Introduction to Big Data and Hadoop 2.x

> 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**

Traditional Processing Systems (RDBMS)

1. Are designed to store only structure data
2. Are vertically scalable
3. RDBMS follow schema on write (Not designed for high velocity data)

Challenges of Big Data

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

Frameworks

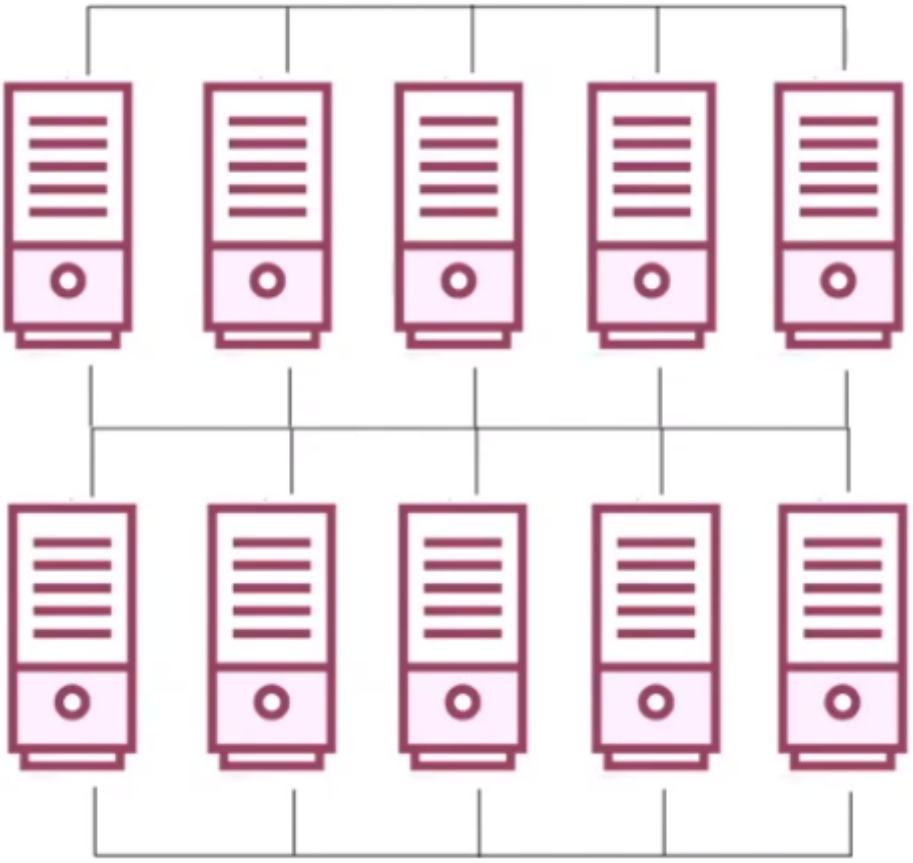
- Hadoop 2.x
- Apache Spark
- Kafka
    - Producer API
    - Consumer API
    - Kafka Streams (Real-time Streaming)
    - Connect
- Azure
- Aws

What is Hadoop

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

Important Point

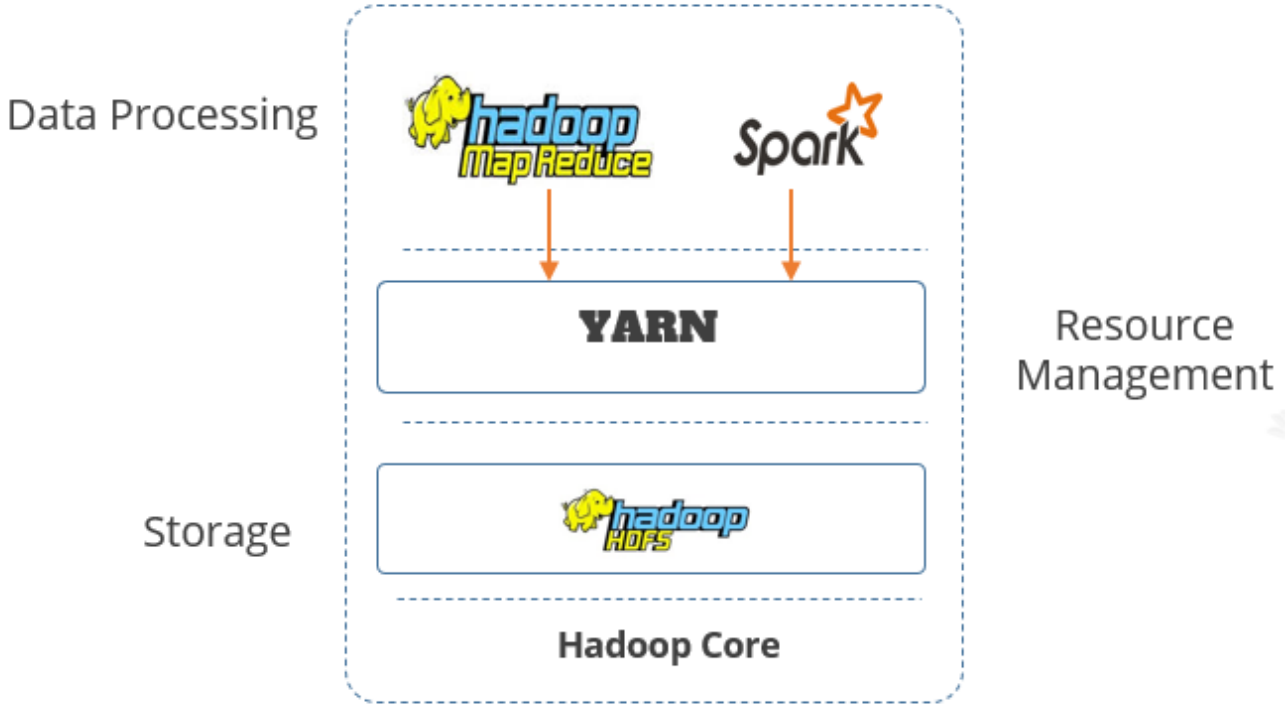
- Hadoop is written in Java



Components of Hadoop

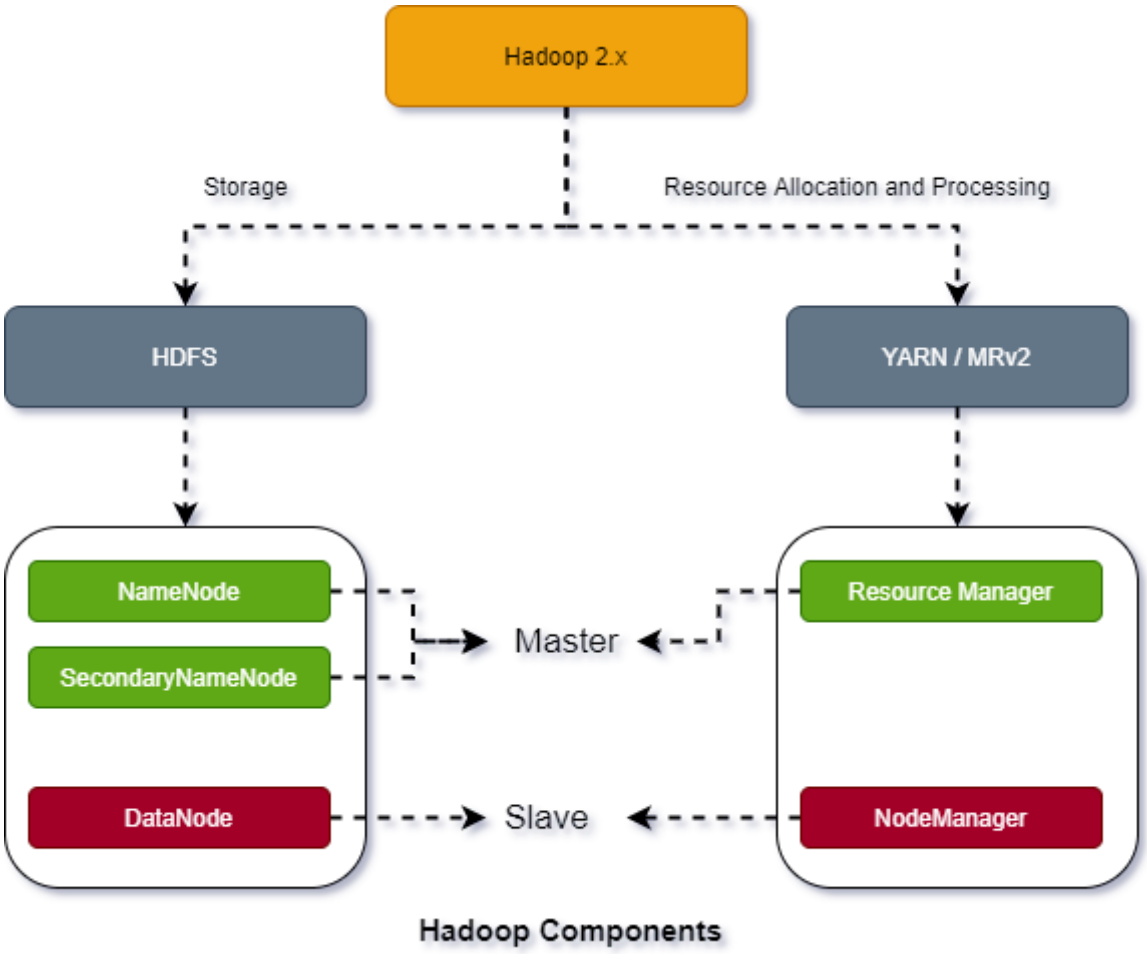
Hadoop consists of three main components

1. Storage Layer : **HDFS**
2. Resource Management Layer : **YARN**
3. Data processing Layer : **MapReduce**



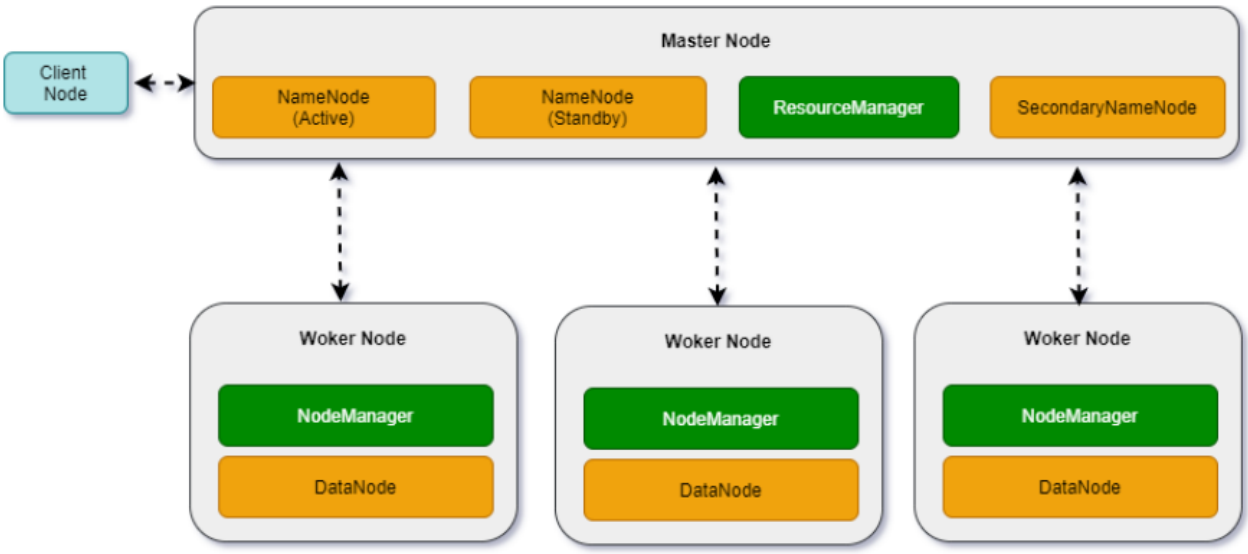
Hadoop Daemon Services

- In Hadoop, a daemon service refers to a long running background process or service that performs some tasks
- Hadoop provides 5 daemon services
  - NameNode
  - SecondaryName
  - DataNode
  - ResourceManager
  - NodeManager



Hadoop Components

Hadoop Master and Slave Architecture



HDFS Architecture

> HDFS (Hadoop Distributed File System) is a distributed and scalable file system designed for storing very large.

- FS is a software which breaks the file into smaller chunks / blocks