CSC 370 – Fall 2017 – Projects
Due: Nov 29, 2017 – e-submission only

Here are two possible projects for this course that you can choose from. You should submit a report describing your work (schema, queries, charts, etc. about 10-15 pages) as well as give a short presentation (5 min) outlining the main points of your project. The projects can be conducted in groups of two, three, or four people.

Project 1.

This is an exploratory project. You are encouraged to collect interesting data sets for an application domain that interests you. The more data you collect the better it is for finding interesting patterns. Your project should consist of the following three phases:

Data Collection (for the domain you like/are interested in)
Data Preprocessing (data transformation into a useful form)
Schema Design (using E/R)
Data Analysis (using SQL queries similar to those you have seen so far)
Data Visualization (using Excel or any other software you prefer)

At least one of these phases should be not trivial. For example, the data collection and preprocessing phase could be non-trivial (the data needs special preprocessing). Or the data analysis queries and visualization could be non-trivial (e.g. queries and charts show some interesting, non-obvious patterns).

Project 2.

Consider one of the following databases:

Musicbrainz (SQLite) about artists, releases, tracks, etc, (until 2012), which can be downloaded from: http://webhome.cs.uvic.ca/~thomo/mblite_post.rar
Even though this is a stripped down version of the http://musicbrainz.org/ database, it still is a big database. It contains 20 tables, some of them with several million tuples.
The schema of the database is:
http://webhome.cs.uvic.ca/~thomo/schema.jpg
and a file with create table statements and explanatory comments is here:
http://webhome.cs.uvic.ca/~thomo/schema_with_comments.txt

Baseball, Hockey, Basketball from
http://www.opensourcesports.com

These are large datasets about these sports spanning many decades since the beginning of the previous century. The datasets are in CSV format. You should load them in your database of choice.

SP 500 database
http://webhome.cs.uvic.ca/~thomo/sp500clean.sqlite
The database includes information about the companies in the SP 500 index. The details about their stock prices for about two years, 2014, 2015 is given in table History. Table SPY contains price information for the index as a whole.

Your project should consist of:

Data Analysis (using SQL queries similar to those you have seen so far)
Data Visualization (using Excel, or other spreadsheet software)

You should have at least the following number of SQL queries depending on the size of the group:

10 for 1-person group
15 for 2-people group
20 for 3-people group
25 for 4-people group

The queries should extract useful information from the database with the goal of exploring and analyzing the database. Some of the queries should be accompanied by charts as appropriate. Be creative and try to derive interesting conclusions from the data.