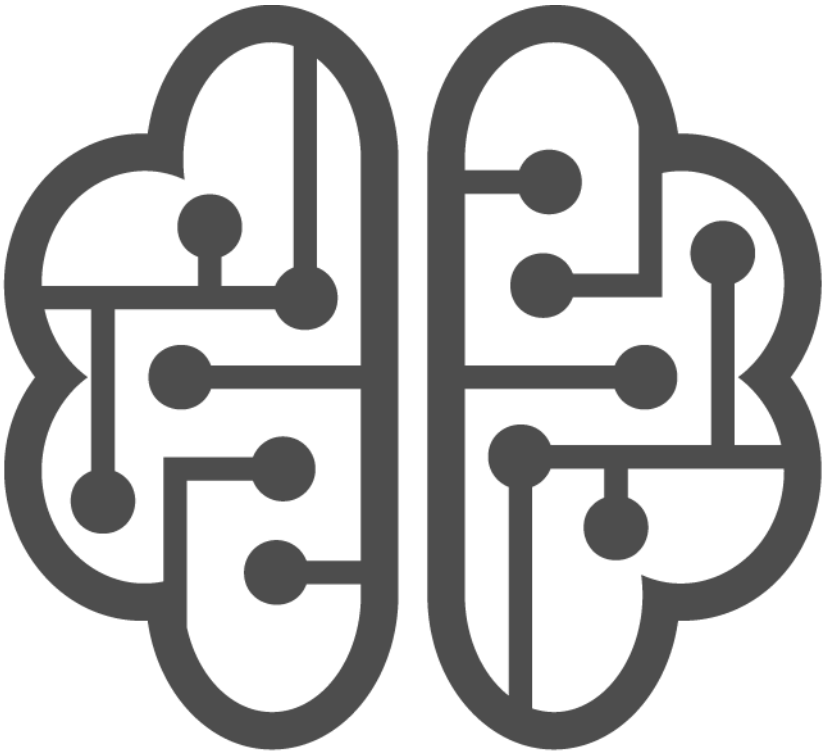


Extraction and cleaning part

December 2018

Contents

Introduction:

This deliverable is dedicated to explain how we did the extraction and the cleaning of the different data source into the main data ware house, this phase is really important because the rest of the data warehouse building blocks will be built on the top of it.

Tools we used:

Talend ETL:

We have used The Talend ETL for the main task of cleaning and extracting the data, it's an Open Studio for Data Integration Features and powerful tool for any integration project

For download the free edition of this software or for more information you can visit the link: https://www.talend.com/products/data-integration/data-integration-open-studio/

DataGrip:

To visualize the result of loading and for testing the resulting data warehouse we used the DataGrip IDE provided by jetbrains[1], it does require a license key to work, more information can be found at the link: https://www.jetbrains.com/datagrip/

[1] https://www.jetbrains.com/

MySQL:

We did used also MySQL database as data warehouse which contain all the data we extract from the different data sources, MySQL is an open-source relational database management system (RDBMS). For more information: https://www.mysql.com/

Data sources:

- We had the university databases which care stored by years (2012 database, 2013 database…), the university database use Microsoft access as a Database management system.
- A Google form[2] Result CSV file which contain some question to the student who did completed their university cycle about some information that are not stored in the database, unfortunately we couldn't use it because the number of response to this form was too small to take it in consideration.

[2] Link to the form here

Extraction:

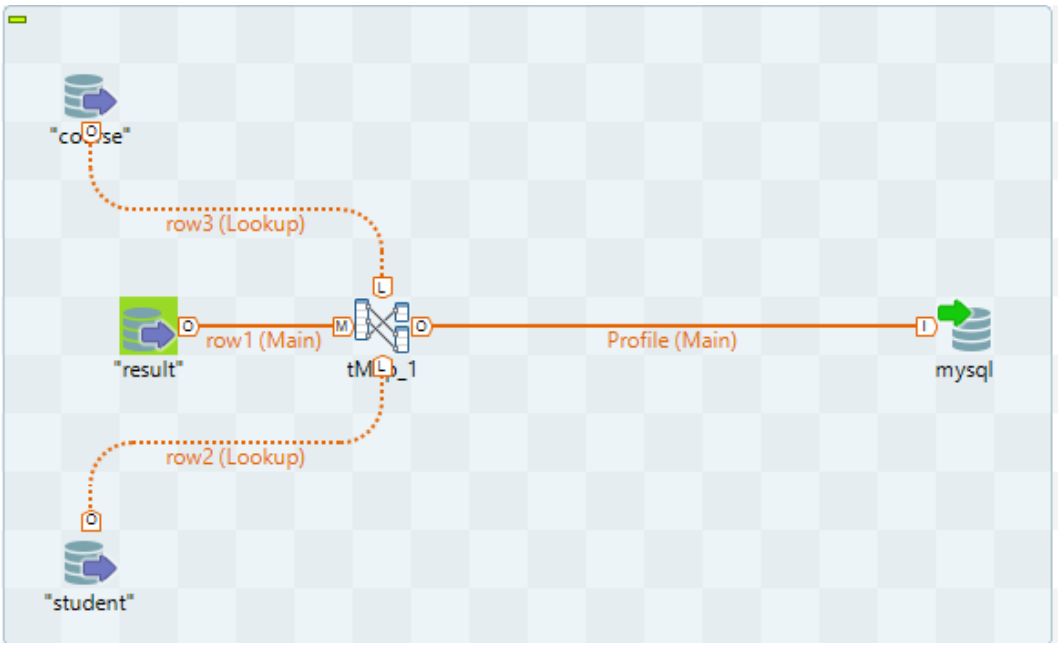
In the extraction part we did simple, concise and easy work, most of the work was handled by the Talend ETL tools, it can be described as follows

1- we designed something called jobs inside the software
2- we told the job how to do the extraction,
3- then we execute the jobs we built

For the job definition we define the inputs which are our tables of the databases of the university, and the output tables of our data warehouse, the we map this two objects with a map, and we tell the map how to takes the input and process it and how to push into the output tables, during this mapping we specified the cleaning options such as filtering and transformation data types.

The jobs we defined are explained in separate HTML document generated by the tool, it does contain the ins and outs how we did the work in detailed way.

Here is a screenshot of one of the jobs:



Cleaning:

In this part (which interrelated with the previous) we did filtering all the outlier data that can cause a problem in the data warehouse analysis and data mining processes, for example we do have some students who has no TD and TP marks and got an average of 12!! , this can cause big problems for data mining algorithm and also for reporting since we are reporting an incorrect data.

The cleaning work is also explained in the past HTML document provided with this delivery.

Conclusion:

Extracting, transforming and cleaning the data before using it is the core task of the data warehouse maker, we did it in our way using Talend data integration tool which makes the loading and transforming the data a piece of cake.