MSc in CS

A Comparative Exploration: MapReduce vs. Apache Spark for Big Data Processing

Name : K.G.D.S.Bandara
Index No : 248211C
Assignment - Video presentation

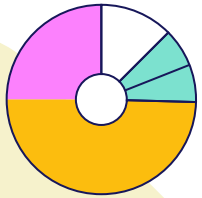
MapReduce

MapReduce is a programming model and software framework introduced by Google in 2004 for processing large datasets in a distributed computing environment.

The MapReduce model consists of two key functions - Map and Reduce.

The **Map** function takes input data and converts it into key-value pairs.

The **Reduce** function aggregates the outputs of the Map function into the final result.



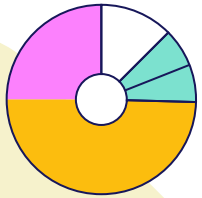
Apache Spark

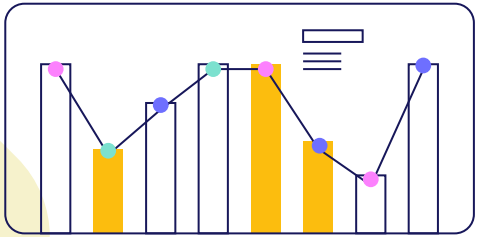
Apache Spark is an open-source cluster computing framework built around speed, ease of use and sophisticated analytics.

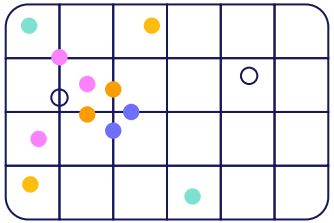
It was originally developed at UC Berkeley in 2009.

Like MapReduce, Spark can distribute data processing tasks across multiple computers.

Spark improves upon the MapReduce model with in-memory data sharing, allowing it to run workloads up to 100x faster than MapReduce in certain situations.



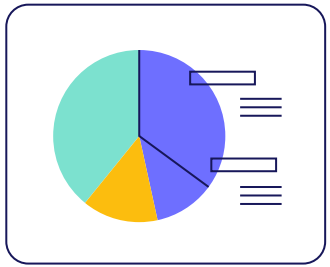


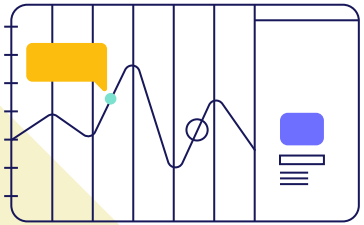


Demo

Let's do this

Compare and contrast





Ease of Use:

MapReduce is a simple and easy-to-use framework that is used for batch processing of large data sets.

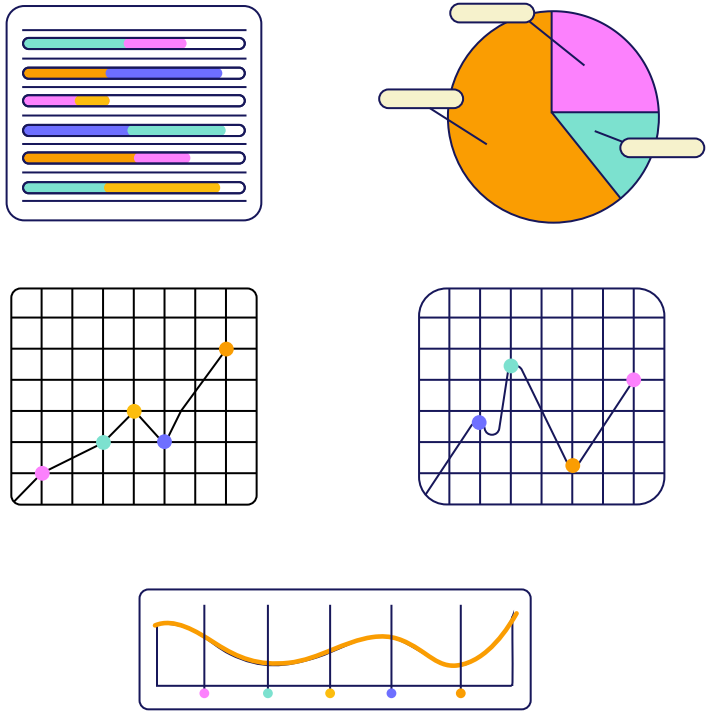
Apache Spark provides a higher-level programming model that makes it easier for developers to work with large data sets

Compare and contrast

Fast Processing:

Apache Spark is generally faster than MapReduce due to its in-memory processing capabilities

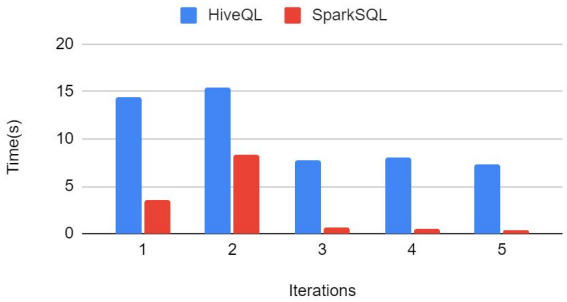
MapReduce reads and writes data to disk for each MapReduce job, therefore it takes more time to execute the queries.



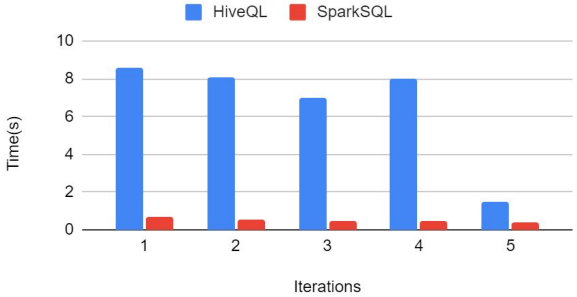
Findings

Time comparison by queries

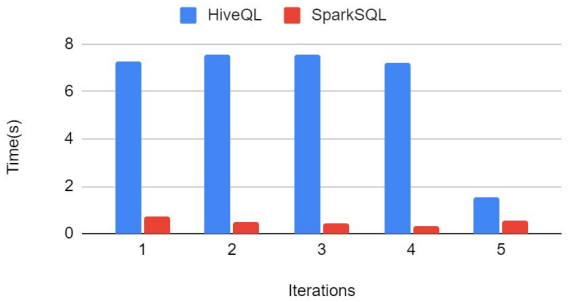
Carrier delay query



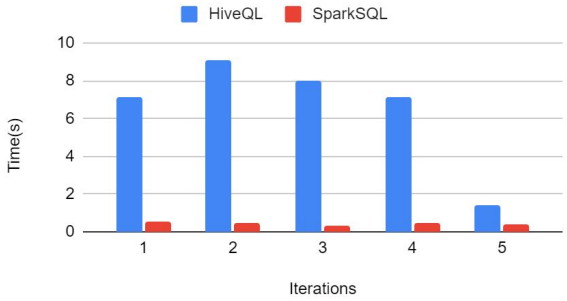
NAS delay query



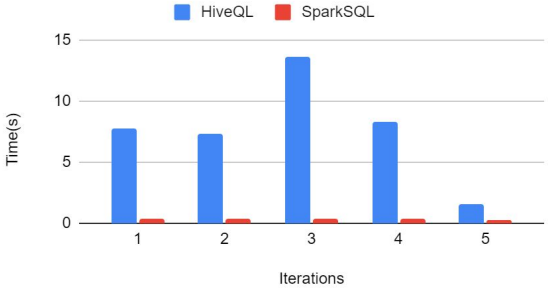
Weather delay query



Late aircraft query

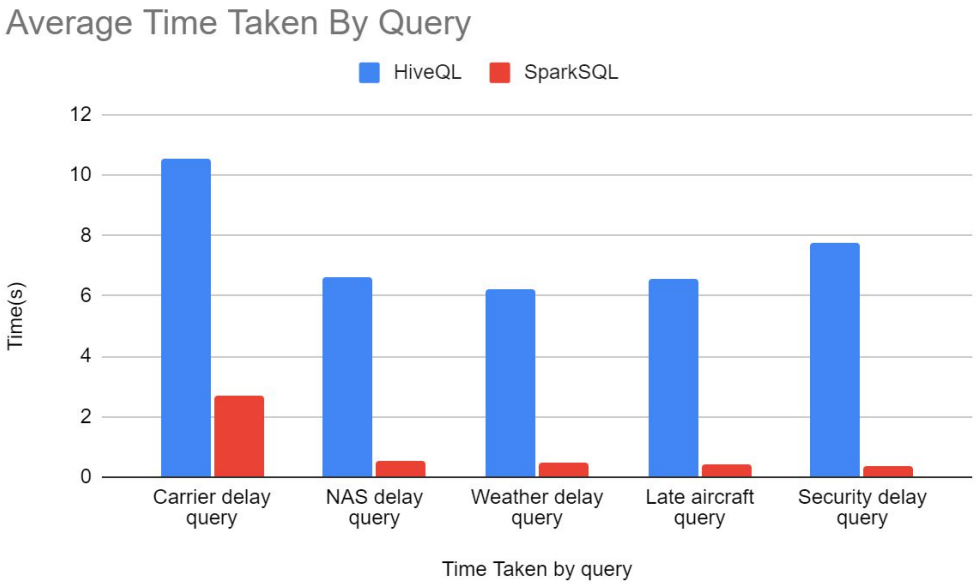


Security delay query



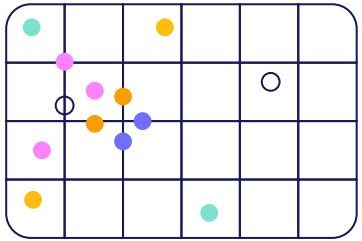
Findings

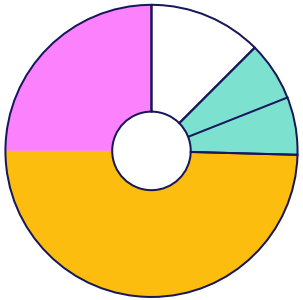
Average Time Taken		
Query	**HiveQL**	**SparkSQL**
Carrier delay query	10.563	2.7118
NAS delay query	6.6252	0.537
Weather delay query	6.2122	0.5008
Late aircraft query	6.546	0.445
Security delay query	7.7316	0.3802



Conclusion

Both MapReduce and Apache Spark enable distributed processing of big data. Spark is generally considered easier to use and faster than the older MapReduce framework. Spark's in-memory data processing engine allows it to perform computations faster by minimizing disk reads and writes. Additionally, Spark's higher level APIs make it more developer friendly compared to MapReduces low level programming model. For most modern big data pipelines, Spark is the preferred distributed computing engine due to its speed and ease of use advantages over MapReduce.





Thank You !