

**Module 1 Lab 0: Grab Azure VM or Install VMWare and Start Cloudera**

CPS Analytics, Northeastern University

ALY 6110: Data Management and Big Data

Dr. Mohammad Shafiqul Islam

November 11, 2022

**INTRODUCTION:**

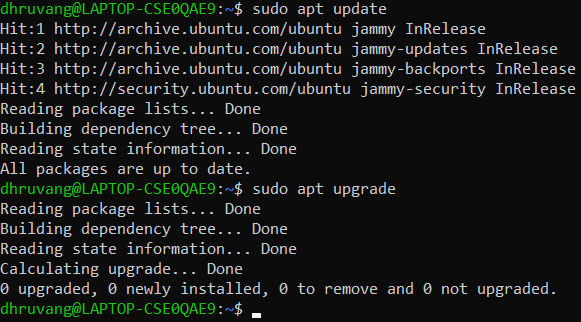
This assignment requires you to successfully create a single node Spark or PySpark cluster using either Azure Lab or your personal laptop, and to include step-by-step screenshots of the key steps with the date and time stamped.

**STEP 1:**

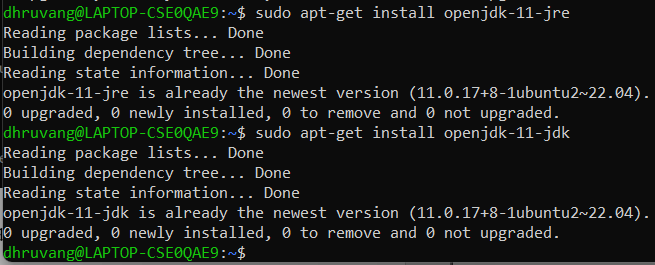
Firstly, Ubuntu VM is downloaded in the laptop and as it is an Windows machine, Windows Subsystem(WLS) for Linux to access Ubuntu VM in windows. So WLS would be used to create single node spark cluster.

**STEP 2:**

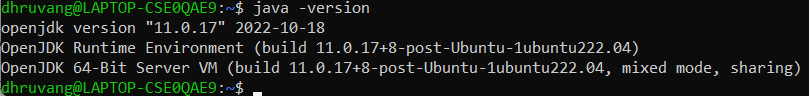
Once Ubuntu has been downloaded, the terminal will open and prompt you to provide a username and password. Once the username and password have been established, Java must be updated and upgraded. JDK and JRE must then be installed following the upgrade. Java programmes may be created and run using any of these. After the installation is complete, we must verify the Java version that is installed on our machine.



**Figure 1: Update and Upgrade**



**Figure 2: Installing JDK and JRE**



**Figure 3: Java Version**

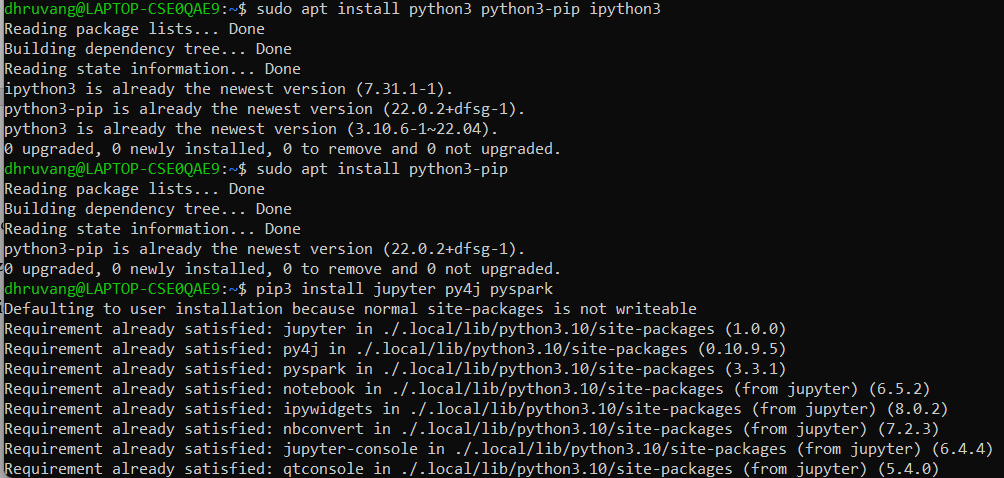
A software development environment for creating Java programmes and applets is called the Java Development Kit (JDK). The Java Runtime Environment (JRE), a Java interpreter/loader, a Java compiler, a Java archiver (jar), a Javadoc documentation generator, and other tools essential for Java development are all included.

The software package known as the Java Runtime Environment (JRE) includes everything needed to run a Java programme. The Java Virtual Machine (JVM), a Java interpreter/loader, libraries (rt.jar, charsets.jar, tools.jar, etc.), and other items required by the JVM are all included in this package.

**STEP 3:**

We must install Python, PYSpark, and Hadoop in our system after Java has been installed. We use the same procedure for installing, updating, and upgrading Python. In other words, PYSpark may be referred to as a Python API.

An open-source web programme called Jupyter Notebooks enables you to create and share documents with real-time code, equations, visuals, and text. Data transformation and cleansing, numerical simulation, statistical modelling, data visualisation, machine learning, and other applications are only a few examples.



**Figure 4: Installing Python, PYSpark and Jupyter**

Spark is a parallel processing platform for running large-scale data analytics applications, and PYSpark is its Python API.

**STEP 4:**

After Java and Python installation, we must establish an alias in the nano /.bashrc file. The bash shell uses a configuration file called /.bashrc to store its settings. It is used to alter the behaviour of the shell and set environment variables like the PATH. Nano is a text editor that is used to edit files in the terminal, to put it another way. A secret file called /.bashrc includes options for the Bash shell.

We must set the alias for the Jupyter notebook, which will be added at the end of the line code, when the nano file has been opened.



**Figure 5: Alias for Jupyter Notebook**

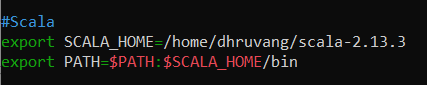
If we don't source the nano file, the command won't run anytime it is called, thus we must source the file.

**STEP 5:**

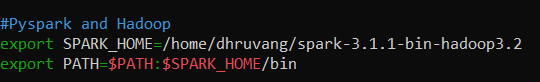
You may either download Scala directly from the web and manually import it into the system, or you can use the command "wget" to download, extract, and automatically import Scala. So I decided to use Ubuntu, which will automatically install Scala and extract it, saving us from having to complete the entire procedure ourselves.

After the installation is complete, we must specify the path of the file in the nano file.

The environment variable "export SCALA HOME=/home/dhruvang/scala-2.13.3" is used to define the location of the Scala installation directory. The location of the Scala installation directory is specified via the PATH environment variable.



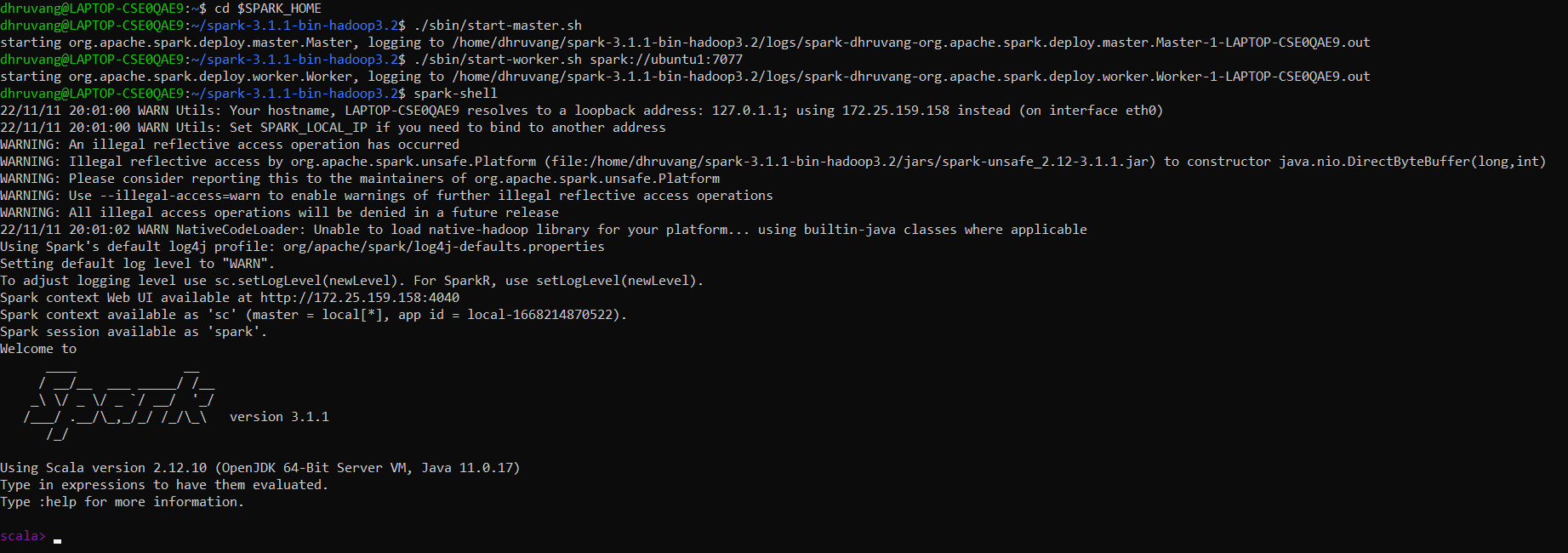
**Figure 6: Setting Path For Scala in Bash File**

****

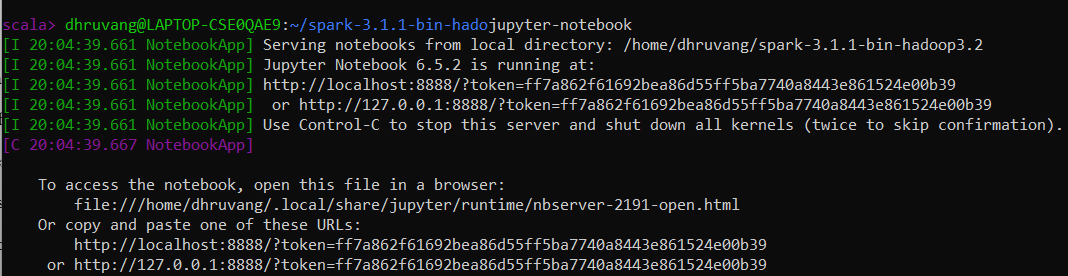
**Figure 7: Spark and Hadoop Environment Variable**

In order to use the SPARK HOME instance, Figure 7 depicts the environment variable configured for Spark and Hadoop. After specifying the location, we must source the file in order to save it using the bash setting. .

**STEP 6:**

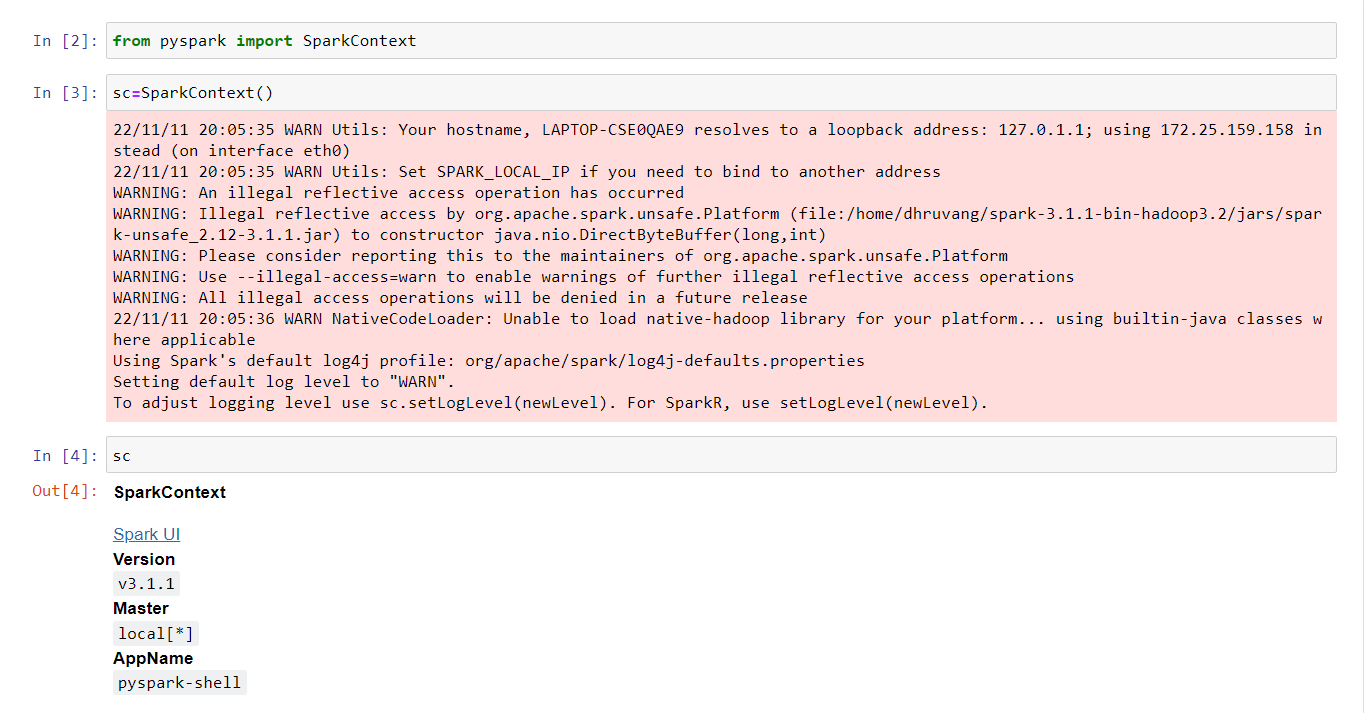
The final step is to activate the spark using the command "cd SPARK HOME" after completing the previous stages. You'll be sent to the spark directory. We must launch Master.sh and Worker.sh after entering the spark directory since these two scripts may be used to manage a Hadoop cluster. The Hadoop Master node is started and stopped using Master.sh, while the Hadoop Slave nodes are started and stopped using Worker.sh. We must examine the spark shell once these two have been initiated. We can find out what version of Spark is presently installed on this device by executing spark shell. ****

**Figure 8: Spark Shell**

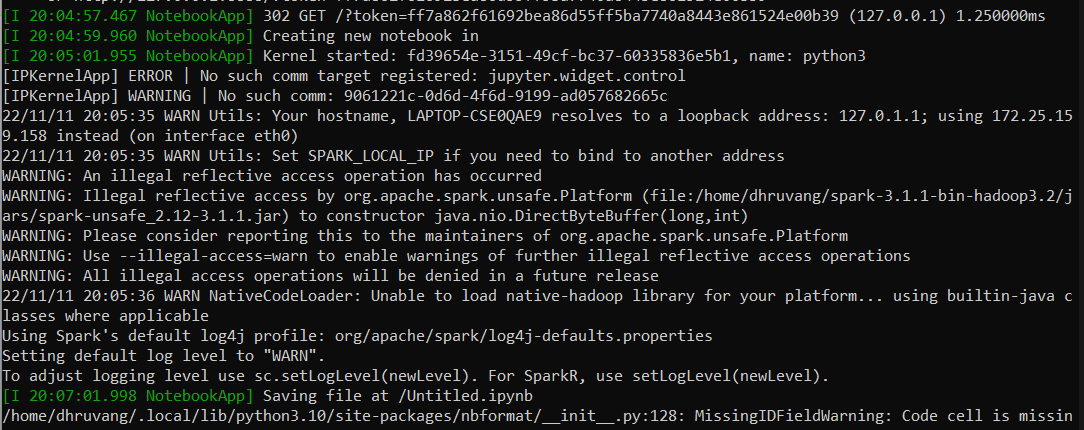


**Figure 9: Jupyter Notebook**

Figure 9 shows that the Jupyter notebook has been executed. Because we previously defined the route for the Jupyter notebook, we can see that there are certain URLs that must be copied and pasted into the web browser in order to view the Jupyter notebook directly.



**Figure 10 : Jupyter Notebook**

****

**Figure 11: Ubuntu Logs**

As can be seen, each time a task is completed in a Jupyter notebook, a log is recorded in Ubuntu and a timestamp is added.

The final stage in stopping all spark home functions is to stop the master and worker since, if we don't, the regular bill will increase. Therefore, it is preferable to end the master-worker relationship.

**REFERNCES:**

1. Islam, S (2022), Single Node Spark/PySpark Cluster on Windows Subsystem for Linux (WSL2), MLearning.ai in Medium. Available at:

<https://medium.com/mlearning-ai/single-node-spark-pyspark-cluster-on-windows-subsystem-for-linux-wsl2-22860888a98d>, Last Accessed: 11th November, 2022

1. Anonyomous. (2018). *Method for Maintaining a Bashrc File Across Multiple Workstations Using Homebrew and Github | Time-based Media & Digital Art.* Retrieved November 11, 2022, from <https://www.si.edu/tbma/resource/method-maintaining-bashrc-file-across-multiple-workstations-using-homebrew-and-github>