Getting Started with Big Data and Frameworks

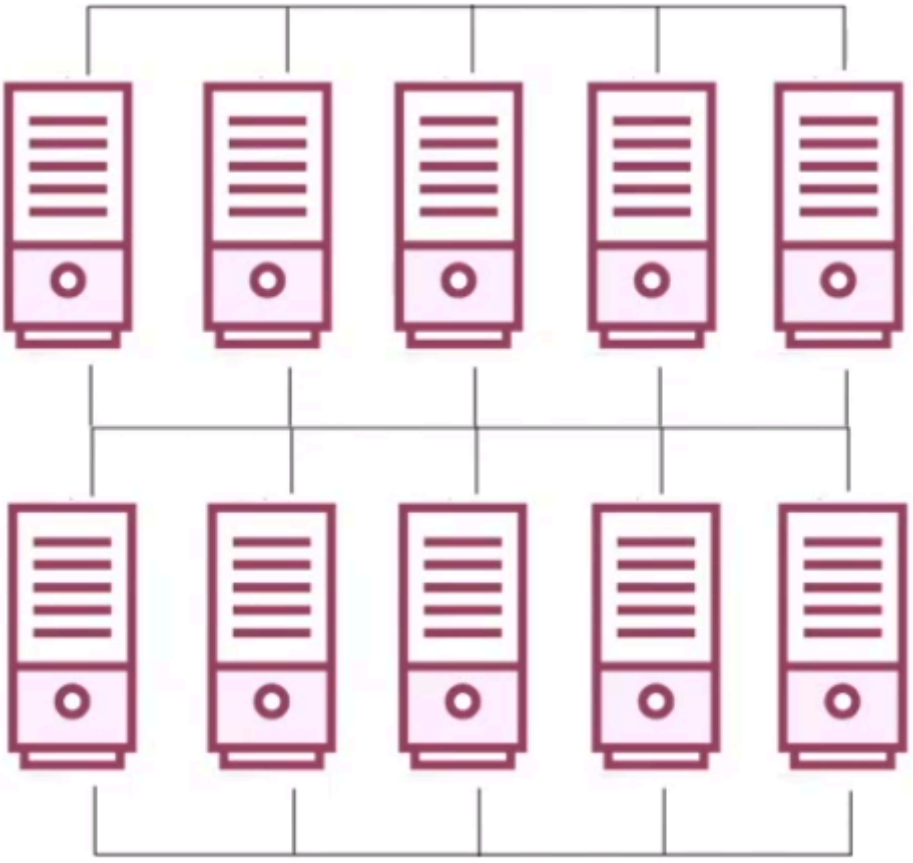
> Big Data refers to the data which is **large, fast and complex type** of structured, semi-structured and unstructured data generated from a variety of different sources, which becomes difficult to store and process using a **traditional processing system (RDBMS)**

Challenges of Big Data

1. Storage : Distributed Storage System
2. Processing : MPP (Massive Parallel Processing)

Distributed Systems

> A DS is a collection of autonomous system that are physically separated but are linked together



Hadoop

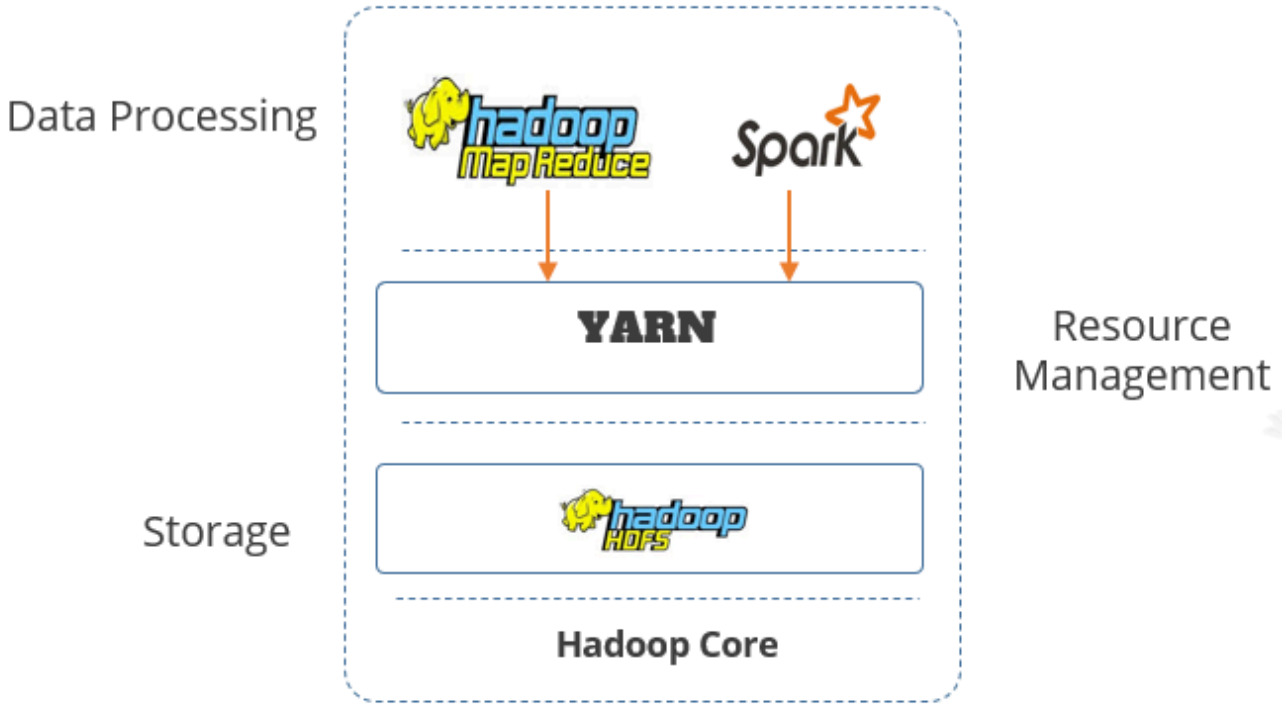
> Apache Hadoop is a software framework that allows us to **store and process large datasets in parallel and**

> **distributed system.**

Components of Hadoop

Hadoop consists of 3 main components

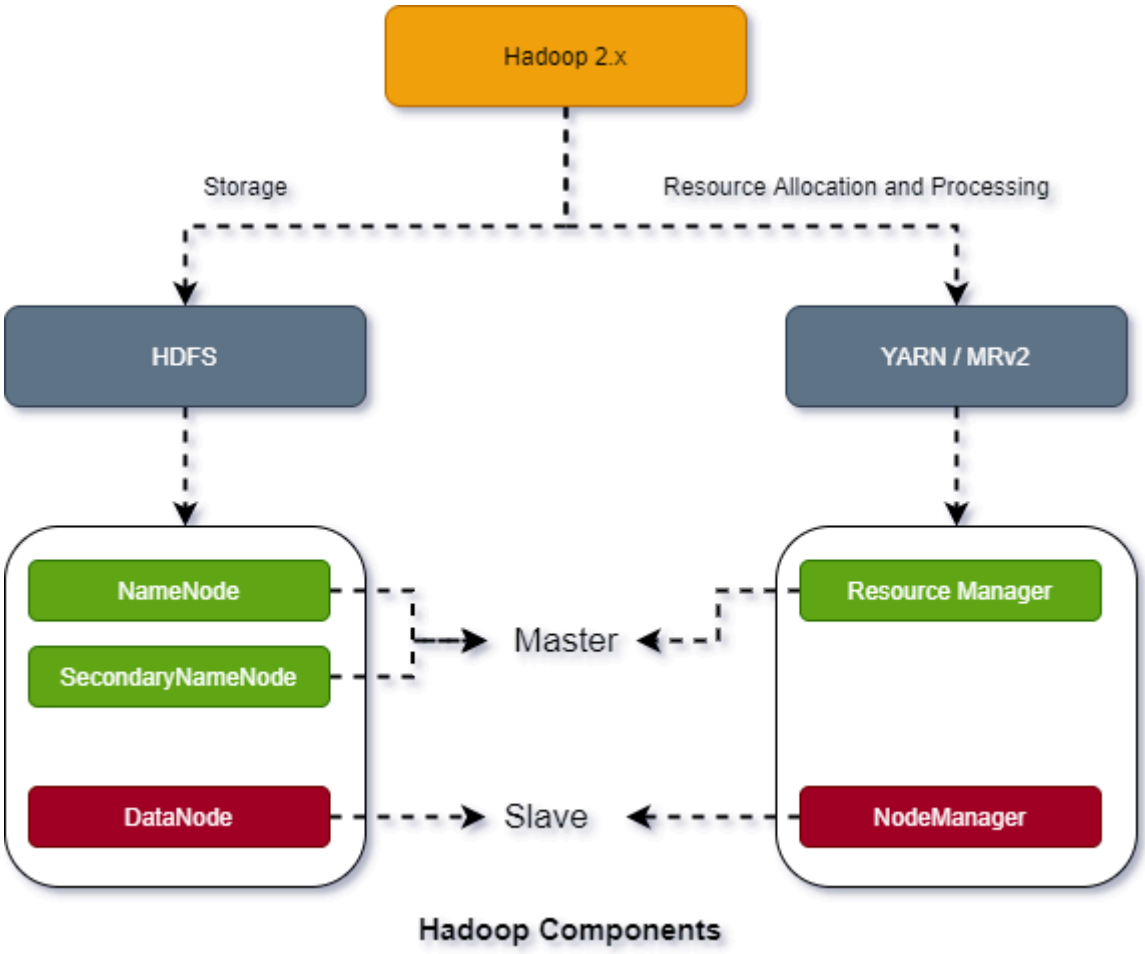
1. Storage Layer : **HDFS (Hadoop Distributed FS)**
2. Resource Management Layer : YARN (**Yet Another Resource Negotiator**)
3. Data Processing Layer : **MapReduce**



- Hadoop MapReduce is **deprecated**.

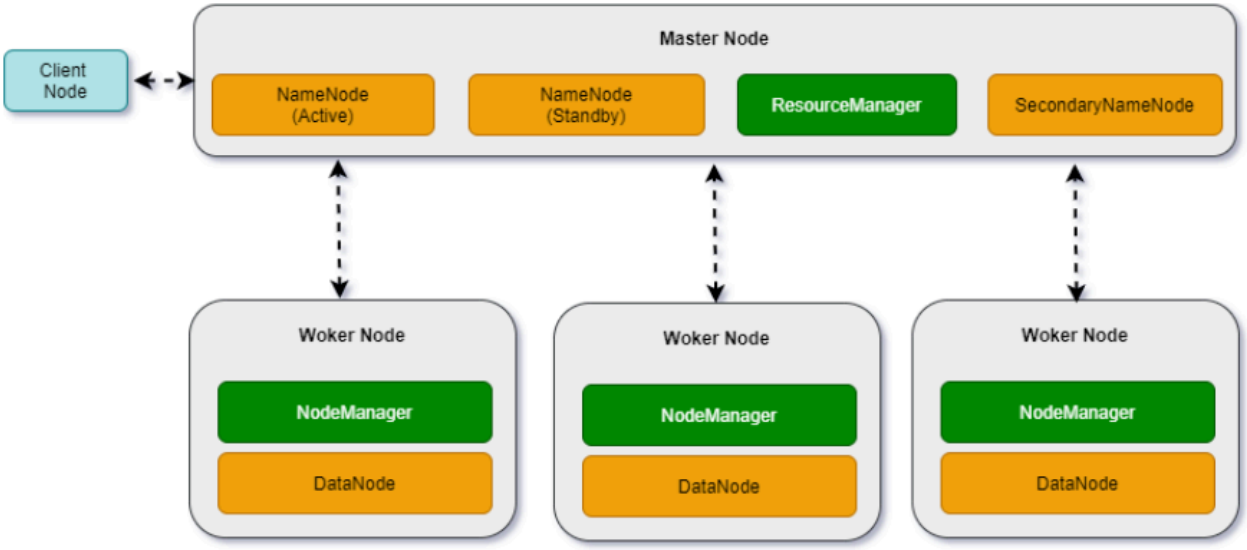
Hadoop Daemon Services

1. NameNode
2. SecondaryNameNode
3. DataNode
4. ResourceManager
5. NodeManager



Hadoop Components

Master and Slave Architecture

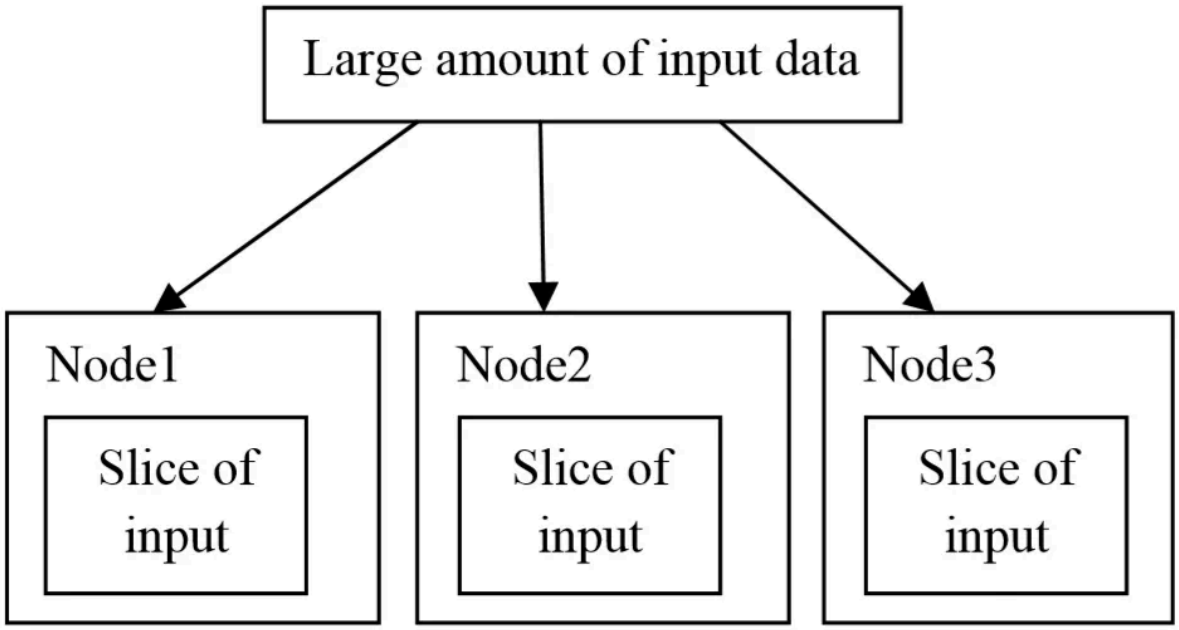


HDFS and Architecture

> HDFS is a **distributed** and **scalable** file system designed for storing very large files.

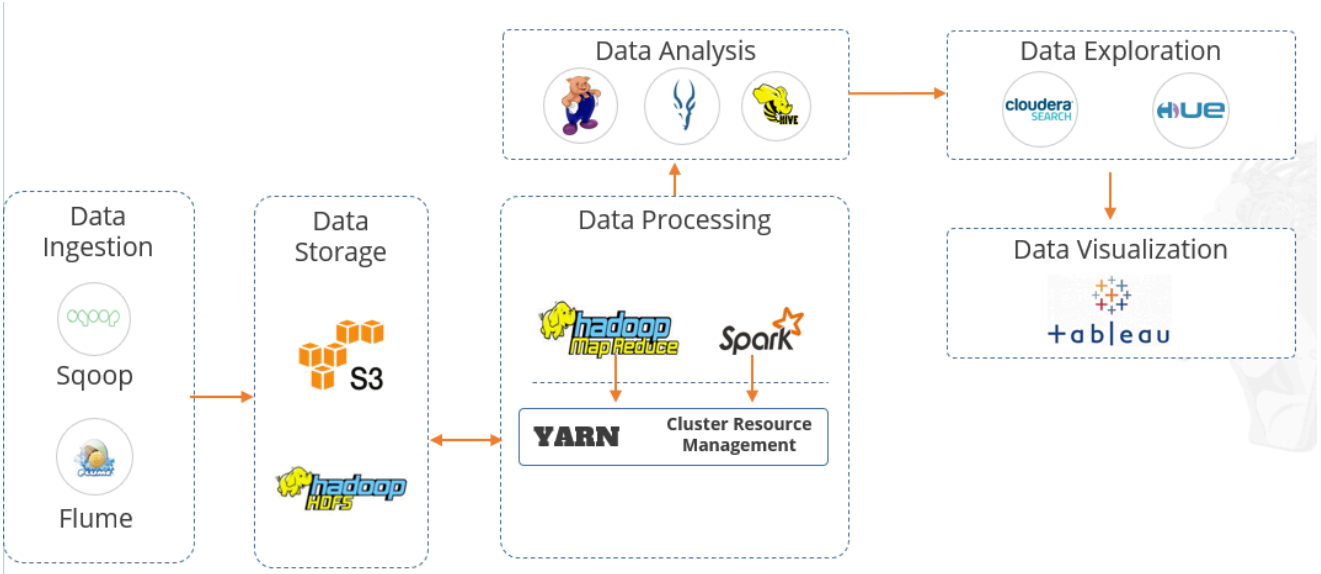
Distributed

- The files are stored across multiple machines, HDFS splits the file into smaller pieces (Blocks), distributes these blocks into multiple machines.



- The block size in 128 MB (configurable)
- 300 MB --> 128 MB 128 MB 44 MB.
- To handle fault tolerance , Hadoop replicates each block thrice

Big Data / Hadoop Ecosystem



Map Reduce

> MR is a programming paradigm to process the data in parallel in multiple machines.

- MR comes with lot of IO operations

YARN

> Yet Another Resource Negotiator

YARN consists of 3 main components

1. Resource Manager
2. Node Manager
3. Application Master

Important Points

- When you work with Hadoop, provides two Web UI
    - NameNode UI : Browse HDFS
    - Resource Manager UI : Track the job exeuction