Universidade Vigo ESEI Enxeñería Informática Escola Superior Ourense
UNIVERSIDADE DA CORUÑA fic facultade de informática da coruña
USC Universidade de Santiago de Compostela ETSE Escola Técnica Superior de Enxeñaría

Data Engineering
Master in Artificial Intelligence

Anália Lourenço (analia@uvigo.es)
Guillermo de Bernardo (guillermo.debernardo@udc.es)
Pedro Celard (pedro.celard@usc.es)

Dimensional modelling and ETL data processing

1. Software and tools

To complete the assignment, students will have to use

- A relational DBMS. The recommended software is MySQL and MySQL Workbench. Other DBMSs are acceptable if you have previous experience with them, but please check with your professor if you would like to use a different DBMS.
- A Python (3.x) installation. You may use libraries and an IDE of your choice.

2. Data sources

The goal of this use case is to implement a dimensional model to support the prediction of flight delays.

The main data source is a Kaggle dataset, available at the following URL: https://www.kaggle.com/code/fabiendaniel/predicting-flight-delays-tutorial/input

Students are strongly encouraged to check the documentation of the dataset available on Kaggle, to better understand the structure of the data. Notice that the dataset contains several CSV files that may be necessary: airlines.csv, airports.csv, flights.csv.

3. Tasks

3.1 Data analysis and modelling

Provide a description of the analytical goals that the model should address, and design a dimensional model that can be used to support those goals. The deliverables from this task include:

- A set of analytical objectives with a brief justification of their interest
- A diagram of the data model.
- An explanation of the data model, describing any relevant decisions taken to design it.
- A database script to create the tables in the DBMS of your choice.

3.2 Data pipeline

Design and implement the data engineering pipeline that populates the data model.
Deliverables from this task are:

- A high-level description of the pipeline, explaining the steps to read the source data, clean and transform the data and load the final data into the database.

- (Optional) Any description of the source or intermediate data, statistics or any other results that justify your design choices.
- The Python source code that implements the pipeline.
- A clear and concise description of how to run the whole pipeline.

4. Submission and evaluation

This lab practice should be conducted in **groups of 2-3 people**.

You are expected to submit the following artifacts for evaluation:

- A PDF report containing:
  - Explanations and diagrams required in each of the tasks.
  - Instructions on how to run the code.
- A .zip file with the complete source code and additional files required to run the whole data pipeline. This should include:
  - (Mandatory) Script to create the tables in the DBMS of your choice.
  - (Mandatory) Python files that implement the data transformations and load the data in the database.
  - (Optional) Any additional files that may be useful to set up and run the code (e.g requirements.txt, additional README files)

Keep in mind that the source code you provide must be execution-ready in the professor's machine. This means that:

- Your choice of DBMS, as well as any unconventional Python libraries used, **must** be clearly documented in your report.
- A small set of configuration parameters **may** be included (preferably in environment or configuration files). This includes the database connection parameters, and possibly the (relative path) location of the input files in the local machine. These configuration steps **must** be clearly documented in your report.
- The source code **must not** contain any absolute local file paths or other elements that prevent it from running in a different machine.
- In summary: make sure that your code can be executed in any other machine following your submitted instructions.

The deadline for submission is **September 28th, 2025, at 23:59.**

You should submit the contents using the available submission task in the online campus of your university.