Yarn - Resource manager in Hadoop

Mesos is an equivalent of yarn, made by Amplab.

Spark can be installed on top of Hadoop and it can make use of HDFS for storage.

MapReduce does Batch processing.

Apache Mahout(uses MapReduce) - ML library

ML is mostly dependent on iteration but with Hadoop we can’t do iteration.

Mapper ----> Reducer -----> Disk. Repeat the process multiple times

This process was very slow

Storm - streaming data analysis

Fling - real time processing

Spark can do the work of tools like pig, hive, mahout, storm, impala etc.,

Spark is a unified processing engine

Credit card Fraud detection should be done in real time. Whenever we swipe the card it has to be captured by the system and it has to process it. For this use case, we can use spark.

Spark does micro-batching. It is not 100% streaming but it is near real time streaming.

Like it collects data for every second but fling collects data for every card swipe(micro seconds). It is 100% streaming.

SparkSql is integrated with Hive. So Hive can be used as a storage and we can use SparkSql to query.

Graph type of data structure will be very suitable for airport data. Like flights between 2 places can be found very fastly. In SQL this might be slow. Graph can also be used in social media. Friend or not

Finding similar images with url can also be done with graph API. i.e unstructured data can also be handled efficiently by spark Graph API (GraphX)

Neo4j is a Nosql database. It stores data in graph format

Spark has no storage. It reads data from somewhere processes it and then stores the output back.

Spark can be run on yarn, standalone, Mesos and on kubernetes.

Zookeeper is responsible for high availability. It is integrated with spark

Zookeeper is a Hadoop ecosystem tool. It is used for communication between the machines.

There are 2 name nodes, Active name node and standby name node. Only the active node will be active and when it fails we go for standby node.

Zookeeper tells which node is active. If a node is down the other node tells that it is active. So thats how zookeeper keeps tabs.

Advantages of spark:

1. Ideal for iterative processes. It takes lots of work to schedule iteration in MapReduce.
2. It doesn’t require any external tool unlike Hadoop which has lots and lots of tools which is tough to install
3. Spark is Fast. 10 to 100 times faster than MapReduce.

Oozie is used by MapReduce to schedule iterations.

Spark does in memory processing. If enough RAM is present for the whole file. It reads the file into the RAM and then processes it, the final output alone is given to the disk. It doesn’t write intermediate cases into the disk.

If enough memory is not present then it reads data part by part.

Driver - Master of the program

Executor - Slave of the program

There are 2 modes to execute spark

1. Local - There will be a container inside that driver and executor will be present.
2. Cluster - In this case there will be a number of nodes(containers). We specify to yarn how many executors(containers) we need and what is the RAM and number of processors in each container. Additionally the driver will run on a separate machine. Initially the driver pushes the program to all executors and finally the output is given back to the driver. We can also direct the output to wherever we want(Hadoop, Cassandra).

Driver is also called App Master.

If the driver crashes ---> we would have set the restart timer in yarn which will restart the machine in another machine. Initially it will try to restart the same machine but if it doesnt work we go for another machine

Yarn(resource manager) will be in constant touch with the driver and It would have the ID of the machine and till where it has executed. So it will be easier for yarn to restart the code in another machine.