Hadoop Ecosystem Fundamentals of Distributed Systems

Author : Amit Singh Chowdhery

Agenda

- HDFS
- YARN

What is distributed Systems?

At a basic level, a distributed system is a **collection of computers** that work together to form a single computer for the end-user. All these distributed machines have one shared state and operate concurrently.

They are able to **fail independently** without damaging the whole system, much like microservices. These interdependent, autonomous computers are linked by a network to share information, communicate, and exchange information easily.

Why do we need distributed systems?

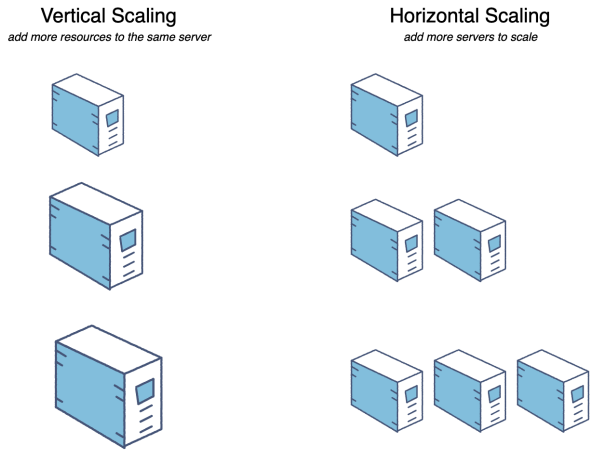
1. **Scalability**: Distributed systems allow us to scale up the resources of a system by adding more computers to the network. This allows us to handle more users and more data without overloading a single computer.

2. **Fault tolerance**: Distributed systems can provide high availability and fault tolerance by replicating data and services across multiple computers. If one computer fails, the system can continue to function without interruption.

3. **Performance**: Distributed systems can improve performance by distributing computation across multiple computers. This allows us to perform complex computations more quickly and handle larger workloads.

4. **Geographical distribution**: Distributed systems can allow users to access services from anywhere in the world. By placing computers in different locations, we can reduce network latency and provide better service to users in different parts of the world.

5. **Cost-effectiveness**: Distributed systems can be more cost-effective than centralized systems because they can use commodity hardware and can scale up or down as needed



Scalability is the biggest benefit of distributed systems. Horizontal scaling means adding more servers into your pool of resources. Vertical scaling means scaling by adding more power (CPU, RAM, Storage, etc.) to your existing servers.

Where/ When we can use distributed systems?

Distributed systems are used in all kinds of things, everything from electronic banking systems to sensor networks to multiplayer online games. Many organizations utilize distributed systems to power content delivery network services.

In the **healthcare industry**, distributed systems are being used for storing and accessing telemedicine. In **finance and commerce**, many online shopping sites use distributed systems for online payments or information dissemination systems in financial trading.

Distributed systems are also used for **transport** in technologies like GPS, route finding systems, and traffic management systems. **Cellular networks** are also examples of distributed network systems due to their base station.

Google utilizes a complex, sophisticated distributed system infrastructure for its search capabilities. Some say it is the most complex distributed system out there currently.

What is Hadoop Distributed File System (HDFS)?

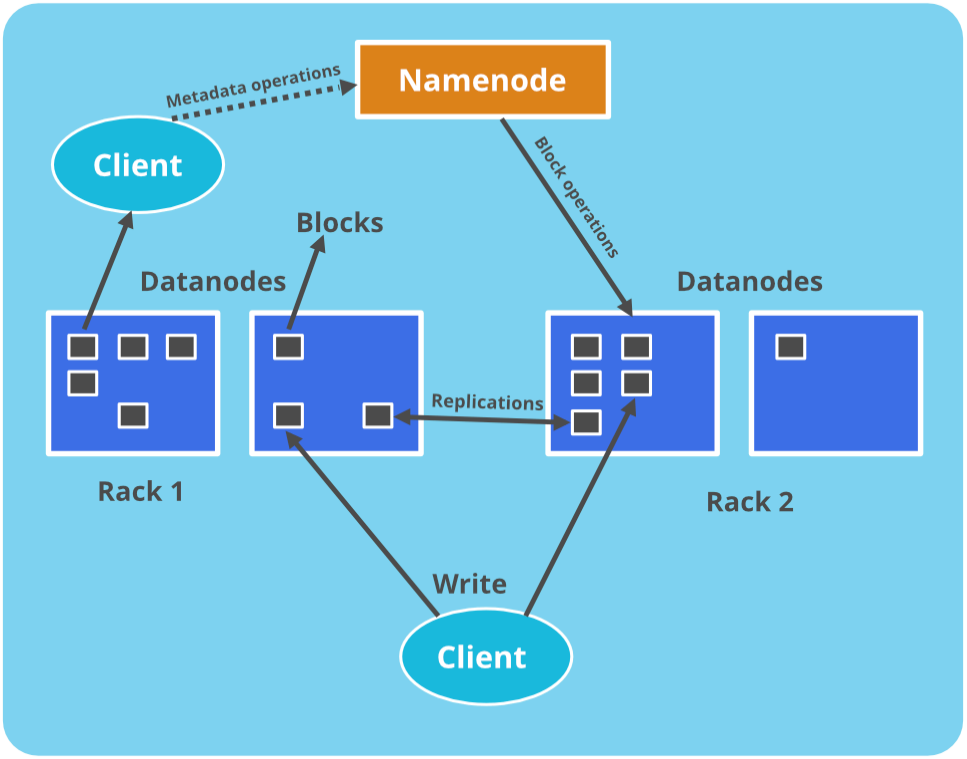
The **Hadoop Distributed File System (HDFS) is a distributed file system** designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant.

HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware, making it suitable for storing and processing large datasets. It achieves scalability by distributing data across multiple nodes in the cluster, and fault tolerance by replicating data across multiple nodes.

In HDFS, files are split into blocks, which are then distributed across the nodes in the cluster. Each block is replicated multiple times to ensure that data is not lost in case of node failure. The replication factor can be configured to ensure that the system remains highly available and fault-tolerant.

Architecture of HDFS:

Hadoop 1.x Architecture



The Hadoop Distributed File System (HDFS) architecture consists of two key components: the **NameNod**e and the **DataNode**. The NameNode manages the file system metadata, while the DataNode stores the actual data.

HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes.

The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes.

The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

The existence of a **single NameNod**e in a cluster greatly simplifies the architecture of the system. The NameNode is the arbitrator and repository for all HDFS metadata.

Secondary NameNode:
When NameNode runs out of disk space, a secondary NameNode is activated to perform a checkpoint.

DataNode
Every slave machine that contains data organizes a DataNode.DataNodes do the following:

- DataNodes store every data.
- It handles all of the requested operations on files, such as reading file content and creating new data

<ask students> what if in HDFS the master goes down? in that case how we are doing or how we are achieving high availability plus what is the algorithm that how these replication are being kept is it just a random order is there something behind it will focus in detail on that topic

To overcome this issue researchers come up with the approach that we have to have a high availability model by **creating 2 name nodes**. One will be in **Active mode** and one will be in **Standby/Passive mode**.

Problem: As you know in Hadoop 1.x architecture Name Node was a single point of failure, which means if your Name Node daemon is down somehow, you don't have access to your Hadoop Cluster. How to deal with this problem?

Solution: Hadoop 2.x is featured with Name Node HA which is referred as HDFS High Availability (HA).

- Hadoop 2.x supports two Name Nodes at a time one node is active and another is standby node
- Active Name Node handles the client operations in the cluster
- StandBy Name Node manages metadata same as Secondary Name Node in Hadoop 1.x
- When Active Name Node is down, Standby Name Node takes over and will handle the client operations then after
- HDFS HA can be configured by two ways
    a. Using Shared NFS Directory
    b. Using Quorum Journal Manager

Now, how the synchronization between Active Namenode and Standby Namenode is going to happen. Who will tell the Standby namenode that now you are an active node and who will tell the data node to send the message to the latest active namenode?

There free tool available which can manage all that synchronization:

1. **Zookeeper:**

ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services. All of these kinds of services are used in some form or another by distributed applications.

It'll also help to elect the leader so inside the zookeeper there is an algorithm which will elect the master node. You have to provide the IP address of all the namenodes present in your cluster to the zookeeper.

How does the Zookeeper elect the master and maintain high availability?

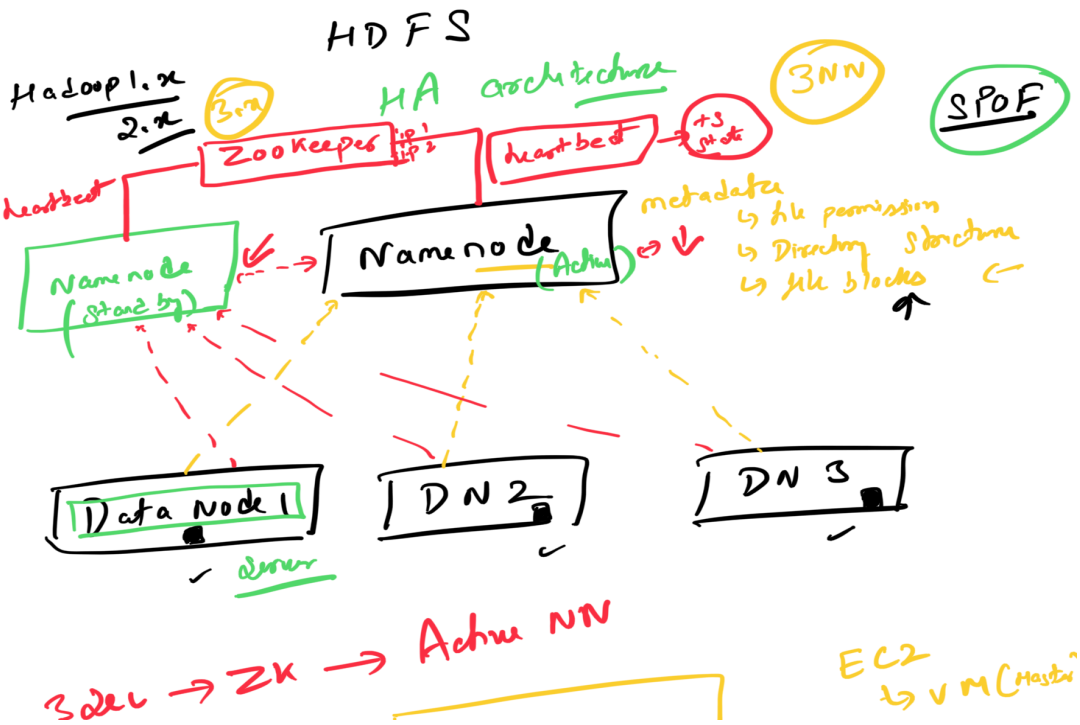
It does this through a simple feedback loop. The Zookeeper, through its Fail-over Controller, monitors the Leader and Follower/Stand-by nodes.

It receives a heartbeat or instant notification of the current health/status of each node. The moment a leader fails, one of the stand-by nodes is elected as the new leader by Zookeeper almost immediately. The election is completed and the new leader sends a message to the data nodes to report to the new master node.

Heartbeat Mechanism:

By default every 3 seconds, zookeeper will send a message to the active node and will do it 10 times by default. If the current Active node is not responding then Zookeeper is going to mark that node as a down.

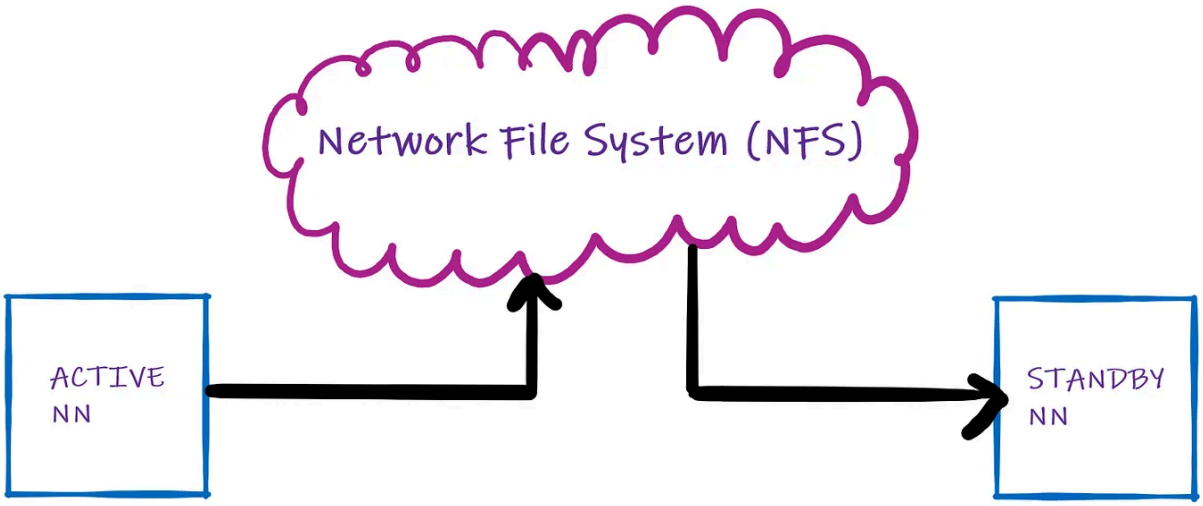
It'll then send the message to the standby node that you are now the Active node.



How does NameNode Metadata Sync?

As you can imagine or see in the image ‘HDFS High Availability Architecture’, the name node metadata is a single point of failure, hence this metadata is replicated to introduce redundancy and enable high-availability (HA).

1) **Shared Storage**

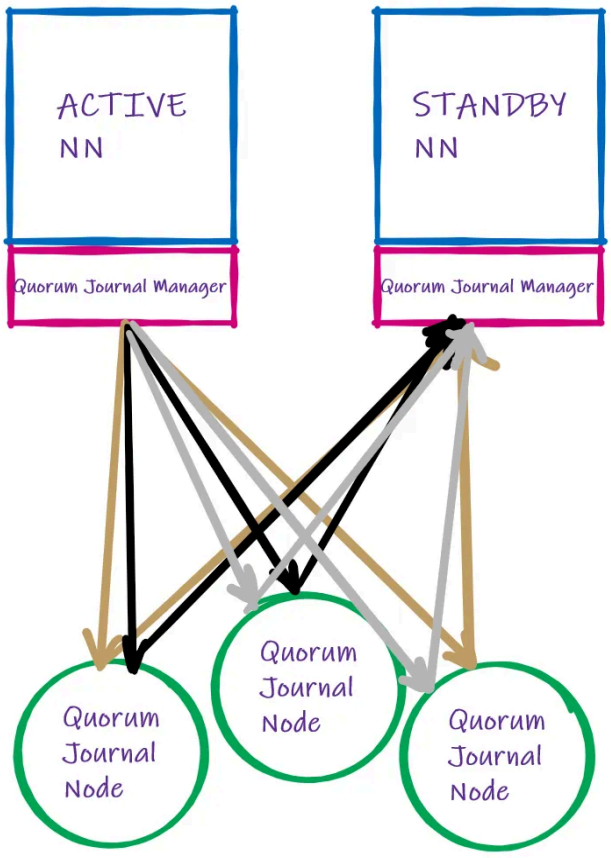


We now know that there exists an Active Name Node and a Standby Name Node. Any change in the *active* is synced in real-time to the shared folder/storage i.e. network file system (NFS).

This NFS is accessible to the *standby,* which downloads all of the relevant incremental information in real-time to maintain the sync between the Namenodes.

Thus, in the event of a failure of the *active*, the *standby* name node already has all the relevant information to continue “business as usual” post fail-over. This is not used in the Production environment.

2) **Quorum Journal Node (QJN)**



“Quorum” means minimum required to facilitate an event.

Here, we use this concept to determine the minimum number of journal nodes aka quorum that is needed to establish a majority and maintain metadata sync.

The image shows three (always odd) journal nodes (process threads not physical nodes) that help to establish metadata sync. When an Active NN receives a change, it pushes it to the majority of the QJ Nodes (follow a single color). The Standby NN, in real-time, requests the majority number of QJ Nodes for the required metadata to establish the sync.

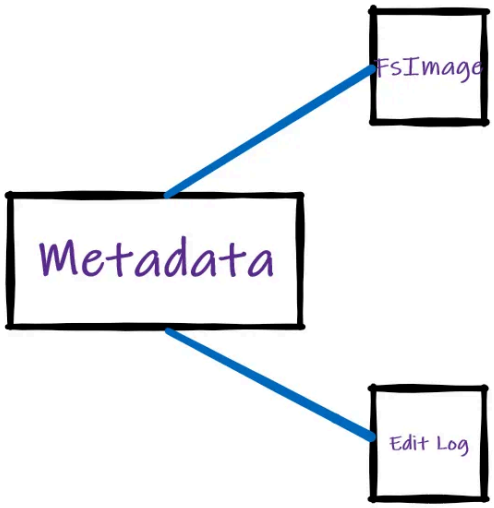
The minimum number for QJN to function is 3 and the quorum/majority is determined by the following formula:

Q = (N+1)/2

where N = total number of Journal Nodes
For example, if we have N=5, the quorum/majority would be established by (5+1)/2 i.e. 3. The metadata change would be written to 3 journal nodes.

The QJN is the preferred Production method of metadata sync as it is also “highly available”.

What is NameNode Metadata?



The name node (NN) metadata consists of two persistent files, namely, **FsImage — namespace** and **Edit logs — transaction logs** (insert, append)

1) **Namespace & FsImage**

In every file system, there is a path to the required files — On Windows: *C:\Users\username\learning\BigData\namenode.txt* and on Unix: */usr/username/learning/BigData/namenode.txt*

HDFS follows the Unix way of namespace. This namespace is stored as part of the FsImage. Every detail of the file i.e. who, what, when, etc. is also stored in the FsImage snapshot. The FsImage is stored on the disk for consistency, durability and security.

2) **Edit logs**

Any real-time changes to all files are logged in what is known as “Edit logs”. These are recorded in-memory (RAM) and contain every little detail of the change and the respective file/block.

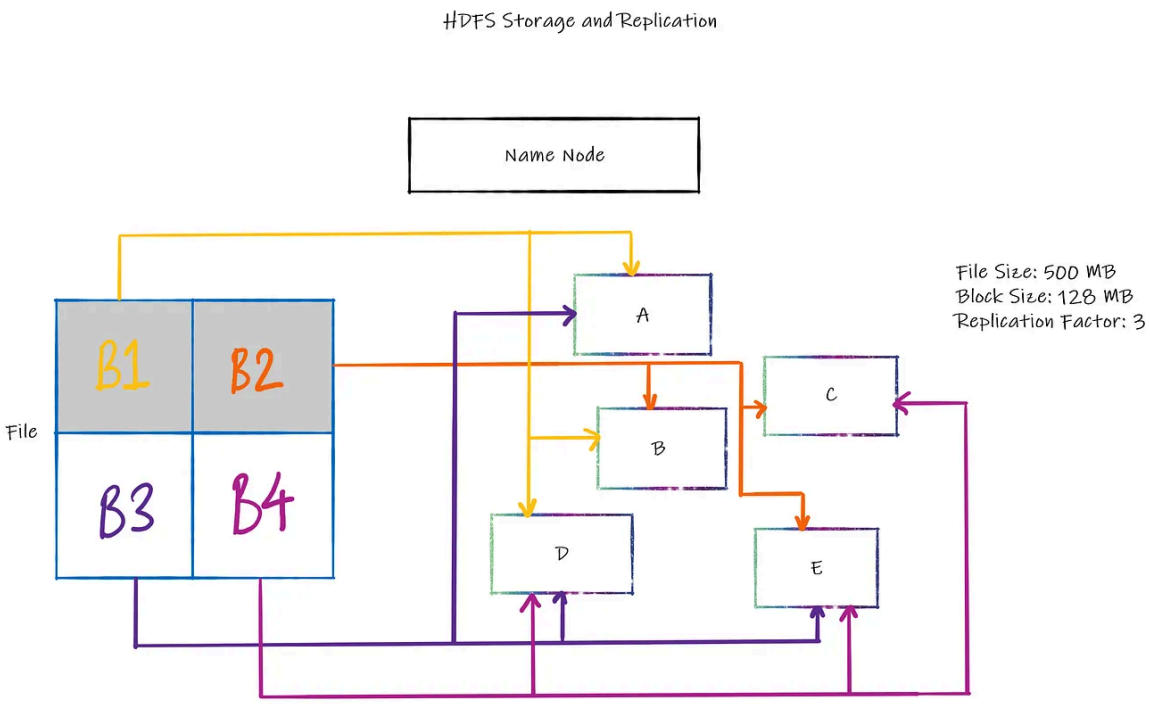
On HDFS startup, the metadata is read from the FsImage and the changes are written to Edit Logs. Once the data is recorded for the day in Edit Logs, it is flushed down onto the FsImage. This is how the two work in tandem.

Replication Factor:

HDFS is a fault-tolerant and resilient system,In order to achieve this, data stored in HDFS is automatically replicated across different nodes.

How many copies are made? This depends on the “replication factor”. By default, it is set to 3 i.e. 1 original and 2 copies.

Storage and Replication Architecture:

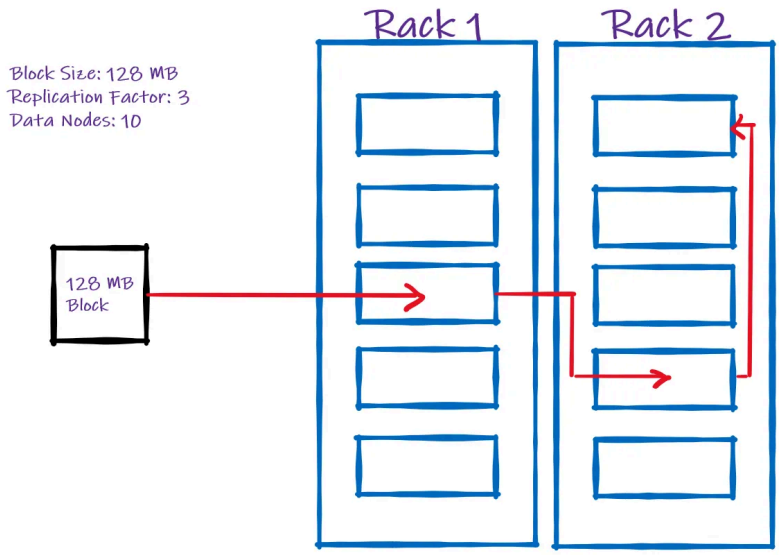


With HDFS’ default block size of 128 MB, this file is broken into 4 blocks B1 — B4. Please note that A — E are our Data Nodes. With HDFS’ default replication factor of 3, the blocks are replicated across our 5 node cluster. Block B1 (in yellow) is replicated across Nodes A, B and D and so on and so forth (follow the coloured lines).

Here, the Name Node maintains the metadata, i.e. data about data. Which replica of which block of which file is stored in which node is maintained in NN — replica 2 of block B1 of file xyz.csv is stored in node B.

So a file of size 500 MB requires a total storage capacity in HDFS of 1500 MB due to its replication. This is abstracted from the end users’ perspective and the user can only see 1 file of size 500 MB stored within HDFS.

The Block Replication Algorithm



The algorithm starts by searching for the topology.map file under HDFS' default configuration folder. This .map file contains metadata information on all the available racks and nodes it contains. In our example case in the image above, we have 2 racks and 10 data nodes.

Once the file is divided into blocks, the first copy of the first block is inserted into the rack and data node which is nearest to the client

The copy of this first block is created and moved onto the next available rack i.e. Rack 2 and stored in any available data node.

Another copy is created here and moved onto the next available rack

this is how Replication algorithm works

Problem Statement:

During the time of Wannacry ransomware attack, Flipkart was running a cluster of 2300 nodes which are on windows. They quickly realize that there is a high chance that they can also face the attack and they can lose all their important data. Microsoft sent them a security patch to install. Data engineers need to figure out when they are facing minimum requests so that they can roll out the patch and Flipkart don't even lose customer interest as well as become more secure.

Solution: suppose we have a jar file which will be working on all the servers parallely and output of this will be what time the requests are minimum.

Now how to execute this code. How hadoop is going to provide the resources across the cluster, we'll focus on that. Here YARN comes into the picture.

YARN (Yet Another Resource Negotiator):

Hadoop 1.x consist of only 2 components

- **Mapreduce**: which is used for processing and resource management
- **HDFS**: for storage

Mapreduce started giving issues:

1) It was only used for doing batch processing. But as technology progressed everybody started progressing on real-time data rather than batch processing but Mapreduce was not able to process real-time data.

They solved this issue in Hadoop 2.x by introducing YARN

Hadoop developed tool which can perform resource management separately from Mapreduce so Hadoop 2.x have 3 components:

1) Mapreduce
2) YARN
3) HDFS

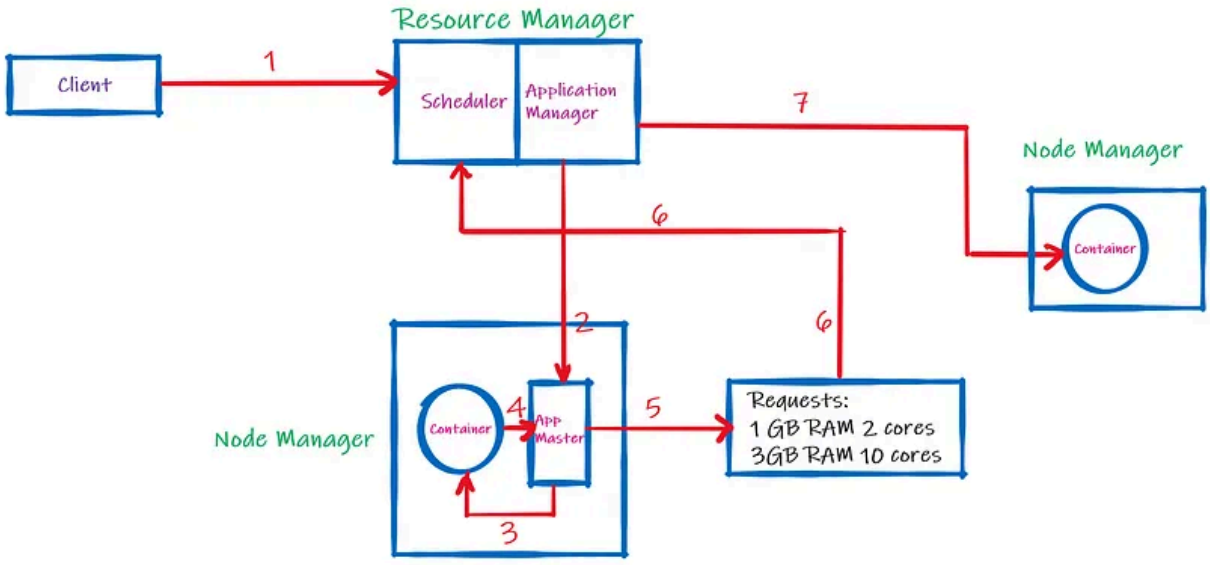
What Exactly is a YARN?

YARN or "Yet Another Resource Negotiator" does exactly as its name says, it negotiates for resources to run a job.

YARN, just like any other Hadoop application, follows a "**Master-Slave**" architecture, wherein the **Resource Manager is the master** and the **Node Manager is the slave**.

The master allocates jobs and resources to the slave and monitors the cycle as a whole. The slave receives the job and requests (additional) resources to complete the job and actually undertakes the execution of the job.

Working of YARN and it's Architecture:



1. The Client sends a job (jar file) to the Resource Manager (RM).

2. The RM contains two parts, namely, **Scheduler** and **Application Manager** (AM). The Scheduler receives the job request and requests the AM to search for available Node Managers (NM). The selected NM spawns the Application Master (App Master).

   Please note, the RM Scheduler only schedules the job. It cannot monitor or restart a failed job. The AM monitors the end-to-end life cycle of the job and can reallocate resources if a NM fails. It can also restart a failed job in App Master.

3. The App Master checks the resources provided for the job in the container. The job now resides in the App Master. Do note, that it communicates the status of the job to AM.

4. It is important to note that the Yarn Container can only be used when a Container Launch Context (CLC) certificate is provided by the App Master. It works like a key to unlock the container's resources. This is internal to YARN. The App Master job is now executed in the container. However, if the resources provided are not sufficient then

5. the App Master creates a list of requests and

6. This list is sent directly to the RM Scheduler. RM again requests the AM for more resources.

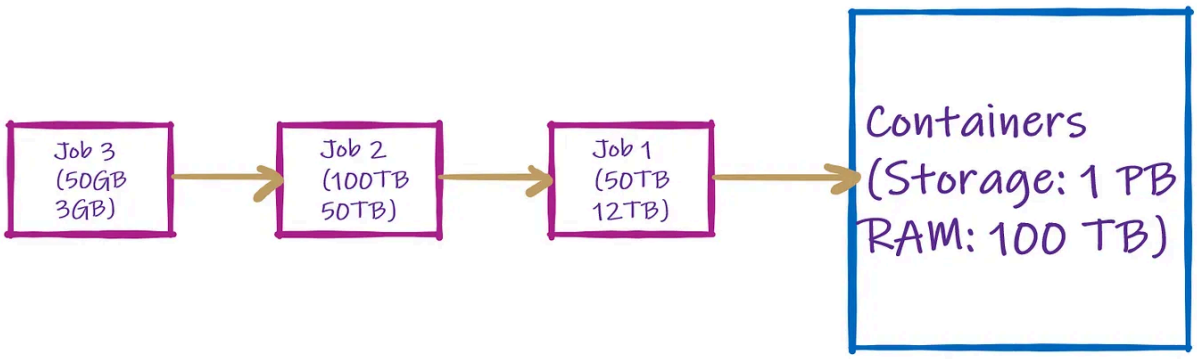
7. A new container is launched through a new NM without an App Master. The job is successfully executed and the resources are released.

Scheduler has an internal policy on How to give resources to the job?

YARN Schedulers:

1) **FIFO Scheduling**
2) **Capacity Scheduling**
3) **Fair Scheduling**

FIFO Scheduling:



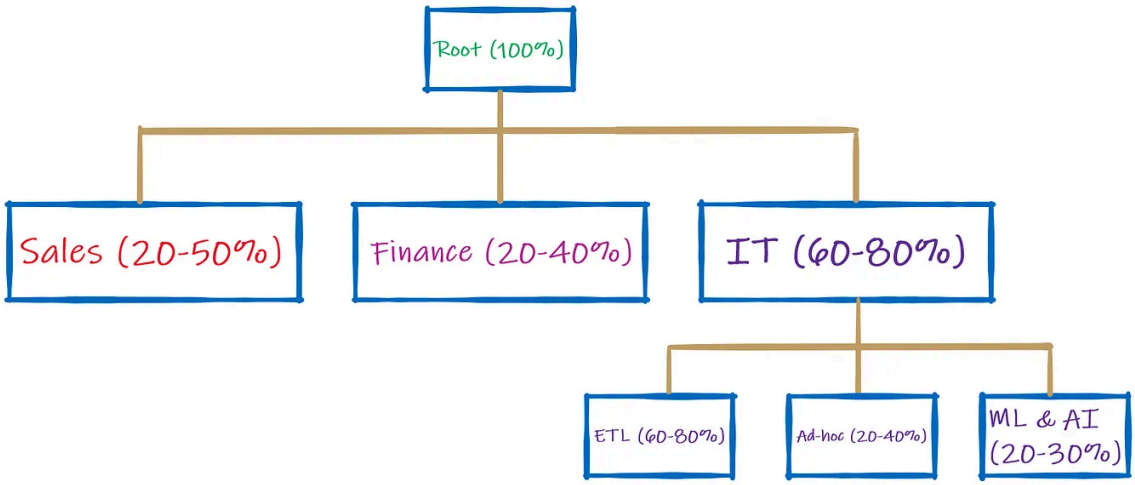
As the name suggests, First in First out or FIFO is the most basic scheduling method provided in YARN.

FIFO places jobs submitted by the client in queues and executes them in a sequential manner on a first-come-first-serve basis.

Although we could run multiple jobs together, in FIFO, they will run sequentially. Such is the waste. Hence, this scheduling methodology is not preferred on a Production/Shared Cluster as it suffers from poor resource utilization.

Capacity Scheduler was introduced to maximize utilization.

Capacity Scheduling:



The central idea is that multiple departments fund the central cluster or “root” denoted as 100% of available resources as seen on top of the above chart. Each department or “leaf” is guaranteed a specific range of capacity to carry out its jobs whenever required without waiting.

Basically, it’ll provide the resources to each job according to its requirement plus it’ll also provide additional runtime if some job is taking more time, like here IT department.

The priority functionality, depending on a case-to-case basis, is a drawback of Capacity Scheduling.

Fair Scheduling:

You need to understand how fair and capacity scheduling works, as Fair scheduling builds upon its drawbacks.

If a single job is run, Fair Scheduling (FS) will throw all its resources at it. If another job is added, it will ensure that a “fair” amount of resources are added to finish that job too.

The main aim is that all jobs receive a “fair” amount of resources over-time to execute successfully. This is especially true when there are 2 jobs and one of them is small. The small job also receives an adequate amount of resources to execute instead of being put on hold until the big job completes

Components of YARN:

1) Resource Manager:

   The Resource Manager is described in the official documentation as the ultimate authority that arbitrates resources among all the applications in the system. The resource manager consists of two parts:

   a. Application Managers
   b. Scheduler

   The astute reader would realize that the Resource Manager acts as a single point of failure. If the machine hosting the RM goes down, no jobs can get scheduled. To mitigate this shortcoming, high availability for YARN was introduced in Hadoop 2.4. A pair of Resource Managers runs in Active/Standby configuration to achieve high availability. If the active Resource Manager dies, then the standby one becomes the active and the cluster continues to function correctly.

2) Node Manager:

   The NodeManager is an agent that runs on every machine in the cluster. It is responsible for launching containers on that machine and managing the use of resources by the containers. It reports the usage back to the Scheduler component of the Resource Manager.

3) Application Master:

   It maintains a registry of running applications and monitors their execution. Whenever a job is submitted to the framework, an Application Master is elected for it. It is in charge of allocating resources from the Resource Manager to the Node Manager, which then monitors and executes the tasks.