Q No.2

What is HDFS? Highlight its features. Explain about its architecture along with appropriate diagram.

Ans:

- A Distributed File System (DFS), is a file system that is distributed on multiple file servers or multiple locations.
- HDFS: HDFS is a distributed file system that handles large data sets running on commodity hardware. It is used to scale up the Apache Hadoop cluster into multiple nodes. HDFS is the major components of Apache hadoop, while other components being MapReduce and YARN. In other words , HDFS is a decentralized file system that stores data across multiple computers in a cluster.
- HDFS is based on GFS (google file system) . It stores data reliably even in the case of hardware failure.

Features of HDFS:

1. Fault Tolerance: when any machine in the cluster goes down, then a client can easily access their data from the other machine which contains the same copy of data blocks.
2. High Availability: During node failures, a user can easily access their data from the other nodes. This is possible because of replication.
3. High Reliability: HDFS by default creates 3 replicas of each block containing data in nodes. So, data is quickly available to the users; which makes sure that the user does not face the problem of data loss.
4. Replication: HDFS makes replicas of data nodes and maintains the process of replication in regular interval of time.
5. Scalability: As it stores data in multiple clusters, we can simply add/ remove clusters for upscaling/ downscaling; called horizontal scaling. It also supports vertical scaling too.
6. Distributed Storage: HDFS stores data in distributed environment.

HDFS Architecture:

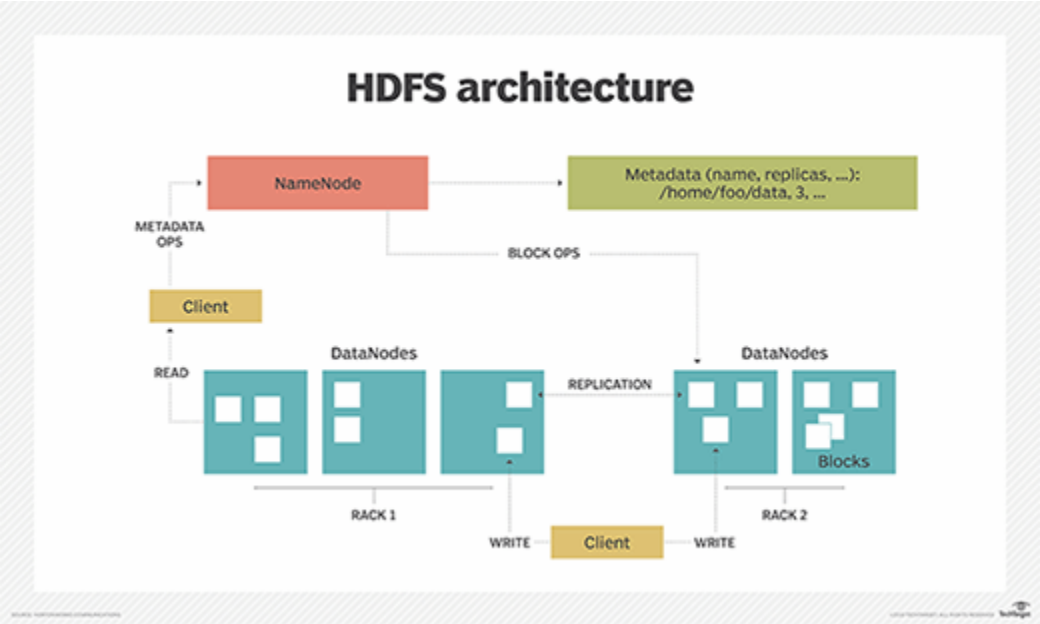

Image Source: techtarget

- HDFS cluster's NameNode is the primary server that manages the file system namespace and controls client access to files.
- NameNode maintains and manages the file system namespace and provides clients with the right access permissions.
- DataNodes manage the storage that's attached to the nodes they run on.
- NameNode performs file system namespace operations, including opening, closing and renaming files and directories.
- DataNodes serve read and write requests from the clients of the file system.The file system namespace hierarchy is like most other file systems -- a user can create, remove, rename or move files from one directory to another.
- NameNode stores the number of copies of a file, called the replication factor of that file.