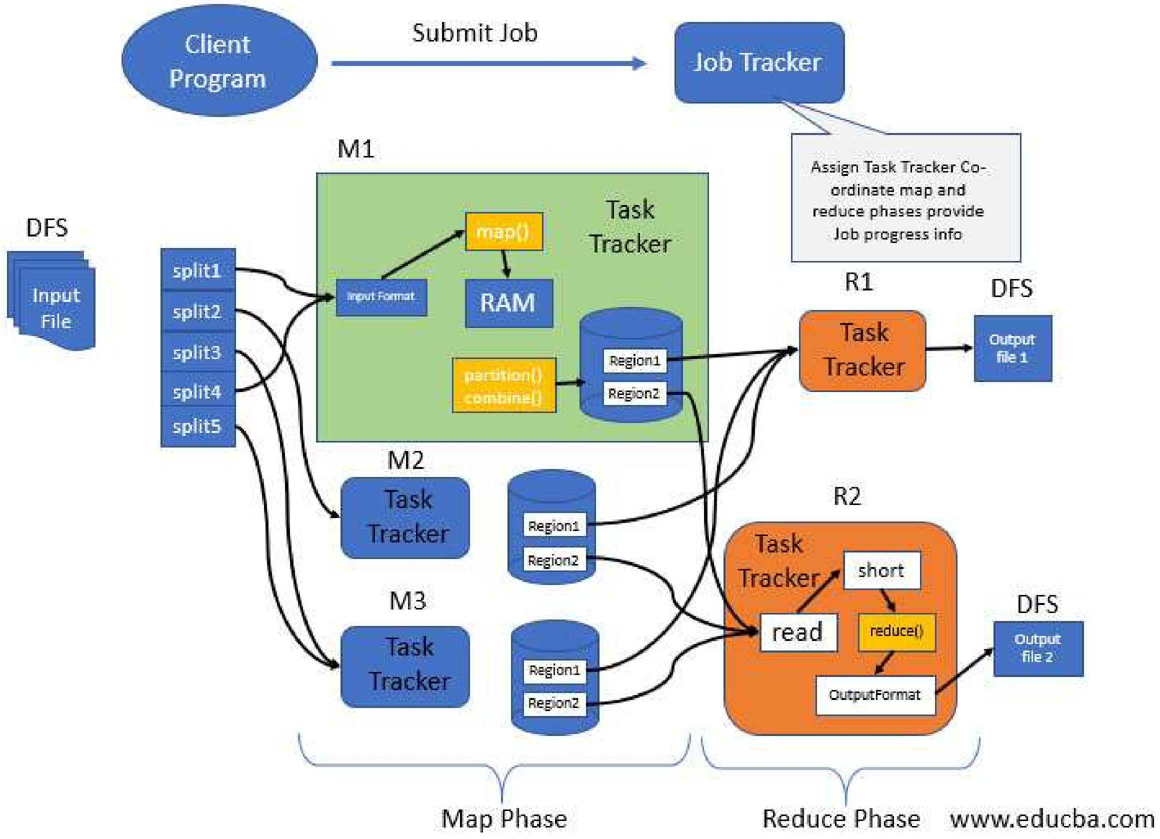
Hadoop MapReduce
or
Apache Spark?

www.educba.com

Presented by Welgama W.I.P.

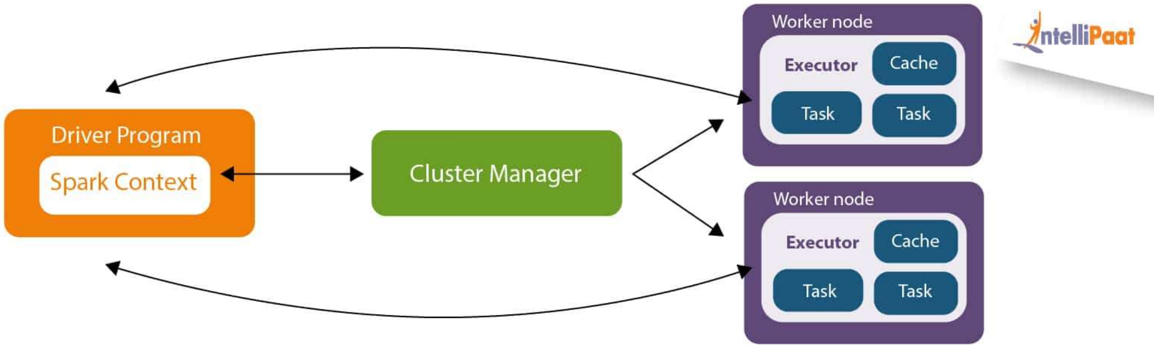
239185V

What is MapReduce?



What is Spark?

Apache Spark is an open-source, distributed computing system designed for big data processing. It provides an interface for programming in various languages such as Java, Python, and Scala, and supports various data sources including Hadoop Distributed File System (HDFS) and Cassandra. Spark's main abstraction is the Resilient Distributed Dataset (RDD), which enables fault-tolerant and parallel processing of data across a cluster of computers.



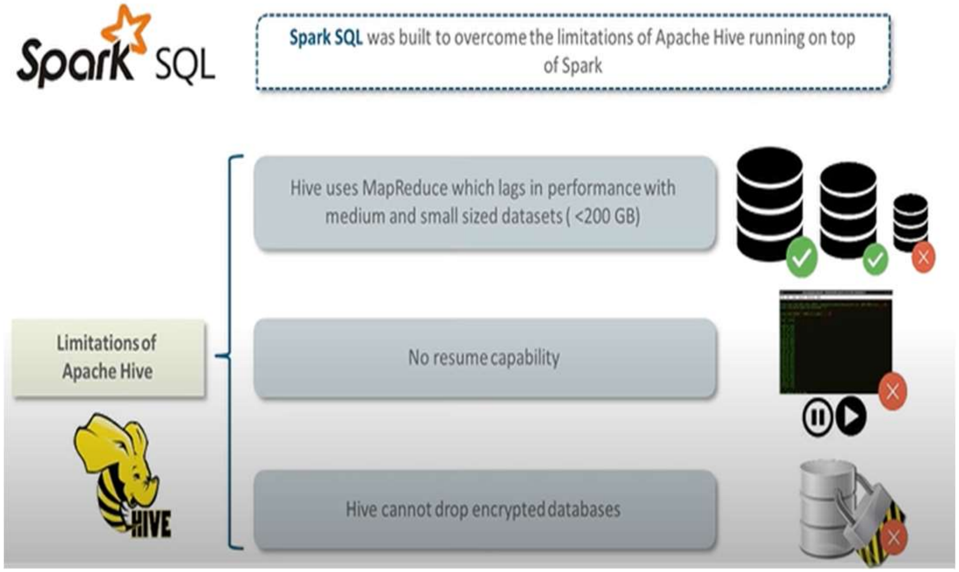
Demo – Part1
MapReduce Framework & HiveQL

Demo – Part2(Using Spark-shell) Apache Spark & Spark-SQL

Demo – Part2(Using Beeline)
Apache Spark & Spark-SQL

Comparison

Limitations of Hive

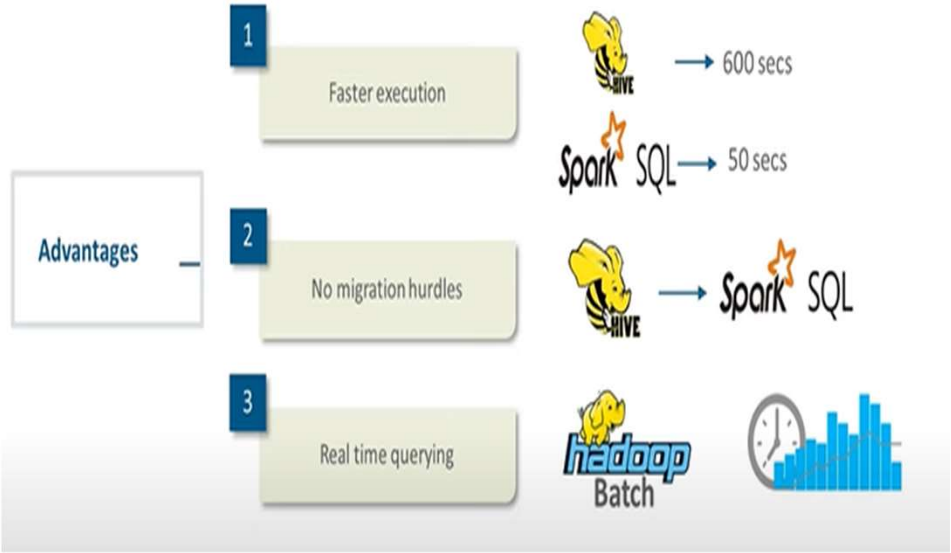


Spark SQL was built to overcome the limitations of Apache Hive running on top of Spark

Limitations of Apache Hive

- Hive uses MapReduce which lags in performance with medium and small sized datasets (<200 GB)
- No resume capability
- Hive cannot drop encrypted databases

Advantages of Apache Spark



Advantages

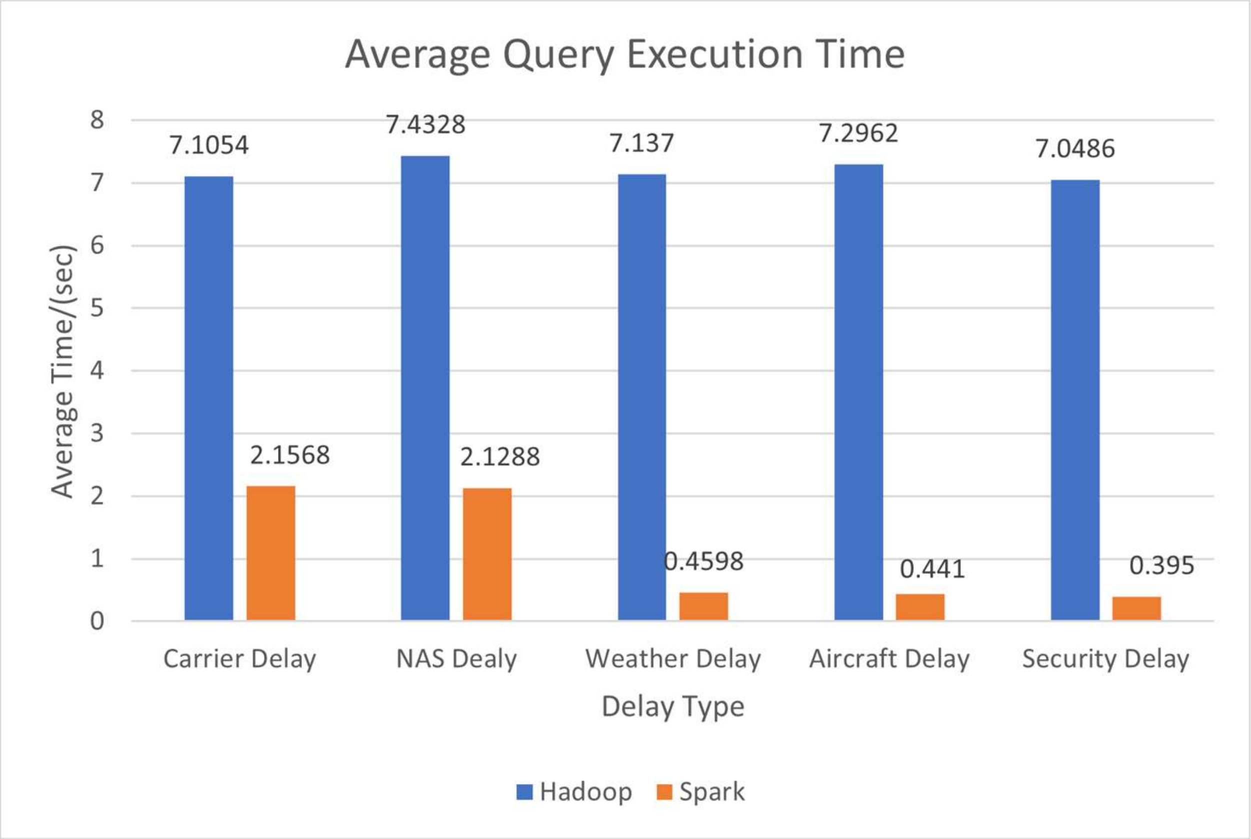
1 Faster execution

2 No migration hurdles

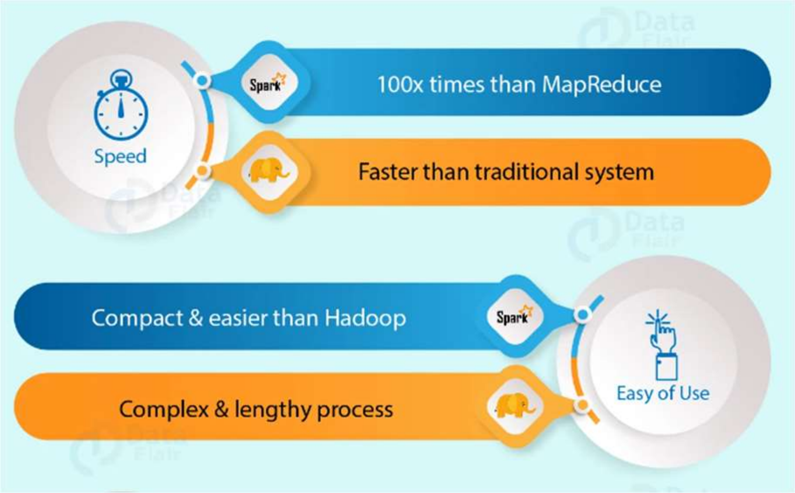
3 Real time querying

Fast Process?

Apache Spark!!



Conclusion



Speed

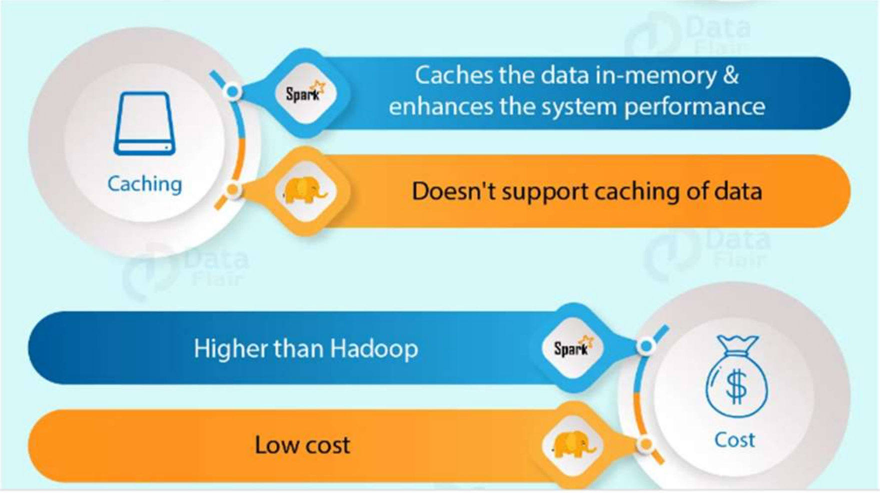
Spark: 100x times than MapReduce

Hadoop: Faster than traditional system

Easy of Use

Spark: Compact & easier than Hadoop

Hadoop: Complex & lengthy process



Caching

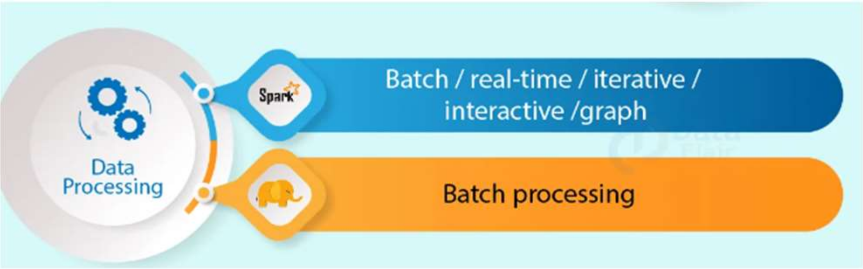
Spark: Caches the data in-memory & enhances the system performance

Hadoop: Doesn't support caching of data

Cost

Spark: Higher than Hadoop

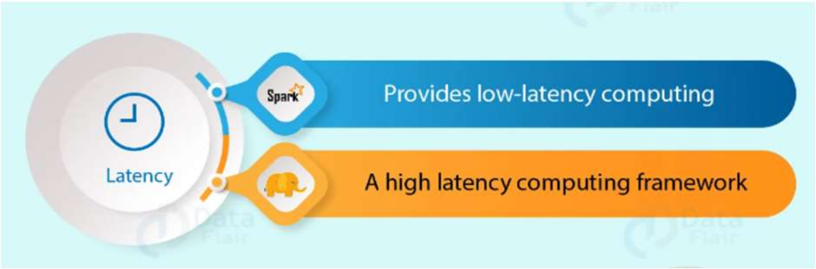
Hadoop: Low cost



Data Processing

Spark: Batch / real-time / iterative / interactive /graph

Hadoop: Batch processing



Latency

Spark: Provides low-latency computing

Hadoop: A high latency computing framework

THANK YOU!