

Họ và tên: Nguyễn Xuân An

Mã Sinh Viên: 21002183

Lớp: K66 Kỹ Thuật Điện Tử Tin Học

Cấu hình và chạy SparkSQL truy vấn cơ sở dữ liệu với DOCKER

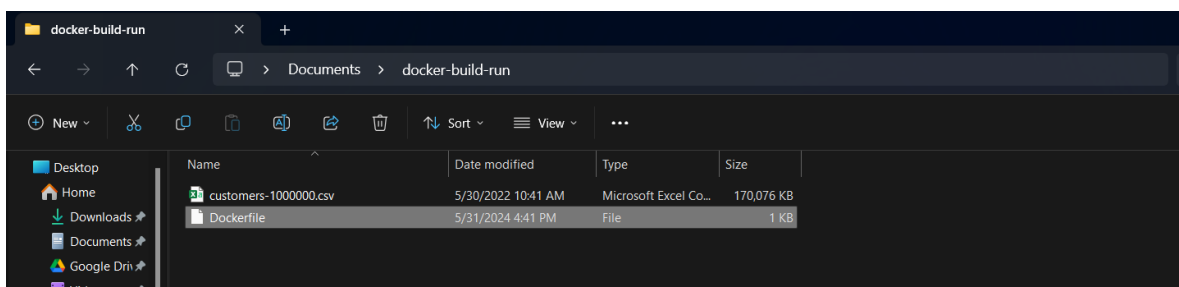
1. Docker Setup

- Thực hiện tải xuống Docker Desktop cho hệ máy Window
- Tạo một thư mục có tên docker-build-run
- Tạo một File với tên Dockerfile không có hậu tố
- Sử dụng trình soạn thảo Notepad thêm vào File đó dòng sau:
#Lựa chọn images gốc cho container

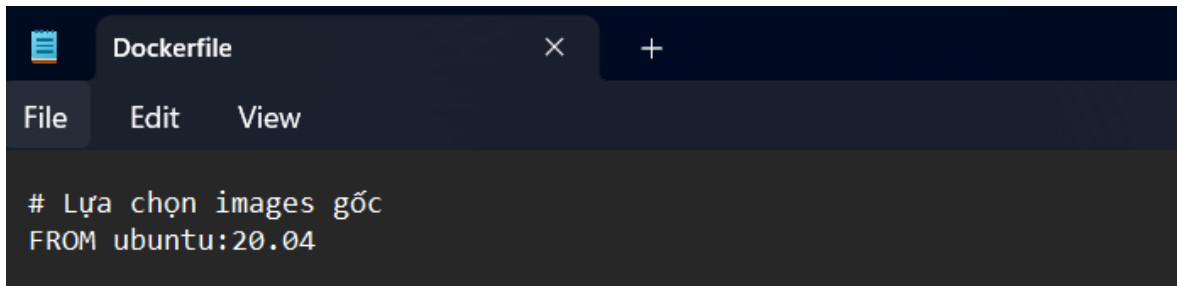
FROM ubuntu:20.04

COPY customers-1000000.csv /home/data.csv

- Sau đó vào thư mục docker-build-run, chạy thư mục với cmd bằng cách chuột phải tại thư mục và chọn "**open in terminal**"
- Tạo một images bằng lệnh: "**docker build -t lab-final .**"
- Sau đó chạy images đó để chạy container bằng lệnh: "**docker run -it lab-final bash**" hoặc chạy trong giao diện của Docker Desktop và tương tác với container qua mục **Exec**

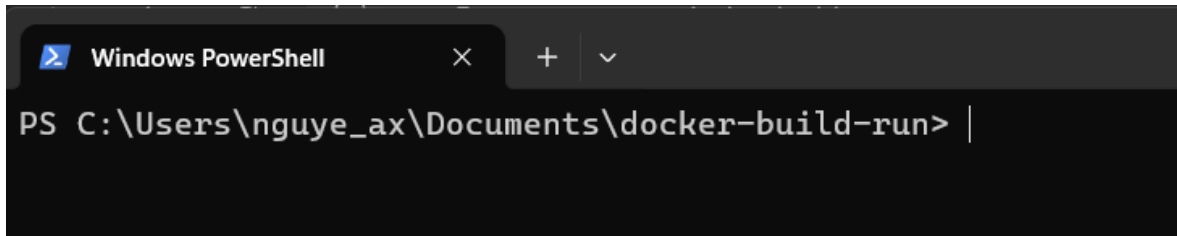


Ảnh 1: Thư mục lưu trữ



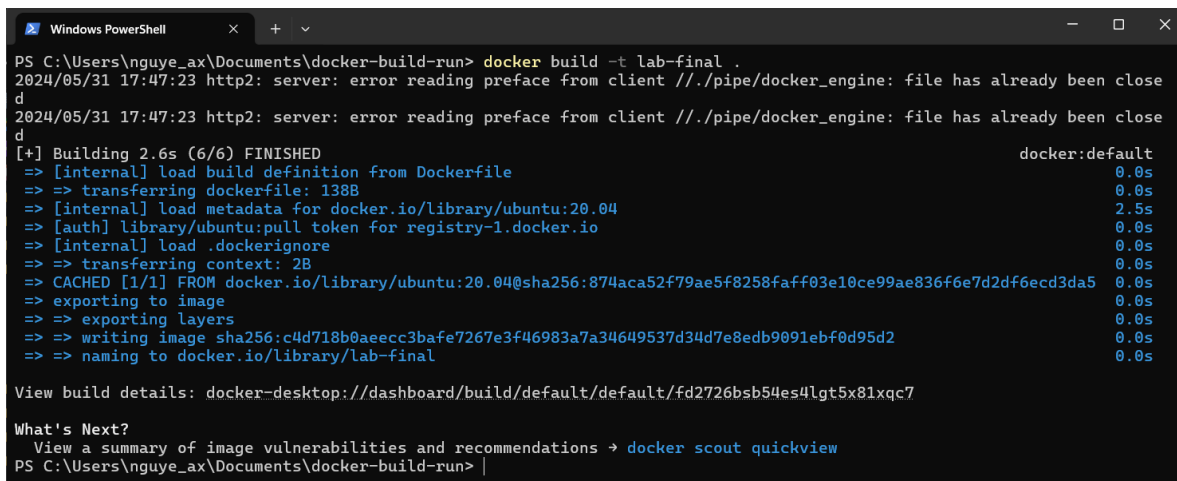
```
Dockerfile
File Edit View
# Lựa chọn images gốc
FROM ubuntu:20.04
```

Ảnh 2: Cấu hình Dockerfile



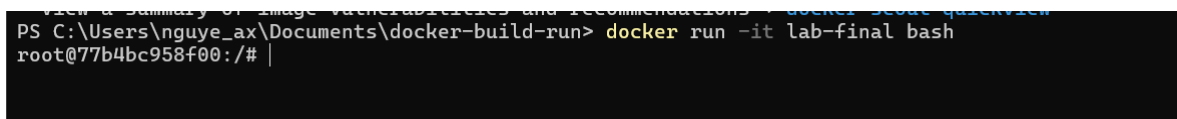
```
Windows PowerShell
PS C:\Users\nguye_ax\Documents\docker-build-run> docker build -t lab-final .
```

Ảnh 3: Chạy với giao diện cmd



```
Windows PowerShell
PS C:\Users\nguye_ax\Documents\docker-build-run> docker build -t lab-final .
2024/05/31 17:47:23 http2: server: error reading preface from client //./pipe/docker_engine: file has already been closed
2024/05/31 17:47:23 http2: server: error reading preface from client //./pipe/docker_engine: file has already been closed
[+] Building 2.6s (6/6) FINISHED
=> [internal] load build definition from Dockerfile
=> => transferring dockerfile: 138B
=> [internal] load metadata for docker.io/library/ubuntu:20.04
=> [auth] library/ubuntu:pull token for registry-1.docker.io
=> [internal] load .dockerignore
=> => transferring context: 2B
=> CACHED [1/1] FROM docker.io/library/ubuntu:20.04@sha256:874aca52f79ae5f8258fa5ff03e10ce99ae836f6e7d2df6ecd3da5
=> exporting to image
=> exporting layers
=> => writing image sha256:c4d718b0aecc3baf7267e3f46983a7a34649537d34d7e8edb9091ebf0d95d2
=> => naming to docker.io/library/lab-final
View build details: docker-desktop://dashboard/build/default/default/fd2726bsb54es4lgt5x81xqc7
What's Next?
View a summary of image vulnerabilities and recommendations -> docker scout quickview
PS C:\Users\nguye_ax\Documents\docker-build-run> |
```

Ảnh 4: Build images



```
PS C:\Users\nguye_ax\Documents\docker-build-run> docker run -it lab-final bash
root@77b4bc958f00:/#
```

Ảnh 5: Chạy và tương tác với container thông qua cmd

2. Spark Installation

-Để chạy Spark trên container vừa tạo ta cần cài đặt một số công cụ cần thiết cho việc sử dụng và chạy. Sử dụng **Dockerfile** tạo ở phần 1 và thêm vào các lệnh sau

RUN apt-get update

RUN apt-get -y install openjdk-8-jdk

```
RUN rm spark-3.5.1-bin-hadoop3.tgz
```

8. Quá trình khởi chạy đã thành công, giờ ta có thể sử dụng ngôn ngữ Scala để thực hiện các thao tác với dữ liệu

```
root@881862b7a8d3:/# spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/05/31 12:48:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
Spark context Web UI available at http://881862b7a8d3:4040
Spark context available as 'sc' (master = local[*], app id = local-1717159720115).
Spark session available as 'spark'.
Welcome to

      /_--/_ _--/_ _--/_ _--/_ _--/_ \
     /  V  -V- -V- -V- -V- -V- -V- \
    /---/_ _--/_ _--/_ _--/_ _--/_ \   version 3.5.1
     /  /  /  /  /  /  /  /  /  /  /
    /  /  /  /  /  /  /  /  /  /  /

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 1.8.0_402)
Type in expressions to have them evaluated.
Type :help for more information.

scala> |
```

Ảnh 6: Xác minh Spark đã chạy

3. Database Setup

-Để truy cập vào container vừa mới build xong ta có thể sử dụng lệnh sau **docker exec -it <name_container> bash**. Quá trình này sẽ cho phép ta tương tác với **Container** thông qua giao diện **cmd**. Lúc này, ta sẽ thao tác và cài đặt thêm một số công cụ cho **Container** này

```
Learn more at https://docs.docker.com/go/debug-cli/  
PS C:\Users\nguye_ax> docker exec -it 881862b7a8d3 bash  
root@881862b7a8d3:/#
```

-Đầu tiên chạy lệnh **apt-get install sqlite3** để tải về SQLite

-Tiếp theo, ở **bước 1** chúng ta đã copy một file có tên customers-1000000.csv vào thư mục **/home** với tên **/data.csv**

-Bây giờ tạo và chạy cơ sở dữ liệu bằng lệnh sau: **sqlite3 /home/sample.db**

-Giao diện dòng lệnh sẽ hiển thị đang tương tác với database, lúc này thực hiện 2 lệnh:

.mode csv

.import /home/data.csv my_database

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

-SELECT * FROM my_database;

-Với câu truy vấn trên toàn bộ dữ liệu, dữ liệu cuối cùng chứa Index 1000000 đã phù hợp với nội dung yêu cầu

```
999986,D5DA9cf9BC389d8,Amy,Barnett,"Pruitt, Herrera and Jackson","New Stephenchester","Saint Pierre and Miquelon",392.53  
1.6256x239,+1-913-005-4831x06431,mcintyrejanice@randolph.com,2020-12-12,http://www.howell-rasmussen.com/  
999987,1c8b26562AcF0cD,Bob,Hall,Bridges-Copeland,"South Saraton",Kuwait,(413)815-7543,001-007-503-7034x9859,neil62@chave  
z-rush.com,2020-08-19,http://beltran.info/  
999988,Dba0E79580Ff18a,Nathaniel,Vargas,Berger-Tyler,"New Sandrachester","Isle of Man",001-984-079-2616,001-447-773-3597  
x072,luisbullock@chambers-ross.com,2021-08-11,http://www.holder-montgomery.info/  
999989,0501d29C32cbC7c,Alejandro,lloyd,"Pierce Inc",Shawburgh,"Western Sahara",+1-734-311-3623x961,001-075-485-5461x443,  
dmaddox@vang-dillon.com,2022-04-15,https://carpenter-trujillo.com/  
999990,Db03Be4E0eC58Fa,Latoya,Ingram,"Wolfe, Lewis and Wilkins","North Carolinetown","United States Minor Outlying Islan  
ds",+1-178-402-3570x56177,464.725.5027x33781,gavinlyons@hayes.com,2022-01-06,http://www.hubbard.com/  
999991,AE5A9A1DBbe29D,Samuel,Weeks,"Mcconnell and Sons",Henryside,"Sierra Leone",084-308-1121,+1-428-821-0397x0147,sava  
gejacqueline@sandoval.com,2020-02-22,http://wong.com/  
999992,561DC620eb44c6a,Drew,Kelley,Estrada-Nelson,Villarrealmouth,Georgia,703-287-8298x88540,803.430.9046x0217,feliciaea  
ton@paul.org,2020-04-06,http://www.bernard.com/  
999993,f9cDFA42Bf458F3,Vincent,Newton,"Macdonald Group",Isabellaberg,Mauritania,(069)934-1961x6086,(869)808-0548,shawn03  
@lloyd-owens.info,2020-08-19,http://www.hansen.net/  
999994,28db47A178aeaB2,Tracey,Roy,"Gonzales and Sons","New Alejandrohaven",Armenia,+1-933-821-0901x692,597.145.4870x659,  
xcollins@carpenter.biz,2020-11-04,https://santana-yang.com/  
999995,Dbdb2bAdfDE0C46,Cheyenne,Vega,"Esparza Inc",Hendrickstown,"Czech Republic",001-986-928-4652x5540,988.459.3161x875  
93,gconway@frazier-moody.com,2022-01-29,http://www.ball-welch.com/  
999996,173F0830C74f3Ab,Jacqueline,Wu,Maxwell-Brennan,Starkhaven,Comoros,+1-957-385-5334x95546,3174863938,dakota95@clark.  
info,2020-07-13,http://www.cantrell.com/  
999997,519F052bE7C94F8,Warren,Donovan,Tate-Hart,Castrobury,Ghana,879.805.1973x228,582.649.3363,joywilkerson@lynn.biz,202  
1-07-20,https://mccconnell-cabrera.com/  
999998,67bed9c4EC3F2a8,Whitney,Lara,"Chandler, Farrell and Macias","West Joanne",Aruba,104.402.9455x883,2895286724,bmaci  
as@gregory-hodges.com,2022-01-26,http://www.gay.net/  
999999,182BC9936E3eCb,Aaron,Mccormick,"Ali, Montoya and Lamb",Hobbsfurt,Tunisia,380.990.7339x526,0024827409,jgreer@long  
-short.com,2021-11-08,http://small.biz/  
1000000,B1DcBaB825d3E63,Craig,Carroll,"Riley, Huerta and Merritt","East Meredithland",Fiji,001-897-468-8999x0265,(125)87  
7-8340x695,annwolf@zhang-castro.com,2021-08-13,https://kim.com/
```

4. Dependencies

-Ở phần này chúng ta sẽ thêm 2 thành phần cần thiết đó là PySpark và JDBC cho sqlite để giúp kết nối và thực hiện truy vấn từ Spark đến cơ sở dữ liệu. Chạy 2 lệnh sau

```
pip install pyspark
```

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

5. Configuration

-Để Spark có thể kết nối đến cơ sở dữ liệu ta cần di chuyển thư viện JDBC vào thư mục jars của Spark, đây là thư mục giúp Spark tự động phát hiện và chạy các thư viện phụ trợ nếu cần thiết. Nhập lệnh

```
mv sqlite-jdbc-3.34.0.jar /opt/spark/jars
```

-Sau khi di chuyển hoàn tất, ta có thể thử truy cập và tương tác với cơ sở dữ liệu bằng PySpark

-Tại giao diện cmd, nhập "**pyspark**" để khởi chạy, lúc này một phiên làm việc của Spark thông qua giao diện dòng lệnh sẽ bắt đầu, nó cung cấp một môi trường tương tác với Spark bằng Python.

```
root@881862b7a8d3:/home# pyspark
Python 3.8.10 (default, Nov 22 2023, 10:22:35)
[GCC 9.4.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/05/31 23:03:12 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to

  ____      _
 / ___|    / \
| |  | |  / _ \
| |  | | / ___ \
| |  | || |___| \
| |  | || |___| \
| |  | || |___| \
|_|  |_| \____/

version 3.5.1

Using Python version 3.8.10 (default, Nov 22 2023 10:22:35)
Spark context Web UI available at http://881862b7a8d3:4040
Spark context available as 'sc' (master = local[*], app id = local-1717196593293).
SparkSession available as 'spark'.
>>>
```

-Sau đó ta nhập lệnh sau

```
df=spark.read.format("jdbc").option("url",f"jdbc:sqlite:/home/sample.db").option("dbtable","my_database").load()
```

```
df.show(13)
```

-Dưới đây là thử nghiệm kết nối đến cơ sở dữ liệu sample.db, bảng my_database, hiển thị 13 dòng đầu tiên

```
>>> df = spark.read.format("jdbc").option("url",f"jdbc:sqlite:/home/sample.db").option("dbtable","my_database").load()
>>> df.show(13)
```

Index	Customer Id	First Name	Last Name	Company	City	Country	P
hone 1	Phone 2	Email	Subscription Date	Website			
1	138fB5315da5fE9	Jeanne	Ferrell	Wilcox-Fox	Tonichester	British Virgin Is...	001-995-820-01
40x...	(757)324-8634	aaronwoods@walter...	2020-03-27	https://www.romer...			
2	b0d61acAc72A388	Ian	Browning	Meadows Inc	Colontown	El Salvador	218-38
3-6764	+1-213-212-0464x0742	zschoultz@blevins-...	2021-03-12	https://ford.com/			
3	1B27Ff7Fd418C89	Taylor	Martinez	Nicholson Inc	East Evelyn	Marshall Islands	+1-455-87
5-7024	(783)689-6710x09859	monicacoffey@mood...	2021-08-29	http://love-moral...			
4	eff8bbcdE3eacD8	Andre	Mccall	Cooper Ltd	New Luis	Colombia	(334)358-51
62x861	016-422-2338	parkerdiana@orr.net	2020-05-29	http://vaughan.biz/			
5	73ee6AaCAcea39C	Alyssa	Mcneil	Arnold, Neal and ...	New Jay	Netherlands Antilles	+1-344-219-80
95x764	001-178-547-0380x...	janePhillips@nich...	2020-11-10	http://ibarra-adk...			
6	e4A1fb3fA732CED	Marisa	West	Stanley PLC	Dalemouth	Morocco	+1-211-391-09
83x663	701-169-4514	matthew73@rush.info	2020-07-09	http://www.hollan...			
7	dC9dbfa601b71b5	Jaclyn	Branch	Ballard LLC	East Jacquelineberg	Bulgaria	3552
675231	+1-769-658-2475x7...	hrobbins@hodge.com	2020-05-07	https://patrick-w...			
8	38ce8fA0e07AdEe	Billy	Stone	Decker, Frederick...	Tammieburgh	Cayman Islands	109.506.040
1x9595	405.634.9674	kmcneil@hendrix-c...	2021-11-06	http://www.stout...			
9	134e7E02Ccb2614	Cristian	Waters	Bright LLC	Shaneville	Uruguay	(569)288-11
86x339	371.412.4418	wanda26@rios.com	2020-03-06	https://www.clark...			

6. Running Queries

-Đầu tiên, để có thể thực hiện các thao tác với cơ sở dữ liệu thông qua các file nội dung, ta cần cấu hình cho session của mỗi file. Tạo một file có tên `get_session_data` để chạy session và khởi động trình truy vấn SQL

-Nhập lệnh "**vi get_session_data.py**" rồi điền vào các nội dung sau

```
from pyspark.sql import SparkSession

def get_session():
    spark = SparkSession.builder.appName("PySpark basic").config("spark.some.config.option", "some-value").getOrCreate()
    return spark

def SQL_Queries(spark, database_path, table_name):
    df = spark.read.jdbc(url = f"jdbc:sqlite:{database_path}", table = table_name)
    df.createOrReplaceTempView(table_name)
    pass
```

-Hàm `get_session()` cho phép khởi tạo các thiết lập và tài nguyên cần thiết cho việc thao tác với Spark

-Hàm `SQL_Queries()` cho phép ta khởi tạo một trình truy vấn SQL với cơ sở dữ liệu được chọn dưới dạng tạm thời và sẽ không làm ảnh hưởng đến cơ sở dữ liệu thực của bạn

6.1. Truy vấn cơ sở dữ liệu

-Sau khi đã tạo file cấu hình xong, giờ ta sẽ tạo một file để chạy truy vấn đồng thời kiểm tra thời gian chạy của nó.

-Tạo 2 file tên **exam1.py** và **exam2.py** truy vấn cơ sở dữ liệu và điền vào nội dung sau. Hàm đo thời gian sẽ chỉ tính trong phạm vi câu truy vấn. Sự khác biệt 2 file là điều kiện **WHERE Index < 10** để xem sự ảnh hưởng của **WHERE** sẽ thế nào đến truy vấn

```
root@881862b7a8d3: /home X + v
from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql("select * from my_database where Index < 10")

time_end = time.time()
|
result.show()

print("thoi gian chay:",time_end - time_start)
~
~
```

Ảnh 1: nội dung exam1.py

```
root@881862b7a8d3: /home X + v
from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql("select * from my_database")

time_end = time.time()

result.show()

print("thoi gian chay:",time_end - time_start)
~
```

Ảnh 2: Nội dung exam2.py

-Kết quả với điều kiện **WHERE**, **exam1** mất khoảng **0.13s** để thực hiện trong khi **exam2.py** không có **WHERE** chỉ mất khoảng **0.08s**

```
thoi gian chay: 0.08145099229431152
root@881862b7a8d3: /home# python3 exam1.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/02 13:15:09 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

[Index]	Customer Id	First Name	Last Name	Company	City	Country	Phone 1	Phone 2
	Email	Subscription Date	Website					
	1 138f85315da5fe9	Jeanne	Ferrell	Wilcox-Fox	Tonichester	British Virgin Is...	[001-995-820-0140x...]	(757)324-8634 aaronwoo
	2 b8d61acAc72A388	ian	Browning	Meadows Inc	Colontown	El Salvador	218-383-6764 +1-213-212-0464x0742 zschultz	
	3 1b27ff7fd418c89	Taylor	Martinez	Nicholson Inc	East Evelyn	Marshall Islands	+1-455-875-7824 (783)689-6710x09859 monicaco	
	4 eff8bbcd3eac08	Andre	Mccall	Cooper Ltd	New Luis	Colombia	(334)358-5162x861	016-422-2338 parkerd
	5 72ee6aAcacea39C	Alyssa	Arnell	Neal and ...	New Jay	Netherlands Antilles	+1-344-219-8095x764 001-178-547-0380x... janephil	
	6 e41fb3fa732CED	Marisa	West	Stanley PLC	Dalemouth	Morocco	+1-211-391-0983x663	701-169-4514 matthew
	7 dc9dbfa601b71b5	Jaclyn	Branch	Ballard LLC	East Jacquelineberg	Bulgaria	3552675231 +1-769-658-2475x7... hrobbi	
	8 38ce8fa0e07Adfe	Billy	Stone	Decker, Frederick...	Tammieburgh	Cayman Islands	189.506.0481x9595	405.634.9674 kmcneil@
	9 134e7E92Ccb2614	Cristian	Waters	Bright LLC	Shaneville	Uruguay	(569)288-1186x339	371.412.4418 wand

```
thoi gian chay: 0.12842392921447754
root@881862b7a8d3: /home#
```

Ảnh 3: Kết quả truy vấn exam1.py


```
ffey@mood...| 2021-08-29|http://love-moral...|
| 4|eff8bbcED3eacD8| Andre| McCall| Cooper Ltd| New Luis| Colombia| (334)358-5162x861| 016-422-2338| parkerd
iana@orr.net| 2020-05-29| http://vaughan.biz/|
| 5|73ee6AaCAcea39C| Alyssa| Mcneil|Arnold, Neal and ...| New Jay|Netherlands Antilles| +1-344-219-8895x764|001-178-547-0380x...|janephil
lips@nich...| 2020-11-18|http://ibarra-adk...|
| 6|e4A1f3FA732CED| Marisa| West| Stanley PLC| Dalemouth| Morocco| +1-211-391-0983x663| 701-169-4514| matthew
73@rush.info| 2020-07-09|http://www.hollan...|
| 7|dC9dbfa601b71b5| Jaclyn| Branch| Ballard LLC|East Jacquelineberg| Bulgaria| 3552675231|+1-769-658-2475x7...| hrobbi
ns@hodge.com| 2020-05-07|https://patrick-w...|
| 8|38ce8FA0e07AdEa| Billy| Stone|Decker, Frederick...| Tammieburgh| Cayman Islands| 109.506.0481x9595| 405.634.9674|kmcneil@
hendrix@c...| 2021-11-06|http://www.stout...|
| 9|134e7E02Ccb2614| Cristian| Waters| Bright LLC| Shaneville| Uruguay| (569)288-1186x339| 371.412.4418| wand
a26@rios.com| 2020-03-06|https://www.clark...|
| 10|BAd847028B01cCF| Edward| Meza| Moore PLC| New Jade| Puerto Rico|+1-753-511-1815x4880| 388.374.1422x28744|sheilari
os@myers.biz| 2021-07-15| https://farley.com/|
| 11|AB9F973e25C48ce| Tonya| Casey| Anthony Ltd| Jonbury| Macedonia| 788.631.6521| 756-377-6482x426|roachger
ald@gomez...| 2020-04-17|http://obrien-hol...|
| 12|097E7318E2BdeE| Franklin| Estes| Ho LLC| Salasville| Sierra Leone| 439.903.8544x933| 913-415-3834| jaclyn4
0@ramsey.com| 2021-04-13|http://www.nelson...|
| 13|aF87BD4c6815163| Candace|Macdonald|Wilkins, Villa an...| New Walter| Montenegro| (205)001-7993x5174| 7247692116|hpeterse
n@whitney...| 2020-01-12|http://www.odom.com/|
| 14|8d3acB34bf34dC2| Gabriella| Nielsen| Pollard-Meyers| Gordonburgh| Sweden| +1-650-214-2790x041| +1-708-649-9945| shane8
3@reyes.info| 2021-09-26|http://www.keith...|
| 15|baf4e1E54ba26Fe| Linda| Velez| Harrell PLC| Terrellport| Guadeloupe| (620)148-8832x4011| (430)644-3960x13772| rebec
ca53@ali.org| 2021-01-29|http://www.gutier...|
| 16|d7cAd0d8B56F531D| Cindy| Haynes| Lozano and Sons| South Tom| Colombia| +1-500-064-0346x793| 001-504-170-7374|rubenmah
oney@eam...| 2020-07-07|http://www.fritz...|
| 17|182AF7ca332C2BA| Shawn| Wheeler| Harmon-Wallace| Kaufmanville| American Samoa| 926.256.7746x07704| +1-942-927-9013|faithfar
rell@barr...| 2022-01-01|http://gallagher-...|
| 18|7ad82F593ACFCac| Joe| Hayden| Suarez PLC| Richardland| Azerbaijan| 645.759.1156x19217|001-099-165-5270x...|sergio33
@fitzpatr...| 2020-08-19| http://holder.com/|
| 19|e08F335e7495f2| Daryl| Rodgers| Gould and Sons| East Brooke| Poland| 914.756.1802x95427| +1-802-269-2026|floresda
wn@herman...| 2021-05-24|https://www.drake...|
| 20|FECF1EFB076917E| Chris| Goodman| Adams-Robbins| Port Henryfurt| Brunei Darussalam| 048.629.0897x1593| 231-457-6144|cmontgom
ery@best.com| 2021-02-11|https://www.calla...|
```

only showing top 20 rows

thời gian chạy: 0.08145499229431152

root@881862b7a8d3:/home# |

Ảnh 4: Kết quả truy vấn exam2.py

-Thử nghiệm với một vài truy vấn khác, tại file **exam3.py** dưới đây, tôi thử nghiệm quá trình truy vấn với việc chỉ chọn 4 cột dữ liệu Index, Customer Id, First Name, Last Name và vẫn với điều kiện Index < 10. Kết quả cho thấy, thời gian chạy chỉ hết khoảng **0.13s**

```
from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql("select Index, 'Customer Id', 'First Name', 'Last Name' from my_database where Index < 10")

time_end = time.time()

result.show()

print("thời gian chạy:", time_end - time_start)
```

Ảnh 5: Nội dung exam3.py

```
root@881862b7a8d3:/home# python3 exam3.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/02 13:14:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
+-----+-----+-----+-----+
|Index| Customer Id|First Name|Last Name|
+-----+-----+-----+-----+
|1|138fB5315da5fE9| Jeanne| Ferrell|
|2|b0d61acAc72A388| Ian| Browning|
|3|1B27Ff7Fd418C89| Taylor| Martinez|
|4|eff8bbcED3eacD8| Andre| McCall|
|5|73ee6AaCAcea39C| Alyssa| Mcneil|
|6|e4A1f3FA732CED| Marisa| West|
|7|dC9dbfa601b71b5| Jaclyn| Branch|
|8|38ce8FA0e07AdEa| Billy| Stone|
|9|134e7E02Ccb2614| Cristian| Waters|
+-----+-----+-----+-----+

thời gian chạy: 0.12875723838806152
root@881862b7a8d3:/home# |
```

Ảnh 6: Kết quả truy vấn exam3.py

-Thử với một truy vấn liên quan đến **GROUP BY**, tôi truy vấn đếm số lượng người đến từ các quốc gia khác nhau từ cơ sở dữ liệu file **exam4.py** và **exam5.py** với sự khác biệt nằm ở **WHERE Index < 1000000**


```

from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql("select count(`Customer Id`), `Country` from my_database group by `Country`")

time_end = time.time()

result.show()

print("thoi gian chay:", time_end - time_start)

```

Ảnh 7: Nội dung exam4.py

```

root@881862b7a8d3:/home# python3 exam4.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/02 13:16:54 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
+-----+
|count(Customer Id)|      Country|
+-----+
|4056|Chad|
|4049|Anguilla|
|3992|Paraguay|
|4211|Macao|
|4095|Heard Island and ...|
|4008|Yemen|
|4005|Senegal|
|4094|Sweden|
|4039|Tokelau|
|4018|French Southern T...|
|4093|Kiribati|
|4134|Guyana|
|4016|Jersey|
|4130|Eritrea|
|4092|Philippines|
|4057|Norfolk Island|
|3993|Tonga|
|4098|Djibouti|
|4122|Malaysia|
|4042|Singapore|
+-----+
only showing top 20 rows
thoi gian chay: 0.12114262580871582
root@881862b7a8d3:/home#

```

Ảnh 8: Kết quả truy vấn exam4.py

```

from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql("select count(`Customer Id`), `Country` from my_database where Index < 100000 group by `Country`")

time_end = time.time()

result.show()

print("thoi gian chay:", time_end - time_start)

```

Ảnh 9: Nội dung exam5.py

```

root@881862b7a8d3:/home# python3 exam5.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/02 13:21:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
+-----+
|count(Customer Id)|      Country|
+-----+
|4056|Chad|
|4049|Anguilla|
|3992|Paraguay|
|4211|Macao|
|4095|Heard Island and ...|
|4008|Yemen|
|4005|Senegal|
|4094|Sweden|
|4039|Tokelau|
|4018|French Southern T...|
|4093|Kiribati|
|4134|Guyana|
|4016|Jersey|
|4130|Eritrea|
|4092|Philippines|
|4057|Norfolk Island|
|3993|Tonga|
|4098|Djibouti|
|4122|Malaysia|
|4042|Singapore|
+-----+
only showing top 20 rows
thoi gian chay: 0.19454007691955566
root@881862b7a8d3:/home#

```

Ảnh 10: Kết quả truy vấn exam5.py

6.2. CRUD

-Tạo bảng chứa dữ liệu tạm thời với PySpark, chúng ta sử dụng cấu trúc lệnh như sau, kết quả thu được sẽ là

```
root@881862b7a8d3: /home x + v
from pyspark.sql import SparkSession
import sqlite3
from get_session_data import *

spark = get_session()

#create dataframe
data = [(1, "Nguyen Xuan An", "21002183"), (2, "Nguyen Bao Ngoc", "@123dc5"), (3, "Le Viet Hung", "$1578@2")]
columns = ["ID", "Name", "Ma Dinh Danh"]
df = spark.createDataFrame(data, columns)
df.show()
```

Ảnh 11: Tạo dữ liệu

```
root@881862b7a8d3:/home# python3 crud.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/07 12:04:25 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
+-----+
| ID |      Name | Ma Dinh Danh |
+-----+
| 1 | Nguyen Xuan An | 21002183 |
| 2 | Nguyen Bao Ngoc | @123dc5 |
| 3 | Le Viet Hung | $1578@2 |
+-----+
```

Ảnh 12: Kết quả crud_create.py

-Đây là một bảng dữ liệu tạm thời, bây giờ để có thể lưu trữ bảng dữ liệu này trong cơ sở dữ liệu, ta cần thực hiện một số thao tác, thêm đoạn mã sau vào file code hiện tại. Thực hiện chạy file và kiểm tra bảng đã được tạo trong cơ sở dữ liệu **sample.db** chưa

```
root@881862b7a8d3: /home x + v
from pyspark.sql import SparkSession
import sqlite3
from get_session_data import *

spark = get_session()

#create dataframe
data = [(1, "Nguyen Xuan An", "21002183"), (2, "Nguyen Bao Ngoc", "@123dc5"), (3, "Le Viet Hung", "$1578@2")]
columns = ["ID", "Name", "Ma Dinh Danh"]
df = spark.createDataFrame(data, columns)
df.show()

con = sqlite3.connect("sample.db")

cur = con.cursor()

create_table = """
CREATE TABLE IF NOT EXISTS example_table(
    ID INTEGER,
    NAME VARCHAR(10) NOT NULL,
    'MA DINH DANH' VARCHAR(10) NOT NULL,
    PRIMARY KEY(ID)
);
"""

cur.execute(create_table)

for Row in df.collect():
    @@@
"crud create nv" 33L 791C 2/1 3 Ton
```

Ảnh 13: Nội dung crud_create.py

-Nhập lệnh **.tables** để hiển thị các bảng dữ liệu hiện có. Ta thấy được, bảng **example_table** đã được tạo trong **sample.db**, sử dụng lệnh **PRAGMA** để kiểm tra thông tin về đặc tính của các cột dữ liệu

```
root@881862b7a8d3:/home# vi crud_create.py
root@881862b7a8d3:/home# sqlite3 /home/sample.db
SQLite version 3.31.1 2020-01-27 19:55:54
Enter ".help" for usage hints.
sqlite> .tables
example_table  my_database
sqlite> PRAGMA table_info(ex
EXCEPT      EXCLUDE      EXCLUSIVE      EXISTS      EXPLAIN      example_table
sqlite> PRAGMA table_info(ex
EXCEPT      EXCLUDE      EXCLUSIVE      EXISTS      EXPLAIN      example_table
sqlite> PRAGMA table_info(example_table)
...> ;
0|ID|INTEGER|0||1
1|NAME|VARCHAR(10)|1||0
2|MA DINH DANH|VARCHAR(10)|1||0
sqlite> SELECT * FROM example_table;
1|Nguyen Xuan An|abc123a
2|Nguyen Bao Ngoc|@123dc5
3|Le Viet Hung|$1578@2
sqlite> |
```

Ảnh 14: Kiểm tra bảng đã được tạo

-Ta thử đọc dữ liệu từ bảng vừa tạo

```
root@881862b7a8d3:/home x + v
from pyspark.sql import SparkSession
import sqlite3
from get_session_data import *

spark = get_session()
SQL_Queries(spark, "sample.db", "example_table")

result = spark.sql("select * from example_table")

result.show
```

Ảnh 15: Nội dung crud_read.py

```
root@881862b7a8d3:/home# python3 crud_read.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/07 14:01:11 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
+-----+-----+-----+
| ID | NAME | MA DINH DANH |
+-----+-----+-----+
| 1.0000000000000000 | Nguyen Xuan An | 21002183 |
| 2.0000000000000000 | Nguyen Bao Ngoc | @123dc5 |
| 3.0000000000000000 | Le Viet Hung | $1578@2 |
+-----+-----+-----+
root@881862b7a8d3:/home# |
```

Ảnh 16: Kết quả crud_read.py

-Bây giờ, ta thử thay đổi dữ liệu trong bảng, có thể thấy **`MA DINH DANH`** của **ID = 1** đã được thay đổi

```

from pyspark.sql import SparkSession
import sqlite3
from get_session_data import *

spark = get_session()
SQL_Queries(spark,"sample.db","example_table")
result = spark.sql("select * from example_table")
result.show()

con = sqlite3.connect("sample.db")
cur = con.cursor()

cur.execute("""UPDATE example_table SET 'MA DINH DANH' = 'abc123a' WHERE ID = 1""")
con.commit()
con.close()
SQL_Queries(spark,"sample.db","example_table")
result = spark.sql("select * from example_table")
result.show()

```

Ảnh 17: Nội dung crud_update.py

```

root@881862b7a8d3:/home# vi crud_update.py
root@881862b7a8d3:/home# python3 crud_update.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/07 15:17:02 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
+----+-----+-----+
| ID|      NAME|MA DINH DANH|
+----+-----+-----+
| 1| Nguyen Xuan An| 21002183|
| 2| Nguyen Bao Ngoc| @123dc5|
| 3| Le Viet Hung| $1578@2|
+----+-----+-----+

+----+-----+-----+
| ID|      NAME|MA DINH DANH|
+----+-----+-----+
| 1| Nguyen Xuan An| abc123a|
| 2| Nguyen Bao Ngoc| @123dc5|
| 3| Le Viet Hung| $1578@2|
+----+-----+-----+

```

Ảnh 18: Kết quả crud_update.py

-Cuối cùng, ta thử nghiệm xóa hàng **ID = 2** trong bảng

```

root@881862b7a8d3:/home # vi crud_delete.py
from pyspark.sql import SparkSession
import sqlite3
from get_session_data import *

spark = get_session()
SQL_Queries(spark,"sample.db","example_table")
result = spark.sql("select * from example_table")
result.show()

con = sqlite3.connect("sample.db")
cur = con.cursor()

cur.execute("""DELETE FROM example_table WHERE ID = 2""")
con.commit()
con.close()
SQL_Queries(spark,"sample.db","example_table")
result = spark.sql("select * from example_table")
result.show()

```

Ảnh 19: Nội dung crud_delete.py

```

root@881862b7a8d3:/home# python3 crud_delete.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/07 15:24:29 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
+----+-----+-----+
| ID|      NAME|MA DINH DANH|
+----+-----+-----+
| 1| Nguyen Xuan An| abc123a|
| 2| Nguyen Bao Ngoc| @123dc5|
| 3| Le Viet Hung| $1578@2|
+----+-----+-----+

+----+-----+-----+
| ID|      NAME|MA DINH DANH|
+----+-----+-----+
| 1| Nguyen Xuan An| abc123a|
| 3| Le Viet Hung| $1578@2|
+----+-----+-----+
root@881862b7a8d3:/home#

```

Ảnh 20: Kết quả crud_delete.py

6.3. Kiểm tra hiệu suất của WHERE với một số câu truy vấn

-Trong nội dung file **exam6.py** dưới đây ta cho đếm số "**Customer Id**" theo "**Country**" với các điều kiện "**First Name**" có kết thúc bằng kí tự 'a' và "**Last Name**" có kết thúc bằng kí tự 'b'. Cùng với đó, ta tạo một file **exam7.py** có nội dung gần tương tự khi thêm một điều kiện **WHERE** nữa là **Index** phải chia hết cho 2. Tốc độ truy vấn của 2 file lần lượt là **0.173s** và **0.241s**. Điều này cho thấy điều kiện liên quan đến tính toán ảnh hưởng khá lớn tới hiệu suất truy vấn khi tốc độ truy vấn tăng tới **39%**

```
root@881862b7a8d3: /home  x  +  v
from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql("select count('Customer Id'),'Country' from my_database where 'First Name' like '%A' and 'Last Name'
like 'B%' group by 'Country'")

time_end = time.time()

result.show()

print("thoi gian chay:",time_end - time_start)
~
~
~
```

Ảnh 21: Nội dung exam6.py

```
asses where applicable
+-----+
|count(Customer Id)|Country|
+-----+
|62|Chad|
|70|Anguilla|
|52|Paraguay|
|59|Macao|
|63|Heard Island and ...|
|66|Yemen|
|76|Senegal|
|71|Tokelau|
|81|Sweden|
|69|French Southern T...|
|82|Kiribati|
|68|Guyana|
|80|Eritrea|
|63|Philippines|
|71|Jersey|
|62|Djibouti|
|64|Norfolk Island|
|72|Tonga|
|70|Singapore|
|68|Malaysia|
+-----+
only showing top 20 rows
thoi gian chay: 0.17397785186767578
```

Ảnh 22: Kết quả exam6.py

```
from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql("select count('Customer Id'),'Country' from my_database where 'First Name' like '%A' and 'Last Name'
like 'B%' and Index % 2 = 0 group by 'Country'")

time_end = time.time()

result.show()

print("thoi gian chay:",time_end - time_start)
~
~
~
```

Ảnh 23: Nội dung exam7.py

```

+-----+-----+
|count(Customer Id)|Country|
+-----+-----+
|24|Chad|
|34|Anguilla|
|27|Paraguay|
|32|Macao|
|28|Heard Island and ...|
|31|Yemen|
|34|Senegal|
|35|Sweden|
|35|Tokelau|
|29|French Southern T...|
|44|Kiribati|
|33|Guyana|
|41|Eritrea|
|32|Philippines|
|39|Jersey|
|38|Djibouti|
|28|Norfolk Island|
|36|Tonga|
|46|Singapore|
|37|Malaysia|
+-----+-----+
only showing top 20 rows
thoi gian chay: 0.24164772033691406
root@881862b7a8d3:/home#

```

Ảnh 24: Kết quả exam7.py

-Tiếp theo, ta kết hợp với một số điều kiện khác như **HAVING** và **ORDER BY** trong 2 file **exam8.py**, **exam9.py** thử so sánh tốc độ truy vấn với một số điều kiện khác biệt, cụ thể ta xây dựng điều kiện **WHERE** của 2 file có chút thay đổi, **exam8.py** sẽ dùng điều kiện liên quan đến kí tự, còn **exam9.py** sẽ dùng điều kiện liên quan đến tính toán

```

from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql(
    """select count('Customer Id') as Number_Person, 'Country'
    from my_database
    where 'First Name' like '%A' and 'Last Name' like 'B%'
    group by 'Country'
    having count('Customer Id') > 50
    order by Number_Person desc
    """
)

time_end = time.time()

result.show()

print("thoi gian chay:", time_end - time_start)
~
~

```

Ảnh 25: Nội dung exam8.py

```

asses where applicable
+-----+-----+
|Number_Person|          Country|
+-----+-----+
|          143|          Congo|
|          132|          Korea|
|           95|    Faroe Islands|
|           88|Sao Tome and Prin...|
|           88|          Oman|
|           86|          Chile|
|           86|    French Polynesia|
|           85|          Rwanda|
|           85|          Benin|
|           85|    Christmas Island|
|           85|          Guam|
|           85|        Australia|
|           84|        Botswana|
|           83|        Reunion|
|           83|        Nigeria|
|           83|Svalbard & Jan Ma...|
|           83|        Greenland|
|           83|        Costa Rica|
|           82|        Kiribati|
|           82|        Slovenia|
+-----+-----+
only showing top 20 rows

thoi gian chay: 0.20789504051208496
root@881862b7a8d3:/home#

```

Ảnh 26: Kết quả exam8.py

-Nội dung của file **exam8.py** sẽ sử dụng các điều kiện **having count(Customer Id) > 50 order by desc** sắp xếp các giá trị từ lớn đến bé. Còn file **exam9.py** sẽ thay thế điều kiện **where 'Last Name' like 'B%'** bằng điều kiện **Index % 2 = 0**. Có thể thấy, khi sử dụng điều kiện liên quan đến tính toán, tốc độ truy vấn tăng đáng kể. Với thời gian **0.207s** của **exam8.py** và **0.234s** của **exam9.py**, tốc độ truy vấn tăng khoảng **18%**

```

from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql(
    """select count('Customer Id') as Number_Person, 'Country'
    from my_database
    where 'First Name' like '%A' and Index % 2 = 0
    group by 'Country'
    having count('Customer Id') > 50
    order by Number_Person desc
    """
)

time_end = time.time()

result.show()

print("thoi gian chay:", time_end - time_start)
~
~

```

Ảnh 27: Nội dung exam9.py


```

24/06/07 19:42:29 WARN NativeCodeLoader: Unable to load native-hadoop lib
asses where applicable
+-----+-----+
|Number_Person|          Country|
+-----+-----+
|          711|          Congo|
|          702|          Korea|
|          411|        Vanuatu|
|          410|         Iceland|
|          408|          Zambia|
|          407|          Qatar|
|          405|Saint Pierre and ...|
|          404|          France|
|          402|Lao People's Demo...|
|          398|        Faroe Islands|
|          395|          Grenada|
|          394|          Namibia|
|          394|          Guyana|
|          392|United States Vir...|
|          392|          Greenland|
|          391|          Kiribati|
|          391|          Zimbabwe|
|          390|Cocos (Keeling) I...|
|          390|          Sri Lanka|
|          389|          Armenia|
+-----+-----+
only showing top 20 rows

thoi gian chay: 0.23412799835205078
root@881862b7a8d3:/home#

```

Ảnh 28: Kết quả exam9.py

-Nội dung 2 file **exam10.py** và **exam11.py** được thay đổi để xem nếu một vài điều kiện khác thay đổi, có ảnh hưởng nhiều đến hiệu suất truy vấn khi cùng một điều kiện **WHERE** không. Ở đây, ta lần lượt sử dụng điều kiện **HAVING** đếm số lượng người phù hợp **>50** và **>40**. Khi chạy thử cho thấy kết quả truy vấn trả về lần lượt là **0.273s** và **0.283s**. Sự khác biệt này là không đáng kể, chỉ khoảng **3%**

```

root@881862b7a8d3:/home x + v
from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql(
    """select sum(Index) as result, 'Country'
    from my_database
    where 'First Name' like '%A' and 'Last Name' like '%B%' and Index % 2 = 0
    group by 'Country'
    having count('Customer Id') > 50
    order by result desc
    """
)

time_end = time.time()

result.show()

print("thoi gian chay:", time_end - time_start)

```

Ảnh 29: Nội dung exam 10.py

```

root@881862b7a8d3:/home# vi exam10.py
root@881862b7a8d3:/home# python3 exam10.py
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/06/07 19:56:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable

+-----+-----+
|      result|Country|
+-----+-----+
|4.2469534E7|Congo|
|3.1079694E7|Korea|
+-----+-----+

thoi gian chay: 0.2736630439758301
root@881862b7a8d3:/home# vi exam10.py

```

Ảnh 10: Kết quả exam10.py

```

from get_session_data import *
import time

spark = get_session()

SQL_Queries(spark, "/home/sample.db", "my_database")

time_start = time.time()

result = spark.sql(
    """select sum('Index') as result, 'Country'
    from my_database
    where 'First Name' like '%A' and 'Last Name' like 'B%' and Index % 2 = 0
    group by 'Country'
    having count('Customer Id') > 40
    order by result desc
    """
)

time_end = time.time()

result.show()

print("thoi gian chay:", time_end - time_start)
~
~

```

Ảnh 11: Nội dung exam11.py

```

asses where applicable

+-----+-----+
|      result|Country|
+-----+-----+
|4.2469534E7|Congo|
|3.1079694E7|Korea|
|2.7503606E7|Niger|
|2.6751912E7|French Polynesia|
|2.5911066E7|Turkey|
|2.5751492E7|Faroe Islands|
|2.5471324E7|Guinea-Bissau|
| 2.516415E7|China|
| 2.465517E7|Mozambique|
|2.4346786E7|Singapore|
|2.4260788E7|Greenland|
|2.4255108E7|Kiribati|
| 2.388116E7|Vanuatu|
|2.3579772E7|Guam|
| 2.348073E7|Comoros|
|2.3412696E7|Svalbard & Jan Ma...|
|2.3314956E7|Togo|
|2.3184804E7|Iraq|
|2.3069556E7|Martinique|
|2.2991596E7|Ethiopia|
+-----+-----+

only showing top 20 rows

thoi gian chay: 0.2834932804107666
root@881862b7a8d3:/home# vi exam11.py
root@881862b7a8d3:/home# |

```

Ảnh 12: Kết quả exam11.py

6.4. Kết luận

-Càng nhiều điều kiện trong **WHERE** thì càng làm ảnh hưởng tới tốc độ truy vấn dữ liệu.

-Các điều kiện trong **WHERE** có liên quan đến tính toán thường có ảnh hưởng lớn tới tốc độ truy vấn.

-Việc **SELECT** cũng làm ảnh hưởng tới tốc độ truy vấn khi có **WHERE**, vậy nên cần lựa chọn các dữ liệu cần thiết để truy vấn.

-Các điều kiện tính toán như **COUNT**, **SUM** với **GROUP BY** cũng gây ra ảnh hưởng khi truy vấn với **WHERE**, ta thấy khi dùng **SUM**, tốc độ truy vấn tăng đáng kể.

-Điều kiện **HAVING**, **ORDER BY** hoạt động ảnh hưởng khá nhỏ tới hiệu suất khi sử dụng **WHERE**