# ECE 656 Winter 2019: Lab 1

Due: February 8th at 7:00 PM

You have been given a MySQL account with on your University of Waterloo UserID, `userid`, as the account ID on the machine "marmoset04.shoshin.uwaterloo.ca" and an associated database, `db656_userid`. To connect to the database server use the following command:

```
mysql -u <userid> -p -h marmoset04.shoshin.uwaterloo.ca
```

This connects you to the database server[1] and starts the MySQL Command-Line Interface. You can connect to your specific database with the CLI command:

```
use db656_<userid>;
```

You may wish to use this database server for Part 1 of this lab. You will need to use it for Part 2 of the lab.

**Part 1**: Baseball Data: Sean Lahman has created a sizable database of baseball statistics, with a detailed description here: http://seanlahman.com/files/database/readme2016.txt. The data at that site is available in two forms, CSV files and `.sql`, though we will provide local copies as necessary. This database comprises 27 tables, with some tables having over 100,000 rows, and more than 20 columns. For this initial lab, you will begin to familiarize yourself with this dataset.

1. Create SQL queries to answer the following questions:

(a) How many players have an unknown birthdate?
(b) Are more players in the Hall of Fame dead or alive? (Output the number alive minus the number dead)
(c) What is the name and total pay of the player with the largest total salary?
(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?

2. The `SQL` file has a very large number of INSERT statements in order to load the data into the database. The CSV files, by contrast, have no associated `SQL` code to load the data into the database. Create a LOAD statement that will load the data for the Fielding CSV (Fielding.csv) into its associated table. You should verify that your LOAD statement operates correctly and issues no warnings.

---

[1] The campus firewall blocks access to port 3306 from outside the campus and so you will need to use the campus VPN if you are trying to access the system from off campus.

3. The SQL file is missing both primary and foreign keys (which is just as well since some of the data causes problems when such keys are included, as you may have discovered when doing the previous questions). We will assume that the baseball database is sufficiently normalized that we do not want to change the basic set of tables. However, we do want to add primary- and foreign-key constraints.

(1) Determine the primary and foreign keys needed for the baseball database
(2) Write the necessary SQL to add the primary and foreign keys to this database. When doing this, keep in mind the fact that there is missing data, as you may have noticed from the previous questions, and so your SQL will need to address this issue.

Part 1 Submission: Write all of the above queries in a single file titled `baseball.sql` and submit that file to the Assignment 1 Dropbox on `Learn` by February 8th at 7:00 PM. In addition, write the answers to the queries for Question 1 and any information for Question 3 about how you resolved the issue of missing data for primary and/or foreign keys.

**Part 2**: Yelp Data: Yelp is a website that maintains consumer reviews of businesses. Its data is very similar to that within a data warehouse: a small number of tables with a very, very large quantity of data. Yelp regularly provides access to a small subset of that data together with a schema for it (https://www.yelp.ca/dataset_challenge). As with the Baseball data, in this lab you will primarily be familiarizing yourself with this data.

Create SQL queries to answer the following questions:

(a) Which user has written the greatest number of reviews?
(b) Which business has received the greatest number of reviews?
(c) What is the average number of reviews written by users?
(d) The average rating written by a user can be determined in two ways:
    a. By direct reading from the Users table "average stars" column
    b. By computing an average of the ratings issued by a user for businesses reviewed

    For how many users is the difference between these two amounts larger than 0.5?
(e) What fraction of users have written more than 10 reviews?
(f) What is the average length of their reviews?

Part 2 Submission: Write all of the above queries in a single file titled `yelp.sql` and submit that file to the Assignment 1 Dropbox on Learn by February 8th at 7:00 PM.