Design Thinking for Data Scientist

GreenStream Energy – Serverless ETL Design

Task A: Conceptual ETL Architecture

Raw smart■meter CSV data is ingested into raw storage. An orchestrator triggers transformation services to clean, validate, and route data into structured storage and analytics archives.

Sample Pseudo■ETL Flow

```
IF new_file_uploaded:
    trigger_ETL_workflow()

ETL_workflow():
    extract_raw_data()
    transform_data()
    load_to_RDS()
    archive_to_parquet()
```

Task B: Transformation Rules

Key transformation and business rules applied:

```
# Unit Standardization
if unit == "W":
    value = value / 1000
    unit = "kW"

# Missing Values
if value is NULL:
    flag_record()
    exclude_from_peak_analysis()

# Faulty Meter Detection
if consumption == 0 for long_period:
    mark_as_faulty_meter()
```

Task C: Single Record Lifecycle

```
1. Record uploaded to raw storage (CSV)
2. Upload event triggers ETL
3. Data cleaned and validated
4. Stored in structured RDS
5. Converted to Parquet
6. Archived for analytics
7. Errors logged and retried automatically
```