**Họ và tên :** Ngô Hoài Nam

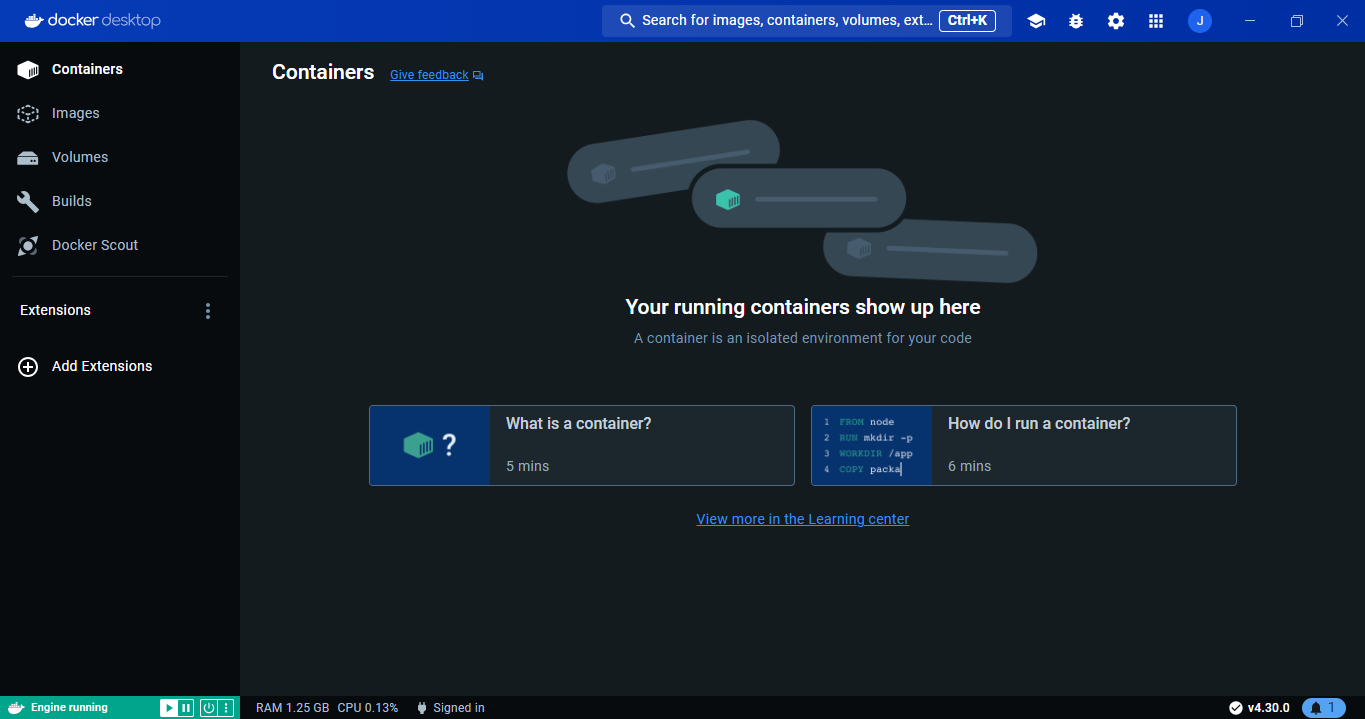
**Mã sinh viên :** 21002219

**BÁO CÁO BÀI TẬP LỚN**

**THỰC HÀNH TIN HỌC ỨNG DỤNG**

1. **Xây dựng và chạy Apache Spark trên Docker Container.**
   1. ***Tải và cài đặt Docker Desktop***

Truy cập đường dẫn [*https://www.docker.com/products/docker-desktop/*](https://www.docker.com/products/docker-desktop/)để tài file cài đặt của Docker Desktop về máy, và cài đặt Docker Desktop. Sau khi cài đặt xong thì ứng dụng sẽ có giao diện như sau :



* 1. ***Cài đặt Apache Spark***

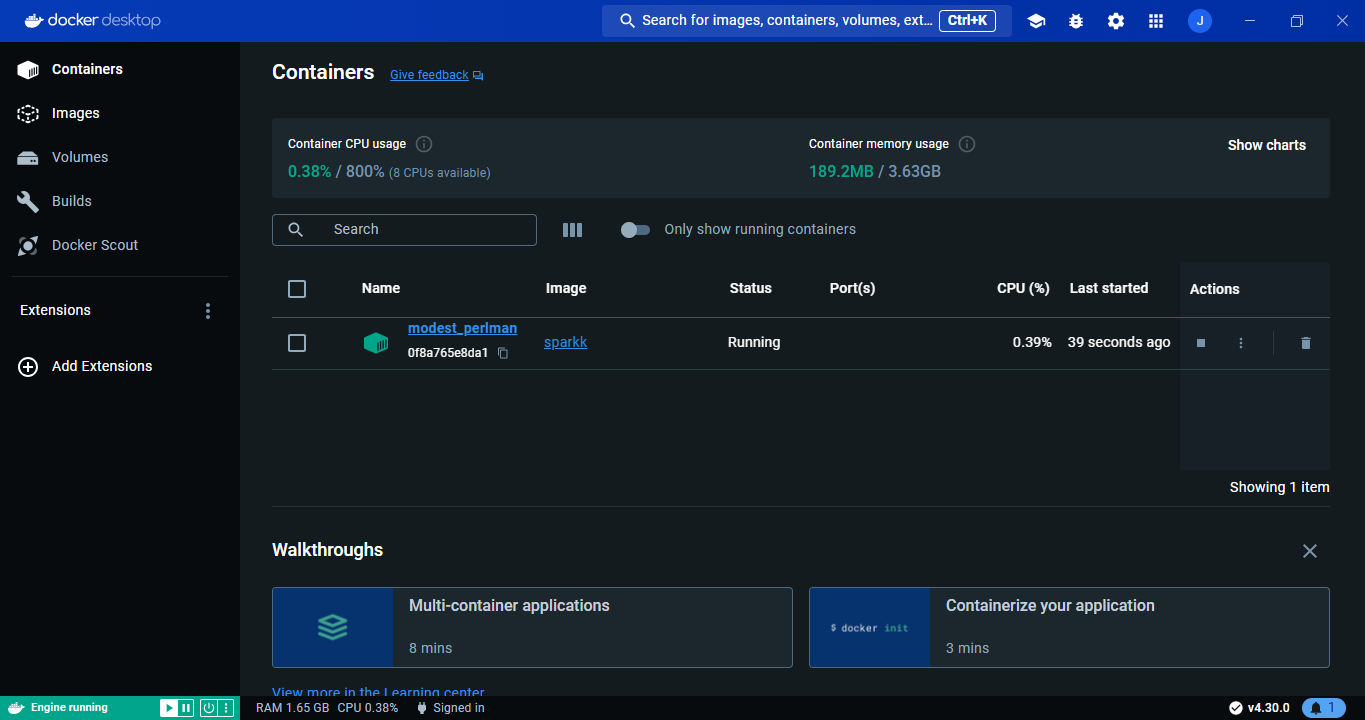
Các bước để build và chạy một Docker Container :

* **Tạo file Dockerfile**: Chứa các chỉ thị để Docker biết cách build image, trong Dockerfile này bao gồm có yêu cầu build image sử dụng base image java11 là

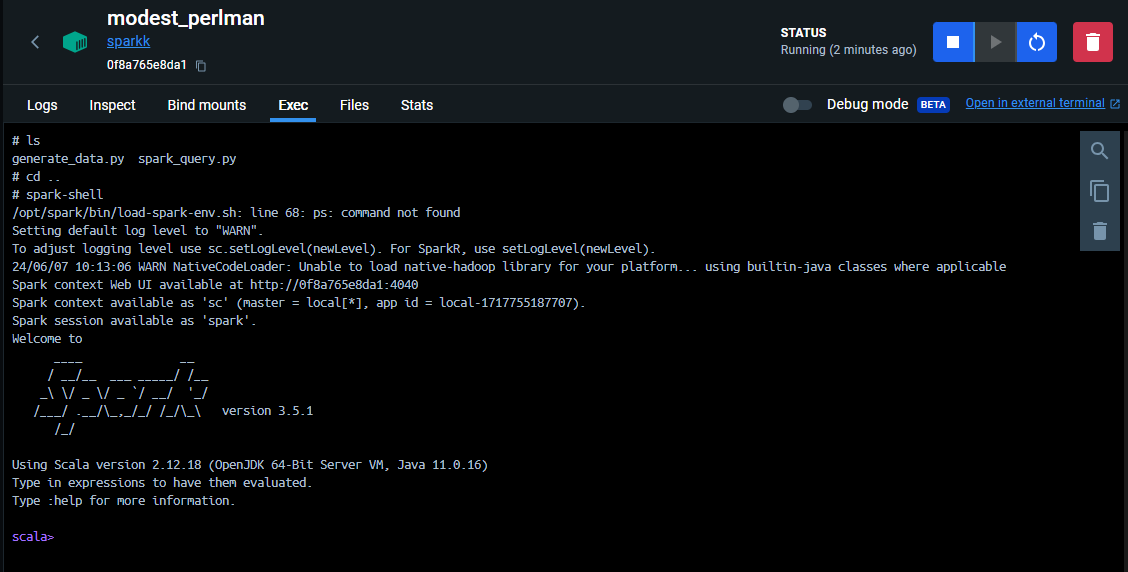
Openjdk:11-jdk-slim, thực hiện cài đặt gói cần thiết python3, sau đó tải Spark và thiết lập môi trường cho Spark và các công cụ để thực hiện project này.

* **Build Docker Image :** Mở terminal và di chuyển đến thư mục chứa Dockerfile và chạy lệnh **docker build -t ten\_cua\_image** để build Docker Image. Để kiểm tra, chạy lệnh **docker images** để hiển thị danh sách các images đã build.
* **Chạy container từ image :** Chạy lệnh **docker run -t ten\_cua\_image** để chạy Docker Container được build từ image hoặc có thể chạy image sau khi build thành công trong Docker Desktop.

Kết quả sau khi build và chạy Docker Container :



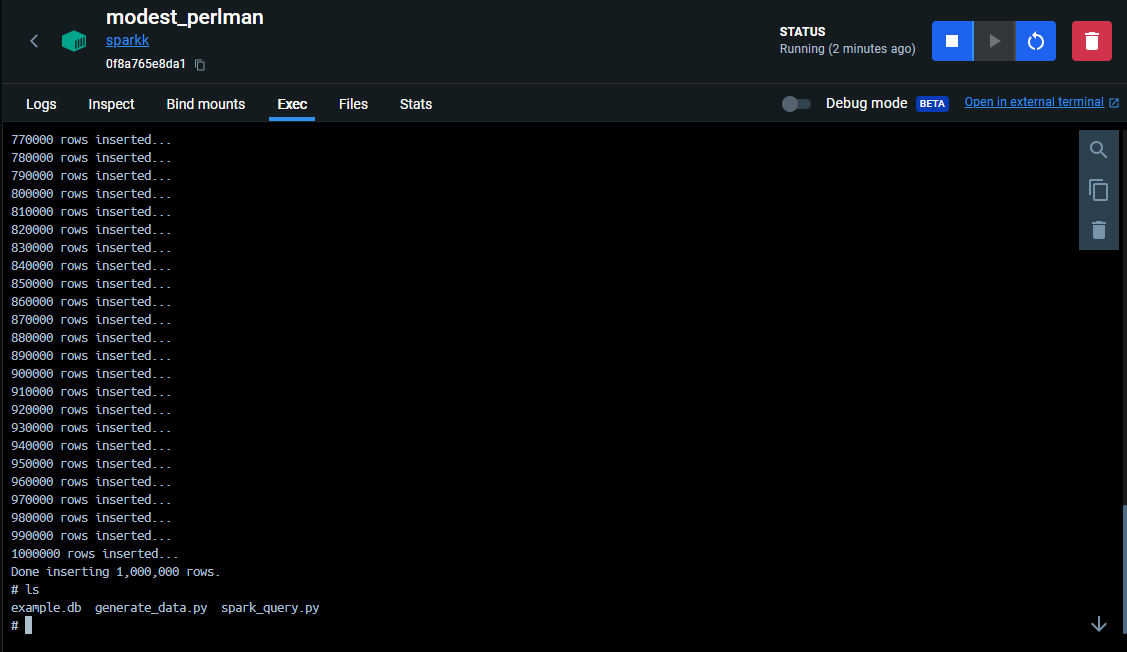
Để xác thực đã cài đặt Spark chưa: Vào container đó 🡪 Exec 🡪 Gõ lệnh spark-shell.



Hiển thị cho kết quả đã cài đặt Spark version 3.5.1 sử dụng Scala 2.12.18, Java11.Như vậy đã hoàn thành xong việc cài đặt Spark trên Docker Container.

1. **Kết nối SparkSQL với cơ sở dữ liệu và thực hiện một số truy vấn cơ bản.**
   1. ***Tạo cơ sở dữ liệu***

Tạo cơ sở dữ liệu với SQLite sử dụng thư viện sqlite3 của python để tạo bảng dữ liệu ngẫu nhiên với 1,000,000 hàng với các cột là id,name,age,address có tên là my\_table và database là example.db. Toàn bộ script tạo database là generate\_data.py và copy script này lên Docker Container thông qua Dockerfile được lưu theo đường dẫn /src/generate\_data.py. Trong Exec của container chạy lệnh **python3 generate\_data.py** để chạy script python sẽ tạo ra database example.db cũng trong cùng thư mục đó. Toàn bộ chi tiết code trong file generate\_data.py.



* 1. ***Kết nối SparkSQL với cơ sở dữ liệu và thực hiện một số truy vấn***

Để kết nối Spark với cơ sở dữ liệu và thực hiện truy vấn, ta sử dụng PySpark và tải JDBC Driver SQLite để có thể connect PySpark với cơ sở dữ liệu SQLite.

Link tải JDBC Driver : <https://www.sqlitetutorial.net/sqlite-java/sqlite-jdbc-driver/>

Kết nối với database qua script python :



Toàn bộ code để kết nối với database và thực hiện truy vấn được thực hiện trong file script spark\_query.py.

Copy JDBC Driver và file script lên container qua Dockerfile. Trong container,sau khi chạy xong lệnh để tạo database, vào Exec nhập **python3 spark\_query.py**.

Kết quả :

