Share on your Social Media

Top 20+ Big Data Hadoop Interview Questions and Answers

Published On: May 24, 2024

Big Data Hadoop Interview Questions and Answers

One of the most widely used frameworks for processing, storing, and analyzing big data is Hadoop. As a result, there is a constant need for specialists in this industry. We offer the frequently asked **Big Data Hadoop interview questions and answers** for succeeding the Hadoop interviews in this blog post.

Download Big Data Hadoop Interview Questions PDF

Featured Articles

Want to know more about becoming an expert in IT?

Click Here to Get Started

100% Placement Assurance

AUTHORISED CERTIFICATION PARTNER IBM

Quick Enquiry

Related Courses at SLA

- Big Data Hadoop Training in OMR
- Big Data Hadoop Training in Chennai

Related Posts

MEAN Stack Interview

Big Data Hadoop Interview Questions and Answers for Freshers

1. What are the various Hadoop distributions tailored to specific vendors?

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

2. Which various Hadoop configuration files are there?

Among the several Hadoop configuration files are:

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

3. Why is HDFS fault-tolerant?

Because HDFS replicates data over several DataNodes, it is resilient to faults. Three DataNodes replicate a block of data by default. There are various DataNodes where the data blocks are kept. The data can still be obtained from other DataNodes if one node crashes.

4. What are the components of Hadoop?

The three parts of Hadoop are as follows:

- Hadoop's resource management component is called **Hadoop YARN.**
- The Hadoop storage unit is called the **Hadoop Distributed File System (HDFS)**.
- **Hadoop MapReduce** is the Hadoop processing engine.

5. Describe Hadoop's fault tolerance.

Questions and Answers
Published On: June 19, 2024

Introduction Since MEAN Stack combines several other applications as part of its functionality, it is...

Top 15 Struts Interview Questions and Answers
Published On: June 18, 2024

Struts Interview Questions and Answers When it comes to developing Java web applications, Struts is...

Top 20 C Sharp Interview Questions and Answers
Published On: June 17, 2024

C Sharp Interview Questions and Answers Microsoft created the general-purpose programming language C# together with...

Top 20 VB.Net

Data is divided into blocks by the Hadoop framework, which then makes multiple copies of each block on different cluster servers. Clients can thus access their data from the other machine that has an identical duplicate of the data blocks, even if one of the cluster's devices fails.

6. How is high availability met in Hadoop?

To duplicate data in an HDFS context, a copy of each block is created. Because duplicate images of blocks are already existing in the other HDFS cluster nodes, users can easily retrieve their data from those nodes.

7. How is high reliability achieved in Hadoop?

HDFS divides the data into blocks, which are then stored on cluster nodes via the Hadoop framework. By creating a duplicate of each block current in the cluster, it preserves data. provides a fault tolerance facility as a result.

By default, it makes three copies of every block that contains data from the nodes. As a result, users can quickly access the data. The user is thereby spared the hassle of data loss. HDFS is hence quite dependable.

8. Why does Hadoop perform replications?

Replication provides a solution to the problem of data loss in the event of unfavorable events, such as device failure or node failures. It oversees the replication process regularly. As a result, the likelihood of losing user data is minimal.

9. What is Hadoop's scalability like?

HDFS stores the information at several nodes. Thus, it can scale the cluster in the event of an increase in demand.

10. Explain NameNode in Hadoop.

The master daemon that manages the master node is called NameNode. It stores the filesystem metadata, which includes file names, information about a file's blocks, block locations, permissions, and so on. It is in charge of the Datanodes.

Download Big Data Hadoop Syllabus PDF

11. Describe DataNode

The slave daemon that manages the slave nodes is called DataNodes. The genuine business data is saved. Based on the NameNode's instructions, it fulfills the client's read and write requests. As it stores the file blocks, NameNode also stores block locations, permissions, and other metadata.

12. What is shuffling in MapReduce?

Shuffling is the process of moving data from the mappers to the crucial reducers in Hadoop MapReduce. This is the procedure by which the system sorts the unstructured data and feeds the reducer's input with the map's output. This is a crucial procedure for reducers.

They wouldn't accept any other information. Furthermore, it helps to save time and finish the process faster because it can start even before the map phase is finished.

13. Give the components of Apache Spark.

The Spark Core Engine, Spark Streaming, GraphX, MLlib, Spark SQL, and Spark R are all parts of Apache Spark.

Any of the other five components listed can be utilized in addition to the Spark Core Engine. Utilizing every Spark component at once is not necessary. One or more of these can be utilized in conjunction with Spark Core, depending on the use case and request.

Big Data Hadoop Salary

Big Data Hadoop Interview Questions and Answers for Experienced Applicants

14. Explain Apache Hive

- Built on top of Hadoop, Hive is an open-source system that aggregates structured data to enable the analysis and querying of big data.
- Additionally, Hive enables **SQL developers** to design statements for data query and analysis in Hive Query Language that are comparable to regular SQL statements.
- It was designed to make programming MapReduce simpler because you don't have to write extensive Java code.

15. Explain Apache Pig

Programs must be transformed into maps and reduced stages to use MapReduce. Because not every data analyst is familiar with MapReduce.

Yahoo researchers created Apache Pig to help close the knowledge gap. On top of **Hadoop**, Apache Pig was developed to provide a high degree of abstraction and free up developers' time to construct intricate MapReduce scripts.

16. Describe Yarn.

- Yarn is an acronym for *Yet Another Negotiator of Resources*. It is Hadoop's layer for resource management. Hadoop 2.x launched the Yarn.
- To execute and process data stored in the Hadoop Distributed File System, Yarn offers a variety of data processing engines, including batch, stream, interactive, and graph processing.
- Yarn provides task scheduling as well. It allows other developing technologies to benefit from HDFS and economic clusters by expanding Hadoop's capabilities.
- The Hadoop 2.x data operating technique is

Apache Yarn. It is made up of three daemons: Application Master, Node Manager, and Resource Manager, the master daemon.

17. Explain Apache ZooKeeper.

- The open-source program Apache Zookeeper facilitates managing a sizable number of hosts. In a distributed system, coordination and management are difficult.
- By automating this process, Zookeeper frees developers from having to worry about the distributed nature of their work and allows them to focus on creating features for the software.
- For distributed apps, Zookeeper aids with the maintenance of configuration information, naming conventions, and group services.
- To prevent the program from executing different protocols on its own, it implements them on the cluster. It offers a unified, cohesive perspective on numerous devices.

18. List the various types of Znode.

Persistent Znodes: In ZooKeeper, the Persistent Znode is the default node. It remains on the Zookeeper server indefinitely unless it is removed by another client.

Ephemeral Znodes: Temporary nodes are known as ephemeral nodes. It is destroyed each time the creator client disconnects from the ZooKeeper server. Let's take an example where client 1 built eznode 1. The eznode1 is deleted when client1 shuts off the ZooKeeper server.

Sequential Znodes: A 10-digit number that is arranged numerically at the end of the name of a sequential Znode. Let's say client 1 created a sznode1. The sznode1 will have the following name in the ZooKeeper server:

sznode0000000001

The next number in the sequence will be displayed on client1 if it generates another sequential znode. The next sequential node is therefore 0000000002.

19. List the attributes of Apache Sqoop.

Robust: It has a lot of strength. It is simple to use and even has community support and contributions.

Full Load: With just one Sqoop command, Sqoop can load the entire table. Additionally, it enables us to load every table in the database with just one Sqoop command.

Incremental Load: It is compatible with the incremental load feature. We may load a portion of the table every time it is updated by using Sqoop.

Parallel import/export: The data is imported and exported using the YARN framework. On top of parallelism, that offers fault tolerance.

Import SQL query results: This feature enables us to import the SQL query's output into the Hadoop Distributed File System.

20. How would you synchronize the HDFS data that Sqoop imports if the source data is periodically updated?

If the source data is updated quickly, incremental parameters are used to synchronize the HDFS data that Sqoop imports.

Even if the table is regularly updated with new entries, we should still utilize incremental import in conjunction with the append option. mostly when a small number of columns' values are checked; only a new row is added if any updated values for those columns are found.

Like incremental import, the origin has a date column that, based on the initial revised column, is reviewed for all the entries that have been modified following the last import. Modern values would be

incorporated.

21. Explain Hadoop's Apache Flume.

For gathering, combining, and transporting massive volumes of streaming data, including record files and events from several references to centralized data storage, Apache Flume is a tool, service, and data intake mechanism.

Flume is a distributed, flexible, and very stable utility. Generally speaking, it is made to copy streaming data—also known as log data—from several web servers to HDFS.

The following elements make up the Apache Flume architecture generally:

- Flume Source
- Flume Channel
- Flume Sink
- Flume Agent
- Flume Event

22. Which file format does Apache Sqoop default to when importing data?

To import data into Sqoop, there are essentially two file formats available:

- Delimited Text File format
- Sequence File Format

Conclusion

Concerning the crucial big data Hadoop interview questions and answers, we hope this blog has been helpful. Enroll in our **Big Data Hadoop Training in Chennai** today to find out more about Hadoop and Big Data.

Big Data Hadoop Online Training

Softlogic Academy

Softlogic Systems

KK Nagar [Corporate Office]

No.10, PT Rajan Salai, K.K. Nagar, Chennai – 600 078.
Landmark: Karnataka Bank Building
Phone: +91 86818 84318
Email: enquiry@softlogicsys.in
Map: Google Maps Link

OMR

No. E1-A10, RTS Food Street
92, Rajiv Gandhi Salai (OMR),
Navalur, Chennai - 600 130.
Landmark: Adj. to AGS Cinemas
Phone: +91 89256 88858
Email: info@softlogicsys.in
Map: Google Maps Link

Courses

- Python
- Software Testing
- Full Stack Developer
- Java
- Power BI
- Clinical SAS
- Data Science
- Embedded
- Cloud Computing

Navigation

- About Us
- Blog Posts
- Careers
- Contact
- Placement Training
- Corporate Training
- Hire With Us
- Job Seekers
- SLA's Recently Placed Students
- Reviews
- Sitemap

Important Links

- Disclaimer
- Privacy Policy
- Terms and Conditions

Social Media Links

Review Sources

- Google
- Trustpilot
- Glassdoor
- Mouthshut
- Sulekha

Hardware and Networking

VBA Macros

Mobile App Development

DevOps

Justdial

Ambitionbox

Indeed

Software Suggest

Sitejabber