Database Systems – Spring 2014
DPT Group, Aalborg University
Handed out: 31.03.2014

Teacher: Katja Hose
Self study 5: mini-project part 4
Deadline: 06.04.2014

The overall goal in the whole mini-project is to create and work with a movie database in five incremental steps. The first step is to identify the information that is going to be represented as well as special requirements. Then, the information is modeled in an ER diagram and mapped to a relational schema. The database schema is then refined and normalized, so that it can be instantiated in PostgreSQL. After filling database with data, it will be queried and optimized.

Filling a database with data and formulating queries

In the previous selfstudy you have refined and normalized your design. In this selfstudy, the main task is to fill your database with data and query it.

The database filled with data will will be the basis for the following self study.

A very common problem when working with databases in companies is that at some point data provided by third parties needs to be integrated. The original format could be an existing data base with a different schema or DBMS, Excel, text, files, etc.

Assume that you want to[1] use PostgreSQL, e.g., because of features such as window functions or recursion, reliability, security, implementing the SQL standard, low costs, etc.

Thus, in short the problem for this self study is creating a database in PostgreSQL, populating it with data provided as an SQL dump originating from MySQL, and finally of course querying your database.

1 Filling a database with data

IMDB data dump

- You can download the IMDB data dump (a subset of all IMDB data) from Moodle (`imdb.sql.gz`). Please note that the data dump is only provided for use within the course; see files found at `www.imdb.com/interfaces` for terms of use.
- This dump is a MySQL dump that cannot directly be loaded into PostgreSQL.
- Thus, there are two obstacles to overcome (i) the dump is not compatible with PostgreSQL without adaptations and (ii) the schema of the dump does not match the schema you have created.

Hints

- Both, MySQL and PostgreSQL, have command line tools[2] that often work much better when working with dumps: `mysql` and `psql`.
- You need to unzip the dump first and work with `imdb.sql`[3].

[1] . . . or because due to company policies you have to. . .
[2] Yes, also for Windows!
[3] Moodle might have double-zipped it for your download.

- PostgreSQL is working with a default user “postgres”.
- You cannot directly import the dump into PostgreSQL because MySQL uses a slightly different SQL syntax. This means you have to edit the statements to make it compatible or find a tool that does the conversion for you – the latter is likely to cost more time.
- You do not necessarily have to import all information contained in the dump as long as the queries in part 2 of this self study can be executed.

Steps

(a) Find out what data is contained in the dump and how it is structured.

(b) Find an appropriate solution to populate your database with the provided IMDB data.

Report

- Explain in detail how you managed to set up and fill a PostgreSQL database with the schema you have developed in the previous selfstudy.
  The explanation should be detailed enough so that a computer scientist is able to understand what you did and reproduce a database with the same content by following your description – you can skip the part that explains how to install the DBMS, of course. In case there was no need to adapt your local schema, you also do not need to repeat the CREATE TABLE statements that were included in the previous report.

2 Querying your database

Now that your database has been populated with data, the main task now is to work with your database and query it.

Please find SQL statements expressing the following queries and execute them on your database[4].

1. How many Danish language movies are in the database?
2. For each year, what is the total number of reviews to movies from that year?
3. Which movies have John Travolta and Uma Thurman starred in together?
4. How many actors and directors have a first name starting with “Q”?
5. How many users rated at least 3 movies?
6. What is the name and birth year of all actors in “Pulp Fiction”? Your query should list the actors in increasing order of birth year.

[4] These are the same queries that you have already seen in self study 2. Hence, your database should contain all necessary information.

7. What are the titles and years of all movies from the 1980s that John Travolta starred in?
8. What are the top-2 highest rated movies (average) from the 1990s according to the users?
9. What are the top-2 highest rated movies (average) from the 1990s according to at least 2 users?
10. In 1994, what was the average rating of a movie for each language?
11. Which actors in Pulp Fiction have never, before or after, starred in the same movie as one of the other actors in "Pulp Fiction"?
12. Which movie starring John Travolta has the highest user ratings?
13. How many actresses have not been alive at the same time as Charles Chaplin?
14. What is the average rating of movies from each genre?
15. What is the average rating of movies from each genre? List only genres with at least 2 ratings.
16. Which movie has the largest number of 2-link references? (If A refers to B, and B refers to C, then we say that A has a 2-link reference, through B, to C. If there are several paths leasing from A to C, we count all of them.)
17. How many actors have also been active as director of at least one movie?
18. Which two genres are most often linked to the same movie? (Note that each movie has a set of genres.)

Steps

(a) Formulate the above queries in SQL in consideration of your database schema and evaluate them on your database.

(b) Optional: If you would like to add some personal ratings of movies, feel free to do so.

Report

- Optional: if you have inserted additional information into your database: please document it, for instance as a list of insert statements.
- List of SQL statements corresponding to the above queries and the results of evaluating them on your database (at most the first 10 tuples).

Course goals covered by this self study

- Make use of SQL to create, modify, and query relational databases

Selfstudy: 31.03.2014
The report must be handed in via Moodle no later than
06.04.2014, 23:55 CET