MSc in Data Science

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

Data Management: Mini-project #2

Deadline: Fri 19 Feb 2021, 00:00

Background: The e-shop that you have designed in your previous mini-project exceeded the expectations of the owners of the small comic shop! They had so many purchases during the previous period that they now want to give you extra budget to build a small data science infrastructure to analyze their customers data and help them setup a better advertisement and sales policy.

General requirements: After a first meeting with your clients, you have concluded that the following part of the database will be useful for the data science infrastructure:

- *Book data:* ISBN, title, authors, publisher, publication year, current price
- *Publisher data:* name, country of headquarters
- *Author data:* full name, biological gender, nationality
- *User data:* username, email, real name
- *Book orders data:* which user made the order, her address of delivery, the order placement timestamp, and the order completion timestamp.
- *Book reviews data: timestamp, score*

Of course, the corresponding primary keys and any relationships between the corresponding entities should also be included. Note that the requirements identified during the first mini-project remain valid.

Assignment 1: Before the final meeting with your clients (in which they will give you examples on the queries they are interested in) you are examining two alternative technologies for your project: MongoDB and Neo4j. You should deliver a small report briefly explaining the differences of the two technologies, their benefits and drawbacks, and the applications for which they are more appropriate. Your report should have the name "xxxx-report1.pdf", where xxxx is your student ID).
(25 points)

Assignment 2: After the final meeting with your clients, you realized that Neo4j fits better to the needs of your project. Thus, you should export all relevant data from your relational database (the one you had constructed and populated in the first mini-project) into CSV files and then use these files to load data in your Neo4j database. Your graph database should have (at least) the following node labels: book, author, publisher, user, order, and review. It should also have the corresponding edge labels that represent the different types of relationships between these types of nodes. You should deliver:

a) A file containing the SQL statements that can be used along with the \copy PostgreSQL command to export the required data in the properly formatted CSV files. Your file should have the name "xxxx-export.sql" (where xxxx is your student ID).
b) A text file containing the Cypher statements that can use as input the CSV files produced in the previous step to insert the corresponding nodes and edges into Neo4j. Your file should have the name "xxxx-import.sql".

c) A small report presenting a figure with your graph databases schema and briefly explaining your important design choices and rationale. The file should be in PDF format and should be named as “xxxx-report2.pdf”.

(50 points)

Assignment 3: After populating the graph database you need to prepare and run a batch of Cypher queries to reveal first insights about the customer data. You should deliver:

a) A Cypher query that returns all comic books published by a publisher with headquarters in France along with the corresponding authors.
b) A Cypher query that returns all users having at least 5 orders ordered according to the number of their orders (from the one with the most to the one with the least order).

For the previous queries you should provide a file named “xxxx-queries.sql” (xxxx being your student ID) containing all the queries, and a PDF document “xxxx-queries.pdf” that contains a screenshot from the Neo4j interface illustrating the results of each query (prefer the graph format, when applicable).

(20 points)

Bonus assignment: Write down something you like very much for this assignment and something you disliked. Include it in a file named “xxxx-feedback.doc” (xxxx is your student ID). You get all the points if you write here something relevant to the question regardless which your opinion is. If your comments are irrelevant or you do not provide any feedback you get no points.

(5 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