MapReduce

Hermann Yepdjio

CWU

May 29, 2019

1. Introduction

2. Hadoop MapReduce

3. Google MapReduce

4. Apache Spark

5. Conclusion

Introduction

- MapReduce is a programming model used to process large amount of data in a distributed fashion over several machines(in a network) or processing units(on a single computer)
- It was invented by Google in 2004
- Many implementations such as Hadoop MapReduce, Amazon Elastic MapReduce, Disco, Apache Spark have been released

Hadoop MapReduce

Hadoop MapReduce process goes through 4 different stages namely

- Input splits,
- Mapping,
- Shuffling,
- Reducing.

Hadoop MapReduce Architecture

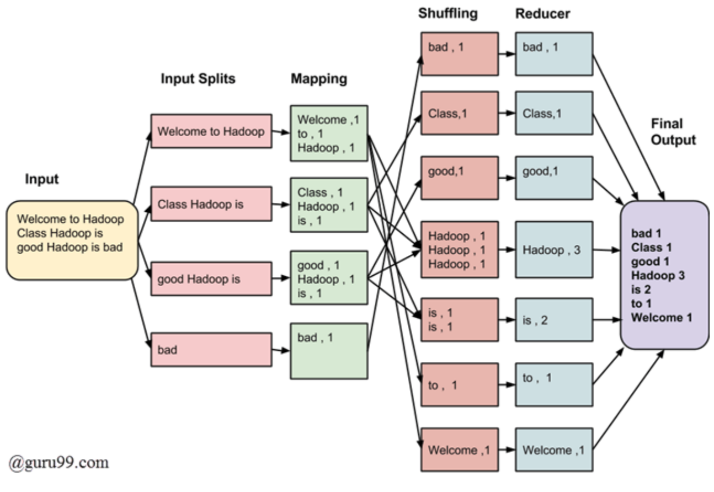


Figure: MapReduce Architecture

Work Organization

- The different tasks or stages can be grouped into two main categories: map tasks (split and mapping) and reduce tasks (shuffling and reducing)
- The whole process is controlled by a job tracker and many task trackers.

Work Organization

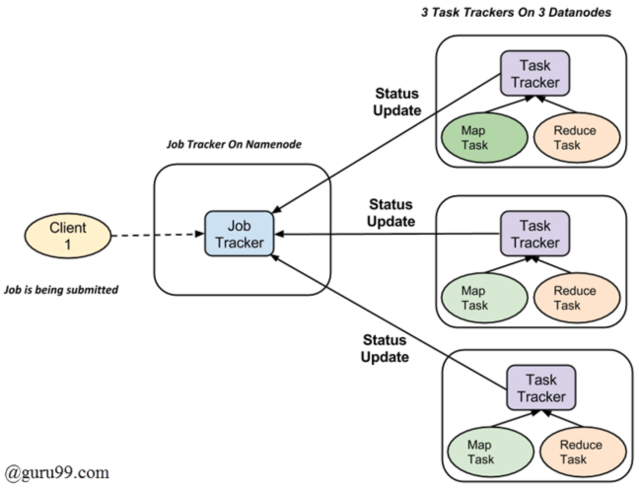
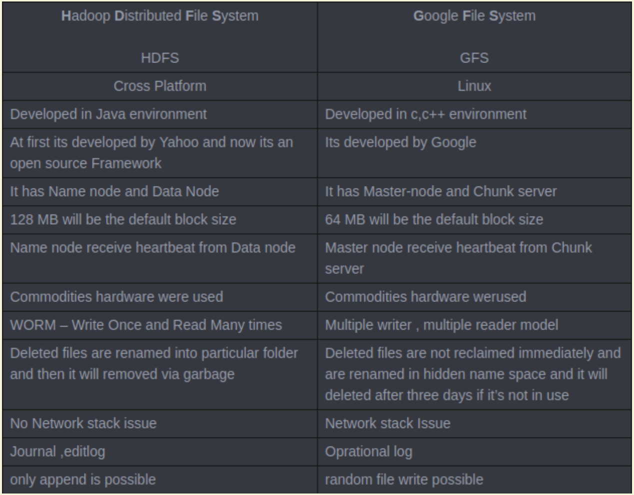


Figure: MapReduce Work Organization

Google MapReduce Vs Hadoop MapReduce

- Implemented using different programming languages (C++ Vs Java),
- Use different file systems (GFS Vs HDFS),
- Proprietary software Vs open source.

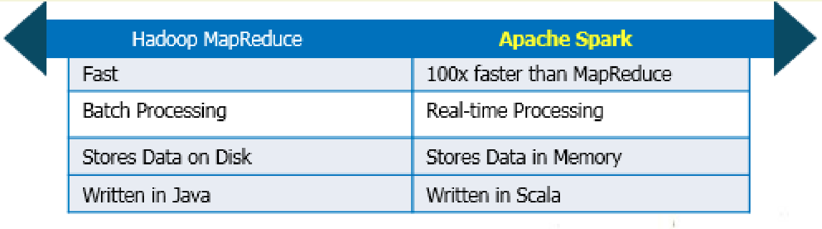
Goggle File System Versus Hadoop Distributed File System



Hadoop Distributed File System HDFS	Google File System GFS
Cross Platform	Linux
Developed in Java environment	Developed in c,c++ environment
At first its developed by Yahoo and now its an open source Framework	Its developed by Google
It has Name node and Data Node	It has Master-node and Chunk server
128 MB will be the default block size	64 MB will be the default block size
Name node receive heartbeat from Data node	Master node receive heartbeat from Chunk server
Commodities hardware were used	Commodities hardware werused
WORM – Write Once and Read Many times	Multiple writer , multiple reader model
Deleted files are renamed into particular folder and then it will removed via garbage	Deleted files are not reclaimed immediately and are renamed in hidden name space and it will deleted after three days if it's not in use
No Network stack issue	Network stack Issue
Journal ,editlog	Oprational log
only append is possible	random file write possible

Figure: GFS Vs HDFS

Apache Spark Vs Hadoop MapReduce



Hadoop MapReduce	Apache Spark
Fast	100x faster than MapReduce
Batch Processing	Real-time Processing
Stores Data on Disk	Stores Data in Memory
Written in Java	Written in Scala

Figure: A Summary of Hadoop MapReduce Vs Spark

Conclusion

- MapReduce is a process which consists in processing large amount of data in a distributed way
- Hadoop MapReduce and Google MapReduce have similar implementations
- Hadoop is open source while Google MapReduce is owned by Google
- Hadoop is fast but Spark is way faster because it processes data from memory while the first processes data from disc