Project 6: Working with Hive

DS 730

Overview

In this project, you will be working with Hive. You will be writing a Hive script for each of the problems below. You will see the difference between solving these problems with Pig and solving them with Hive. Some of them may be easy; some may be harder.

Project Tasks

We will be using three different files for this project. You have seen all of them in a previous project:

- **Master.csv**
- **Batting.csv**
- **Fielding.csv**

You can get all of these files from the online course.

You must write a Hive script for each of the following problems. If there is a tie for any of the questions (e.g. number 3 may have multiple weights that are second most common), you should print out all of them. You can assume the data is stored in the HDFS folder of: **/home/hduser/hive/**

Problem #	Description
1	Output the birth city of the player who had the most at bats (AB) in his career.
2	Output the top three birthdates that had the most players born. I am only looking for day and month combinations. For instance, how many were born on February 3rd, how many were born on March 8th, how many were born on July 20th… print out the top three dates.
3	Output the second most common weight.
4	Output the team that had the most errors in 2001.
5	Output the name of the player who had the most errors in all seasons combined.

Problem #	Description
6	A player who hits well and doesn't commit a lot of errors is obviously a player you want on your team. Who were the top 3 players from 2005 through 2009 who maximized the following criterion: $$\left(\frac{number\ of\ hits\ (H)}{number\ of\ at\ bats\ (AB)}\right) - \left(\frac{number\ of\ errors\ (E)}{number\ of\ games\ (G)}\right) ?$$ The above equation might be skewed by a player who only had three at bats but got two hits. To account for that, *only consider players who had at least 40 at bats and played in at least 20 games*. Be aware that: • Some players played for multiple teams during that 5-year span. • A player could have played multiple positions during that span. See, for instance, **buchabr01** in 2004 who played LF, OF and RF in the same season. You may need a UDF for this problem.
7	Sum up the number of doubles and triples for each birthCity/birthState combination. Output the top 5 birthCity/birthState combinations that produced the players who had the most doubles and triples.
8	Output the birthMonth/ birthState combination that produced the worst players. The worst players are defined by the lowest of: $$\frac{number\ of\ hits\ (H)}{number\ of\ at\ bats\ (AB)}$$ To ensure one player does not skew the data, *make sure that at least three people came from the same state and were born in the same month*. For this problem, the year does not matter. For examples, a player born in December, 1970 in Detroit and a player born in December, 1982 in Detroit are in the same group because they were both born in December and are from Detroit.

Submitting Your Work

When you are finished, create a file called **output.txt** and store your answers to each problem in that file. You should create a Hive script for each question and store your code in those files so that they can be easily tested. Be sure you are reading in from the correct **/home/hduser/hive/*.csv** location in your code. Submit *a .zip file* containing these items to the Project 6 dropbox:

- Your code stored in Hive scripts
- The output file, which should be named **output.txt**