**LAB 1**

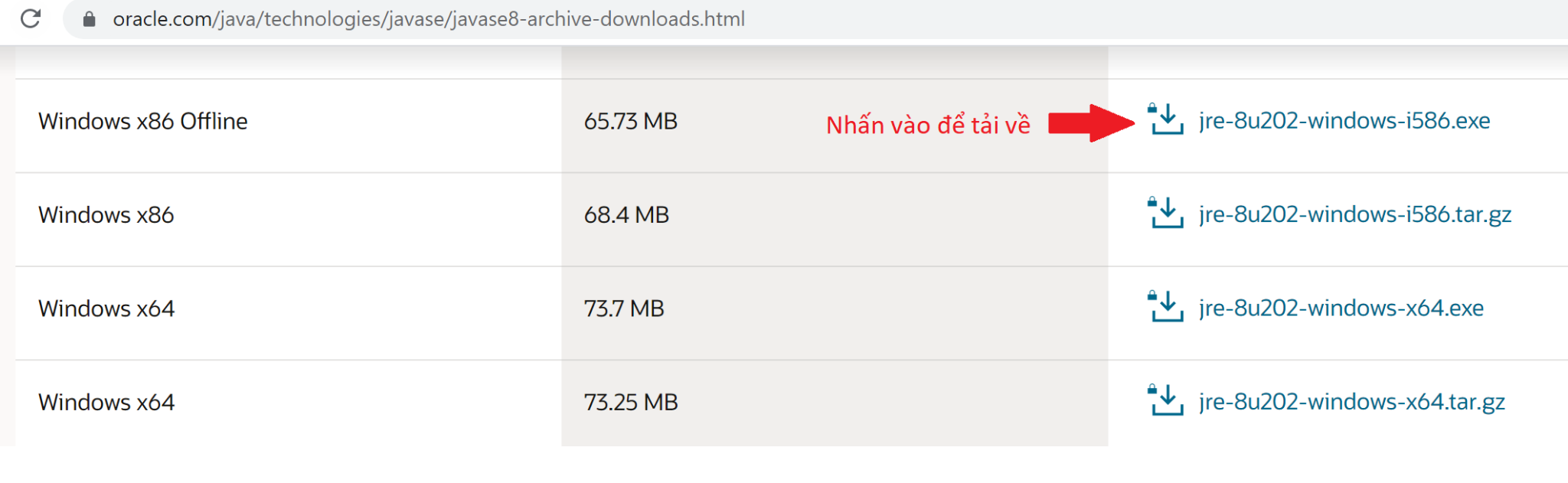
***Cài đặt Hadoop – Spark trên Windows***

1. Chuẩn bị

* Hệ điều hành Windows
* Cài đặt Java 8s
* Cài đặt Python 3
* **Lưu ý**: Nếu máy đã cài đặt rồi có thể bỏ qua

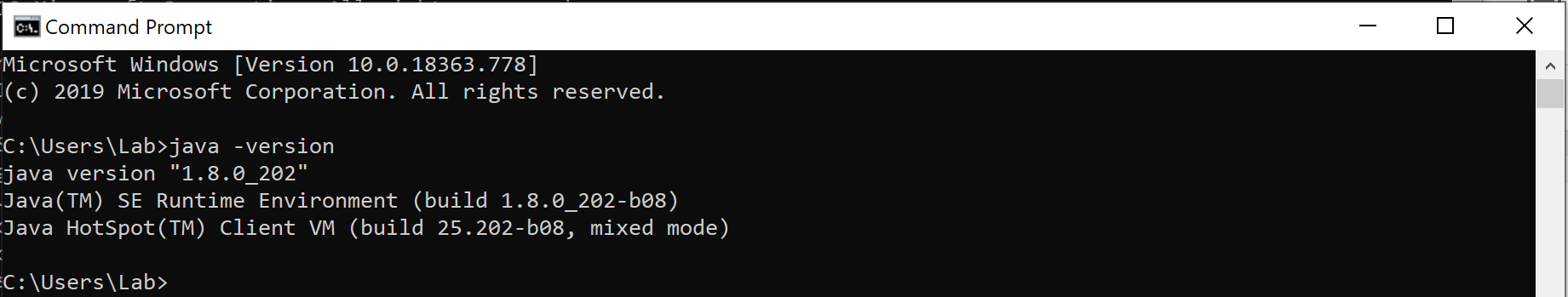
1. Cài đặt Java

* Truy cập đường dẫn <https://www.oracle.com/java/technologies/javase/javase8-archive-downloads.html>



* Chạy **jre-8u202-windows-i586.exe** để tiến hành cài đặt
* Sau khi cài đặt xong chạy lệnh bên dưới để kiểm tra

java -version



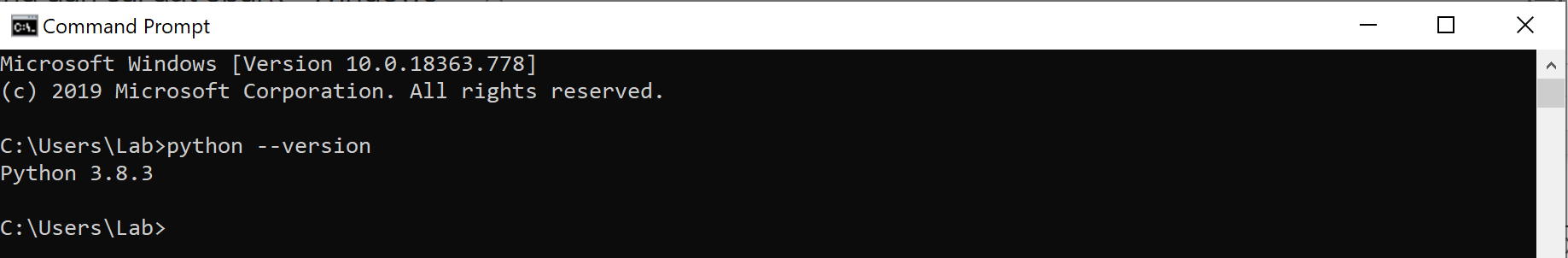
1. Cài đặt Python

* Truy cập đường dẫn <https://www.python.org/downloads/release/python-383/> và tải tập tin **Windows x86 executable installer**
* Tiến hành cài đặt, chọn option: Add Python 3.8 to PATH



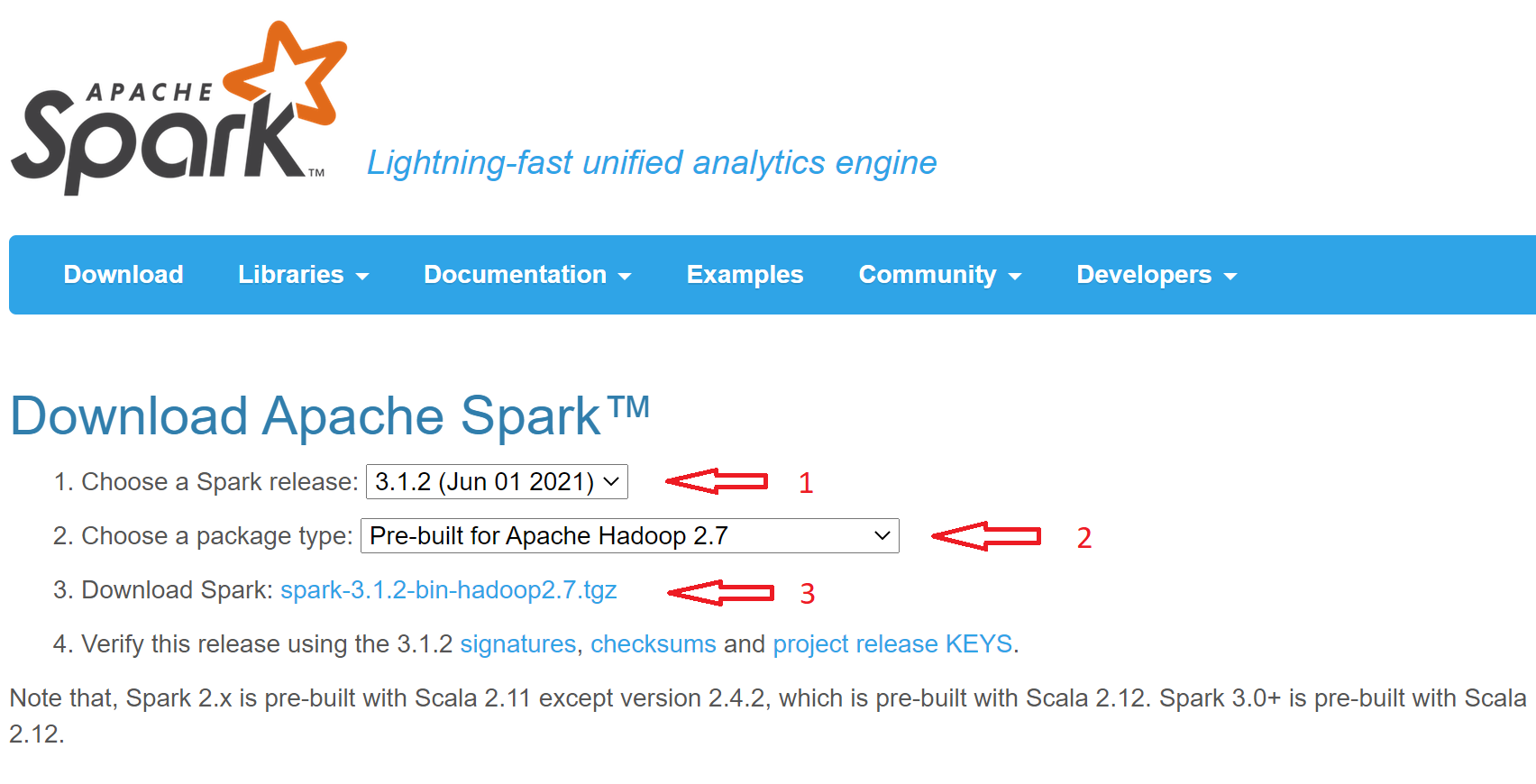
* Sau khi cài đặt xong chạy lệnh để kiểm tra

python --version

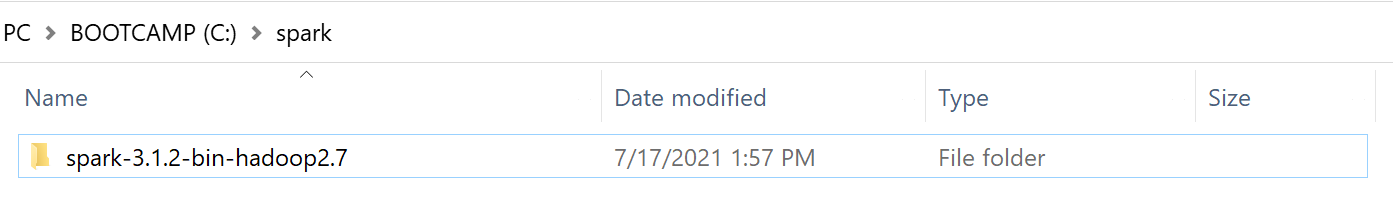


1. Cài đặt Apache Spark

* Truy cập đường dẫn <https://spark.apache.org/downloads.html>
* Chọn phiên bản Spark: **3.1.2 (có thể chọn phiên bản khác)**
* Chọn gói **Pre-built for Apache Hadoop 2.7** để nhận bản build sẵn
* Nhấp vào liên kết **spark-3.1.2-bin-hadoop2.7.tgz** để tải về

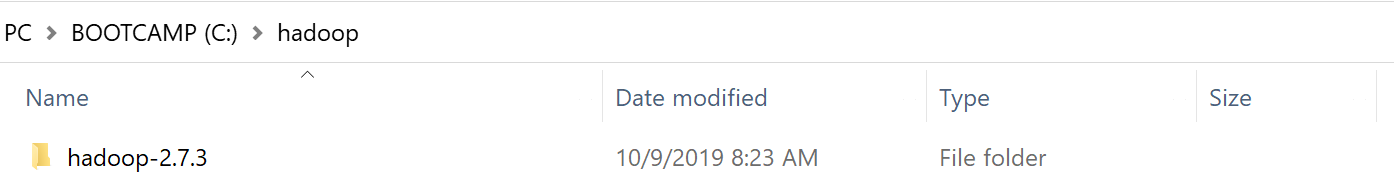
****

* Giải nén và copy vào folder **C:\spark\spark-3.1.2-bin-hadoop2.7**



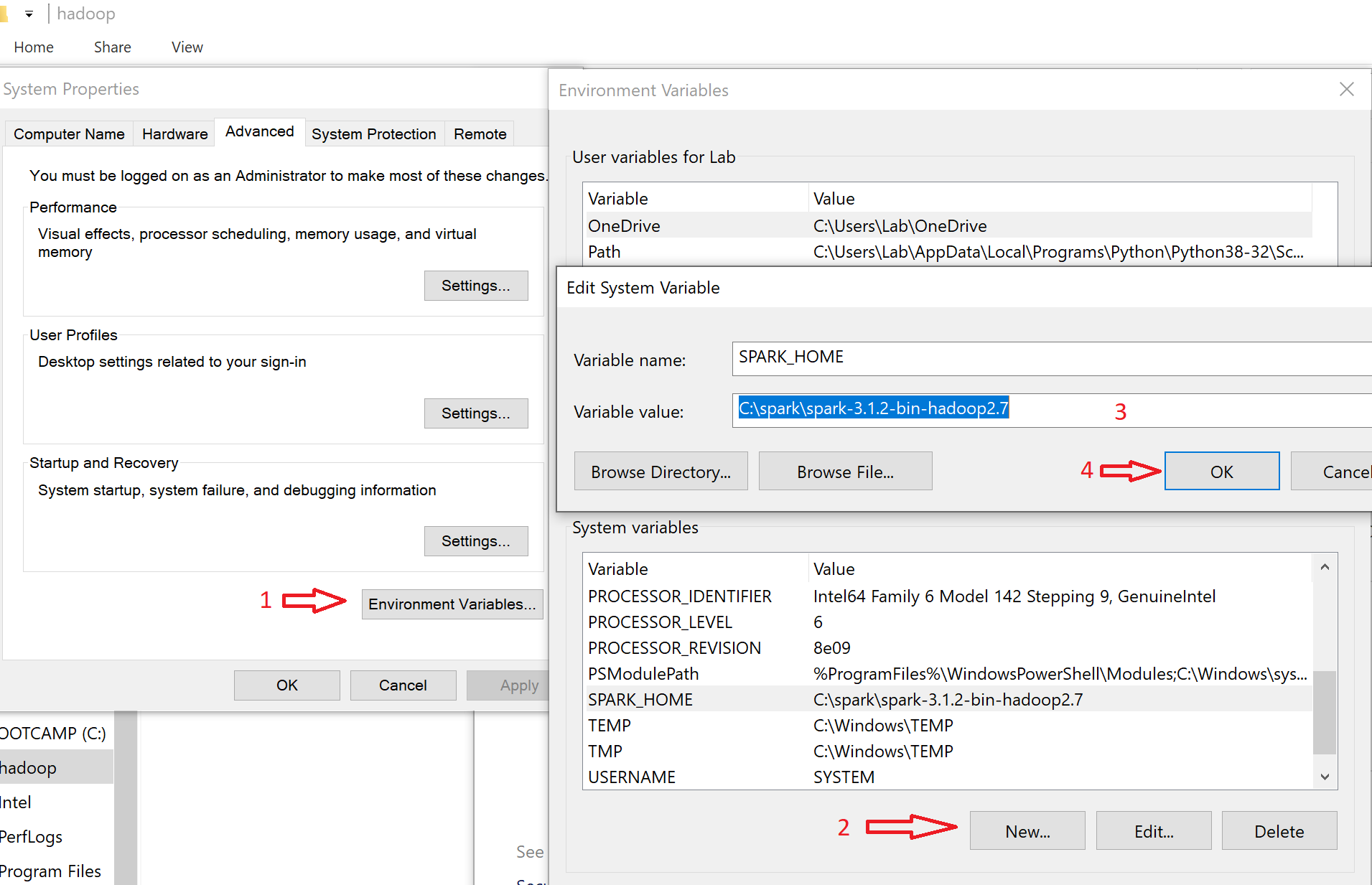
1. Cài đặt Hadoop

* Tạo folder chứa hadoop: **C:/hadoop/**
* Truy cập đường dẫn và tải về <https://github.com/cdarlint/winutils/blob/master/>
* Giải nén tập tin **winutils-master.zip**
* Chọn Hadoop có version tương ứng với Spark bên trên
* Copy vào **C:/hadoop/**

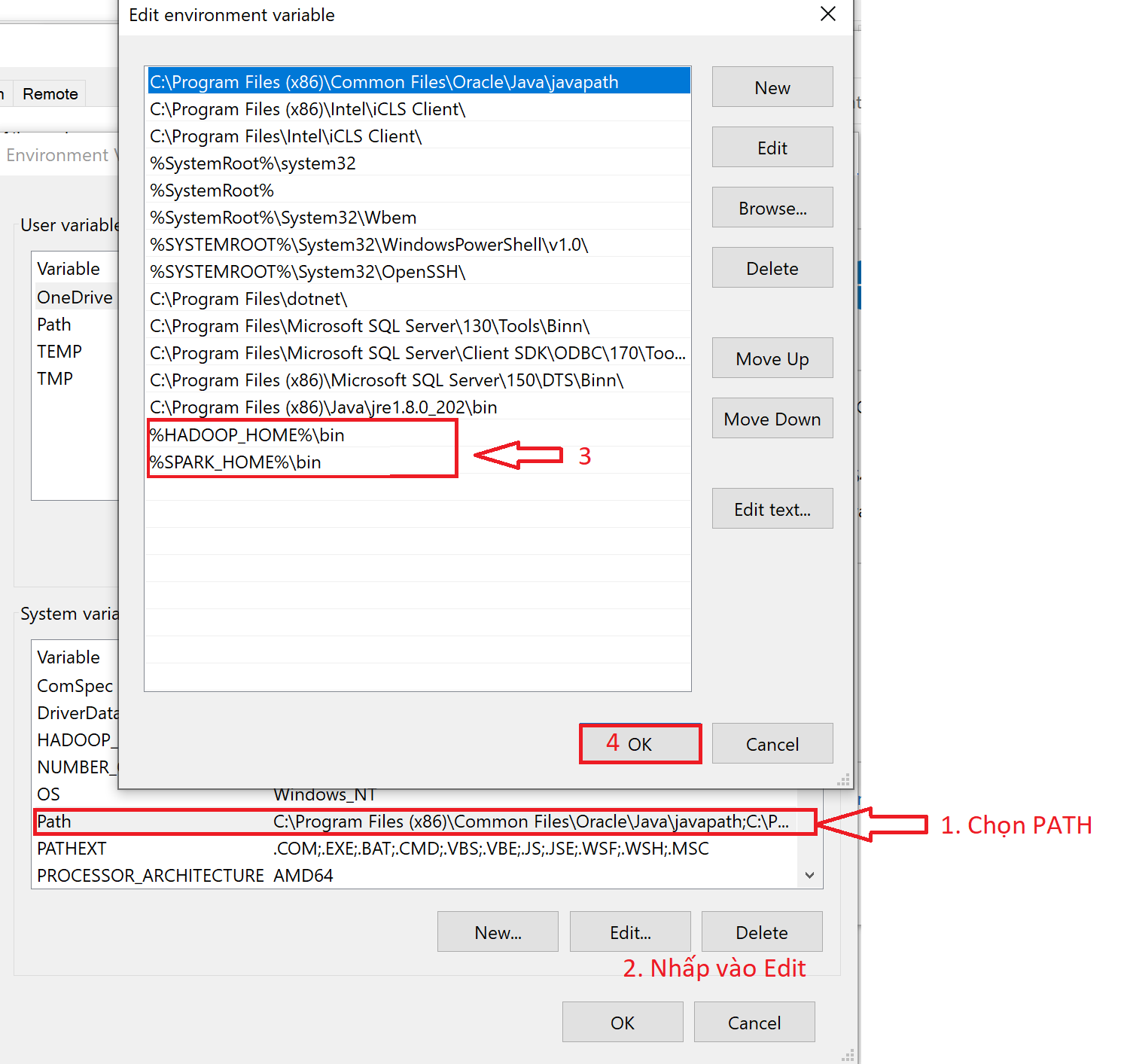
****

1. Thiết lập biến môi trường

* **SPARK\_HOME**=C:\spark\spark-3.1.2-bin-hadoop2.7
* **HADOOP\_HOME**=C:\hadoop\hadoop-2.7.3



* Thêm Spark và Hadoop vào system PATH



1. Kiểm tra và khởi động Apache Spark

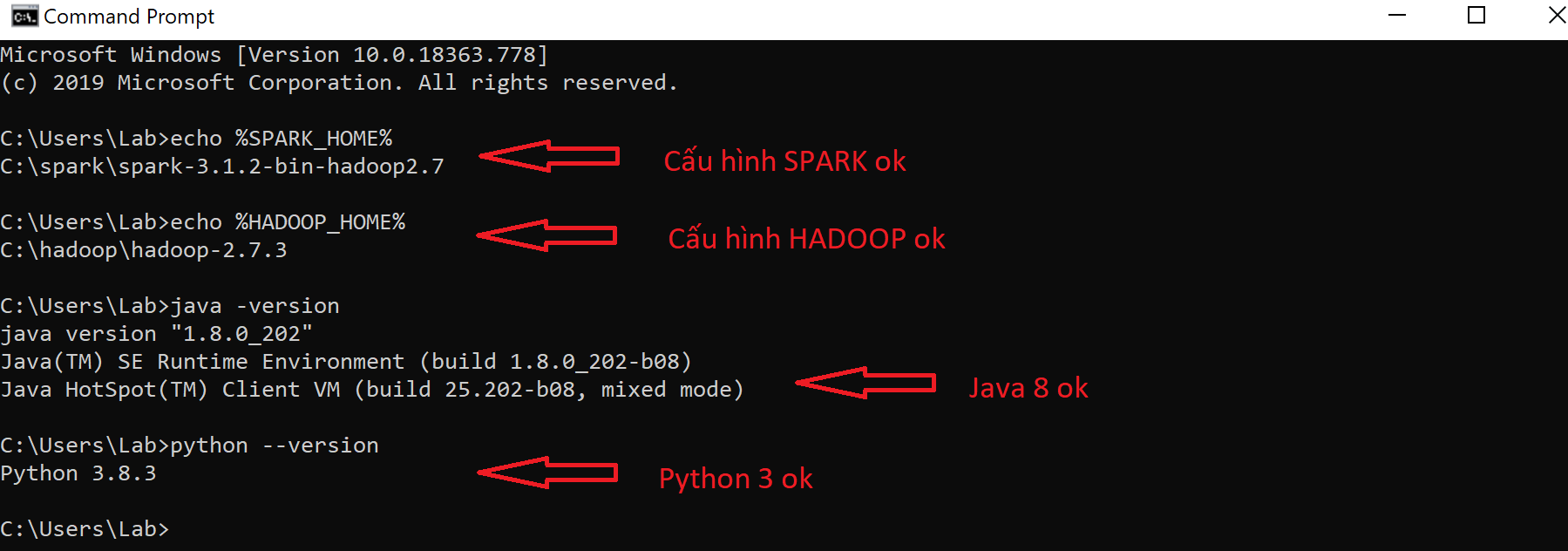
* Tắt và mở lại Command Prompt
* Chạy lệnh

echo %SPARK\_HOME%

echo %HADOOP\_HOME%

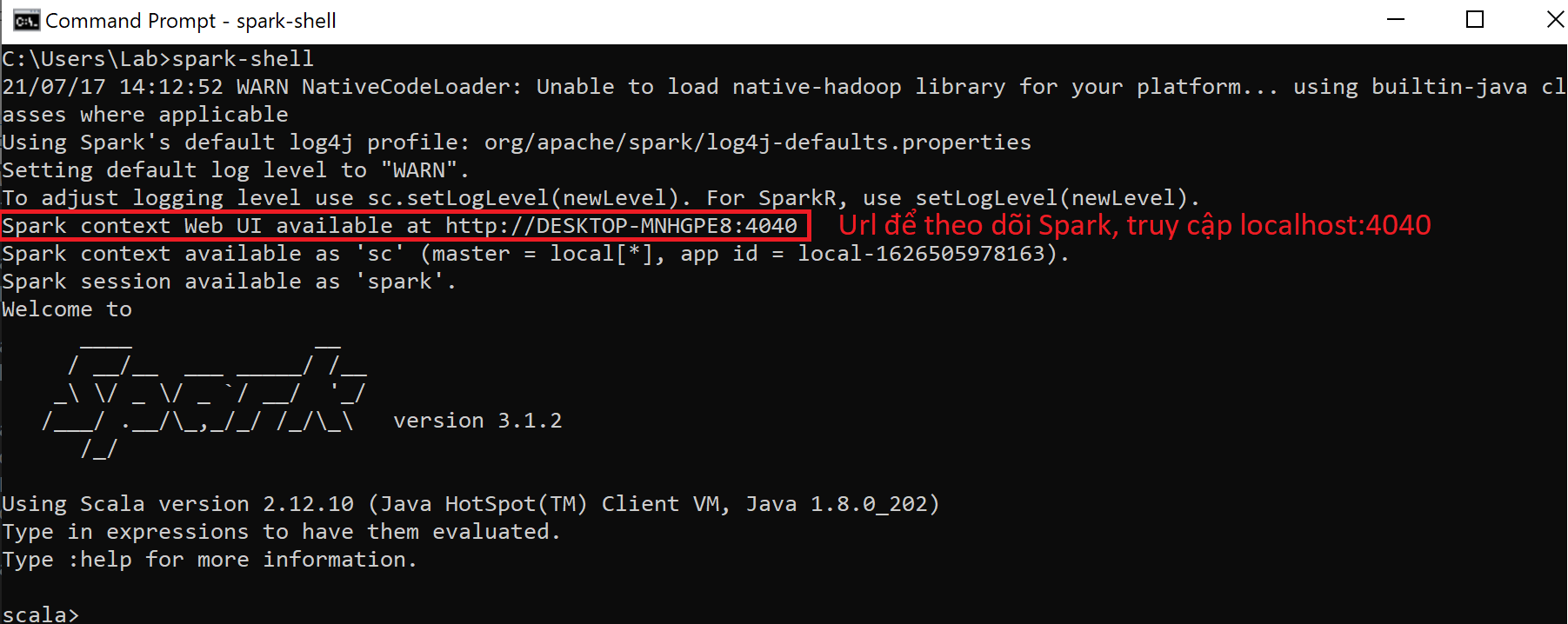
java -version

python --version



* Khởi động Apache Spark, chạy lệnh:

**spark-shell** làm việc với Apache Spark bằng scala code



**pyspark** làm việc với Apache Spark bằng python code



1. Ví dụ: Word Count <https://spark.apache.org/examples.html>

Mô tả: Đếm tất cả các từ trong file input.

// dữ liệu đầu vào README.md

val textFile = sc.textFile("file:///C:/spark/spark-3.1.2-bin-hadoop2.7/README.md")

// thực hiện cắt chuỗi và đếm

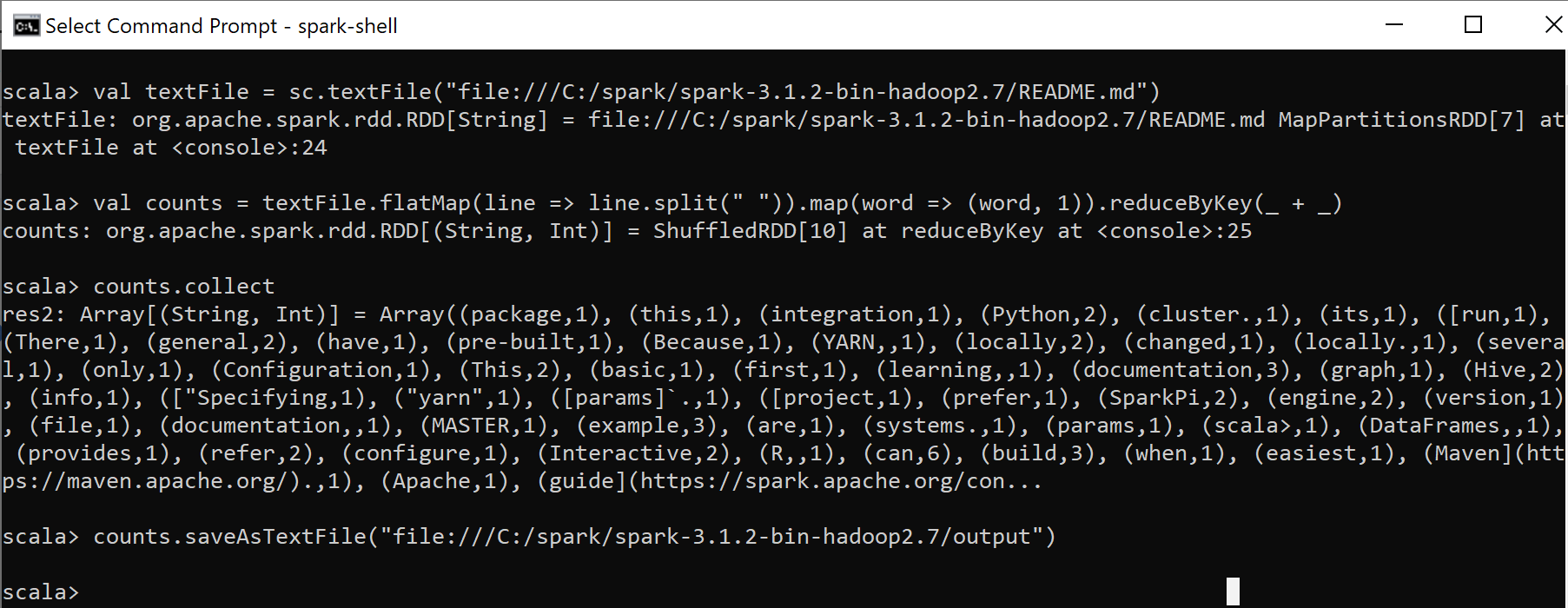
val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(\_ + \_)

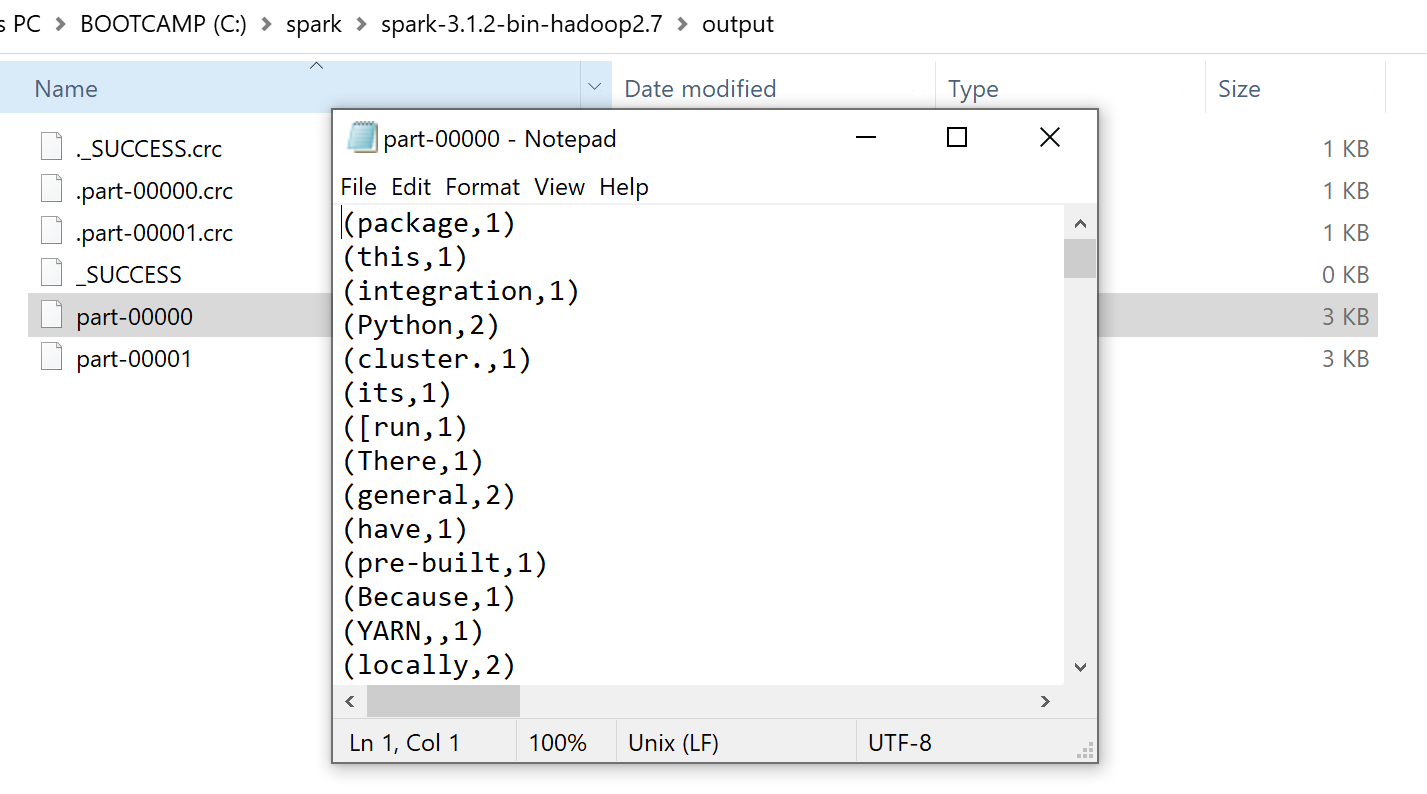
// in output ra màn hình

counts.collect

// lưu kết quả vào file

counts.saveAsTextFile("file:///C:/spark/spark-3.1.2-bin-hadoop2.7/output")





KETQUA

A computer screen with white text

Description automatically generated

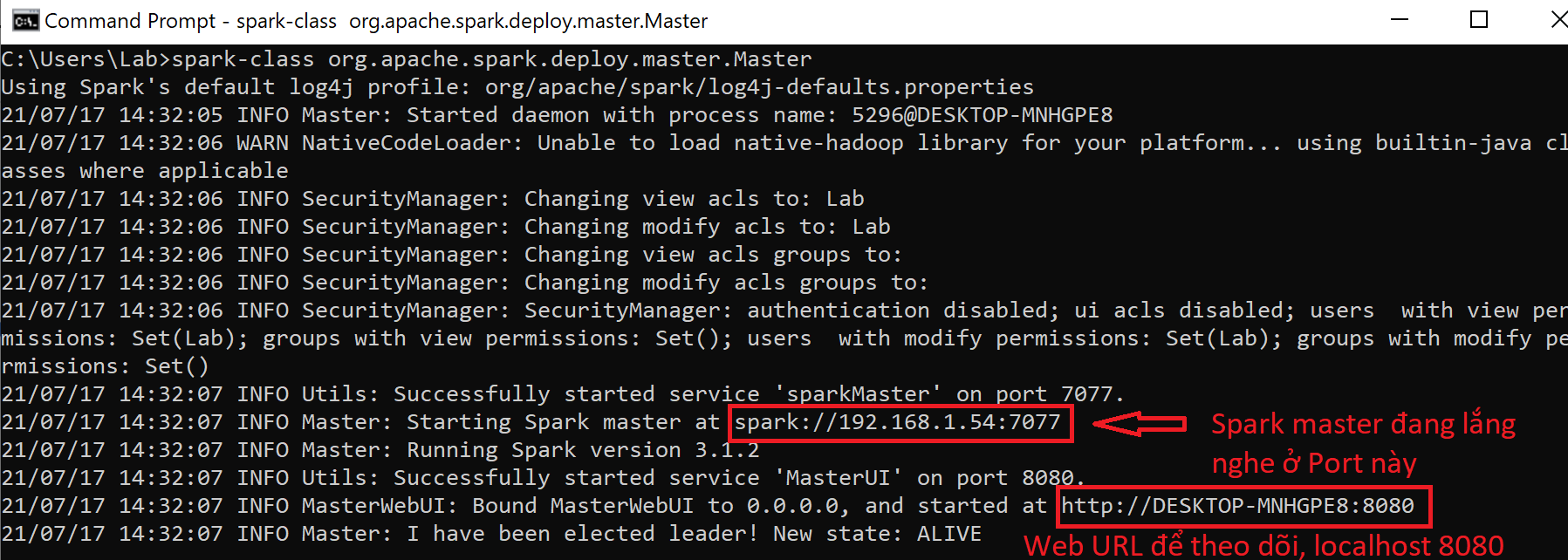
A screenshot of a computer

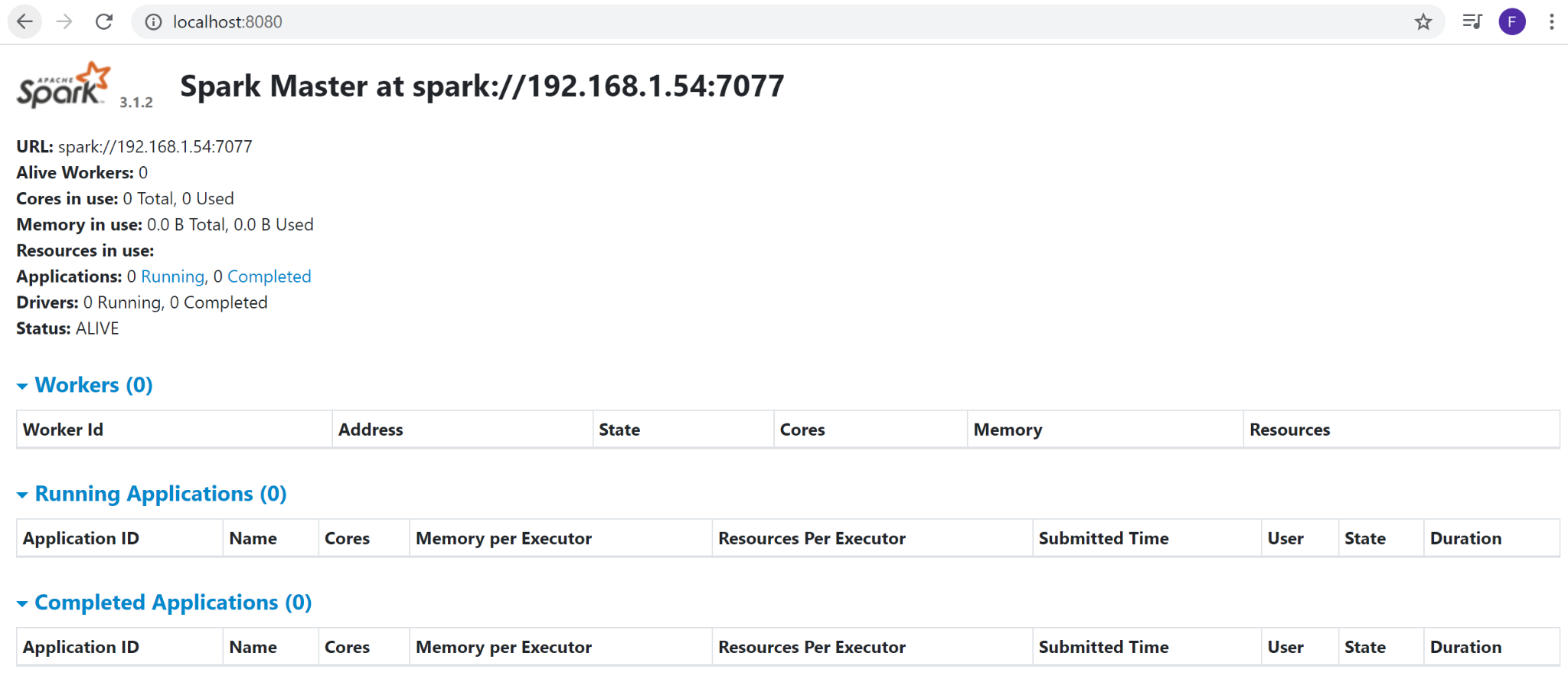
Description automatically generated

1. Khởi động SPARK Master

* Mở 1 cửa sổ Command Prompt mới, chạy lệnh:

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

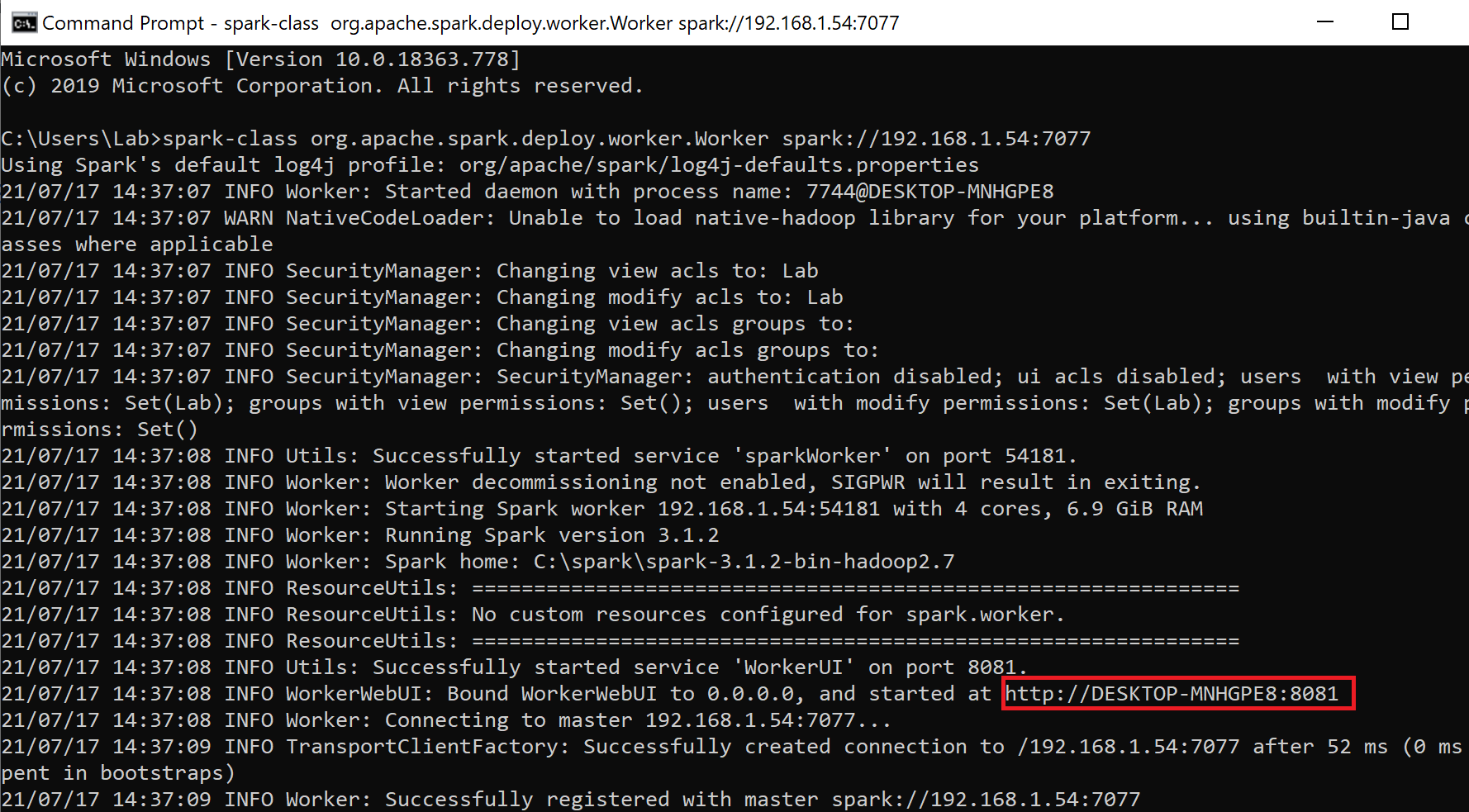




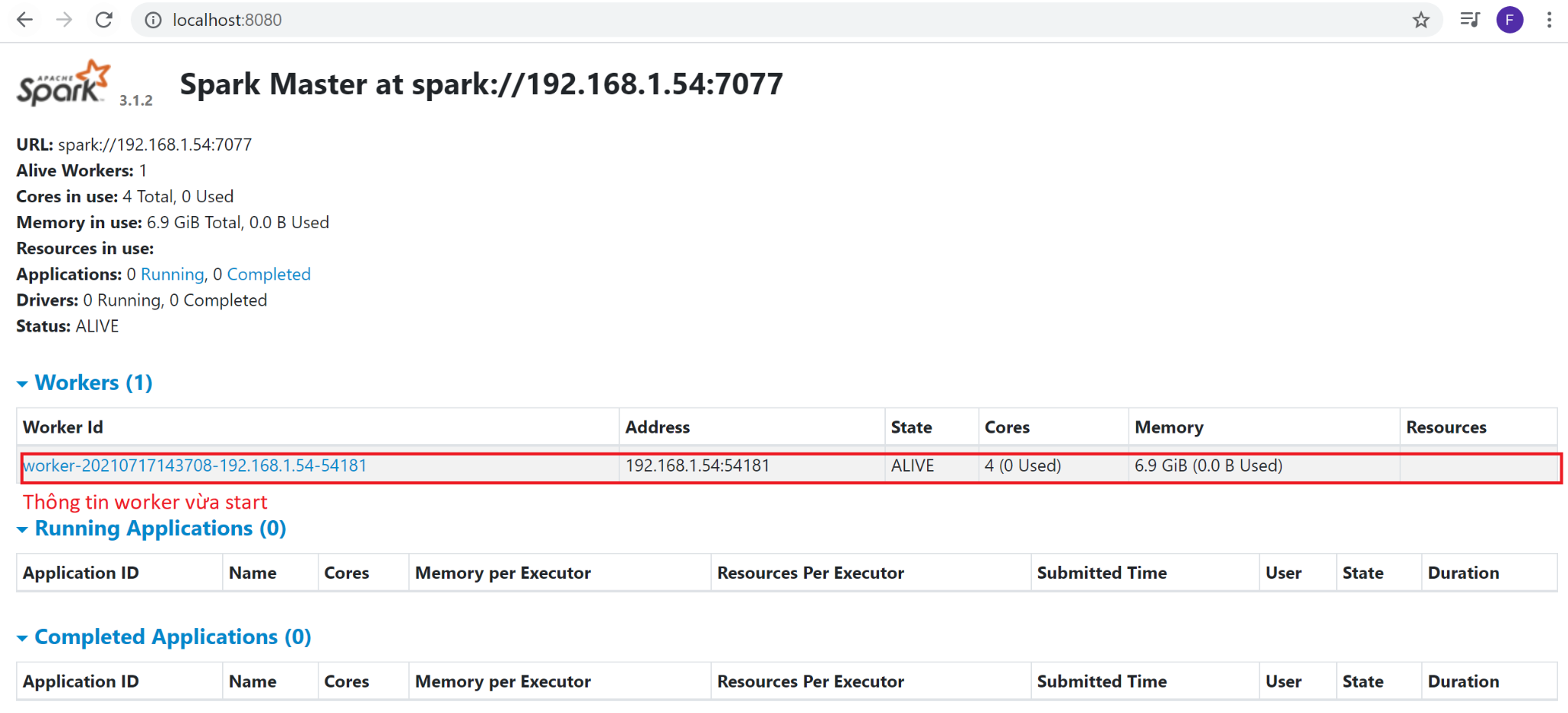
1. Khởi động SPARK Worker

* Mở 1 cửa sổ Command Prompt mới, chạy lệnh:

spark-class org.apache.spark.deploy.worker.Worker spark://192.168.1.54:7077 (chỗ này phải sửa ip theo trên máy mình)



* Reload lại trang Spark Master, localhost:8080 để xem thông tin, trạng thái Worker vừa được thêm vào



KETQUA

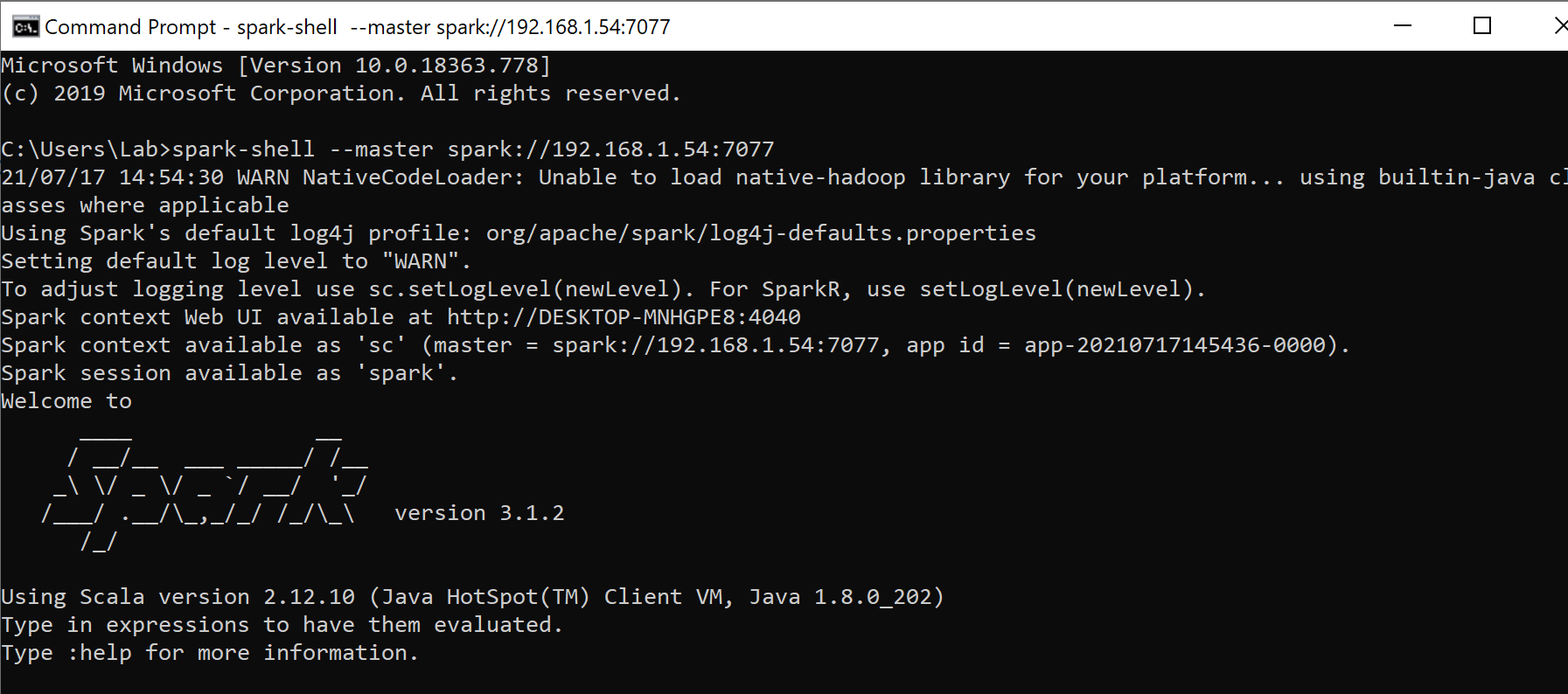
A screenshot of a computer

Description automatically generated

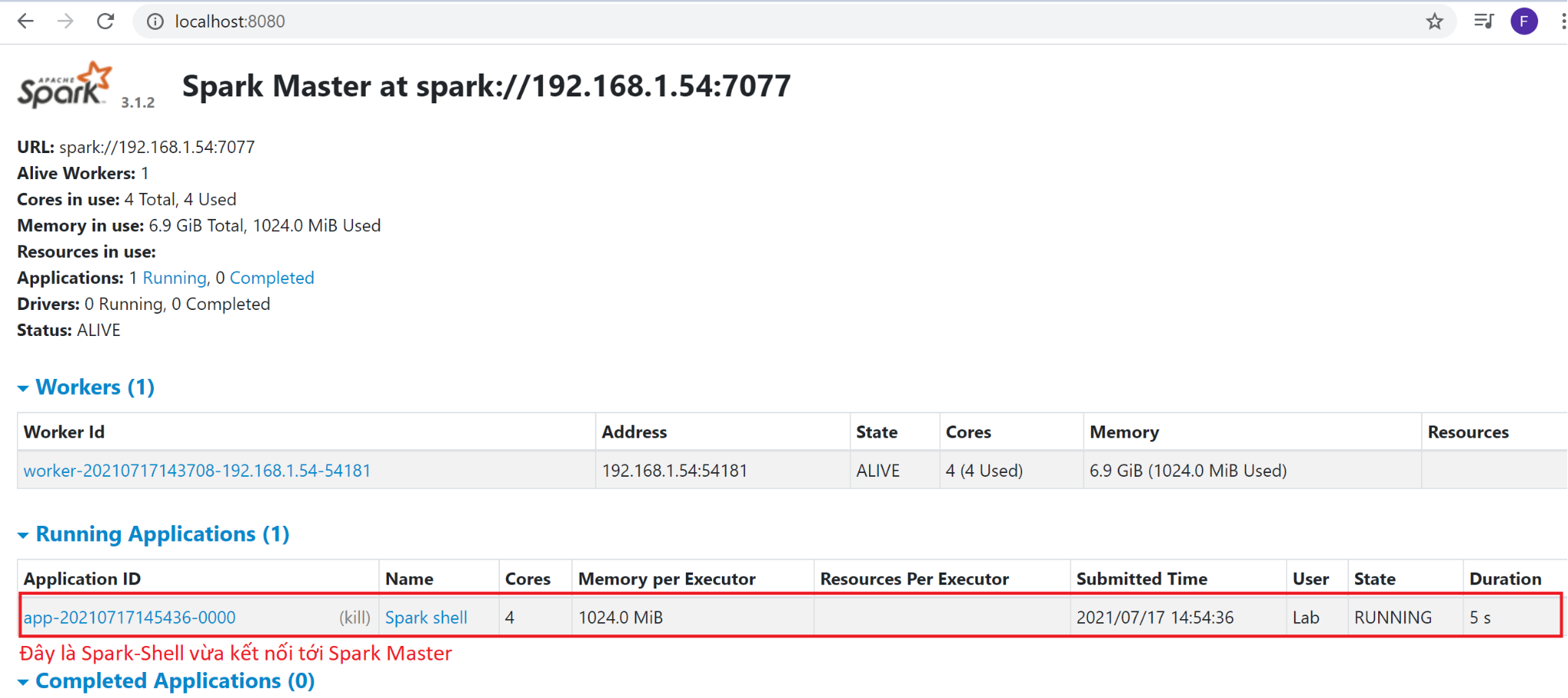
1. Kết nối Shell với Spark Master

* spark-shell --master spark://IP:PORT
* Ví dụ:

spark-shell --master spark://192.168.1.54:7077



* Reload lại trang Spark Master, localhost:8080 để xem thông tin, trạng thái Spark Shell



KETQUA

A screenshot of a computer

Description automatically generated

Tài liệu tham khảo:

<https://spark.apache.org/docs/latest/>

<https://spark.apache.org/docs/latest/spark-standalone.html>