Data Engineering Assignment

(kudos to the Atmos Squad for this material)

CASE

As a marketing manager working for Nike, I want to have an overview of our most important customers so that the company can send them each a very special present for Christmas.

ASSIGNMENT

You are going to build a small data pipeline that outputs an overview of Nike's VIP customers, ordered by the total sales value of their purchases. For this promotion, the marketing department is only interested in VIPs currently located in The Netherlands. VIPs that have not purchased any products are still eligible – they are VIP, after all. This ask is not a one-off, as Marketing is already thinking about running this promotion again next year if it is successful.

In the overview, Marketing is looking for:

- The name of the VIP
- The email address of the VIP
- The total sales value of the VIPs purchases

There are 3 data systems in scope of this assignment:

- From the Sales domain team working on our next-gen Data Lake, you receive a file with the transactions done in the last month. This data is provided to you in the Parquet format.

- From the Customer domain team working on a legacy CRM, you receive 3 snapshots of VIP dimensional master data as CSV files. The filenames contain the export date from the CRM, and each file represents the truth at that point in time. A new snapshot is generated each time VIP data changes in the system. The overview you generate for Marketing should respect this temporal truth.

- Unfortunately, the Sales and Customer domain teams operate separately. Thus, from the Marketing team itself, you receive a file containing the latest user-managed mapping between profiles (from the Sales domain) and VIPs (from the Customer domain). The Marketing team keeps this mapping up to date in Excel but was able to export a CSV file.

FILES RECEIVED

transactions.parquet
vips_2020-11-01.csv, vips_2020-11-15.csv & vips_2020-11-25.csv
umd_vip_to_profile_mapping.csv

WHAT WE EXPECT

- We see this assignment as simulation of working environment where we discuss requirements, design and implementation details with diverse group of stakeholders containing both technology and business experts. The goal of is not to grade the solution on formal correctness, but to use it as a basis for discussion to assess candidate's proficiency level
- You can build the data pipeline in any language you want, but we prefer if it's done using PySpark
- Take notes of all assumptions you make, that will help to cover them all in the discussion
- Prove that your output is correct with unit tests
- Explore the data and assess it, end users from the business should only see clean data. If you don't have time to create logic to deal with all data issues, note them and mention in the discussion
- Put the code you created in the private repository in your private Github account and share with the Github users of the interviewers (user IDs will be provided). DO NOT make the assignment code available to wider audience by settings repository visibility to "public" or in any other way.
- You're expected to drive the discussion. While preparing, consider the following structure
    - Setting the scene (summary of what the assignment is about)
    - Solution design (which technologies you used and how you structured your application)
    - Assessment of the source data (issues, proposed cleansing actions)
    - Assumptions made and steps you would take to tackle them

FOLLOW-UP QUESTIONS

(you do not have to consider these for your implementation)

- The Brand and Merchandise teams like the success the Marketing team had because of using your data pipeline. How would you abstract this system so that Nike can implement it generically across departments?

- The Sales and Customer teams currently provide you files for batching, but Marketing is investigating sending direct 'thank you' emails every time VIP makes a purchase instead of physical presents to reduce operating costs. How would you redesign this system to work for streaming data?