CSE 487/587 Programming Assignment #3

Due Date: March 30, 2015
Weight: 12%

Problem: (a) Use PIG and HIVE to compute volatility of stocks in NASDAQ
(b) Compare the performance of PIG, HIVE and MapReduce implementation.

Description

In this assignment, you will use PIG and HIVE in CCR to compute the monthly volatility of stocks, using the data and volatility equation in hw#1. Find the top 10 stocks with the lowest (min) volatility and the top 10 stocks with the highest (max) volatility.

Data

Using the data of hw#1, the first column Date and the last column Adj Close will be used in this assignment; other columns are neglected.

Note:

- If the stock has volatility of 0, skip it.
- Number N in equation should NOT be fixed as 36, as some stocks don't have 36 months of data.
- For large and medium dataset, each file is considered as different stock, although the name of the stock is the same. For example, AAPL-1 and AAPL-2 are different stocks.

What you need to do:

In your report, include following aspects:

- Your rationale for PIG and HIVE computation.
- Speed-up plot for your PIG and HIVE computation (including any preprocessing time) for varying problem size and varying cores. Sample test cases are given in submission guidelines.
- Comparison of the performance of PIG, HIVE and MapReduce in same settings.
- Discussion of all the experimental results and comparison results.

Specific Submission Guidelines: Assignment 3

1. Files should be strictly organized as following structure in your own directory (**/gpfs/courses/cse587/spring2015/students/username/hw3/**) and the naming of the directory should be followed exactly (case sensitive):

NOTE THAT YOU SHOULD NOT MAKE ANY CHANGES TO THE DIRECTORY AFTER THE SUBMISSION DEADLINE, AS THE TIME STAMP OF THE FILES WILL BE USED FOR TIMELY SUBMISSION.

hw3/pig/src/
(include the source code of your job, e.x. python script, pig script)

hw3/pig/SLURMmyHadoop
(the sample slurm script of your mapreduce job)
hw3/hive/src/
(include the source code of your job, e.x. python script, pig script)
hw3/hive/SLURMmyHadoop
(the sample slurm script of your mapreduce job)
hw3/*username*.pdf
(your assignment report)
hw3/misc/ (optional)
(include any other files you may want to submit)

2. Your code will be evaluated using automated script in following fashion.
For pig, [username@rush:~] sbatch SLURMmyHadoop <input directory> <output directory>
For hive, [username@rush:~] sbatch SLURMmyHadoop <input directory>

3. SLURMmyHadoop specifications:
- use partition debug
- set the time to 20 mins
- set the nodes as 2 and 2 tasks-per-node
- email as your own email
- do not change HADOOP_CONF_DIR in your final submission, i.e.,
  export HADOOP_CONF_DIR=$SLURM_SUBMIT_DIR/config-$SLURM_JOBID
- For pig, the output directory should use command line such as
  $HADOOP_HOME/bin/hdfs --config $HADOOP_CONF_DIR dfs -get /pigdata/hw3_out
$2

4. In your report, include execution time for the following cases.

Problem Size	Execution Time: 1 node (12 cores)	Execution Time: 2 nodes (24 cores)	Execution Time: 3 nodes (48 cores)
Small			
Medium			
Large			

- Execution time is the time taken for your entire computation (Including any preprocessing you might have used, exclude any time taken for configuring, logging and cleaning hadoop)
- You should also submit a compressed tar file of your entire hw3 directory to UBLearns.

Grading Criteria

For each part (i.e., PIG and HIVE)

- Program correctness (working program): 25%
- Data Scaling: 6%
- Node Scaling: 6%
- Performance: 6%
- Discussion and the report: 7 %