BDA VIVA QUESTIONS

*******************These are the imp viva ques*******************

1) Question: Write Big Data Characteristics.

Volume, Velocity, Variety, Veracity, and Value.

2) Question: Explain Types of Big Data.

Structured, Semi-structured, Unstructured.

3) Question: Difference Traditional vs. Big Data business approach.

Traditional relies on structured data, while Big Data includes unstructured data for deeper insights.

4) Question: Discuss MapReduce.

Programming model for processing large datasets in parallel.

5) Question: What is noSQL?

A type of database management system designed for unstructured and semi-structured data.

6) Question: NoSQL master-slave versus peer-to-peer.

Master-Slave has a single primary node, while Peer-to-Peer has equal nodes with no single point of control.

7) Question: What are the sampling data techniques?

Random sampling, stratified sampling, systematic sampling, and cluster sampling.

8) Question: What is the Count Distinct Problem?

Finding the number of unique elements in a dataset efficiently.

9) Question: What are decaying windows?

Windows in streaming data that give more weight to recent data.

10) Question: Describe built-in functions in R.

Functions like mean, median, sum, and sd for data analysis in R.

*******************These were the imp viva ques*******************

1. Difference between Task Tracker and Job Tracker

The Task Tracker is responsible for executing individual tasks on DataNodes in a Hadoop cluster, while the Job Tracker manages and coordinates these tasks, maintaining information about job progress and overall cluster resource management.

2. Any 2 components of Hadoop ecosystem

Two components of the Hadoop ecosystem are HDFS (Hadoop Distributed File System) and MapReduce.

3. Who created Hadoop

Hadoop was created by Doug Cutting and Mike Cafarella.

4. What is Hive

Hive is a data warehousing and SQL-like query language for Hadoop. It provides a higher-level abstraction for querying and managing data stored in Hadoop.

5. What is HDFS

HDFS (Hadoop Distributed File System) is the primary storage system used by Hadoop. It is designed to store and manage vast amounts of data across a distributed cluster of commodity hardware.

6. What is DIS

It's unclear what "DIS" refers to in the context of Big Data Analytics. Please provide more information or clarify.

7. What is MapReduce

MapReduce is a programming model and processing framework used in Hadoop for processing and generating large datasets. It divides tasks into smaller sub-tasks that can be processed in parallel.

8. What is Oozie

Oozie is a workflow scheduler for Hadoop jobs. It allows you to create and manage workflows that coordinate the execution of various Hadoop processes.

9. What is the purpose of using the MapReduce approach

The purpose of using MapReduce is to process and analyze large datasets in a parallel and distributed manner, making it suitable for big data analytics. It enables efficient and scalable data processing.

10. Applications of BDA (Big Data Analytics)

Applications of BDA include customer analytics, fraud detection, recommendation systems, sentiment analysis, healthcare analytics, and more.

11. What is the use of a Combiner in MapReduce

A Combiner in MapReduce is used to perform a local aggregation of data before sending it to the reducer. It helps reduce data transfer between the map and reduce tasks, improving overall performance.

12. Differentiate between normal RDBMS and Big Data DBMS

RDBMS (Relational Database Management System) follows a structured schema, whereas Big Data DBMS is schema-less and designed to handle unstructured and semi-structured data. Big Data DBMS is also distributed and highly scalable.

13. What is the use of ZooKeeper in Hadoop

ZooKeeper is used in Hadoop for distributed coordination, synchronization, and maintaining configuration information. It helps manage the distributed aspects of Hadoop clusters.

14. What is YARN

YARN (Yet Another Resource Negotiator) is the resource management and job scheduling component in Hadoop. It separates the resource management and job scheduling functions, improving cluster resource utilization.

15. What was the flaw in Hadoop 1 that led to the introduction of YARN

Hadoop 1 had a centralized Job Tracker that became a bottleneck for large clusters. The introduction of YARN addressed this limitation by decentralizing resource management and job scheduling.

16. Advantages of NoSQL over SQL

Advantages of NoSQL databases over SQL databases include flexibility in handling unstructured data, horizontal scalability, and better performance for certain types of applications.

17. Explain any NoSQL database (e.g., MongoDB)

MongoDB is a document-oriented NoSQL database. It stores data in flexible, JSON-like BSON format documents and is known for its scalability, flexibility, and ease of use.

18. What is a key-value store

A key-value store is a NoSQL database that stores data in a simple key-value format, making it efficient for retrieval. Examples include Redis and Amazon DynamoDB.

19. Difference between RDBMS and NoSQL

RDBMS enforces a fixed schema, while NoSQL databases have a flexible or schema-less approach. RDBMS is best suited for structured data, while NoSQL databases handle unstructured and semi-structured data.

20. Difference between ACID and BASE

ACID (Atomicity, Consistency, Isolation, Durability) is a set of properties for ensuring data integrity, while BASE (Basically Available, Soft state, Eventually consistent) is a set of properties for distributed systems that prioritize availability and performance over strict consistency.

21. What is "B" in BASE, and what does "Basic Availability" mean

The "B" in BASE stands for "Basic Availability." It means that the system remains available for read and write operations, even when some components or nodes may fail. It doesn't guarantee immediate consistency.

22. Features of Graph Stores

Graph stores are designed for managing and querying graph data. Their features include the ability to represent complex relationships, perform graph traversals, and support queries that involve nodes and edges.

23. Where is Graph Store used (e.g., social media)

Graph stores are commonly used in social media applications to represent and analyze social networks, friendships, and connections between users.

24. Difference between Row Stores and Column Stores

Row stores store data in rows, which is efficient for retrieving entire records. Column stores store data in columns, which is better for analytics where only specific columns are needed. The choice depends on the use case.

25. Significance of a Stream and Difference between a Database and a Data Stream

A stream is a continuous flow of data in real-time. Unlike databases, which store historical data, data streams process and analyze data as it arrives, making them suitable for real-time analytics and monitoring.

26. CAP Properties

CAP properties refer to Consistency, Availability, and Partition Tolerance. In distributed systems, the CAP theorem states that it's impossible to achieve all three properties simultaneously; you can prioritize two out of the three.

27. What is a Bloom Filter, and where is it used

A Bloom Filter is a data structure used to test whether an element is a member of a set. It's often used in applications like caching, network routers, and databases to quickly eliminate non-existent elements.

28. Amazon uses which NoSQL database in the cart

Amazon uses Amazon DynamoDB, a managed NoSQL database service, for its shopping cart and other applications.

29. WhatsApp uses which NoSQL database

WhatsApp uses WhatsApp uses a custom-built, Erlang-based, and highly distributed NoSQL database for its messaging platform.

30. What is HBASE

HBase is a NoSQL database that provides real-time, random read and write access to large datasets. It's built on top of HDFS and is often used for applications requiring low-latency data access.

31. What do you mean when we say NoSQL follows a schema-less architecture

When we say NoSQL follows a schema-less architecture, it means that NoSQL databases allow you to store data without a predefined schema, offering flexibility to handle various data structures without rigid constraints.

32. MongoDB is what kind of NoSQL database

MongoDB is a document-oriented NoSQL database. It stores data in BSON (Binary JSON) documents and is known for its flexibility and scalability.

33. Use of Flajolet-Martin algorithm

The Flajolet-Martin algorithm is used for estimating the cardinality(unique values of dataset).

34. What is a stream data model, and how does it differ from traditional batch processing?

A stream data model is a way of processing data continuously as it arrives, in contrast to traditional batch processing, which processes data in fixed-sized chunks.

35. Can you explain the concept of data streams and give examples of applications where they are used?

Data streams are continuous, potentially infinite sequences of data elements, and they are used in applications like social media analytics and IoT sensor data.

36. How do you handle out-of-order data in a stream data model?

Out-of-order data in stream processing can be handled using techniques like event time timestamps and watermarking.

37. How can you evaluate the performance of a stream processing system?

Performance of a stream processing system can be evaluated by measuring latency, throughput, and scalability, among other factors.

38. Various Ways to Find Distance Measures:

Euclidean Distance:

Euclidean distance is a measure of the straight-line distance between two points in a Euclidean space.

Formula: $\sqrt((x2 - x1)^2 + (y2 - y1)^2$

Jaccard Distance:

Jaccard distance is a metric used to measure the dissimilarity between two sets.

Formula: Similarity = $|A \cap B| / |A \cup B|$

Cosine Distance:

Cosine similarity is a metric used to measure the similarity between two non-zero vectors.

Formula: Similarity = $(a \cdot b) / (|a| * |b|)$

It calculates the cosine of the angle between vectors a and b in a multidimensional space.

Edit Distance:

Edit distance, also known as Levenshtein distance, measures the minimum number of edit operations (insertion, deletion, substitution) required to transform one string into another.

There are two primary ways to calculate edit distance:

Longest Common Subsequence: The edit distance is calculated by finding the longest common subsequence between two strings. The formula is the sum of the lengths of the two strings minus twice the length of their longest common subsequence

.Classical Method: The edit distance is calculated by considering individual character insertions, deletions, and substitutions. It involves dynamic programming to find the minimum number of operations.

Hamming Distance:

Hamming distance is a metric used to measure the dissimilarity between two binary strings of equal length.

Formula: $H(a, b) = \sum(i=1 \text{ to } n)\ (a\_i \neq b\_i)$

39. What is the primary objective of the CURE algorithm, and how does it contribute to addressing challenges in data clustering?

The CURE algorithm aims to improve data clustering by addressing the limitations of traditional hierarchical clustering methods. It introduces representative points to efficiently cluster large datasets.

40. Walk me through the key steps of the CURE algorithm, including how it identifies representative points in clusters.

Key steps in the CURE algorithm include selecting representative points, clustering these representatives, and creating a hierarchy of clusters. Outliers are addressed by considering the nearest cluster.

41. In what ways does CURE overcome the scalability issues commonly associated with clustering large datasets?

CURE overcomes scalability issues through representative-based clustering and hierarchical structure.

It is suitable for applications with large datasets, such as spatial databases and image processing

42. Can you discuss the advantages and potential limitations of the CURE algorithm, and how it handles outliers in clustering?

Advantages of the CURE algorithm include improved scalability and the ability to handle arbitrary shaped clusters. Limitations include the need for parameter tuning and sensitivity to noise in data.

43. What makes the PageRank algorithm significant in the context of web search and ranking, and how does it work in the context of hyperlink structures?

The PageRank algorithm is significant in web search and ranking because it determines the importance of web pages by analyzing their links and the structure of the web. It plays a fundamental role in providing users with relevant search results.

44.. Explain the role of the damping factor in the PageRank algorithm and how it influences ranking.

The damping factor is a parameter in the PageRank algorithm that models the likelihood that a user may randomly jump to any page rather than following links. It helps prevent infinite loops and ensures more accurate rankings.

45. How does PageRank handle the challenge of link spam and black-hat SEO techniques, and what measures can be taken to ensure the algorithm's integrity?

PageRank handles the challenge of link spam and black-hat SEO techniques by assigning lower importance to low-quality or manipulated web pages. Measures like link quality and trustworthiness are used to counteract spam.

45. Are there any modern variations or extensions of the PageRank algorithm, and how do they address evolving challenges in web ranking and recommendation systems?

Modern variations of the PageRank algorithm include personalized PageRank, topic-sensitive PageRank, and PageRank with trust features, which enhance the algorithm's performance in specialized search and recommendation systems.

46. DGIM:

1. D-G-I-M algorithm estimates distinct elements in data streams efficiently with limited memory usage.

2. It uses "logarithmic counting" with exponentially increasing bucket sizes to provide approximate results.

3. The algorithm maintains buckets to track frequency counts, efficiently estimating distinct items in the stream.

4. D-G-I-M offers low memory usage and scalability but provides only approximate results, with trade-offs between memory efficiency and accuracy.

50. Creators, Datar, Gionis, Indyk, and Motwani, is used for approximate counting of distinct elements in data streams.