Research Assignment

Section A

① **Relational DAtabases (RDBMS)**

→ store data in tables (row and columns)

→ use SQL to mange and query data.

NoSQL Databases

→ Store unstructured or semi-structured data

Cloud Databases

→ Hosted and managed on cloud platforms

→ Provide scalability, security and remote access

② A Relational Database Management is a type of database software that store, manage and organize data in tables made up of rows and colamns each table represents a specific type of data

③ A primary key is a unique identifier for each record (row) in a database table. It ensures that no two rows have the same value in the column. (NO duplicates)

④ Database normalization is the process of organizing data in a database to reduce data redundancy (repetition) and improve data integrity.

⑤ A database schema is the structure of a database. It defines how data is organized including the tables. It basically the plan that show what data exists how its stored and how different part of the database connect.

⑥ Structured Data – organized in fixed formats like tables with rows and columns easy to store and query using SQL
Example: Databases, excel sheets

Semi-structured Data – Has some structured bits not in a strict table Formats uses tag or marker to separate data.
Example: CSV, email.

Unstructured Data – No predefined structure difficult to store in tables and analyze directly
Example: videos, image

- is connects to dimension tables using foreign keys.

(7) Fact table - is store measurable quantitative data

Example: - is contain numbers like sales amount, quantity or profit

record each sale { Order ID, Product ID, Customer ID, Sale Amount }

- Each record represent a specific event (eg: one sale)

Dimension Table - it contain descriptive attribute (like customer name, product categories or date)

Example:

- contain Primary key [to find the details who, what, where, when]

(8) A data model is a visual or logical way of showing how data is structured, stored and related in a database. is define WHAT DATA will be stored (entities or tables)

How data relates, what rules apply (keys)

Importance

→ Helps organize data clearly before building the database

→ Make is easier to understand relationship between data

→ Guide analysts when designing or querying the database

is helps plan and structure data properly so that database work efficiently and accurately

(9) • Database - store current, day-to-day operational data is is mostly structured data and is organized in tables [row and column]

Example: Banking system, online shopping apps

• Data warehouse - store historical data for analysis and reporting it is structured and cleaned data it organized and optimize for queries

Example: Business intelligence dashboards, trend analysis

• Data lake - store all types of data (raw, unprocessed, is is structure semi-structured and unstructured. NO fixed structure, very flexible [Example: Big data analytic, machine learning]

(10) A data mart is a smaller, focused part of a data warehous that stores data for a specific business area or department.

Difference between Data Mart and Data Warehouse

Aspect	Data warehouse	Data Mart
(1) Scope	cover the entire organization data	Focuses on a specific departments or function [Eg: Sale, finance]
(2) Data size	very large - store data from many source	smaller - contain only relevant data for one team
(3) Purpose	used for company-wide analysis and reporting	used for departments-level analysis.
(4) Data source	integrate data from multiple system	usually gets data from the data warehouse

Section B SQL and Data Processing

(11) A query language is a language used to communicate with a database to retrieve add, update, or delete data

→ Structured Query language is commonly used because its the standard language for most relational databases.

(12) Indexes in databases are special data structure that make it faster to find and retrieve data from a table [Just like a book index helps you find information quickly trading a bit of space for much better speed.]

(13) A transaction in a database is a single unit of work that group one or more operations (like Inserts, update or delete) so they either all happen successfully or none happen at all

ACID Properties

Atomicity - All parts of a transaction happen or none do if one steps fail the whole transaction is rolled back

Consistency - The database remain valid before and after the transaction. Balance must still add up correctly after transfer

Isolation - Transaction don't interfere with each other

Durability - Once committed, data is permanently saved even after a crash

(14) A database engine is like the brain of the database it manage how data is stored, processed and retrieved in a database. It's what execute SQL queries and handles reading and writing data.

→ The impact on Performance is manage caching and memory use, reducing disk access. A good engine optimize queries and uses index efficiently making searches faster.

(15) Views - is a virtual table created from the results of an SQL query
- used to simplify complex queries, secure data and present customized results

→ Store Procedure = A stored procedure is a predefined set of SQL statements save in the database
- It can take inputs, perform operation and return results

Example → CREATE PROCEDURE GetCustomerOrders(@CustomerID INT)
AS
SELECT * FROM orders WHERE customer_id = @CustomerID;

→ Triggers is a piece of SQL code that automatically runs when a specific event happen on a table like INSERT, UPDATE, or DELETE (automatic action when data changes)

(16) ETL = Transform before loading (clean first, then store)
ELT = load first then transform inside the warehouse

(17) Batch processing = handles data in bulk, later it work with historical data
Stream processing = handles data continuously in real time it work with real-time

(18) A Join in SQL is used to combine data from two or more table based on a related column between them ① Inner Join = Return only the rows that match in both table ② left Join = Return all rows from the left table and matching rows from the right table ③ Right Join = return all row from the right table

(19) Referential integrity means that the relationship between table in a relational database remain consistent and accurate.
→ if maybe there is a foreign key in a table it prevent add link it as something that doesn't exist

(20) Data Redundancy mean storing the same piece of data in multiple place within a database. eg: if a customer's address appear in both the customer table and the order table that's redundancy the same data is repeated unnecessarily.
→ Storage = it Repeated data consume extra disk space especially in large database with million of record
→ Performance = More data mean slower queries, inserts, updates, and backup the database engine has to process more rows

---

Section C:

(21) Cloud Database is Hosted on remote server managed by a cloud provider (eg: AWS, Google Cloud). The cloud provider handles hardware, backups, scaling and updates.
→ On-Premise Database is installed and maintained on organization own Physical servers. The organization IT team is responsible for setup

(22) Data governance is the process of managing the availability, usability, integrity and security of data within an organization. It is important because it maintain data quality, supports compliance, improve decision-making and reduce risk related to data misuse or errors.

(23) Data integrity means the accuracy, consistency and reliability of data throughout its entire lifecycle from when it created to when it used or deleted. It can be maintained by using validation rule, primary and foreign keys access control backup and audit trail to ensure that data remain reliable and unchanged.

(24) Data quality refer to how accurate, complete, consistent and reliable data is for its intended purpose. It critical for analytic because reliable data leads to accurate insight better decision and greater trust in analytical result.

(25) A data Analyst manages and analyze data from databases by collecting, cleaning and interpreting information to find trends and supports decision-making

(26) A Database Administrator is responsible for managing and maintaining database Their key duties include installing and configuring database system ensuring data security, performing backup nd recovery.

(27) The main steps in designing a data pipline include identifying data source, defining data requirement extracting data, transforming it for consistency and accuracy, loading it into the target system, automating the workflow and continously monitoring and maintaining the pipline

(28) common challenges in managing large-scale database include maintaining performance and scalability, ensuring data security and integrity, managing backups and storage, handling concurrency and integrating data from multiple source

(29) MySQL = open-source relational database widely used for web application (Wordpress)
Postgresql = Avanced open-source relational database that support complex queries
Snowflake = cloud-based data warehouse built for big data and analytic.
Handles structured and semi-structured data with scalability and high performance

(30) CSV(comma-separated value = simple text format for tabular data easy to use and read
Parquet = columnar format optimized for analytic and large datasets
Json (Javascript object Notation) flexible text format for semi-structured data widely in API and web applications.

Unstructured data
Example: videos, image and analyze directly