Analyzing Flight Interconnected Data

Data Processing and Analytics (DPA) Course - International Master DISS UCBL 23/24

Please upload the files in one ZIP on TOMUSS (column TP3) by 27 Nov. 2023 midnight CET.

The objective of this project is to use Spark's APIs to analyze the flight interconnected data to find which are the most popular airports. To do so, you'll need to implement the **PageRank** algorithm natively on Spark SQL.

You can find the data (in csv format) here::
https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018/data

Development Environment:

For this project , please use the same environment you used for the TP related to Spark and Spark Streaming.

- Start the environment with **docker-compose up**

Requirements:

- You need to decide how to encode the dataset as a graph. What are the vertices? What are the edges? How do you compute the weight?
- Implement the **PageRank** algorithm natively on Spark using Spark SQL. In this case, you will have to use relational constructs to implement a graph and build the PageRank algorithm on top of these.

Expected Deliverables:

- A working Jupyter notebook that can be executed to perform the analysis. Please use the markdown cells of the notebook to document your code.