ECOLE INTERNATIONALE
DES SCIENCES DU
TRAITEMENT DE L'INFORMATION

EISTI

ADEO1 PROJECT

Geographical statistics of engineering internships

Authors:
Gustavo FLEURY
Cylia BERKANE
Fagnuo CAI
INDURAJ PR
Quoc Viet PHAM
Yen Chu CHEN

Advisers:
Esma TALHI
Rachid CHELOUAH

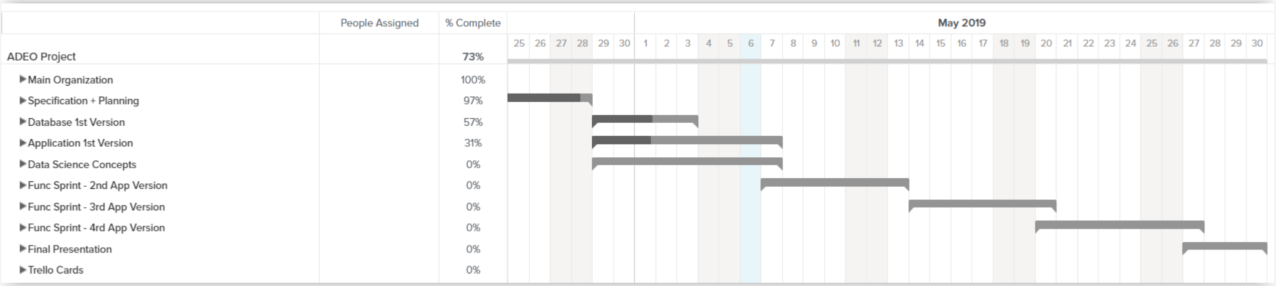
Cergy - May 6, 2019

Contents

List of Figures

1 Planning of the project

The project steps done are showed in figure below. We update the date of Final Presentation and add one more Sprint for development. For the previous activities, just the first example of infrastructure is not finished.



	People Assigned	% Complete
ADEO Project		**73%**
▶ Main Organization		100%
▶ Specification + Planning		97%
▶ Database 1st Version		57%
▶ Application 1st Version		31%
▶ Data Science Concepts		0%
▶ Func Sprint - 2nd App Version		0%
▶ Func Sprint - 3rd App Version		0%
▶ Func Sprint - 4rd App Version		0%
▶ Final Presentation		0%
▶ Trello Cards		0%

Figure 1: Schedule and current status of the final grade project.

2 First Data Base Model

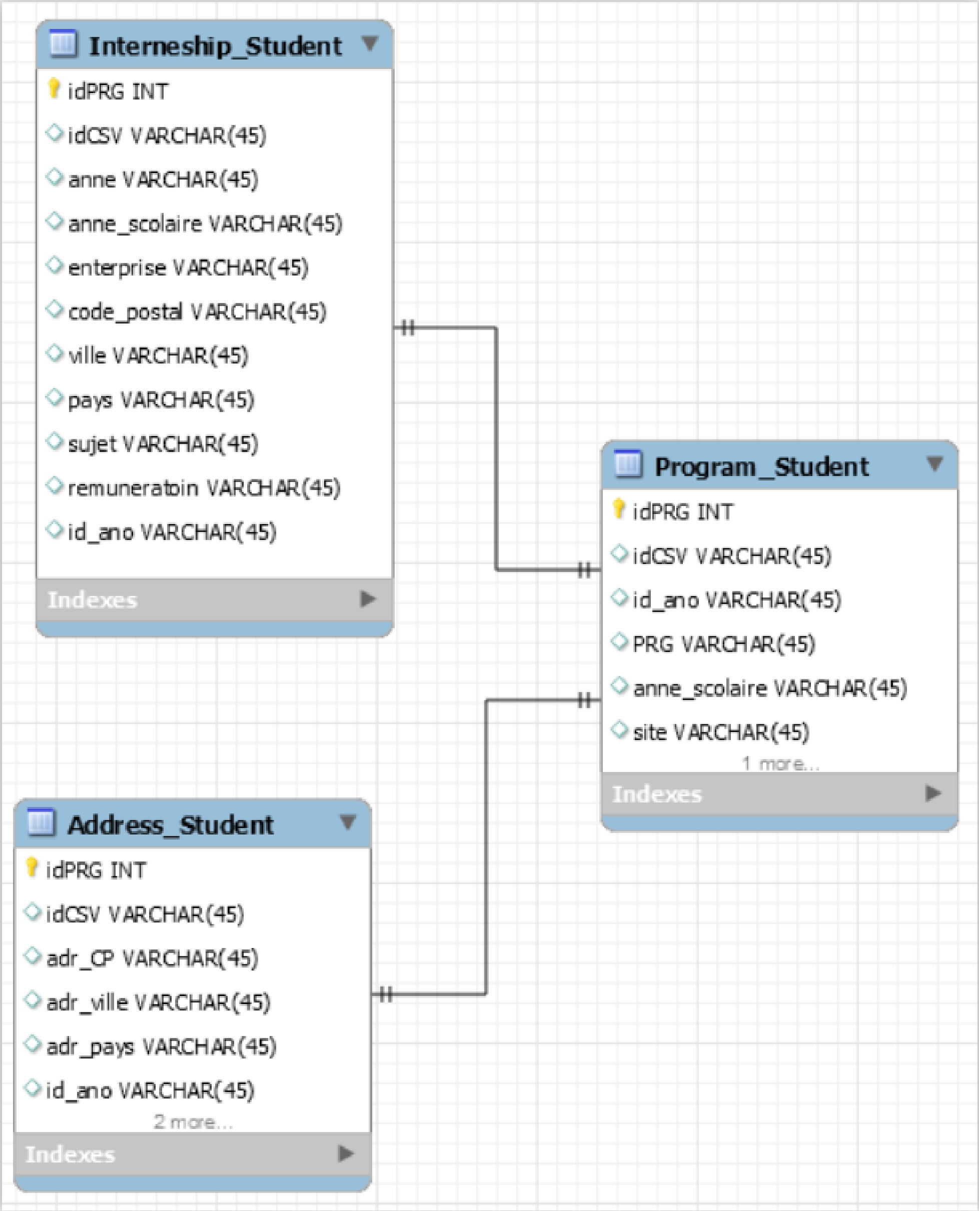
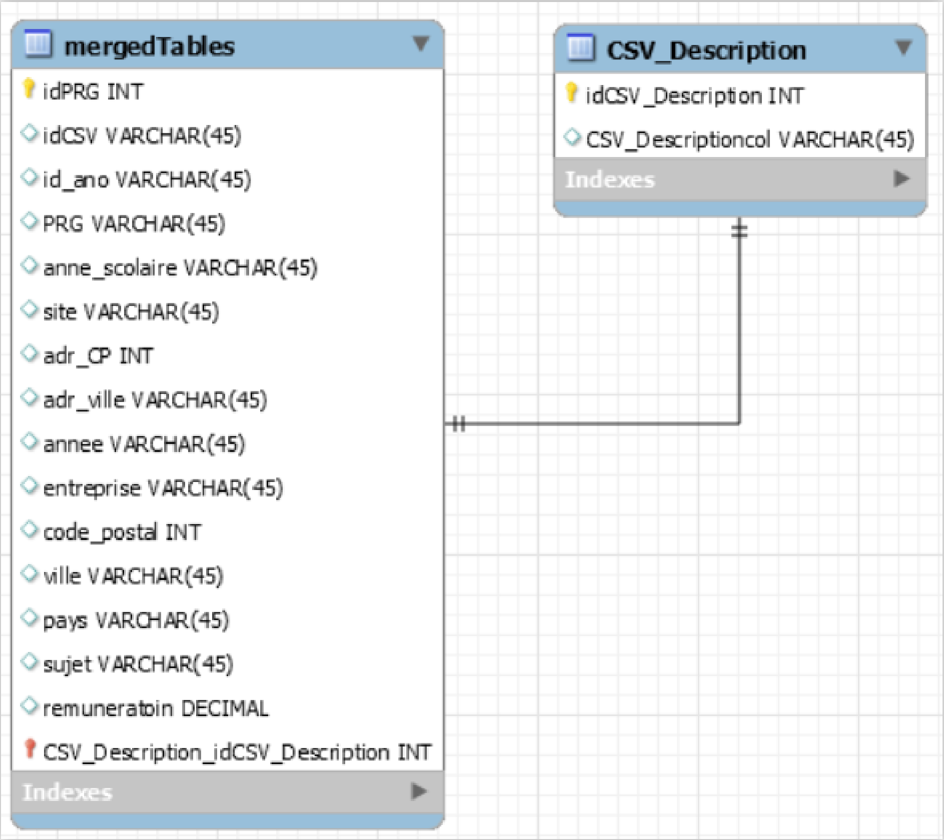


Figure 2: CSV files First Tables .

In the previous figure we can see the raw data from the three CSV files

- PRG_STUDENT_SITE_2017_2018_DATA_TABLE.txt
- STUDENT_INTERNSHIP_2013_2018_DATA_TABLE.txt
- ADR_STUDENTS_2017_2018_DATA_TABLE.txt

After ETL we merged like showed below :



mergedTables

idPRG INT
idCSV VARCHAR(45)
id_ano VARCHAR(45)
PRG VARCHAR(45)
anne_scolaire VARCHAR(45)
site VARCHAR(45)
adr_CP INT
adr_ville VARCHAR(45)
annee VARCHAR(45)
entreprise VARCHAR(45)
code_postal INT
ville VARCHAR(45)
pays VARCHAR(45)
sujet VARCHAR(45)
remuneratoin DECIMAL
CSV_Description_idCSV_Description INT

Indexes

CSV_Description

idCSV_Description INT
CSV_Descriptioncol VARCHAR(45)

Indexes

Figure 3: First Data Model.

3 ETL - Extract Transform Load

The role of an ETL software is to collect relevant data from both files or systems, transform them to make them compatible with the Data Model, and finally load them into the Data Base.

The operation of the ETL platform is divided into three phases. The Extraction phase consists of collecting data from one or more sources.

The transformation phase consists of reformatting and transforming the data. Finally, the loading phase consists of transferring the transformed data to the target database.

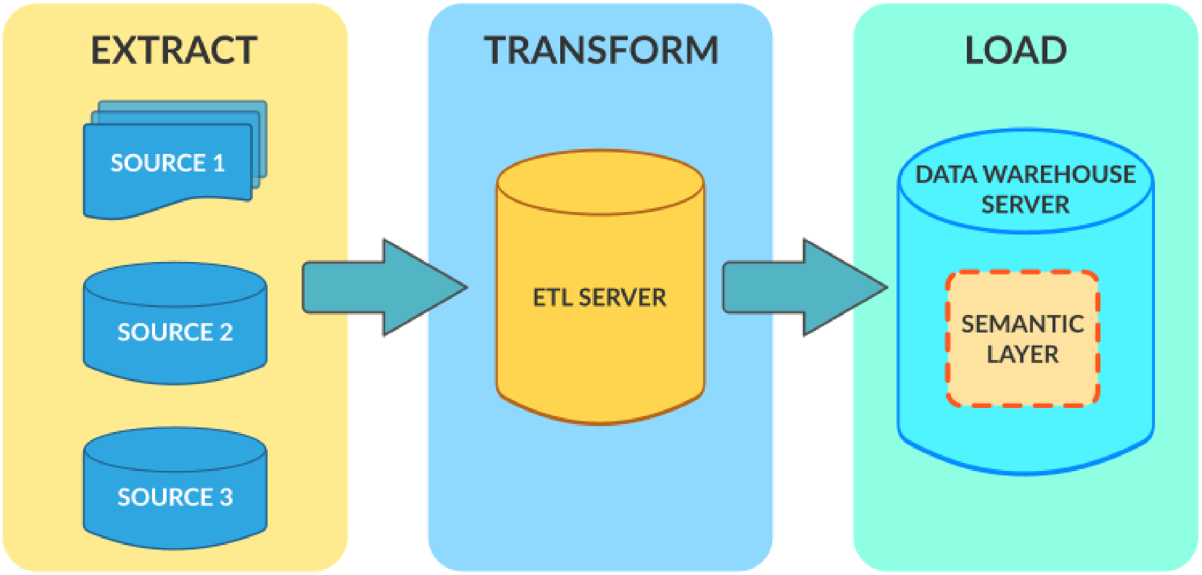
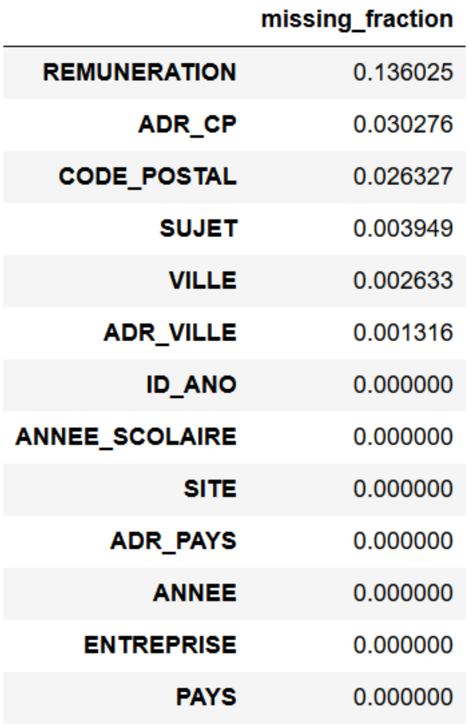


Figure 4: How ETL Works.

During the cleaning process of the data we went through several steps, one of them was the discovery of the missing values. All the steps are found in the Jupyter File.



	missing_fraction
REMUNERATION	0.136025
ADR_CP	0.030276
CODE_POSTAL	0.026327
SUJET	0.003949
VILLE	0.002633
ADR_VILLE	0.001316
ID_ANO	0.000000
ANNEE_SCOLAIRE	0.000000
SITE	0.000000
ADR_PAYS	0.000000
ANNEE	0.000000
ENTREPRISE	0.000000
PAYS	0.000000

Figure 5: Missisng Values.