EAT SLEEP & DRINK
WHITBREAD

Python Test Overview

This python task has been designed to test candidate knowledge on creation of data pipelines, with emphasis on the techniques applied for the following components:

- Data Ingestion
- Data Pre-processing
- Data Transformation and Aggregation
- Data Export

Using the sample dataset, you will be required to process existing data, where attention should be paid to good engineering practices such as scalability, performance, automated processing and testing. As a part of this task we will be looking to assess the following:

- Problem-Solving approach
- Syntax quality and code readability/structure
- Ability to evaluate alternate approaches

If you do not know the solution, then pseudo code is acceptable, as our main goal is to understand your problem-solving abilities. You can complete this test using any language and are welcome to use any available resources such as the internet.
Please provide the source code, tests, documentation or assumptions that you have made.

Test Instructions

Transactions

The sample sales order dataset consists of three years transactions, and each year's file will be provided in a json file format.
You are required to produce the following two outputs:

1) A transformed extract to be saved in a parquet format partitioned by the existing 'OrderDate' column into daily partitions. E.g.
{base_dir}/Year=yyyy/Month=mm/Day=dd/{filename}.parquet
2) A summarised extract that is queryable to answer the following:
    - What is the total sales value of the cancelled orders?
    - What is the total sales value of the orders currently on hold for the year 2005?
    - What is the count of distinct products per product line?
    - What is the total sales variance for sales calculated at both sales price and MSRP (Manufacturer Suggested Retail Price)?
    - What has been the percentage change in sales YoY for classic cars, for years 2004 and 2005 where the status is shipped?

Please apply the following transformations to both extracts:

1) Dataset should be filtered for the following product lines; ‘Vintage Cars’, ‘Classic Cars’, ‘Motorcycles’, ‘Trucks and Buses’

2) Add calculated column by applying quantity-based price discounts using the following thresholds and recalculate sales:

Qty Threshold	Discount Applied
0-30	0%
30-60	2.5%
60-80	4.0%
80-100	6.0%
> 100	10%

(apply discount to the sales price)

3) Add calculated column by recalculating sales using MSRP.
(pricing calculation is price = sales/Qty)