In Previous Years Questions

Hadoop Distributed File System (HDFS) is a special file system used to store and manage large amounts of data across a group of computers. It is important for working with Big Data.

Key Characteristics and Features of HDFS

1. Distributed Architecture: HDFS distributes data across multiple nodes in a cluster, allowing it to scale horizontally as the data volume grows. Each node (also known as a DataNode) is a commodity hardware machine that contributes storage capacity to the file system.
2. Replication and Fault Tolerance: To ensure data availability and fault tolerance, HDFS replicates data blocks across different DataNodes. By default, HDFS replicates each data block three times, storing each copy on different machines. If one DataNode fails, the replicas on other nodes can be used to serve data, providing fault tolerance.
3. Data Blocks: HDFS stores files in large blocks, typically ranging from 64 MB to 128 MB in size. These large block sizes optimize data processing, as it minimizes the overhead of seeking and reading small blocks.
4. Write Once, Read Many (WORM): HDFS follows the Write Once, Read Many principle, which means that data is written to HDFS once and becomes immutable (read-only) afterward. This design simplifies data management and enhances data reliability.
5. Data Locality: HDFS aims to process data where it resides, known as data locality. When a task (such as a MapReduce job) is scheduled on a node, HDFS tries to schedule the task on the node where the data is located. This reduces data movement and improves processing efficiency.
6. Namespace and Metadata: HDFS uses a single namespace to represent the file and directory hierarchy. Metadata about files (e.g., file name, file size, replication factor) is stored in a separate component called the NameNode.

7. Secondary NameNode: The Secondary NameNode is not a backup or failover NameNode but assists the primary NameNode by periodically merging the changes to the file system’s metadata. This helps the system recover from failures more quickly.

Suggested Readings:

1. Explain Hadoop architecture and its components with proper diagram ?
2. Explain the process of data storage in Hadoop Distributed File System (HDFS) with the help of a suitable example.
3. Write down the goals of HDFS ?
4. Hadoop’s Parallel World