Explain the process of data storage in Hadoop Distributed File System
(HDFS) with the help of a suitable example.

In Previous Years Questions

HDFS is a distributed file system designed to store and manage large data sets across a
cluster of machines.

It adopts a simple but effective approach to data storage:

1. Data Splitting

- Large files are broken down into fixed-size blocks, typically 64MB or 128MB.
- This partitioning enables parallel processing, where each block can be processed
  independently across different nodes in the cluster.

2. Block Replication

- Each data block is replicated across multiple nodes in the cluster, ensuring data
  availability even if one node fails.
- Replication factor is configurable, allowing for a balance between data redundancy and
  storage efficiency.

3. Metadata Management

- The NameNode acts as the central authority, storing metadata about all files and
  blocks in the system.
- This metadata includes block locations, replication factors, and file permissions.
- The DataNodes store the actual data blocks and report their health status to the
  NameNode.

Explain the process of data storage in Hadoop Distributed File System
(HDFS) with the help of a suitable example.

4. Data Read and Write Operations

- Clients interact with the NameNode to locate the desired data blocks.
- The NameNode directs the client to the DataNodes where the blocks are located.
- Clients can then read or write data directly to the DataNodes.

Example

Imagine you want to store a 1GB file containing weather data in HDFS.

The process would be as follows:

1. File Splitting: The file is split into 16 blocks of 64MB each.
2. Block Replication: Each block is replicated 3 times across different DataNodes in the cluster.
3. Metadata Management: The NameNode stores the information about the file, including the block locations and replication factors.
4. Data Storage: Each DataNode stores three copies of each block, resulting in a total of 48 blocks stored across the cluster.