Problem Statement Project 2

Input Dataset: Input is a MySQL database table which is getting data from some old traditional system which can’t be changed. Here is schema for the table

http://www.infochimps.com/datasets/nyse-daily-1970-2010-open-close-high-low-and-volume

Other datasets related to NYSE - http://www.infochimps.com/tags/nyse

```
CREATE TABLE nasdaq_daily_prices (
        exchange VARCHAR(35) ,
        stock_symbol VARCHAR(35) ,
        date VARCHAR(35) ,
        stock_price_open DOUBLE ,
        stock_price_high DOUBLE ,
        stock_price_low DOUBLE,
        stock_price_close DOUBLE,
        stock_volume BIGINT,
        stock_price_adj_close DOUBLE
);
```

To insert data in table please run following command:

```
LOAD DATA local INFILE 'C:\\Windows\\Temp\\NASDAQ_daily_prices_A.txt' INTO TABLE
nasdaq_daily_prices FIELDS TERMINATED BY ',' ENCLOSED BY '"'
LINES TERMINATED BY '\n';
```

Problem Statement: Assume that MySQL table has huge data and this cannot be processed using MySQL. We will need to put this data in HDFS and then process the data using Map-Reduce. Once the data is processed, processed data need to be put into MySQL back for reporting purpose.

We will use Sqoop for import and export of data from and to MySQL and hadoop map reduce for doing the processing.

Output: To find out total volume for each stock_symbol. Output table would look like this:

```
CREATE TABLE stock_volume (
            stock_symbol VARCHAR(35) ,
            total_stock_volume BIGINT,
);
```