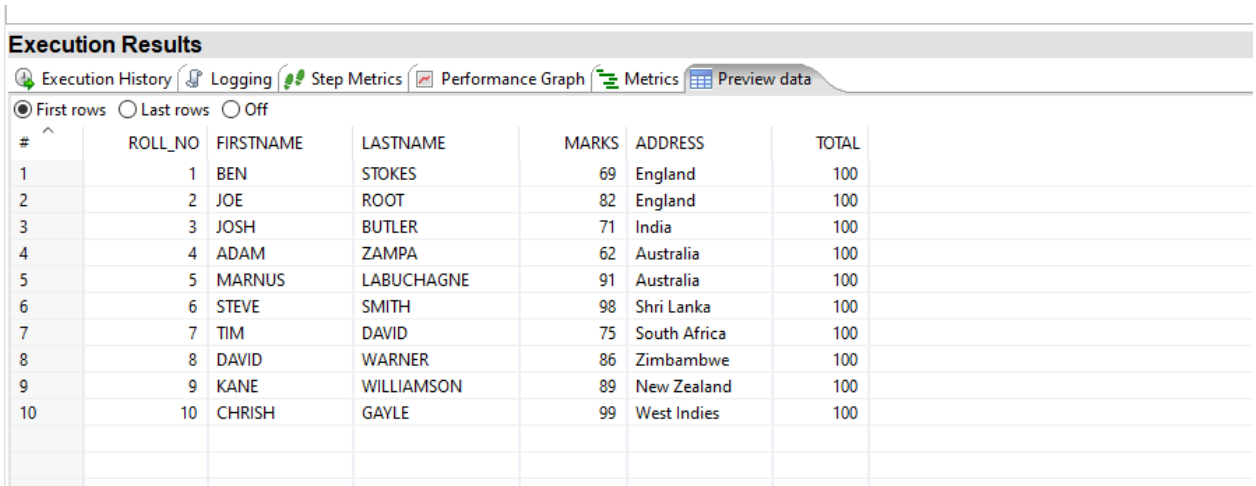
Using this transformation, we can perform different string operations on any field.

In my case I am transforming First Name and Last Name into uppercase.





#	In stream field	Out stream field	Trim type	Lower/Upper	Padding	Pad char	Pad Length	InitCap	Escape	Digits	Remove Special character
1	FIRSTNAME		none	upper	none			N	None	none	none
2	LASTNAME		none	upper	none			N	None	none	none
3	ADDRESS		none	none	none			N	None	none	none
4											

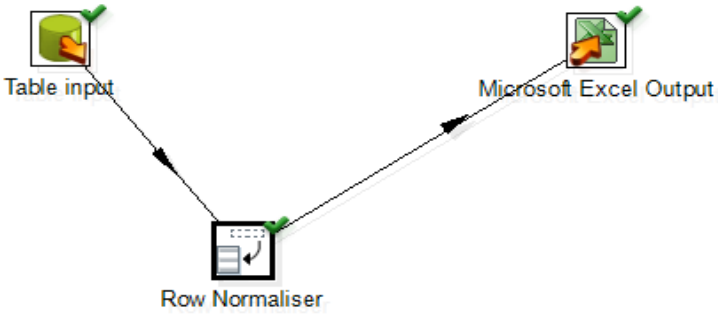


#	ROLL_NO	FIRSTNAME	LASTNAME	MARKS	ADDRESS	TOTAL
1	1	BEN	STOKES	69	England	100
2	2	JOE	ROOT	82	England	100
3	3	JOSH	BUTLER	71	India	100
4	4	ADAM	ZAMPA	62	Australia	100
5	5	MARNUS	LABUCHAGNE	91	Australia	100
6	6	STEVE	SMITH	98	Shri Lanka	100
7	7	TIM	DAVID	75	South Africa	100
8	8	DAVID	WARNER	86	Zimbambwe	100
9	9	KANE	WILLIAMSON	89	New Zealand	100
10	10	CHRISH	GAYLE	99	West Indies	100

3. Row Normaliser Transformation

This will normalise mentioned field.

In my case I am normalising Total Marks.



Execution Results

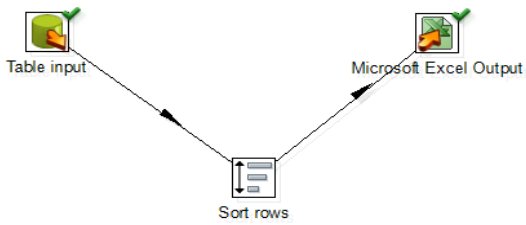
Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

#	ROLL_NO	FIRSTNAME	LASTNAME	ADDRESS	TOTAL	typefield	Normalize_Marks
1	1	Ben	Stokes	England	100	MARKS	69
2	2	Joe	Root	England	100	MARKS	82
3	3	Josh	Butler	India	100	MARKS	71
4	4	Adam	Zampa	Australia	100	MARKS	62
5	5	Marnus	Labuchagne	Australia	100	MARKS	91
6	6	Steve	Smith	Shri Lanka	100	MARKS	98
7	7	Tim	David	South Africa	100	MARKS	75
8	8	David	Warner	Zimbambwe	100	MARKS	86
9	9	Kane	Williamson	New Zealand	100	MARKS	89
10	10	Chrish	Gayle	West Indies	100	MARKS	99

4. Sort Rows Transformation

I am sorting my table using Marks field.



Sort rows

Step name: Sort rows
Sort directory: %%java.io.tmpdir%% Browse...
TMP-file prefix: out
Sort size (rows in memory): 1000000
Free memory threshold (in %):
Compress TMP Files? ☐
Only pass unique rows? (verifies keys only) ☐

Fields :

#	Fieldname	Ascending	Case sensitive compare?	Presorted?
1	MARKS	Y		
2	ROLL_NO	Y		
3	FIRSTNAME	Y		
4	LASTNAME	Y		
5	ADDRESS	Y		
6	TOTAL	Y		

Execution Results

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

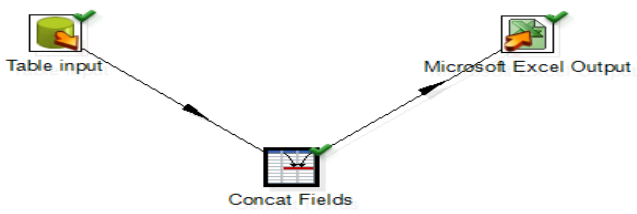
First rows | Last rows | Off

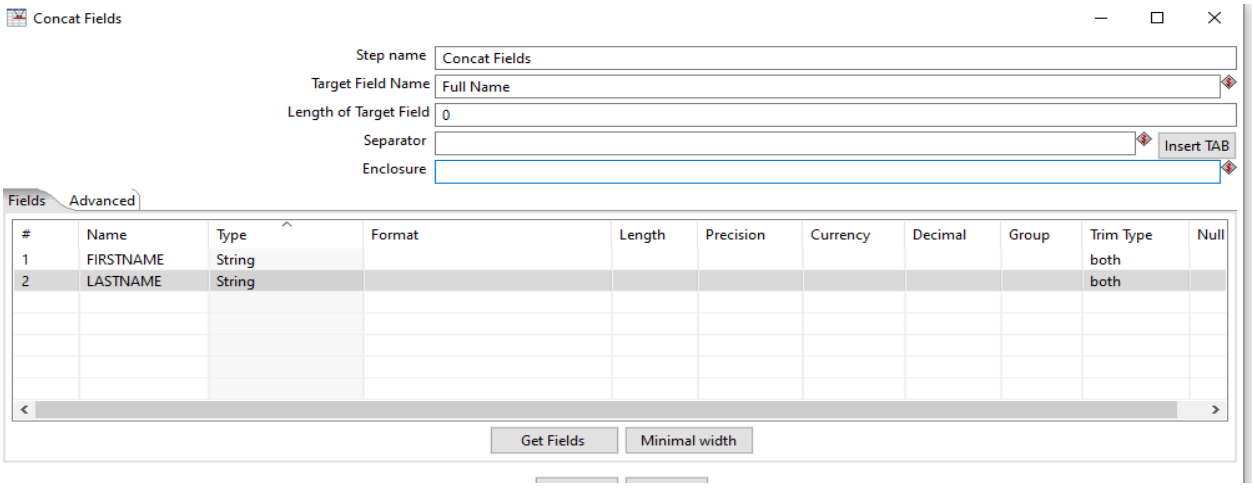
#	ROLL_NO	FIRSTNAME	LASTNAME	MARKS	ADDRESS	TOTAL
1	4	Adam	Zampa	62	Australia	100
2	1	Ben	Stokes	69	England	100
3	3	Josh	Butler	71	India	100
4	7	Tim	David	75	South Africa	100
5	2	Joe	Root	82	England	100
6	8	David	Warner	86	Zimbambwe	100
7	9	Kane	Williamson	89	New Zealand	100
8	5	Marnus	Labuchagne	91	Australia	100
9	6	Steve	Smith	98	Shri Lanka	100
10	10	Chrish	Gayle	99	West Indies	100

5. Concat Fields Transformation

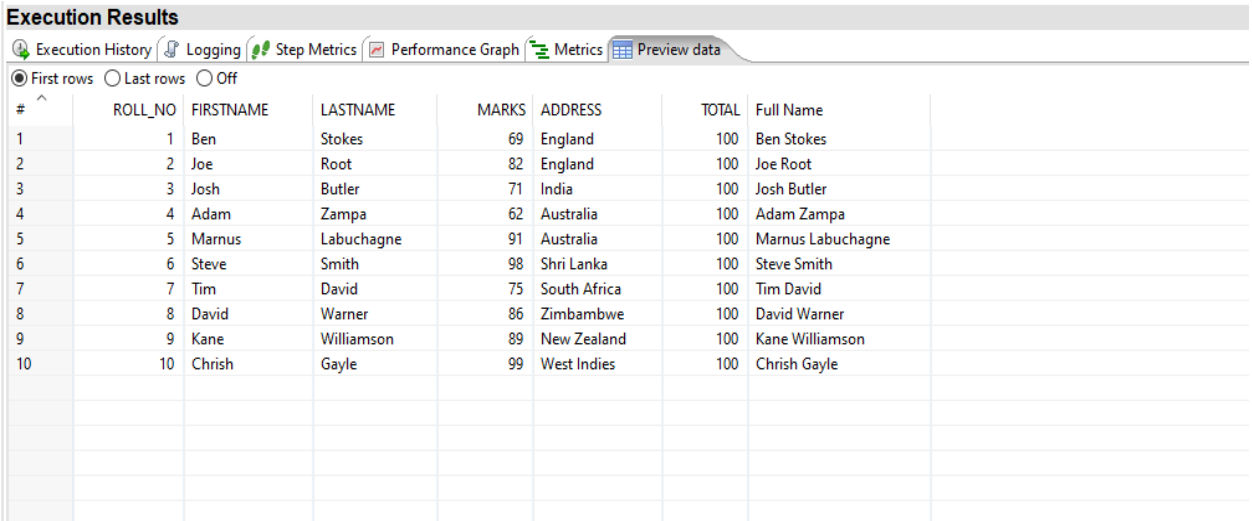
This transformation is used to concat two different field and store data into a new field.

Here I am concating First Name and Last Name, storing into Full Name.





#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim Type	Null
1	FIRSTNAME	String							both	
2	LASTNAME	String							both	



#	ROLL_NO	FIRSTNAME	LASTNAME	MARKS	ADDRESS	TOTAL	Full Name
1	1	Ben	Stokes	69	England	100	Ben Stokes
2	2	Joe	Root	82	England	100	Joe Root
3	3	Josh	Butler	71	India	100	Josh Butler
4	4	Adam	Zampa	62	Australia	100	Adam Zampa
5	5	Marnus	Labuchagne	91	Australia	100	Marnus Labuchagne
6	6	Steve	Smith	98	Shri Lanka	100	Steve Smith
7	7	Tim	David	75	South Africa	100	Tim David
8	8	David	Warner	86	Zimbambwe	100	David Warner
9	9	Kane	Williamson	89	New Zealand	100	Kane Williamson
10	10	Chrish	Gayle	99	West Indies	100	Chrish Gayle

Conclusion:

I have successfully understood Pentaho and performed ETL transformation.