Hadoop Interview Questions and Answers

Introduction

Apache Hadoop skills remain the foundation of big data engineering, as it serves as the fundamental framework for distributed storage and processing. This Hadoop interview questions and answers examines the key components of Hadoop, including HDFS, YARN, and MapReduce, as well as the integration of Hadoop with contemporary tools such as Spark and Hive. The ability to clearly demonstrate knowledge of fault tolerance and data locality is crucial to succeeding in technical interviews.

Are you ready to conquer the Big Data world? Get practical experience by registering for our industry-relevant Hadoop Certification Course in Chennai today.

List of Hadoop Interview Questions for Freshers

1. What makes Hadoop a tool for big data analytics?
2. What is the command to launch every Hadoop daemon simultaneously?
3. Which input formats are most frequently used with Hadoop?
4. List the most widely used data management applications for Hadoop Edge Nodes.

5. What kind of file formats are compatible with Hadoop?
6. List the many operating modes for Hadoop.
7. Define NAS
8. Explain Hadoop streaming.
9. Define Mapper
10. What can be done with the ‘jps’ command?
11. What does Hadoop’s Avro serialization mean?
12. What is HDFS, and what parts make it up?

Get started with our Hadoop course syllabus.

Hadoop Interview Questions and Answers for Freshers

1. What makes Hadoop a tool for big data analytics?

The open-source Hadoop framework, written in Java, handles large amounts of data processing on a cluster of inexpensive hardware. Additionally, it permits the execution of numerous exploratory data analysis activities on entire datasets without sampling.

The following characteristics of Hadoop make it a necessary prerequisite for Big Data:

- Massive data collection
- Exceptional data storage
- Data processing
- Independent

2. What is the command to launch every Hadoop daemon simultaneously?

The following command launches each Hadoop daemon simultaneously:

./sbin/start-all.sh

3. Which input formats are most frequently used with Hadoop?

In Hadoop, the most commonly used input formats are:

- Key-value input structure
- Input format for sequence files
- Format for text input

4. List the most widely used data management applications for Hadoop Edge Nodes.

The most popular data management programs that are compatible with Hadoop's Edge Nodes are Plume, Oozie, Ambari, and Pig.

5. What kind of file formats are compatible with Hadoop?

The following file types are utilized with Hadoop:

- JSON
- Columnar
- Sequence files
- CSV format
- Parquet file

6. List the many operating modes for Hadoop.

There are three ways in which Hadoop can operate:

- Standalone mode
- Pseudo Distributed mode (Single node cluster)

- Fully distributed mode (Multiple node cluster)

7. Define NAS.

Network-attached storage (NAS) is often shortened to NAS. It is a computer data storage server that stores files at the file level and is network-connected. It provides a diverse group with access to data.

8. Explain Hadoop streaming.

A user can construct and execute Map/Reduce tasks using any executable, script, or programming language, such as **Python**, Perl, Ruby, etc., using Hadoop Streaming, a generic API. The newest tool for Hadoop streaming is called Spark.

9. Define Mapper.

The initial piece of code that migrates or manipulates HDFS block-stored data into key-value pairs is called the mapper. On HDFS, there is a single mapper for each data block.

10. What can be done with the ‘jps’ command?

We may verify whether the Hadoop daemons, such as name node, data node, resource manager, node manager, etc., are operating on the system by using the ‘jps’ command.

11. What does Hadoop’s Avro serialization mean?

In Hadoop, the process of translating object or data structure states into binary or textual representation is called Avro serialization. This is done to move the data across a network or save it on a permanent storage device. Avro deserialization is referred to as unmarshalling, whereas Avro serialization is known as marshaling.

12. What is HDFS, and what parts make it up?

The Hadoop Distributed File System, or HDFS, is extremely fault-tolerant and operates on commodity hardware. HDFS is appropriate for distributed processing and storage since it offers file permissions and authentication. The name node, the data node, and the secondary node are its three constituent parts.

List of Hadoop Interview Questions for Experienced

1. Explain “Name Nodes” that are active and passive.
2. Why would one use the commands dfsadmin -refreshNodes and rmadmin -refreshNodes?
3. When copying data from the local system to HDFS, which command will you use?
4. What commands will you use to ascertain the health of the FileSystem and the status of the blocks?
5. List the main setup parameters that a MapReduce program needs.
6. What are the various parts of a hive architecture?
7. What are the main elements of HBase?
8. Which tombstone markers in HBase are available for deletion?

Explore what the Hadoop salary is for freshers and experienced professionals.

Hadoop Technical Interview Questions and Answers for Experienced

1. Explain “Name Nodes” that are active and passive.

All of the data nodes’ metadata is kept up to date by a name node. In a High Availability (HA) architecture, there are two Name Nodes: the Active Name Node and the Passive or Standby Name Node.

While the Passive Name Node is a standby Name Node with data that is comparable to that of the Active Name Node, the Active Name Node functions and operates within the cluster.

The cluster’s passive Name Node will take over as the active Name Node if the active Name Node fails. As a result, the cluster never fails and never lacks a Name Node.

2. Why would one use the commands dfsadmin -refreshNodes and rmadmin -refreshNodes?

The commands dfsadmin and rmadmin -refreshNodes are used for:

- The HDFS client is executed using the dfsadmin –refreshNodes command. It updates the NameNode's node settings.
- ResourceManager administration is done with the rmadmin –refreshNodes command.

3. When copying data from the local system to HDFS, which command will you use?

To copy data from the local system onto HDFS, use the following command:

- The file will be copied to the HDFS from the local file system using the *Hadoop copyFromLocal* command.
- Format: hadoop fs –copyFromLocal [source] [destination]

4. What commands will you use to ascertain the health of the FileSystem and the status of the blocks?

The command to verify the blocks' status is as follows:

hdfs fsck -files -blocks

To examine the FileSystem's health, run the following command: hdfs fsck / -files –blocks –locations > dfs-fsck.log.

5. List the main setup parameters that a MapReduce program needs.

The primary configuration parameters in a MapReduce program are as follows:

- Enter the jobs' locations in HDFS.
- The jobs' output location in HDFS
- The data's input format
- The data's output format
- Classes with a map function in them
- Classes with a reduction function in them

6. What are the various parts of a hive architecture?

The various parts of the Hive architecture are:

- ***User Interface:*** It provides a means of communication between the user and the colony. It allows users to ask the system questions. To construct an execution plan for the query, the user interface first creates a session handle and sends it to the compiler.
- ***Compiler:*** Produces the plan of execution.
- ***Execute Engine:*** To perform the query, it functions as a bridge between Hadoop and Hive.
- ***Metastore:*** Upon receiving a request to submit metadata, it stores the metadata and forwards it to the compiler so that the query can be executed.

7. What are the main elements of HBase?

The major elements of HBase comprise:

- ***Region server:*** Based on their key values, the HBase tables are arranged into regions that are separated horizontally. As worker nodes, each region server handles client read, write, update, and delete requests.
- ***HMaster:*** For load balancing, it gives RegionServers regions. HMaster watches over the Hadoop cluster. When a client wants to modify the metadata operations and schema, it is utilized.
- ***ZooKeeper:*** To keep the cluster's servers in good condition, it provides a distributed coordination service. It notifies users of server failures and indicates which servers are up and running.

8. Which tombstone markers in HBase are available for deletion?

The three kinds of tombstone markers that can be removed from HBase are:

- Family Delete Marker: Indicates every column in the family.

- Version Delete Marker: Identifies a single-column version that should be removed.
- Column Delete Marker: Identifies every iteration of a certain column.

Conclusion

Passing a Hadoop interview means having an in-depth understanding of the HDFS storage system, MapReduce processing, and YARN resource management. Employers seek data engineers who can optimize cluster performance, data locality, and fault tolerance. Your ability to describe how Hadoop works in a Spark, Hive, and Kafka ecosystem will demonstrate your worth in designing scalable big data solutions. As companies shift to hybrid cloud infrastructures, your expertise in handling large amounts of data in distributed systems will make you stand out as a senior big data professional.
Ready to conquer the Big Data world? Learn distributed computing on Hadoop & Big Data Engineering in our software training institute in Chennai.