Assignments - Data Engineering & ETL

1.Research and Compare Data Storage Methods like RDBMS, DBMS.
Write a short paragraph about each method, explaining its key features, advantages,
and disadvantages.

Ans:
RDBMS

RDBMS (Relational Database Management System): RDBMS ensures data integrity and ACID compliance by storing data in organised tables with clear relationships. It allows for complicated operations and reporting by using SQL for querying. Data integrity and effective querying are benefits, however it may not be as scalable for really large datasets.

DBMS
The word "DBMS" (Database Management System) refers to a broad category of data storing techniques. It is not constrained to the relational model, unlike RDBMS, and can support various data formats. Flexibility is provided, but structure and ACID conformance may be compromised. suitable for data that is less structured or more diverse, yet can be difficult to maintain.

2.Go through database fundamentals like Normalisation, Data Inconsistency, Data Deduplications.

Ans:

Normalization: Organizing data into tables and reducing redundancy to enhance data integrity and efficiency.

Data Inconsistency: Discrepancies in the same data across a database, leading to errors and confusion.

Data Deduplication: Identifying and removing duplicate data to improve storage efficiency and accuracy.

3.Based on your understanding, create an example table for these two topics Data Inconsistency, Data Deduplications.

Ans:

Data Inconsistancy

Student_id	Student_name	Course
101	Gopi	Java
101	Gopi nath	C++
102	Allwin	Java
Inconsistent data in the "Student_Name" column for Student_ID 101.		

Data Duplication

Customer_id	Name	Mail
201	Gopi	gopi@gmail.com
202	Allwin	allwin@gmail.com
203	Gopi	gopi@gmail.com
Duplicate "Gopi" entries with the same email address reduced to a single entry through data deduplication.		

4.Where actually the data is getting stored, if we use MySQL and Psql dbs. Do some R&D and give explanations based on your understanding.

Ans:

MySQL: Each database's accompanying directory contains files that contain the data. The data, indexes, and schema information for each table are kept within individual files on the server's file system.

PostgreSQL (PSQL): For each database cluster, PostgreSQL (PSQL) stores data in a collection of files contained in a single data directory. These files, which are arranged according to tablespaces, include tables, indexes, and system data that enable effective data administration and access.

5.Explore where we are using DAS in the database? What is the algorithm?

Ans:

Databases frequently employ Direct-Attached Storage (DAS) for local storage of database systems on a single server. It works well for small-scale applications with concentrated and constrained data needs. Databases use various storage management techniques like B-trees and indexing to optimise data retrieval and administration on DAS; DAS itself does not have a specific algorithm connected with it.

6.Understand about Distributed Systems?

Ans:

Distributed Systems: Networks of linked computers work together to accomplish a common objective in distributed systems. They make it possible to distribute data, processing jobs, and resources across several machines, improving scalability and fault tolerance. Peer-to-peer networks and cloud computing are two examples of how to create effective and resilient computer environments.

7.Go through the basic SQL queries like select and update.

Ans:

SELECT Query:

```
SELECT column1, column2 FROM table WHERE condition;
```

INSERT Query:

```
INSERT INTO table (column1, column2) VALUES (value1, value2);
```

UPDATE Query:

```
UPDATE table SET column1 = value1, column2 = value2 WHERE condition;
```

DELETE Query:

```
DELETE FROM table WHERE condition;
```

CREATE Table Query:

```
CREATE TABLE table_name (
    column1 datatype,
    column2 datatype,
    ...
);
```

ALTER Table Query:

```
ALTER TABLE table_name
ADD column_name datatype;
```

DROP Table Query:

```
DROP TABLE table_name;
```

8.Create a Simple ETL Flow after doing some R&D.Need flow diagram or steps followed in actual ETL.

Ans:

Simple ETL Flow:

Extraction: Retrieve data from various sources like databases, APIs, or files.
Transformation: Clean, validate, and transform data with calculations, formatting, and logic.
Loading: Load transformed data into a target database or data warehouse for analysis.
This sequential process ensures data quality and accessibility for insights.

9.Explain the difference between structured and unstructured data.

Ans:

Structured data:
Well-organized and able to fit into tables with set columns and rows is structured data. Utilising conventional databases allows for simple processing and querying.

Unstructured data:
Lacks a consistent framework and comes in a variety of media, including text, graphics, and audio. because to its complexity and diversity, requires the examination of cutting-edge methods like AI and machine learning.

10. Do some R&D on raw data sources, like what are the data sources available,
structured data and unstructured data.

Ans:
Sources of Unprocessed Data: Data comes from a variety of sources, including databases, spreadsheets, APIs, sensors, logs, websites, social media, and IoT devices.

Sources of Structured Data: For straightforward organisation and analysis, structured data is found in relational databases, spreadsheets, and CSV files.

Unstructured Data Sources: Data without a set format, such as that found in unstructured texts, photos, videos, social media posts, emails, and documents, requires specialised tools for analysis.