MASTERING
DATA ENGINEERING
THE
ULTIMATE
TOP 25 GUIDE!

Prepared By
Afrin Ahamed

1.Differentiate between relational and non-relational database management systems.

Relational Database Management Systems (RDBMS)	Non-relational Database Management Systems
Relational Databases primarily work with structured data using SQL (Structured Query Language). SQL works on data arranged in a predefined schema.	Non-relational databases support dynamic schema for unstructured data. Data can be graph-based, column-oriented, document-oriented, or even stored as a Key store.
RDBMS follow the ACID properties - atomicity, consistency, isolation, and durability.	Non-RDBMS follow the Brewers Cap theorem - consistency, availability, and partition tolerance.
RDBMS are usually vertically scalable. A single server can handle more load by increasing resources such as RAM, CPU, or SSD.	Non-RDBMS are horizontally scalable and can handle more traffic by adding more servers to handle the data.
Relational Databases are a better option if the data requires multi-row transactions to be performed on it since relational databases are table-oriented.	Non-relational databases are ideal if you need flexibility for storing the data since you cannot create documents without having a fixed schema. Since non-RDBMS are horizontally scalable, they can become more

	powerful and suitable for large or constantly changing datasets.
E.g. PostgreSQL, MySQL, Oracle, Microsoft SQL Server.	E.g. Redis, MongoDB, Cassandra, HBase, Neo4j, CouchDB

2. What is data modeling?

Data modeling is a technique that defines and analyzes the data requirements needed to support business processes. It involves creating a visual representation of an entire system of data or a part of it. The process of data modeling begins with stakeholders providing business requirements to the data engineering team.

3. How is a data warehouse different from an operational database?

Data warehouse	Operational database
Data warehouses generally support high-volume analytical data processing - OLAP.	Operational databases support high-volume transaction processing, typically - OLTP.
You may add new data regularly, but once you add the data, it does not change very frequently.	Data is regularly updated.

Data warehouses are optimized to handle complex queries, which can access multiple rows across many tables.	Operational databases are ideal for queries that return single rows at a time per table.
There is a large amount of data involved.	The amount of data is usually less.
A data warehouse is usually suitable for fast retrieval of data from relatively large volumes of data.	Operational databases are optimized to handle fast inserts and updates on a smaller scale of data.

4.What are the big four V's of big data?

- Volume: refers to the size of the data sets to be analyzed or processed. The size is generally in terabytes and petabytes.
- Velocity: the speed at which you generate data. The data generates faster than traditional data handling techniques can handle it.
- Variety: the data can come from various sources and contain structured, semi-structured, or unstructured data.
- Veracity: the quality of the data to be analyzed. The data has to be able to contribute in a meaningful way to generate results.

5.Differentiate between Star schema and Snowflake schema.

Star schema	Snowflake Schema

Star schema is a simple top-down data warehouse schema that contains the fact tables and the dimension tables.	The snowflake schema is a bottom-up data warehouse schema that contains fact tables, dimension tables, and sub-dimension tables.
Takes up more space.	Takes up less space.
Takes less time for query execution.	Takes more time for query execution than star schema.
Normalization is not useful in a star schema, and there is high data redundancy.	Normalization and denormalization are useful in this data warehouse schema, and there is less data redundancy.
The design and understanding are simpler than the Snowflake schema, and the Star schema has low query complexity.	The design and understanding are a little more complex. Snowflake schema has higher query complexity than Star schema.
There are fewer foreign keys.	There are many foreign keys.

6. What are the differences between OLTP and OLAP?

OLTP (Online Transaction Processing) Systems	OLAP (Online Analytical Processing) Systems
System for modification of online databases.	System for querying online databases.
Supports insert, update and delete transformations on the database.	Supports extraction of data from the database for further analysis.
OLTP systems generally have simpler queries that require less transactional time.	OLAP queries generally have more complex queries which require more transactional time.
Tables in OLTP are normalized.	Tables in OLAP are not normalized.

7.What are some differences between a data engineer and a data scientist?

Data engineers and data scientists work very closely together, but there are some differences in their roles and responsibilities.

Data Engineer	Data scientist
The primary role is to design and implement highly maintainable database management systems.	The primary role of a data scientist is to take raw data presented on the data and apply analytic tools and modeling techniques to analyze the data and provide insights to the business.
Data engineers transform the big data into a structure that one can analyze.	Data scientists perform the actual analysis of Big Data.
They must ensure that the infrastructure of the databases meets industry requirements and caters to the business.	They must analyze the data and develop problem statements that can process the data to help the business.
Data engineers have to take care of the safety, security and backing up of the data, and they work as gatekeepers of the data.	Data scientists should have good data visualization and communication skills to convey the results of their data analysis to various stakeholders.
Proficiency in the field of big data, and strong database management skills.	Proficiency in machine learning is a requirement.

A data scientist and data engineer role require professionals with a computer science and engineering background, or a closely related field such as mathematics, statistics, or economics. A sound command over software and programming languages is important for a data scientist and a data engineer.

8. How is a data architect different from a data engineer?

Data architect	Data engineers
Data architects visualize and conceptualize data frameworks.	Data engineers build and maintain data frameworks.
Data architects provide the organizational blueprint of data.	Data engineers use the organizational data blueprint to collect, maintain and prepare the required data.
Data architects require practical skills with data management tools including data modeling, ETL tools, and data warehousing.	Data engineers must possess skills in software engineering and be able to maintain and build database management systems.
Data architects help the organization understand how changes in data acquisitions will impact the data in use.	Data engineers take the vision of the data architects and use this to build, maintain and process the architecture for further use by other data professionals.

9. Differentiate between structured and unstructured data.

Structured Data	Unstructured Data
Structured data usually fits into a predefined model.	Unstructured data does not fit into a predefined data model.
Structured data usually consists of only text.	Unstructured data can be text, images, sounds, videos, or other formats.
It is easy to query structured data and perform further analysis on it.	It is difficult to query the required unstructured data.
Relational databases and data warehouses contain structured data.	Data lakes and non-relational databases can contain unstructured data. A data warehouse can contain unstructured data too.

10. How does Network File System (NFS) differ from Hadoop Distributed File System (HDFS)?

Network File System	Hadoop Distributed File System

NFS can store and process only small volumes of data.	~~Hadoop Distributed File System~~, or HDFS, primarily stores and processes large amounts of data or Big Data.
The data in an NFS exists in a single dedicated hardware.	The data blocks exist in a distributed format on local hardware drives.
NFS is not very fault tolerant. In case of a machine failure, you cannot recover the data.	HDFS is fault tolerant and you may recover the data if one of the nodes fails.
There is no data redundancy as NFS runs on a single machine.	Due to replication across machines on a cluster, there is data redundancy in HDFS.

11. What is meant by feature selection?

~~Feature selection~~ is identifying and selecting only the features relevant to the prediction variable or desired output for the model creation. A subset of the features that contribute the most to the desired output must be selected automatically or manually.

12. How can missing values be handled in Big Data?

Some ways you can handle missing values in Big Data are as follows:

- Deleting rows with missing values: You simply delete the rows or columns in a table with missing values from the dataset. You can drop the entire column from the analysis if a column has more than half of the rows with null values. You can use a similar method for rows with missing values in more than half of the columns. This method may not work very well in cases where a large number of values are missing.

- Using Mean/Medians for missing values: In a dataset, the columns with missing values and the column's data type are numeric; you can fill in the missing values by using the median or mode of the remaining values in the column.
- Imputation method for categorical data: If you can classify the data in a column, you can replace the missing values with the most frequently used category in that particular column. If more than half of the column values are empty, you can use a new categorical variable to place the missing values.
- Predicting missing values: Regression or classification techniques can predict the values based on the nature of the missing values.
- Last Observation carried Forward (LCOF) method: The last valid observation can fill in the missing value in data variables that display a longitudinal behavior.
- Using Algorithms that support missing values: Some algorithms, such as the k-NN algorithm, can ignore a column if values are missing. Another such algorithm is Naive Bayes. The RandomForest algorithm can work with non-linear and categorical data.

13. What is meant by outliers?

In a dataset, an outlier is an observation that lies at an abnormal distance from the other values in a random sample from a particular data set. It is left up to the analyst to determine what can be considered abnormal. Before you classify data points as abnormal, you must first identify and categorize the normal observations. Outliers may occur due to variability in measurement or a particular experimental error. Outliers must be identified and removed before further analysis of the data not to cause any problems.

14. What is meant by logistic regression?

Logistic regression is a classification rather than a regression model, which involves modeling the probability of a discrete outcome given an input variable. It is a simple and efficient method that can approach binary and linear classification problems. Logistic regression is a statistical method that works well with binary classifications but can be generalized to multiclass classifications.

15. Briefly define the Star Schema.

The star join schema, one of the most basic design schemas in the Data Warehousing concept, is also known as the star schema. It looks like a star, with fact tables and related dimension tables. The star schema is useful when handling huge amounts of data.

16. Briefly define the Snowflake Schema.

The snowflake schema, one of the popular design schemas, is a basic extension of the star schema that includes additional dimensions. The term comes from the way it resembles the structure of a snowflake. In the snowflake schema, the data is organized and, after normalization, divided into additional tables

.

17. What is the difference between the KNN and k-means methods?

- The k-means method is an unsupervised learning algorithm used as a clustering technique, whereas the K-nearest-neighbor is a supervised learning algorithm for classification and regression problems.
- KNN algorithm uses feature similarity, whereas the K-means algorithm refers to dividing data points into clusters so that each data point is placed precisely in one cluster and not across many.

18. What is the purpose of A/B testing?

A/B testing is a randomized experiment performed on two variants, ‘A’ and ‘B.’ It is a statistics-based process involving applying statistical hypothesis testing, also known as “two-sample hypothesis testing.” In this process, the goal is to evaluate a subject’s response to variant A against its response to variant B to determine which variants are more effective in achieving a particular outcome.

19. What do you mean by collaborative filtering?

Collaborative filtering is a method used by recommendation engines. In the narrow sense, collaborative filtering is a technique used to automatically predict a user's tastes by collecting various information regarding the interests or preferences of many other users. This technique works on the logic that if person 1 and person 2 have the same opinion on one particular issue, then person 1 is likely to have the same opinion as person 2 on another issue than another random person. In general, collaborative filtering is the process that filters information using techniques involving collaboration among multiple data sources and viewpoints.

20. What are some biases that can happen while sampling?

Some popular type of bias that occurs while sampling is

- Undercoverage- The undercoverage bias occurs when there is an inadequate representation of some members of a particular population in the sample.
- Observer Bias- Observer bias occurs when researchers unintentionally project their expectations on the research. There may be occurrences where the researcher unintentionally influences surveys or interviews.
- Self-Selection Bias- Self-selection bias, also known as volunteer response bias, happens when the research study participants take control over the decision to participate in the survey. The individuals may be biased and are likely to share some opinions that are different from those who choose not to participate. In such cases, the survey will not represent the entire population.
- Survivorship Bias- The survivorship bias occurs when a sample is more concentrated on subjects that passed the selection process or criterion and ignore the subjects who did not pass the selection criteria. Survivorship biases can lead to overly optimistic results.
- Recall Bias- Recall bias occurs when a respondent fails to remember things correctly.

- Exclusion Bias- The exclusion bias occurs due to the exclusion of certain groups while building the sample.

21. What is a distributed cache?

A distributed cache pools the RAM in multiple computers networked into a single in-memory data store to provide fast access to data. Most traditional caches tend to be in a single physical server or hardware component. Distributed caches, however, grow beyond the memory limits of a single computer as they link multiple computers, providing larger and more efficient processing power. Distributed caches are useful in environments that involve large data loads and volumes. They allow scaling by adding more computers to the cluster and allowing the cache to grow based on requirements.

22. Explain how Big Data and Hadoop are related to each other.

Apache Hadoop is a collection of open-source libraries for processing large amounts of data. Hadoop supports distributed computing, where you process data across multiple computers in clusters. Previously, if an organization had to process large volumes of data, it had to buy expensive hardware. Hadoop has made it possible to shift the dependency from hardware to achieve high performance, reliability, and fault tolerance through the software itself. Hadoop can be useful when there is Big Data and insights generated from the Big Data. Hadoop also has robust community support and is evolving to process, manage, manipulate and visualize Big Data in new ways.

23. Briefly define COSHH.

COSHH is an acronym for Classification and Optimization-based Scheduling for Heterogeneous Hadoop systems. As the name implies, it offers scheduling at both the cluster and application levels to speed up job completion.

24. Give a brief overview of the major Hadoop components.

Working with Hadoop involves many different components, some of which are listed below:

- Hadoop Common: This comprises all the tools and libraries typically used by the Hadoop application.
- Hadoop Distributed File System (HDFS): When using Hadoop, all data is present in the HDFS, or Hadoop Distributed File System. It offers an extremely high bandwidth distributed file system.
- Hadoop YARN: The Hadoop system uses YARN, or Yet Another Resource Negotiator, to manage resources. YARN can also be useful for task scheduling.
- Hadoop MapReduce: Hadoop MapReduce is a framework for large-scale data processing that gives users access.

25. List some of the essential features of Hadoop.

- Hadoop is a user-friendly open source framework.
- Hadoop is highly scalable. Hadoop can handle any sort of dataset effectively, including unstructured (MySQL Data), semi-structured (XML, JSON), and structured (MySQL Data) (Images and Videos).
- Parallel computing ensures efficient data processing in Hadoop.
- Hadoop ensures data availability even if one of your systems crashes by copying data across several DataNodes in a Hadoop cluster.

THANK

Y🌍U