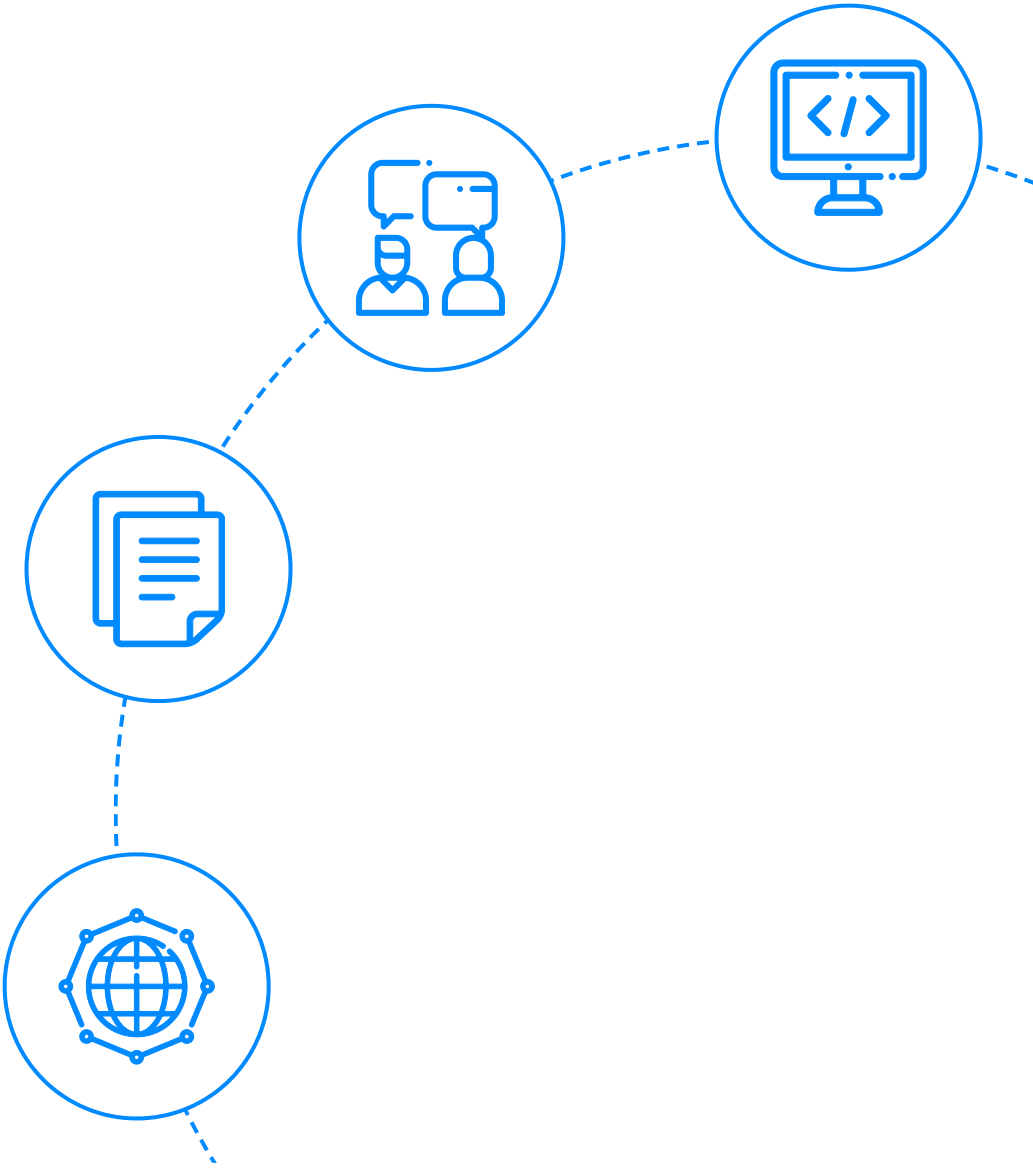
InterviewBit

Big Data Interview Questions



To view the live version of the page, click here.

Contents

Big Data Interview Questions for Freshers

Big Data Interview Questions for Experienced

Big Data Interview Questions for Experienced

(.....Continued)

19. Explain the distributed Cache in the MapReduce framework.
20. Explain overfitting in big data? How to avoid the same.
21. What is a Zookeeper? What are the benefits of using a zookeeper?
22. What is the default replication factor in HDFS?
23. Mention features of Apache sqoop.
24. Write the command used to copy data from the local system onto HDFS?
25. What is partitioning in Hive?
26. Explain Features Selection.
27. How can you restart NameNode and all the daemons in Hadoop?
28. What is the use of the -compress-codec parameter?
29. What are missing values in Big data? And how to deal with it?
30. What are the things to consider when using distributed cache in Hadoop MapReduce?
31. Mention the main configuration parameters that has to be specified by the user to run MapReduce.
32. How can you skip bad records in Hadoop ?
33. Explain Outliers.
34. What is Distcp?
35. Explain Persistent, Ephemeral and Sequential Znodes.
36. Explain the Pros and Cons of Big Data?
37. How do you convert unstructured data to structured data?
38. What is data preparation?

Big Data Interview Questions for Experienced

(.....Continued)

39. Steps for Data preparation.

Let's get Started

Introduction to Big Data:

> "Information is the oil of the 21st century, and analytics is the combustion engine."
> -Peter Sondergaard

Big data is all about vast amounts of data, i.e., large datasets measured in terabytes or petabytes or even more. According to a study, around 90% of today's data was generated in the last two years. Well, the term big data was coined in the mid's of **2005 by O'Reilly Media**.

Big data helps numerous companies to generate valuable insights about the products or the services they offer. In recent years, almost every company uses big data technology to refine its marketing campaigns and techniques.

Well, what about a profession in Big data? Because big data is still or maybe ever-growing technology, it is among the world's ten highest paying jobs in the year 2020 in technology. There are tremendous opportunities available across the globe. Thus Big data proves to be an attractive career for this generation.

Big Data Interview Questions for Freshers

1. What is Big Data, and where does it come from? How does it work?

Big Data refers to extensive and often complicated data sets so huge that they're beyond the capacity of managing with conventional software tools. Big Data comprises unstructured and structured data sets such as videos, photos, audio, websites, and multimedia content.

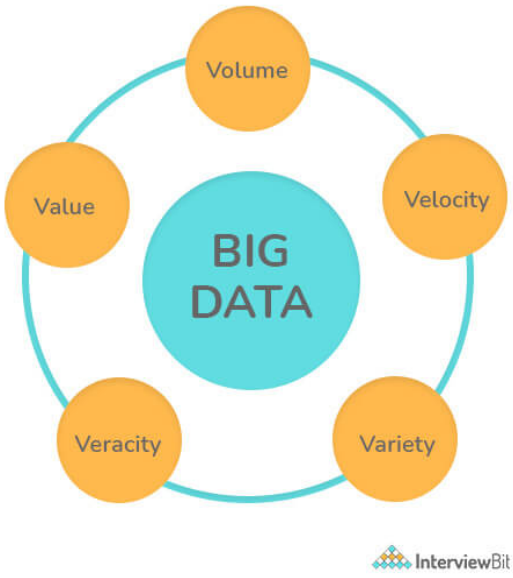
Businesses collect the data they need in countless ways, such as:

- Internet cookies
- Email tracking
- Smartphones
- Smartwatches
- Online purchase transaction forms
- Website interactions
- Transaction histories
- Social media posts
- Third-party trackers -companies that collect and sell clients and profitable data

Working with big data involves three sets of activities:

- **Integration:** This involves merging data often from different sources – and molding it into a form that can be analysed in a way to provide insights.
- **Management:** Big data must be stored in a repository where it can be collected and readily reached. The largest amount of Big Data is unstructured, causing it ill-suited for conventional relational databases, which need data in tables-and-rows format.
- **Analysis:** The Big Data investment return is a spectrum of worthy market insights, including details on buying patterns and customer choices. These are represented by examining large data sets with tools driven by AI and machine learning.

2. What are the 5 V's in Big Data?



- **Volume:** A considerable amount of data stored in data warehouses reflects the volume. The data may reach random heights; these large volumes of data need to be examined and processed. Which may exist up to or more than terabytes and petabytes.
- **Velocity:** Velocity basically introduces the pace at which data is being produced in real-time. To give a simple example for recognition, imagine the rate at which Facebook, Instagram, or Twitter posts are generated per second, an hour or more.
- **Variety:** Big Data comprises structured, unstructured, and semi-structured data collected from varied sources. This different variety of data requires very different and specific analyzing and processing techniques with unique and appropriate algorithms.
- **Veracity:** Data veracity basically relates to how reliable the data is, or in a fundamental way, we can define it as the quality of the data analyzed.
- **Value:** Raw data is of no use or meaning but once converted into something valuable. We can extract helpful information.

3. Why businesses are using Big Data for competitive advantage.

Irrespective of the division and scope of the firm, data is now an essential tool for businesses to utilise. Companies are frequently using big data to gain a competing edge over business rivals.

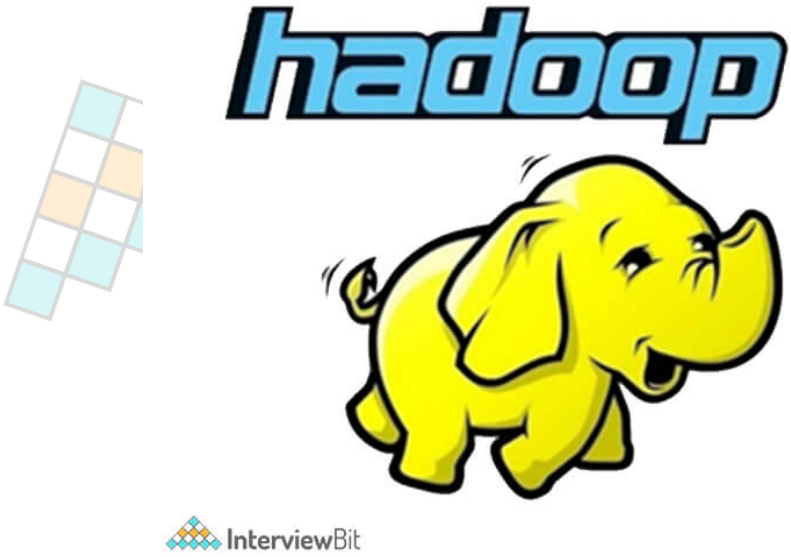
Checking the datasets a company collects is just one part of the big data process. Big data professionals also need to know what the company requires from the application and how they plan to use the data to their advantage.

- **Confident decision-building:** Analytics aims to develop decision building, and big data endures to sustain this. Big data can help enterprises speed up their decision-making method with so much data available while still being assured of their choice. Nowadays, moving fast and reacting to broader trends and operational changes is a huge business benefit in a quick-paced society.
- **Asset optimisation:** Big data signifies that businesses can control assets at a personal level. This implies they can adequately optimise assets depending on the data source, improve productivity, extend the lifespan of help, and reduce the downtime some assets may require. This gives a competing advantage by assuring the company is getting the most out of its assets and links with decreasing costs.
- **Cost reduction:** Big data can support businesses to reduce their outgoings. From analysing energy usage to assessing the effectiveness of staff operating patterns, data collected by companies can help them recognise where they can make cost savings without having a negative impact on company operations.
- **Improve customer engagement:** When surveying online, consumers make confident choices indicating their decisions, habits, and tendencies that can then be used to develop and tailor consumer dialogue, which could then be interpreted into increased sales. Understanding what each client is looking for through the data collected on them means you can target them with specific products, but it also gives a personal feel that many consumers today have come to await.
- **Identify new revenue streams:** Analytics can further assist companies in identifying new revenue streams and expanding into other areas. For example, knowing customer trends and decisions allow firms to decide the way they should go. The data companies accumulate can also likely be sold, adding income streams and the potential to build alliances with other businesses.

4. How is Hadoop and Big Data related?

If we talk about Big Data, we do talk about Hadoop as well. So, this is one of the most critical questions from an interview perspective. That you might surely face. Hadoop is an open-source framework for saving, processing, and interpreting complex, disorganized data sets for obtaining insights and knowledge. So, that is how Hadoop and Big Data are related to each other.

5. Explain the importance of Hadoop technology in Big data analytics.



Since big data includes a large volume of data, i.e., structured, semi-structured, and unstructured data, analyzing and processing this data is quite a big task. There was a need for a tool or technology to help process the data at a rapid speed. Therefore, Hadoop is used because of its capabilities like storage, processing capability. Moreover, Hadoop is an open-source software. If you want to consider the cost, it's beneficial for business solutions.
The main reason for its popularity in recent years is that this framework permits distributed processing of enormous data sets using crosswise clusters of computers practicing simple programming models.

6. Explain the core components of Hadoop.

Hadoop is an open-source framework intending to store and process big data in a distributed manner.

Hadoop's Essential Components:

- **HDFS (Hadoop Distributed File System)** – Hadoop's key storage system is HDFS. The extensive data is stored on HDFS. It is mainly devised for storing massive datasets in commodity hardware.
- **Hadoop MapReduce** – The responsible layer of Hadoop for data processing is MapReduce. It puts a request for processing of structured and unstructured data which is already stored in HDFS. It is liable for the parallel processing of a high volume of data by distributing data into detached tasks. There are two stages of processing: Map and Reduce. In simple terms, Map is a stage where data blocks are read and made available to the executors (computers /nodes /containers) for processing. Reduce is a stage where all processed data is collected and collated.
- **YARN** – The framework which is used to process in Hadoop is YARN. For resource management and to provide multiple data processing engines like real-time streaming, data science, and batch processing is done by YARN.

7. Explain the features of Hadoop.

Hadoop assists in not only store data but also processing big data. It is the most reliable way to handle significant data hurdles. Some salient features of Hadoop are –

- **Distributed Processing** – Hadoop helps in distributed processing of data, i.e., quicker processing. In Hadoop HDFS, the data is collected in a distributed manner, and the data is parallel processing, and MapReduce is liable for the same.
- **Open Source** – Hadoop is independent of cost as it is an open-source framework. Changes are allowed in the source code as per the user's requirements.
- **Fault Tolerance** – Hadoop is highly fault-tolerant. By default, for every block, it creates three replicas at distinct nodes. This number of replicas can be modified according to the requirement. So, we can retrieve the data from a different node if one of the nodes fails. The discovery of node failure and restoration of data is made automatically.
- **Scalability** – It is fitted with different hardware, and we can promptly access the new device.
- **Reliability** – The data in Hadoop is stored on the cluster in a safe manner that is autonomous of the machine. So, the data stored in the Hadoop ecosystem's data does not get affected by any machine breakdowns.

8. How is HDFS different from traditional NFS?

NFS (Network File system): A protocol that enables customers to access files over the network. NFS clients would allow files to be accessed as if the files live on the local device, even though they live on the disk of a networked device.

HDFS (Hadoop Distributed File System): A distributed file system is shared between multiple networked machines or nodes. HDFS is fault-tolerant because it saves various copies of files on the file system; the default replication level is 3.
The notable difference between the two is Replication/Fault Tolerance. HDFS was intended to withstand failures. NFS does not possess any fault tolerance built-in.

Benefits of HDFS over NFS:
Apart from fault tolerance, HDFS helps to create multiple replicas of files. This reduces the traditional bottleneck of many clients accessing a single file. In addition, since files have multiple images on various physical disks, reading performance scales better than NFS.

9. What is data modelling and what is the need for it.

Data Modeling as a business has been practiced in the IT sector for many decades. As an idea, the data model is a means to arrive at the diagram by examining the data in question and getting a deep knowledge. The method of representing the data visually encourages the business and the technology specialists to understand the data and understand how it will get used.

Kinds Of Data Models

The three principal types of data models are conceptual, logical, and physical. Think of them as an improvement from an abstract layout to a detailed mapping of the database setup and final form:

- **Conceptual Data Model:**
  Conceptual data models are the most simplistic and abstract. Minor annotation happens in this model, but the overall layout and controls of the data relationships are set. You'll find elements like basic market rules that need to be applied, the levels or entity classes of data that you plan to cover, and any other regulations that may limit layout options. Data models are usually used in the development stage of a project.
- **Logical Data Model:**
  The logical data model extends on the basic framework laid out in the conceptual model, but it counts more relational factors. Thus, some basic annotations are related to overall properties or data attributes, but not many annotations concentrate on actual data units. Hence, this model is beneficial in data warehousing projects.
- **Physical Data Model:**
  The physical data model is the most comprehensive and the last step before database production, it usually accounts for database management system-specific properties and rules.

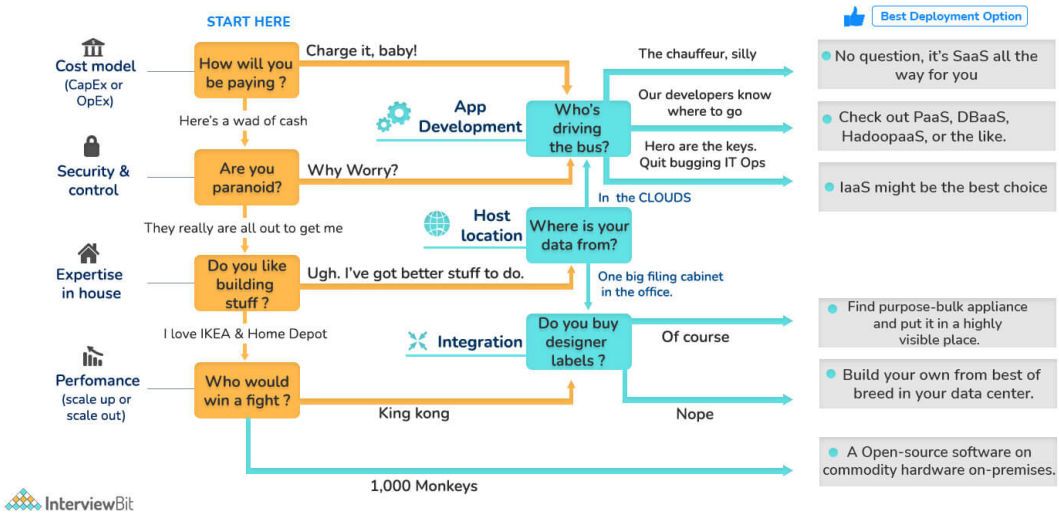
Advantages Of Data Modeling:

- Data modeling offers several different benefits to companies as part of their data management:
- Before you even build a database, you've cleaned, organized, and modeled your data to project what your next step should look like. Data modeling advances data quality and makes databases limited, prone to mistakes and bad design.
- Data modeling produces a visual flow of data and how you plan to organize it. This supports employees know what's happening with data and how they relate to the data management puzzle. It also develops data-related communication across departments in an organization.
- Data modeling allows more profound database design, bringing forth more useful applications and data-based business insights down the line.

10. How to deploy a Big Data Model? Mention the key steps involved.

Deploying a model into a **Big Data Platform** involves mainly three key steps they are,

- Data ingestion
- Data Storage
- Data Processing



Let's have a look at what these are,

- **Data Ingestion:** This process involves collecting data from different sources like social media platforms, business applications, log files, etc.
- **Data Storage:** When data extraction is completed, the challenge is to store this large volume of data in the database in which the Hadoop Distributed File system (HDFS) plays a vital role.
- **Data Processing:** After storing the data in HDFS or HBase, the next task is to analyze and visualize these large amounts of data using specific algorithms for better data processing. And yet again, this task is more straightforward if we use Hadoop, Apache Spark, Pig, etc.

After performing these essential steps, one can deploy a big data model successfully.

11. What is fsck?

The term fsck stands for File System Check, used by HDFS. It is used to check discrepancies and if there is any difficulty in the file. For instance, if there are any missing blocks in the file, HDFS gets reported through this command.

12. What are the three modes that Hadoop can run?

- **Local Mode or Standalone Mode:**
  By default, Hadoop is configured to operate in a no distributed mode. It runs as a single Java process. Instead of HDFS, this mode utilizes the local file system. This mode is more helpful for debugging, and there isn't any requirement to configure core-site.xml, hdfs-site.xml, mapred-site.xml, masters & slaves. Standalone mode is ordinarily the quickest mode in Hadoop.
- **Pseudo-distributed Mode:**
  In this mode, each daemon runs on a separate java process. This mode requires custom configuration (core-site.xml, hdfs-site.xml, mapred-site.xml). The HDFS is used for input and output. This mode of deployment is beneficial for testing and debugging purposes.
- **Fully Distributed Mode:**
  It is the production mode of Hadoop. One machine in the cluster is assigned as NameNode and another as Resource Manager exclusively. These are masters. Rest nodes act as Data Node and Node Manager. These are the slaves. Configuration parameters and environment need to be defined for Hadoop Daemons. This mode gives fully distributed computing capacity, security, fault endurance, and scalability.

13. Mention the common input formats in Hadoop.

The common input formats in Hadoop are -

- **Text Input Format:** This is the default input format in Hadoop.
- **Key-Value Input Format:** Used to read Plain Text Files in Hadoop.
- **Sequence File Input format:** This is used to read Files in a sequence in Hadoop.

14. What are the different Output formats in Hadoop?

The different Output formats in Hadoop are -

- **Textoutputformat:** TextOutputFormat is the default output format in Hadoop.
- **Mapfileoutputformat:** Mapfileoutputformat is used to write the output as map files in Hadoop.
- **DBoutputformat:** DBoutputformat is just used for writing output in relational databases and Hbase.
- **Sequencefileoutputformat:** Sequencefileoutputformat is used for writing sequence files.
- **SequencefileAsBinaryoutputformat:** SequencefileAsBinaryoutputformat is used to write keys to a sequence file in binary format.

Big Data Interview Questions for Experienced

15. What are the different big data processing techniques?

Big Data processing methods analyze big data sets at a massive scale. Offline batch data processing is typically full power and full scale, tackling arbitrary BI scenarios. In contrast, real-time stream processing is conducted on the most recent slice of data for data profiling to pick outliers, impostor transaction exposures, safety monitoring, etc. However, the most challenging task is to do fast or real-time ad-hoc analytics on a big comprehensive data set. It substantially means you need to scan tons of data within seconds. This is only probable when data is processed with high parallelism.

Different techniques of Big Data Processing are:

- Batch Processing of Big Data
- Big Data Stream Processing
- Real-Time Big Data Processing
- Map Reduce

16. What is Map Reduce in Hadoop?

Hadoop MapReduce is a software framework for processing enormous data sets. It is the main component for data processing in the Hadoop framework. It divides the input data into several parts and runs a program on every data component parallel. The word MapReduce refers to two separate and different tasks. The first is the map operation, which takes a set of data and transforms it into a diverse collection of data, where individual elements are divided into tuples. The reduce operation consolidates those data tuples based on the key and subsequently modifies the value of the key.

17. When to use MapReduce with Big Data.

MapReduce is a programming model created for distributed computation on big data sets in parallel. A MapReduce model has a map function that performs filtering and sorting and a reduced function, which serves as a summary operation.

MapReduce is an important part of the Apache Hadoop open-source ecosystem, and it's extensively used for querying and selecting data in the Hadoop Distributed File System (HDFS). A variety of queries may be done depending on the broad spectrum of MapReduce algorithms possible for creating data selections. In addition, MapReduce is fit for iterative computation involving large quantities of data requiring parallel processing. This is because it represents a data flow rather than a procedure.

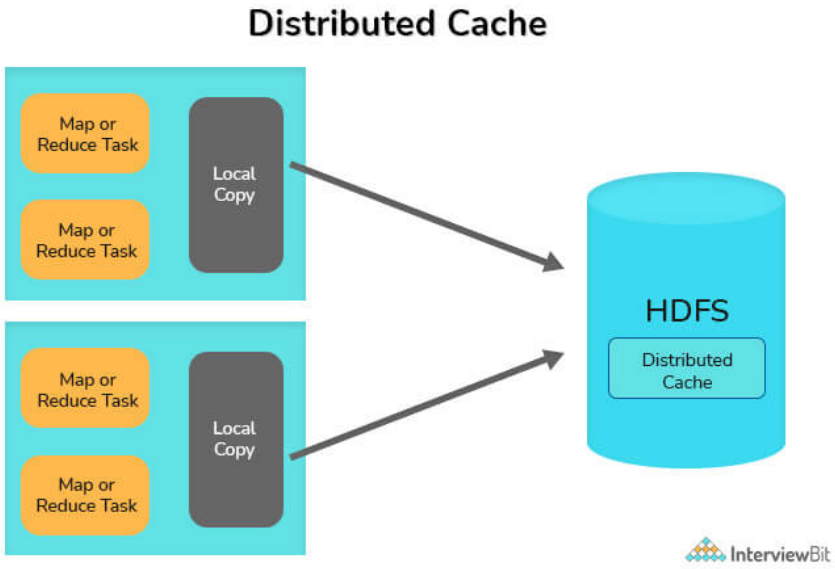
The more enhanced data we produce and accumulate, the higher the need to process all that data to make it usable. MapReduce's iterative, parallel processing programming model is a good tool for creating a sense of big data.

18. Mention the core methods of Reducer.

The core methods of a Reducer are:

- **setup():** setup is a method called just to configure different parameters for the reducer.
- **reduce():** reduce is the primary operation of the reducer. The specific function of this method includes defining the task that has to be worked on for a distinct set of values that share a key.
- **cleanup():** cleanup is used to clean or delete any temporary files or data after performing reduce() task.

19. Explain the distributed Cache in the MapReduce framework.



Distributed Cache is a significant feature provided by the MapReduce Framework, practiced when you want to share the files across all nodes in a Hadoop cluster. These files can be jar files or simple properties files. Hadoop's MapReduce framework allows the facility to cache small to moderate read-only files such as text files, zip files, jar files, etc., and distribute them to all the Datanodes(worker-nodes) MapReduce jobs are running. All Datanode gets a copy of the file(local-copy), which Distributed Cache sends.

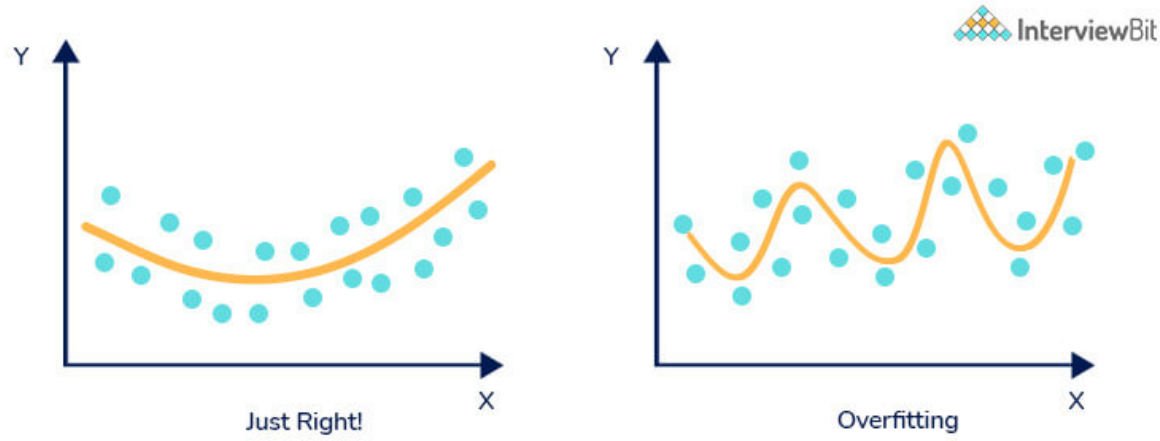
20. Explain overfitting in big data? How to avoid the same.

Overfitting is generally a modeling error referring to a model that is tightly fitted to the data, i.e. When a modeling function is closely fitted to a limited data set. Due to Overfitting, the predictivity of such models gets reduced. This effect leads to a decrease in generalization ability failing to generalize when applied outside the sample data.

There are several Methods to avoid Overfitting; some of them are:

- **Cross-validation:** A cross-validation method refers to dividing the data into multiple small test data sets, which can be used to tune the model.
- **Early stopping:** After a certain number of iterations, the generalizing capacity of the model weakens; in order to avoid that, a method called early stopping is used in order to avoid Overfitting before the model crosses that point.

- **Regularization:** this method is used to penalize all the parameters except intercept so that the model generalizes the data instead of Overfitting.



21. What is a Zookeeper? What are the benefits of using a zookeeper?

Hadoop's most remarkable technique for addressing big data challenges is its capability to divide and conquer with Zookeeper. After the problem has been divided, the conquering relies on employing distributed and parallel processing methods across the Hadoop cluster.
The interactive tools cannot provide the insights or timeliness needed to make business judgments for big data problems. In those cases, you need to build distributed applications to solve those big data problems. Zookeeper is Hadoop's way of coordinating all the elements of these distributed applications.
Zookeeper as technology is simple, but its features are powerful. Arguably, it would be difficult, if not impossible, to create resilient, fault-tolerant distributed Hadoop applications without it.

Benefits of using a Zookeeper are:

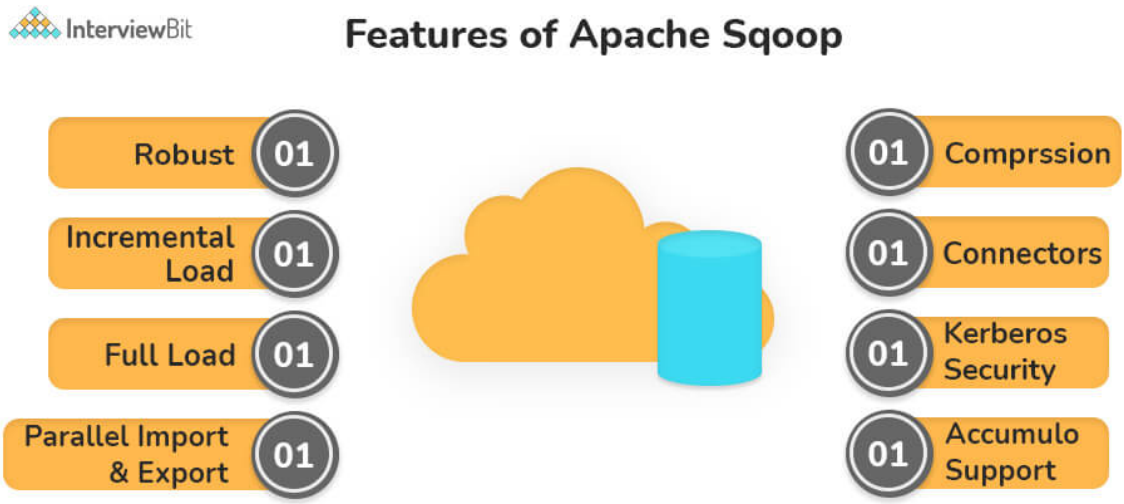
- **Simple distributed coordination process:** The coordination process among all nodes in Zookeeper is straightforward.
- **Synchronization:** Mutual exclusion and co-operation among server processes.
- **Ordered Messages:** Zookeeper tracks with a number by denoting its order with the stamping of each update; with the help of all this, messages are ordered here.
- **Serialization:** Encode the data according to specific rules. Ensure your application runs consistently.
- **Reliability:** The zookeeper is very reliable. In case of an update, it keeps all the data until forwarded.
- **Atomicity:** Data transfer either succeeds or fails, but no transaction is partial.

22. What is the default replication factor in HDFS?

By default, **the replication factor is 3**. There are no two copies that will be on the same data node. Usually, the first two copies will be on the same rack, and the third copy will be off the shelf. It is advised to set the replication factor to at least three so that one copy is always safe, even if something happens to the rack.

We can set the default replication factor of the file system and each file and directory exclusively. We can lower the replication factor for files that are **not essential**, and critical files should have a high replication factor.

23. Mention features of Apache sqoop.



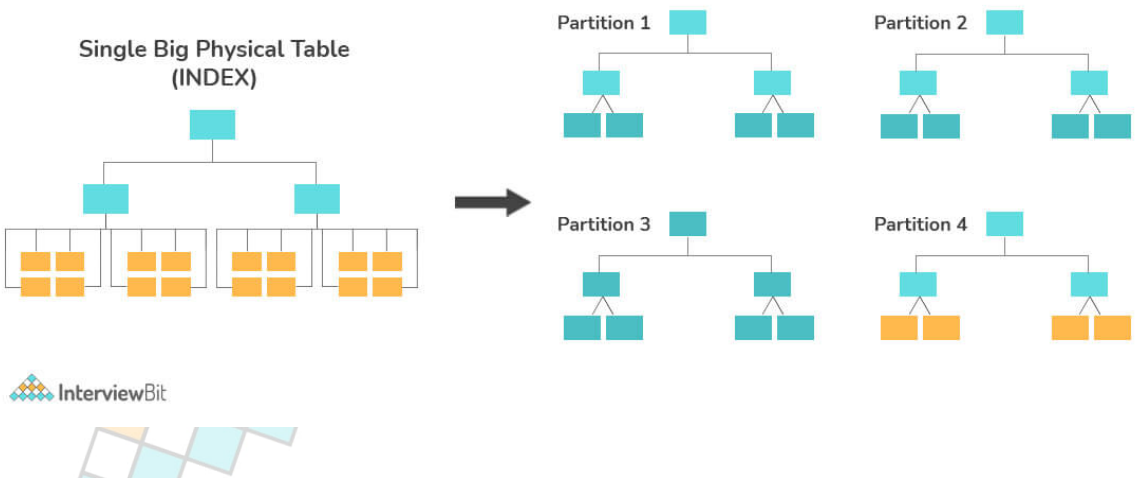
- **Robust:** It is extremely robust and easy to use. In addition, it has community support and contribution.
- **Full Load:** Loading a table in Sqoop can be done in one command. Multiple tables can also be loaded in the same process.
- **Incremental Load:** Incremental load functionality is also supported. Whenever the table is updated, with the help of Sqoop, it can be loaded in parts too.
- **Parallel import/export:** Importing and exporting of data is done by the YARN framework. It also provides fault tolerance too.
- **Import results of SQL query:** It allows us to import the output from the SQL query into the Hadoop Distributed File System.

24. Write the command used to copy data from the local system onto HDFS?

The command used for copying data from the Local system to HDFS is:
hadoop fs –copyFromLocal [source][destination]

25. What is partitioning in Hive?

In general partitioning in Hive is a logical division of tables into related columns such as date, city, and department based on the values of partitioned columns. Then these partitions are subdivided into buckets so that they provide extra structure to the data that may be used for more efficient querying.



Now let's experience data partitioning in Hive with an instance. Consider a table named Table1. The table contains client details like id, name, dept, and year of joining. Assume we need to retrieve the details of all the clients who joined in 2014.

Then, the query examines the whole table for the necessary data. But if we partition the client data by the year and save it in a different file, this will decrease the query processing time.

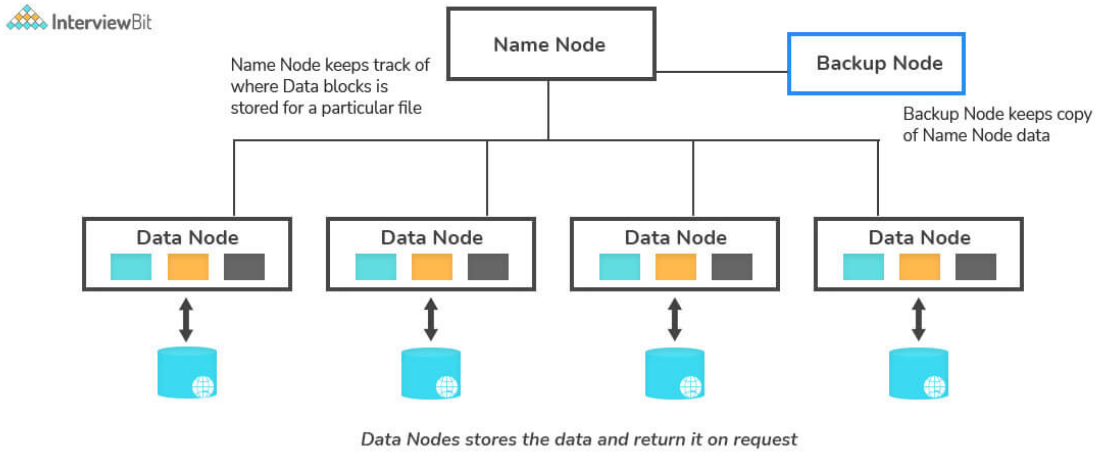
26. Explain Features Selection.

During processing, Big data may contain a large amount of data that is not required at a particular time, So we may be required to select only some specific features that we are interested in. The process of extracting only the needed features from the Big data is called Feature selection.

Feature selection Methods are -

- **Filters Method:** In this method of variable ranking, we only consider the importance and usefulness of a feature.
- **Wrappers Method:** In this method, 'induction algorithm' is used, Which can be used to produce a classifier.
- **Embedded Method:** This method is a combination of efficiencies of both Filters and wrappers methods.

27. How can you restart NameNode and all the daemons in Hadoop?



The following commands will help you restart NameNode and all the daemons:

You can stop the NameNode with **./sbin /Hadoop-daemon.sh** stop NameNode command and then start the NameNode using **./sbin/Hadoop-daemon.sh** start NameNode command.You can stop all the daemons with the **./sbin/stop-all.sh** command and then start the daemons using the **./sbin/start-all.sh** command.

28. What is the use of the -compress-codec parameter?

-compress-codec parameter is generally used to get the output file of a sqoop import in formats other than .gz.

29. What are missing values in Big data? And how to deal with it?

Missing values in Big Data generally refer to the values which aren't present in a particular column, in the worst case they may lead to erroneous data and may provide incorrect results. There are several techniques used to deal with the missing values they are -

- **Mean or Median Imputation:**
  When data is missing at irregular intervals, we can use list-wise or pair-wise deletion of the missing observations. Still, there can be multiple reasons why this may not be the most workable option:
  - There may not be enough notes with non-missing data to produce a reliable analysis
  - In predictive analytics, missing data can prevent the forecasts for those observations which have missing data
  - External factors may require specific observations to be part of the analysis
    In such cases, we impute values for missing data. A simple technique is to use the mean or median of the non-missing observations. This can be useful in cases where the number of missing observations is low. However, for many missing values, using mean or median can result in loss of variation in data, and it is better to use imputations.
- **Multivariate Imputation by Chained Equations (MICE):**
  MICE believes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.
  To set up the data for MICE, it is essential to note that the algorithm uses all the variables in the data for predictions. In this case, variables that may not be useful for predictions, like the ID variable, should be removed before implementing this algorithm.

  ```
  Data$ID <- NULL
  ```

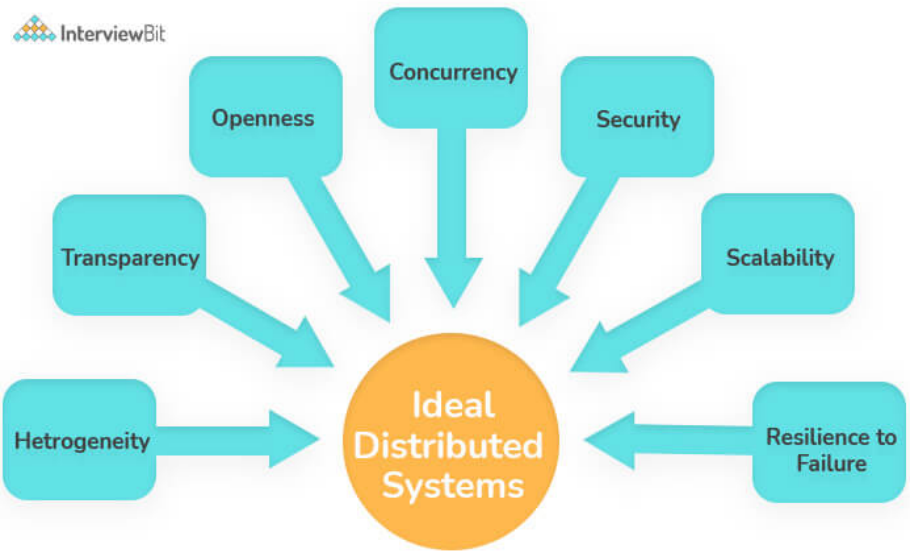
  Secondly, as mentioned above, the algorithm treats different variables differently. So, all categorical variables should be treated as factor variables before implementing MICE.

  ```
  Data$year <- as.factor(Data$year)
  Data$gender <- as.factor(Data$gender)
  ```

  Then you can implement the algorithm using the MICE library in R

  ```
  library(mice)
  init = mice(Data, maxit=0)
  method = init$method
  ```

30. What are the things to consider when using distributed cache in Hadoop MapReduce?



- **Heterogeneity:** The design of applications should allow the users to access services and run applications over a heterogeneous collection of computers and networks, considering hardware devices, OS, Network, Programming languages.
- **Transparency:** Distributed system Designers must hide the complexity of the system as much as they can. Some Terms of transparency are location, access, migration, relocation, and so on.
- **Openness:** It is a characteristic that determines whether the system can be extended and reimplemented in various ways.
- **Security:** Distributed system Designers must take care of confidentiality, integrity, and availability.
- **Scalability:** A system is said to be scalable if it can manage the increase of users and resources without undergoing a striking loss of performance.

31. Mention the main configuration parameters that has to be specified by the user to run MapReduce.

The chief configuration parameters that the user of the MapReduce framework needs to mention is:

- Job's input Location
- Job's Output Location
- The Input format
- The Output format
- The Class including the Map function
- The Class including the reduce function
- JAR file, which includes the mapper, the Reducer, and the driver classes.

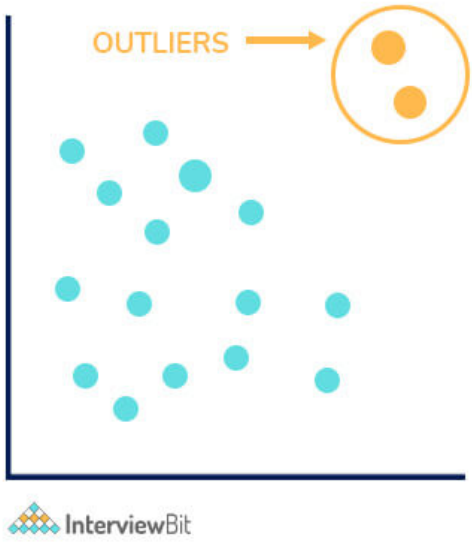
32. How can you skip bad records in Hadoop ?

Hadoop can provide an option wherein a particular set of lousy input records could be skipped while processing map inputs. SkipBadRecords class in Hadoop offers an optional mode of execution in which the bad records can be detected and neglected in multiple attempts. This may happen due to the presence of some bugs in the map function. The user has to manually fix it, which may sometimes be possible because the bug may be in third-party libraries. With the help of this feature, only a small amount of data is lost, which may be acceptable because we are dealing with a large amount of data.

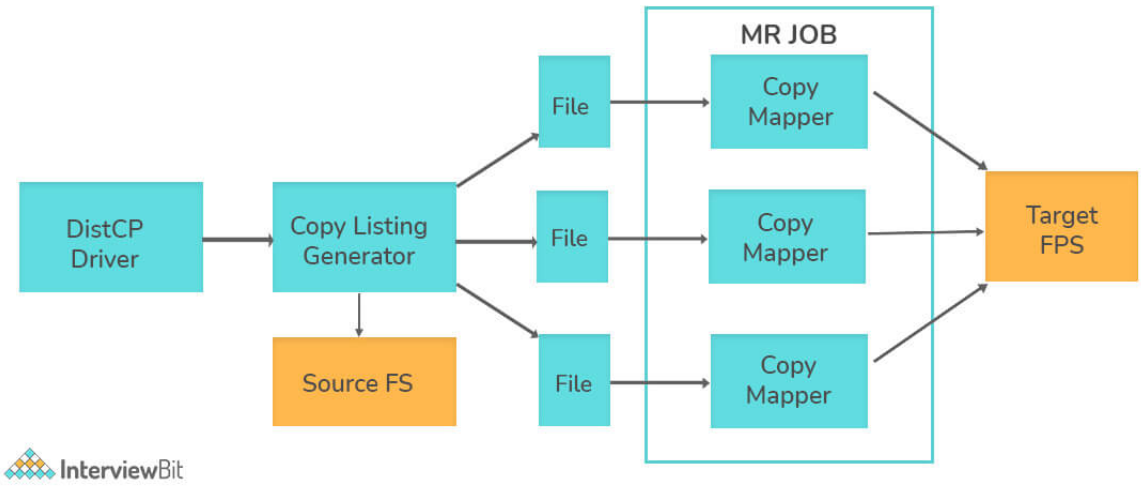
33. Explain Outliers.

Outliers are the data points that are very far from the group, which is not a part of any group or cluster. This may affect the behavior of the model, they may predict wrong results, or their accuracy will be very low. Therefore Outliers must be handled carefully as they may also contain some helpful information. The presence of these outliers may lead to misleading a Big Data model or a Machine Learning Model. The results of this may be,

- Poor Results
- Lower accuracy
- Longer Training Time



34. What is Distcp?



It is a Tool which is used for copying a very large amount of data to and from Hadoop file systems in parallel. It uses MapReduce to affect its distribution, error handling, recovery, and reporting. It expands a list of files and directories into input to map tasks, each of which will copy a partition of the files specified in the source list.

35. Explain Persistent, Ephemeral and Sequential Znodes.

- **Persistent znodes:** The default znode in ZooKeeper is the Persistent Znode. It permanently stays in the zookeeper server until any other clients leave it apart.
- **Ephemeral znodes:** These are the temporary znodes. It is smashed whenever the creator client logs out of the ZooKeeper server. For example, assume client1 created eznode1. Once client1 logs out of the ZooKeeper server, the eznode1 gets destroyed.
- **Sequential znodes:** Sequential znode is assigned a 10-digit number in numerical order at the end of its name. Assume client1 produced a sznode1. In the ZooKeeper server, the sznode1 will be named like this: sznode0000000001.
  If client1 generates another sequential znode, it will bear the following number in a sequence. So the subsequent sequential znode is <znode name>0000000002.

36. Explain the Pros and Cons of Big Data?

Pros of Big Data are:

- **Increased productivity:** Recently, it was found that 59.9% of businesses use big data tools like Hadoop and Spark to develop their sales. Current big data tools enable analysts to examine instantly, which enhances their productivity. Also, the insights inferred from the analysis of big data can be used by organizations to increase productivity in different forms throughout the company.
- **Reduce costs:** Big data analytics help businesses reduce their costs. In most companies, big data tools had served them to enhance operational performance and decrease costs, and in few other companies had started using big data to decrease expenses. Interestingly, very few companies selected cost reduction as their primary goal for big data analytics, suggesting that this is merely a very welcome side benefit for many.
- **Improved customer service:** Improving customer service has always been one of the primary goals for big data analytics projects, and it has been a success for many companies with the help of this. Various customer contact points like Social media, customer relationship management systems, etc., transfer a lot of information about their customers. And this analysis and data is used to improve the services for the customers
- **Fraud detection:** The primary purpose of using Big data analytics is in the financial services industry for detecting frauds. The advantage of big data analytics systems is that it depends on machine learning, because of which they are great at recognizing patterns and irregularities. As a result, these techniques can give banks and credit card companies the capacity to detect stolen credit cards or deceitful purchases, usually before the cardholder knows that something is wrong.
- **More significant innovation:** A few companies have started investing in analytics with the sole purpose to bring new things and disturb their markets. The reason behind this is if they can see the future of the market with the help of insights before their competitors, they can come out strong from that situation with a few new goods and services and capture the market quickly.

On the other hand, implementing big data analytics is not as easy as we think; there are a few difficulties too when it comes to implementing it.

Cons of Big Data are:

- **Need for talent:** The number one big data challenge that we have been facing for the past three years is the skill set required for it. A lot of companies also face difficulty when designing a data lake. Hiring or training staff will only increase the cost considerably, and also, imbibing big data skills takes a lot of time.
- **Cybersecurity risks:** Storing, especially sensitive big data will make those businesses a prime target for cyberattackers. Security is one of the top big data challenges, and cybersecurity breaches are the single greatest data threat that enterprises encounter.
- **Hardware needs:** Another critical concern for businesses is the IT base necessary to help big data analytics drives. Storage space for storing the data, networking bandwidth for transferring it to and from analytics systems, and calculating resources to achieve those analytics are costly to buy and keep.
- **Data quality:** The disadvantage in working with big data was the requirement to address data quality problems. Before companies can use big data for analytics purposes, data scientists and analysts need to ensure that the data they are working with is accurate, appropriate, and in the proper format for analysis. This slows the process, but if companies don't take care of data quality issues, they may find that the insights produced by their analytics are useless or even harmful if performed.

37. How do you convert unstructured data to structured data?

An open-ended question and there are many ways to achieve this.

- **Programming:** Coding/ Programming is the most tried out method to transform unstructured data into a structured form. Programming is advantageous to accomplish because we get independence with it, which you can use to change the structure of the data in any form possible. Several programming languages, such as Python, Java, etc., can be used.
- **Data/Business Tools:** Many BI (Business Intelligence) tools support the drag and drop functionality for converting unstructured data into structured data. One cautious thing before using BI tools is that most of these tools are paid, and we have to be financially capable to support these tools. For people who lack both experience and skills needed for option 1, this is the way to go.

38. What is data preparation?

Data preparation is the method of cleansing and modifying raw data before processing and analyzing it. It is a crucial step before processing and usually requires reformatting data, making improvements to data, and consolidating data sets to enrich data.

Data preparation is an unending task for data specialists or business users. But, it is essential to convert data into context to get insights and then, can eliminate the biased results found due to poor data quality.

For instance, the data construction process typically includes standardizing data formats, enhancing source data, and/or eliminating outliers.

39. Steps for Data preparation.

Steps for data preparation are:

- **Gather data:** The data preparation process starts with obtaining the correct data. This can originate from a current data catalogue or can be appended ad-hoc.
- **Discover and assess data:** After assembling the data, it is essential for each dataset to be identified. This step is about learning to understand the data and knowing what must be done before the data becomes valuable in a distinct context. Discovery is a big task but can be done with the help of data visualization tools that assist users and help them browse their data.
- **Clean and verify data:**
  Even though cleaning and verifying data takes a lot of time, it is the most important step, because this step not only eliminates the incorrect data but also fills the rifts. Significant tasks here include:
  - Eliminating alien data and outliers.
  - Filling in missing values.
  - Adjusting data to a regulated pattern.
  - Masking private or sensitive data entries.
    After cleaning the data, the mistakes that we came across during the data development process have to be examined and approved. Generally, an error in the system will become apparent during this step and need to be fixed before proceeding.
- **Transform and enrich data:**
  Transforming data modernizes the arrangement or value entries to reach a well-defined result or make the data more quickly recognized by broader viewers. Improving data refers to adding and connecting data with other similar information to provide deeper insights.
- **Store data:**
  Lastly, the data can be collected or channeled into a third-party application like a business intelligence tool-making technique for processing and analysis.

Links to More Interview Questions

C Interview Questions	Php Interview Questions	C Sharp Interview Questions
Web Api Interview Questions	Hibernate Interview Questions	Node Js Interview Questions
Cpp Interview Questions	Oops Interview Questions	Devops Interview Questions
Machine Learning Interview Questions	Docker Interview Questions	Mysql Interview Questions
Css Interview Questions	Laravel Interview Questions	Asp Net Interview Questions
Django Interview Questions	Dot Net Interview Questions	Kubernetes Interview Questions
Operating System Interview Questions	React Native Interview Questions	Aws Interview Questions
Git Interview Questions	Java 8 Interview Questions	Mongodb Interview Questions
Dbms Interview Questions	Spring Boot Interview Questions	Power Bi Interview Questions
Pl Sql Interview Questions	Tableau Interview Questions	Linux Interview Questions
Ansible Interview Questions	Java Interview Questions	Jenkins Interview Questions