**Data Engineer Coding Home Assignment**

This exercise consists of developing a distributed Extract-Transform-Load (ETL) application.

Your application should ingest the data from a source relational database system and use Apache Spark to compute some statistics and output them in a form that can be loaded into some destination storage system for consumption.

Please write your application in Scala. The application must be buildable from the command line; it should not require an IDE to build or run.

The exercise should generally take 3-4 hours to complete, and should not take more than 3 days at most to complete. If you don't finish within this time frame, that's okay; submit what you've got anyway.

### **Source database**

Build the source database by provisioning an instance of MySQL or equivalent. Use this SQL script to create the database schema and load it with seed data. Note that if you choose a different RDBMS than MySQL, you may need to edit the SQL script to make it work with the chosen system.

The source database is based on Sean Lahman's baseball database.

* <https://www.dropbox.com/s/hflxqja7xq87ffv/lahman2016-sql.zip?st=dohz4qvi&dl=0>

### **Extract data and compute statistics**

The application should extract data from the source database and perform the following computations on the extracted data:

1. **Calculate the average salary** for infielders and pitchers for each year.
2. **Calculate the number of all-star appearances** for each Hall of Fame pitcher and their average ERA in their all-star years, and list the year they were inducted into the Hall of Fame.
3. **Calculate the top 10 pitchers' average** regular season and post-season ERAs and average win/loss (w/(w+l)) percentages.
   * i.e. The top 10 pitchers' ERAs ((0.65(player1) + 0.72(player2) + ...) / 10) and (win/loss of player1 + win/loss of player2 + ...)/10
   * The pitchers in the top 10 may not have been on a team that made it to the post-season, so average the ERAs & win/loss of the pitchers that made it into the post-season.
4. **List the first and last place teams** and their number of at-bats for each year.

**N.B.** These computations should be performed in code using the extracted dataset. They should not be performed as SQL queries against the source database.

### **Prepare data for loading**

The application should format the output dataset into a number of CSV files as specified in the below examples and upload them to a storage system of your choosing (e.g., S3) for daily download by consumers.

#### **1. Average Salaries**

Year, Fielding, Pitching

1985, "2,028,571", "1,713,333"

1990, "2,100,000", "2,600,000"

2000, "3,111,000", "4,500,000"

#### **2. Hall of Fame All Star Pitchers**

Player, ERA, # All Star Appearances, Hall of Fame Induction Year

abcdef01, 3.11, 8, 1999

defghi01, 2.31, 8, 1988

ghijkl01, 1.91, 11, 2006

#### **3. Pitching**

Year, Player, Regular Season ERA, Regular Season Win/Loss, Post-season ERA, Post-season Win/Loss

1990, defgei01, 1.74, 73, 1.14, 100

1991, abcdhi01, 1.36, 71, 2.14, 85

1992, fdwesi01, 2.06, 70, 1.85, 90

1993, sdfwei01, 1.90, 65, 0.85, 87

#### **4. Rankings**

Team ID, Year, Rank, At Bats

PH1, 1871, 1, 1281

RC1, 1871, 9, 1036

LAA, 2014, 1, 5652

CHN, 2014, 5, 5508

(*These are examples for illustrative purposes only. They are not meant to demonstrate the actual data points that the application will output.*)

### **Other considerations**

* Include whatever documentation you believe is necessary for others to understand and maintain the application.
* Include tests that verify the behavior of your application, if time permits.

### **Bonus points**

Implement your application in a **stream-oriented fashion**, or explain how you would do so in addition to implementing a batch-oriented application.

### **Submission**

Please choose one of the following methods to submit your completed exercise:

1. **Online file sharing**: Upload the completed exercise to a file sharing service such as Google Drive, Dropbox, or OneDrive and share the link in this email thread.
2. **Online code repository**: Push the completed exercise to a code repository such as GitHub or GitLab and share the link in this email thread.

Please follow all of the above instructions and send code that you are proud to show us!