**Data Processing Pipeline**

* This project is taking Retail stores data across the globe and doing aggregations and sending the outcome to downstream consumers (internal or external) for reporting and analysis.
* **DailyDataIngestAndRefine.py**

This script collects the sales data from landing zone and segregates the data into valid and invalid based on some conditions set by business and store them at different locations. Further it is adding hold reason column to the invalid data.

* **EnrichProductReference.py**

This script is reading product reference data from different source and merging with valid data to derive sale amount by multiplying product price and qty sold, storing to different location.

* **VendorEnrichment.py**

This script is reading vendor data and USD reference data from different sources and deriving the performance of each vendor and converting all the currency to base USD currency. This data is stored in different locations.

* **Live\_Stream.py**

This is a sample script to read the sales data as a stream (micro batches) for live sales data. It streams the live data and calculates sale amount for live data. Gets aggregated values of total sales and quantities of products being sold by each vendor in streaming data.

* **Refer to orchestration document for pipeline orchestration.**