MapReduce Vs. Apache Spark

Presenter : Liyanage D. P. 239162A

MapReduce

A programming model or pattern within the Hadoop framework that is used to access big data stored in the Hadoop File System (HDFS). It is a core component, integral to the functioning of the Hadoop framework.
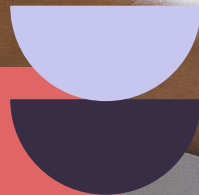
Apache Spark

Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size.
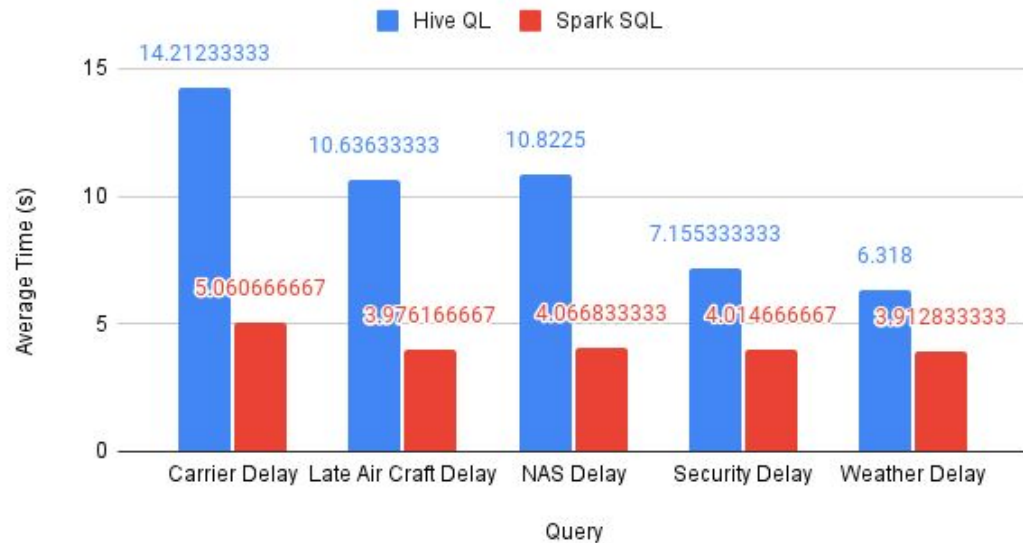
Demonstration

MapReduce Vs. Spark

Fast process

Apache Spark

Average Time Taken By

■ Hive QL   ■ Spark SQL

Query	Hive QL	Spark SQL
Carrier Delay	14.21233333	5.060666667
Late Air Craft Delay	10.63633333	3.976166667
NAS Delay	10.8225	4.066833333
Security Delay	7.155333333	4.014666667
Weather Delay	6.318	3.912833333

Average Time (s)
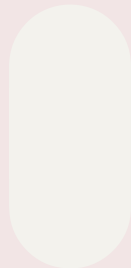
Query

Ease of Use

Apache Spark

Easy to use and support interactive mode

MapReduce

Difficult to use with no interactive mode

Performance

Apache Spark

Runs 100 times faster in-memory and 10 times faster on disk than Hadoop

MapReduce

Faster than traditional system

Data Processing

Apache Spark

Can handle all data processing requirements (batch, real time, graph etc.)

MapReduce

Is ideal for batch processing

Conclusion

- Apache Spark is very much popular for its speed

- Apache Spark is easier to use than MapReduce

- Even from performance wise, Apache Spark takes the lead

Thanks!