

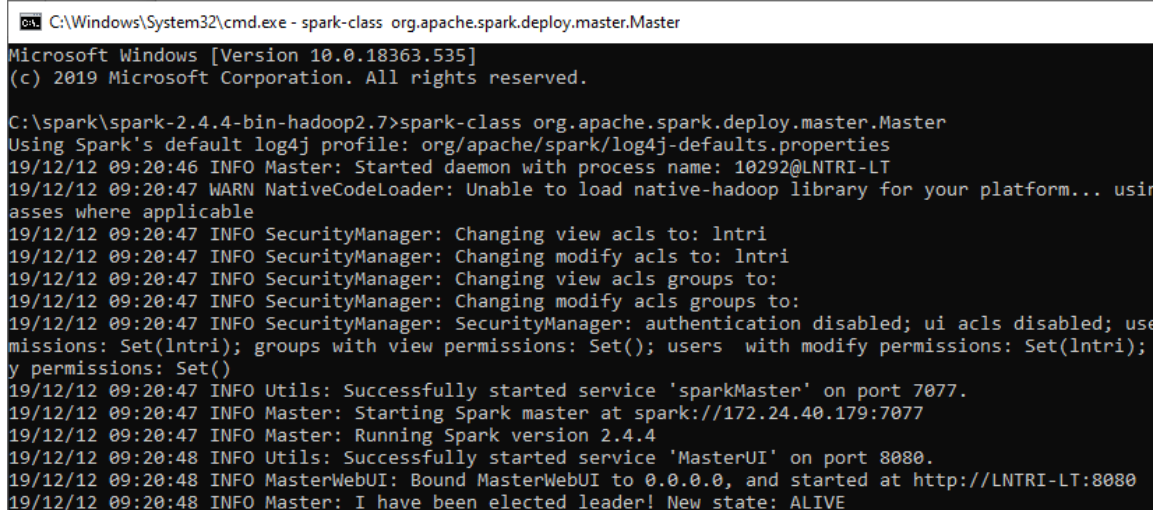
HƯỚNG DẪN CẤU HÌNH APACHE SPARK STANDALONE CLUSTER KẾT HỢP VỚI JUPYTER

Yêu cầu: Đã cài đặt PySpark thành công.

1. THỰC THI MASTER SERVER

- Mở cửa sổ **cmd**, chuyển đường dẫn đến thư mục bin của spark, sau đó thực hiện lệnh sau:

```
spark-class org.apache.spark.deploy.master.Master
```



```
C:\Windows\System32\cmd.exe - spark-class org.apache.spark.deploy.master.Master
Microsoft Windows [Version 10.0.18363.535]
(c) 2019 Microsoft Corporation. All rights reserved.

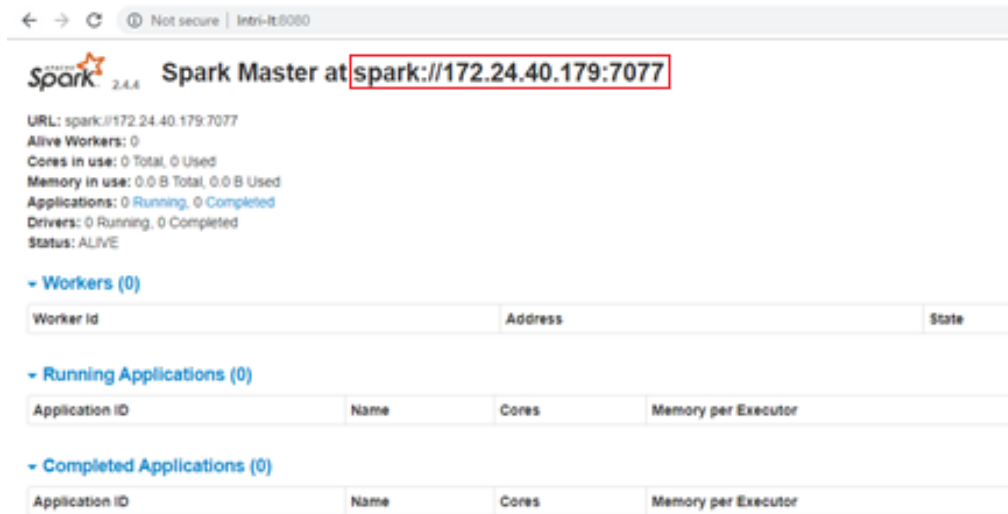
C:\spark\spark-2.4.4-bin-hadoop2.7>spark-class org.apache.spark.deploy.master.Master
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/12/12 09:20:46 INFO Master: Started daemon with process name: 10292@LNTRI-LT
19/12/12 09:20:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
  asses where applicable
19/12/12 09:20:47 INFO SecurityManager: Changing view acls to: lntri
19/12/12 09:20:47 INFO SecurityManager: Changing modify acls to: lntri
19/12/12 09:20:47 INFO SecurityManager: Changing view acls groups to:
19/12/12 09:20:47 INFO SecurityManager: Changing modify acls groups to:
19/12/12 09:20:47 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; use
missions: Set(lntri); groups with view permissions: Set(); users with modify permissions: Set(lntri);
y permissions: Set()
19/12/12 09:20:47 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
19/12/12 09:20:47 INFO Master: Starting Spark master at spark://172.24.40.179:7077
19/12/12 09:20:47 INFO Master: Running Spark version 2.4.4
19/12/12 09:20:48 INFO Utils: Successfully started service 'MasterUI' on port 8080.
19/12/12 09:20:48 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://LNTRI-LT:8080
19/12/12 09:20:48 INFO Master: I have been elected leader! New state: ALIVE
```

- Quan sát trên cửa sổ cmd, tìm url có cấu trúc sau:

http://<tên máy>:8080

Ví dụ ở đây là: <http://LNTRI-LT:8080>

- Nhập url trên vào trình duyệt:



Ghi chú: **spark://<IP>:7077** là thông tin sẽ cung cấp cho Slave để kết nối đến Master

2. KẾT NỐI CÁC MÁY SLAVE ĐẾN MASTER

- Mở cửa sổ cmd, chuyển đường dẫn đến thư mục bin của spark, sau đó thực hiện lệnh sau:

```
spark-class org.apache.spark.deploy.worker.Worker spark://ip:port
```

Trong đó: **ip:port** được cung cấp từ Master Server

```
C:\Windows\System32\cmd.exe - spark-class org.apache.spark.deploy.worker.Worker spark://172.24.40.179:7077
Microsoft Windows [Version 10.0.18362.239]
(c) 2019 Microsoft Corporation. All rights reserved.

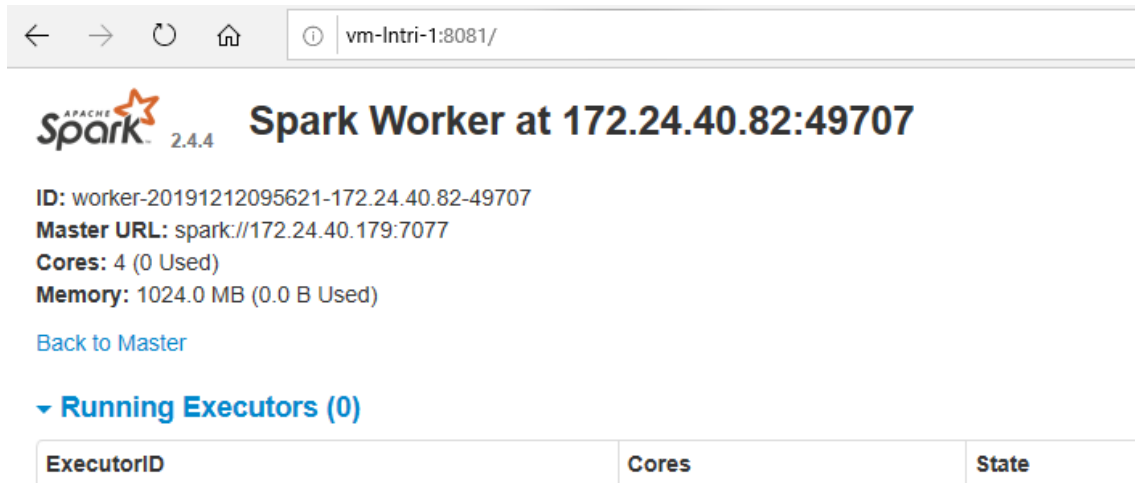
C:\spark\spark-2.4.4-bin-hadoop2.7\bin>spark-class org.apache.spark.deploy.worker.Worker spark://172.24.40.179:7077
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/12/12 09:56:18 INFO Worker: Started daemon with process name: 6688@VM-LNTRI-1
19/12/12 09:56:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
19/12/12 09:56:19 INFO SecurityManager: Changing view acls to: lntri
19/12/12 09:56:19 INFO SecurityManager: Changing modify acls to: lntri
19/12/12 09:56:19 INFO SecurityManager: Changing view acls groups to:
19/12/12 09:56:19 INFO SecurityManager: Changing modify acls groups to:
19/12/12 09:56:19 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view per
missions: Set(lntri); groups with view permissions: Set(); users with modify permissions: Set(lntri); groups with modif
y permissions: Set()
19/12/12 09:56:21 INFO Utils: Successfully started service 'sparkWorker' on port 49707.
19/12/12 09:56:21 INFO Worker: Starting Spark worker 172.24.40.82:49707 with 4 cores, 1024.0 MB RAM
19/12/12 09:56:21 INFO Worker: Running Spark version 2.4.4
19/12/12 09:56:21 INFO Worker: Spark home: C:\spark\spark-2.4.4-bin-hadoop2.7
19/12/12 09:56:22 INFO Utils: Successfully started service 'WorkerUI' on port 8081.
19/12/12 09:56:22 INFO WorkerWebUI: Bound WorkerWebUI to 0.0.0.0, and started at http://VM-LNTRI-1:8081
19/12/12 09:56:22 INFO Worker: Connecting to master 172.24.40.179:7077...
19/12/12 09:56:22 INFO TransportClientFactory: Successfully created connection to /172.24.40.179:7077 after 109 ms (0 ms
spent in bootstraps)
19/12/12 09:56:23 INFO Worker: Successfully registered with master spark://172.24.40.179:7077
```

- Quan sát trên cửa sổ cmd, tìm url có cấu trúc sau:

http://<tên máy>:8081

Ví dụ ở đây là: <http://VM-LNTRI-1:8081>

- Nhập url trên vào trình duyệt để đảm bảo máy Slave đã làm việc:



← → ↻ 🏠 ⓘ vm-lntri-1:8081/

Spark 2.4.4 **Spark Worker at 172.24.40.82:49707**

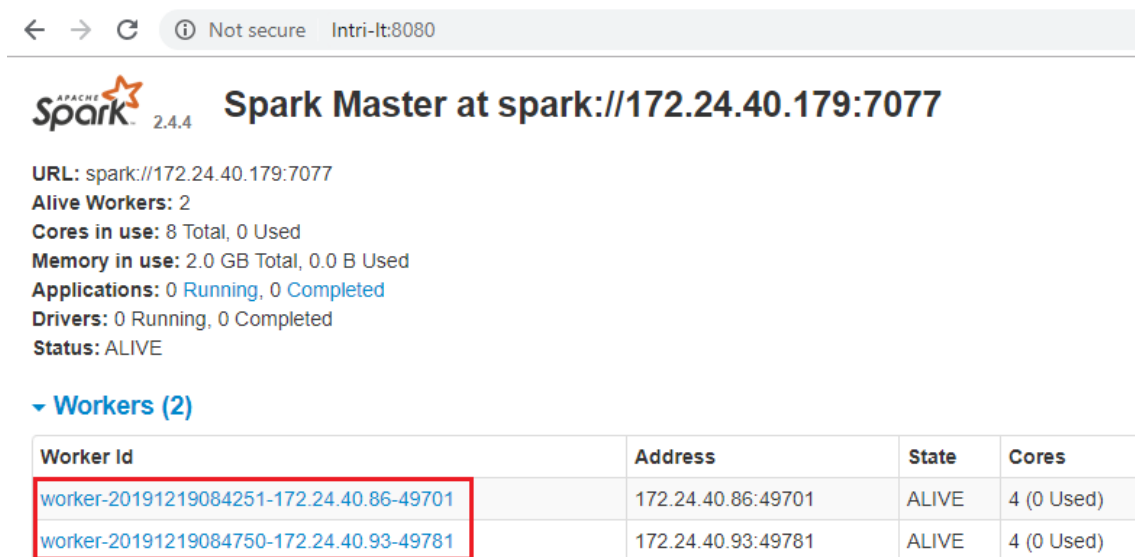
ID: worker-20191212095621-172.24.40.82-49707
Master URL: spark://172.24.40.179:7077
Cores: 4 (0 Used)
Memory: 1024.0 MB (0.0 B Used)

[Back to Master](#)

▼ **Running Executors (0)**

ExecutorID	Cores	State
------------	-------	-------

- Trở lại máy Master, quan sát thấy có 2 máy Slave đã kết nối đến:



← → ↻ ⓘ Not secure Intri-It:8080

Spark 2.4.4 **Spark Master at spark://172.24.40.179:7077**

URL: spark://172.24.40.179:7077
Alive Workers: 2
Cores in use: 8 Total, 0 Used
Memory in use: 2.0 GB Total, 0.0 B Used
Applications: 0 [Running](#), 0 [Completed](#)
Drivers: 0 [Running](#), 0 [Completed](#)
Status: ALIVE

▼ **Workers (2)**

Worker Id	Address	State	Cores
worker-20191219084251-172.24.40.86-49701	172.24.40.86:49701	ALIVE	4 (0 Used)
worker-20191219084750-172.24.40.93-49781	172.24.40.93:49781	ALIVE	4 (0 Used)

3. Triển khai Project

- Khởi động Project bằng Jupyter Notebook, sau đó khai báo các thông số kết nối đến Master như hình sau:

```

jupyter demo_pyspark_3 Last Checkpoint: a minute ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 C
In [ ]: import findspark
        findspark.init()

In [ ]: import pyspark

In [ ]: from pyspark import SparkContext
        from pyspark.conf import SparkConf
        from pyspark.sql import SparkSession

        sc = SparkContext(master='spark://172.24.40.179:7077', appName='pyspark_3')
        spark = SparkSession(sc)
        sc

In [ ]: # Create an DataFrame from file path
        people_df = spark.read.csv("hdfs://SRVLT2:19000/people.csv", header=True, inferSchema=True)

        # Check the type of people_df
        print("The type of people_df is", type(people_df))
  
```

- Tiến hành chạy chương trình, và sau đó quay lại màn hình Master, quan sát thấy ứng dụng đã kết nối đến Master và các máy Worker đang làm việc:

Spark Master at spark://172.24.40.179:7077

URL: spark://172.24.40.179:7077
 Alive Workers: 2
 Cores in use: 8 Total, 8 Used
 Memory in use: 2.0 GB Total, 2.0 GB Used
 Applications: 1 Running, 0 Completed
 Drivers: 0 Running, 0 Completed
 Status: ALIVE

▼ Workers (2)

Worker Id	Address	State	Cores	Memory
worker-20191219084251-172.24.40.86-49701	172.24.40.86:49701	ALIVE	4 (4 Used)	1024.0 MB (1024.0 MB Used)
worker-20191219084750-172.24.40.93-49781	172.24.40.93:49781	ALIVE	4 (4 Used)	1024.0 MB (1024.0 MB Used)

▼ Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20191219091141-0000	(kill) pyspark_3	8	1024.0 MB	2019/12/19 09:11:41	Intri	RUNNING	84 ms

▼ Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	----------------	------	-------	----------

--- Hết ---