

Athens University of Economics and Business

MSc in Business Analytics

Assignment 2

Deadline: 3/7/2022

Group assignment (groups of up to 3 people)

The assignment corresponds to 25% of the total grade of the course.

Discussions between groups are recommended but collaborating on the actual solutions is considered cheating and will be reported.

There will be no extension of the assignment deadline!

Professor: Y.Kotidis ([kotidis@aueb.gr](mailto:kotidis@aueb.gr))

Assistant responsible for this assignment: I.Filippidou ([filippidou@aueb.gr](mailto:filippidou@aueb.gr))

## **Assignment Description: Neo4j Graph database**

### **Dataset:**

You are given a part of OpenFlights Airports network, which contains airports, airlines and flights between airports. In particular, the dataset contains 7698 Airports, 6161 Airlines, 6956 Cities, 237 Countries and 65.935 Flights between Airports. You can download the dataset (Airline Dataset) from moodle in csv format. The dataset contains 3 csv files and a description for the attributes in each file.

### **Property graph model**

You are asked to model the data as a property graph by designing the appropriate entities and assigning the relevant labels, types and properties. For your modeling, you need to study the details of all the files that describe the airline network and represent the required attributes on nodes and edges of a graph (Not all attributes in the csv files are required in the model). In your model you should include only the attributes that describe each node and edge type, without repetitions of elements (e.g. same property being displayed on both a node and an edge). Finally, nodes should not be connected when this is not required by the model.

### **Importing the dataset into Neo4j**

Based on your model, **you should create a graph database on Neo4j and load the airline network elements (nodes, edges, attributes)**. You can load the dataset directly from the provided csv files, by using either the neo4j browser or the neo4j import tool, or any programming language that is supported by neo4j. To speed up loading and query response times, you could also create proper indexes on your model properties.

## Querying the database

After the creation of your database, **you are asked to write and execute the following queries using the Cypher language.**

### Queries:

- 1) Which are the top 5 airports with the most flights. Return airport name and number of flights.
- 2) Which are the top 5 countries with the most airports. Return country name and number of airports.
- 3) Which are the top 5 airlines with international flights from/to 'Greece'. Return airline name and number of flights.
- 4) Which are the top 5 airlines with local flights inside 'Germany'. Return airline name and number of flights.
- 5) Which are the top 10 countries with flights to Greece. Return country name and number of flights.
- 6) Find the percentage of air traffic (inbound and outbound) for every city in Greece. Return city name and the corresponding traffic percentage in descending order.
- 7) Find the number of international flights to Greece with plane types '738' and '320'. Return for each plane type the number of flights.
- 8) Which are the top 5 flights that cover the biggest distance between two airports (use function `point({ longitude: s1.longitude, latitude: s1.latitude })` and function `distance(point1, point2)`). Return From (airport), To (airport) and distance in km.
- 9) Find 5 cities that are not connected with direct flights to 'Berlin'. Score the cities in descending order with the total number of flights to other destinations. Return city name and score.
- 10) Find all shortest paths from 'Athens' to 'Sydney'. Use only relations between flights and city airports.

## Assignment handout

Your deliverable should be a compressed file that you will upload to moodle and include:

### **1. Report.pdf**

- a. Detailed description of your graph model using a chart and a verbal description of the elements.
- b. The commands you used in order to import the files to the database.
- c. The Cypher code for the required queries with their **respective results**.

2. **The program/script you implemented:** for any step of this assignment.
3. **queries.cy:** A text file with the queries you expressed in Cypher language.