Hadoop is a Big data Frame work to solve problems related to Hadoop.

Hadoop 1.0 🡺 It consists of Map Reduce, HDFS

Hadoop 2.0 🡺 It consists of Map Reduce, YARN, HDFS

YARN:

It is Yet Another Resource Negotiator.

It helps to mainly responsible for resource management, it is just like an operating system over multiple machines (cluster).

That means if some node in the cluster needs resources it will connect YARN for it.

As Hadoop 2.0 consists of HDFS, Map Reduce, YARN

HDFS for Distributed Storage

Map Reduce for Distributed Processing

YARN for Resource Management.

There are some eco system technologies around this Hadoop such as Hive, HBase, Sqoop, OOZIE, PIG, Spark together it is called Hadoop ecosystem.

Code written in Map reduce is Java, Hive is like a wrapper over Map Reduce like if we don’t know the Java using Hive, we can write queries in SQL in which it can hard code those queries into Java to Map Reduce. The Query language which we used in Hive is not SQL it is HQL (Hive Query Language).

Sqoop: Sqoop is tool used to transfer Data/Data Ingestion from Traditional Database to Hadoop using Sqoop Commands. It is a command-line interface application for transferring data between relational databases and Hadoop. Sqoop is also a Map Reduce Job only. Only Mappers do this job in real time.

PIG: 1. It helps to clean data 2. It helps to convert unstructured data to Structured data. Now a days PIG is not used in Industry as these things can be done by using Apache spark.

HBase: It is a column Oriented NoSQL database that runs on top of HDFS.

OOZIE: It is a Workflow Scheduler system to manage Apache Hadoop Jobs.

SPARK: A Distributed general purpose in-memory compute engine.

So as we seen that in Hadoop cluster there are mainly three components

1. HDFS is like Storage Unit, MapReduce is like compute Engine, Yarn is like Resource manager.

That means we can say HDFS🡺HDFS | MapReduce🡺SPARK | YARN🡺 YARN

From this we can say owe can’t say Hadoop is same as spark but we can say MapReduce is same as SPARK. SPARK is much more better than MapReduce.

So current industry uses HDFS|SPARK|YARN

SPARK is like plug and play thing but it need additionally two things Storage system and Resource Manager

SPARK as Compute Engine

Plug it with any storage system (Local Storage/HDFS/Amazon S3)

Plug it with any Resource Manager (YARN/MESOS/Kubernetes)

And the options among them can be decided by the Hadoop admin.

SPARK is written in SCALA however, spark officially support Java, Scala Python and R.