**Hadoop vs Spark: A Comparison**

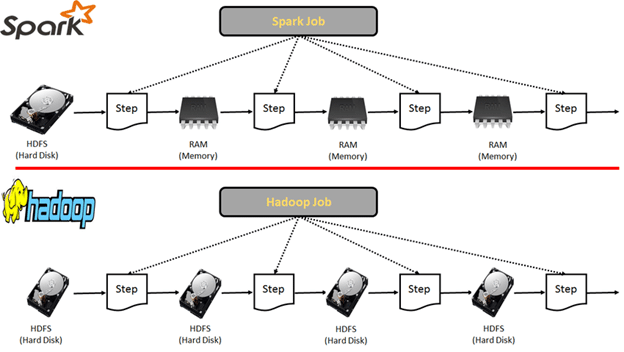
**1. Speed**

In Hadoop, all the data is stored in Hard disks of DataNodes. Whenever the data is required for processing, it is read from hard disk and saved into the hard disk. Moreover, the data is read sequentially from the beginning, so the entire dataset would be read from the disk, not just the portion that is required.

While in Spark, the data is stored in RAM which makes reading and writing data highly faster. Spark is 100 times faster than Hadoop.

Suppose there is a task that requires a chain of jobs, where the output of first is input for second and so on. In MapReduce, the data is fetched from disk and output is stored to disk. Then for the second job, the output of first is fetched from disk and then saved into the disk and so on. Reading and writing data from the disk repeatedly for a task will take a lot of time.

But in Spark, it will initially read from disk and save the output in RAM, so in the second job, the input is read from RAM and output stored in RAM and so on. This reduces the time taken by Spark as compared to MapReduce.



**2. Data Processing**

Hadoop cannot be used for providing immediate results but is highly suitable for data collected over a period of time. Since it is more suitable for batch processing, it can be used for output forecasting, supply planning, predicting the consumer tastes, research, identify patterns in data, calculating aggregates over a period of time etc.

Spark can be used both for both batch processing and real-time processing of data. Even if data is stored in a disk, Spark performs faster. It is suitable for real-time analysis like trending hashtags on Twitter, digital marketing, stock market analysis, fraud detection, etc.

**3. Cost**

Both Hadoop and Spark are open source Apache products, so they are free software. But they have hardware costs associated with them. They are designed to run on low cost, easy to use hardware. Since Hadoop is disk-based, it requires faster disks while Spark can work with standard disks but requires a large amount of RAM, thus it costs more.

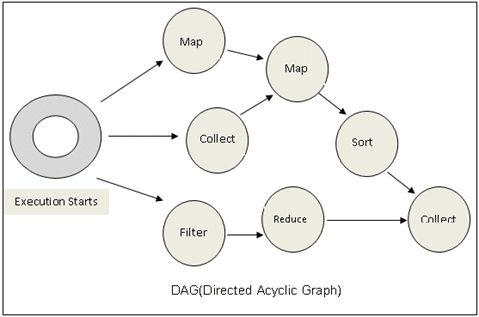
**4. Simplicity**

Spark programming framework is much simpler than MapReduce. It’s APIs in Java, Python, Scala, and R are user-friendly. But Hadoop also has various components which don’t require complex MapReduce programming like Hive, Pig, Sqoop, HBase which are very easy to use.

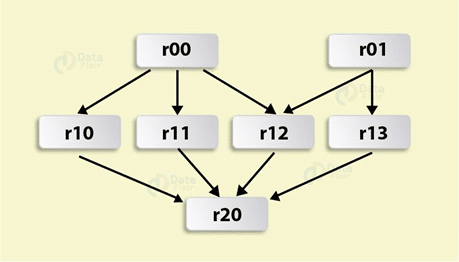
**5. Fault Tolerance**

In Hadoop, the data is divided into blocks which are stored in DataNodes. Those blocks have duplicate copies stored in other nodes with the default replication factor as 3. So, if a node goes down, the data can be retrieved from other nodes. This way, Hadoop achieves fault tolerance.

Spark follows a Directed Acyclic Graph (DAG) which is a set of vertices and edges where vertices represent RDDs and edges represents the operations to be applied on RDDs. In this way, a graph of consecutive computation stages is formed.



Spark builds a lineage which remembers the RDDs involved in computation and its dependent RDDs.



So if a node fails, the task will be assigned to another node based on DAG. Since RDDs are immutable, so if any RDD partition is lost, it can be recomputed from the original dataset using lineage graph. This way Spark achieves fault tolerance.

But for processes that are streaming in real time, a more efficient way to achieve fault tolerance is by saving the state of spark application in reliable storage. This is called checkpointing. Spark can recover the data from the checkpoint directory when a node crashes and continue the process.

**6. Scalability**

Hadoop has its own storage system HDFS while Spark requires a storage system like HDFS which can be easily grown by adding more nodes. They both are highly scalable as HDFS storage can go more than hundreds of thousands of nodes. Spark can also integrate with other storage systems like S3 bucket.

It is predicted that 75% of Fortune 2000 companies will have a 1000 node Hadoop cluster.

Facebook has 2 major Hadoop clusters with one of them being an 1100 machine cluster with 8800 cores and 12 PB raw storage.

Yahoo has one of the biggest Hadoop clusters with 4500 nodes. It has more than 100,000 CPUs in greater than 40,000 computers running Hadoop.

*Source:*[*https://wiki.apache.org/hadoop/PoweredBy*](https://wiki.apache.org/hadoop/PoweredBy)

**7. Security**

Spark only supports authentication via shared secret password authentication.

While Hadoop supports Kerberos network authentication protocol and HDFS also supports Access Control Lists (ACLs) permissions. It provides service level authorization which is the initial authorization mechanism to ensure the client has the right permissions before connecting to Hadoop service.

So Spark is little less secure than Hadoop. But if it is integrated with Hadoop, then it can use its security features.