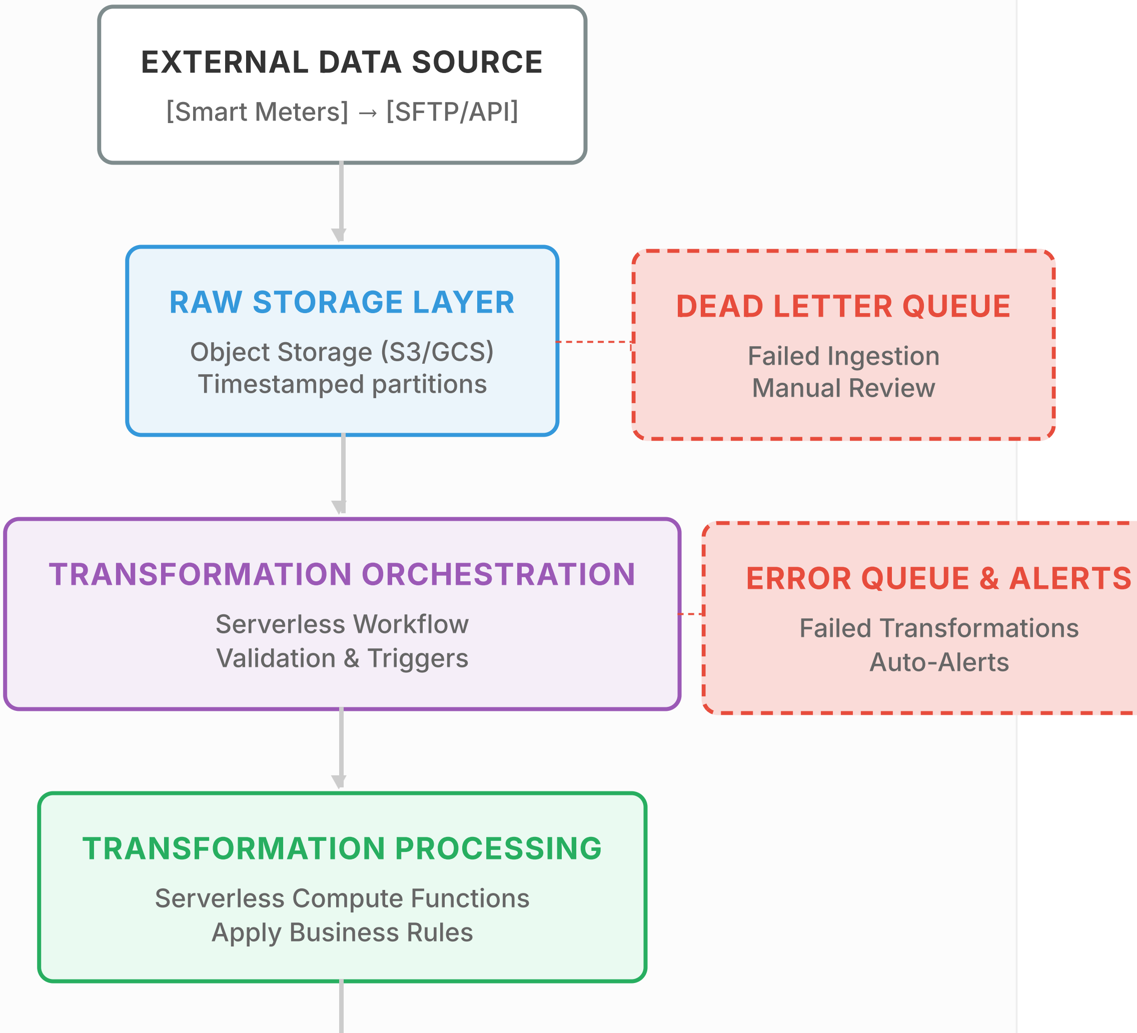
Hands-on Task: Design Thinking for Data Scientist

ETL Architecture & Data Transformation Pipeline

GREENSTREAM ENERGY PROJECT SUBMISSION

Task A: ETL Architecture Diagram

Conceptual Serverless ETL Pipeline for GreenStream Energy



EXTERNAL DATA SOURCE
[Smart Meters] → [SFTP/API]

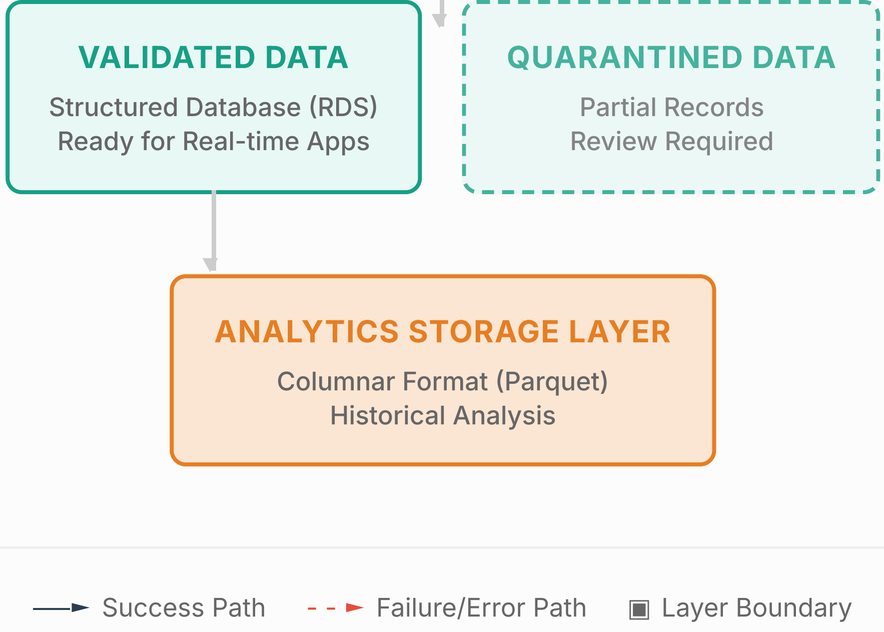
RAW STORAGE LAYER
Object Storage (S3/GCS)
Timestamped partitions

DEAD LETTER QUEUE
Failed Ingestion
Manual Review

TRANSFORMATION ORCHESTRATION
Serverless Workflow
Validation & Triggers

ERROR QUEUE & ALERTS
Failed Transformations
Auto-Alerts

TRANSFORMATION PROCESSING
Serverless Compute Functions
Apply Business Rules



Task B: Transformation Logic & Business Rules

⚡ Unit Standardization

- If unit is "W" or "watts", convert to kW (÷1000).
- If unit is "kWh", treat as kW for hourly readings.
- **Inference:** If unit missing: value < 10 = kW; value > 1000 = W.

🔍 Missing Value Handling

- Gap < 4h: Interpolate linearly. Auto-fix
- Gap 4h - 24h: Flag missing_short, exclude from peak.
- Gap ≥ 24h: Flag missing_extended, trigger alert.
- Reduce quality score proportionally to gap duration.

✓ Data Validation

- Verify `meter_id` against master registry.

⚠️ Faulty Meter Detection

- **Zero Value:** 0 kW for ≥ 48h (non-vacant).
- **Stuck Value:** No variance for ≥ 24h.

- Timestamp sanity check (±1 hour window).
- **Physical Limits:** Min 0 kW, Max 20 kW.
- **Rate Check:** >500% increase or >90% drop flagged.

- **Overload:** >20kW for ≥ 6 consecutive readings.
- **Low Variance:** 7-day CV < 0.01.
- **Atypical:** Correlation with neighborhood < 0.3.

📊 Peak Period ID

- Calculate hourly aggregates across all meters.
- Identify top 3 hours of consumption.
- Apply seasonal adjustments (Summer/Winter).
- Exclude holidays & extreme weather days.

⭐ Data Quality Scoring

- Start Score: **100**.
- **Deductions:**
  -5 (Unit Conversion)
  -10/hr (Interpolation)
  -15 (Warnings)
- **Threshold:** Score < 70 excluded from analytics.

Task C: Journey of a Single Smart Meter Reading

Scenario: Meter #MTR-78910 records 1250 Watts at 14:00:00

1

t_0 (14:00)

Data Generation & Upload

Meter records 1250W. Internal clock timestamps 2024-03-15 14:00:00. Transmits CSV string.

2

t_0 + 1 min

Raw Storage Ingestion

File arrives at S3 endpoint. Format validation occurs.

✓ Validation Success: Saved to `raw/2024/03/15/14/`

3

t_0 + 2 min

Orchestration Trigger

S3 event triggers Orchestrator. Job ID `TRN-20240315-1405-001` assigned.

4

t_0 + 3 min

Transformation Execution

Application of Business Rules (Task B).

- **Unit:** 1250 W converted to 1.25 kW
- **Limits:** 1.25 < 20kW (Pass)
- **Faults:** No historical fault patterns

Final Quality Score: 95/100 (-5 for unit conversion)

!

Alternative Path: Failure Scenario

If transformation fails (e.g. timeout):

1. Auto-retry triggered (up to 3x)
2. If fails: Alert sent to Data Engineering
3. File moved to Error Queue for manual review

5

t_0 + 4 min

Structured Storage (RDS)

Clean record inserted into SQL database. Immediately available for real-time dashboards.

6

t_0 + 60 min

Batch Aggregation

Hourly job runs at 15:00. Groups 14:00-14:59 records. Calculates Avg/Min/Max stats.

7

t_0 + 65 min

Parquet Archival

Data written to columnar format (Parquet) in Analytics bucket. Optimized for long-term query performance.

8

End State

Data Utilization

Immediate: Real-time dashboard usage.
Long-term: Data Scientist access via Athena/Presto for ML modeling.

Final Summary

⏱️

Efficiency

~65 mins end-to-end latency for analytics; <4 mins for real-time.

🛡️

Resilience

Dead Letter Queues and Auto-Retries ensure 0% data loss.

📈

Quality

Automated scoring ensures only high-trust data (Score >70) is used.

💾

Optimization

Parquet compression reduces long-term storage costs by ~70%.