STSCI 5065 HW 3

(Assigned: 3/22/2019; Due: 3/29/2019 at 11:59PM)

What and how to turn in:

Submit electronically to the course website: a PDF file named STSCI5065-HW3-LastName-FirstName, containing all the steps you performed, including the code (which cannot be an image so that your graders can copy your code into their system to test your code), and your answers to the questions, listed in the order of the questions.

In this homework, there are two problem sets. You will practice how to use Hive to establish a database and process the data in HDFS according to the requirements specified below in the problem sets. Unless otherwise directed, list all your code in a separate paragraph with blue font, for example:

```
select flight_delay
from delays
where carrier = “NW”;
```

Problem Set 1

You are given a set of two text files of New York Stock Exchange (NYSE) data, NYSE_daily_prices.csv and NYSE_dividends.csv. You are required to create a Hive database containing two tables based on the text files, and then do some analysis of the data using Hive queries. (70 points total)

A. (6 points) Download the data file, stock.zip, from the course website and unzip it in a local OS directory. Open the files in a text editor to see how the data fields are separated. Create an HDFS directory, **HW3**, right below the CentOS root. In HW3, create a directory called **stockdata**. Use the File View in Ambari to upload the two data files into stockdata. Display a screenshot (in Ambari) showing all these files in the /HW3/stockdata directory (your screenshot should show this directory).

B. (13 points) Create a Hive database called **stocksdb** located in HW3. In the stocksdb database, use Hive command line interface to create two tables **stock_prices** and **stock_dividends**. The stock_prices table contains columns in the following order: exchng, symbol, ymd, price_open, price_high, price_low, price_close, price_volume, and price_adj_close. The data types of exchng, symbol and ymd are string (same in the stock_dividends table) and other columns are of the double data type. The stock_dividends contains the columns in the following order: exchng, symbol, ymd, and dividend. The data type of dividend is double. Describe these two tables in Hive command line interface, and attach one single screenshot, showing the commands and the results.

C. (5 points) Load the data files into the respective tables created in step B.

D. (6 points) Query the stock_prices table to find out the number of entries of each stock symbol. You are only required to show two screenshots of the first 10 rows and the last 10 rows of your result, including the line starting with "Time taken" in your output.

E. (6 points) Calculate the average stock opening price of each stock symbol rounded to 4 decimal points (same below). Use an alias, ap, for the average price column. Attach a screenshot of the last ten rows of the output.

F. (3 points) Create a Hive view, **average_price_v**, based on step E.

G. (12 points) Find out the stock symbols that have the highest and lowest average opening prices, respectively. You are required to use the view created in step F and to complete this query in Ambari using the Hive View. Attach the screenshots of the query results.

H. (12 points) Find out the stock symbol that has the highest dividends. What are the date and the opening price of that stock when that happened? Complete this query also in Ambari's Hive View. Attach the screenshots of the query results.

Problem Set 2

You are given a text file (flight12.csv) of flight delays and are required to create a Hive table using regular expressions, and then do some queries on the table. (30 points total)

A. (5 points) Download the data file and use the File View in Ambari to load the file into HDFS in the HW3 directory. Create a Hive database called **flightsdb** located in HW3, and then in flightsdb create a Hive table, **flight_delays_hw3**. This table has the following columns: ymd, flight_num, carrier_delay, weather_delay, nas_delay, security_delay, and late_aircraft_delay. The first two columns are of the string data type and the other columns are of the double data type.

B. (3 points) Create a temporary one-column Hive table, **temp_flight**, to read in the data from the text file so that a line of text becomes a row in the temp_flight table, and then load the text data file into temp_flight.

C. (15 points) Insert data into flight_delays_hw3 by extracting it from temp_flight with regular expression (regexp_extract()) calls. The data fields are separated by commas. You are required to extract fields 2, 5,19, 20, 21, 22, and 23 to populate your table.

D. (2 points) Query the flight_delays_hw3 table with Hive command line and by only display the first 10 rows. Attach a screenshot of Hive command line and its result.

E. (5 points) Query the table and find out the longest delay of each category of delay. Use a meaningful alias for each column. Do this in Ambari Hive View and attach a screenshot of the results.