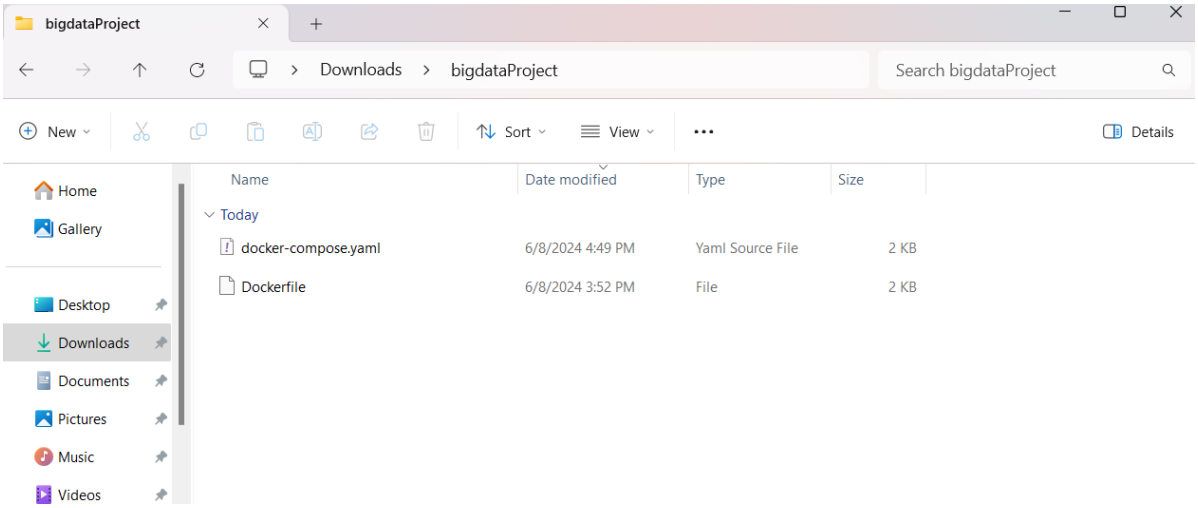
Project kết thúc môn học Thực hành tin học ứng dụng

1. Docker Setup

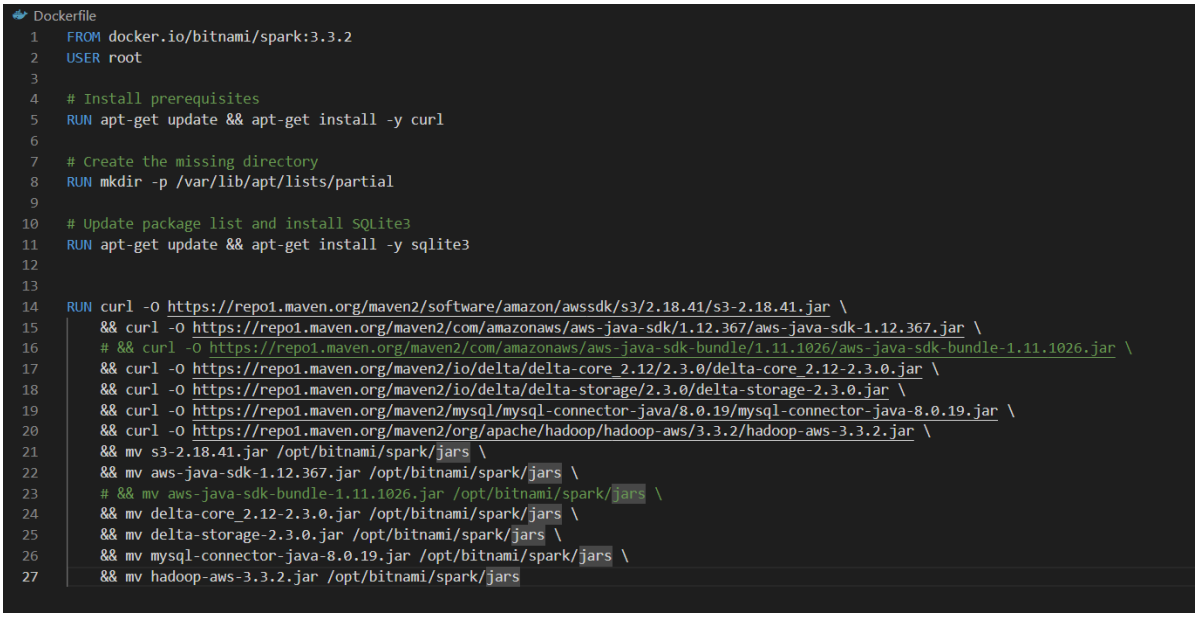
- Thực hiện tải xuống Docker Desktop cho hệ máy Window

Hình 1: Docker Desktop download

- Tạo một thư mục có tên là bigdataProject, bên trong tạo một file Dockerfile và 1 file docker_compose.yaml để phục vụ cho quản lý đa container.



Hình 2. Thư mục của project



```
Dockerfile
FROM docker.io/bitnami/spark:3.3.2
USER root

# Install prerequisites
RUN apt-get update && apt-get install -y curl

# Create the missing directory
RUN mkdir -p /var/lib/apt/lists/partial

# Update package list and install SQLite3
RUN apt-get update && apt-get install -y sqlite3


RUN curl -O https://repo1.maven.org/maven2/software/amazon/awssdk/s3/2.18.41/s3-2.18.41.jar \
    && curl -O https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk/1.12.367/aws-java-sdk-1.12.367.jar \
    # && curl -O https://repo1.maven.org/maven2/com/amazonaws/aws-java-sdk-bundle/1.11.1026/aws-java-sdk-bundle-1.11.1026.jar \
    && curl -O https://repo1.maven.org/maven2/io/delta/delta-core_2.12/2.3.0/delta-core_2.12-2.3.0.jar \
    && curl -O https://repo1.maven.org/maven2/io/delta/delta-storage/2.3.0/delta-storage-2.3.0.jar \
    && curl -O https://repo1.maven.org/maven2/mysql/mysql-connector-java/8.0.19/mysql-connector-java-8.0.19.jar \
    && curl -O https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-aws/3.3.2/hadoop-aws-3.3.2.jar \
    && mv s3-2.18.41.jar /opt/bitnami/spark/jars \
    && mv aws-java-sdk-1.12.367.jar /opt/bitnami/spark/jars \
    # && mv aws-java-sdk-bundle-1.11.1026.jar /opt/bitnami/spark/jars \
    && mv delta-core_2.12-2.3.0.jar /opt/bitnami/spark/jars \
    && mv delta-storage-2.3.0.jar /opt/bitnami/spark/jars \
    && mv mysql-connector-java-8.0.19.jar /opt/bitnami/spark/jars \
    && mv hadoop-aws-3.3.2.jar /opt/bitnami/spark/jars
```

Hình 3. File Dockerfile

- Mở terminal của thư mục trên hình 2, thực hiện tạo một images bằng lệnh: "**docker built –t project_image .**"
- Sau đó chạy images đó để chạy container bằng lệnh: "**docker run -it project_image bash**"

docker-compose.yaml

D: > THUD > docker-compose.yaml

```yaml
version: "3.9"

services:
  spark-master:
    build:
      context: ./docker_image/spark
      dockerfile: ./Dockerfile
    container_name: "spark-master"
    ports:
      - "7077:7077"  # Spark master port
      - "8081:8080"  # Spark master web UI port
    expose:
      - "7077"
    environment:
      - SPARK_MODE=master
      - SPARK_RPC_AUTHENTICATION_ENABLED=no
      - SPARK_RPC_ENCRYPTION_ENABLED=no
      - SPARK_LOCAL_STORAGE_ENCRYPTION_ENABLED=no
      - SPARK_SSL_ENABLED=no
      - SPARK_USER=spark
    volumes:
      - ./docker_image/spark/conf/spark-defaults.conf:/opt/bitnami/spark/conf/spark-defaults.conf
      - ./docker_image/spark/conf/log4j.properties:/opt/bitnami/spark/conf/log4j.properties
      - ./data:/opt/spark
    networks:
      - data_network

  spark-worker-1:
    image: docker.io/bitnami/spark:3.3.2
    container_name: "spark-worker-1"
    env_file:
      - .env
    depends_on:
      - spark-master
    networks:
      - data_network
```

Hình 4: Docker-Compose

```
PS D:\Download\bigdataProject> docker run -it project_image bash
spark 10:12:10.72
spark 10:12:10.72 Welcome to the Bitnami spark container
spark 10:12:10.72 Subscribe to project updates by watching https://github.com/bitnami/containers
spark 10:12:10.73 Submit issues and feature requests at https://github.com/bitnami/containers/issues
spark 10:12:10.73

root@9487d4a31935:/opt/bitnami/spark#
```

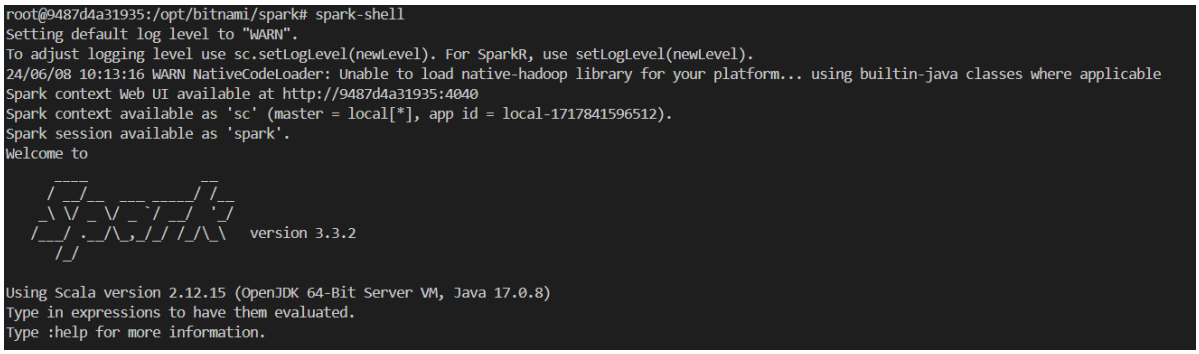
Hình 5: Xác nhận build thành công

2. Spark Installation

Trong Dockerfile chứa các câu lệnh cho việc cài đặt các công cụ cần thiết cho việc sử dụng và chạy Spark trên container. Ngoài ra có một số công cụ và thư viện để dễ dàng hơn trong quá trình làm việc với Docker như sau:

- SQLite3: Một hệ quản trị cơ sở dữ liệu nhúng (embedded database) nhẹ, không cần server, thường được sử dụng cho các ứng dụng nhỏ hoặc phát triển.
- AWS SDK for Java (s3-2.18.41.jar): Bộ công cụ phát triển phần mềm của Amazon cho phép Spark tương tác với các dịch vụ lưu trữ Amazon S3.
- AWS Java SDK (aws-java-sdk-1.12.367.jar): Bộ công cụ toàn diện hơn của Amazon cho Java, bao gồm nhiều dịch vụ AWS khác ngoài S3.
- Delta Lake (delta-core_2.12-2.3.0.jar & delta-storage-2.3.0.jar): Một framework mã nguồn mở giúp xây dựng hồ dữ liệu (data lake) trên nền tảng Apache Spark, cung cấp các tính năng ACID và khả năng xử lý dữ liệu theo thời gian thực (real-time).
- MySQL Connector/J (mysql-connector-java-8.0.19.jar): Trình điều khiển JDBC để kết nối Spark với cơ sở dữ liệu MySQL.
- Hadoop AWS (hadoop-aws-3.3.2.jar): Thư viện Hadoop cung cấp tích hợp giữa Hadoop và các dịch vụ AWS, cho phép Spark đọc và ghi dữ liệu từ các dịch vụ lưu trữ AWS như S3.

Sau đó, thực hiện **build** lại images với tên my-task và chạy container và cuối cùng, nhập lệnh "**spark-shell**" để xác minh Spark đã chạy



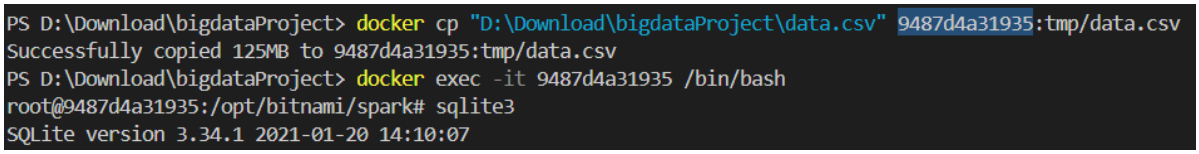
Hình 6: Xác nhận Spark thành công

3. Database Setup

Trong Dockerfile, chúng ta đã có câu lệnh thực hiện việc update và install Sqlite để đảm bảo việc Sqlite được cài đặt khi chúng ta thực hiện tạo Image và chạy Container

Tiếp theo, chúng ta thực hiện việc copy 1 file có tên 1000000 Sales Records.csv được lấy từ Excel BI Analytics, sau đó ta đổi tên thành data.csv.

Truy cập vào terminal của container bằng câu lệnh `docker exec -it <spark-id> /bin/bash` , sau đó tạo và chạy cơ sở dữ liệu sqlite 3. <spark-id> hiện ở vị trí root@<spark-id> ở màn hình build.



Ảnh 7: Tạo file data.csv và cơ sở dữ liệu sqlite 3

Sau khi thực hiện bước trên, giao diện sẽ hiển thị tương tác với database, lúc này ta sẽ thực hiện 2 lệnh:

- .mode csv
- .import data.csv my_table

Quá trình này import file data.csv gồm 1 triệu dòng vào my_table

-Thử một câu truy vấn hiển thị 10 dòng dữ liệu đầu

```
sqlite> .mode csv
sqlite> .import data.csv my_table

sqlite>
sqlite> sqlite> SELECT COUNT(*) FROM my_table;
Error: near "sqlite": syntax error
sqlite> SELECT COUNT(*) FROM my_table;
1000000
sqlite> SELECT * FROM my_table LIMIT 10;
"Sub-Saharan Africa","South Africa",Fruits,Offline,M,7/27/2012,443368995,7/28/2012,1593,9.33,6.92,14862.69,11023.56,3839.13
"Middle East and North Africa",Morocco,Clothes,Online,M,9/14/2013,667593514,10/19/2013,4611,109.28,35.84,503890.08,165258.24,338631.84
"Australia and Oceania","Papua New Guinea",Meat,Offline,M,5/15/2015,940995585,6/4/2015,360,421.89,364.69,151880.40,131288.40,20592.00
"Sub-Saharan Africa",Djibouti,Clothes,Offline,H,5/17/2017,880811536,7/2/2017,562,109.28,35.84,61415.36,20142.08,41273.28
Europe,Slovakia,Beverages,Offline,L,10/26/2016,174590194,12/4/2016,3973,47.45,31.79,188518.85,126301.67,62217.18
Asia,"Sri Lanka",Fruits,Online,L,11/7/2011,830192887,12/18/2011,1379,9.33,6.92,12866.07,9542.68,3323.39
"Sub-Saharan Africa","Seychelles ",Beverages,Online,M,1/18/2013,425793445,2/16/2013,597,47.45,31.79,28327.65,18978.63,9349.02
"Sub-Saharan Africa",Tanzania,Beverages,Online,L,11/30/2016,659878194,1/16/2017,1476,47.45,31.79,70036.20,46922.04,23114.16
"Sub-Saharan Africa",Ghana,"Office Supplies",Online,L,3/23/2017,601245963,4/15/2017,896,651.21,524.96,583484.16,470364.16,113120.00
"Sub-Saharan Africa",Tanzania,Cosmetics,Offline,L,5/23/2016,739008080,5/24/2016,7768,437.20,263.33,3396169.60,2045547.44,1350622.16
sqlite>
```

4. Dependencies

Chúng ta sẽ thêm 2 thành phần cần thiết là PySpark và JDBC cho sqlite để giúp kết nối và thực hiện truy vấn từ Spark đến cơ sở dữ liệu. Để thực hiện được tôi đã sử dụng 3 câu lệnh sau:

pip install --upgrade setuptools
pip install pyspark

wget https://repo1.maven.org/maven2/org/xerial/sqlite-jdbc/3.34.0/sqlite-jdbc-3.34.0.jar

```
PS D:\THUD> pip install pyspark
Requirement already satisfied: pyspark in c:\users\pro 2004\appdata\local\programs\python\python311\lib\site-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in c:\users\pro 2004\appdata\local\programs\python\python311\lib\site-packages (from pyspark) (0.10.9.7)

[notice] A new release of pip is available: 23.3.1 -> 24.0
[notice] To update, run: python.exe -m pip install --upgrade pip
PS D:\THUD> pip install pyspark
Requirement already satisfied: pyspark in c:\users\pro 2004\appdata\local\programs\python\python311\lib\site-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in c:\users\pro 2004\appdata\local\programs\python\python311\lib\site-packages (from pyspark) (0.10.9.7)

[notice] To update, run: python.exe -m pip install --upgrade pip
PS D:\THUD> python -c "import pyspark; print(pyspark.__version__)"
>>
3.5.1
PS D:\THUD> wget https://repo1.maven.org/maven2/org/xerial/sqlite-jdbc/3.34.0/sqlite-jdbc-3.34.0.jar


StatusCode        : 200
StatusDescription : OK
Content           : {80, 75, 3, 4...}
RawContent        : HTTP/1.1 200 OK
                    Connection: keep-alive
                    X-Checksum-MD5: 743bacfa02e66cad1027e80b065c45ad
                    X-Checksum-SHA1: fd29bb0124e3f79c80b2753162a6a3873c240bcf
                    Age: 2656032
                    X-Served-By: cache-iad-kiad7000141-I...
Headers           : {[Connection, keep-alive], [X-Checksum-MD5, 743bacfa02e66cad1027e80b065c45ad], [X-Checksum-SHA1, fd29bb0124e3f79c80b2753162a6a3873c240bcf], [Age, 2656032].
..}
RawContentLength  : 7296329
```

5. Configuration