Research Assignment 1

Section A

1. Main types of databases

- Relational databases - Store data in tables with rows & columns, using a predefined schema. They are ideal for structured data
Eg SQL Server, MySQL, PostgreSQL

- NoSQL databases - Store data in various flexible formats, making them suitable for large volumes of unstructured or semi-structured data.
Eg MongoDB

- Object-Oriented Databases - Store data as objects in programming.

2.
RDBMS (Relational Database Management System) is software that creates, manages, and provides an interface to relational databases, which organize data into tables with relationships between them.
- It supports data storage and management
- Supports data retrieval and querying
- Data Integrity

3. Primary - Is a unique identifier.
Foreign key - The key that do not belongs in the table, referred to PK in another table

4. Database normalization - It is the process of organizing a database into tables and relationships to minimize redundancy and improve data integrity - It is important because it prevents issues like duplicate data, data inconsistencies, reduces storage space & makes it easier to update, query & maintain data.

5.
Database Schema - It is the blueprint or logical structure of a database, defining how data is organized, structured and related.

6. Structured Vs. Semi-Structured Vs. Unstructured data
Structured - Fixed Schema (e.g SQL tables).
Semi-Structured - Partial structure (e.g., JSON, XML).
Unstructured - No fixed format (e.g Images, videos, text).

7. Fact Table Vs. Dimension Table:
Fact Table: Stores measurable quantitave data
Dimension table: Contains descriptive attributes
- because the fact table

8.
Data Model: A conceptual representation of how data is stored and related.

It is important in database design because it acts as a roadmap to ensure data is structured logically, consistently and efficiently, which helps reduce errors.

9. Database vs. Data Warehouse vs. Data Lake
Database - Storage place for our data
Data Warehouse - ~~Stores historical~~,
- Is a Central place where we can Multiple databases
Data Lake - Stores raw data (Structured & unstructured

10. Data Mart: A subset of a data warehouse focused on a specific department or function (e.g. IT, Marketing) Smaller & More Specialized than a full data warehouse.

Section B. SQL & Data processing

11. A Query language - Is a Specialized Computer programming language designed for retrieving & managing data from databases

SQL is Mostly used because it's Standardized, powerful & easy to learn.

12.
Indexes - data structures that speed up data retrieval
How indexes improves performance.
- Speed up queries
- Reduce disk I/O
- Optimize sorting and grouping
- Ensure data integrity

Section C: Data Management AND Analytics Concepts

21. Cloud database VS on-premise database

Feature	Cloud Database Mana	On-premise Database
Cost	Subscription based	Requires significant initial investment in Software & hardware.
Scalability	Easily scale resources up or down in response	Scaling requires purchasing and configuring new hardware, which can be slow & costly.
Maintenance Control & Security	Managed by Provider The provider secures the infrastructure, but you are responsible for data access & configuration	Managed by internal IT Full control: You have control over the entire system including data, hardware & security configuration.
Accessibility	Remote access- Accessible from anywhere with internet	Local access: Access is typically limited to the local network

22. Data governance is the system of system of policies processes, standards, & roles that ensures an organization data is managed secured, accurately & consistently making it available, usable & trustworthy.

- It is important in data management because it improves data quality, enhances decision-making, ensures compliance & can lower costs by eliminating redundancy

23. Data integrity is the accuracy, consistency, & reliability of data throughout its lifecycle, ensuring it is complete & trustworthy
Maintained using constraints, validation & backup.

24.
Data quality is degree to which data is accurate, complete, consistent, and reliable for its intended use. High data quality results in better decision and analytics outcomes.

25. Role of a Data analyst is managing & analyzing database information.
Data Analyst manages & analyzes database information by collecting, cleaning, & interpreting data to find trends & patterns that information business decision

26.
Database Administrator key responsibilities include managing, securing, and maintaining database systems to ensure data is stored, retrieved & protected efficiently & securely.

27. Steps in designing a data pipeline:
1. Data Collection (Extract)
2. Data transformation (clean, format)
3. Data loading (store in warehouse)
4. Automation & Monitoring

28 Challenges in Managing large-scale databases:
- Scalability
- Performance tuning

- Data Security
- Backup & disaster recovery
- Backup & disaster recovery
- Integration & diverse sources

29. Popular database & their use cases:
- MySQL: Web applications
- PostgreSQL: Advanced analytics
- Oracle: Large enterprise systems
- SQL Server: Microsoft environments
- Snowflake: Cloud data warehousing

30 Data storage formats in analytics:
CSV: Simple text-based tabular format
Json: Semi-structured web data
Parquet: Columnar compressed format (used in BIS data)
Avro: Binary row-based format for fast serialization