Ciências ULisboa

Tecnologias de Processamento de dados - 2019/2020

Phase II (deadline: April 26)

Make a Python notebook containing the following sections

1. Dimensions and facts tables of the data warehouse
    a. Define and model them in SQL
    b. Identify hierarchies and fact granularity
    c. Create the dimensions and facts tables in the DBMS (postgreSQL)
2. Define an ETL workflow
    a. Identify all data sources for all dimensions. Add URL links to all data that should be available. If not public data, point to dropbox files, Google drive, or whatever
    b. For each dimension show the code used for modeling, filtering and inserting data
    c. Describe the process for inserting facts data
3. Do a critical assessment of the work
    a. Describe potential issues with the ETL procedure used
    b. Compare your schema to the one previously defined in phase I
    c. Discuss the issues for updating the data warehouse with novel data

The notebook should be produced with full output (that should not be longer than necessary!).

The Notebook can be written in Portuguese or English. It is to be submitted on the course’s Moodle **until 23:59 of April 26**. On the Section Zero it must contain the **group number, the Student Id’s and names, and an estimation of how many hours each student contributed to this phase**