02 HDFS Interview Questions -KirkYagami

1) What is the difference between HDFS and GFS?

HDFS (Hadoop Distributed File System)	GFS (Google File System)
Default block size in HDFS is 128 MB.	Default block size in GFS is 64 MB.
Only data append operation is possible in HDFS.	GFS allows random file writes.
Data is represented in blocks.	Data is represented in chunks.
HDFS has the edit log and journal. (JournalNode service in HA configuration)	GFS has the operation log.
Works on Single Write and Multiple Read Model.	Works on Multiple Write and Multiple Read Model.

2) How will you measure HDFS space consumed?

The two popular utilities or commands to measure HDFS space consumed are `hdfs dfs -du` and `hdfs dfsadmin -report`. HDFS provides reliable storage by copying data to multiple nodes. The number of copies it creates is usually referred to as the replication factor, which is greater than one.

- `hdfs dfs -du` - This command shows the space consumed by data without replications.
- `hdfs dfsadmin -report` - This command shows the real disk usage by considering data replication. Therefore, the output of `hdfs dfsadmin -report` will always be greater than the output of the `hdfs dfs -du` command.

3) Is it a good practice to use HDFS for multiple small files?

No, it is not a good practice to use HDFS for multiple small files because the NameNode, which stores metadata, is an expensive high-performance system. Occupying the NameNode space with unnecessary metadata generated for each small file is inefficient. HDFS is optimized for large files, where it can efficiently manage space and provide better performance.

4) I have a file "Sample" on HDFS. How can I copy this file to the local file system?

This can be accomplished using the following command:

```
bin/hadoop fs -copyToLocal /hdfs/source/path /localfs/destination/path
```

5) Replication causes data redundancy then why is it still preferred in HDFS?

Replication is preferred in HDFS despite data redundancy because Hadoop works on commodity hardware, which has a higher probability of failure. Replication ensures high availability and fault tolerance. Data in HDFS is typically stored in at least three locations. If one copy becomes corrupted and another is unavailable, the third copy ensures data accessibility without loss.

6) Data is replicated at least thrice on HDFS. Does it imply that any alterations or calculations done on one copy of the data will be reflected in the other two copies also?

No, calculations or transformations are performed on the primary replica of the data, and changes are not automatically reflected in the other replicas. The master node identifies where the original data is located and performs the calculations. Replication in HDFS is for redundancy and fault tolerance, not for real-time synchronization of data changes across replicas.

7) What do you understand by Safe Mode in Hadoop?

Safe Mode in Hadoop is a state where the NameNode does not perform replication or deletion of blocks. In this mode, the NameNode only collects block reports information from the DataNodes. Safe Mode is typically used during startup to ensure that a sufficient number of replicas are available before allowing the system to operate normally.

8) If a DataNode is marked as decommissioned, can it be chosen for replica placement?

No, when a DataNode is marked as decommissioned, it cannot be considered for replica placement. However, it continues to serve read requests until it is fully decommissioned, meaning until all the blocks on that DataNode are successfully replicated to other nodes.

9) What is the role of the NameNode in HDFS?

The NameNode manages the metadata and the directory structure of HDFS. It maintains the file system namespace and the mapping of data blocks to DataNodes. The NameNode is a critical component, as it is the single point of failure in HDFS.

10) What happens if the NameNode fails?

If the NameNode fails, the HDFS becomes inaccessible as it is the central repository of all the metadata. However, HDFS provides mechanisms such as Secondary NameNode, Checkpoint Node, and Backup Node to recover from such failures. In production environments, High Availability (HA) configurations are often used to avoid downtime due to NameNode failure.

11. What is the role of the Secondary NameNode?

- **Answer:** The Secondary NameNode is not a backup for the NameNode. Its primary role is to periodically merge the NameNode's transaction log (edit log) with the file system's image (FsImage) to prevent the NameNode from running out of memory. It helps in reducing the load on the NameNode.

12. How does HDFS handle large files?

- **Answer:** HDFS is designed to handle large files efficiently by splitting them into blocks and storing each block on different DataNodes. This allows for parallel processing of large datasets, improving throughput and fault tolerance.

13. What are the differences between HDFS and traditional file systems like ext4 or NTFS?

- **Answer:** Key differences include:
  - **Scalability:** HDFS is designed to scale across thousands of nodes, while traditional file systems are limited to single machines.
  - **Fault Tolerance:** HDFS replicates data across multiple nodes, whereas traditional file systems typically do not have built-in replication.
  - **Write-once-read-many:** HDFS is optimized for batch processing with a write-once-read-many access model, while traditional file systems allow for more flexible read/write operations.

14. What is a heartbeat in HDFS, and what role does it play in the functioning of the cluster?

- **Answer:** A heartbeat is a signal sent by DataNodes to the NameNode at regular intervals to indicate that they are functioning properly. The heartbeat mechanism allows the NameNode to monitor the health and status of DataNodes. If a DataNode fails to send a heartbeat within a specified time, the NameNode marks it as dead,

initiates the replication of its blocks to other DataNodes, and ensures that data remains available and fault-tolerant in the cluster.

HDFS Commands

1) How do you list all the files in a directory on HDFS?

- **Answer:** You can list all the files in a directory on HDFS using the command:

```
hdfs dfs -ls /path/to/directory
```

2) How can you create a new directory on HDFS?

- **Answer:** To create a new directory on HDFS, use the following command:

```
hdfs dfs -mkdir /path/to/new/directory
```

3) What command would you use to upload a file from the local file system to HDFS?

- **Answer:** You can upload a file from the local file system to HDFS using the command:

```
hdfs dfs -put /localfs/path/to/file /hdfs/path/to/destination
```

4) How do you download a file from HDFS to your local file system?

- **Answer:** To download a file from HDFS to your local file system, use:

```
hdfs dfs -get /hdfs/path/to/file /localfs/path/to/destination
```

5) How can you check the disk usage of a directory in HDFS?

- **Answer:** You can check the disk usage of a directory in HDFS with:

```
hdfs dfs -du -h /path/to/directory
```

6) Which command is used to display the contents of a file stored in HDFS?

- **Answer:** To display the contents of a file in HDFS, you can use:

```
hdfs dfs -cat /hdfs/path/to/file
```

7) How can you move a file within HDFS from one directory to another?

- **Answer:** You can move a file from one directory to another within HDFS using:

```
hdfs dfs -mv /hdfs/source/path/file /hdfs/destination/path/
```

8) What command would you use to delete a file or directory in HDFS?

- **Answer:** To delete a file or directory in HDFS, use:

```
hdfs dfs -rm /hdfs/path/to/file
```

or

```
hdfs dfs -rm -r /hdfs/path/to/directory
```

9) How do you set replication factor for a file in HDFS?

- **Answer:** To set the replication factor for a file in HDFS, use:

```
hdfs dfs -setrep -w 3 /hdfs/path/to/file
```

Here, `3` is the replication factor.

10) Which command would you use to find the block locations of a file in HDFS?

- **Answer:** To find the block locations of a file in HDFS, you can use:

```
hdfs fsck /hdfs/path/to/file -files -blocks -locations
```

11) How do you check the file permissions in HDFS?

- **Answer:** You can check file permissions in HDFS using the command:

```
hdfs dfs -ls -l /path/to/file
```

12) What command would you use to change the owner of a file or directory in HDFS?

- **Answer:** To change the owner of a file or directory in HDFS, use:

```
hdfs dfs -chown newowner:newgroup /path/to/file_or_directory
```

13) How can you copy a file within HDFS?

- **Answer:** You can copy a file within HDFS using the command:

```
hdfs dfs -cp /hdfs/source/path/file /hdfs/destination/path/
```

14) What command would you use to display the last few lines of a file in HDFS?

- **Answer:** To display the last few lines of a file in HDFS, use:

```
hdfs dfs -tail /hdfs/path/to/file
```

15) How do you check the health of HDFS?

- **Answer:** You can check the health of HDFS using the command:

```
hdfs fsck /
```

16) What command would you use to display the filesystem usage statistics?

- **Answer:** To display the filesystem usage statistics, use:

```
hdfs dfs -df -h
```