Data Engineer - Assessment

HEMA, is a Dutch variety store-chain.

It began operations as a variety store.

The chain is characterized by relatively low pricing of generic household goods, which are mostly made by and for the chain itself, often with an original design.

About Assessment

This assessment is intended to evaluate:

- python programming level
- pyspark knowledge
- best practices to get spark performance
- code reproducibility

Case

A retail sales dataset was made available to Hema's new data engineer to ingest into the company's Datalake.

It is expected that some business rules will be added to the final dataset, in addition to some transformations.

This dataset may contain new attributes in the short term, requiring your pipeline to be flexible to support this schema evolution and display it transparently to users

Dataset

Retail dataset of a global superstore for 4 years

Here, you can find the dataset available on kaggle

Column	Description
Row ID	Unique identifier of row
Order ID	Order number of sales
Order Date	Day , Month and Year of Order Date
Ship Date	Day , Month and Year of Ship Date
Ship mode	Classification of Ship Mode
Customer ID	Customer Identification Number
Customer Name	
Segment	Customer Classification
Country	Name of Country
City	Name of City

Requirements

Technical Requirement

- Column name should be renamed in caseCamel format
- Data should be stored in raw, curated and consumption formats (Data Lake structure)
- Data Ingestion Pipeline should be written in Python, using Pyspark framework
- Metadata like filename, ingestionDate, loadingTime should be part of the dataset in each data lake layer
- Consistent log should be added to the solution in order to be interpreted and debugged when necessary
- Data should be partitioned by order data based on (Year, Month and Day)
- Code should be pushed to some git repository (Git, Bitbucket, Gitlab) and shared the code after the end of assessment
- Data should be stored in curated and consumption layer in parquet file format

Functional Requirement

- Dataset should be splitted in 2 different one in consumption layer (Sales, Customers)
- Sales dataset might contains the attributes below
  - orderId
  - orderDate (YYYY/MM/DD format)
  - shipDate (YYYY/MM/DD format)
  - shipMode
  - city
- Customer dataset might contains the attributes below, considering that quantity of orders should be calculated filed based on raw data:
  - customerId
  - customerName
  - customerFirstName
  - customeLastName
  - customerSegment
  - country
  - city
  - quantityOfOrders(last5Days)
  - quantityOfOrders(last15Days)
  - quantityOfOrders(last30Days)
  - totalQuantityOfOrders
- Customer dataset should be rewritten every run of the pipeline

Advices

- Consider to use virtual environment (venv, conda) and then install pyspark inside of you env.
- As nice to have, consider user Docker as part of your solution (Will be considered as plus)
- Git repository and code structure is up to you, just remember to share your repository after the end of assessment share in the email below:
    - email: marcus.azevedo@hema.nl
    - Subject: Your Name