DATA 514 Final Project Specification

Learning Goals

Data Management

Students will exercise database design and query execution skills on a real-world dataset. Skills from multiple units during the course will be required, including writing queries, designing schema and indexing, DBMS selection, and data governance.

Database Communication

Students will practice using formal tools (ER diagrams) and presentation skills to communicate fluently about database design and DBMS use.

Teamwork

Students will complete this project in a team requiring the use of internal team communication and workload balancing.

Overview

This project will be completed in teams of 2-3 students. (Students may self-select teams. No single person teams permitted.) Each team will select a real world dataset with which they have some interaction. Over the course of a couple of weeks teams will

- Describe this dataset in detail
- Identify data governance issues with this dataset
- Design a schema for storing this dataset
- Write >=3 example use cases with accompanying queries
- Analyze a query's logical plan
- Identify any indexes that support the use cases above
- Identify an appropriate DBMS system for use with this dataset
- Note: Teams do not need to execute the above with a full dataset and DBMS

Datasets

Teams may identify any dataset in which they hold an interest. This may be a personal dataset, such as an MSDS lending library, or it may be an industrial dataset, such as ORCA card usage. It is not necessary for a team to have access to the actual data, although they should be able to generate some example records. If a team chooses to use an existing dataset they must ensure that they have publication rights to the data.

Acceptable datasets will have at least four relations (tables), with at least one one-to-many and one many-to-many relationship.

Grading

This project will be graded according to the point values specified below. All together, the project is worth 120 points, weighted in the final grade at 10%. The project presentation and the slide deck will be graded as team products, with the same score being assigned to each team member. The reflection assignment will be graded individually.

Deliverables

A presentation will be delivered in which each team member will present some aspect of the design and the team will field appropriate questions (**100 points**). This presentation will be graded based on presentation quality and the accompanying slide deck. The slide deck will also serve as documentation for the overall project (it is possible, but not required, that slides beyond those presented are included in the deck for further documentation).

Pre-work (**5 points**, graded collectively)

Students will submit their team choices and selected topic area approximately 2 weeks before the presentation date. However, students may choose to begin work on their presentation before this due date, if they desire more time.

Presentation (**30 points**, graded collectively)

Teams will have 10-15 minutes to present their work. Teams are expected to collaborate on presentation. Grades will reflect

- Quality of visual aids (**10 points**)
- Balance of presentation responsibilities (**5 points**)
- Appropriate use of technical terminology (**5 points**)
- Clarity of communication (**5 points**)
- Response to questions (**5 points**)

Slide Deck & Documentation (**70 points**, graded collectively)

- Describe this dataset in detail (**10 points**)
    - Summary description of data
        - Be thoughtful about datatypes
    - Likely audience and use
- Identify data governance issues with this dataset (**10 points**)
    - What is sensitive information in this dataset?
        - How will you protect this information?
        - If the data is re-used in the future, what concerns are there?

- Design a schema for storing this dataset
    - Include an ER diagram (**10 points**)
- Write >= 3 example use cases with accompanying queries
    - Include text description of the use case (**6 points**)
        - Including frequency
    - Include sample query language (**12 points**)
- Analyze the logical plan of a query
    - Present the "naive" RA trees for >= 3 of your queries (**6 points**)
    - For at least one query, present an alternative RA tree (**4 points**)
    - Briefly address any expected performance differences between the two trees (**2 points**)
        - You do not need to have a full-tree cardinality estimate, but you should be able to justify the differences, if any
- Identify any indexes that support the use cases above (**5 points**)
    - Include text description of reasoning
- Identify an appropriate DBMS system for use with this dataset (**5 points**)
    - What are you taking into consideration with this proposal?

Reflection (**15 points**, graded individually)

A reflection document will capture an individual's experiences in this project. These are due at the same time as the presentation and its related documents.

Timeline

Team and topic selection due by `May 23, 2024`.* (**5 points**)
Final presentations take place on `Jun 4, 2024` 5-9pm. (**30 points**)
Documentation, presentation materials submitted by `Jun 6, 2024`.* (**70 points**)
Reflection assignment (individual) due by `Jun 6, 2024`.* (**15 points**)

* Canvas assignments for collection