MSc in Data Science

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

Data Management: Mini-project #1

Deadline: Mon 14 Dec 2020, 00:00

Background: As a freelancer you agreed to undertake the job to design and implement a relational database for a small comic shop that due to the COVID-19 outbreak wants to build an e-shop.

Database requirements: The comic shop owners have told you that the database should contain all the required information for their e-shop to work:

- First of all, the database will be hosted in a PostgreSQL relational DBMS.
- For each comic book, the database should contain its ISBN, its title, its authors, its publisher, its publication year, a short description, and its current price in the shop.
- ISBN is a number with 10 digits and it is unique for each book.
- The title is a large string containing at most 200 characters.
- A book can have multiple authors, and each of them can author multiple books. Also, each author has a role for each book, that can be "Writer", "Illustrations", etc. It is likely that there are comic books with multiple authors sharing the same role. For each author, it is desired to also keep her nationality and biological gender.
- Each book has one publisher. For each publisher, it is desired to keep its name, its address, the country of its headquarters, and its contact phone.
- The publication year is the year during which the book has been published.
- The short description is a small text of unknown length (it can be a couple of paragraphs long).
- The price is a fixed-point number with two decimal digits and it represents the current price of the comic book in euros.
- The database should also contain information about book orders. In particular, for each book order, the database should keep the user that made it, her address of delivery, the order placement timestamp, and the order completion timestamp. Completion timestamp remains NULL until the completion of the order.
- For each user, the database should keep a username (that should be unique in the database), a password, an email (unique), the real name (first/middle/last name combined in one string field), a phone number, and multiple addresses (she may use different of them for different orders).
- Finally, the database should also include information about comic book reviews. In particular, each review is anonymous (i.e., is not directly connected to the users mentioned above) and keeps a nickname (a string up to 40 characters), a creation timestamp, a score (which is an integer between 1 to 5), and a small text (it can be a couple of paragraphs long).

Assignment 1: Design the database based on the provided requirements. You should deliver:

a) The SQL statements to create your database in a file named "xxxx-schema.sql", where xxxx is your student ID in MSc. You should also name all of your tables as "xxxx_[table-name]" (e.g., "38_book"). The file should be ready to be loaded in a PostgreSQL relational DBMS. Take extra care so that the database is well-designed (with proper primary & foreign keys) and follows all the requirements. Do not forget to implement all implied data constraints.

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”.

(40 points)

Assignment 2: The comic store owners want to include all book metadata and all relevant review information from the UCSD Book Graph dataset. All the information required is contained in the following files: file 1, file 2, file 3. Of course, these files contain more information than needed, hence you should keep only those data required based on the provided requirements. These files are in JSON format, however general-purpose tools to load JSON data in relational databases won’t work in this case, since UCSD’s JSON files do not follow your SQL schema. For attributes that are not present in the JSON files (e.g., info about the orders) use automatically generated integers/strings, NULL values, and DEFAULT values in a way convenient to you. This is why you are required to implement a proper Python3 script to parse the JSON files and create the SQL statements to include the required data into your database tables. You should deliver:

a) A file containing the SQL statements to insert the required data into the database. Your file should have the name “xxxx-data.sql” (where xxxx is your student ID). This file should be ready to be loaded in PostgreSQL.

b) The Python3 script you have used to create the file in 2a) having as name “xxxx-parser.py”.

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

(35 points)

Assignment 3: After creating the database and inserting all the relevant data you need to prepare and run a few useful SQL queries. You should deliver:

a) An SQL query that counts all the comic books in your database.

b) An SQL query that returns the average review score for the comic book with title “Feynman”.

c) An SQL query that returns the ISBNs and titles for all books authored by “Alan Moore” (in any role).

d) An SQL transaction that modifies a user’s order by removing a previous order with a new one of the same user.

For each of the previous queries you should provide a file named “xxxx-queries.sql” containing all the queries, where xxxx is your student ID.

(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 os 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