DATA ENGINEERING ASSIGNMENT

Welcome to the Rockfeather Data engineering assignment

You have been given multiple data files containing sales product data. The data provided consists of three different data folders. Each folder contains data files in a specific data file format (CVS, JSON or AVRO). You are also been provided with an Excel file that contains a product portfolio and a manufacturer portfolio. Use the data files provided, combine them and create a usable dataset on which you can perform analyses to answer the questions below.

REQUIREMENTS

You should do the assignment as if you are working for a customer that is primarily used to working with low code tools. For this reason, you should solve the assignment to the best of your abilities in a low code tool (e.g. Azure Data Factory, Azure Logic Apps, FiveTran, Airbyte and Alteryx). If necessary you can also use programs such as Python.

The following should apply

- You might experience that some orders are duplicated because there is an overlap in the data from the different source systems (each folder is data coming from a different source system). Before performing the analysis the duplicates should be removed.
    - Note: each order has a unique Order ID
- The JSON and AVRO files only contain data from the United States, the CSV files contain data from other countries.
- Some column have nested JSON Values, for those column you can use the function OPENJSON or other native JSON functions.

QUESTIONS (AD-HOC ANALYSES)

When data changes all (ad-hoc) analyses should be able to run without having to change anything – so no hardcoded values.
Use SQL queries for the (ad-hoc) analyses.

Find the answer to the following questions

- Which zip code has the highest sales per order for the latest month in the data set?
- What is the market share of each country across all years?
- For each month, which country generated the highest revenue?
    - Show also which product that contributed mostly to the revenue for that particular country and month.

Bonus questions

- How big a share of the products that exist (products in the product portfolio) are sold each month?
- For each product (products sold more than two times) find the longest duration (measured in days) between two consecutive transactions.
    - Example: if a product is sold on January 1st 2022, February 1st 2022 and October 1st 2022 then the longest duration between two transactions is the days between February 1st 2022 and on October 1st 2022
- For each month, which country has the highest year-to-date (YTD) growth since previous year's YTD.
    - The growth calculation is: [Revenue YTD] - [Revenue YTD previous year] / [Revenue YTD previous year]