Module	HIVE
Instructor	Dr. Vinay Kulkarni
Assignment Start Date	March 11, 2017
Assignment Submission Date	April 2, 2017

Goal of these exercises

- To ensure that you gain good understanding of working with Hive

About the data that has been provided

- See explanation of the fields given towards the end of this document
- NSE data has been provided for multiple years. On a typical day about 1700 companies' information is tracked and prices such as OPEN, CLOSE, HIGH, LOW, Traded Quantity, etc. are provided for every tracked company.
- The exercises given below require the use of this data.

Exercise 1: MapReduce [10 Marks]

1. With the help of table(s) with about 5 lines of data each clearly illustrate – showing <key, value> transformations that should happen during mapping and reducing stages to implement the following operations:
    a. Filter rows of a table based on specified criteria (eg. Age > 35)
    b. Select specific columns from those available (eg. SELECT name, age, height)
    c. FULL OUTER JOIN of two tables
    d. Generate a count of the number of rows in a given table

Exercise 2: Calculation of various statistical quantities and decision making <u>using HIVE [20 Marks]</u>

1. Only lines with value "EQ" in the "Series" column should be processed. As the first stage, filter out all lines that do not fulfil this criteria.
2. For every stock, for every year, calculate the following statistical parameters and store the generated information in properly designed tables:
    - Min, Max, Mean, Standard Deviation
3. Select any year for which data is available. In your report clearly mention the year selected.
    - For that year, create a table that contains data only for those stocks that have an average total traded quantity of 3 Lakhs or more per day
    - Find out the Pearsons Correlation Coefficient for EVERY PAIR of stocks contained in this table. Final output should be in decreasing order of the coefficient.
4. Use the information generated in step 3 in the following way:
    - Assume you have Rs 10 Lakh to invest.
    - Assume you have to invest in six stocks on the first working day of January of the chosen year.
    - By using logic / simulation / etc. Identify the stocks that you will invest in, such that at the end of the year:
        i. At least your overall capital (Rs 10 Lakh) is protected.
        ii. You make good profit.

Exercise Steps for HIVE

1. Create a directory for executing this exercise
2. Copy the provided stock market data into this directory
    - Column details are provided later in this document
3. Create Hive SQL programs to answer all the questions
4. Work in this directory so that Hive can create and maintain its metadata in this directory (and the same will be available to you across sessions)

Your submissions should include:

- HiveQL Programs written to answer all the exercise questions. Programs should be well commented.
- Outputs in CSV format of the final table(s) that contain the answers to the questions
- A document / presentation outlining the approach taken by you to solve the problem. Such a document should include logic behind the design of tables and samples and explanation of the generated output.

Evaluation Criteria and marks

Correctness and colpleteness of the solution(s)	30
Content and quality of final report / presentation	10
Submission of code and output. Marks will depend on how well the code is commented, in addition to its correctness and completeness.	10

Note

- Solutions may be discussed amongst yourselves. However, the code and presentations **should be individually created**. Copied code / presentations will not get any credit (for all parties involved!
- As a pre-requisite to solving the problems in Hive, you will have to learn and understand how SQL programs can be written in external files (source code) and called and executed from within the Hive environment.
- Please refer to documentation available at the following location for HiveQL:
    - https://cwiki.apache.org/confluence/display/Hive/LanguageManual

Format of the CSV File

SYMBOL	SERIES	OPEN	HIGH	LOW	CLOSE	LAST	PREVCLOSE	TOTTRDQTY	TOTTRDVAL	TIMESTAMP	TOTALTRA	ISIN
20MICRONS	EQ	28.2	28.2	27.15	27.4	27.2	27.8	30252	831503.95	18-Aug-15	243	INE144J01027
3IINFOTECH	EQ	3.05	3.05	3.05	3.05	3.05	3.2	596734	1820038.7	18-Aug-15	380	INE748C01020
3MINDIA	EQ	10547.3	11000	10310.1	10832.25	10900	10399.75	1261	13411045.05	18-Aug-15	537	INE470A01017
3RDROCK	IT	203	203.5	203	203.5	0	205	51726	10508999	18-Aug-15	6	INE768P01012
8KMILES	EQ	880	896	865.1	887.85	890	878.6	23572	20920236.85	18-Aug-15	1400	INE650K01013
A2ZINFRA	EQ	27.4	28.9	25.8	26.9	26.65	27.35	396620	10966299	18-Aug-15	2152	INE619I01012
AARTIDRUGS	EQ	645.2	653	639.95	643.4	647	641.1	35362	22796290.95	18-Aug-15	1381	INE767A01016

Field	Description
SYMBOL	Company symbol provided by National Stock Exchange (NSE)
SERIES	We should only process records with EQ in this field (Equities)
OPEN	Opening price of the stock on a given day
HIGH	Highest price reached by the stock on a given day
LOW	Lowest price reached by the stock on a given day
CLOSE	Closing price of the stock on a given day
LAST	Last bid value (ignore this field)
PREVCLOSE	Closing price of the previous day
TOTTRDQTY	Total traded quantity on a given day
TOTTRDVAL	Total value of all trades of the stock on the given day
TIMESTAMP	Date of the trades
TOTALTRADES	Number of distinct trades during the day
ISIN	NSE Stock code