University of Nebraska Lincoln

Data Modeling for Systems Development
CSCE 411/811

Programming Assignment 4

Spring 2020

Build An End-to-End E-R Data Analysis Workflow

Basic Info

The programming code will be graded on **both implementation and correctness**.

Submit a written report as a PDF file to address the questions in the handout. Please use question number to document your response on the report. You don't need to use the IEEE format.

Assignment Goals

This assignment is intended to build the following skills:

- Building web-based interactive data analysis system
- Understand the three types of data modeling techniques for Database design

---

Score Distribution

Conceptual Modeling: 411 & 811: 25 pts
Logical Modeling: 411 & 811: 10 pts
Physical Modeling: 411 & 811: 30 pts
End-to-End E-R Workflow: 411 (25 pts) & 811 (35 pts)

Total: 411 (90 pts) & 811 (100 pts)

This assignment is designed to help you gain a deeper understanding of E-R (Entity-Relationship) modeling. You will design and implement a mini data query system using E-R modeling.

Dataset:
In real-world settings, a dataset may not be directly used with a relational database management system (RDBMS). Different datasets can be provided with different formats (e.g., text, binary, images, and so on). In order to use these datasets in RDBMS, **data processing and modeling techniques** are typically required.

You are provided with a dataset in a zip archive file (*Assignment4_Dataset_Code.zip*). After decompress, there are about 2,000 data files, named from *record_000000.dat* to *record_001999.dat*. Each data file corresponds to a synthesized record containing User ID, User Name, User Location, Number of Messages, and Message information (Send Time and Send Text) in a binary format.

For an illustration purpose, 4 source code files (*process_record.c, read_record.c, record.c* and *record.h*) are provided in the zip file to show **how to process these binary data files** using C under Linux. You can certainly use any other programming language(s) and operating system(s) in this assignment.

i) *record.h* defines the record format in each data file.
ii) *record.c* provides the (not fully optimized) functions to access the content of each data file.
iii) *read_record.c* prints the content of a record. To build this program, type: *gcc -o read_record read_record.c record.c*. It will generate an executable named read_record. To run the program, type *./read_record record_number*. For example, running ./read_record 999 will print the record of User ID: 999. The program also shows the processing time at the end of its output.
iv) *process_record.c* provides a framework to process all records within a range of record numbers. To build this program, type: *gcc -o process_record process_record.c record.c*. To run the program, type *./process_record 14 223*. In the example, the program simply goes through all records from 14 to 223.

Note that *read_record* and *process_record* need to be executed in the same directory as the dataset.

Database Design: These raw dataset cannot be directly imported into RDBMS yet without E-R (Conceptual) Modeling and Logical Modeling.

Conceptual Modeling (411 & 811: 25 pts)

1. Provide your E-R diagrammatic representation(s). **[10 pts]**

2. Explain the reasons leading to your design. **[5 pts]**

3. How does your E-R diagrammatic representation help to **comprehend** the business requirements and **verify** your design? **[5 pts]**

4. Give at least one alternative E-R diagrammatic representation and compare it with your first representation. **[5 pts]**

Logical Modeling (411 & 811: 10 pts)

5. Provide your tables in relational notation and explain the reasons leading to your design. **[5 pts]**

6. Did you do any normalization? Why or why not? Justify. **[5 pts]**

Physical Modeling (411 & 811: 30 pts)

7. To populate your MySQL database first you need to store your data for each table into separate CSV files. For example, if there are 3 tables in your database, then there should be 3 CSV files that contain the data. Write scripts to extract data from the binary files and store those on the respective CSV files. You may use any programming language. You need to submit the codes and the CSV files. **[20 pts]**

8. Create your tables in MySQL and import data from CSV files into your MySQL database. Submit the sqldump file. **[5 pts]**

9. In the report document your approach for the physical modeling. **[5 pts]**

End-to-End Data Analysis Workflow (411 & 811: 25 pts)

10. For the end-to-end workflow the 411 students are not required to perform data analysis and visualization. Just create a **web interface** such that a user could input following queries through the webpage and then simply displays the query results in the text format. You can use PHP or any web programming language. You need to submit your code for the web application.

[15 pts]

11. Conduct the following four queries in through this simplified workflow. **[10 pts]**

- Find all users of Nebraska. Record the total number of user records that meet the requirement. Record the processing time without printing the records on your screen. (But you can certainly print them for verification.)
- Find all users who sent messages between 8 am-9 am. Record the total number of user records that meet the requirement. Record the processing time without printing the records on your screen. (But you can certainly print them for verification.)
- Find all users who sent messages between 8 am-9 am ***from Nebraska***. Record the total number of user records that meet the requirement. Record the processing time without printing the records on your screen. (But you can certainly print them for verification.)
- Find the user who sent the maximum number of messages between 8 am-9 am ***from Nebraska***. Record the User ID whose record meets the requirement. Record the processing time without printing the records on your screen.

[**Following question is Mandatory for 811 and Extra Credit for 411**]

12. Create at least one visualization on the data stored in your database. Use the D3 JavaScript library for the visualization. On you web page there should be a with a button/link with meaningful name, upon clicking on which the visual representation should be displayed. Submit your code for visualization.

[10 pts]