Research Assignment 1

SECTION A : Database Fundamentals

1. Relational Database (SQL)

NoSQL

2. Software that stores data in tables (rows & columns) & allows you to create, update and manage a relational database using SQL.

It maintains data integrity and relationships

3 Primary key: A unique identifier for each record in a table (e.g. Student_ID).

Foreign key: A field in one table that refers to the Primary key (unique) in another table, creating a link between them.

4. It is a process for organizing information to avoid redundancy with the aim of enhancing data integrity.

It reduces duplicate data, avoids update anomalies of the database, & makes it more efficient to manage queries.

5. This is the plan for a database.

It provides the tables, columns, data types, relationships & constraints.

6. Structured data:

Arranged within a fixed schema

(example: Excel, application forms)

- Semi-structured:

Has some organizational properties, but not fixed

example: emails, HTML

- Unstructured data

No predefined format (eg. images, videos)

7. FACT Table

Holds quantitative data (that can measured)

For example: Sales amount

Dimension Table

Holds descriptive data that provide context to the facts (

example: product details, customer information)

8. Illustrates data structures, relationships & rules visually.

It provides as a good foundation to construct the database,

making sure everything is clear & preventing errors.

9. Database

This for transational processing - day-to-day operations

Data Warehouse

Used for analytical processing (OLAP)

Historical data for reporting

Data Lake:

stores vast amounts of raw data in its native format for future use.

10. Data Mart

is a condensed part of a data warehouse, that helps a particular business domain or department of data storage

Unlike the enterprise data warehouse, it is focussed on the needs of specific teams (like sales or marketing).

Section B: SQL and Data Processing

11. A specialized computer language used to retrieve, manipulate, and manage data from databases & other information systems.

It enables interaction with databases & other structured data systems to request & access specific information.

SQL is the most widely used because its standardized, easy to learn & powerful.

Based on simple English commands, SQL lets users state what data they need, not how to get it.

12. Indexes are specific lookup tables for faster access to data retrieval.

They increase performance due to the ability for the database to locate data without scanning the entire table

13. A set of operations that must all succeed or fail together.

Properties:

Atomicity	All operations must be complete or none at all.
Consistency	keeps the database in a valid state
Isolation	It guarentees that transactions are protected from meddling
Durability	ensures permanency after a change is made

14. The fundamental software which stores, retrieve & manages data.
The engine selection directly affects speed, concurrently & reliably among workloads.

is. View
A virtual table based on a SQL query
Store Procedure
A pre-written SQL code that can be executed
Trigger
A procedure that runs automatically due to an event occuring on some table.

16. Extract Transform Load (ETL)
Data is transformed before loading into warehouse

Extract Load Tranform (ELT)
The data gets looded to the warehouse first, & then transformed within it.

17. ~~Batth~~ Batch Processing
Processes large volumes of data at scheduled intervals

Stream Processing
Processing data in real-time as it generated.

18. A JOIN merges rows from 2 or more tables based on a common column.
Types of JOINS
INNER JOIN - Generates records that have rows that match in both table.
LEFT JOIN - Returns all the records of the left table & any matches from the right table
RIGHT JOIN - returns all the records from the right table, including matched records from the left table
FULL Outer JOIN - provides all records when there is a match in either of the tables.
Referential integrity is the principle that a foreign key always means a primary

key that already exists.
This idea is very vital to maintain correct & regular relationship between the tables

example

Table: students

student_id	name
1	Alice
2	Brian
3	Carol

Table: courses

course_id	student_id	course_name
101	1	Math
102	2	English
103	4	Science

- INNER JOIN

```
SELECT    students.name, courses.course_name
FROM      students
INNER JOIN  courses
ON        students.student_id = courses.student_id;
```

* Only students with mathing student id in both table appear.

- LEFT JOIN

only if they dont have a course

```
SELECT    students.name, courses.coure_name
FROM      students
LEFT JOIN   courses
ON        students.student_id = courses.student_id;
```

* Carol appears but with NULL because she has no matching course

RIGHT JOIN

shows all courses, even if no matching student exists.

```
SELECT      students.name, courses.course_name
FROM        students
RIGHT JOIN  courses
ON          students.student_id = course.student_id;
```

- The science course appears, even though no student with student_id = 4 exists

FULL JOIN

```
SELECT      students.name, courses.course_name
FROM        students
FULL OUTER JOIN  courses
ON          students.student_id = course.student_id;
```

- Combines the results of left and RIGHT JOINs - includes all records

19. It is vital for maintaining accurate & consistent relationships between tables.
Ensures a foreign key value always points to an existing primary key.

20. Strategic redundancy can sometimes improve read performance for complex queries
- It also wastes storage space & can lead to data inconsistencies (update abnormalities).

Section C: Data Management & Analytics concepts

21. Cloud databases are hosted online (scalable) less maintence).
On-premise database are installed locally (more control but more cost).

22. Data
The rules & procedural practices for maintin accurate, secure & responsible data

It is important for ensuring data is trustworthy, secure & used properly across the organization.

23. Means data stays correct & complete through validations, keys & constraints.
Data integrity is maintained by using keys, constraints, validation, access control & backups to keep data accurate & reliable

24. This pertains to how good or reliable the data is, including, but not limited accuracy, completeness & consistency.
Its necessary in that with poor quality data, it can lead to misleading analyses and poor business decisions.

29. They start by quering databases to find the required information, then they clean this data & organize it to ensure that it is usable.
Then the data analysis looks for trends & patterns.
They than produce reports & visualizations that guide business in decisions.

26. Manages database performance, backups, user access & security.
-

27. Date Ingestion - Identify data sources
Extract data
Transform / clean data
Load into destination
Monitor & maintain

28. Performance issues
storage costs
data security
back-up
scaling

29. Relational Databases such as:
Microsoft SQL Server
~~MySQL~~ Business intelligence with which to gain insights about your business & customers supported by machine learning

Oracle Database
centralises and consolidates large amounts of data from multiple sources.

30. CSV
simple, human-readable

- Parquet
Columnar format, highly efficient for quering & storage

- Avro
Row-based format, good for serialization & schema evolution