MSc in Data Science

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

Data Management: Mini-project #1

Deadline: Mon 5 Dec 2022, 20:00 EET (Greek time zone)

Background: As a freelancer you agreed to undertake the job to design and implement a relational database for the owners of a small bookstore that want to build an e-shop.

Database requirements: The bookstore 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 book, the database should contain its ISBN, its title, its authors, its publisher, its publication year, and its current price in the shop.
- ISBN is a code consisting of 10 digits and it is unique for each book. Although there are books without an ISBN (e.g., self-published books), the bookstore only focuses on books with ISBN, hence every book should have a such code.
- The title is a large string containing at most 200 characters. All books in the store should have a title.
- 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 books with multiple authors sharing the same role. For each author, it is desired to also keep their nationality and biological gender (for those cases that this information is available).
- 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 price is a fixed-point number with two decimal digits and it represents the current price of the book in euros.
- Finally, the database should also include information about book reviews. In particular, each review contains 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 "schema-xx.sql", where xx is your student ID in MSc. You should also name all of your tables as "[table-name]_xx" (e.g., "book_42"). 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 showing a diagram summarizing your database schema and briefly explaining your important design choices and rationale. The file should be in PDF format and should be named as "report1-xx.pdf".

(40 points)

Assignment 2: The bookstore 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: file1, file2, and file3 (the book store will load only children books, at this point). 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, you need to use a JSON parsing library to read the files line-by-line and get the data you need. For attributes that are not present in the JSON files use automatically generated integers/strings, NULL values, and DEFAULT values in a way convenient to you. Finally, keep in mind that the average rating values should not be stored in your database.
For this assignment, you are required to implement a proper Python3 script to parse the JSON files and create the "INSERT INTO" 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 for the first 100 tuples into the database. Your file should have the name "data-xx.sql" (where xx 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 "parser-xx.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 "report2-xx.pdf".

(20 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 modifies a book's publication year to "1950" for a book having the code "0434961604" as ISBN.
b) An SQL query that counts all the books in your database.
c) An SQL query that returns the ISBNs, titles, and publication years for all books authored by "Antoine de Saint-Exupery" (in any role) having a price lower than 20 euros.
d) An SQL query the ISBNs, titles, publication years, and author name for all books having a price lower than 2 euros.
e) An SQL query that returns the average review score for the book with ISBN "0434961604".
f) An SQL query that returns the ISBN and titles for all books having average review score greater than 4.5 and at least 3 reviews.

For the previous, you should provide a file named "queries-xx.sql" containing all the queries, where xx is your student ID.
(40 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