

# INDEX

Sr. No.	Name of the Practical	Date	Sign.
1	K-Means Clustering		
2	Apriori Algorithm		
3	Simple Linear Regression and Logistic Regression		
4	Decision Tree Classification		
5	Naïve Bayes Classification		
6	Text Analysis		
7	Virtual Box Installation		
8	Ubuntu Installation		
9	Hadoop Installation		
10	WordCount in Hadoop		

**Practical 1****AIM:** K-Means Clustering**Code:**

```
install.packages("plyr")
install.packages("ggplot2")
install.packages("cluster")
install.packages("lattice")
install.packages("grid")
install.packages("gridExtra")

library(plyr)
library(ggplot2)
library(cluster)
library(lattice)
library(grid)
library(gridExtra)

grade_input=as.data.frame(read.csv("D:\\grades_km_input.csv"))
kmdata_orig=as.matrix(grade_input[, c("Student","English","Math","Science")])
kmdata=kmdata_orig[,2:4]
kmdata[1:10,]
wss=numeric(15)
for(k in 1:15)wss[k]=sum(kmeans(kmdata,centers = k,nstart = 25)$withinss)
plot(1:15,wss,type = "b",xlab = "Number of Clusters",ylab = "Within sum of Square")
km = kmeans(kmdata,3,nstart = 25)
km
c( wss[3] , sum(km$withinss))
df=as.data.frame(kmdata_orig[,2:4])
df$cluster=factor(km$cluster)
centers=as.data.frame(km$centers)
g1=ggplot(data=df, aes(x=English, y=Math, color=cluster ))
+geom_point() + theme(legend.position="right") +
geom_point(data=centers,aes(x=English,y=Math,
```

```

color=as.factor(c(1,2,3))),size=10, alpha=.3, show.legend =FALSE)

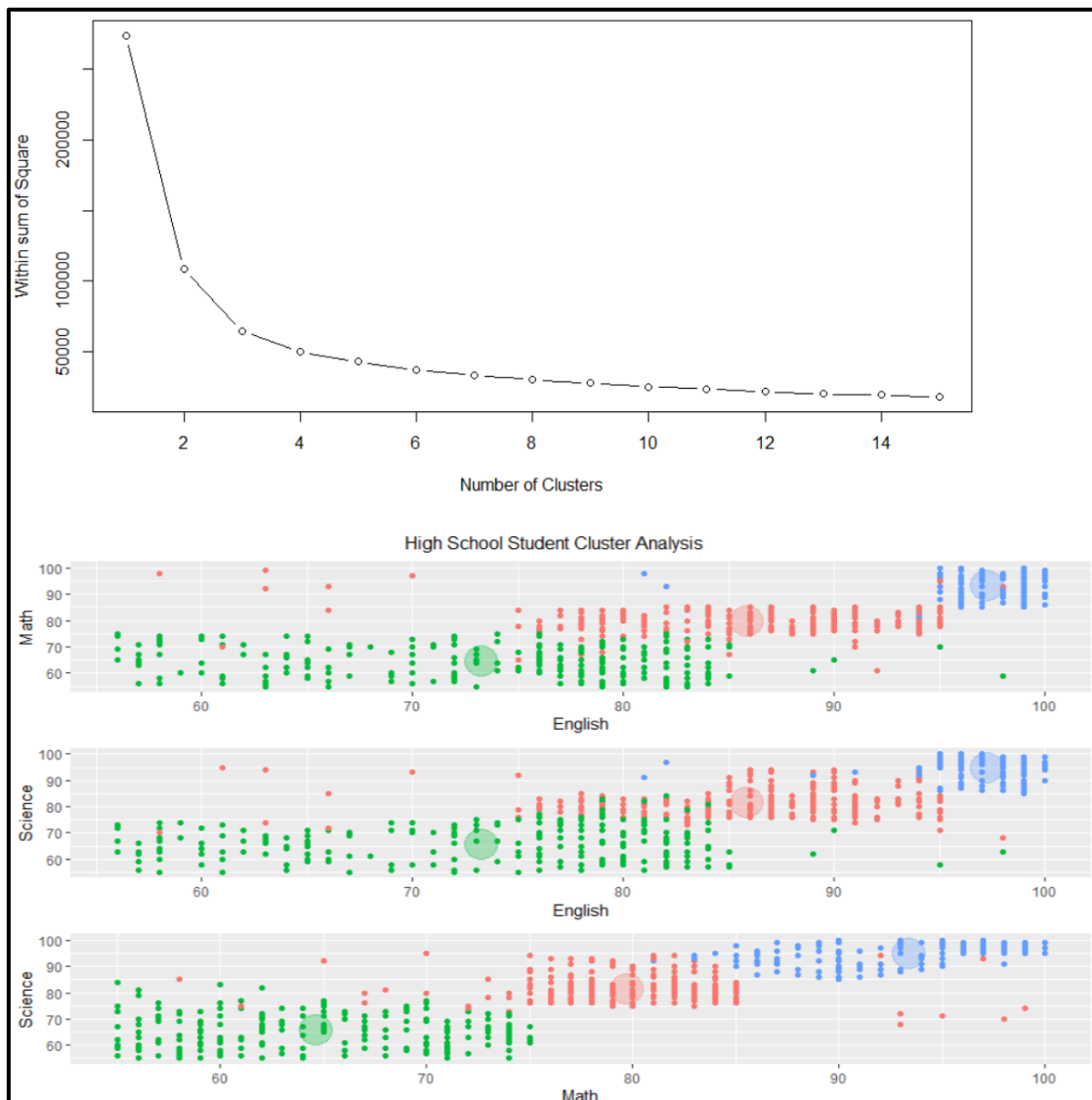
g2=ggplot(data=df, aes(x=English, y=Science, color=cluster )) + geom_point ()
+geom_point(data=centers,aes(x=English,y=Science,
color=as.factor(c(1,2,3))),size=10, alpha=.3, show.legend=FALSE)

g3 = ggplot(data=df, aes(x=Math, y=Science, color=cluster )) + geom_point () +
geom_point(data=centers,aes(x=Math,y=Science,
color=as.factor(c(1,2,3))),size=10, alpha=.3, show.legend=FALSE)

tmp=ggplot_gtable(ggplot_build(g1))

grid.arrange(arrangeGrob(g1 + theme(legend.position="none"),g2 +
theme(legend.position="none"),g3 + theme(legend.position="none"),top ="High
School Student Cluster Analysis" ,ncol=1))

```

**Output:**

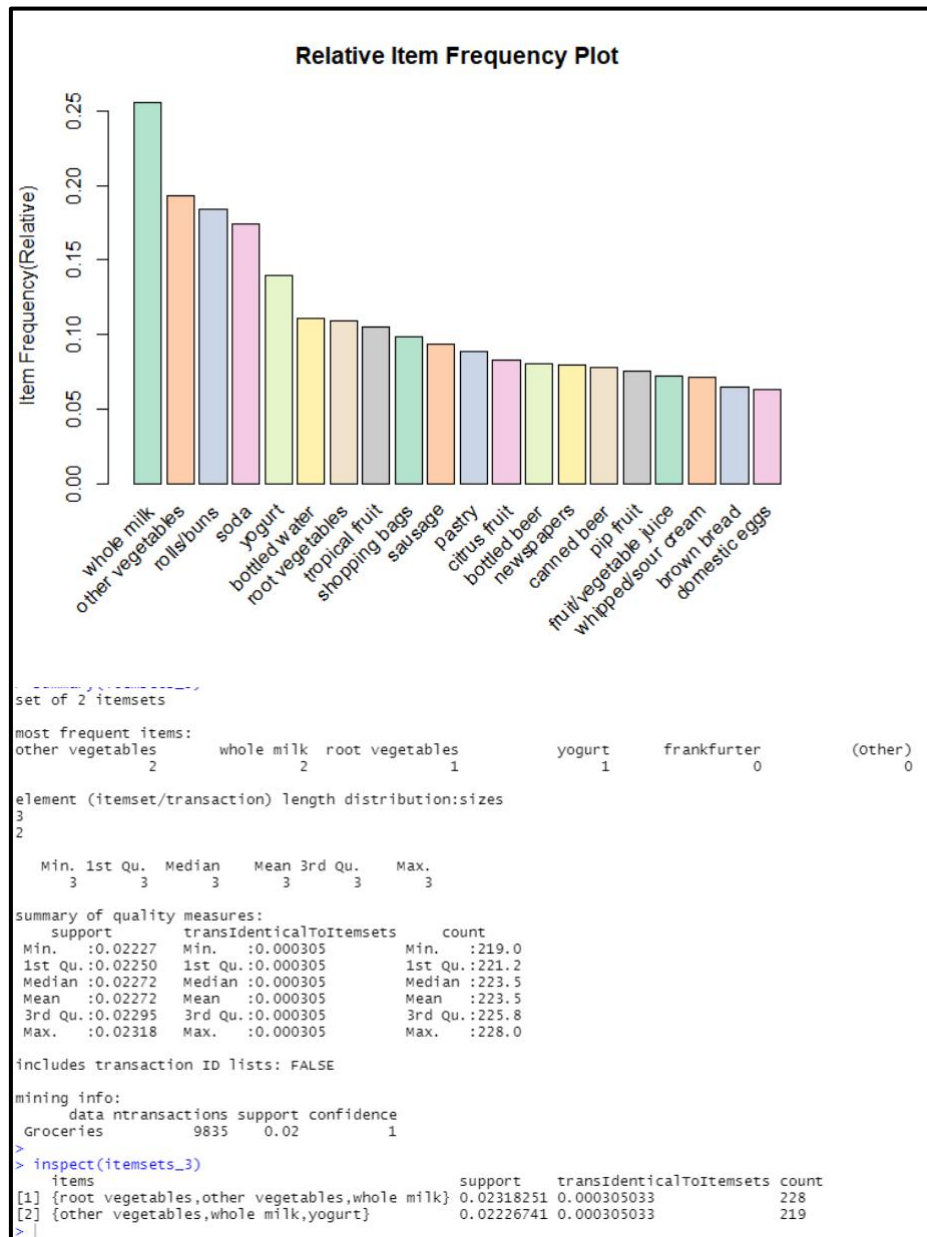
**Practical 2**

**Aim:** Apriori Algorithm

**Code:**

```
install.packages("arules")
install.packages("arulesViz")
install.packages("RColorBrewer")
library(arules)
library(arulesViz)
library(RColorBrewer)
data("Groceries")
Groceries
summary(Groceries)
class(Groceries)
rules = apriori(Groceries, parameter = list(supp = 0.02, conf = 0.2))
summary(rules)
inspect(rules[1:10])
arules::itemFrequencyPlot(Groceries, topN = 20,
                           col = brewer.pal(8, 'Pastel2'),
                           main = 'Relative Item Frequency Plot',
                           type = "relative",
                           ylab = "Item Frequency(Relative)")
itemset = apriori(Groceries, parameter = list(minlen=2, maxlen=2,
support=0.02, target="frequent itemset") )
summary(itemset)
inspect(itemset[1:10])
itemsets_3 = apriori(Groceries, parameter = list(minlen=3, maxlen=3,
support=0.02, target="frequent itemset"))
summary(itemsets_3)
inspect(itemsets_3)
```

## Output:



**Code:****# Apriori****I. Data Preprocessing**

```
install.packages('arules')  
install.packages("RColorBrewer")  
library(arules)  
library(RColorBrewer)  
dataset = read.csv('D:\\Market_Basket_Optimisation.csv', header = FALSE)  
dataset = read.transactions('D:\\Market_Basket_Optimisation.csv', sep = ',',  
rm.duplicates = TRUE) summary(dataset)
```

**II. Training Apriori on the dataset**

```
rules = apriori(data = dataset, parameter = list(support = 0.004, confidence =  
0.2))  
  
# Visualising the results  
inspect(sort(rules, by =  
'lift')[1:10])  
itemFrequencyPlot(dataset,  
topN = 10,  
col = brewer.pal(8, 'Pastel2'),  
main = 'Relative Item Frequency Plot',  
type = "relative",  
ylab = "Item Frequency (Relative)")  
  
itemsets = apriori(dataset, parameter = list(minlen=2,  
maxlen=2,support=0.02, target="frequent itemsets"))  
summary(itemsets)
```

# using

inspect() function

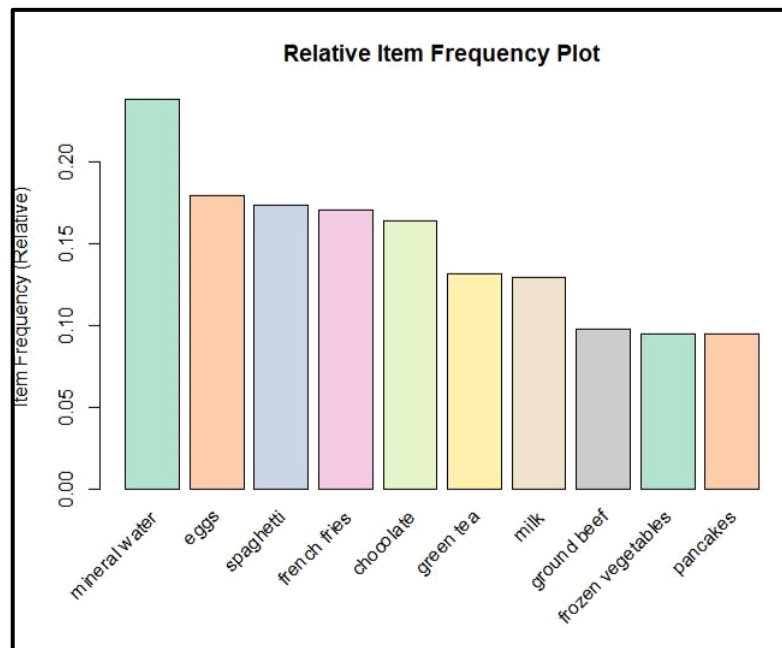
```
inspect(itemsets[1:
10])
```

```
itemsets_3 = apriori(dataset, parameter = list(minlen=3,
maxlen=3,support=0.02, target="frequent itemsets"))
```

```
summary(itemsets_3)
```

```
print ("Candidate list with 3 itemsets is not possible for this dataset")
```

Output:



```
> itemsets_3 = apriori(dataset, parameter = list(minlen=3, maxlen=3, support=0.02, target="frequent itemsets"))
Apriori
Parameter specification:
confidence minval smax arem aval originalsupport maxtime support minlen maxlen target ext
NA 0.1 1 none FALSE TRUE 5 0.02 3 3 frequent itemsets TRUE

Algorithmic control:
filter tree heap memopt load sort verbose
0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 150

set item appearances ... [0 item(s)] done [0.00s].
set transactions ... [119 item(s), 7501 transaction(s)] done [0.00s].
sorting and recoding items ... [53 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
sorting transactions ... done [0.00s].
writing ... [0 set(s)] done [0.00s].
creating S4 object ... done [0.00s].
Warning message:
In apriori(dataset, parameter = list(minlen = 3, maxlen = 3, support = 0.02, :
Mining stopped (maxlen reached). Only patterns up to a length of 3 returned!
> summary(itemsets_3)
set of 0 itemsets
>
> print ("Candidate list with 3 itemsets is not possible for this dataset")
[1] "Candidate list with 3 itemsets is not possible for this dataset"
```

### Practical 3

**AIM:** Simple Linear Regression and Logistic Regression

**Code:**

```
years_of_exp = c(7,5,1,3)
```

```
salary_in_lakhs = c(21,13,6,8)
```

```
employee.data = data.frame(years_of_exp,salary_in_lakhs)
```

```
employee.data
```

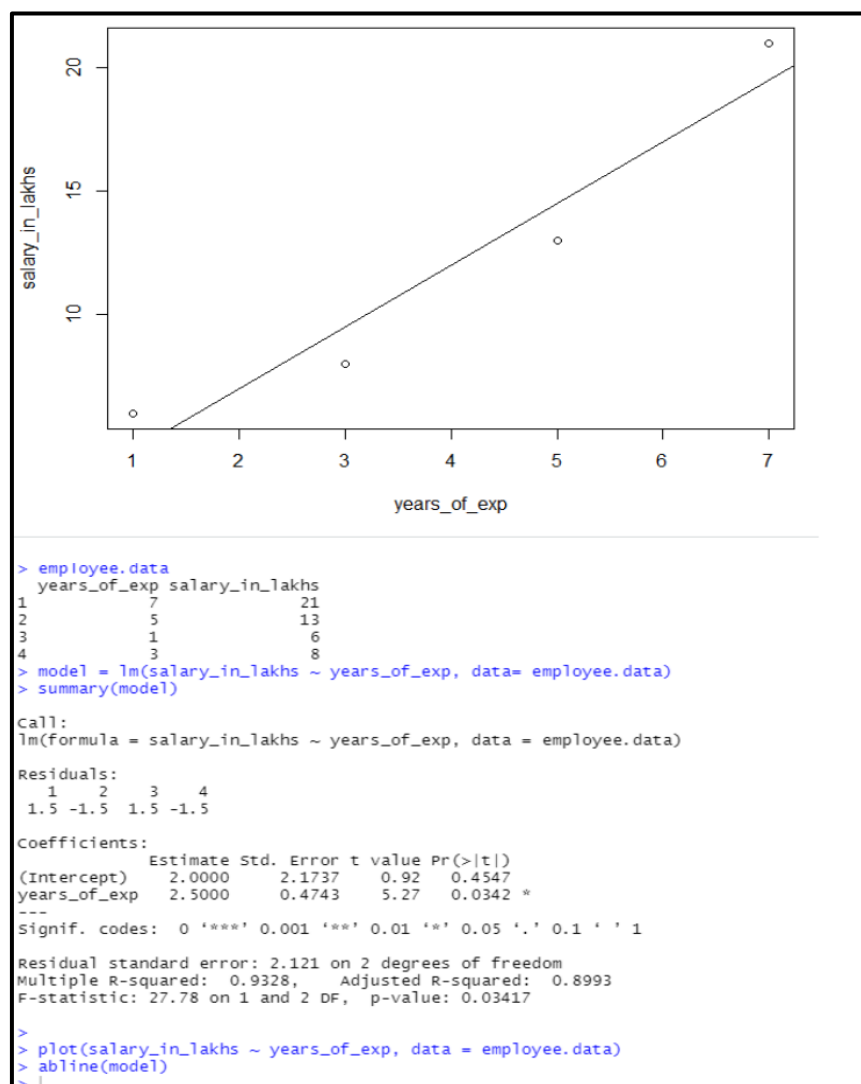
```
model = lm(salary_in_lakhs ~ years_of_exp, data= employee.data)
```

```
summary(model)
```

```
plot(salary_in_lakhs ~ years_of_exp, data = employee.data)
```

```
abline(model)
```

**Output:**





**Logistic Linear Regression:****Code:**

```
install.packages("ISLR")
library(ISLR)
data <- ISLR::Default
print(head(ISLR::Default))
summary(data)
nrow(data)
set.seed(1)
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.7,0.3))
print(sample)
train <- data[sample, ]
test <- data[!sample, ]
nrow(train)
nrow(test)
model <- glm(default~student+balance+income, family = "binomial", data = train)
summary(model)
install.packages("InformationValue")
library(InformationValue)
predicted <- predict(model,test,type="response")
confusionMatrix(test$default,predicted)
```

**Output:**

```
Call:
glm(formula = default ~ student + balance + income, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5586   -0.1353   -0.0519   -0.0177    3.7973

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.148e+01  6.234e-01 -18.412  <2e-16 ***
studentYes   -4.933e-01  2.857e-01  -1.726   0.0843 .
balance       5.988e-03  2.938e-04  20.384  <2e-16 ***
income       7.857e-06  9.965e-06   0.788   0.4304
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2021.1  on 6963  degrees of freedom
Residual deviance: 1065.4  on 6960  degrees of freedom
AIC: 1073.4

Number of Fisher Scoring iterations: 8
> confusionMatrix(test$default,predicted)
      No Yes
0 2912   64
1   21   39
> |
```

**Practical 4****AIM:** Decision Tree Classification**Code:**

```
dataset = read.csv('D:\\Social_Network_Ads.csv')
dataset = dataset[3:5]
print(dataset)
dataset$Purchased = factor(dataset$Purchased, levels = c(0, 1))
install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$Purchased, SplitRatio = 0.75)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
training_set[-3] = scale(training_set[-3])
test_set[-3] = scale(test_set[-3])
print(training_set[-3])
print(test_set[-3])
install.packages('rpart')
library(rpart)
classifier = rpart(formula = Purchased ~ ., data = training_set)
y_pred = predict(classifier, newdata = test_set[-3], type = 'class')
cm = table(test_set[, 3], y_pred)
print(cm)

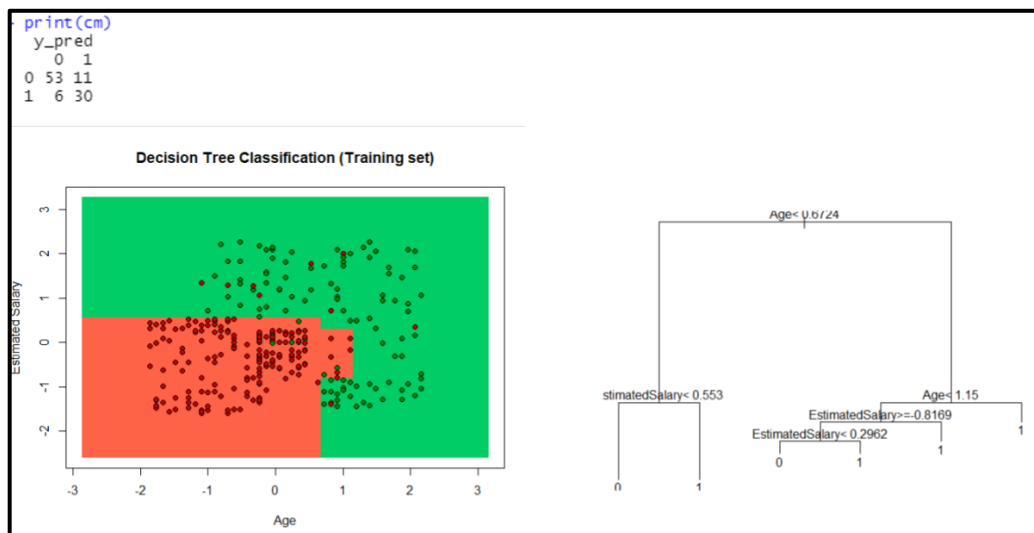
install.packages("ElemStatLearn")
library(ElemStatLearn)
set = training_set
X1 = seq(min(set[, 1]) - 1, max(set[, 1]) + 1, by = 0.01)
X2 = seq(min(set[, 2]) - 1, max(set[, 2]) + 1, by = 0.01)
grid_set = expand.grid(X1,X2)
colnames(grid_set) = c('Age','EstimatedSalary')
y_grid = predict(classifier, newdata = grid_set, type = 'class')
```

```

plot(set[, -3],
     main = 'Decision Tree Classification (Training set)',
     xlab = 'Age', ylab = 'Estimated Salary',
     xlim = range(X1), ylim = range(X2))
contour(X1, X2, matrix(as.numeric(y_grid), length(X1), length(X2)), add = TRUE)
points(grid_set, pch = '.', col = ifelse(y_grid == 1, 'springgreen3', 'tomato'))
points(set, pch = 21, bg = ifelse(set[, 3] == 1, 'green4', 'red3'))
set = test_set

X1 = seq(min(set[, 1]) - 1, max(set[, 1]) + 1, by = 0.01)
X2 = seq(min(set[, 2]) - 1, max(set[, 2]) + 1, by = 0.01)
grid_set = expand.grid(X1, X2)
colnames(grid_set) = c('Age', 'EstimatedSalary')
y_grid = predict(classifier, newdata = grid_set, type = 'class')
plot(set[, -3],
     main = 'Decision Tree Classification (Test set)',
     xlab = 'Age', ylab = 'Estimated Salary',
     xlim = range(X1), ylim = range(X2))
contour(X1, X2, matrix(as.numeric(y_grid), length(X1), length(X2)), add = TRUE)
points(grid_set, pch = '.', col = ifelse(y_grid == 1, 'springgreen3', 'tomato'))
points(set, pch = 21, bg = ifelse(set[, 3] == 1, 'green4', 'red3'))
plot(classifier)
text(classifier)

```

**Output:**

**Practical 5****AIM: Naïve Bayes Classification****Code:**

```
#Naive Bayes
#Importing the dataset
dataset = read.csv('D:\\Social_Network_Ads.csv')
dataset = dataset[3:5]
# Encoding the target feature as factor
dataset$Purchased = factor(dataset$Purchased, levels = c(0, 1))
#Splitting the dataset into the Training set and
Test set #install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$Purchased,
SplitRatio = 0.75) training_set = subset(dataset,
split == TRUE)
test_set = subset(dataset, split == FALSE)
#Feature Scaling
training_set[-3] = scale(training_set[-3])
test_set[-3] = scale(test_set[-3])
#Fitting Naive Bayes to the
Training set
install.packages('e1071')
library(e1071)

classifier = naiveBayes(x =
training_set[-3], y =
training_set$Purchased)
#Predicting the Test set results
y_pred = predict(classifier, newdata = test_set[-3])
```

**#Making the Confusion**

```
Matrix cm = table(test_set[,  
3], y_pred) print(cm)
```

**#Visualising the Training set  
results**

```
install.packages("ElemStatLea  
rn")  
  
library(ElemStatLearn)  
set = training_set  
print(set)  
X1 = seq(min(set[, 1]) - 1, max(set[, 1]) + 1, by = 0.01)  
X2 = seq(min(set[, 2]) - 1, max(set[, 2]) + 1, by = 0.01)  
grid_set = expand.grid(X1, X2)  
colnames(grid_set) = c('Age', 'EstimatedSalary')  
y_grid = predict(classifier, newdata = grid_set)  
plot(set[, -3],  
      main = 'Naive Bayes (Training set)',  
      xlab = 'Age', ylab = 'Estimated Salary',  
      xlim = range(X1), ylim = range(X2))  
contour(X1, X2, matrix(as.numeric(y_grid), length(X1), length(X2)), add = TRUE)  
points(grid_set, pch = '.', col = ifelse(y_grid == 1, 'springgreen3', 'tomato'))  
points(set, pch = 21, bg = ifelse(set[, 3] == 1, 'green4', 'red3'))
```

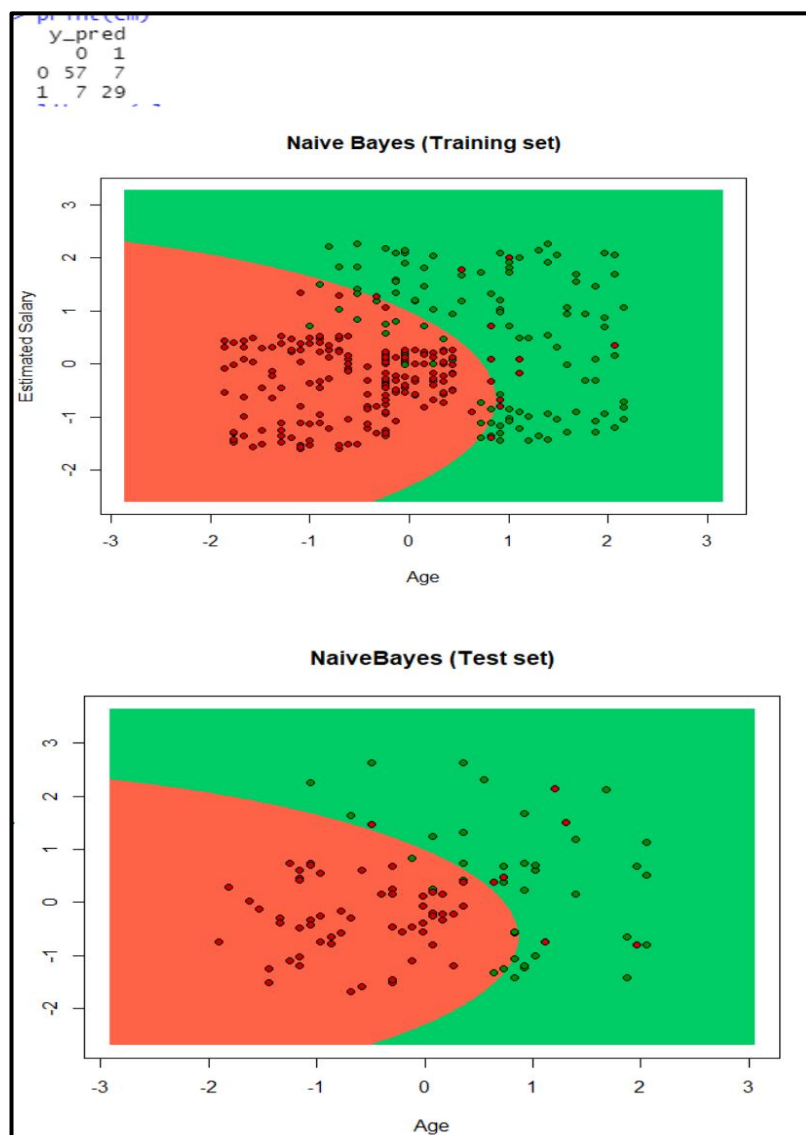
**#Visualising the Test set**

```
results  
library(ElemStatLearn)  
set = test_set  
X1 = seq(min(set[, 1]) - 1, max(set[, 1]) + 1, by = 0.01)  
X2 = seq(min(set[, 2]) - 1, max(set[, 2]) + 1, by  
= 0.01) grid_set = expand.grid(X1, X2)
```

```

colnames(grid_set) = c('Age',
'EstimatedSalary') y_grid = predict(classifier,
newdata = grid_set) plot(set[, -3], main =
'NaiveBayes (Test set)',
xlab = 'Age', ylab = 'Estimated
Salary', xlim = range(X1), ylim =
range(X2))
contour(X1, X2, matrix(as.numeric(y_grid), length(X1), length(X2)),
add = TRUE) points(grid_set, pch = '.', col = ifelse(y_grid == 1,
'springgreen3', 'tomato')) points(set, pch = 21, bg = ifelse(set[, 3] ==
1, 'green4', 'red3'))

```

**Output:**

### Practical 6

#### **AIM: Text Analysis**

##### **Code:**

```
dataset_original = read.delim('D:\\\\Restaurant_Reviews.tsv', quote = '',
stringsAsFactors = FALSE)

install.packages('tm')

install.packages('SnowballC')

library(tm)

library(SnowballC)

corpus = VCorpus(VectorSource(dataset_original$Review))

corpus = tm_map(corpus, content_transformer(tolower))

corpus = tm_map(corpus, removeNumbers)

corpus = tm_map(corpus, removePunctuation)

corpus = tm_map(corpus, removeWords, stopwords())

corpus = tm_map(corpus, stemDocument)

corpus = tm_map(corpus, stripWhitespace)

dtm = DocumentTermMatrix(corpus)

dtm = removeSparseTerms(dtm, 0.999)

dataset = as.data.frame(as.matrix(dtm))

dataset$Liked = dataset_original$Liked

print(dataset$Liked)

dataset$Liked = factor(dataset$Liked, levels = c(0,1))

install.packages(caTools)

library(caTools)

set.seed(123)

split = sample.split(dataset$Liked, SplitRatio = 0.8)

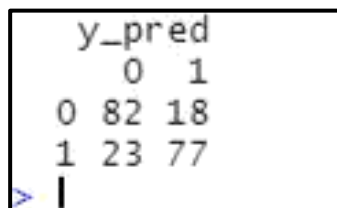
training_set = subset(dataset, split == TRUE)

test_set = subset(dataset, split == FALSE)

install.packages('randomForest')
```

```
library(randomForest)
classifier = randomForest(x = training_set[-692],
                          y =
                          training_set$Liked,
                          ntree = 10)
y_pred = predict(classifier, newdata = test_set[-692])
cm = table(test_set[,692], y_pred)
print(cm)
```

**Output:**



```
y_pred
  0  1
0 82 18
1 23 77
> |
```



## Practical 7

**AIM: Virtual Box Installation**

### **Step 1: Download and install VirtualBox**

Go to the website of Oracle VirtualBox and get the latest stable version from the following site

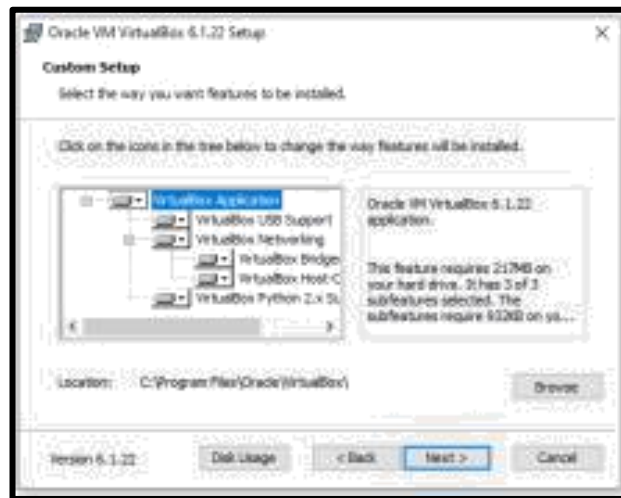
<https://www.virtualbox.org/>  
click on 'Download'



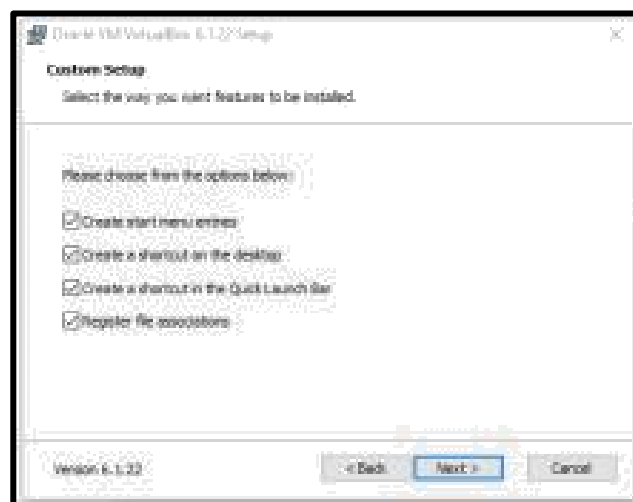
You will get VirtualBox-6.1.22-144080-Win.exe file downloaded. Double click and run it. Click on next.



Click on 'next' without changing the default folder as shown below:



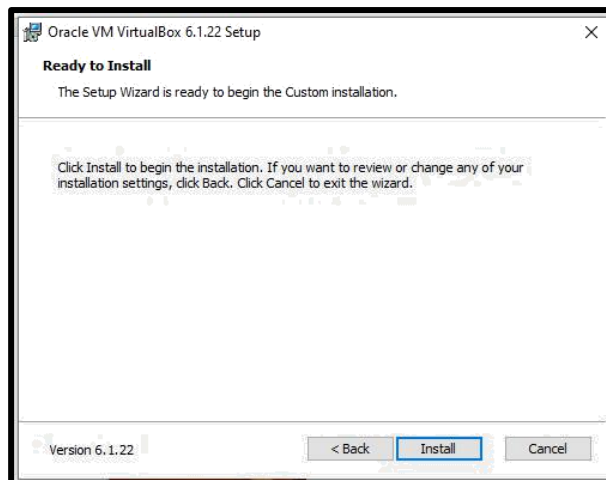
Again, click on next as shown below:



Finally, click on 'Yes'.



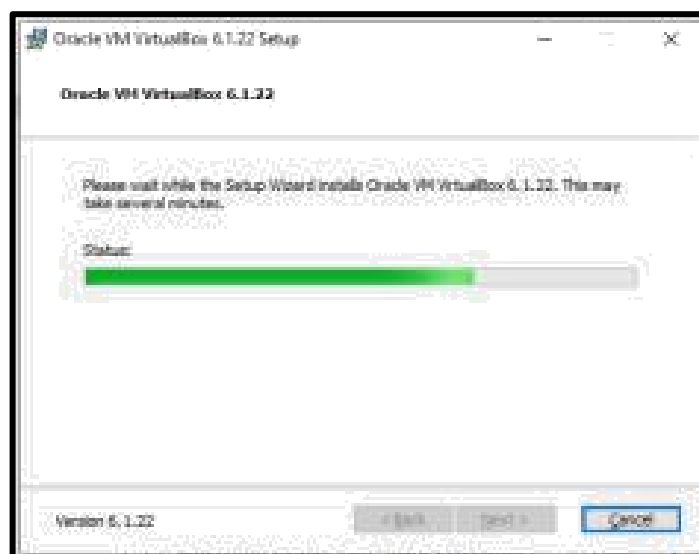
Click on 'Install'.



It may ask you for the permission to install, click 'yes' to allow. Select 'Install' as shown below:



You will get the screen as shown below:



Click on 'Finish' to finish Installation of virtual box.



## Practical 8

### AIM: Ubuntu Installation

Download iso file ubuntu-20.04.2.0-desktop-amd64; which is required to install Ubuntu.

Browse ubuntu.com

Click on download and 20.04 LTS as shown below:

LTS stands for Long term support

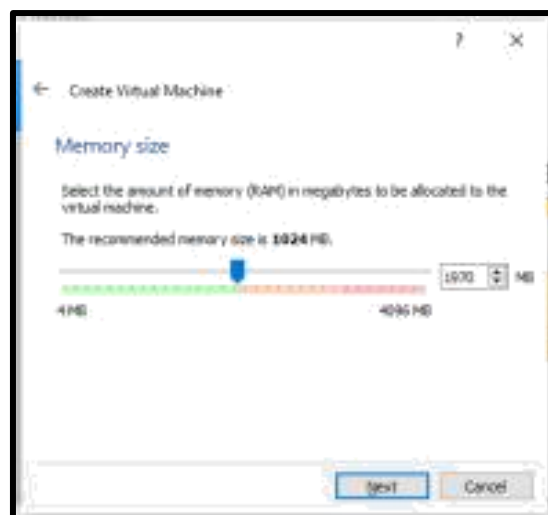


You will get file, which may take few minutes to download.

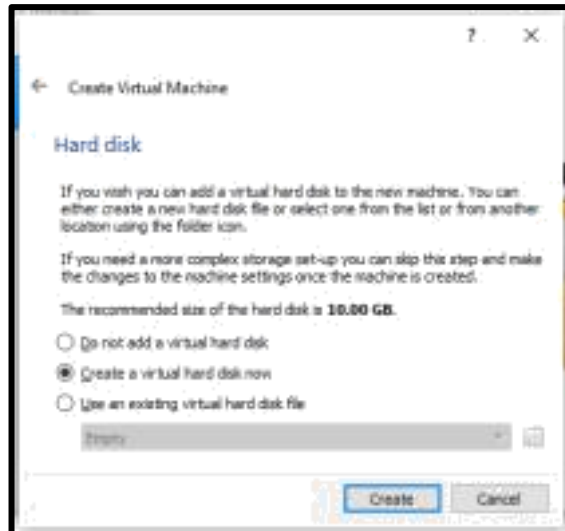
Now, click on 'New' to virtual box and write Name as 'Ubuntu' as shown below:



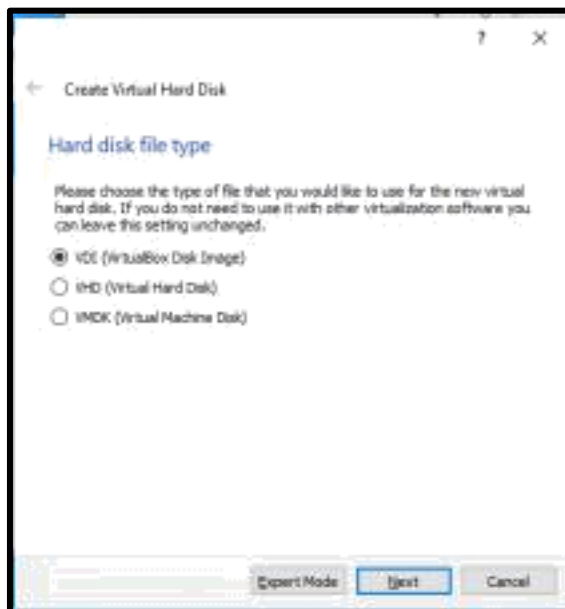
Click on 'Next'.



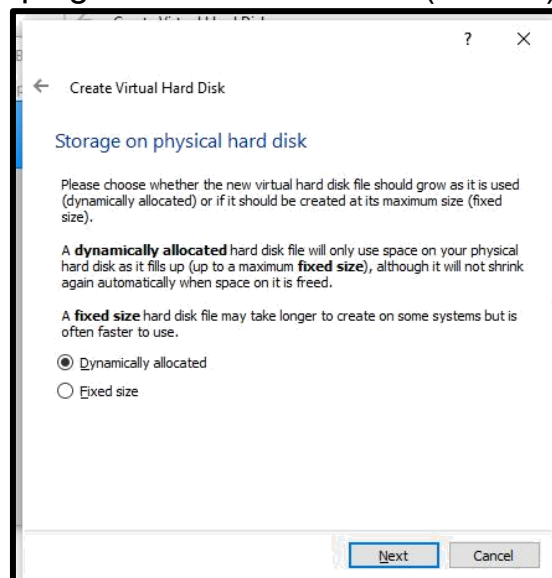
Here, you allow memory size up to green indicator (1970 MB). Click on 'Next'



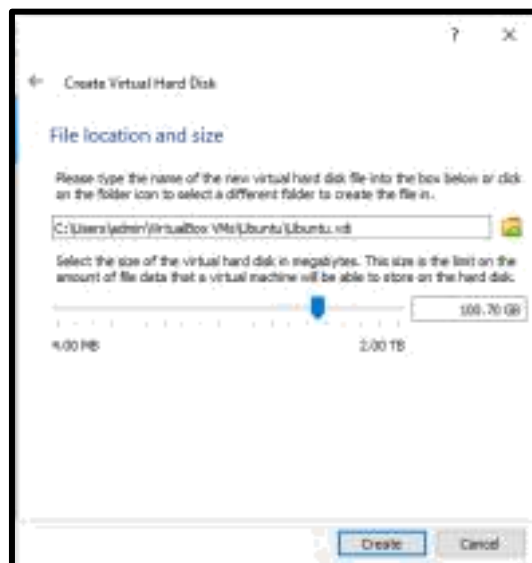
Don't change anything in this screen and click on 'Create'.



Click on 'Next', keeping the selection as it is (on VDI).'



Keep this screen also as it is and click on 'Next'.



Keep the file location as it is but preferably keep size 100 GB and click on 'Create'.

You may see the following screen having Ubuntu on Virtual Machine.

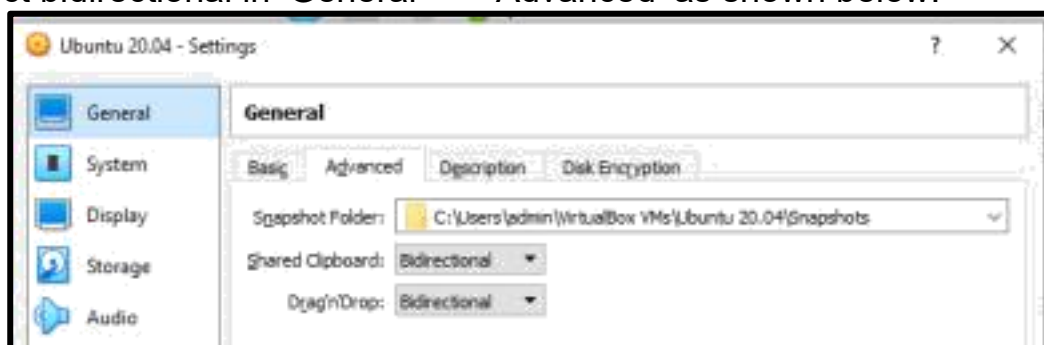


Select 'settings'

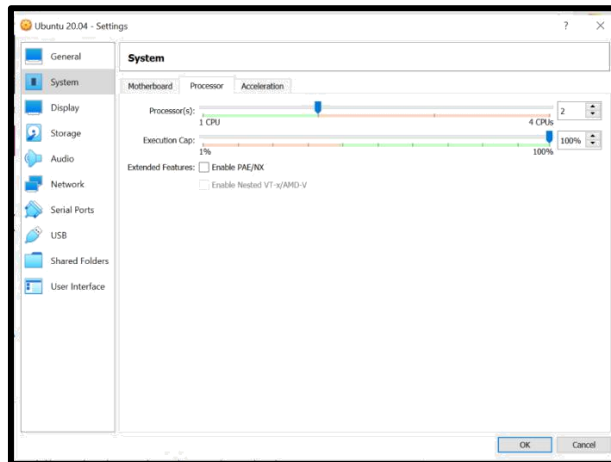
Select 'General' -> 'Basic' as shown below:

You may change the name from Ubuntu to Ubuntu 20.04

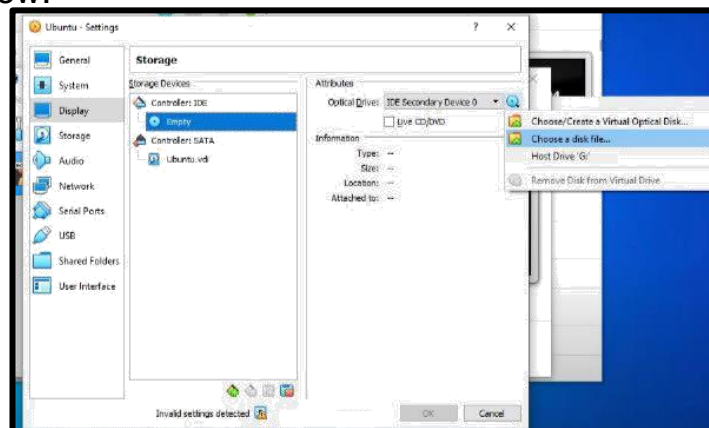
Select bidirectional in 'General' -> 'Advanced' as shown below:



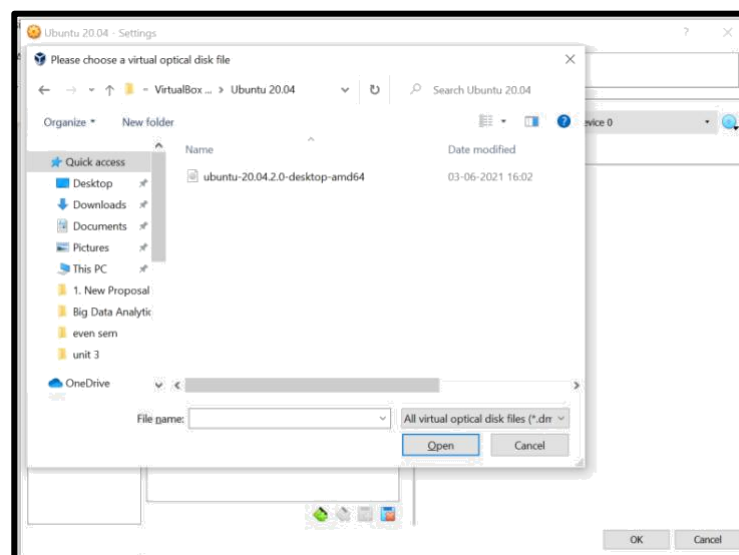
Go to 'System' option and change the processor up to green bar, usually 4.(if it allows)



and paste your ubuntu .iso file from current folder to C:\Users\ADMIN\VirtualBox VMs\Ubuntu 20.04 folder.  
Click on 'Storage' and click on 'Empty' followed by 'Choose a disk file' as shown below:

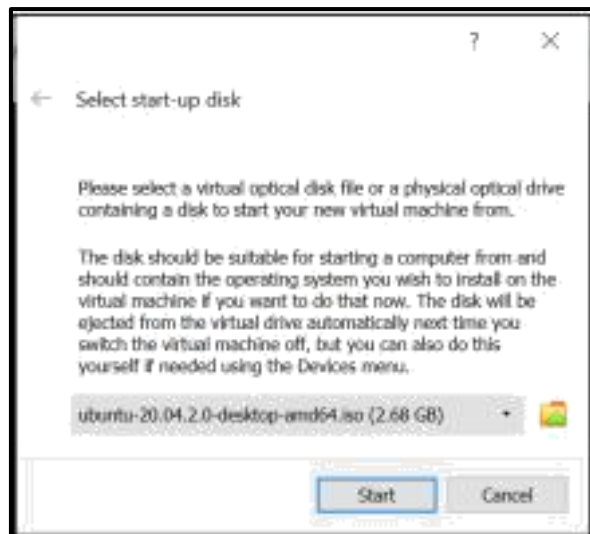


Browse the folder where you have selected ubuntu iso file.





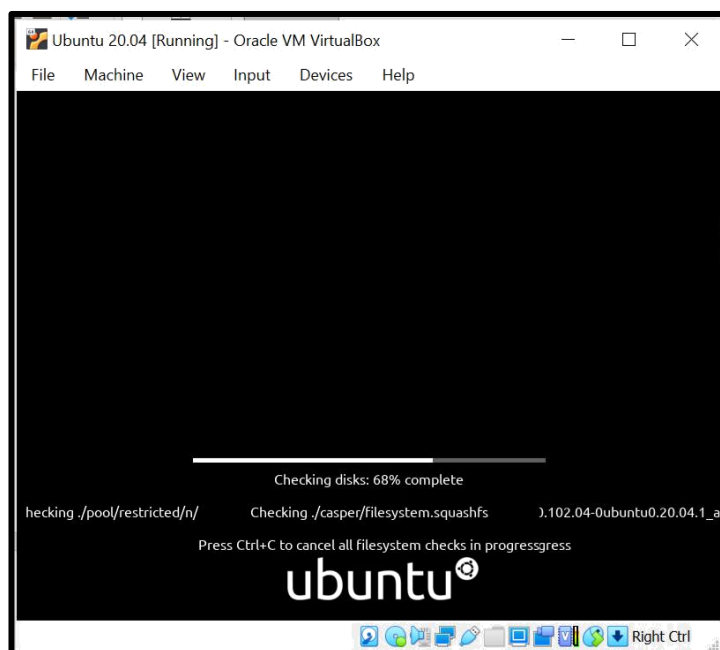
Click on Ubuntu....iso file and click on open and then click on ok.  
Click on Ubuntu -> start button.



Again, click on 'Start' button. It will show you the following screen.

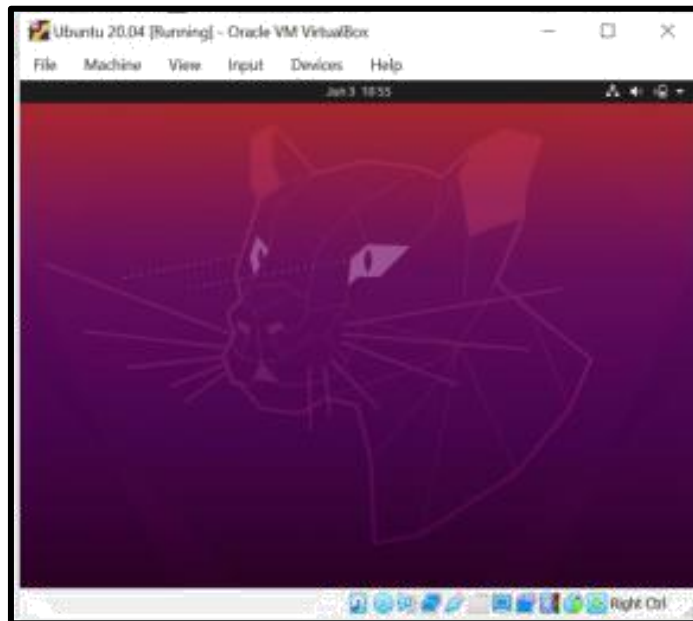


And simultaneously one more screen as follows:



Keep on closing all warnings.

Next you will get following screen automatically.



Select language -> English and click on 'Install Ubuntu'. In 'Keyboard Layout' screen, select 'English US'. Click on 'Continue'. Click on 'Continue'. (if you will select 'English UK', then some key will be changed as follows:

**\*\*Note:**

**Some Keys for Ubuntu under UK keyboard layout**

“->@

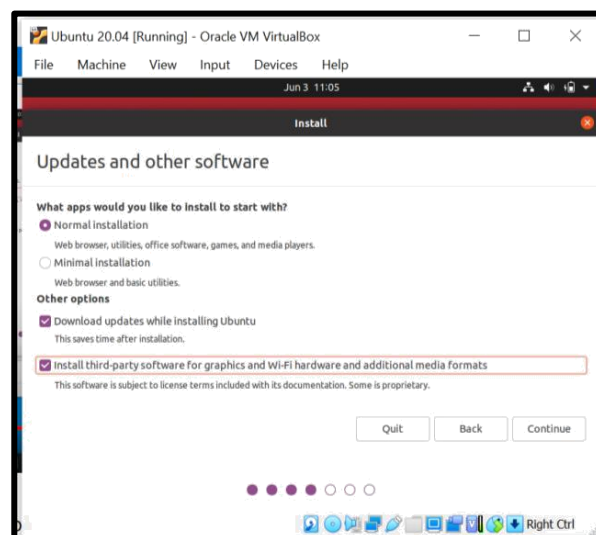
@->“

pipe -> take from this file or on google search for pipe in linux

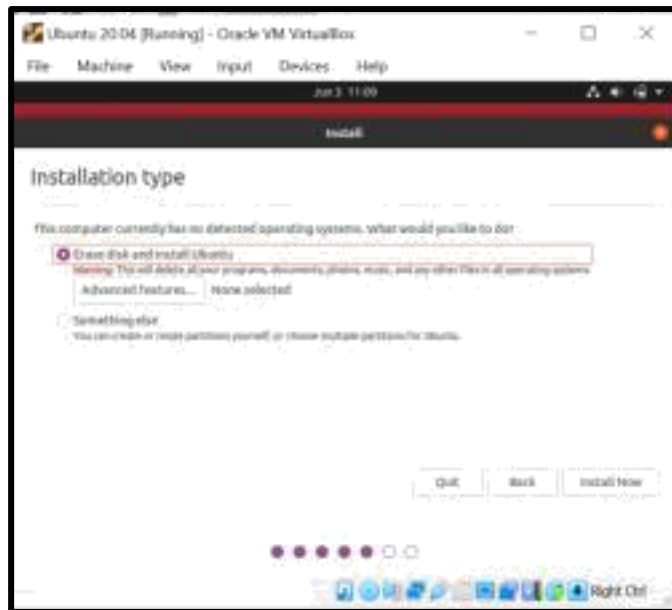
~ -> pipe

)

Select the checkbox for third party software as shown below:



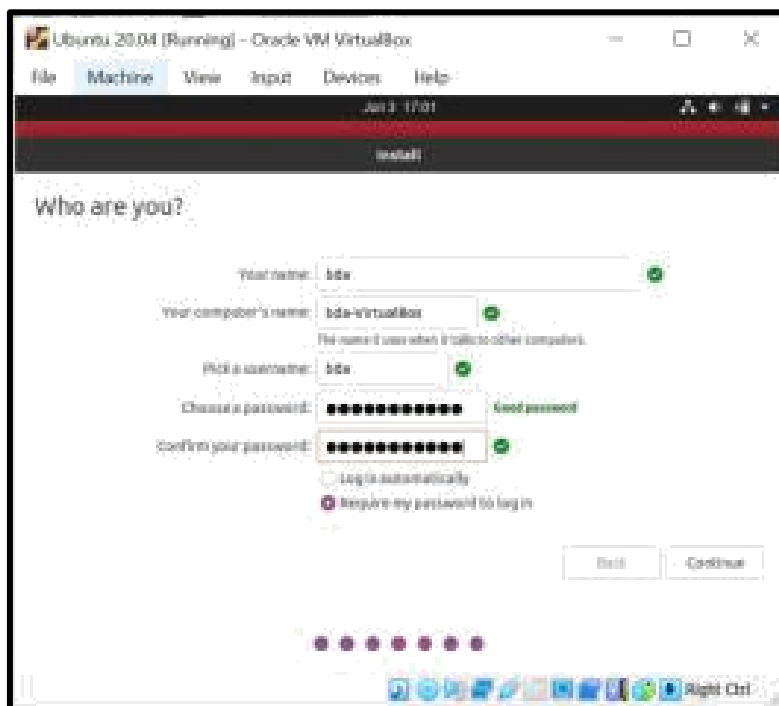
Click on 'continue'.



Select Erase disk and Install Ubuntu and click on '**Install Now**'.

Click on '**Continue**' on the next screen.

Select "Kolkata" for "where are you?" and click on '**Continue**'.



Click on continue after entering name, company name, username, password and confirm your password.



Installation of Ubuntu started. Click on finish once installation done. Click on restart and press Enter key.

## Practical 9

### **AIM: Hadoop Installation**

Login to ubuntu

Some keys may change like you try to type @ and it types “.

\*\* please refer to note -Some Keys for Ubuntu under UK keyboard layout –at the end. Search for Ubuntu terminal on search bar, after login done.

Apply following commands from ubuntu terminal

**Prerequisite bda@bda-**

**VirtualBox:~\$ sudo apt update**

Ign:1 cdrom://Ubuntu 20.04.2.0 LTS \_Focal Fossa\_ -Release amd64  
(20210209.1) focal InRelease

Hit:2 cdrom://Ubuntu 20.04.2.0 LTS \_Focal Fossa\_ -Release amd64 (20210209.1) focal

Release Hit:4 http://archive.ubuntu.com/ubuntu focal InRelease

Hit:5 http://archive.ubuntu.com/ubuntu focal-updates

InRelease Hit:6 http://security.ubuntu.com/ubuntu focal-  
security InRelease Reading package lists... Done

Building dependency tree

Reading state information...

Done

291 packages can be upgraded. Run 'apt list --upgradable' to see them.

**bda@bda-VirtualBox:~\$ sudo apt install**

**default-jdk** Reading package lists... Done Building  
dependency tree

Setting up default-jdk (2:1.11-72) ...

Setting up libxt-dev:amd64 (1:1.1.5-1) ...

**bda@bda-VirtualBox:~\$ java -version**

openjdk version "11.0.11" 2021-04-20

OpenJDK Runtime Environment (build 11.0.11+9-Ubuntu-0ubuntu2.20.04)

OpenJDK 64-Bit Server VM (build 11.0.11+9-Ubuntu-0ubuntu2.20.04, mixed mode,  
sharing) open ssh server

**bda@bda-VirtualBox:~\$ sudo apt install openssh-server openssh-client -y**

Reading package lists... Done

Building dependency tree

:

Processing triggers for ufw (0.36-6) ...

**bda@bda-VirtualBox:~\$ sudo adduser hadoop**

Adding user `hadoop' ...

Adding new group `hadoop' (1000) ...

Adding new user `hadoop' (1000) with group `hadoop' ...

Creating home directory `/home/hadoop' ...

Copying files from `/etc/skel' ...

New password: hadoop

Retype new password:

passwd: password updated successfully

```

Changing the user information for hdoop
Enter the new value, or press ENTER for the default
Full Name []:
Room Number []:
Work Phone []:
Home Phone []:
Other []:
Is the information correct? [Y/n] y
bda@bda-VirtualBox:~$ su -hdoop
Password: hdoop
hdoop@bda-VirtualBox:~$ ssh-keygen -t rsa -P '' -f
~/ssh/id_rsa Generating public/private rsa key pair.
Created directory '/home/hdoop/.ssh'.
Your identification has been saved in /home/hdoop/.ssh/id_rsa
Your public key has been saved in /home/hdoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:EDxiHTL1r3LUCdKFWc0moPHUh1D8tU6Y0b2rnXuWUtQ hdoop@bda-
VirtualBox The key's randomart image is:
  ---[RSA 3072]----
+  +
|  o+=.X++ .. |
|  oo+Oo.= * + .|
|  ..+. =*E.|
|    o + . = o. |
|    S+=. |
|    ..+. |
|    . o . ... |
|    o  .. o |

|    .+. |
+----[SHA256]-----+
hdoop@bda-VirtualBox:~$ cat ~/ssh/id_rsa.pub >>
~/ssh/authorized_keys hdoop@bda-VirtualBox:~$ chmod 0600
~/ssh/authorized_keys hdoop@bda-VirtualBox:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is
SHA256:4TE4DDAv14vhARPWjZcW3C5UM3X94B7wUudPrT+ZmF0.
Are you sure you want to continue connecting (yes/no/[fingerprint])? Yes
: Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.
Downloading Hadoop
hdoop@bda-VirtualBox:~$
wgethttps://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop-
3.3.1.tar.gz
--2021-06-14 08:52:00--https://downloads.apache.org/hadoop/common/hadoop-
3.3.1/hadoop-3.3.1.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219,
135.181.209.10, 135.181.214.104, ...

```

```
connected. HTTP request sent, awaiting response... 200 OK
Length: 359196911 (343M) [application/x-gzip]
Saving to: 'hadoop-3.3.1.tar.gz'
hadoop-3.3.1.tar.gz 100%[=====>] 342.56M 15.4MB/s in 33s 2021-
06-14 08:52:34 (10.2 MB/s) -'hadoop-3.3.1.tar.gz' saved [359196911/359196911]
```

```
hadoop@bda-VirtualBox:~$ ls
hadoop@bda-VirtualBox:~$ tar xzf hadoop-3.3.1.tar.gz
hadoop@bda-VirtualBox:~$ ls
hadoop-3.3.1 hadoop-3.3.1.tar.gz
```

**Editing 6 important files for creating a single**

**cluster** hadoop@bda-VirtualBox:~\$ su -bda

bda@bda-VirtualBox:~\$ sudo adduser hadoop sudo

Adding user `hadoop' to group `sudo' ...

Adding user hadoop to group sudo

Done.

bda@bda-VirtualBox:~\$ su -hadoop

1.

hadoop@bda-VirtualBox:~\$ sudo nano .bashrc

File will be opened and add following lines at the end of the file:

#Hadoop Related Options

export HADOOP\_HOME=/home/hadoop/hadoop-3.3.1

export HADOOP\_INSTALL=\$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=\$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=\$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=\$HADOOP\_HOME

export YARN\_HOME=\$HADOOP\_HOME

export

HADOOP\_COMMON\_LIB\_NATIVE\_DIR=\$HADOOP\_HOME/lib/nat

ive export

PATH=\$PATH:\$HADOOP\_HOME/sbin:\$HADOOP\_HOME/bin

export HADOOP\_OPTS="-

Djava.library.path=\$HADOOP\_HOME/lib/native" save this file as

ctrl x and y. Press enter. hadoop@bda-VirtualBox:~\$ source

~/bashrc

## 2. Edit hadoop-env.sh File

The *hadoop-env.sh* file serves as a master file to configure YARN, HDFS, MapReduce, and Hadoop-related project settings.

When setting up a **single node Hadoop cluster**, you need to define which Java implementation is to be utilized. Use the previously created **\$HADOOP\_HOME** variable to access the *hadoop-env.sh* file:

hadoop@bda-VirtualBox:~\$ sudo nano

**\$HADOOP\_HOME/etc/hadoop/hadoop-env.sh** at the end of the file add the following line

**export JAVA\_HOME=/usr/lib/jvm/java-11-openjdk-amd64/** save it.

### 3. Edit core-site.xml File

The *core-site.xml* file defines HDFS and Hadoop core properties.

To set up Hadoop in a pseudo-distributed mode, you need to **specify the URL** for your NameNode, and the temporary directory Hadoop uses for the map and reduce process.

Open the *core-site.xml* file in a text editor:

```
hadoop@bda-VirtualBox:~$ sudo nano
$HADOOP_HOME/etc/hadoop/core-site.xml <configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/home/hadoop/tmpdata</value>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9870</value>
</property>
</configuration>
```

4

```
hadoop@bda-VirtualBox:~$ sudo nano
$HADOOP_HOME/etc/hadoop/hdfs-site.xml <configuration>
<property>
<name>dfs.data.dir</name>
<value>/home/hadoop/dfsdata/namenode</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/home/hadoop/dfsdata/datanode</value>
</property>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
</configuration>
```

5

```
hadoop@bda-VirtualBox:~$ sudo nano $HADOOP_HOME/etc/hadoop/mapred-
site.xml
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```



6

**hadoop@bda-VirtualBox:~\$ sudo nano \$HADOOP\_HOME/etc/hadoop/yarn-site.xml**

```
<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>
<name>yarn.resourcemanager.hostname</name>
<value>127.0.0.1</value>
</property>
<property>
<name>yarn.acl.enable</name>
<value>0</value>
</property>
<property>
<name>yarn.nodemanager.env-whitelist</name>
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_C
ONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MA
PRED_HOME</value>
</property>
</configuration>
```

Format HDFS NameNode

**hadoop@bda-VirtualBox:~\$ hdfs namenode -format**

:

xid=0 when meet shutdown.

2021-06-18 14:16:33,353 INFO namenode.NameNode: SHUTDOWN\_MSG:

/\*\*\*\*\*

SHUTDOWN\_MSG: Shutting down NameNode at bda-VirtualBox/127.0.1.1

\*\*\*\*\*/

Start Hadoop Cluster (services) **hadoop@bda-**

**VirtualBox:~\$ cd Hadoop-3.3.1**

**hadoop@bda-**

**VirtualBox:~/Hadoop-3.3.1\$ cd sbin**

**hadoop@bda-**

**VirtualBox:~/hadoop-3.3.1/sbin\$ ./start-dfs.sh**

Starting namenodes on [localhost] Starting datanodes

Starting secondary namenodes [bda-VirtualBox] bda-VirtualBox: Warning:  
Permanently added 'bda-virtualbox' (ECDSA) to the list of known hosts.

2021-06-18 14:26:34,962 WARN util.NativeCodeLoader: Unable to load native-  
hadoop library for your platform... using builtin-java classes where applicable

**hadoop@bda-VirtualBox:~/hadoop-3.3.1/sbin\$ ./start-yarn.sh**

Starting resourcemanager

Starting nodemanagers

To see all components, we use jps command:

```
hdoop@bda-VirtualBox:~/hadoop-3.3.1/sbin$ jps
```

```
11744 NodeManager
```

```
11616 ResourceManager
```

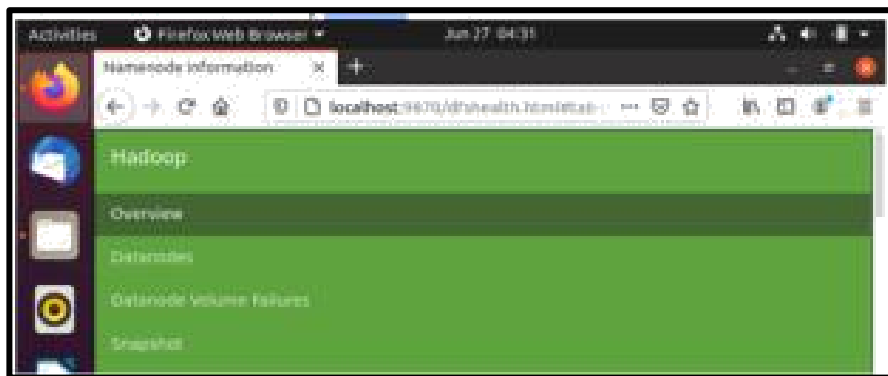
```
12192 Jps
```

```
11268 SecondaryNameNode
```

```
11077 DataNode
```

```
10954 NameNode
```

Browse localhost:9870 on any browser:



```
hdoop@bda-VirtualBox:~/hadoop-3.3.1/sbin$ hdfs dfs -ls /
```

```
2021-06-18 14:33:24,698 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
hdoop@bda-VirtualBox:~/hadoop-3.3.1/sbin$ sudo nano
```

```
/home/bda/sample.txt
```

```
[sudo] password for hdoop:
```

```
edit the file by adding some text and save and exit
```

```
hdoop@bda-VirtualBox:~/hadoop-3.3.1/sbin$ ls
```

```
/home/bda/ Desktop Downloads Pictures sample.txt
```

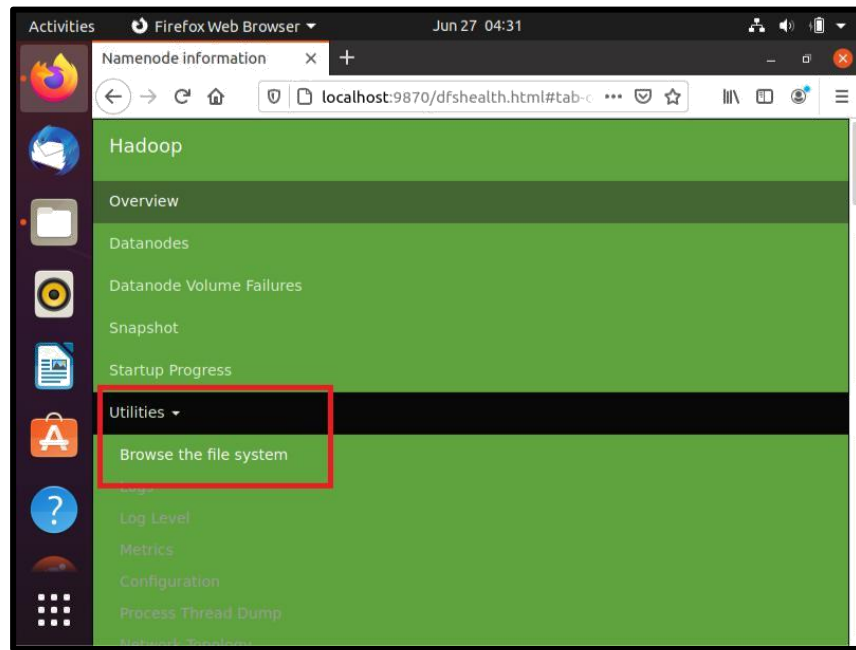
```
Videos Documents Music Public Templates
```

```
hdoop@bda-VirtualBox:~/hadoop-3.3.1/sbin$ hdfs dfs -put
```

```
/home/bda/sample.txt /
```

```
2021-06-18 14:44:24,257 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

Browse localhost:9870 on any browser and click on **utility** and **select browse the file system** you can see your folder there.



```
hdoop@bda-VirtualBox:~/hadoop-3.3.1/sbin$ hdfs dfs -ls /
```

```
2021-06-18 14:48:17,221 WARN util.NativeCodeLoader: Unable to load native-hadooplibrary  
for your platform... using builtin-java classes where applicable
```

```
Found 1 items
```

```
-rw-r--r--1 hdoop supergroup      6 2021-06-18 14:44 /sample.txt
```

### Practical 10

#### **AIM: WordCount in Hadoop**

**Login to bda user of Ubuntu**

**create a folder wordcount under home folder of Ubuntu create file WordCount.java and store in that folder: code:**

```
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import
org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import
org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {
public static class TokenizerMapper extends Mapper<Object, Text, Text,
IntWritable>{ private final static IntWritable one = new IntWritable(1); private
Text word = new Text();

public void map(Object key, Text value, Context
context ) throws IOException, InterruptedException {

StringTokenizer itr = new StringTokenizer(value.toString());
while (itr.hasMoreTokens()) {
word.set(itr.nextToken());
context.write(word, one);
}
}
}

public static class IntSumReducer extends
Reducer<Text,IntWritable,Text,IntWritable> { private IntWritable result = new
IntWritable();

public void reduce(Text key, Iterable<IntWritable> values, Context context) throws
IOException,
InterruptedException
{
int sum = 0;
for (IntWritable val : values) {
sum += val.get();
}
}
```

```

result.set(sum);
context.write(key, result);
}
}

```

```

public static void main(String[] args) throws
Exception { Configuration conf = new
Configuration();
Job job = Job.getInstance(conf, "word count");
job.setJarByClass(WordCount.class);
job.setMapperClass(TokenizerMapper.class);
job.setCombinerClass(IntSumReducer.class);
job.setReducerClass(IntSumReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
FileInputFormat.addInputPath(job, new
Path(args[0]));
FileOutputFormat.setOutputPath(job, new
Path(args[1]));
System.exit(job.waitForCompletion(true)?0:1);
}
}

```

**create a file input.txt with text editor and store it in wordcount folder.**

input.txt should have list of names in it, with each name in one line.

**create a folder wordcount\_classes in wordcount folder.**

**create env variable using hdoop terminal as follows and apply subsequent commands:**

**bda@bda-VirtualBox:~\$ su hdoop**

Password:

**hdoop@bda-VirtualBox:/home/bda\$ export HADOOP\_CLASSPATH=\$(hadoop classpath) hdoop@bda-VirtualBox:/home/bda\$ echo \$HADOOP\_CLASSPATH**  
/home/hdoop/hadoop-3.3.1/etc/hadoop:/home/hdoop/hadoop-3.3.1/share/hadoop/common/lib/\*:/home/hdoop/hadoop-

3.3.1/share/hadoop/common/\*:/home/hdoop/hadoop-

3.3.1/share/hadoop/hdfs:/home/hdoop/hadoop-

3.3.1/share/hadoop/hdfs/lib/\*:/home/hdoop/hadoop-

3.3.1/share/hadoop/hdfs/\*:/home/hdoop/hadoop-

3.3.1/share/hadoop/mapreduce/\*:/home/hdoop/hadoop-

3.3.1/share/hadoop/yarn:/home/hdoop/hadoop-

3.3.1/share/hadoop/yarn/lib/\*:/home/hdoop/hadoop-3.3.1/share/hadoop/yarn/\*

**hdoop@bda-VirtualBox:/home/bda\$ start-**

**dfs.sh** Starting namenodes on [localhost]

Starting datanodes

Starting secondary namenodes [bda-VirtualBox] 2021-06-26 13:13:58,694 WARN

util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using

builtin-java classes where applicable **hadoop@bda-VirtualBox:/home/bda\$ start-yarn.sh**

Starting resourcemanager

Starting nodemanagers

**hadoop@bda-VirtualBox:/home/bda\$ jps**

3248 ResourceManager

3779 Jps

3382 NodeManager

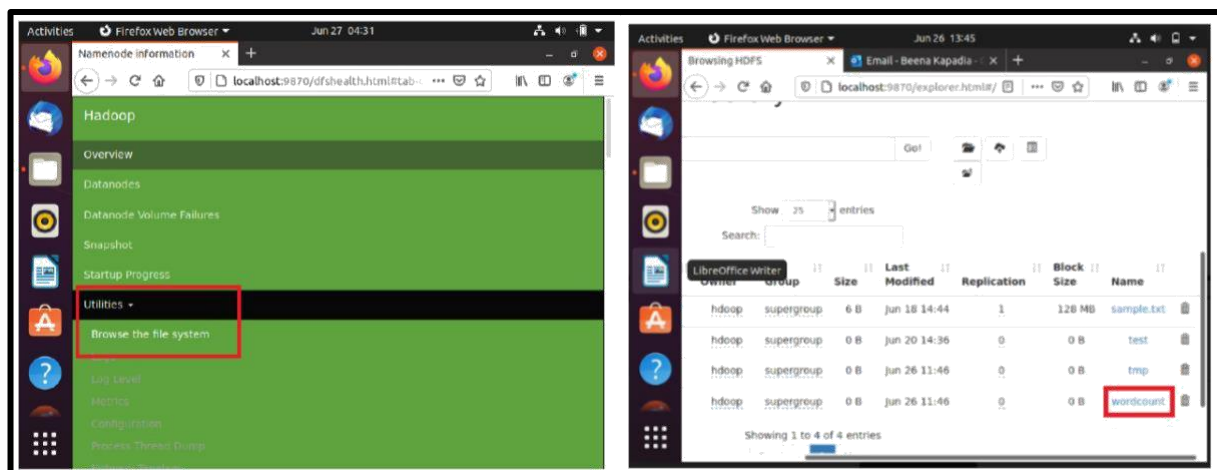
2631 NameNode

2760 DataNode

2953 SecondaryNameNode

**hadoop@bda-VirtualBox:/home/bda\$ hdfs dfs -mkdir /wordcount/**

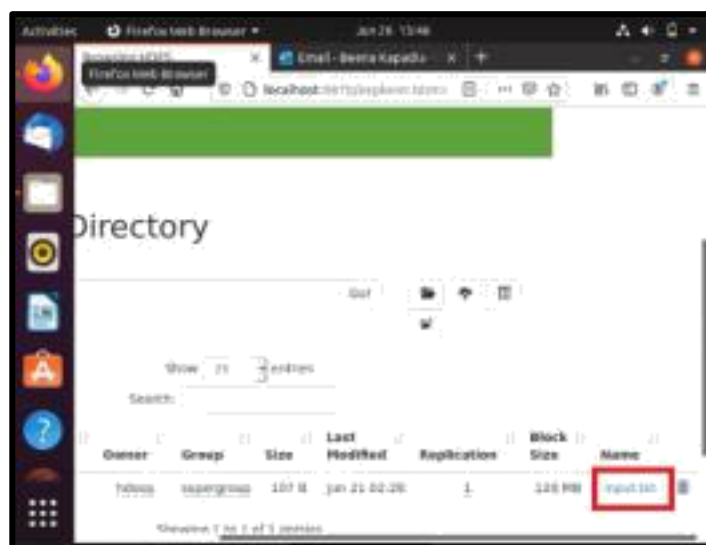
(Browse localhost:9870 on any browser and click on **utility** and **select browse the file system** you can see your folder there)



**hadoop@bda-VirtualBox:/home/bda\$ hdfs dfs -mkdir**

**/wordcount/input/ (now, move local file to hadoop folder)**

**hadoop@bda-VirtualBox:/home/bda\$ hdfs dfs -put '/home/bda/wordcount/input.txt' /wordcount/in**



```
hadoop@bda-VirtualBox:/home/bda$ hdfs dfs -cat /wordcount/input/*
```

```
2021-06-26 13:16:53,156 WARN util.NativeCodeLoader: Unable to load native-hadoop  
library for your platform... using builtin-java classes where applicable beena
```

```
yashesh  
vinod  
beena  
kruti  
nidhi  
vinod  
yashesh  
dhavan  
kruti  
nidhi  
komal  
jeenisha  
mitish  
nidhi  
yashesh
```

```
hadoop@bda-VirtualBox:/home/bda$ su bda
```

```
password
```

```
bda@bda-VirtualBox:~$ cd wordcount
```

```
bda@bda-VirtualBox:~/wordcount$ javac -classpath ${HADOOP_CLASSPATH} -d
```

```
'/home/bda/wordcount/wordcount_classes' '/home/bda/wordcount/WordCount.java'
```

```
check your wordcount_classes folder, which now has three classes in it: WordCount.class,  
WordCount$IntSumReducer.class and WordCount$TokenizerMapper.class
```

```
javac -classpath ${HADOOP_CLASSPATH} -d
```

```
'/home/bda/wordcount/classes' '/home/bda/wordcount/WordCount.java'
```

```
check your wordcount_classes folder, which now has three classes in it: WordCount.class,  
WordCount$IntSumReducer.class and WordCount$TokenizerMapper.class
```

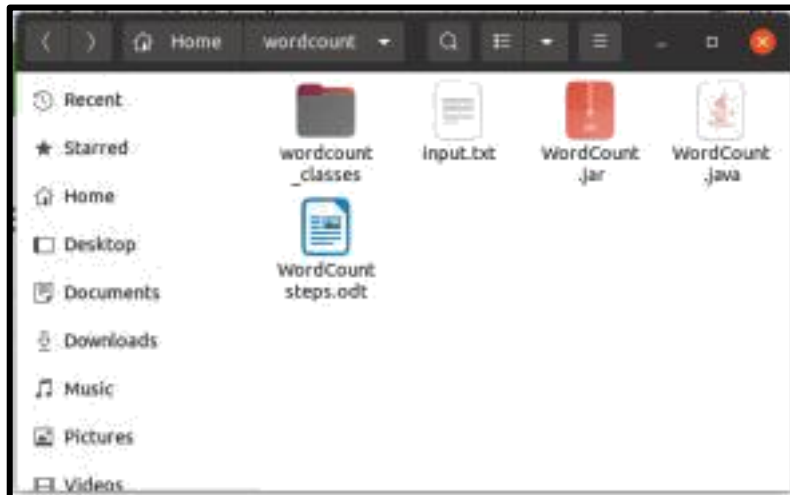


```
bda@bda-VirtualBox:~/wordcount$ jar -cvf WordCount.jar -C '/home/bda/wordcount/classes'/.
added manifest
```

```
adding: WordCount$IntSumReducer.class(in = 1755) (out= 750)(deflated 57%)
```

```
adding: WordCount$TokenizerMapper.class(in = 1752) (out= 761)(deflated 56%)
```

```
adding: WordCount.class(in = 1511) (out= 832)(deflated 44%)
```



```
bda@bda-VirtualBox:~/wordcount$ jar -cvf WordCount.jar -C '/home/bda/wordcount/wordcount_classes'/.
added manifest
```

```
adding: WordCount$IntSumReducer.class(in = 1755) (out= 750)(deflated 57%)
```

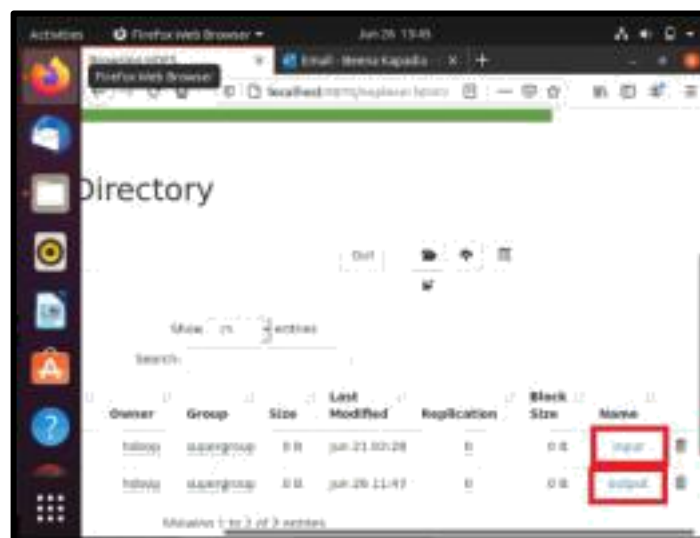
```
adding: WordCount$TokenizerMapper.class(in = 1752) (out= 761)(deflated 56%)
```

```
adding: WordCount.class(in = 1511) (out= 832)(deflated 44%)
```

Compiled By: Ms. Beena Kapadia Vidyalankar School of Information Technology 6

```
hadoop@bda-VirtualBox:/home/bda/wordcount$ hadoop jar
```

```
'/home/bda/wordcount/WordCount.jar' WordCount /wordcount/input/ /wordcount/output output is created.
```





**Get the output:**

```
hadoop@bda-VirtualBox:/home/bda/wordcount$ hdfs dfs -cat /wordcount/output/*
2021-06-26 13:33:57,781 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
beena 2
dhavan 1
jeenisha 1
komal 1
kruti 2
mitish 1
nidhi 3
vinod 2
yashesh 3
hadoop@bda-VirtualBox:/home/bda/wordcount$
```