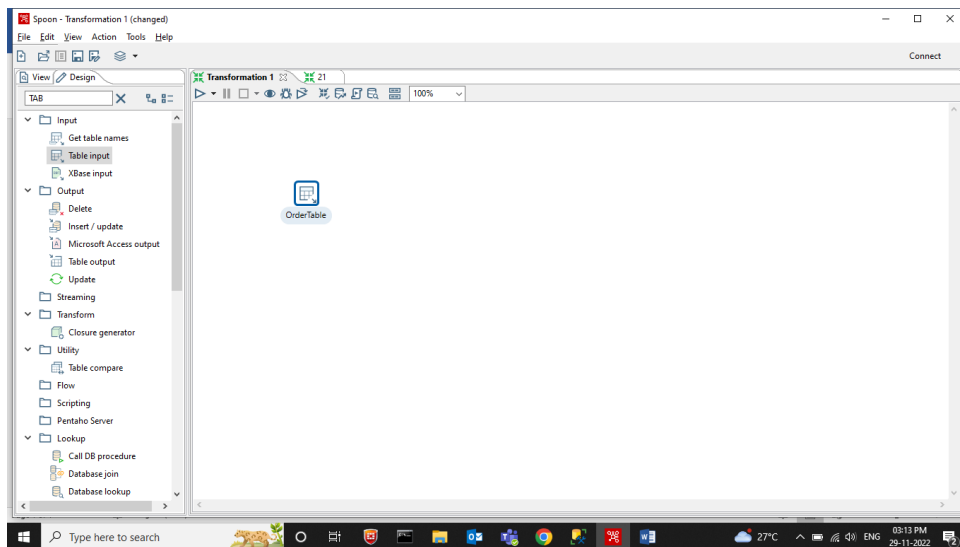


Pentaho Data Validation :

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

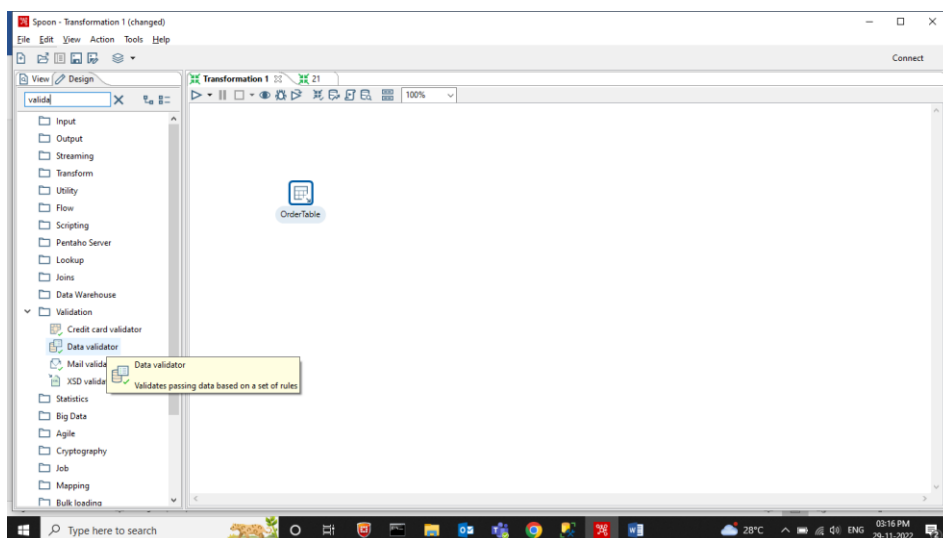
Now your screen will look like below :



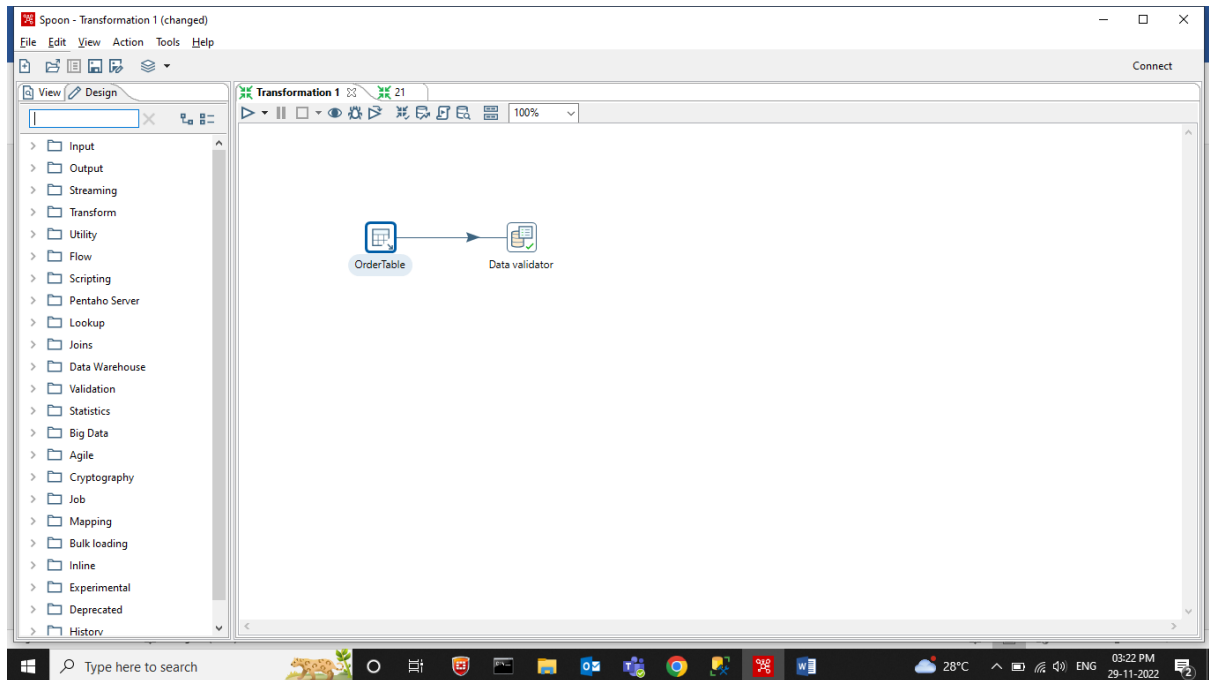
Here my table name is “Order” and it has 830 rows.

Now we need to validate the data’s inside the table based on some conditions. For that, We can do the below steps :

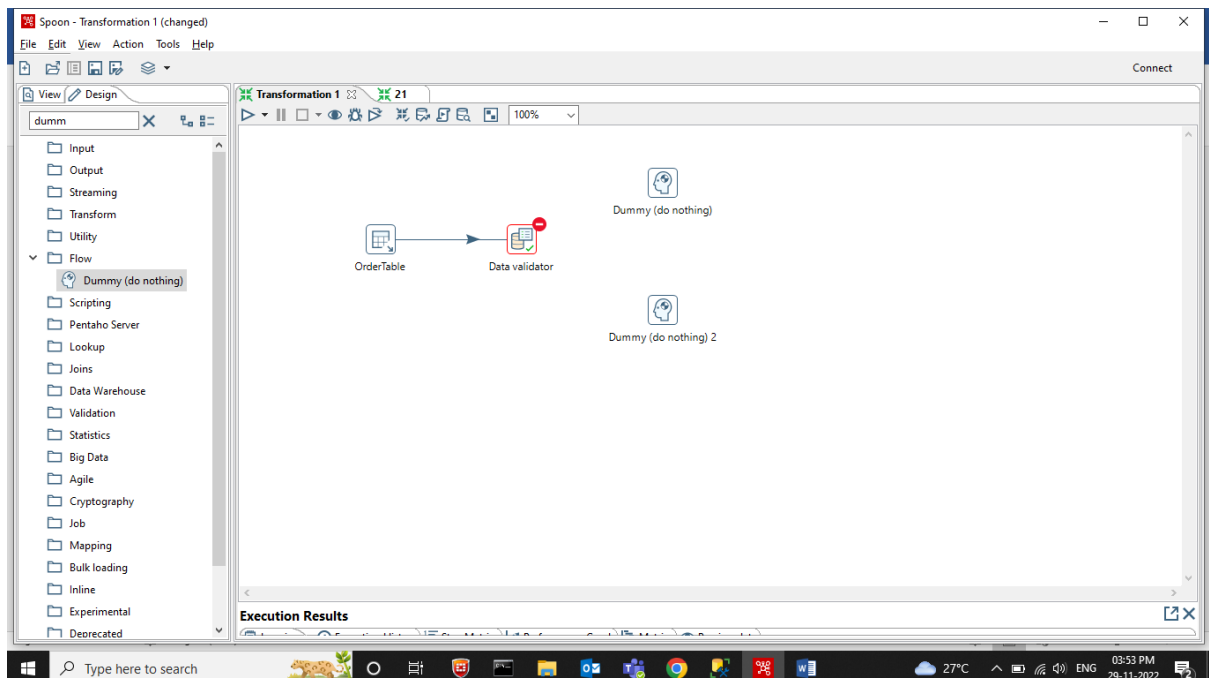
Step 1 : Search like “**validator**” in “**Design**” section, and you will be able to see the “**Data validator**” under “**Validation**” as below :



Step 2: Drag and drop the “Data validator” and make connection from the “Table Input” to the “Data validator” as below :

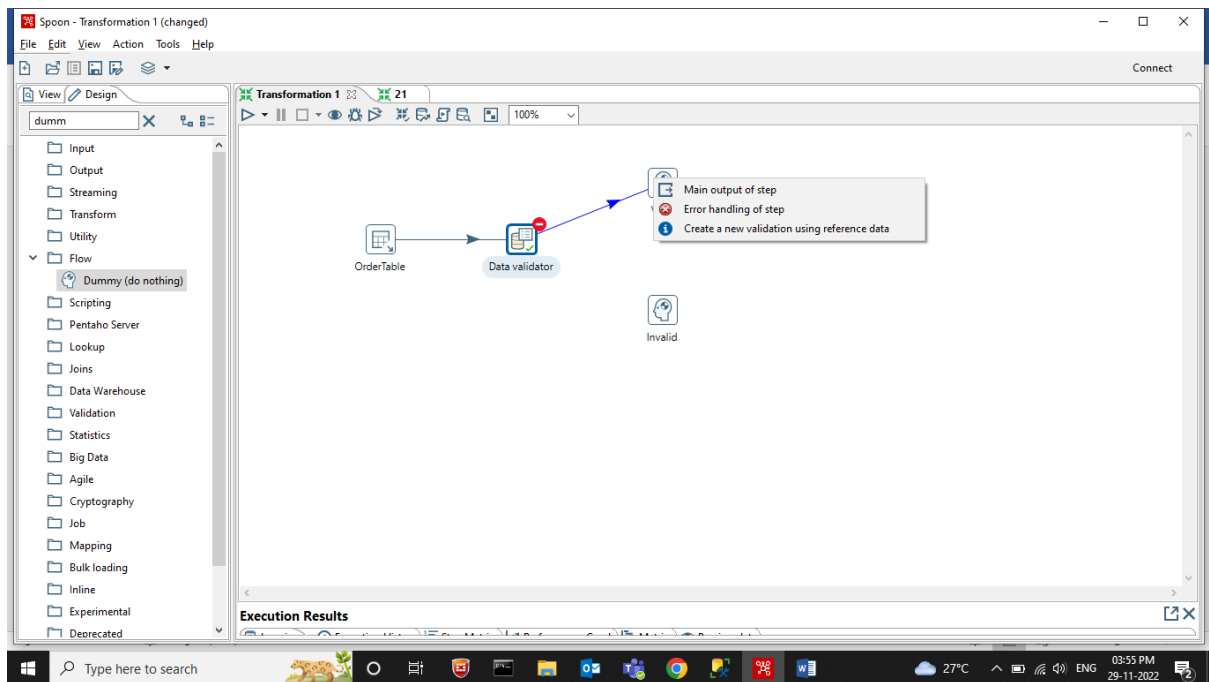


Step 3 : Search like “dummy” in the “Design” section and add 2 dummy as below to store the valid and invalid data’s based on the condition :



Step 4: Double click on the dummy and change the name as “Valid” and “Invalid” to understand easily:

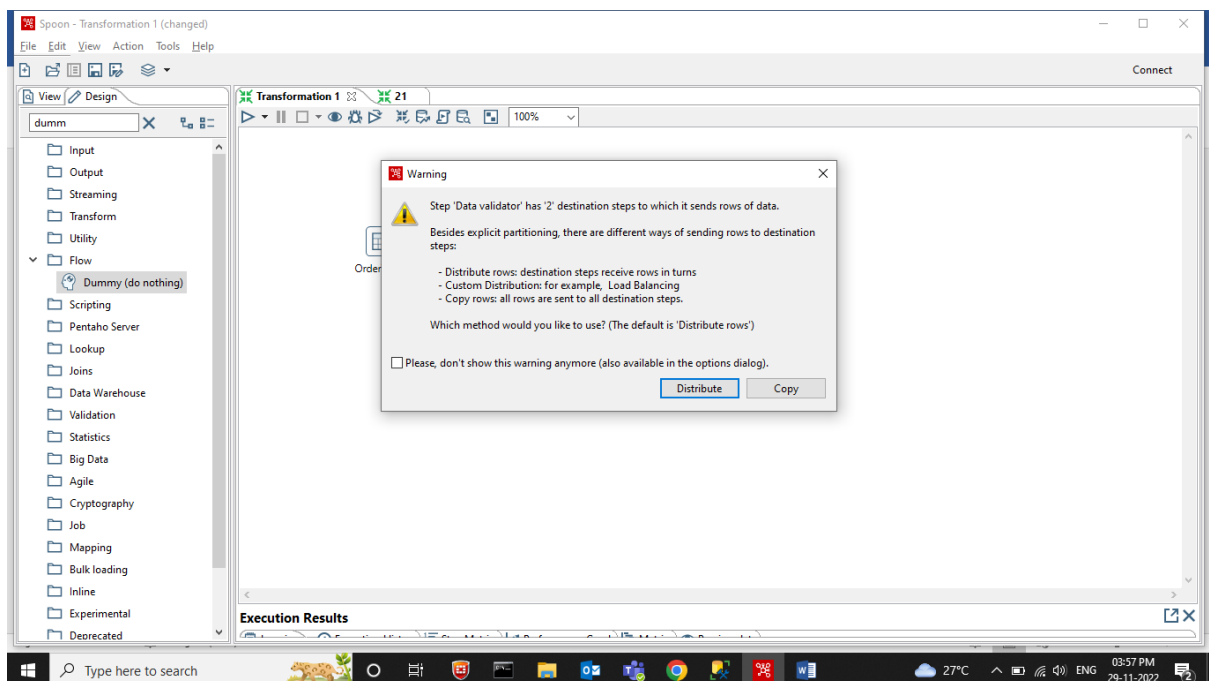
Step 5 : Now make connection from “**Data validator**” to the “**Valid**” and you will be able to see as below :



Select “**Main output of step**”

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

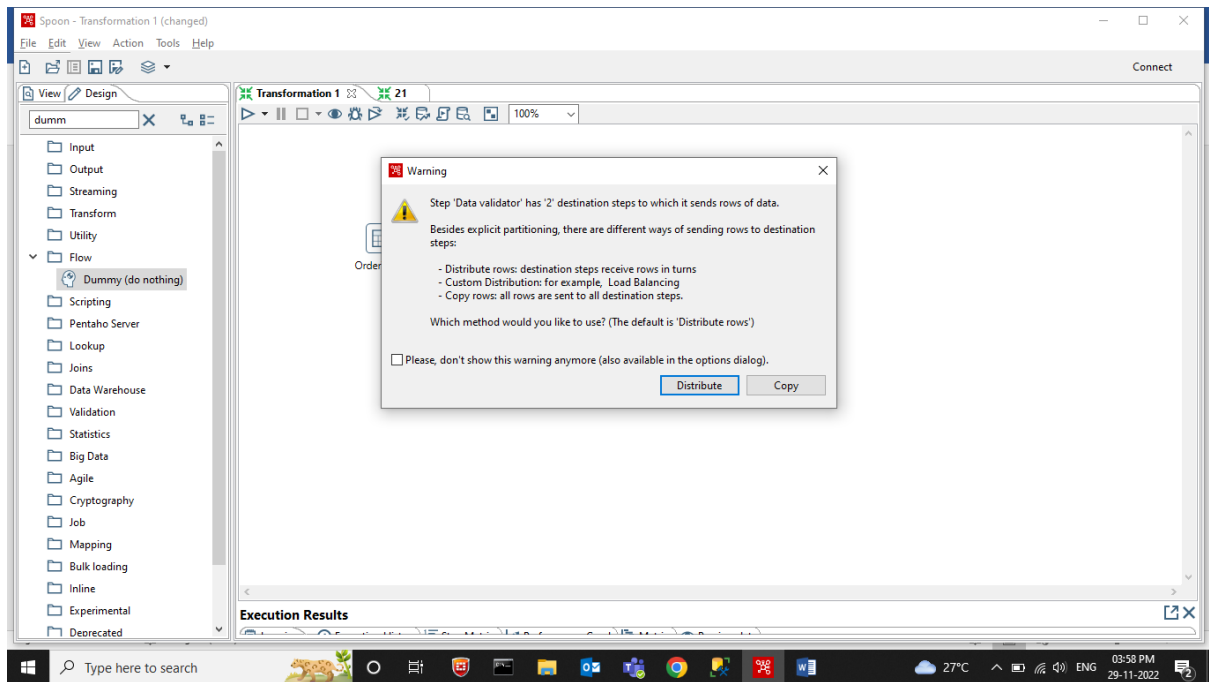
You will be able to see the below window:



Click on “**Distribute**”

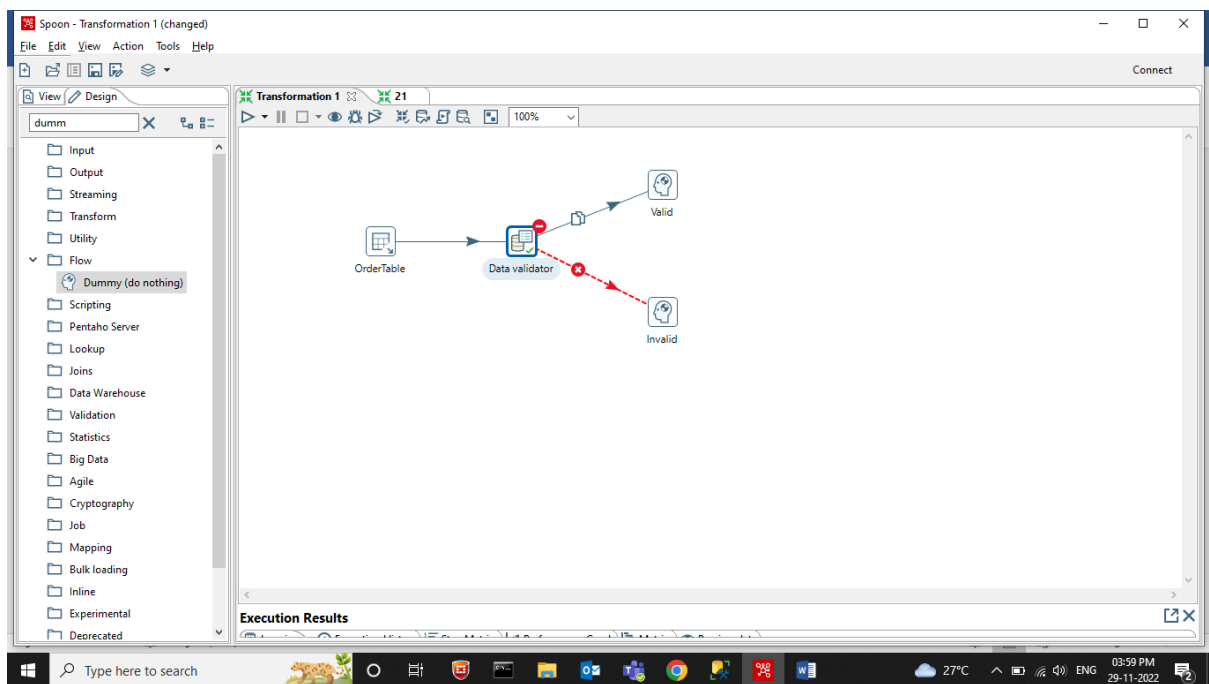
Now delete the connection given to the “**Valid**” one and again make connection.

Now Select “**Main output of step**”, and you will be able to see the below window again :

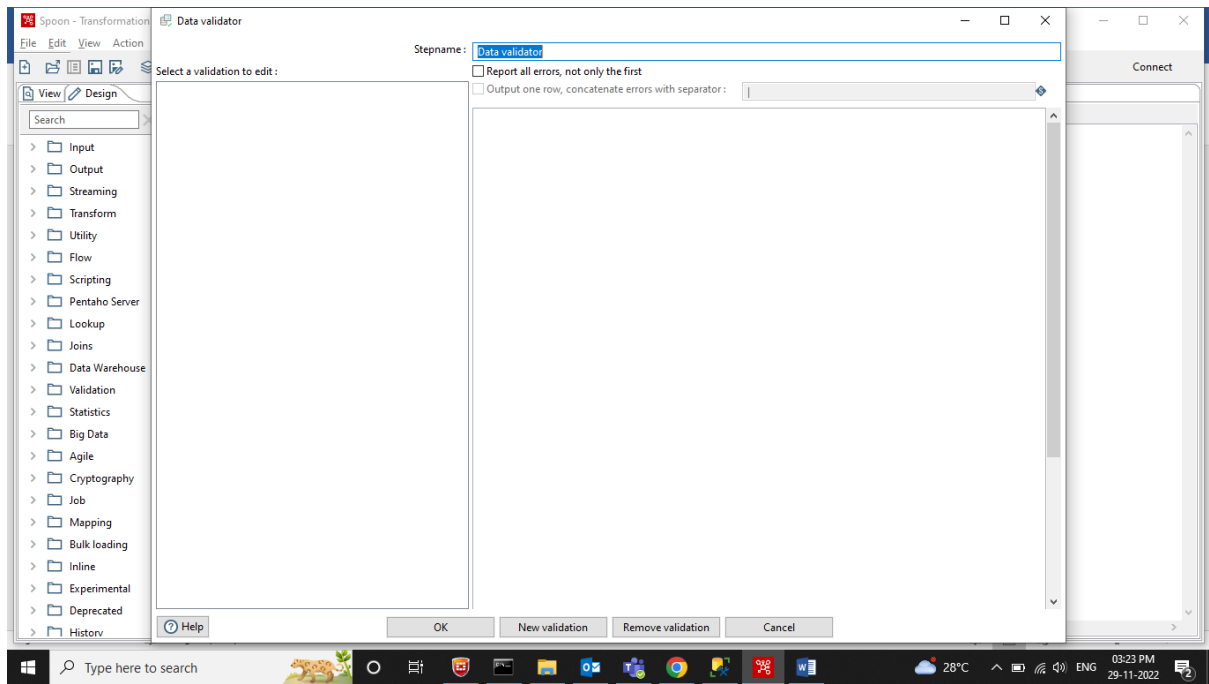


Then click on **“Copy”** option

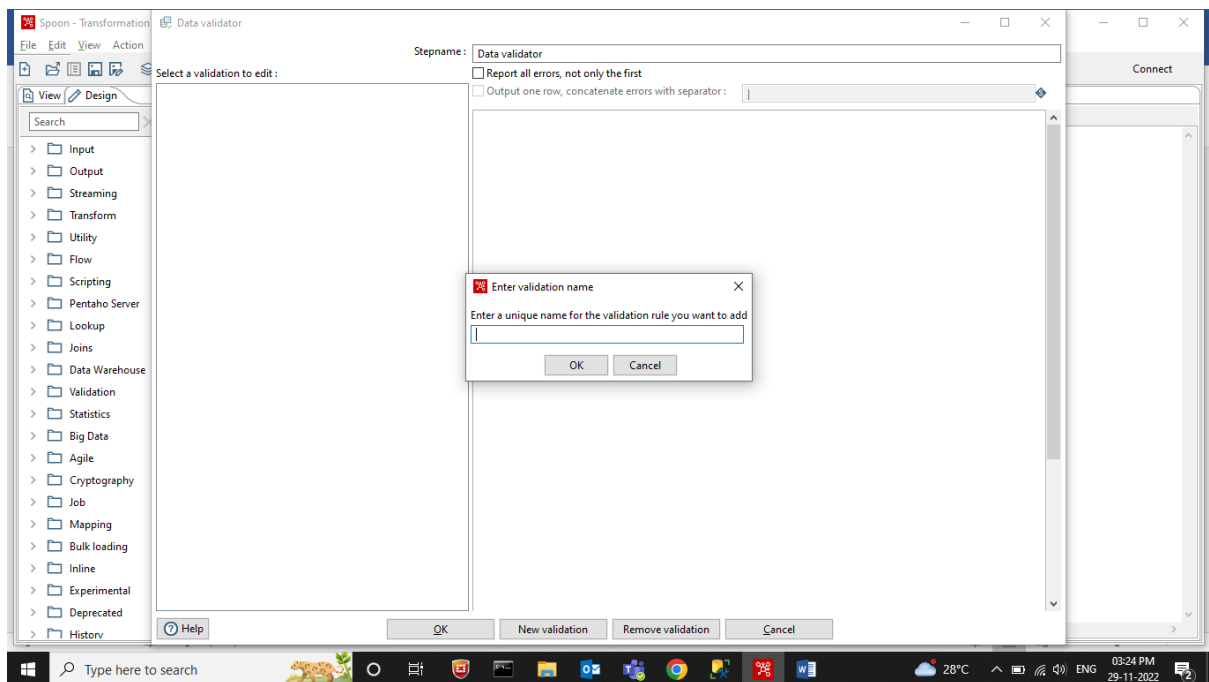
Now your screen will looks like below :



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 :

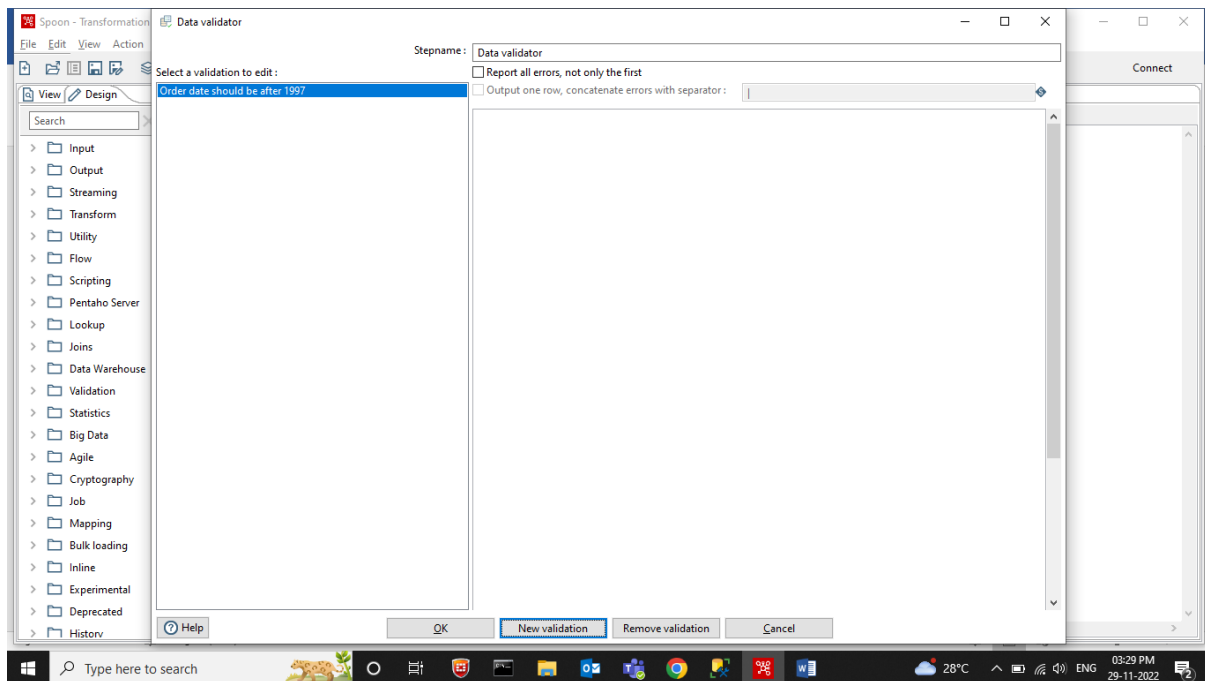


Step 8 : Enter a name for the validation that you want to do.

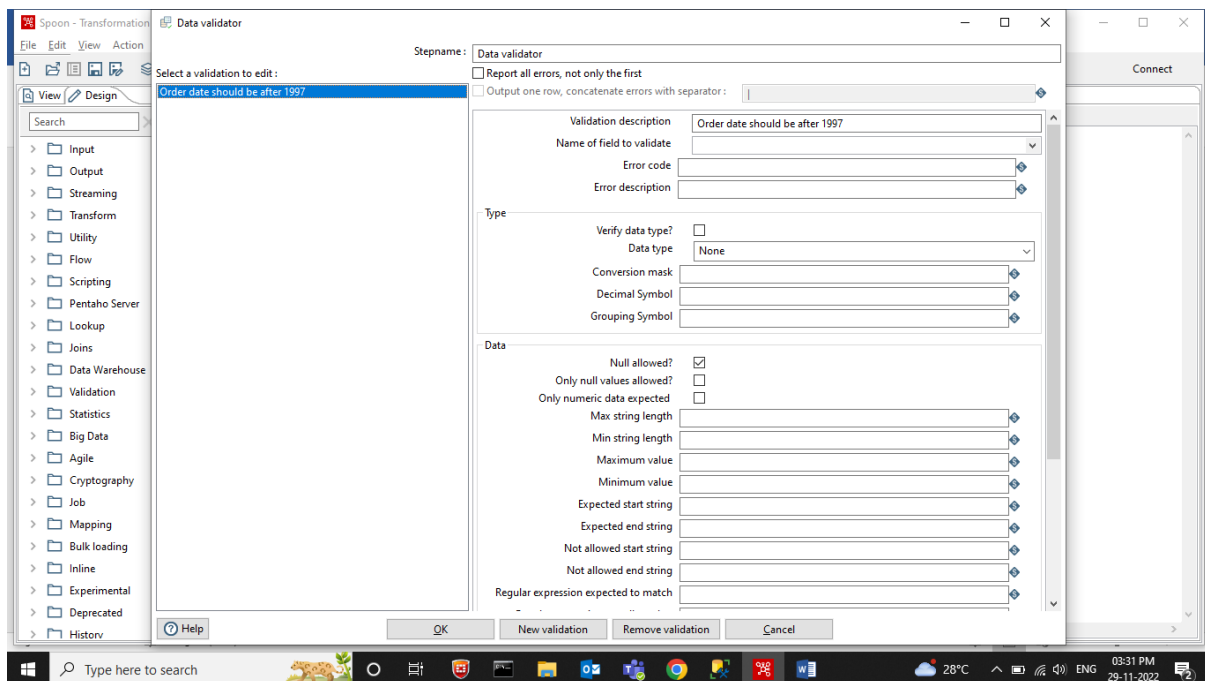
Ex : “Order date should be after 1997”

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:



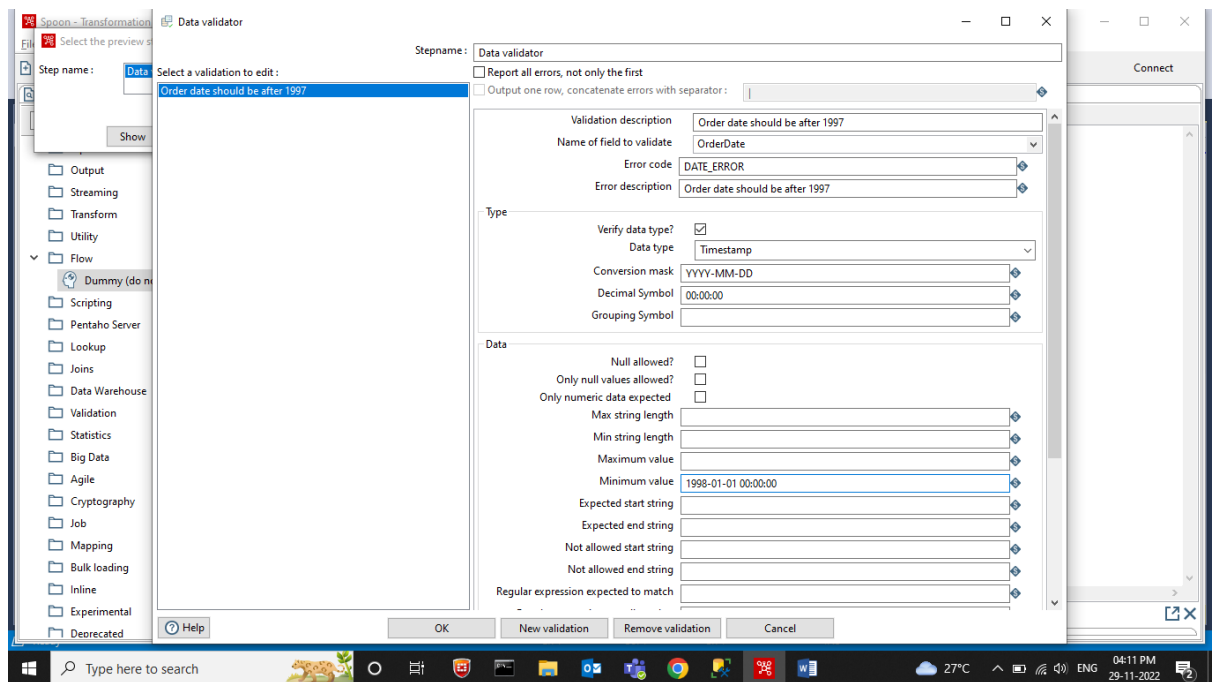
Step 9 : Click on the newly added validation name, and you will be able to see the fields to enter some details like below :



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 :

Examine preview data

Rows of step: Data validator (270 rows)

#	OrderID	CustomerID	EmployeeID	OrderDate	RequiredDate	ShippedDate	ShipVia	Freight	ShipName
242	11049	GOURL	3	1998/04/24 00:00:00.0000000000	1998/05/22 00:00:00.0000000000	1998/05/04 00:00:00.0000000000	1	8.34	Gourmet Lanchonetes
243	11050	FOLKO	8	1998/04/27 00:00:00.0000000000	1998/05/25 00:00:00.0000000000	1998/05/05 00:00:00.0000000000	2	59.41	Folk och fa H8
244	11051	LAMAI	7	1998/04/27 00:00:00.0000000000	1998/05/25 00:00:00.0000000000	<null>	3	2.79	La maison d'Asie
245	11052	HANAR	3	1998/04/27 00:00:00.0000000000	1998/05/25 00:00:00.0000000000	1998/05/01 00:00:00.0000000000	1	67.26	Hanan Carnes
246	11053	PICCO	2	1998/04/27 00:00:00.0000000000	1998/05/25 00:00:00.0000000000	1998/04/29 00:00:00.0000000000	2	53.05	Piccolo und mehr
247	11054	CACTU	8	1998/04/28 00:00:00.0000000000	1998/05/26 00:00:00.0000000000	<null>	1	0.33	Cactus Comidas para llevar
248	11055	HILAA	7	1998/04/28 00:00:00.0000000000	1998/05/26 00:00:00.0000000000	1998/05/05 00:00:00.0000000000	2	120.92	HILARION-Abastos
249	11056	EASTC	8	1998/04/28 00:00:00.0000000000	1998/05/12 00:00:00.0000000000	1998/05/01 00:00:00.0000000000	2	278.96	Eastern Connection
250	11057	NORTS	3	1998/04/29 00:00:00.0000000000	1998/05/27 00:00:00.0000000000	1998/05/01 00:00:00.0000000000	3	4.13	North/South
251	11058	BLAUS	9	1998/04/29 00:00:00.0000000000	1998/05/27 00:00:00.0000000000	<null>	3	31.14	Blauer See Delikatessen
252	11059	RICAR	2	1998/04/29 00:00:00.0000000000	1998/06/10 00:00:00.0000000000	<null>	2	85.8	Ricardo Adocicados
253	11060	FRANS	2	1998/04/30 00:00:00.0000000000	1998/05/28 00:00:00.0000000000	1998/05/04 00:00:00.0000000000	2	10.98	Franchi S.p.A.
254	11061	GREAL	4	1998/04/30 00:00:00.0000000000	1998/06/11 00:00:00.0000000000	<null>	3	14.01	Great Lakes Food Market
255	11062	REGGC	4	1998/04/30 00:00:00.0000000000	1998/05/28 00:00:00.0000000000	<null>	2	29.93	Reggiani Caseifici
256	11063	HUNGO	3	1998/04/30 00:00:00.0000000000	1998/05/28 00:00:00.0000000000	1998/05/06 00:00:00.0000000000	2	81.73	Hungry Owl All-Night Grocers
257	11064	SAVEA	1	1998/05/01 00:00:00.0000000000	1998/05/29 00:00:00.0000000000	1998/05/04 00:00:00.0000000000	1	30.09	Save-a-lot Markets
258	11065	LILAS	8	1998/05/01 00:00:00.0000000000	1998/05/29 00:00:00.0000000000	<null>	1	12.91	LILA-Supermercado
259	11066	WHITC	7	1998/05/01 00:00:00.0000000000	1998/05/29 00:00:00.0000000000	1998/05/04 00:00:00.0000000000	2	44.72	White Clover Markets
260	11067	DRACD	1	1998/05/04 00:00:00.0000000000	1998/05/18 00:00:00.0000000000	1998/05/06 00:00:00.0000000000	2	7.98	Drachenblut Delikatessen
261	11068	QUEEN	8	1998/05/04 00:00:00.0000000000	1998/06/01 00:00:00.0000000000	<null>	2	81.75	Queen Cozinha
262	11069	TORTU	1	1998/05/04 00:00:00.0000000000	1998/06/01 00:00:00.0000000000	1998/05/06 00:00:00.0000000000	2	15.67	Tortuga Restaurante
263	11070	LEHMS	2	1998/05/05 00:00:00.0000000000	1998/06/02 00:00:00.0000000000	<null>	1	136.0	Lehmanns Marktstand
264	11071	LILAS	1	1998/05/05 00:00:00.0000000000	1998/06/02 00:00:00.0000000000	<null>	1	0.93	LILA-Supermercado
265	11072	ERNSH	4	1998/05/05 00:00:00.0000000000	1998/06/02 00:00:00.0000000000	<null>	2	256.64	Ernst Handel
266	11073	PERIC	2	1998/05/05 00:00:00.0000000000	1998/06/02 00:00:00.0000000000	<null>	2	24.95	Pericles Comidas clásicas
267	11074	SIMOB	7	1998/05/06 00:00:00.0000000000	1998/06/03 00:00:00.0000000000	<null>	2	18.44	Simons bistro
268	11075	RICSU	8	1998/05/06 00:00:00.0000000000	1998/06/03 00:00:00.0000000000	<null>	2	6.19	Richter Supermarkt
269	11076	BONAP	4	1998/05/06 00:00:00.0000000000	1998/06/03 00:00:00.0000000000	<null>	2	38.28	Bon app'
270	11077	RATTC	1	1998/05/06 00:00:00.0000000000	1998/06/03 00:00:00.0000000000	<null>	2	8.53	Rattlesnake Canyon Grocery

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