Design Thinking for Data Scientist – ETL Pipeline Solution

Task A: ETL Architecture (Conceptual Design)

Source:
• Smart meters from 50,000 households upload raw CSV files to Raw Data Storage.

Transformation Layer:
• A serverless ETL function is triggered automatically on new data arrival.
• Data is cleaned, standardized, validated, and enriched.

Destinations:
• Clean structured data is stored in a relational database for querying and validation.
• Optimized Parquet files are archived in analytics storage for long-term analysis.

Orchestration & Error Handling:
• Workflow orchestration manages execution order and retries.
• Failed records are logged and sent to a Dead Letter Queue for inspection.
• Automatic retries occur for transient failures.

Logical Data Flow:

Smart Meters → Raw Storage → ETL Trigger → Transform & Validate → RDS (Success Path) → Parquet Archive
Transform Failure → Error Logs / Retry → Dead Letter Queue

Task B: Transformation Logic & Business Rules

- Rule 1: If energy unit = 'W', divide the value by 1000 and convert the unit to 'kW'.
- Rule 2: If energy unit = 'kW', keep the value unchanged.
- Rule 3: If an energy reading is NULL, flag the record and exclude it from peak consumption analysis.
- Rule 4: If missing values are short gaps, interpolate using the previous valid reading.
- Rule 5: Reject records with negative or extremely high values beyond physical limits.
- Rule 6: If a meter reports zero or constant consumption for an unusually long time, flag it as potentially faulty.
- Rule 7: Store validation status and fault flags as additional columns for traceability.

Task C: Single Record Lifecycle

1. A smart-meter record is uploaded in CSV format to raw storage.
2. The upload event automatically triggers the serverless ETL process.
3. The transformation layer checks units, converts values to kW, and validates ranges.
4. Missing or abnormal values are flagged according to business rules.
5. Clean and validated records are written to a structured relational database.
6. The same data is converted into Parquet format and archived for analytics.
7. If processing succeeds, the workflow is marked complete.
8. If processing fails, the system retries automatically; persistent failures are logged and sent to a Dead Letter Queue.