1. Explain what is a cluster and what is a hadoop cluster

* CLUSTER:
* Cluster can generally be defined as a group or a collection of similar items put together.
* In a computer system, a cluster can be defined as a group of servers and other resources that act like a single system and enable high availability and, in some cases, load balancing and parallel processing.
* In terms of storage they can be viewed as a logical unit of file storage on a hard disk etc.

HADOOP CLUSTER:

|  |
| --- |
|  |
|  |
|  |  | |
|  |  | |
| * They can be viewed as a collection of a no. of loosely coupled computers connected together by a network used in hadoop. | |
|  | |
| * They were designed for storing and analyzing huge amounts of unstructured data in a distributed environment. | |
|  | |
| * These clusters are cost-Efficient as they run on low cost computers. | |
|  | |
| * With the introduction of hadoop clusters, the data reading speed and the time taken to read TB's of data is reduced. * They are similar to that of normal clusters in structure but the difference being that These are computer collection used to store unstructured data employed in hadoop! * These clusters consists of 3 components: * Client * Master * Slave | |
|  | |
|  | |

1. **Components of Hadoop 1.x**

   => NameNode

   => Secondary Namenode

 => DataNode

   => JobTracker

=> TaskTracker

 => NameNode:

* It is often referred to as the Master while its slaves are the datanodes.
* It is often used to store the metadata information of HDFS and contains 2 information – fsimage and edits.
* This node is responsible for checking whether or not a file can be stored in hdfs, i.e it will have the information of how much space each data node contains! So when a file needs to be stored this name node will analyse its space whether the data nodes have enough space to store them .
* It keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept.
* It follows a centralized structure so, there are chances for single point failure.

 => Secondary Namenode

* This is a component that performs the house-keeping activities for the name node.
* The fsimage contains the HDFS metadata and edits contain the changes in the HDFS metadata, so to get the exact view of the HDFS structure, both must be merged which will be done by this secondary namenode! Which is often referred to as checkpointing.
* Moreover ,once the fsimage and the edits are merged this component can also act as a name node .
* But, it will not contain any data of the edits.

 => DataNode

* This is the main component of hadoop which stores the actual blocks of data.
* They are often referred to as the slaves or workhorse of the system which is controlled by the namenode.
* They periodically keep sending signals to their master to ensure that they are active often referred to as heart beats.
* They perform all the block operations, including periodic the Checksum. They receive instructions from the namenode of where to put the blocks and how to put the blocks.

   => JobTracker

* This is the component responsible for coordinating the execution of data i.e dataprocessing or mapreduce jobs.
* It often communicates with the namenode to identify the location of the data.
* It accounts to single point failure ,all the jobs get halted if something goes wrong with this component.
* This component identifies the nodes present near the data via task tracker and assigns the job to it.
* Once all the work is completed it will update its status.

=> TaskTracker

* This component runs the individual mapreduce jobs on the data nodes.
* Once the job tracker assigns the job to it ,it will locate the necessary datanodes and do the processing and finally reports to the jobtracker when done.
* It often receives signals from the job tracker and it gives updates to it in return.