Задание

С помощью Docker-compose развернуть среду из 4 контейнеров. Контейнер 1. PostgreSQL

Контейнер 2. Spark – master Контейнер 3. Spark – worker1 Контейнер 4. Spark – worker2

Источник данных – файл house\_prices.csv

Схема обработки данных и взаимодействия контейнеров:

1. «Контейнер 1. PostgreSQL» запускается и создает (подключается к существующей) БД.
2. В БД подгружаются данные из файла house\_prices.csv в соответствующую таблицу.
3. «Контейнер 2. Spark – master» имеет возможность подключения к БД.
4. Контейнеры 2, 3 и 4 настроены как кластер Spark.
5. Приложение PySpark в виде файла my\_spark.py загружено в «Контейнер 2. Spark – master». Его задача:

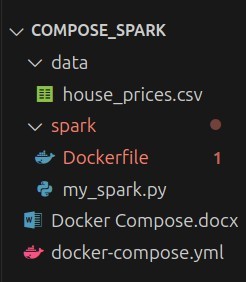
*Из данных* house\_prices *вывести информацию: Средние стоимости квартир и домов в зависимости от района и количества комнат.*

ОТЧЕТНОСТЬ:

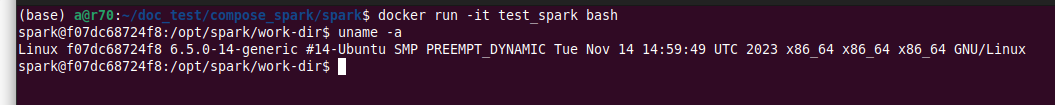
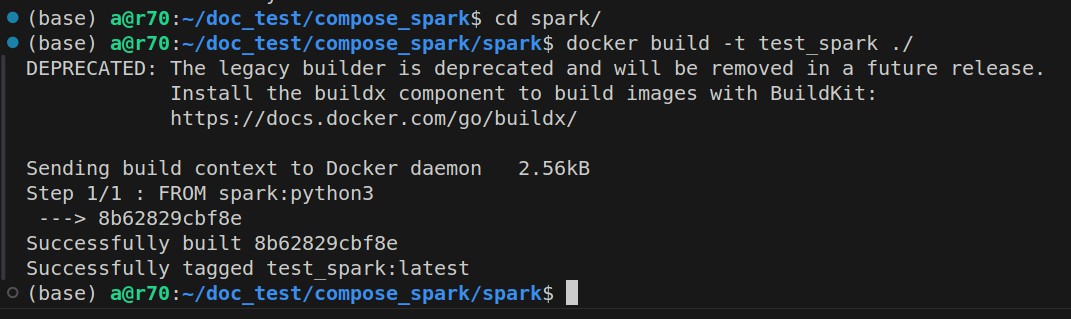
На почту преподавателю необходимо отправить ссылку на Ваш проект в GitHub и скрин экрана с результатом работы приложения my\_spark.py.

========================================

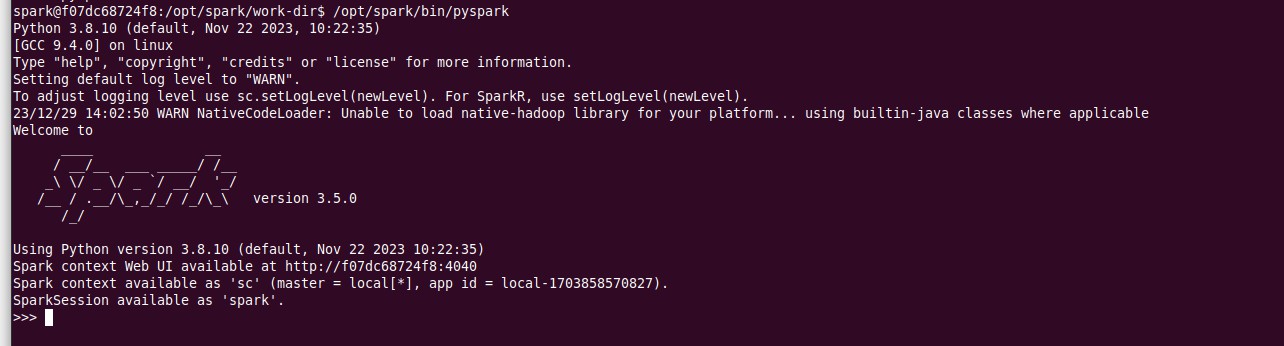
## Поднимаем Spark master контейнер

Делаем такую структуру папок.

Создадим контейнер spark

 docker run -it test\_spark bash

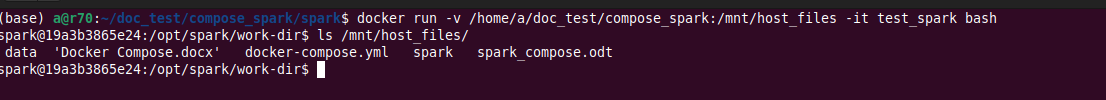
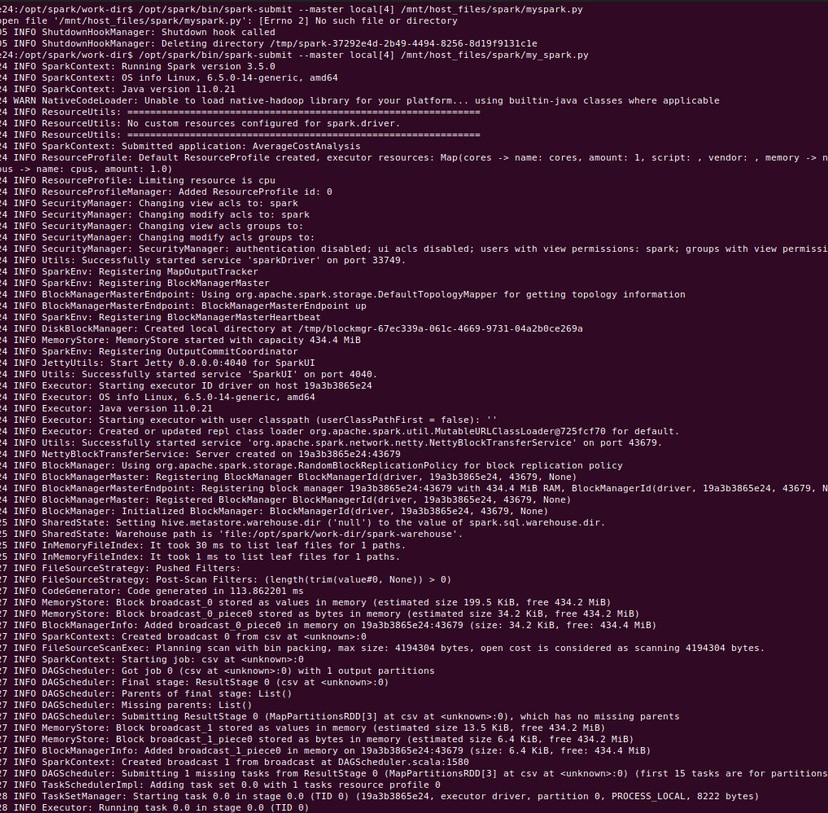
pyspark "/opt/spark/bin/pyspark"

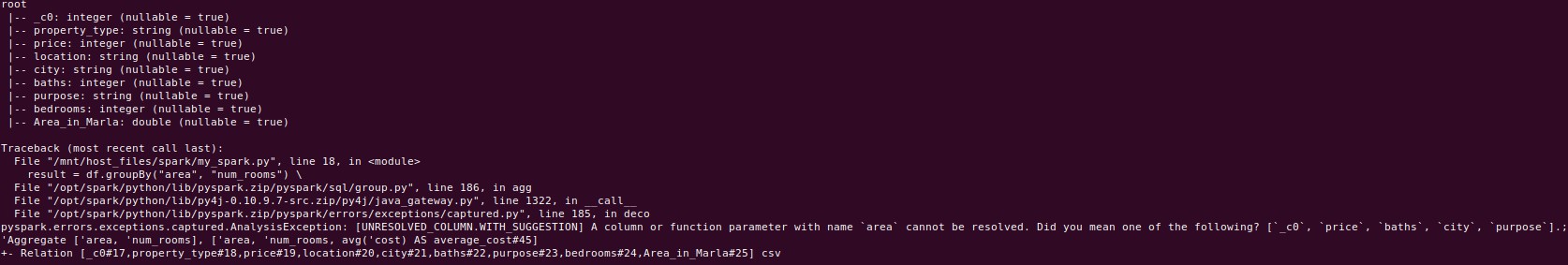


# Пишем и тестируем my\_spark.py с загрузкой из файла

Перезапускаем докер, так чтобы можно было читать файл .csv

docker run -v /home/a/doc\_test/compose\_spark:/mnt/host\_files -it test\_spark bash





Traceback (most recent call last):

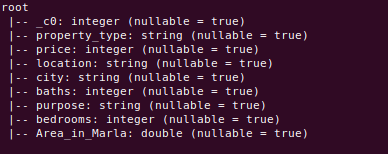
File "/mnt/host\_files/spark/my\_spark.py", line 18, in <module> result = df.groupBy("area", "num\_rooms") \ **В этой строчке ошибка**

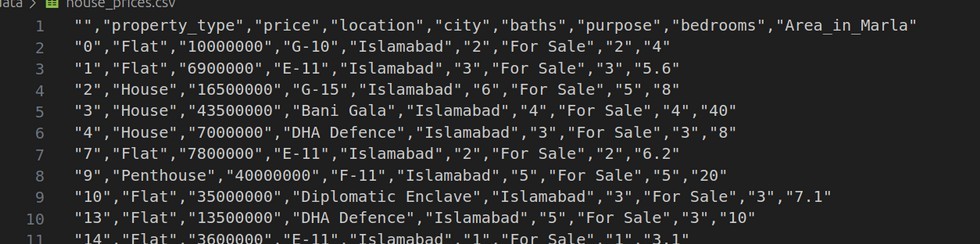
File "/opt/spark/python/lib/pyspark.zip/pyspark/sql/group.py", line 186, in agg

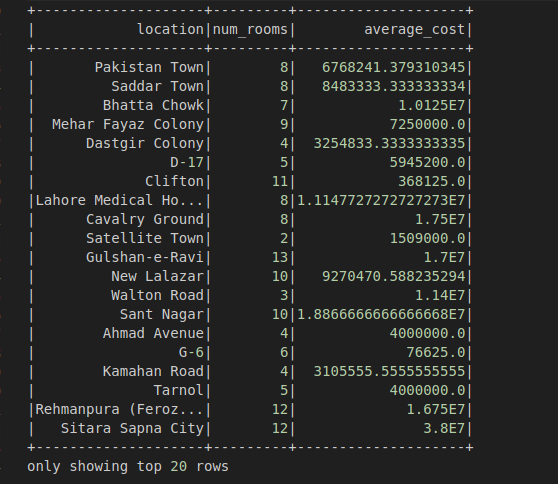
File "/opt/spark/python/lib/py4j-0.10.9.7-src.zip/py4j/java\_gateway.py", line 1322, in call File "/opt/spark/python/lib/pyspark.zip/pyspark/errors/exceptions/captured.py", line 185, in deco

pyspark.errors.exceptions.captured.AnalysisException: [UNRESOLVED\_COLUMN.WITH\_SUGGESTION] A column or function parameter with name

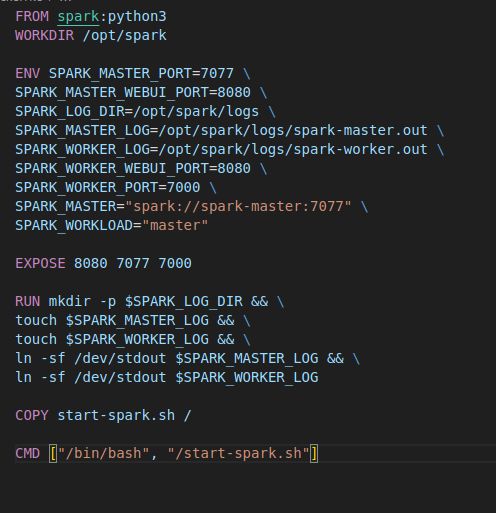
`area` cannot be resolved. Did you mean one of the following? [`\_c0`, `price`, `baths`, `city`, `purpose`].; 'Aggregate ['area, 'num\_rooms], ['area, 'num\_rooms, avg('cost) AS average\_cost#45]

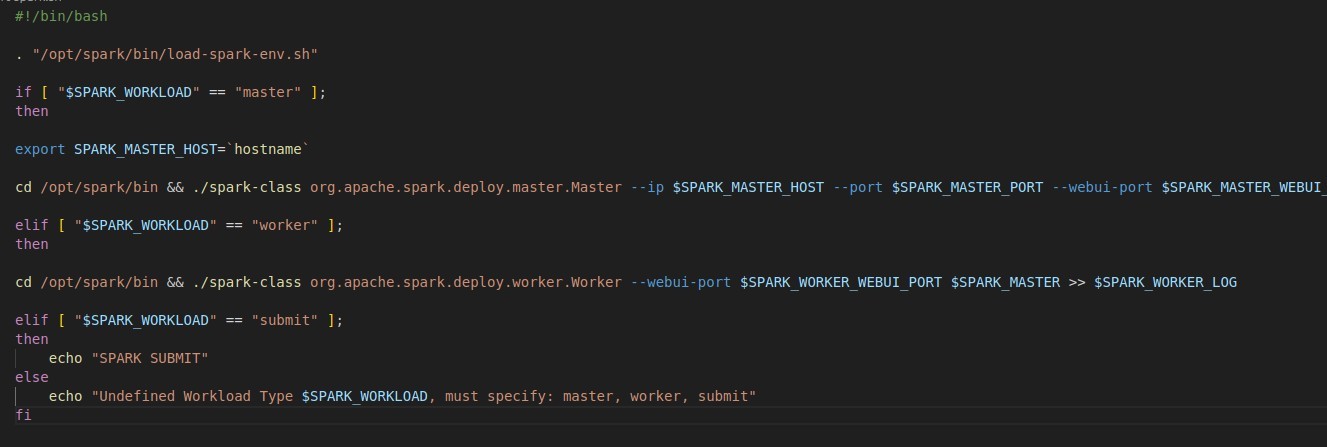
.

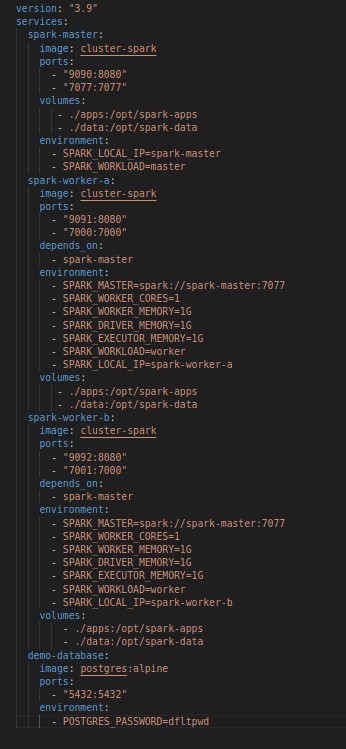


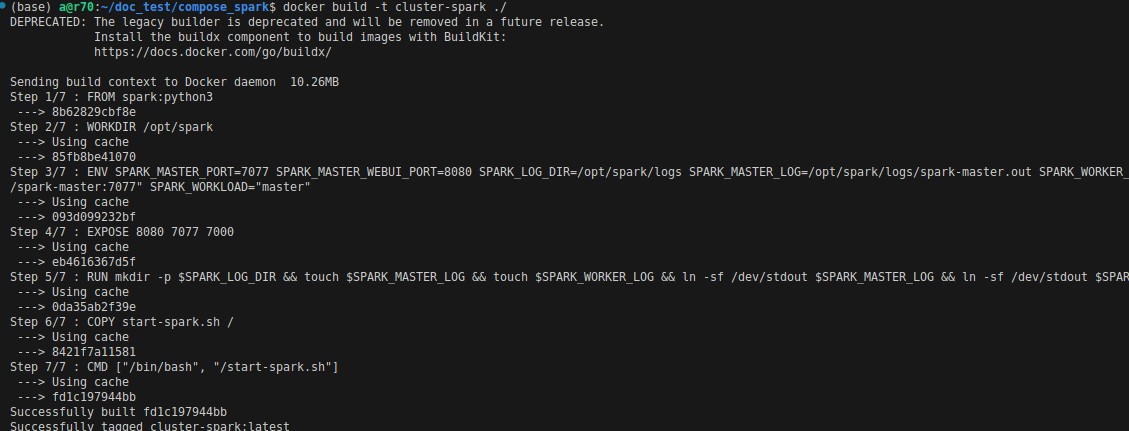


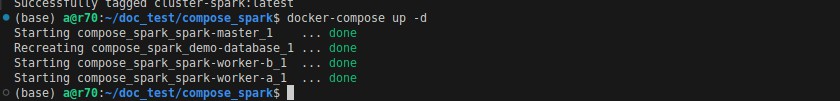
## Делаем чтобы спарк работал в кластере master+2workers

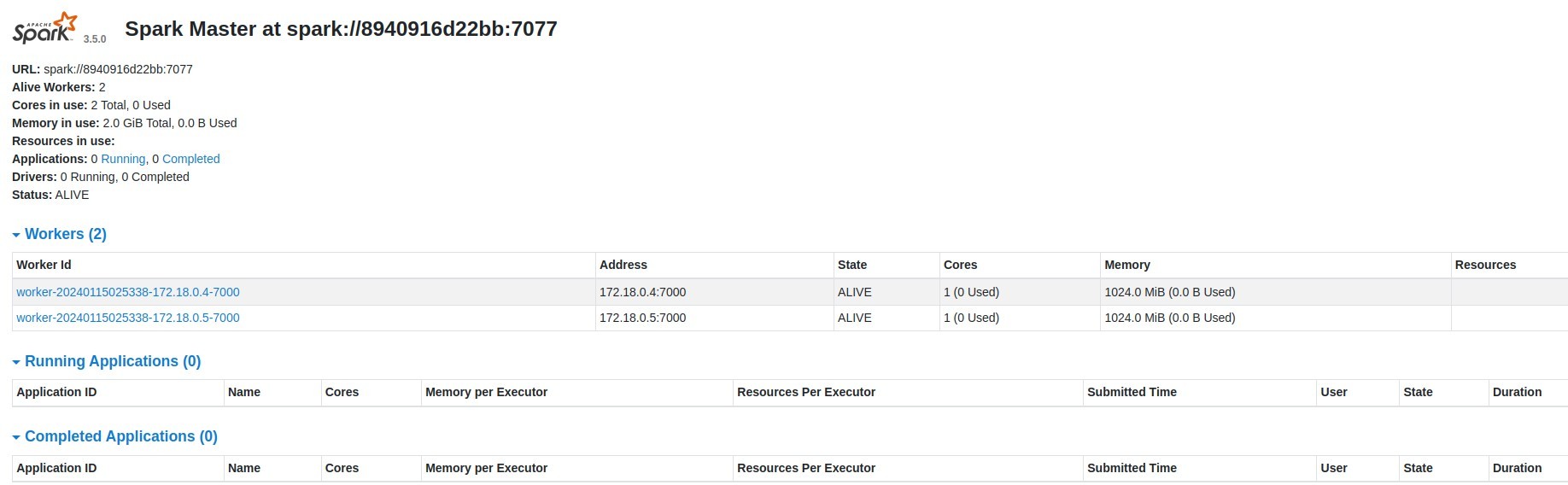






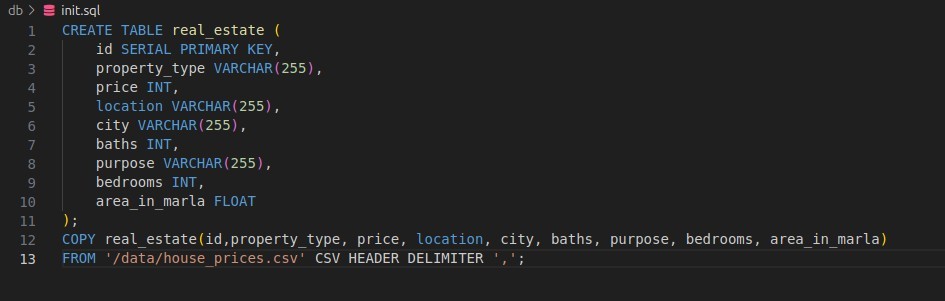


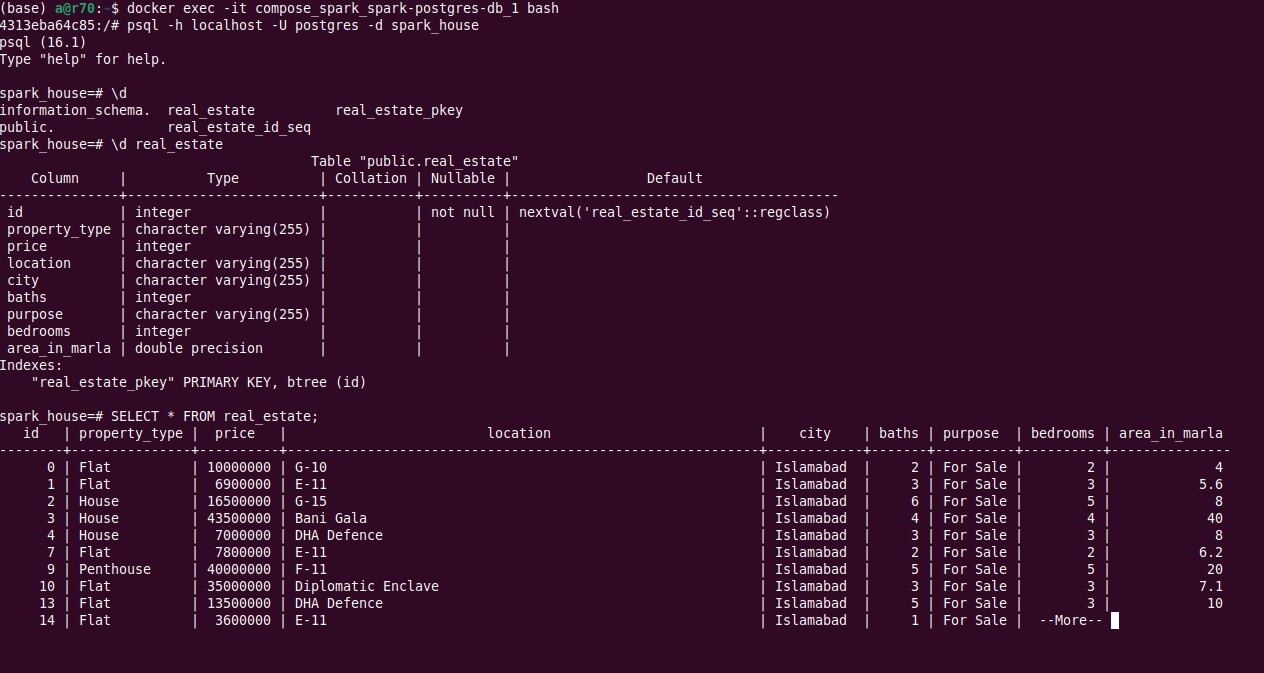
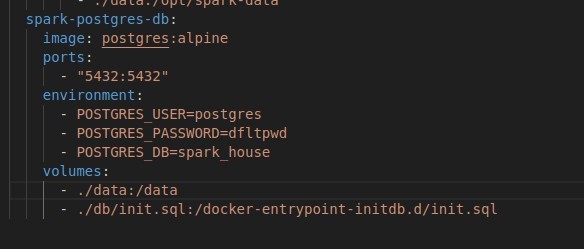






# Заменяем подгрузку с файла на подгрузку из бд PostgreSQL и переписываем основу

init.sql



docker-compose up -d

/opt/spark/bin/spark-submit --master spark://spark-master:7077 --jars /opt/spark-apps/postgresql-

42.7.1.jar --driver-memory 1G --executor-memory 1G /opt/spark-apps/my\_spark.py

