Take-home challenge: py-spark

Task:

*In a nutshell, Analyze the Crude Oil dataset.*

Requirements

*• Create a dockerized stand-alone single node spark cluster and run spark jobs to answer the following questions on a Crude oil dataset.*

• What are the top 5 destinations for oil produced in Albania?

• For UK, which destinations have a total quantity greater than 100,000?

• What was the most exported grade for each year and origin?

Data

Please use the crude oil dataset from kaggle

https://www.kaggle.com/datasets/alistairking/u-s-crude-oil-imports/data

Delivery

Please post your project on GitHub.

Delivery

**2-3 hours (we do not want you to spend more than 3 hours on this exercise)**

**Bonus question: Get the data from Question 1 and convert it to Apache iceberg format.**

**For Bonus question, we are not expecting the final solution. Any work in progress solution can also be added to GitHub.**