DATA INTENSIVE COMPUTING CSE587

PROGRAMMING ASSIGNMENT-3 REPORT

By
Name: Vishwas Shanbhog
UBIT No: 50135707

Introduction:

In this assignment, calculation of the volatility for stocks was used to evaluate the performance of Apache Pig and Hive. There were three sets of stocks given. The goal of this assignment was to calculate the top 10 stocks with maximum volatility and bottom 10 stocks with lowest volatility. The platform used was Apache Hive and Apache Pig. All the datasets (small, medium and large) were run on Hive and Pig with different node configurations which were 1 nodes (12 cores), 2 nodes (24 cores) and 4 nodes (48 cores) on both Pig and Hive. The results of the runtime are described below.

Results:

PIG:

The runtime of calculating in pig for small, medium and large datasets for 1, 2 and 4 nodes are as follows:

Problem size	Execution time(1 node- 12cores)	Execution time(2 nodes- 24 cores)	Execution time(4 nodes- 48 cores)
Small	657 sec	643 sec	630 sec
medium	1369 sec	1350 sec	1347 sec
Large	2704 sec	2678 sec	2669 sec

HIVE:

The runtime of calculating in pig for small, medium and large datasets for 1,2 and 4 nodes are as follows. :

Problem size	Execution time(1 node- 12cores)	Execution time(2 nodes- 24 cores)	Execution time(4 nodes- 48 cores)
small	787 sec	768 sec	760 sec
medium	1465 sec	1432 sec	1430 sec
Large	General compute error	General compute error	General compute error

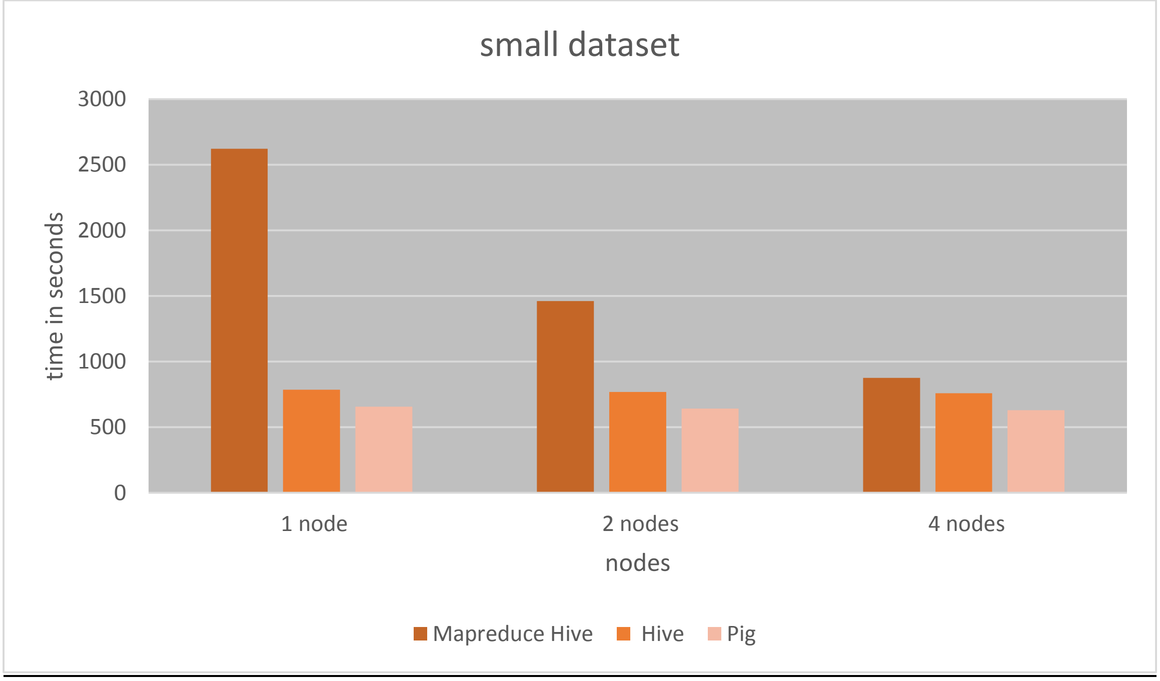
HADOOP MAPREDUCE:

The runtime of calculating in pig for small, medium and large datasets for 1,2 and 4 nodes are as follows. The results:

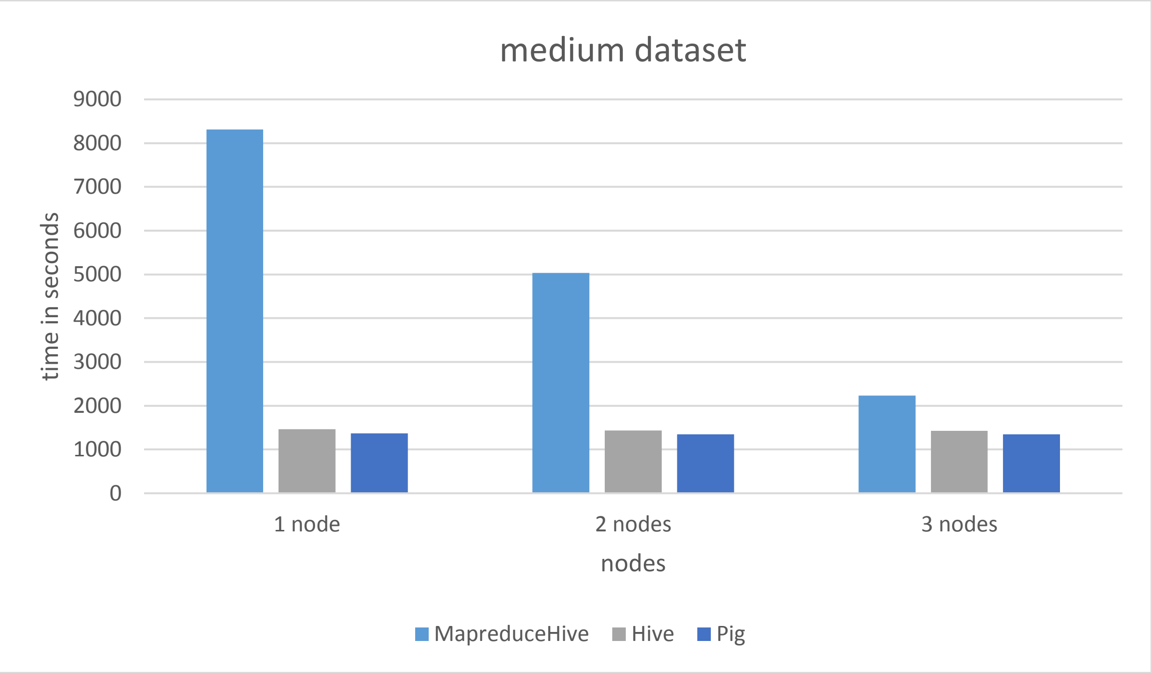
Problem size	Execution time(1 node- 12cores)	Execution time(2 nodes- 24 cores)	Execution time(4 nodes- 48 cores)
small	2623 sec	1461 sec	875 sec
medium	8313 sec	5032 sec	2231 sec
Large	26568 sec	12967 sec	7199 sec

PERFORMANCE COMPARISON GRAPHS:

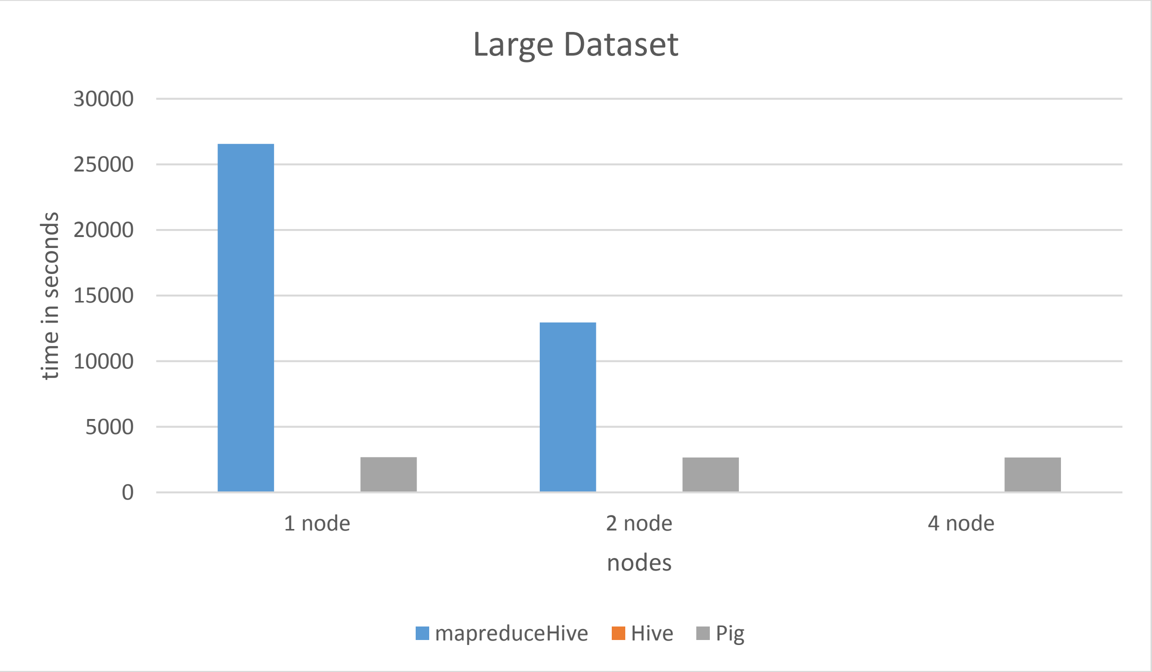
Small data set performance comparison:



Medium Dataset Performance Comparison:



Large Dataset Performance Comparison:



*Runtime: time taken to execute to execute the entire program

Conclusion:

From the results Hive and Pig performance are very similar for different configurations of nodes and perform better than map reduce because of the optimizations made. The Hive and Pig are by default set single node configuration which is why their performance is same for different configurations.