## CSCI 620/Section 02/Mior, Introduction to Big Data, Spring 2241

# 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 a link to the dataset, a detailed description of the data, a meaningful relational model to faithfully represent the dataset, and a program to load the dataset. (Note that phase III of the project will involve data cleaning, so you do not need to be overly concerned with cleaning your data at this stage.) Some suggestions on where to find datasets are given in the Resources section of myCourses.
**Do not include your dataset in your submission.**

### Phase II

Propose a document-oriented model for your dataset and compare it with your relational model. You should aim to make effective use of the document model as discussed in class. This means your model will likely be different from your relational model. You should also provide code to load your data into this model. Your report should include details on the structure of each collection you create.

Provide a program that issues at least five interesting SQL queries over the previous relational model and propose and explain indexes to speed up query execution. Report your timings before and after indexing. Discover and explain functional dependencies and discuss normalization with respect to the relational model you provided in Phase I.

### Phase III

Provide a program that cleans your dataset. Consider possible issues with dirty data as discussed in class and explain in your report how you addressed these issues. You should also apply frequent itemset mining and association rule 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 be able to translate data to the chosen model where necessary. You do *not* need to implement the above algorithms using both models, but only the final one you select.

*(continued on next page)*

## Final presentations

Prepare a presentation of your project approximately 15-20 minutes in length. 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. You should aim to explain your dataset as well as the specific insights you found while completing each phase of your project. Note that it is not necessary to explain concepts taught in class since your classmates should already know this material.

## Feedback

You need to provide feedback about the performance of each of your teammates and any two other team presentations. The feedback you submit should consist of a grade out of 100 points and comments: around 100 words for each teammate, and 250 words for each other team's presentations. **Each group member will *individually* submit feedback written as a single PDF file.**