Research Assignment 1

Section A : Database Fundamentals

1. (A) Snowflake - This is a Cloud-based data warehouse for large-scale data & analytics

(B) Graph Databases - focused on managing data with intricate node and edge interactions such as Amazon Neptune.

(C) Relational (RDBMS) - Stores SQL for searches and stores structured data in rows and columns. For instance, SQL Server, Oracle and MySQL.

2. A Relational Database Management System (RDBMS) is a data warehouse which stores data in tables with rows and columns. This data warehouse utilises SQL to handle and query data. Its ideal for systems that require data consistency and complex querying.

3. - A primary key is the unique identifier of each row in a table. For instance, passport number.
- A foreign key links back to the primary key from the other table. It acts like a reference.

4. - Database Normalization is the process of organising data in a database to reduce duplicates and improve accuracy and consistency.
- It involves breaking down large tables into smaller, related tables and defining relationships between them using primary and foreign keys.
- It is important due to the following:
  ① Prevents storing duplicates in multiple places
  ② Improves consistency by ensuring that data is updated in one place only.
  ③ It saves space by not repeating unnecessary data.

5. A schema exists inside a database. It is where a table is created. It is structure of a database and defines how data is organized.

6. (A) Structured data: This is data stored in a table with rows and columns. For instance, an Excel spreedsheet.
(B) Semi-Structured data: This is data that does not follow a strict table format but still has some structure. For instance NoSQL databases.

© Unstructured data: This is data with no predefined structure - it is not stored in a table format. For instance, Emails

7. - Fact table: It stores measurable quantitative data about business processes. It contains numerical data. For instance, Revenue, date or customer id.
- Dimension table: It stores descriptive data about business processes. It describes the 'Fact table' and it consists of primary keys.

8. - A data model is a conceptual framework that defines the following:
① How data is structured
② How data elements relate to each other
③ What rules govern the data.
- It is important because it organizes data clearly, it improves communication, it prevents redundancies and eases maintenance.

9. - Database: Stores current operational data that is structured. It is fast for real-time transactions. For example, Oracle.

- Data Warehouse: It stores historical analytical structured data. It is optimized for complex queries such as OLAP (Online Analytical Processing). For example, Snowflake.
- Data Lake: It stores raw structured and unstructured data. It is slower and depends on processing engine. For example, Azure Data Lake Hadoop.

10. A data Mart is a smaller, more focused version of a data warehouse designed for a specific department or business function such as Sales, HR, Finance and Marketing. It exacts only sales-related data from the data warehouse, making it easy and quick for the sales team to generate reports.
On the other hand, a data warehouse stores data for all departments such as HR, Finance, Sales, Marketing, etc.

Section B: SQL and Data Processing

11. A query language is used to interact with databases to retrieve, insert, update or delete data.
- SQL (Structured Query Language) is the most commonly used because it is standardized, it is easy to learn, it supports complex queries and data manipulation.

12. An index is like a book's table of contents - it allows the database to find data faster. Indexes improves query speed, especially for SELECTs. For instance, if you search for a student by ID without an index, the system checks every row. With an index on Student ID, it jumps straight to the right row.

13. A transaction is a group of one or more operations that must succeed or fail together.
- ACID ensures safe and reliable transactions. ACID stands for Atomicity, Consistency, Isolation, Durability.

14. A database engine is the core software that handles storage, retrieval and management of data. It decides how data is stored, indexed and queried. For example, MySQL engines.

15. - View: A virtual table based on a SQL query that is used to simplify complex joins or filter data.
- Stored Procedure: A saved set of SQL statements that can be executed repeatedly (like a function).
- Trigger: An automatic SQL code that runs when certain events happen (INSERT, UPDATE, DELETE).

16. - ETL (Extract, Transform, Load): This is data that is transformed before loading into the warehouse. It is slower with large unstructured data.
- ELT (Extract, Load, Transform): This is where data is transformed after loading into the warehouse. It is faster and better for big data.

17. - Batch Processing: It processes large volumes of data at once. It is used in the cases of payroll and billing. For instance, AWS Batch
- Stream Processing: It processes data in real-time as it arrives. It is used to detect fraud. An example, Spark Streaming.

18. A JOIN combines rows from 2 or more tables based on related columns (usually a foreign key).
Types of JOINS are:
① INNER JOIN: Returns matching rows in both tables. It is used when you want records that exists in both tables
② LEFT JOIN: Returns all rows from the left table, plus matching rows from the right table.
③ RIGHT JOIN: Returns all rows from the right table, plus matching rows from the left table.

④ FULL OUTER JOIN: Returns all rows when there is a match in either table including NULL values.

⑤ CROSS JOIN: Returns all possible combinations (cartesian product)

19. - Referencial integrity ensures that relationship between tables remain valid. For instance, if a foreign key points to a record, that record must exist in the parent table.
- It is essential because it prevents orphan records, maintains consistent and reliable data.

20. - Effects of Data Redundancy includes increased storage usage which creates duplicated data that takes up unnecessary space.
- It slows down performance meaning larger tables take longer to query and data.
- Data becomes inconsistent whereby the same data in multiple places can become conflicting.

Section C: Data Management and Analytics Concepts

21. Cloud Database: This database is hosted on cloud such as AWS. It is highly scalable and accessible over the internet For instance, Snowflake.

- On - Premise Database : This database is installed and managed on local servers. This is handled by in-house IT Team. For example, Oracle.

22. - Data governance is the framework for managing data policies, standard, and responsibilities to ensure data is secure, accurate, accessible and compliant with regulations.
- Importance : It prevents data misuse, ensures data privacy and compliance, it improves data trust and accountability.

23. Data Integrity means ensuring that data is accurate, consistent and reliable over its lifecycle. It can be maintained through the following :
① Access controls
② Audit logs
③ Backups
④ Constraints (Primary & Foreign key)
⑤ Validation rules.

24. Data quality refers to how useful and trustworthy data is for analysis and decision-making.

- It is critical for analytics for the following reasons:

① It promotes accuracy because wrong data leads to wrong decisions

② It promotes completeness because missing data limits analysis.

③ It promotes consistency because conflicting data causes confusion.

④ It promotes timeliness because ~~data~~ outdated data loses value

25. Data Analyst extracts data using SQL. The analyst cleans and prepares data, performs statistical analysis, create dashboards and reports and lastly, helps the business teams make data-driven decisions. A data analyst uses tools such as Excel, SQL, PowerBI, Tableau, R, Python.

26 (A) Database Setup - it installs and configure DBMS.

(B) Performance tuning - it optimizes queries and indexing.

(C) Security - it manages users, permissions, encryption.

(D) Backup & Recovery - it ensures data can be restored after a failure.

(E) Monitoring & Maintenance - it keeps the database running smoothly.

27. Step 1: Data Ingestion - Pull data from sources such as APIs, databases, files.

Step 2: Data Validation - Checks for errors or missing values

Step 3: Data Transformation - Cleans, enrich or reformat data

Step 4: Data Storage - Loads into target systems (For instance, data warehouse).

Step 5: Orchestration & Scheduling - Automates and manages tasks (For example, using Airflow).

Step 6: Monitoring & logging - Tracks performance and handles failures.

28. (A) Performance issues - it slows queries due to data size.

(B) Data Consistency - it keeps data synchronized across systems

(C) Backup & recovery - it manages large backups efficiently

(D) Cost control - it ~~storage~~ stores and computes expenses.

(E) Security - it ~~process~~ protects sensitive or personal data

(F) Scaling - it handles growing data and user demands.

29. (A) MySQL - Best used for web applications, e-commerce, cms (For example, WordPress), small-to-medium businesses.

(C) PostgreSQL - Best used for advanced analytics, geospatial data (Post GIS), enterprice apps and financial systems.

(D) Oracle Database - Best used for large enterprise systems ERP, banking, high-volume transaction processing.

(E) Snowflake - Best used for big data analytics, AI reporting, scalable cloud data storage

30. (A) CSV - This is a simple tabular data, consists of structured data, easy to read and export, it is not efficient for large-scale analytics.

(B) JSON - This is human-readable, consists of Web APIs and it consists on semi-structured data.

(C) Parquet - Includes big data, schema evolution support and this used with Kafka and Hadoop.

(D) ORC - It is optimized for Hive-based systems, high performance and compression.