Research Assignment 1

① Main types of databases:
   ① Rational databases
   ② Graph databases
   ③ Document databases
   ④ ~~etc~~ NOSQL databases

② Relational Database Management System (RDBMS)
- A software that stores structured data in table formal. It allows us to create, read, update & delete data using SQL

③ Primary key: uniquely identifies a record in a table
foreign key: A field in a table that links to a primary key in another table.

④ Data normalization is the process of organising data. It reduces data duplication and improves data consistency.

⑤ A schema is a structure of a database, defining how data is organized into tables, colums, and relationships

⑥ Structured data: Tabulated data like spreadsheets
Semi-structured data: Data with some structure but not rigid. e.g XML.
Unstructured data: No predefined structure e.g images and text documents.

⑦ Fact table: Contains measurable data e.g
① sales amount.

Dimension table: Contains descriptive data e.g
customer names, product categories.

⑧ A data model is a visual representation of data structures and relationships. It helps in designing efficient databases and ensures data integrity.

⑨ Data mart: A subset of a data warehouse focused on a specific subject e.g sales.

Data warehouse: A central place where we can have multiple databases.

Data lake: A storage repository that holds raw data in its native format.

⑩ Difference between a data mart and datawarehouse.
Data Mart: A specific to a department or business function

Data warehouse: integrates data across the entire organization.

Section B

(11) A query language is used to interact with databases, most common for relational databases because it's powerful for queries and managing data

(12) Indexes are database objects that speed up data retrival by allowing quick lookups on specific columns. They improve query performance but can slow down data modifications.

(13) Transactions are units of work in a database
ACID properties: Atomicity: All or nothing
Consistency: Valid data after transactions
Isolation: Transactions dont interfere
Durability: Commited changes persist.

(14) A database trigger is a stored procedure that automatically execcutes when certain events occur
e.g insert, update, delete

(15) Views: Virtual tables based on queries
Stored procedures: Precompiled SQL code for reuse
Triggers: Automated actions on events.

(16) Differences: ① ETL: Transforms data before loading into a target system.
② ELT: load data first, then transform in the target systems.

(17) Differences: ① Batch processing: Process data in bulk in intervals.
② Stream processing: Process data in real-time as it arrives.

(18) A join combines rows from tables based on related columns. Types of Joins: ① Inner Join - Matching rows.
② LEFT JOIN: All left table rows.
③ Right JOIN: All right table rows.
④ FULL JOIN: All rows from both.

e.g INNER JOIN - SELECT *
FROM customers - table
INNER JOIN orders ON customers.ID = orders.Customer ID;

(19) Referential integrity ensures relationships between tables are consistent (foreign keys).

(20) Data redundancy can increase storage usage and affect performance negatively due to extra data to manage.

Section C

(21) Differences: ① Cloud-based databases: Hosted on cloud platforms Scalable and managed. eg AWS RDS Azure SQL Database.
② On-premise databases: Hosted locally on a company's own servers. Requires maintanance.

(22) Data governance is the managment of data availability, usability, integrity and security. It is important for ensuring data quality, Compliance with regulations and protecting sensitive data.

(23) Data intergrity is the accuracy and consistency of data. It is maintained through contraints, validation rules, and regular data audits.

(24) Poor data quality leads to incorrect insights, which can impact business decisions negatively.

(25) A data analyst interprets data to help businesses make informed decisions. They analyze trends, create reports and visualize data.

(26) A Database Administrator (DBA) manages database performance security backups and ensure data integrity.

(27) ① Identify data sources
② Define data transformations needed.
③ Choose a destination
④ Schedule and automate the pipeline.

(28) ① Scalability issues
② Data security concerns
③ Maintaining performances with growing data volumes.

(29) ① MySQL
② Snowflake
③ PostgreSQL
④ Oracle

(30) ① CSV (comma-Seperated Values)
② Parquet (columnar storage)
③ JSON (Javascript Object Notation)
④ AVro (row-based format)