**Hadoop Single Node Setup:**

Hadoop can be installed in windows and Linux. Since Predominantly the industry usage is in Linux, we will be setting up environment in Linux.

**VMware Workstation Player Setup:**

Open your browser and search for VMWare Workstation player or [Click Here](https://www.vmware.com/products/workstation-player/workstation-player-evaluation.html) . Select the latest Version if your Mac User You should download VMware Fusion. Install the VM Ware workstation application from downloaded file.

We are downloading Ubuntu 20.04 LTS on VMware Workstation Player:

Go to Download section and choose Ubuntu Desktop version of ubuntu from <https://releases.ubuntu.com/20.04/>, or search Ubuntu 20.04 LTS the download Process take few minutes.

Graphical user interface, diagram, application

Description automatically generated

Now Click on **Create a New Virtual machine** and Select I will Install Operating System Later and Click on **Next**

Graphical user interface, text, application

Description automatically generated

Select the **guest Operating System as Linux and Version is Ubuntu 64 bit**

**Graphical user interface, text, application, email

Description automatically generated**

**Name the Virtual Machine:**

Name virtual machine as DSCI5350\_Training and Select file browser location or leave it default to avoid any errors.

Graphical user interface, text, application, email

Description automatically generated

**Specify Disk Capacity:**

Recommended Size is 20GB, but to be on safe side allocate 50GB and **select split virtual disk into multiple files**

Graphical user interface, text, application, email

Description automatically generated

**Ready to create Virtual Machine:**

Click on Customize Hardware and Set your RAM to 4GB, if your laptop supports multi core allocate 2 cores and Select New CD/DVD (SATA) Option, here in connections select Use ISO Image file and Browse the Downloaded ISO file, Go to Display and Disable Accelerated 3D graphics, and close configuration.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

And Click on Finish. You can Play your Virtual Machine by clicking on Play Button or Power on Top.

Start Virtual Machine, Click on Try or Install Ubuntu. Click On Install

Select the Language, keyboard layout and click Continue.

Table

Description automatically generated with low confidence

Click on Normal Installation, Select Download update while Installing Ubuntu and Install Third-Party Software for graphics and Wi-Fi Hardware and additional media formats.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

In Installation Type: Select Erase disk and Install Ubuntu, don’t stress out this step will not harm your system.

When the dialogue box of write changes to disks opens just click Continue.

Graphical user interface, text, application, email

Description automatically generated

Next Select the Location and your Create Username Password. It takes few moments to install. Click or skip online account and location enabling.

Graphical user interface, text, application

Description automatically generated

**Hadoop and Hive Installation:**

Install Editors command: sudo apt-get install vim

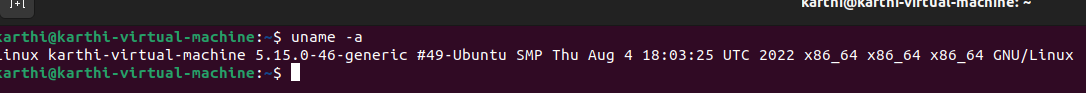
Text

Description automatically generated

Command ‘ls -a’ is used to view the hidden files also.

uname -a

Is a command to note which type of system configuration is they’re that is 64 or 32 bits. Here my system is 64-bit configuration.



This is Single node setup because it is on your system only, Hadoop is distributed technology it can be setup in more than one machine, A machine is called has Node.

Search For ‘Hadoop Download’ in your Browser:

You will land on the page like this

Graphical user interface, text, application, email

Description automatically generated

Click on the link their, select Hadoop 2.9.1 version and download Hadoop-2.9.1.tar.gz transfer downloaded file from downloads to home.

Search For JDK 1.8 or greater Version and download that, transfer the downloaded file from downloads to home.

A screenshot of a computer

Description automatically generated

Download Apache Hive from archive.apache.org we are going with version 3.1.1. Hive comes with Embedded Database called Derbie (Which is not Centralized). This is When connected has cluster we can’t access the table created in some other machine because they have their own meta data. So, we will be using non embedded database My SQL we will be using that has meta-store.

Command to install MySQL: sudo apt-get install mysql-server



Text

Description automatically generated

Hive uses MySQL or Oracle just to store meta store other than that there is no other relation, the Queries are run by map-reduce but not by MySQL or Oracle (RDBMS).

Download the MySQL Connector Jar file from maven repository

Next extract the hive downloaded file via GUI or Via Command Line ‘tar -zxvf’

And paste the downloaded jar file in the ‘lib’ of the extracted hive folder.



Install an Editor

Text

Description automatically generated with low confidence

Now In ‘Hive’ folder go into conf and create hive-site.xml file with the code given below



**Spark standalone Setup**

Open your comfortable Browser and Search for Apache Spark Download, that will take us to first link spark.apache.org, And here we can download Recent Version or Recent Version minus 1 can be downloaded.

Here Spark standalone is not same having single node, Spark processing engine is sitting on Linux not on Hadoop. For practice everything is similar



Now Open the .bashrc file in home directory .bashrc is a hidden file it will be seen in terminal if we use command ls -a this is file to specify path in Linux. Enter ESC+SHIFT+G or go the last line, Enter this code over their and save the file(ESC and enter :wq ).

Text

Description automatically generated

Execute the file by writing ‘source .bashrc’



Go to Hadoop extracted folder, navigate to etc then Hadoop. Open the following list of files in the editor.



We modify the files

Core-site.xml which defines Name node properties

And paste these lines in between configure:

<property>

<name>fs.default.name</name>

<value>hdfs://localhost:50000</value>

</property>

This is for single node in multi node we configure this with real ip-Address. we configure node manager by yarn-site.xml, paste this code in between configuration.

<property>

<name>yarn.nodemanager.aux-services</name> <value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name> <value>org.apache.hadoop.mapred.ShuffleHandler</value>

</property>

<property>

<description>The hostname of the RM.</description>

<name>yarn.resourcemanager.hostname</name>

<value>localhost</value>

</property>

<property>

<description>The address of the applications manager interface in the RM.</description>

<name>yarn.resourcemanager.address</name>

<value>{yarn.resourcemanager.hostname}:8032</value>

</property>

hdfs-site.xml is used to configure HDFS related properties for example block size, replication factor.

Graphical user interface, text, application, email

Description automatically generated

Next configure mapred-site.xml

<property>

<name>mapreduce.framework.name</name>

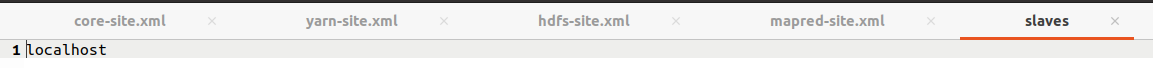
<value>yarn</value>

</property>

Create normal file slaves, using these commands

vi etc/hadoop/slaves

localhost



Now set the environment variables in following env.sh file, Add a export jdk line in the below files

Command: export JAVA\_HOME=/home/dsci5350/jdk



Text

Description automatically generated

Install SSH : ssh is installed to set password less ssh to localhost and to slaves.



Text

Description automatically generated

**Formatting is only done at time of installation not every time**

**A picture containing text

Description automatically generated**

hadoop-dir file is created in their namenode-dir host meta data and data node host the rest blocks of data.

Command: jps Java process Status

To check all the five demons are running

Text

Description automatically generated

Web UI

Name Node - <http://localhost:50070/>

Graphical user interface, text, application, email

Description automatically generated

Resource Manager - <http://localhost:8088/>

Graphical user interface, application, Word

Description automatically generated