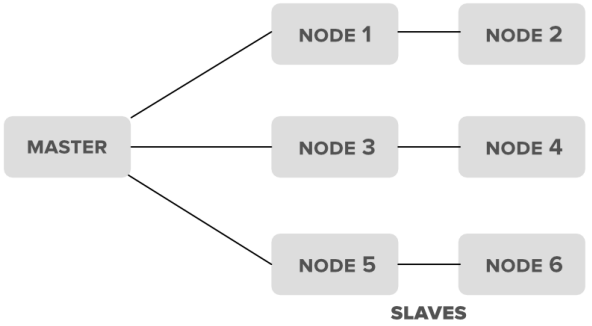
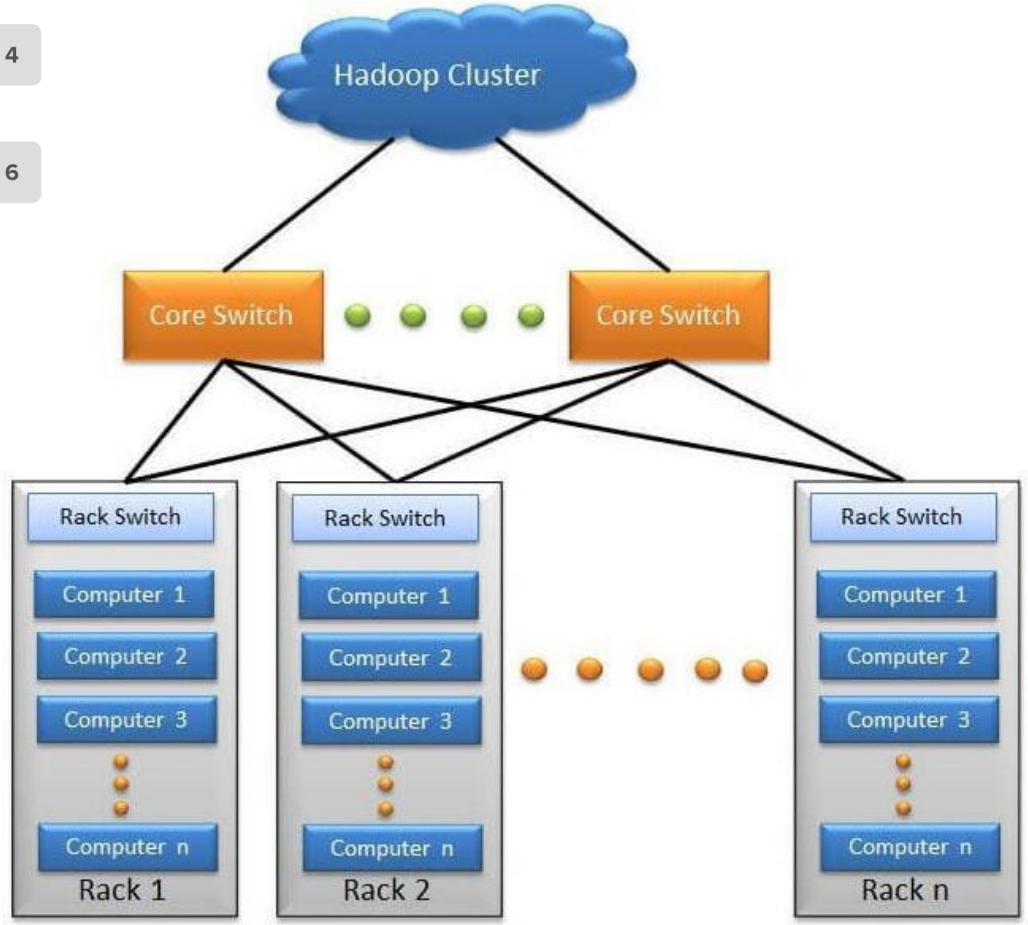
Hadoop Cluster

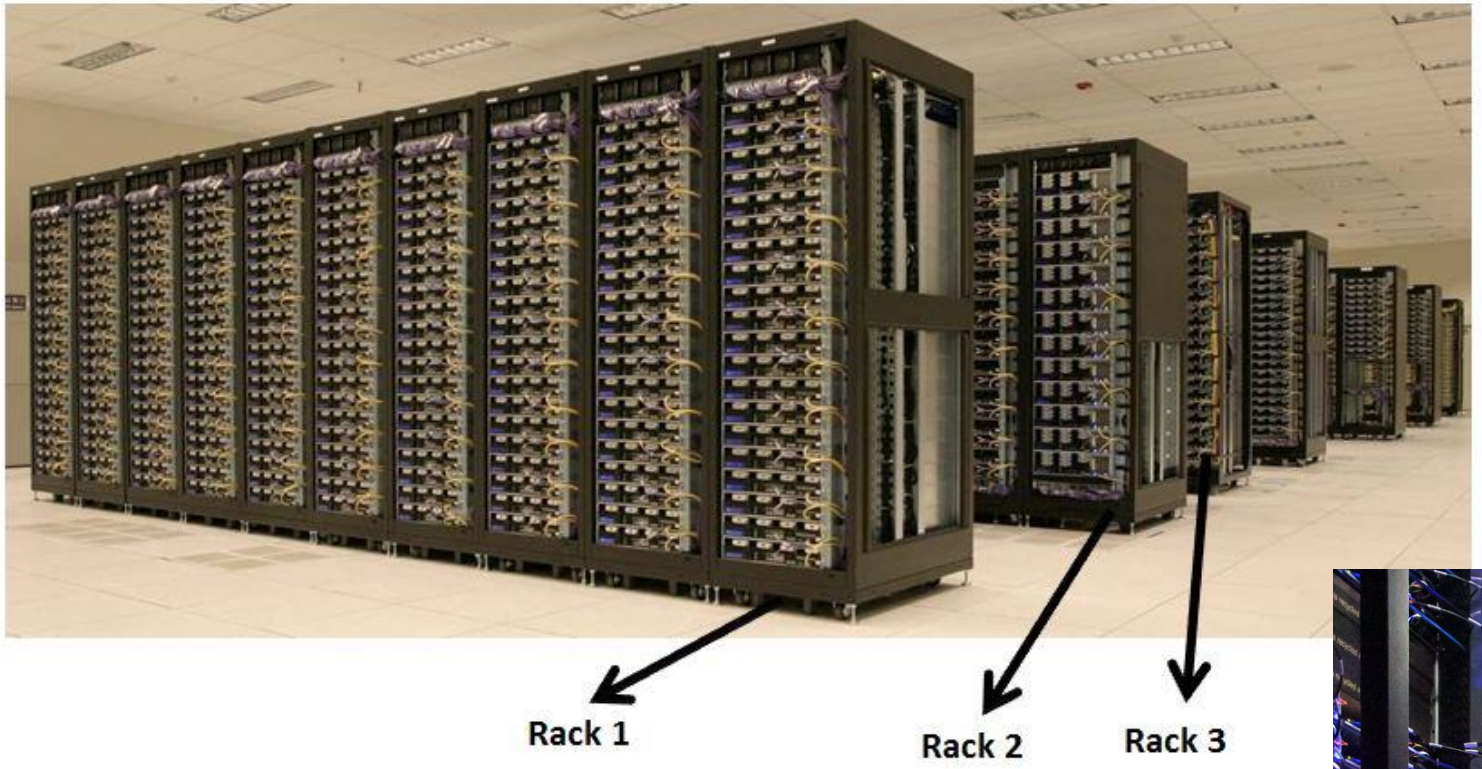
A Hadoop cluster is **a collection of computers, known as nodes, that are networked together to perform these kinds of parallel computations on big data sets.**



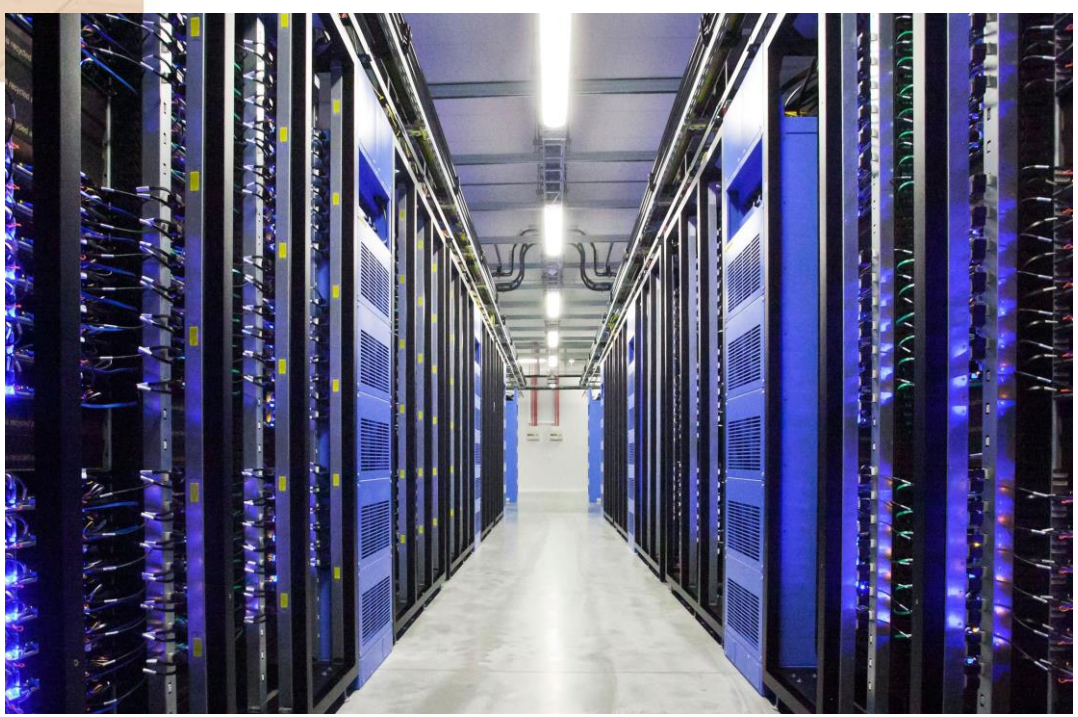
Hadoop Cluster Architecture

- Cluster: Loosely/Tightly connected computers work together as a single system
- Hadoop Cluster: Storing & analysing large amount of unstructured data in distributed environment.
- These clusters run on low cost on commodity computers





A Picture of Yahoo's Hadoop Cluster

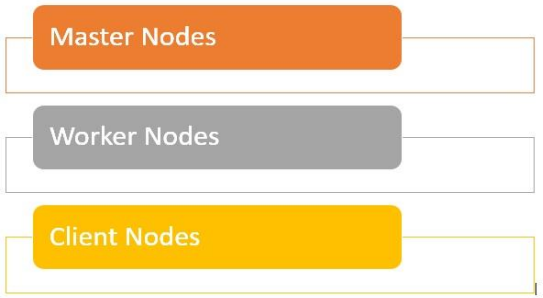


A Hadoop cluster is **a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment**. Such clusters run Hadoop's open source distributed processing software on low-cost commodity computers

Hadoop Cluster Architecture

Hadoop clusters are composed of a network of master and worker nodes that orchestrate and execute the various jobs across the Hadoop distributed file system. The master nodes typically utilize higher quality hardware and include a NameNode, Secondary NameNode, and JobTracker, with each running on a separate machine. The workers consist of virtual machines, running both **DataNode and TaskTracker** services on commodity hardware, and do the actual work of storing and processing the jobs as directed by the master nodes. The final part of the system are the Client Nodes, which are responsible for loading the data and fetching the results.

- Master nodes are responsible for storing data in **HDFS** and overseeing key operations, such as running parallel computations on the data using MapReduce.
- The worker nodes comprise most of the virtual machines in a Hadoop cluster, and perform the job of storing the data and running computations. Each worker node runs the DataNode and TaskTracker services, which are used to receive the instructions from the master nodes.
- Client nodes are in charge of loading the data into the cluster. Client nodes first submit MapReduce jobs describing how data needs to be processed and then fetch the results once the processing is finished.



What is cluster size in Hadoop?

- A Hadoop cluster size is a set of metrics that defines storage and compute capabilities to run Hadoop workloads, namely :Number of nodes : number of Master nodes, number of Edge Nodes, number of Worker Nodes.
- Configuration of each type node: number of cores per node, RAM and Disk Volume.

What are the advantages of a Hadoop Cluster?

- Hadoop clusters can boost the processing speed of many big data analytics jobs, given their ability to break down large computational tasks into smaller tasks that can be run in a parallel, distributed fashion.
- Hadoop clusters are easily scalable and can quickly add nodes to increase throughput, and maintain processing speed, when faced with increasing data blocks.
- The use of low cost, high availability commodity hardware makes Hadoop clusters relatively easy and inexpensive to set up and maintain.
- Hadoop clusters replicate a data set across the distributed file system, making them resilient to data loss and cluster failure.
- Hadoop clusters make it possible to integrate and leverage data from multiple different source systems and data formats.
- It is possible to deploy Hadoop using a single-node installation, for evaluation purposes.

What are the challenges of a Hadoop Cluster?

- Issue with small files - Hadoop struggles with large volumes of small files - smaller than the Hadoop block size of 128MB or 256MB by default. It wasn’t designed to support big data in a scalable way. Instead, Hadoop works well when there are a small number of large files. Ultimately when you increase the volume of small files, it overloads the Namenode as it stores namespace for the system.
- High processing overhead - reading and writing operations in Hadoop can get very expensive quickly especially when processing large amounts of data. This all comes down to Hadoop's inability to do in-memory processing and instead data is read and written from and to the disk.
- Only batch processing is supported - Hadoop is built for small volumes of large files in batches. This goes back to the way data is collected and stored which all has to be done before processing starts. What this ultimately means is that streaming data is not supported and it cannot do real-time processing with low latency.
- Iterative Processing - Hadoop has a data flow structure is set-up in sequential stages which makes it impossible to do iterative processing or use for ML.

Hadoop Cluster Modes

- Standalone mode
- Single node cluster/Pseudo distributed mode
- Multi node cluster/ Fully Distributed mode

Standalone mode

- Default mode
- HDFS is not utilized in this mode
- Local file system is used for input and output
- Used for debugging purpose
- No custom configuration is required in 3 Hadoop files
  - ➢mapred- site.xml
  - ➢core-site.xml
  - ➢hdfs-site.xml
- Standalone mode is much faster than pseudo distributed mode

Pseudo distributed mode / Single Node Cluster

A single node cluster means **only one DataNode running and setting up all the NameNode, DataNode, ResourceManager, and NodeManager on a single machine**.

- Configuration is required in given 3 files
- Replication factor is one for HDFS
- Here, one node is used as Master node/data node/Job tracker/Task tracker
- Used for real code to test in HDFS

Pseudo distributed cluster is a cluster where all daemons are running on one node itself

Fully Distributed mode/Multiple node cluster.

A Multi Node Cluster in Hadoop **contains two or more DataNodes in a distributed Hadoop environment**. This is practically used in organizations to store and analyze their Petabytes and Exabytes of data.

- Production phase
- This mode involves the code running on an actual Hadoop cluster.
- It is mode in which you see the actual power of Hadoop, when you run your code against a very large input on 1000s of servers.
- It is always difficult to debug a MapReduce program as you have Mappers running on different machine with different piece of input.
- You can never know where the Mappers are going to run eventually.
- Also, with large inputs, it is likely that the data will be irregular in its format.

1. Explain Hadoop Cluster Modes?
2. Write Hadoop Multinode Cluster configuration steps and Yarn Comfiguration?
3. Write a JAVA MapReduce programe along with execution steps?
4. Explain any 8 commands for Interacting HDFS using command line?
5. What are Java Classes and methods to interacting HDFS using JAVA API?
6. Explain File handling and Administrative Hadoop commands?