CSCI 620/Section 02/Mior, Introduction to Big Data, Spring 2195

Project description

The goal of this project is for students to practice the concepts studied in the course exploiting different datasets than the ones used in class. The project consists of three phases:

- Phase I: Select one or more datasets to be used in the rest of the project. You may use existing datasets, or may compile your own. The final dataset needs to be large (~50M tuples in a relational database), and interesting enough so you can perform meaningful queries and mine meaningful information from it. You need to provide the dataset (contact the instructor on how to share your dataset if, in zip format, it exceeds 100Mb), a description of the data, a meaningful relational model to faithfully represent the dataset, and a program to load the dataset.
- Phase II: Propose a document-oriented model for your dataset and compare it with your relational model. Provide a program tha, issues at least five interesting SQL queries over the previous relational model and propose indexes to speed up query execution (report your timings). Discover functional dependencies and study normalization with respect to the relational model you provided in Phase I.
- Phase III: Provide a program that cleans and integrates your dataset and applies itemset mining to discover interesting association rules. Briefly describe the steps taken by your program. You need to elaborate which model is a better fit for this task (relational or document-oriented), so your program should also perform the necessary translations.

Final presentations

Prepare a presentation of your project with 15-20 slides including comments. Note that other teams will evaluate your slides based on these comments as well. You have complete freedom for this task about format and contents: You need to make sure to talk about the inner details and concepts, but you also need to make it attractive to other students. The quality of the final presentation should be comparable to demo presentations at international conferences. See Foofah and DBLOC for examples.

Feedback

You need to provide feedback about the performance of your teammates and two other team presentations. These will be the presentations of the group numbers before and after you. (Also, group 9 will evaluate groups 1 and 8. Group 1 will evaluate groups 2 and 9). The feedback will consist on a grade out of 100 points and comments: Around 100 words for teammates, and 500 words for other team presentations.