Configuration Manual

MSc Internship
Cyber Security

Ritesh Naresh Gohil
Student ID: X18205836

School of Computing
National College of Ireland

Supervisor: Mr. Vikas Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	**Ritesh Naresh Gohil**		
Student ID:	**X18205836**		
Programme:	**Cyber Security**	**Year:**	**2020**
Module:	**MSc Internship**		
Lecturer:	**Mr. Vikas Sahni**		
Submission Due Date:	**17/08/2020**		
Project Title:	**Fast and accurate classfication of threats in IDS using Distributed Machine Learning techniques.**		
Word Count:	**1038**	**Page Count: 15**	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.
I agree to an electronic copy of my thesis being made publicly available on NORMA the National College of Ireland's Institutional Repository for consultation.

Signature: Ritesh Naresh Gohil
……

Date: 17/08/2020
……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Ritesh Naresh Gohil
Student ID: X18205836

1. Introduction

This configuration manual will provide in depth information about configuration and implementation of these Research Project which includes system specifications and software requirements. Python is used for Traditional Machine Learning whereis Hadoop Spark framework and Pyspark language is used for classification of threats in IDS using Distributed Machine Learning Techniques.

2. Project Setup

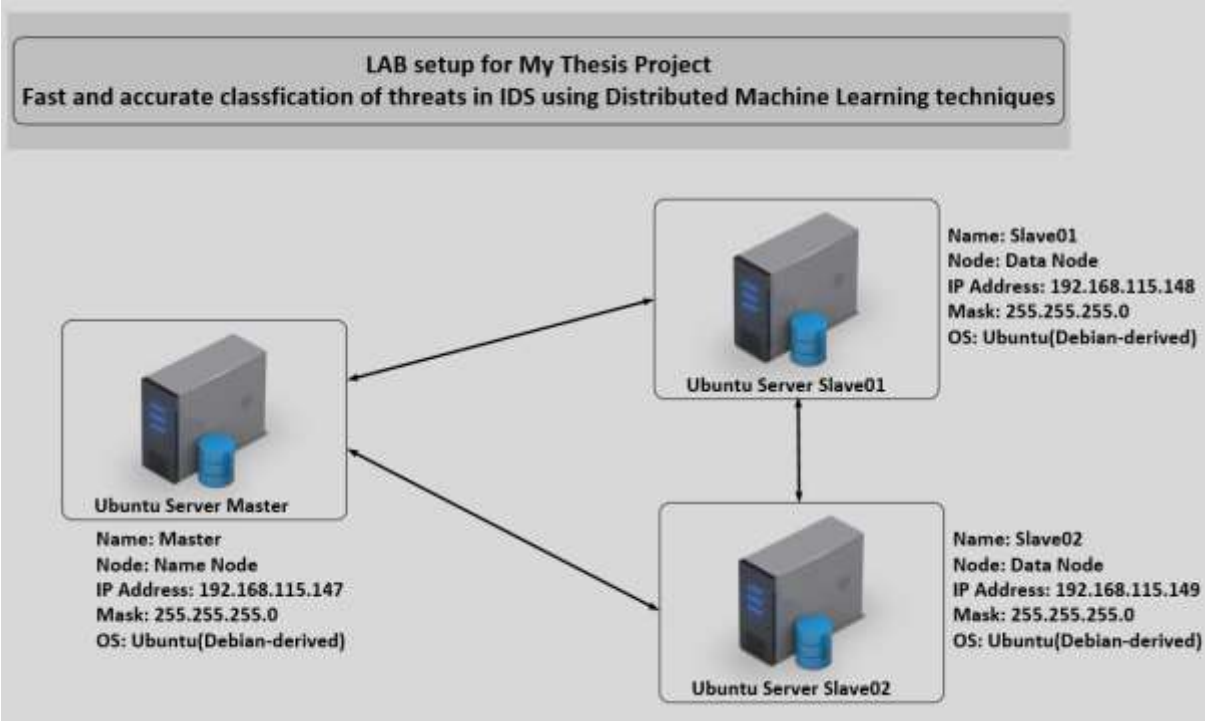

Figure 1: Project Setup

2.1 System Requirements

Ubuntu Server (Master)	Ubuntu Server (Slave01)	Ubuntu Server (Slave02)
4 GB RAM	4 GB RAM	4 GB RAM
40GB HDD	40GB HDD	40GB HDD
Number of Processor 2	Number of Processor 2	Number of Processor 2
Number of Cores per Processor 2	Number of Cores per Processor 2	Number of Cores per Processor 2
Network Adapter: NAT	Network Adapter: NAT	Network Adapter: NAT

Table 1: System Requirements

3. Software and Tools Used

- **VMWare Workstation 15.0 Pro**
- **Anaconda 2020**
- **Python 3.8**
- **Jupyter Notebook 6.0**
- **Pyspark**

4. Configuration Steps

Below are the steps followed for Hadoop Distributed File System (HDFS) installation and configuration:-

Step 1: Installation of Java JDK version 8.
***Command*: sudo apt-get install openjdk-8-jdk**

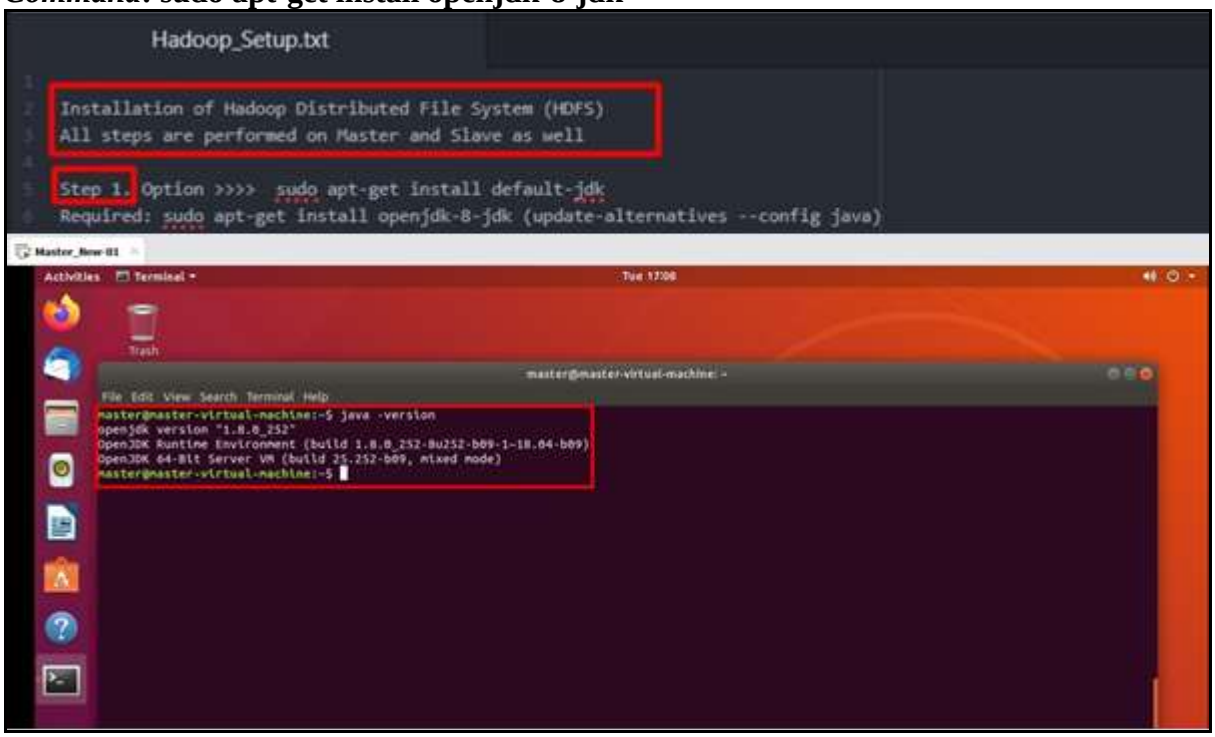

Figure 2: Validate Java Version

Step 2: Creating a user for Hadoop group whose name is hduser.

Figure 3: Created hduser

Step 3: Add all hostname and IP addresses in /etc/hosts file
***Command*: vim /etc/hosts**

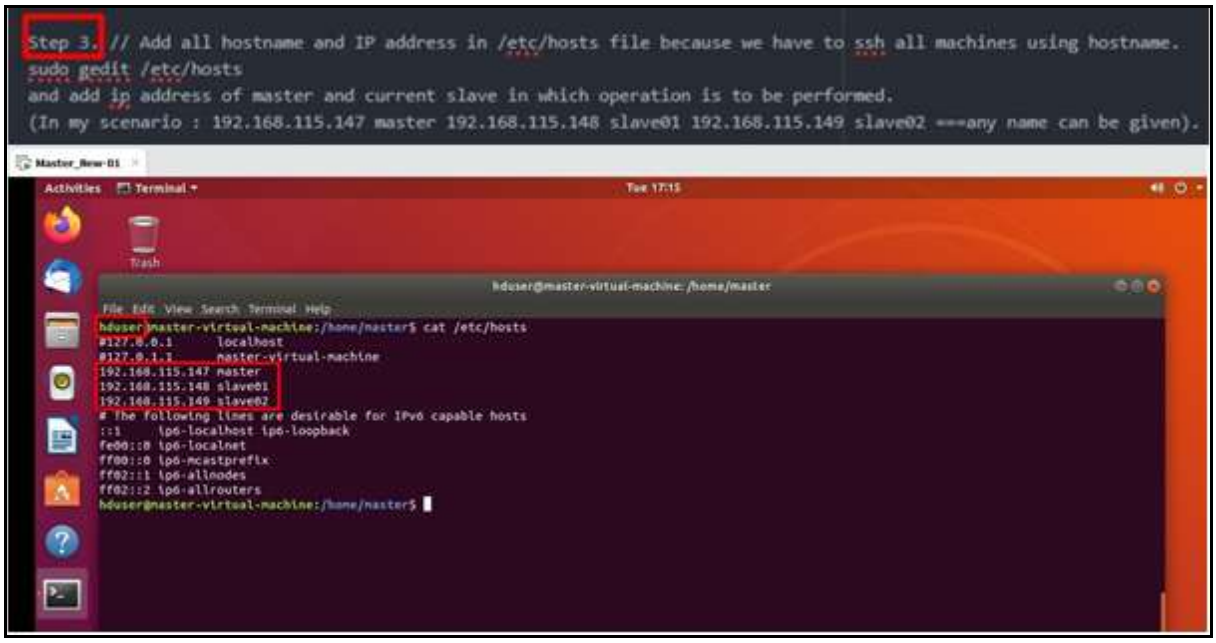

Figure 4: Entry of hosts in /etc/hosts file

Step 4: Install OpenSSH and Share SSH public key to other machines
***Command*: sudo apt-get install openssh-server**

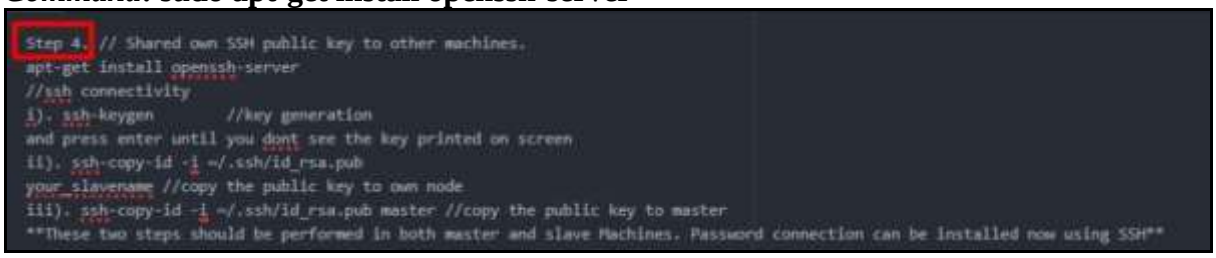

Figure 5: Generate and share SSH public key

Step 5: Global variable JAVA_HOME, HADOOP_HOME, SPARK_HOME in .bashrc file.

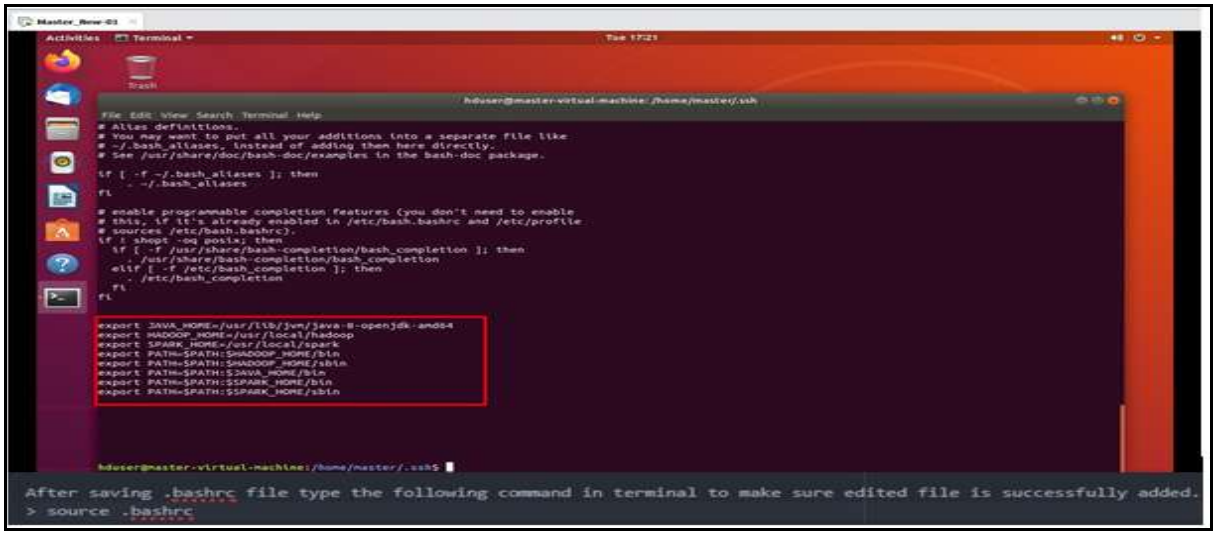

Figure 6: .bashrc file

Step 6: Install Hadoop and verify $HADOOP_HOME.
***Command*: sudo wget https://archive.apache.org/dist/hadoop/common/hadoop-3.2.0/**

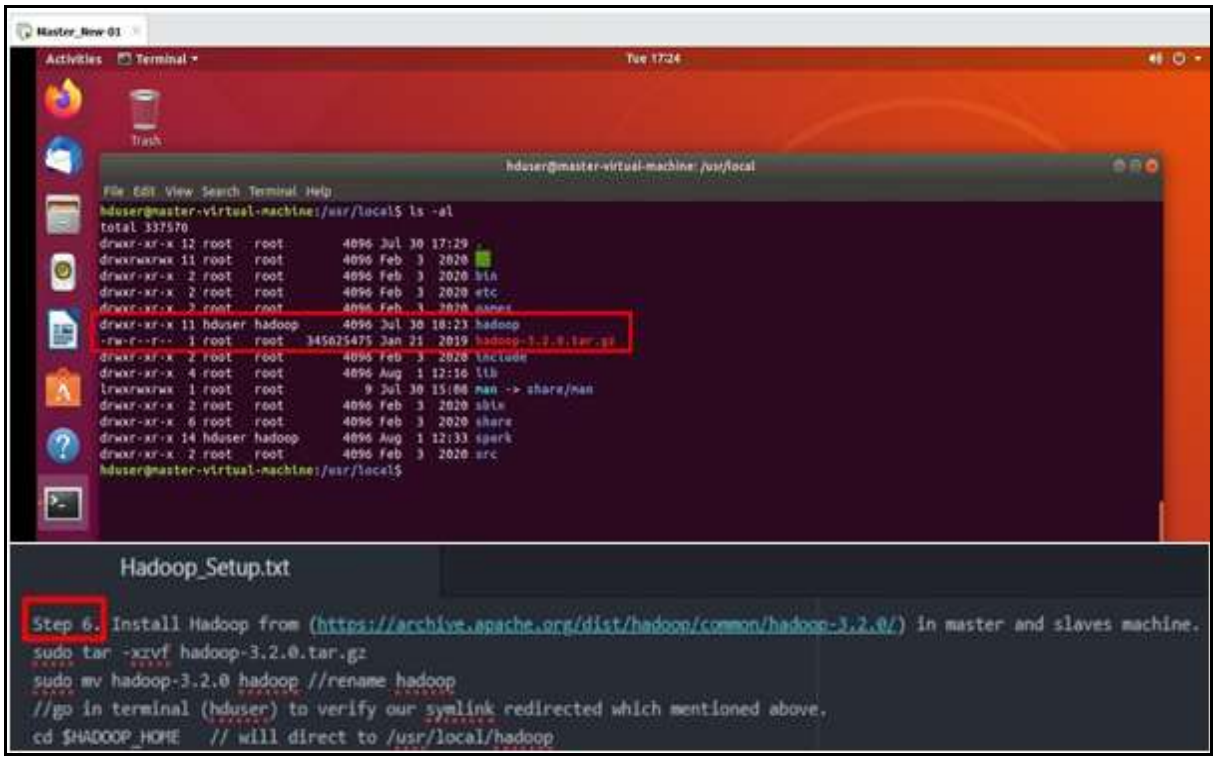

Figure 7: Install Hadoop

Step 7: edit hadoop-env.sh file which is located at /hadoop/etc/hadoop/.
***Command*: vim /hadoop.hadoop-env.sh**

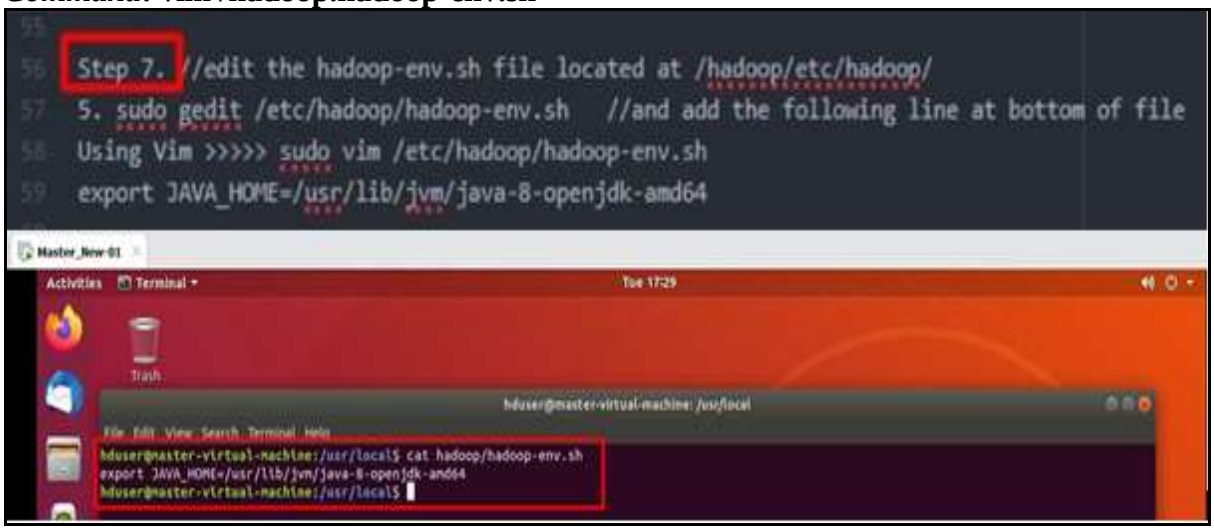

Figure 8: hadoop-env.sh

Step 8: Make directories and provide permissions for Hadoop hduser.

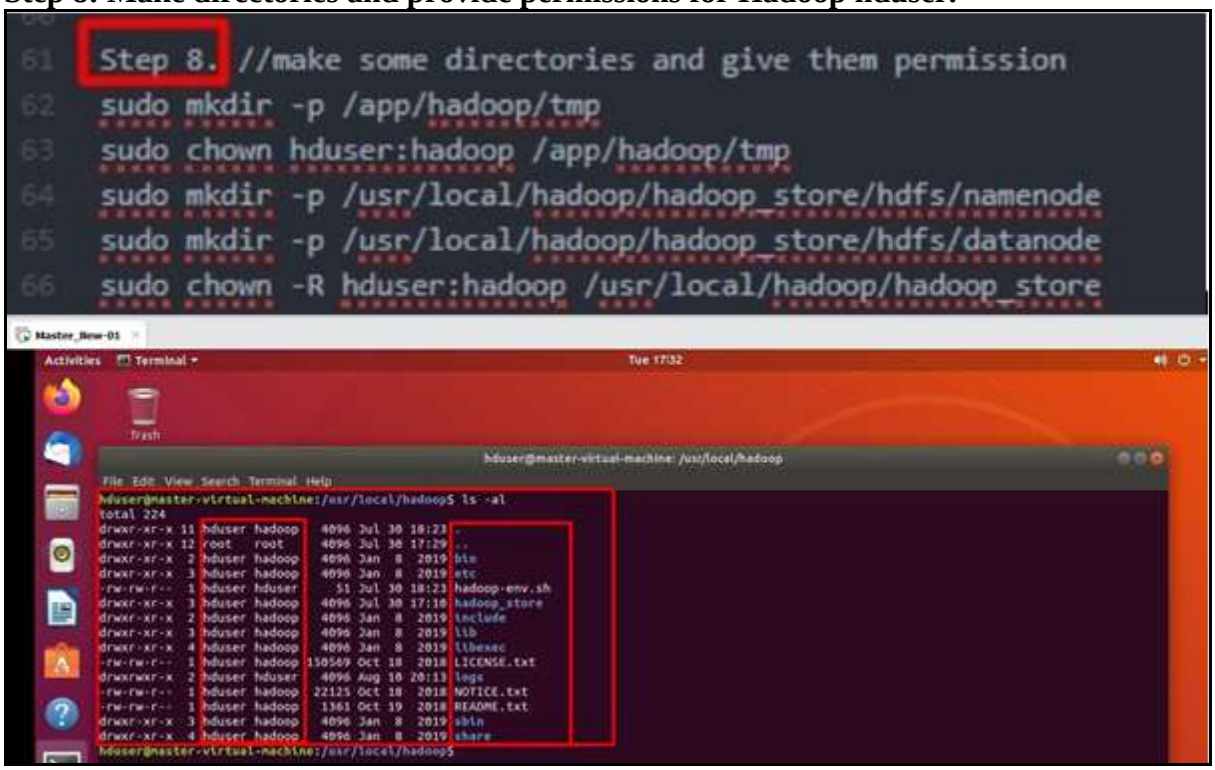

Figure 9: Permission for Hadoop hduser

Step 9: Installation of Spark (Master and Slaves).

Command:

sudo wget https://mirrors.estointernet.in/apache/spark/spark-3.0.0/spark-3.0.0-bin-hadoop2.7.tgz

sudo tar -xzvf spark-3.0.0-bin-hadoop2.7.tgz

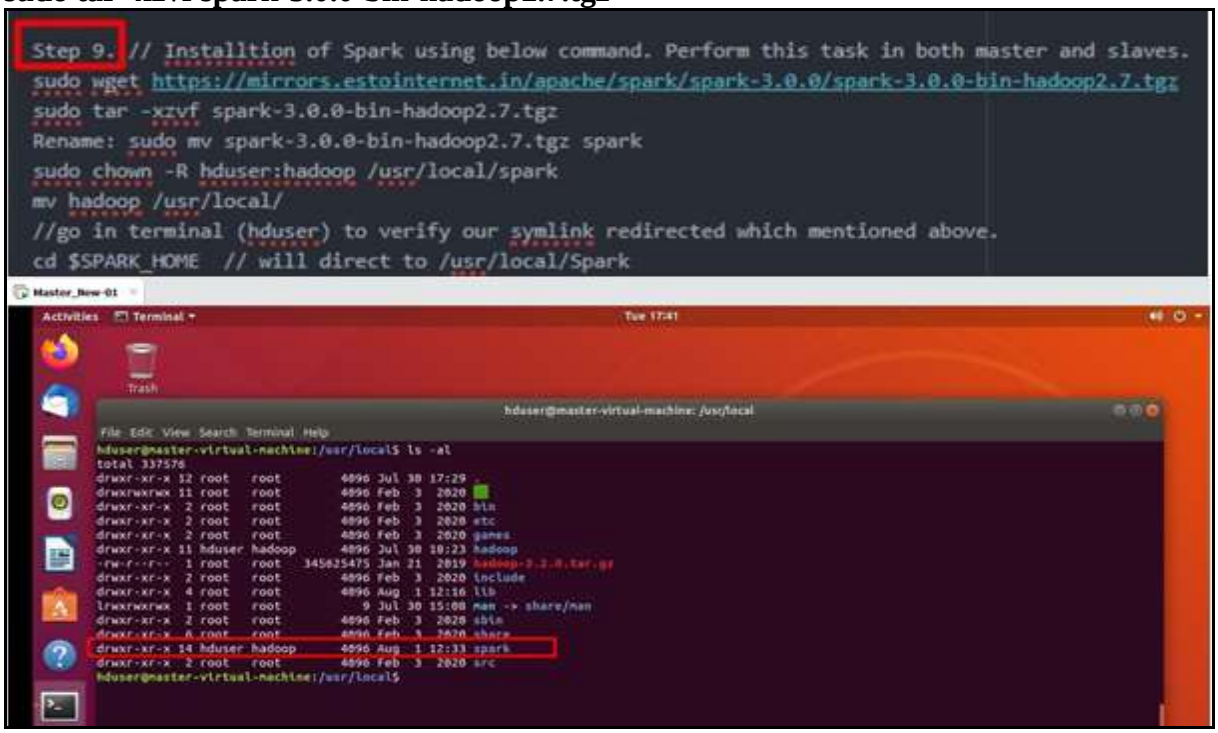

Figure 10: Spark Install

Step 10 to 14: Modification in Hadoop main config (xml) file.

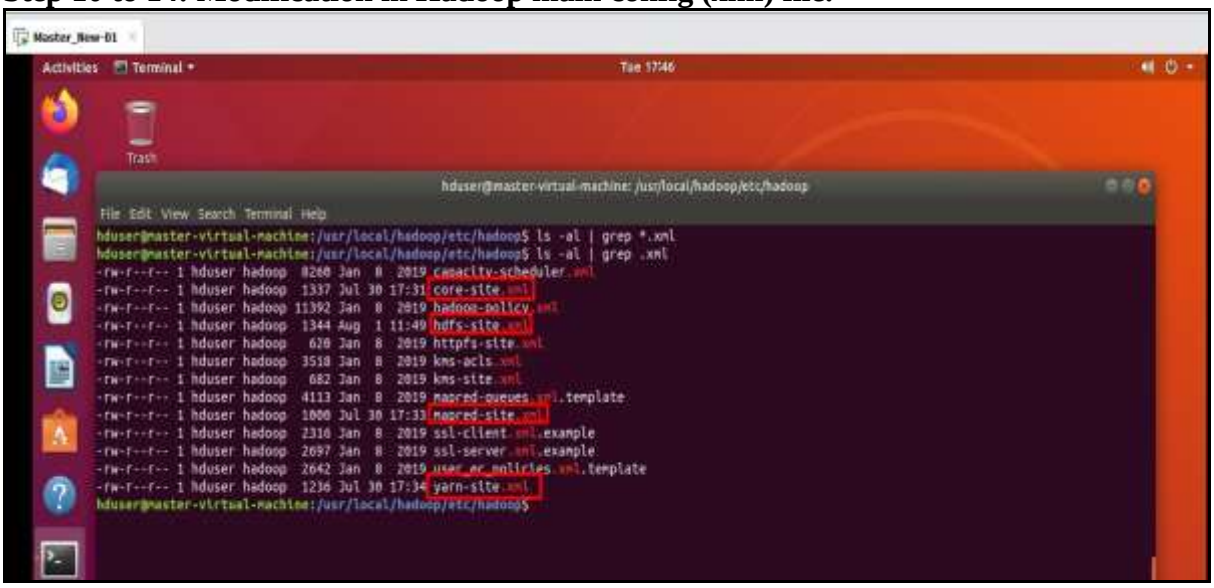

Figure 11: List of Hadoop Configuration File

Make changes in **core-site.xml** file under /usr/local/Hadoop/etc/Hadoop directory.
***Command*: vim core-site.xml**

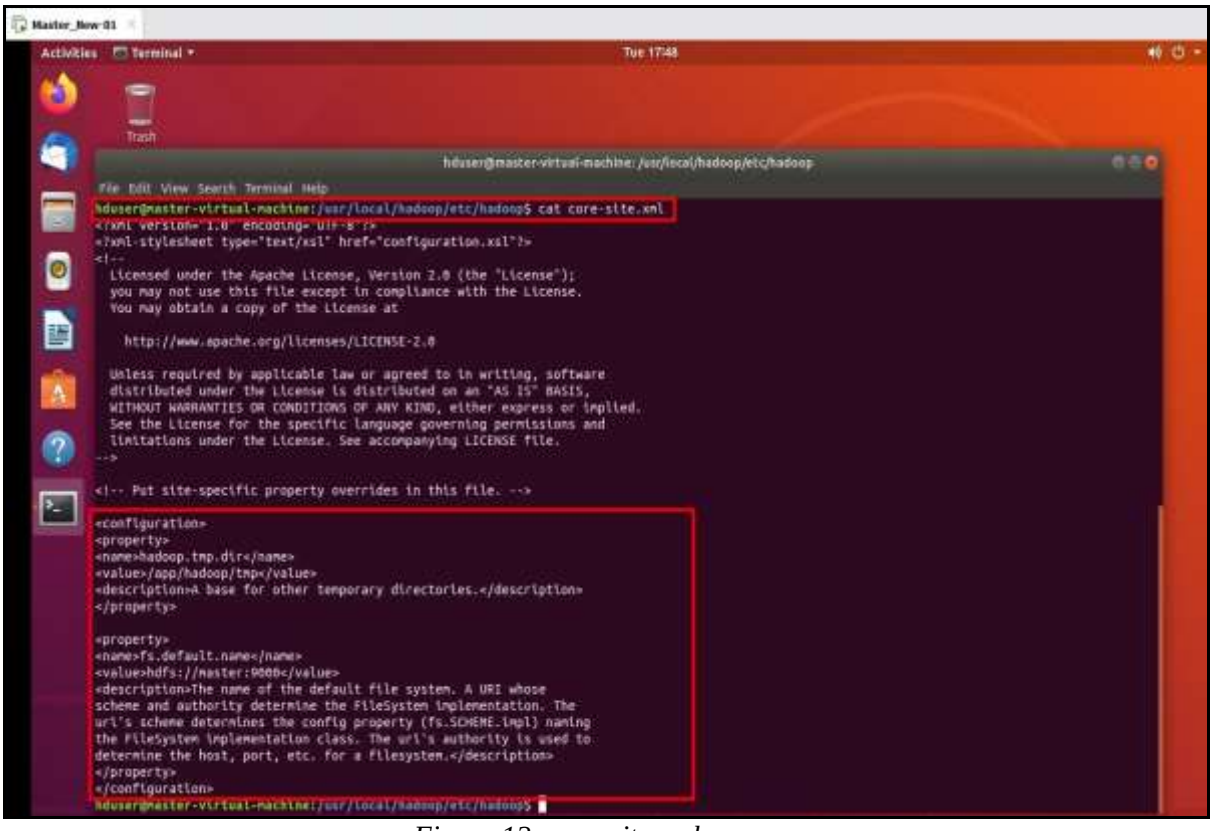

Figure 12: core-site.xml

Make changes in **hdfs-site.xml** file under /usr/local/Hadoop/etc/Hadoop directory.
***Command*: vim hdfs-site.xml**

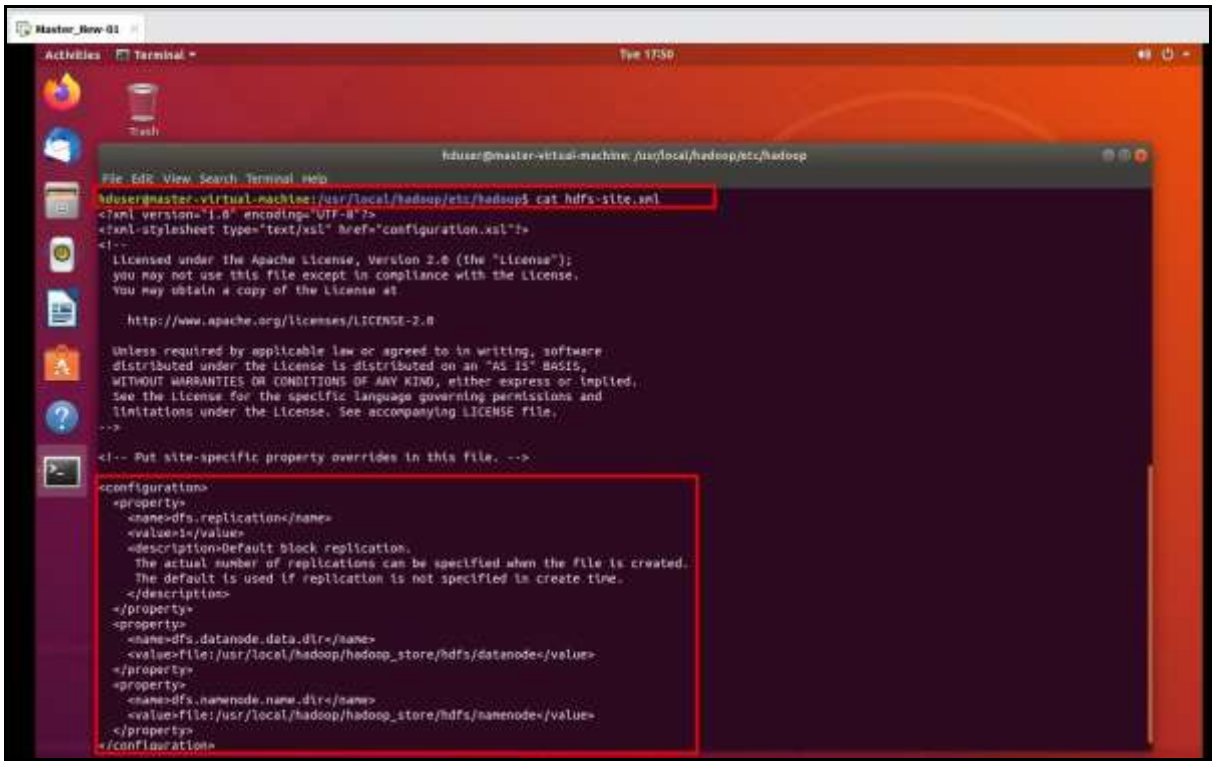

Figure 13: hdfs-site.xml

Make changes in **mapred-site.xml** file under /usr/local/Hadoop/etc/Hadoop directory.
***Command*: vim mapred-site.xml**

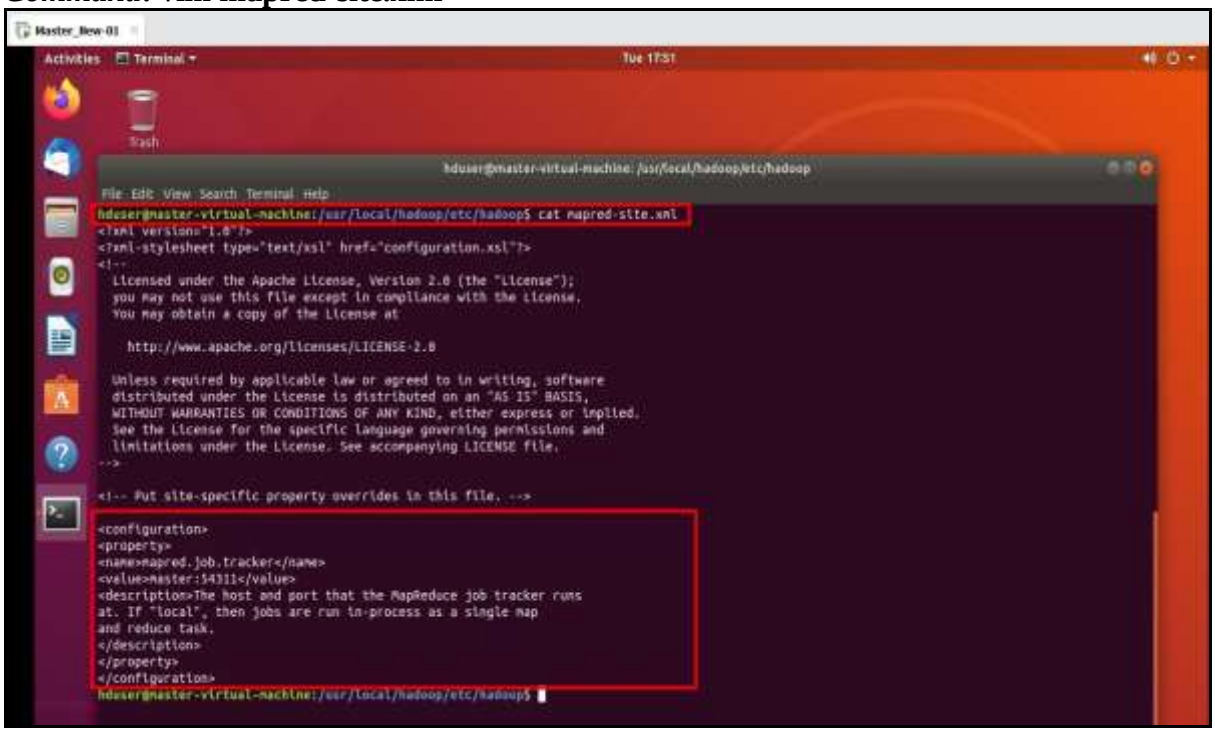

Figure 14: mapred-site.xml

Make changes in **yarn-site.xml** file under /usr/local/Hadoop/etc/Hadoop directory.
***Command*: vim yarn-site.xml**

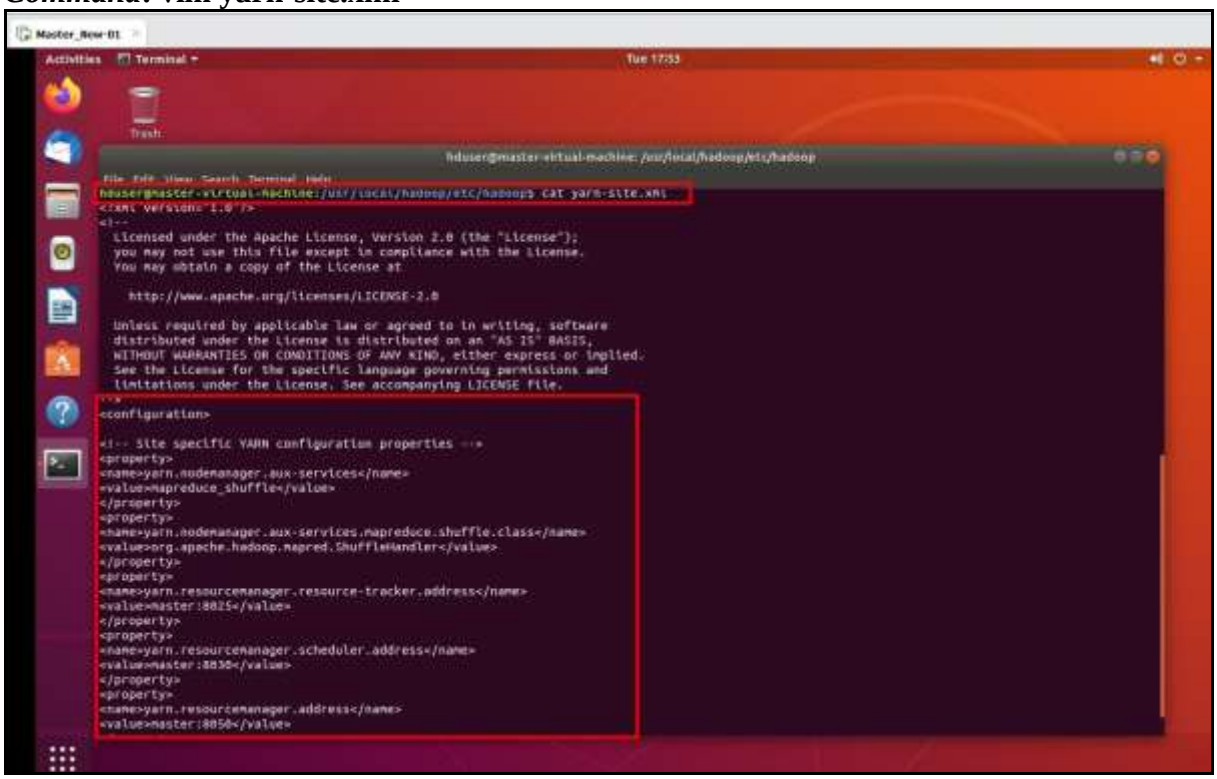

Figure 15: yarn-site.xml

5. Data Processing and Visualisation of Dataset

Data set contains 23 different types of attack which can be mapped in 4 various categories i.e Dos, User to root attack (U2R), Remote to local attack (R2L) and probing attack. Also, prepared column data which include protocol types and flag**.**

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import OneHotEncoder

#preparing column data
cols="""duration,
protocol_type,
service,
flag,
src_bytes,
dst_bytes,
land,
wrong_fragment,
urgent,
hot,
num_failed_logins,
logged_in,
num_compromised,
root_shell,
su_attempted,
num_root,
num_file_creations,
num_shells,
num_access_files,
num_outbound_cmds,
is_host_login,
is_guest_login,
count,
srv_count,
```

```
# renaming attack types
attack_types = {
    'normal': 'normal',
'back': 'dos',
'buffer_overflow': 'u2r',
'ftp_write': 'r2l',
'guess_passwd': 'r2l',
'imap': 'r2l',
'ipsweep': 'probe',
'land': 'dos',
'loadmodule': 'u2r',
'multihop': 'r2l',
'neptune': 'dos',
```

Figure 16: Data Processing

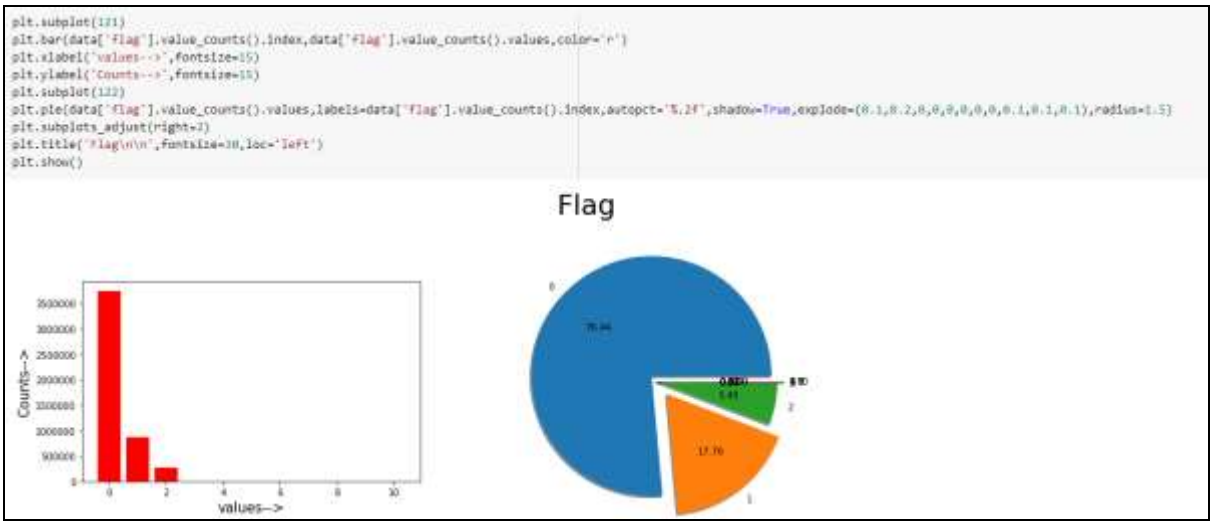

Figure 17: Data Visualisation

6. Implementation and Validation by Hadoop GUI portal

- **Command for loading data in Hadoop**
  ***Command*: hadoop fs -put *.csv /IDS_data**

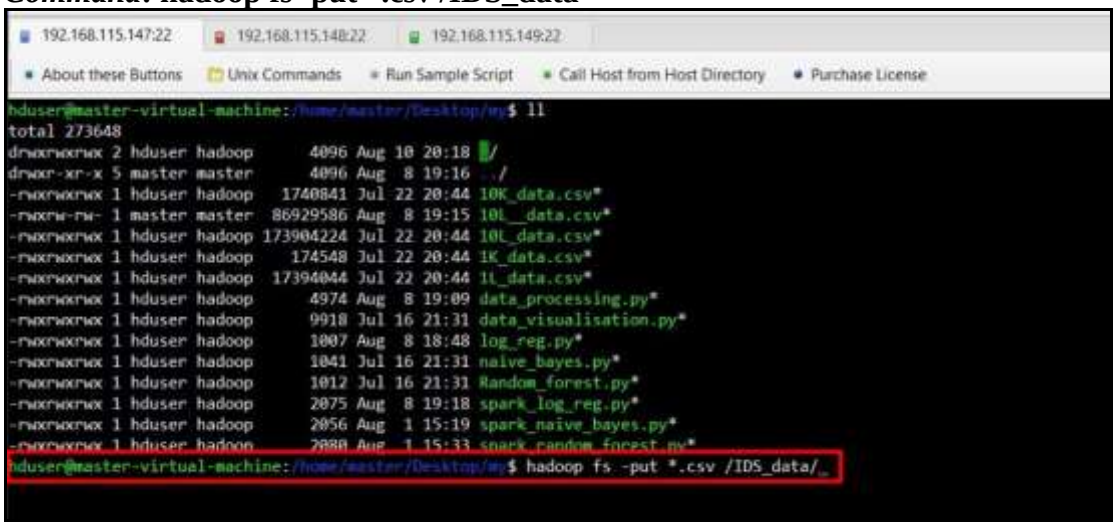

Figure 18: Command for loading data in Hadoop

- **Starting Hadoop (HDFS)**
  ***Command*: start-dfs.sh**

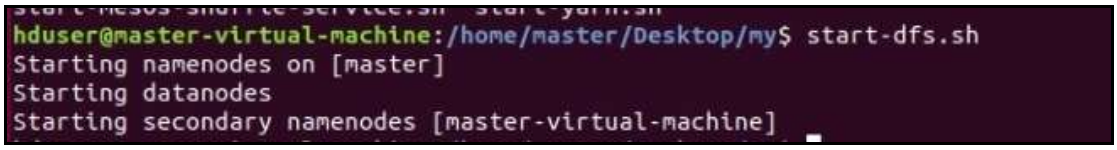

Figure 19: Command for satring Hadoop

- **Starting Spark for Master and Slaves (User: hduser)**
  ***Command*: start-master.sh**

Figure 20: Command for starting Spark

- **Hadoop GUI portal (http://ip_address:9870)**

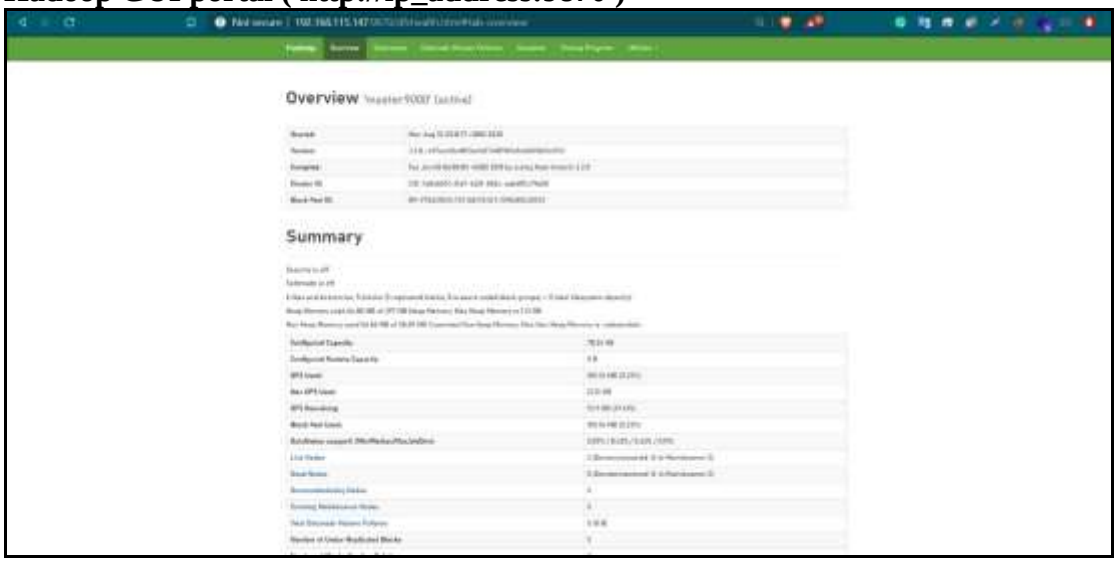

Figure 21: Hadoop GUI Portal

- **Data Node information of Slave01 and Slave02 (http://ip_address:9870)**

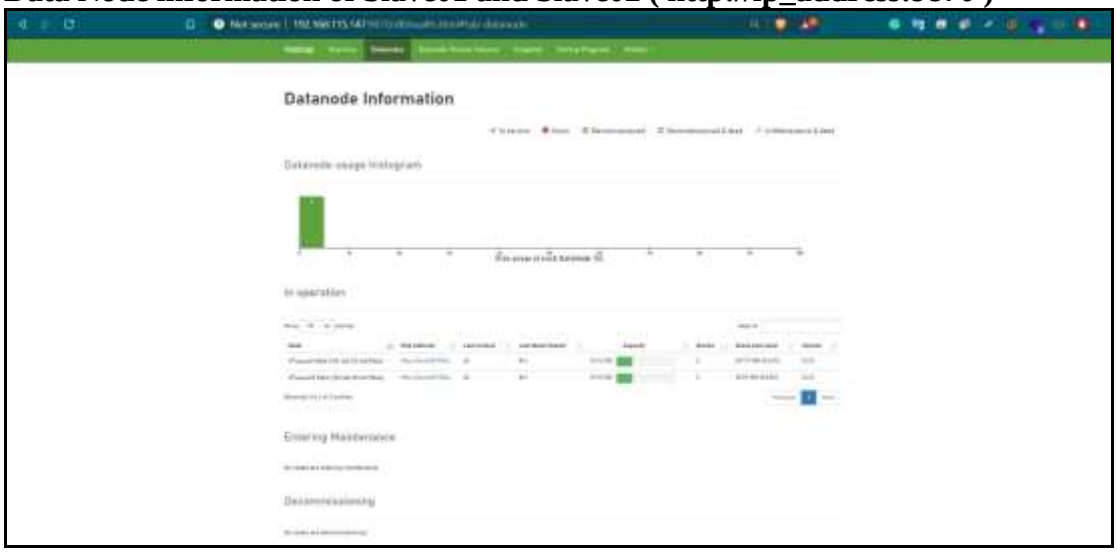

Figure 22: Data Node Information

- **Validating Spark GUI portal and Justification of Slave - Masters connection (http://ip_address:8080)**

Figure 23: Spark GUI Portal

- **Job completed by logistic regression algorithm. (http://ip_address:8080)**

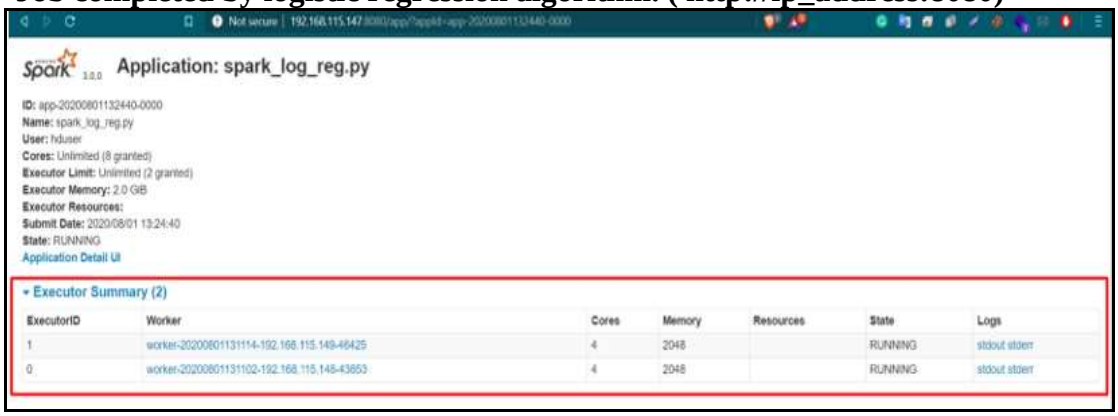

Figure 24: Spark Logistic Regression

7. Steps for executing command to analyse output.

- **Execute python file data_processing.py.**
  python3 **data_processing.py**

After executing **data_processing.py.** It will create 1k, 10k, 1L and 10L csv. These files used in to analyse Logistic Regression, Naïve Bayes and Random Forest results.

List of Algorithm used:

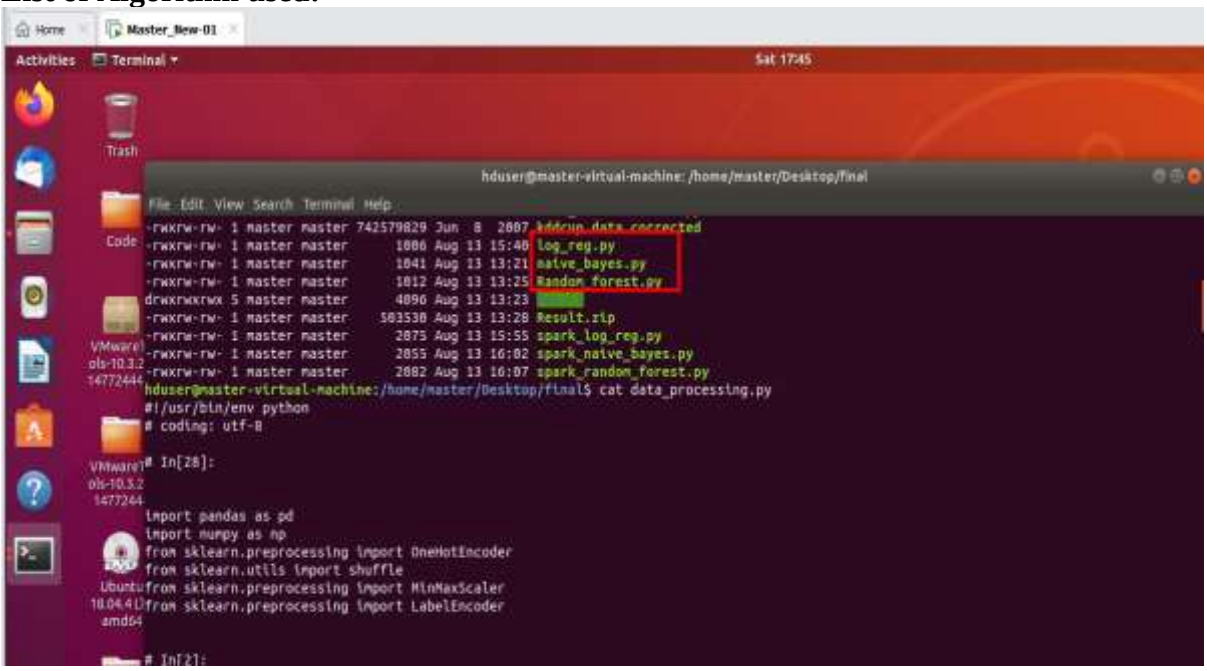

Figure 25: Algorithm Scripts

To perform every algorithm output, I must define 1k, 10k, 1L and 10L csv file to see result individually for both machine learning techniques. As per below snapshot:

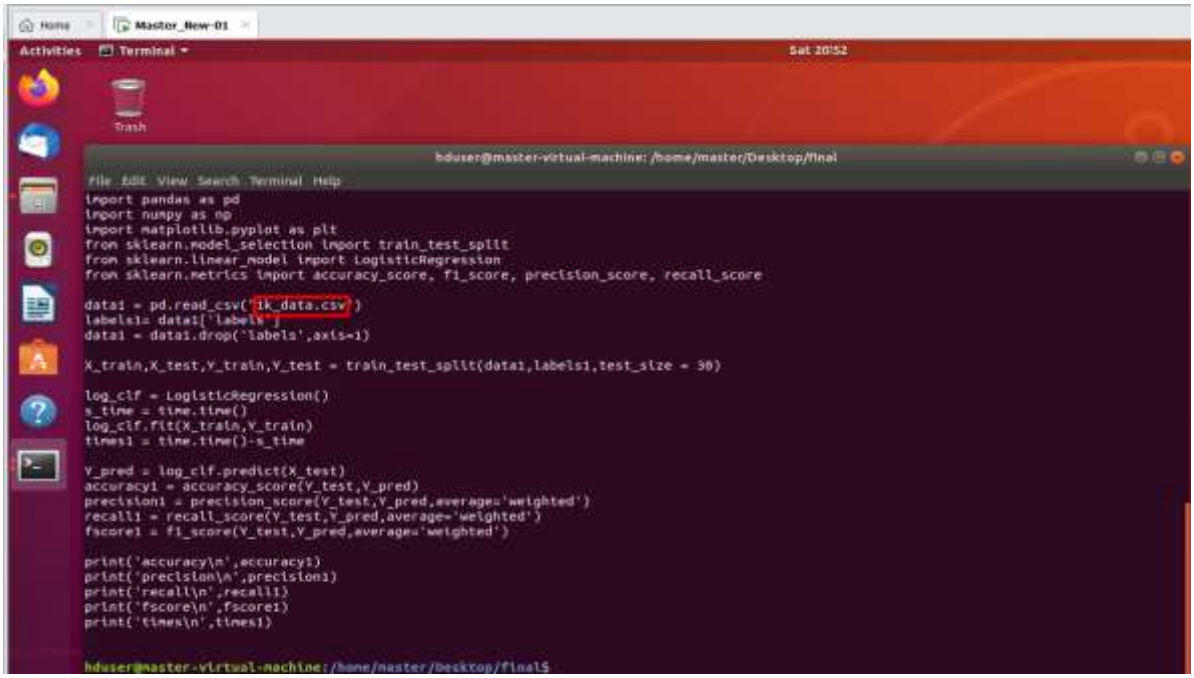

Figure 26: Define sample data file in Algortithm

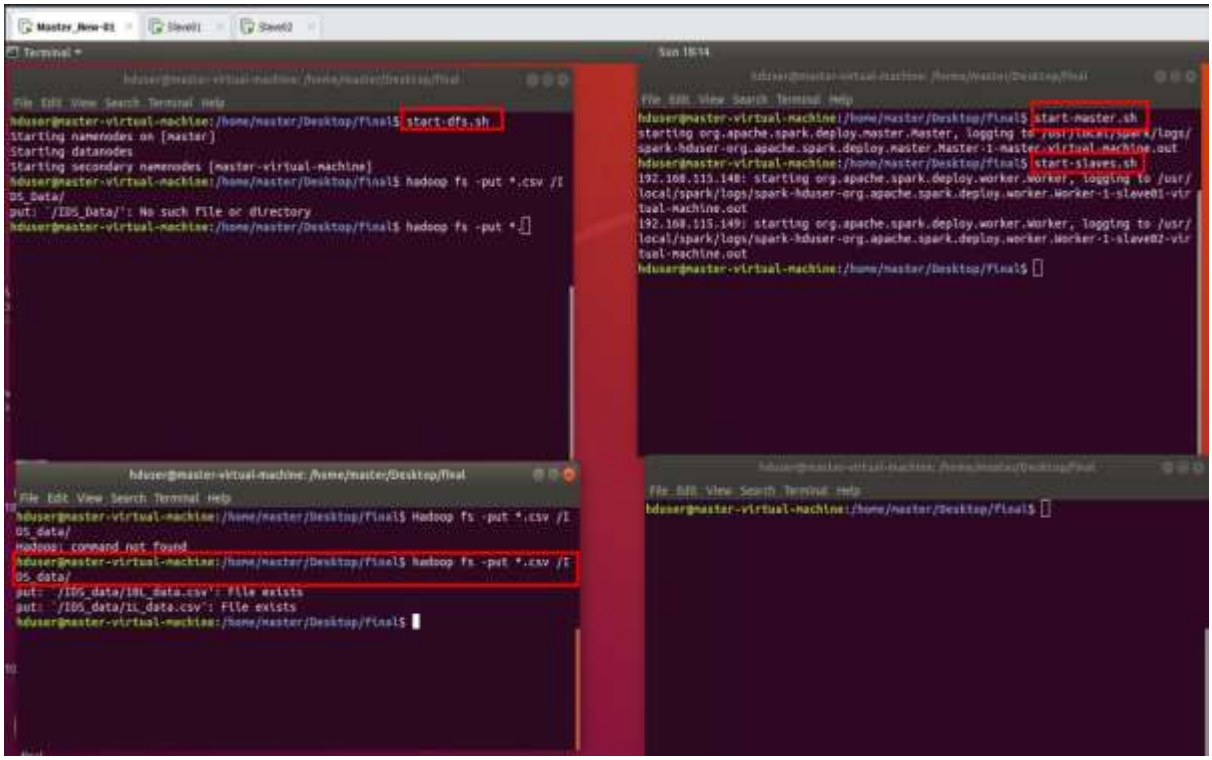

Figure 27: All in one snapshot to start Hadoop and spark

- Execute below command to run algorithm in **tradition machine learning**.
  python3 spark_log_reg.py
- Execute below command to run algorithm in **distributed machine learning**
  spark-submit --master spark://192.168.115.147:7077 --executor-memory 2G spark_log_reg.py
- After that, analyse and compare the result for all algorithm i.e Logistic Regression, Naïve Bayes and Random Forest for 1k,10k,1L and 10L sample data.

References:

[1] “Installing Python 3 on Linux — The Hitchhiker’s Guide to Python,” *docs.python-guide.org*. Available:docs.pythonguide.org/starting/install3/linux/.
[2] “Anaconda,” *Anaconda*,2018. anaconda.com/
[3] “VMware Maintenance,” *maintenance.vmware.com*. my.vmware.com/en/web/ vmware/ downloads/info/slug/desktop_end_user_computing/vmware_workstation_pro/15_0
[4] “KDD Cup 1999 Data,” *kdd.ics.uci.edu*. kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.