**IDS 521 Homework 2**
**Prof. Ali Tafti**
**Fall 2015**
This version: October 16, 2015

## Real-time Diffusion of Information on Twitter and the Financial Markets

In this assignment you will use SQL to do some variants of computations described in the attached research paper entitled "Real-time Diffusion of Information on Twitter and the Financial Markets." In particular, one of the challenges of this research project is to compute $99^{th}$ percentile threshold values of Twitter mentions or trading volume for each firm that change over time. This may be done with relative elegance using SQL as a declarative language; and by taking advantage of the relational database's capabilities to perform joins or other computations efficiently.

In this homework assignment, you will extend the research described in the attached paper by considering the impact of earnings releases. You are given a SQL script that does the following:

1) Loads the main dataset in which each tuple is identified by a combination of the firm's symbol and ten-minute time increment during data collection hours. This set contains a total of 1,223,906 tuples, from May 21, 2012 through September 18, 2013.
2) Loads a file containing a comprehensive list of earnings releases; and then creates a new table in which earnings releases are matched to the appropriate firm and period-number combination of the main data set. Matches are based on the firm, date, and ten-minute time increment.

Building on the provided SQL script, you will write a complete SQL script that, as detailed in the steps below, measures changes in financial trading volume for stocks after earnings releases and also after peaks in Twitter activity.

Please submit the following to show your work for Parts 1 - 3:

1) A SQL script (a .sql file) that will load the source .csv files into tables and then do all the processing to create the final tables of results. The script needs to run without breaking from beginning to end in MySQL. Please build upon the SQL script provided to you for this exercise. Please indicate clearly the names of all project partners that worked with you, on top of the file.
2) Any comma-separated value files used in your SQL scripts besides the files provided to you. For example, if you decide to create a Date-Hour database table, provide the .csv used to populate it and make sure its name matches the name used in your Load Data command.
3) An MS Word or PDF file in which you describe your final results. Please present any ***small*** tables of final results here. Clearly list the name of each database table which contains your results for each intermediate step. Briefly describe each of

the SQL statements used in the process; and refer to specific SQL statements in your SQL script file. ***Please do not paste large tables in your final report; simply refer to them in your database by name and describe how your script creates them.*** Please indicate clearly the names of all project partners that worked with you, on top of the file.

**Part 1: Trading volume levels following Earnings Releases**

Steps:

1) The following query shows the number of unique firms present in the main dataset (it should be 100). Please modify this query (or write a new query) to show the number of firms that are represented in at least 30 days. Based on your query, how many different firms are represented in at least 30 days in the main dataset? Which firms in the dataset (if any) do not meet this criterion?

```
Select count(*) from  (
        Select null from Tweets  group by symbol) t;
```

2) **Write a script of SQL statements to** do the following:
    a. For each firm that has had at least one earnings release, show a column with the average trading volume for the firm on days in which it had an earnings release, and another column showing the average trading volume on days for which there was no earnings release. On average, how do these two numbers compare? Does the difference seem to depend on the number of earnings releases? Include in your script the SQL statement that populates a view or table with this data.

    b. Create and populate a table that shows changes in trading volume immediately following each earnings release. For each earnings release, show the average volume of shares traded in the immediate one hour following the earnings release. This is calculated as a difference in trading volume over the hour, specifically as $volumeend_{t+6}$ - $volumeend_t$ where the $volumeend_t$ represents the firm's trading volume at the end of the ten-minute increment t. Likewise, $volumeend_{t+6}$ represents trading volume for the same firm after sixty minutes (or six periods into the future).