Architecture of Hadoop intro

Hadoop consists of three component:

🡪HDFS (Hadoop Distributed File System)

🡪Yarn

🡪MapReduce

Let have some intro about above three components, will discuss this component well in detail in further word files of each component.

**HDFS**

🡪It is the Storage layer of Hadoop

🡪It consists of Name Node and Data Node where Name Node consists of meta data and Data Node consists of actual data placed in various blocks of each block size up to 64MB or 128 MB. Basically Name Node will assign the distributed content of the file in various block whereas each block contain only 1 content of the individual file automatically.

**Yarn**

🡪It is the resource management layer of Hadoop

🡪It provides resource management and job scheduling

**MapReduce**

🡪It is the processing layer of Hadoop

🡪It is a software framework for writing an application that perform distributed processing.

**Features of Hadoop**

🡪**Hadoop is an open source**

Hadoop is an open-source project, which means its source code is available free of cost for inspection, modification, and analyses that allows enterprises to modify the code as per their requirements.

### 🡪Hadoop cluster is Highly Scalable

Hadoop cluster is scalable means we can add any number of nodes (**horizontal scalable**) or increase the hardware capacity of nodes (**vertical scalable**) to achieve high computation power. This provides horizontal as well as vertical scalability to the Hadoop framework.

🡪**Hadoop provides Fault Tolerance**

Fault tolerance is the most important feature of Hadoop. [HDFS](https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html) in Hadoop 2 uses a **replication mechanism** to provide fault tolerance.

It creates a replica of each block on the different machines depending on the replication factor (by default, it is 3). So if any machine in a cluster goes down, data can be accessed from the other machines containing a replica of the same data.

Hadoop 3 has replaced this replication mechanism by **erasure coding**. Erasure coding provides the same level of fault tolerance with less space. With Erasure coding, the storage overhead is not more than 50%.

Read [**Erasure coding**](https://data-flair.training/blogs/hadoop-hdfs-erasure-coding/) article to learn the erasure coding algorithm.

🡪 **Hadoop provides High Availability**

This feature of Hadoop ensures the high availability of the data, even in unfavorable conditions.

Due to the fault tolerance feature of Hadoop, if any of the DataNodes goes down, the data is available to the user from different DataNodes containing a copy of the same data.

Also, the high availability Hadoop cluster consists of 2 or more running NameNodes (active and passive) in a hot standby configuration. The active node is the NameNode, which is active. Passive node is the standby node that reads edit logs modification of active NameNode and applies them to its own namespace.

If an active node fails, the passive node takes over the responsibility of the active node. Thus even if the NameNode goes down, files are available and accessible to users.

### 🡪 Hadoop is very Cost-Effective

Since the Hadoop cluster consists of nodes of commodity hardware that are inexpensive, thus provides a cost-effective solution for storing and processing big data. Being an open-source product, Hadoop doesn’t need any license.

### 🡪 Hadoop is Faster in Data Processing

Hadoop stores data in a distributed fashion, which allows data to be processed distributedly on a cluster of nodes. Thus it provides lightning-fast processing capability to the Hadoop framework.

### 🡪 Hadoop is based on Data Locality concept

Hadoop is popularly known for its data locality feature means moving computation logic to the data, rather than moving data to the computation logic. This features of Hadoop reduces the bandwidth utilization in a system.

### 🡪 Hadoop provides Feasibility

Unlike the traditional system, Hadoop can process unstructured data. Thus provide feasibility to the users to analyze data of any formats and size.

### 🡪 Hadoop is Easy to use

Hadoop is easy to use as the clients don’t have to worry about distributing computing. The processing is handled by the framework itself.

### 🡪 Hadoop ensures Data Reliability

In Hadoop due to the replication of data in the cluster, data is stored reliably on the cluster machines despite machine failures.

The framework itself provides a mechanism to ensure data reliability by Block Scanner, Volume Scanner, Disk Checker, and Directory Scanner. If your machine goes down or data gets corrupted, then also your data is stored reliably in the cluster and is accessible from the other machine containing a copy of data.