This exercise is designed to evaluate your skills in **data modeling, ETL, big data processing, and database optimization**

Assignment Overview

The goal of this exercise is to assess your ability to:

- Design efficient data models for both analytical and operational needs.
- Implement ETL processes to load and transform data.
- Work with large datasets using big data frameworks (e.g Hadoop, Spark).
- Optimize database performance and ensure data integrity.

Assignment Details

You will be provided with two datasets:

1. **Float Data**: This csv contains staffing and allocation information for projects, including details like team member name, project name, role, estimated hours, and project dates.
2. **ClickUp Data**: This csv contains task and time tracking information, including details like team member name, task name, project name, date, hours logged, and billable hours.

Your tasks are:

1. Data Warehousing & ETL Process

Task: Design a data warehouse schema that can handle these datasets and support both reporting and operational needs.

Instructions:

- Create a **dimensional model** for the data (star schema).
- Implement an **ETL process** to load the datasets into your schema. You may use any tool or language (SQL, Python, Apache Airflow, etc.). - Use Airflow to load into a DB within a docker container. Build an airflow pipeline that reads and loads to respective tables.
- Ensure data integrity and cleanliness throughout the process.

Deliverables:

- A detailed explanation of your data warehouse design. - Will be included in Overall PowerPoint Slide
- ETL scripts or processes used to load the data.
- A description of how you ensure data integrity and cleanliness.

2. Database Query Optimization

Task: In the below pre-written SQL query that operates on a complex dataset provided to you. Your job is to optimize the query for performance.

```
SELECT
   c.Name,
   f.Role,
   SUM(c.hours) AS Total_Tracked_Hours
   SUM(f.Estimed Hours) AS Total_Allocated_Hours,
   Date
FROM
   ClickUp c
JOIN
   Float f on c.Name = f.Name
GROUP BY
   c.Name, f.Role
HAVING
   SUM(c.hours) > 100
ORDER BY
   Total_Allocated_Hours DESC;
```

Instructions:

- Analyze and refactor the query for efficiency.
- Implement indexing, partitioning, or other techniques as needed.
- Provide a brief explanation of the steps you took, and the performance improvements achieved. - To be added to powerpoint Slide.

Deliverables:

- The optimized SQL query.
- A description of the optimization steps and their impact.

3. Big Data Processing with Spark or Hive

Task: Using either Apache Spark or Hive, process and analyze a large dataset of your choice (e.g any task log records of over 20 million).

Instructions:

- Perform a transformation on the dataset (e.g., aggregating hours logged by project).
- Ensure scalability and efficient processing of large volumes of data.

Deliverables:

- The code used to process the data in Spark or Hive.
- An explanation of your approach and the performance considerations.

4. Data Modeling Techniques

Task: Using the provided datasets, design data models for both analytical and operational purposes.

Instructions:

- Create both a dimensional model (for analytical purposes) and an entity-relationship model (for operational use)

- Explain why you chose this model and how it addresses operational needs.

Deliverables:

- The data model diagram.
- A brief report explaining your design decisions.

Submission

Please submit your ETL scripts, SQL queries, diagrams,dockerized solution with instructions in a readme file and any related files. Share any code in a GitHub repository, Jupyter notebook, or equivalent and screenshots of your process and output can also be helpful.

Any written explanations or reports should be submitted as a Google Document.