* **Hadoop**

1. **What is Hadoop?**

Hadoop is an open-source framework that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.

It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

1. **Why do we use Hadoop?**

Hadoop is used to develop applications that could perform complete statistical analysis on huge amounts of data.

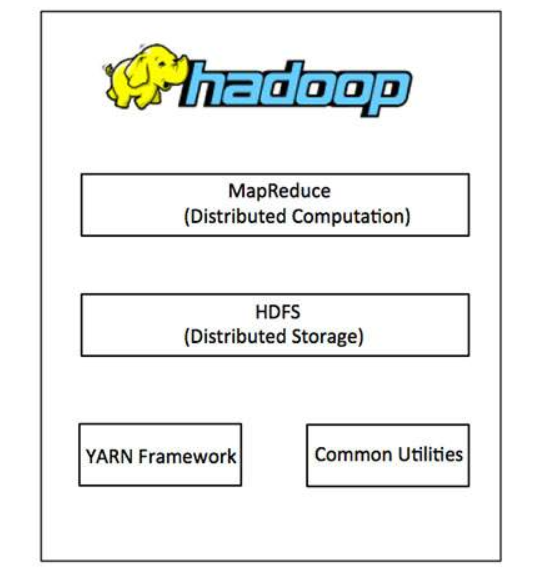
1. **In which language Hadoop is developed?**

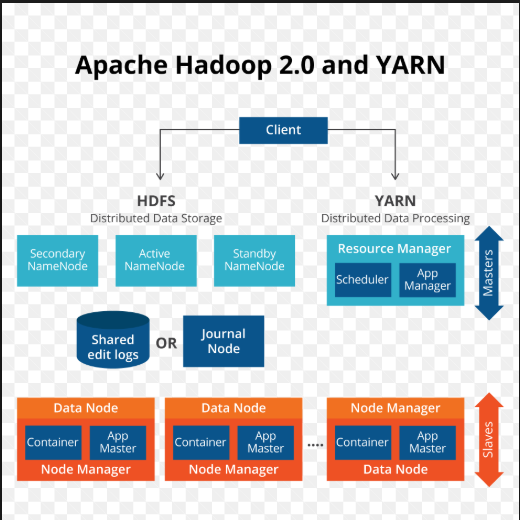
Java

1. **Explain Hadoop Architecture? (I)**

At its core, Hadoop has two major layers namely-

1. Processing/Computation layer: MapReduce.
2. Storage layer: Hadoop Distributed File System





**MapReduce:** is a processing technique and a program model using map reduce we can process/compute huge set of data in distributed manner based on java in a reliable, fault-tolerant manner.

**HDFS (H**adoop **D**istributed **F**ile **S**ystem**):**

It is distributed file system based on Google File System (GFS) which is designed to run on commodity hardware.

**Hadoop YARN:**

This framework is for job scheduling and cluster resource management.

**Hadoop Common:**

These are Java libraries and utilities required by other Hadoop modules

1. **How does Hadoop Works?**

We know that Hadoop is cluster of commodity hardware. Hadoop stores the data and runs the code across the cluster.

Below are the important tasks that Hadoop does

1. Data is initially divided into directories and files. Files are divided into uniform sized blocks of 128M and 64M (preferably 128M).
2. These files are then distributed across various cluster nodes for further processing.
3. HDFS, being on top of the local file system, supervises the processing.
4. Blocks are replicated for handling hardware failure.
5. Checking that the code was executed successfully.
6. Performing the sort that takes place between the map and reduce stages.
7. Sending the sorted data to a certain computer.
8. Writing the debugging logs for each job.
9. **What are Hadoop operation modes?**

We can operate Hadoop cluster in one of the three supported modes.

**Local/Standalone Mode:** by default, Hadoop is configured in a standalone mode and can be run as a single java process.

**Pseudo Distributed Mode** − It is a distributed simulation on single machine. Each Hadoop daemon such as hdfs, yarn, MapReduce etc., will run as a separate java process. This mode is useful for development.

**Fully Distributed Mode** − this mode is fully distributed with minimum two or more machines as a cluster. We will come across this mode in detail in the coming chapters.