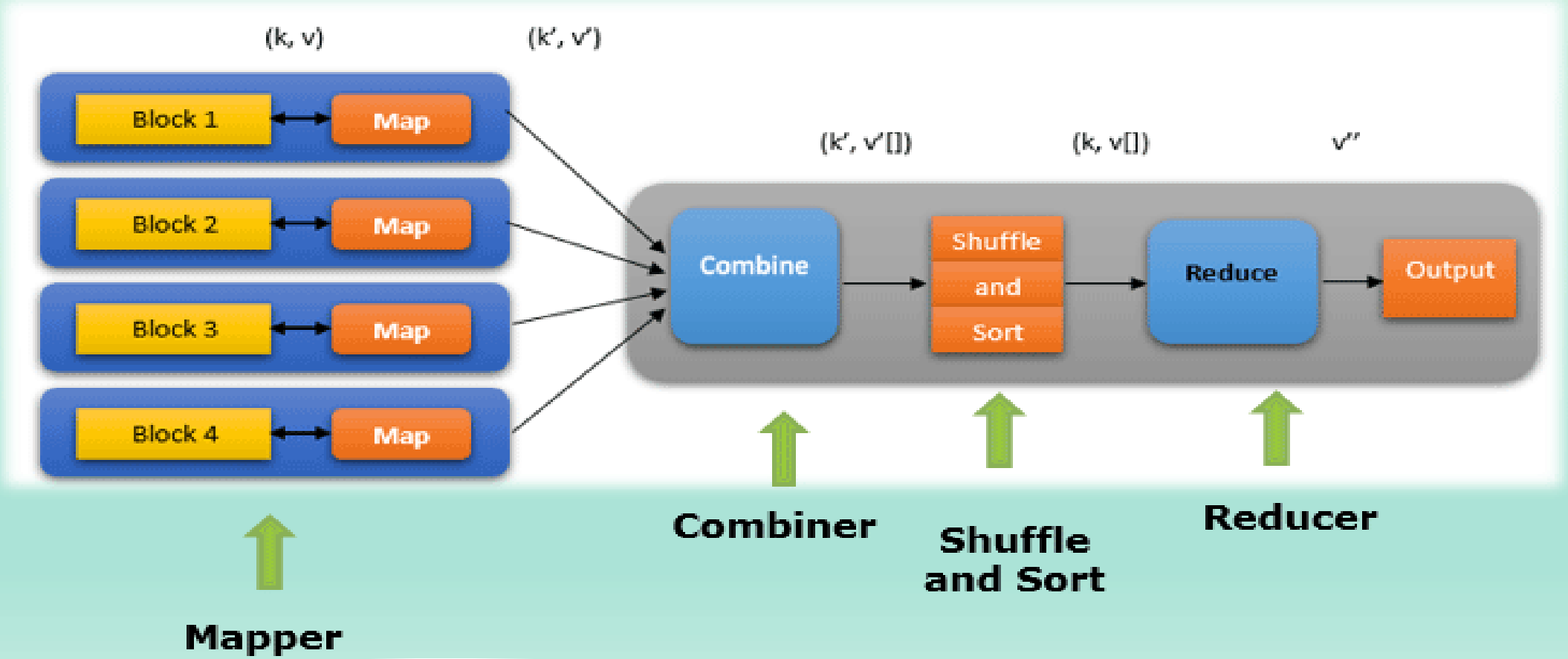
MapReduce and Apache Spark

CS5229 - Big Data Analytics Technologies

Hasini Weerasooriya
248287L
MSc in Computer Science
University of Moratuwa



MapReduce

- MapReduce is a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster.
- It performs the processing of large data sets in a distributed and parallel manner.
- MapReduce consists of two distinct tasks – **Map** and **Reduce**

Apache Spark

- Apache Spark is an open-source data processing engine to store and process big data.
- It is built on top of the Hadoop distributed file system.

Support various
Programing Languages

Developers and data scientists incorporate spark into their applications to rapidly query, analyze, and transform data at scale.


DEMO

mazon EMR
MR Serverless
MR on EC2
usters
otebooks and Git repos
vents
ock public access
ecurity configurations
MR on EKS
rtual clusters
MR Studio
etting Started
udios
orkspaces (Notebooks)
hat's New
deo tour
Compact mode

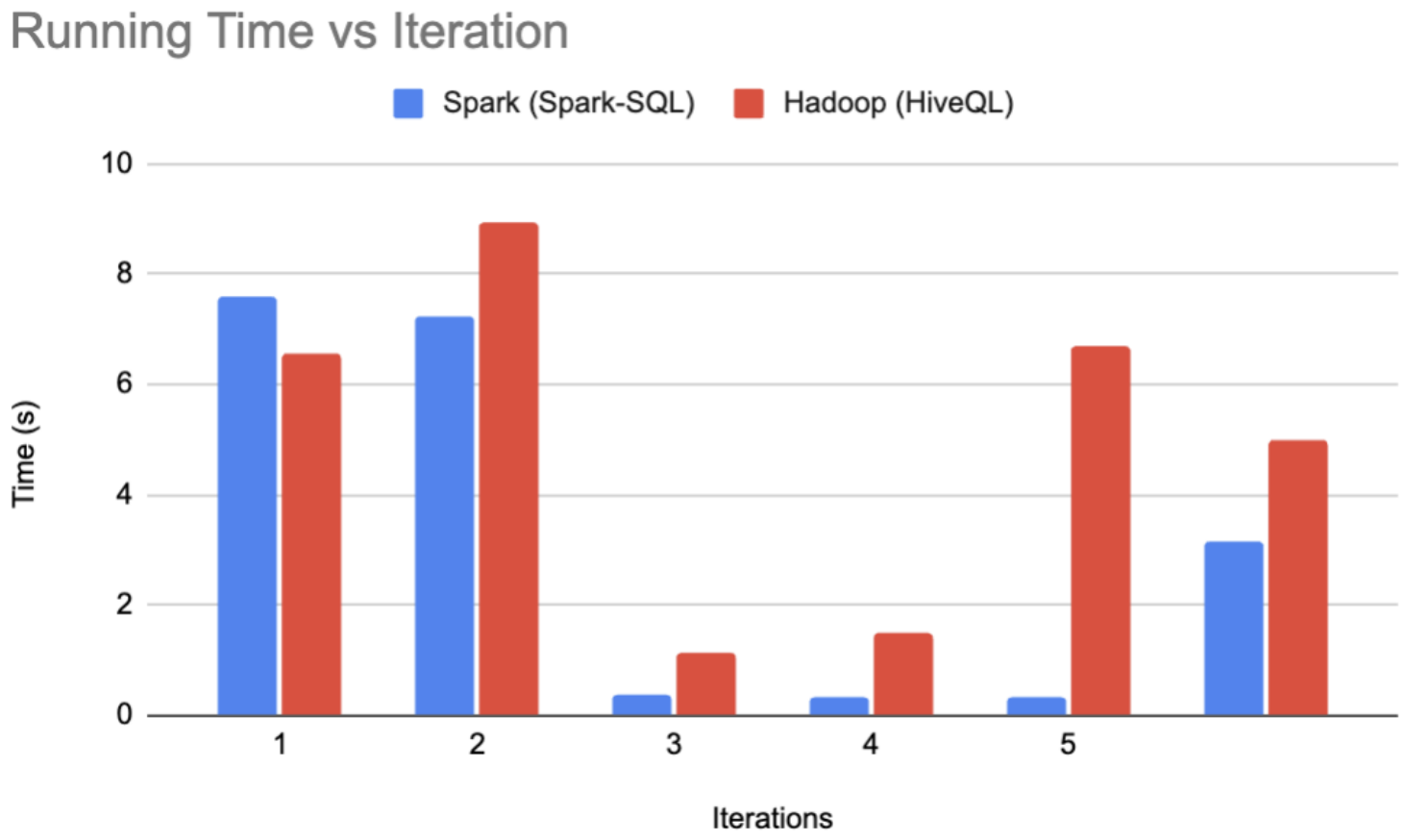
Clusters (5) Info

View details | Terminate | Clone | Create cluster

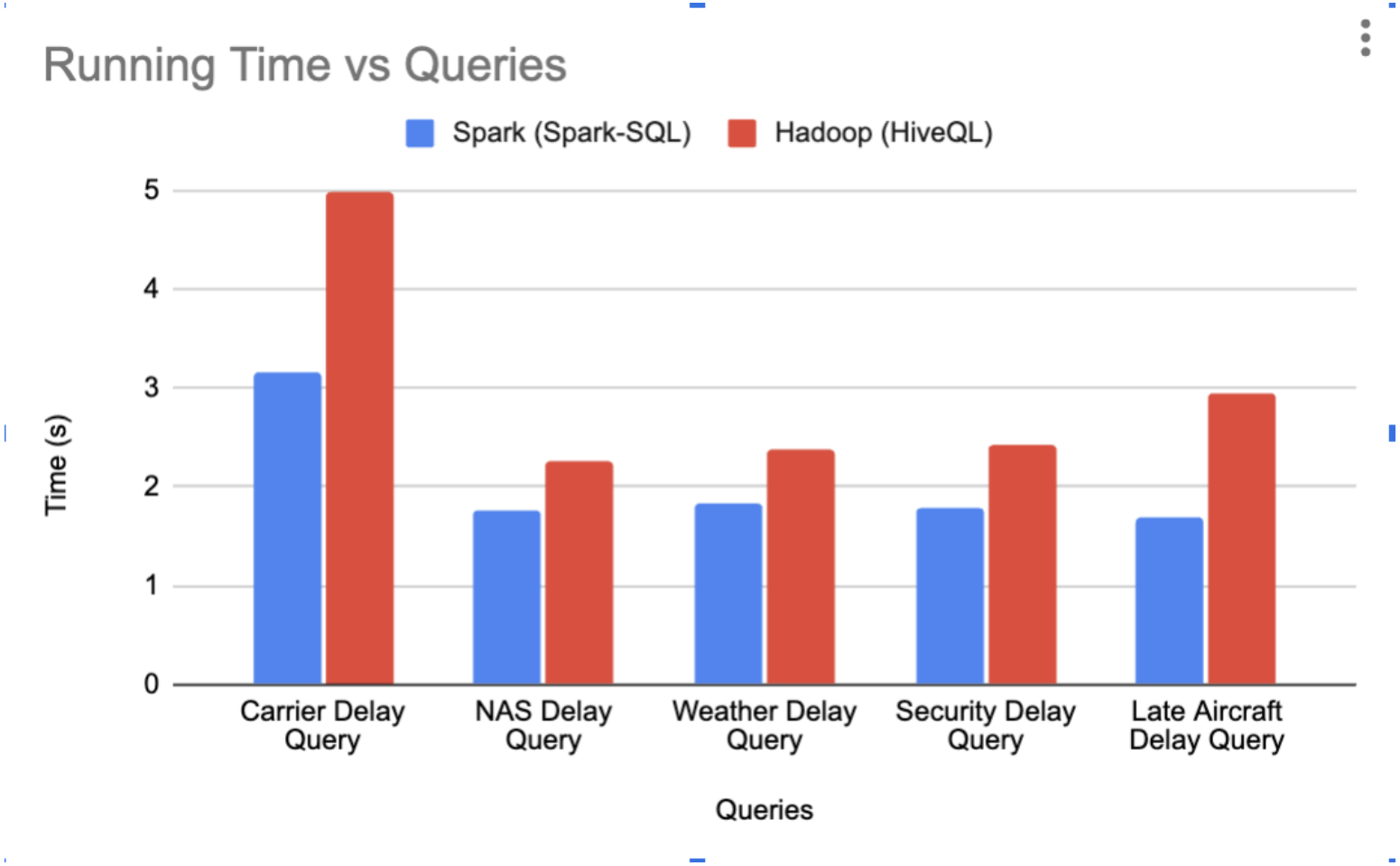
Filter clusters by status | Find clusters | Filter clusters by creation date-time | 1

Cluster ID	Cluster name	Status	Creation time (UTC+05:30)	Elapsed time	Normalized instance hours
j-161LNYOIV743O	emr_cluster_bigdata	Waiting Ready to run steps	March 04, 2024, 21:16	32 minutes	0
j-17KVIY6C5MKPW	bigdata_emr_cluster	Terminated User request	March 04, 2024, 20:40	53 minutes, 39 seconds	24
j-13RNOUDUCUTHQ	bigdata_emr	Terminated User request	March 04, 2024, 19:40	58 minutes, 23 seconds	24
j-QEJA2MS9RP2B	Sample cluster	Terminated User request	February 13, 2024, 22:54	41 minutes, 33 seconds	24
j-JG68VN4UDV9D	My sample cluster	Terminated User request	February 13, 2024, 22:12	26 minutes, 55 seconds	24

Comparison Results



Comparison Results (Cont.)



MapReduce vs Apache Spark

MapReduce	Apache Spark
Processing data using MapReduce in Hadoop is slow	Spark processes data 100 times faster than MapReduce as it is done in-memory
Performs batch processing of data	Performs both batch processing and real-time processing of data
Compact and Lengthy	Compact and easier than Hadoop
Written in Java with more lines of code and takes more time to execute	Implemented in Scala with fewer lines of code
Doesn’t support caching of data	Caches the data in-memory & enhances the system performance

Conclusion

- In here we did the comparison using Hadoop and Spark.
- Both used to process big data in different ways.
- Hadoop was created to delegate data processing to several servers instead of running the workload on a single machine.
- Apache Spark is a newer data processing system that overcomes key limitations of Hadoop.


THANK YOU!