Q1.> Explain Hadoop Daemon's.

ANS. In the Hadoop Ecosystem, a daemon is a background process that runs continuously to perform specific functions related to data processing, storage, and Management. There are several key Hadoop daemons that are essential for the functioning of a Hadoop cluster.

1. NameNode.

The Master daemon for HDFS (Hadoop Distributed File System) It manages the metadata and file directory structure of the system. It keeps track of which data blocks are stored on which machines and the overall structure of the file system, but it does not store actual data itself. Responsible for overall health for HDFS.

2. DataNode

The worker daemon for HDFS. It is responsible for actually storing and retrieving the data blocks on the local data blocks of the machine. DataNodes manage the data storage on individual nodes and report the status of the stored data to the NameNode periodically.

3. Resource Manager.

The Master daemon for the YARN (Yet Another Resource Negotiator) framework, responsible for managing resources and scheduling tasks across the cluster. It decides where and when to execute the job based on resource availability.

4. Node Manager

A slave daemon for YARN that runs on each node in the cluster. It is responsible for managing the individual node resources, like memory and CPU, and reporting

this information back to the Resource Manager. It also overseas the execution of containers (tasks) on that Node.

5. JobHistoryServer
Tracks the History of completed MapReduce Jobs. It allows users to check the status and logs of Past Jobs.

6. Secondary NameNode:
A Helper daemon that periodically merges the NameNode's metadata with the transaction log to prevent the NameNode for running out of memory It does not serve as a backup for the NameNode but helps in reducing the workload on it.

Q2) Explain HDFS.

ANS. HDFS stands for Hadoop Distributed File System.
. It is a storage layer of the Hadoop Ecosystem.
. It is designed to store Large amounts of data across multiple machines in a distributed Fashion.
The key Features of HDFS are:-

1. Distributed Architecture.
HDFS is designed to run on commodity hardware, using a cluster of machines (DataNodes) to store data in a distributed manner.

2. Fault Tolerance.
Data is split into large blocks (typically 128 MB or 256 MB), and each block is replicated across multiple DataNodes to ensure fault Tolerance. If one DataNode Fails, another replica can serve the data.

3. Master-Slave Model
HDFS uses a Master slave Architecture where the NameNode is the master and Data Nodes are the slaves.

4. High Throughput.
HDFS is optimized For High throughput data access, especially for Large files, as opposed to low latency access.

5. Write Once, Read Many.
Files in HDFS are generally written once and read many times, making it efficient for Applications like data Analytics and Batch processing.

6. Block-level Storage.
Files are broken into fixed size blocks, and these are blocks are distributed across various DataNodes.

Q3) Explain MapReduce Framework.

ANS. MapReduce is a programming Model and processing framework in Hadoop that allows For the parallel processing of Lorge datasets across a distributed cluster. It has two main stages.

1. Map
2. Reduce.

2. Map
In the Map phase, the input data is divided into smaller chunks, which are processed in parallel by individual mappers. Each Mapper processes a portion of data, performing Operations like Filtering, transForming or extracting useful information from the Input.
The Output of each mapper is a set of key-value pairs, which are then passed to the next stage (Reduce)

2 Shuffle and shrt:-

After the map phase the intermediate key value pairs are shuffled and sorted based on the key. This ensures that all values for the same key are grouped together and sent to the same reducer.

3 Reduce:-

In Reduce phase, the grouped key-value pairs are processed by reducers. The reducer aggregates, Filters or further processes the Intermediate results. 1a
The output from the Reducers is the final Result which is Typically written to HDFS.

⟹ Key-Features:-

i. Parallel Processing.
ii. Fault Tolerance.
iii. Scalability.
iv. Data locality.

or

MapReduce jobs can be written in Java, Python, or other programming Languages, and they are executed by the Hadoop Framework on a distributed Cluster. 1