ECE 356 Winter 2017: Lab 1

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, `db356-userid`. To connect to the database server use the following command:

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

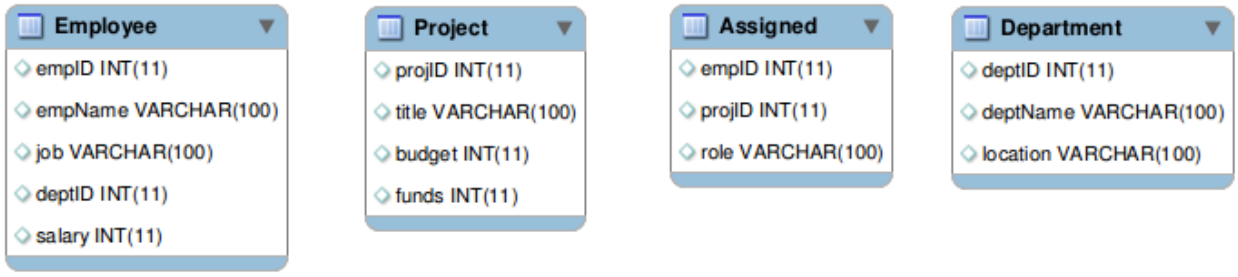
This connects you to the database server[1] and starts the MySQL Command-Line Interface. You can connect to your specific database with the CLI command:

```
use db356-<userid>;
```

Part 1: Basic SQL with few tables and a small quantity of data: using the createTables.sql code, you will create four tables with a small amount of data in them. To execute the createTables.sql code, use the following CLI command:

```
source createTables.sql;
```

The schema for the set of tables is as follows:



1. Single-Table Queries: Create SQL queries to answer the following:

(a) Which departments are in located Waterloo?
(b) How many people are employed in each job types in the company, ordered by job type?
(c) What are the names and salaries of the engineers?
(d) What is the average salary by job type?
(e) What is the ID of the department with the most engineers?
(f) What is the ID of the department where the greatest percentage of employees in the department are engineers?

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.

2. Multi-table Queries:
(a) What are the names and IDs of employees who are currently not assigned to any project?
(b) What is the name, job and role of all employees whose role is different than their job?
(c) What is the number of employees, by job, who's role is identical to their job?
(d) What is the total of the salaries of all employees assigned to each project?
(e) Repeat (d) but include a row for the total of the salaries of all employees not assigned to any project.
(f) Identify any employees (if any), by name and ID, who are assigned to more than one project.

3. Data insertion, deletion, and update: Create the necessary SQL to achieve the following:
(a) Raise the salary of anyone working on the compiler project by 10%
(b) We wish to raise the salary of all janitors by 5%, and all Waterloo employees by 8%; if there are any janitors in Waterloo, their pay should be raised by 8%, not by 13%.
(c) Close the Kitchener location. Anyone currently assigned to a project should be moved to the Waterloo location. Anyone not currently assigned to a project should be deleted from the Employee table. Delete the Kitchener location from the Department table.

Write all of the above queries in a single file titled `basicSQL.sql` and submit that file to `http://marmoset03.shoshin.uwaterloo.ca/` to project `356-1-basicSQL`. Submission must be within two weeks of your scheduled lab date.

Part 2: Baseball Data: Sean Lahman has created a sizable database of baseball statistics, with a detailed description here: http://seanlahman.com/files/database/readme2014.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 24 tables, with some tables having over 100,000 rows, and more than 20 columns. For this initial lab, we will begin to familiarize 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.

Write all of the above queries in a single file titled `baseball.sql` and submit that file to `marmoset03` to project `356-1-baseball` within two weeks of your scheduled lab.

Part 3: 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?

Write all of the above queries in a single file titled `yelp.sql` and submit that file to `marmoset03` to project `356-1-yelp` within two weeks of your scheduled lab.