**SparkSQL trong một Docker container**

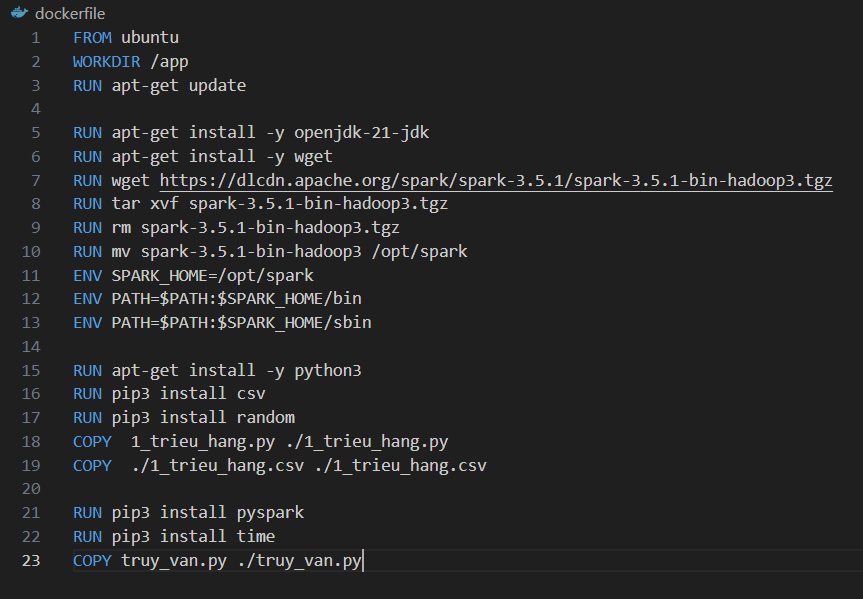
**Mục tiêu**

Mục tiêu củaproject là thiết lập và chạy SparkSQL trong Docker container, đồng thời thực hiện các truy vấn SQL trên database được lưu trong container này.

**Cài đặt**

Để tạo docker, ta sử dụng Docker Desktop với giao diện đơn giản và dễ dung. Ta cài đặt phần mềm này từ [*https://www.docker.com/*](https://www.docker.com/)tuỳ theo hệ điều hành mà máy sử dụng.

Tiếp theo, ta tạo một Dockerfile với nội dung như sau:



Ở đây, dự án sử dụng một Docker image từ Docker hub là ubuntu, một image cung cấp cho ta dữ liệu và giao diện để xây dựng một Docker image theo ý ta muốn. Tiếp theo, để có môi trường làm việc, ta cần phải cài jdk từ openjdk-21 để thiết lập môi trường. Sau đó thì ta mới có thể cài Apache Spark được tải từ link: https://dlcdn.apache.org/spark/spark-3.5.1/spark-3.5.1-bin-hadoop3.tgz và thiết lập môi trường trên Apache Spark để chạy. Tiếp đó, ta cài python3 làm công cụ để xây dựng project và cài đặt các thư viện là csv, random, pyspark, time để tạo database và truy vấn dữ liệu. Đặc biệt, pýpark là thư viện cần thiết để chạy SparkSQL

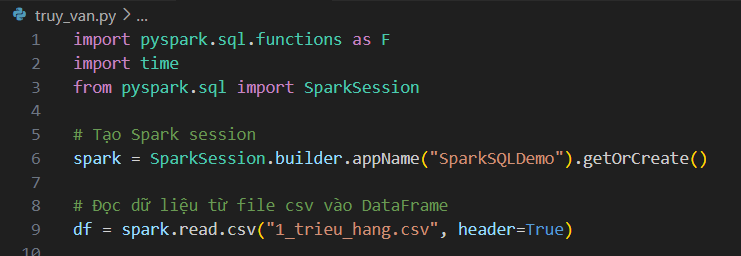
**Thực thi**

Bước 1: Tạo database



Ta tạo một database có tên là 1\_trieu\_hang. Database này gồm 3 cột có tên lần lượt là cot1, cot2 và cot3, gồm 1 triệu hàng với các số ngẫu nhiên.

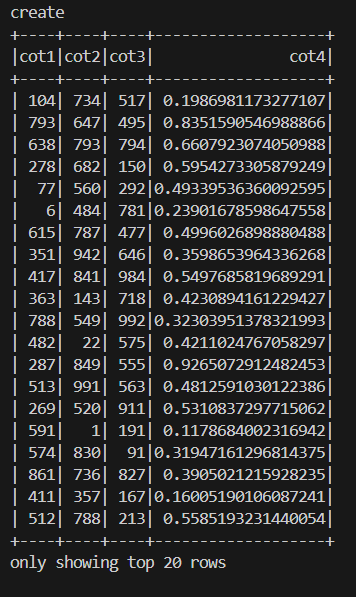
Bước 2: kết nối với database



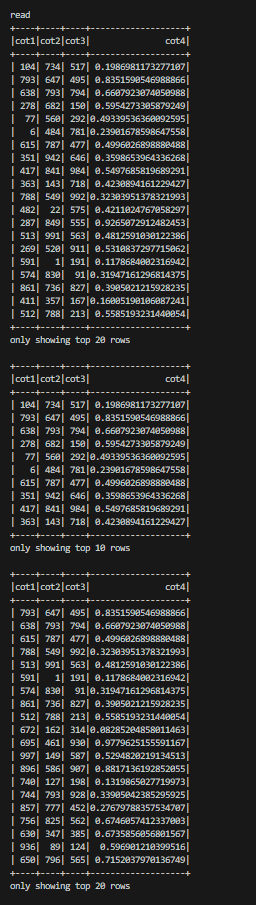
Bước 3: Thực hiện các thao táo CRUD

Với thao tác Create:

Thêm một cột mới `cot4` với giá trị ngẫu nhiên cho mỗi hàng



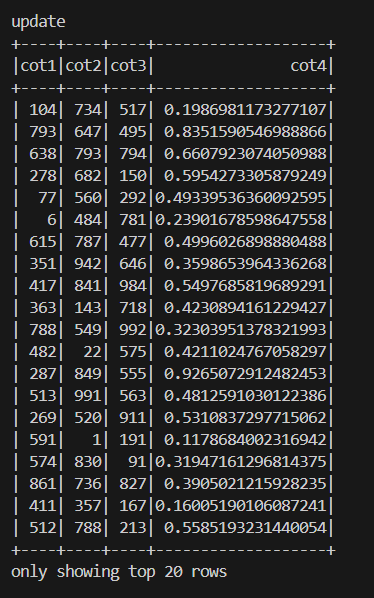
Với thao tác Read:



Với thao tác Update:

Lọc các hàng có cot1 lớn hơn 700

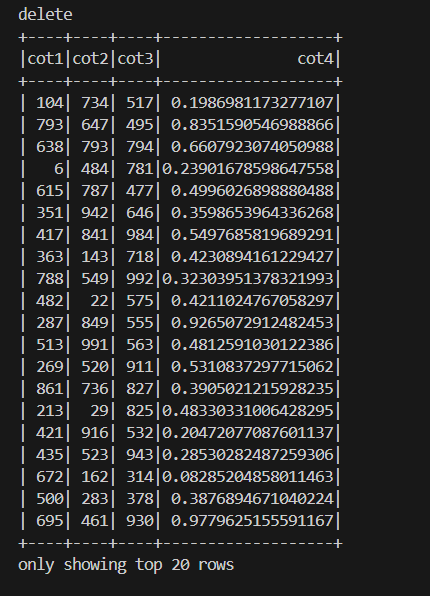
Cập nhật cột cot2 với giá trị cố định (1000) cho các hàng được lọc



Với thao tác Delete:

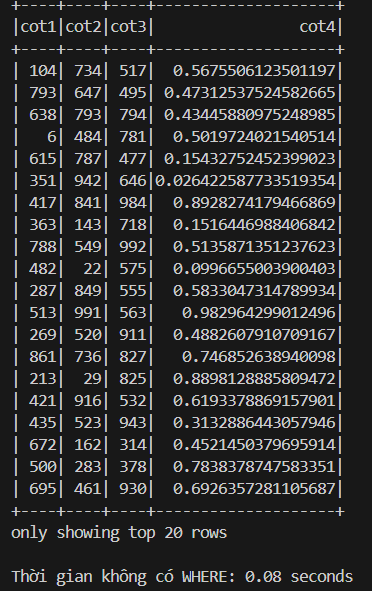
Lọc các hàng có cot3 nhỏ hơn 300

Giữ các hàng có cot3 >= 300, bỏ các hàng đã lọc (tức là các hàng có cot3 < 300)

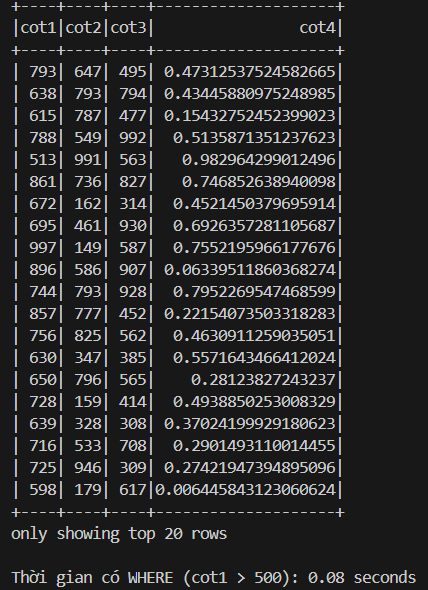


Bước 4: So sánh thời gian chạy các câu lệnh truy vấn:

Khi không có WHERE:



Khi có WHERE:



**Kết luận**

Như vậy, documentation này đã hướng dẫn đầy đủ chi tiết về việc cài đặt và sử dụng SparkSQL trong một Docker container, thực hiện được những thao tác CRUD cơ bản và so sánh thời gian thực hiện các truy vấn.