DSBDA Question bank

1. Define Big Data and explain the Vs of Big Data.

Big Data can be defined as a collection of complex unstructured or semi-structured data sets which have the potential to deliver actionable insights.

The four Vs of Big Data are –

Volume – Talks about the amount of data

Variety – Talks about the various formats of data

Velocity – Talks about the ever increasing speed at which the data is growing

Veracity – Talks about the degree of accuracy of data available

2. How is Hadoop related to Big Data?

Hadoop is an open-source framework for storing, processing, and analyzing complex unstructured data sets for deriving insights and intelligence.

3. Define HDFS and YARN, and talk about their respective components.

The HDFS is Hadoop's default storage unit and is responsible for storing different types of data in a distributed environment.

HDFS has the following two components:

NameNode – This is the master node that has the metadata information for all the data blocks in the HDFS.

DataNode – These are the nodes that act as slave nodes and are responsible for storing the data.

YARN, short for Yet Another Resource Negotiator, is responsible for managing resources and providing an execution environment for the said processes.

The two main components of YARN are –

ResourceManager – Responsible for allocating resources to respective NodeManagers based on the needs.

NodeManager – Executes tasks on every DataNode.

4. What do you mean by commodity hardware?

Commodity Hardware refers to the minimal hardware resources needed to run the Apache Hadoop framework. Any hardware that supports Hadoop's minimum requirements is known as 'Commodity Hardware.'

5. Define and describe the term FSCK.

FSCK stands for Filesystem Check. It is a command used to run a Hadoop summary report that describes the state of HDFS. It only checks for errors and does not correct them. This command can be executed on either the whole system or a subset of files.

6. What is the purpose of the JPS command in Hadoop?

The JPS command is used for testing the working of all the Hadoop daemons. It specifically tests daemons like NameNode, DataNode, ResourceManager, NodeManager and more.

7. Name the different commands for starting up and shutting down Hadoop Daemons.

This is one of the most important Big Data interview questions to help the interviewer gauge your knowledge of commands.

To start all the daemons:

```
./sbin/start-all.sh
```

To shut down all the daemons:

```
./sbin/stop-all.sh
```

8. Why do we need Hadoop for Big Data Analytics?

In most cases, Hadoop helps in exploring and analyzing large and unstructured data sets. Hadoop offers storage, processing and data collection capabilities that help in analytics.

9. Explain the different features of Hadoop.

Open-Source – Hadoop is an open-sourced platform. It allows the code to be rewritten or modified according to user and analytics requirements.

Scalability – Hadoop supports the addition of hardware resources to the new nodes.

Data Recovery – Hadoop follows replication which allows the recovery of data in the case of any failure.

Data Locality – This means that Hadoop moves the computation to the data and not the other way round. This way, the whole process speeds up.

10. Define the Port Numbers for NameNode, Task Tracker and Job Tracker.

NameNode – Port 50070

Task Tracker – Port 50060

Job Tracker – Port 50030

11. What do you mean by indexing in HDFS?

HDFS indexes data blocks based on their sizes. The end of a data block points to the address of where the next chunk of data blocks get stored. The DataNodes store the blocks of data while NameNode stores these data blocks.

12. What are Edge Nodes in Hadoop?

Edge nodes refer to the gateway nodes which act as an interface between Hadoop cluster and the external network. These nodes run client applications and cluster management tools and are used as staging areas as well. Enterprise-class storage capabilities are required for Edge Nodes, and a single edge node usually suffices for multiple Hadoop clusters.

13. What are some of the data management tools used with Edge Nodes in Hadoop?

This Big Data interview question aims to test your awareness regarding various tools and frameworks.

Oozie, Ambari, Pig and Flume are the most common data management tools that work with Edge Nodes in Hadoop.

14. Explain the core methods of a Reducer.

setup() – This is used to configure different parameters like heap size, distributed cache and input data.

reduce() – A parameter that is called once per key with the concerned reduce task
cleanup() – Clears all temporary files and called only at the end of a reducer task.

15. How can Big Data add value to businesses?

In the present scenario, Big Data is everything. If you have data, you have the most powerful tool at your disposal. Big Data Analytics helps businesses to transform raw data into meaningful and actionable insights that can shape their business strategies. The most important contribution of Big Data to business is data-driven business decisions. Big Data makes it possible for organizations to base their decisions on tangible information and insights.

Furthermore, Predictive Analytics allows companies to craft customized recommendations and marketing strategies for different buyer personas. Together, Big Data tools and technologies help boost revenue, streamline business operations, increase productivity, and enhance customer satisfaction. In fact, anyone who's not leveraging Big Data today is losing out on an ocean of opportunities.

16. How do you deploy a Big Data solution?

- **Data Ingestion** – This is the first step in the deployment of a Big Data solution. You begin by collecting data from multiple sources, be it social media platforms, log files, business documents, anything relevant to your business. Data can either be extracted through real-time streaming or in batch jobs.
- **Data Storage** – Once the data is extracted, you must store the data in a database. It can be HDFS or HBase. While HDFS storage is perfect for sequential access, HBase is ideal for random read/write access.
- **Data Processing** – The last step in the deployment of the solution is data processing. Usually, data processing is done via frameworks like Hadoop, Spark, MapReduce, Flink, and Pig, to name a few.

17. How is NFS different from HDFS?

The Network File System (NFS) is one of the oldest distributed file storage systems, while Hadoop Distributed File System (HDFS) came to the spotlight only recently after the upsurge of Big Data.
The table below highlights some of the most notable differences between NFS and HDFS:

NFS	**HDFS**
It can both store and process small volumes of data.	It is explicitly designed to store and process Big Data.
The data is stored in dedicated hardware.	Data is divided into data blocks that are distributed on the local drives of the hardware.
In the case of system failure, you cannot access the data.	Data can be accessed even in the case of a system failure.
Since NFS runs on a single machine, there's no chance for data redundancy.	HDFS runs on a cluster of machines, and hence, the replication protocol may lead to redundant data.

18. List the different file permissions in HDFS for files or directory levels.

One of the common big data interview questions. The Hadoop distributed file system (HDFS) has specific permissions for files and directories. There are three user levels in HDFS – Owner, Group, and Others. For each of the user levels, there are three available permissions:

- read (r)
- write (w)
- execute(x).

These three permissions work uniquely for files and directories.
For files –

- The r permission is for reading a file
- The w permission is for writing a file.

Although there's an execute(x) permission, you cannot execute HDFS files.
For directories –

- The r permission lists the contents of a specific directory.
- The w permission creates or deletes a directory.
- The X permission is for accessing a child directory.

19. Name the three modes in which you can run Hadoop.

- **Standalone mode** – This is Hadoop's default mode that uses the local file system for both input and output operations. The main purpose of the standalone mode is debugging. It does not support HDFS and also lacks custom configuration required for mapred-site.xml, core-site.xml, and hdfs-site.xml files.
- **Pseudo-distributed mode** – Also known as the single-node cluster, the pseudo-distributed mode includes both NameNode and DataNode within the same machine. In this mode, all the Hadoop daemons will run on a single node, and hence, the Master and Slave nodes are the same.
- **Fully distributed mode** – This mode is known as the multi-node cluster wherein multiple nodes function simultaneously to execute Hadoop jobs. Here, all the Hadoop daemons run on different nodes. So, the Master and Slave nodes run separately.

20. Explain "Overfitting."

Overfitting refers to a modeling error that occurs when a function is tightly fit (influenced) by a limited set of data points. Overfitting results in an overly complex model that makes it further difficult to explain the peculiarities or idiosyncrasies in the data at hand. As it adversely affects the generalization ability of the model, it becomes challenging to determine the predictive quotient of overfitted models. These models fail to perform when applied to external data (data that is not part of the sample data) or new datasets.

Overfitting is one of the most common problems in Machine Learning. A model is considered to be overfitted when it performs better on the training set but fails miserably on the test set. However, there are many methods to prevent the problem of overfitting, such as cross-validation, pruning, early stopping, regularization, and assembling.

21. What is Feature Selection?

Feature selection refers to the process of extracting only the required features from a specific dataset. When data is extracted from disparate sources, not all data is useful at all times – different business needs call for different data insights. This is where feature selection comes in to identify and select only those features that are relevant for a particular business requirement or stage of data processing.

The main goal of feature selection is to simplify ML models to make their analysis and interpretation easier. Feature selection enhances the generalization abilities of a model and eliminates the problems of dimensionality, thereby, preventing the possibilities of overfitting. Thus, feature selection provides a better understanding of the data under study, improves the prediction performance of the model, and reduces the computation time significantly.

Feature selection can be done via three techniques:

- **Filters method**
  In this method, the features selected are not dependent on the designated classifiers. A variable ranking technique is used to select variables for ordering purposes. During the classification process, the variable ranking technique takes into consideration the importance and usefulness of a feature. The Chi-Square Test, Variance Threshold, and Information Gain are some examples of the filters method.
- **Wrappers method**
  In this method, the algorithm used for feature subset selection exists as a 'wrapper' around the induction algorithm. The induction algorithm functions like a 'Black Box' that produces a classifier that will be further used in the classification of features. The major drawback or limitation of the wrappers method is that to obtain the feature subset, you need to perform heavy computation work. Genetic Algorithms, Sequential Feature Selection, and Recursive Feature Elimination are examples of the wrappers method.
- **Embedded method**
  The embedded method combines the best of both worlds – it includes the best features of the filters and wrappers methods. In this method, the variable selection is done during the training process, thereby allowing you to identify the features that are the most accurate for a given model. L1 Regularisation Technique and Ridge Regression are two popular examples of the embedded method.

22. Define "Outliers."

An outlier refers to a data point or an observation that lies at an abnormal distance from other values in a random sample. In other words, outliers are the values that are far removed from the group; they do not belong to any specific cluster or group in the dataset. The presence of outliers usually affects the behavior of the model – they can mislead the training process of ML algorithms. Some of the adverse impacts of outliers include longer training time, inaccurate models, and poor outcomes.

However, outliers may sometimes contain valuable information. This is why they must be investigated thoroughly and treated accordingly.

23. Explain Rack Awareness in Hadoop.

Rack awareness is an algorithm that identifies and selects DataNodes closer to the NameNode based on their rack information. It is applied to the NameNode to determine how data blocks and their replicas will be placed. During the installation process, the default assumption is that all nodes belong to the same rack.

Rack awareness helps to:

- Improve data reliability and accessibility.
- Improve cluster performance.
- Improve network bandwidth.
- Keep the bulk flow in-rack as and when possible.
- Prevent data loss in case of a complete rack failure.

24. Can you recover a NameNode when it is down? If so, how?

Yes, it is possible to recover a NameNode when it is down. Here's how you can do it:

- Use the FsImage (the file system metadata replica) to launch a new NameNode.
- Configure DataNodes along with the clients so that they can acknowledge and refer to newly started NameNode.
- When the newly created NameNode completes loading the last checkpoint of the FsImage (that has now received enough block reports from the DataNodes) loading process, it will be ready to start serving the client.

However, the recovery process of a NameNode is feasible only for smaller clusters. For large Hadoop clusters, the recovery process usually consumes a substantial amount of time, thereby making it quite a challenging task.

25. What is a Distributed Cache? What are its benefits?

Any Big Data Interview Question and Answers guide won't complete without this question. Distributed cache in Hadoop is a service offered by the MapReduce framework used for caching files. If a file is cached for a specific job, Hadoop makes it available on individual DataNodes both in memory and in system where the map and reduce tasks are simultaneously executing. This allows you to quickly access and read cached files to populate any collection (like arrays, hashmaps, etc.) in a code.

Distributed cache offers the following benefits:

- It distributes simple, read-only text/data files and other complex types like jars, archives, etc.
- It tracks the modification timestamps of cache files which highlight the files that should not be modified until a job is executed successfully.

26. What is a SequenceFile in Hadoop?

In Hadoop, a SequenceFile is a flat-file that contains binary key-value pairs. It is most commonly used in MapReduce I/O formats. The map outputs are stored internally as a SequenceFile which provides the reader, writer, and sorter classes.

There are three SequenceFile formats:

- Uncompressed key-value records
- Record compressed key-value records (only 'values' are compressed).
- Block compressed key-value records (here, both keys and values are collected in 'blocks' separately and then compressed).

27. Explain the role of a JobTracker.

One of the common big data interview questions. The primary function of the JobTracker is resource management, which essentially means managing the TaskTrackers. Apart from this, JobTracker also tracks resource availability and handles task life cycle management (track the progress of tasks and their fault tolerance).

Some crucial features of the JobTracker are:

- It is a process that runs on a separate node (not on a DataNode).
- It communicates with the NameNode to identify data location.
- It tracks the execution of MapReduce workloads.
- It allocates TaskTracker nodes based on the available slots.
- It monitors each TaskTracker and submits the overall job report to the client.
- It finds the best TaskTracker nodes to execute specific tasks on particular nodes.

28. Name the common input formats in Hadoop.

Hadoop has three common input formats:

- Text Input Format – This is the default input format in Hadoop.
- Sequence File Input Format – This input format is used to read files in a sequence.
- Key-Value Input Format – This input format is used for plain text files (files broken into lines).

29. What is the need for Data Locality in Hadoop?

One of the important big data interview questions. In HDFS, datasets are stored as blocks in DataNodes in the Hadoop cluster. When a MapReduce job is executing, the individual Mapper processes the data blocks (Input Splits). If the data does is not present in the same node where the Mapper executes the job, the data must be copied from the DataNode where it resides over the network to the Mapper DataNode. When a MapReduce job has over a hundred Mappers and each Mapper DataNode tries to copy the data from another DataNode in the cluster simultaneously, it will lead to network congestion, thereby having a negative impact on the system's overall performance. This is where Data Locality enters the scenario. Instead of moving a large chunk of data to the computation, Data Locality moves the data computation close to where the actual data resides on the DataNode. This helps improve the overall performance of the system, without causing unnecessary delay.

30. What are the steps to achieve security in Hadoop?

In Hadoop, Kerberos – a network authentication protocol – is used to achieve security. Kerberos is designed to offer robust authentication for client/server applications via secret-key cryptography.

When you use Kerberos to access a service, you have to undergo three steps, each of which involves a message exchange with a server. The steps are as follows:

- **Authentication** – This is the first step wherein the client is authenticated via the authentication server, after which a time-stamped TGT (Ticket Granting Ticket) is given to the client.
- **Authorization** – In the second step, the client uses the TGT for requesting a service ticket from the TGS (Ticket Granting Server).
- **Service Request** – In the final step, the client uses the service ticket to authenticate themselves to the server.

31. How can you handle missing values in Big Data?

Final question in our big data interview questions and answers guide. Missing values refer to the values that are not present in a column. It occurs when there's is no data value for a variable in an observation. If missing values are not handled properly, it is bound to lead to erroneous data which in turn will generate incorrect outcomes. Thus, it is highly recommended to treat missing values correctly before processing the datasets. Usually, if the number of missing values is small, the data is dropped, but if there's a bulk of missing values, data imputation is the preferred course of action.

In Statistics, there are different ways to estimate the missing values. These include regression, multiple data imputation, listwise/pairwise deletion, maximum likelihood estimation, and approximate Bayesian bootstrap.

32) Mention what are the various steps in an analytics project?

- Problem definition
- Data exploration
- Data preparation
- Modelling
- Validation of data
- Implementation and tracking

33) Mention what is data cleansing?

Data cleaning also referred as data cleansing, deals with identifying and removing errors and inconsistencies from data in order to enhance the quality of data.

34) List of some best tools that can be useful for data-analysis?

- Tableau
- RapidMiner
- OpenRefine
- KNIME
- Google Search Operators

- Solver
- NodeXL
- io
- Wolfram Alpha's
- Google Fusion tables

35) what is the difference between data mining and data profiling?

- **Data profiling:** It targets on the instance analysis of individual attributes. It gives information on various attributes like value range, discrete value and their frequency, occurrence of null values, data type, length, etc.
- **Data mining:** It focuses on cluster analysis, detection of unusual records, dependencies, sequence discovery, relation holding between several attributes, etc.

36) **List out some common problems faced by data analyst?**

- Common misspelling
- Duplicate entries
- Missing values
- Illegal values
- Varying value representations
- Identifying overlapping data

37) **What is a hash table?**

In computing, a hash table is a map of keys to values. It is a data structure used to implement an associative array. It uses a hash function to compute an index into an array of slots, from which desired value can be fetched.

38) **What are the different data types supported by Tableau?**

- Boolean- True/False
- Date- Date values
- Date and time- Timestamp values and date values
- Geographical Values- Geographical Mapping
- Text/ String
- Number- Decimal and Whole numbers

39) What is data visualization?

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

40) Mention the use of the split function in Python?

The use of the split function in Python is that it breaks a string into shorter strings using the defined separator. It gives a list of all words present in the string.

41) Mention five benefits of using Python?

- Python comprises of a huge standard library for most Internet platforms like Email, HTML, etc.
- Python does not require explicit memory management as the interpreter itself allocates the memory to new variables and free them automatically
- Provide easy readability due to use of square brackets
- Easy-to-learn for beginners

- Having the built-in data types saves programming time and effort from declaring variables

42) List different data types in python

Text Type:	**str**
Numeric **Types**:	int , float , complex
Sequence **Types**:	list , tuple , range
Mapping **Type**:	dict
Set **Types**:	set , frozenset

43) how to define a function in python?

Def function_name(parameters)

“”” docstring “””

Statements in function

………

reutrn() # not mandatory

44) Name reading and printing statement in python?

print - output

input – reading

45) Strings in python are immutable

Yes # cannot be modified.

46) write syntax of conditional and iterative statements in python.

Note : refer python ppt given

48) Does Hadoop support parallel processing
Yes
49) Name few file handling functions in python.

Refer PPT given

50) Name Different NoSQL Databases

1. MongoDB
2. Cassandra
3. ElasticSearch
4. Amazon DynamoDB
5. HBase