Khushboo Tekchandani

CSE587 Homework 4

Solution:

The problem requires us to calculate the volatilities of NASDAQ stocks and find the stocks with top 10 maximum and the top 10 minimum volatilities.

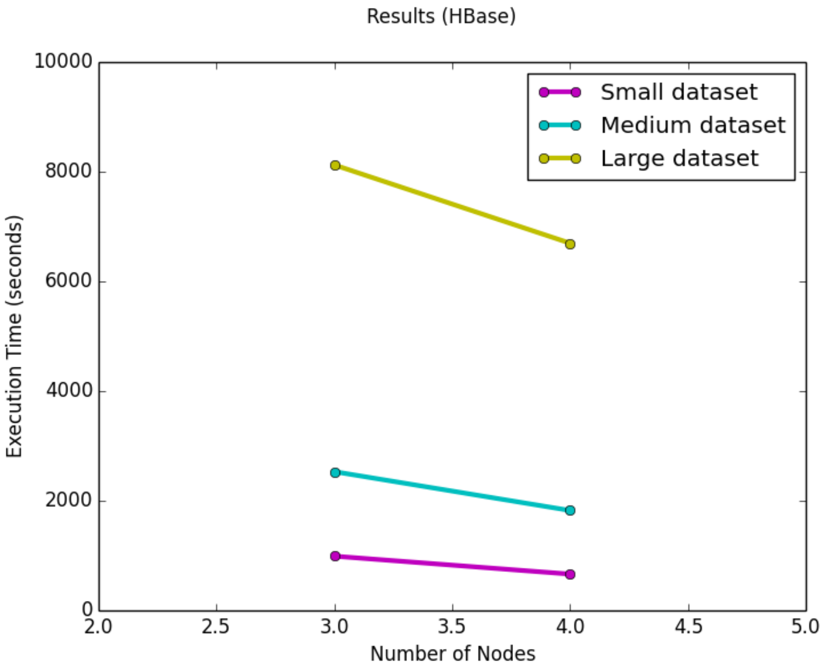
The problem requires us to use Hbase, a distributed, non-relational database system. HBase is used when you need random, real-time read/write access to your Big Data.

Evaluation:

Using these implementations, I executed my code on CCR to understand the scalability and performance of my solution using Hbase. The experiment was done using different data sets and different number of nodes. Following data was obtained by carrying out these experiments:

	Execution Time in Seconds	
Problem Size	3 node (36 cores)	4 nodes (48 cores)
Small	990	663
Medium	2531	1823
Large	8124	6697

Table 1: Execution time of the program on different number of nodes, using multiple data sets.



Graph1: Execution time for different problem sizes on different number of compute nodes

Comparisons/Observations:

- MapReduce is a computing framework. The first assignment required us to implement the solution to the NASDAQ problem using the Hadoop which is nothing but a combination of the Hadoop Distributed File System and the Map Reduce framework. HDFS lacks **random read and write access**.
- This is why HBase is used. It is a **distributed, scalable, big data store.** It stores data as key/value pairs. Thus Hbase can help in efficiently reading and writing data to the file system and increases performances of Map reduce tasks.
- Hive is a **data warehouse** technology that works on top of Hadoop cluster and provides and SQL like interface. Whereas Pig is a **data flow language**. It allows you to write SQL queries which are the interpreted and optimized by the Pig Interpreter. Both Hive and Pig use MapReduce as the underlying framework. And as noted in the previous assignment, Pig performs better than Hive.
- Conclusion:
  Performance of
  Pig > Hive > Hbase > Hadoop.