HDFS - Hadoop Distributed File System

One of the major components of the Hadoop ecosystem is the Hadoop Distributed File System, generally known as HDFS (storage) and the other essential part is the processing or compute. The major factors that set it apart from other file systems are the scalability and distribution. Hadoop stores various duplicates of large files over various data nodes, which appears like it is in its default mode. This is an architecture that can work perfectly with two indispensable functions: a) Data is available on various nodes, making it possible for the compute engine to work in parallel across respective nodes and then the results are summarized later, by that means, enabling very large parallelism. b) Another vital advantage is the fault tolerance. Hadoop is known as a system geared towards the processing of very large amount of jobs over various machines, thus, it is crucial that a single lost node does not affect the job to result into failure or data loss. c). This is one of the major ideas to comprehend in a distributed environment. The loss of just one node ought not to negatively affect the operation of the system or result into making a long-running task become a failure. HDFS likewise provides some commands, just like Unix, for the manipulation of file system.

Example:

```
hdfs dfs -ls /home/hellobigdatauser/datafile1
hdfs dfs -mkdir hellobigdata
```

For every Hadoop cluster, there is a single master as well as a group of data nodes clustered together. A master node has the regular components for the cluster just like job tracker, metastore, software, etc. and there will be some data nodes also known as the worker nodes – they are the real data which does the required computations. There is the availability of other distributed file systems which provide the same function, such as MapR-FS – a MapR distributed file system and Amazon S3.

The Hadoop project includes these modules:

- Hadoop Common: The common utilities that support the other Hadoop modules.
- Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.

http://hadoop.apache.org/