MapReduce vs. Apache Spark for Big Data Processing

> Weligalla W.N.C. 248288P
> Assignment - Video presentation

MapReduce

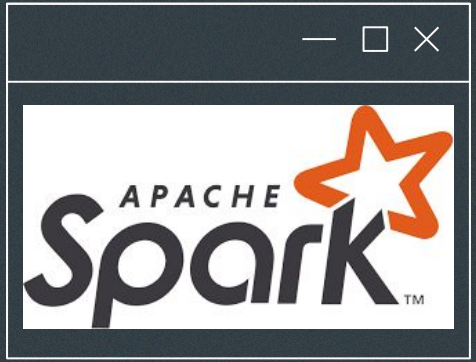
* Introduced by Google in 2004, MapReduce is a programming model and software framework designed for the distributed processing of large datasets within a computing environment.

/ The MapReduce model is centered around two primary functions: Map and Reduce.

/ The Map function processes input data, transforming it into key-value pairs.

/ The Reduce function consolidates the outputs generated by the Map function, producing the final result.

Apache Spark



* Apache Spark is an open-source cluster computing framework designed with a focus on speed, user-friendly features, and advanced analytics capabilities.

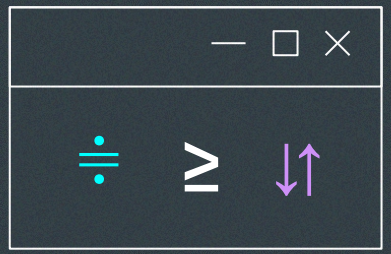
* Originally developed at UC Berkeley in 2009, Apache Spark, similar to MapReduce, has the capability to distribute data processing tasks across multiple computers.

* Spark enhances the MapReduce model by incorporating in-memory data sharing, enabling it to execute workloads up to 100 times faster than MapReduce in specific scenarios.

DEMO

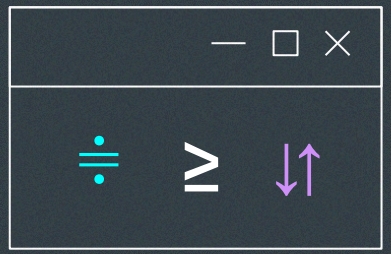
/loading, processing and applying queries

Compare and contrast
/Ease of Use



* MapReduce serves as a straightforward and user-friendly framework designed for the batch processing of extensive datasets.

* Apache Spark offers a more advanced programming model, simplifying the task for developers working with substantial datasets.



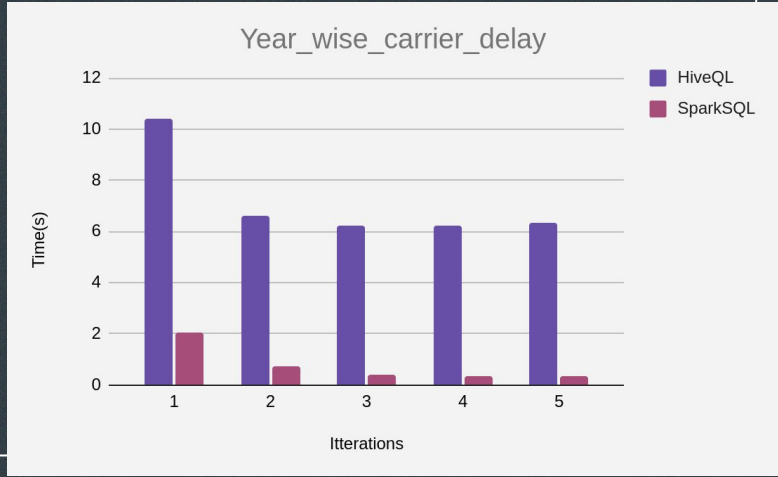
Compare and contrast /Fast Processing

* Apache Spark typically outpaces MapReduce in speed, thanks to its ability to perform in-memory processing.

* Each MapReduce job involves the reading and writing of data to disk, causing an increase in the time required for query execution.

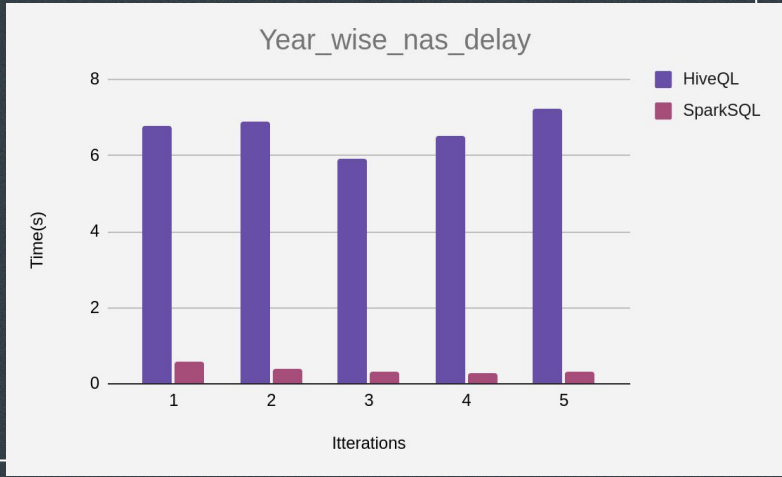
Findings

Year_wise_carrier_delay		
Itterations	HiveQL	SparkSQL
1	10.405	2.07
2	6.6	0.738
3	6.235	0.396
4	6.224	0.361
5	6.35	0.337



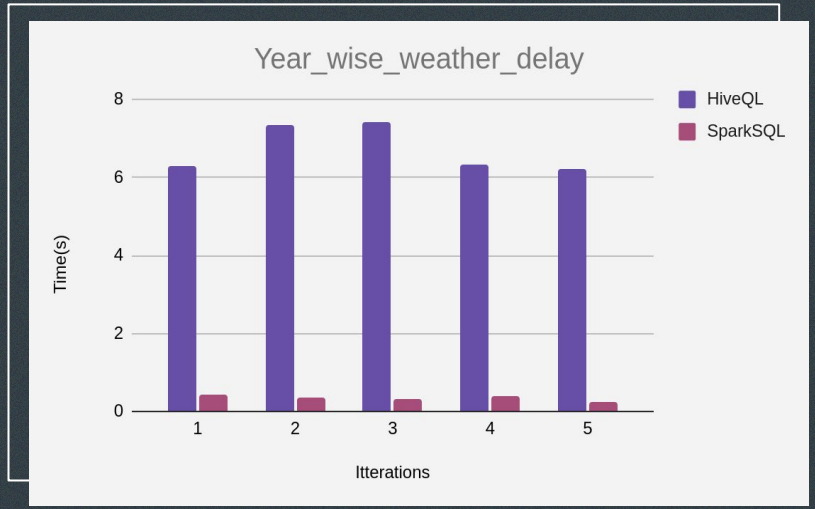
Findings

Year_wise_nas_delay		
Itterations	HiveQL	SparkSQL
1	6.764	0.571
2	6.881	0.404
3	5.912	0.331
4	6.521	0.296
5	7.22	0.305



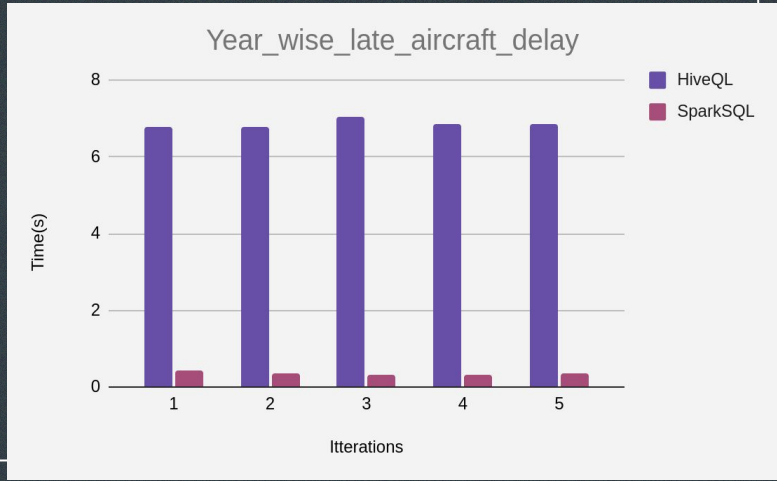
Findings

Year_wise_weather_delay		
Itterations	HiveQL	SparkSQL
1	6.306	0.424
2	7.336	0.37
3	7.414	0.308
4	6.328	0.389
5	6.222	0.257



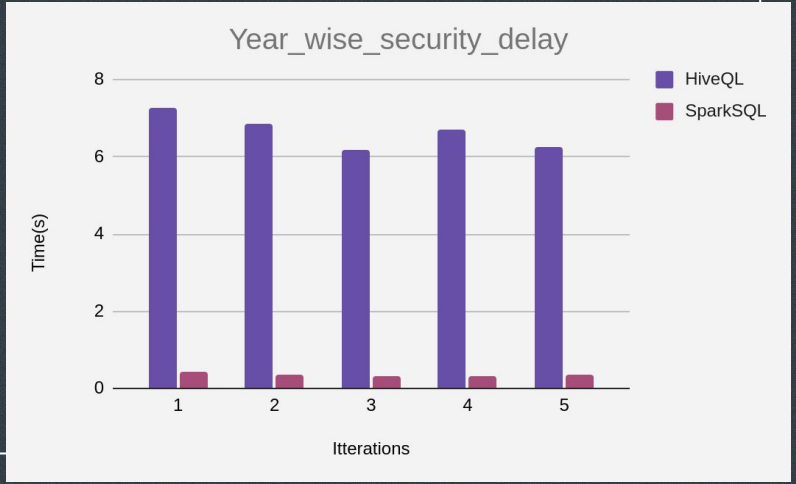
Findings

Year_wise_late_aircraft_delay		
Itterations	HiveQL	SparkSQL
1	6.796	0.448
2	6.783	0.361
3	7.027	0.311
4	6.872	0.325
5	6.844	0.337



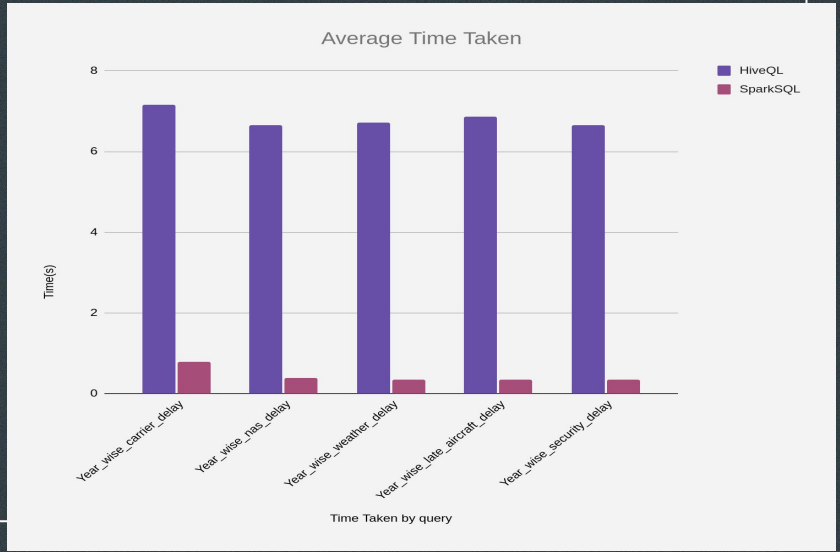
Findings

Year_wise_security_delay		
Itterations	HiveQL	SparkSQL
1	7.272	0.448
2	6.85	0.361
3	6.178	0.311
4	6.694	0.325
5	6.246	0.337



Findings

Average Time Taken		
Query	HiveQL	SparkSQL
Year_wise_carrier_delay	7.1628	0.7804
Year_wise_nas_delay	6.6596	0.3814
Year_wise_weather_delay	6.7212	0.3496
Year_wise_late_aircraft_delay	6.8644	0.3564
Year_wise_security_delay	6.648	0.3564



Conclusion

* Distributed Processing: Both MapReduce and Apache Spark enabled distributed processing of big data.

* Performance Comparison: Spark is generally considered easier to use and faster than the older MapReduce framework.

* In-Memory Processing: Spark's in-memory data processing engine enhances computational speed by minimizing disk reads and writes.

* APIs and Developer Friendliness: Spark offers higher-level APIs, making it more developer-friendly compared to MapReduce low-level programming model.

* Modern Big Data Pipelines: For most modern big data pipelines, Spark is the preferred distributed computing engine.

* Advantages: Spark is favored for its speed and ease of use advantages over MapReduce in contemporary big data processing scenarios.


Thank You!