Data Engineer test - Aptitude

Purpose

This test is designed to showcase your understanding of databases and data processing with Spark, together with your aptitude in Python.

There are two stages to this code test:

1. Preparing code at home ahead of your interview.
2. Pairing with us at the interview, making additions to your code.

The pairing phase is to give us an indication of what it will be like to work together, and should be regarded as a collaborative effort.

Problem

Using the information in TLC Trip Record Data you should be able to get all the parquet files related to Yellow Taxi top 10% trips based on distance travelled.

You should be able to:

- Build a python application that is able to get the parquet files dynamically based on starting and ending dates.
- Process the downloaded parquet files with PySpark.
- Find any data quality issues in **total_amount** field, find a suitable way to work with those issues.

Notes on completing these tasks

- There is no right way to do this. We are interested in the choices that you make, how you justify them, and your development process.
- You should not use any third party service for solving this problem. We are evaluating your capacity to solve the problem, not to use third party tools.
- Consider what kind of error handling and testing is appropriate.

What should I send?

A repo with all the code you used and a readme explaining your approach, the steps to reproduce your test and everything you feel is important.