Laboratory 5 - Graph analytics with Spark GraphFrames

In this lab, we will learn to perform graph analysis applied to a Big Data framework. For this lab, you are required to **use GraphFrames Spark library with the API for DataFrames** (GraphFrames Spark library) to analyze a dataset containing information about flight connections and airports in the whole world. You have to read the data, build a graph, and perform different analyses on it.

In order to use the library in our cluster through a jupyter notebook at `jupyter.polito.it` , you have to select the kernel `GraphFrames (Yarn)` when creating a new notebook

1. Input Data

You will use a dataset containing informations about airports, airlines and flights world-wide. Consider these three `csv` files:

- `/data/students/bigdata_internet/lab5/airports.csv` : contains one line for each airport in the world. Among the others, it provides the columns: `id` , `name` , `city` , `country` , `iata` , `latitude` and `longitude` .
- `/data/students/bigdata_internet/lab5/airlines.csv` : provides some information for each airline. Among the others, it provides the columns: `airline_id` , `name` , `country` , `icao` .
- `/data/students/bigdata_internet/lab5/routes.csv` : enumerates the flights provided by each airline between two airports. Among the others, it provides the columns: `airline_id` , `airport_source_id` , `airport_destination_id` .

2. Top airports and airlines

Read the `csv` files to DataFrames, and answer to the following questions:

Task 1: Which are the countries in the world with more than 100 airports? Reports these countries and their number of airports (in descending order).

Task 2: Which are the Top-15 airlines by total number of flights? For each airline in the Top-15, provide airline name, airline icao code and number of flights (in descending order).

Task 3: Which are the Top-15 routes in the world? Routes are identified by airport source and airport destination, independently on the airline. For each route in the Top-15, provide the source airport name and city, the destination airport name and city, and the number of routes (in descending order).

3. Create the graph of fligth connections

Build a graph using GraphFrames where vertices are the airports in `airports.csv` , and edges are the flights from one airport to another contained in `routes.csv` .

Note: there are some missing values on flights. These are indicated with `\\\\N` either in the source airport or in the destination airport. Filter out the lines that contains these values, otherwise you will get an error.

Note: due to a bug in GraphFrames, vertex `id` , and edges `src` and `dst` columns should be converted to string for some algorithms to work. You can convert column `a` from e.g., `int` to `str` with:

```
df = df.withColumn("a", df.a.cast("string"))
```

4. Analyze and process the graph

Task 1: Show top-15 airports by in and by out degree. For the two results, provide the name of the airport , its ID and its degree (out and in) (in descending order).

Task 2: How many airports are reachable from Turin taking exactly 1 flights? What about taking exactly 2 flights? And exactly 3 flights? **Hint**: Use the motif finding functionality. Turin has `id = 1526`

Task 3: Compute the shortest path length from each airport in the dataset to Turin airport (`id = 1526`). Which are the 12 airports that are farther from Turin, in terms of number of hops? For each of these airports, report its name, its city and country, and the shortest path length to Turin (i.e., number of hops).

Task 4: Given Turin airport and the airport in the city of "Puerto Montt", compute:

- from how many airports in the world you can reach Turin using less hops than to reach Puerto Montt
- from how many airports in the world you can reach Puerto Montt using less hops than to reach Turin
- from how many airports in the world you can reach with the same number of hops Turin and Puerto Montt

Task 5: How many connected components of at least two airports are there in the graph? Report the number of connected components and their sizes. **Hint**: Drop the isolated vertices.

Task 6: Consider only the subgraph of the flights that are performed by two different airlines (identified by the ICAO), each involving at least 6 cities. Report the two chosen airlines and plot the subgraph of these flights. Report the name of the cities (of the airports) in the graph. **Hint**: use Graphvix and choose companies with a number of flights not too big.

5. Bonus Task

Task: Starting from the "Tancredo Neves International Airport" in the city of "Belo Horizonte" compute the maximum total distance you can flight taking exactly two fligths. Consider the total distance in kilometers (km), considering the distance from Turin to the first airport summed to the distance from the first airport to the second one. Use the constraint that you cannot come back to Turin (i.e., the second airport cannot be Turin).

Hint: Assume airplanes are always flying using the shortest path on a sphere (the Earth). Use the haversine formula to compute distance travelled with a flight in km (https://en.wikipedia.org/wiki/Haversine_formula).

How to write and submit the report

In your report, you must answer to all questions, report the code you have written, and show the output.

You are are required to comment each instruction (or group of instructions) - i.e., what it is the goal of the piece of code. You must follow the order in which questions and exercises are posed.

Your report must be a PDF. It can be directly generated from a Jupyter notebook. Go on `File-> Export Notebook As ... -> Export Notebook As PDF` . It must be submitted through the Teaching Portal course page (didattica.polito.it).

Naming convention: The file must be named with the following schema:

```
s<id>_lab<n>.pdf
```

For example, if you student ID is 123456 and you are submitting lab 5, the file must be: `s123456_lab5.pdf` Reports with a wrong file name will **NOT** be considered.

Report lenght: the report length must not exceed 10 pages, otherwise it will **NOT** be considered valid.