Working with Hive

DS730

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.
We will be using three files for this project. You have seen all of them in a previous project: Master.csv, Batting.csv and Fielding.csv. 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 must have Batting.csv stored in the HDFS folder of:
/user/maria_dev/hivetest/batting/

Master.csv is stored in the HDFS folder of:
/user/maria_dev/hivetest/master/

and Fielding.csv is stored in the HDFS folder of:
/user/maria_dev/hivetest/fielding/

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[1]. Whenever a question asks for a top K rank, it is asking for all answers that are in that particular rank. For example, consider number 3. Assume the number of people for each weight are as follows:

70: 20
71: 22
72: 20
73: 20
74: 22
75: 21
76: 18

[1] What is described with respect to ties is something called a DENSE_RANK.

The top weight would be 71 and 74 because they each have 22 people with that weight. The second most common weight would be 75. The third most common height would be 70, 72 and 73. Finally, the fourth most common weight would be 76.

I have italicized the data that must be output for each problem. 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 at bats that player had. As another example, do not output the top 10 cities of players with the most at bats for question 1. Only output what the answer is and nothing else. Also be sure to output them in the correct order if necessary. Lastly, do not save your answer to a file. Simply output your answer to the terminal window.

Do not worry about any specific format for any of these problems. For example, for question 2, if you output mm/dd or mm:dd or simply mm,dd, these are all acceptable. As long as your output is obvious, it is fine.

A few tips before getting started:

1. You should never use the **LOAD DATA INPATH** command in any of your answers. This defeats the purpose of using Hive. The goal is to take your schema to the data and not move your data to a new location. All of your solutions should start off with **CREATE EXTERNAL TABLE…**
2. You should never use the **INSERT OVERWRITE DIRECTORY** command in any of your answers. You only need to output your answer to the terminal window. If you use this command, it may overwrite something important in my own Hortonworks HDFS and I don't want that.
3. Ensure that each script for each problem is self contained and doesn't rely on any previous script. In other words, do not assume a table has been created in a previous script.

These are the questions you are to solve:

1. Output the *birth city* (or cities) of the player(s) who had the most at bats (AB) in his career.
2. Output the *top three ranked 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 *mm/dd* combinations. You must output the information in mm/dd form (it is ok to print out 5 instead of 05).

3. Output the *second most common weight by rank*.
4. Output the *team(s)* that had the most errors in 2001.
5. Output the *playerID(s)* of the player who had the most errors in all seasons combined.
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 ranked 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 that you are using is from the Fielding file.** For this problem, be sure to ignore rows in the Fielding 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 ranked birthCity/birthState* combinations that produced the players who had the most doubles and triples (i.e. combine the doubles and triples for all players with that city/state combination). A *birthState* is any non-empty value in the birthState column.
8. Output the *birthMonth/birthState* combination(s) 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 1 player who barely played does not skew the data, make sure that:
   a. at least 5 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 month/state exceeds 100.

   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 5 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 10th most at-bats is Donora.
2. The birthday that had the 5th most people born on it is 10/4, or October 4th.
3. The 3rd most common weight is 175.
4. The teams with the 6th most errors in 2001 was Oakland (OAK) and Baltimore (BAL).
5. The player who had the 4th most errors in his entire career is Germany Smith (smithge01).
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 Brooklyn/NY.
8. The birthMonth/birthState combination that produced the 5th worst players is 2/Colorado.

When you are finished, upload the following files to the Project 3 dropbox in a single zipped file called `p3.zip` containing:

1. One Hive script file for each problem. Use the names `P1.q`, …, `P8.q` for these files.
2. A text file called `answers.txt` that contains all of your properly labelled answers to each of the problems.