WEEK 6 ASSIGNMENT 1

Large-Scale Data Storage Systems – DATA-5400 | Spring 2020

Christina Morgenstern

The goal of this assignment is to understand the data analysis process better. Before running queries, a Data Scientist should get acquainted with the data first through looking at it manually. While this helps with understanding the problem it will also save time and thus money because in this process no additional resources are required to run.
The data for this assignment deals with baseball statistics and was downloaded as *.csv* file from http://www.seanlahman.com/baseball-archive/statistics/. The file baseballdatabank-master contains two folders, *core* and *upstream* as well as a README.txt. The README.txt in this directory provides general information on the dataset such as the license and distribution details. The *core* folder contains the databank itself and is of interest to us. While the *core* folder holds all tables, the README.txt file within the *core* folder will tell you what the tables hold i.e. the column names and their explanation.

(Q1) Use the information from the readme file to answer the following questions:

-
    - Which CSV file/table would you use to determine the total number of players?

The total number of baseball players should be retrieved from the *People.csv* table (In the README.txt this table is named Master table). This table holds 19878 rows of player information like name, date of birth, weight and height as well as starting and end game dates. I can just scroll down to the bottom of the table and retrieve the number of rows.

    - Which CSV file/tablewould you use to determine a player such as Derek Jeter's salary for the year 2010?

For retrieving the salary of a player, the table *Salaries.csv* is most appropriate. It holds the following information: year, team, league, player ID code and salary. With the help of the *People.csv* table, Derek Jeter´s player ID can be made out and then applied in a filter together with the year within the *Salaries.csv* table.

    - Which CSV file/tablewould you use to determine the player's date of birth and country of birth?

This information can be found in the *People.csv* table.

    - Which CSV file/tablewould you use to determine whether the player was inducted into the Hall of Fame?

The *HallOfFame.csv* table holds information which players and in which year he was awarded to the hall of fame. It also contains information on the voting method, the number of votes received and in which category he was honored.

- Which CSV file/table would you use to determine the name of the team a player played in, in the year 2000?

Using the *Teams.csv* table information of players and their teams in the respective years can be found.

- Which CSV file/table would you use to determine the number of home runs scored by a player such as Derek Jeter in 2010?

The number of home runs scored by a player in a year can be determined using the *Batting.csv* file.

- Which CSV file/table and which column would you use to check if the player is still alive?

The *People.csv* table again holds this sort of personal information

(Q2) Provide the Hive query language (HQL) commands for the following. You will run these HQL commands in the Azure cluster in Week 6 Assignment 2. Provide the names of the data file(s) that you will need to include in the queries and the results that you expect to get from the queries. NOTE : The HQL commands look similar to SQL queries.

3.

a. What is the total number of baseball players?

```
SELECT * from People.csv;
```

This command counts the number of rows in the People.csv table and should return the number of baseball players.

b. How many players were born in the year 1960 and earlier?

```
SELECT * FROM People.csv WHERE birthYear <= "1960";
```

c. How many players were born in the USA?

```
SELECT * FROM People.csv WHERE birthCountry = "USA";
```

d. How many players were born outside the USA?

```
SELECT * FROM People.csv WHERE birthCountry != "USA";
```

e. Display the number of players born in each year starting from 1960 thru 2000. For example, the output should show: 1980 4 (where 4 is the number of players born in 1980)

```
SELECT birthYear, COUNT(*) FROM People.csv GROUP BY  birthyear;
```

f. How many players and managers were inducted into the Hall of Fame?

```
SELECT * from fame where inducted = "Y" and category = "Player" or category "Manager";
```

g. Provide a list of all players for any team and from any year. For example, print the list of players who played for Chicago Cubs in 2000.

```
SELECT playerID, Count(*) FROM allstar GROUP BY yearID GROUP BY teamID;
```