CSE 4/560 Data Models and Query Language Semester Project

Instructor: Dr. Sreyasee Das Bhattacharjee

February 6, 2021

1 Overall Project Objective

Build up a database to demonstrate interesting searches. Up to 2 students are allowed in a team.

1.1 Task

The task details are as follows:

1. Select one interesting usecase domain, building your database using SQL. It should be relatively substantial, but not too enormous. Several project ideas are described at the end of this document. However, these ideas are just to support you start thinking, and you are encouraged to come up with your own choice of usecase. Please keep the following points in mind: a) while real datasets are highly recommended, you may also use program-generated "fake" datasets if real ones are too difficult to obtain; b) How are you going to use the data? What kind of queries do you want to ask? How is the data updated? Your application should support both queries and updates.

2. Design the database schema. Start with an E/R diagram and convert it to a relational schema. Identify any constraints that may be applicable in your usecase problem and implement them using database constraints. If you plan to work with real datasets, it is important to go over some real data samples to validate your design (in fact, you should start Task 7 below as early as possible, in parallel to Tasks 3-6). Do not forget to apply database design theory and check for redundancies.

3. Create a sample database using a small subset by hand to facilitate debugging and testing because large datasets make debugging difficult. It is a good idea for different scripts to create/load/alter/update/destroy the sample database automatically.

4. Acquire the large "production" dataset, either by downloading it from a real data source or generating it using a program. Make sure the dataset fits your schema. For real datasets, you might need to write programs/scripts to transform them into a suitable form for loading into a database. For program-generated datasets, make sure they contain interesting enough "links" across rows of different tables to show the results of different Advanced SQL queries learned in class.

5. You are required to make sure all of your relations are in Boyce-Codd normal form. Provide a list of dependencies for each relation. Decompose them if the tables are not in BCNF. If you decide to keep it in 3NF instead of BCNF, justify the decision for a particular relation. Your report for this milestone should contain a separate section with the details of the transformation from the initial schema to the final schema where all the relations are in BCNF. This file should also contain all the functional dependencies you started with.
   Note: This is quite possible that your initial schema is already in BCNF, and in that case, you need to provide us the functional dependencies and convince us that the relations are already in BCNF.

6. Do you specifically run into any problem while handling the larger dataset? Did you try to adopt some indexing concept to resolve?

7. Test your result with the smaller sample database first. You may need to iterate the design and implementation several times in order to correct any unforeseen problems.

8. **Bonus Question [20]** Query execution analysis: identify the queries (show their cost) implemented in this project, where the performance can improve. Provide a detailed execution plan (you may use EXPLAIN in PostgreSQL) on how do you plan to improve these queries.

1.2 Expected Result & Grade Distribution

1. Presentation or Demo: record a 5-10 minutes video about your project's demo to UBLearn. The demo has to be a live demo, not screenshots. We will then fix each checkpoint's demo dates after the corresponding deadline.

2. Project Report(group): write the report and state clearly the contribution from each team member. It should be 4-6 pages in ACM format (https://www.overleaf.com/latex/templates/acm-conference-proceedings-master-template/pnrfvrrdbfwt). More details to follow within the milestone details.

3. Please submit your demo video, database SQL dump, and all other files within a zip file. The report should be separately uploaded as a pdf file.

4. Implementation/demo 60% and report 40%.

2 Deliverables

2.1 Milestone 1: Due date: 03/12/2021 [100]

You are required to hand in a report which contains the overview of your project proposal. The overview can change slightly as we go over the course, but the main theme should be intact. The proposal should consist of two or more pages describing the problem you plan to solve, outlining how you plan to solve it, and describing what you will "deliver" for the final project. Your report should contain the following sections:

1. Project details: Name of your project, your team, and all team members, everyone's UB id(not the UB number);
2. Problem Statement: Describe the problem that your proposed database system will solve. Why do you need a database instead of an excel file?
3. Target user: Who will use your database? Who will administer the database? You are encouraged to give a real-life scenario;
4. Task 1-4 should be complete, and you should start planning for tasks 5-7. The detailed description and demonstration of your work on each of these tasks should be presented in the Project description and demo presentation video.
5. Any code for downloading/scraping/transforming real data that you have written for data acquiring should also be reported if applicable.

Save your report as Milestone1.pdf and upload it to UBLearn. Only the team leader (Member 1) need to submit the file. For example usecases, please see the section 3; while you may want to choose your own use-case, it should be equivalently detailed.

NOTE: Define a list of relations and their attributes.

1. Indicate the primary key and foreign keys (if any) for each relation. Justify your choice;
2. Write the detailed description of each attribute (for each table), its purpose, and datatype;
3. Indicate each attribute's default value (if any) or if the attribute can be set to 'null';
4. Explain the actions taken on any foreign key when the primary key (that the foreign key refer to) is deleted (e.g., no action, delete cascade, set null, set default).

2.2 Milestone 2: Due date: 04/30/2021 [100]

This is your final submission of the project. The complete report should be there.
Your complete report should contain:

1. Details of all the tasks that you are required to perform in this project. Please highlight any new assumptions, E/R diagram, and list of tables (if they have changed since Milestone 1 that you have added/edited).

2. Create a file *create.sql* which will create all the tables in your database. Load these relations from data files (tab or comma-separated files). The tab or comma-separated files can be created by you (dummy values) or other sources. Create a *load.sql* file for bulk loading. Create a *readme.txt* file that states your data source. Put *create.sql*, *load.sql*, all the '.dat' files (or *.csv* files, or data files in any other format) and a *readme.txt* file into a sub-directory.

3. Make a separate section for **Query Execution Analysis** if you plan to attempt.

4. Demo Video, as mentioned in **Expected Result**

Save your final report as *Milestone2.pdf* and upload it to UBLearn. Only the team leader (Member 1) need to submit the file.

Final Project Demo. You will need to present your system's working demo at the end of the semester. Instructions on how to sign up for the demo will be given during the second to last week of the class. You are also encouraged to stay in touch with the TAs (Keyan and Zhenyi) to discuss your project and get their feedback on how to improve.

3 Example Application Scenarios:

Please follow the links for the details.

- IMDb makes their movie database available http://www.imdb.com/interfaces.
- historical stock quote can be downloaded and scraped from many sites such as Yahoo! and Google Finance.
- Data.gov http://www.data.gov/ has a huge compilation of data sets produced by the US government.
- The Supreme Court Database (http://scdb.wustl.edu/data.php tracks all cases decided by the US Supreme Court.
- US government spending data (http://usaspending.gov/data) has information about government contracts and awards.
- Federal Election Commission (http://www.fec.gov/disclosure.shtml) has campaign finance data to download; their "disclosure portal" (http://www.fec.gov/pindex.shtml) also provide nice interfaces for exploring the data.
- Google Fusion Table (http://www.google.com/fusiontables/Home/) hosts quite a number of datasets of public interest. It is a good place to find datasets or data sources to work on, and you can consider using it as a method of hosting your data for public access.