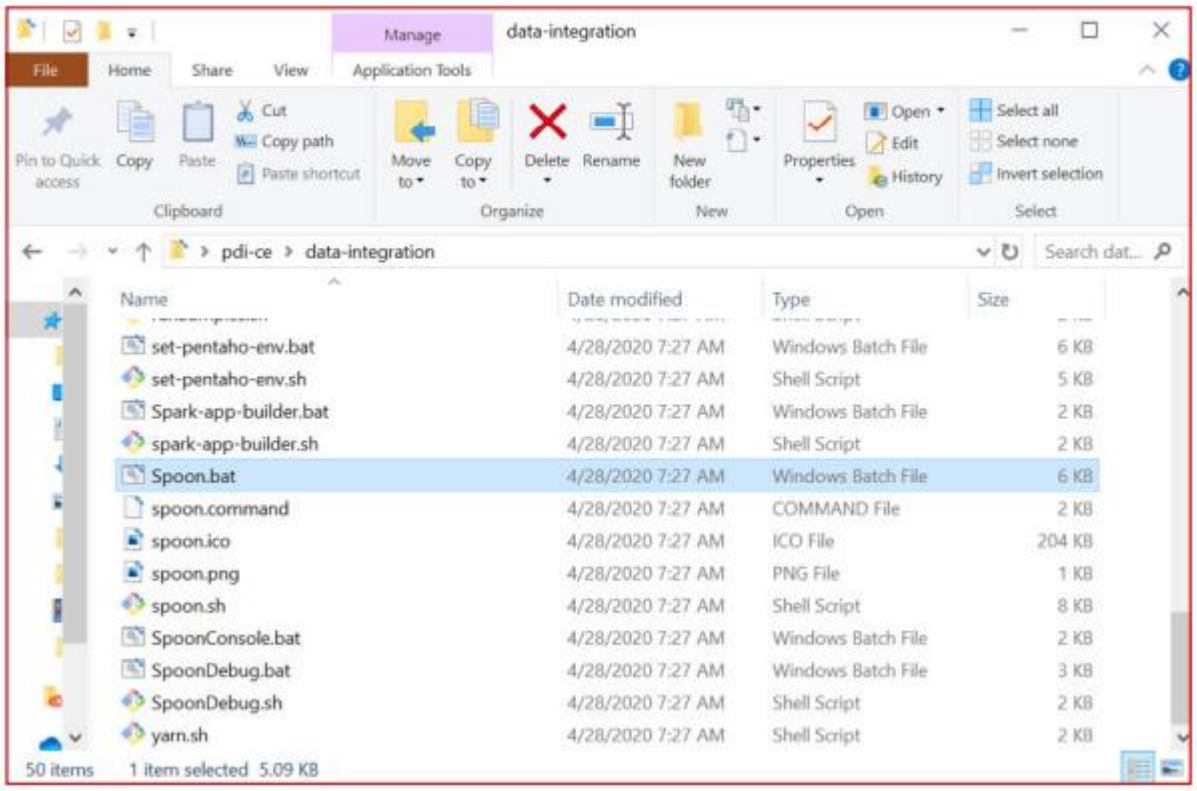
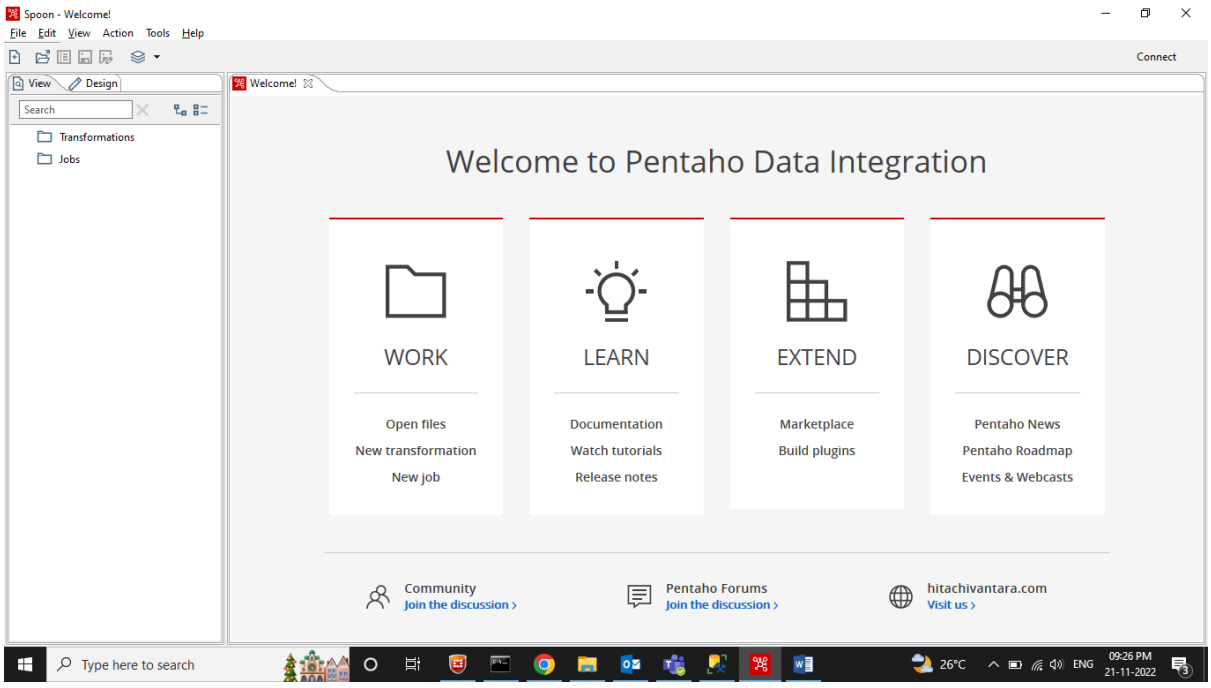
Pentaho ETL Tool - User Guide :

Open the Pentaho by double clicking the “**Spoon.bat”** file :



The app will looks like below :



In Pentaho, for doing a comparison between ”**Source table data**” to “**Target table data**”, we have to do the following steps :

1. Create a connection with the database (Connect Source DB and Target DB)
2. Fetch data from the DB table (Fetch data from both Source and Target tables)
3. Compare the Source Table to Target Table
4. Verify the generated Report (To find the mismatches)

1.Create a connection with the DB :

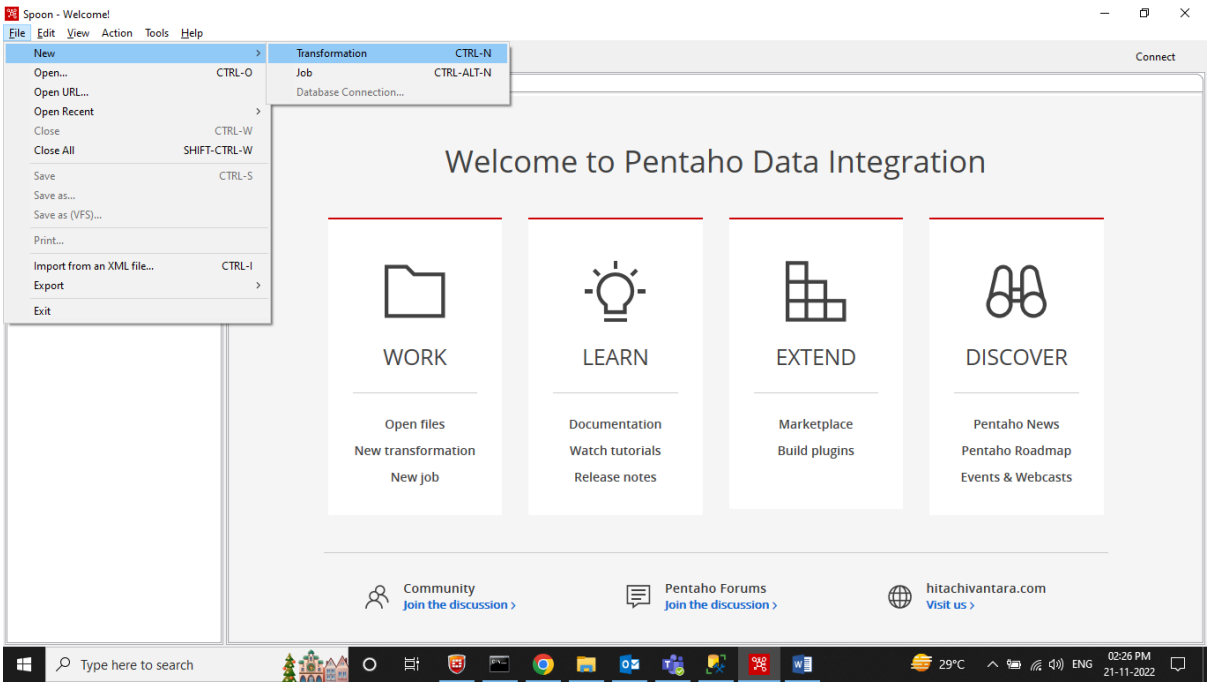
If you want to work with a DB, either read, write, view data, etc, in Pentaho, the first thing you will have to do is to create a connection with that database.

In order to set up the connection, we need the following:

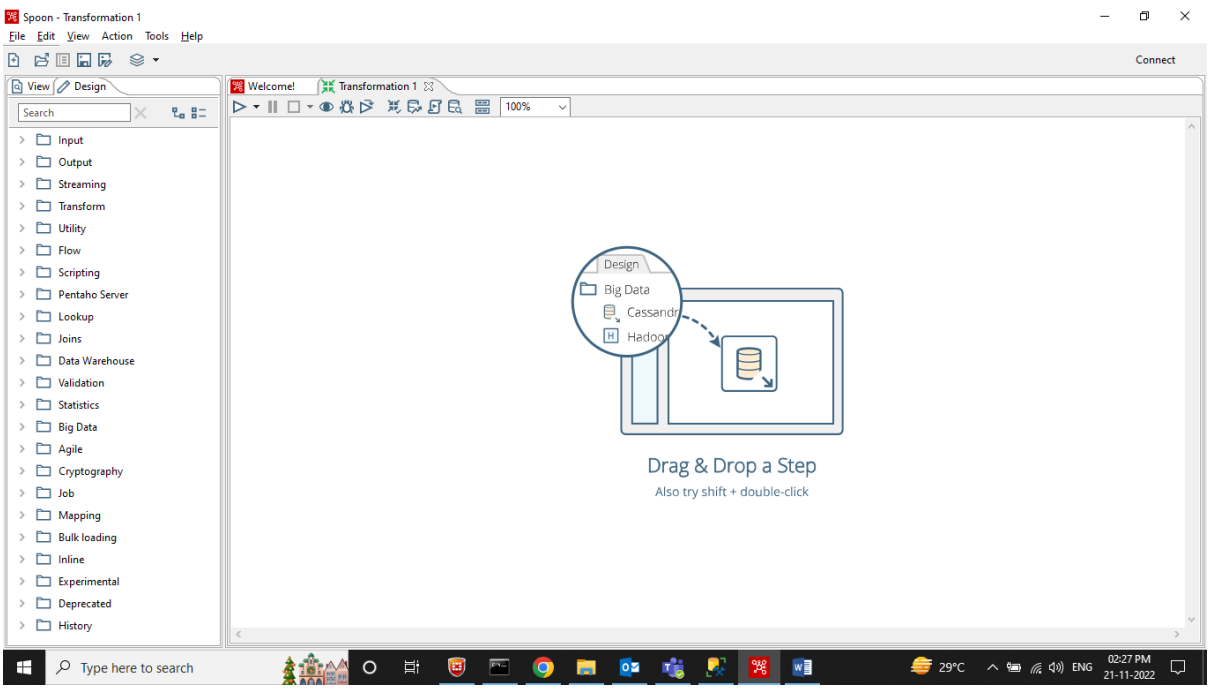
- **Host name:** Domain name or IP address of the database server.
- **Database name:** The schema or other database identifier.
- **Port number:** The port the database connects to. Each database has its own default port.
- **Username:** The username to access the database.
- **Password:** The password to access the database.

PFB the Steps to Connect to the DB:

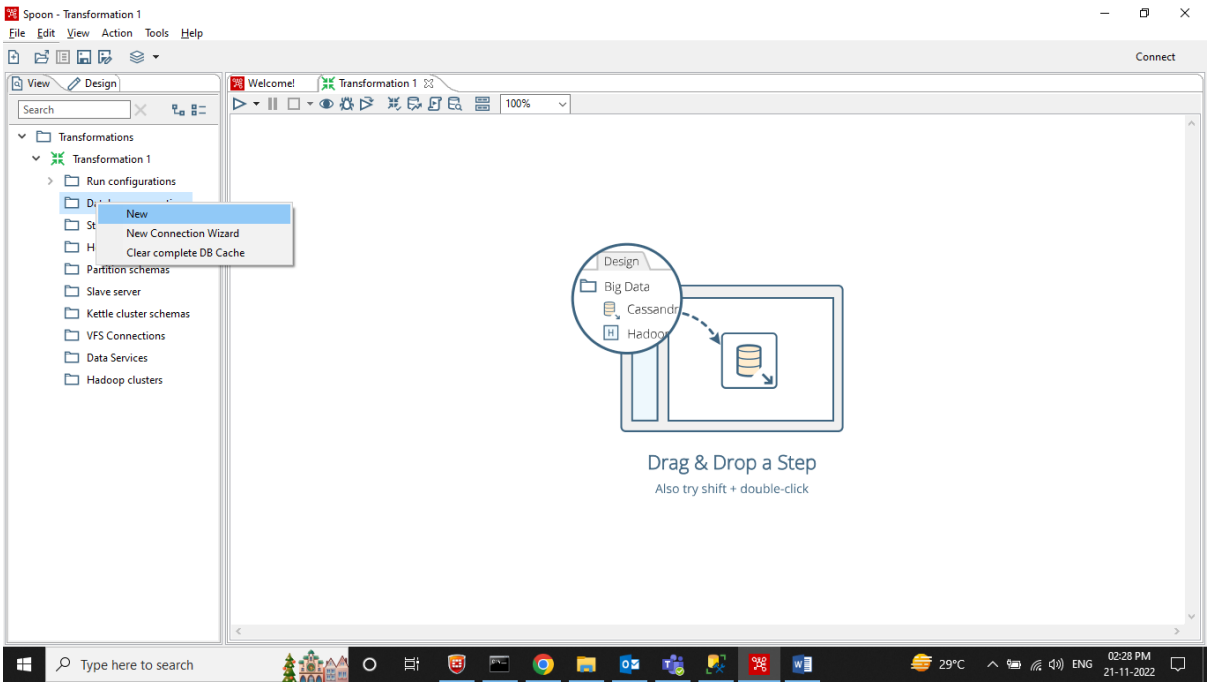
Step 1 : Open Spoon and create a new Transformation as below by clicking on the “**File**” icon:



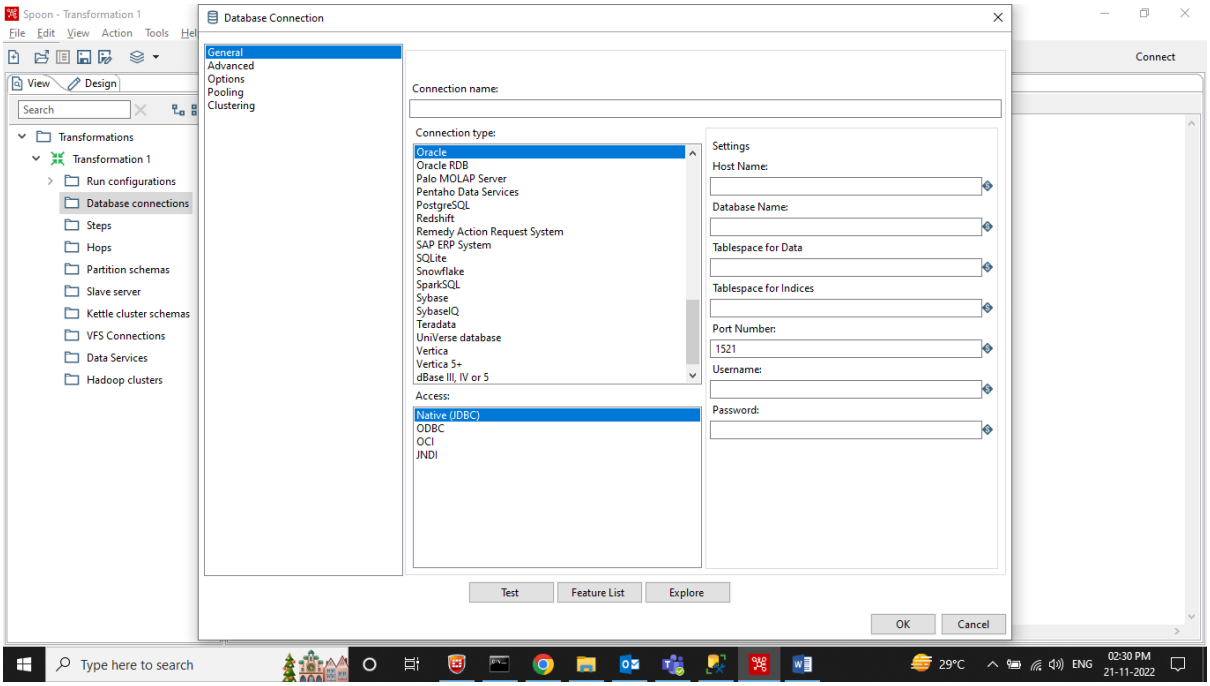
It will create a new Transformation and looks like below:



Step 2: Select the "**View**" option that appears in the upper-left corner of the screen, right-click on the "**Database connections**" option, and select "**New**".



The "**Database Connection"** dialog window will appears as below :

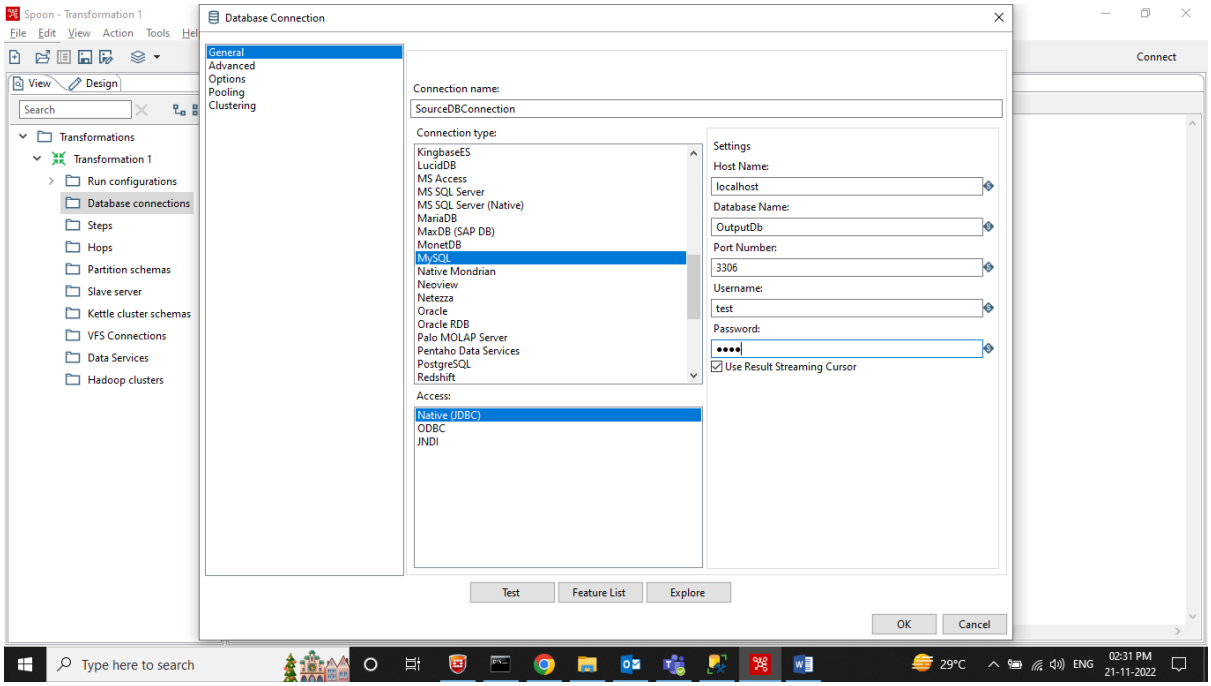


Step 3 :Provide Connection Name by typing it in the "**Connection Name:"** textbox.

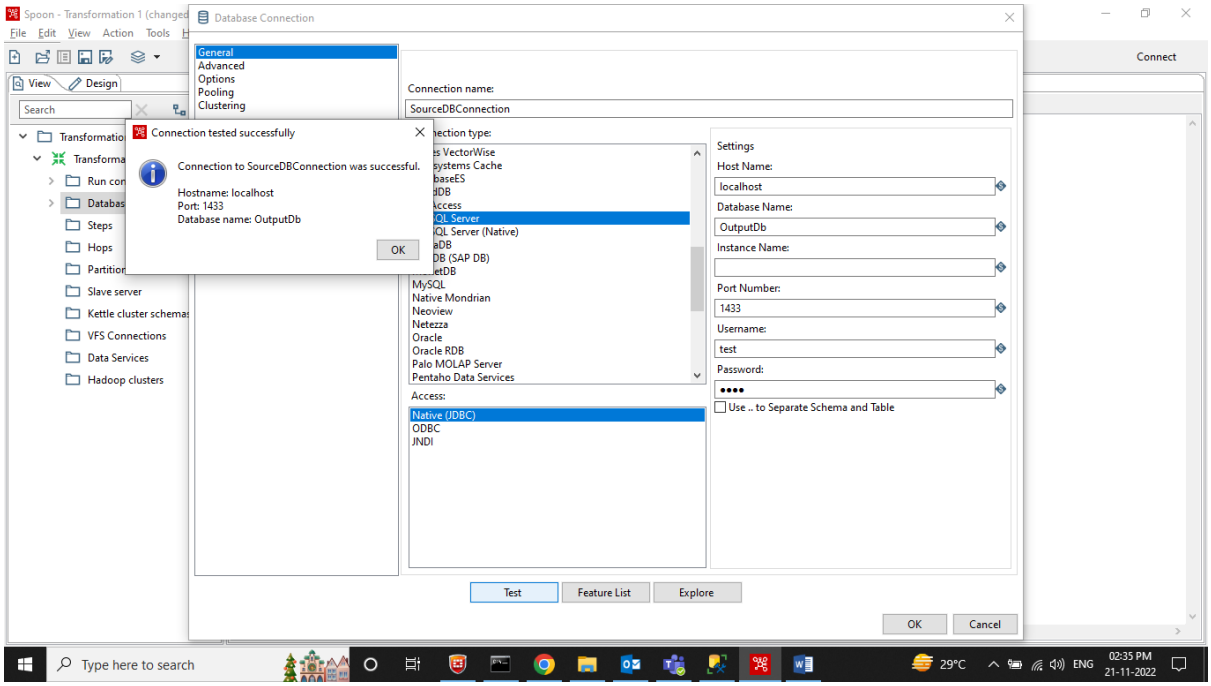
Under "**Connection Type"**, select the database engine that matches your DBMS.

Fill all the required details in the "**Settings"** options

Your Database Connection window should looks like below:

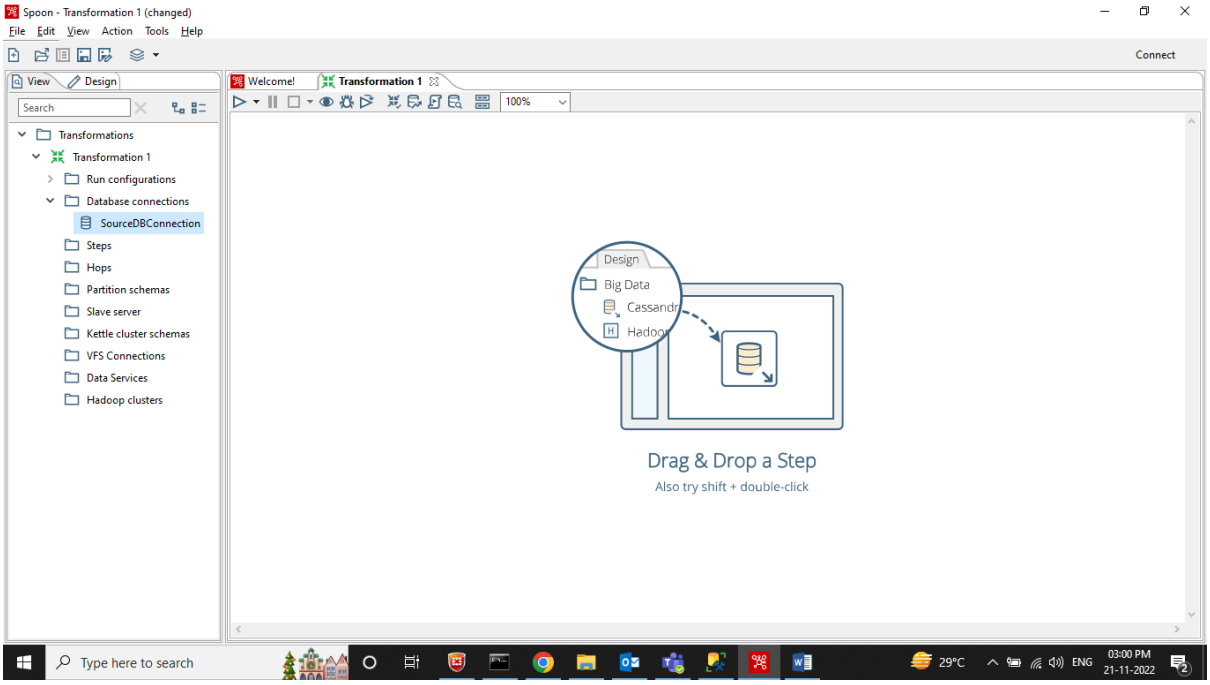


Step 4 : Press the "**Test"** button. A message should appear as below, informing you that the connection to your database is OK.



Step 5 : Now Tap on the "**OK**" button, the window will close and your DB Connection is ready.

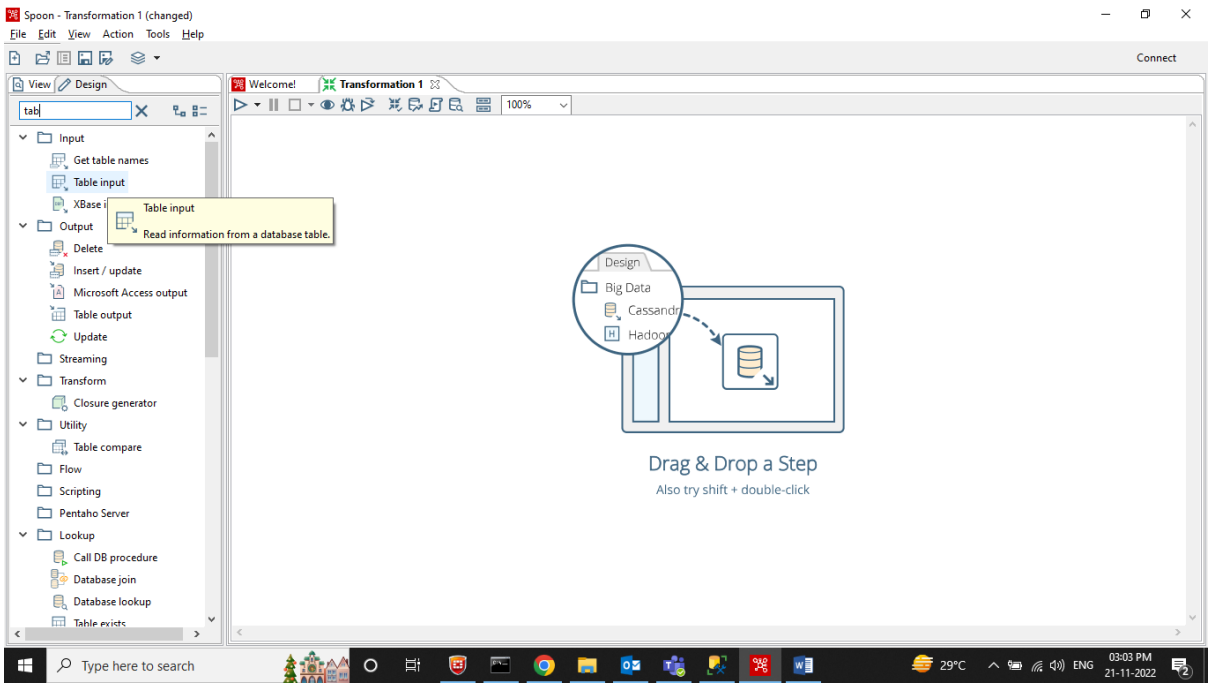
You will be able to see the created connection under **View->Database Connection** as below :



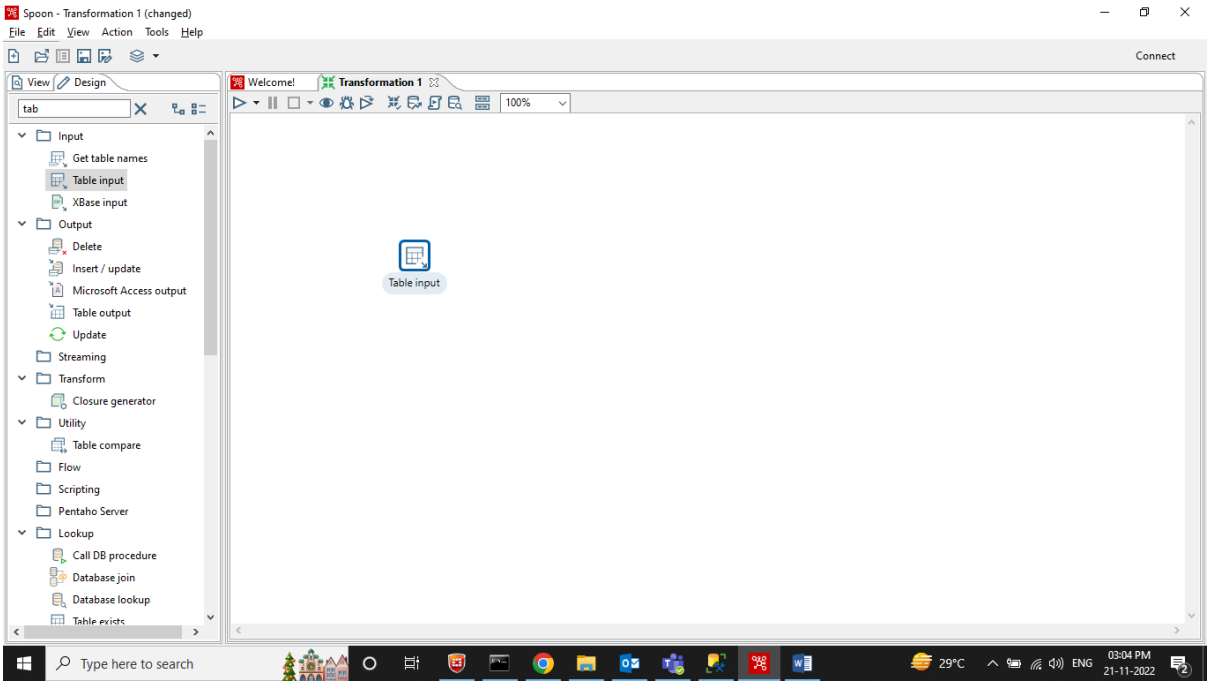
2.Fetch Data from the Database Table :

PFB the Steps for fetching data from the DB :

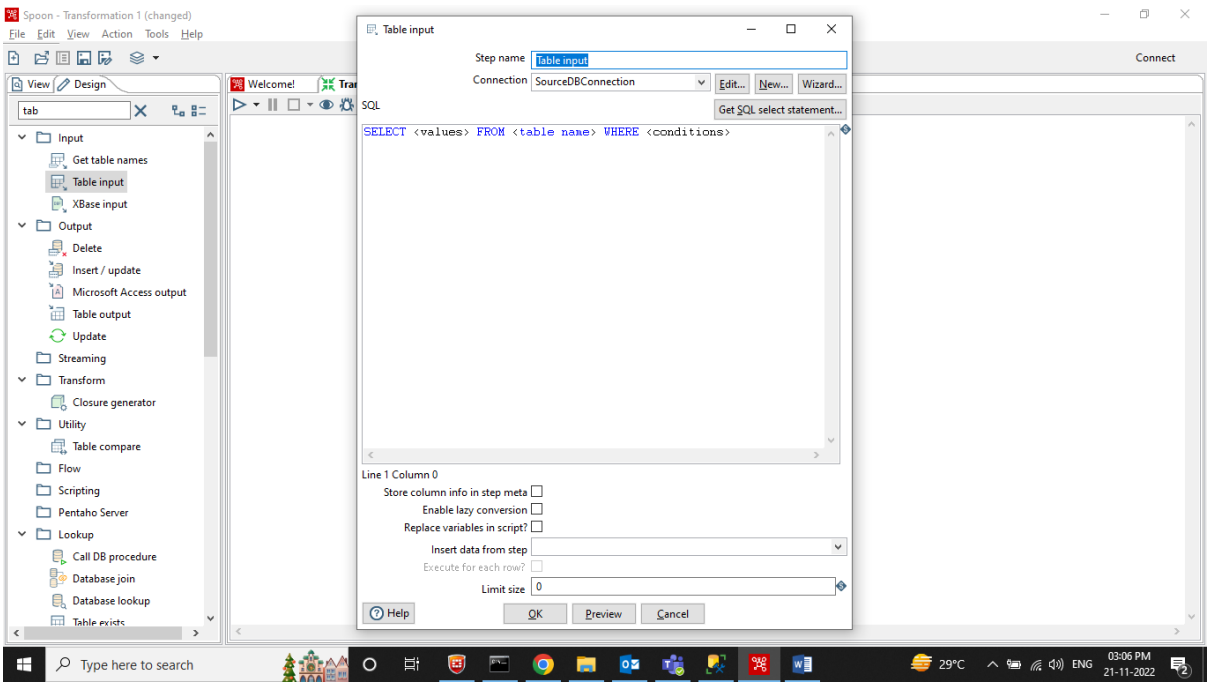
Step 1 : Go to the "**Design**" tab and type as "**Table**", you will be able to see the "**Table Input**" option as below :



Step 2 : Drag and drop the "**Table input**" and it will looks like below :



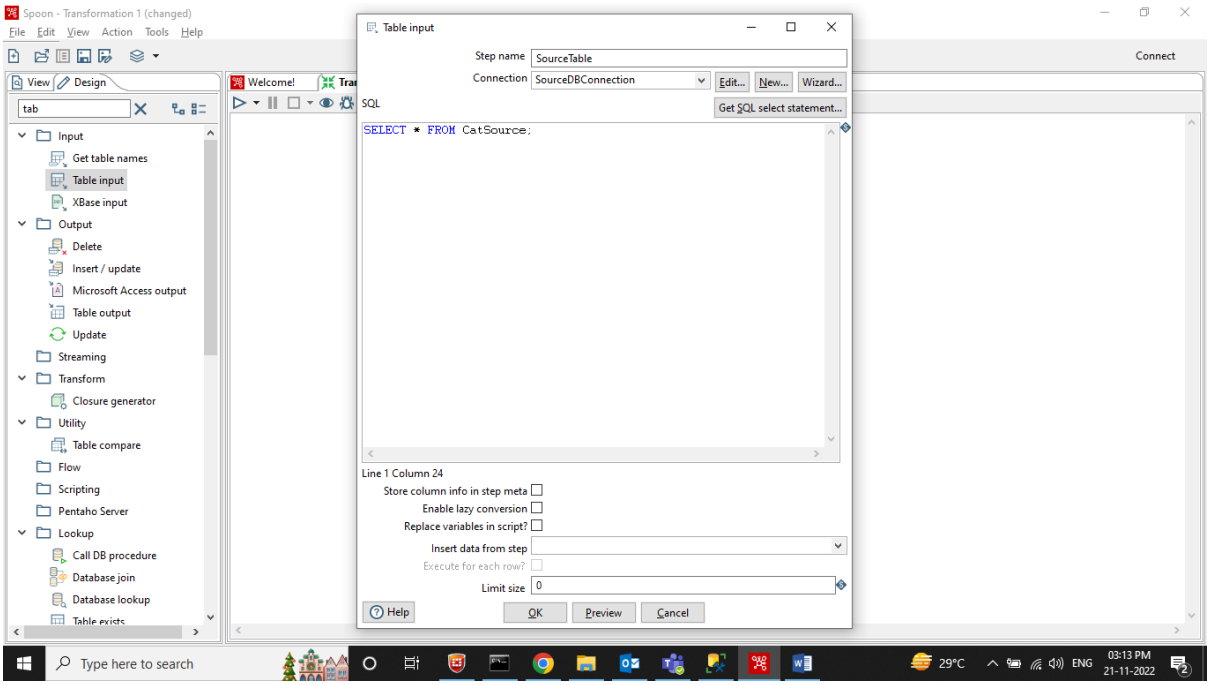
Step 3 : Double click on the "**Table input**" and you will be able to see the below window :



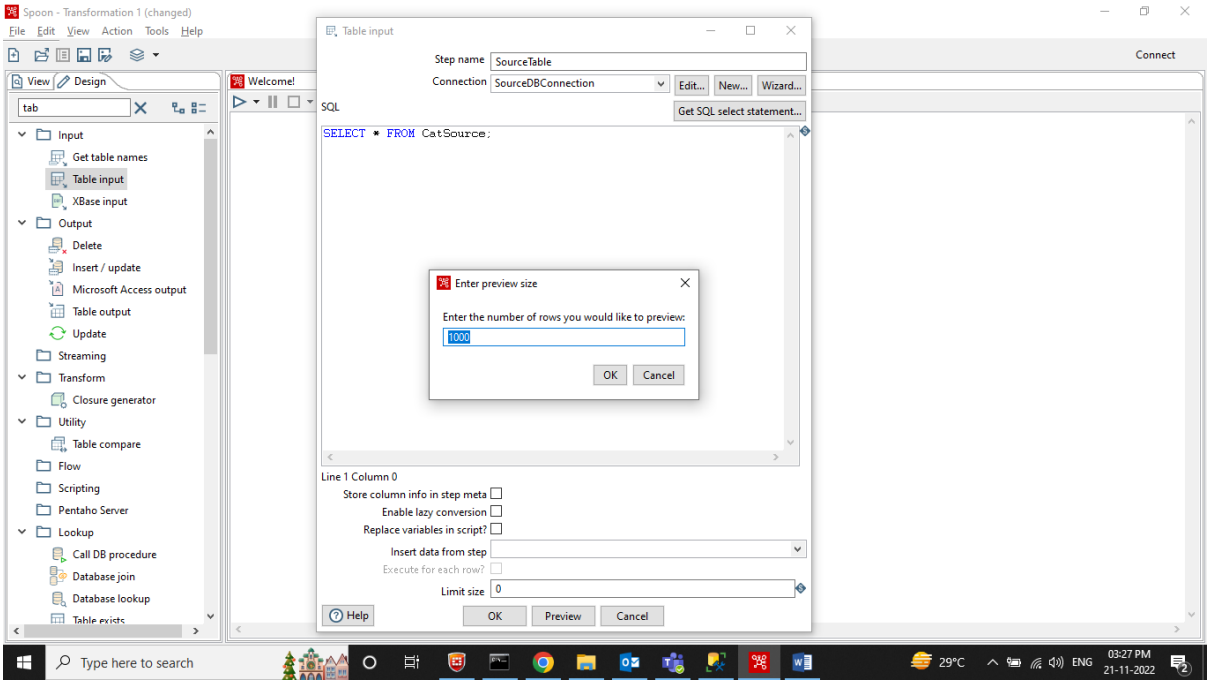
Step 4 : Enter the below details :

- **Step name** : If you want to change the name, you can change it. Or keep it as the same(optional)
- **Connection** : By default, you will be able to see the newly created connection there. If we have more than 1 connection, then we have to select the connection from the drop down list based on the DB from where we are fetching the table data
- **SQL** : Now you have to write the **SQL Query** to fetch data from the DB Table.

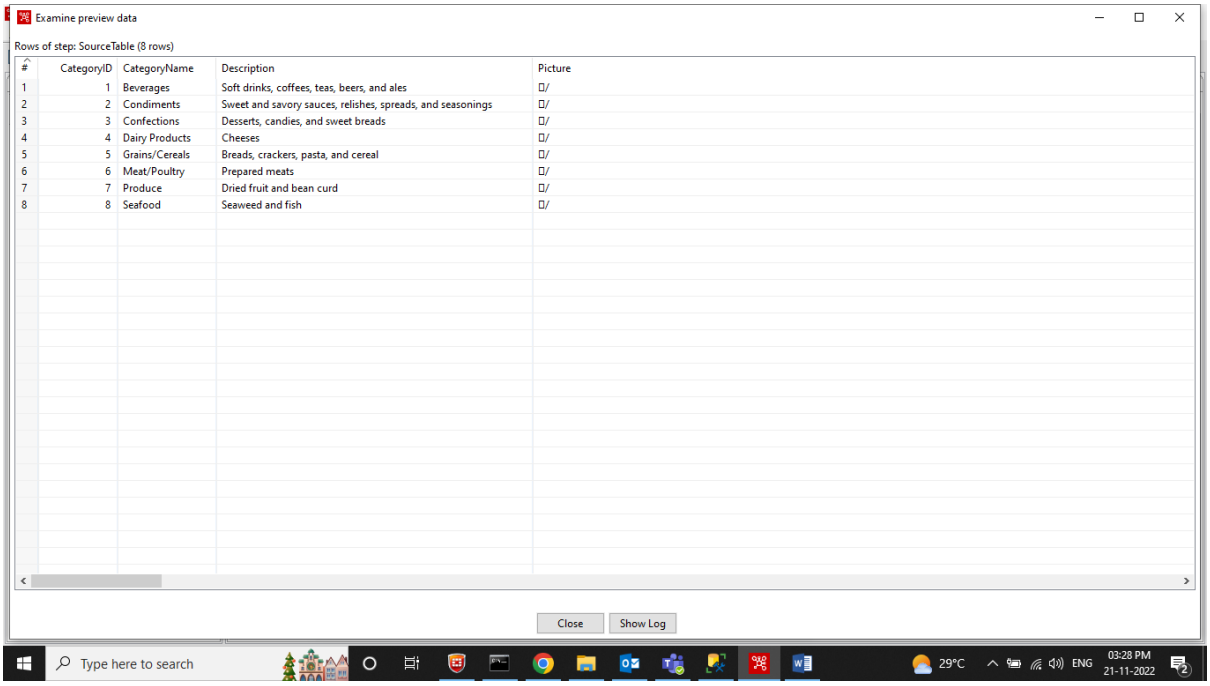
Now the window will looks like below :



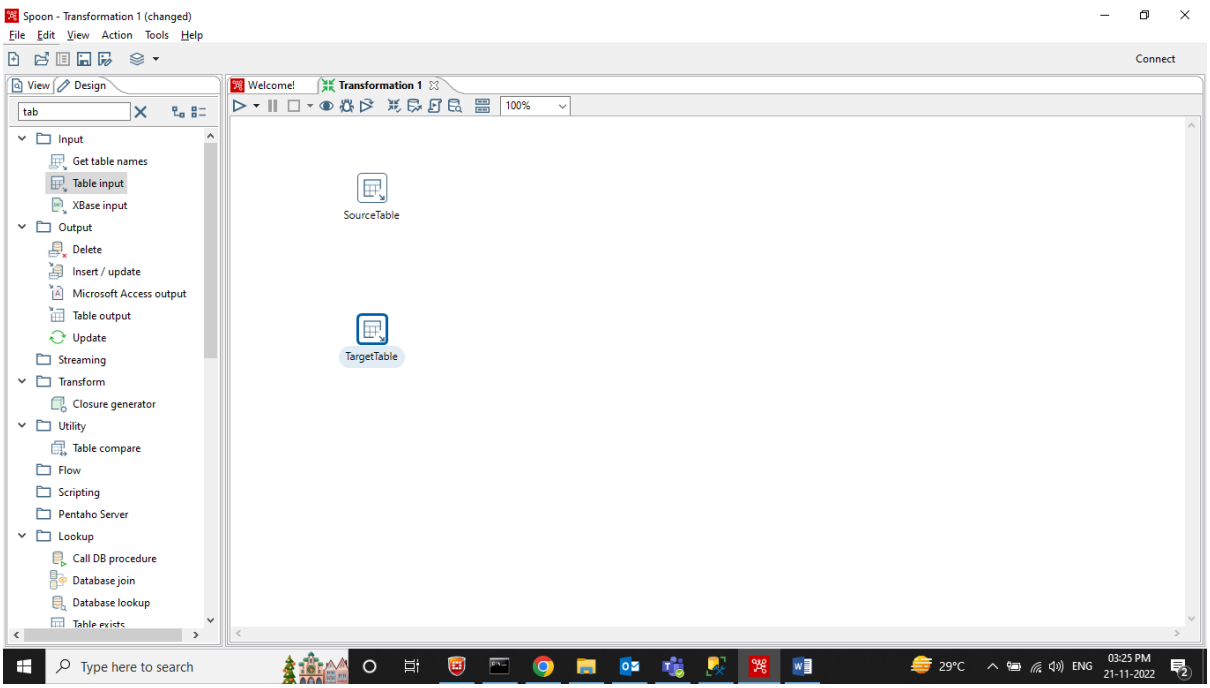
Step 5 : Now click on the "**Preview**" Button.



Step 6 : Click on the "**OK**" button and you will be able to see the fetched data from the table as below :



Here, to do the comparison, we need to fetch data from the **Target table** also. So, create a "**Table input**" for that by using the above steps. And then it will look like below:



3.Compare the Source table to Target table :

In Pentaho, for comparing 2 tables, we are using the "**Merge rows(diff)**".

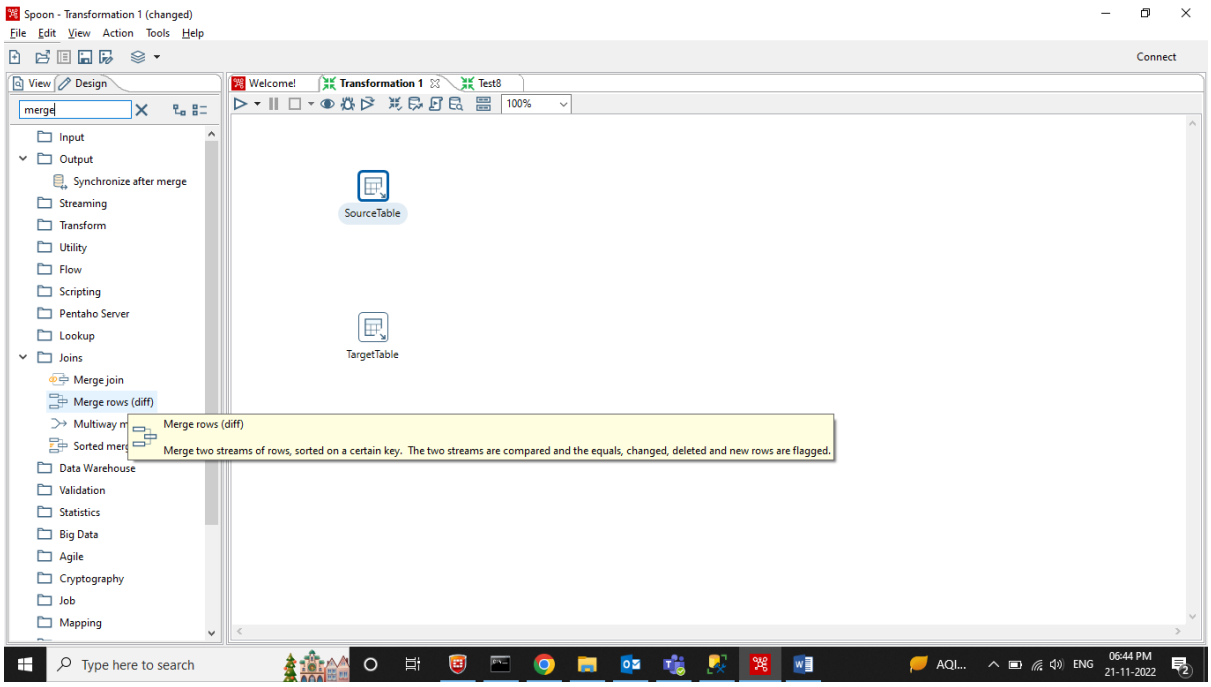
The **Merge rows (diff)** step compares and merges data within two rows of data. This step is useful for comparing data collected at two different times.

Based on keys for comparison, this step merges **reference rows (previous data)** with **compare rows (new data)** and creates merged output rows. A **flag** in the row indicates how the values were compared and merged. Flag values include:

- **Identical** : The key was found in both rows, and the compared values are identical.
- **Changed** : The key was found in both rows, but one or more compared values are different.
- **New** : The key was not found in the reference rows.
- **Deleted** : The key was not found in the compare rows.

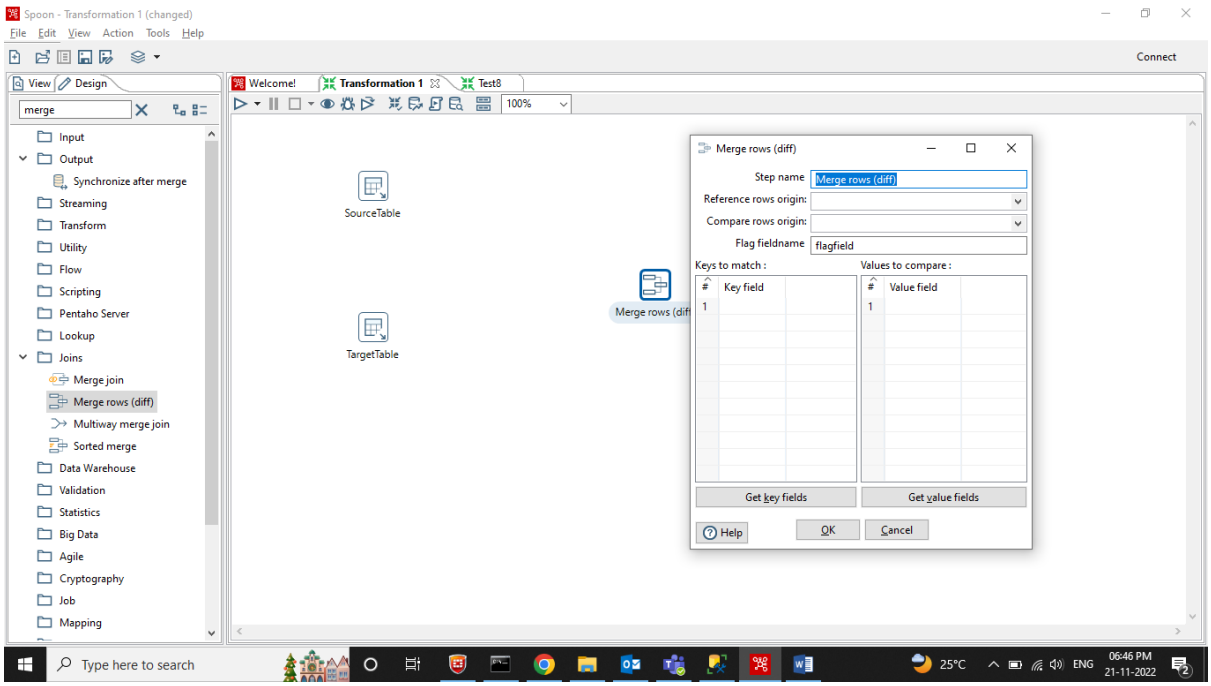
PFB the Steps to compare the Tables :

Step 1 : In " **Design**" tab, search like "**merge**", and you will be able to see the **"Merge rows(diff)**" under "**Joins**" as below :



Step 2 : Drag and drop the "**Merge rows(diff)**"

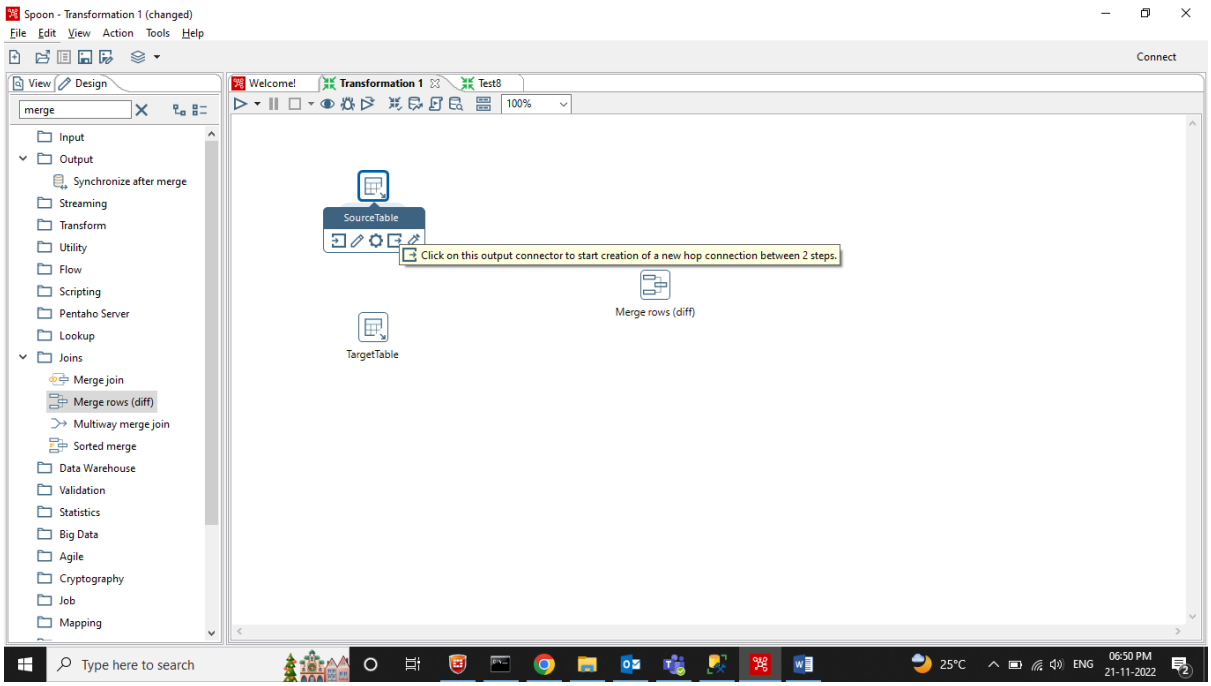
Step 3 : Double click on the "**Merge rows(diff)**", and you will get the below window :



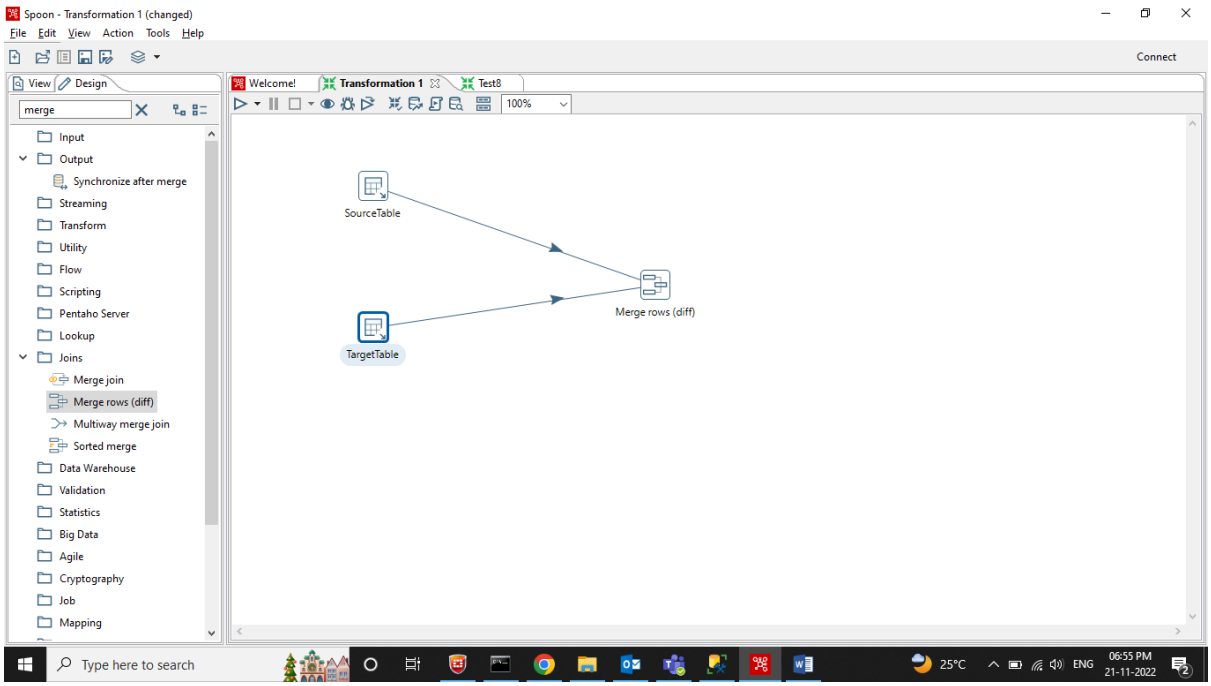
Step 4 : Enter the below details :

- **Step name** : Give a name to the Step(optional)
- **Reference rows origin** : To select this, we need to make connection from the Source Table
- **Compare rows origin** : To select this, we need to make connection from the Target Table

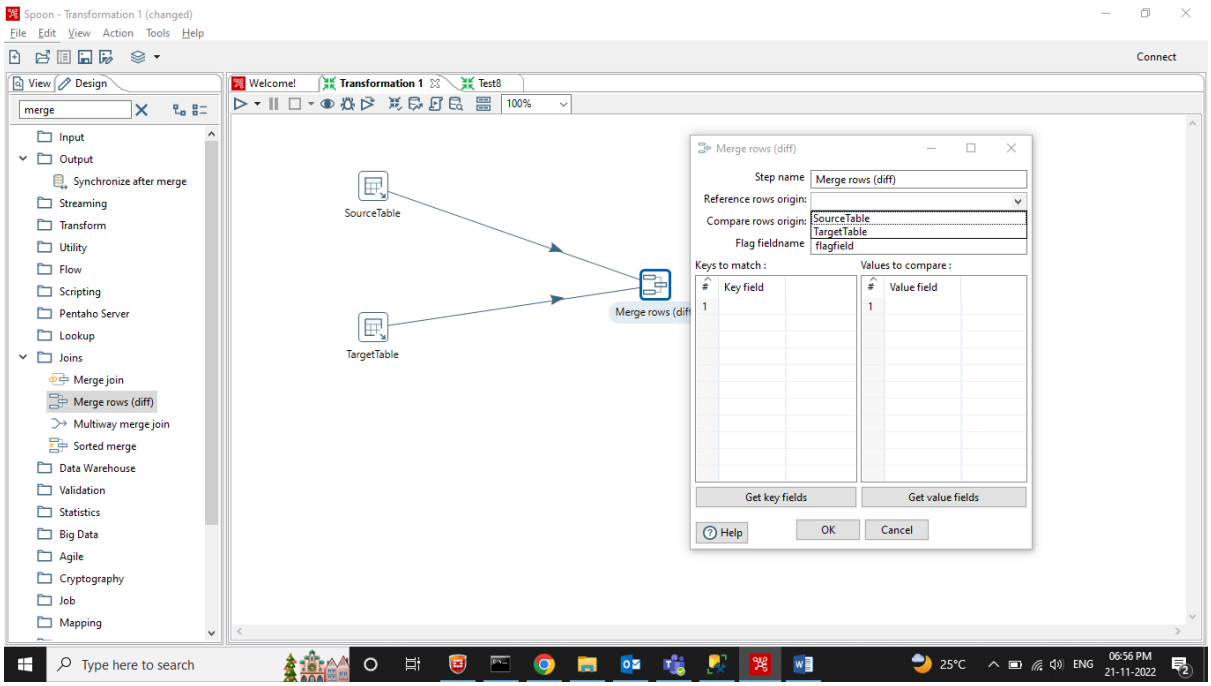
  For making the connection, tap on the "**Table input**" created for the **Source and Target**. You will be able to see as below :



Make connection from both Source and Target "**Table input**" to the "**Merge rows(diff)"** as below :



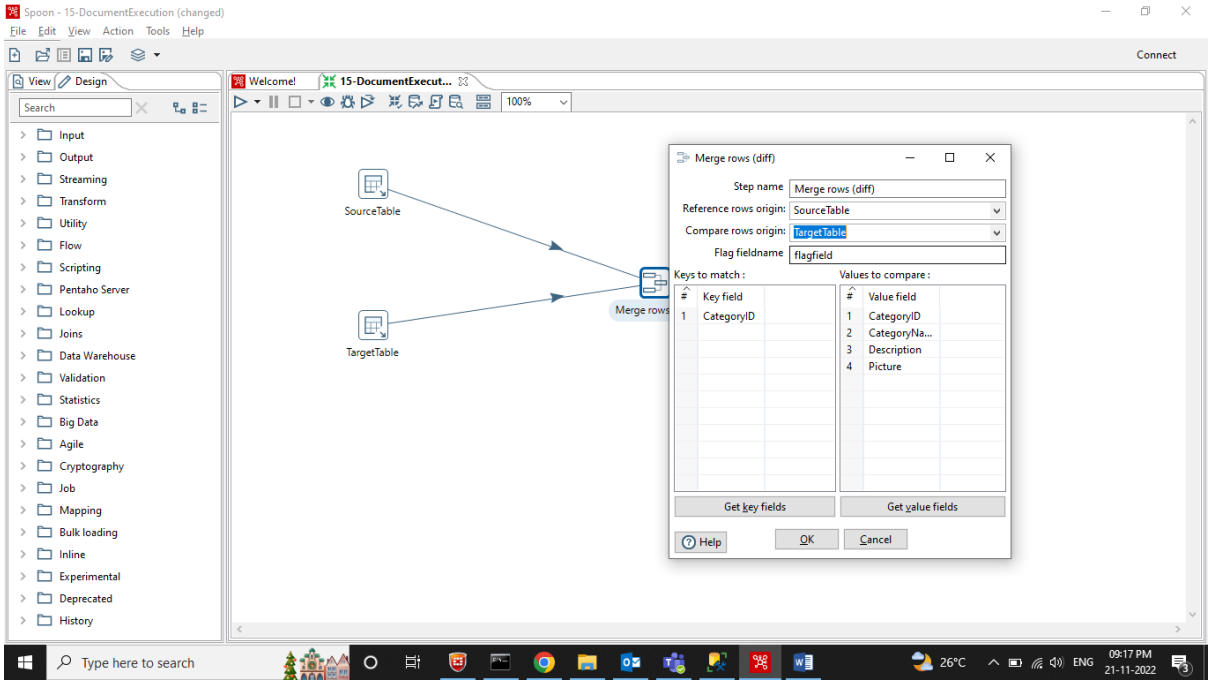
Now, when double click on the "**Merge rows(diff)",** you will be able to see as below in the drop down list :



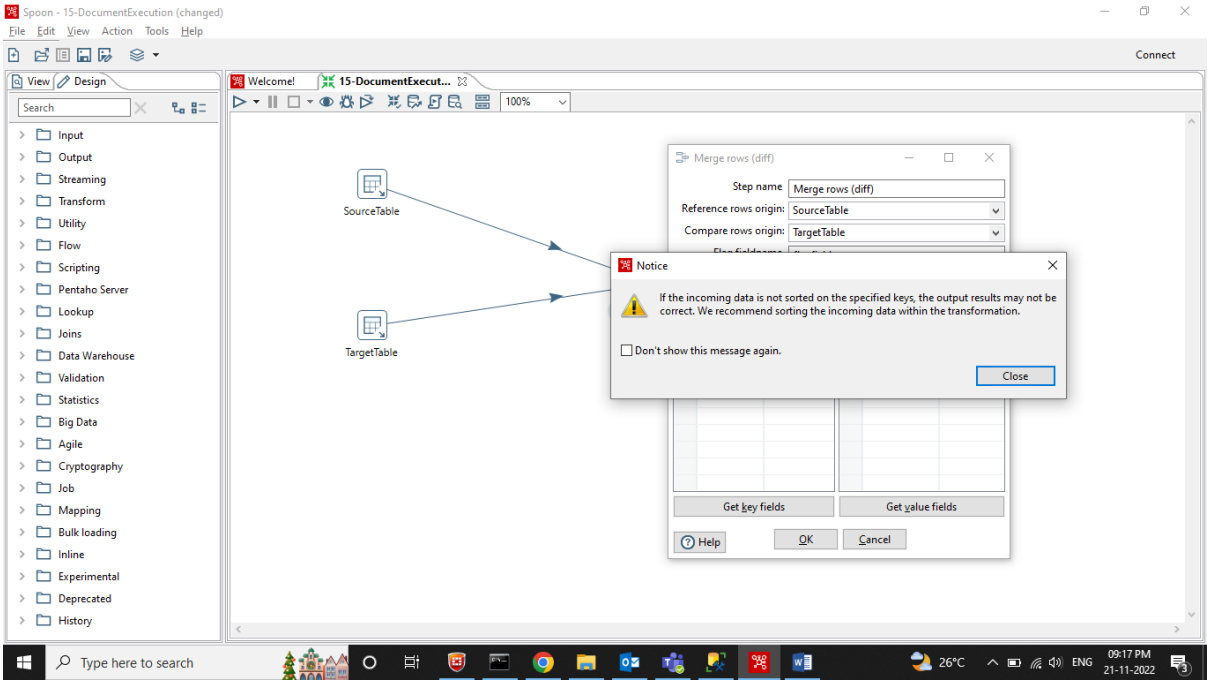
So, select the "**Source table input**" as "**Reference rows origin**" and select the "**Target table input**" as "**Comparison rows origin**".

Now Tap on the "**Get key fields**" and select the key to match(The key value which column we need to match on).

Tap on the "**Get value fields**" and select the value to compare(Select all the column names there since we need to compare all the data in all the columns).

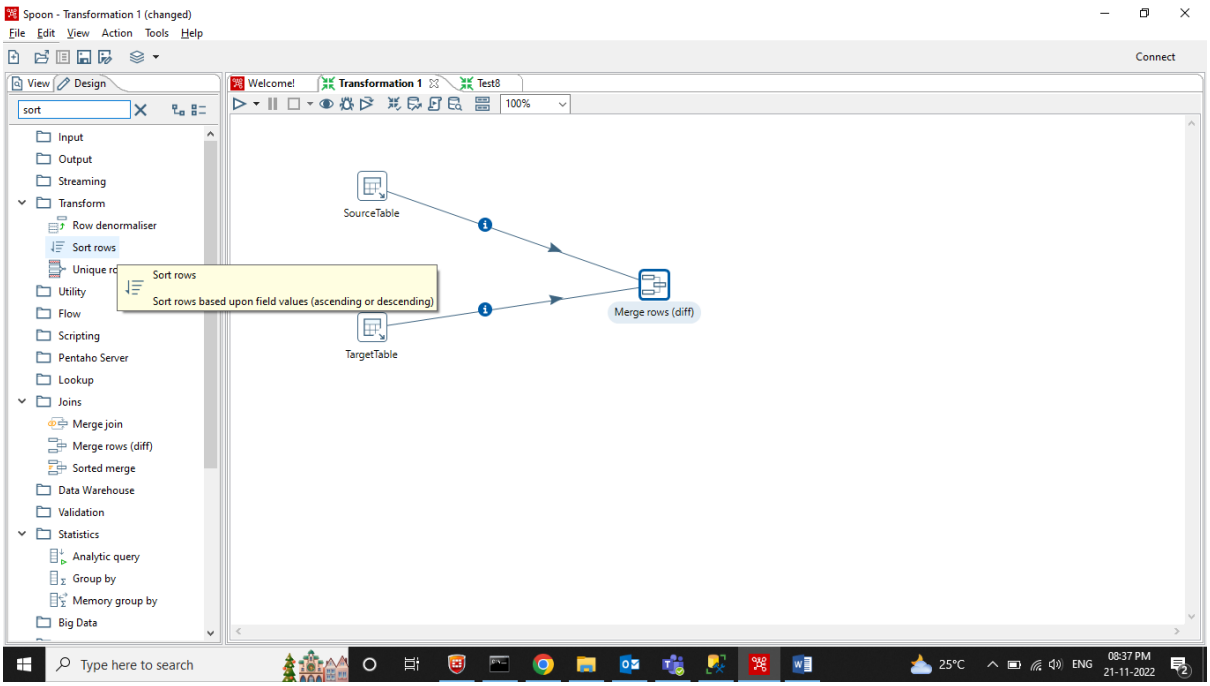


Now when click on "**OK**", we will get the below message :



So we need to Sort the data inside the tables before using the **Merge rows(diff)**

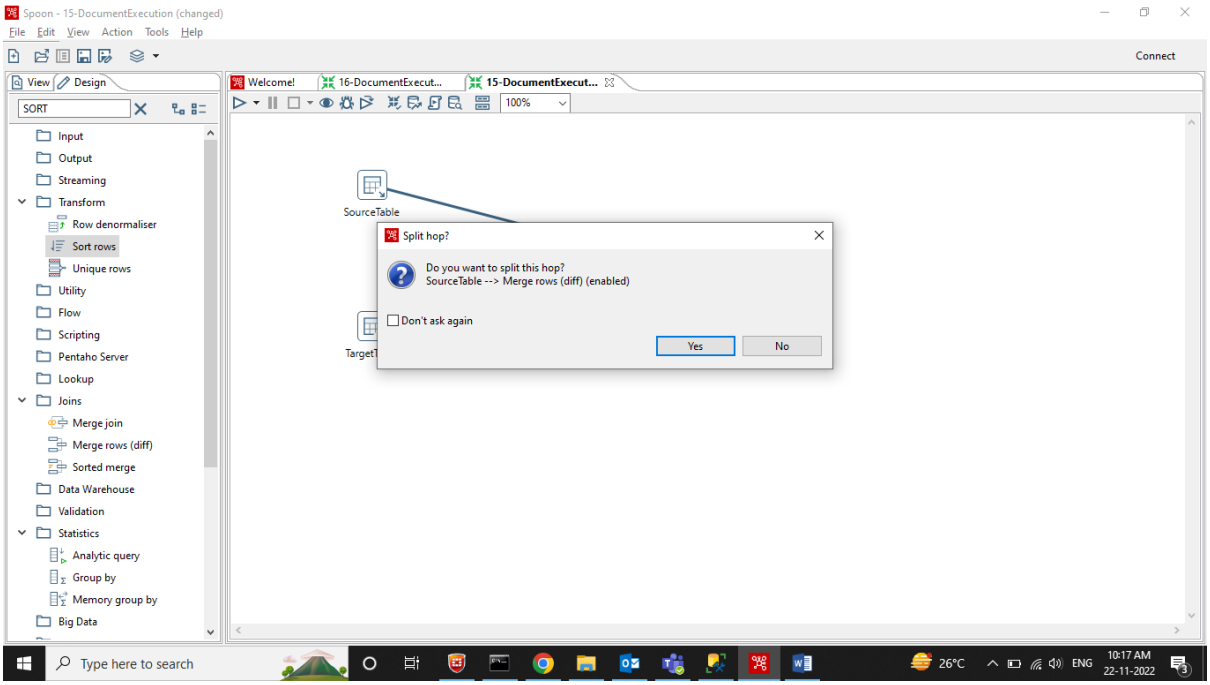
Step 5 : In "**Design**" view, search like "**Sort**", you will be able to see as below :



Step 6 : Drag and drop the "**Sort rows**" in between the

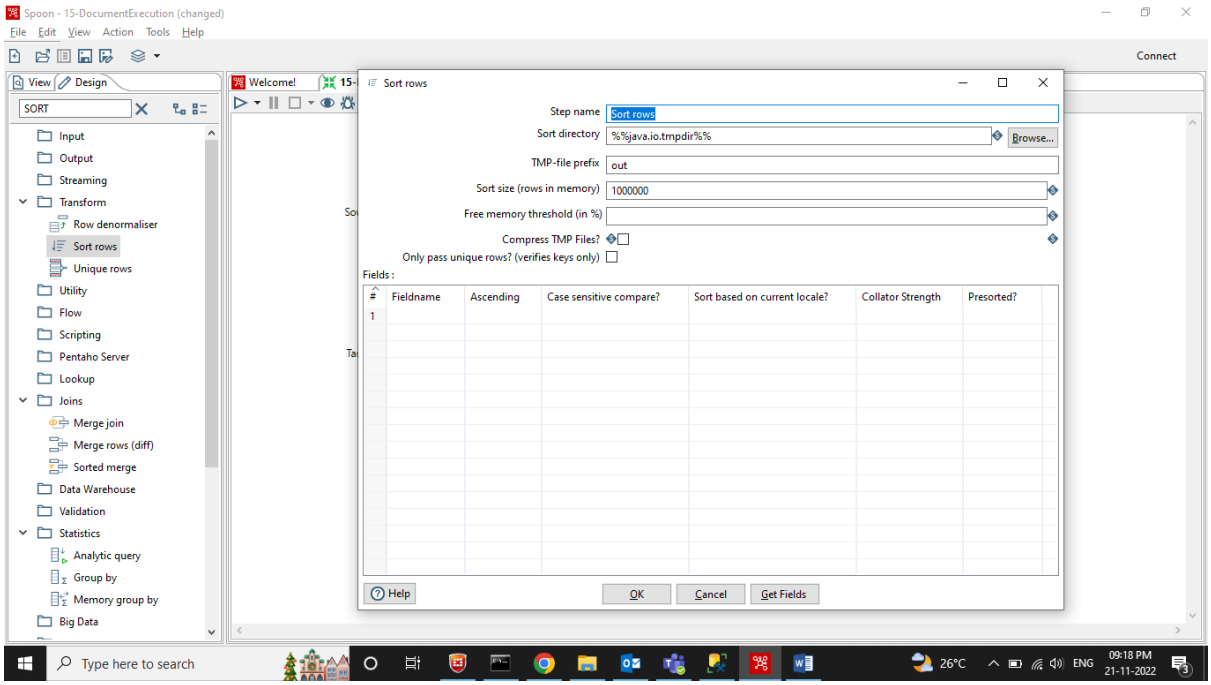
- "**Source table input**" and "**Merge rows(diff)"**
- "**Target table input**" and "**Merge rows(diff)"**

And we will be able to see a message like below :



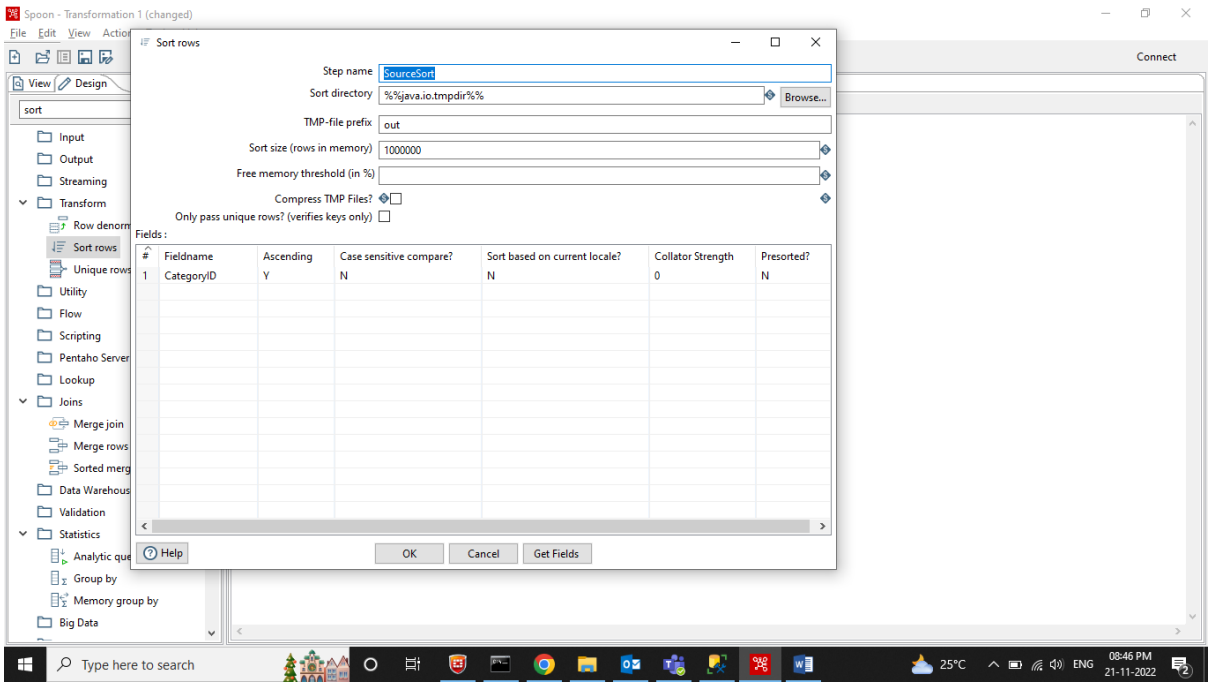
Click on "**Yes**"

Step 7 : Now double click on the "**Sort rows**"



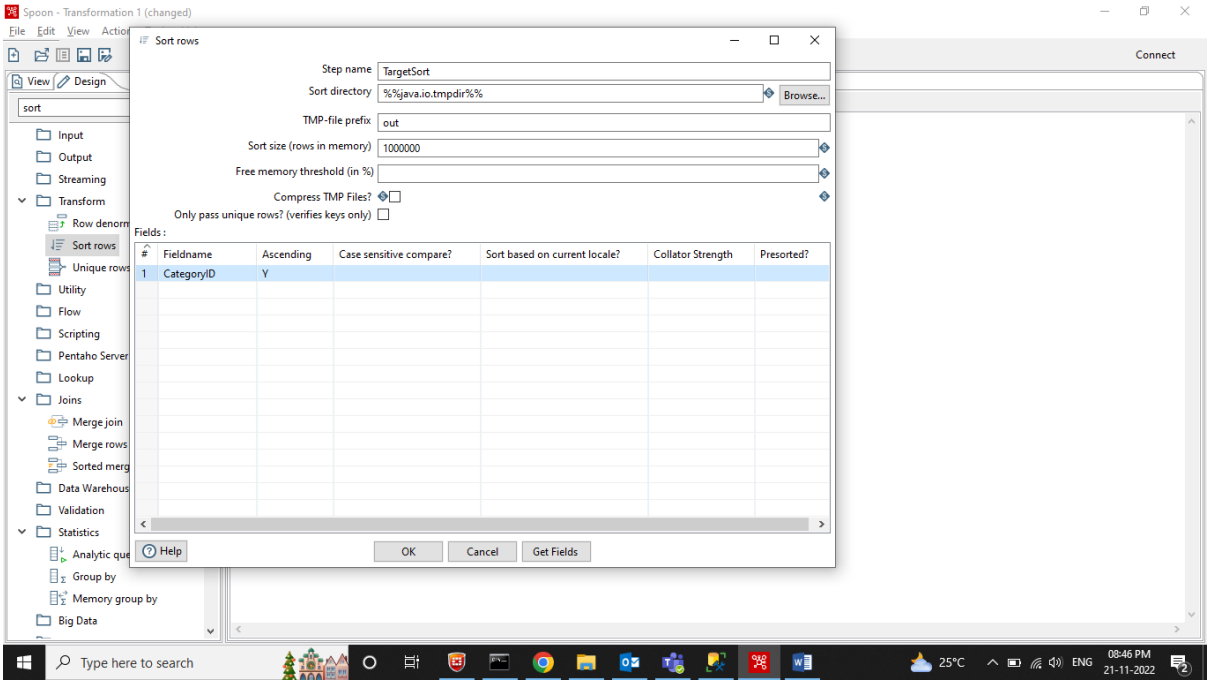
Step 8 : Enter the below details :

- Write the "**Step name**" (optional. If you want to change the name, you can change)
- Click on "**Get fields**" and you will be able to see all the column names of the table there. Select the field that you want to sort and you will be able to see the Ascending by default



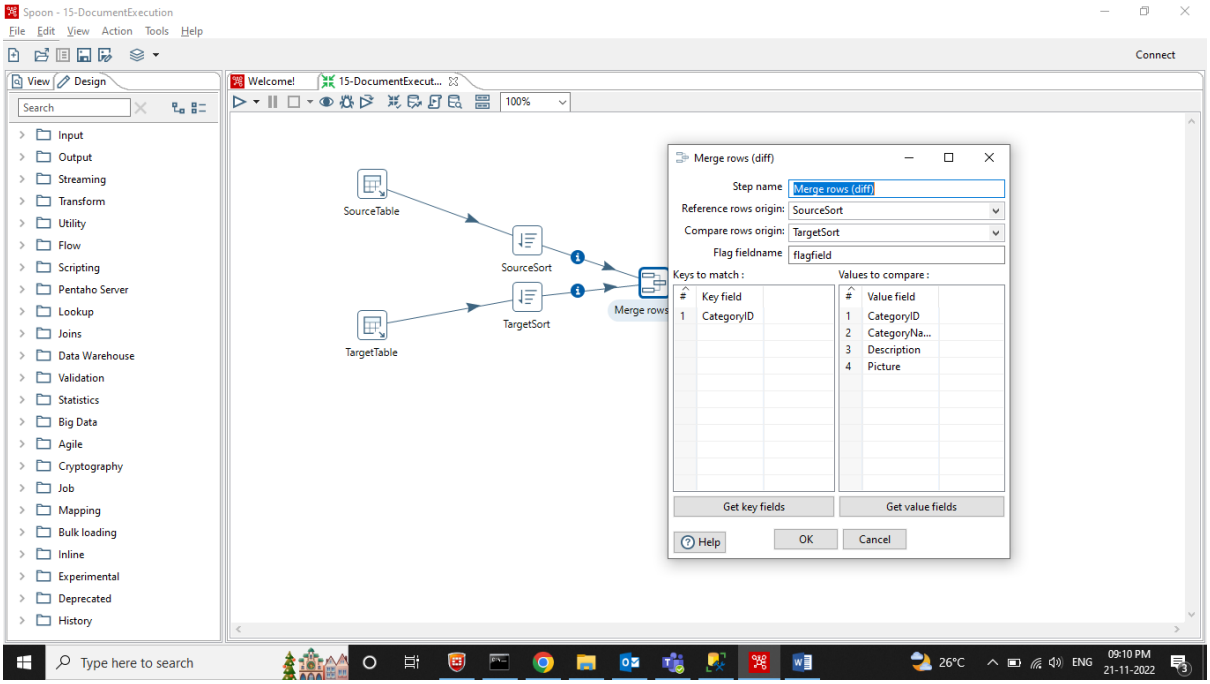
Step 9 : Now click on **“OK”**

Do the same for other “**Sort rows**”(Target Table input)

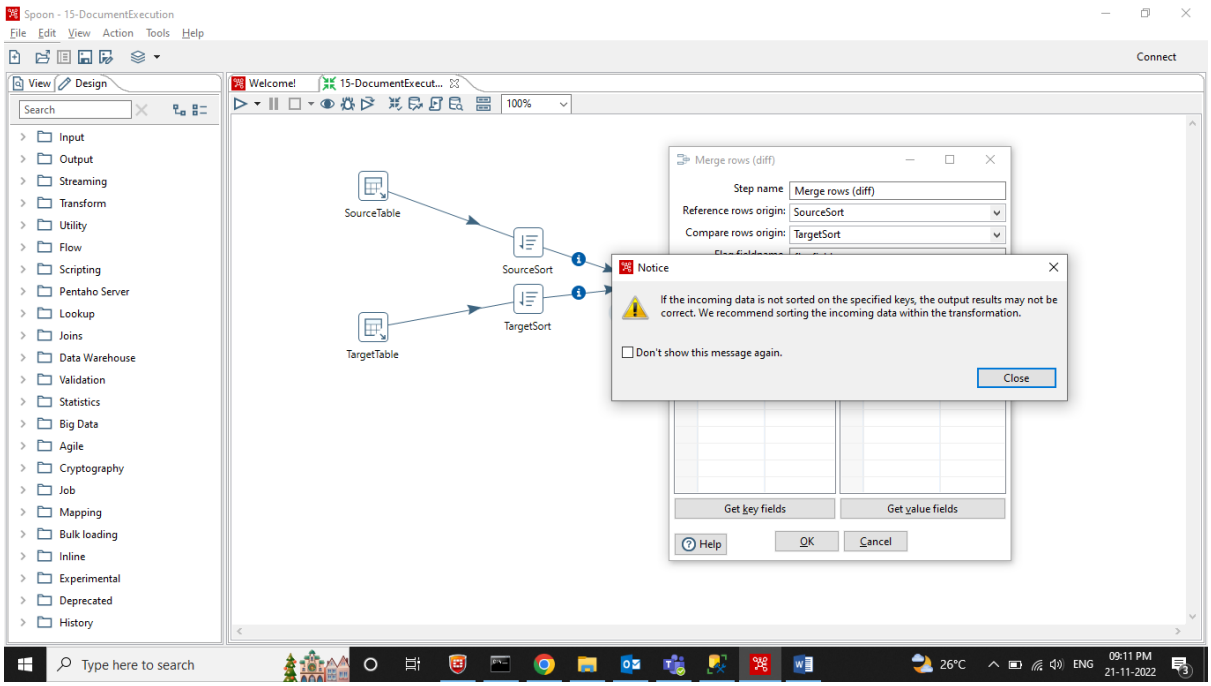


Step 10 : Now double click the **Merge rows(diff)**

The “**Reference rows origin**” and “**Compare rows origin**” will automatically changed here to the new values as below :

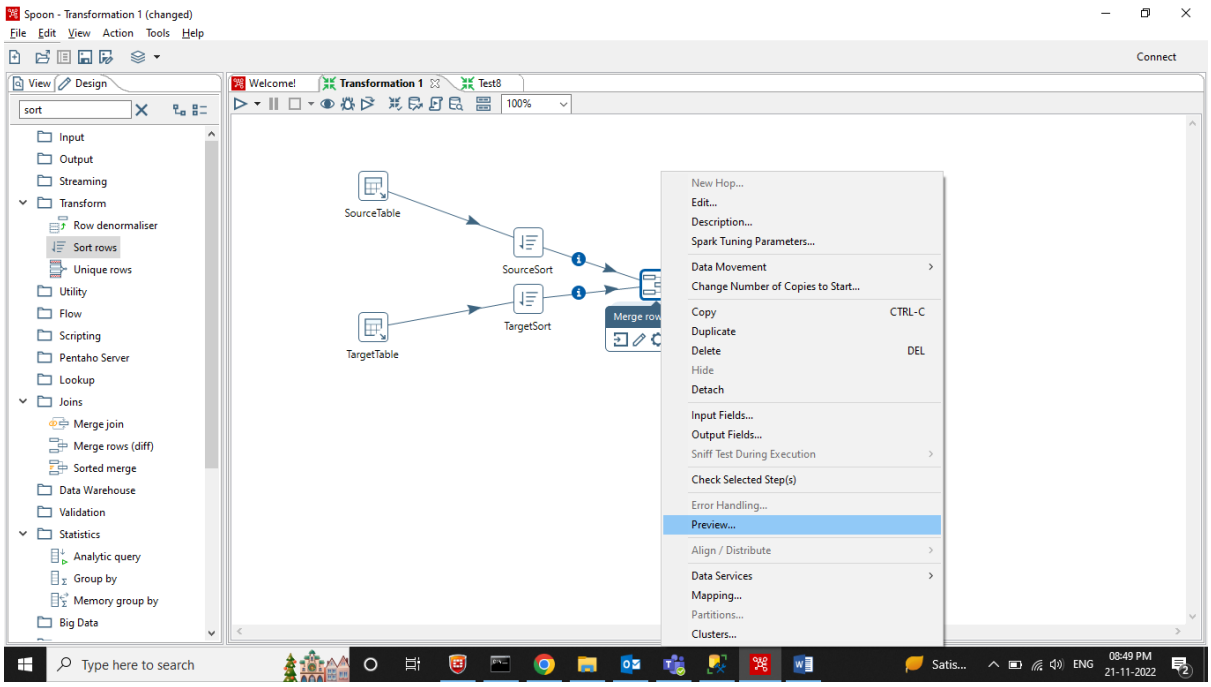


Step 11 : When click on "**OK**", you will be able to see this message again.

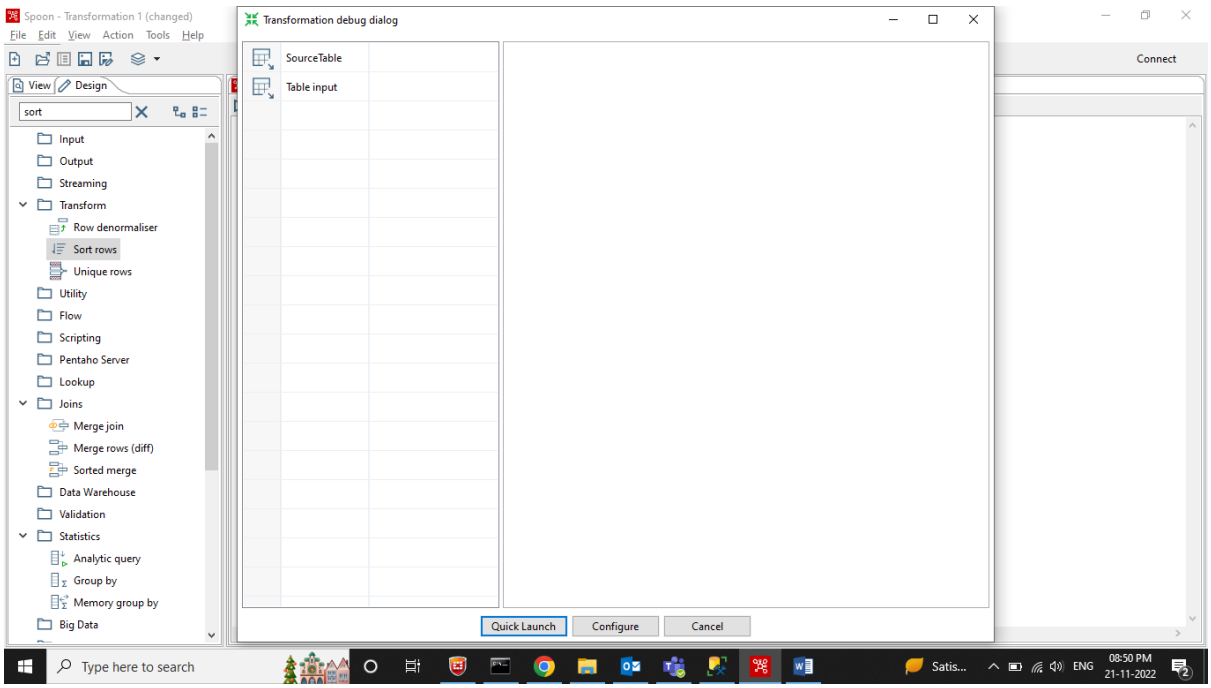


Step 12 : Click on "**Close**"

Step 13 : Now right click on the **Merge rows(diff)** and "**Preview**"



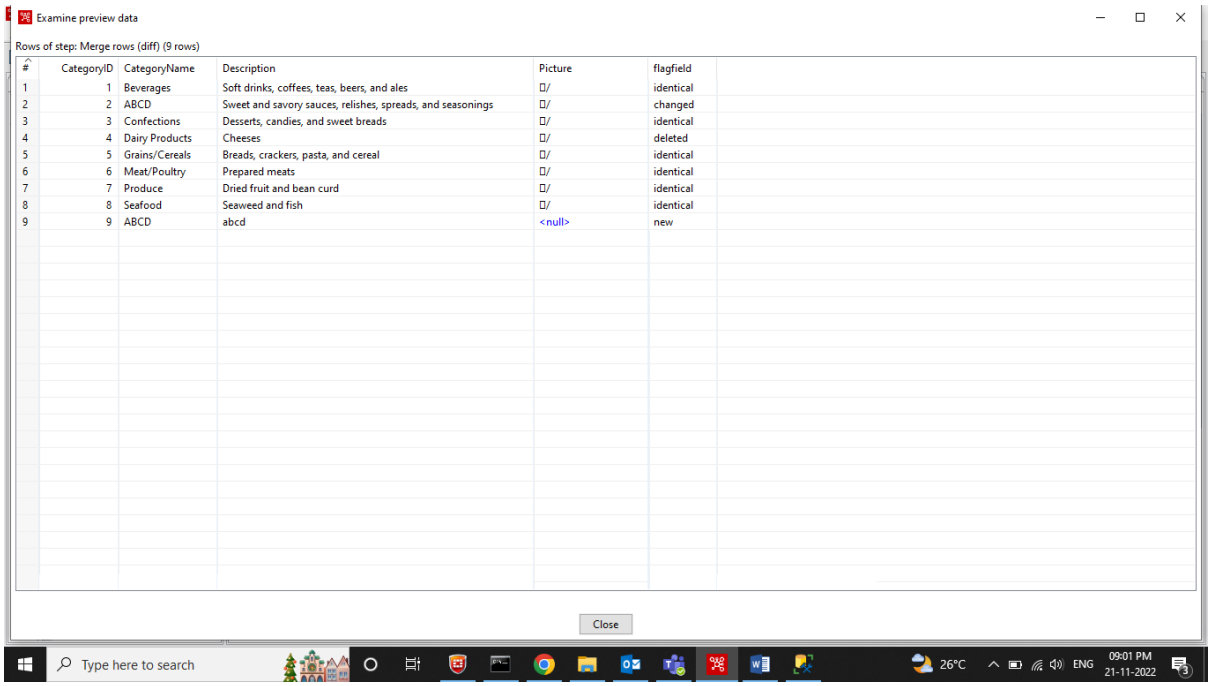
Step 14 : Click on “**Quick Launch**”



Then you will be able to see the compared result

4.Verify the generated report(To find the mismatches) :

When click on “Quick Launch”, then you will be able to see the compared result as below :

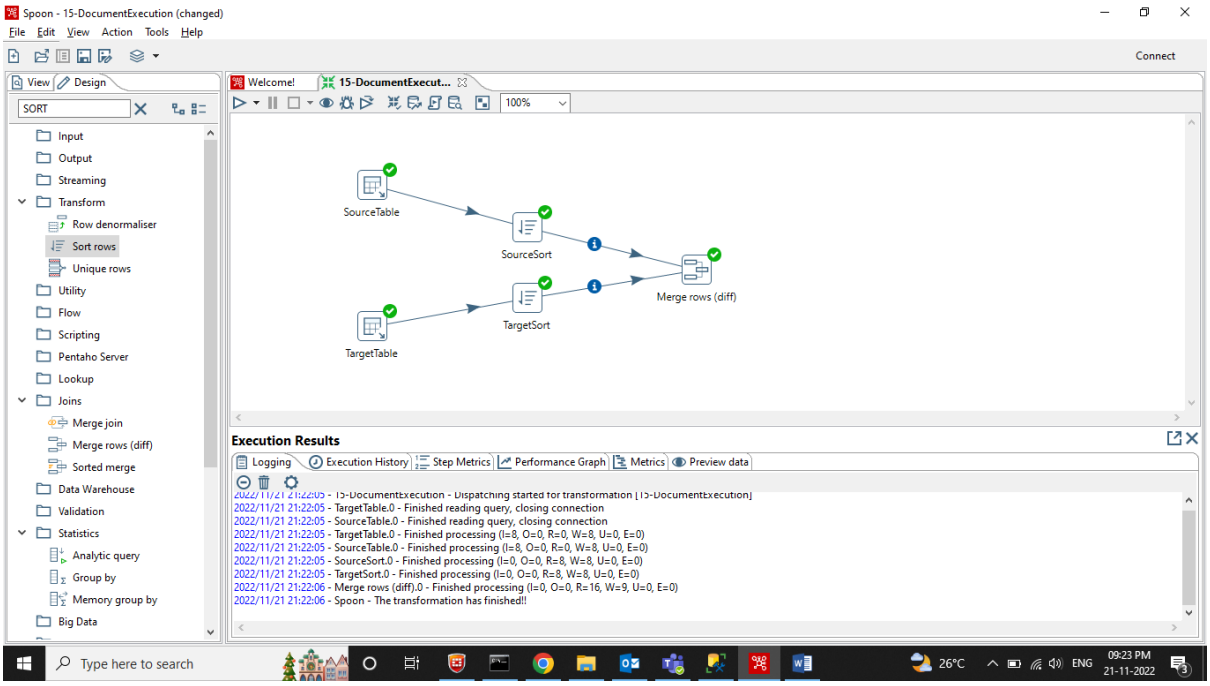


#	CategoryID	CategoryName	Description	Picture	flagfield
1	1	Beverages	Soft drinks, coffees, teas, beers, and ales	□/	identical
2	2	ABCD	Sweet and savory sauces, relishes, spreads, and seasonings	□/	changed
3	3	Confections	Desserts, candies, and sweet breads	□/	identical
4	4	Dairy Products	Cheeses	□/	deleted
5	5	Grains/Cereals	Breads, crackers, pasta, and cereal	□/	identical
6	6	Meat/Poultry	Prepared meats	□/	identical
7	7	Produce	Dried fruit and bean curd	□/	identical
8	8	Seafood	Seaweed and fish	□/	identical
9	9	ABCD	abcd	<null>	new

As mentioned earlier, the flag field indicates the changes as below :

- **Identical** : The key was found in both rows, and the compared values are identical.
- **Changed** : The key was found in both rows, but one or more compared values are different.
- **New** : The key was not found in the reference rows.
- **Deleted** : The key was not found in the compare rows.

Now when close this window, you will be able to see the icons with green tick(indicates executed without any errors) as below :



This is the comparison of 1 Source table to 1 Target table in Pentaho.

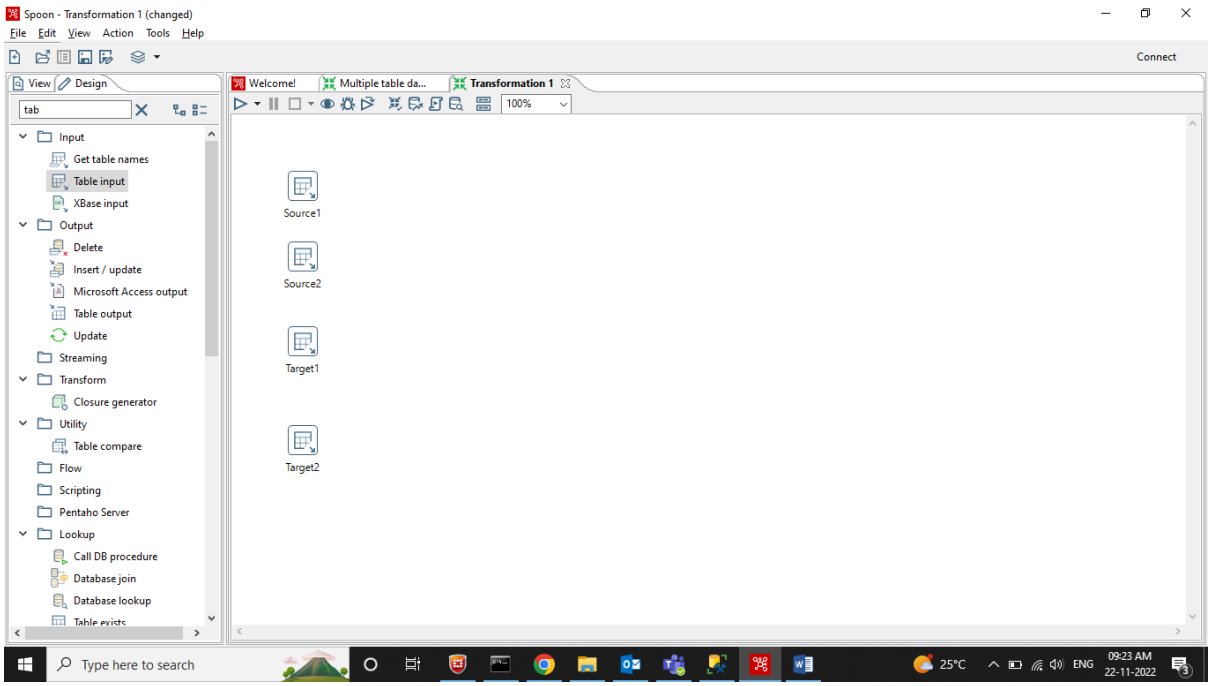
Like this, we can compare multiple source with multiple target tables.

Multiple Table Comparison:

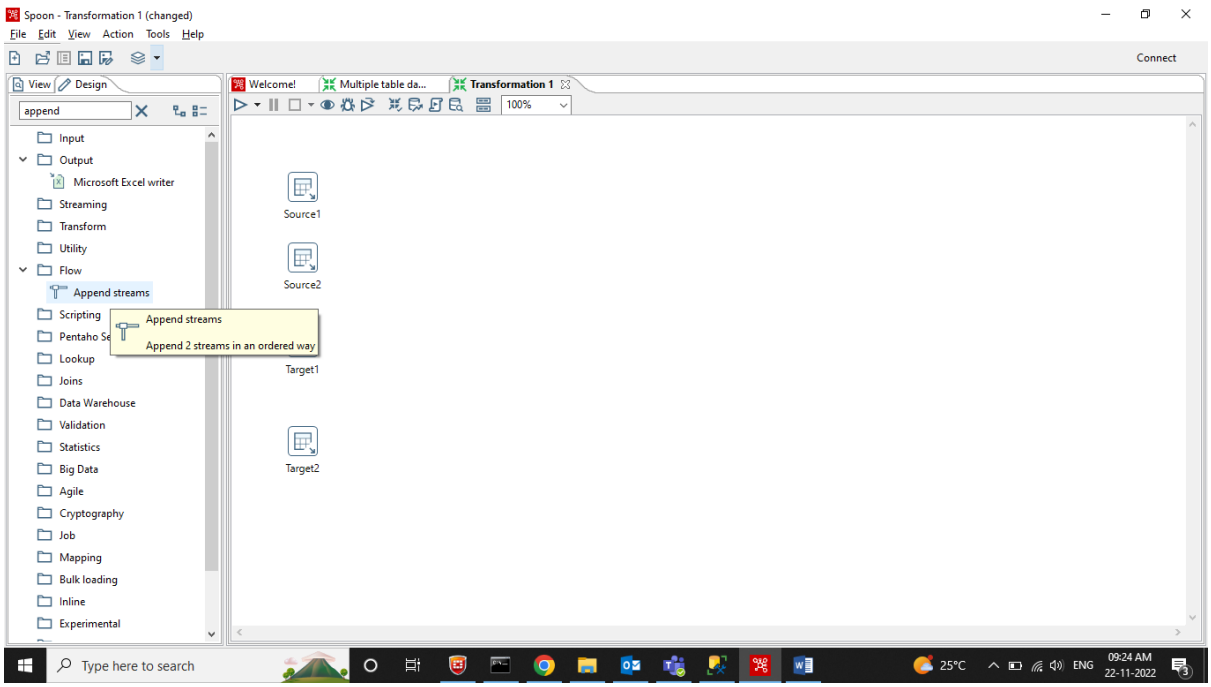
For multiple table comparison, we are using "**Append Streams**".

PFB the steps to compare multiple tables:

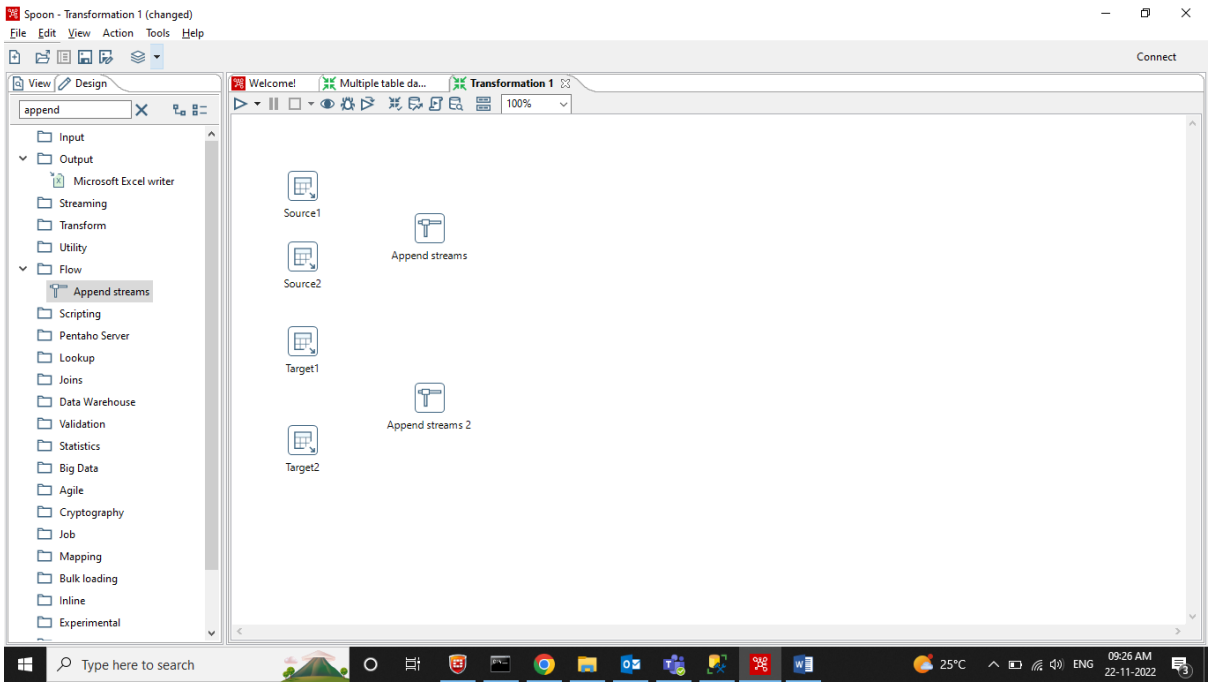
Step 1 : Like in "**Fetch Data from the Database Table"** ,fetch data from 2 source table and 2 target table and it will looks like below :



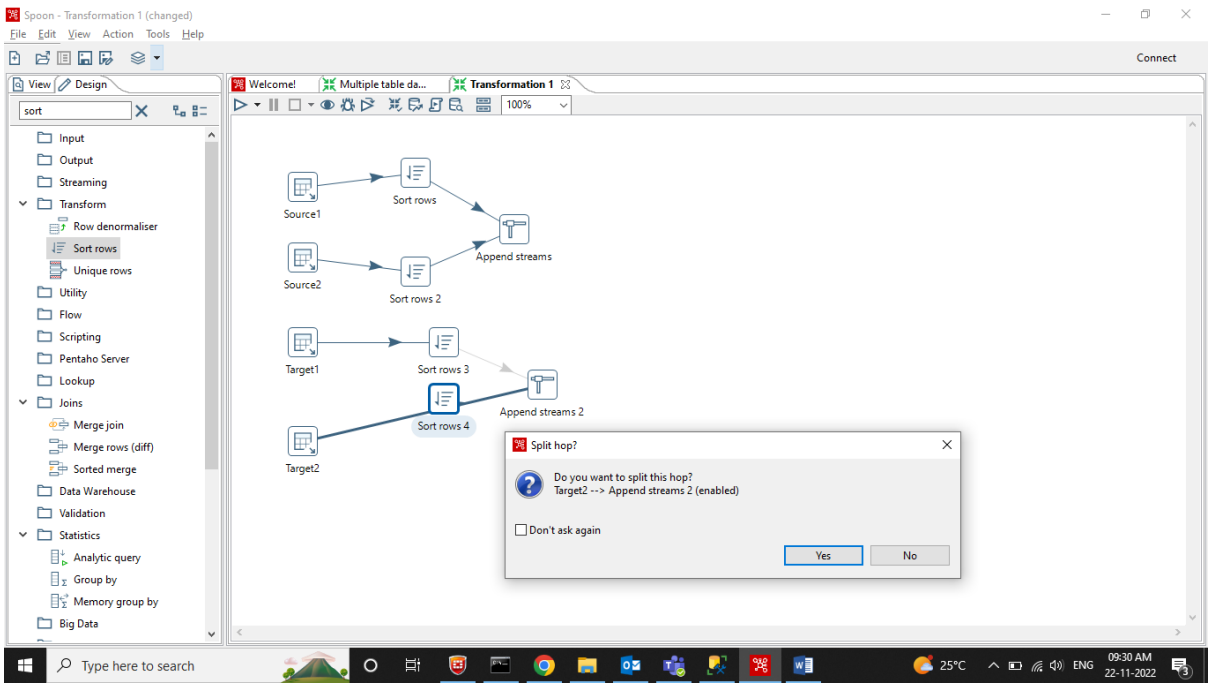
Step 2 : In "**Design"** tab, search like "**Append"**, then you will be able to see the **"Append Streams"** under "**Flow"** as below :



Step 3 : Drag and drop the "**Append Streams**" between the 2 tables that you want to merge. It will look like below :

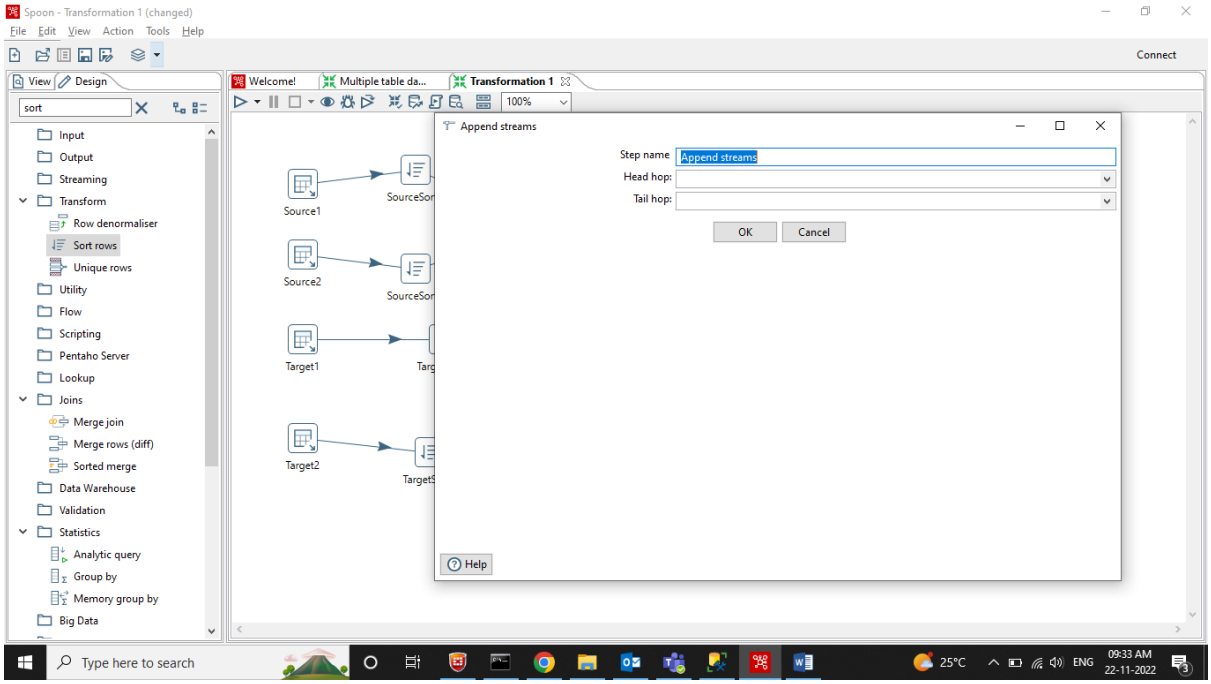


Step 4 : Make connection from both the table to "**Append Steams**" and then drag and drop the "**Sort rows**" also for sorting the data. It will looks like below :



Enter details in all the "**Sort rows**"

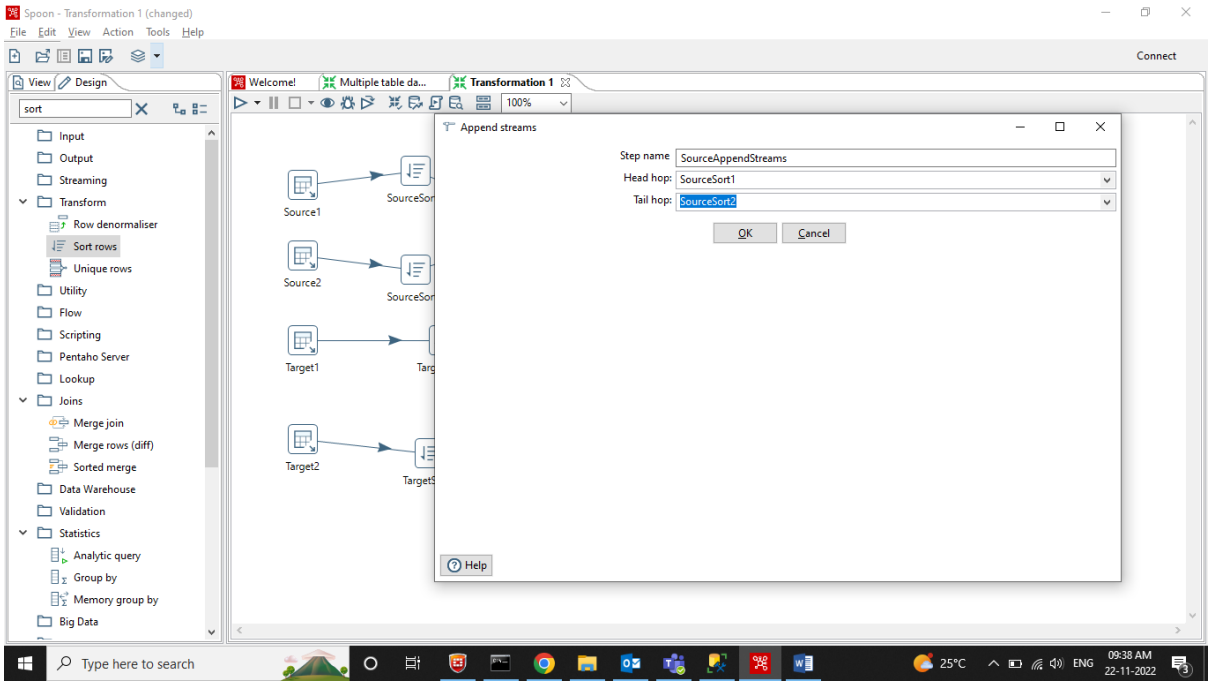
Step 5 : Now double click on the "**Append Streams**" and you will be able to see the window as below :



Steps 6 : Enter the below details :

- **Step Name** : If you want to give a name, you can give. It's optional.
- **Head Hop** : Select the name of the **Sort row**
- **Tail Hop** : Select the name of the **Sort row**

And tap on "**OK**"

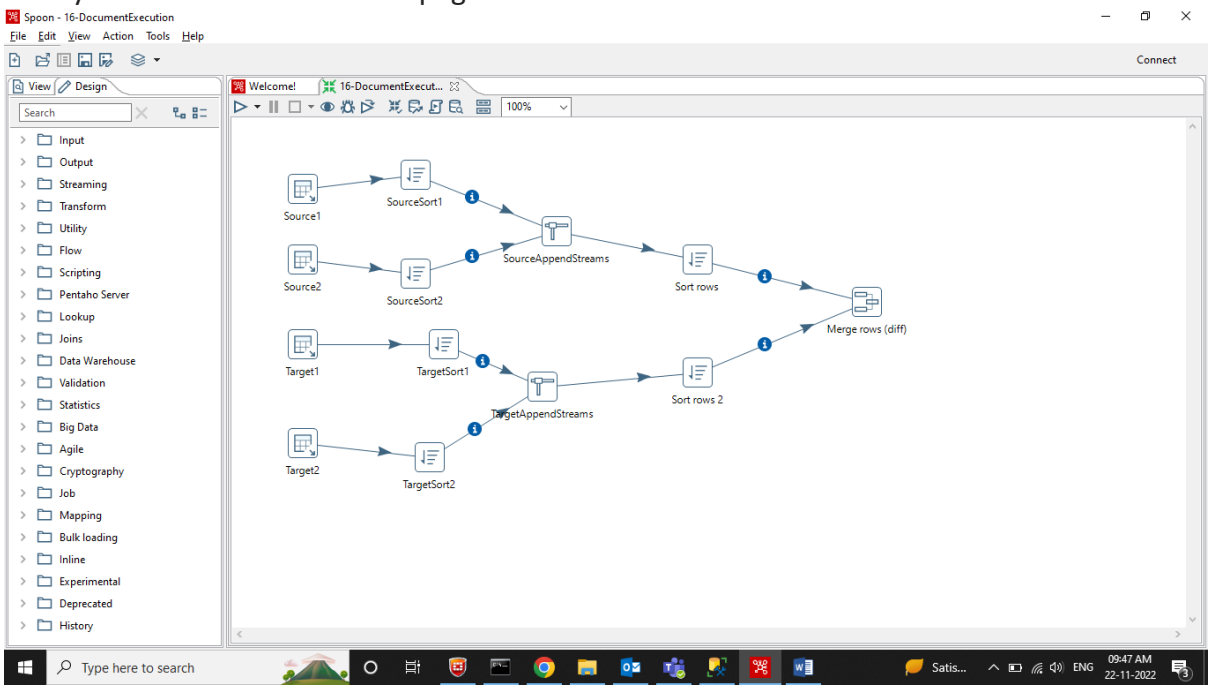


Do the same for **Target Append Stream** also.

Now the **Source tables** are combined and data stored in one "**Append Streams**" and **Target tables** are combined and data stored in other "**Append Streams**"

Step 7 : Now use the **"Merge rows(diff)**" as mentioned in "**Compare the Source table to Target tables**"

Now you will be able to see the page as below :



When "**Quick Launch**" the "**Merge rows(diff)**", you will be able to see the result of the combined data.