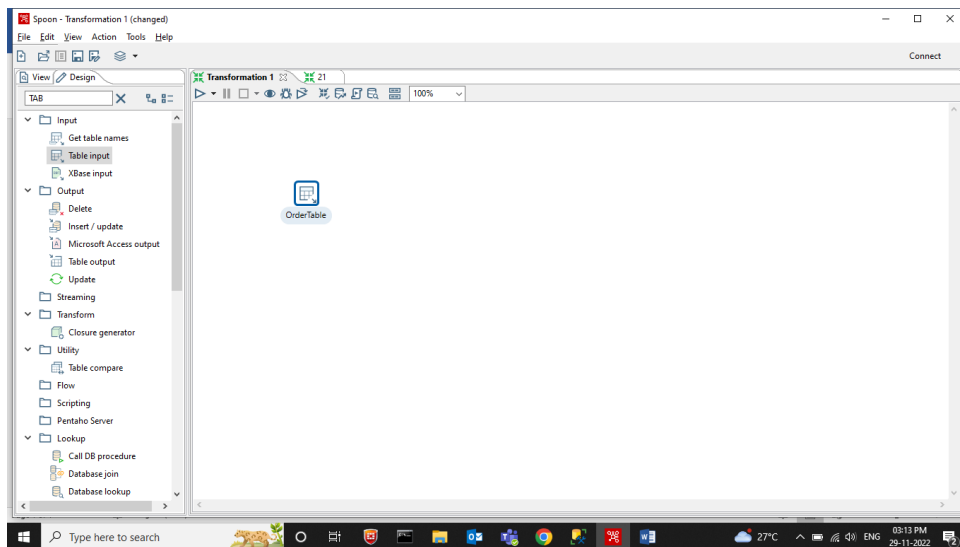# Pentaho Data Validation :

As explained in the document : "**Pentaho ETL Tool - Testing - User Guide**", Open **Pentaho** and do the following steps first :

1. Create a **Transformation**
2. Create the **Database Connection** to fetch data from DB
3. Drag and drop the "**Table input**" and write query in that to fetch data from the database table

Now your screen will looks like below :

Here my table name is "Order" and it has 830 rows.

Now we need to validate the data's inside the table based on some conditions. For that, We can do the below steps :
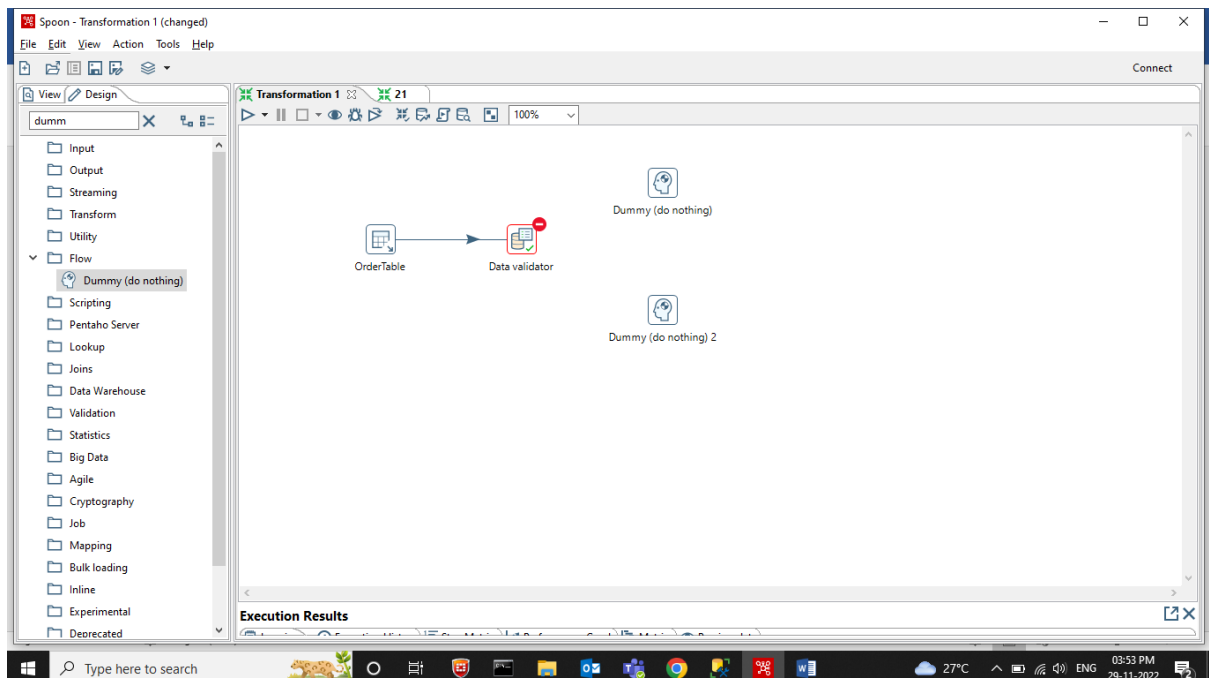
Step 1 : Search like "**validator**" in "**Design**" section, and you will be able to see the "**Data validator**" under "**Validation**" as below :
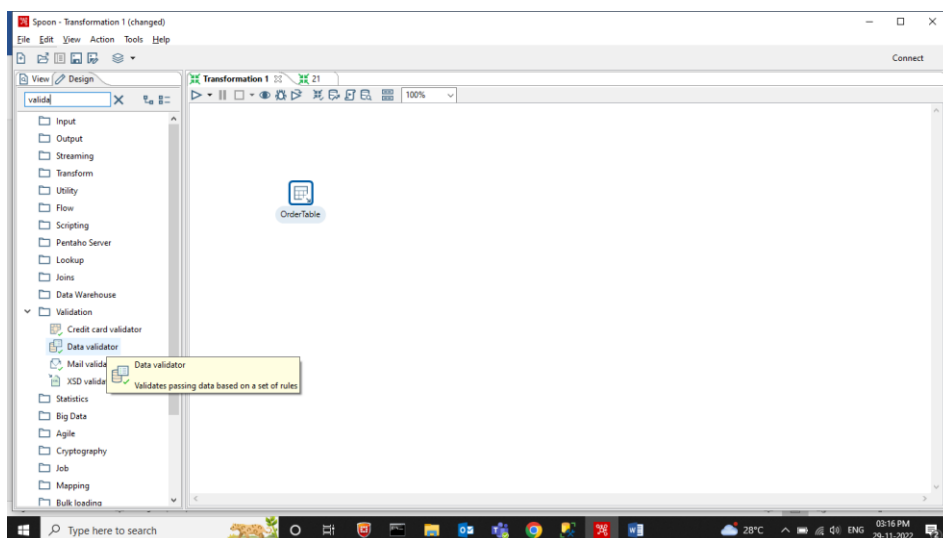
Step 2: Drag and drop the "**Data validator**" and make connection from the "**Table Input**" to the "**Data validator**" as below :
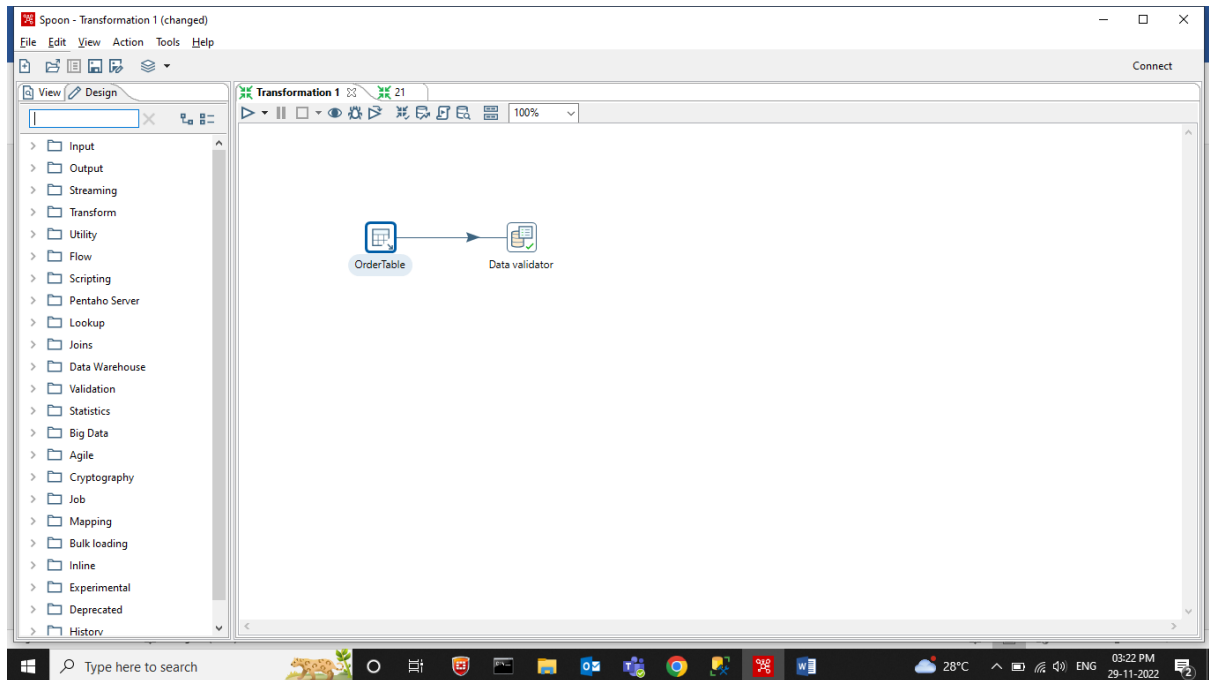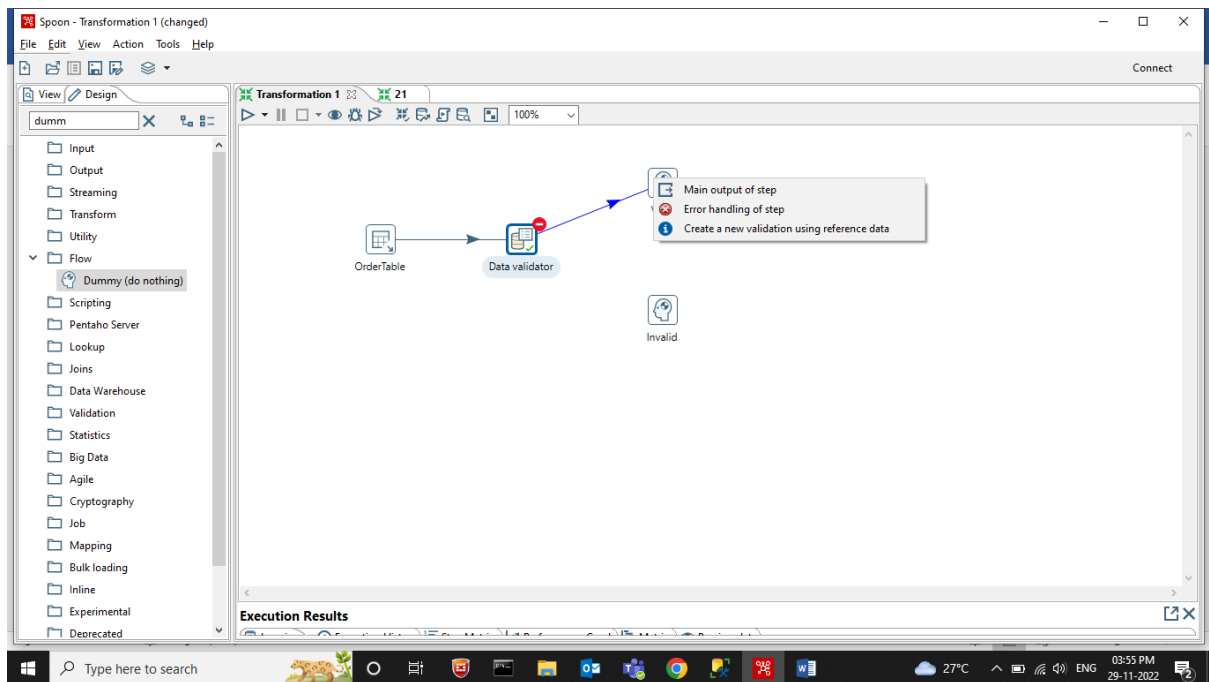

Step 3 : Search like "**dummy**" in the "**Design**" section and add 2 dummy as below to store the valid and invalid data's based on the condition :

Step 4: Double click on the dummy and change the name as "Valid" and "Invalid" to understand easily:
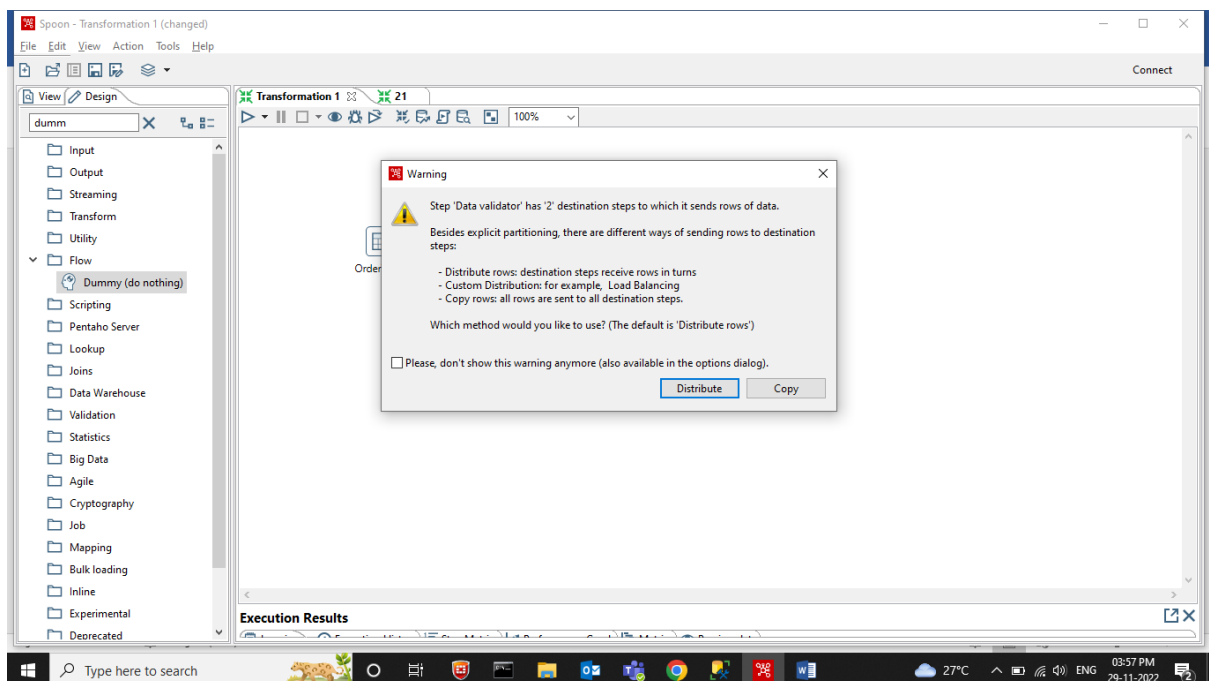
Step 5 : Now make connection from "**Data validator**" to the "**Valid**" and you will be able to see as below :

Select "**Main output of step**"

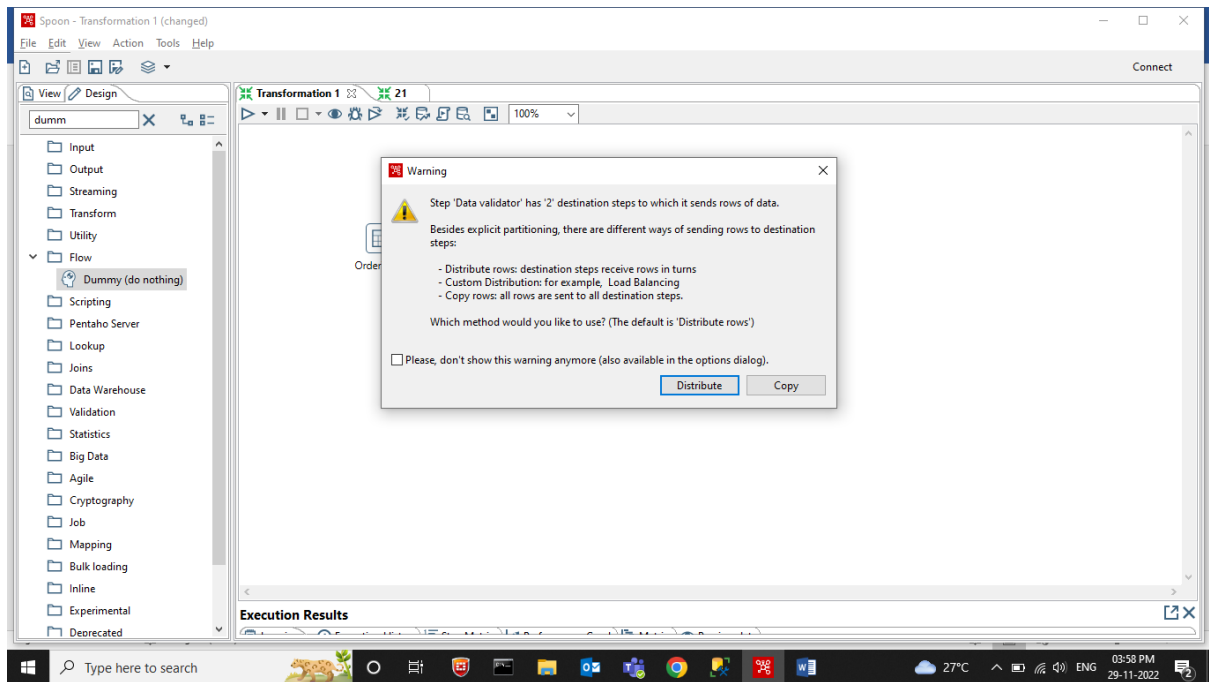Now make connection to the "**Invalid**" one and select "**Error handling of step**"

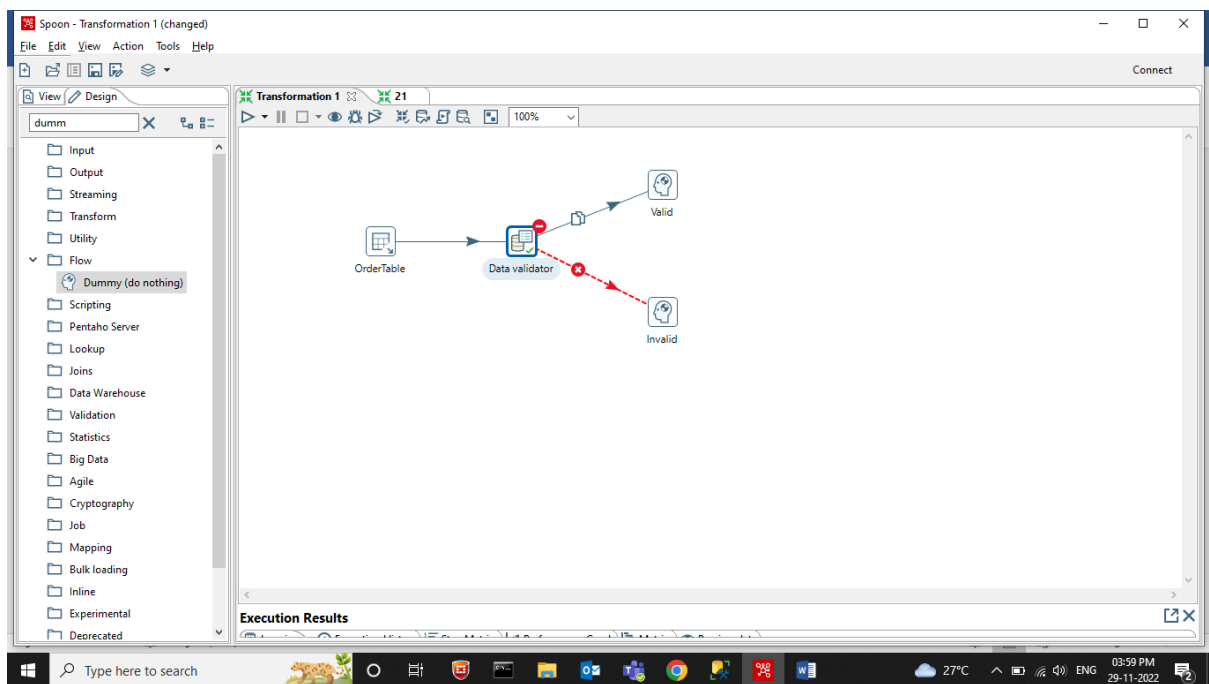You will be able to see the below window:

Click on "**Distribute**"

Now delete the connection given to the "**Valid**" one and again make connection.

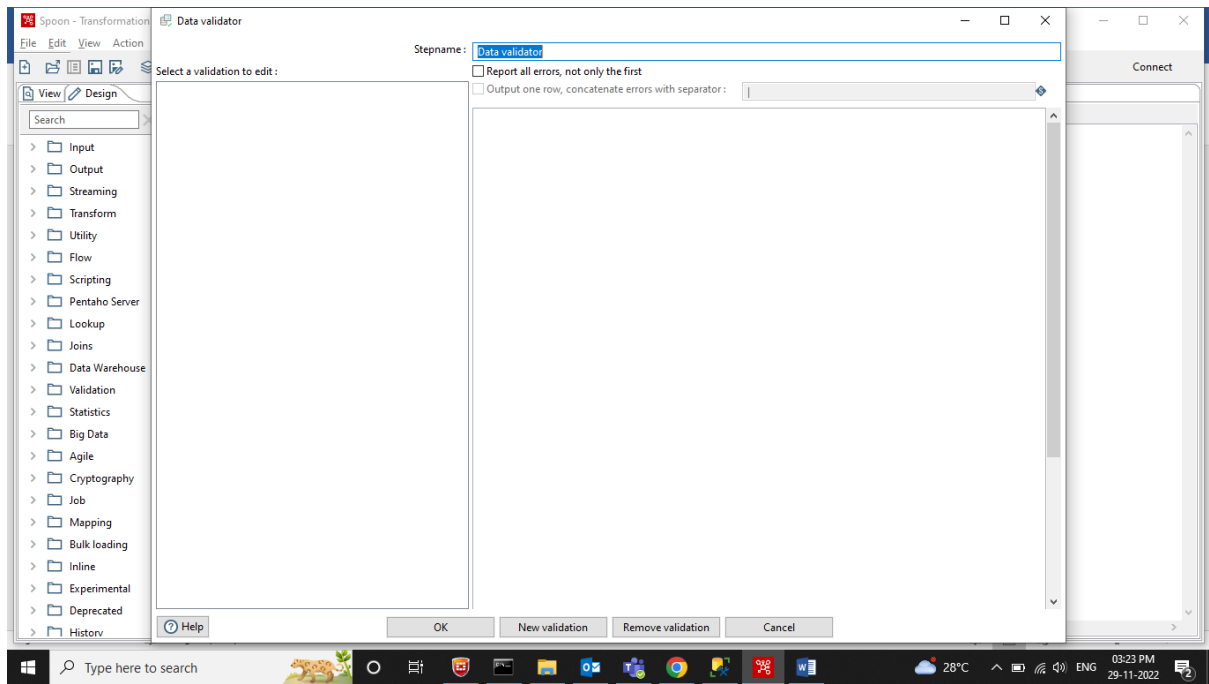Now Select "**Main output of step**", and you will be able to see the below window again :
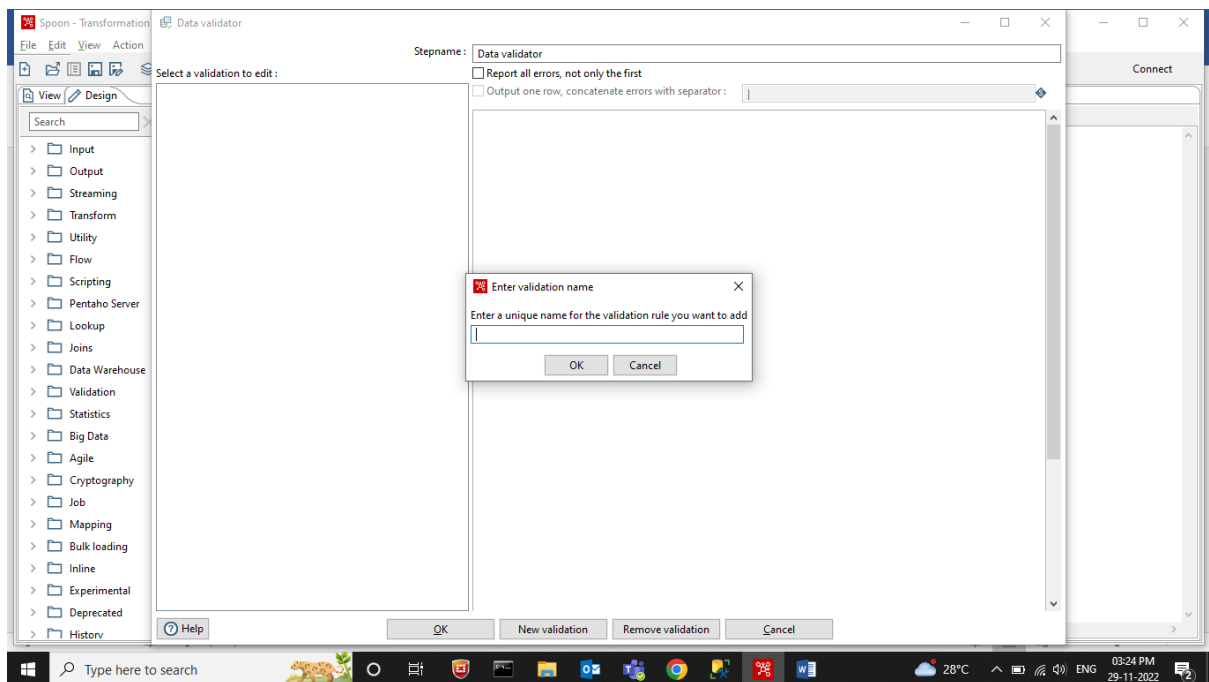
Then click on "**Copy**" option

Now your screen will looks like below :

Step 6 : Now double click on the "**Data validator**" and the window will open as below :

Step 7 : Click on the "**New Validation**" button, and you will be able to see the below screen :
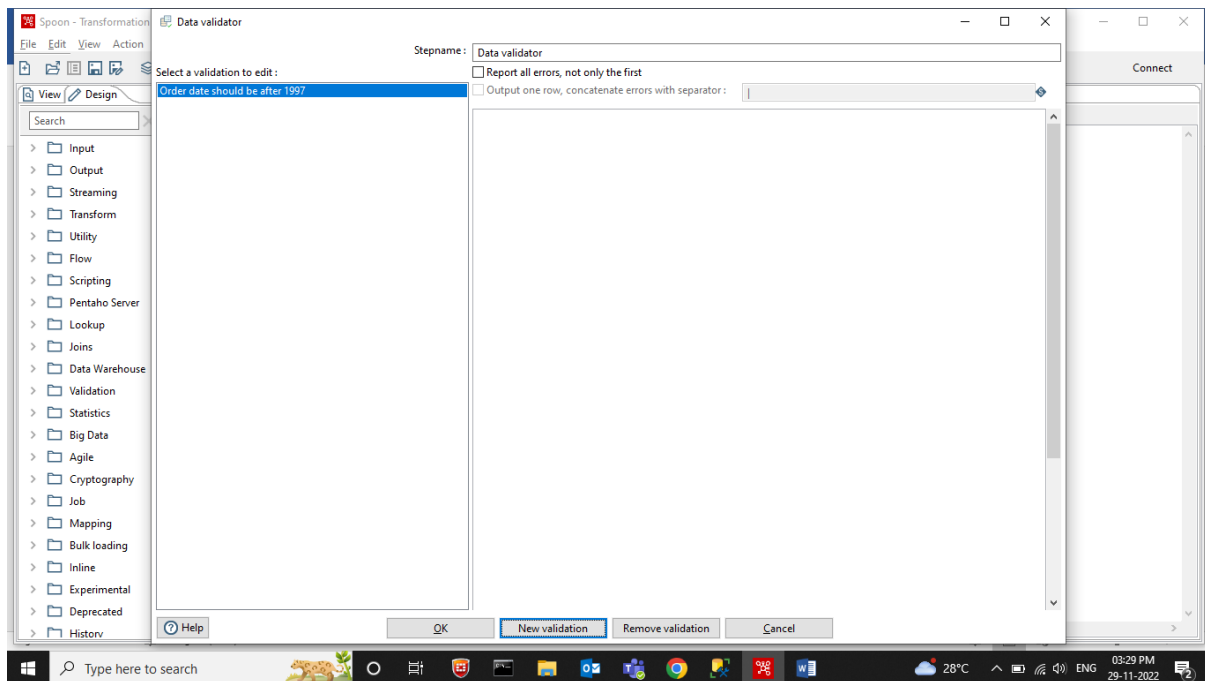

**Step 8** : Enter a name for the validation that you want to do.

Ex : "Order date should be after 1997"

Here, "Order date" is a field inside my table "Order". And I want to fetch the data from the table based on the condition applied on the "Order Date"

Now click on "**OK**", and you will be able to see the below screen:

Step 9 : Click on the newly added validation name, and you will be able to see the fields to enter some details like below :
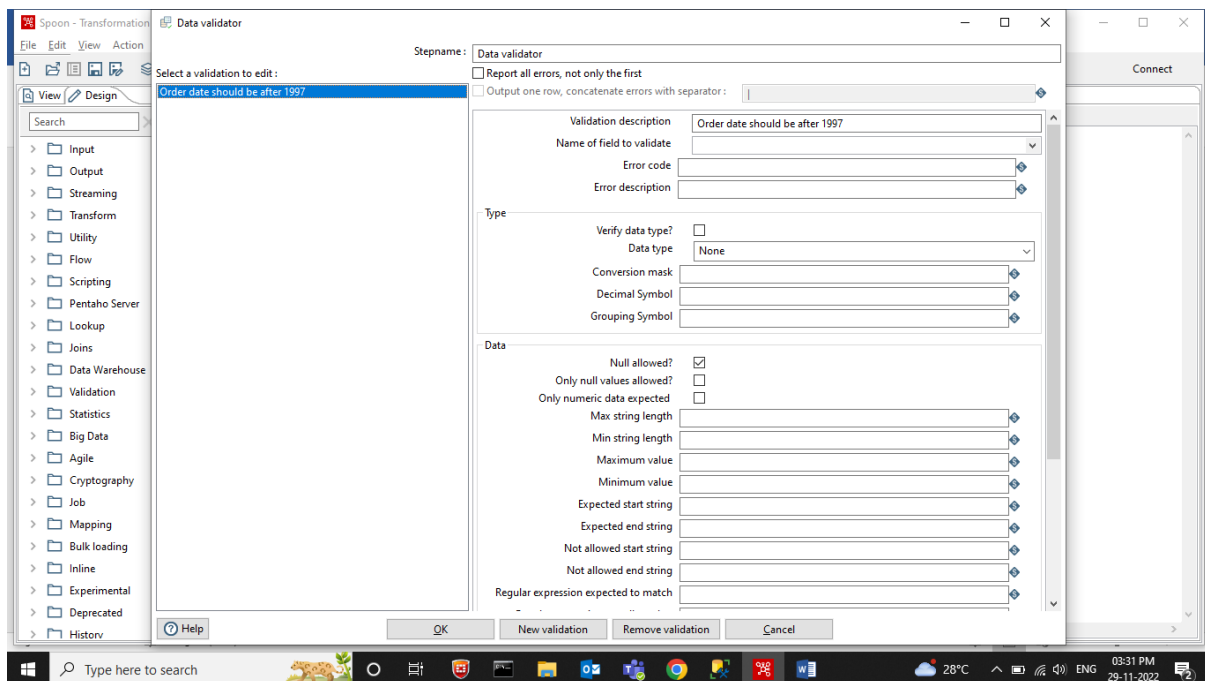
Step 10: Enter the below details(The details need to give will be based on your condition) :

1. Validation description : Auto populated from the name that we have given
2. Name of field to validate : Select the field to validate from the drop down. It will show all the column names from your table. Here, for me, it's "OrderDate"
3. Error code : Give any name like "DATE_ERROR"

4. Error description : Give the details of the error like "Order date should be after 1997"
5. Verify data type : If you want to verify the data type also, check the box and select the data type from the drop down
   Here, I need to verify the data type. So I entered the details based on that.
6. In "Data" section, if you want to include the "Null values" also, you can check the checkbox
7. In, "Data" section, fill the details based on your condition.
   Here, I need to enter the "Minimum value" as "1998-01-01 00:00:00". Because I am validating the Order date after 1997
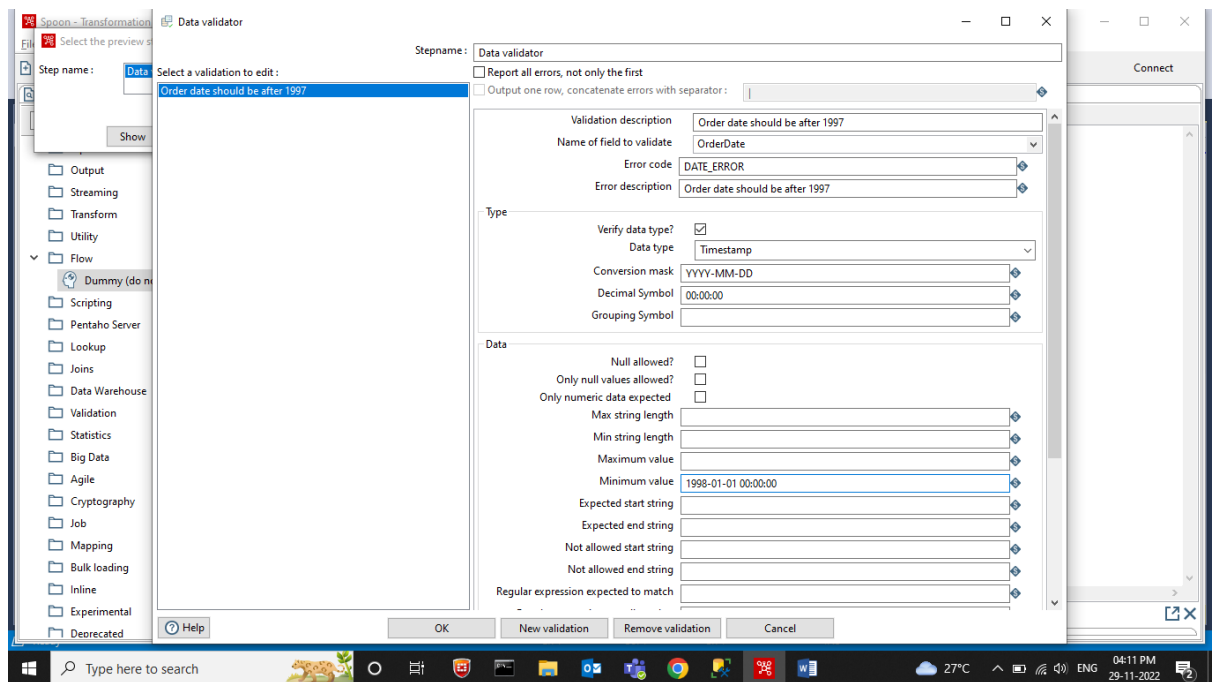
Now click on "**OK**" button

Step 11 : Now "**Preview** "and "**Quick Launch**" the "**Data validator**" by right clicking on it. And you will be able to see the result based on the condition as below :

I have 270 rows from 830 in my result(that matches my condition).