**Part 2: Data Preparation Using Talend Data Integration**

Before loading both datasets (D7005AA1_DataA and D7005AA1_DataB) into SAS Enterprise Miner (SAS EM) for analysis, the preprocessing of data is performed on Talend Data Integration (TDI) and Talend Data Preparation (TDP). First of all, the dataset is loaded into Talend Data Integration using the following schema:

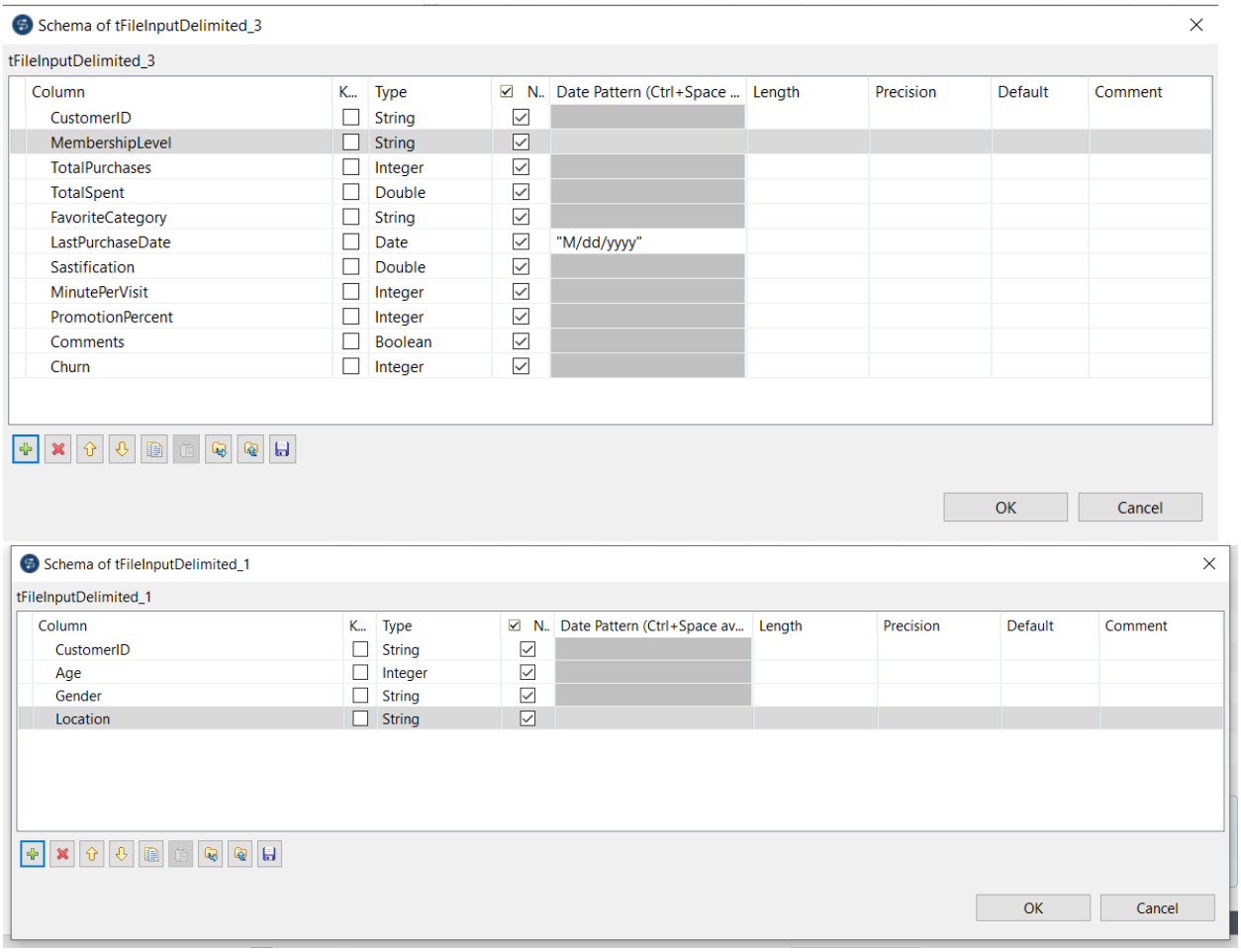


*Figure 2: Schema when loading the dataset into Talend Data Integration.*

Although the CustomerID is made up of 10 to 11-digit integers, however when the data type of CustomerID is set as Integer, some errors exist as the system identifies some of the numbers as strings. As the CustomerID is only used to check for duplicates, which is not involved much in analysis, the string data type is acceptable for Customer ID. The error is shown in Figure 3.

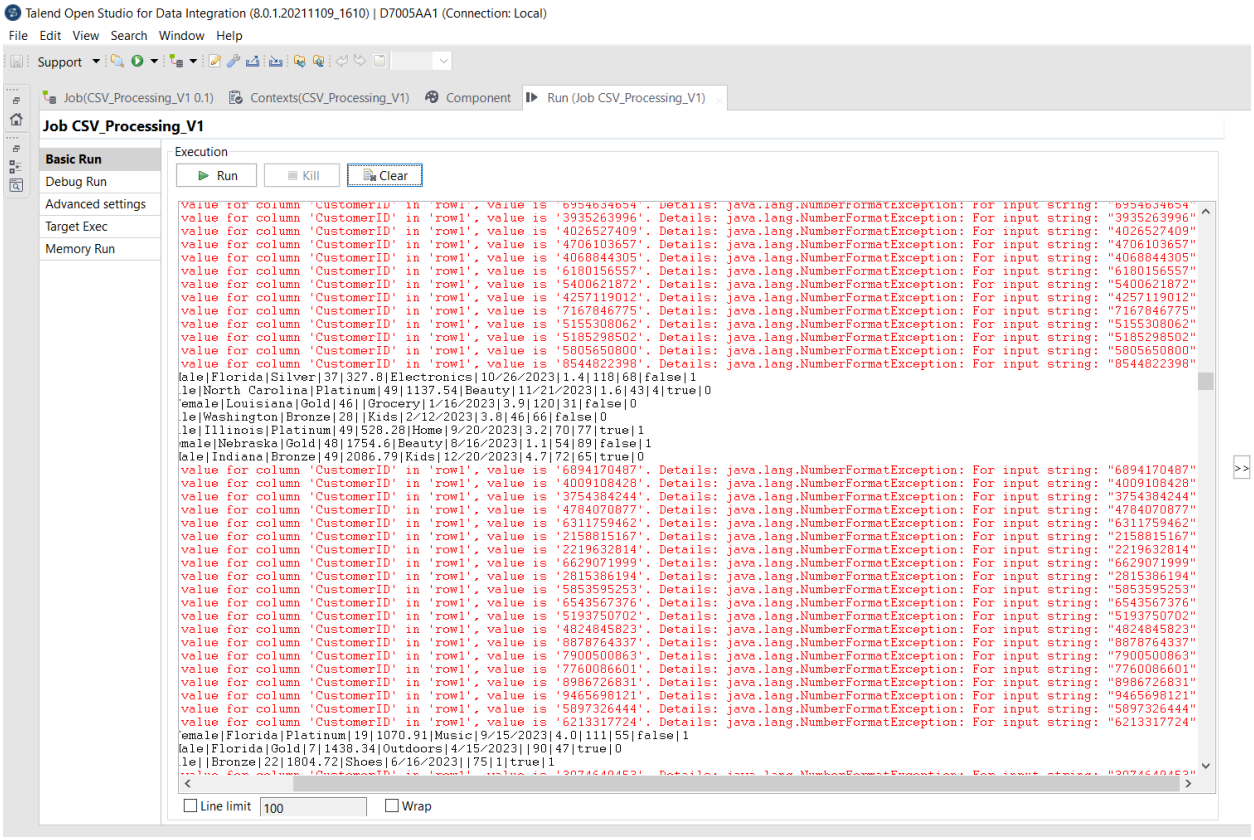


*Figure 3: Error when loading the dataset to TDI when CustomerID type is set as integer.*

First of all, both datasets are related so they needed to be joined together. The tLogRow component provides an output that enables us to determine whether the data under processing is correct. The tMap is used to join both datasets.

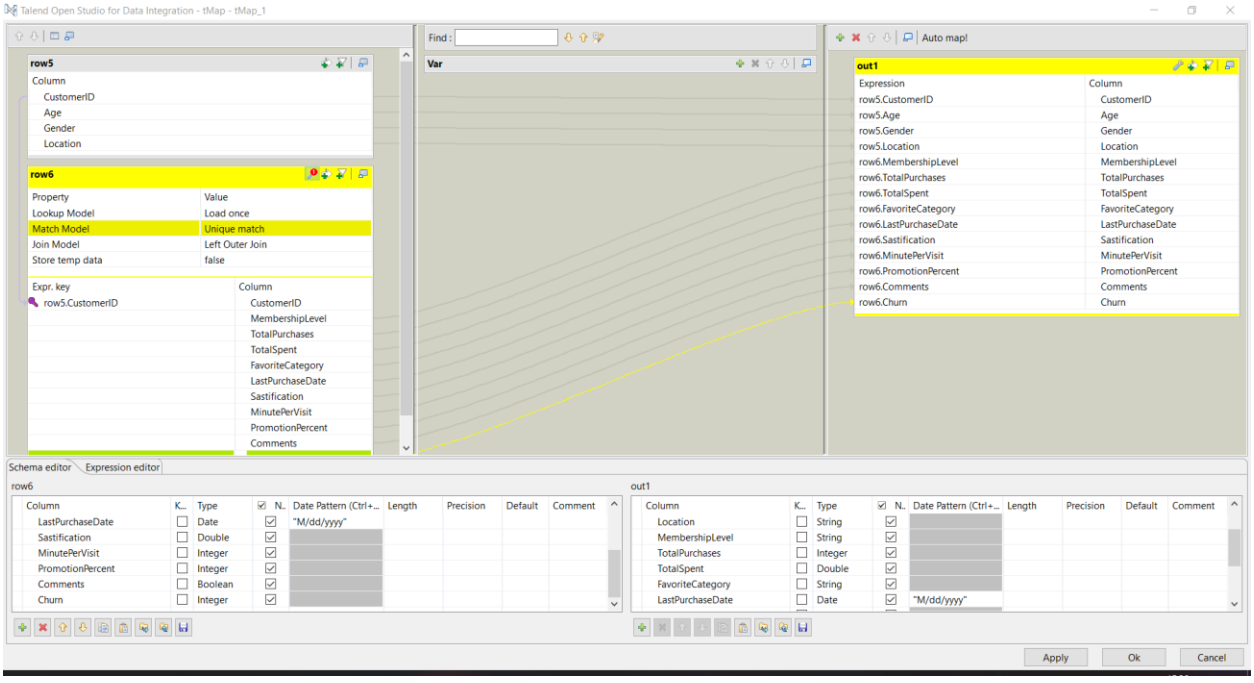

*Figure 4: Setting of tMap.*

Figure 4 shows both datasets are joined together based on the unique CustomerID match. After the output schema is determined, the input column is linked to the desired output column.

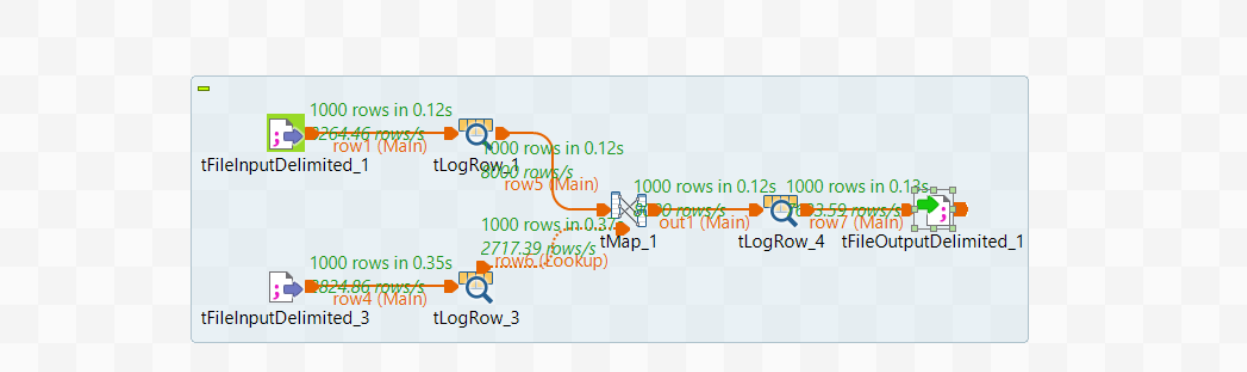


*Figure 5: Workflow of producing combined CSV dataset.*

As shown in Figure 5, after the datasets are joined, the combined dataset is now ready to be exported for further usage. To ensure smoother loading of data in the further step, the settings as shown in Figure 6 are applied.

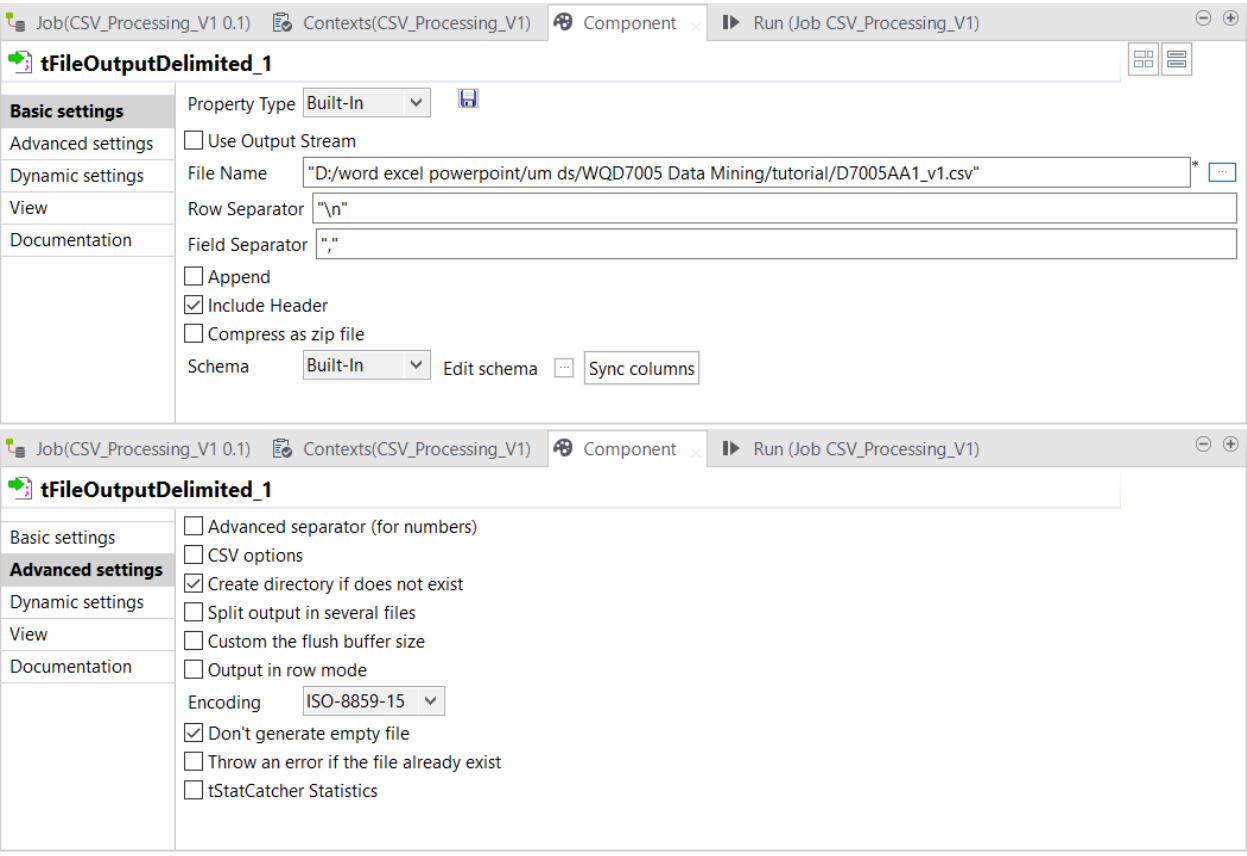


*Figure 6: Setting of export of combined CSV dataset.*

The newly exported dataset is now able to be used. After loading the combined dataset into TDI, the first attempt of handling missing values is operated, which is dropping all the rows with missing values. By using the tFilterRow component with the setting as shown in Figure 7, the remaining number of rows is shown.

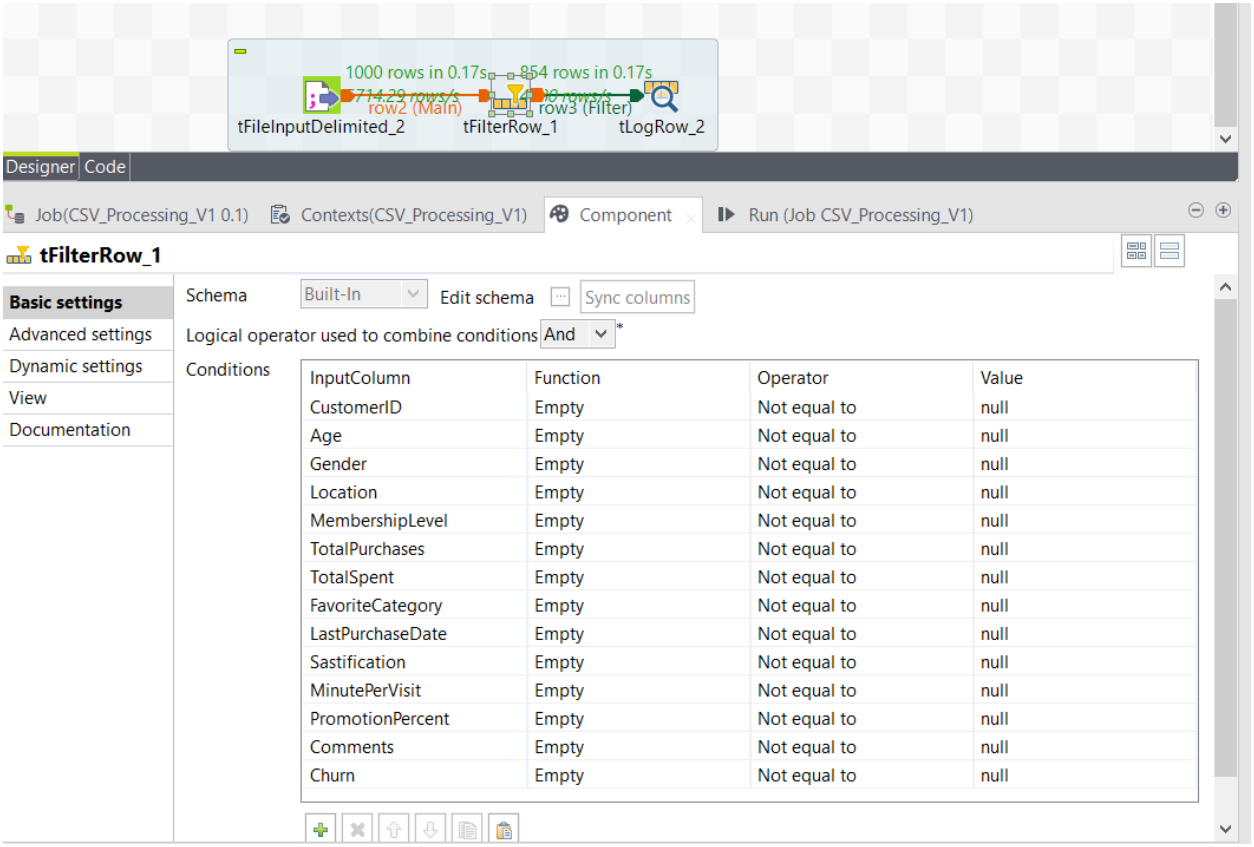


*Figure 7: Filtering out all rows with missing values.*

As shown in the figure above, only 854 rows are remaining in the dataset, which means 146 records (14.6%) are dropped. As the proportion of rows with missing values is not small enough to be removed, data imputation seems to be the better option to handle missing values.

The data preprocessing is now moved to Talend Data Preparation (TDP) as TDP provide an easier yet robust way to process the data.