**HADOOP**

1. Scalability: Hadoop clusters are very much capable of scaling-up and scaling-down the number of nodes i.e. servers or commodity hardware. Let’s see with an example of what actually this scalable property means. Suppose an organization wants to analyze or maintain around 5PB of data for the upcoming 2 months so he used 10 nodes(servers) in his Hadoop cluster to maintain all of this data. But now what happens is, in between this month the organization has received extra data of 2PB, in that case, the organization has to set up or upgrade the number of servers in his Hadoop cluster system from 10 to 12(let’s consider) in order to maintain it. The process of scaling up or scaling down the number of servers in the Hadoop cluster is called scalability.

2. Flexibility: This is one of the important properties that a Hadoop cluster possesses. According to this property, the Hadoop cluster is very much Flexible means they can handle any type of data irrespective of its type and structure. With the help of this property, Hadoop can process any type of data from online web platforms.

3. Speed: Hadoop clusters are very much efficient to work with a very fast speed because the data is distributed among the cluster and also because of its data mapping capability’s i.e. the MapReduce architecture which works on the Master-Slave phenomena.

4. No Data-loss: There is no chance of loss of data from any node in a Hadoop cluster because Hadoop clusters have the ability to replicate the data in some other node. So in case of failure of any node no data is lost as it keeps track of backup for that data.

5. Economical: The Hadoop clusters are very much cost-efficient as they possess the distributed storage technique in their clusters i.e. the data is distributed in a cluster among all the nodes. So in the case to increase the storage we only need to add one more another hardware storage which is not that much costliest.

1. Single Node Hadoop Cluster: In Single Node Hadoop Cluster as the name suggests the cluster is of an only single node which means all our Hadoop Daemons i.e. Name Node, Data Node, Secondary Name Node, Resource Manager, Node Manager will run on the same system or on the same machine. It also means that all of our processes will be handled by only single JVM(Java Virtual Machine) Process Instance.

2. Multiple Node Hadoop Cluster: In multiple node Hadoop clusters as the name suggests it contains multiple nodes. In this kind of cluster set up all of our Hadoop Daemons, will store in different-different nodes in the same cluster setup. In general, in multiple node Hadoop cluster setup we try to utilize our higher processing nodes for Master i.e. Name node and Resource Manager and we utilize the cheaper system for the slave Daemon’s i.e.Node Manager and Data Node.

there are “5Vs” of Big Data which are also termed as the characteristics of Big Data.

Volume: With increasing dependence on technology, data is producing at a large volume. Common examples are data being produced by various social networking sites, sensors, scanners, airlines and other organizations.

Velocity: Huge amount of data is generated per second. It is estimated that by the end of 2020, every individual will produce 3mb data per second. This large volume of data is being generated with a great velocity.

Variety: The data being produced by different means is of three types:

Structured Data: It is the relational data which is stored in the form of rows and columns.

Unstructured Data: Texts, pictures, videos etc. are the examples of unstructured data which can’t be stored in the form of rows and columns.

Semi Structured Data: Log files are the examples of this type of data.

Veracity: The term Veracity is coined for the inconsistent or incomplete data which results in the generation of doubtful or uncertain Information. Often data inconsistency arises because of the volume or amount of data e.g. data in bulk could create confusion whereas less amount of data could convey half or incomplete Information.

Value: After having the 4 V’s into account there comes one more V which stands for Value!. Bulk of Data having no Value is of no good to the company, unless you turn it into something useful. Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V’s

Evolution of Hadoop: Hadoop was designed by Doug Cutting and Michael Cafarella in 2005. The design of Hadoop is inspired by Google. Hadoop stores the huge amount of data through a system called Hadoop Distributed File System (HDFS) and processes this data with the technology of Map Reduce. The designs of HDFS and Map Reduce are inspired by the Google File System (GFS) and Map Reduce. In the year 2000 Google suddenly overtook all existing search engines and became the most popular and profitable search engine. The success of Google was attributed to its unique Google File System and Map Reduce. No one except Google knew about this, till that time. So, in the year 2003 Google released some papers on GFS. But it was not enough to understand the overall working of Google. So in 2004, Google again released the remaining papers. The two enthusiasts Doug Cutting and Michael Cafarella studied those papers and designed what is called, Hadoop in the year 2005. Doug’s son had a toy elephant whose name was Hadoop and thus Doug and Michael gave their new creation, the name “Hadoop” and hence the symbol “toy elephant.” This is how Hadoop evolved. Thus the designs of HDFS and Map Reduced though created by Doug Cutting and Michael Cafarella, but are originally inspired by Google. For more details about the evolution of Hadoop, you can refer to Hadoop | History or Evolution.

Traditional Approach: Suppose we want to process a data. In the traditional approach, we used to store data on local machines. This data was then processed. Now as data started increasing, the local machines or computers were not capable enough to store this huge data set. So, data was then started to be stored on remote servers. Now suppose we need to process that data. So, in the traditional approach, this data has to be fetched from the servers and then processed upon. Suppose this data is of 500 GB. Now, practically it is very complex and expensive to fetch this data. This approach is also called Enterprise Approach.

In the new Hadoop Approach, instead of fetching the data on local machines we send the query to the data. Obviously, the query to process the data will not be as huge as the data itself. Moreover, at the server, the query is divided into several parts. All these parts process the data simultaneously. This is called parallel execution and is possible because of Map Reduce. So, now not only there is no need to fetch the data, but also the processing takes lesser time. The result of the query is then sent to the user. Thus the Hadoop makes data storage, processing and analyzing way easier than its traditional approach.

Components of Hadoop: Hadoop has three components:

HDFS: Hadoop Distributed File System is a dedicated file system to store big data with a cluster of commodity hardware or cheaper hardware with streaming access pattern. It enables data to be stored at multiple nodes in the cluster which ensures data security and fault tolerance.

Map Reduce : Data once stored in the HDFS also needs to be processed upon. Now suppose a query is sent to process a data set in the HDFS. Now, Hadoop identifies where this data is stored, this is called Mapping. Now the query is broken into multiple parts and the results of all these multiple parts are combined and the overall result is sent back to the user. This is called reduce process. Thus while HDFS is used to store the data, Map Reduce is used to process the data.

YARN : YARN stands for Yet Another Resource Negotiator. It is a dedicated operating system for Hadoop which manages the resources of the cluster and also functions as a framework for job scheduling in Hadoop. The various types of scheduling are First Come First Serve, Fair Share Scheduler and Capacity Scheduler etc. The First Come First Serve scheduling is set by default in YARN.

Hadoop 3: This is the recent version of Hadoop. Along with the merits of the first two versions, Hadoop 3 has one most important merit. It has resolved the issue of single point failure by having multiple name nodes. Various other advantages like erasure coding, use of GPU hardware and Dockers makes it superior to the earlier versions of Hadoop.

Economically Feasible: It is cheaper to store data and process it than it was in the traditional approach. Since the actual machines used to store data are only commodity hardware.

Easy to Use: The projects or set of tools provided by Apache Hadoop are easy to work upon in order to analyze complex data sets.

Open Source: Since Hadoop is distributed as an open source software under Apache License, so one does not need to pay for it, just download it and use it.

Fault Tolerance: Since Hadoop stores three copies of data, so even if one copy is lost because of any commodity hardware failure, the data is safe. Moreover, as Hadoop version 3 has multiple name nodes, so even the single point of failure of Hadoop has also been removed.

Scalability: Hadoop is highly scalable in nature. If one needs to scale up or scale down the cluster, one only needs to change the number of commodity hardware in the cluster.

Distributed Processing: HDFS and Map Reduce ensures distributed storage and processing of the data.

Locality of Data: This is one of the most alluring and promising features of Hadoop. In Hadoop, to process a query over a data set, instead of bringing the data to the local computer we send the query to the server and fetch the final result from there. This is called data locality.