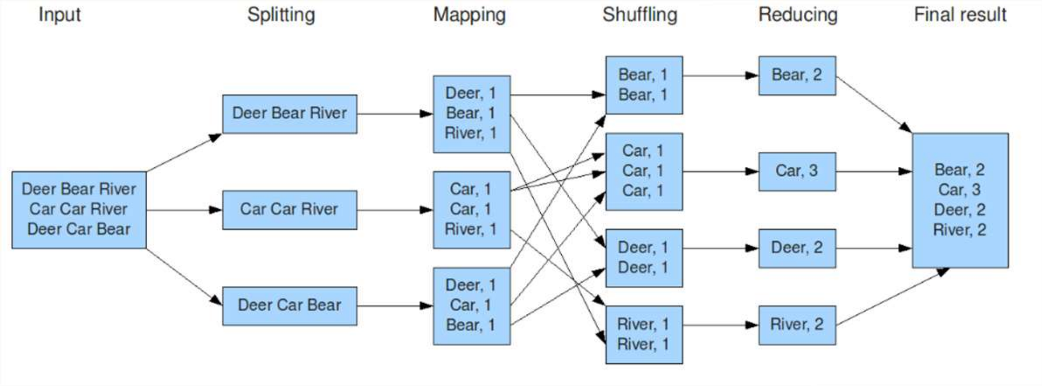
CS5229 - Big Data Analytics Technologies

Kokularaj B.

239329T

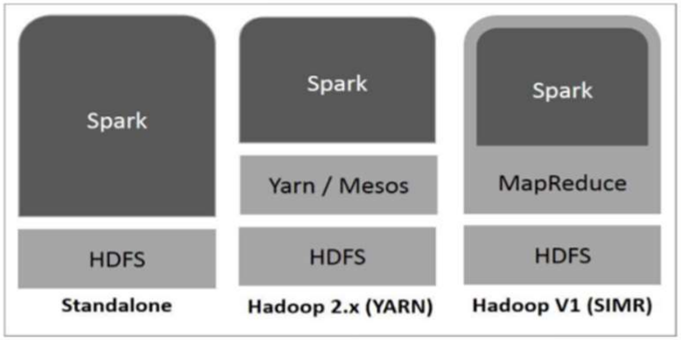
MapReduse

- MapReduce is a software framework and programming model used for processing huge amounts of data in a distributed fashion over several machines.
- Two phases of MapReduce are
  - ➢ Map: Mapping the data set into a collection of key-value pairs. (Map Script)
  - ➢ Reduce: Reducing over all pairs with the same key. (Reduce Script)
- The rest work will be handled by Amazon Elastic MapReduce(EMR) framework.



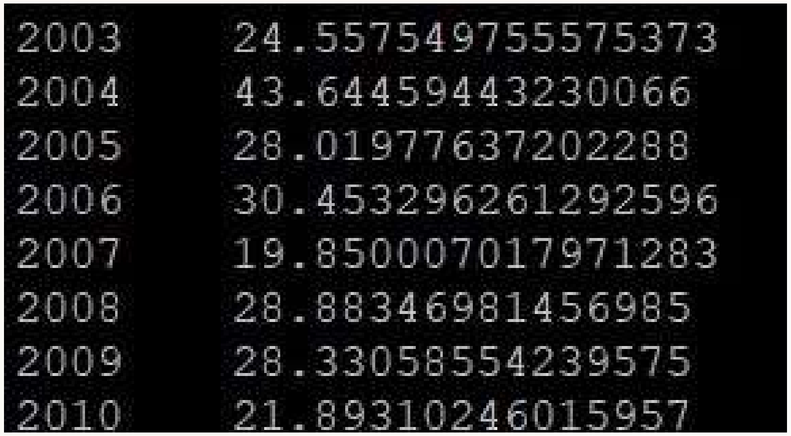
Apache Spark

- It is based on Hadoop MapReduce and extends the MapReduce model to efficiently use it for more types of computations, including interactive queries and stream processing.
- The main feature of Spark is its in-memory cluster computing which increases the processing speed of an application.
- Features of Apache Spark
  - Speed: Up to 100 times faster in memory and 10 times faster in running on disk
  - Supports multiple languages: Java, Scala, Python
  - Advanced Analytics: Support SQL queries, Streaming Data, ML, and Graph Algorithms
- Three ways of Spark deployment
  - Standalone
  - Hadoop Yarn
  - Spark in MapReduce(SIMR)



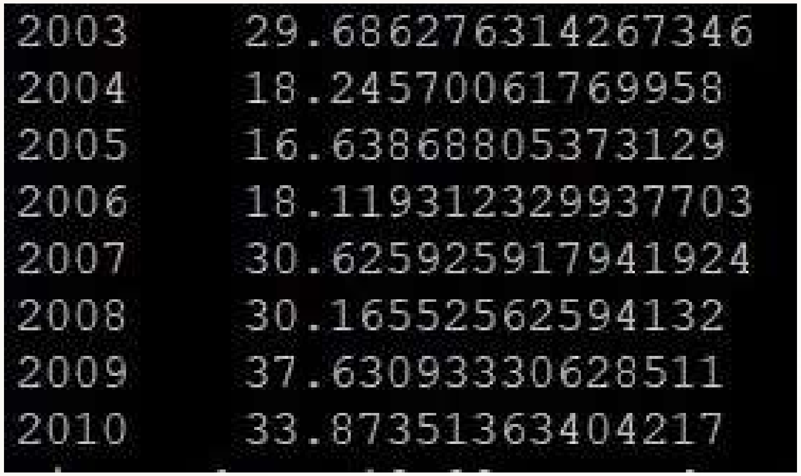
Query and Results

Carrier Delay



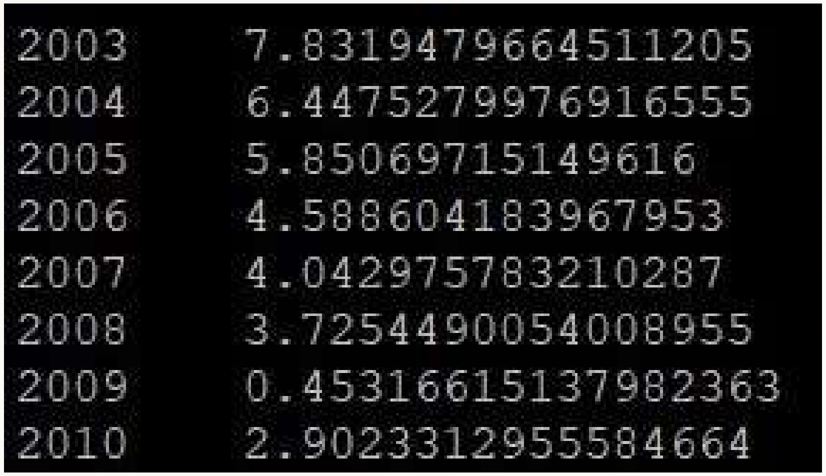
```
2003      24.557549755575373
2004      43.64459443230066
2005      28.01977637202288
2006      30.453296261292596
2007      19.850007017971283
2008      28.88346981456985
2009      28.33058554239575
2010      21.89310246015957
```

NAS Delay



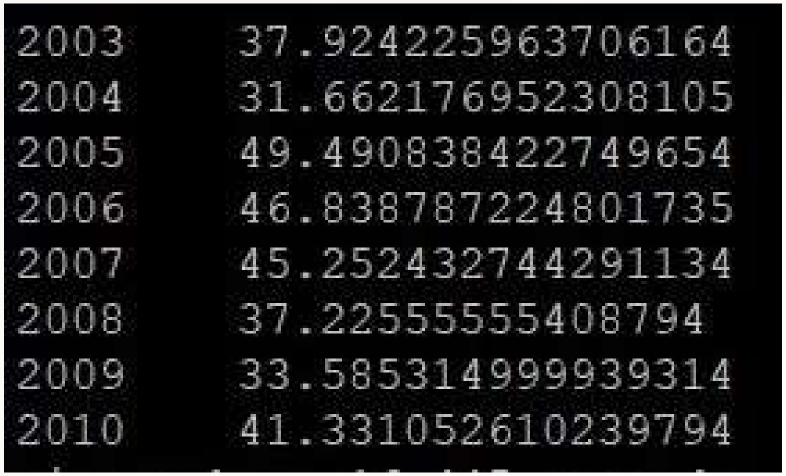
```
2003     29.686276314267346
2004     18.24570061769958
2005     16.63868805373129
2006     18.119312329937703
2007     30.625925917941924
2008     30.16552562594132
2009     37.63093330628511
2010     33.87351363404217
```

Weather Delay



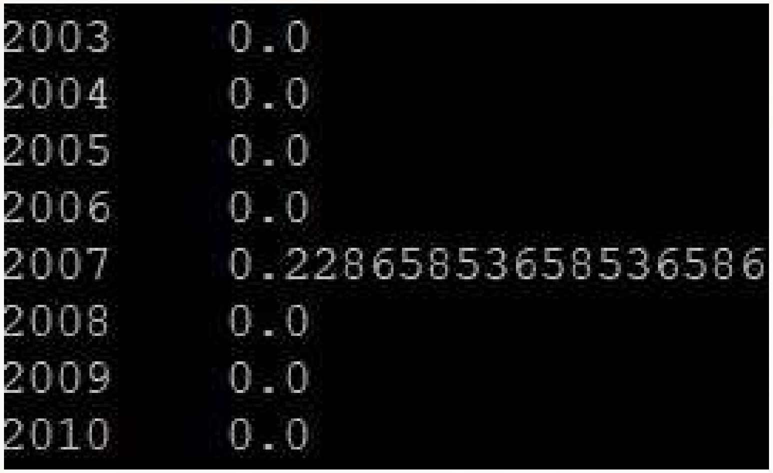
```
2003    7.8319479664511205
2004    6.4475279976916555
2005    5.85069715149616
2006    4.588604183967953
2007    4.042975783210287
2008    3.7254490054008955
2009    0.45316615137982363
2010    2.9023312955584664
```

Late Aircraft Delay



```
2003    37.924225963706164
2004    31.662176952308105
2005    49.490838422749654
2006    46.838787224801735
2007    45.252432744291134
2008    37.22555555408794
2009    33.585314999939314
2010    41.331052610239794
```

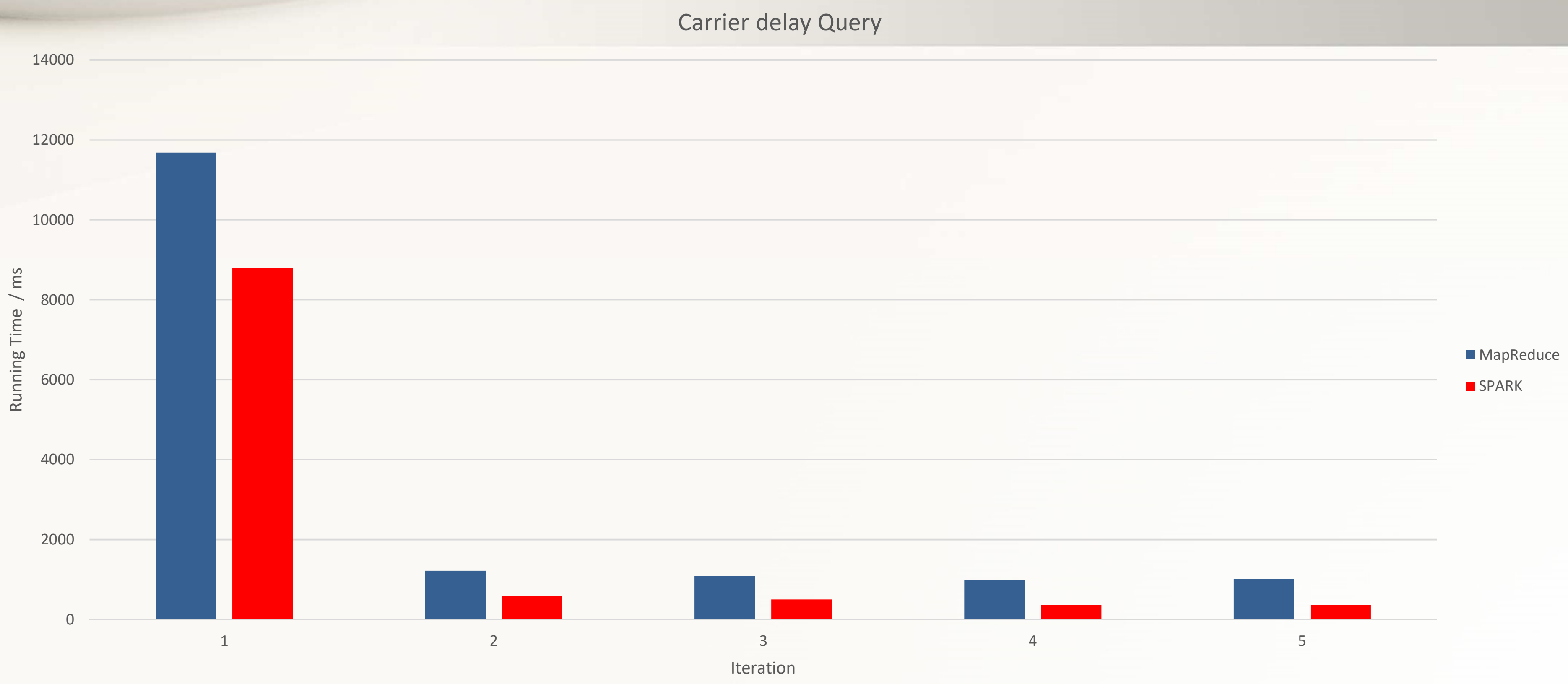
Security Delay



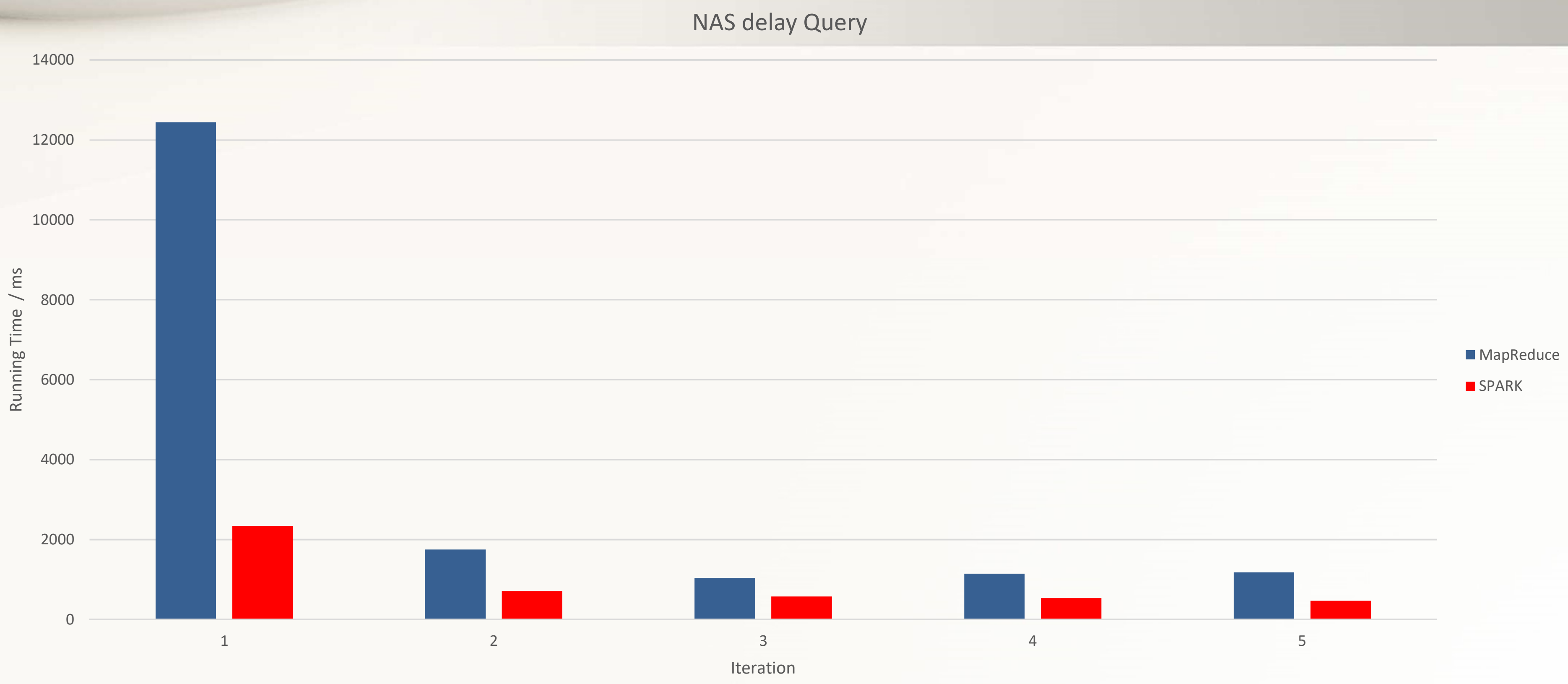
```
2003    0.0
2004    0.0
2005    0.0
2006    0.0
2007    0.22865853658536586
2008    0.0
2009    0.0
2010    0.0
```

Query Vs Running Time

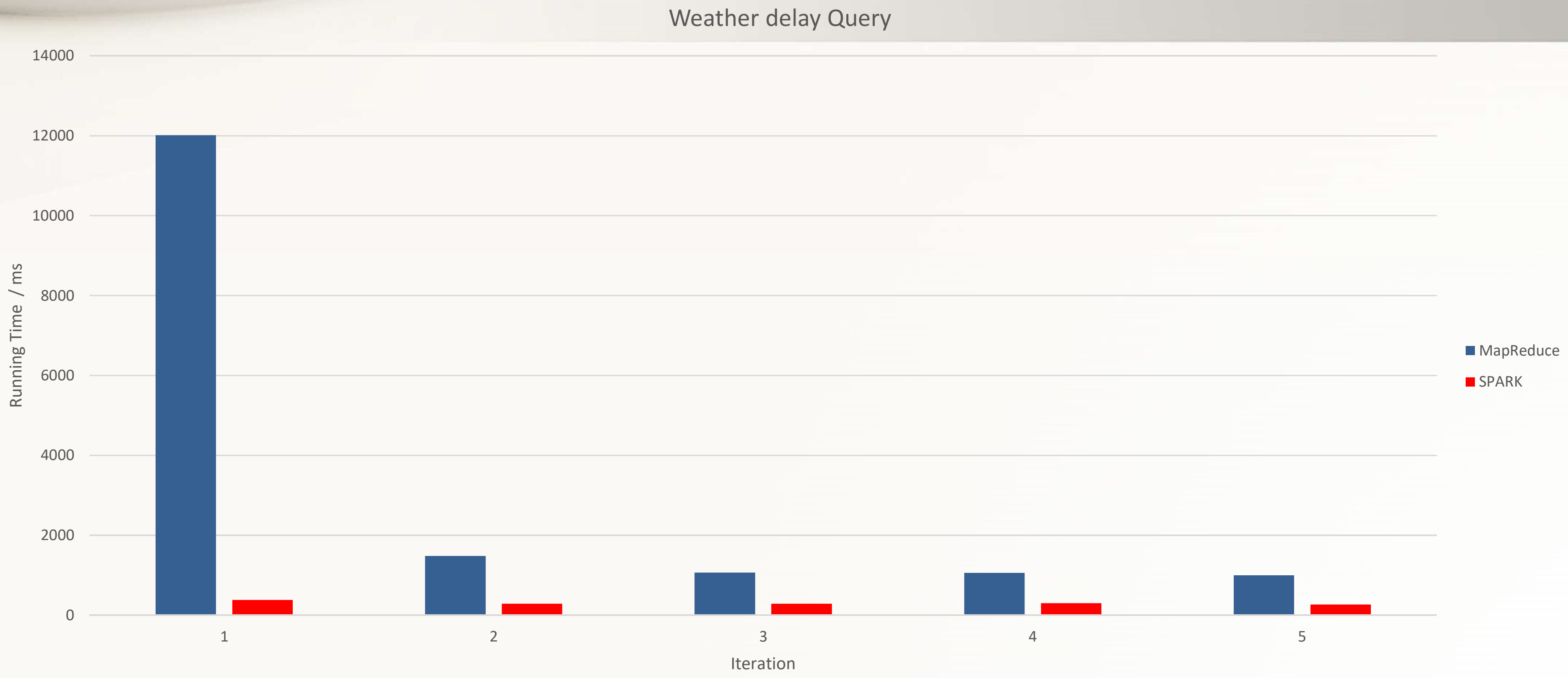
Carrier delay Query



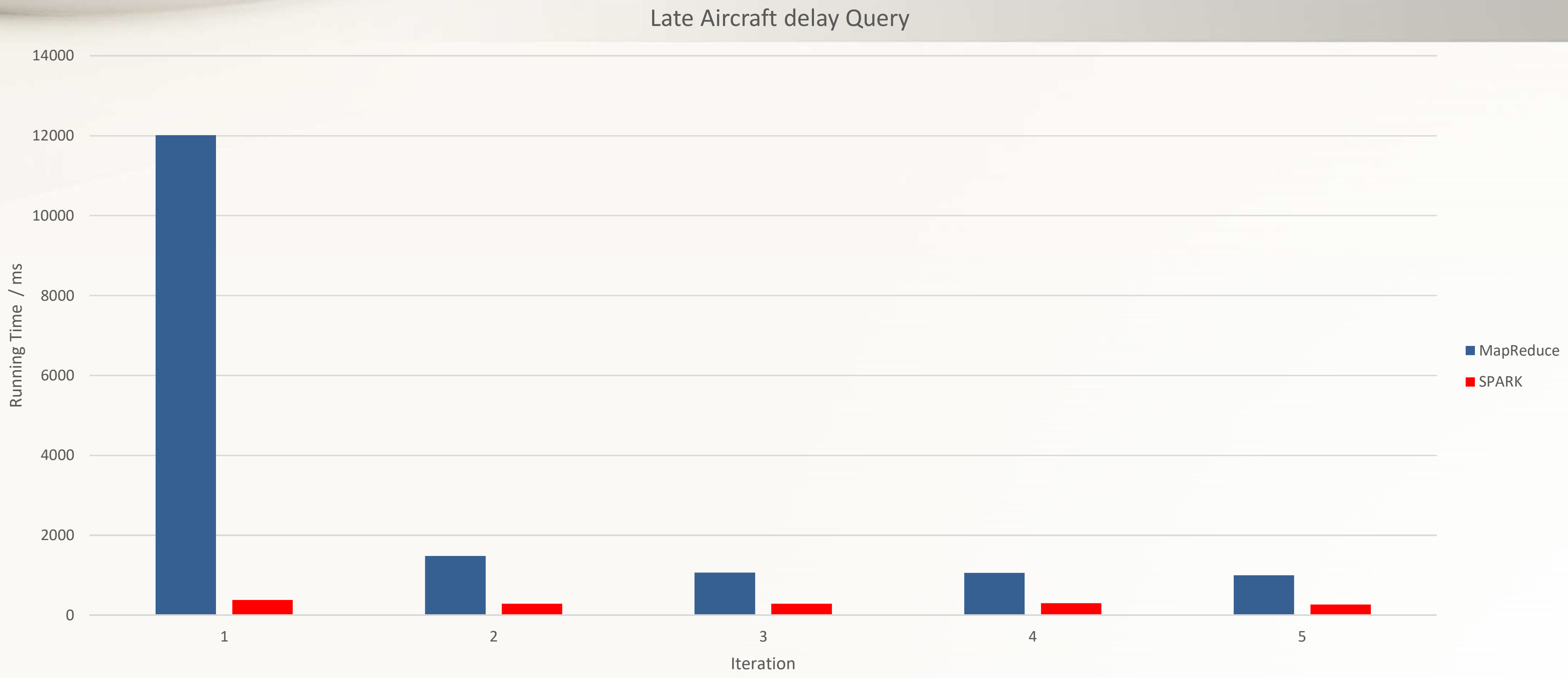
NAS delay Query



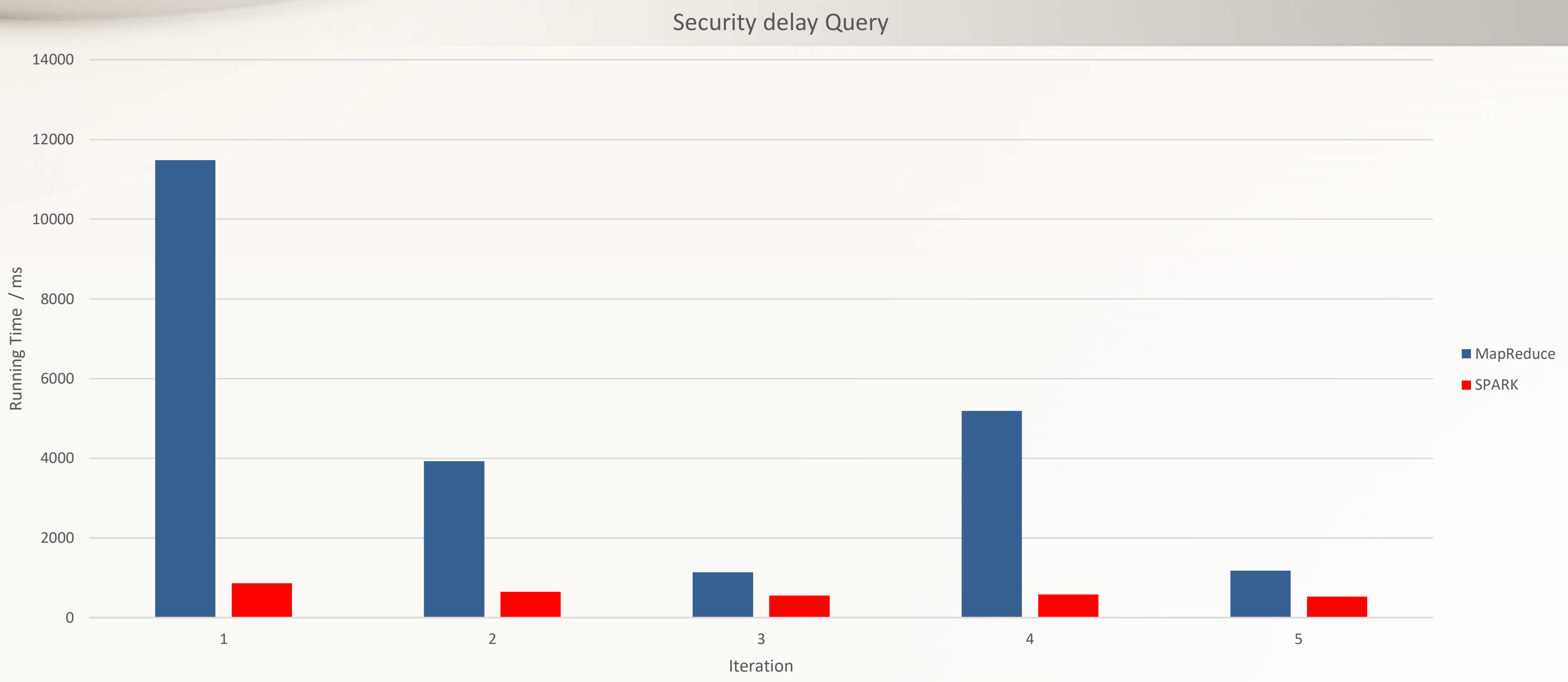
Weather delay Query



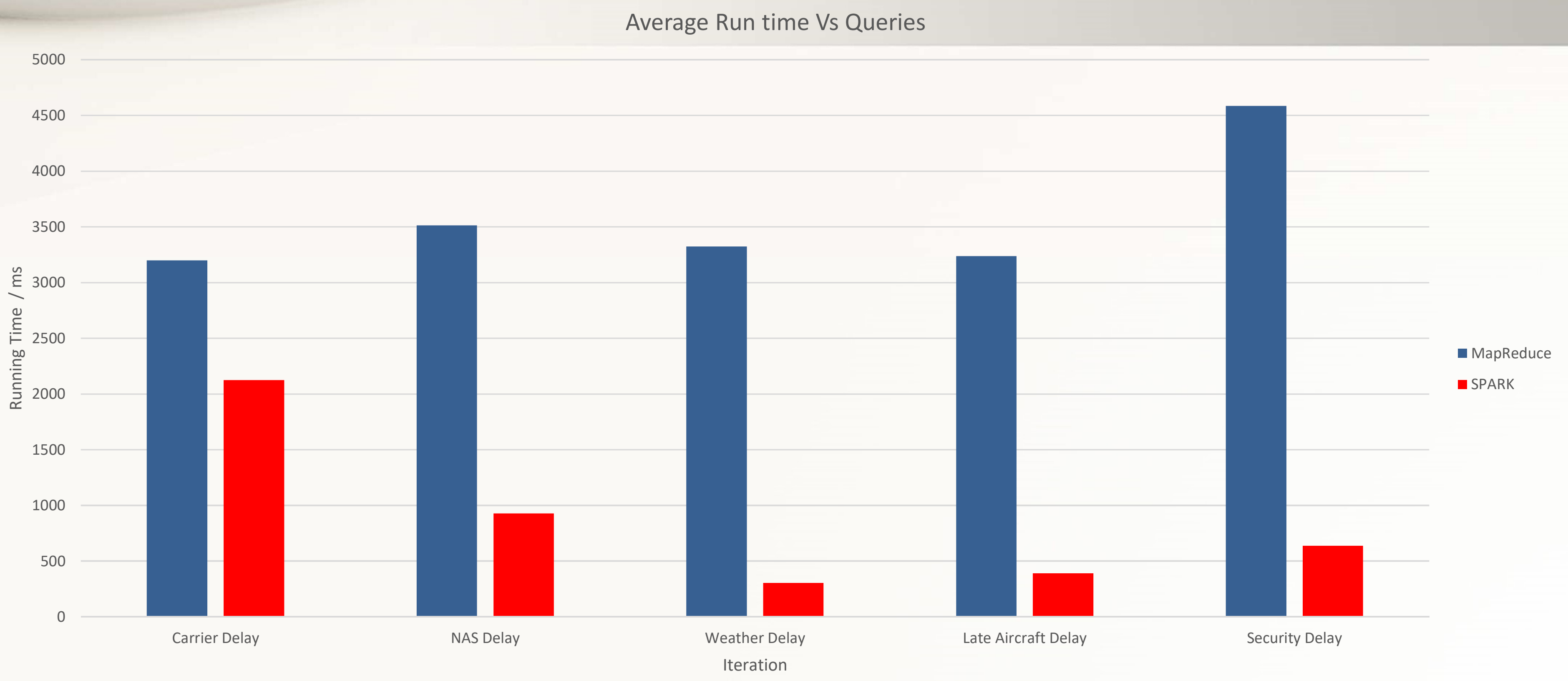
Late Aircraft delay Query



Security delay Query



Run Time Vs Queries



	Apache Spark	MapReduce
Ease of Use	• Apache Spark contains APIs for Scala, Java, and Python and Spark SQL for SQL users. • It can be use in interactive mode when running commands to get an instant response.	• Hadoop MapReduce was developed in Java and is difficult to program. • here is no interactive mode with Hadoop MapReduce
Fast Processing	• Faster • Design for in-memory computing model.	• Comparatively Slow • Design for Large data set in Batch-orient process

Conclusion

- Apache Spark is easier because of its high-level programming model
- Apache Spark is fast processing due its in-memory computing model design.
-   MapReduce can be used for batch process for larger data-set