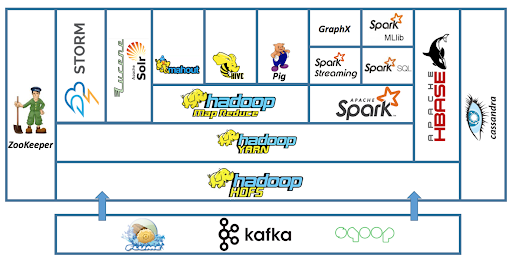
**Hadoop**

Hadoop is a framework that uses distributed storage and parallel processing to store and manage big data. It is the software most used by data analysts to handle big data



## Components of Hadoop

Hadoop is a framework that uses distributed storage and parallel processing to store and manage Big Data. It is the most commonly used software to handle Big Data. There are three components of Hadoop.

1. Hadoop HDFS - Hadoop Distributed File System (HDFS) is the storage unit of Hadoop.
2. Hadoop MapReduce - Hadoop MapReduce is the processing unit of Hadoop.
3. Hadoop YARN - Hadoop YARN is a resource management unit of Hadoop.

### Hadoop HDFS

Data is stored in a distributed manner in HDFS. There are two components of HDFS - name node and [data](https://www.simplilearn.com/what-is-data-article) node. While there is only one name node, there can be multiple data nodes.

### Name Node:

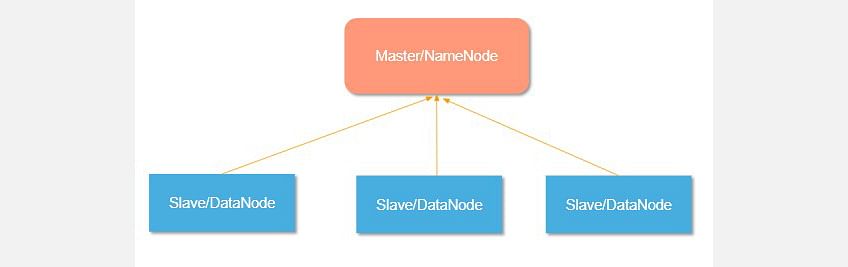
Name Node is the primary component of HDFS. Name Node maintains the file systems along with namespaces. Actual data cannot be stored in the Name Node. The modified data, such as Metadata, block data etc., can be stored here.

### Data Node:

Data Node follows the instructions given by the Name Node. Data Nodes are also known as ‘slave Nodes’. These nodes store the actual data provided by the client and simply follow the commands of the Name Node.

#### Master and Slave Nodes

Master and slave nodes form the [HDFS cluster](https://www.simplilearn.com/what-is-a-hadoop-cluster-article). The name node is called the master, and the data nodes are called the slaves.



The name node is responsible for the workings of the data nodes. It also stores the metadata.

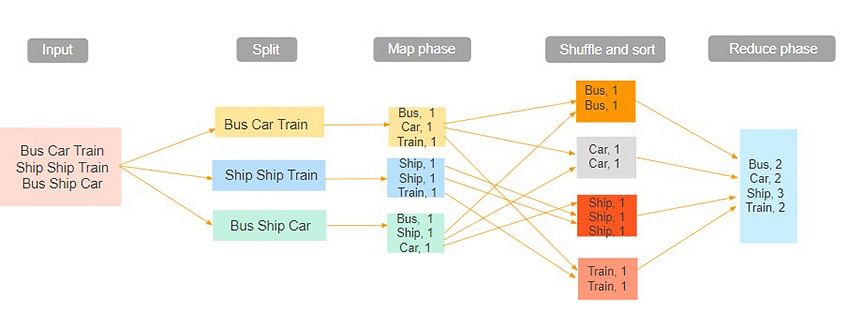
The data nodes read, write, process, and replicate the data. They also send signals, known as heartbeats, to the name node. These heartbeats show the status of the data node.

### Hadoop MapReduce

Hadoop MapReduce is the processing unit of Hadoop. In the MapReduce approach, the processing is done at the slave nodes, and the final result is sent to the master node.

MapReduce program executes in three stages, namely

* map stage,
* shuffle stage,
* reduce stage.



The input dataset is first split into chunks of data. In this example, the input has three lines of text with three separate entities - “bus car train,” “ship ship train,” “bus ship car.” The dataset is then split into three chunks, based on these entities, and processed parallelly.

In the map phase, the data is assigned a key and a value of 1. In this case, we have one bus, one car, one ship, and one train.

These key-value pairs are then shuffled and sorted together based on their keys. At the reduce phase, the aggregation takes place, and the final output is obtained.

### Hadoop YARN

Hadoop YARN stands for Yet Another Resource Negotiator. It is the resource management unit of Hadoop

* Hadoop YARN acts like an OS to Hadoop. It is a file system that is built on top of HDFS.
* It is responsible for managing cluster resources to make sure you don't overload one machine.
* It performs job scheduling to make sure that the jobs are scheduled in the right place

## How Hadoop Works:

The primary function of Hadoop is to process the data in an organised manner among the cluster of commodity software. The client should submit the data or program that needs to be processed. Hadoop HDFS stores the data. YARN, MapReduce divides the resources and assigns the tasks to the data. Let’s know the working of Hadoop in detail.

* The client input data is divided into 128 MB blocks by HDFS. Blocks are replicated according to the replication factor: various DataNodes house the unions and their duplicates.
* The user can process the data once all blocks have been put on HDFS DataNodes.
* The client sends Hadoop the MapReduce programme to process the data.
* The user-submitted software was then scheduled by ResourceManager on particular cluster nodes.
* The output is written back to the HDFS once processing has been completed by all nodes.

## Advantages of Hadoop

Hadoop was created to deal with big data, so it’s hardly surprising that it offers so many benefits. The five main benefits are:

* Speed. Hadoop’s concurrent processing, MapReduce model, and HDFS lets users run complex queries in just a few seconds.
* Diversity. Hadoop’s HDFS can store different data formats, like structured, semi-structured, and unstructured.
* Cost-Effective. Hadoop is an open-source data framework.
* Resilient. Data stored in a node is replicated in other cluster nodes, ensuring fault tolerance.
* Scalable. Since Hadoop functions in a distributed environment, you can easily add more servers.