oetker digital

Data Engineer Challenge

Dear future colleague,

This challenge was created to evaluate your skills so please be as explicit as possible and note that some points are vaguely described so you can propose possible strategies/improvements.

Bonus points are not hard requirements to succeed but will give us a better understanding of your skills. Again, don't worry if you don't have enough time to complete them.

Requirements:

1) Download the sample data from: https://storage.googleapis.com/datascience-public/data-eng-challenge/MOCK_DATA.json

Given the data sample we ask you to build a simple batch processing application. You will be required to install apache-airflow locally.
Check: https://airflow.apache.org/docs/apache-airflow/stable/start/local.html for references.

2) Write a simple Python app that:
    a) Retrieves the data from the given url and stores it in a local path.
    b) Write a simple transform in PySpark that reads the file and does the following:
        i) Cleans the data. For example (you may come up with different ones): country fields always start with capital letter, IP address format is correct, date field has always the same format
    c) Computes some simple measures of the data (you may come up with different ones): most and least represented country, number of unique users in the current run.

3) Bonus points if you:
    a) Use Docker for your local development
    b) Use correct Python best practices. docstrings, PEP-8, tests, etc.

Present your solution by uploading the code to GitHub and the instructions on how to install the tools you used.

We are curious to see and discuss your solution with you so good luck and happy coding!