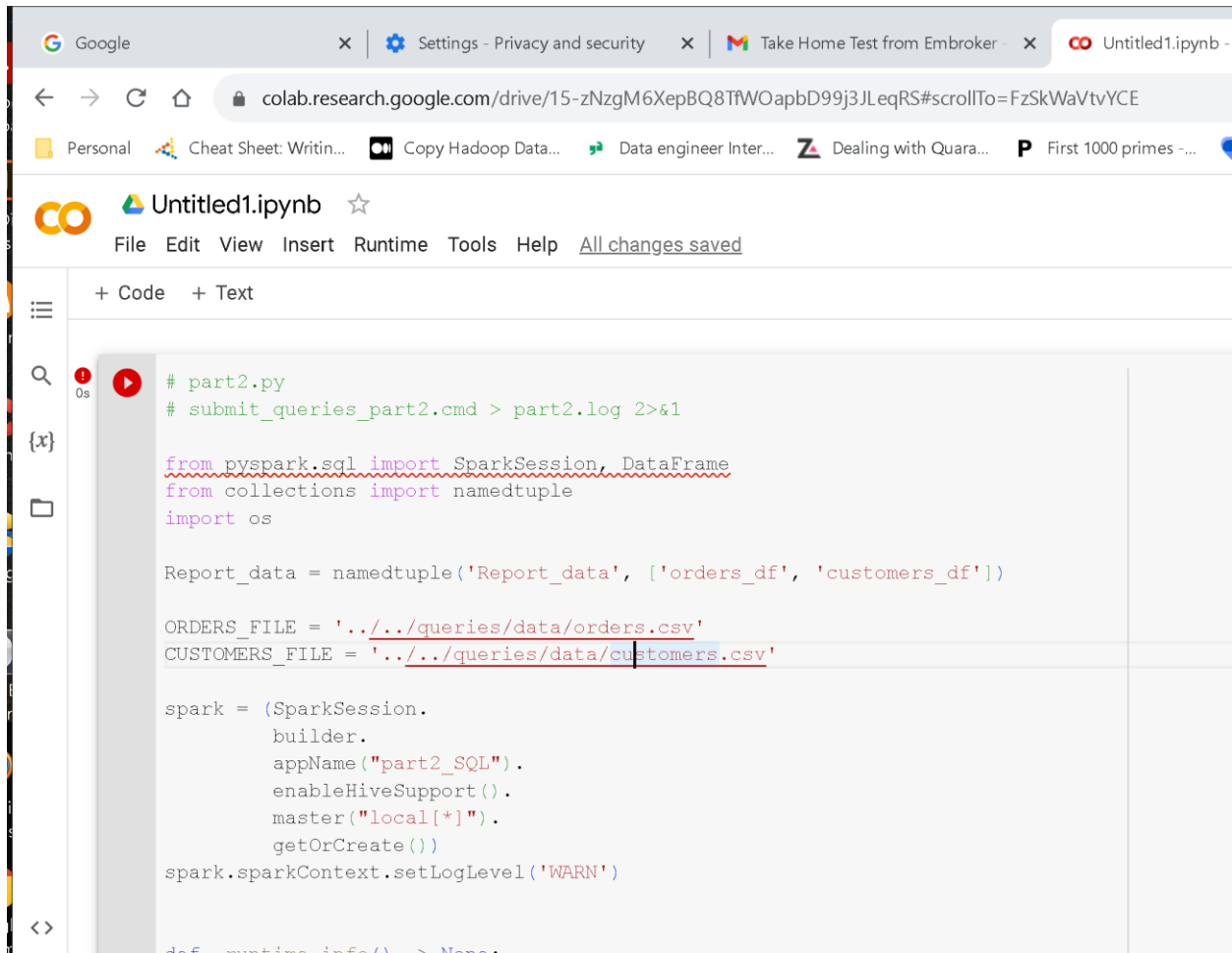


Findings for Embroker (8/15/23)

Thanks for the assignment, I learned a lot. I was unable to access a working development environment. You kindly gave me a URL on 8/14 where Spark was thought to be preinstalled. However, the Spark python libraries seemed to be inaccessible. See the image:



```
# part2.py
# submit_queries_part2.cmd > part2.log 2>&1

from pyspark.sql import SparkSession, DataFrame
from collections import namedtuple
import os

Report_data = namedtuple('Report_data', ['orders_df', 'customers_df'])

ORDERS_FILE = '../queries/data/orders.csv'
CUSTOMERS_FILE = '../queries/data/customers.csv'

spark = (SparkSession.
    builder.
    appName("part2_SQL").
    enableHiveSupport().
    master("local[*]").
    getOrCreate())
spark.sparkContext.setLogLevel('WARN')

def runtime_info() -> None:
```

The difficulties of using a new environment would take even more time. I did not want to wait any longer to complete the assignment, so I loaded Spark on my laptop and did most of the exercises using open-source Apache Spark. I used standard Python code files and the `spark-submit` command to run the exercises. Here is a summary of the files I uploaded:

Code:

1. `part1.py` – Code for programming assignment (DE_Part_1) [partially complete]
2. `part2.py` – Code for SQL assignment () [complete]
3. `TestPart1.py` – Code for Test module for individual methods in `part1.py` (TDD)

Logs:

1. Findings.pdf – this document
2. part2.log – The SQL outputs (fully succeeded)(
3. test1.log – Test module output for part1.py (partially successful)
4. test1_Failed_JSON_read_.log – the failed read_json method (details below)

You can review the code in the Python files and the Spark output in the log file.

Programming Assignment (part1.py is DE_Part_1) Details

I did not have access to a database in my local environment, so I was going to write to a Parquet file locally instead. However, the schema for the JSON file failed to work in the read JSON method. As a result, I was unable to read and write the data. I created the TestPart1.py file, in the spirit of TDD, so I could get some of the methods in part1.py working.

High Level Feedback

- A properly working development environment is problematic.
- The assignment is very “literal” in that the files and tasks seem to be “real work”.
- The assignment seems to indicate you require only “data wrangling” and no analytical work.

Personal Notes

- I learned a lot about JSON and StructType; I primarily used HDFS files.
- Notebooks have many deficiencies (see below.)
- I would try offering the assignment on (say) the Databricks community edition (full Spark ready to go!)
- Provide an account to access the assignment.
- I primarily work with “Big Data”, and with Hadoop or AWS S3 files, using Parquet or ORC formats.

Tech Stack Comments

A good IDE, like PyCharm, offers static error checking, PEP-8 formatting, auto completion, code base navigation, and many more features not in Notebooks . . . and not in Google. Python modules are the unit of code management in Python, not notebooks. Python modules can be packaged and distributed for reuse. Regular Python files are easy to manage in GIT where DIFFs work well.

Spark shines in the “Big Data” arena with data science workloads. Those workloads are mostly “full table scans”. Databases are good at transactions but less well adapted by the scale of Big Data. The mix of Spark and the Delta database seems curious.