Share on your Social Media

Top 20 Big Data Analytics Interview Questions and Answers

Published On: June 3, 2024

Big Data Analytics Interview Questions and Answers

Big data analysts with skills in organizing and storing vast volumes of data usually fill these roles. We go over a few commonly asked **big data interview questions and answers** to increase your confidence.

Request to Download PDF

Big Data Analytics Interview Questions

Related Courses at SLA

- Big Data Analytics Training in OMR
- Big Data Analytics Training in Chennai

Related Posts

and Answers for Freshers

1. What is big data?

Large, complex datasets produced rapidly by machines, organizations, and people make up big data. It contains information gathered from various sources, such as mobile devices, social media, sensors, and more. The five V's are volume, velocity, variety, veracity, and value, which distinguish big data.

2. What does big data mean, and how does it get started? How is it operated?

Big Data Analytics is the term used to describe enormous volumes of data from people and organizations, both structured and unstructured. It comes from devices, sensors, and social media, among other places.

Big data processing and analysis depend heavily on technologies like Spark and Hadoop. It involves more than just data amount; it also involves data complexity and generating speed.

3. Why are companies utilizing big data analytics to gain a competitive edge?

Businesses utilize big data analytics to spot patterns, get strategic insights, and make data-driven choices that improve customer experiences.

Using big data to optimize operations, improve product development, and increase customer interaction gives businesses a competitive advantage.

4. Describe the role that Hadoop technology plays in the analysis of big data.

For several reasons, **Hadoop** technology is extremely important to big data analytics. It offers a

MEAN Stack Interview 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...

scalable and affordable way to handle and process large amounts of data. Hadoop's fundamental function in big data analytics is established by its ability to facilitate parallel processing, guarantee fault tolerance, and disseminate data effectively.

5. What exactly is data modeling, and why is it necessary?

In HDFS, three is the default replication factor. This suggests that to provide fault tolerance, data is saved in triplicate across different cluster nodes. The replication factor can be changed according to particular requirements and cluster setups.

Big Data Analytics Syllabus

6. How is a big data model deployed? Mention the important actions that need to be taken.

Putting a big data model to practical use is known as deployment. Training, testing, validation, and continual observation are among the steps. Make sure that the model adjusts to new data and works well in real-world circumstances.

7. Describe fsck.

A Hadoop utility is called File System Check, or fsck. It assesses the health of the Hadoop Distributed File System (HDFS). It looks for problems, such as corrupted data blocks, and attempts to resolve them.

8. Which of the three operating modes does Hadoop support?

Hadoop functions in three ways:

- When developing and testing on a single workstation without the use of a cluster or distributed file system, local (*standalone*) mode is employed.
- *Pseudo-Distributed Mode*: Generates a test environment akin to a cluster by simulating a

small cluster on a single machine.

- *Fully Distributed Mode*: Hadoop manages real-world workloads on a multi-node cluster and is suitable for production use.

9. Which output formats are available in Hadoop?

- **Text:** For files with plain text, the default.
- **SequenceFile:** A binary key-value pair format.
- **Avro:** A condensed, effective format that facilitates schema development.

10. What does the term "collaborative filtering" mean to you?

The group of technologies known as collaborative filtering forecasts and predicts what products a certain client would prefer. Depending on each person's choices, this filtering is carried out.

Big Data Analytics Salary

Big Data Analytics Interview Questions and Answers for Experienced

11. Which big data processing approaches are there?

Several techniques are used in big data processing to organize and examine large datasets. These techniques are:

Batch processing: managing substantial amounts of data, mainly for offline analysis, at predetermined times.

Real-time data analysis while it is being generated is known as "***stream processing***," which enables prompt insights and action.

Enabling real-time searches and interactive data exploration through ***interactive processing.*** These methods address various kinds of data and analytical requirements.

12. When to apply MapReduce to large-scale data.

Batch processing jobs like log analysis, data transformation, and ETL (Extract, Transform, Load) are where MapReduce shines. When data can be split into discrete components for processing in parallel, it performs exceptionally well.

13. What does overfitting in big data mean? How to stay away from similar situations.

When a sophisticated machine learning model fits training data too closely, it is said to be overfitting, which impairs the model's capacity to generalize to new, unknown data. To reduce overfitting, you can use the methods listed below:

Cross-validation: To evaluate the generalization of the model, divide the data into training and validation sets.

Regularization: Putting penalties on intricate models to prevent them from overfitting.

Feature Selection: To make the model simpler, select relevant features and eliminate unnecessary ones.

14. List the features of Apache Sqoop.

Apache Sqoop facilitates efficient data transfer between relational databases and Hadoop. Among its features are:

Parallel Data Transfer: To improve performance, Sqoop transfers data in parallel.

Incremental Load Support: Only newly added or updated data from the previous transfer can be moved using incremental load support.

Data Compression: To lower storage and bandwidth requirements, Sqoop facilitates data compression.

15. Describe the feature selection process.

A crucial stage in machine learning is feature selection, which involves selecting pertinent features from a dataset. This lowers complexity and improves model performance.

Duplicate or superfluous characteristics can reduce interpretability, increase processing requirements, and compromise accuracy.

Methods for selecting features Determine which features have the most predictive power for the model by weighing their value.

16. How do you restart all of Hadoop's daemons, including NameNode?

You can use the following commands to restart all daemons in a Hadoop cluster, including the NameNode:

```
hadoop-daemon.sh stop namenode

hadoop-daemon.sh stop datanode

hadoop-daemon.sh stop secondarynamenode
```

Next, to launch the daemons:

```
hadoop-daemon.sh start namenode

hadoop-daemon.sh start datanode

hadoop-daemon.sh start secondarynamenode
```

These commands control several Hadoop daemons, including the NameNode, DataNode, and Secondary NameNode.

17. In big data, what values are missing? And how should one handle it?

Data analysis must handle missing, undefined, or unrecorded values in datasets. These holes may impair insight and model accuracy. Typical approaches to dealing with them include:

Imputation: Replace missing values (such as mean imputation) with estimated or statistical values.

Deletion: Delete any rows or columns that have a small number of missing values.

Based on Models Imputation: Predict and fill in missing data by using machine learning models.

18. List the main configuration parameters that the user must set for MapReduce to work.

The performance of data processing in MapReduce processes can be greatly increased by properly setting and optimizing the distributed cache.

Key configuration parameters that users must define to perform MapReduce tasks include:

- Paths for Input and Output: Establish the directories' paths.
- Indicate which classes define the map and reduce jobs (mapper and reducer classes).
- Number of Reducers: Ascertain how many parallel reduction tasks need to be carried out.

These parameters define how the job behaves and flows data.

19. In Hadoop, how may poor records be skipped?

Two important configurations in Hadoop allow you to skip bad records:

- mapreduce.map.skip.maxrecords: Finds the maximum number of records to skip before the task fails.
- mapreduce.map.skip.procure: Determines whether the job should be ended when the maximum number of skipped records is reached.

These characteristics allow tasks to skip records that contain mistakes and go on processing.

20. Explain Distcp

Distributed Copy, or Distcp, is a Hadoop utility for transferring massive amounts of data between HDFS clusters. Its goal is to improve data transport through effective copy management and parallelization. It is useful for backups and data transfers between Hadoop clusters.

Conclusion

You can prepare to face big data interviews by familiarizing yourself with these top **Big Data analytics interview questions and answers**. Transform your career by enrolling in our **big data analytics training in Chennai**.

Big Data Analytics Training

Share on your Social Media

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

Disclaimer

Privacy Policy

Terms and Conditions

Courses

- Python
- Software Testing
- Full Stack Developer
- Java
- Power BI
- Clinical SAS
- Data Science
- Embedded
- Cloud Computing
- Hardware and Networking
- VBA Macros
- Mobile App Development
- DevOps

Social Media Links

Review Sources

- Google
- Trustpilot
- Glassdoor
- Mouthshut
- Sulekha
- Justdial
- Ambitionbox
- Indeed
- Software Suggest
- Sitejabber