Research Assignment

SECTION A : Database Fundamentals

(1) What are the main types of databases?

→ Relational Database RDBMS

→ NON - Relational Database

→ Object - oriented Database

→ Time - series Database

→ Graph Database

→ Hierarchical and Network Database

(2) What is a Relational Database system (RDBMS)?

→ is a database that organizes data into structured tables made up of rows and columns, where relationships between data points are defined using keys.

(3) What is a primary key and foreign key in a database?

→ Primary key → ensures data integrity by preventing duplicate records.

- it uniquely identifies each record in a table.

→ foreign key → is a system that links one table to another establishing a relationship between them.

→ maintains referential integrity, ensuring that relationships between tables are valid.

(4) What is database normalization and why is it important?

→ Database normalization is a design technique used in relational database to structure data efficiently.

→ it is a systematic process of organizing the data in a relational database to minimize redundancy and improve data integrity and consistency.

(4) → Its important because it reduces data redundancy and prevents storing the same data in multiple places, which saves space and avoids duplication
→ Prevents Data Anomalies

(5) What is a database schema?
→ It is the blueprint that defines how data is organized in a database, including tables, fields and contraints.
→ it acts as a structural framework for managing and accessing data efficiently

(6) Differentiate between structured, semi-structured and unstructured data?
→ Structured data → is highly organized and stored into rows and columns in a tabular format.
→ Semi-structured Data - has some organization but does not fit neatly into tables.
→ Unstructured Data → has no consistent format meaning it exist in its raw, native form.

(7) What is the difference between a fact table and a Dimension Table in a data warehouse?
→ A Fact Table - stores measurable quantative data

→ A Dimension Table - provides descriptive context for those attributes Together the fact and dimensional tables form the backbone of data Warehouse schemas like star and snowflake.

(8) What is a data model and why is it important in data warehouse and a data lake?

⇒ A data model is a conceptual map or blueprint that organizers data elements and defines their relationships to one another and to the properties of real-world entities.

* Its important in the data warehouse and data lake because
  (i) it organizes data into fact and dimension tables
  (ii) improves query performance by optimizing joins and indexes.
  (iii) it adds structure to chaos by tagging and cataloging data.
  (iv) it enables data (~~integrity~~) discovery through metadata and schema on read.
  (v) supports governance and compliance by defining data lineage and access rules.

(9) Explain the difference between a database, a data warehouse and a data lake

⇒ A data warehouse organizes structured data for analytics and reporting.

⇒ A data lake holds raw, unstructured or semi-structured data for flexible exploration and advance use cases like machine learning

(10) What is a data mart and how does it differ from a data warehouse

⇒ A data mart is a smaller, focused version of a data warehouse that serves the needs of a specific business unit like marketing or finance

→ While a data warehouse is a large, centralized system that stores data from across the entire organisation for broad analytics.

Section B SQL AND Data Processing:

(11) What is a query language and why is SQL the most commonly used?

→ A query language is a way to communicate with databases to retriere, insert, update or delete data.

→ SQL is most commonly used because its standardized, powerful and supported by nearly all relational databases.

(12) What are indexes in databases and how do they improve performance?

→ Indexes in databases are special data structures that help speed up data retrieval.

→ They improve performances by allowing the database to locate rows without scanning the entire table.

(13) What are transactions in databases and what are the ACID properties?

→ Transactions in database are sequences of operation that are treated as a single unit of work.

→ And ACID properties ensure reliable and consistent transactions

(14) What is a database engine and how does it impact performance?

→ A database engine is the core software component that handles how data is stored, retrieved and managed in a database.

- It directly impacts performance by determining how efficiently queries run, how data is indexed and how transtions are processed.

(15) What are views, stored procedures and triggers in SQL?

→ View is a virtual table based on a SQL query, it simplifies complex queries by saving them as reusable objects

-> Stored Procedures is a saved block of SQL code that performs a task or set of tasks.
-> It automates repetitive operations like inserting data, updating records or of generating reports.

-> Triggers are a piece of code that runs automatically in response to certain events
-> It enforces rules or logs changes without manually intervention.

(16) Differentiate between ETL (Extract, Transform, Load) and ELT Extract, Load, Transform)?

-> ETL	ELT
- It requires a specialized and independent transformation engine. - extracts data, transforms it using a secondary processor and loads it	- The back end data warehouse must have the computational and processing capabilities to perform transformations. - Extracts data, loads it and transforms it within the system

(17) Differentiate between batch processing and stream processing in data pipelines?
- Batch processing handles large volumes of data at once.
- While stream processing handles data continuously in real time.
* The key difference is timing - batch is delayed -> Stream is immediate.

(18) Explain what a join is in SQL and list different types of joins with examples?
- A JOIN in SQL is used to combine rows from two or more tables based on a related column between them.
- It helps you retrieve meaningful data spread across multiple tables.

Types of JOINS.

(i) INNER JOIN → Returns only matching rows for both tables.

(ii) LEFT JOIN → Returns all rows from the left table and matching rows from the left.

(iii) FULL OUTER JOIN → Returns all rows whether there is a match or not.

(iv) Right JOIN → Returns all rows from the right table and matching rows from the left.

(v) CROSS JOIN → Returns the Cartesian product - every row from one table paired with every row from the other.

(vi) SELF JOIN → JOINS a table to itself.

(19) What is referential integrity and why is it important in relational databases?

→ Referential integrity ensures that relationships between tables in a relational database remain accurate and consistent.

- It prevents broken links between records by enforcing rules around foreign keys.

(20) How does data redundancy affect database performance and storage?

→ Data redundancy negatively affects database performance by increasing storage usage, slowing down updates and risking data inconsistencies.

SECTION C : Data Management and Analytics Concepts

(21) How does cloud database management differ from on-premise database?

⇒ Cloud database management is hosted and maintained by third-party providers over the internet, offering scalability and flexibility, while on-premise databases are installed and managed locally by an organization

→ offering more control but requiring more resources.

(22) What is data governance, and why is it important in data management?

⇒ Data governance is the discipline of managing~~t~~ data availibity, quality, security and usability across an organization

⇒ It's essential in data management because it ensures that data is trustworthy, compliant and aligned with business goals

(23) What is data integrity and how can it be maintained?

⇒ Data integrity means ensuring that data is accurate, consistent and reliable throughout its lifecycle.

⇒ it can be maintained through validation rules, access controls, backups and strong database design.

(24) What is data quality and why is it critical for analytics?

- Data quality refers to how accurate, complete, consistent, timely and relevant your data is
- It's critical for analytics because poor-quality data leads to misleading insights, flawed decisions and wasted resources.

(25) Explain the role of a Data Analyst in managing and analyzing database information?

-> A data analyst plays a pivotal role in transforming raw database information into actionable insights that drive business decisions.
-> Their work bridges the gap between data storage and strategic execution.

(26) What are the key responsibilities of a Database Administrator (DBA)?

-> A Database Administrator (DBA) -> is responsible for the performance, integrity, security, and availability of a database system.
-> They ensure that data is stored efficiently, accessed reliably and protected from loss or unauthorized access.

(27) What are the main steps involved in designing a data pipeline?

-> Designing a data pipeline involves creating a structured flow that moves from source to destination while transforming it into usable format.
-> Whether you're building a marketing dashboard or syncing campaign metrics.

(28) What are some common challenges in managing large-scale database?

-> Managing large-scale databases comes with unique set of challenges that span performance, reliability, scalability and governance.

(29) What are some popular database platforms (eg, Snowflake, MySQL, PostgreSQL, Oracle) and their use cases?

-> Snowflake -> ideal for cloud-native analytics and data sharing across teams
-> Supports semi-structured data and integrates with tools like dbt, Tableau and Power BI

-> MySQL -> lightweight and fast for transactional systems.
-> common in WordPress, shopify and custom web apps

-> PostgreSQL -> known for data integrity, custom functions and geospatial support (PostGIS).
-> Excellent for systems requiring strict relational logic and advanced queries.

-> Oracle -> Enterprise-grade with robust security, high availability and scalability.
-> Supports multi-model data, including relational, document and graph.

(30) What are the main data storage formats used in analytics (e.g CSV, Parquet, JSON, Avro)?

-> The main data storage formats used in analytics include CSV, JSON, Parquet and Avro -> each optimized for different use cases like human readability, compression and schema evolution.