Big Data Analytics Assignment - 2

Abhishek Agarwal - 2016126
Kaustav Vats - 2016048

Average Execution time for each storage engine with Apache Spark

No of Executors \ Apache Spark with	Postgres storage engine	MongoDB storage engine	HDFS storage engine	Indexing on Query 10	Without indexing on Query 10
1 Executor	0.694 ms	1.129 ms	1.064 ms	0.314 ms (Postgres SE) 0.653 ms (MongoDB SE)	0.452 ms (Postgres SE) 1.121 ms (MongoDB SE)
2 Executors	0.672 ms	1.047 ms	1.005 ms	0.244 ms (Postgres SE) 0.507 ms (MongoDB SE)	0.379 ms (Postgres SE) 1.104 ms (MongoDB SE)

Some observations -

1. Distributed workload
2. Better query execution time for all engines on increasing no of executors
3. MongoDB contains unstructured data, that affects the query execution time with MongoDB as storage engine

Learnings

- Setting up Apache Spark
- Got familiar with spark python api ‘pyspark’.
- Integrating Apache Spark with different data storage engines like Postgres, MongoDB and HDFS.
- Setting up above mentioned data storage engines in both linux and windows.
- Parallelizing tasks using RDD
- Learned various api calls of the apache spark
- Learned about parameters of spark, which can be used to improve execution time.

Challenges in solving

- Setting up Apache Spark -
  - JDK issue
  - Scala version compatibility with python
  - Spark home and hadoop path
- Created MongoDB data, faced issue because of huge size of data.
- Regarding NoSQL MongoDB, faced issue in understanding query exection in MongoDB.
- HDFS installation on windows, faced issues with some unknown exceptions. Switched to linux os after multiple attempts on windows.
- Indexing queries execution.

References

- https://www.knowledgehut.com/blog/big-data/how-to-install-apache-spark-on-windows
- https://www.guru99.com/download-install-postgresql.html
- https://itnext.io/apache-spark-and-hadoop-hdfs-hello-world-ed6fb2077c20
- https://linuxconfig.org/how-to-install-hadoop-on-ubuntu-18-04-bionic-beaver-linux
- https://docs.mongodb.com/spark-connector/master/
- https://databricks.com/blog/2015/03/20/using-mongodb-with-spark.html
- https://github.com/mongodb/mongo-spark