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. We will be using three files for this project: etfs.csv and meta.csv. You must write a Hive script for each of the following problems.

You will have etfs.csv stored in the HDFS folder of:
/user/maria_dev/hivetest/financial/

meta.csv is stored in the HDFS folder of:
/user/maria_dev/hivetest/metadata/

If there is a tie for any of the questions, similar to the Pig project, 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. See the Pig project for more information on this.

I have bolded 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. 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 things to follow 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]…**

[1] What is described with respect to ties is something called a DENSE_RANK.
[2] If the table already exists, then you'll want to drop it first.

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. Make sure to load your tables correctly in each script given that information. You should create 1 script per problem and call your script **PX.q** where **X** is the question number you are solving.
4. Your script should be fully self-contained. Do **not** have separate scripts for creating/dropping tables/views.
5. Do not assume that tables have already been created in each script. In other words, create table XYZ before using it.
6. Do not assume that no tables have been created. In other words, drop table XYZ before creating table XYZ.

These are the questions you are to solve:

1. Output the **total number** of distinct city names that are in the meta.csv file.
2. Output the **fund** that has the third highest average volume in 2013.
3. Output the **highest closing price** and the **year** for the aadr.us fund each year it was trading. Be sure to provide an answer for each year it was trading.
4. Output the **founding state** that had the highest total volume in 2014.
5. Output the **year** and **month** that had the third least amount of trading days.
6. Output the **first** and **last name** of the founder whose etf saw the most number of profitable days. A profitable day is defined as a day where the close was higher than the open.
7. Output the **fund** and the **year** that had the lowest average closing price for that fund in the first half of that year (January 1 - June 30).

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. Not much of a hint here. I guess one could open it up in a spreadsheet software and count all of the unique cities… or create a smaller file and test it against that.
2. The fund with the second highest trading volume in 2013, on average, is eem.us.
3. The highest closing price for aadr.us in 2010 was 28.928.
4. The second highest volume state was Montana.
5. The month with the second least number of trading days was September, 2001 with 15 trading days. The month with the lowest number of trading days was the end of the data so that particular month has no meaning.

6. The second highest fund was founded by James Anderson.
7. The 5th lowest fund was (spxl.us, 2009) which had an average closing price in the first half of that year of 2.071270967741935.

What should you do if you are not getting the same answers as the hints above? A good way to debug is to look at the output that your code gives and compare that to the input. Does your output make sense? You may have to print out more information when debugging to see where the issue is.

What to Submit

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, …, P7.q` for these files.
2. A text file called `answers.txt` that contains all of your properly labeled answers to each of the problems.