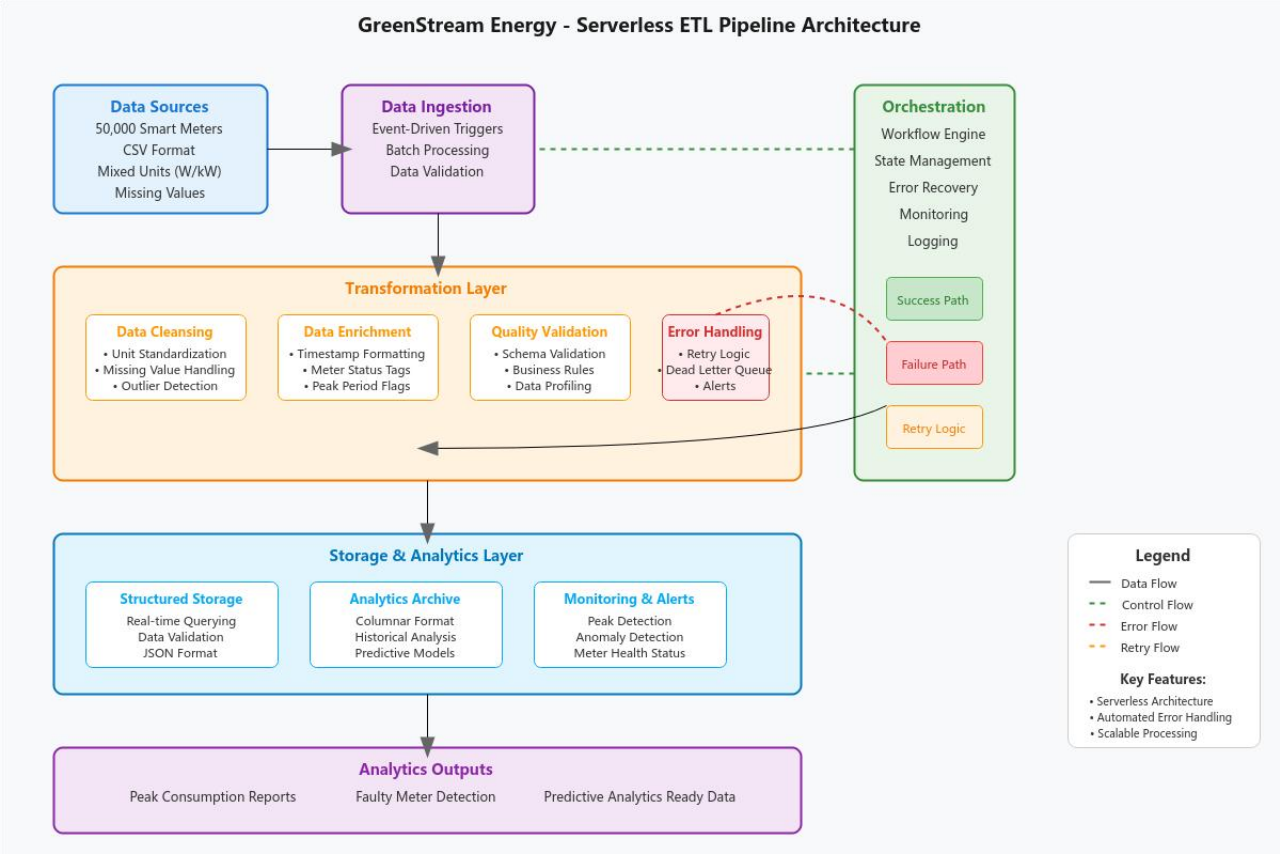
Name : Amany Abdelmonem

ID : 412200012

Task A: ETL Architecture Diagram (System Design)



Task B: Transformation Logic & Business Rules Design .

During the Transform phase, raw smart-meter data is cleaned, standardized, and validated to ensure it is accurate and analytics-ready. The following business rules are applied to resolve the data quality issues described in the case study.

Rule 1: Unit Standardization

Description:

Smart meters report energy consumption using different units (Watts and Kilowatts), which creates inconsistency in the dataset.

Logic Applied:

- If the energy unit is **"W"**, divide the energy value by 1000 and convert the unit to **"kW"**.
- If the energy unit is neither **W** nor **kW**, flag the record as invalid.

Reason:
Standardizing units ensures accurate comparison and aggregation of energy consumption values.

Rule 2: Missing Values Handling

Description:
Temporary network outages may cause missing energy readings.

Logic Applied:

- If the energy reading is **NULL**, flag the record and exclude it from peak energy consumption calculations.

Reason:
Missing values can distort analytics results if they are included in calculations.

Rule 3: Timestamp Validation

Description:
Each smart-meter reading must be associated with a valid and unique timestamp.

Logic Applied:
- If the timestamp is missing, mark the record as invalid.
- If duplicate timestamps exist for the same meter, flag the record as invalid.

Reason:
Accurate time-series analysis depends on valid and non-duplicated timestamps.

Rule 4: Data Range Validation

Description:
Energy consumption values must fall within realistic and acceptable limits.

Logic Applied:

- If the energy value is less than zero, mark the record as invalid.
- If the energy value exceeds a predefined maximum threshold, flag the record as an anomaly.

Reason:
This rule helps detect erroneous readings and sensor malfunctions.

Rule 5: Faulty Meter Detection

Description:
A smart meter that reports zero or near-zero consumption for an unusually long continuous period may be faulty.

Logic Applied:

- If a meter reports zero or near-zero energy consumption over a prolonged period, mark the meter as **potentially faulty**.

Reason:
Early detection of faulty meters improves data quality and system reliability.

Rule 6: Schema Validation

Description:
Each record must follow the required data schema.

Required Fields:

- meter_id
- timestamp
- energy_value
- energy_unit **Logic Applied:**

- If any required field is missing, route the record to the failure path.

Reason:
Schema validation ensures consistency and prevents corrupted data from entering the system.

Task C: Single Record Lifecycle Explanation

This section explains the complete lifecycle of a single smart-meter record from ingestion to archival.

---

1. Upload to Raw Storage

A smart meter generates an electricity usage record and uploads it as part of a CSV file to the raw data storage.
The data is stored in its original form without any modification for backup and traceability purposes.

---

2. Triggering the Transformation Process

The arrival of a new file in raw storage automatically triggers the ETL orchestrator, which initiates the transformation workflow.

---

3. Data Cleaning and Validation

During the transformation phase, the system applies all defined business rules to the record, including:

- Standardizing energy units
- Handling missing values
- Validating timestamps
- Checking data ranges
- Detecting potentially faulty meters

---

4. Storage in Structured Format (RDS)

If the record passes all validation and cleaning rules, it is stored in a structured relational database (RDS) as a clean, query-ready record.

---

5. Conversion and Archival in Parquet Format

Clean records are periodically batched, converted into **Parquet format** , and archived in the analytics data lake.
This format is optimized for large-scale analytics, forecasting, and machine learning tasks.

---

6. Success and Failure Handling Success Case:

- The record is successfully stored in the structured database.
- It is archived in Parquet format.
- The operation is logged as a successful process.

Failure Case:

- The record is routed to the failed records storage.
- Error details are logged.
- Automatic retry mechanisms are applied.

- If retries fail, the record is flagged for manual review.