**Retail Sales Data Pipeline**

**1. Project Overview**

The **Retail Sales Data Pipeline** automates the extraction, transformation, and loading (ETL) of sales data from multiple sources into Azure-based storage, enabling efficient business intelligence reporting in Power BI.

**2. Architecture Diagram**

**A diagram of a retail sales pipeline architecture

AI-generated content may be incorrect.**

**3. Data Sources & Storage Layers**

**Data Sources**

* **API (HTTP/RESTful)** - Real-time sales transactions
* **CSV Files** – Customer information
* **On-Prem SQL Server** - product information

**Storage Layers**

* **Bronze Layer (Raw Storage)** - Stores unprocessed data in **Azure Data Lake Storage (ADLS)**.
* **Silver Layer (Cleansed & Transformed)** - Standardized and cleaned data using **ADF Data Flow**.
* **Gold Layer (Curated Data)** - Aggregated data optimized for reporting in **Azure SQL Database**.

**4. Data Ingestion Process**

* **API Data**: Extracted using Azure Data Factory (ADF) via RestAPI Connector.
* **CSV Files**: Uploaded to ADLS and processed using ADF pipelines.
* **SQL Server Data**: Extracted using ADF’s Copy Data activity with self hosted integration runtime.

**Role:**

As part of data engineering team, I mainly worked on creation of self hosted integration runtime and data flow implementation part for moving data from raw layer to silver.

**Self-Hosted Integration Runtime (SHIR) Setup:**

Since the project involved extracting data from **On-Prem SQL Server**, we configured **Self-Hosted Integration Runtime (SHIR)** in Azure Data Factory.

**Steps for SHIR Setup:**

1. **Installation**: Installed SHIR on an on-prem server.
2. **Configuration**:
   * Registered SHIR in **Azure Data Factory**.
   * Configured network settings for secure data movement.
3. **Pipeline Setup**:
   * Created **Linked Service** in ADF to connect to On-Prem SQL Server.
   * Validated secure connectivity and access controls.

**Outcome:** Enabled secure **on-prem to cloud** data transfer for historical sales data.

**Business Logic Transformations**

Implemented logic as per requirements:

1. **Revenue per Product** 
   * Aggregated total revenue per product.
2. **Top Customers by Total Spend** 
   * Summed totals spend per customer.
3. **Inventory Turnover Calculation** 
   * Tracked stock levels and product movement.

**Outcome: The Silver Layer contained clean, structured, and enriched data ready for analytics.**

**Sales Retail Workspace (Resource Group):**

Holds necessary configured resources for this project as below

A screenshot of a computer

AI-generated content may be incorrect.

**Service Principal:**

Retail\_App service principal is created and assigned role of Storage Blob data contributor to both backednd adls and bronze layer adlsgen2 storage account.

**A screenshot of a chat

AI-generated content may be incorrect.**

**Azure Key Vault:**

As part of this process, linked services and datasets were created as needed. For linked services, authentication type Service Principal is utilized for better security and Azure Key vault for storing service Principal Secrets.

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Datasets:**

Necessary datasets for this process are created leveraging linked services as below.

**A screenshot of a computer

AI-generated content may be incorrect.**

**Step 1 – (Ingesting data to raw layer)**

Purpose of this step is to get data from backend data sources to raw layer as its raw form. Get Metadata1 activity will fetch list of files from backend adls container and forEach1 will loop through each file from get metadata1 activity and copy activity inside forEach will copy each file to raw layer in its raw form. Pipeline sales to raw will read sales data, which holds necessary connection details for other sales file in GitHub and copy files into raw layer in its raw form.

**Sales to raw:**

**A screenshot of a computer

AI-generated content may be incorrect.**

This pipeline is configured with a tumbling trigger to run every 8 hours.

A screenshot of a computer

AI-generated content may be incorrect.

**Sales retail pipeline:**

This will copy data from both on-premises SQL server and backend blob storage into raw layer.

A screenshot of a computer

AI-generated content may be incorrect.

Schedule trigger is configured to this pipeline to run every one hour.

A screenshot of a computer screen

AI-generated content may be incorrect.

**Raw container:**

As expected after a pipeline run, all raw data from all 3 data sources are ingested into bronze layer.

A screenshot of a computer

AI-generated content may be incorrect.

**Step 2: Raw to Silver (Data Cleaning)**

Data flow was leveraged to do necessary cleaning, and validations are done as below

A screenshot of a computer

AI-generated content may be incorrect.

**Silver Container:**

As expected after a pipeline run, all data from raw layer is cleaned and ingested into silver layer.

A screenshot of a computer

AI-generated content may be incorrect.

**Step 3: Business logic transformation**

* + **Revenue per Product**: Total sales revenue calculated in Gold Layer.
  + **Top Customers**: Ranking customers based on total spend.
  + **Inventory Turnover**: Measuring stock movement against sales.

**A screenshot of a computer

AI-generated content may be incorrect.**

**Azure SQL data warehouse:**

As expected after a pipeline run, all silver data from transformed aggregated and ingested into respective tables as below:

**Product Performance Table:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Top Customers Table:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Revenue per product table:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Pipeline Run:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Power Bi report:**

Power Bi reporting tool is used to generate reports. Power Bi is connected to azure SQL database and pulled records from it as below

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

GitHub link : **https://github.com/Arunkumar-Senthilkumar/Retail-Sales-ETL**