P2 Project - D&A - Nikhil Sharma @Revtaure

Project Overview: E-Com Insight Pipeline

The E-Com Insight Pipeline is a comprehensive data engineering project designed to streamline the process of generating, cleansing, analyzing, and visualizing E-Commerce data. The goal is to equip the Analytics team with actionable insights derived from large-scale data, enabling stakeholders to make well-informed business decisions.

Project Components:

1. `Data Generation`: Develop a Python program to generate CSV data that adheres to the specified E-Commerce data schema. This data will simulate **real-world transactions** and include rogue records to enhance the robustness of analysis.

2. `Exploratory Data Analysis (EDA) and Data Cleansing`: Before loading the data into GCS, perform EDA to identify patterns, inconsistencies, or outliers. Cleanse the data by:
   - Removing duplicates or irrelevant records.
   - Standardizing formats (e.g., dates, numeric values).
   - Addressing rogue or incomplete records.

   Upload both the files to GCS, raw as well as cleaned.

3. `Data Upload`: Implement functionality within the Python program to upload the generated CSV files to Google Cloud Storage (GCS). This step ensures that the data is securely stored and accessible for further processing.

4. `Data Loading`: Utilize BigQuery to load the cleansed CSV data from GCS. This stage involves configuring BigQuery tables to accommodate the E-Commerce data structure and facilitate efficient querying.

5. `Data Analysis`: Perform data analysis within BigQuery to address specific marketing questions:
   - Identify the top-selling product categories by country.
   - Analyze product popularity trends throughout the year by country.
   - Determine locations with the highest sales traffic.
   - Assess peak sales times by country.

6. `Data Visualization`: Connect BigQuery to Looker to create dynamic visualizations that provide clear and actionable answers to the marketing team's questions. These visualizations will support data-driven decision-making and help in understanding sales patterns and customer behavior.

Objective:

The E-Com Insight Pipeline aims to build a robust ETL (Extract, Transform, Load) pipeline that seamlessly integrates data generation, cleansing, storage, analysis, and visualization processes. By implementing this pipeline, the project will facilitate efficient data handling and insightful reporting, contributing to strategic business decisions and improved operational efficiency.

Data Analysis Questions – Marketing Department

1. What is the top-selling category of items? Per Country?
2. How does the popularity of products change throughout the year? Per Country?
3. Which locations see the highest traffic of sales?
4. What times have the highest traffic of sales? Per Country?
5. What is the average order value across different product categories? Per Country?
6. How do payment methods impact sales volume and success rates? Per Country?
7. What are the common reasons for payment failures, and how do they vary by country?

E-Commerce Data Structure

Fields (Schema)

Field name	Description
`order_id`	Order Id
`customer_id`	Customer Id
`customer_name`	Customer Name
`product_id`	Product Id
`product_name`	Product Name
`product_category`	Product Category
`payment_type`	Payment Type (Card, Internet Banking, UPI, Wallet)
`qty`	Quantity ordered
`price`	Price of the product
`datetime`	Date and time when the order was placed
`country`	Customer Country
`city`	Customer City
`ecommerce_website_name`	Site from where order was placed
`payment_txn_id`	Payment Transaction Confirmation Id
`payment_txn_success`	Payment Success or Failure (Y=Success, N=Failed)
`failure_reason`	Reason for payment failure

Sample Data (CSV)

```
1,101,John Smith,201,Pen,Stationery,Card,24,10,2021-01-10
10:12,India,Mumbai,www.amazon.com,36766,Y,
2,102,Mary Jane,202,Pencil,Stationery,Internet Banking,36,5,2021-10-31
13:45,USA,Boston,www.flipkart.com,37167,Y,
3,103,Joe Smith,203,Some mobile,Electronics,UPI,1,4999,2021-04-23
11:32,UK,Oxford,www.tatacliq.com,90383,Y,
4,104,Neo,204,Some laptop,Electronics,Wallet,1,59999,2021-06-13
15:20,India,Indore,www.amazon.in,12224,N,Invalid CVV.
5,105,Trinity,205,Some book,Books,Card,1,259,2021-08-26
19:54,India,Bengaluru,www.ebay.in,99958,Y,
```

INSTRUCTIONS

1. Standard Functional Scope

The scope of the E-Com Insight Pipeline includes the following key functions:

- **Data Generation**: A Python program capable of generating a CSV file with ~10,000 records, adhering to the defined E-Commerce schema, and introducing rogue records for testing purposes, the generated data should depict real world transactions.
- **EDA:** Perform EDA and Data cleansing and upload both the raw as well as cleansed file.
- **Data Upload**: Automatic upload of the generated CSV file to Google Cloud Storage (GCS) using the Python program.
- **Data Loading**: Loading data from GCS into BigQuery, ensuring it follows the E-Commerce schema. All relevant transformations and data validation steps will be handled to guarantee data quality.
- **Data Analysis**: Running SQL queries in BigQuery to answer specific marketing questions, providing insights such as product popularity, sales traffic, and peak sales times by country.
- **Data Visualization**: Connecting BigQuery to Looker to generate dynamic and informative visualizations that display insights graphically for the stakeholders.

2. Definition of Done

The project is considered complete when the following criteria are met:

- The Python program successfully generates a CSV file with 10,000 records and rogue entries.
- The generated CSV data is uploaded to GCS and loaded into BigQuery without errors.
- Data analysis queries are executed in BigQuery, providing accurate results for all predefined marketing questions.
- Visualizations in Looker display the results of the data analysis clearly and effectively.

- All relevant documentation (architecture, ETL process, and data models) is provided.
- The code repository is finalized and shared for review.

3. Submission

The final submission should include the following:

1. **Test Script**: A Python or SQL script that fetches 5 records from each BigQuery table, ensuring that data has been loaded correctly.
2. **Code Repository**: The associates' code repository must be shared for technical review, including:
   - **Architecture**: Clear documentation of the project architecture, showing the flow from data generation to visualization.
   - **Data Models**: Diagrams or tables representing the data structure in BigQuery.
   - **ETL Documentation**: Comprehensive documentation explaining the data extraction, transformation, and loading processes, including any data quality checks or validations.

4. Non-Functional Expectations

- **Version Control**: The project codebase should be managed using a version control system, such as Git. All team members are expected to contribute using feature branches, with regular commits and code reviews.
- **Scrum Process**: The project should follow Scrum methodology. This includes sprint planning, daily stand-ups, backlog grooming, and sprint retrospectives to ensure collaborative and iterative progress toward project completion.