**HDFS**

\* HDFS is popularly known as Hadoop Distributed File System,which is the core component of Hadoop. HDFS is a java-based file system and is the place where all the data in the Hadoop cluster resides.In typical terms, Hadoop has the Master-Slave architecture. This is named in perspective to the HDFS.

\* It is called as Master-Slave architecture because there is a Master which takes control of all the Slaves. Here the Master is named as Namenodes and the Slaves are named as Datanodes.

\* HDFS is a scalable, fault-tolerant, distributed storage system that works closely with a wide variety of concurrent data access applications, coordinated by YARN. HDFS will “just work” under a variety of physical and systemic circumstances. By distributing storage and computation across many servers, the combined storage resource can grow linearly at every amount of storage.

**Hadoop cluster**

A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amounts of unstructured data in a distributed computing environment.

Such clusters run Hadoop's open source distributed processing software on low-cost commodity computers. Typically one machine in the cluster is designated as the NameNode and another machine as JobTracker; these are the masters. The rest of the machines in the cluster act as both DataNode and TaskTracker; these are the slaves. Hadoop clusters are often referred to as "shared nothing" systems because the only thing that is shared between nodes is the network that connects them.

Hadoop clusters are known for boosting the speed of data analysis applications. They also are highly scalable: If a cluster's processing power is overwhelmed by growing volumes of data, additional cluster nodes can be added to increase throughput. Hadoop clusters also are highly resistant to failure because each piece of data is copied onto other cluster nodes, which ensures that the data is not lost if one node fails.

**Benefits of Hadoop Cluster:**

\* Hadoop clusters are ideally suited to analyze the data.

\* Hadoop works by breaking the data into pieces. Each pieces of data is assigned to a cluster, so that the data will be no loss.

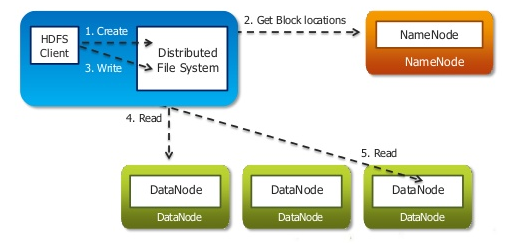
\* The data does not have to be uniform because each piece of data is being handled by a separate process on a separate cluster node.

\* scalability.

\* Hadoop clusters are cost effective solution.

**HDFS Blocks**

Hadoop distributed file system also stores the data in terms of blocks. However the block size in HDFS is very large. The default size of HDFS block is 64MB. The files are split into 64MB blocks and then stored into the hadoop filesystem. The hadoop application is responsible for distributing the data blocks across multiple nodes.



**The benefits with HDFS block are:**

\* The blocks are of fixed size, so it is very easy to calculate the number of blocks that can be stored on a disk.

\* HDFS block concept simplifies the storage of the datanodes. The datanodes doesn’t need to concern about the blocks metadata data like file permissions etc. The namenode maintains the metadata of all the blocks.

\* If the size of the file is less than the HDFS block size, then the file does not occupy the complete block storage.

\* As the file is chunked into blocks, it is easy to store a file that is larger than the disk size as the data blocks are distributed and stored on multiple nodes in a hadoop cluster.

\* Blocks are easy to replicate between the datanodes and thus provide fault tolerance and high availability. Hadoop framework replicates each block across multiple nodes (default replication factor is 3). In case of any node failure or block corruption, the same block can be read from another node.