Roll No. 171347

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

Ans.

HDFS stands for Hadoop Distributed File System. It is really hard to maintain huge volumes of data in a single machine. Thus, it is essential to break down the data into smaller chunks and store it on multiple machines.

Distributed file systems are filesystems that control storage over a network of machines.

Hadoop's storage component is called Hadoop Distributed File System (HDFS). Distributed over a cluster of computers, Hadoop stores all of its data. However, it has a few characteristics that make it what it is.

a. **Huge volumes** – Being a distributed file system, it is highly capable of storing petabytes of data without any glitches.
b. **Data access** – It is based on the philosophy that "the most effective data processing pattern is write-once, the read-many-times pattern".
c. **Cost-effective** – HDFS runs on a cluster of commodity hardware. These are inexpensive machines that can be bought from any vendor.

Following are the highlighted features of HDFS:

1. **Fault Tolerance**

The ability of a system to function under challenging circumstances is known as fault tolerance in Hadoop HDFS. It can withstand many faults. Blocks of data are created using the Hadoop framework. Create several copies of the blocks on other cluster machines after that.

Therefore, a client may quickly retrieve their data from another machine that has the identical copy of the data blocks when any node in the cluster goes down.

2. **High Availability**

A highly accessible file system is Hadoop HDFS. By making a replica of the blocks on the other slave nodes in the HDFS cluster, data is replicated among the nodes in the Hadoop cluster. Consequently, whenever a user needs to retrieve this data, they can do so via the slaves that store its blocks.

A user can readily retrieve their data from other nodes in the event of undesirable circumstances, such as a node failure. because the other nodes in the HDFS cluster have duplicate copies of the blocks.

3. **High Reliability**

HDFS offers trustworthy data storage. It has a data storage capacity of 100s of petabytes. On a cluster, HDFS dependably stores data. The data is divided into blocks by it. These blocks are stored by the Hadoop framework on the HDFS cluster's nodes.

By making a replica of each and every block in the cluster, HDFS stores data with reliability. As a result, it offers fault tolerance. A user can quickly access the data from the other nodes in the cluster if the node that contains the data in the cluster goes down.

By default, each block containing data present on the nodes in HDFS is replicated three times. Data is therefore rapidly made available to users. As a result, the user is not concerned about data loss. As a result, HDFS is quite dependable.

4. **Replication**

Data Replication is unique features of HDFS. Replication solves the problem of data loss in an unfavorable condition like hardware failure, crashing of nodes etc. HDFS maintain the process of replication at regular interval of time.

HDFS also keeps creating replicas of user data on different machine present in the cluster. So, when any node goes down, the user can access the data from other machines. Thus, there is no possibility of losing of user data.

5. **Scalability**

Hadoop HDFS stores data on multiple nodes in the cluster. So, whenever requirements increase you can scale the cluster. Two scalability mechanisms are available in HDFS: Vertical and Horizontal Scalability.

6. **Distributed Storage**

Distributed storage and replication are used to implement all of HDFS's functionalities. Data is spread among the nodes and stored by HDFS. Data in Hadoop is organized into blocks and stored on the HDFS cluster's nodes.

Then HDFS creates a copy of every block and stores it on other nodes. We can quickly access our data from the other nodes that house its replica when one machine in the cluster crashes.

Architecture of HDFS:

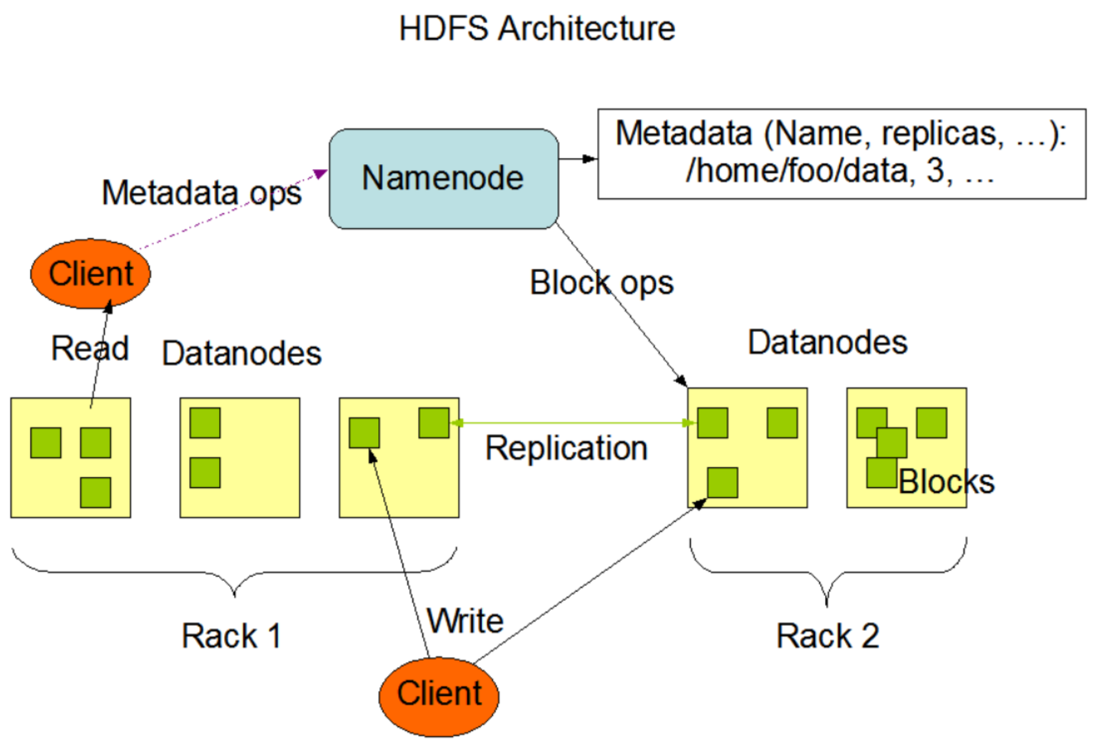
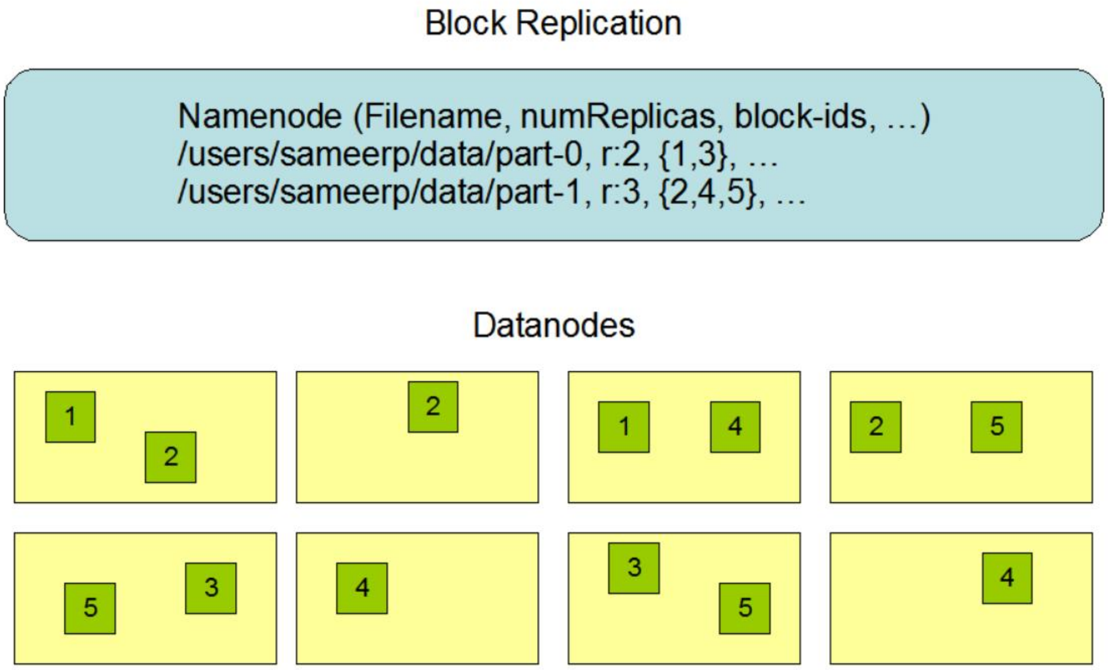


Figure 1. HDFS Architecture


Block Replication
Namenode (Filename, numReplicas, block-ids, …)
/users/sameerp/data/part-0, r:2, {1,3}, …
/users/sameerp/data/part-1, r:3, {2,4,5}, …
Datanodes
1
2
2
1
4
2
5
5
3
4
3
5
4

Datanodes in HDFS

Datanodes are the worker nodes. They are inexpensive commodity hardware that can be easily added to the cluster.

Datanodes are responsible for storing, retrieving, replicating, deletion, etc. of blocks when asked by the Namenode.

They periodically send heartbeats to the Namenode so that it is aware of their health. With that, a DataNode also sends a list of blocks that are stored on it so that the Namenode can maintain the mapping of blocks to Datanodes in its memory.

But in addition to these two types of nodes in the cluster, there is also another node called the Secondary Namenode. Let’s look at what that is.