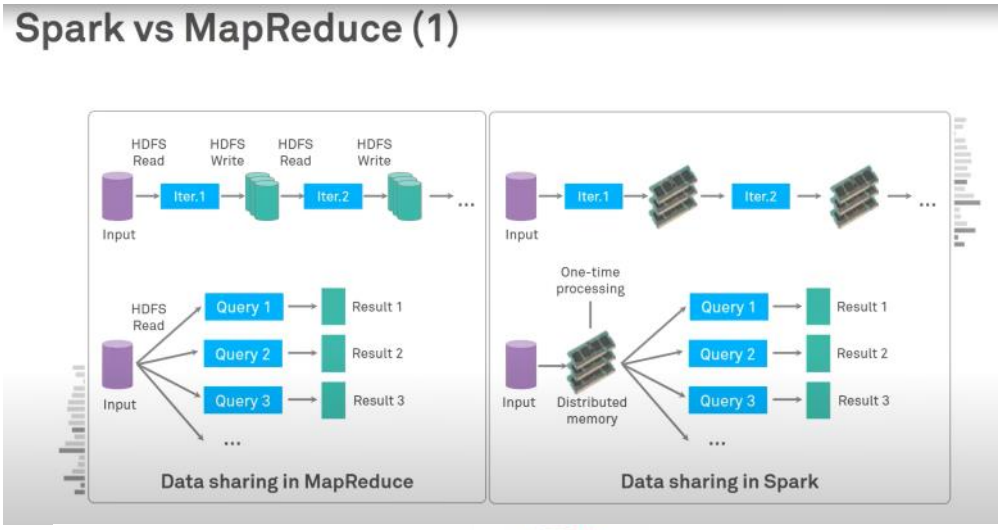
Apache Spark

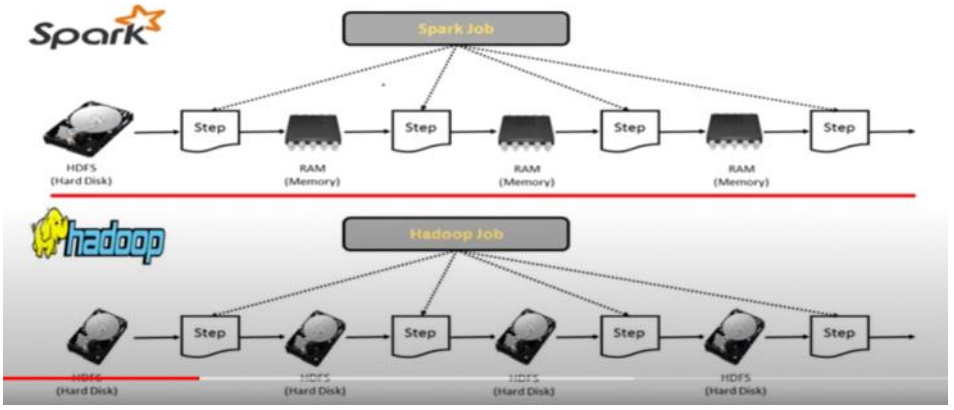
20 December 2022 10:25

Spark is an open-source computing framework engine managed by Apache, comparable to Hadoop MapReduce.

What makes Apache Spark fast?

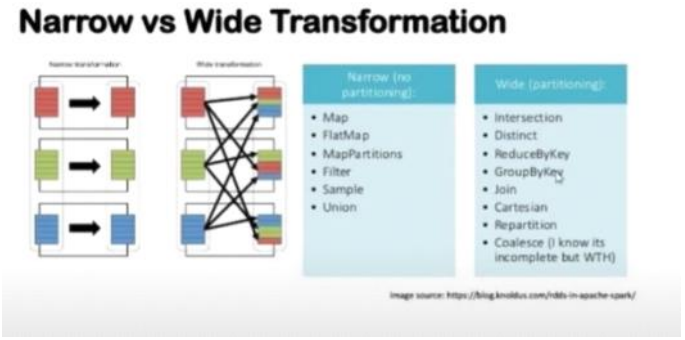
- **In-memory computation**
  - In MapReduce, intermediate results generated by Map tasks are written to local disk as MOF (MapOutFiles) waiting to be obtained by Reduce tasks
  - In Spark, the intermediate data of Spark is stored in distributed memory (data is only stored in disk as the final destination) by caching the partial results across its memory of distributed hardware
  - Spark ensures lower latency computations (reduces I/O from reading/writing data to and from RAM and Disk)



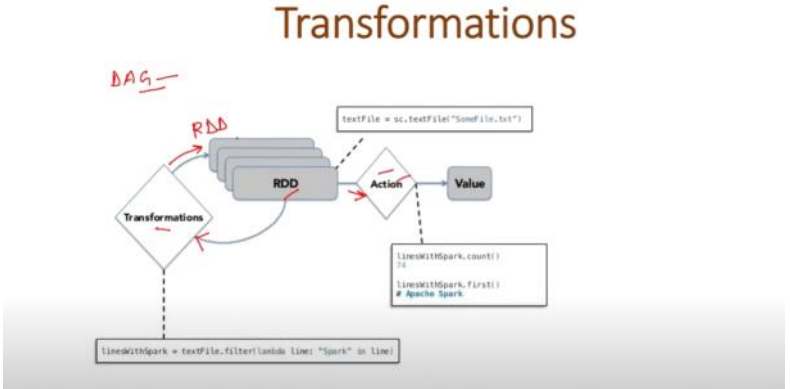


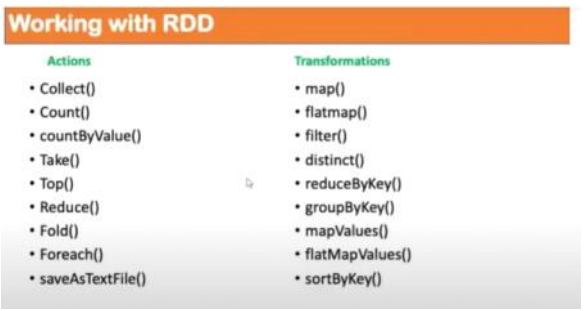
MAP REDUCE	SPARK
Computing Framework Engine, open source managed by Apache	Computing Framework Engine, open source managed by Apache
Yes , Map Reduce is Faster than traditional system but it does not leverage the memory of hadoop cluster to the maximum	spark has been proved to execute the batch processing jobs 10 to 100 times faster
Map Reduce is disk Oriented completely. Higher latency. No caching support.	Spark ensures lower latency computations by caching the partials results across its memory of distributed hardware. Stores data in memory
MapReduce is a cheaper option available while comparing it in terms of cost.	As spark requires a lot of RAM to run in-memory. Thus, increases the cluster, and also its cost.
Writing Map reduce pipelines is complex and lengthy as it is purely Java	Writing Spark code is always easy and we can write in 4 languages
Batch Processing	Batch/Iterative/ Real Time /Interactive Processing
Fault Tolerance and Highly Scalable and Cross platform	Fault Tolerance and Highly Scalable and Cross platform
Map Reduce has been tested on 15000 nodes	Spark has been tested on 8000 nodes
it has not inbuilt support to various things like SQL,ML,RT	it has in built support to various things like SQL,ML,RT
It is basic data processing engine.	It is data analytics engine. Hence, it is a choice for Data Scientist.
MapReduce runs very well on commodity hardware.	Spark needs mid to high-level hardware.

- Lazy Execution
  - Transformations:
    - Takes RDD as input and recreates a new RDD as an output. It does not perform the operation immediately, instead it checks the syntax and builds up a DAG to run only when an **action** is run
    - There are 2 types of transformation:
      - **Narrow** - no reshuffling needed (i.e. Filter)
      - **Wide** - reshuffling needed (i.e. Intersection)



  - Action
    - Returns a value as output





- Parallel Processing

What is RDD?

RDD stands for Resilient Distributed Dataset. A RDD in Spark is an **immutable** distributed collection of objects. Each RDD is split into multiple partitions, which may be computed on different nodes of the cluster.

RDD can be created in two ways:

- By loading an external dataset
- By transforming one RDD into another