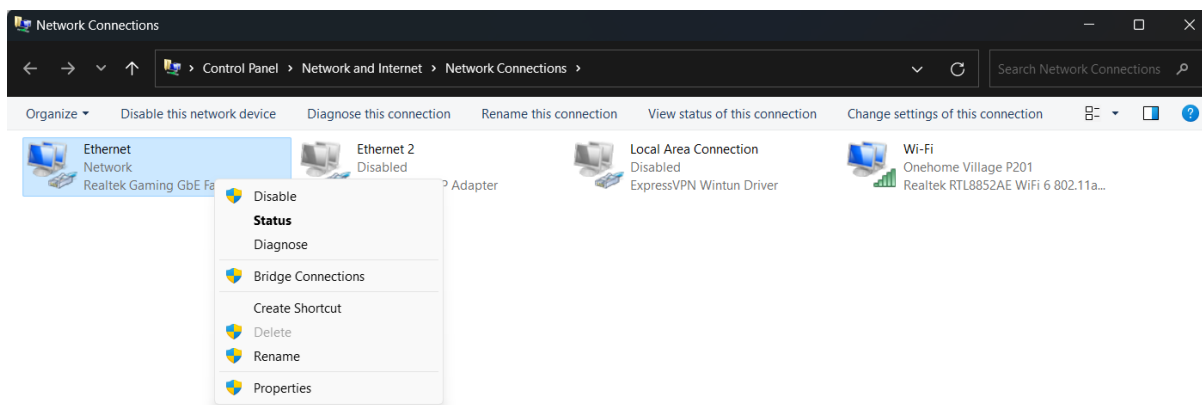


THIẾT LẬP STANDALONE CLUSTER APACHE SPARK THÔNG QUA DÂY MẠNG LAN TRÊN WINDOWS

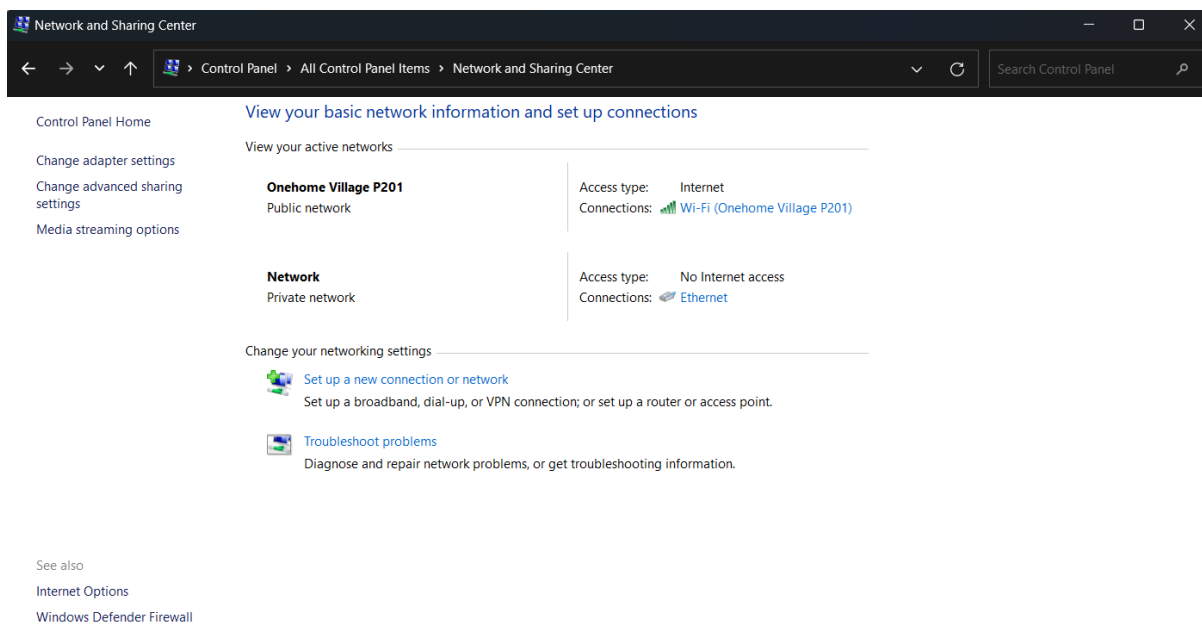
1. Thiết Lập Mạng LAN:

Bước 1: Việc đầu tiên chuẩn bị một đoạn dây cáp LAN. Cắm dây vào cổng kết nối của mỗi máy tính muốn thực hiện kết nối.

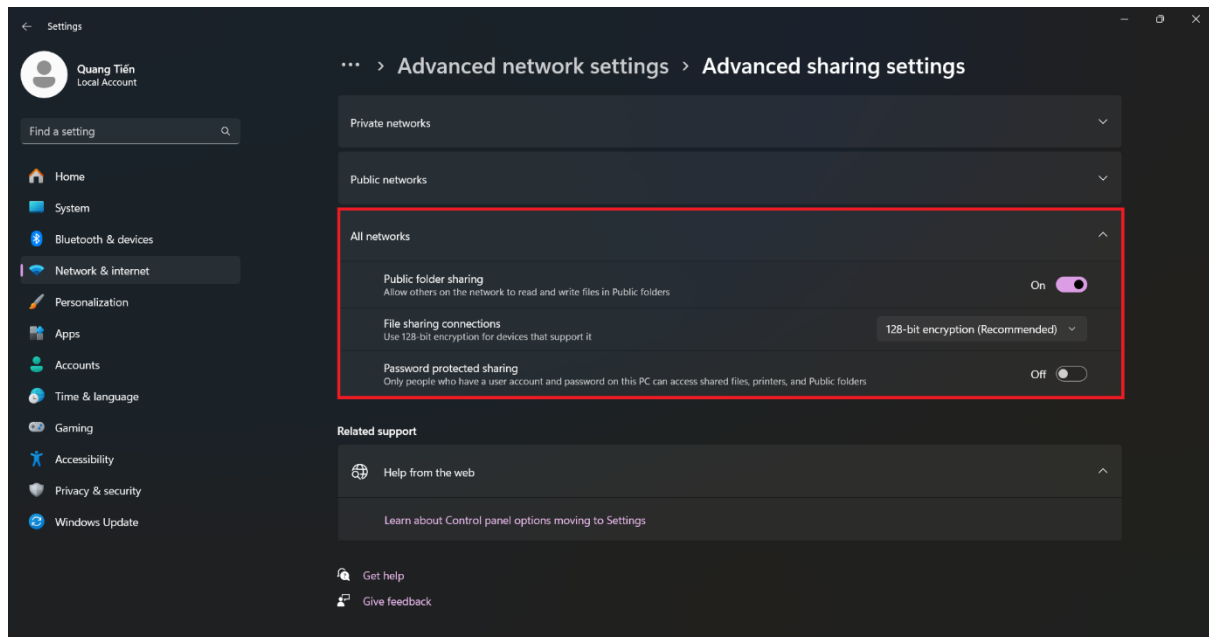
Bước 2: Tại tìm kiếm, gõ “View Network Connections”. Kiểm tra Ethernet có được bật không. Nếu không được bật thì tiến hành bật bằng cách nhấp chuột phải chọn “Enable”.



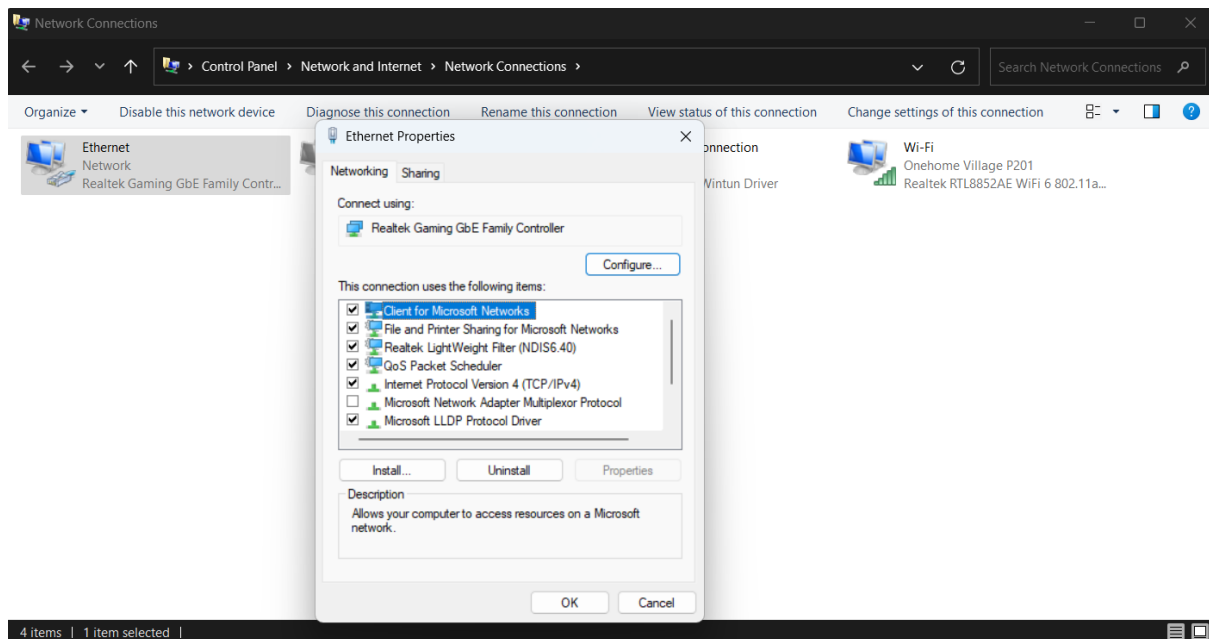
Bước 3: Tại tìm kiếm, gõ Control Panel, vào Control Panel > Chọn mục Network and Sharing Center.



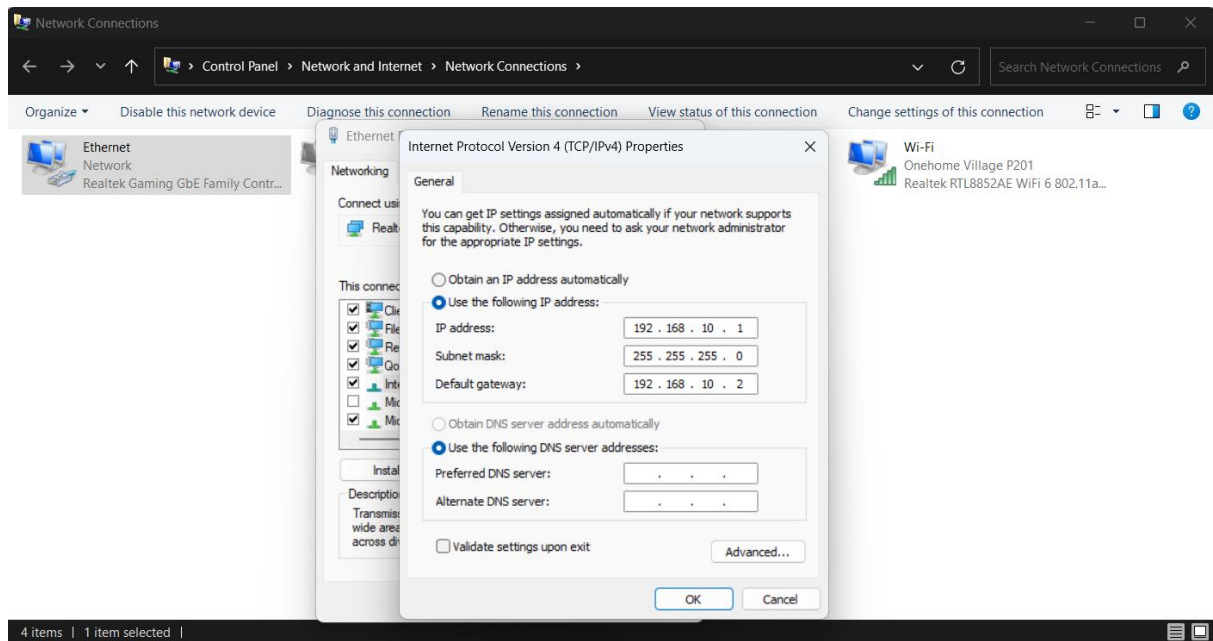
Bước 4: Tùy chỉnh Change advanced sharing settings > Kéo xuống tìm All Networks > Thiết lập như hình bên dưới.



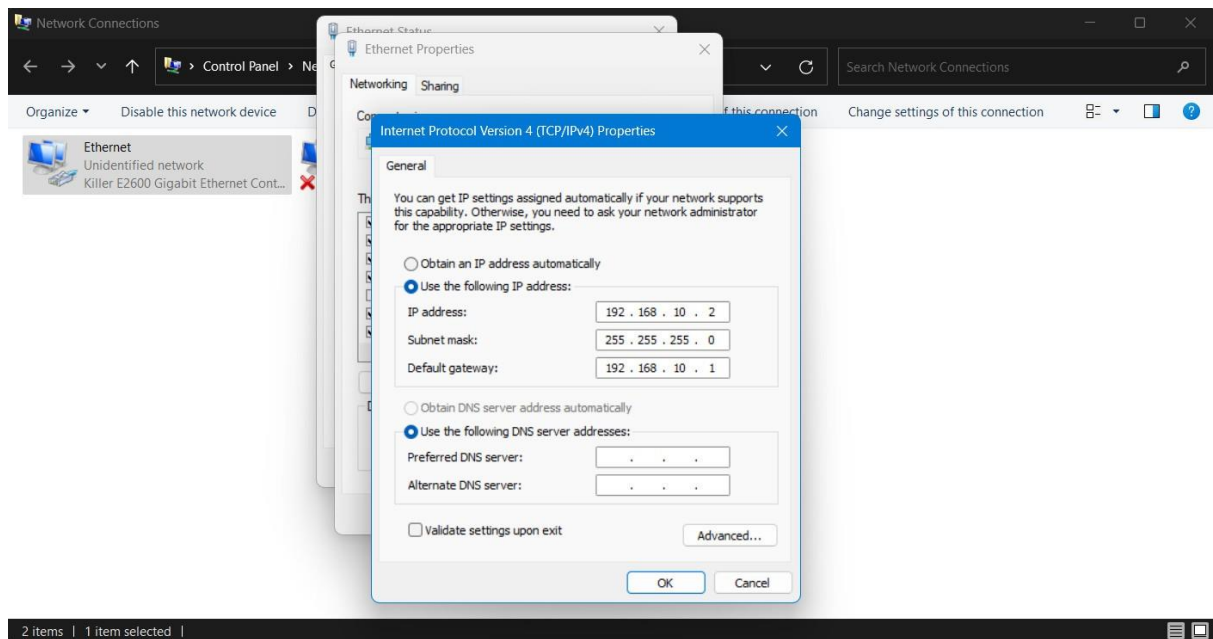
Bước 5: Vào “View Network Connections”. Nhấn vào mạng LAN cần thiết lập > Vào Properties.



Bước 6: Nhấp chọn mục Internet Protocol Version 4 (TCP/IPv4). Thiết lập IP cho mạng LAN.



Bước 7: Thiết lập máy còn lại như các bước trên.



Chú ý: Việc giữ ba số đầu tiên giống nhau cho cả hai máy tính là cần thiết để chúng có thể giao tiếp với nhau trên cùng một mạng LAN. Nếu chúng không có ba số đầu tiên giống nhau, chúng sẽ không nằm trong cùng một mạng LAN và sẽ không thể giao tiếp trực tiếp với nhau trên mạng cục bộ (LAN).

2. Kiểm Tra Kết Nối Mạng:

Trên cả hai máy tính, mở CMD. Thực hiện lệnh ping và xem kết quả trả về từ cả hai lệnh ping sẽ cho biết liệu hai máy tính có thể giao tiếp với nhau thông qua mạng hay không.

```
C:\WINDOWS\system32\cmd. x + v
Microsoft Windows [Version 10.0.22631.3296]
(c) Microsoft Corporation. All rights reserved.

C:\Users\min>ping 192.168.10.2

Pinging 192.168.10.2 with 32 bytes of data:
Reply from 192.168.10.2: bytes=32 time=1ms TTL=128
Reply from 192.168.10.2: bytes=32 time=1ms TTL=128
Reply from 192.168.10.2: bytes=32 time=1ms TTL=128
Reply from 192.168.10.2: bytes=32 time=1ms TTL=128

Ping statistics for 192.168.10.2:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
    Approximate round trip times in milli-seconds:
        Minimum = 1ms, Maximum = 1ms, Average = 1ms
```

```
C:\WINDOWS\system32\cmd. x + v
Microsoft Windows [Version 10.0.22631.3296]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Admin>ping 192.168.10.1

Pinging 192.168.10.1 with 32 bytes of data:
Reply from 192.168.10.1: bytes=32 time=1ms TTL=128
Reply from 192.168.10.1: bytes=32 time=1ms TTL=128
Reply from 192.168.10.1: bytes=32 time=1ms TTL=128
Reply from 192.168.10.1: bytes=32 time=1ms TTL=128

Ping statistics for 192.168.10.1:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
    Approximate round trip times in milli-seconds:
        Minimum = 1ms, Maximum = 1ms, Average = 1ms
```

Chú ý: Lệnh này thực hiện sau khi sau chạy Cluster để kiểm tra kết nối của cả 02 máy với nhau. Trên cả hai máy tính, mở Windows PowerShell. Thực hiện lệnh “Test-NetConnection -ComputerName <IP> -Port <Port>” và xem kết quả trả về từ cả hai lệnh sẽ cho biết liệu hai máy tính có thể giao tiếp với nhau thông qua mạng hay không.

```
Windows PowerShell x + v
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\min> Test-NetConnection -ComputerName 192.168.10.2 -Port 8081

ComputerName      : 192.168.10.2
RemoteAddress     : 192.168.10.2
RemotePort        : 8081
InterfaceAlias    : Ethernet
SourceAddress     : 192.168.10.1
TcpTestSucceeded  : True
```

```
Windows PowerShell x + v
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\Admin> Test-NetConnection -ComputerName 192.168.10.1 -Port 7077

ComputerName      : 192.168.10.1
RemoteAddress     : 192.168.10.1
RemotePort        : 7077
InterfaceAlias    : Ethernet
SourceAddress     : 192.168.10.2
TcpTestSucceeded  : True
```

3. Chạy Cluster:

Bước 1: Máy Master. Vào CMD chạy lệnh “cd %SPARK_HOME%” và “spark-class org.apache.spark.deploy.master.Master”.

```
C:\WINDOWS\system32\cmd. X + v
Microsoft Windows [Version 10.0.22631.3296]
(c) Microsoft Corporation. All rights reserved.

C:\Users\min>cd %SPARK_HOME%

C:\SPARK>spark-class org.apache.spark.deploy.master.Master
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
24/03/19 00:27:27 INFO Master: Started daemon with process name: 24332@QuangTien
24/03/19 00:27:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
24/03/19 00:27:32 INFO SecurityManager: Changing view acls to: min
24/03/19 00:27:32 INFO SecurityManager: Changing modify acls to: min
24/03/19 00:27:32 INFO SecurityManager: Changing view acls groups to:
24/03/19 00:27:32 INFO SecurityManager: Changing modify acls groups to:
24/03/19 00:27:32 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view perm
issions: min; groups with view permissions: EMPTY; users with modify permissions: min; groups with modify permissions: E
MPTY
24/03/19 00:27:33 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
24/03/19 00:27:33 INFO Master: Starting Spark master at spark://192.168.10.1:7077
24/03/19 00:27:33 INFO Master: Running Spark version 3.5.1
24/03/19 00:27:33 INFO JettyUtils: Start Jetty 0.0.0.0:8080 for MasterUI
24/03/19 00:27:33 INFO Utils: Successfully started service 'MasterUI' on port 8080.
24/03/19 00:27:33 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://QuangTien:8080
24/03/19 00:27:33 INFO Master: I have been elected leader! New state: ALIVE
```

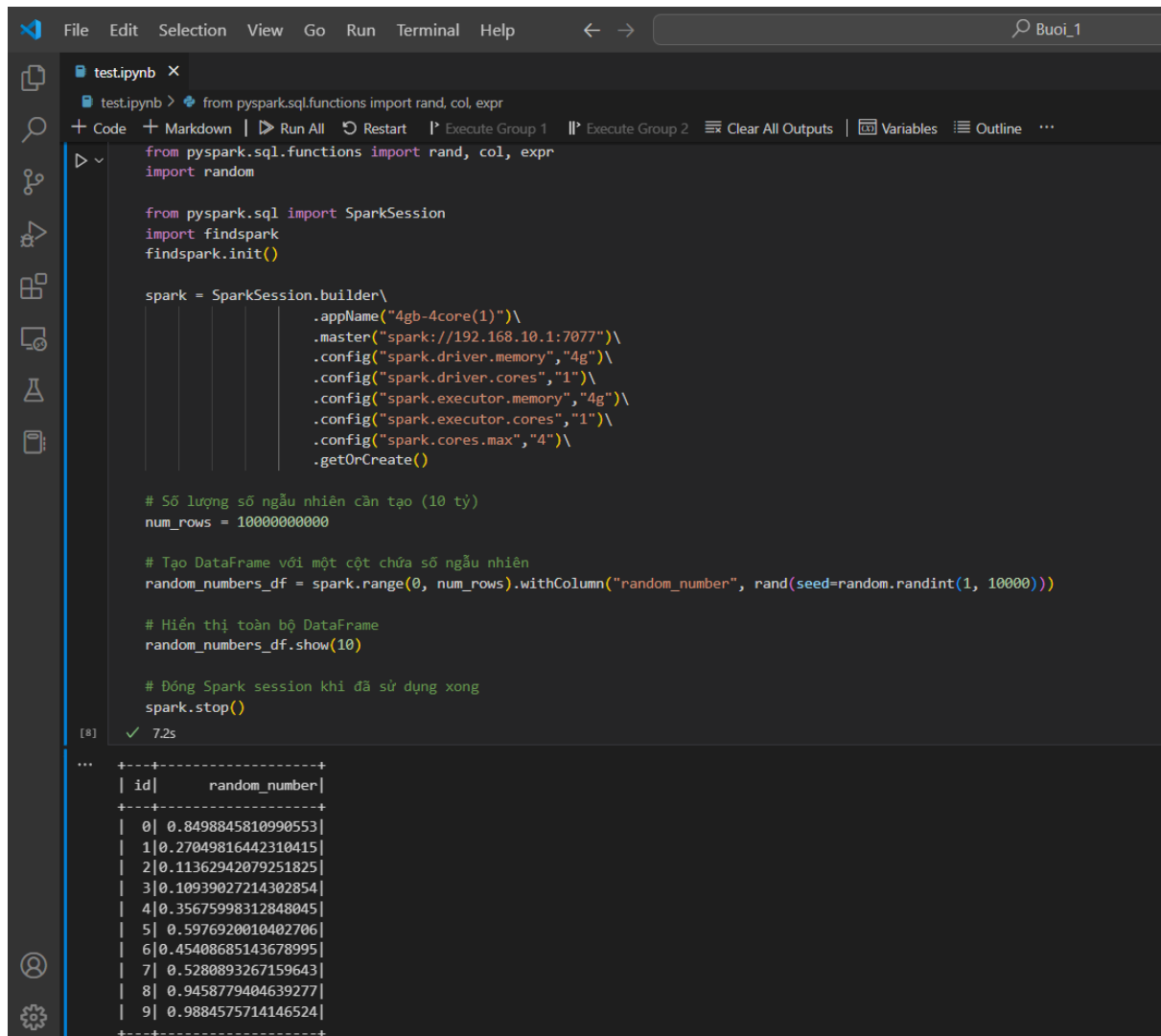
Bước 2: Máy Worker. Vào CMD chạy lệnh “cd %SPARK_HOME%” và “spark-class org.apache.spark.deploy.worker.Worker spark://<MASTER-IP>:7077”.

```
C:\WINDOWS\system32\cmd. X + v
Microsoft Windows [Version 10.0.22631.3296]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Admin>cd %SPARK_HOME%

C:\SPARK>spark-class org.apache.spark.deploy.worker.Worker spark://192.168.10.1:7077
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties
24/03/19 00:38:33 INFO Worker: Started daemon with process name: 14864@LAPTOP-MA41906E
24/03/19 00:38:33 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
24/03/19 00:38:33 INFO SecurityManager: Changing view acls to: Admin
24/03/19 00:38:33 INFO SecurityManager: Changing modify acls to: Admin
24/03/19 00:38:33 INFO SecurityManager: Changing view acls groups to:
24/03/19 00:38:33 INFO SecurityManager: Changing modify acls groups to:
24/03/19 00:38:33 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view perm
issions: Admin; groups with view permissions: EMPTY; users with modify permissions: Admin; groups with modify permission
s: EMPTY
24/03/19 00:38:34 INFO Utils: Successfully started service 'sparkWorker' on port 57374.
24/03/19 00:38:34 INFO Worker: Worker decommissioning not enabled.
24/03/19 00:38:34 INFO Worker: Starting Spark worker 192.168.10.2:57374 with 8 cores, 30.8 GiB RAM
24/03/19 00:38:34 INFO Worker: Running Spark version 3.5.1
24/03/19 00:38:34 INFO Worker: Spark home: C:\SPARK
24/03/19 00:38:34 INFO ResourceUtils: =====
24/03/19 00:38:34 INFO ResourceUtils: No custom resources configured for spark.worker.
24/03/19 00:38:34 INFO ResourceUtils: =====
24/03/19 00:38:34 INFO JettyUtils: Start Jetty 0.0.0.0:8081 for WorkerUI
24/03/19 00:38:34 INFO Utils: Successfully started service 'WorkerUI' on port 8081.
24/03/19 00:38:34 INFO WorkerWebUI: Bound WorkerWebUI to 0.0.0.0, and started at http://LAPTOP-MA41906E:8081
24/03/19 00:38:34 INFO Worker: Connecting to master 192.168.10.1:7077...
24/03/19 00:38:34 INFO TransportClientFactory: Successfully created connection to /192.168.10.1:7077 after 40 ms (0 ms s
pent in bootstraps)
24/03/19 00:38:34 INFO Worker: Successfully registered with master spark://192.168.10.1:7077
```

4. Chạy Ứng Dụng Trong Cluster:



```
testipyb > from pyspark.sql.functions import rand, col, expr
+ Code + Markdown | ▶ Run All ⏮ Restart ▶ Execute Group 1 ||▶ Execute Group 2 ⌵ Clear All Outputs | 📄 Variables 📖 Outline ...

from pyspark.sql.functions import rand, col, expr
import random

from pyspark.sql import SparkSession
import findspark
findspark.init()

spark = SparkSession.builder\
    .appName("4gb-4core(1)")\
    .master("spark://192.168.10.1:7077")\
    .config("spark.driver.memory", "4g")\
    .config("spark.driver.cores", "1")\
    .config("spark.executor.memory", "4g")\
    .config("spark.executor.cores", "1")\
    .config("spark.cores.max", "4")\
    .getOrCreate()

# Số lượng số ngẫu nhiên cần tạo (10 tỷ)
num_rows = 10000000000

# Tạo DataFrame với một cột chứa số ngẫu nhiên
random_numbers_df = spark.range(0, num_rows).withColumn("random_number", rand(seed=random.randint(1, 10000)))

# Hiển thị toàn bộ DataFrame
random_numbers_df.show(10)

# Đóng Spark session khi đã sử dụng xong
spark.stop()

[8] ✓ 7.2s

... +---+-----+
| id|      random_number|
+---+-----+
| 0| 0.8498845810990553|
| 1| 0.27049816442310415|
| 2| 0.11362942079251825|
| 3| 0.10939027214302854|
| 4| 0.35675998312848045|
| 5| 0.5976920010402706|
| 6| 0.45408685143678995|
| 7| 0.5280893267159643|
| 8| 0.9458779404639277|
| 9| 0.9884575714146524|
+---+-----+
```

test.ipynb

test.ipynb > from pyspark.sql.functions import rand, col, expr

+ Code + Markdown | ▶ Run All ↺ Restart | ▶ Execute Group 1 | ▶ Execute Group 2 | Clear All Outputs | Variables | Outline | ...

▶

```
from pyspark.sql.functions import rand, col, expr
import random

from pyspark.sql import SparkSession
import findspark
findspark.init()

spark = SparkSession.builder\
    .appName("8gb-6core(2)")\
    .master("spark://192.168.10.1:7077")\
    .config("spark.driver.memory", "8g")\
    .config("spark.driver.cores", "3")\
    .config("spark.executor.memory", "8g")\
    .config("spark.executor.cores", "2")\
    .config("spark.cores.max", "6")\
    .getOrCreate()

# Số lượng số ngẫu nhiên cần tạo (10 tỷ)
num_rows = 10000000000

# Tạo DataFrame với một cột chứa số ngẫu nhiên
random_numbers_df = spark.range(0, num_rows).withColumn("random_number", rand(seed=random.randint(1, 10000)))


# Hiển thị toàn bộ DataFrame
random_numbers_df.show(10)

# Đóng Spark session khi đã sử dụng xong
spark.stop()
```

[9] ✓ 6.2s

+---+-----+
| id | random_number |
+---+-----+
0	0.9882500222775016
1	0.131965678069781
2	0.5507026934681072
3	0.14582958693474013
4	0.8982143212167923
5	0.37963781836072363
6	0.6796873241382057
7	0.07327605764474843
8	0.20464584040564704
9	0.2937267150030958
+---+-----+

← → ↻ ⚠ Not secure quangtien.0080 ☆ 🚫 ⋮

 3.5.1

Spark Master at spark://192.168.10.1:7077

URL: spark://192.168.10.1:7077

Alive Workers: 1

Cores in use: 8 Total, 0 Used

Memory in use: 30.8 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 Running, 2 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20240318231235-192.168.10.2-56885	192.168.10.2:56885	ALIVE	8 (0 Used)	30.8 GiB (0.0 B Used)	

Running Applications (0)

Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20240318232000-0001	8gb-6core(2)	6	8.0 GiB		2024/03/18 23:20:00	min	FINISHED	6 s
app-20240318231952-0000	4gb-4core(1)	4	4.0 GiB		2024/03/18 23:19:52	min	FINISHED	7 s