**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**

* The primary goal of this lab is to become familiar with installing Single Node Cluster and PySpark on hardware such that Java, Python, and Jupyter notebook may be installed.

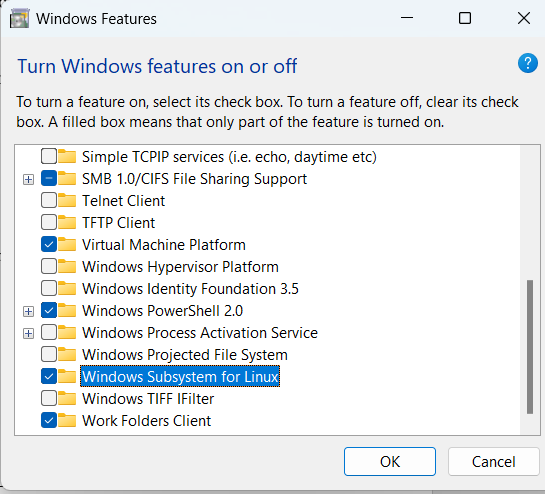
* In an Ubuntu virtual machine, we will set up a single node cluster to run PySpark (VM). We can set up an Ubuntu virtual machine in accordance with preferences and operating system.
* A cluster with only an Apache Spark driver and no Spark workers is known as a single node cluster. Spark tasks and all Spark data sources, including Delta Lake, are supported by a single node cluster. To perform Spark tasks on a Standard cluster, there must be at least one Spark worker.
* Single Node clusters are helpful for:
* Workloads for single-node machine learning that load and store data using Spark.
* Exploratory data analysis that is simple.

**Limitation of Single Node Cluster**

* A Single Node cluster's resources will run out while processing large amounts of data. Databricks advises utilising a Standard mode cluster for these workloads.
* Clusters with a single node are not intended to be shared. When a cluster must be shared, Databricks advises utilising a Standard mode cluster to prevent resource conflicts.
* It is not possible to scale a Standard mode cluster to 0 workers. Instead, employ a single node cluster.
* Process isolation is incompatible with single node clusters.
* Clusters with a single node do not have GPU scheduling enabled.

**Steps To Installation**

* Turning on the Windows Feature for the Linux and Virtual Machine Platform was the first thing I did after seeing a YouTube video.

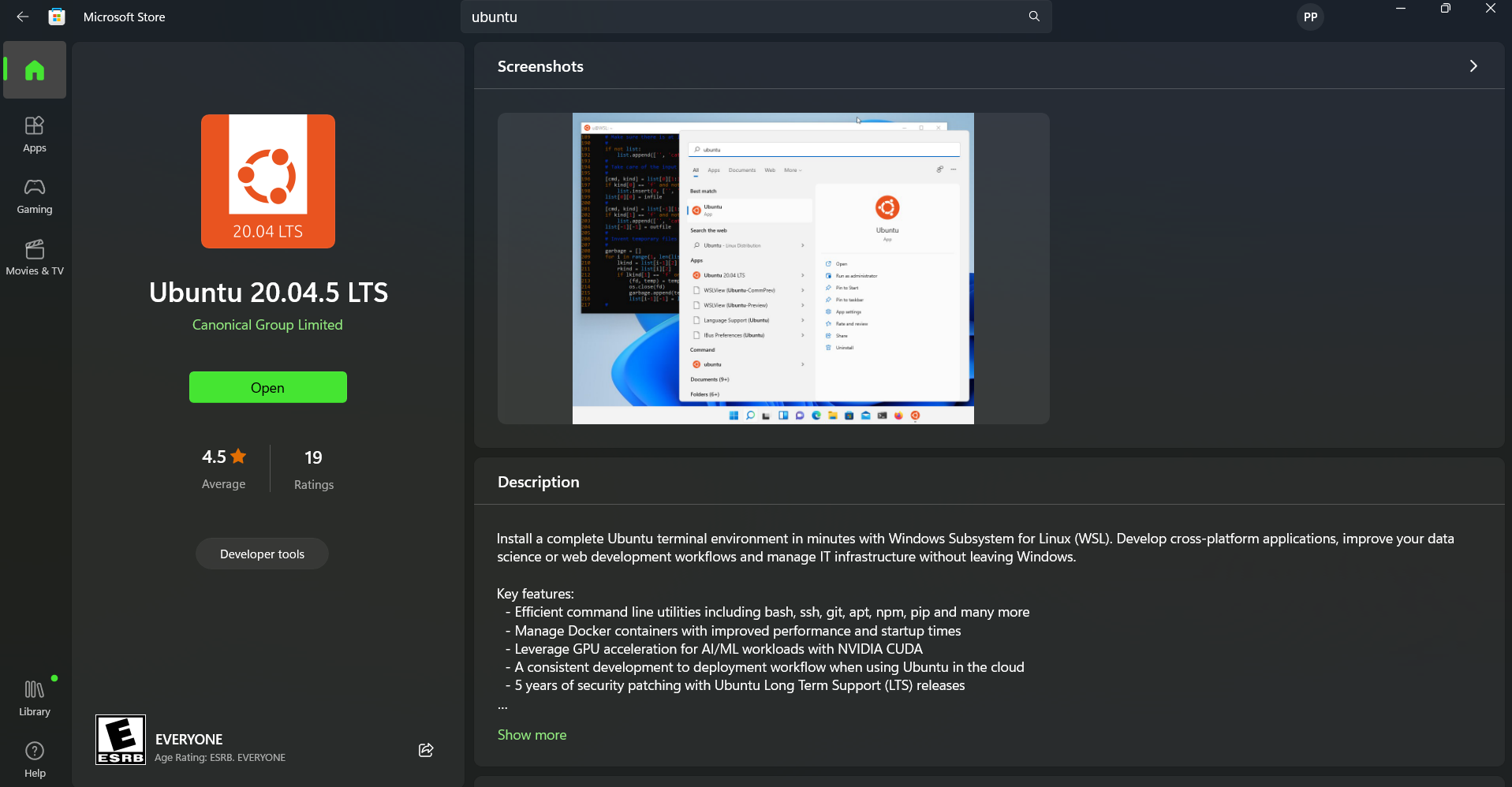


**Figure 1: Windows Feature On or Off**

* Once the modifications have taken effect, restart computer by clicking the Restart now option.

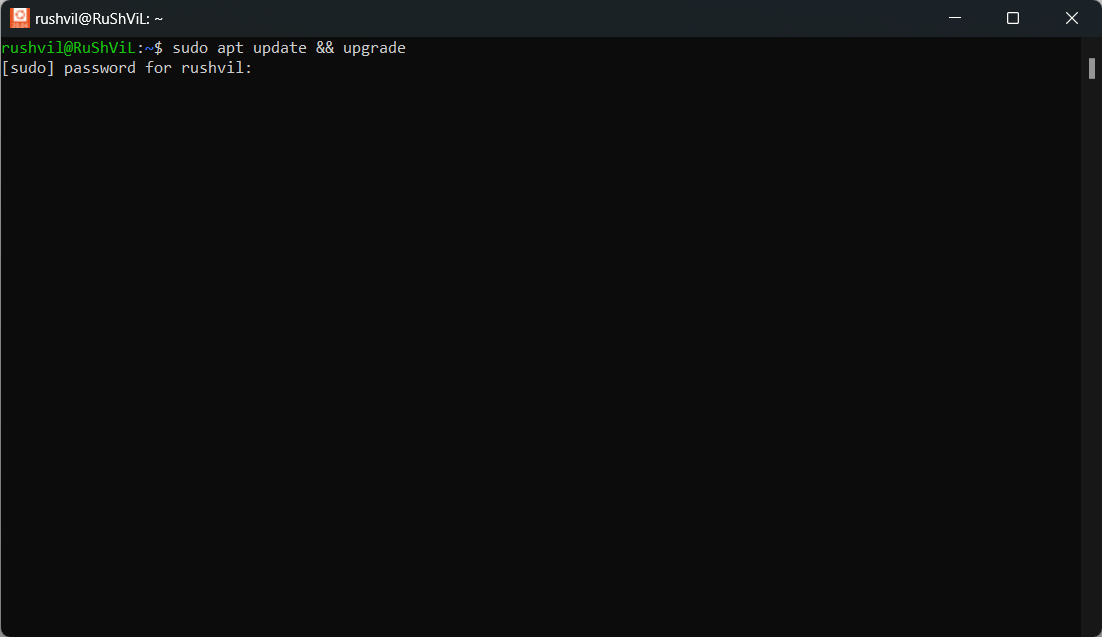
1. **Install and Update Ubuntu Linux**

* It is necessary to install a Linux distribution first. Navigate to Start > Microsoft Store.
* Find Ubuntu using the Microsoft Store search bar.
* Ubuntu will be downloaded and installed by Windows. Click the Launch button when it's completed.



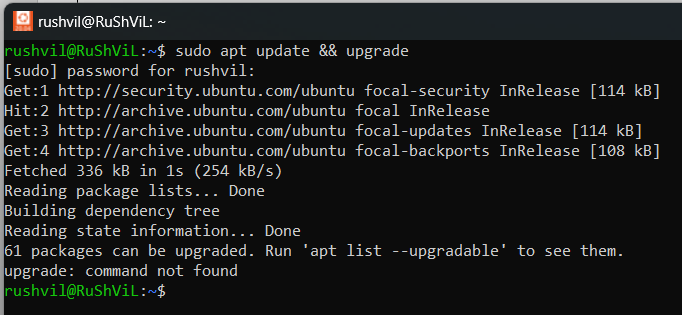
**Figure 2: Microsoft Store to Install Ubuntu**

* A window on the terminal will appear. When asked, establish a new Linux username and password after waiting for Ubuntu to finish installing. The new Linux command line is currently active.

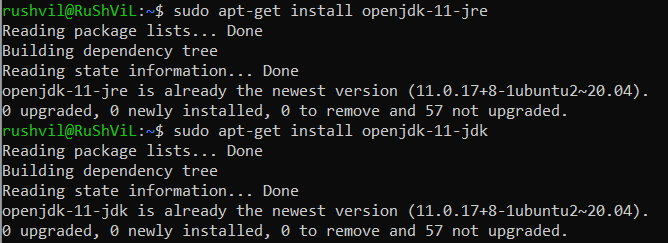


**Figure 3: Install and Update the Package**

* Substituting "sudo apt update && upgrade" for the command After being given the password, it will prompt you to begin downloading the package.
* The "sudo apt-get upgrade" programme downloads and instals the most recent versions of all out-of-date system dependencies and packages.

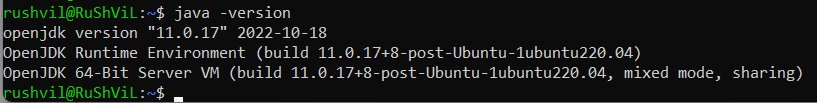
`

**Figure 4: Installed Java**

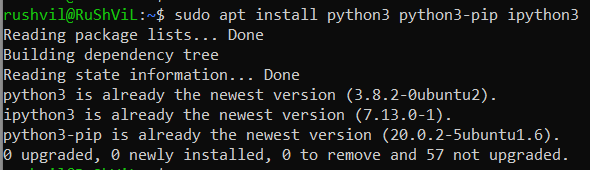
****

**Figure 5: Installing Jdk and Jre**

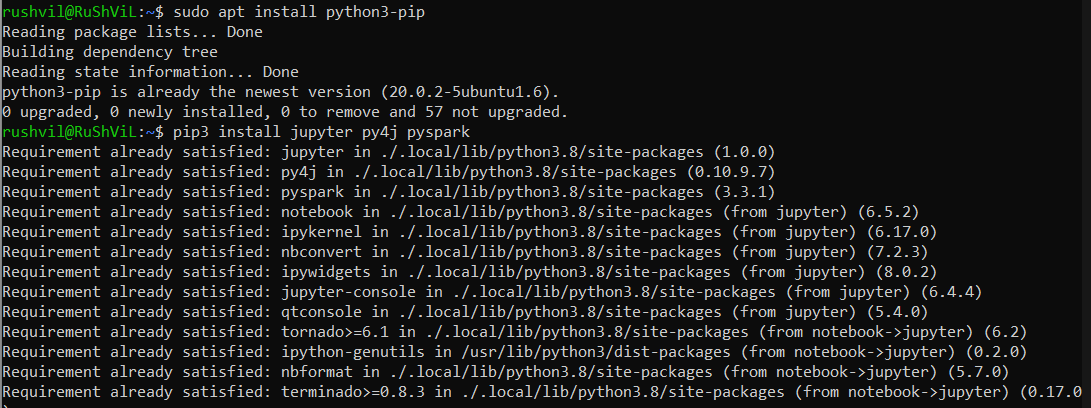
* The jdk and jre are installing, as seen in the above figure.
* JDK, or Java Development Kit, stands for. It is a setting for creating software that is used to create Java applets and apps.
* The Java Runtime Environment is known as JRE. It is the JVM's (Java Virtual Machine) implementation and was created specifically to offer a setting in which Java programmes might be run. Like JDK, it is likewise platform-specific. JVM, Java binaries, and other classes are included for the program's efficient operation.
* We can find out what version of Java is installed on the device by executing "java -version".



**Figure 6: Version of Java**

****

**Figure 7: Installing Python3**

****

**Figure 8: Installing PySpark and Jupyter**

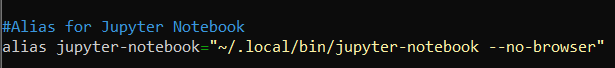
* Apache Spark is an open source, distributed computing platform and collection of tools for real-time, massive data processing, and PySpark is its Python API.
* It is simpler to explain the full project's process to intended audience when using Jupyter Notebook to gather all the components of a data project in one location. Users may develop data visualisations and other project components using the web-based application to share with others via the platform.

1. **Nano Bashrc File**

* Every time Bash is launched interactively, it executes a script called ".bashrc" in the Bash shell. An interactive shell session is set up.
* We may set values to that file in this hidden directory, which is another name for it.
* When a user signs in, a script file called ".bashrc" is launched. A number of settings for the terminal session are contained in the file itself. Setting up or activating these features involves creating command aliases, shell history, colouring, and completion.
* The nano ~/.bashrc file is a text editor that is used to modify various settings for the Bash shell.
* There are some benefits of Nano bashrc file such as
* For modifying configuration files, use the straightforward text editor nano.
* Customizations for the Bash shell are stored in the /.bashrc file.
* You may quickly add, remove, or modify settings in the /.bashrc file using nano.
* The syntax highlighting feature of nano makes it easier to spot configuration file issues.
* Other configuration files, such /.vimrc or /.screenrc, may be edited using nano.

1. **Setting Alias for Jupyter Notebook**

* After installing Python 3 and Java on Ubuntu, we must set an alias for Jupyter notebook in the /.bashrc file's nano editor.

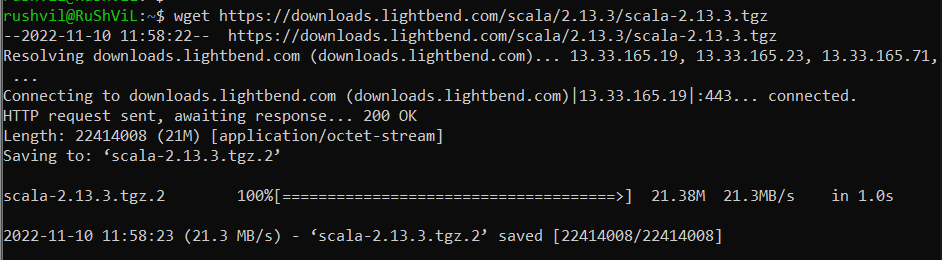


**Figure 9: Setting alias for Jupyter Notebook**

* An alias can save you a tonne of typing if often launch Jupyter Notebook from the command line.
* It may be simpler to recall the command to start Jupyter Notebook if it has an alias.

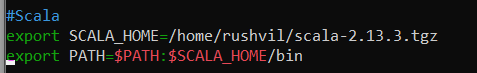
1. **Installing Scala, Spark and Hadoop**

* There are 2 methods to download Scala: directly from the web or using the Ubuntu command, which will download the most recent version of Scala.
* It is best to get it directly from Ubuntu since we don't need to extract anything; installation and extraction are handled for us.



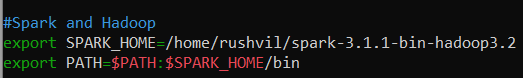
**Figure 10: Downloading Scala**

* After the installation is complete, the Scala version is verified and the path is specified in "nano /.bashrc".



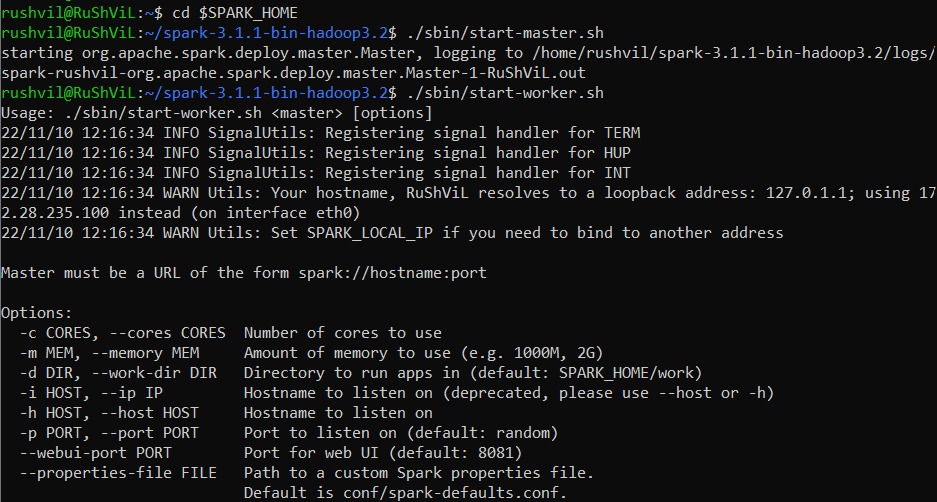
**Figure 11: Setting Path in Nano File**

* This changes the environment variable SCALA HOME to point to /home/rushvil/scala-2.13.3.



**Figure 12: Setting Path for Spark and Hadoop**

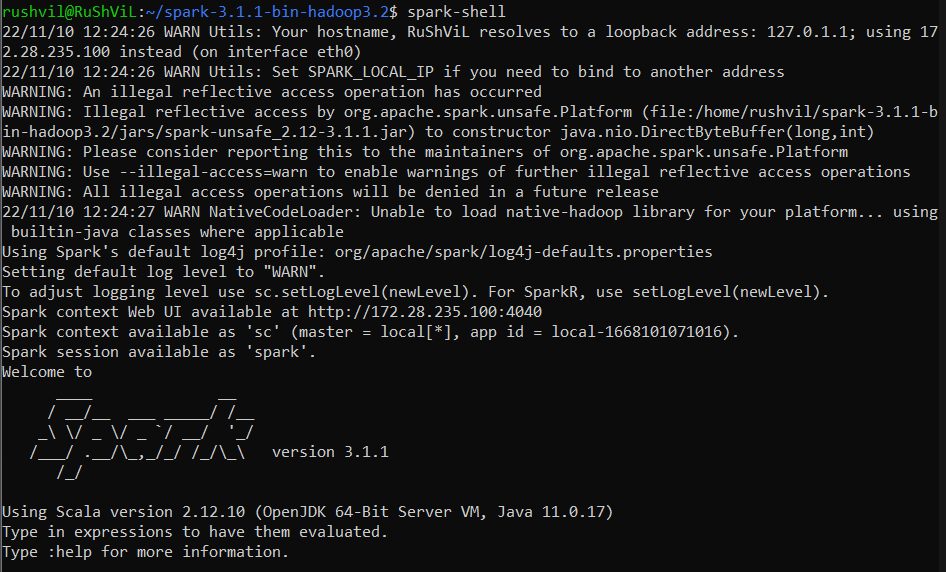
* In order to set environment variables, use the export command. The path "/home/rushvil/spark-3.1.1-bin-hadoop3.2" is being used as the value for the environment variable SPARK HOME in this instance. The Spark programme may most likely be found and utilised as a result of this.
* After setting all the path for Scala, Spark and Hadoop we need to source the file which is used to save the file.
* The source /.bashrc command reloads the .bashrc file's settings into the active shell. When making changes to the .bashrc file and wanting them to take effect right away without having to close and reopen your terminal, this is handy.



**Figure 13: Setting Directory for Spark and Starting Master and Worker**

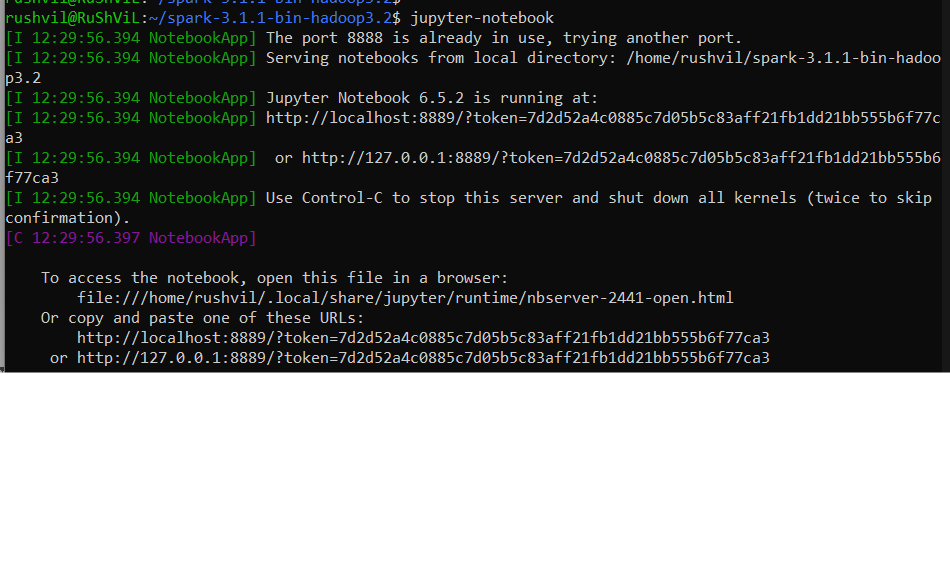
* The Spark installation directory is assigned as the current directory as a result.
* A master instance will be launched on the local system using this script. Additionally, it will add a new directory named "spark-events" to the current working directory.
* A Spark worker node is started using the start-worker.sh script. The URL of the master node is the only argument accepted by the script.

1. **Spark Shell**



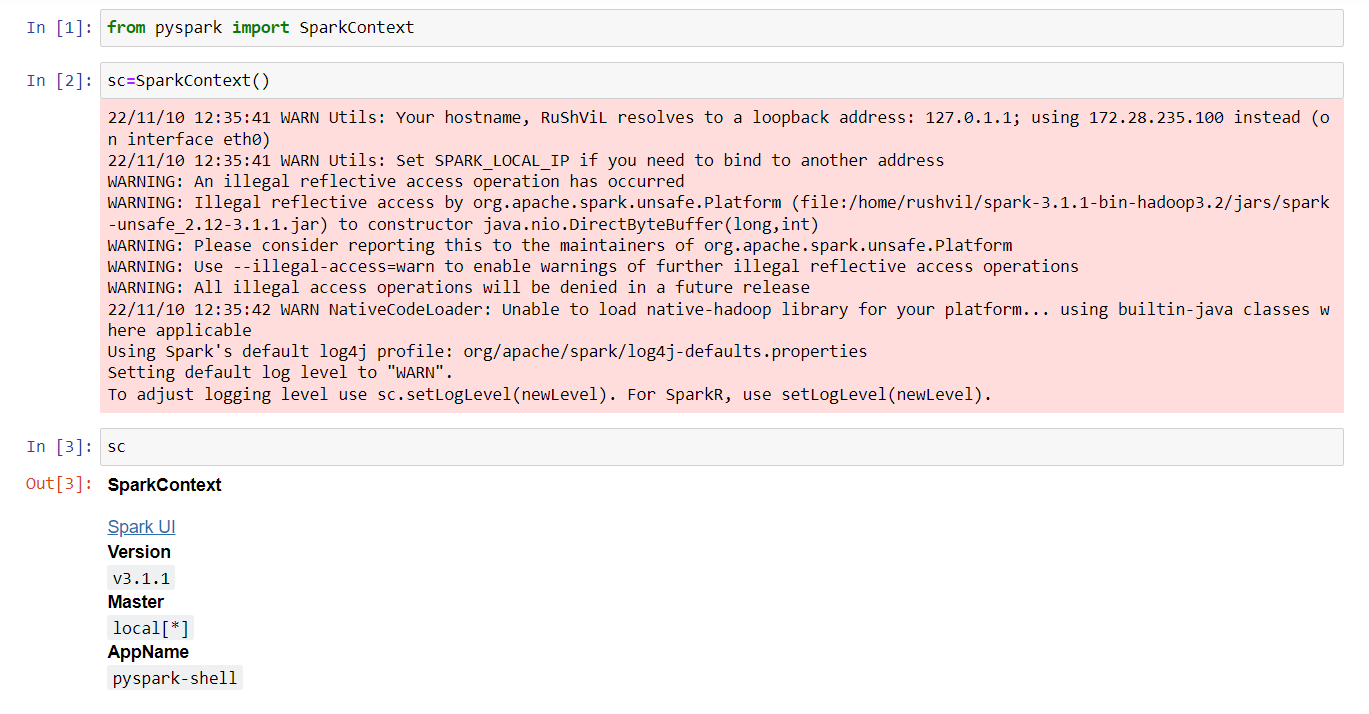
**Figure 14: Spark Shell**

* The command line interface for working with Spark is called Spark-shell. It offers a straightforward method of accessing the Spark UI and submitting Spark tasks.
* To run Spark commands and applications interactively, use the Spark shell. It has an integrated Scala interpreter and enables interactive data querying.

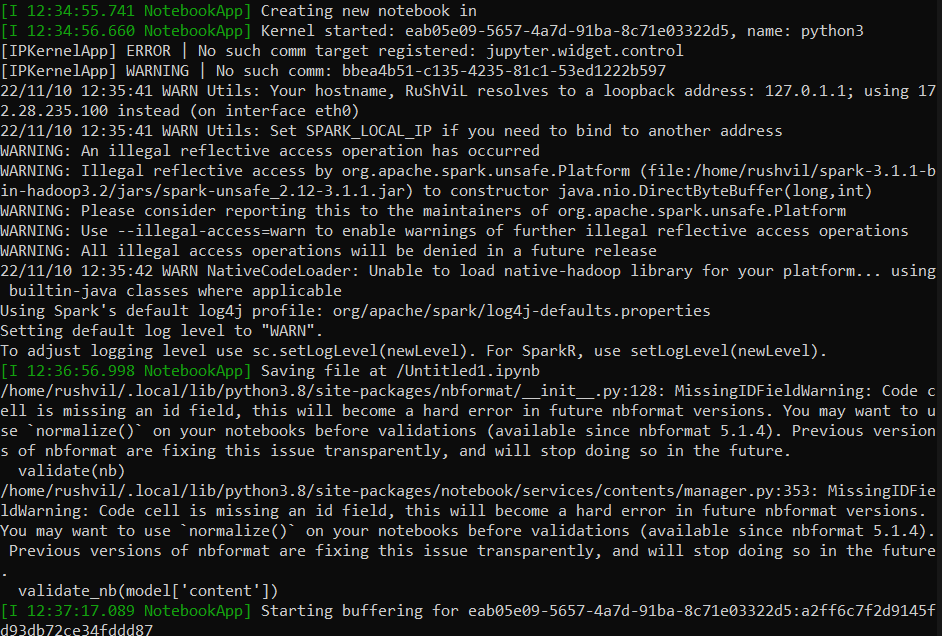


**Figure 15: Running Jupyter Notebook in Spark**

* From the code above, it is clear that when Jupyter code is executed, a URL for the notebook is returned.
* The Jupyter lab starts after copying the URL and launching it in the web browser. The logs will show up in the Ubuntu console if we do any actions in the Jupyter notebook.



**Figure 16: Jupyter Notebook and Version Check**

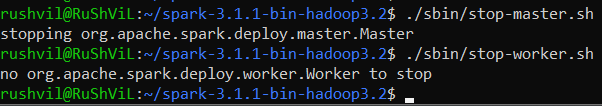
****

**Figure 17: Logs**

* As seen in the above graphic, whenever a job is carried out in the Jupyter notebook, the logs are appearing in Ubuntu.
* I first produced a new kernel, and as a result, the log for kernel construction is visible in the terminal.

1. **Stopping Master and Worker**

* If you don't stop the master and worker, you'll be charged more than normal.

****

**Figure 18: Stopping Master.sh and Worker.sh**

**References**

*Windows 10 tutorial: install WSL2 — Windows Subsystem for Linux 2*. (2020, November 9). YouTube. Retrieved November 10, 2022, from <https://www.youtube.com/watch?v=n-J9438Mv-s>

*What is .bashrc file in Linux? | DigitalOcean*. (2022, August 3). What Is .Bashrc File in Linux? | DigitalOcean. Retrieved November 10, 2022, from <https://www.digitalocean.com/community/tutorials/bashrc-file-in-linux>

*What is PySpark? | Domino Data Science Dictionary*. (n.d.). What Is PySpark? | Domino Data Science Dictionary. Retrieved November 10, 2022, from <https://www.dominodatalab.com/data-science-dictionary/pyspark>