# Analyzing Flight Interconnected Data

**Data Processing and Analytics (DPA) Course - International Master DISS UCBL 25/26**

**Please upload the files in one ZIP on Moodle by 23ed Dec. 2025 midnight CET.**

The objective of this project is to use Spark's APIs to analyze the flight interconnected data to understand the popularity of the airports and flight patterns.

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 Dataframes and Spark Streaming.

- Download the graphframes docker container from moodle
- Start the environment with **docker-compose up**

**On alternative,** you find a shell script**:**

- Download the script from moodle
- Give required permissions to the script file and run it
- Start the environment with **docker-compose up**

# Requirements:

- Read the csv file and create a graph using Graphframes in the Spark environment.
- Compute different statistics : in-degree, out-degree, total degree and triangle count.
- Compute the total number of triangles in the graph.
- Compute a centrality measure of your choice natively on Spark using Graphframes.
- Implement the PageRank algorithm natively on Spark using Graphframes.
- Find the group of the most connected airports
- Find the most important airport
- Compare the results using different metrics (degree, centrality, PageRank)
- (Optional) Draw the heatmap of the flights between the airports.

Note: Graphframes provide functions for computing degree, pagerank score and triangles as part of the package, which should not be used for the purpose of this assignment. You can use the graphframes functions to compare your results.

# 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.
- A document (pdf) containing
  - An explanation of what you did
  - The results of the analysis for each task
  - Bonus point if you manage to use some visualizations (OPTIONAL)

- A user guide on how to run the notebook (e.g., which cells do what, or any other information that makes it easy for us to understand how to make your code work)

# Evaluation :

- 3 pts for the statistics
- 3 pts for triangles counting
- 3 pts for the centrality
- 5 pts for the PageRank
- 4 pts for the most connected airports
- 1 pt for the most important airport
- 1 pt for the comparison
- 1 pt (bonus) for visualizations
- 1 pt (bonus) for the quality of the report
- 22 total points

**p.s. Non commented code, unreadable notebook, short or incomplete report, project uploaded not in zip format will cause a reduction in the evaluation (-3 points).**