Research Assignment

SECTION A: DATABASE Fundamentals

1. There a various kinds of databases:
   * relational database - which stores data in tables with rows and colouns (e.g MS SQL, Postgresql
   * Non-relational database are designed for flexibility and scale
   * Data warehouse - for analytical queries and BI (Snowflake, Redshift
   * In memory database: Stores data in RAM for fast access
   * Time-Series Databases: optimize storage to time-stamped data.
   * Oject-oriented database; Stores data as objects

2. A Relational database managerent System is → a software that stores & manage data using tables with relationships defined by keys. It supports SQL queries, and structured tabular da e.g MS SQL Server

3. A primary key: is a unique identifier for each record in the table. It cannot be null.
   A foreign key is a field that links one table to another by referencing the primary key of the related table.

(8) A data model is a blueprint that shows how data is structured and related
- It guides a database design
- Supports efficient queries &
- Ensures consistency.

(9) Database: Supports daily operations / or transactions (e.g OLTP → Mario scanning at Chopnte, all records will record on an OLTP & that is a database). A relational database

Data warehouse: is designed for analytic
- processes (An OLAP).
- Stores current & historical data & data is stored & refreshed from source system
- It stores cleaned and structured data (Snowflake)

Data lake: Stores raw, unstructured, semi-structured and structured for large scale analytics

(10) A data mart is a smaller, (Maybe) department-based / department-specific subset of a data house (e.g HR mart in a company database)
- It differs from a database in that a database is an enterprise wide analytical storage serving multiple functions

Section B: SQL and Data Processing

(11) A query language is the way in which users retrieve and manipulate data in SQL. This will be syntax (How we communicate with data to make sense of it.
SQL is the most common query language for structured data
- It is standardized
- Works on most RDBMS

(12) Indexes in a database are data structures (like pointers) that speed up search queries by avoiding full table scans. They improve read performance

(13) A transaction in a database is a unit of work executed as a whole (e.g insert, update, delete)

Acid Properties
- A set of guarantees that ensures transactions are reliable. E.g

ACID has the following elements
A- Atomicity (All or nothing)
C- Consistency (valid before & After)
I- Isolation (No interference between transactions)
D- Durability (Permanent results after commit

14. A database engine is the underlying software responsible for storing, indexing and retrieving data. It affects
- speed
- Storage method
- concurrency control
- query optimization

15. A VIEW: A virtual table based on a query
Stored procedure: A saved SQL script executed on demand
Trigger: A procedure executed when certain events occur (e.g. insert, delete)

16. The difference between ETL & ELT

This is a pipeline through which data gets taken from OLTP to OLAP

~~ETL~~: Extract - Transform - Load.

ETL: Transformation is done before data is loaded. Most used in traditional warehouses

ELT: Extract - Load - Transform.
: Transformation of data is done inside the data warehouse, after data is loaded. Most used with cloud systems like Bigquery & Snowflake.

17) Batch processing: Processes large chunks of data at intervals - (nearly, daily)
Stream Processing: (processes data in real time (This could be in an OLTP

18) A Join in SQL is a way to combine rows from two or more tables based on related columns. There are various types of Joins
↳ inner Join : matches all common records
↳ left Join : ALL from left table & only matched from right
↳ right Join : ALL from right & only matched from left.
↳ full Join : ALL records when match exists
↳ Cross Join : Cartesian product - (meaning ALL rows possible combinate of rows.

19) Referential integrity: ensures relationships between tables remain valid by enforcing rules via foreign keys
Importance: ↳ prevents orphan values
↳ Ensures consistency
↳ Maintains data accuracy

20 Redundancy caused increased storage costs, slowers query performance and causes more higher chance of inconsistencies and complex updates.

Section C: Data Management and Analytics Concepts

21 cloud database management: is scalable on demand, accessible anywhere and its managed services with lower upfront costs

on-premises database management: Has/ offers full control, high setup/ maintanance costs and hard ware dependancy.

22 Data governance: A Framework of polices and processes ensuring the correct Management of data.
- It ensures compliance.
- protects data quality
- Manages risk &
- Defines roles & responsibility in terms of accountability.

23 Data integrity refers to accuracy, consistency, and reliability of data. It is Maintained through
- Validation rules
- keys, normalization and

(24) Data quality
- means data is accurate, complete, consistent, timely and valid. This is critical because poor data affects decision making, providing misleading insights and impacts reporting accuracy.

(25) Role of Data Analysts is to
- Extract & clean data.
- Perform SQL queries
- Build Dashboards that provide visualisation & Story telling of data for insights for decision makers
- Analyze trends
- Supports decision making with insights

(26) A database Administrator (DBA) -
- installs & maintains databases
- Performs tuning
- Back up + recovery
- Access control of database
- Security management.
- Ensures data integrity/assurance.

(27) Steps in designing a data pipeline
1. requirements gathering
2. Source identification
3. Data extraction
4. Data transformation/cleaning.

5. Load into Storage (warehouse/lake).
6. Scheduling.
7. Monitoring & Maintainance.

28. There are various challenges that could present in managing large scale databases including:
- Performance Bottlenecks
- Storage Scalability
- Data quality issues
- Concurrency conflicts
- Back up complexity
- Security threats
- Cost Management (cloud compute)

29. Popular databases are amongst others
- MSSQL - web applications, OLTP systems.
- Snowflake - Advanced analytics of Big data, Cloud warehousing
- Oracle - large enterprise transactional systems

30. Main types of data Storage formats used in analytics include.
- CSV (Simple text format) & it is the most used
- JSON - Semi Structured, APIs & NO SQL