DYNAMIC PLANNER™

ENSURING INVESTMENT SUITABILITY

Dynamic Planner

Data Engineer Case Study

A. Introduction

You're a Data Engineer working for an eCommerce company called Stenelem Superstore and you've been asked to ingest, process, and present data for the Head of Data and the Head of Engineering about the performance of their business.

The senior management is keen to understand the performance their business, including revenue, products, & orders. As they plan to increase their customer base, any information about customer behaviour will also be insightful them.

B. Data

All data has been provided by Kaggle. *(CC BY-NC-SA 4.0)*

- Orders dataset: This dataset has information about each item that was ordered.
- Order items dataset: This dataset includes data about the items purchased within each order.
- Order payments dataset: This dataset includes data about the orders payment options.
- Product dataset: This dataset includes data about the products sold by Stenelem.
- Product category name translated dataset: This dataset provides translation of product_category_name to English.
- Order reviews dataset: This dataset includes data about the reviews made by the customers.
- Customers dataset: This dataset has information about the customer and its location. Use it to identify unique customers in the orders dataset and to find the orders delivery location.
- Geolocation dataset: This dataset has information Brazilian zip codes and its lat/lng coordinates.

C. Outcome

Based on the data provided under the folder /input, develop the codebase, and answer the questions below

a. **What is the total amount spent by each customer* at the Stenelem store?**
   Output Column Names: *Customer_UID, Total_Amount*

b. **How many orders has each customer* made at the Stenelem store based on order status (each status must be a separate column)?**
   Output Column Names: *Customer_UID, Created, Invoiced, etc*

c. **What was each customer's* first item/product category (English name) in a successfully delivered order?**
   Output Column Names: *Customer_UID, Product_Category_Name, Product_Category_Name_English*

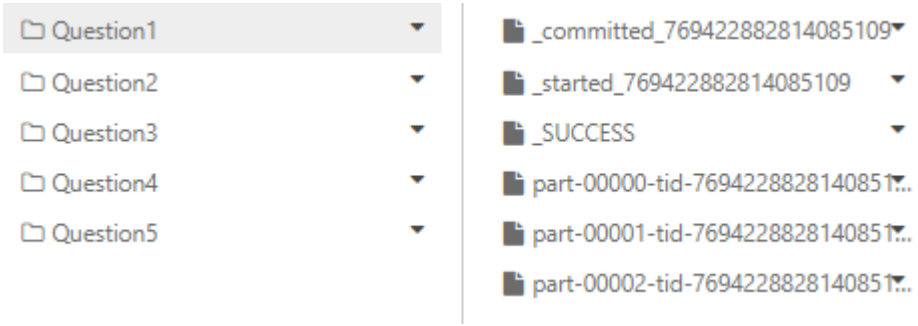
d. **What is the average time taken (days) to complete a review (after creation) for each customer*?**
   Output Column Names: *Customer_UID, Avg_Time_To_Complete, Time_Period*
   Time_Period = [Within 1 day, 2 to 5 days, More than 5 days]

e. **What are the product categories (english name), total customers, and total orders for each city in brazil? Only a single row is allowed for each city.**
   Output Column Names: *City_Name, Product_Category_Array, Total_Customers, Total_Orders*

**Unique customer*

The output of each of the questions above should be saved as parquet files under /output/Question<number>

An example of this is provided below.



- The codebase should be developed in Python and the outputs should be derived using Spark dataframes. *(Pandas isn't required).*
- Any helper/utility functions may be encapsulated as a separate file under a folder named /util
- The actual classes may be in separate file under a folder named /model
- The object instantiation may be in a separate file under a folder named /run

You have the freedom to enhance the codebase and the points above are meant to guide you in your development.

Please bundle your project files with the completed code and outputs, and send it back to us for our review.

D. Additional Points that will enhance your submission

- Data cleansing
- Adherence to object-oriented design patterns
- Logging
- Code formatting and comments
- Spark Optimization