ETL – Data processing using Talend

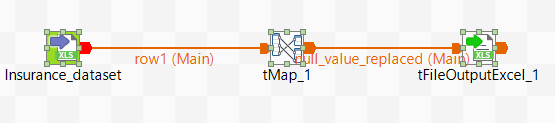
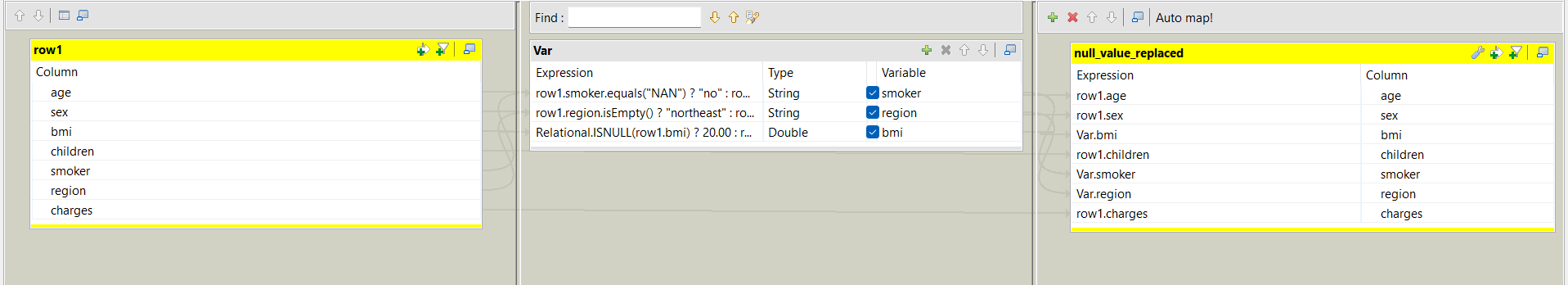
I’ll be using the the [Insurance\_dataset](<https://www.kaggle.com/utkarshakathayat/insurance-dataset>

I have used below 5 processing Talend components:  
tLogRow  
tMap  
tFilterRow  
tAggregateRow  
tReplace

1. Replacing missing or null values

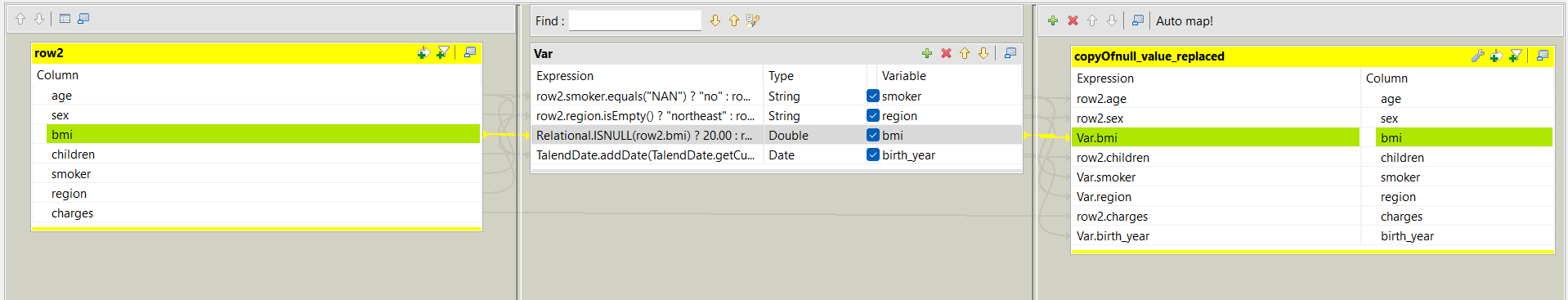
tmap component from the processing family. We’ll form expressions in the tMap component window using the available functions to transform data as per requirement.

row1.smoker.equals("NAN") ? "no" : row1.smoker  
row1.region.isEmpty() ? "northeast" : row1.region  
Relational.ISNULL(row1.bmi) ? 20.00 : row1.bmi

[](https://private-user-images.githubusercontent.com/67043458/348155704-a43299ef-9fd3-4ab1-aabd-1cf15f7227ad.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..buuZI2uelZL1pBvGhFckA6rGkNu5M9-KgjEARTv3OgE)  
[](https://private-user-images.githubusercontent.com/67043458/348156399-6b950e30-ea9b-4e09-8300-c445b908832a.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..acnOuDpC_TPtXeaWbzFQge4RiGHb84AUnjM7XJj7L9I)

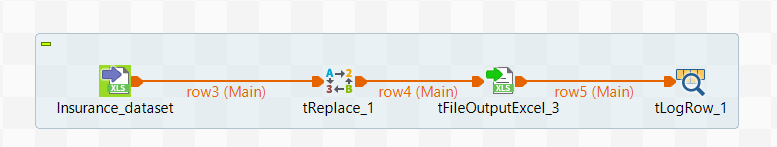
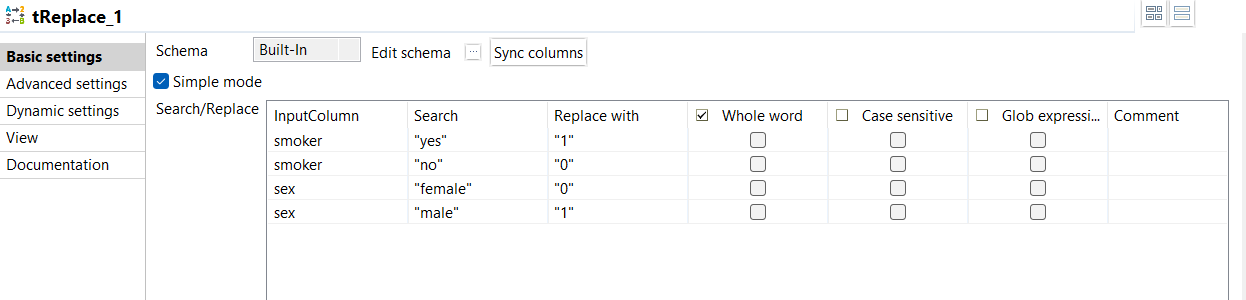
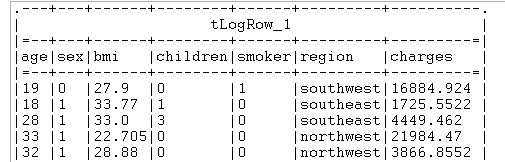
1. Creating a column — Birth Year

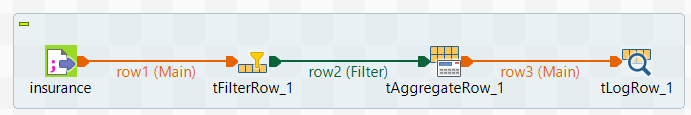
the first parameter passed in the addDate function is another in-built function provided by Talend date category- getCurrentDate(), which as the name suggests returns the current date. We need to subtract the age so the second parameter passed is minus of age and third is yyyy since the year is required.  
TalendDate.addDate(TalendDate.getCurrentDate(),-row1.age, "yyyy")

[](https://private-user-images.githubusercontent.com/67043458/348157452-8ef5de71-3dfa-4ba7-b292-db212102f8ac.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..58ge8qD74XpBBzHryMNaazPZWU_cGV3hnIh1-FWuKRM)  
[](https://private-user-images.githubusercontent.com/67043458/348157886-3c72da8e-bafb-4bc7-9868-0c1f1ace2f6e.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..PIFMAqdbpmKvL2u781dcp-D0LEgBZZWjNNfR_S8HhOM)

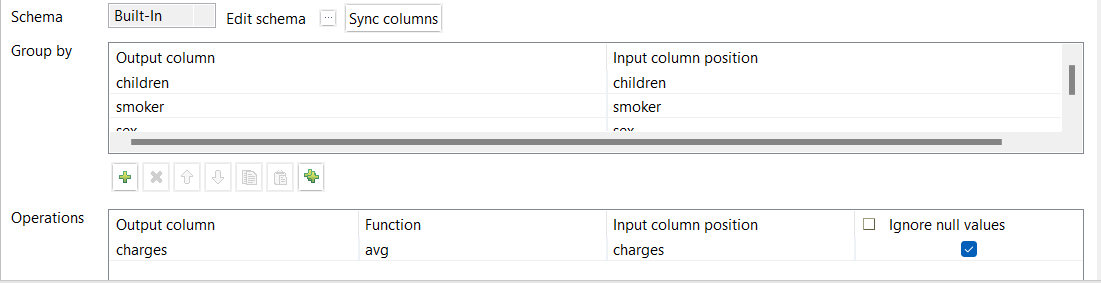
1. Categorical

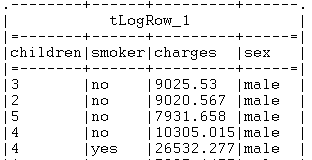
tReplace component from the processing family which searches and replaces in the input columns defined and helps to clean the file before further processing.

[](https://private-user-images.githubusercontent.com/67043458/348159346-48080b30-6d07-4df6-90bb-b359208c2e8b.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..5YidUq2N7hBoRrllJY7bcsCiGC3uzFMIFN4YRKa5y3w)  
[](https://private-user-images.githubusercontent.com/67043458/348160811-2ae64f0a-a731-4c0a-aafe-cbefd0e326fc.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..NEu44ai-mzfc4P5iaq3yMkle08ZkdLcIaLisspm9Q7g)  
[](https://private-user-images.githubusercontent.com/67043458/348161420-3ea4d6c8-77df-4879-a423-84ccb9936bd6.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..hZWHRC2uHxQehpHLjOx2Z5tDh_I6yXZxSh0oHzg9Jpo)

1. Filtering and Aggregating Data  
   We will compare the average insurance charges for smoker and non-smoker males and females with children  
   [](https://private-user-images.githubusercontent.com/67043458/348162711-987f2b4e-a1ba-4365-a893-a28431c4ef24.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..DMfIVmQjsx5L1uJh5T2PUGAa4bWD1U76PaQ1ezUxgyg)

a) We will filter out the family with children first , so we will apply the filter as children > 0  
[](https://private-user-images.githubusercontent.com/67043458/348162921-5ab43766-faa4-4bfd-8a1f-c28e71769ecd.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..LF-MGinLi0lbmQ9M5DRYv8zyONUfZNgcNJ8KTW1AcVE)

b)Once the filtered data is passed we will group the data by sex and smoker and apply the average function to the charges column  
[](https://private-user-images.githubusercontent.com/67043458/348163491-a9e5c9a2-d8c5-4780-a04e-1285d6d7b111.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..WUO4N0dqWIIsvQm-6CfoFABy-LFBe2KO8fh7-6cWItE)

family having children and non smoker is having less insurance charges as compared to smoker  
[](https://private-user-images.githubusercontent.com/67043458/348163955-b7ff77c6-65c2-4041-8002-0d565ac406a0.png?jwt=eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9..exGRahHutyESJFVSPrkj_rpejtsITt7AG36F7Qnlcfs)