Databricks Project Architecture

Overview

This project outlines the end-to-end architecture and implementation of a data processing pipeline using **Azure Databricks**, **Azure Data Lake Storage Gen2 (ADLS Gen2)**, **Snowflake**, and **Azure Data Factory (ADF)**. The goal is to process raw data, clean and transform it, and generate key performance indicators (KPIs) for business use.

Architecture Workflow

Source Layer (Raw Data Ingestion)
The source data is stored in ADLS Gen2 under a designated Raw Zone path. This data should be copied to your Azure account in ADLS Gen2 i.e Bronze Zone using Azure Data Factory (ADF).

Bronze Layer (Raw Zone in Databricks)
In the Raw Zone, basic data quality checks are performed. Null values and duplicate records are identified and handled. The cleaned data is then written to the Silver Zone for further transformations.

Silver Layer (Data Modelling and Preparation for KPIs)
This layer handles further cleaning, enrichment, and transformation. Data modelling is performed to prepare the dataset for KPI calculation. Relevant **dimension** and **fact tables** are created.

Gold Layer (KPI Computation)
In the Gold Zone, cleaned and modelled data is joined and transformed to derive KPIs. All KPI logic is implemented in Databricks notebooks, and the final results are stored back in ADLS Gen2.

Data Export to Snowflake
Using Azure Data Factory (ADF), the final KPI dataset is moved from ADLS Gen2 to Snowflake for business reporting and dashboard integration.

ADF Orchestration

In the Azure Data Factory (ADF) pipeline, each KPI dataset is first stored in a gold layer with sub-folders within Azure Data Lake Storage Gen2 (ADLS Gen2) to maintain logical separation and facilitate organized data management. Subsequently, ADF orchestrates the movement of each individual KPI dataset from its respective container into Snowflake, where the data is utilized for business reporting and dashboard integration.

List of KPIs:

KPI 1:

Q1: Customer Retention Rate (Returning Customers)

Description: Measures the percentage of customers who return and make repeat purchases within a year compared to those who made purchases in the previous year.

Expected O/P:

- Year
- percentage

Q2: Monthly Churn Rate (Customers Who Didn't Come Back)

Description: Indicates the percentage of customers who made a purchase in the previous month but did not return in the current month.

Expected O/P:

- Month
- Percentage

KPI 2:

Q: Product Return/ Cancelled Rate by Brand

Description: Measures the percentage of products returned or cancelled for each brand relative to total orders

Expected O/P:

- Product Brand
- Return_Rate

KPI 3:

Q: Average Product Rating by Category

Description: Calculates the average customer rating for products within each category.

Expected O/P:

Product Category

- Avg_rating

KPI 4:

Q: Payment Method Preference

Description: Identifies the most frequently used payment methods by customers, helping to understand user preferences and optimize checkout experiences.

Expected O/P:

- Payment_Method
- Frequency

KPI 5:

Q: Most Product Purchases by Day, Weekday, Month, and Year

Description: This KPI analyzes when customers are most actively making purchases by breaking down transaction data across multiple timeframes. It examines:

- **Day-wise trends** to spot specific high-activity dates.
- **Weekday patterns** (e.g., Mondays vs. weekends) to understand shopping behavior throughout the week.
- **Monthly performance** to identify seasonal spikes or dips.
- **Yearly comparisons** to observe growth or decline over time.
- Expected O/P:

  [Daywise/weekdays/months/year]

  Frequency of product purchase

KPI 6:

Q: Average Delivery Time by Shipping Method (in days)

Description: This KPI calculates the average number of days taken to deliver a product, from the time of order placement to delivery. The output provides the average delivery time for each year and month, allowing for a detailed analysis of delivery performance over time.

Expected O/P:

- Year
- Month
- Shipping_Method
- Avg_delivery_time

KPI 7:

Q: Find the customers who bought the same product repeatedly over time.

Description: Identify customers who have purchased the same product multiple times over a period.

Expected O/P:

- Customer_name
- product_name
- order_count

KPI 8:

Q: Revenue Risk Due to Returns Purpose: Detect how much revenue is being lost by specific locations.

Description: Measure the financial impact of product returns on revenue for each country. This metric helps detect how much revenue is being lost due to returns in specific countries.

Expected O/P:

- Country
- State
- Total_revenue
- revenue_lost
- revenue_risk_percentage)

KPI 9:

Q: Most Product Purchase Frequency by Timeslot.

Description: Measure the frequency of product purchases within specific 6-hour timeslots.

Timeslot Breakdown:

- 12 AM - 6 AM
- 6 AM - 12 PM
- 12 PM - 6 PM
- 6 PM - 12 AM

Expected O/P:

- Timeslot
- Frequency

KPI 10:

Q: Product Type Revenue by Gender & Age Group

Description: Analyze which product type generates more revenue segmented by gender and age groups (e.g., 18–25, 26–35, etc.).

Expected Output Columns:

- `Product_Type`
- `Gender`

- Age_Group
- Total_Revenue
- Avg_Rating

KPI 11:

Q: Repeat Purchase Score by Product

Description: Identify products with a high repeat purchase rate (same customer buying the same product more than once).

Expected Output Columns:

- Product_name
- Product_Category
- Repeat_Purchase_Count
- Repeat_Purchase_Percentage
- Top_Repeat_Customers(count of cust)

KPI 12:

Q: Brand Loyalty Score

Description: For each customer, calculate how many times they purchased from the same brand. Helps identify brand loyalists.

Expected Output:

- Customer_ID
- Brand
- Repeated_Purchase_Count
- Last_Purchase_Date
- Avg_Rating_Given