Working with Pig

DS730

In this project, you will be working with Pig. You will be writing a Pig script to solve the following problems. You should solve the problems using the Pig system on your Hortonworks system.

We will be using three different files for this project. The first two (Batting and Master) are from a previous activity so you should know them well. The third file is a Fielding.csv file that deals with fielding statistics for a player for any given year. You can download the fielding file from:

http://www.uwosh.edu/faculty_staff/krohne/ds730/Fielding.csv

All of your input files must be in the following hierarchy on the HDFS:

/user/maria_dev/pigtest/Batting.csv
/user/maria_dev/pigtest/Master.csv
/user/maria_dev/pigtest/Fielding.csv

You must write one Pig script for each one of the following problems. Your final submission will include 8 pig script files. Be sure to DUMP your answer to the terminal window. Do not STORE your answer to a file. 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 output all of them. You should also assume that for ties, all of the ones that are tied have the same rank. For example, consider number 3. Assume the number of people for each height are as follows:

70: 200
71: 220
72: 200
73: 200
74: 220
75: 210
76: 180

The top height would be 71 and 74 because they each have 220 people with that height. The second most common height would be 75. The third most common height would be 70, 72 and 73. Finally, the fourth most common height would be 76.

I have italicized the data that must be output for each problem. Only output the answer to the problem. Do not output any extra information. For example, for question 1, do not output the name of the player or how many runs batted in that player had. As another example, do not output the top 10 cities of players with the most runs batted in for question 1. Only output what the answer is and nothing else. Also be sure to output them in the correct order if necessary. For example, for problem 2, the most common birthMonth/birthYear is first, the second most common birthMonth/birthYear is second and so on.

1. Output the *birth city* of the player who had the most runs batted in (RBI) in his career.
2. Output the *top three birthMonth/birthYear* that had the most players born. I am only looking for month and year combinations. For instance, how many were born in February, 1956, how many were born in March, 1975, and so on. Print out the top three *mm/yyyy* combinations. You must report the information in mm/yyyy form (i.e. it is ok to print out 5 instead of 05).
3. There are 2 people that had unique heights. Who are they? You should output their first and last names.
4. Which *team*, after 1950, had the most errors in any 1 season. A season is denoted by the year.
5. Output the *playerID* and *team* of the player who had the most errors with any 1 team in all seasons combined. Only consider seasons after 1950.
6. A player who hits well and doesn't commit a lot of errors is obviously a player you want on your team. Output the *playerID's* of the top 3 players from 2005 through 2009 (including 2005 and 2009) who maximized the following criterion:

(number of hits (H) / number of at bats (AB)) – (number of errors (E) / number of games (G))

The above equation might be skewed by a player who only had 3 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 **over that entire 5 year span**. You should note that both files contain a "number of games" column. **The 20 game minimum and the games values that you are using must come from the Fielding.csv file.** For this problem, be sure to ignore rows in the Fielding.csv file that are in the file for *informational* purposes only. An *informational* row contains no data in the

7th-17th columns (start counting at column 1). In other words, if all of the 7th, 8th, 9th, … 16th and 17th columns are empty, the row is informational and should be ignored.

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 combined (i.e., combine the doubles and triples for all players with that city/state combination). Some caveats:
    a. A *birthState* is any non-empty value in the birthState column.
    b. The *birthCity* must start with a vowel (i.e an A, E, I, O or U).
8. Output the *birthMonth/birthState* combination that produced the worst players. The worst players are defined by the lowest of:

    (number of hits (H) / number of at bats (AB))

    To ensure a small number of people who hardly played don't skew the data, make sure that:
    a. at least 10 people came from the same state and were born in the same month and
    b. the sum of the at-bats for all of the players from the same birthMonth/birthState exceeds 1500.

    For this problem, the year does not matter. A player born in December, 1970 in Michigan and a player born in December, 1982 in Michigan are in the same group because they were both born in December and were born in Michigan. A *birthState* is any non-empty value in the birthState column. In terms of condition a., you should count a player as one of your 10 players even if the player has no at-bats and/or no hits. You should ignore all players who do not have a birthMonth or who do not have a birthState.

To give you some sense of whether or not you are doing this correctly, some of the output for each question is included below:

1. The birth city of the person who had the 5th most runs batted in is Sudlersville.
2. The birth month/birth year combination that had the 7th most people born in it is a tie between 10/1960, 08/1978, 8/1974, and 10/1969.
3. There isn't much of a hint to give here. If you have the correct 2 people, it's easy to lookup their heights in the file and verify that no one else has the same height.
4. The team with the 3rd most errors in any season after 1950 was LAA in 1961.

5. The player who had the 4th most errors was schmimi01 and he played for PHI when he committed all of those errors. Oddly enough, 7 of the first 10 people on this list are in the hall of fame.
6. The player who had the 4th best “score” for the equation is Joe Mauer (mauerjo01). His value for the equation was .299850… One can manually verify this score by checking his stats:
   ((144+181+119+176+191)/(489+521+406+536+523)) -
   ((5+4+1+3+3)/(116+120+91+139+109))
7. The birthCity/birthState combination that produced the players with the 7th most doubles and triples is Indianapolis/IN.
8. The birthMonth/birthState combination that produced the 4th worst players is 7/LA.

When you are finished, upload the following to the Project 2 dropbox in a zipped file called `p2.zip` containing:

1. One pig script file for each problem. Use the names `P1.pig`, ..., `P8.pig` for these files. Be sure to use the DUMP command to produce your answer. Do not store the answer in a file.
2. If you created a Python UDF for this project, be sure to upload that as a separate Python file. This is not a required file. In fact, we recommend that you do not use a Python UDF for this project.
3. A text file called `answers.txt` that contains all of your properly labelled answers to each of the problems. There is no specific format for this file. Be sure to specify the order of your answers for questions that ask for an ordering (e.g the top 5 or the top 3).