Hands-on Task: Design Thinking for Data Scientist

1. Case Study Background

GreenStream Energy is a smart-utility provider that collects electricity usage data from 50,000 households using smart meters.
Although large volumes of data are continuously collected, the data is currently "dark data"—stored but not transformed into actionable insights.

The company's strategic goals are to:

- Identify peak energy consumption periods
- Detect abnormal or faulty smart meters
- Prepare data for future predictive analytics and forecasting

To achieve these goals, GreenStream Energy plans to design an automated, serverless data pipeline that transforms raw operational data into analytics-ready datasets.

2. Data Challenges in the Current System

The raw smart-meter data suffers from multiple real-world data quality issues:

- Inconsistent measurement units
    - Some meters report energy in Watts (W)
    - Others report in Kilowatts (kW)
- Missing readings (null values)
    - Temporary Wi-Fi outages cause gaps in the time series
- Inefficient data format
    - Data arrives in CSV format, which is not optimized for large-scale historical analytics

3. Design Goal (The design should reflect data science thinking, not cloud configuration skills)

Your task is to **design—not implement—**a conceptual serverless ETL (Extract, Transform, Load) pipeline that:

1. Cleans and standardizes energy data

2. Stores structured data for querying and validation
3. Archives analytics-optimized data for long-term analysis
4. Handles failures and retries automatically

5. Apply the following to the case study:

Task A: ETL Architecture Diagram (System Design)

Create a detailed flow diagram of the ETL pipeline that includes:

- Source
- Transformation Layer
- Destinations
- Orchestration

Your diagram must:

- Clearly show data flow direction
- Include success and failure paths
- Indicate retry or error-handling behavior
- Use standard icons or clearly labeled blocks

Focus on logic and flow, not implementation details.

Task B: Transformation Logic & Business Rules Design

List and explain the business rules that must be applied during the Transform phase to resolve the data issues described in the case study.

Your rules should address:

- Unit standardization
- Missing values
- Data validation
- Faulty meter detection (basic logic)

Example Format:

- Rule 1: If energy unit = "W", divide value by 1000 and convert to "kW".

- Rule 2: If energy reading is NULL, flag the record and exclude it from peak-usage calculations.
- Rule 3: If a meter reports zero consumption for an unusually long period, mark it as a potential faulty meter.

Task C: Single Record Lifecycle Explanation

Explain step by step what happens to one single smart-meter record from the moment it is uploaded until it is archived.

Your explanation should include:

1. Upload to raw storage
2. Triggering of the transformation process
3. Data cleaning and validation steps
4. Storage in structured format (RDS)
5. Conversion and archival in Parquet format
6. How success or failure is handled