Data Engineering
Tech Assessment

Thank you for taking the time to accept this challenge. We hope you enjoy the challenge we have prepared and that it piques your curiosity for the work we do at Loka. We wish to give you the best possible preview of the kind of work we do and the skills needed.

If possible, document/explain your code and thought process so we can understand it better and prepare relevant questions to make the discussion more interesting.

Challenge

Introduction

door2door collects the live position of all vehicles in its fleet in real-time via a GPS sensor in each vehicle. These vehicles run in operating periods that are managed by door2door's operators. An API is responsible for collecting information from the vehicles and place it on an S3 bucket, in raw format, to be consumed.

The goal of this challenge is to automate the build of a simple yet scalable data lake and data warehouse that will enable our BI team to answer questions like:

What is the average distance traveled by our vehicles during an operating period?

We would like to ask you to develop a solution that:

1. Fetches the data from the bucket on a daily basis and stores it on a data lake;
2. Processes and extracts the main events that occurred during operating periods;
3. Store the transformed data on a data warehouse. The data warehouse should be SQL-queriable (SQL database or using something like AWS Athena).

Data

The data for this challenge lives on the S3 bucket `s3://de-tech-assessment-2022`. Inside you can find:

1. a folder named `data` that contains all the data;
2. a file named `DE_Tech_Assessment_Metadata.pdf` that describes the data.

In case you want to use a different cloud provider or develop a local solution you can download the data from here.

Technical assumptions

- The fetching process should only get data from a certain day on each run and should run every day;
- Files on the "raw" S3 bucket can disappear but we might want to process them differently in the future;
- No need to answer the question stated in the introduction;
- If your solution is setup to run locally, it must be containerized;
- There is no need for paid, expensive and highly performant data warehouses. You can use a "standard" SQL database.

Bonus points

- Sketch how you would set up the application on the cloud (AWS, GCP, etc);
- It is encouraged to simplify the data by a data model on the data warehouse layer.

Delivery of your solution

The output of this exercise is expected to be a **private** GitHub repository and so, please add the following users to the repo:

- `henriqueribeiro`
- `peter-ram`
- `taxuspt`

Reviewing

There are no right or wrong answers to this challenge, the results you achieve are of secondary importance. We will mostly focus on: your work ethics, your skills, how you justify your design choices and your thought process. With this in mind, it's not mandatory that you provide a full fledged solution.

Below we leave some bullet points that we are looking for in the solution, so it can guide you through this challenge and help you to prepare for our discussion on your solution (we might not cover all but it should help):

- Data Engineering: Data formats, Ingestion techniques, Data Modelling, etc...
- Documentation: Is the project and the code properly documented?
- Technology: Which libraries or approaches are used? Do they make sense for the task? Justify why you've decided to use those technologies to solve the code challenge
- Code quality: Is the code understandable and maintainable?