Homework 7

Due: Wednesday April 3 at 11:59pm in compass2g.illinois.edu

For this assignment, submit one `.hive` script file containing all the necessary Hive code to generate your results and one text-based report file (Word doc, pdf, or .txt file). The **Hive script file should be an executable `.hive` file** (If using the Hue editor or running your code in steps at command line, you may copy the code and paste it into a Notepad++ file with extension `.hive`.).

Use Hive for all exercises. Any code based on code from elsewhere (e.g. code provided with the text) must reference in comments the source of the original code.

All exercises are based on the stocks data set in `stocks.csv` which accompanies *Programming Hive: Data Warehouse and Query Language for Hadoop* by Edward Capriolo, Dean Wampler, and Jason Rutherglen [Safari Online]. The data is available in the zip file available from the book's github page. You will need to download the book source from github (use the green "Clone or download" button, or click on the `.zip` file and then click "Download"), extract the files and get the `stocks.csv` file from the data folder.
Note: If you open `stocks.csv` in Excel, **do not** save the file. Large files may not open entirely in Excel and saving an incompletely opened file will result in a loss of data.

Exercises for All Students

Exercise 1:

Create and populate a `stocks` table for the stocks data. The table should contain the same columns as the data set (`market, stocksymbol, datemdy, price_open, price_high, price_low, price_close, volume, price_adj_close`). The first three fields should be `STRING`; the price fields should be `FLOAT`; and `volume` should be `INT`.

Show 5 records from the table.

Exercise 2:

Create a table for the records for IBM (stock symbol "IBM" in the NYSE `market`). Show 5 results from that table.

Using Hive queries find the highest daily high price and the lowest daily low price for IBM within the data set, and the dates on which those max high and min low prices occurred.

Exercise 3:

Create a view for the max daily spread (high price – low price) for each stock symbol and include the market in the view (group by symbol and market in case the same symbol occurs on multiple markets).

Obtain the minimum, average, and maximum daily spreads from the data in this view.

Additional Exercise for Graduate Students

Exercise 4:

For each market, determine the company with the largest daily price spread, the value of that company's max daily price spread, and the date on which it occurred. The view from exercise 3 and the original table should be useful for this.