MIT Academy of Engineering

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

TALEND WORKSHOP

TITLE :- TALEND PROCESSING COMPONENTS

SBMITTED BY:

Akhila Pillai – 83

Lubhavani Singh - 85

tReplace

tReplace component is used to cleanse all files before further processing.
It carries out a Search and Replace operation in the input columns defined.
It performs string substitution on a input column.

In the **Simple Mode** -> Search/Replace :

- **Input column**: Select the column of the schema the search & replace is to be operated on
- **Search**: Type in the value to search in the input column
- **Replace with**: Type in the substitution value.
- **Whole word**: Select this check box if the searched value is to be considered as whole.
- **Case sensitive**: Select this check box to care about the case.

In **Advanced Mode** :
Select this check box when the operation you want to perform cannot be carried out through the simple mode. In the text field, type in the regular expression as required.

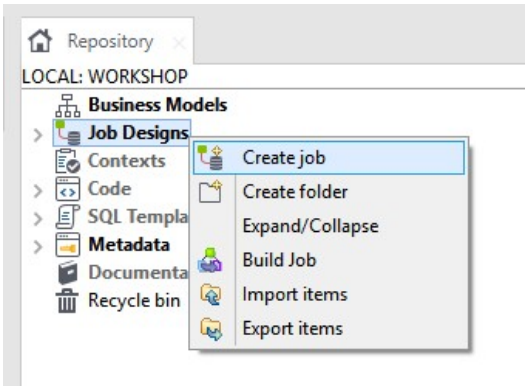
Input Customer File :-

Cust_First_Name,Cust_Address,Cust_City,Cust_Country
Lee,722 Bur Oak Avenue ,Berlin,Germany
William,543 ANJOU ANJOU H1K 2P4,Mexico,Mexico
Nick,40 Toronto M8Z 3Z7,Mexico,Mexico
Jian,11 Collaroy 2093,London,UK
Ming,223 Wellington 5011,Berlin,Germany
Joseph,324 Tman Street Stafford Heights ,Mannheim,Germany
Justin,90 Ruapehu Road Ohakune ,Strasbourg,France
Ming,32 Wilfred road,Madrid,Spain
Lee,Williamstown Road,Marseille,France

Copy the input customer file into a text editor and save it as filename.csv.

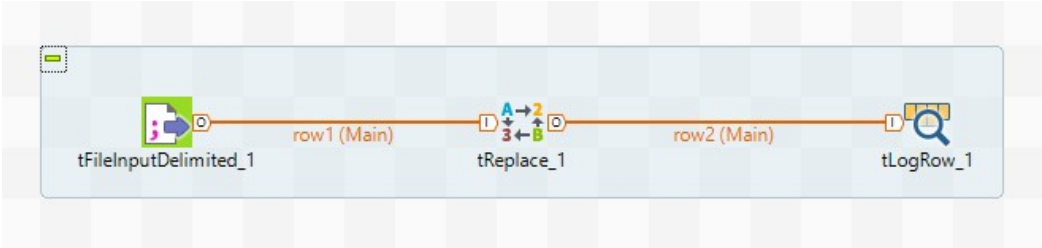
Steps :-

1. Creating a new job: In the Repository right click on *Job Designs* an select *Create Job.*

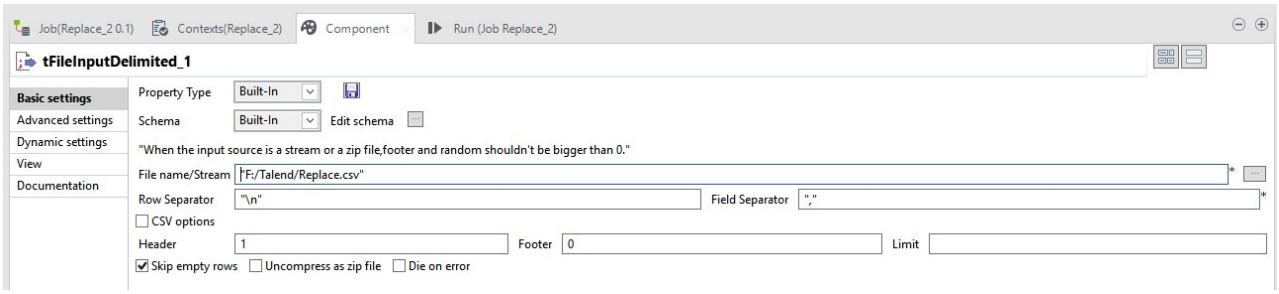


   Fill in the pop-up window that appears: Write the name of your job, its purpose and its description.

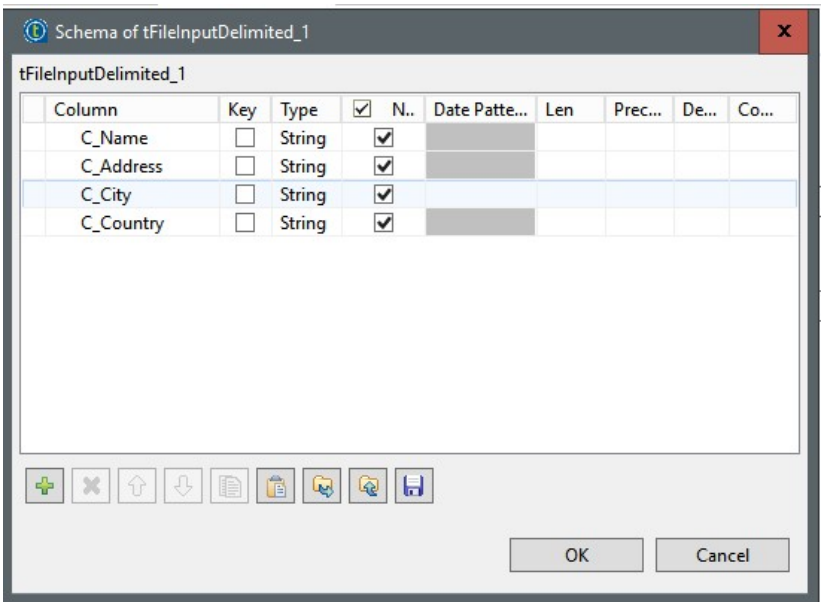
2. Drop the ***tFileInputDelimited, tReplace*** and ***tLogRow*** from the palette onto the design workspace.
3. Connect the components by *right-click -> **Row** -> **Main***.



4. Double click the ***tFileInputDelimited*** component to open its basic settings. In the ***File name/Stream*** specify the path to the input file.

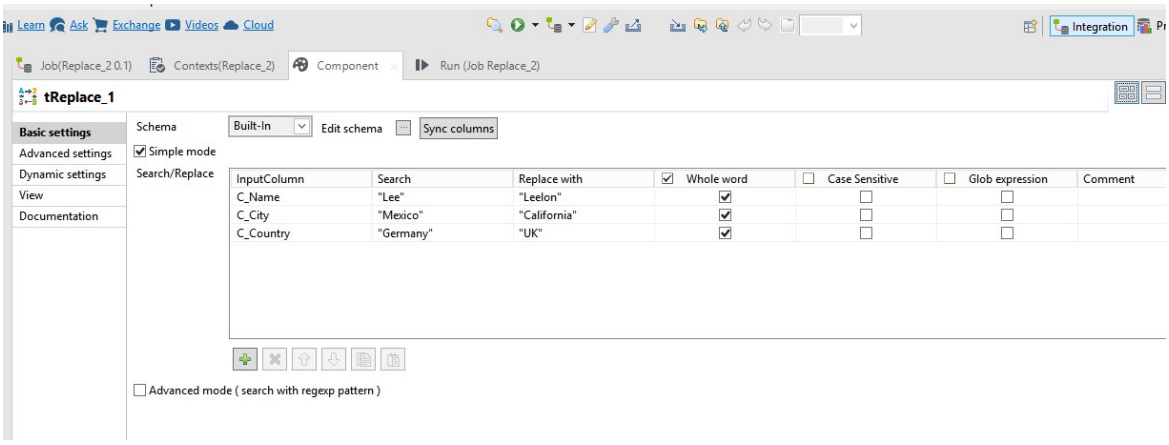


5. Click the **[...]** button next to ***Edit schema*** to open the Schema dialog box, and set up the input schema by adding four columns as following:

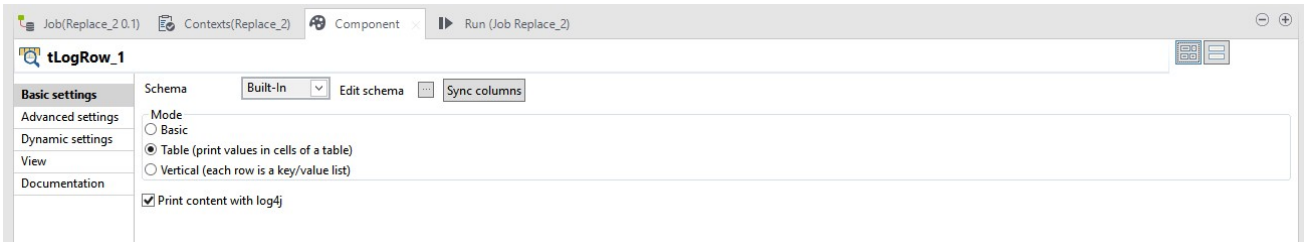


When done, click ***OK*** to validate your schema setup and close the dialog box.

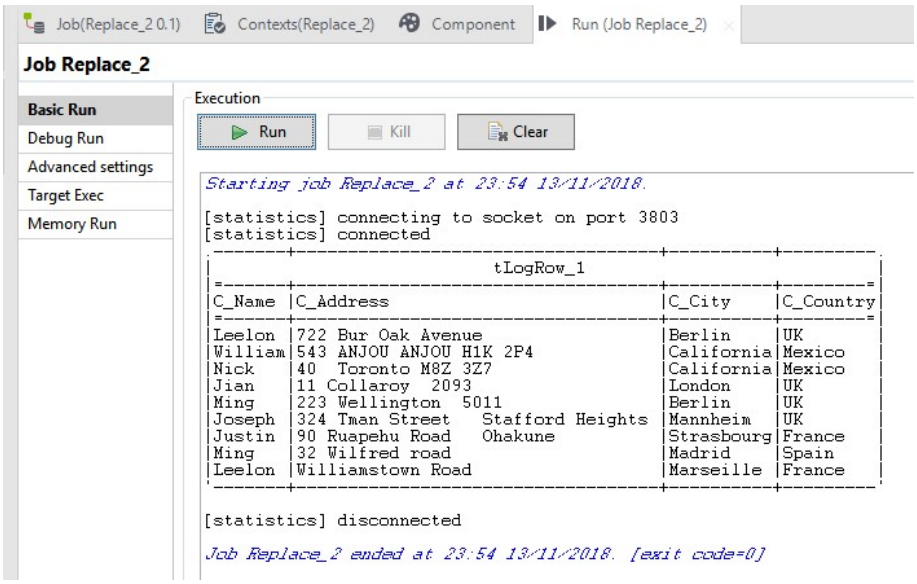
6. Double click the ***tReplace*** component to open its Basic Settings. Check the schema, and if necessary, click **Sync columns** to get the schema synchronized with the input component. Select the ***Simple mode*** check box. Click the plus[+] sign to add some lines to the parameters table.
    - On the first line, select **C_Name** as ***InputColumn***. Type "Lee" in the ***Search*** field, and "Leelon" in the ***Replace*** field.
    - On the second line, select **C_City** as ***InputColumn***. Type "Mexico" in the ***Search*** field, and "California" in the ***Replace*** field.
    - On the third line, select **C_Country** as ***InputColumn***. Type "Germany" in the ***Search*** field, and "UK" in the ***Replace*** field.



7. Double click the ***tLogRow*** component to open its **Basic Settings**, select the ***Table*** radio button in the **Mode** section.



8. Save and Run your job.



Here in C_Name "Lee" is replaced by "Leelon". In C_City "Mexico" is replaced by "California". In C_Country "Germany" is replaced by "UK".

tNormalize

tNormalize helps improve data quality and thus eases the data update.
Normalized data is cleaner and easier to maintain and change as your needs change.

Input :-

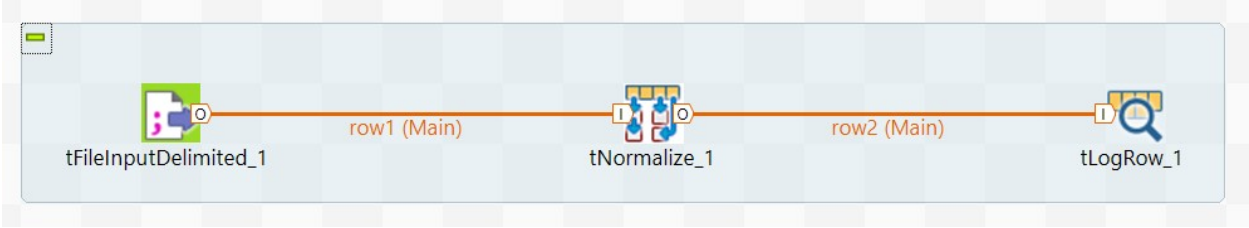
arya,
ankita, rahul ,
ana roy, tanya ,,, kumar ,
aushman, bhumika ,,
kapoor,,
isha,
aishwarya, ankit,divyansh,
kapoor,
aishwarya,ankit, veena,
virat,,

swati,
aishwarya, deepika,ankit

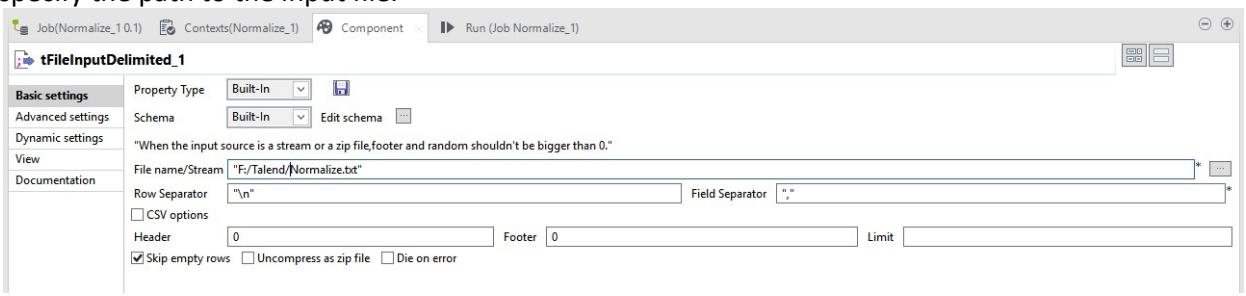
Copy the input customer file into a text editor and save it as filename.csv.

Steps :-

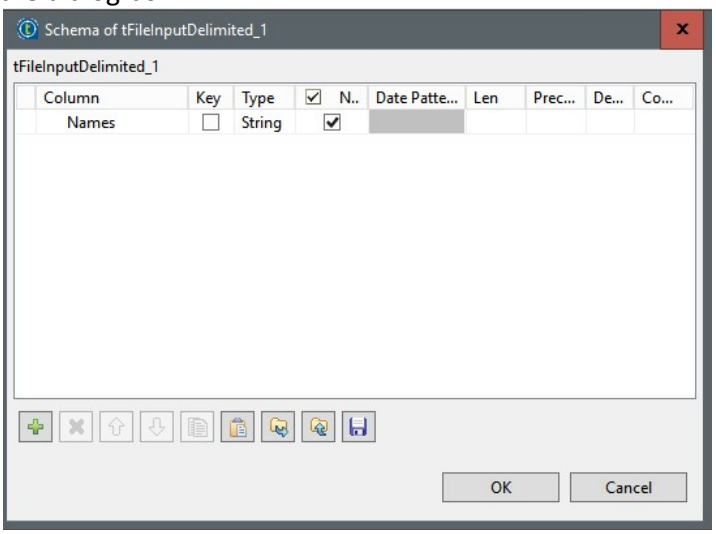
1. Creating a new job: In the Repository right click on ***Job Designs*** an select ***Create Job.*** Fill in the pop-up window that appears: Write the name of your job, its purpose and its description.
2. Drop the ***tFileInputDelimited, tNormalize*** and ***tLogRow*** from the palette onto the design workspace.
3. Connect the components by *right-click -> **Row** -> **Main***.



4. Double click the ***tFileInputDelimited*** component to open its basic settings. In the ***File name/Stream*** specify the path to the input file.

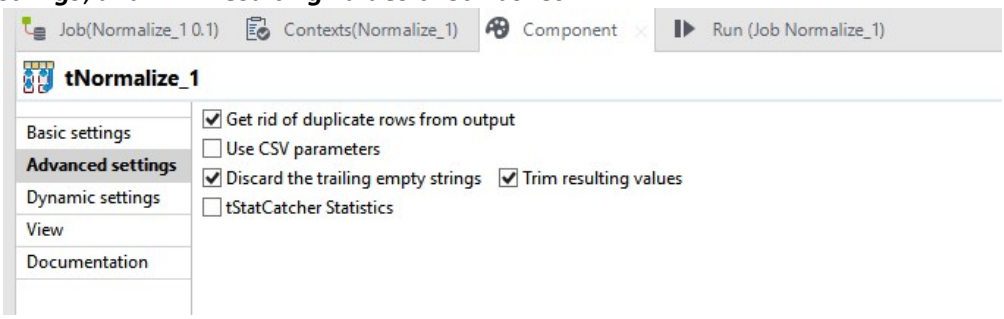


5. Click the **[...]** button next to ***Edit schema*** to open the Schema dialog box, and set up the input schema by adding one column named *Names*. When done, click **OK** to validate your schema setup and close the dialog box.

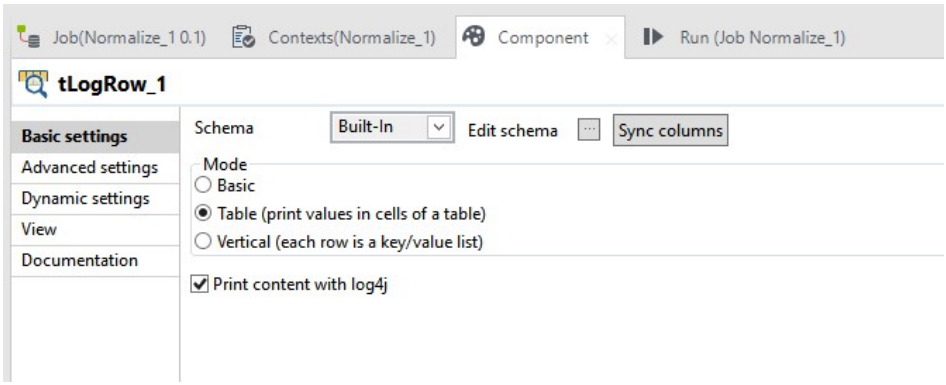


6. Double-click the ***tNormalize*** component to open ***Basic settings*** view. Check the schema, and if necessary, click ***Sync columns*** to get the schema synchronized with the input component.

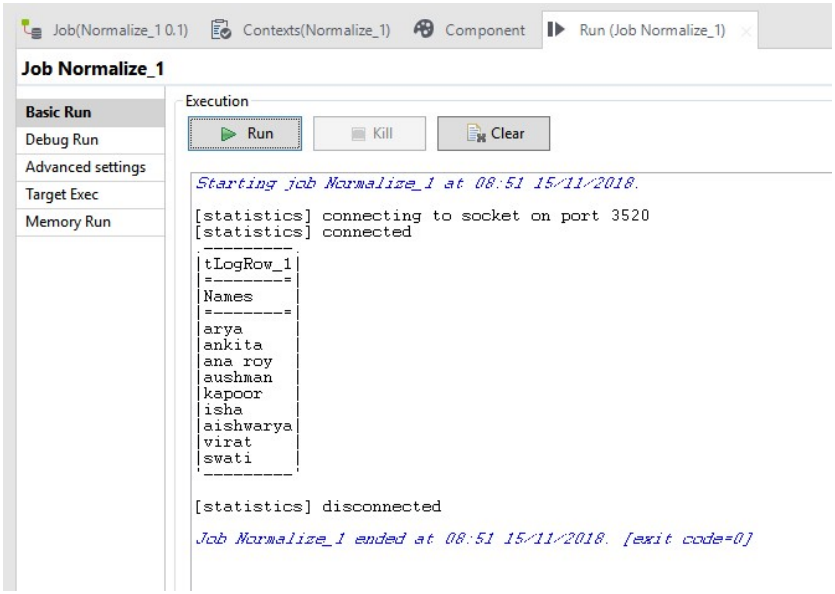
7. In the ***Advanced settings***, select the ***Get rid of duplicate rows from output***, ***Discard the trailing empty strings***, and ***Trim resulting values*** check boxes.



8. Double click the ***tLogRow*** component to open its **Basic Settings**, select the ***Table*** radio button in the **Mode** section.



9. Save and Run your job.

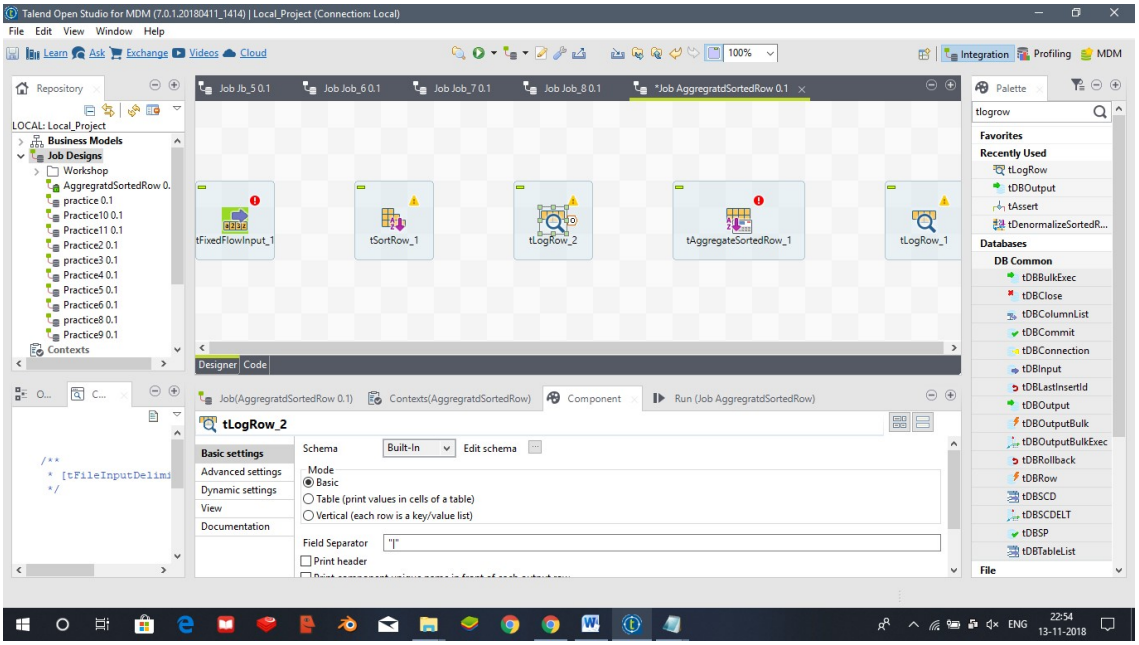


The list is tidied up, with duplicate tags, leading and trailing whitespace and trailing empty strings removed, and the result is displayed in a table cell.

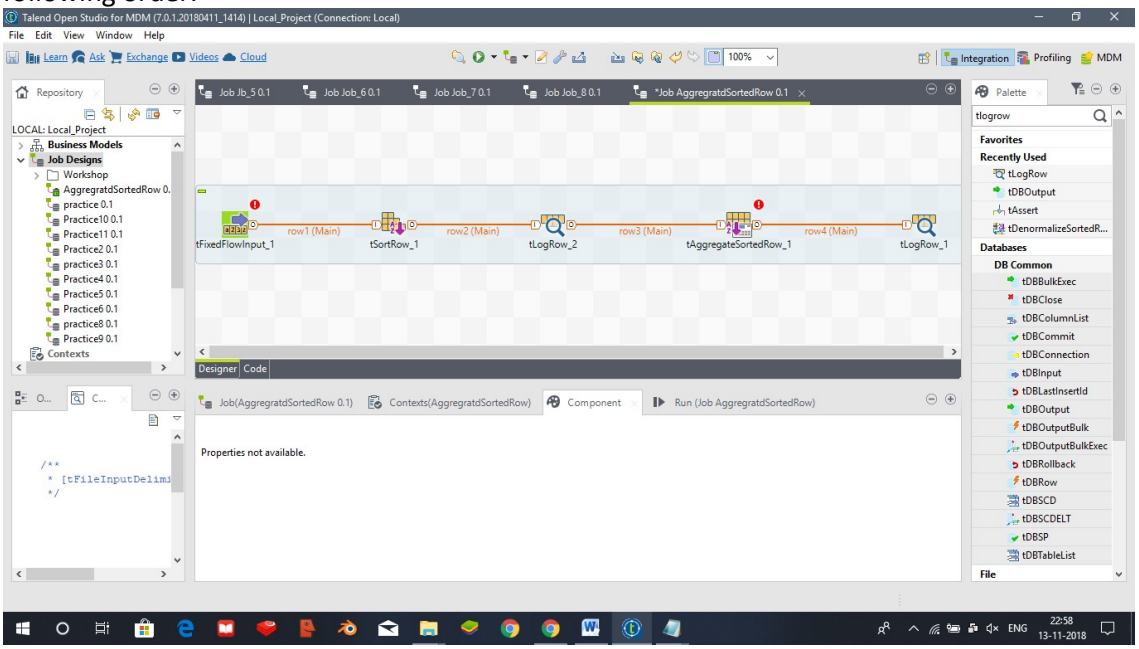
tAggregatedSortedRow

With this function we aggregate the already sorted input data, based on a set of operations for the desired output. We can customize the output in columns as required and the output obtained is the better aggregated data of the input.

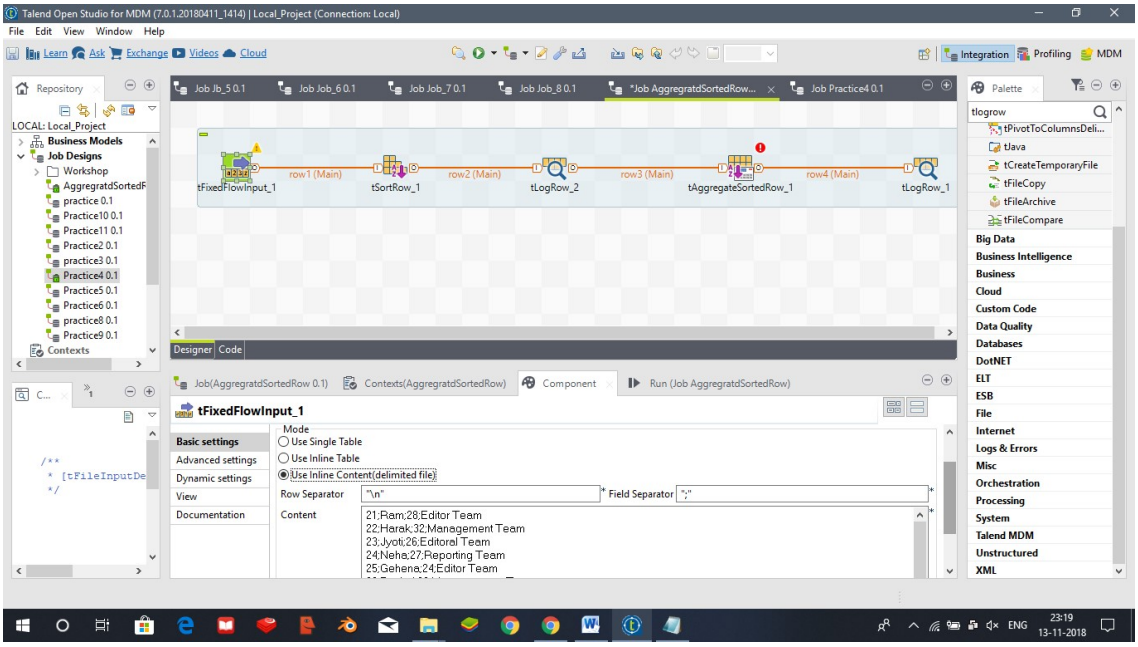
STEPS:



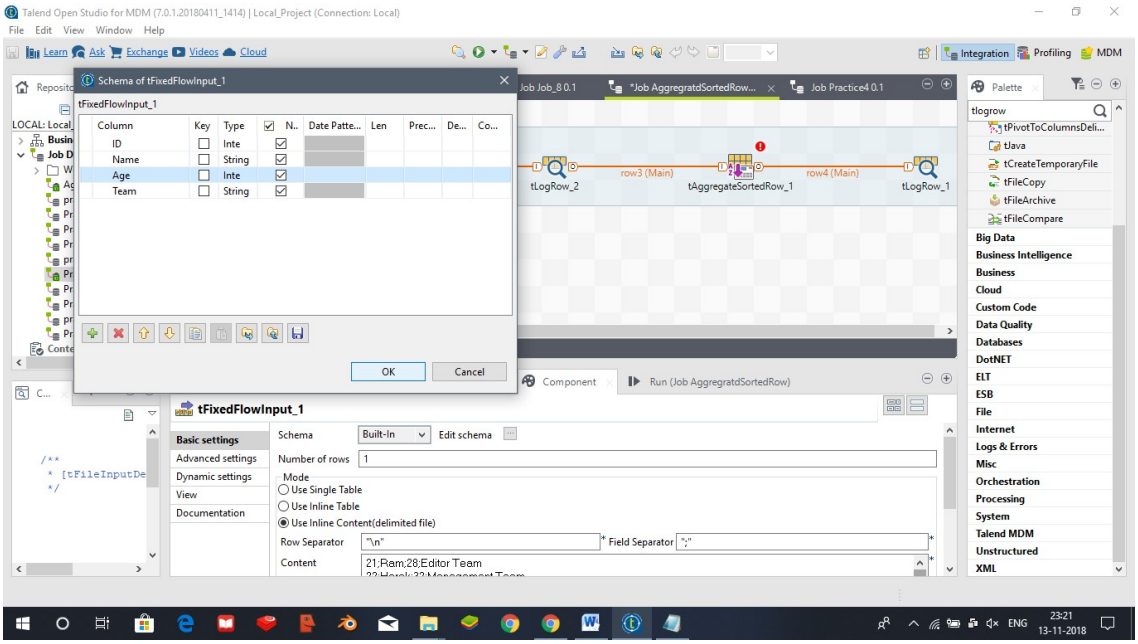
Get the components: tFixedFlowInput, tSortRow, tAggregatedSortedRow, tLogRow. And connect them in the following order.



Now we configure tFixedFlowInput by selecting Mode as Use Inline Content (delimited file) like this and entering the required data.

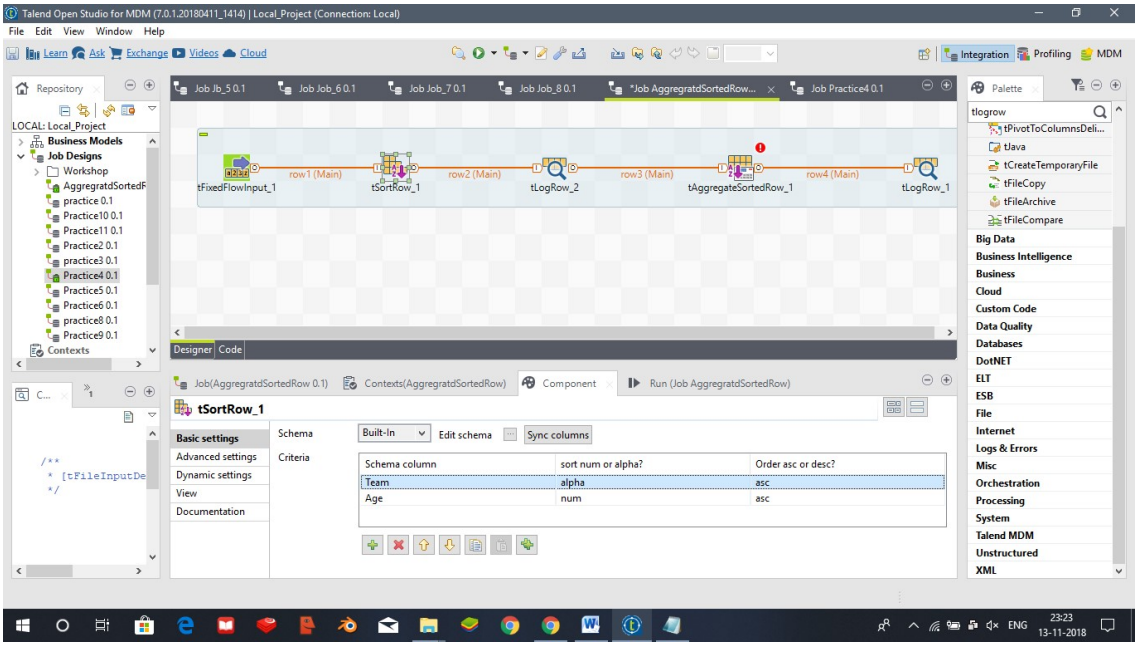


Then we edit the schema like this

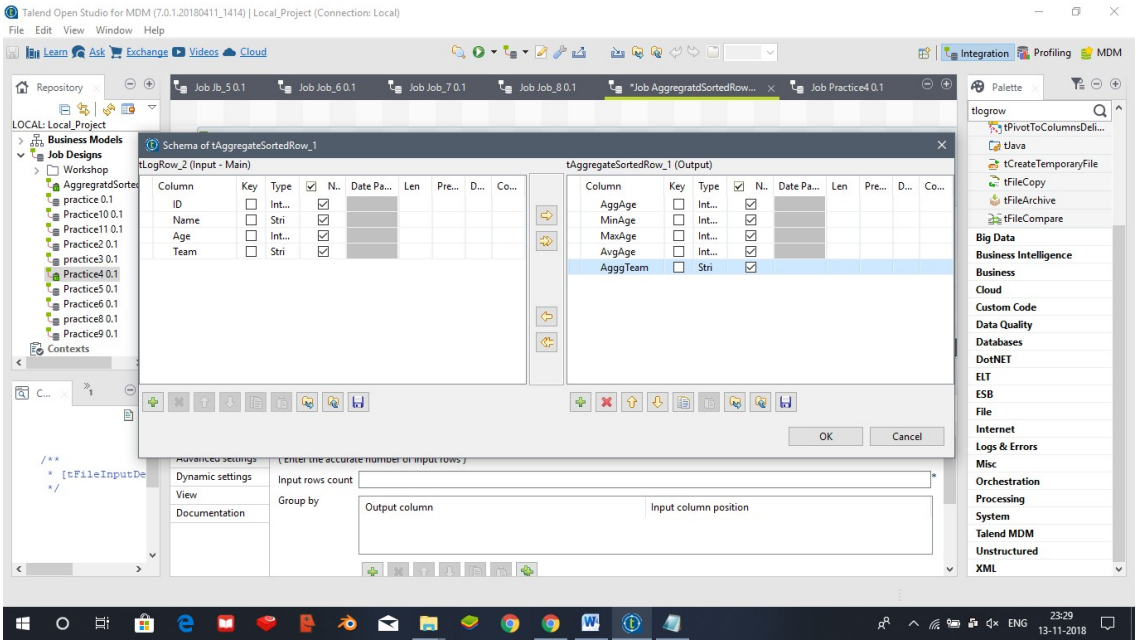


And propagate the changes.

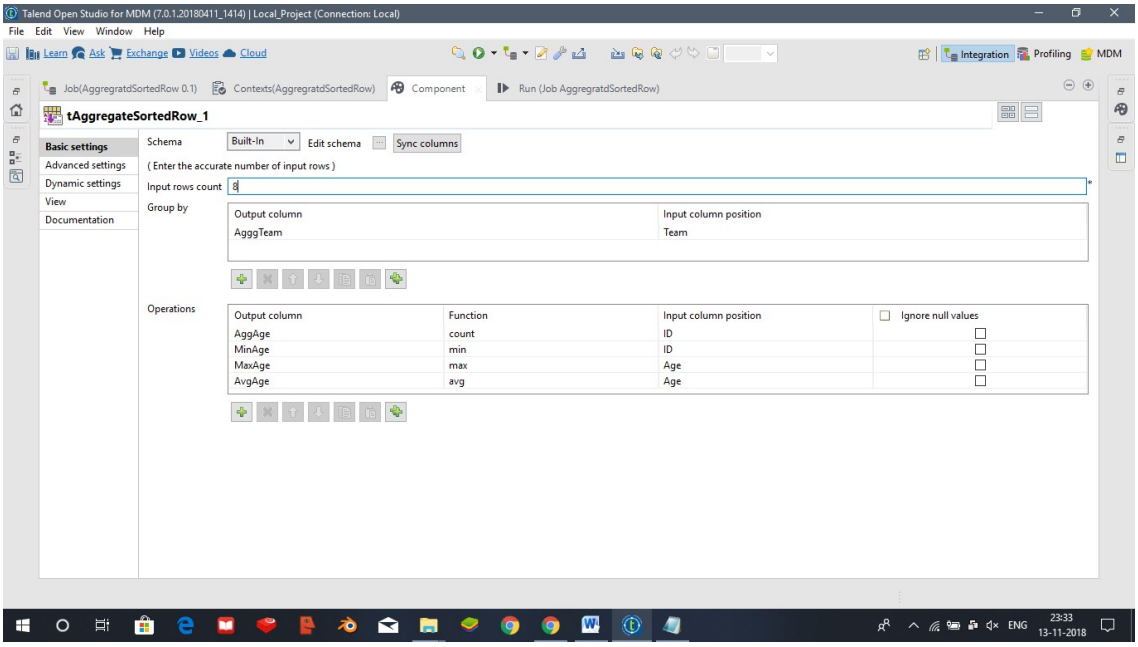
Now we configure the tSortRow component and we add the criteria with which we want to sort the data accordingly.



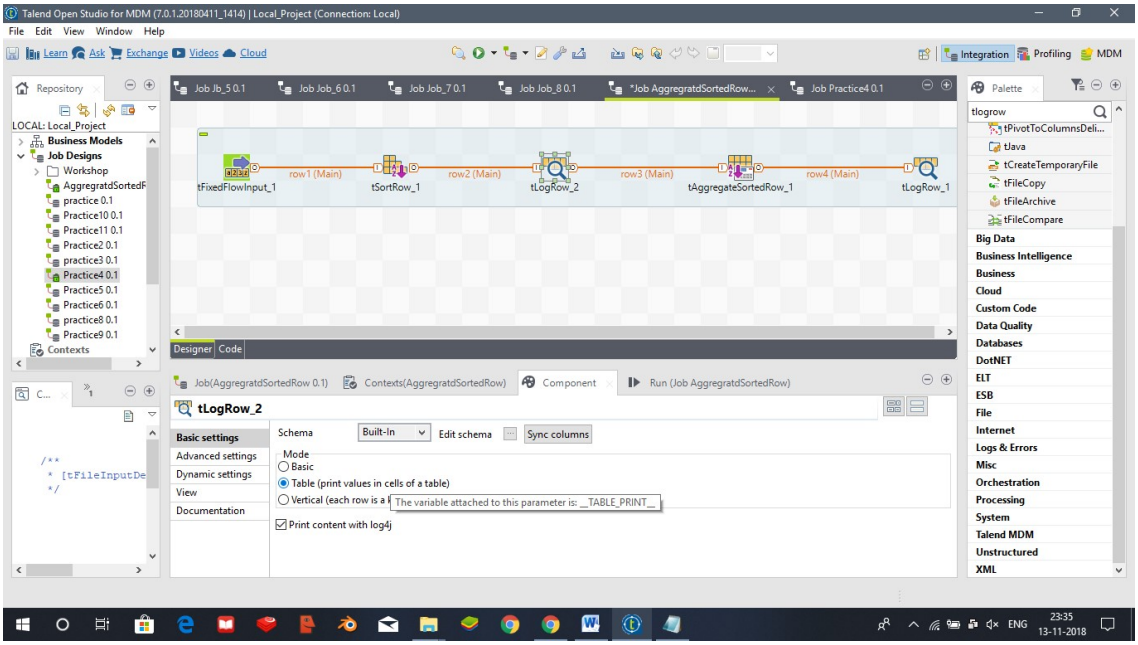
tAggregatedSortedRows schema is edited as



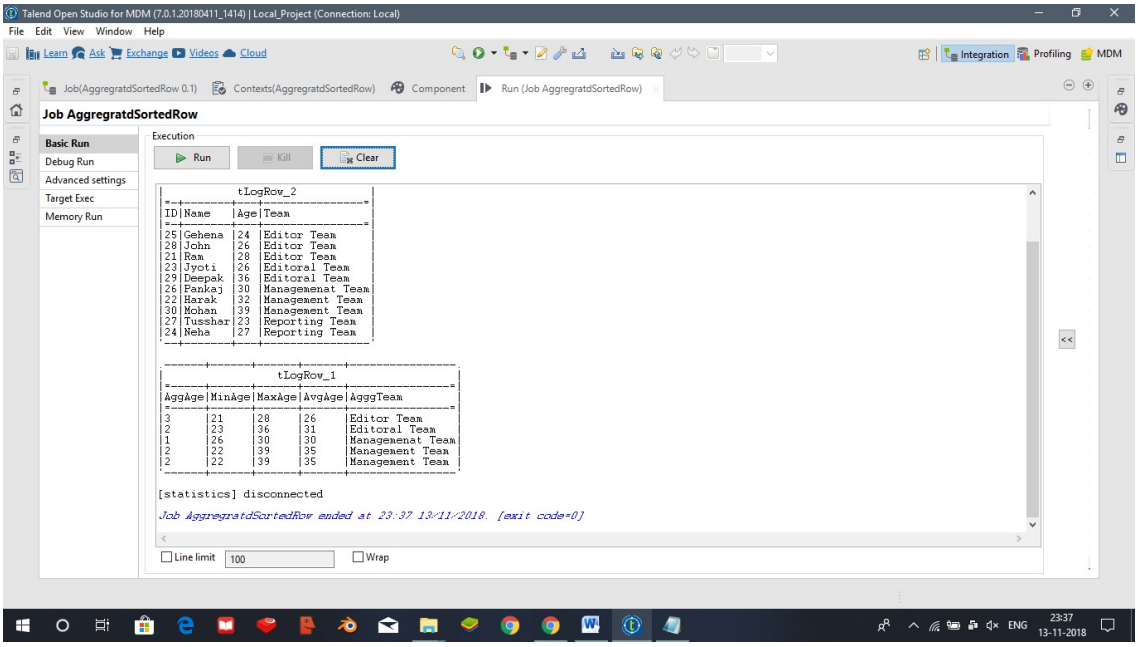
And click ok to propagate the changes. We can group by accordingly and operations as follows



And we finally configure the tLogRow component as

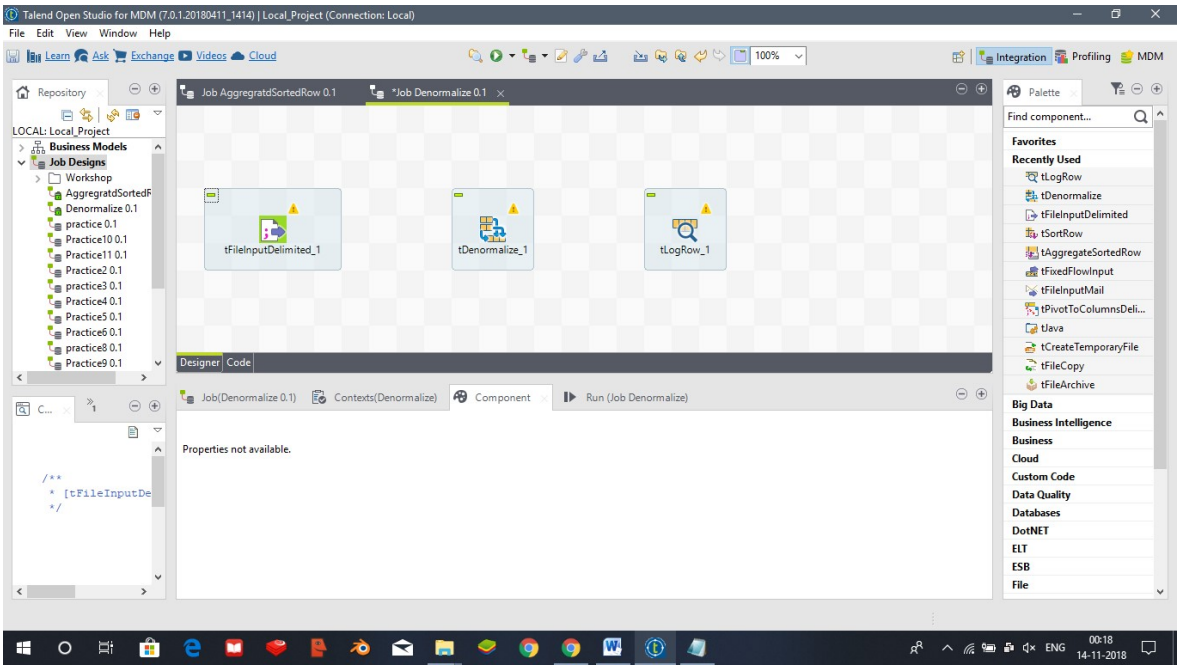


The output we get is as follows as after saving and executing as

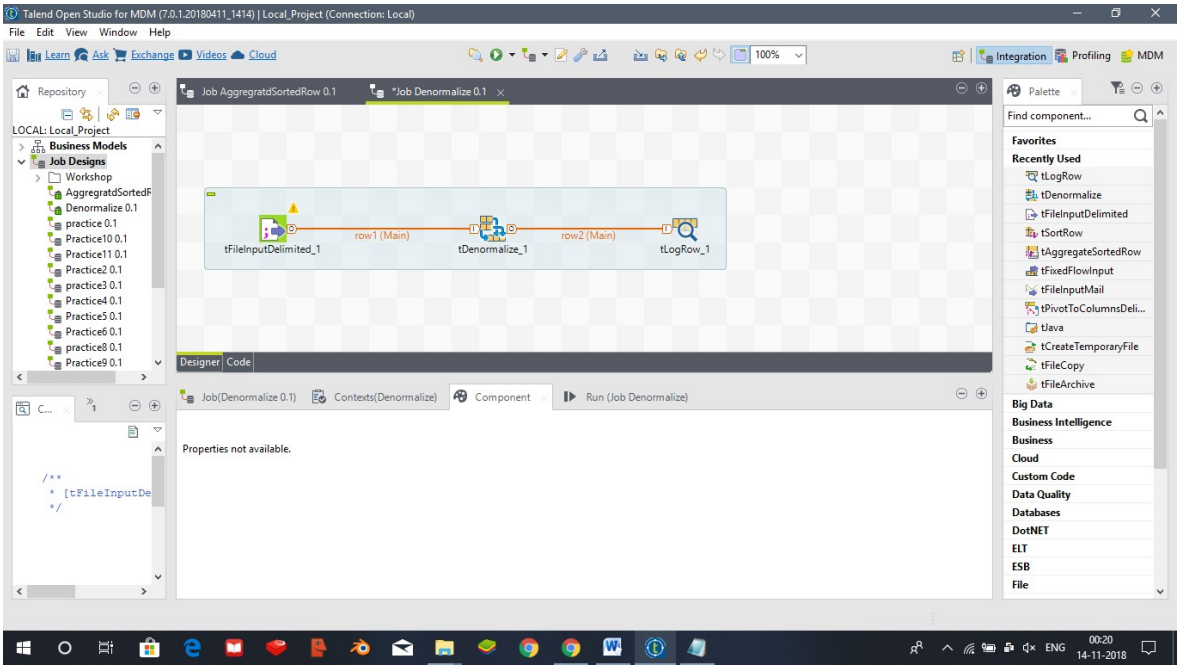


tDenormalize

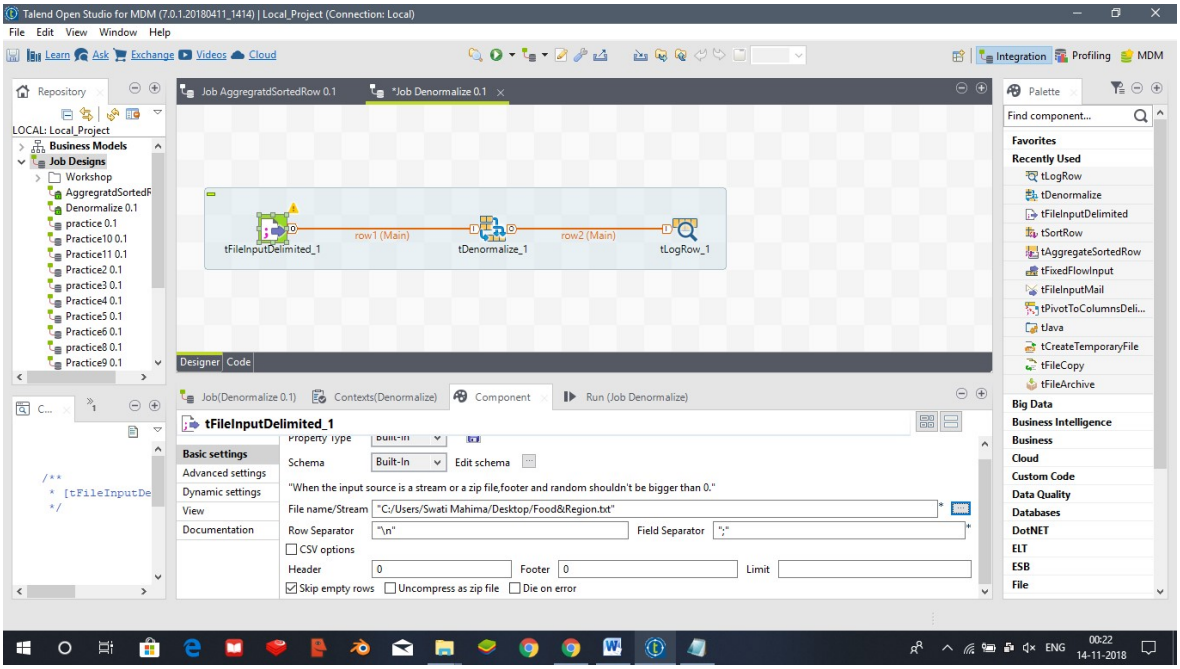
In computing, **denormalization** is the process of trying to improve the read performance of a database, at the expense of losing some write performance, by adding redundant copies of data or by grouping data.
We add the following components: tInputDelimited, tDenormalize, tLogRow.



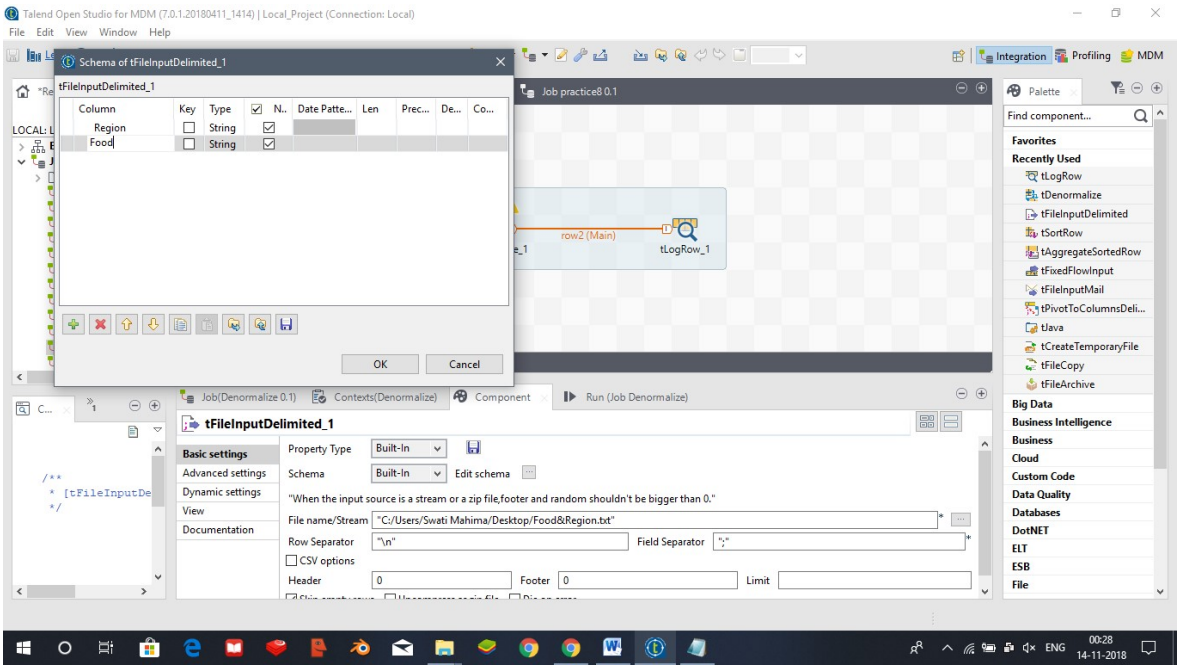
And connect them using Row->Main.



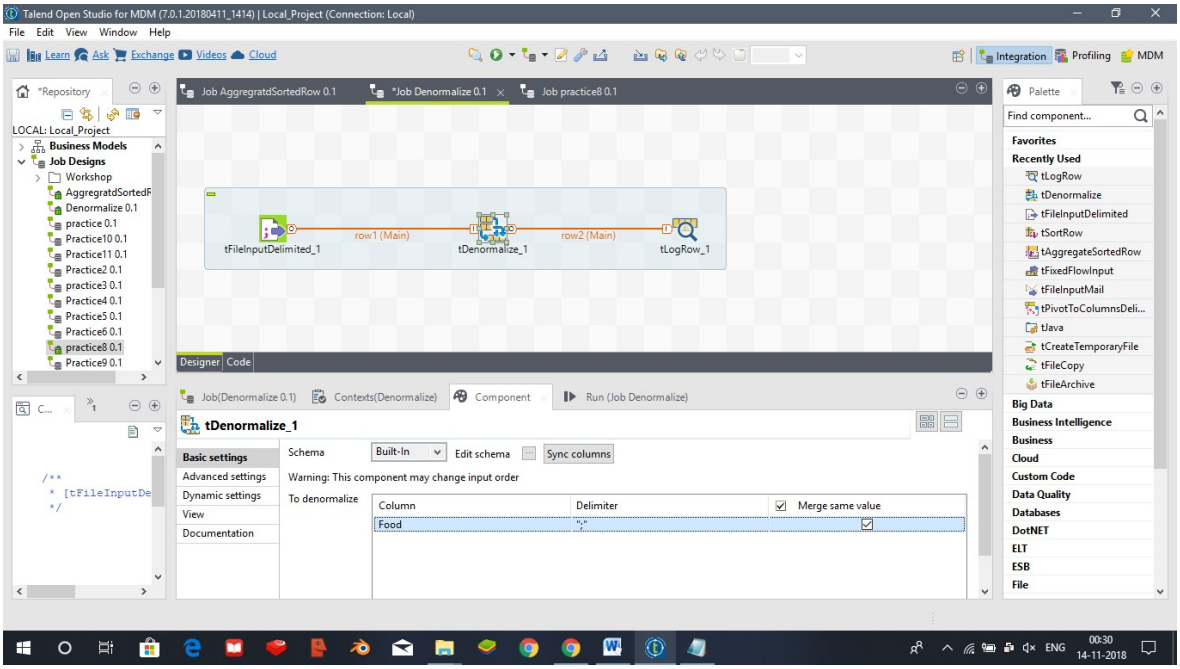
Configure the tFileInputDelimited and add the required data file.



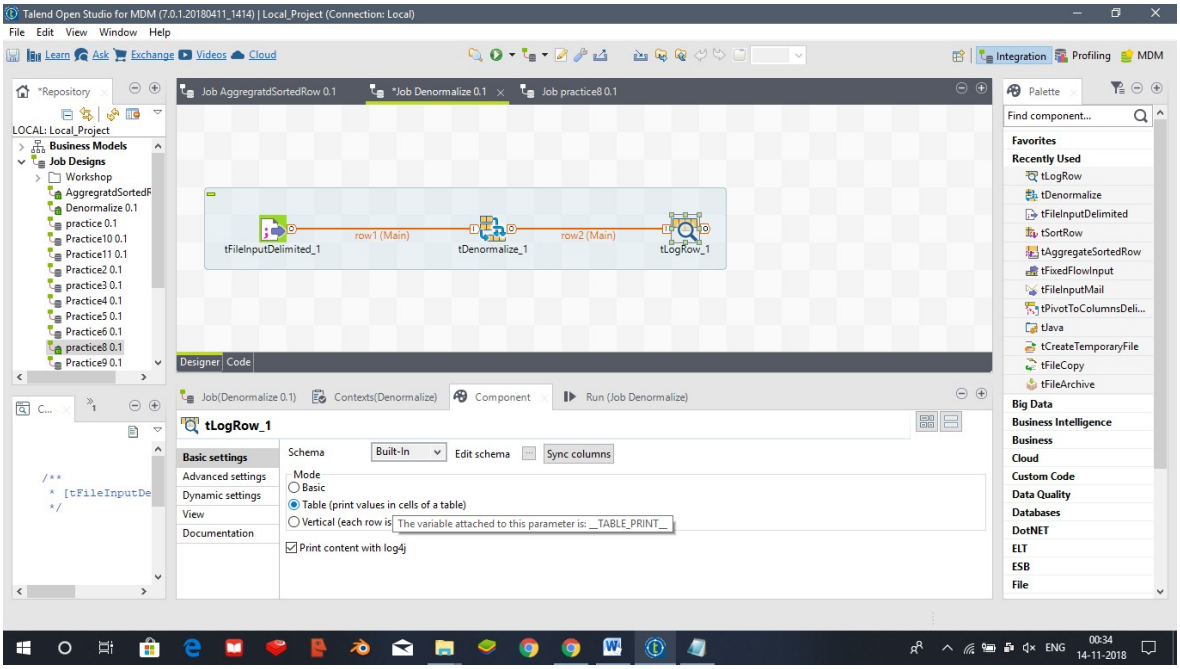
And edit schema as follows



we configure the tDenormalize component as follows



Similarly tLogRow component



Now we save and execute the job and obtain the following result