CMPSC 312
Database Systems
Spring 2019

Final Project Assignment:
Advanced Topics in Data Management
Submit deliverables through your assignment GitHib repository.

Objectives

To produce a project where databases are integral part of the design and implementation. To gain experience working with real-world applications where database systems are used.

GitHub Starter Link for Groups

STOP! STOP!
Not everyone will be clicking this link!
Only the team leader will be clicking the link to create the repository!!
`https://classroom.github.com/g/hKjGN3lC`
To use this link, please follow the steps below.

Pre-Introduction

Since this is your first team-based assignment we will be using a group assignment functionality of GitHub Classroom. For group assignments **only one person will be creating the team while the other team members will join that team.**

The selected person of the team should go into the link to the lab in the assignment sheet. Copy this link and paste it into your web browser. Now, you should accept the laboratory assignment and create a new team with a unique and descriptive team name (under "Or Create a new team").

Now the other members of the team can click on the assignment link and select their team from the list under "Join an Existing Team." When other team members join their group in GitHub Classroom, a team is created in our GitHub organization. Teams have amazing functionality, including threaded comments and emoji support. Every team member will be able to push and pull to their teams repository. Your teams project manager should be the one to resolve any conflicts or merge pull requests.

Please work in groups: Unless you provide the instructor with documentation of the extenuating circumstances that you are facing, not working in a team and not accepting the assignment means that you automatically receive a failing grade for it.

To use the GitHub link, please follow the steps below.

- Click on the link and accept the assignment.
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.
- Clone this repository (bearing your name) and work on the lab locally.

- As you are working on your lab, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):
  - `git commit <nameOfFile> -m ‘‘Your notes about commit here’’`
  - `git push`

  Alternatively, you can use the following commands to add multiple files from your repository:
  - `git add -A`
  - `git commit -m ‘‘Your notes about commit here’’`
  - `git push`

Introduction

Throughout the semester, we learned much about the construction and usages of databases. During our class and lab time, we discussed concepts, theoretical mechanisms, management and general technologies which bring databases to life in academia and industry. The topics were applied to structured query language (SQL and NoSQL-based systems), the implementation of database applications (i.e., using Python and Django), and also the creation and parsing of XML databases and files. During this work, you learned the differences between database types and gained a deeper understanding of what types of data may be used, while ensuring an efficient database where you involved: integrity constraints, attributes, table dimensions, queries and etc. In this final project, you are to combine all your knowledge of database construction, with a research question, to create a database that will be able to provide answers for your research.

Overview

This project invites you to sufficiently explore an advanced topic in databases which will be helpful to storing and ordering your data for your research. You will first create research question(s) which can be answered from database queries (for example, statistically-focused questions or other types). You will use collected data to determine queries which will address your overall research. This data must come from real-life or has be gathered from public databases where it is not artificial. For example, data may be found at the Allegheny College Library Scholar Databases (found at the following address: `https://allegheny.libguides.com/az.php`) or other online sites such as the *World Health Organization*, `http://www.who.int/gho/en/`, *UniProt* `http://www.uniprot.org/` and similar types of outlets.

Group Work

It is suggested that this project be completed in teams (of two to three people in a group), however, you may still choose to work alone. Be sure that one of the people in your group acts as the “spokesperson” for your group who will communicate with the instructor and submit the deliverables for the project. Be sure that each member has a copy of the deliverables for his or her own records.

You and your group must decide on all factors of the project: research topic, data, data source, database, database design, construction, implementation, querying and interpretation of results

to answer the research question. Once collected you will have to then choose the database technologies to process your data (i.e., storing and querying to be able to respond to your research question). These are all decisions which must be discussed between your group members due to their importance to your work.

Description of Project

Your project will concern a responding to research questions which you will use a database to answer. **Choose project which presents several questions (at least five) for which the data you collect will allow you to answer.** Remember, that this assignment is not so much about being able to analyze data to resolve some issue, but rather, it is to showcase your ability to organize collected data so as to be in the position to perform such an analysis. Your project is therefore to respond to your research questions by use of a database system to address the necessary research to understand or resolve the problem.

At least Five Research Questions for your Queries: Choose a research problem and locate your source of data. Then choose at least five intelligent questions to answer from your collected data concerning your research topic. Remember, your work in queries should be sufficient to determine responses to your five research questions.

Examples of Research Questions: An example of a general research problem is to determine whether films that yield high earnings are actually inexpensive to make. Another example may be to determine whether students who get regular exercise are generally more successful in school (have higher GPA) than students who do not get regular exercise. To answer these kinds of questions, correlations or other methods may be applied if you wish, however the queries should be the primary tool to produce the solution to each question.

Report: Your project should result in a detailed report written in Markdown. The report should include a description of the research question and explain its relevance to databases. You must discuss the your data as it was found, what you had to do to be able to use it (i.e., cleaning data) along with a reference to where and when it was obtained. The report must also fully discuss the implementation steps and types of usage and querying of your database. Please spend time to argue why your technology and database software was appropriately applied to respond to your research question.

Deliverables

You are to submit your report, code and data for your final project deliverables. The written report should be precise, formal, appropriately formatted, grammatically correct, informative, and interesting. The source code that you write to build your database must be carefully documented and tested. If you install and use a data management framework, the steps for installation and use should be clearly documented. Your data should be the cleaned version. Please note: if your data set is over 5 megabytes, please consider compressing it into a zip file for your submission.

Your project (and its deliverables) must conform to the following guidelines.

1. **Report:** You are to submit a report document that details the project. For this, the five (related) research questions are to be discussed along with the methods (i.e., data, software packages and etc.) that your group utilized to address the research question. Your report will

contain screen shots, plots and other visual tools with discussion to explain your conclusions. This is a technical document and will have the same types of sections as other technical documents that you have come to know during your exploration in the literature. In the past, final project reports from this class had between 15 to 20 pages.

2. **Data:** You are to use real data which you have either, amassed from experimental techniques (i.e., you acquire your data by your own designs), or has been obtained from public repositories (mentioned above) where the data has been collected from real events.

3. **Database software:** Your project must be written for one of the software packages that we have discussed in class. You may use SQL, NoSQL, Python, Django, SQLBrowser, XML or any of the software packages that we have discussed in class.

4. **Schema:** Once you have acquired your data, you must build the database. The schema of how your database will be created is extremely important to your project and must be graphically displayed in your final report. Each table must be discussed in the report to justify its existence in your project. If you are using a database that does require building a schema, then please provide some *map* to show how the data is formulated (i.e., a depiction of the form of the data as it was entered into your database).

5. **Source Code:** Any code must be submitted for the instructor to evaluate. Please include comments and print statements, where necessary, to help with reading the code and to provide details of its execution.

6. **Queries:** In your project, you will be using this database to extract meaningful information (patterns) from your data to explain some phenomenon. Exactly which patterns to find will be left to you and your group. However, these patterns must be well thought-out and relevant to the theme of your project. For this task, queries must be used and each query is to be outlined with a short justification about why the particular query was used in your project.

7. **Interpretation of Information from the Queries:** Your written project must reiterate the research question and give your formalized conclusion which is based off your queries of the database containing your data.

8. **Testing Your Database:** In your report, you must provide some detail about how you know that the information form the queries is actually correct. Argue for these details by clearly explaining (justifying) why the returned details were correct, in light of your project research question.

9. **Interface:** The user-side of your database will contain tables, not visible query-programming code, to help the user navigate and use the database. For this reason, Django is highly recommended as the interface to use for the results display. If you choose not to use Django, then you must use some other graphical interface (HTML, plots, slides, formatted pages containing tables, etc) to describe the results of your queries. Imagine that this lay-out will be the only connection to your database system. Please include screen-shots of your interfaces in your report.

10. **References:** Please be sure to cite your facts in your work. For each claim that you make which is not necessarily public knowledge, you ought to have some reference to a primary,

peer-reviewed article, to support your statement. If you are mentioning a novel tool, please search for its supporting article to add to your reference section and tie this reference to your discussion of the tool.

Final Project Deadlines

This assignment invites you to submit the following deliverables:

1. **Status Update Due Date**: Friday, 15 April 2019. The file `writing/statusUpdate.md` will contain your writing about the following details.

   (a) The subject of your project
   (b) The main questions are you that you are addressing,
   (c) The citations of your data and some leading references to support the project
   (d) The steps that you have already taken
   (e) The step left to complete the project
   (f) The names of the group members with their assigned tasks to the project

   Please be specific and provide a complete list names of all group members in this document so that your instructor can give credit for this work.

2. **Presentations Due Dates**: 19-24 April 2019.

   During the last lab and class sessions teams will present their projects to the class audience. Please note that your project does not need to be complete at this time, however, you should have enough of the project implemented so that you have something to demonstrate during your presentation. Each team should prepare a few slides explaining the overall project and its research questions. You should also highlight all database related concepts (i.e., schema) and software used in your project and the data you have obtained. Finally, you should give a demonstration of your preliminary database system. Your talk should not be longer than about 8 minutes.

3. **Final Project Due Date**: Monday, 6 May, 2019 at 9am.

   On this date, your group spokesperson should submit the final version of your project, in the Git repository. Make your directories clear: use *finalReport/* to contain the final report. Use *featuredWork/* to contain all other files of your project: including all of the relevant source code and output, and any additional materials that will demonstrate the success of your project. Students must submit the completed assignment before 5:30 pm on the due date. **This due date cannot be extended for any reason.** Please see the instructor if you have any questions about this assignment.

In adherence to the Honor Code, students should complete this assignment individually. While it is appropriate for students in this class to have high-level conversations about the assignment with other class members, it is necessary to distinguish carefully between an individual who discusses the principles underlying a problem with others and the student who produces an assignment that

is identical to, or merely a variation on, the work of someone else. As such, deliverables that are nearly identical to the work of others will be taken as evidence of violating the Honor Code. Students should contact the course instructor with questions about this course policy.