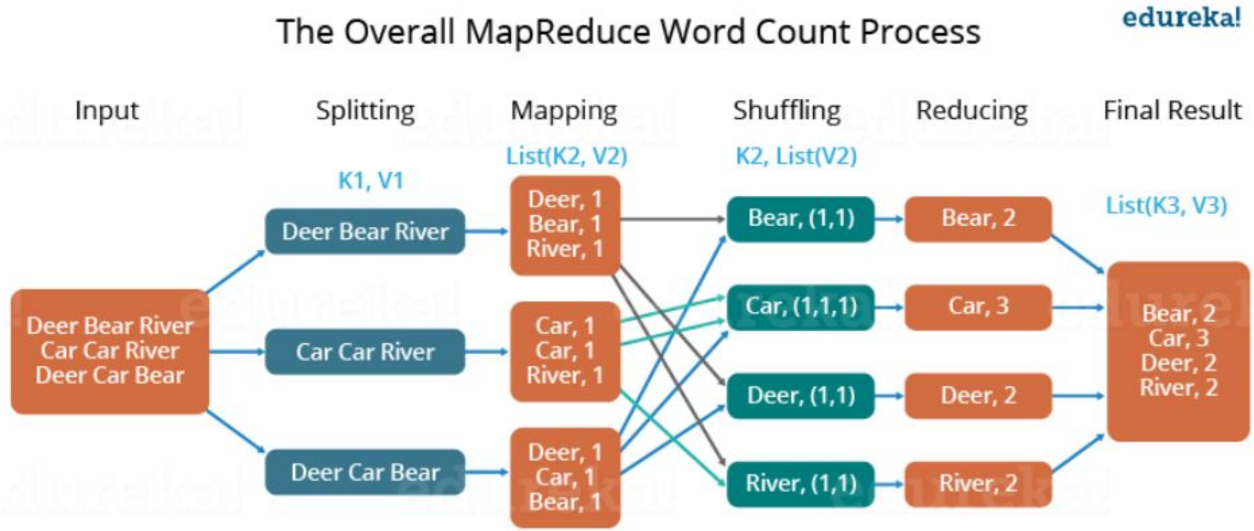
Comparison of MapReduce and Apache Spark

K. Kovishwakarunya - 248357D

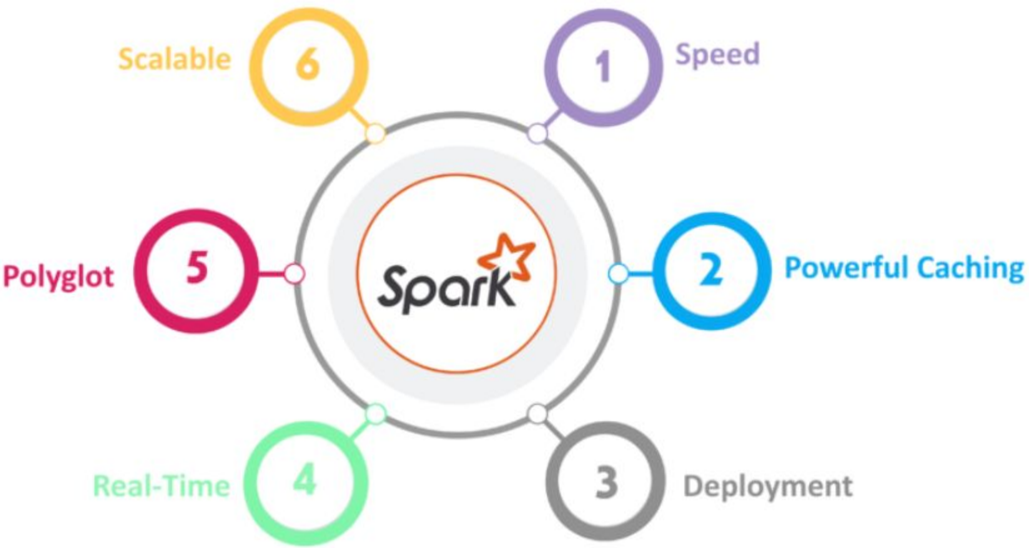
What is MapReduce?

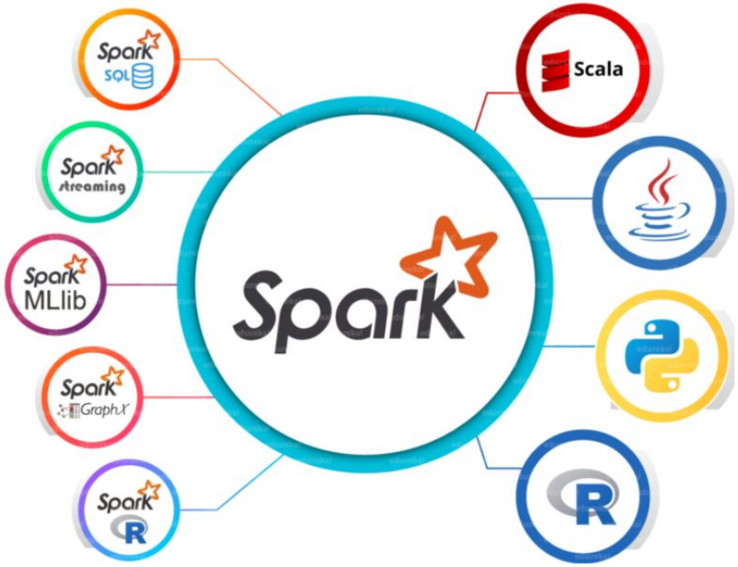
MapReduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment.


The Overall MapReduce Word Count Process

What is Apache Spark?

Apache Spark is an open source cluster computing framework for real-time big data work loads. It utilizes in-memory caching and optimized query execution for fast analytical queries against data of any size.

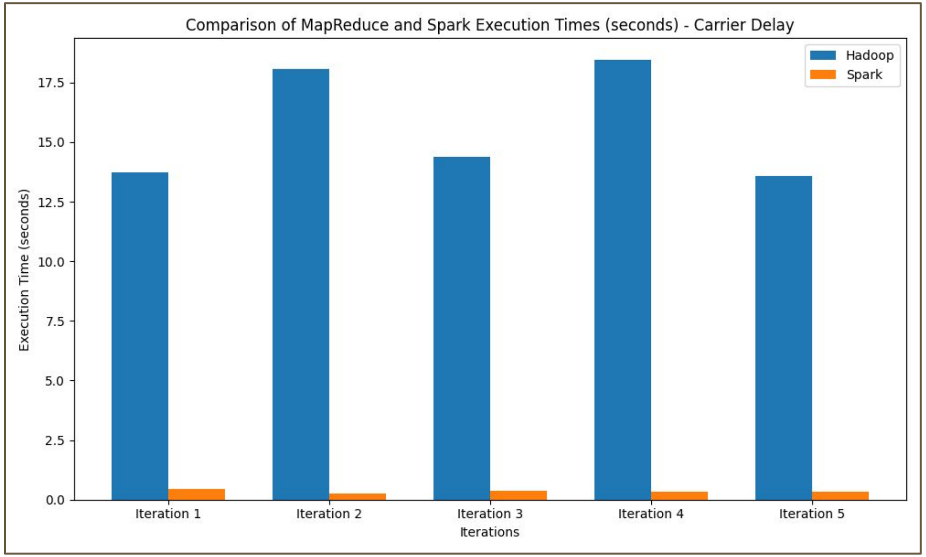




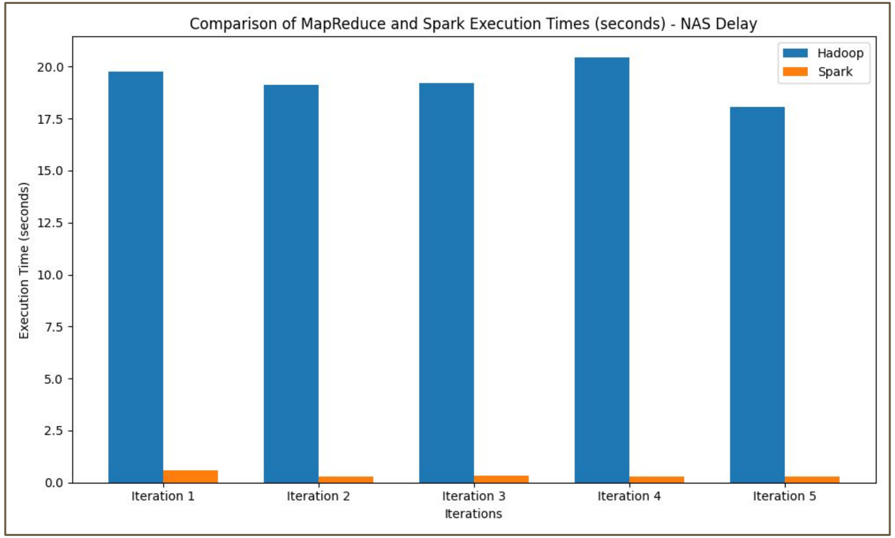
Demo

1. Create AWS EMR Cluster.
2. Uploading the data source to S3 bucket.
3. Processing and Applying queries to the given data.
    a.Hadoop & MapReduce - HiveQL
    b.Apache Spark - Spark-SQL

Results - RunTime Vs Iteration

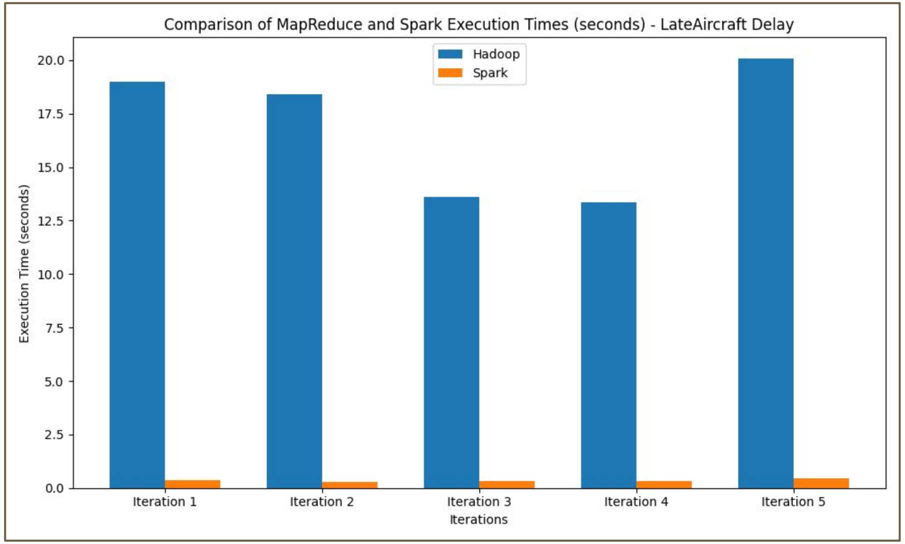


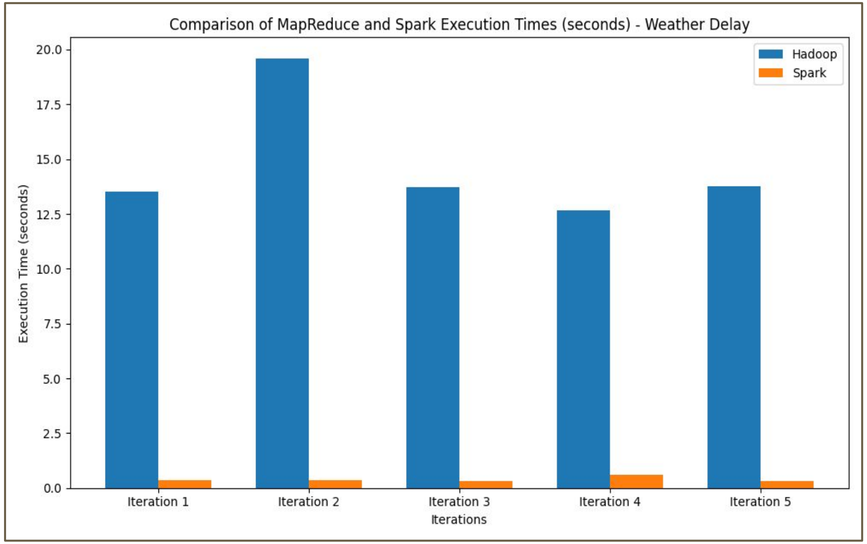
CARRIER DELAY



NAS DELAY

Results - RunTime Vs Iteration

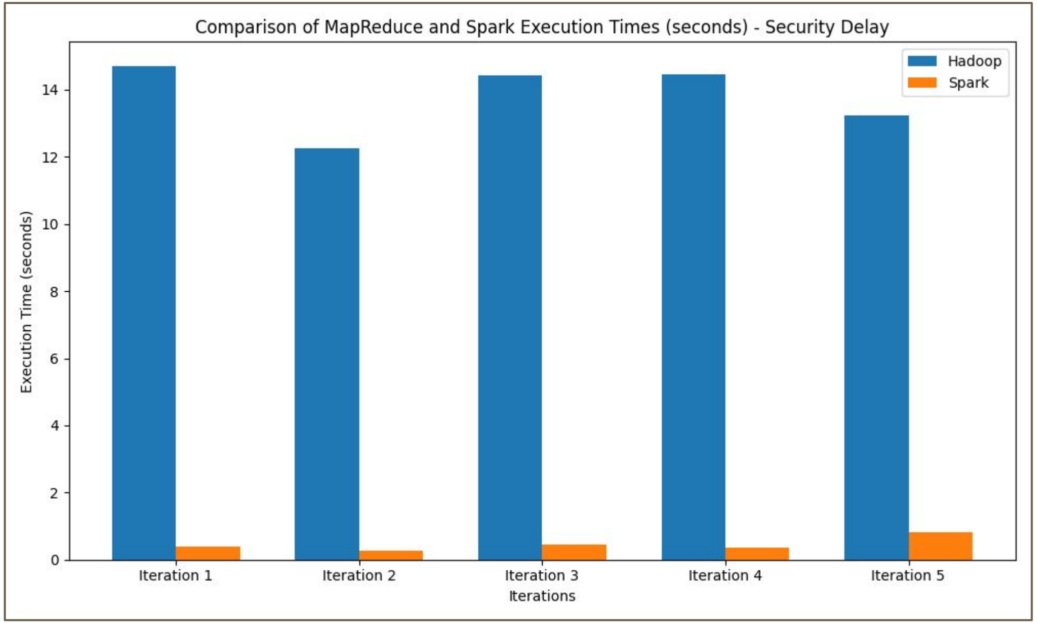




LATE AIRCRAFT DELAY

WEATHER DELAY

Results - RunTime Vs Iteration

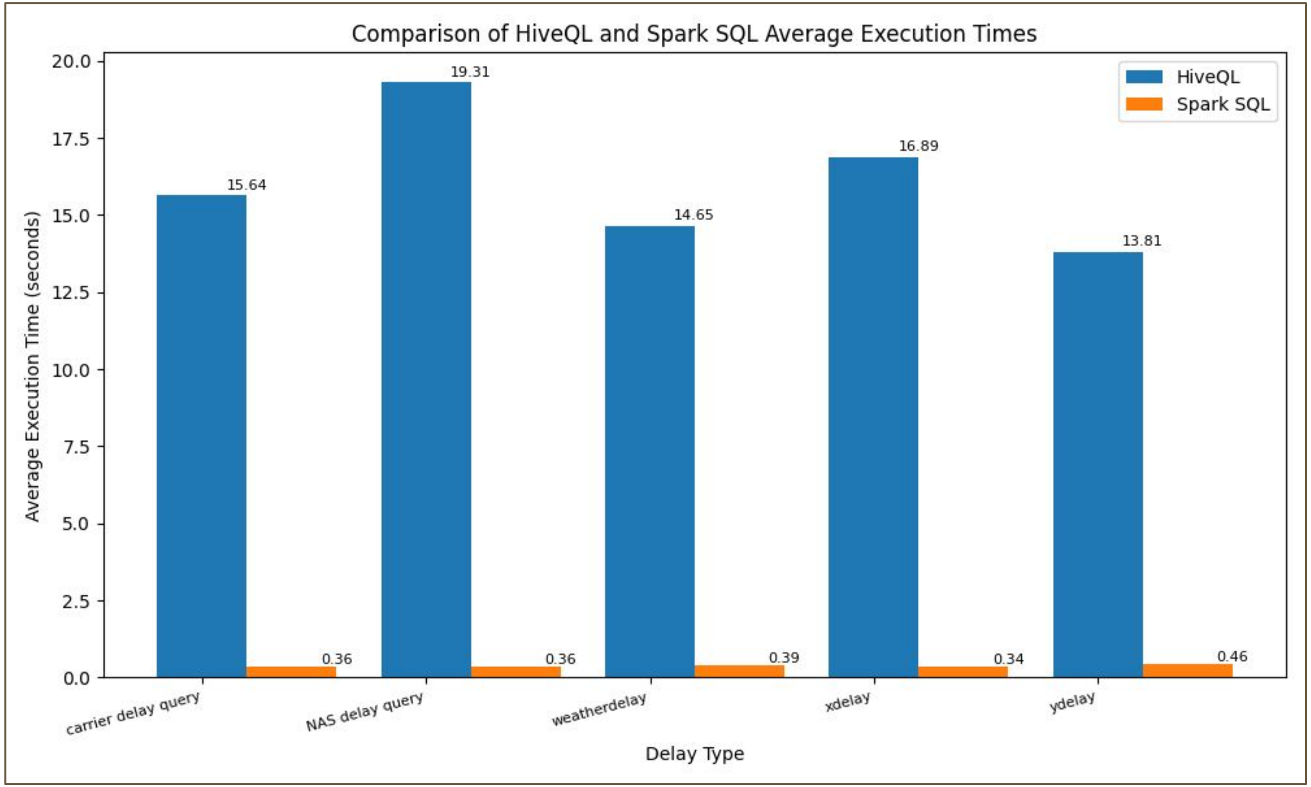


SECURITY DELAY

Results - Average Time Comparison

Time taken by Query in sec	HiveQL	Spark-SQL
Carrier Delay Query	15.63	0.36
NAS Delay Query	19.30	0.35
Weather Delay Query	14.65	0.39
Late Aircraft Delay Query	16.88	0.34
Security Delay Query	13.80	0.45

Results - Average Time Comparison



Hadoop MapReduce Vs Apache Spark

	Hadoop MapReduce	Apache Spark
Ease of use	Complex to use as there is no interactive shell and need to handle low level APIs to process the data.	Supports user friendly APIs for many programming languages. Lower learning curve for developers. Provides an interactive shell to query as well as have immediate feedback
Fast Process	Slow since it performs operations on the disk. Cannot deliver near real time analytics from the data.	Fast because it has in memory processing. It can deliver near real time analytics.