SETUP Spark Cluster

# 1. Installing environments

## 1.1 Installing Java 11

Spark uses JDK 11, installs Java 11 packages and sets environment variables

Open the "edit environment variable" tool

A blue rectangular object with orange edges

Description automatically generated

A blue rectangle with white text

Description automatically generated

## 1.2 Installing Python

Download python from <https://www.python.org/downloads/> website and install them. Then open "edit environment variable" tool and declare environment variable

A blue and white rectangle with text

Description automatically generated

A blue rectangle with white letters

Description automatically generated

# 2. Installing Spark

## 2.1 Download Spark

Download spark at the following address <https://www.apache.org/dyn/closer.lua/spark/spark-3.5.3/spark-3.5.3-bin-hadoop3.tgz>

After downloading, we unzip the folder

A black background with white text

Description automatically generated

## 2.2 Setup Spark

Next we need to open the "edit environment variable " tool and declare environment variables for spark



A blue rectangle with white text

Description automatically generated

# 3. Testing spark

Use the command cd %SPARK\_HOME% to move to the Spark directory and type the command "Spark-Shell"

A computer screen shot of a black screen

Description automatically generated

A screenshot of a computer

Description automatically generated

# 4. Spark Cluster

## 4.1 Connecting to a common network (master and workers)

A screen shot of a phone

Description automatically generated

Master's ip address

A screen shot of a computer

Description automatically generated

## 4.2 Renaming the hosts

Rename the host file of the master and slave machines for easy management (perform on master and workers)

Run command prompt as admin

A screenshot of a computer command prompt

Description automatically generated

To easily manage slave machine addresses by name instead of numbers. Type the command “notepad C:\Windows\System32\drivers\etc\hosts” to name each ip

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

This will help us manage the machine easily when pinging or checking. Instead of pinging the machine address directly, we only need to ping the machine name.

## 4.3 Turning off the firewall

Turn off the firewall of the master and slave machines (master and workers)

Open “Control Panel”

A black and white sign with white text

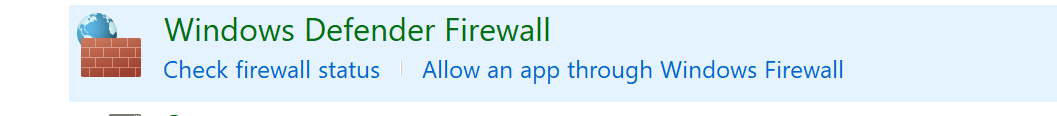
Description automatically generated

Click “System and Security”

A close-up of a computer security message

Description automatically generated

Click “Windows Defender Firewall”



Click “Turn Windows Defender Firewall on or off”

A screenshot of a computer

Description automatically generated

Turn off and click ok

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

## 4.4 Check connection

Use the ngocphung-master machine to ping the ngocphung-slave machine

A screenshot of a computer program

Description automatically generated

## 4.5 Deploy on the master machine (execute on the master machine)

Navigate to Spark\_Home using the following command “cd %SPARK\_HOME%\bin”

Start the master machine using the following command “spark-class2.cmd org.apache.spark.deploy.master.Master”

A black screen with white text

Description automatically generated

A screenshot of a computer

Description automatically generated

## 4.6 Deploy on the master machine (execute on the master machine)