Advanced Databases for Finance
Project Assignment
Part 1

In this project, we will go through the entire process of modeling a database from a set of business requirements and delivering a functional system capable of delivering insights from data.

The objective is to apply the different concepts we have seen during the lectures and tutorials in order to gain a practical knowledge of the SQL language and the psycopg2 python library.

Environment setup:

Before start working on your solution, create a new notebook and write your full name and student number in the first cell.

You should write the full solution using this notebook; it will be your rendering for the project.

Before starting your notebook, be sure to install the psycopg2 library.

For each step, we want you to have two cells, one markdown cell to explain your approach, and the other one for the code.

Problem statement:

In order to have a clear view of the employment market in France, the ministry of economy, in a joint initiative with the “Pôle Emploi” institution, decided to create a database containing:

- The information about the people, the companies that employ them, the sector of activity of those companies, the departments where they are located and where the people live
- In addition, information about the regions to which the departments belong

1- Entity-Relation Diagram

To help you design the ER diagram that represents well the problem you are solving, let’s consider the following specifications:

- A person is identified by its medical security number (ssn) and characterized by its first name, last name, age, education level (Bachlor, Master, Doctorate or Professional qualification), main sector activity, and employment status (employed or unemployed)
- A person lives in a department
- A person works in one company or is unemployed
- A person cannot work in multiple companies
- A company is identified by a company identification number, called KBIS, and characterized by its name, year of creation, turnover and number of employees
- A company is located in a department and has only one main sector activity
- A sector activity is identified by a string of four characters and has a name

- A department is identified by its number and characterized by its name, area, population and density. A department belongs to one and only region
- A region is composed of multiple departments and has a name, geographical situation (North, South, East or West)

Using the editor of your choice, draw the ER diagram to model your database and respect the different specifications.
Once drawn, export the image and add it to a markdown cell of your notebook.

2- Relational schema

Once you finish the ER-Diagram, write down the relational schema in this format:
table_1(<u>id:</u> string (primary key), name:string, …, attribute_fk: integer (foreign key (table_name))

The relational schema should be written in a cell on your notebook and indicated the different constraints (primary key, foreign key, …)

3- Tables creation :

Write SQL statements that create tables based on your relational schema

The SQL queries should be executed only from the notebook using pyscopg2

4- Data creation :
Populate you database with a minimum of 10 lines in each table, the data you create should be consistent with the constraints you defined.

5- Insights :

Write SQL queries to answer the following questions:

- How many companies there is in a specific department (choose one department, let's say department number 75)
- Get the number of companies present in each sector activity
- How many people hold a master degree and are in same time unemployed
- Get the regions name and geographical situation and order them by the number of companies situated in them, from the region with the greatest number of companies to the region with the least number
- What is the department with the highest number of people having a bachelor or a master and working in a company created after 2000 and operating in the activity sector "aeronautics"

<u>Important:</u> Some queries can have impact on your relational schema; in this case, you have to modify your schema in order to be compliant with the queries.

Submission instructions:

Project submission deadline:

- 11th April - 23:55
- No late submission is allowed.
- Submission after the deadline => 0

Groups rule:

- One project per student
- No group allowed
- Two students per a project => 0 for each student

Submission format and place:

- The submission should be in the form of a python notebook (.ipynb) (all the cells must be tested before submission)
- Only submission in the form of a python notebook will be considered
- The comments and explanations has to be added as a markdown cell
- The Entity Relationship Diagram has to be inserted in a markdown cell (.jpeg or .png file)
- All the queries have to be executed from the notebook, no query or manipulation should be done outside the notebook
- The notebook has to be self-contained and the cells in order, which means that when opening the notebook, we should be able to execute the cells one after another and get all the results.
- A particular attention will be offered to the quality of your code: Making functions instead of using a script style, adding comments, using the good functions, adding default values to some parameters …