# MSc in Data Science

NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS" UNIVERSITY OF THE PELOPONNESE

**Data Management: Mini-project #2**

*Deadline: Fri Feb 16th 2023 EoD*

*Background:* Capitalizing the expertise you have gained in the previous mini-project, you decide to build a Neo4j-based data science infrastructure for a large online bookstore. After the requirement analysis you have decided that you can start with a graph that will include:

- Node types/labels:
  - **Book**, with properties: ISBN, title, publication year, language, and price
  - **Publisher**, with properties: ID, name, and country of headquarters
  - **Author**, with properties: ID, full name, biological gender, and nationality
  - **User**, with properties: username and email
- Edge types/labels:
  - **Published**, that connects Books with their Publishers (a Nx1 relation) and has no edge properties
  - **Authored**, that connects Authors with their Books (a NxM relation) that has the following additional properties: author order and role
  - **Ordered**, that connects Users with Books (a NxM relation) that has the following additional properties: order placement timestamp, order completion timestamp, and address of delivery.

ISBNs and usernames are considered to be unique and can identify Books and Users, respectively. One of your colleagues has already created one CSV file for each node and edge label, having the aforementioned properties (and the required IDs of the involved nodes for the edges) in the determined order. The name of each file has the following format:

"<Node/Edge label>.csv".

*Assignment 1:* Your first job is load the CSV files into Neo4j. You should write the appropriate Cypher expressions and you should deliver:

a) A text file containing the Cypher statements that will insert the corresponding nodes and edges (encoded in the CSV files) into Neo4j. Your file should have the name "xxxx-assignment1.txt", where xxxx is your student ID.
b) A small report briefly explaining your important design choices and rationale. The file should be in PDF format and should be named as "xxxx-report1.pdf", where xxxx is your student ID.

(40 points)

*Assignment 2:* After loading the data into Neo4j you are expected to run a first set of useful Cypher queries. You should deliver:

a) A Cypher statement that counts all the books in your database.
b) A Cypher statement that returns the IDs and the names of the authors with Greek nationality.
c) A Cypher statement that will retrieve all properties of all Books that have an author named "John Doe".

d) A Cypher statement that will retrieve the names of all Authors who have authored Books that have been ordered by a user having the “shybear” username.

For the previous queries you should provide a file named “xxxx-assignment2.txt” containing all the queries (again, xxxx is your student ID).
(40 points)

*Assignment 3:* A young data scientist is interested to help you in the project. Explain with as many arguments as you can which are the advantages of using a graph database system, like Neo4j. Include your response in a document named “xxxx-report3.pdf” (xxxx is your student ID).
(20 points)

NOTE: YOU SHOULD CREATE A COMPRESSED FILE CONTAINING ALL THE FILES OF YOUR ASSIGNMENT AND UPLOAD THEM ON THE ECLASS.

**GOOD LUCK!**

Thanasis Vergoulis