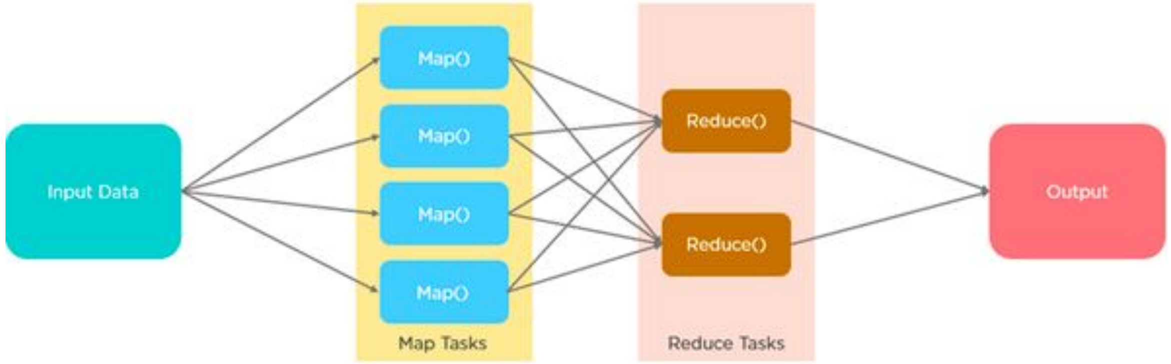
Mapreduce VS Apache Spark

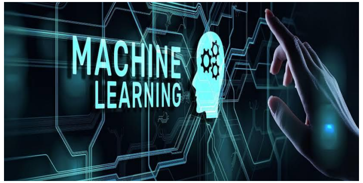
T.H.Sachini Tharinda Chandrasena
239150K

Mapreduce

- Program Paradigm
- Enables massive scalability across hundred/thousands of servers in hadoop cluster
- Heart of Apache Hadoop as the processing component



Apache Spark



- Data processing framework with high performance
- easy -to-use APIs

python™

R Programming

Scala

Java™

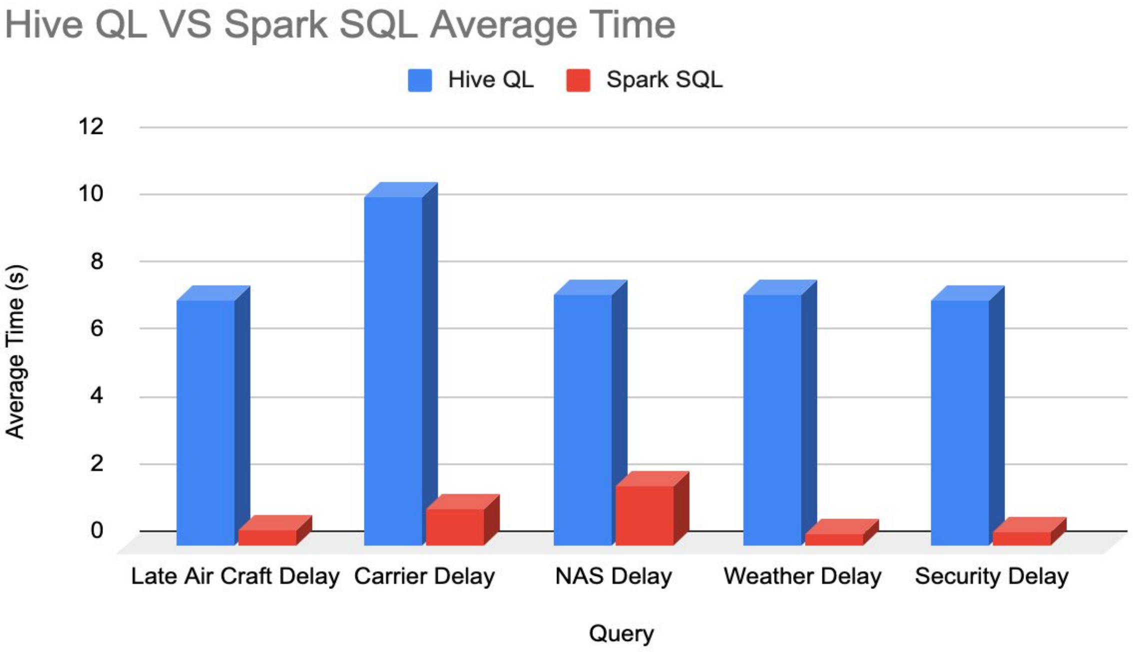
SQL

Demonstration

Fast Process

Spark	Mapreduce
In memory data engine	Designed for batch processing
Deliver near real-time analytics	Not fast as spark but faster than traditional systems
Multi stage jobs	Two stage executions
	Suitable for memory limited big data problems

Experiment


Hive QL VS Spark SQL Average Time
Hive QL
Spark SQL
12
10
8
6
4
2
0
Average Time (s)
Late Air Craft Delay
Carrier Delay
NAS Delay
Weather Delay
Security Delay
Query

Easy to Use

Spark	Mapreduce
Developer Friendly Spark APIs - Scala, Java, Python	Bit complex to write
Rich spark SQL APIs and SQL functions	Difficult to use due to lack of interactive mode
Scope for writing user defined analytical functions	

Conclusion

Spark is way more faster and easy to use than traditional Mapreduce model

Framework of choice when processing big data