By OnlineInterviewQuestions.com

Big Data interview questions

Big Data is one of the recently and greatly used solution systems in different organizations. Some of the common job opportunities available in this field are in Data Analyst, Database administrator, Big Data Engineer, Data Scientist, Database administrator, Hadoop Big Data Engineer, etc.

The biggest benefit Big Data provides companies is that it increases their revenue and interaction with customers and clients. Some of the other advantages include its efficient way of resolving various business glitches. With regard to this, many recruiters are in the hunt for individuals who have the right technical knowledge along with adequate work experience.

In order to find the right candidate companies ask a diverse range of **Big Data interview questions** to not only freshers but also to the experienced individuals wishing to display their talent and knowledge in this field. Here are some important **Big Data interview questions** that will not only give you a basic idea of the field but also help to clear the interview.

Q1. What do you mean by Big Data and what is its importance?

Big Data is a term related to large and complex data sets. Big Data is required in order to manage and perform different operation on a wide set of data.

Q2. List the five important V's of Big Data.

The five important V's of Big Data are:

1. Value – It refers to changing data into value, which allows businesses to generate revenue.
2. Velocity – Any data growing at an increasing rate is known as its variety. Social media is an important factor contributing to the growth of data.
3. Variety – Data can be of different types such as texts, audios, videos, etc. which are known as variety.
4. Volume – It refers to the amount of any data that is growing at an exponential rate.
5. Veracity – It refers to the uncertainty found in the availability of data. It mainly arises due to the high demand for data which results in inconsistency and incompleteness.

Q3. What is the connection between Hadoop and Big Data?

Hadoop and Big Data are nearly equivalent terms with respect to each other. However, with the ascent of Big Data, Hadoop has also been commonly used. It is a system, which has practical experience in Big Data and also performs additional tasks. Experts can utilize this system in order to break down Big Data and help

organizations to make further decisions.

Q4. How does Big Data help in increasing business revenue?

Big Data has been widely used by a number of organizations in order to increase their business revenue. It is done by helping organizations to distinguish themselves from other competitors in the market. Big Data provides organizations with customized suggestions and recommendations through a series of predictive analysis. Big Data also allows organizations to release new products in accordance with the needs of the customer and their preferences. All these factors contribute to the increase in revenue of a particular business.

Q5. What are the three steps involved in Big Data?

The three essential steps involved in Big Data are:

- Data Ingestion
- Data Storage
- Data Processing

Q6. Explain the first step in Big Data Solutions.

Data Ingestion is the first step of Big Data Solutions. This step refers to the extraction of data from different sources. Different sources data could include CRM, for instance, Salesforce; RDBMS such as MySQL, various Enterprise Resource Planning Systems such as SAP other with other log files, social media feeds, documents, papers, etc. All the data that is extracted is then stored in HDFS.

Q7. What do you understand by the term Data Storage?

Data Storage is the next step in Big Data Solutions. In this step, the data is extracted from the first step is stored in HDFS or NoSQL database, also known as HBase. The HDFS storage is widely used for sequential access. On the contrary, HBase is used for random read or write access.

Q8. What do you mean by Data Processing?

Data Processing is the final step of Big Data Solutions. In this step, with the help of different processing frameworks, the data is processed. Various processing frameworks used are Pig, MapReduce, Spark, etc.

Q9. Name the components of HDFS and YARN respectively

The components of HDFS include:

- NameNode
- DataNode or Slave node

The components of YARN include:

- ResourceManager
- NodeManager

Q10. What is the purpose of using Hadoop for Big Data Analytics?

Hadoop is mainly used for Big Data Analysis for the following benefits:

- Storage
- Processing
- Data Collection
- Ease of dealing with varied structured, semi-structured and unstructured data
- Cost-benefit

Q11. Differentiate between NAS and HDFS

1. In the case of HDFS, data storage is achieved in the form of data blocks within local drivers. On the contrary, data storage in NAS is achieved in the form of dedicated hardware.
2. HDFS works with the help of machines in the form of clusters while NAS works with the help of individual machines.
3. Data dismissal is a common issue in case of HDFS; no such problem is encountered while using NAS.

Q12. What is the procedure to recover a NameNode when it is slow?

In order to recover a NameNode, following steps need to be carried out:

- Using the file system metadata replica FsImage start a new NameNode.
- Configure different DataNodes along with the clients in order to make them recognize the newly initiated NameNode.
- As soon as the new NameNode has completed the checkpoint using FsImage, it will start helping the clients. This is achieved when FsImage has received enough amount of block reports from DataNodes.

Q13. List the common input formats used in Hadoop.

Some of the common input formats used in Hadoop include:

- Key Value Input Format
- Sequence File Input Format
- Text Input Format

Q14. What are some of the different modes used in Hadoop.

Some of the different modes used in Hadoop are:

- Standalone Mode, also known as Local Mode
- Pseudo – Distributed Mode
- Fully – Distributed Mode

Q15. What are the core components that are utilized in Hadoop?

The core components used in Hadoop include:

- Hadoop Distributed File System (HDFS)
- Hadoop MapReduce
- YARN

Q16. What is a cluster in big data?

Clustering in Bigdata is a well-established unsupervised data mining approach that groups data points based on similarities. Clustering entities will give insights into the characteristics of different groups and results in the minimization of the dimensionality of data set when you are dealing with a myriad number of data. The higher the homogeneity within the cluster and the higher the differences between the clusters, the finer the cluster will be. Clusters are mainly of two types; soft clustering, based on the probability that a data point will belong to a specific cluster and, hard clustering, data points are separated into independent clusters. Among hundreds of clustering algorithms, they can be labeled into one of the following models such as connectivity, density, distribution, and centroid model.