Hadoop

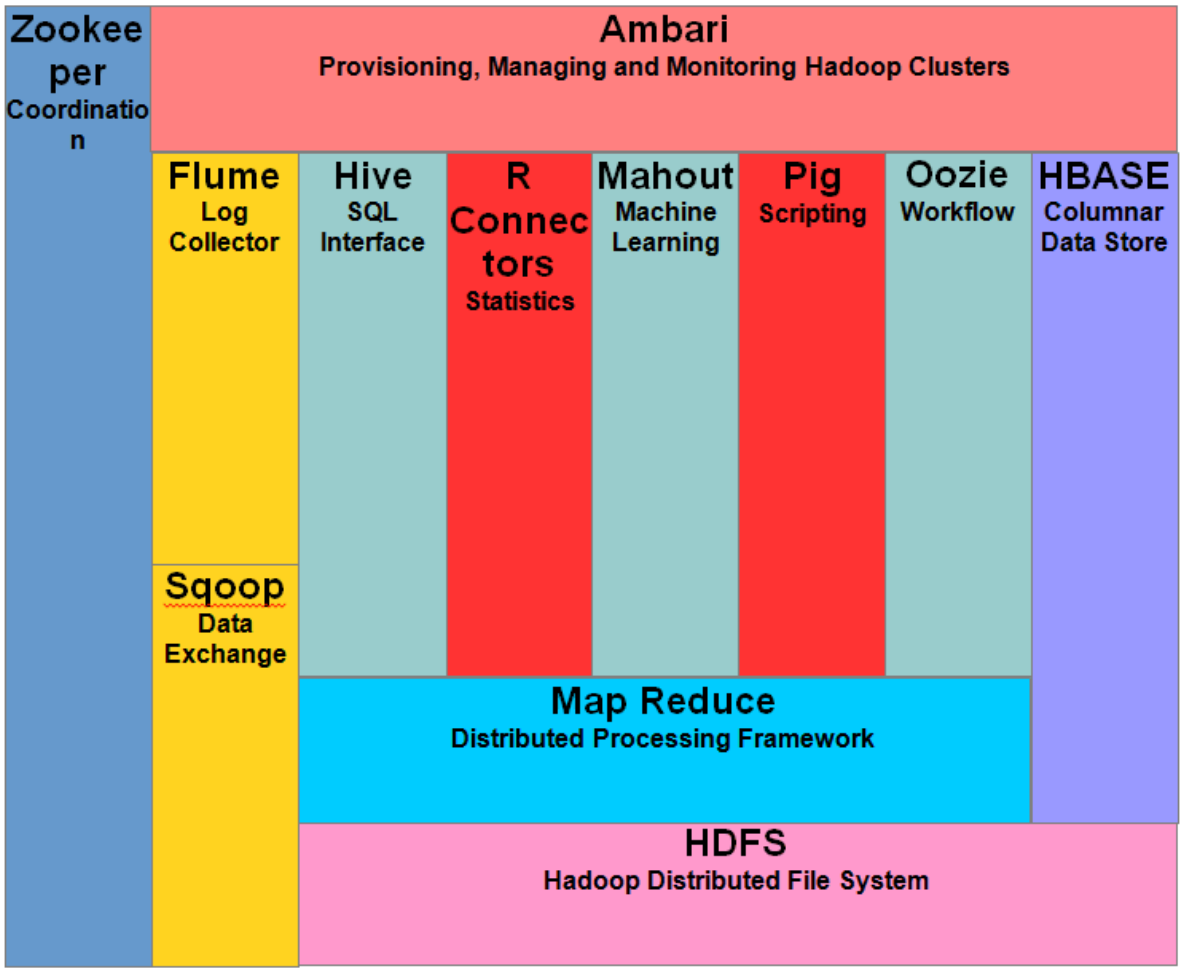
What is Hadoop?

Apache Hadoop is a free and open source software framework for developing data processing applications that run on distributed computing platforms.
HADOOP-based applications run on massive data sets distributed across clusters of commodity PCs. Computers classified as commodities are inexpensive and generally available. These are mostly beneficial for increasing computational power at a reasonable cost.

Similar to how data is stored on a personal computer's local file system, Hadoop data is stored on a distributed file system called the Hadoop Distributed File System. Based on the 'Data Locality' paradigm, computing logic is spread to cluster nodes (servers) that store data. Nothing more than a compiled version of a programme written in a high-level language such as Java constitutes this computational logic. Such a programme processes Hadoop HDFS data.

Hadoop EcoSystem and Components

The figure below illustrates the many components of the Hadoop ecosystem.



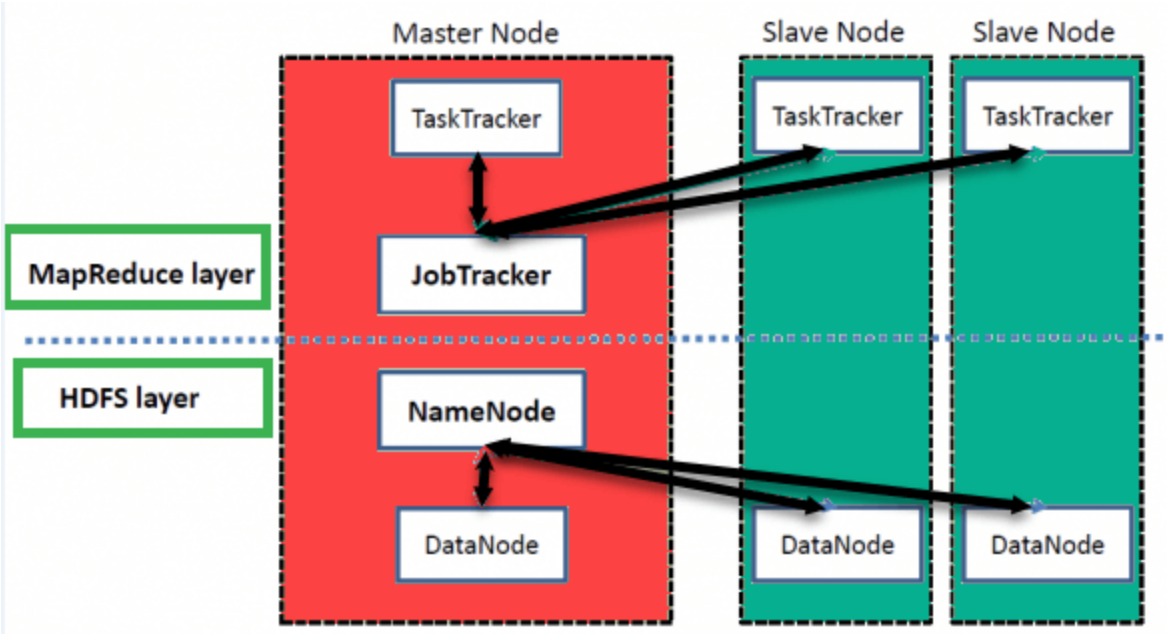
Apache Hadoop is divided into two sub-projects:

- Hadoop MapReduce: MapReduce is a computational methodology and software framework for developing Hadoop-based applications. These MapReduce applications are capable of parallelizing the processing of massive amounts of data on big clusters of compute nodes.

- HDFS (Hadoop Distributed File System): HDFS (Hadoop Distributed File System): HDFS is used to manage the storage for Hadoop applications. HDFS replicates data blocks and distributes them across a cluster's compute nodes. This distribution enables computations to be both trustworthy and incredibly fast.

While Hadoop is most commonly associated with MapReduce and its distributed file system, HDFS, the term is also used to refer to a set of related projects that fall under the field of distributed computing and large-scale data processing. Hive, HBase, Mahout, Sqoop, Flume, and ZooKeeper are several further Apache Hadoop projects.

Hadoop Architecture



High Level Hadoop Architecture

Hadoop uses the MapReduce and HDFS methods to store and process data in a Master-Slave architecture.

NameNode:

A NameNode has been used to represent each namespace's files and directories.

DataNode:

DataNode enables you to manage the state of an HDFS node and communicate with individual blocks.

MasterNode:

The master node enables concurrent data processing using Hadoop MapReduce.

Slave node:

Slave nodes are additional machines in the Hadoop cluster that enable the storage of data for sophisticated calculations. Additionally, each slave node includes a Task Tracker and a DataNode. This enables the processes to be synchronised with the NameNode and Job Tracker, respectively.

You can deploy a master or slave system in Hadoop either in the cloud or on-premises.

The Characteristics Of 'Hadoop'

• Suitable for Big Data Analysis

Due to the distributed and unstructured nature of Big Data, HADOOP clusters are ideally suited for Big Data analysis. Because processing logic (rather than real data) is sent to computing nodes, network traffic is conserved. This is referred to as the data locality idea, and it contributes to the efficiency of Hadoop-based systems.

• Extensibility

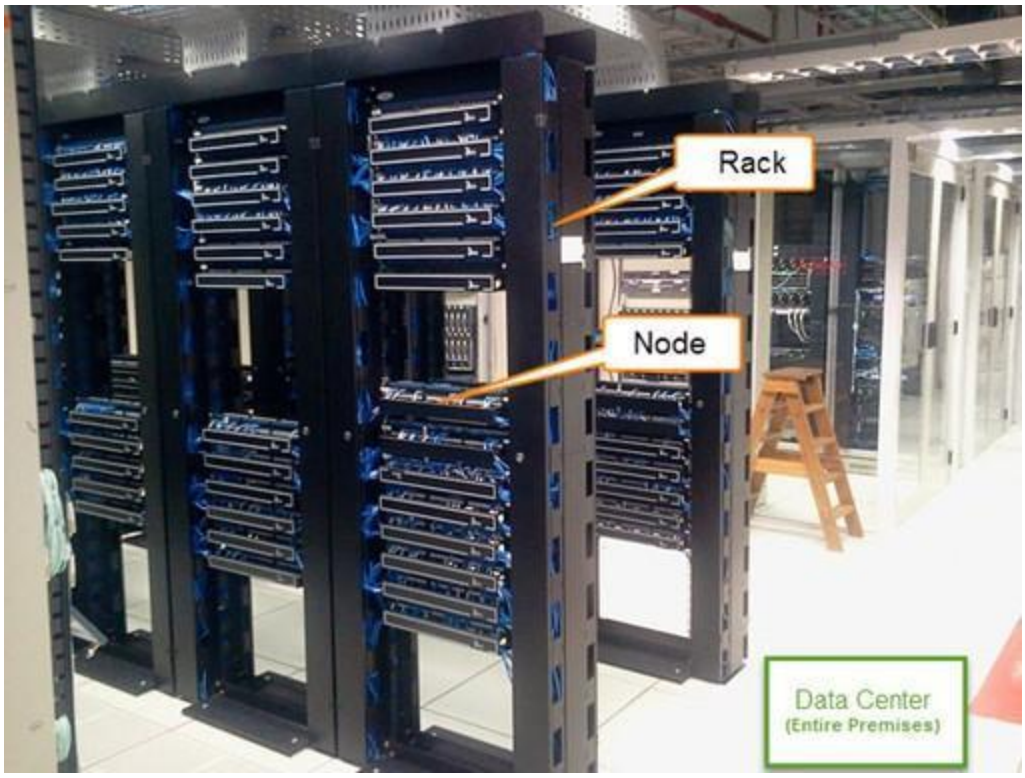
HADOOP clusters may be easily grown to any size by adding extra cluster nodes, which enables Big Data to grow. It's also not necessary to change the application's logic to achieve scalability.

• Fault Tolerance

As part of the HADOOP ecosystem, input data can be replicated over many cluster nodes. As a result, in the event of a cluster node failure, data stored on a different cluster node can be used to continue processing.

Network Topology In Hadoop

When the Hadoop cluster develops in size, the topology (organisation) of the network has an effect on its performance. Along with performance, one must consider high availability and failure handling. Cluster formation in Hadoop makes use of network topology to do this.



Typically, network bandwidth is a critical consideration while establishing any network. However, because measuring bandwidth can be challenging, Hadoop represents a network as a tree, and the distance between the nodes of this tree (number of hops) is a significant aspect in the development of a Hadoop cluster. The distance between two nodes in this case is equal to the total of their distances to their nearest common ancestor.

A Hadoop cluster is made up of three components: a data centre, a rack, and the node that actually runs operations. The data centre is composed of racks, and each rack is composed of nodes. The network bandwidth accessible to processes varies according to their location. That is, the available bandwidth decreases as we go away from-

- On the same node, many processes
- Numerous nodes on a single rack
- Nodes located in distinct racks inside the same data centre
- Nodes located in several data centres