# HW3

Deadline: April. 4th 2:59 P.M. (before class)

## ***Description***

The tick_data.csv file is tick data of stocks. **Please re-download the file from Blackboard since the column name is corrected.** The file includes 5 columns:

- DATE: (int) date when transaction made
- TIME_S: (str) time in seconds when transaction made
- SYM_ROOT: (str) stock name of transaction
- SIZE: (int) transaction volume
- TRADE: (double) price of transaction

In this part, you need to

1. [4 pts] Using Spark Dataframe, calculate total trading volume for each stock in a certain hour. For example,

   ```
   SYM_ROOT,DATE,TIME_H,SIZE_H
   Stock_A,Day1,09,size1
   Stock_A,Day1,10,size2
   …
   ```

   where TIME_H represents time in hours and SIZE_H is the total volume in this hour.

2. [5 pts] Assuming TRADE reflects the stock price at the time, calculate hourly return of each stock with the following formula,

   $$r_t = \frac{p_t^n - p_t^1}{p_t^1}$$

   where in a certain hour $t$, $p_t^n$ represents the price of last trade and $p_t^1$ represents the price of first trade.

   The dataframe should be built on top of part 1 and include

   ```
   SYM_ROOT,DATE,TIME_H, SIZE_H,RETURN_H
   ```

   where RETURN_H represents the hourly return that is calculated from the formula.

3. [1 pt] Sort the output by SYM_ROOT, DATE, and TIME_H.

## ***Submission***

1. PySpark code (part 1, 2, 3)
2. Final output that is generated from part 3.