

**Retail Sales Data Pipeline - Project Documentation**

Boot Camp Project01



Student Name – Sai Sravan Kumar Repaka

Trainer – Dr. Praveen Sharma

**1. Project Overview**

This project demonstrates the development of an end-to-end retail sales data pipeline using **Azure Data Factory**, **Azure Data Lake**, and **Power BI**. The pipeline extracts data from multiple sources including a REST API, on-prem SQL Server, and CSV files, performs transformations, and visualizes the data for insights.

**2. Architecture Diagram**A diagram of a data flow

AI-generated content may be incorrect.The architecture includes:

* REST API via Ngrok tunnel
* CSV and JSON files from Blob Storage
* On-Prem SQL Server via Self-hosted IR
* Azure Data Factory for orchestration and transformation
* Azure Data Lake Storage (Bronze/Silver)
* Azure SQL DB for curated data
* Power BI for reporting

| **3. Technologies Used** |  |
| --- | --- |
| |  |  | | --- | --- | | **Component** | **Technology** | | Data Orchestration | Azure Data Factory | | Storage | Azure Data Lake (Gen2) | | Integration Runtime | Self-hosted Integration Runtime | | Source Data | REST API, Blob CSV, On Prem SQL Server tables | | Target | Azure SQL Database | | Visualization | Power BI | |  |

**4. Source Data**

Three datasets were used:

* **Customers**: Customer details (Name, Email, Location, etc.)
* **Products**: Product catalog
* **Sales**: Transactional sales data

**5. Ingestion Pipelines**

**a. REST API (ngrok)**

Since on the internet I did not find any publically Api’s which are available to use based on our schema, I’ve decided to run the Json server locally on my computer with db.json file which as Customers, Products and Sales records in the form of Json file and using Ngrok I publically exposed this file and used the ADF lookup to pass the variables as relative URL for the Rest API.

**A black screen with white text

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

* Used a Web activity to call REST API
* Loaded data into Bronze layer in ADLS

**b. On-Prem SQL Server**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

* Used Self-hosted IR for secure data access
* Extracted data into ADLS Bronze

**c. CSV from BlobA screenshot of a computer

AI-generated content may be incorrect.**

* CSV data ingested using dataset and copy activity

The pipeline which I have created to copy all three sources to the bronze container is attached belowA screenshot of a computer

AI-generated content may be incorrect.

**6. Transformation (Data Flow)**

**A screenshot of a computer

AI-generated content may be incorrect.**

* Applied data cleaning and transformations such as removing nulls, filtering out the duplicates rows based on row-number, sorting the columns, applying joins etc. in ADF Data Flow
* Output saved to **Silver Layer** in ADLS Gen2 for all the individual files and the tables which are joined after transforming are saved Azure SQL Database.

A screenshot of a computer

AI-generated content may be incorrect.

**7. Loading to Azure SQL DB**

**A screenshot of a computer

AI-generated content may be incorrect.**

* Final transformed data moved from Silver layer to Azure SQL Database

**8. Power BI Reporting**

**A screenshot of a graph

AI-generated content may be incorrect.**

* Connected to Azure SQL DB
* Created visuals for sales trends, top customers, top products, etc.

**9. Data Validations**

* Row count comparison before and after transformation
* Schema validation through pipeline monitoring

**Challenges faced**

1. Finding out a way to create our own API dataset and expose them publicly is the biggest challenge I faced in this project since our other datasets from different sources have different schemas.
2. In dataflows, we can’t add multiples files with different sources to a single source. So, I have taken only source from On-prem and performed the transformations, otherwise for 3 files from 3 different sources for bronze folder, would mean 9 sources in data flow.