HDFS Interview Questions - HDFS

1. What are the different vendor-specific distributions of Hadoop?

The different vendor-specific distributions of Hadoop are Cloudera, MAPR, Amazon EMR, Microsoft Azure, IBM InfoSphere, and Hortonworks (Cloudera).

2. What are the different Hadoop configuration files?

The different Hadoop configuration files include:

* hadoop-env.sh
* mapred-site.xml
* core-site.xml
* yarn-site.xml
* hdfs-site.xml
* Master and Slaves

4. What are the differences between regular FileSystem and HDFS?

1. Regular FileSystem: In regular FileSystem, data is maintained in a single system. If the machine crashes, data recovery is challenging due to low fault tolerance. Seek time is more and hence it takes more time to process the data.
2. HDFS: Data is distributed and maintained on multiple systems. If a DataNode crashes, data can still be recovered from other nodes in the cluster. Time taken to read data is comparatively more, as there is local data read to the disc and coordination of data from multiple systems.

5. Why is HDFS fault-tolerant?

HDFS is fault-tolerant because it replicates data on different DataNodes. By default, a block of data is replicated on three DataNodes. The data blocks are stored in different DataNodes. If one node crashes, the data can still be retrieved from other DataNodes.

### 9. If you have an input file of 350 MB, how many input splits would HDFS create and what would be the size of each input split?

By default, each block in HDFS is divided into 128 MB. The size of all the blocks, except the last block, will be 128 MB. For an input file of 350 MB, there are three input splits in total. The size of each split is 128 MB, 128MB, and 94 MB.

### 10.  Which command will help you find the status of blocks and FileSystem health?

To check the status of the blocks, use the command:

**hdfs fsck <path> -files -blocks**

To check the health status of FileSystem, use the command:

**hdfs fsck / -files –blocks –locations > dfs-fsck.log**

### 11. How can you restart NameNode and all the daemons in Hadoop?

The following commands will help you restart NameNode and all the daemons:

You can stop the NameNode with **./sbin /Hadoop-daemon.sh stop NameNode** command and then start the NameNode using **./sbin/Hadoop-daemon.sh start NameNode**command.

You can stop all the daemons with **./sbin /stop-all.sh** command and then start the daemons using the **./sbin/start-all.sh**command.

### 14. How do you copy data from the local system onto HDFS?

The following command will copy data from the local file system onto HDFS:

**hadoop fs –copyFromLocal [source] [destination]**

Example: **hadoop fs –copyFromLocal /tmp/data.csv /user/test/data.csv**

In the above syntax, the source is the local path and destination is the HDFS path. Copy from the local system using a -f option (flag option), which allows you to write the same file or a new file to HDFS.

### 15. When do you use the dfsadmin -refreshNodes and rmadmin -refreshNodes commands?

The commands below are used to refresh the node information while commissioning, or when the decommissioning of nodes is completed.

**dfsadmin -refreshNodes**

This is used to run the HDFS client and it refreshes node configuration for the NameNode.

**rmadmin -refreshNodes**

This is used to perform administrative tasks for ResourceManager.

### 16. Is there any way to change the replication of files on HDFS after they are already written to HDFS?

Yes, the following are ways to change the replication of files on HDFS:

We can change the dfs.replication value to a particular number in the **$HADOOP\_HOME/conf/hadoop-site.xml**file, which will start replicating to the factor of that number for any new content that comes in.

If you want to change the replication factor for a particular file or directory, use:

**$HADOOP\_HOME/bin/Hadoop dfs –setrep –w4 /path of the file**

**Example: $HADOOP\_HOME/bin/Hadoop dfs –setrep –w4 /user/temp/test.csv**

17. Name some Hadoop-specific data types that are used in a MapReduce program.

This is an important question, as you would need to know the different data types if you are getting into the field of Big Data.

For every data type in Java, you have an equivalent in Hadoop. Therefore, the following are some Hadoop-specific data types that you could use in your MapReduce program:

* IntWritable
* FloatWritable
* LongWritable
* DoubleWritable
* BooleanWritable
* ArrayWritable
* MapWritable
* ObjectWritable

34. Which of the following has replaced JobTracker from MapReduce v1?

1. NodeManager
2. ApplicationManager
3. ResourceManager
4. Scheduler

The answer is ResourceManager. It is the name of the master process in Hadoop v2.

19. Write the YARN commands to check the status of an application and kill an application.

The commands are as follows:

a) To check the status of an application:

yarn application -status ApplicationID

b) To kill or terminate an application:

yarn application -kill ApplicationID