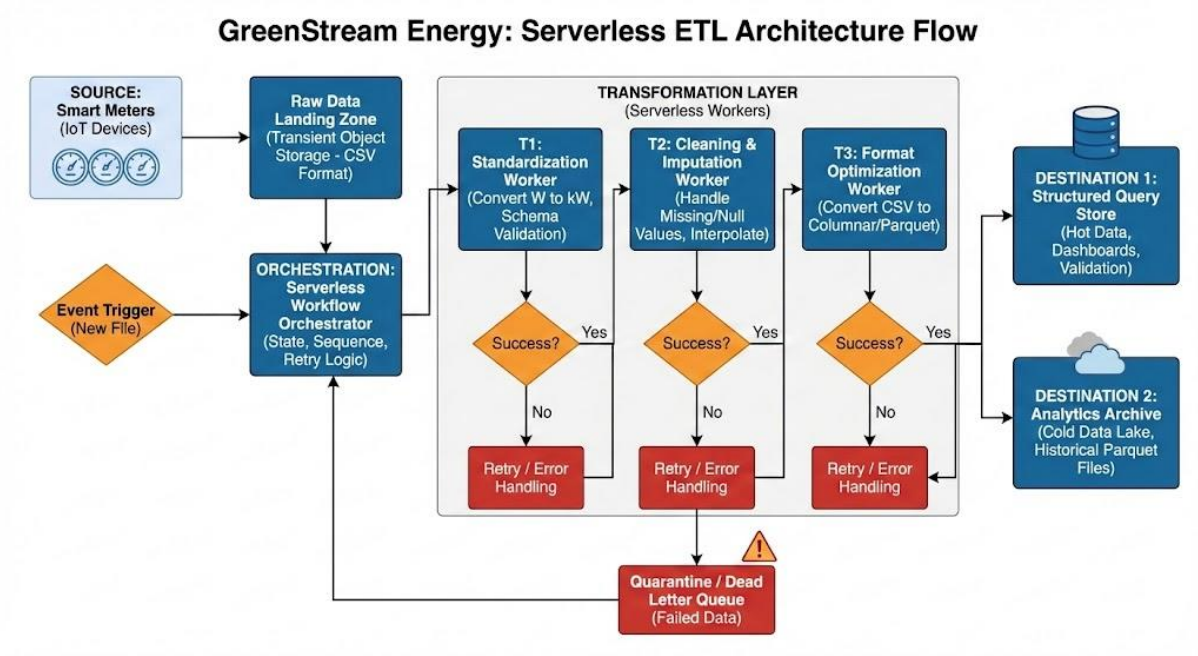
1-


GreenStream Energy: Serverless ETL Architecture Flow
SOURCE: Smart Meters (IoT Devices)
Raw Data Landing Zone (Transient Object Storage - CSV Format)
TRANSFORMATION LAYER (Serverless Workers)
T1: Standardization Worker (Convert W to kW, Schema Validation)
T2: Cleaning & Imputation Worker (Handle Missing/Null Values, Interpolate)
T3: Format Optimization Worker (Convert CSV to Columnar/Parquet)
DESTINATION 1: Structured Query Store (Hot Data, Dashboards, Validation)
Event Trigger (New File)
ORCHESTRATION: Serverless Workflow Orchestrator (State, Sequence, Retry Logic)
Success?
Yes
No
Retry / Error Handling
DESTINATION 2: Analytics Archive (Cold Data Lake, Historical Parquet Files)
Quarantine / Dead Letter Queue (Failed Data)

Transformation Logic & Business Rules Design:

This phase acts as the "intelligence" of the pipeline, ensuring that "dark data" is refined into high-quality assets for decision-making.

Rule Category	Business Rule Logic	Data Science Justification
Unit Standardization	If unit == "W", then value = value / 1000 and unit = "kW".	Essential for mathematical aggregation. You cannot calculate total grid load if units are mismatched.
Missing Values	If reading is NULL: 1. For gaps < 1 hour: Apply **Linear Interpolation**.	Prevents artificial "dips" in energy consumption charts caused by Wi-Fi outages rather than actual low usage.

Rule Category	Business Rule Logic	Data Science Justification
	2. For gaps > 1 hour: Mark as "Missing" and exclude from peak-load sums.	
Data Validation	If value < 0 or value > 50kW (per household hourly limit), move to **Quarantine**.	Filters out sensor noise or extreme outliers that would skew predictive forecasting models.
Faulty Meter Detection	If value == 0 for > 24 consecutive hours while status == 'online', flag record with is_faulty = TRUE.	Identifies hardware malfunctions or "dead" meters that need physical maintenance.

Single Record Lifecycle Explanation :

To understand how the system works, let's follow a **single data point** (one meter reading) from a household in the GreenStream network:

1. **Upload to Raw Storage:** The smart meter sends a small CSV snippet containing a reading (e.g., `MeterID: 101, Value: 1500, Unit: W, Time: 2025-12-22T10:00Z`). This file is uploaded to the

   Raw Landing Zone (Object Storage).

2. **Triggering the Process:** The arrival of this file creates an **S3/Object Event**. This event instantly triggers the **Serverless Orchestrator**, which spins up the first

transformation worker. No servers stay running; they only "wake up" to process this record.

3. **Data Cleaning & Validation:** The worker applies the business rules from Task B:
    - It sees `1500 W` and converts it to `1.5 kW`.
    - It checks the timestamp format.
    - It checks if the previous reading was missing and performs interpolation if necessary.
4. **Storage in Structured Format (RDS/Serving Layer):** The cleaned record (`MeterID: 101, Value: 1.5, Unit: kW...`) is inserted into a **Structured SQL Database (RDS)**. It is now "Live" and will immediately appear on the company's dashboard to help identify **peak energy periods**.
5. **Conversion and Archival (Parquet):** Once the "Hot" analysis is done, the record is bundled with others and converted into a **Parquet file**. This file is compressed and stored in the **Analytics Archive**. Because Parquet is columnar, a data scientist can later query 5 years of data for *only* "MeterID 101" without scanning the entire database.
6. **Success or Failure Handling:**
    - **Success:** The orchestrator logs a "Job Complete" status and moves the original raw CSV to a "Processed" folder.
    - **Failure:** If the record was malformed (e.g., text in the value field), the system attempts **3 retries**. If it still fails, the record is moved to a **Dead Letter Queue (DLQ)** for manual inspection, ensuring no data is ever "silently" lost.