**Realistic Scenario-Based Problem Statement: ETL Pipeline for Retail Sales Analytics**

**Business Scenario**

You are a data engineer at **RetailMart**, a mid-sized e-commerce company specializing in electronics and apparel. The company processes daily sales data from multiple online stores, but the raw data ingested into Azure Data Lake Storage Gen2 (ADLS2) is messy due to legacy systems, manual entries, and third-party integrations. This leads to inaccurate reporting, such as inflated revenue figures from duplicate transactions or lost insights from incomplete customer profiles.

**Problem Statement**: Design and implement an ETL (Extract, Transform, Load) pipeline in Azure Data Factory (ADF) using **mapping data flows** to:

1. **Extract** raw sales transaction data and customer profile data from ADLS2.
2. **Transform** (clean and enrich) the data: Handle missing values, remove duplicates, standardize formats (e.g., dates, currencies), join datasets, and aggregate key metrics (e.g., total sales per customer segment).
3. **Load** the cleaned and aggregated data into a structured table in **Azure Synapse Analytics** (dedicated SQL pool) for downstream analytics, such as Power BI dashboards for sales forecasting and customer segmentation.

**Key Challenges**:

* Data quality issues: Null values in customer emails, duplicate transactions (e.g., same order ID processed multiple times), invalid dates (e.g., future dates or non-standard formats), and inconsistent currency codes.
* Scalability: Handle up to 1 million rows daily, with incremental loads for new data.
* Compliance: Ensure PII (e.g., emails) is anonymized or hashed for privacy (GDPR compliance).
* Scheduling: Run the pipeline daily at 2 AM, with alerts on failures.

**Success Criteria**:

* Pipeline processes data end-to-end in under 30 minutes.
* Output table in Synapse contains accurate, deduplicated data with aggregates (e.g., average order value by region).
* Verify with sample queries showing clean data (e.g., no duplicates, all dates valid).

This scenario mirrors real-world retail ETL workflows, where clean data drives decisions like inventory restocking or targeted marketing.

**Sample Datasets**

We'll use two CSV files stored in ADLS2 (container: retail-data, paths: /raw/sales-transactions.csv and /raw/customer-profiles.csv). These are small (under 100 rows) for demo purposes but scalable. Upload them to your ADLS2 account for testing. The datasets include intentional issues for cleaning.

**1. Sales Transactions (sales-transactions.csv)**

Raw data from point-of-sale systems. Issues: Duplicates (rows 3 and 7), null quantities, invalid dates (e.g., "2025-02-30"), mixed currency (USD/EUR).

| **OrderID** | **CustomerID** | **Product** | **Quantity** | **UnitPrice** | **OrderDate** | **Currency** |
| --- | --- | --- | --- | --- | --- | --- |
| 1001 | CUST001 | Laptop | 1 | 999.99 | 2025-10-01 | USD |
| 1002 | CUST002 | Shirt | null | 29.99 | 2025-10-02 | USD |
| 1003 | CUST003 | Phone | 2 | 699.00 | 2025-10-03 | USD |
| 1004 | CUST004 | Jeans | 1 | 49.99 | InvalidDate | EUR |
| 1005 | CUST001 | Headphones | 1 | 149.99 | 2025-10-05 | USD |
| 1003 | CUST003 | Phone | 2 | 699.00 | 2025-10-03 | USD |
| 1006 | CUST005 | Watch | 1 | 199.00 | 2025-02-30 | USD |
| 1007 | CUST006 | Shoes | 3 | 89.99 | 2025-10-07 | EUR |

**2. Customer Profiles (customer-profiles.csv)**

Demographic data from CRM. Issues: Missing emails, inconsistent regions (e.g., "USA" vs. "US"), null ages.

| **CustomerID** | **Name** | **Email** | **Age** | **Region** | **Segment** |
| --- | --- | --- | --- | --- | --- |
| CUST001 | Alice Johnson | [alice@retailmart.com](mailto:alice@retailmart.com) | 28 | North America | Premium |
| CUST002 | Bob Smith | null | 35 | Europe | Standard |
| CUST003 | Charlie Lee | [charlie@retailmart.com](mailto:charlie@retailmart.com) | null | Asia | Premium |
| CUST004 | Diana Wong | [diana.wong@email.com](mailto:diana.wong@email.com) | 42 | North America | Standard |
| CUST005 | Eve Garcia | [eve@retailmart.com](mailto:eve@retailmart.com) | 31 | South America | Budget |
| CUST006 | Frank Kim | [frank.kim@email.com](mailto:frank.kim@email.com) | 29 | Asia | Premium |
| CUST007 | Grace Patel | [grace@retailmart.com](mailto:grace@retailmart.com) | 37 | Europe | Standard |

**Expected Dataflow Transformations**

In your ADF dataflow:

* **Sources**: Load both CSVs from ADLS2.
* **Join**: Inner join on CustomerID.
* **Clean**:
  + Filter: Remove rows with null Quantity or invalid dates (use isDate(OrderDate)).
  + Derived Column: Standardize Currency to USD (convert EUR: multiply UnitPrice by exchange rate, e.g., 1.1), hash Email (use sha2(Email, 256) for anonymization), fix Region (map "North America" to "NA", etc.).
  + Remove Duplicates: On OrderID.
  + Fill Nulls: Default Age to 0 or median (e.g., 32).
* **Aggregate**: Group by Region and Segment; compute TotalRevenue (sum(Quantity \* UnitPrice)), AvgOrderValue (avg(Quantity \* UnitPrice)).
* **Sink**: Load to Synapse table dbo.SalesSummary (columns: Region, Segment, TotalRevenue, AvgOrderValue).

**Expected Output Table (After Cleaning)**

After running the pipeline, query SELECT \* FROM dbo.SalesSummary; in Synapse:

| **Region** | **Segment** | **TotalRevenue** | **AvgOrderValue** |
| --- | --- | --- | --- |
| NA | Premium | 1148.98 | 1148.98 |
| Europe | Standard | 89.99 | 89.99 |
| Asia | Premium | 1398.00 | 1398.00 |
| South America | Budget | 199.00 | 199.00 |