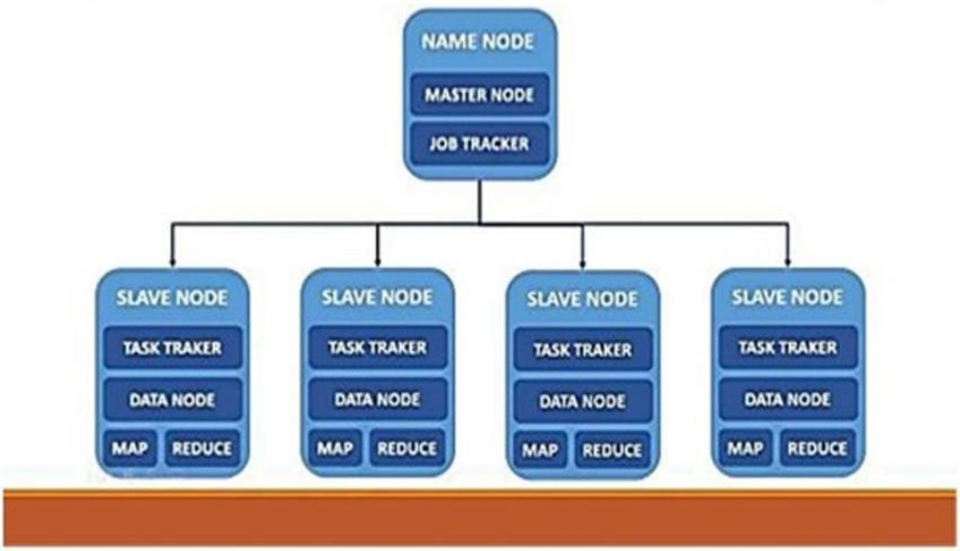
- When we have to handle large amounts of data, pandas is not enough. then we have to use something else, like PySpark

- 5 V's of Big Data
    - Volume - The size of the data is high
    - Velocity - The speed at which the data is generated. If this speed is high then its Big Data
    - Variety - The different types of data
    - Veracity - this is classify of the trustworthiness of the data in terms of accuracy
    - Value - we are doing to abstract some value from the data.

- Hadoop Cluster

1. HDFS - This is not a database. This is the File System.
2. MapReduce - This is built in two ways. Map and Reduce.
    a. Map : Operations which are performed in the parallel on the small portions of the dataset.
    b. Reduce : Combine all the results which are extracted from the Map. this will be given to the user.
3. YARN - resource manager. Hadoop has two things.
    Resource Manager
    Node Manager

- Hadoop works on Master Slave architecture. This is the Architecture,



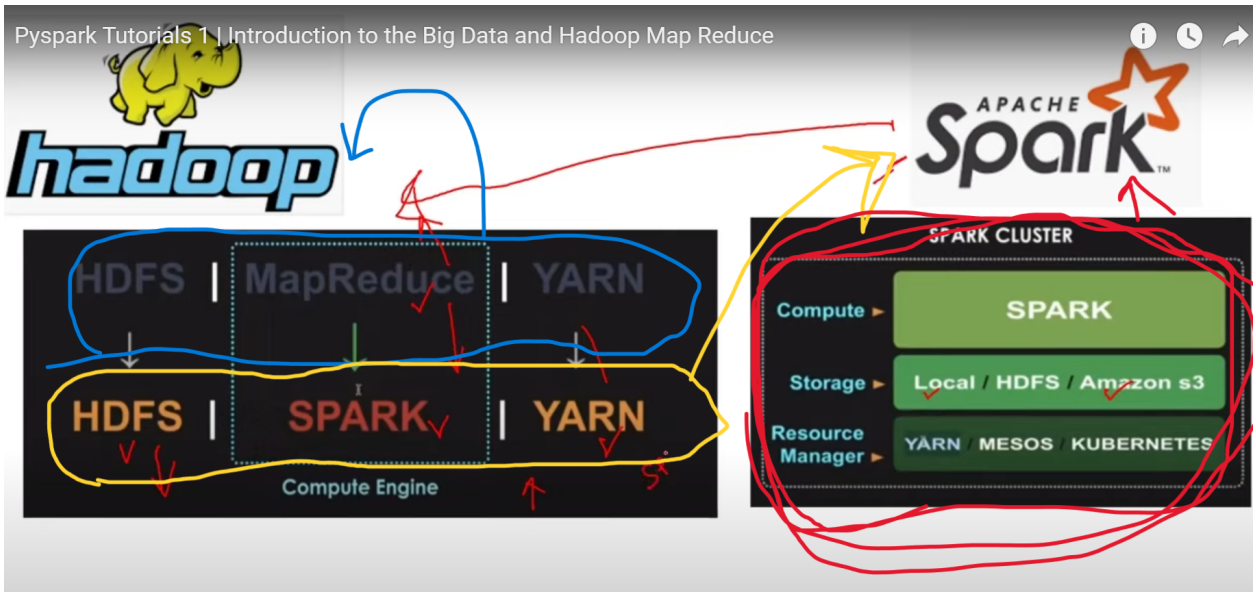
Resource Manager -> NAME NODE Node Manager -> SLAVE NODE

- In apache spark, eliminate the MapReduce component. For that in Apache Spark use Spark. So components are HDFS, Spark and YARN.

- Typical process,
  Source → Ingest → Process → Store → Serve Data

  Source : Define source of the data
  Ingest : We have to integrate the data and store in one place
  Process : Process the operations on that data. In here we do MapReduce or Spark
  Store : store the data in some format
  Serve Data : serve the data to user or some application

- Difference between Hadoop and Apache Spark



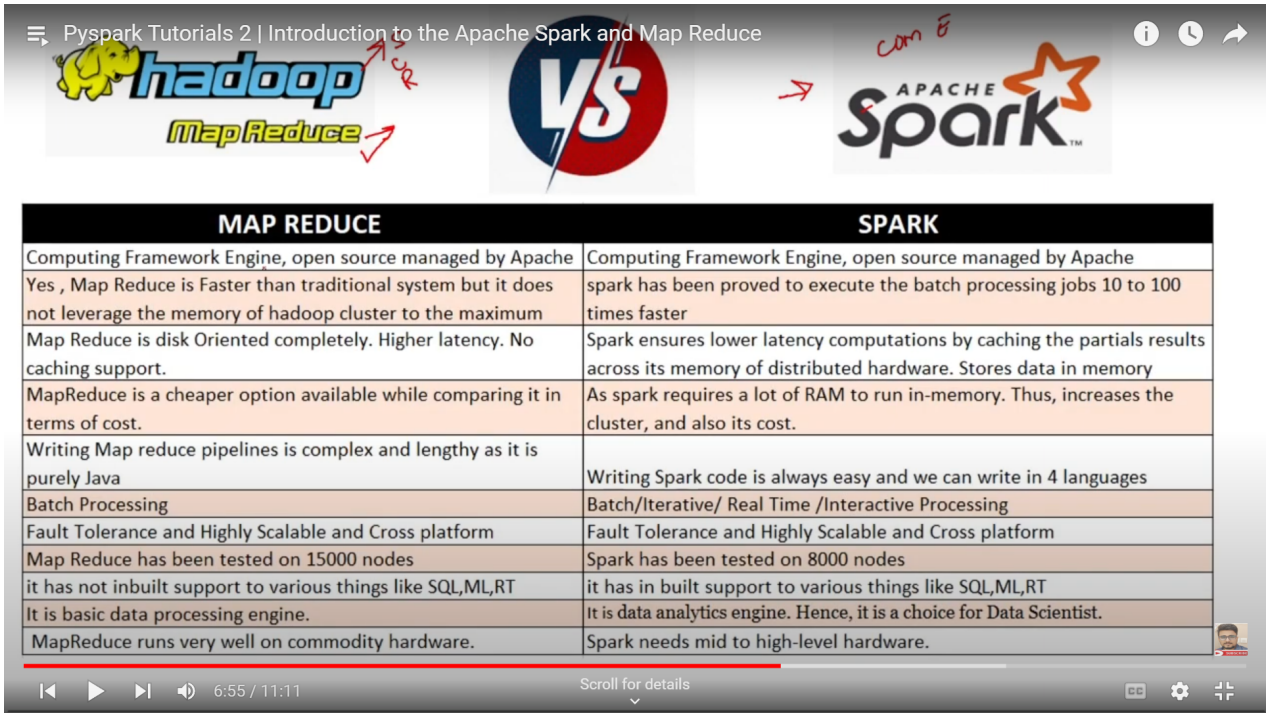
- Spark Cluster can be in two modes,
  1. Distributed environment mode - If we use HDFS and YARN
  2. Standalone environment mode - if we are not use HDFS and YARN

- Hadoop is only Java based. Only for batch processing. We can not schedule anything. If we want to use machine learning libraries we have to specify what library we want to use. There are a lot of libraries. And a lot of SQL libraries as well.

- Spark has everything. Inbuilt machine learning libraries, Inbuilt SQL, Inbuilt Real Time data processing

- What is Spark?
    1. Fast, real-time processing framework, written in Scala
    2. To convert Java language to Python we use the Py4J library we used.
    3. In-memory computations, Lazy execution and parallel processing
    4. Apache Hadoop MapReduce is performing batch processing only. But Spark does Batch processing and Real time processing as well.
    5. It leverages Apache Hadoop for both storage and processing. It uses HDFS for storage and it can run Spark applications on YARN as well.
    6. Spark can load data directly from disk, memory and other data storage technologies such as Amazon S3, Hadoop Distributed File System, HBase, Cassandra.

- We can not compare Apache Hadoop and Apache Spark. Because Hadoop is a complete package. It has HDFS, MapReduce and YARN. Spark is a compute engine. So we can compare MapReduce and Spark.



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.