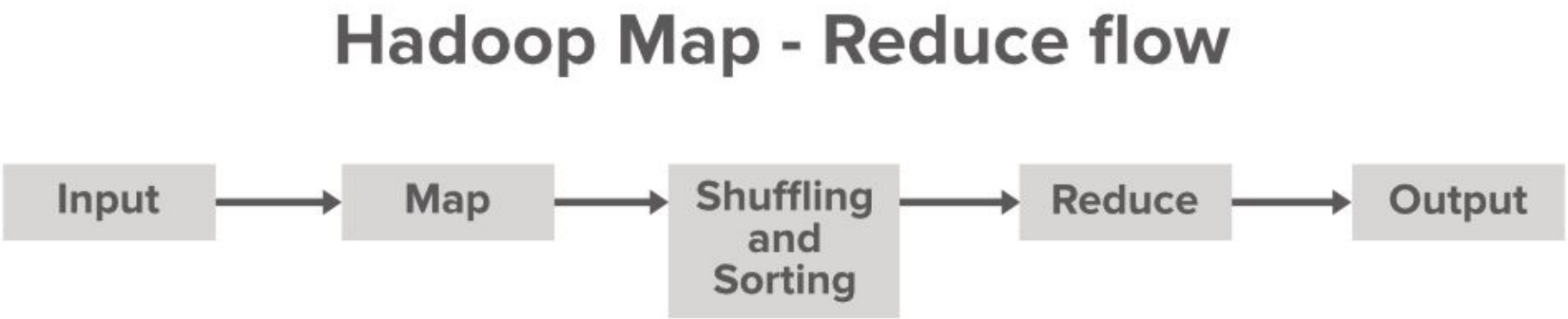
Apache Spark vs Hadoop Map Reduce

229366T W.K.R. Perera

Hadoop Map Reduce



- Hadoop MapReduce is a programming model and framework for processing large datasets in parallel across a cluster of computers
- Two main phases
  - **Map Phase:** Taking input data and transforming it into a set of key-value pairs
  - **Reduce Phase:** Aggregates and summarizes the output of the Map function based on the keys

Apache Spark

- An open-source, distributed computing system that is used for processing large datasets in parallel across a cluster of computers.
- Fast, flexible, and easy to use
- Provides a unified platform for a wide range of data processing tasks, including batch processing, machine learning, and stream processing
- Resilient Distributed Datasets (RDDs) for fault tolerance and efficient data processing
- Interactive and Supports a lot of languages such as Python, Scala etc

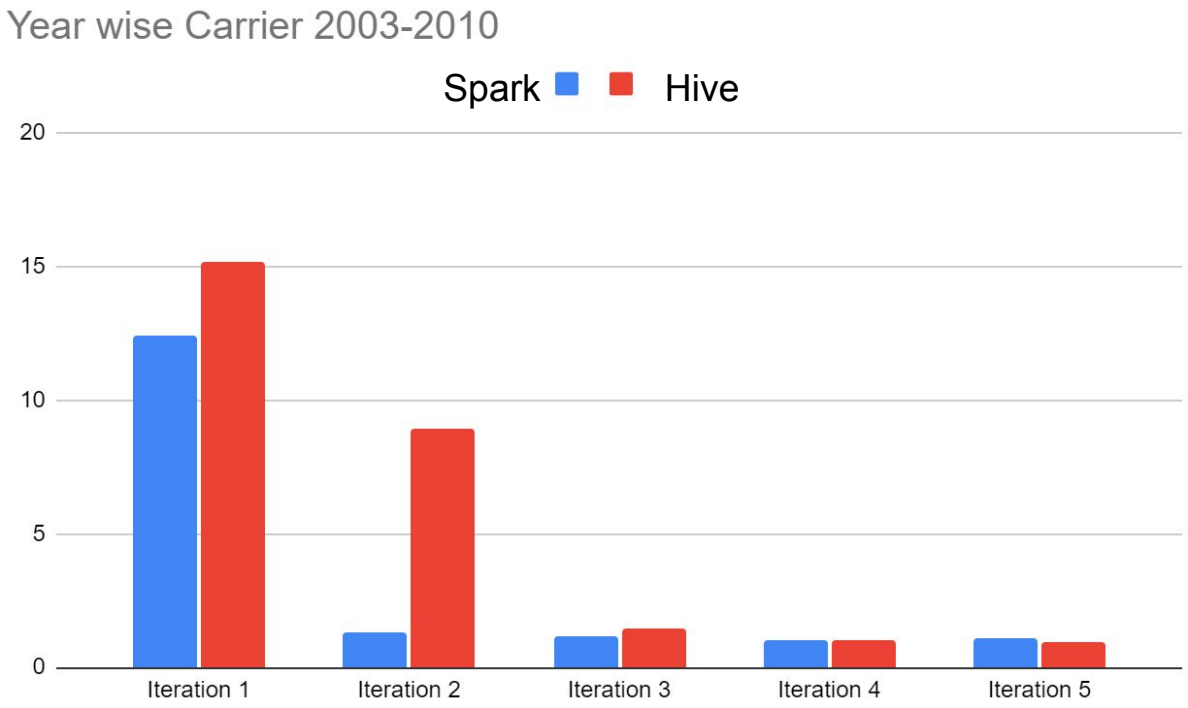
Mapreduce vs Spark

Spark	**Mapreduce**
Easy usage and interactive mode	Difficult usge and no interactive mode
Faster in memory mode (100 x hadoop) and Faster in disk mode (10 x hadoop)	Faster than traditional systems
Can self schedule tasks	Dependent on external job scheduler ex:Oozie
Less secure	More secure
Need to start from scratch if restarted	Can resume work after restarted
Can handle all data processing requirements	Ideal for batch processing
Cost is high	Cost is low

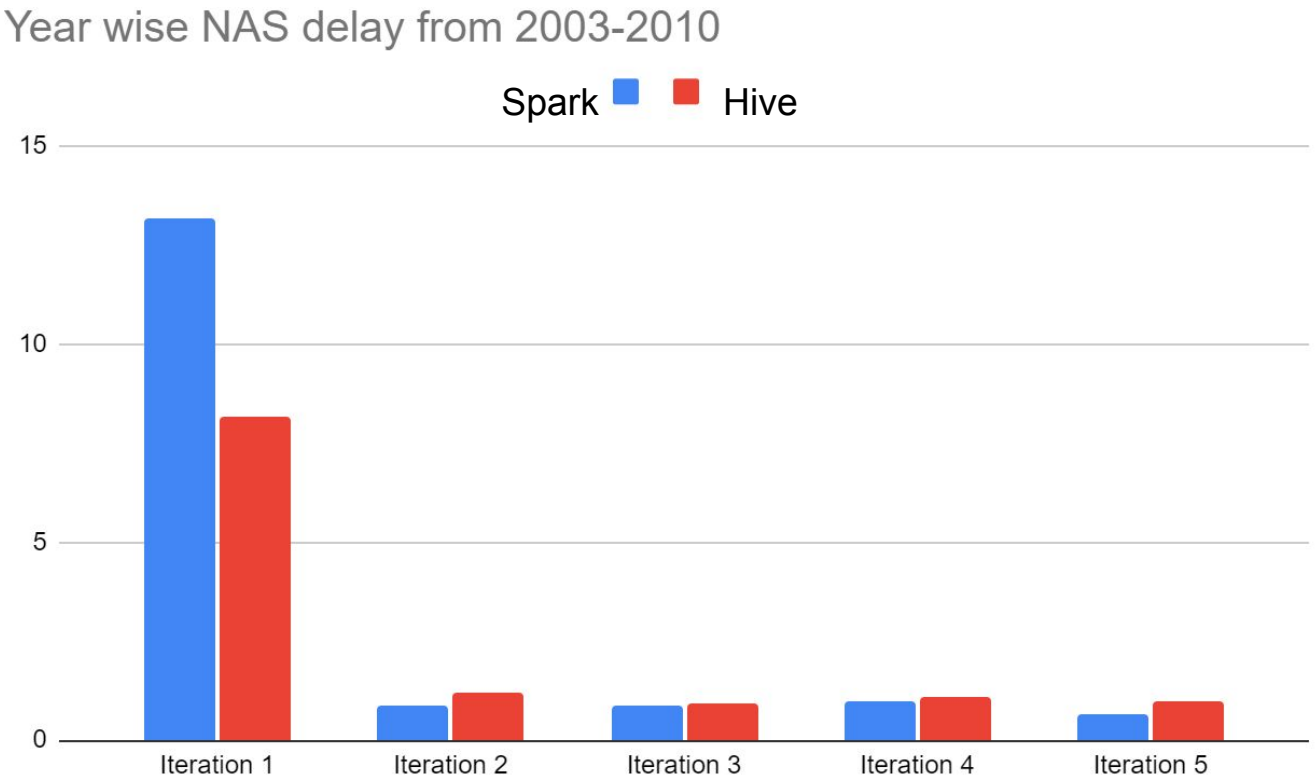
Setup

- Setup:
  - Single Cluster
    - Spark
    - Jupyter
    - Hadoop
  - Single Dataset: Uploaded through a S3 bucket
- Spark Analysis: Analyzed through PySpark
- Hadoop Mapreduce Analysis: Hive Querying

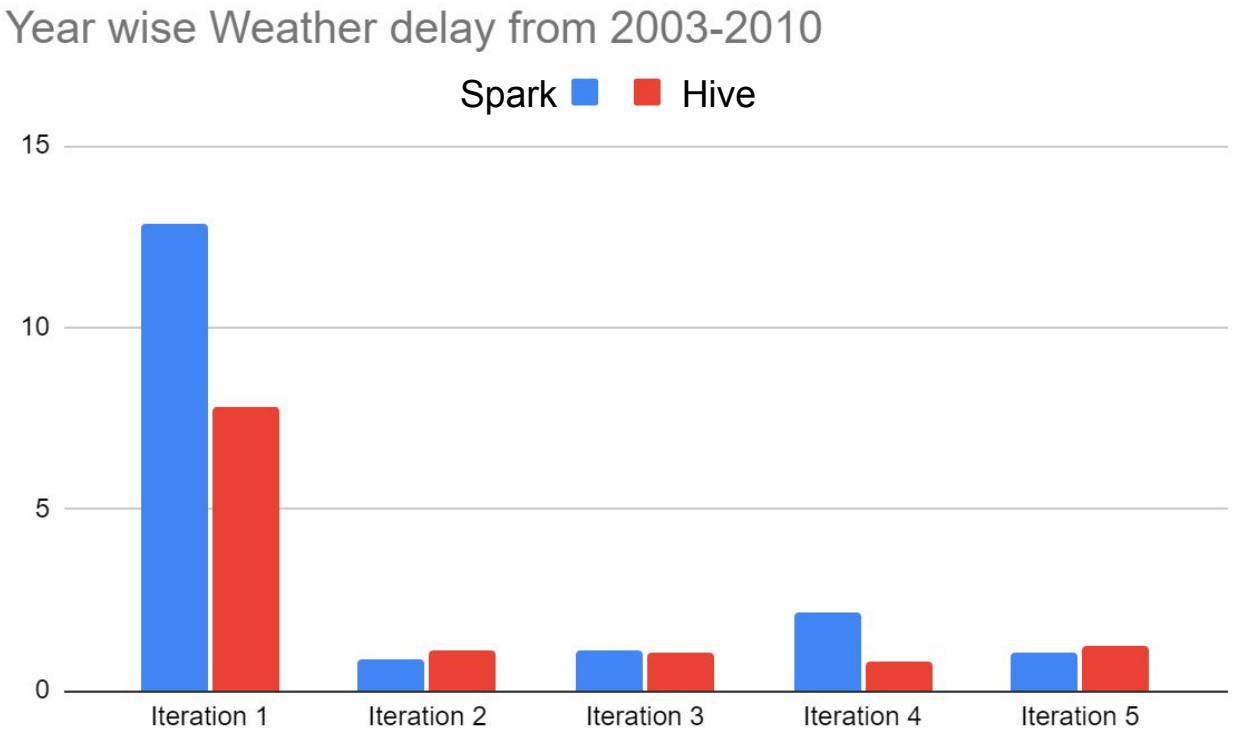
Carrier Delay Query



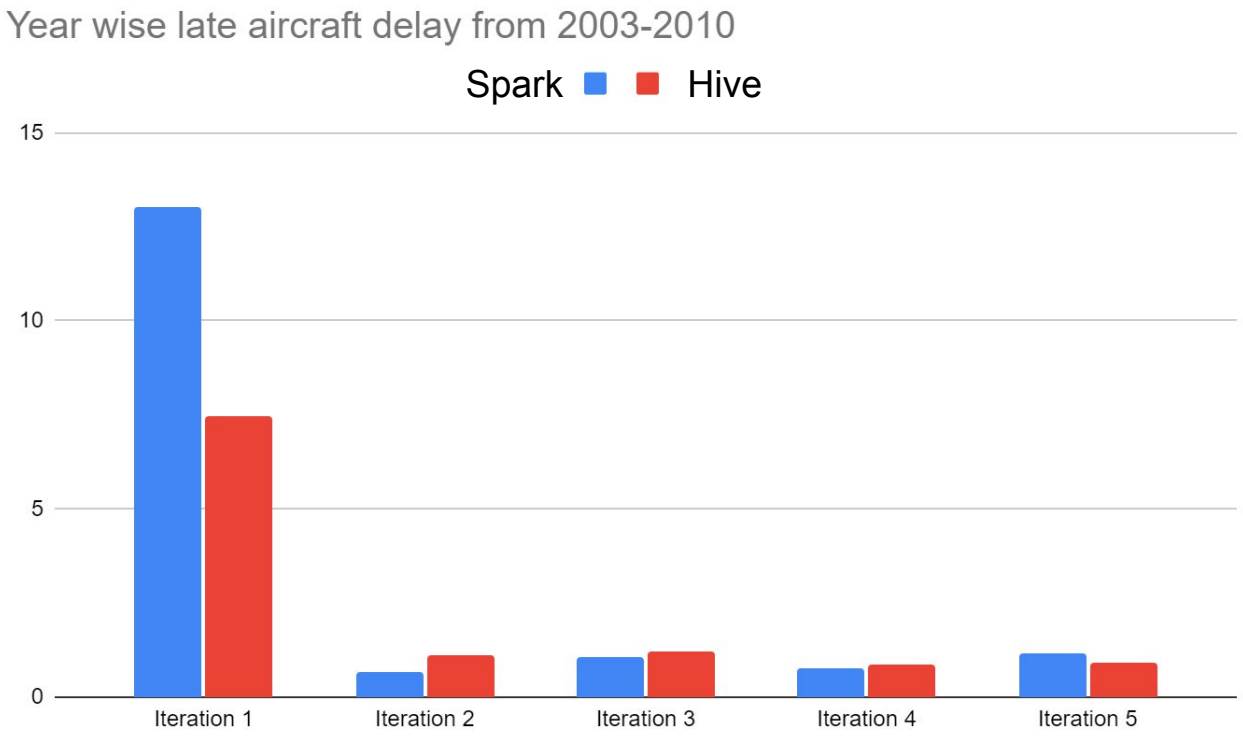
NAS Delay Query



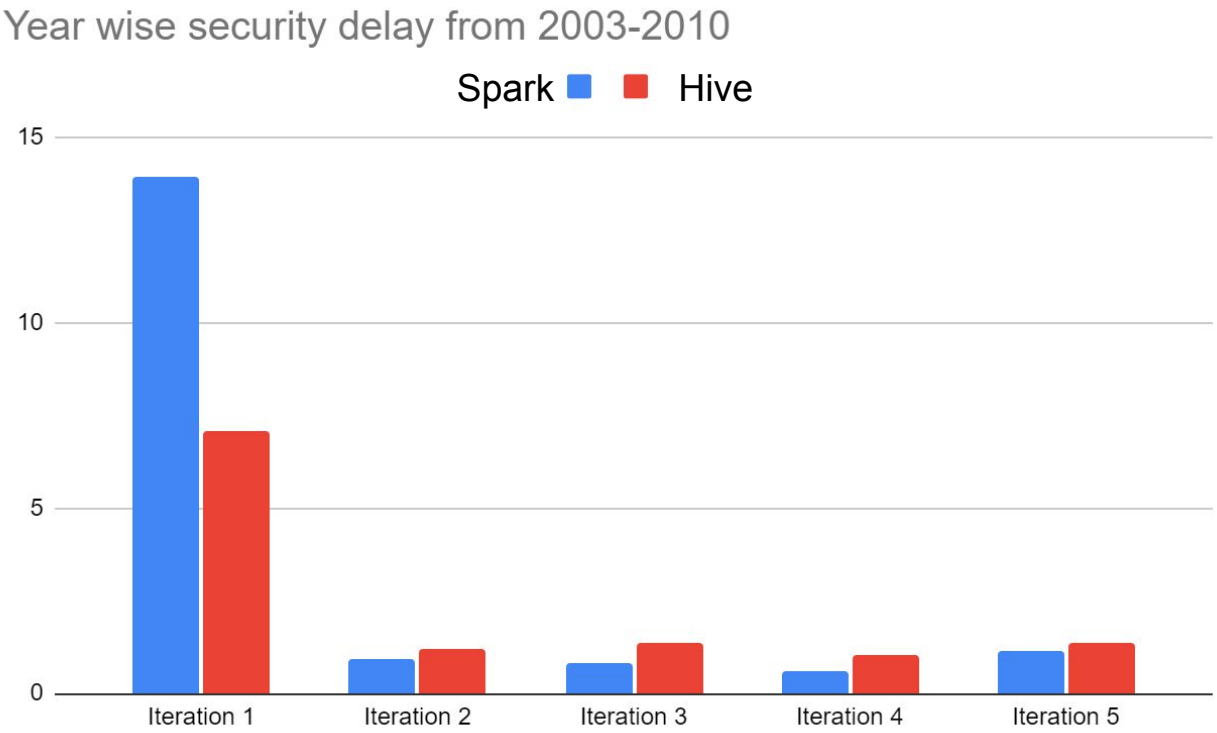
Weather Delay Query



Late Delay Query

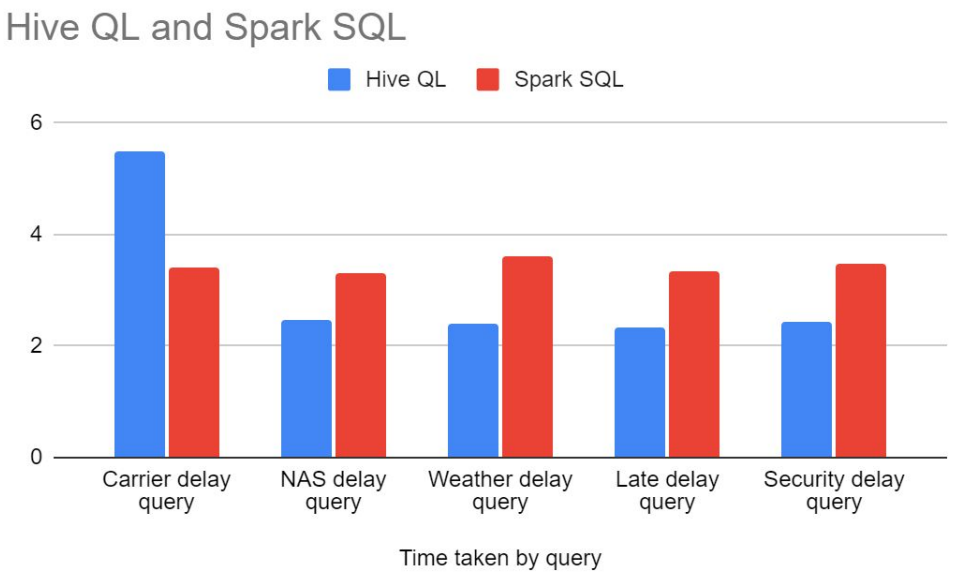


Security Delay Query



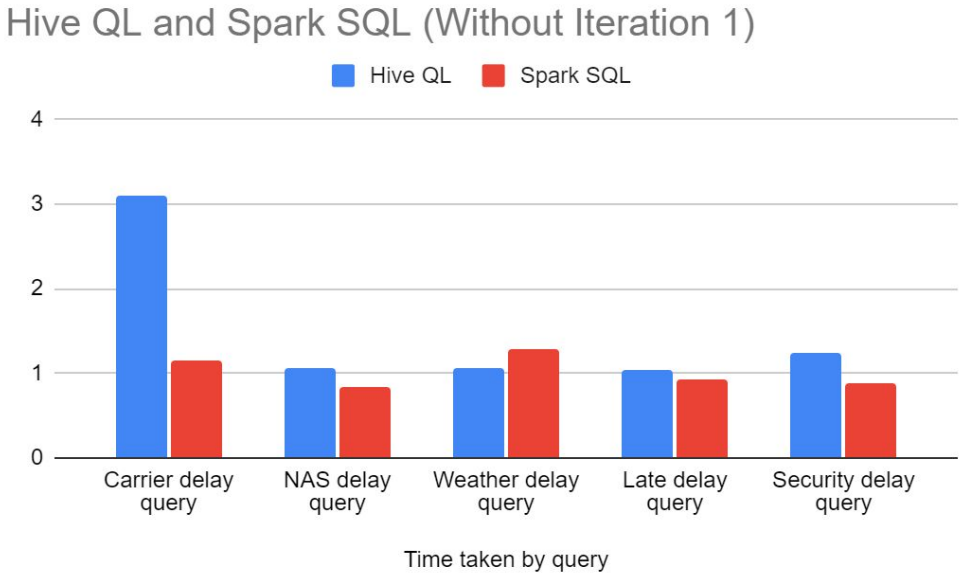
Average in iterations

Time taken by query	Hive QL	Spark SQL
Carrier delay query	5.4996	3.395
NAS delay query	2.4786	3.308
Weather delay query	2.3974	3.605
Late delay query	2.3174	3.345
Security delay query	2.4196	3.487



Averages ignoring first iteration

Time taken by query	Hive QL	Spark SQL
Carrier delay query	3.08775	1.142
NAS delay query	1.0595	0.847
Weather delay query	1.0515	1.289
Late delay query	1.03125	0.919
Security delay query	1.24875	0.871



Conclusion

Spark is a lot more user friendly than Hive because of its interface and integrations

Since we were using a small data set external factors seem to affect more on the results

Overall average of hive was faster but when investigated properly we can see that spark is faster when considering execution time after first iteration.

We can observe that since edata is cached in memory and spark performs better