CSC2141 Project Part 3

Designing and Building Your Database

Due date: Monday, April 8, 11:59PM

Overview and Grading

In the first part of the course assignment, you assembled a dataset comprising at least four tables that will be used for design and query purposes. In the second part you designed and created your database and provided a few basic database queries.

- Part 3a is **optional** and only needs to be completed if you wish to replace your grades on Parts 1 and 2 of the assignment. If your score on Part 3a exceeds the aggregated score of Assignment Parts 1 and 2, the Part 3a score can replace this 15% of your final grade. It is scored out of 30.
- Part 3b is **mandatory**, is scored out of 25 and worth 10% of your final grade. Please refer to the main assignment document for general rules and guidelines.

The assignment components are as follows. Details are provided on the following pages.

Read the rubric at the end! It outlines the expectations.

3a (optional): A **single PDF file** with the following information

(1) Introduce and describe your dataset
(2) Show the internal model of your database
(3) Show a dependency diagram of each table in your database, explaining why each table is in 3NF

3b (mandatory): A **.zip file or .tar.gz** archive containing the following:

(1) **SQL file #1:** A **single** file that contains all your data definition statements (CREATE TABLE / VIEW, INSERT INTO / LOAD DATA statements)
(2) **SQL file #2:** A **single** file that contains a stored procedure
(3) **SQL file #3:** A **single** file that contains the SQL queries specified below.
(4) Any **source files** you used in LOAD DATA statements
(5) A **documentation file** in .PDF format

Part 3a: Describing Your Database (optional)

This part has three sections. It mirrors some of the tasks from Parts 1 and 2, but also has some important modifications.

Section 1: Overview of Your Dataset (10 points)

Provide a description of your dataset, about **two pages in length**. We will not be strict about the submission length, but half a page is definitely too short, and five pages definitely too long.

- *What is the dataset?* Describe the data you have obtained (including the parts you may have generated), answering the following questions:
  o What is the source of the data? Provide a URL or other reference. Please also <u>clearly state the license information</u>.
  o What have the data been used for in the past? Has anyone else done anything useful or fun with the dataset? Provide citations / URLs where appropriate.
  o Provide complete details of how you generated any simulated part of your data, including scripts, ChatGPT prompts, etc.
- *What do you plan to do with the dataset?* Explain the key questions you would like to ask using SQL operations on the dataset.

Section 2: Description of Your Tables (10 points)

For each of your tables, provide the following information:

- The name of the table
- The attributes it contains, including brief explanations of any attributes that are particularly important or non-obvious
- Any primary keys / foreign keys, including brief explanations of how the foreign keys link the tables to each other
- The dimensions (rows and columns)

Section 3: Internal Schema and Normalization (10 points)

Note that these figures <u>must not be hand drawn</u>; they must be generated using tools such as the MySQL reverse engineering tool, diagrams.net, PowerPoint, etc.

- Internal schema, including all tables, attributes, datatypes, relationships including cardinality information
- Dependency diagrams for each table, with primary keys identified, and arrows showing determinant / dependent relationships
- Explanation of why these tables are in 3NF, or cannot be converted to 3NF.

To submit: A single PDF file, text in a legible font (e.g., 12 point Times New Roman), with reasonable margins (e.g., Word or LaTeX default, or narrow)

Part 3b: Using the Database (4 sections)

At long last, it is time to submit the database and any supporting data (if necessary). We will attempt to build your database, run your supplied queries, and test your stored procedures. Our intention is to test these in MySQL; if for some reason you used a different DBMS or made other nonstandard alterations or have a large file as part of your dataset, <u>please let us know before the deadline</u>.

Section 1: Everything needed to construct the database (10 points)

The first part of your submission is an .sql file that contains everything needed to construct your database. This includes all the DDL statements (which will probably be CREATE TABLEs) as well as the DMLs that populate those tables with data using INSERT INTO and/or LOAD DATA statements.

Requirements:

- The SQL file must run from start to finish, so for example if we load it into MySQL Workbench and hit Ctrl-Shift-Enter, your entire database should be created.
- We should also be able to successfully **re-run** the script, so it should contain the appropriate DROP TABLE IF EXISTS statements and anything else needed to "clean" the database before it is re-run.
- Every statement (CREATE TABLE, etc.) must be commented.

Section 2: Stored procedures (5 points)

You must submit a second .sql file that contains **two stored procedures and CALL() statements to invoke them**. These must satisfy the following requirements:

- They must contain at least two separate DML statements that update, insert, and / or delete data in a related way, for example adding a row to a table and updating every other row in the table, or deleting a row if it is no longer needed.
- They should use IN, OUT, or INOUT parameters
- They should have basic transaction control to ensure that the entire sequence of statements either completes successfully or is rolled back.
- The file should also contain at least one CALL() statement for each of the stored procedures. Since the procedures need to have at least one parameter, you will need to declare / use variables in this scope as well.
- Each stored procedure must be commented.

Section 3: SQL queries (5 points)

The third .sql file **must** contain five queries.

Query 1. SELECT from a single table with a WHERE clause, producing a derived attribute.

Query 2. A NATURAL, INNER, or OUTER JOIN between two of your tables.

Query 3. A query covering one or more tables that uses a GROUP BY statement on at least one of your variables.

Query 4: A query that makes use of at least one subquery in the FROM clause.

Query 5: A sequence of queries that:

- Creates a VIEW from two or more tables, including derived attributes
- Runs a SELECT query on the view
- Modifies one of the underlying tables
- Re-runs the SELECT query on the view, reflecting changes in the underlying tables and the derived attributes

Section 4: Documentation (5 points)

The final task of the assignment is to write documentation for your database. This should consist of the following:

- A brief summary of the database: data source, license information, number of tables, number of attributes.
- Document at least **three** business rules that are enforced by your database. Explain how these rules are expressed as table constraints.
- Briefly explain the five queries you have submitted. What is their purpose?
- Explain how to use the stored procedures. The reader should be able to understand enough to know what information needs to be provided in the CALL(), what the procedure does, and what (if any) information is returned by OUT or INOUT variables.

Rubric (Assignment 3a):

	Excellent (100%)	Very Good (80%)	Acceptable (60%)	Borderline (40%)	Unacceptable (0% - 20%)	Notes
Section 1: Dataset Overview	All necessary elements are present. Descriptions are clear and professionally structured. Descriptions are also sufficient to allow others to retrieve and recreate the datasets. Normalization descriptions are accurate.	All necessary elements are present. Descriptions are clear and professionally structured. Descriptions are also sufficient to allow others to retrieve and recreate the datasets. Normalization descriptions are accurate.	All necessary elements are present. Descriptions have minor clarity and structural issues. Descriptions are lacking some details necessary to reproduce the dataset. Normalization descriptions are mostly accurate.	Some required elements are missing. Descriptions have significant clarity and/or structural issues. Descriptions are lacking significant details necessary to reproduce the dataset. Normalization descriptions are inaccurate / contain omissions.	Many required elements are missing. Clarity and structure are poor or lacking. Descriptions are not useful for dataset retrieval and generation. Normalization explanations are missing or wrong.	
Section 2: Table Descriptions	All requested components are present. Descriptions of important attributes and keys are comprehensive and clear.	All requested components are present. Descriptions of important attributes and keys are clear.	All requested components are present. Some important attributes and keys are not described but descriptions are still clear.	All requested components are present. Many important attributes and keys are not described, and descriptions are lacking clarity.	Some requested components are present. Descriptions of important attributes and keys are rudimentary and unclear.	
Section 3: Internal Schema and Normalization	The diagrams adhere to the stated requirements. Presentation of elements (primary keys, relationships, dependencies, etc) is comprehensive and correct.	The diagrams adhere to the stated requirements. Presentation of elements (primary keys, relationships, dependencies, etc) is comprehensive and correct, with minor omissions.	The diagrams adhere to the stated requirements, with minimal deviations. Presentation of elements (primary keys, relationships, dependencies, etc) is correct, but with some important elements unclear or missing.	The diagrams are missing important elements to the stated requirements. Presentation of elements (primary keys, relationships, dependencies, etc) has significant omissions.	Diagrams are poorly laid out and missing critical elements (attributes, dependencies, etc). Elements are largely incorrect and/or missing.	Hand-drawn diagrams will be assigned a score of zero.

Rubric (Assignment 3b):

	Excellent (100%)	Very Good (80%)	Acceptable (60%)	Borderline (40%)	Unacceptable (0% - 20%)	Notes
Section 1: Database construction file	All database creation SQL statements can be run multiple times without errors. SQL statements are correct and have a consistent style. Statements have comments that clarify their purpose.	All database creation SQL statements can be run multiple times without errors. SQL statements are correct and have a largely consistent style. Statements have comments that clarify their purpose.	All database creation SQL statements run once correctly, re-running may generate errors. SQL statements are correct and have some style inconsistencies. Statements have comments that clarify their purpose.	All database creation SQL statements run once correctly, re-running may generate errors. SQL statements are correct with inconsistent style. Comments are inadequate.	Creation script fails to run. SQL statements are a mess. Comments are absent.	
Section 2: Stored procedures	Stored procedures meet requirements in terms of structure, number of queries, and transaction control. CALL() statements execute correctly.	Stored procedures meet most requirements in terms of structure, number of queries, and transaction control. CALL() statements execute correctly.	Stored procedures are missing some elements of structure, number of queries, and transaction control. CALL() statements execute correctly.	Stored do not adhere to required structure, number of queries, and transaction control. CALL() statements execute only with modifications to input / output parameters.	Stored procedures meet requirements in terms of structure, number of queries, and transaction control. CALL() statements fail to run.	
Section 3: SQL queries	All five queries meet the requirements, execute correctly, and are well documented. Style is consistent.	All five queries meet the requirements, execute correctly, and are well documented. Style is consistent.	Most queries meet the requirements, execute correctly, and are well documented. Style is largely consistent.	Some elements are missing. Style is inconsistent and a minority of queries fail to run.	Several elements are missing. Most or all queries show inconsistent style and fail to run.	
Section 4: Documentation	Database summary, queries, and stored procedure documentation is complete, well structured, and clear.	Database summary, queries, and stored procedure documentation is mostly complete, well structured, and clear.	Documentation is not sufficient to understand all aspects of the database, queries, and stored procedures.	Significant deficiencies in the documentation structure and clarity.	Documentation is poorly structured, unclear, or absent.	