**Babu Banarasi Das**

**University**



Predictive Analytics

**(BCADS15301)**

**LAB FILE**

**SUBMITTED TO: SUBMITTED BY:-**

**Mr. AYUSHMAN BHADURIA Name: Udit Agrahari**

**Roll No: 1230258453**

**Class: BCADS36**

**SPSS Modeler Practical Workbook: Data Integration and Analysis**

**Definition:** You work for a telecommunications firm and have to combine a number of datasets into a single dataset for analyses and modeling later.

**Outcomes/Learning:** Learning how to join different data sets

How to join records of different data sets

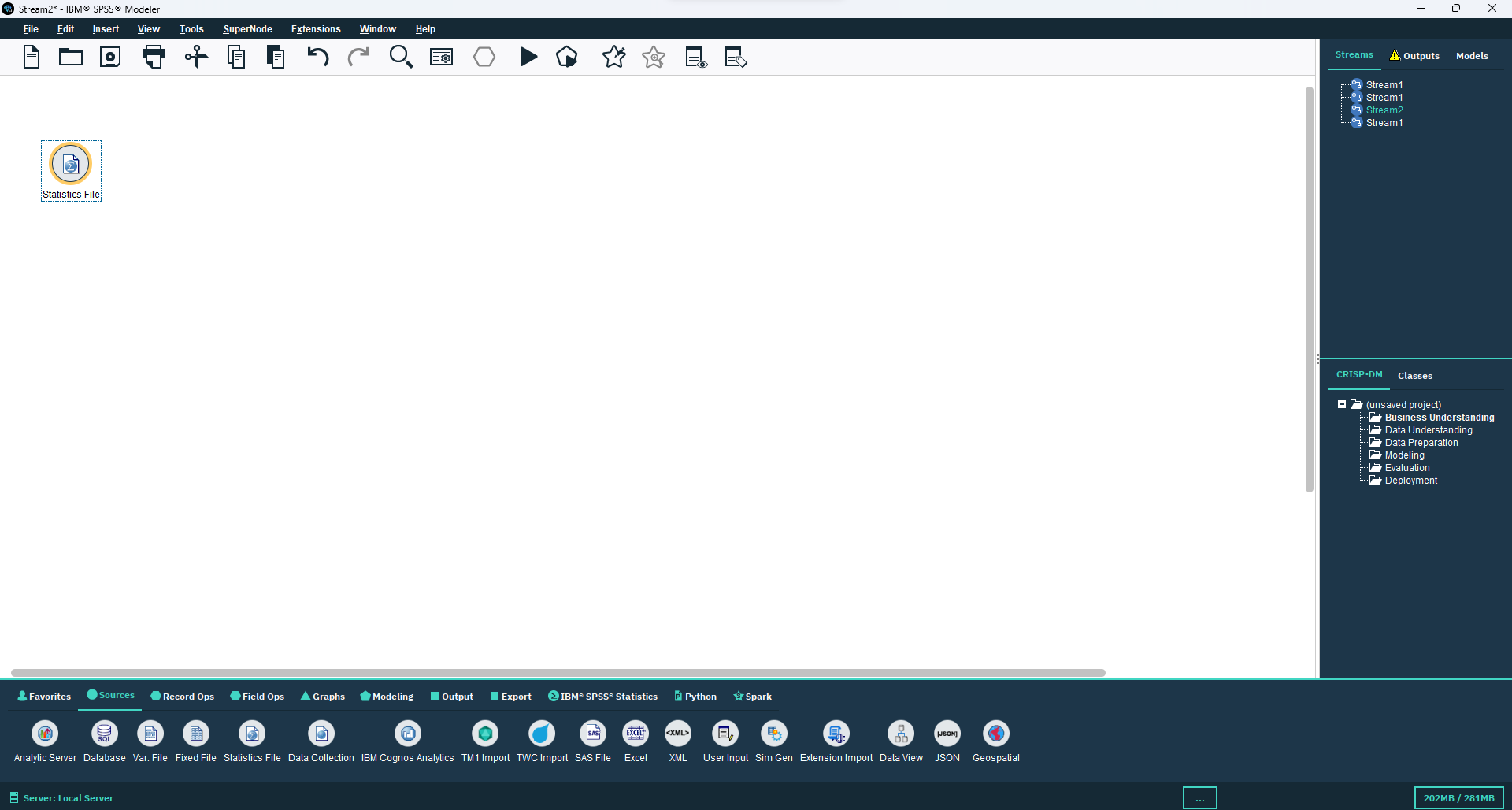
How to sample data from data sets

**Required Tool:**  IBM SPSS Modeler Tool

**Working:** Using Merge, Append, Sample nodes to merge two data sets, to append records from two or more data sets, To sample data from whole data sets.

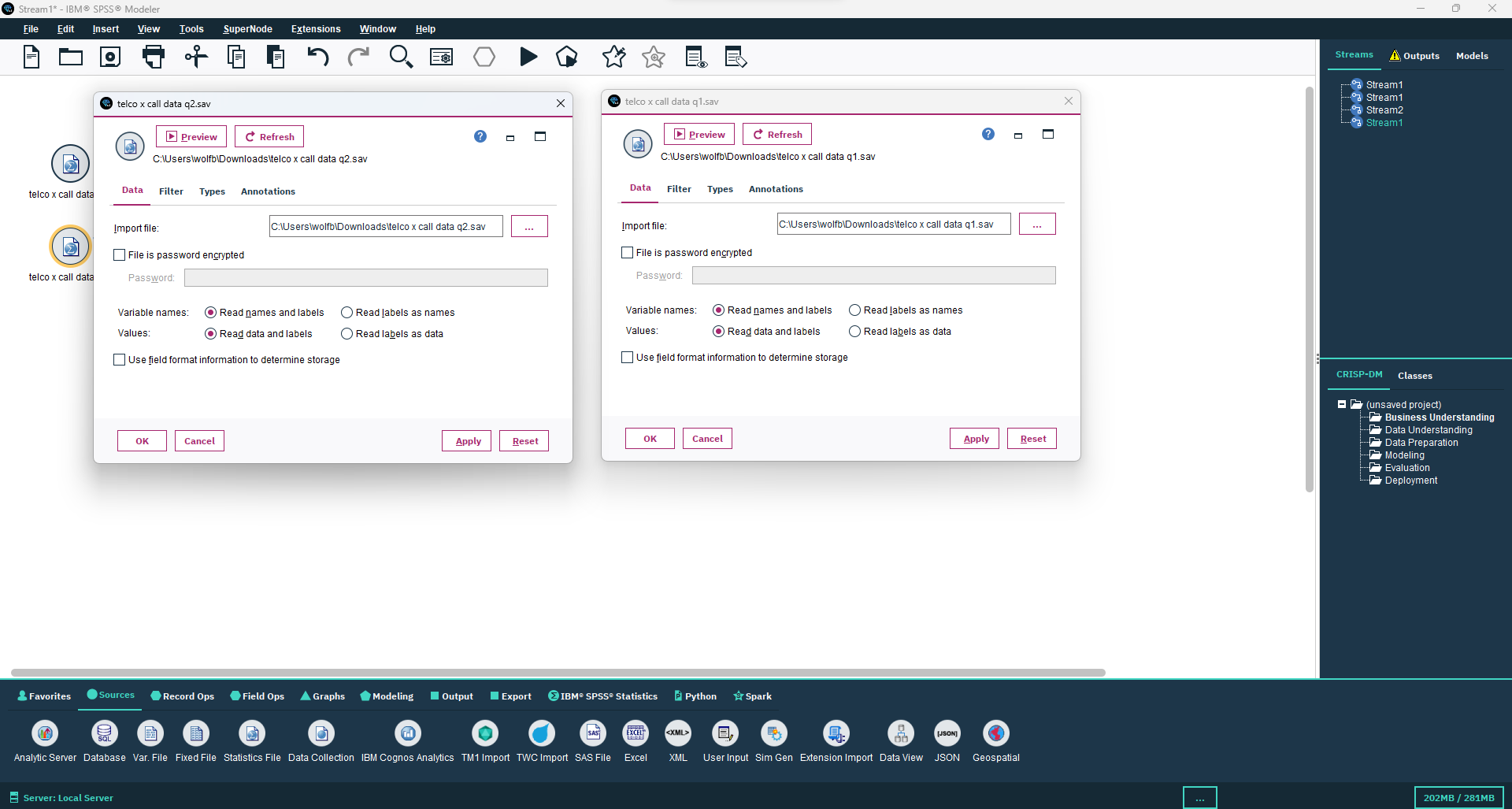
**Step 1:** Open SPSS Modeler tool then on Source category select statistics File node (we are selecting Statistics node because the data set we are using is an sav file.)

Double clicking on statistics node will make it appear on canvas

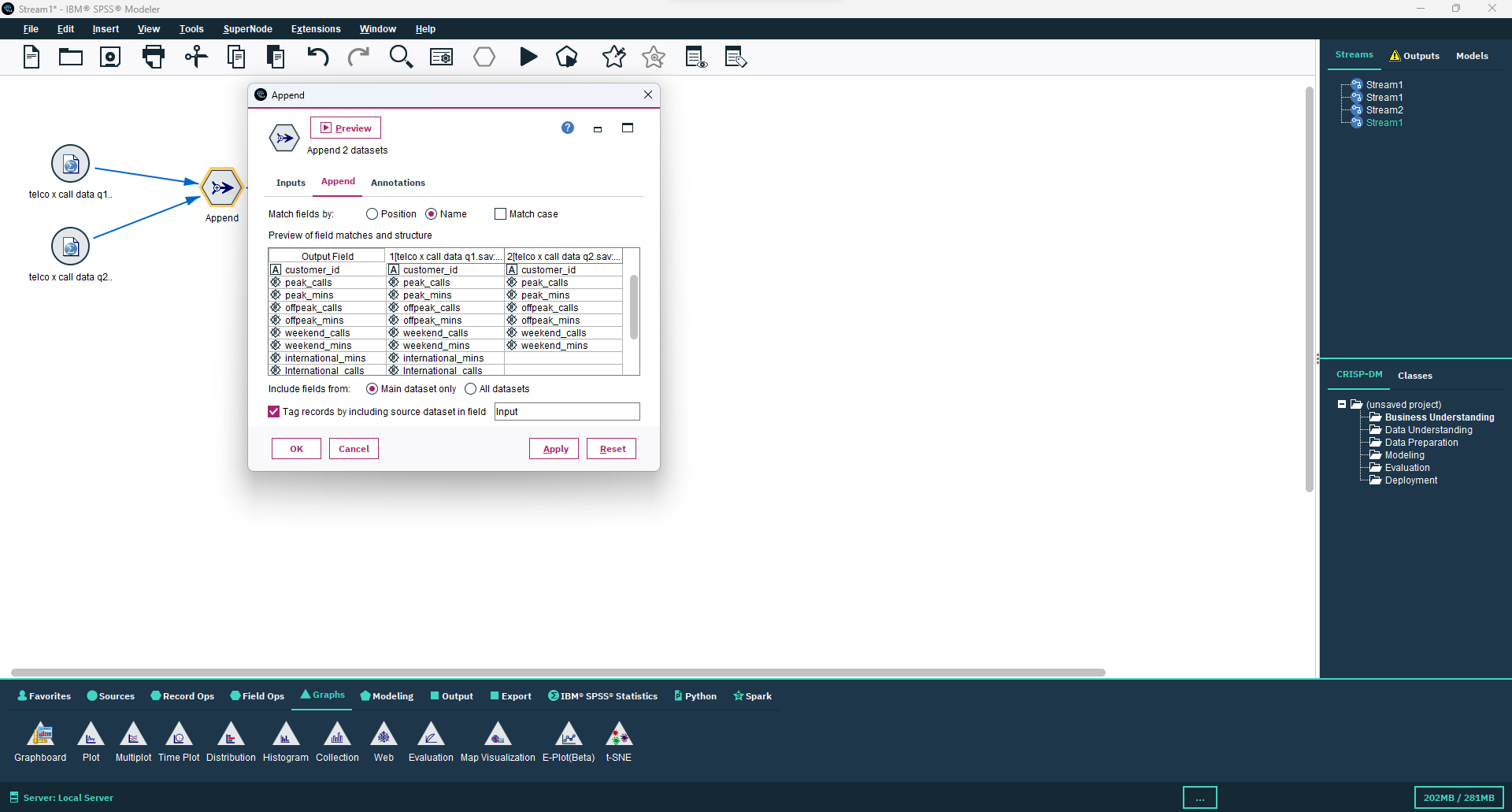
****

**Step 2:** Now we import a data set using the import option which can be accessed by double clicking on Statistics icon on the canvas.

We import a data set telco x call data q1.sav then import one more statistics file tele x call data q2.sav.

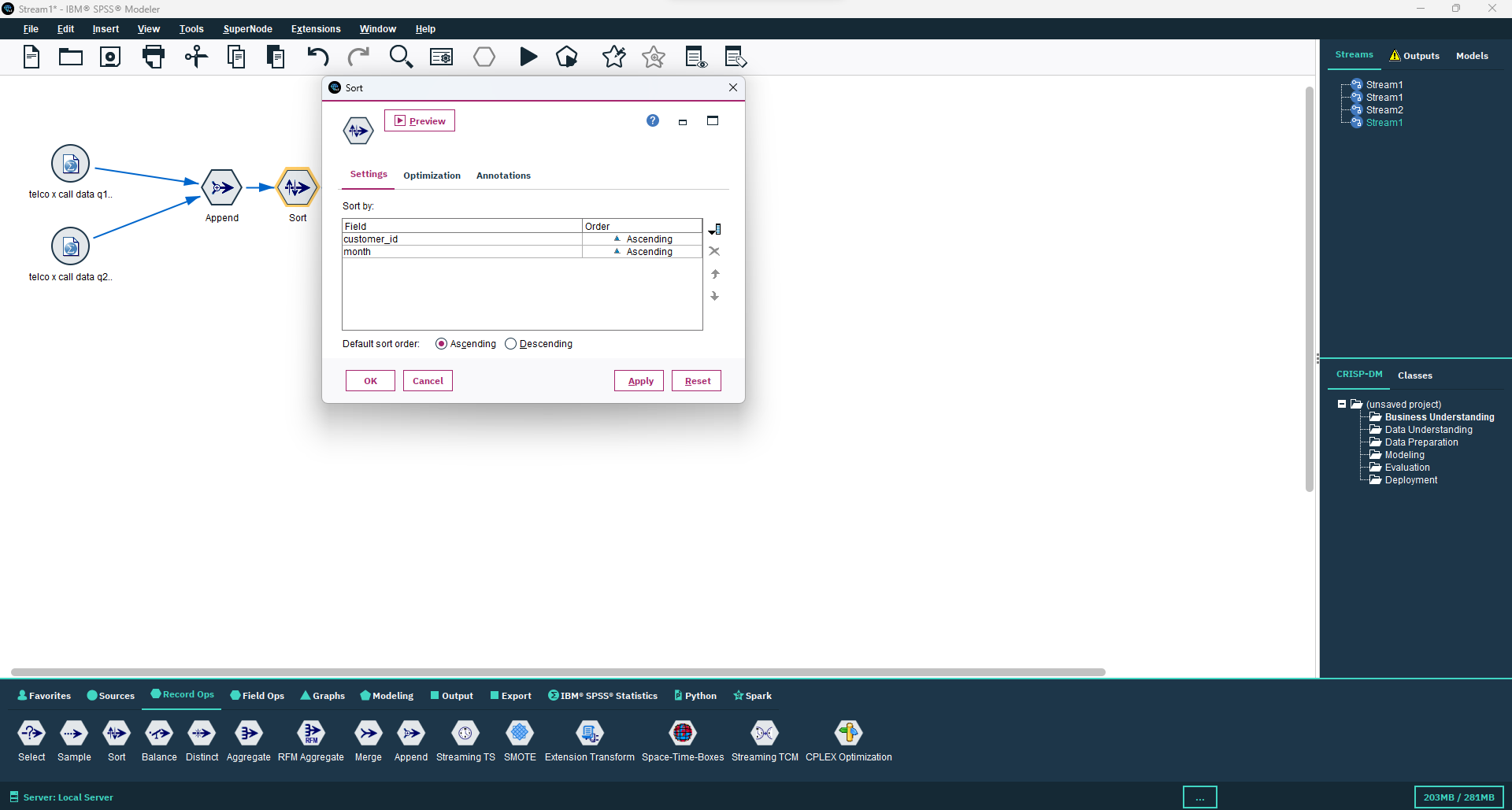
****

**Step 3:** We connect both of these files to Append node from Records Ops. click on apply and ok.



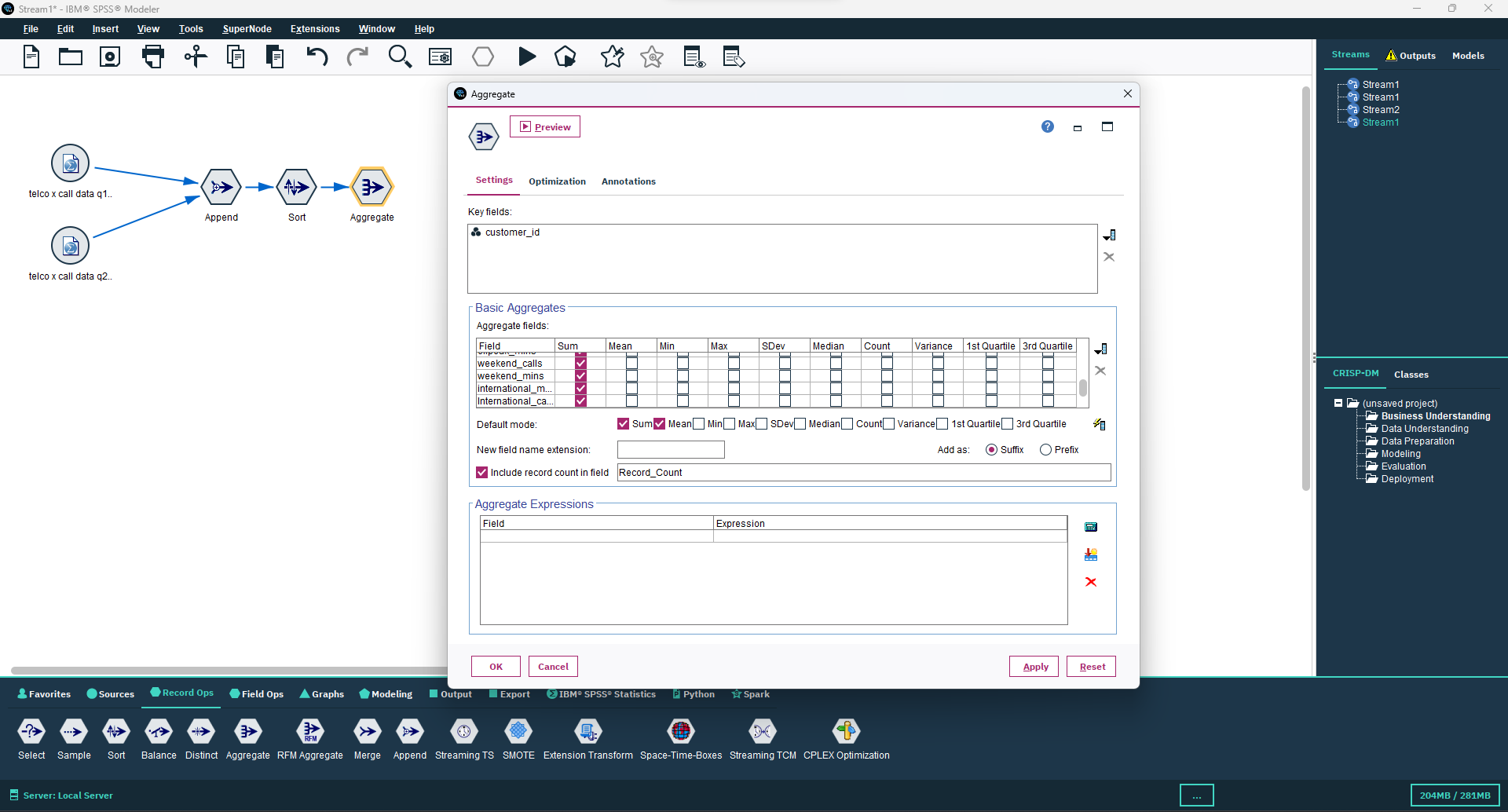
**Step 4:** Now we add sort node from Record Ops and connect it to Append node.

By double clicking on sort node a window opens in which under Sort by we select customer\_id and month fields and sort them in ascending order.



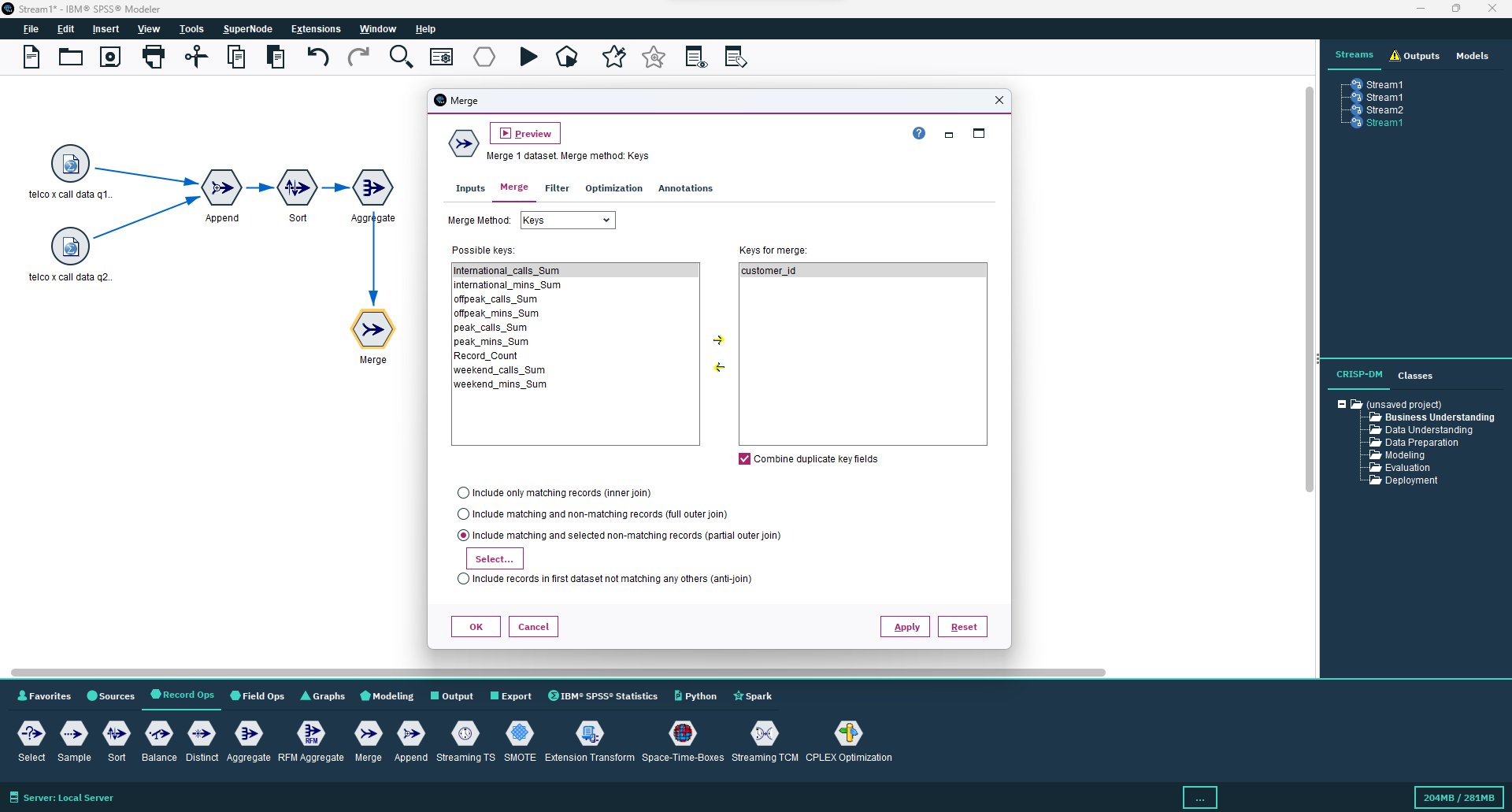
**Step 5:** We add Aggregate node from Record ops and connect it to sort node.

We customer\_id as Key Field and all other fields should be selected in Aggregate fields and all should have only sum box as checked.

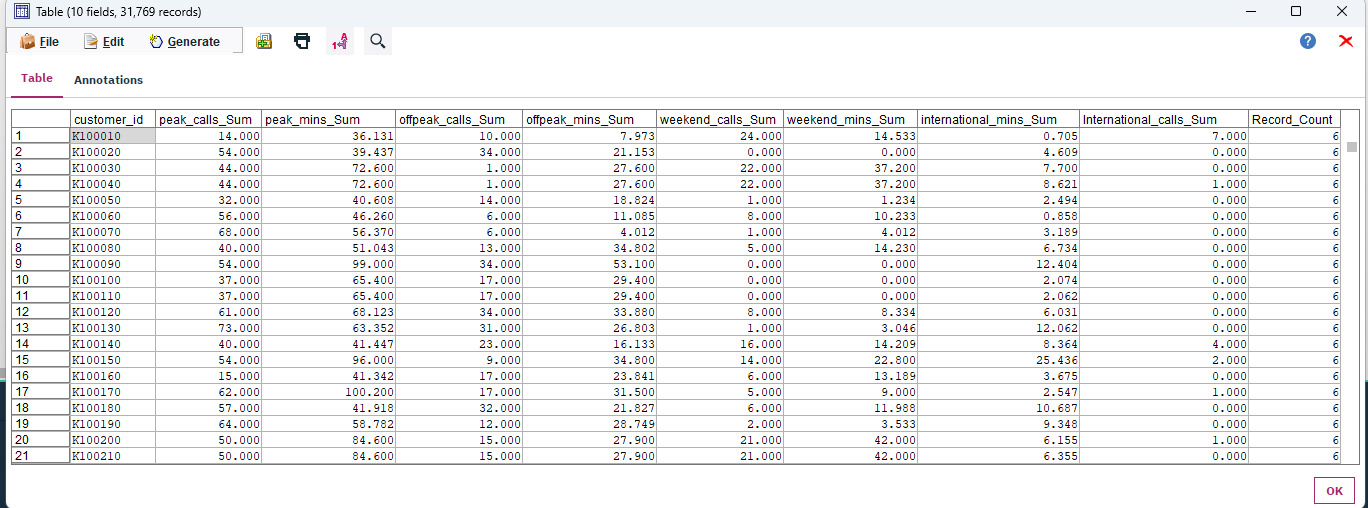


**Step 6:** Now add Merge node from Record ops and connect it to Aggregate node.

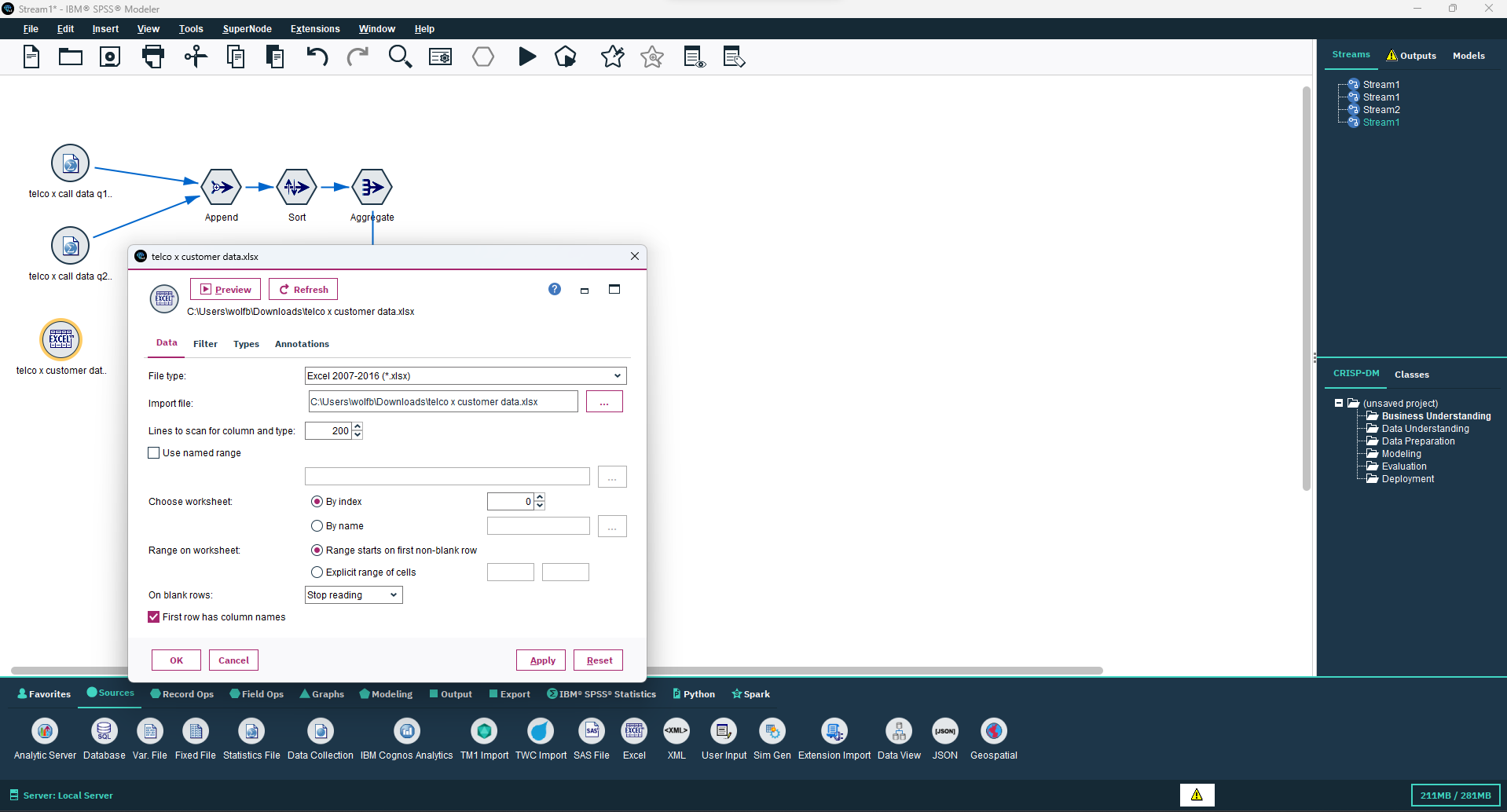
Select customer\_id as the key for merge and partial outer join by double clicking on Merge node.



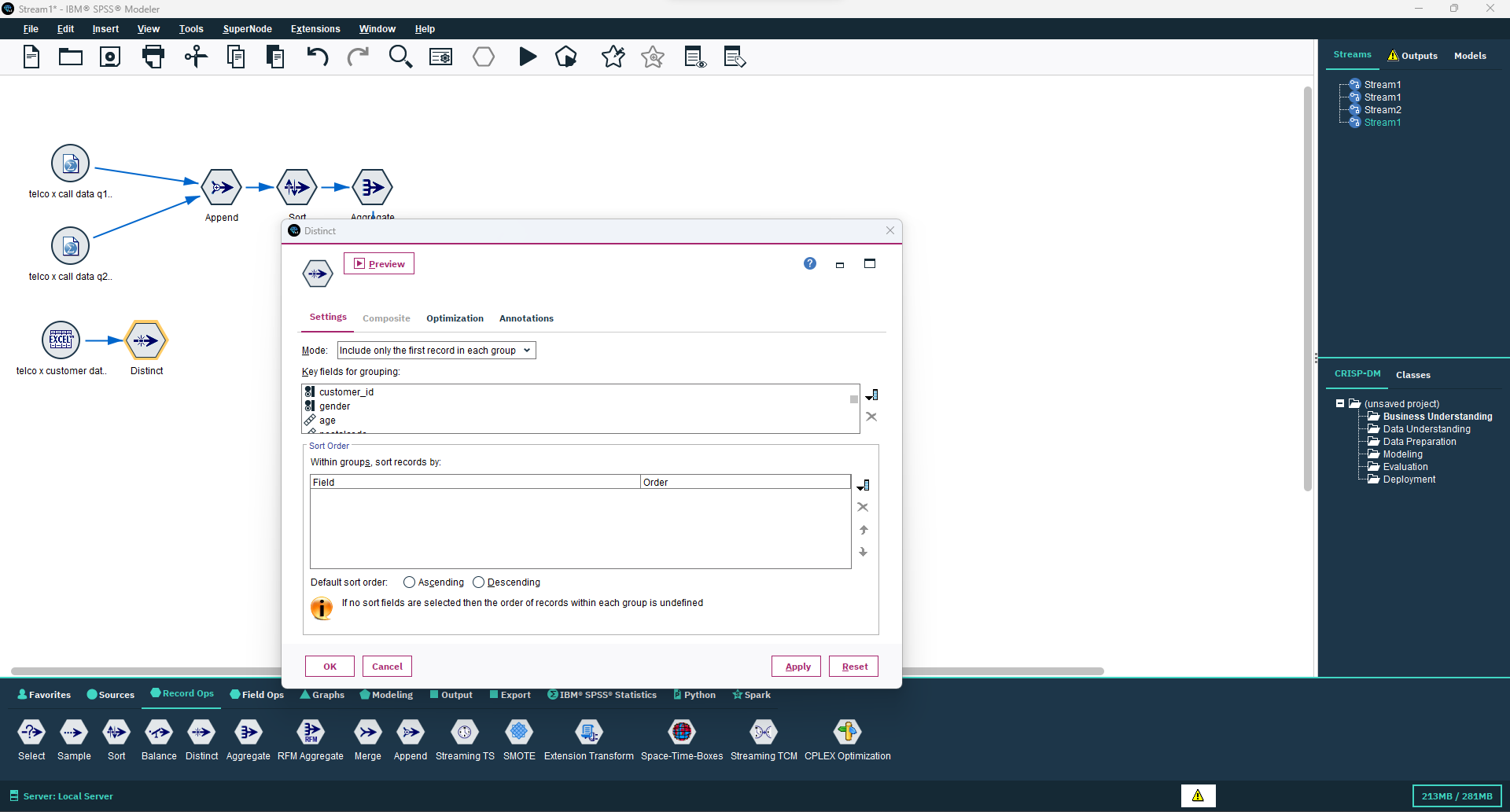
**Step 7:** We connect table node to see any change in data set.



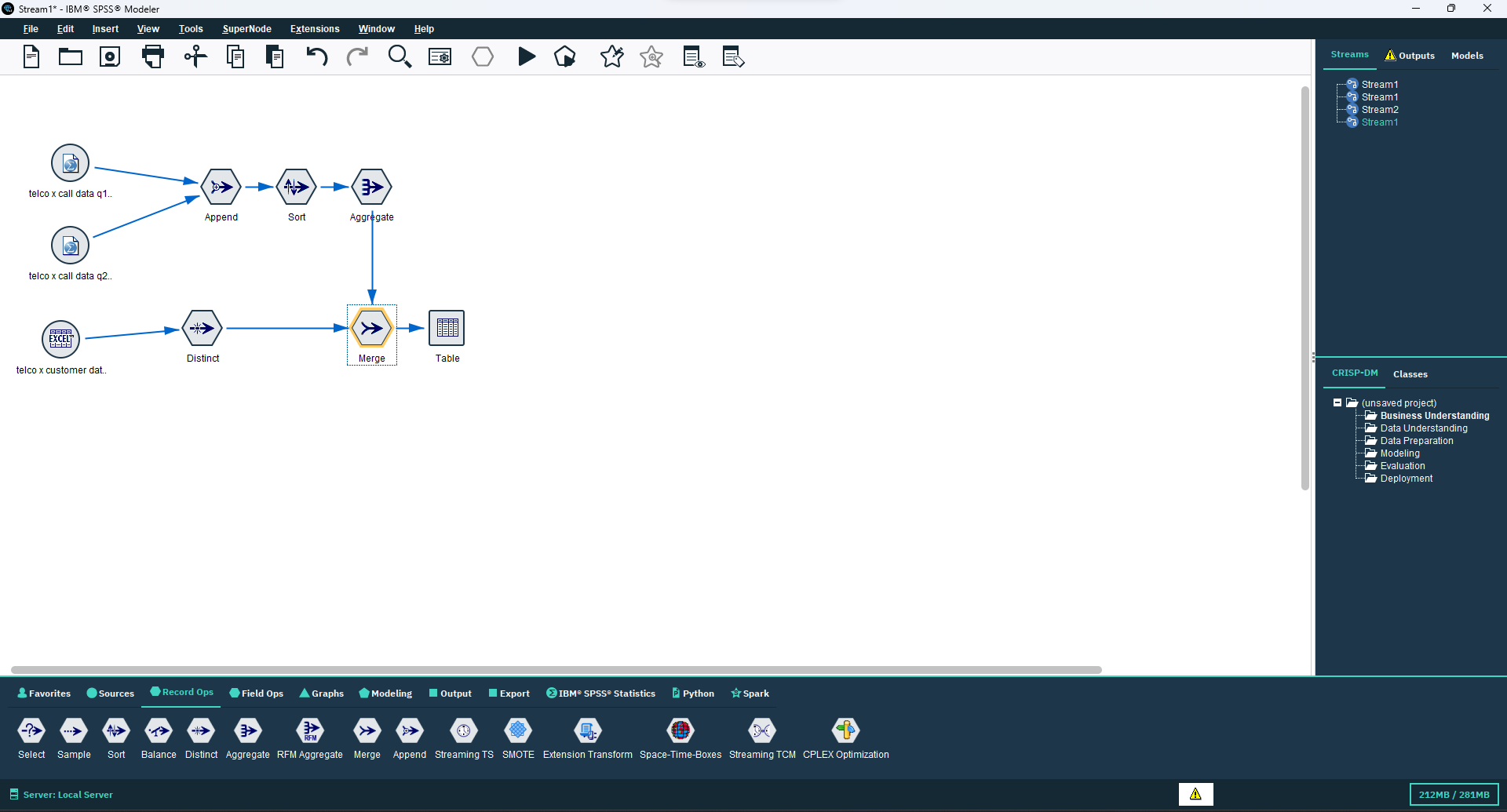
**Step 8:** Now we import a new data set telco x customer data.xlxs which is an excel file.



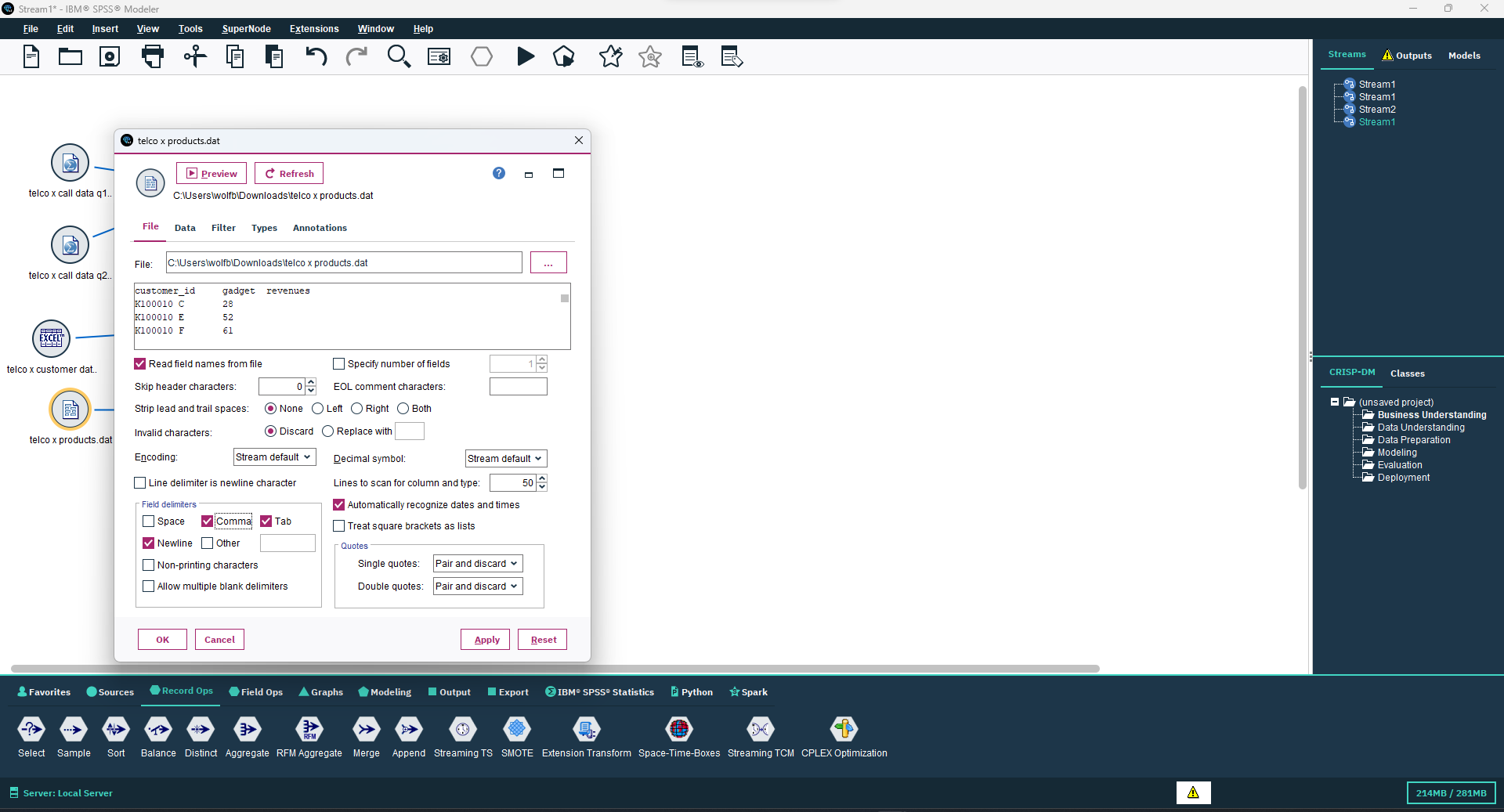
**Step 9:** Connect a Distinct node from Record ops and take all the fields of the data set as key fields in Distinct node by double clicking the Distinct node. And select Include only first record from each group.



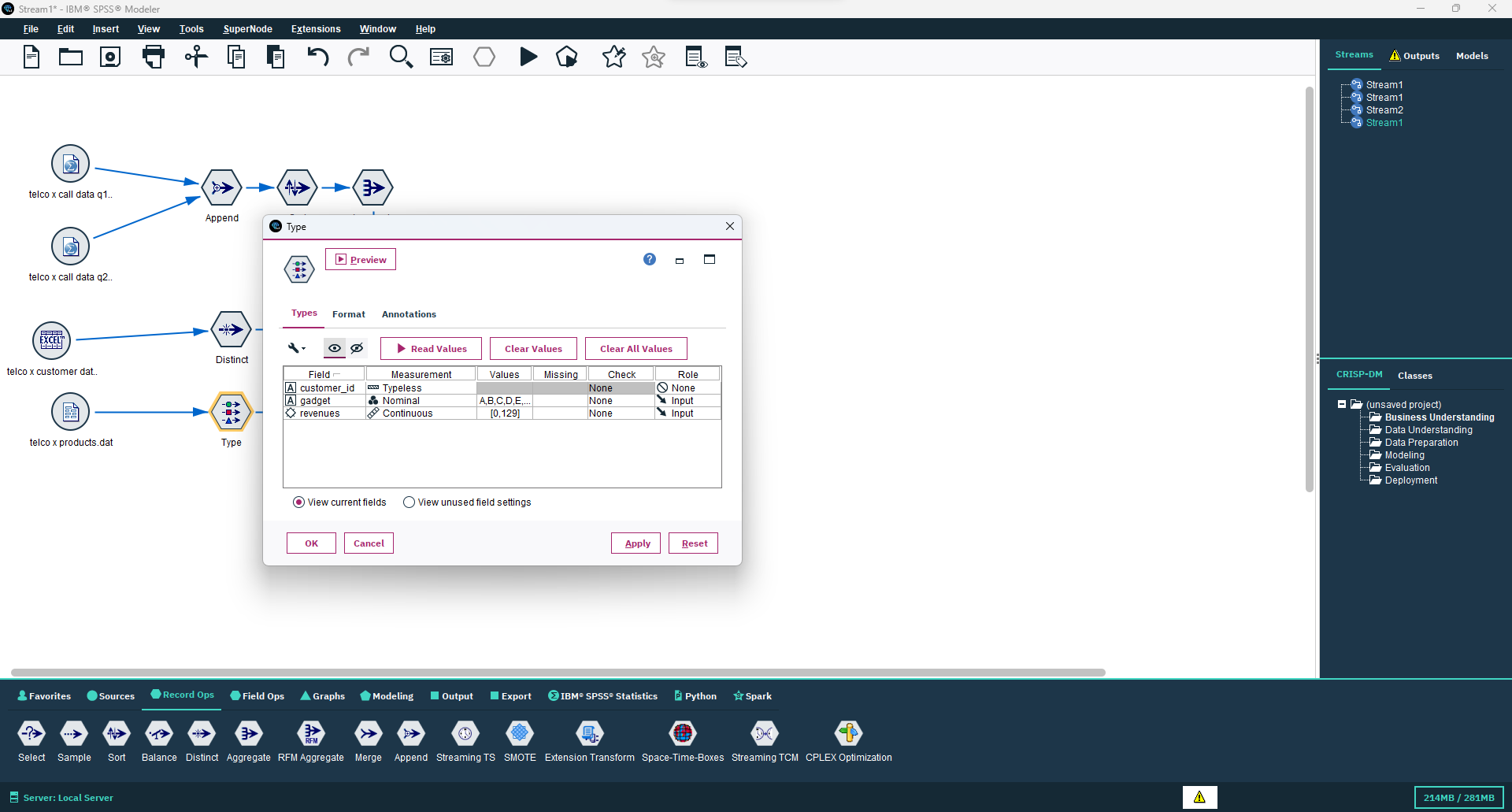
**Step 10:** Connect the Distinct node to Merge node.



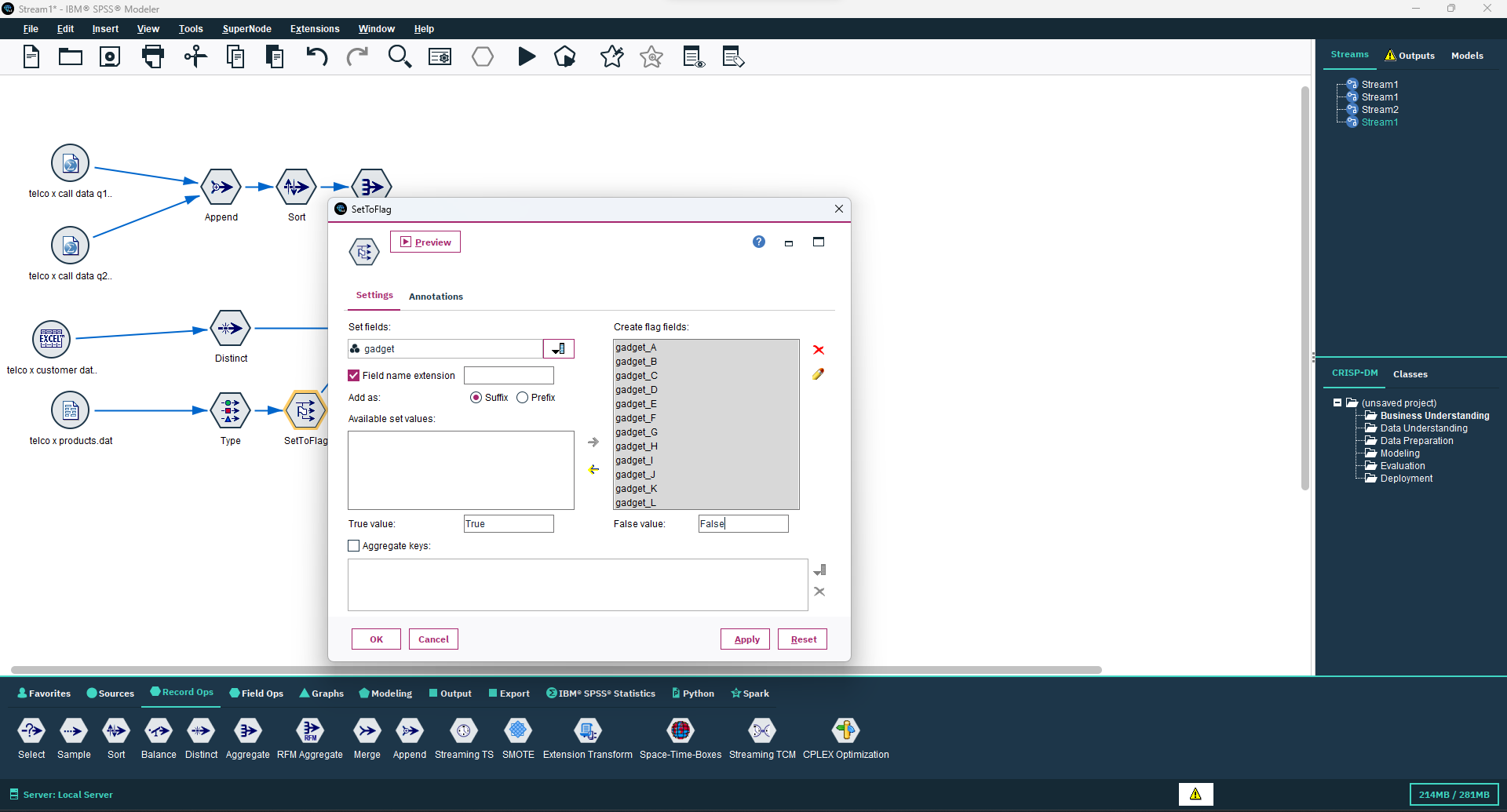
**Step 11:** Take a var file node from Sources Category and import telco x products.dat.



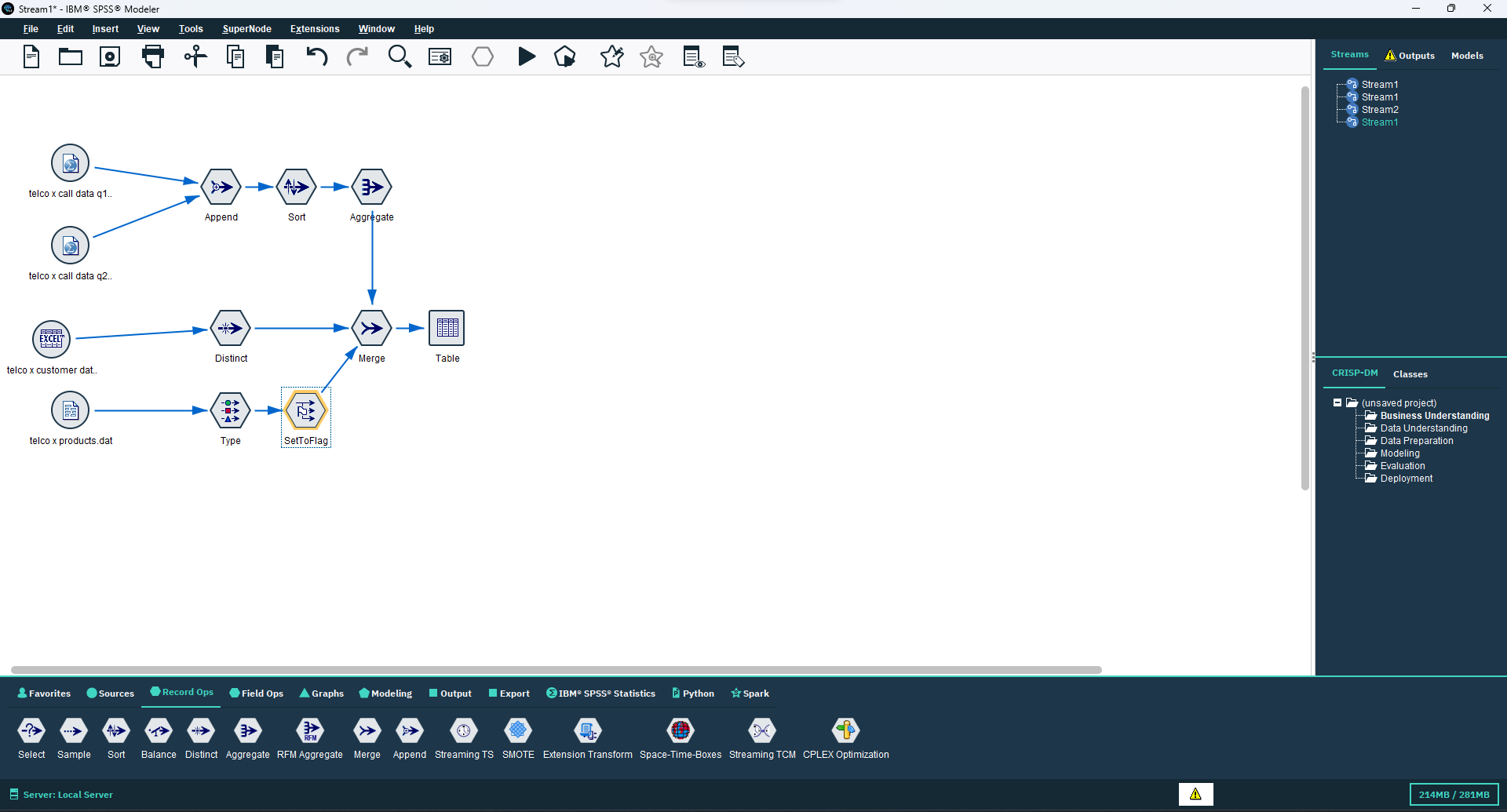
**Step 12:** Connect Type node to the new var file and get the specific category of all values by clicking on Read Values button in Type node window.



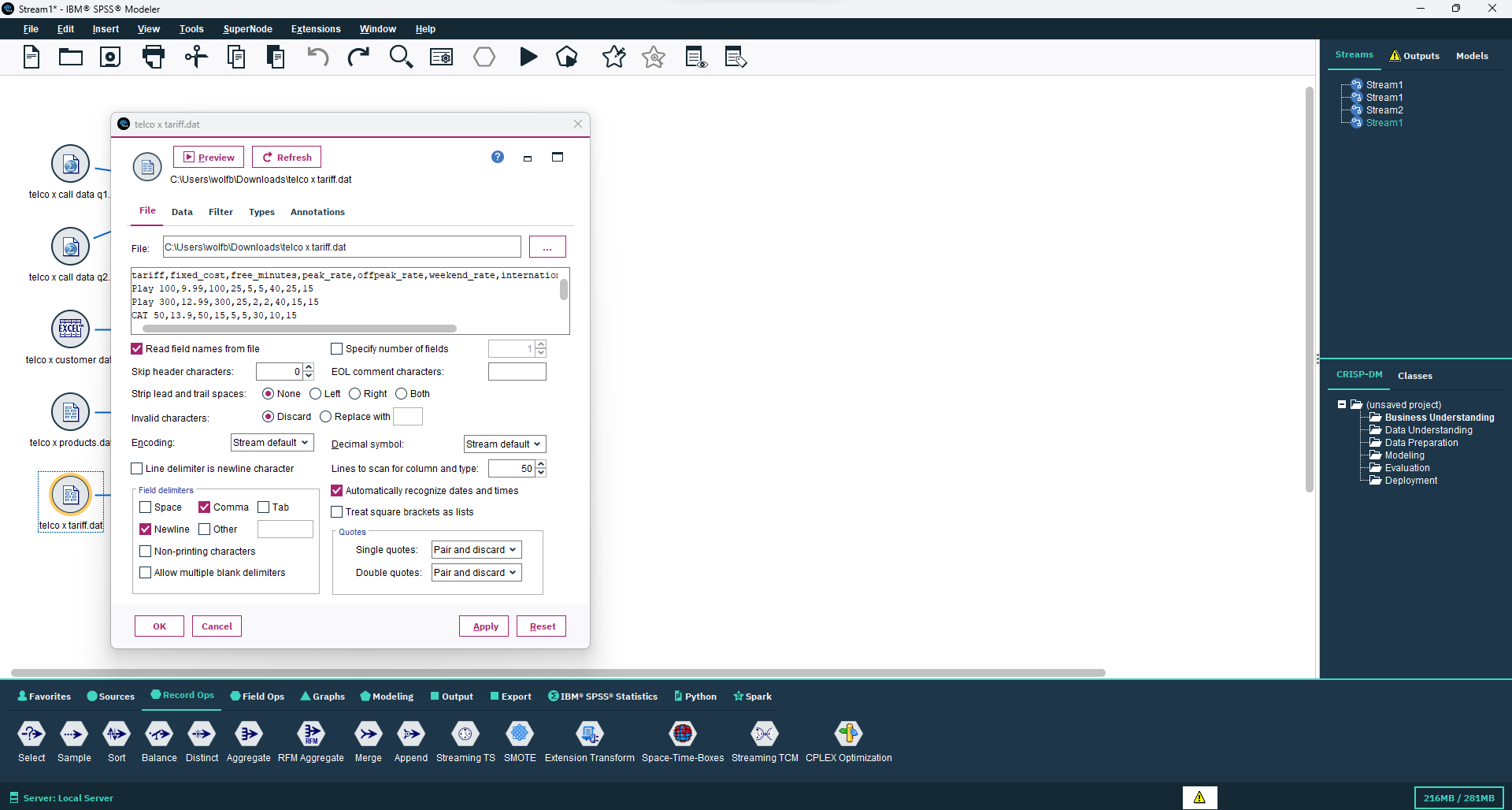
**Step 13:** Now we connect type node to Set to flag node and select gadget in Set fields section in Set to flag window and all the values should be selected and sent to Create Flag Fields.



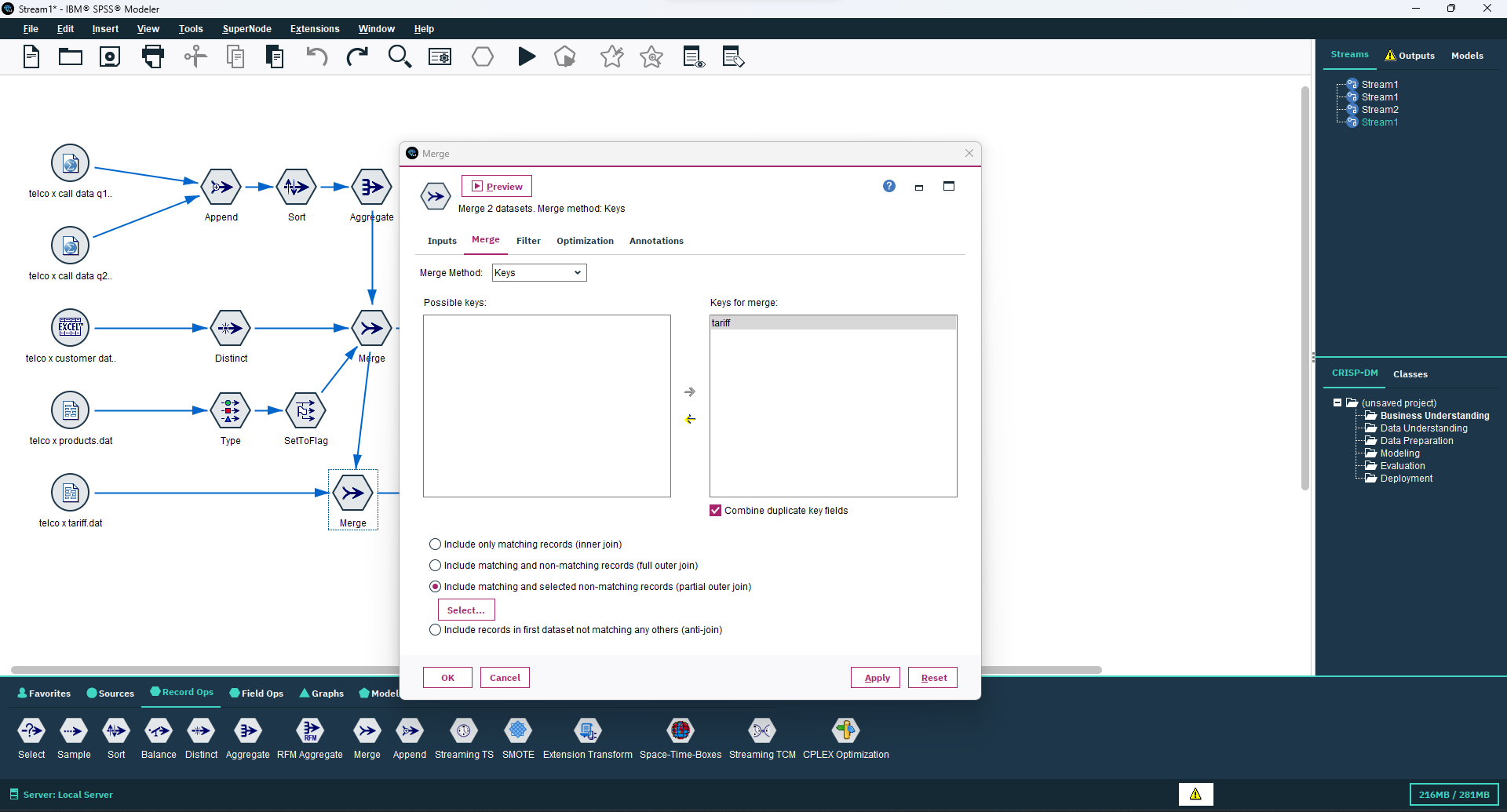
**Step 14:** Connect the Set to Flag node to Merge node.

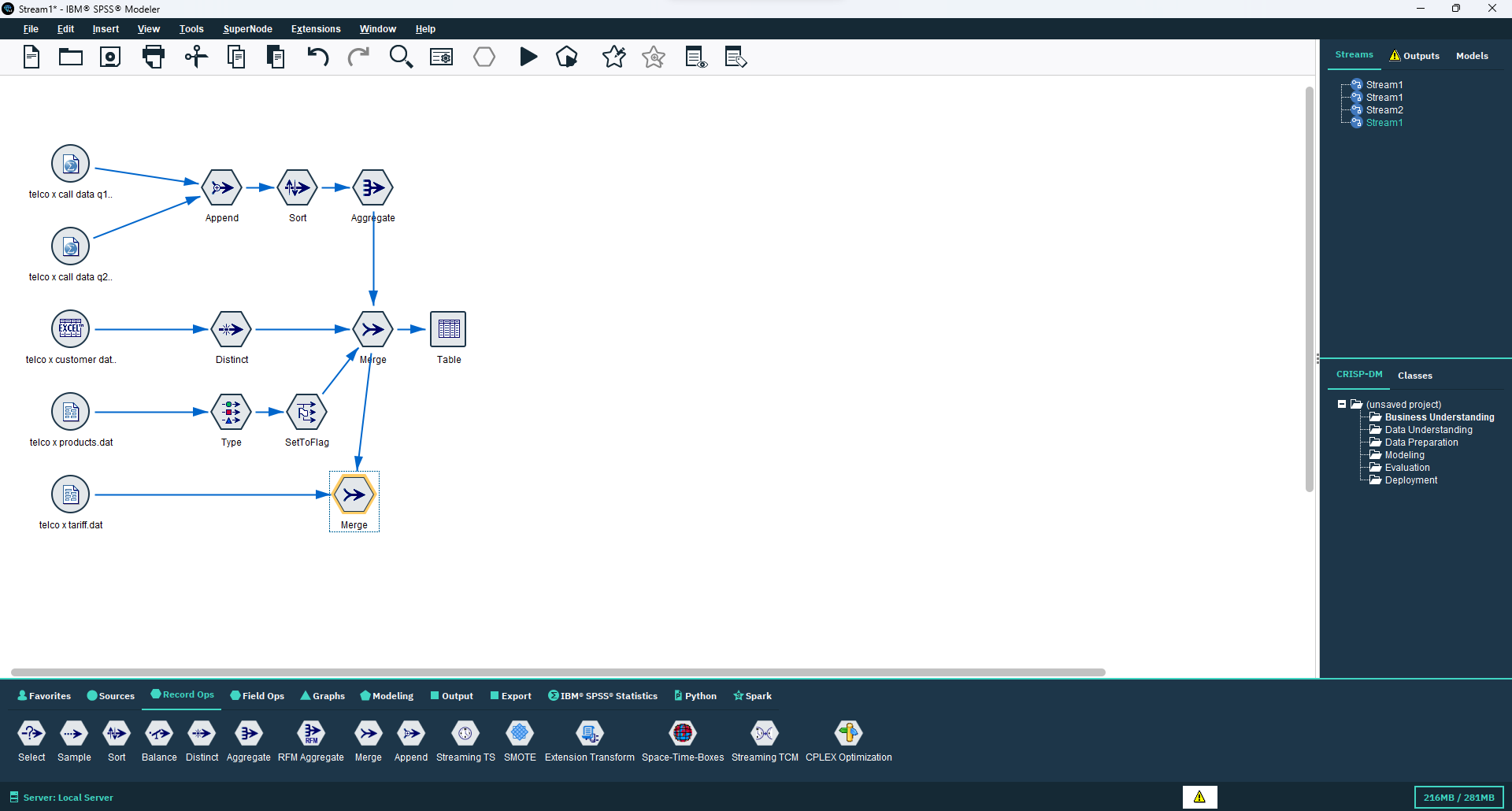


**Step 15:** Now we add another var file telco x tariff.dat



**Step 16:** Connect Merge node to the new var file and take tariff as key field and partial outer join in the Merge node window then connect the previous Merge Node to the new one Which is connected to telco x tariff.dat.



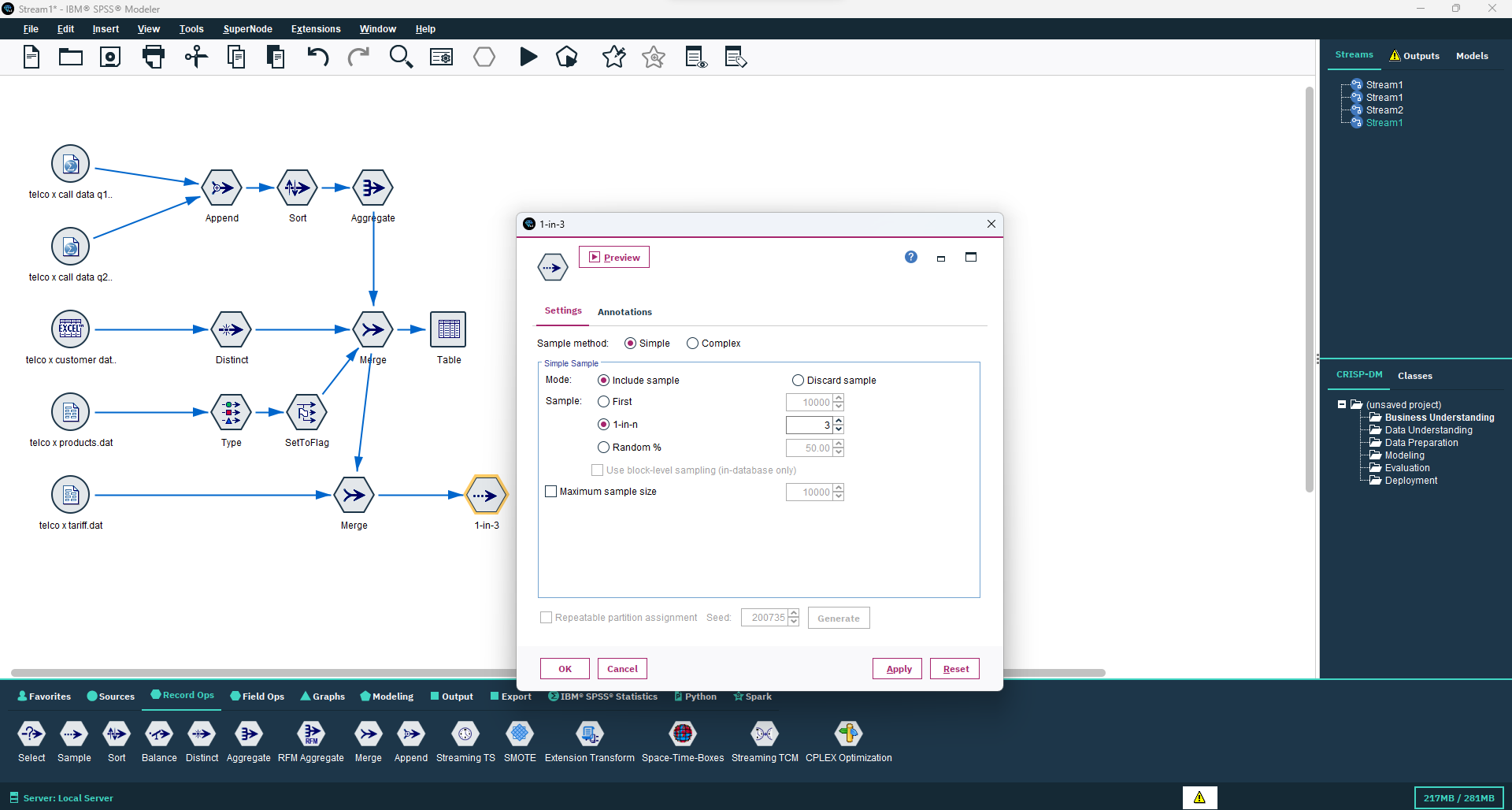


**Step 17:** Then connect the sample node from the Record ops to the Merge node connected to telco x tariff.dat.

Sample node is used to take a part of data from a data set to test the data set.

Open the Sample node window by double clicking on it and now select 1 in n rule now we choose a integer in place of n.

If we choose 3 in place of n the result will show the first record and then the third record skipping the record on the second place.



**Step 18:** Connect Table node to Sample node to view the final result which will show sample data of the four joint data sets.

