ECE 656 Winter 2020: Assignment 1

Due: 28th January at 9:00 PM

Setup a MySQL Server instance, per the discussion in the tutorial and the setup notes on Piazza. If you are having any difficulty with this, ask questions sooner rather than later, because without this, you will not be able to even start this assignment.

Create a database on your server, with a name of your choosing; since this assignment will be using the Sean Lahman Baseball data, I will assume that the name of the DB is BASEBALL.

Sean Lahman has created a sizable database of baseball statistics, with a detailed description of the data in the "readme" under the Assignment 1 module on Learn. You will need to familiarize yourself with this readme. You can create the necessary tables and load the data using the SQL source file "lahman2016.sql"; however, that file drops existing tables, including any data. Therefore, there are two variants of this SQL source: one "-tables" simple creates the tables, with the other "-data" assumes the tables exist and inserts the data.

1. Cleaning up the database: The database table creation commands may exhibit warning messages; ("show warnings" will display all warnings from the last command executed); if there are any warnings caused by those table-creation commands, you should modify the lahman2016-tables.sql source so as to ensure there are no warnings. Further, the SQL file is missing both primary and foreign keys.

- Determine appropriate primary and foreign keys needed for the baseball database.
- Modify lahman2016-tables.sql to add the primary and foreign keys to this database.
- It is possible that in doing this there may be a conflict with some of the baseball data and/or there is missing data, which means that the primary and/or foreign keys may prevent you from inserting some of the data. There are three ways you can resolve this problem. What are they (one sentence each, maximum)? What is the necessary SQL to implement the solution in each case.

2. The SQL file has a very large number of INSERT statements in order to load the data into the database. It is typically preferred to load data directly from source files. In the case of the Baseball data, the course files are "Comma-Separated Variable" (or CSV) files. Create a LOAD statement that will load the data for the Batting CSV (Batting.csv) into its associated table. You should verify that your LOAD statement operates correctly and issues no warnings.

- Where is the CSV data located relative to the CLI and to the DB Server?
- Time how long it takes to LOAD the CSV vs. Using the equivalent INSERT statement method.

3. Create RA and SQL queries to answer each of the following questions:

(a) How many players have an unknown birthdate?

(b) How many people are in the Hall of Fame? What fraction of each category of person are in the Hall Of Fame? Are more people in the Hall Of Fame alive or dead? Does this vary by category?

(c) What are the names and total pay (individually) of the three people with the three largest total salaries? What category are these people (Players? Managers? Other?)? What are the top three by category?

(d) What is the average number of Home Runs a player has?

(e) If we only count players who got at least 1 Home Run, what is the average number of Home Runs a player has?

(f) If we define a player as a good batter if they have more than the average number of Home Runs, and a player is a good Pitcher if they have more than the average number of ShutOut games, then how many players are *both* good batters *and* good pitchers?

Submission:

For any code questions you should upload the relevant .sql source code to the Assignment 1 DropBox on Learn. For question 1, this will be your modified version of lahman2016-tables.sql" for question 2 this should be "loadBatting.sql" and for question this should be "baseballQueries.sql"

In addition, written answers, as required, should be submitted as a PDF with the name "assignment1.pdf"