INF0 6210
Data Management and Database Design
Physical Data Model and Normalization
Assignment 2

Professor: Nik Bear Brown

Due February 22, 2020

In this assignment, you are assumed to be working for a company called Nerd Analytics and that you are completely in charge of the database. You will create a physical database, normalize, actively stream social media data to the database you created in Assignment 1. You can choose to create a new database as well.

Another group of statisticians and machine learning experts will be using the data that you model, gather, clean and database analyze Social Media for a particular domain (e.g. Games, Film, Databases, Cartoons, Baseball, Pokemon, Music, etc.). Each domain must have entities that represent consumers, producers and companies. For example, for games one must be able to model gamers, game developers and game companies. For music one must be able to model music lovers, musicians and music companies.

This assignment can be done in pairs, or individually.

Database normalization

Database normalization is the process of organizing the attributes and tables of a relational database to minimize data redundancy.

Normalization involves refactoring a table into smaller (and less redundant) tables but without losing information; defining foreign keys in the old table referencing the primary keys of the new ones. The objective is to isolate data so that additions, deletions, and modifications of an attribute can be made in just one table and then propagated through the rest of the database using the defined foreign keys.

Edgar F. Codd, the inventor of the relational model (RM), introduced the concept of normalization and what we now know as the First normal form (1NF) in 1970. Codd went on to define the Second normal form (2NF) and Third normal form (3NF) in 1971.

Design Requirements

You will check whether your tables are in First normal form (1NF), Second normal form (2NF) and Third normal form (3NF). If not, you'll restructure your database so that all of the tables are in Third normal form; that is, you normalize the database.

First normal form (1NF)

- Each table has a primary key: minimal set of attributes which can uniquely identify a record
- The values in each column of a table are atomic (No multi-value attributes allowed).
- There are no repeating groups: two columns do not store similar information in the same table.

Second normal form (2NF)

- All requirements for 1st NF must be met.
- No partial dependencies.
- No calculated data

Third normal form (3NF)

- All requirements for 2nd NF must be met.
- Eliminate fields that do not directly depend on the primary key; that is no transitive dependencies.

Conceptual Model

Design Requirements

This can be an extension of your first conceptual model or a new one.

Your submission must include:

A domain (e.g. games, film, databases, cartoons, etc.)
Conceptual models (entities) for a tweet/post, a Social Media user, a person, and a company.
Conceptual models (entities) that represent consumers, producers and companies in your chosen domain.
Conceptual models (entities) for at least two things specific to the domain (e.g. a game, a film, a song, etc.)
Relationships that connect the entities.
Appropriate attributes and keys.
ER diagrams that illustrate the entire conceptual model.
The ER diagrams can use standard ER symbols or UML.
Physical models (i.e. SQL and data) that implement your conceptual models.

Questions

Questions you must answer about your conceptual model:

1. What are the ranges, data types and format of all of the attributes in your entities?
2. When should you use an entity versus attribute? (Example: address of a person could be modeled as either)
3. When should you use an entity or relationship, and placement of attributes? (Example: a

manager could be modeled as either)
4. How did you choose your keys? Which are unique?
5. Did you model hierarchies using the "ISA" design element? Why or why not?
6. Were there design alternatives? What are their tradeoffs: entity vs. attribute, entity vs. relationship, binary vs. ternary relationships?
7. Where are you going find real-world data populate your model?

Questions you must answer about your physical model:
1. Are all the tables in 1NF?
2. Are all the tables in 2NF?
3. Are all the tables in 3NF?

Social Media Account

You need a Social Media account (e.g. Twitter, Facebook, Instagram, etc.) It is recommended that you create a Social Media account separate from your personal one for this class as it will be used for interacting with the Social Media API.

Audit

- Audit Validity/ Accuracy
- Audit Completeness
- Audit Consistency/Uniformity

You must clean the data or show that it doesn't need cleaning (10 points)

Write a report explaining all of the files, the tests and their results and code. (30 points)

Professionalism

- Consistent Naming Scheme
- Tests
- Commenting & Documentation
- File and Folder Organization
- Separation of Code and Data
- Use Efficient Data Structures and Algorithms
- Keep Your Code Portable
- Contribution statement

Design Requirements

Your submission must include:

- Sample data from every source
- A conceptual schema explaining the data relations
- Any code and scripts you used
- A report explaining all of the files, the tests and their results and code

Scoring Rubric

- (10 points) You must find sources of data
- (20 points) Create a conceptual schema
- (20 points) Create a physical schema
- (10 points) You must download and reformat the data fit your conceptual schema
- (10 points) Audit
- (20 points) Report
- (10 points) Professionalism

Submission of Assignments

You will submit your assignments via Blackboard ***and*** Github. Click the title of assignment (blackboard -> assignment -> <Title of Assignment>), tgtthe submission page. You will know your score on an assignment, project or test via BlackBoard. BlackBoard only represents only the raw scores. Not normalized or curved grades. A jupyter notebook file ALONG with either a .DOC or .PDF rendering of that jupyter notebook file must be submitted with each assignment.

Multiple files must be zipped. N.RAR, .bz, .7z or other extensions.

Assignment file names MUST start with students last name then first name OR the groups name and include the class number and assignment number.

Assignment MUST estimate the percentage of code written by the student and that which came from external sources.