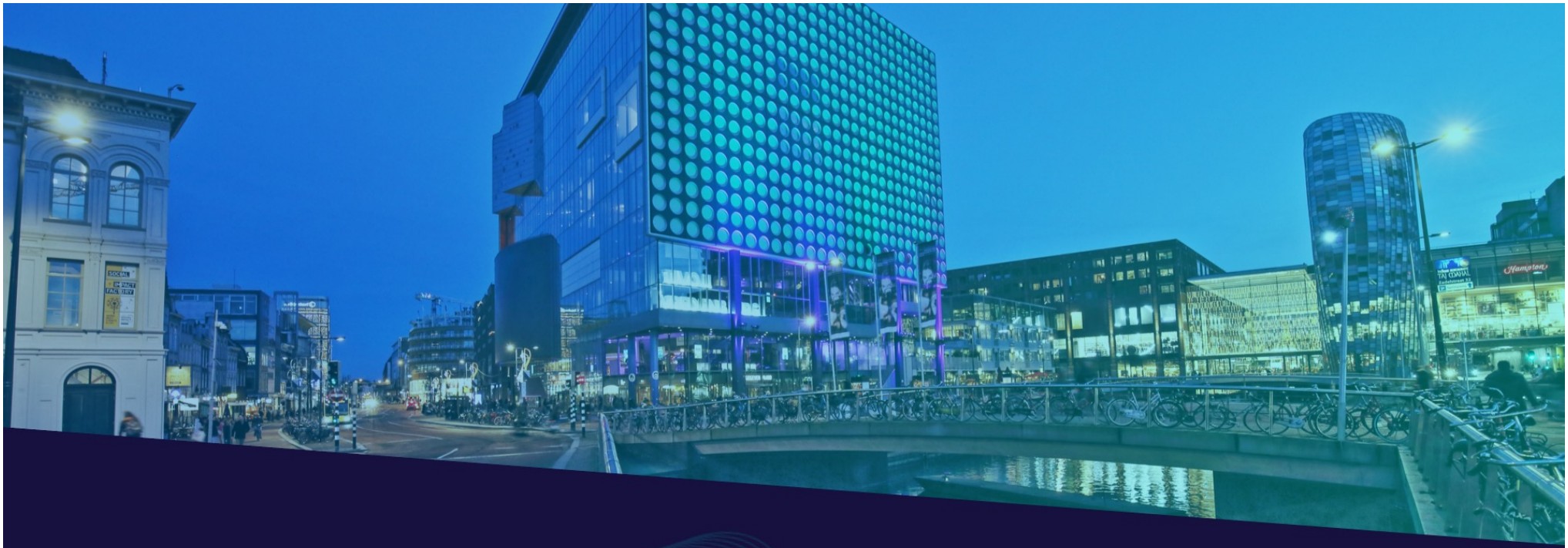


Assignment Data Engineer

Brainbay, July 2022

brainbay

www.brainbay.nl

Introduction

Brainbay was founded in 2018 by the NVW – the Nederlandse Vereniging van Makelaars (Dutch association of real estate brokers and valuers).

As a young company, brainbay is the platform for insights and data-driven services for the Dutch real estate market.

We provide the real estate market with the best data and information products. The data exchange system of NVM members and the database are housed at brainbay. This gives brainbay an extensive database with information about homes, commercial real estate and agricultural companies.

We are a pure data driven organisation. A great place to work. You contribution as a Data Engineer is at the heart of our company!

Brainbay is the leading platform for datadriven services in real estate, and an indispensible partner for NVM and her members

Brainbay mission

brainbay

www.brainbay.nl

Assignment

- This assignment is intended to evaluate:
  - Python programming level
  - Use of Python libraries (like pandas) and modules (venv, conda)
  - Cleverness and performance

- The assignment is used as a starting point for the Technical interview.

- The assignment is a simulation of a different sector than real estate but aims to be a first impression of the role and responsibilities.

brainbay

Case

You are asked to ingest a dataset into the database of a retail company by the implementation of a pipeline you create from scratch.

This dataset with sales information will be available every month.

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

This dataset may or may not contain new attributes in future months

- Your pipeline to be flexible to support this data schema evolution
- New attributes needs to be available transparently to users

Data set

- The dataset contains 4 year of sales data of a global retail store

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

- The dataset (in CSV-format) can be found here: https://github.com/brainbaynl/DE_Assignment

Technical requirements

- Column names 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
- 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)
- Data should be stored in curated and consumption layer in parquet file format
- The dataset should be rewritten every run of the pipeline

Functional requirements

- Dataset should be split in two different consumption layers and may (or may not) contain the following attributes per layer

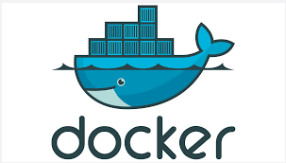
1. Sales
    - orderId
    - orderDate (YYYY/MM/DD format)
    - shipDate (YYYY/MM/DD format)
    - shipMode
    - city

2. Customer
    - customerId
    - customerName
    - customerFirstName
    - customeLastName
    - Custo+merSegment
    - country
    - city
    - quantityOfOrders(last5Days)
    - quantityOfOrders(last15Days)
    - quantityOfOrders(last30Days)
    - totalQuantityOfOrders

note: quantity of orders should be calculated based on raw data

Advices

- Consider to use virtual environment (venv, conda) and then install python packages inside of your environment.
- It is advised to use the pandas library.
- Consider user Docker as part of your solution.
- You can fork your solution from our Git repository. How you structure your code is up to you, we do expect however that you can explain your decisions.



brainbay

Finished?

- Push your code to Github and share the link with us!

brainbay