Lab 10

In this laboratory, we will use GraphFrames to analyze graphs. Specifically, we will 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.

To use the library on the cluster BigData@Polito, through a jupyter notebook at jupyter.polito.it, **you have to select the kernel “GraphFrames(Yarn)”** when creating a new notebook.

To use GraphFram locally on your PC run the following command if you installed Spark 3:

```
pyspark --packages graphframes:graphframes:0.8.1-spark3.0-s_2.12
--repositories https://repos.spark-packages.org
```

Use package `graphframes:graphframes:0.8.0-spark2.4-s_2.11` if you locally installed Spark 2.

Input Data

You will use a dataset containing information about airports, airlines and flights world-wide.

Consider these three csv files:

- /data/students/bigdata-01QYD/Lab10_DBD/airports.csv
    - It 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-01QYD/Lab10_DBD/airlines.csv
    - It provides some information for each airline.
    - Among the others, it provides the columns: airline_id, name, country, icao
- /data/students/bigdata-01QYD/Lab10_DBD/routes.csv
    - It 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.

Step 1 - Create the graph of flight connections

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

Note: There are some missing values in routes.csv. Specifically, there are some missing values either in the source airport or in the destination airport. Filter out the lines that contain these values, otherwise you will get an error.

Note: Due to a bug in GraphFrames, vertex id, and edges src and dst columns must be converted to string before applying some of the available algorithms. You can convert the data type of a DataFrame column by using the cast method. For example, suppose the input DataFrame df is characterized by Column id of type integer and you need to cast Column id to the data type string. The following code can be used to perform this cast operation:

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

Step 2 - Analyze and process the graph

Task 1

Show on the standard output the top-10 airports by in degree. For each of selected airports show the name of the airport, its ID, and in degree.

The transformation **limit(n)** can be used to select the first n records/rows from a Dataframe. The result is stored in a new Dataframe. limit(n) is similar to take(n) but it is specifically designed for managing Dataframes.

Task 2

The Turin airport has id = 1526

How many airports are reachable from Turin taking exactly 1 flight?
How many airports are reachable from Turin taking exactly 2 flights?
How many airports are reachable from Turin taking exactly 3 flights?

Print the results on the standard output.

Hint: Use the motif finding functionality of the GraphFrames library.

Task 3

Compute the shortest path length from each airport in the dataset to the Turin airport (id = 1526).

What is the farther airport(s) from Turin, in terms of number of hops?

For the selected airport(s), show on the standard output its name, its city and country, and the shortest path length to Turin (i.e., number of hops).

Task 4

How many connected components of at least two airports are there in the graph?