Azure Data Factory End-to-End Pipeline Project

Trainer / Creator: Dr. Sandeep Kumar Sharma

Project Title: Sales Analytics Mini Project using Azure Data Factory

This project walks a beginner through creating a complete end-to-end Azure Data Factory (ADF) pipeline using a clean, meaningful dataset. The dataset simulates real sales data (about 5–8 columns, not too small, not too complex). The pipeline extracts raw data from storage, transforms it using Data Flow, and loads the final curated report into a clean zone.

All steps are written in very simple language, keeping the project beginner-friendly.

Learning Objectives

- Learn how to design a simple but real-world ADF pipeline.
- Create structured data zones: **Raw Zone** → **Clean Zone** → **Curated Zone**.
- Build datasets (source + transformed outputs).
- Use **Copy Activity** + **Data Flow** for light transformation.
- Publish and run a fully functional ADF pipeline.

Business Scenario

Imagine you work in a retail company. Every day the sales team uploads a CSV file containing the following fields:

Column	Meaning
OrderID	Unique ID of each sale
OrderDate	Date of sale
CustomerName	Name of customer
ProductName	Item purchased
Quantity	Items sold
UnitPrice	Price per item
TotalAmount	Quantity × UnitPrice

This dataset is simple but meaningful for reporting.

Your goal:

1. Upload this dataset to the **Raw** folder in Azure Storage.
2. Use ADF to **copy** the raw data to a **Clean** folder.
3. Use a **Mapping Data Flow** to calculate total sales by product.
4. Store the final report in a **Curated** folder.

Architecture Overview

```
        [Raw Zone]
           |
           | Copy Activity
           v
        [Clean Zone]
           |
           | Data Flow Transformation
           v
        [Curated Zone]
```

Step-by-Step Instructions

STEP 1 — Create folders inside Azure Storage

1. Go to your Storage Account.
2. Open **Containers** → create a container named: `salesdata`.
3. Inside this container, create three folders:
4. `raw`
5. `clean`
6. `curated`
7. Download or create a CSV file named `daily-sales.csv` with columns as described above.
8. Upload it into the **raw** folder.

STEP 2 — Open Azure Data Factory Studio

1. Go to your Azure Data Factory → Click **Launch ADF Studio** / **Author & Monitor**.
2. Once opened, go to the **Manage** tab.
3. Make sure your Storage Linked Service is already created.

STEP 3 — Create Datasets

Dataset 1: Raw Sales Dataset

1. Go to **Author → Datasets → + New dataset**.
2. Select **Azure Blob Storage → CSV**.
3. Name: `DS_RawSales`.
4. Linked service: select your storage linked service.
5. File path: `salesdata/raw/daily-sales.csv`.
6. First row as header: **Yes**. Click **OK**.

Dataset 2: Clean Sales Dataset

1. Create another CSV dataset.
2. Name: `DS_CleanSales`.
3. Path: `salesdata/clean/` (folder only — ADF will create file automatically).

Dataset 3: Curated Final Report

1. New CSV dataset.
2. Name: `DS_CuratedReport`.
3. Path: `salesdata/curated/`.

STEP 4 — Create the Pipeline

1. Go to **Author → Pipelines → + New Pipeline**.
2. Name it: `PL_SalesAnalytics`.

STEP 5 — Add Copy Activity (Raw → Clean)

1. Drag **Copy Data** activity.
2. Name: `Copy_Raw_to_Clean`.
3. Source: select `DS_RawSales`.
4. Sink: select `DS_CleanSales`.
5. Leave all transformations as default.
6. This step simply copies raw data to a clean folder.

STEP 6 — Create a Mapping Data Flow

6.1 Create Data Flow

1. Go to **Author → Data flows → + New Mapping Data Flow**.
2. Name: `DF_SalesAggregation`.

3. Add **Source** → Dataset: `DS_CleanSales`.
4. Add **Aggregate** transformation.

6.2 Configure Aggregate

Group by:

- `ProductName`

Aggregates:

- `TotalQuantity = sum(Quantity)`
- `TotalRevenue = sum(TotalAmount)`

6.3 Add Sink

1. Add a Sink transformation.
2. Dataset: `DS_CuratedReport`.
3. File naming option: Auto.

6.4 Publish the Data Flow

Click **Publish**.

STEP 7 — Add Data Flow to Pipeline

1. Go back to pipeline `PL_SalesAnalytics`.
2. Drag **Data Flow** activity below the Copy Activity.
3. Select Data Flow: `DF_SalesAggregation`.
4. Connect the Copy Activity output → Data Flow input.

Pipeline now looks like:

```
Copy_Raw_to_Clean → DF_SalesAggregation
```

STEP 8 — Debug and Publish

1. Click **Debug** to test.
2. If pipeline succeeds:
3. The raw file is copied to `/clean/`
4. The aggregated report is created in `/curated/`
5. Click **Publish All** to save changes.

FINAL OUTPUT (Result)

Inside `curated` folder, a CSV file will be generated with the following columns:

ProductName	TotalQuantity	TotalRevenue

This file gives a clear summary of which product generated how much revenue.

Learning Outcome

By completing this project, participants understand:

- How to design simple but real-world ADF pipelines.
- How Raw → Clean → Curated zones work.
- How to use **Copy Activity** and **Mapping Data Flow**.
- How to work with meaningful datasets.
- How to generate useful business insights from raw data.
