In my previous blog i.e. [***Big Data Tutorial***](http://www.edureka.co/blog/big-data-tutorial?utm_source=blog&utm_medium=related-posts&utm_campaign=hadoop-tutorial), I discussed about Big Data in detail and the challenges with Big Data. I introduced Hadoop as a solution to Big Data problems. So in this Hadoop Tutorial, I will take you through Hadoop in detail. Hadoop is a huge concept, and looking at the interest and demand of Hadoop, I have posted a series of blogs on Hadoop which you can see in the menu on the left side.

This is the 1st blog of Hadoop Tutorial series in which I will cover the following topics:

1. What is Hadoop?
2. Hadoop Characteristics
3. Hadoop Core Components

If we drill down to the problems that Big Data has, we need something other than RDBMS which can store huge amount of data and process it as well. This is where Hadoop comes into the picture. It is truly said that ‘Necessity is the mother of all inventions’ and ‘Hadoop’ is among the finest inventions in the world of Big Data!

**Hadoop Tutorial: What is Hadoop?**

Hadoop is an open-source software framework used for storing and processing Big Data in a distributed manner on large clusters of commodity hardware. Hadoop is licensed under the Apache v2 license. It is therefore generally known as Apache Hadoop. Hadoop has been developed, based on a paper originally written by Google on MapReduce system and applies concepts of functional programming. Hadoop is written in the Java programming language and ranks among the highest-level Apache projects. Hadoop was developed by Doug Cutting and Michael J. Cafarella and the charming yellow elephant you see is basically named after Doug’s son’s toy elephant!

[Check out our Hadoop Course](http://www.edureka.co/big-data-and-hadoop?utm_source=blog&utm_medium=blog-cta&utm_campaign=hadoop-tutorial-1)

**Hadoop Tutorial: Hadoop Features  
**

***Flexibility:***Hadoop is very Flexible in terms of ability to deal with all kinds of data. We discussed “Variety” in our previous blog on [***Big Data Tutorial***](http://www.edureka.co/blog/big-data-tutorial?utm_source=blog&utm_medium=related-posts&utm_campaign=hadoop-tutorial), where data can be of any kind and Hadoop can store and process them all, whether it is structured, semi-structured or unstructured data.

***Reliability:***When machines are working in tandem, if one of the machine fails, another machine will take over the responsibility and work in a reliable and fault tolerant fashion. Hadoop infrastructure has in-built fault tolerance features and hence, Hadoop is highly reliable. We will understand this feature in more detail in upcoming blogs on HDFS.

***Economical:*** Hadoop uses commodity hardware (for example your PC, laptop) For example, in a small Hadoop cluster, all your data nodes can have normal configurations like 8-16 GB RAM with 5-10 TB hard disk and Xeon processors, but if I would have used hardware-based RAID with Oracle for the same purpose, I would end up spending 5x times more at least. So the cost of ownership of a Hadoop based project is pretty minimized. It is easier to maintain the Hadoop environment and is economical as well. Also, Hadoop is an open source software and hence there is no licensing cost.

***Scalability:***At last, but not the least, we talk about the scalability factor. Hadoop has in-built capability of integrating seamlessly with cloud based services. So if you are installing Hadoop on a cloud, you don’t need to worry about the scalability factor because you can go ahead and procure more hardware and expand your setup within minutes whenever required.

These 4 characteristics make Hadoop a front runner in terms of Big Data solution. Now that we know what is Hadoop, we can explore the core components of Hadoop.

But before talking about Hadoop core components, I will explain what led to the creation of these components.

There were two major challenges with Big Data:

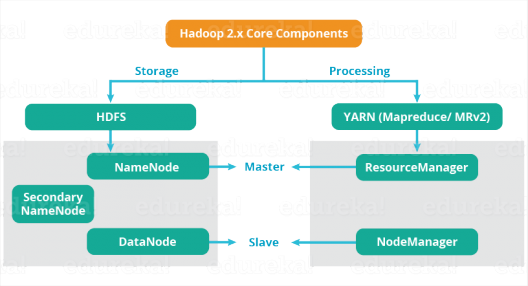
* ***Big Data Storage:*** To store Big Data, in a flexible infrastructure that scales up in a cost effective manner, was critical.
* ***Big Data Processing:*** Even if a part of Big Data is Stored, processing it would take years.

To solve the storage issue and processing issue, two core components were created in Hadoop – **HDFS** and **YARN**. HDFS solved the storage issue as it stores the data in a distributed fashion and is easily scalable, you will get to know more about HDFS in my blog on [***HDFS Tutorial***](https://www.edureka.co/blog/hdfs-tutorial?utm_source=blog&utm_medium=content-link&utm_campaign=hadoop-tutorial). And, YARN solved the processing issue by reducing the processing time drastically.

Let us now understand, what are these core components of Hadoop.

**Hadoop Tutorial: Hadoop Core Components**

When you are setting up a Hadoop cluster, you can choose a lot of services that can be part of Hadoop platform, but there are two services which are always mandatory as a part of Hadoop set up. One is **HDFS (storage)** and the other is **YARN (processing)**. HDFS stands for **Hadoop Distributed File System**, which is used for storing Big Data. It is highly scalable. Once you have brought the data onto HDFS, MapReduce can run jobs to process this data. YARN does all the resource management and scheduling of these jobs.

**Hadoop Master Slave Architecture**

The main components of ***HDFS*** are **NameNode** and **DataNode**.

**NameNode**

* It is the master daemon that maintains and manages the DataNodes (slave nodes)
* It records the metadata of all the files stored in the cluster, e.g. location of blocks stored, the size of the files, permissions, hierarchy, etc.
* It records each and every change that takes place to the file system metadata
* For example, if a file is deleted in HDFS, the NameNode will immediately record this in the EditLog
* It regularly receives a Heartbeat and a block report from all the DataNodes in the cluster to ensure that the DataNodes are live
* It keeps a record of all the blocks in HDFS and in which nodes these blocks are stored
* It has high availability and federation features which I will discuss in HDFS architecture in detail

**DataNode**

* These are slave daemons which runs on each slave machine
* The actual data is stored on DataNodes
* They are responsible for serving read and write requests from the clients
* They are also responsible for creating blocks, deleting blocks and replicating the same based on the decisions taken by the NameNode
* They send heartbeats to the NameNode periodically to report the overall health of HDFS, by default, this frequency is set to 3 seconds

The components of ***YARN*** are **ResourceManager** and **NodeManager**.

**ResourceManager**

* It is a cluster level (one for each cluster) component and runs on the master machine
* It manages resources and schedule applications running on top of YARN
* It has two components: Scheduler & ApplicationManager
* The Scheduler is responsible for allocating resources to the various running applications
* The ApplicationManager is responsible for accepting job submissions and negotiating the first container for executing the application
* It keeps a track of the heartbeats from the Node Manager

**NodeManager**

* It is a node level component (one on each node) and runs on each slave machine
* It is responsible for managing containers and monitoring resource utilization in each container
* It also keeps track of node health and log management
* It continuously communicates with ResourceManager to remain up-to-date

Go through our **Hadoop Tutorial** **video** below to know more about Big Data and Hadoop:

**Apache Hadoop Tutorial | Hadoop Tutorial For Beginners | Edureka**

This Edureka Hadoop tutorial video helps you understand Big Data and Hadoop in detail. This Hadoop tutorial is ideal for both beginners as well as professionals who want to learn or brush up their Hadoop concepts.

This is the end of the 1st blog of the Hadoop Tutorial blog series. We now know what is Hadoop, its features and the core components of Hadoop.

In my next blog in the Hadoop Tutorial series, I will talk about the [***Hadoop Ecosystem***](http://www.edureka.co/blog/hadoop-ecosystem?utm_source=blog&utm_medium=content-link&utm_campaign=hadoop-tutorial) which is a suite of several tools and platforms working on top of Hadoop.