A picture containing shape

Description automatically generated

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

Parva Pareshbhai Patel (002195186)

CPS Analytics, Northeastern University

ALY6110 : Data Management and Big Data

Dr. Mohammad Shafiqul Islam

November 11, 2022

**Introduction**

* In this project the main objective is to create the environment where big data knowledge can be practice in an actual setting. PySpark will be used to demonstrate the business use case.
* In PySpark, one will leverage parallel computation with large datasets, and get ready for high-performance machine learning. From cleaning data to creating features and implementing machine learning models, you'll execute end-to-end workflows with Spark.
* Single node cluster will be created in Ubuntu Virtual Machine in order to access the PySpark. During the project, code such as sudo, nano, alias and wget would be used.

**Creating Spark Cluster**

* Initially the Ubuntu Virtual Machine were installed. Once Ubuntu VM are configured java and python will be installed.
* The sudo apt update && upgrade code was used to install available upgrades of all packages currently installed on the system from the sources configured via sources. list file. New packages will be installed if required to satisfy dependencies, but existing packages will never be removed.

**Text

Description automatically generated**

Figure 1 : update and upgrade

**Installing JAVA**

* Now, JAVA JDK and JRE were installed so Spark Cluster can be used.
* Java Runtime Environment (JRE) is software that Java programs require to run correctly. The JRE is the underlying technology that communicates between the Java program and the operating system. Whereas, the JDK is a development environment for building applications, applets, and components using the Java programming language.

Text

Description automatically generated

Figure 2 : Installing JDE and JDK

* Now the version of JAVA is checked to know that it is the latest version.

**Text

Description automatically generated**

Figure 3 : Version of JAVA

**Installing Python**

* After installing JAVA packages for the python3 will be installed to run the Spark Cluster. The python3-pip package is used to install pip3, which is a tool for installing and managing Python3 packages and IPython is an interactive shell for Python that allows you to type Python code and execute it immediately. Moreover, Jupiter notebook is also installed.

**Text

Description automatically generated**

Figure 4 : Installing python3 and Jupiter notebook

* The nano command was use to open the editor. Through this editor the hidden bash file was accessed which was created when the terminal runs. In this hidden bash file the path for the Jupiter notebook, Scala and Hadoop would be given.



Figure 5 : nano and source command

* Once using the nano command the editor was open now the path for the Jupiter notebook was exported.



Figure 6 : Set alias for Jupiter Notebook

* Now to save the changes made in the editor shell the source ~/.bashrc command was used.

**Installing Scala**

* Scala is a modern multi-paradigm programming language designed to express common programming patterns in a concise, elegant, and type-safe way. Scala is designed to interoperate well with the popular Java Runtime Environment (JRE).
* Now to install Scala the wget command is used. Here the link to download Scala is given in the wget command. So, executing this command the zip file of the Scala would be downloaded.

Text

Description automatically generated

Figure 8 : Downloading Scala

* As the downloaded file was zip file, it was necessary to unzip it in order to use it. To do so, tar xvf scala-2.13.3.tgz command is used. The tar command will unzip the xvf (x: extract, v: verbose, f: file) Scala file.
* Once the file was extracted its path was set into the hidden bash file using the export command and finally to save the changes in bash file source ~/.bashrc command was executed.

Text

Description automatically generated

Figure 9 : Set Variable for Scala

* Now to check the version of Scala the command Scala -version was used.

**Installing Spark with Hadoop**

* Now to install Spark with Hadoop the wget command is used. Here the link to download Spark with Hadoop is given in the wget command. So, executing this command the zip file of the Hadoop would be downloaded.

Text

Description automatically generated

Figure 10 : Downloading Spark with Hadoop

* As the downloaded file was zip file, it was necessary to unzip it in order to use it. To do so, tar xvf spark-3.1.1-bin-hadoop3.2.tgz command is used. The tar command will unzip the xvf (x: extract, v: verbose, f: file) Hadoop file.
* Once the file was extracted its path was set into the bash file using the export command. Further, source ~/.bashrc command was used to save the changes in the bash file.

Text

Description automatically generated

Figure 11 : Setting Path for the Spark.

**Creating Cluster**

* To activate the cluster first the current working directory of the Spark was accessed using cd $SPARK\_HOME command. After opening the directory of the Spark the code sbin/start-master.sh is executed which will start a master instance on the machine. Whereas on. sbin/start-workers.sh code will start a worker instance on each machine specified in the conf/workers file.

A picture containing graphical user interface

Description automatically generated

Figure 12 : Creating master and worker.

* The $spark-shell is a command line tool used to interact with the Spark cluster. It allows you to submit Spark jobs, monitor their progress, and debug them if necessary.

Text

Description automatically generated

Figure 13 : Spark on Shell

* Now to run the jupyter notebook the code jupyter notebook is executed which in return provide some URL. When this URL is copied into the browser, it will direct use to the jupyter lab where the code can be executed.

Graphical user interface, text

Description automatically generated

Figure 14 : Jupyter notebook

Graphical user interface, text, application, Word

Description automatically generated

Figure 15 : Code in jupyter Notebook

* Now whenever the code is performed in the jupyter notebook the logs of that code will be dropped into the terminal.

Text

Description automatically generated

Figure 16 : Jupyter notebook logs in terminal

* Now when all the work is done then it is necessary to stop the master and worker cluster. The main reason is to free up resources on the cluster so that other applications can use them. Additionally, stopping the cluster will also stop any running applications on the cluster.

Text

Description automatically generated

Figure 17 : Stopping the master and worker.

**Reference :**

* 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>
* Verma. (2022, August 3). *What is .bashrc file in Linux? | DigitalOcean*. What Is .Bashrc File in Linux? | DigitalOcean. Retrieved November 11, 2022, from <https://www.digitalocean.com/community/tutorials/bashrc-file-in-linux>
* Ashir. (n.d.). *What is Kubernetes cluster? What are worker and master nodes?* Educative: Interactive Courses for Software Developers. Retrieved November 11, 2022, from <https://www.educative.io/answers/what-is-kubernetes-cluster-what-are-worker-and-master-nodes>