SANTOSH KASHINATH PARSE
Application No: 113366

Big Data Analytics
Practical Journal
Subject Code: PSIT2P1

M.Sc. (IT)
Part-1 / SEM 2

“VIDYALANKAR SCHOOL OF INFORMATION TECHNOLOGY, WADALA”

AFFILIATED
TO
UNIVERSITY OF MUMBAI

INSTITUTE OF DISTANCE AND OPEN LEARNING (IDOL)

CERTIFICATE

This is to certify that, **Santosh Parse** of M.Sc. (I.T.) Semester - II with Application ID **1 1 3 3 6 6** has completed the practical of **‘Big Data Analytics’** in this college during the academic year **2022 - 2023**.

Subject In-Charge **Coordinator -In-Charge**

Prof. Ujwala Sav

Examined By:

Table of Contents

Practical No: 1

Aim: K-Means Clustering using R Studio

Description:

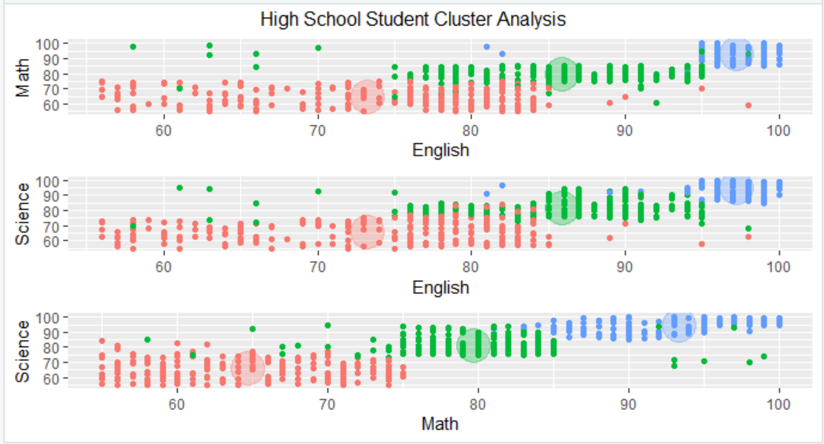
Practical No: 1

Aim: K-Means Clustering using R Studio

Code:

```
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 No: 2

Aim: Apriori Algorithm using R Studio

Description:

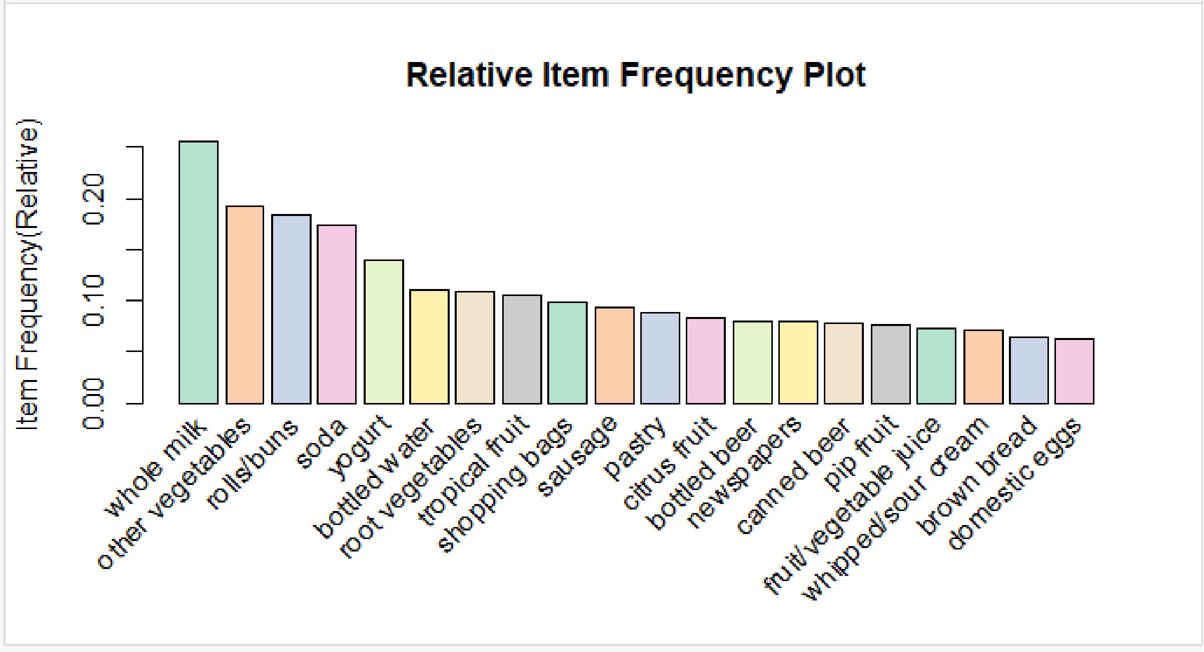
Practical No: 2

Aim: Apriori Algorithm using R Studio

Code:

```
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:



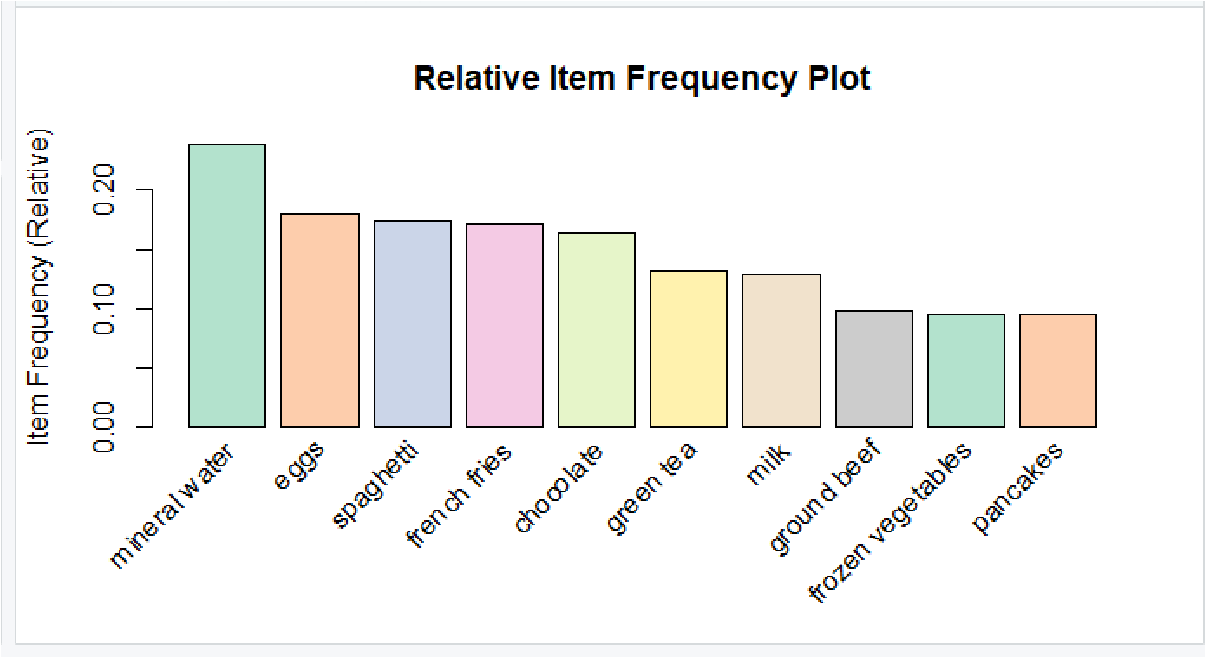
Practical No: 2

Aim: Apriori Algorithm using R Studio

Code:

```
# Apriori
# 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)
# 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:



Practical No: 3

Aim: Simple Linear Regression and Logistic Regression using R Studio

Description:

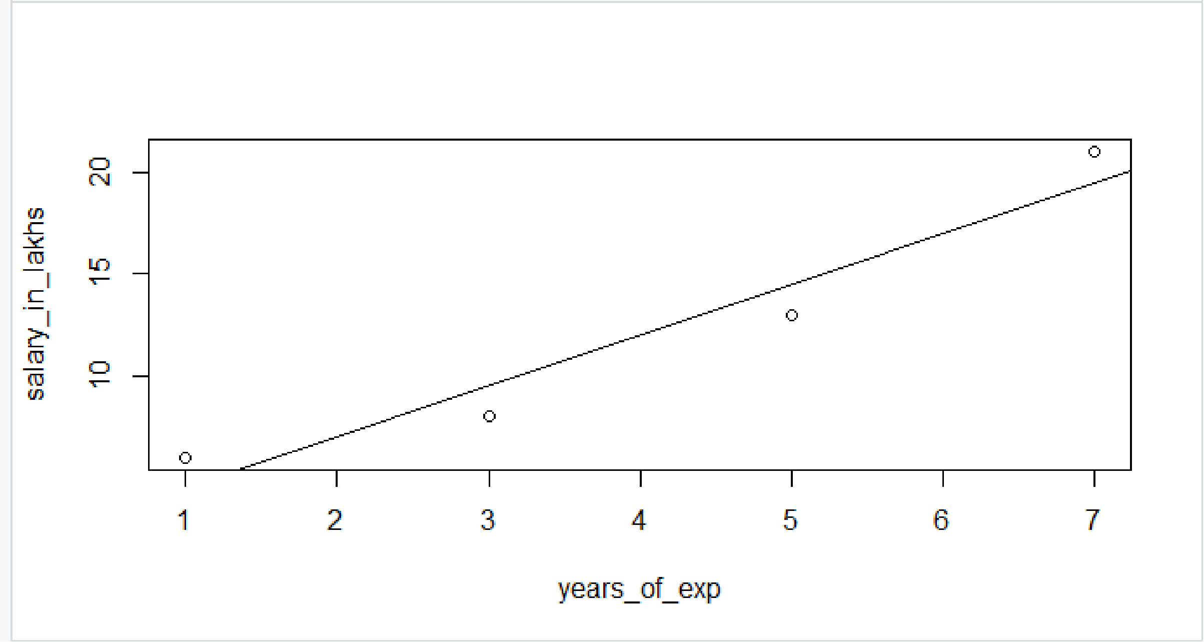
Practical No: 3

Aim: Simple Linear Regression and Logistic Regression using R Studio

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:



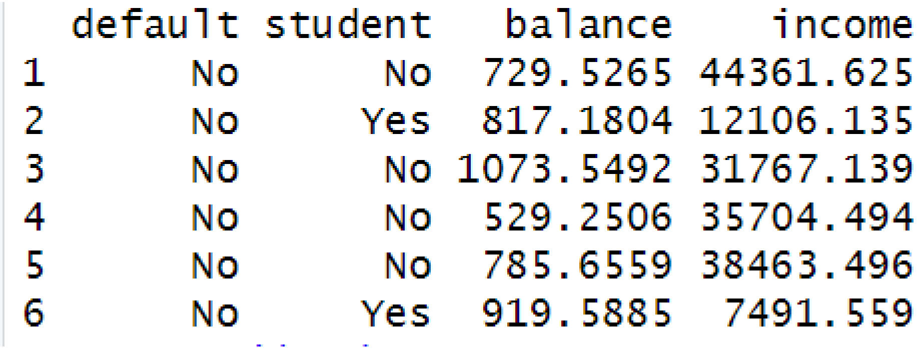
Practical No: 3

Aim: Simple Linear Regression and Logistic Regression using R Studio

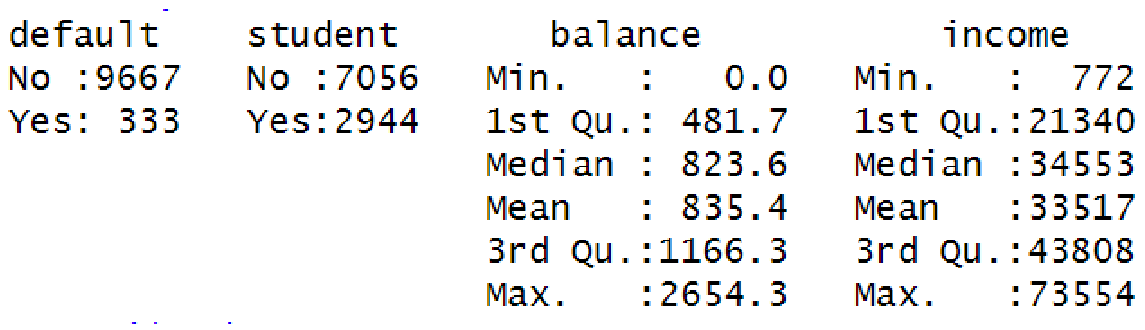
Code :

```
install.packages("InformationValue")
install.packages("devtools")
devtools::install_github("selva86/InformationValue")
library(ISLR)
library(InformationValue)
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)


predicted <- predict(model,test,type="response")
confusionMatrix(test$default,predicted)
```

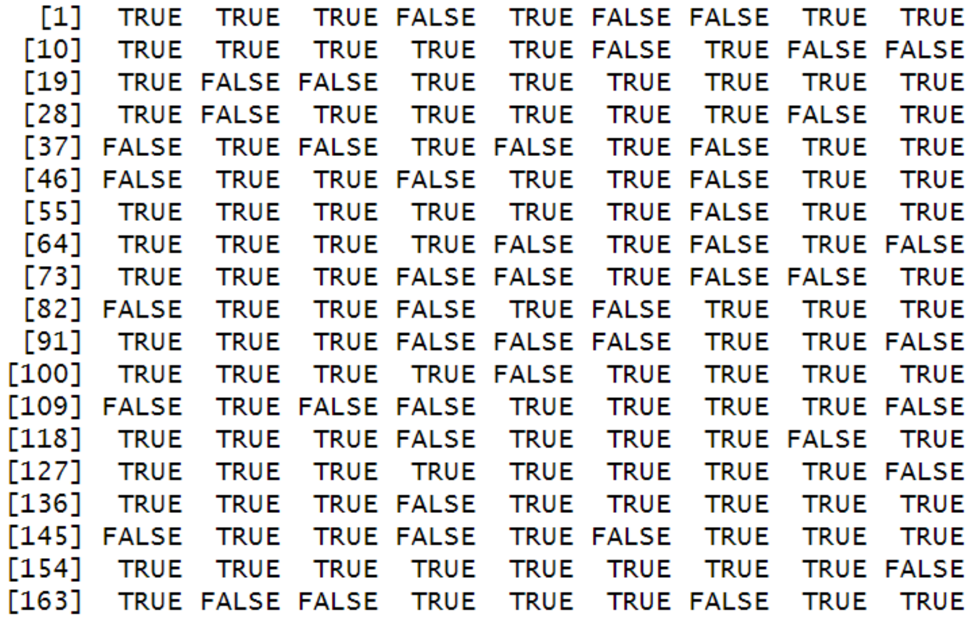
Output:



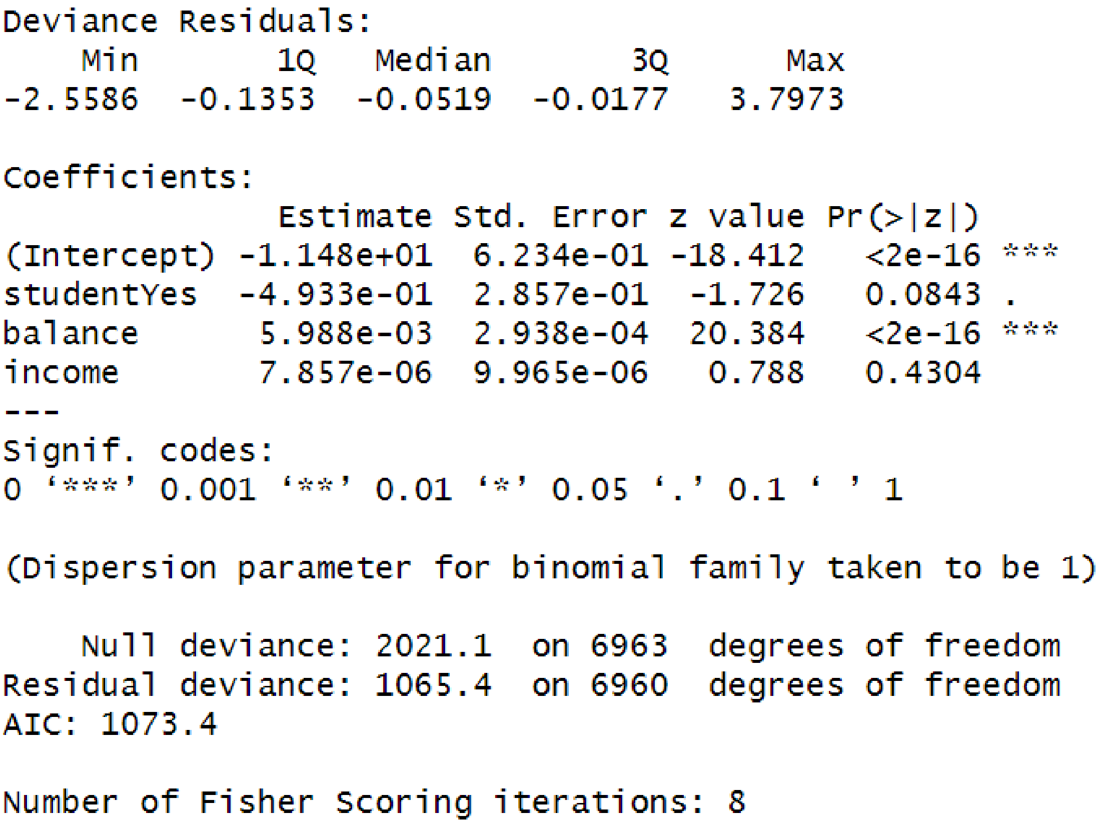
```
  default student   balance    income
1      No      No  729.5265 44361.625
2      No     Yes  817.1804 12106.135
3      No      No 1073.5492 31767.139
4      No      No  529.2506 35704.494
5      No      No  785.6559 38463.496
6      No     Yes  919.5885  7491.559
```



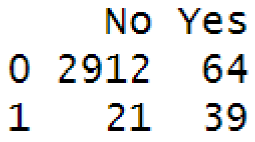
```
 default     student       balance           income     
 No :9667   No :7056   Min.   :   0.0   Min.   :  772  
 Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340  
                       Median : 823.6   Median :34553  
                       Mean   : 835.4   Mean   :33517  
                       3rd Qu.:1166.3   3rd Qu.:43808  
                       Max.   :2654.3   Max.   :73554  
```



```
  [1]  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE
 [10]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE
 [19]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
 [28]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
 [37] FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
 [46] FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE
 [55]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
 [64]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE
 [73]  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE
 [82] FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
 [91]  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE
[100]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
[109] FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
[118]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE
[127]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
[136]  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
[145] FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE
[154]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
[163]  TRUE FALSE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```



```
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
```



```
    No  Yes
0 2912   64
1   21   39
```

Practical No: 4

Aim: Decision tree classification using R studio.

Description:

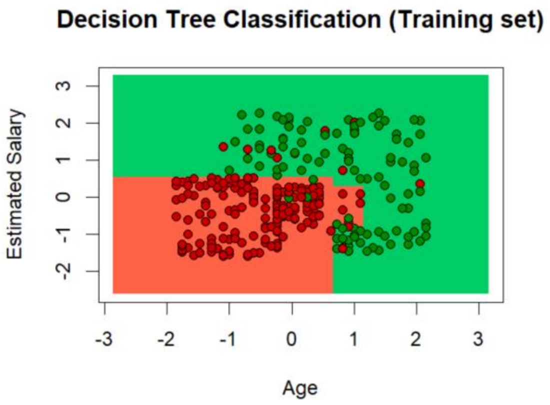
Practical No: 4

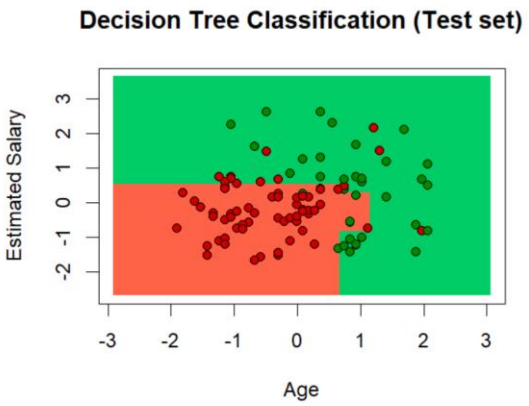
Aim: Decision tree classification using R studio.

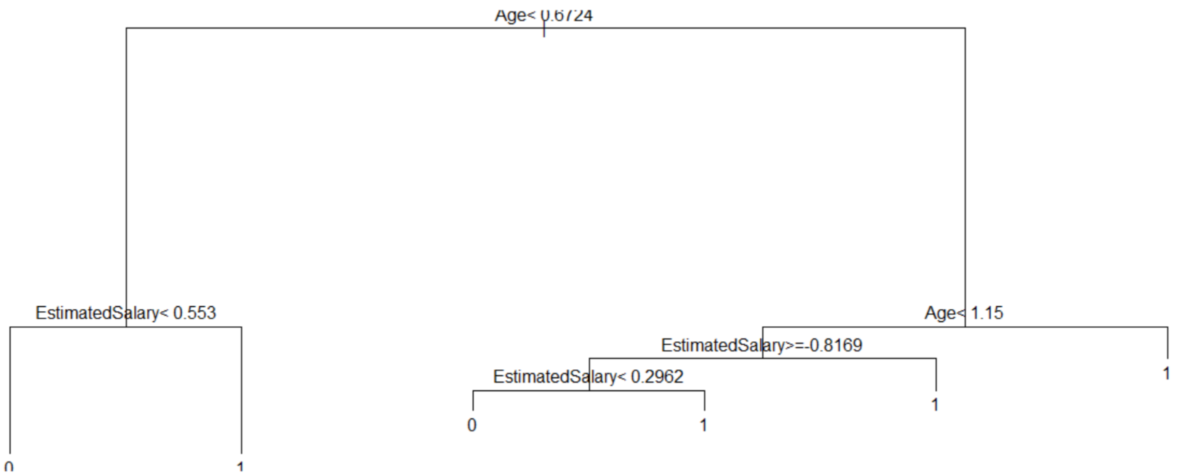
Code:

```
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:


Decision Tree Classification (Training set)
Estimated Salary
-2
-1
0
1
2
3
-3
-2
-1
0
1
2
3
Age


Decision Tree Classification (Test set)
Estimated Salary
-2
-1
0
1
2
3
-3
-2
-1
0
1
2
3
Age


Age< 0.6724
EstimatedSalary< 0.553
Age< 1.15
EstimatedSalary>=-0.8169
EstimatedSalary< 0.2962
0
1
0
1
1
1

Practical No: 5

Aim: Naive Bayes Classification using R Studio

Description:

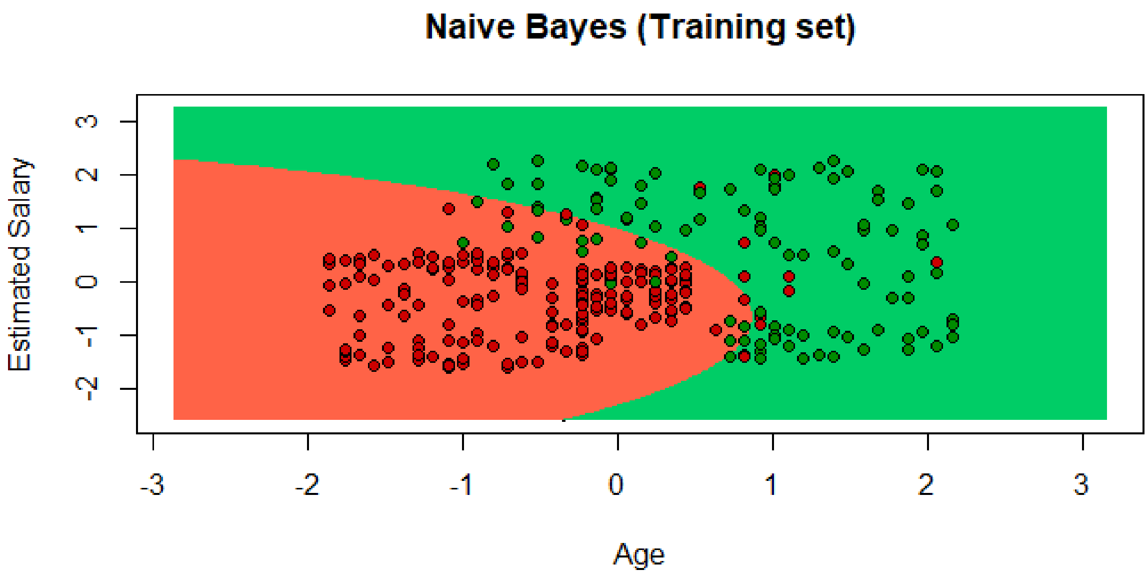
Practical No: 5

Aim: Naive Bayes Classification using R Studio

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("ElemStatLearn")
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 No: 6

Aim: Text Analysis using R Studio

Description:

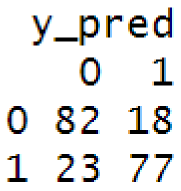
Practical No: 6

Aim: Text Analysis using R Studio

Code:

```
dataset_original =
read.delim('C:\\Playground\\msc_practical\\sem2\\big_data_analytics\\data\\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 No: 7

Aim: Virtual Box Installation

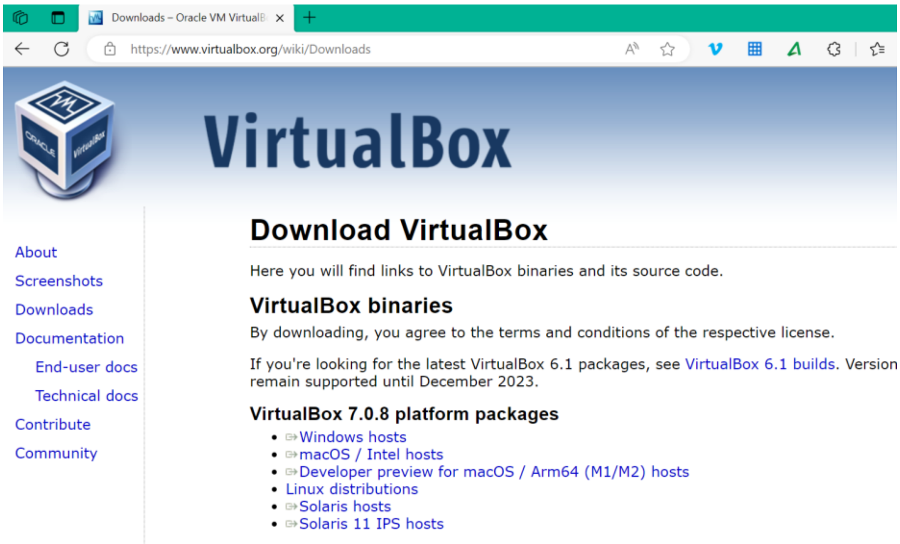
Description:

Practical No: 7

Aim: Virtual Box Installation

Installation Steps:

Step 1: Download Virtual Box from https://www.virtualbox.org/wiki/Downloads depending on platform you want to install (here Windows)



Download VirtualBox

Here you will find links to VirtualBox binaries and its source code.

VirtualBox binaries

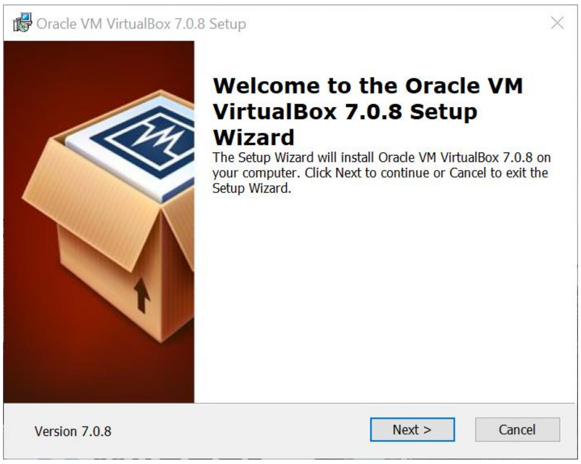
By downloading, you agree to the terms and conditions of the respective license.

If you're looking for the latest VirtualBox 6.1 packages, see VirtualBox 6.1 builds. Version remain supported until December 2023.

VirtualBox 7.0.8 platform packages

- Windows hosts
- macOS / Intel hosts
- Developer preview for macOS / Arm64 (M1/M2) hosts
- Linux distributions
- Solaris hosts
- Solaris 11 IPS hosts

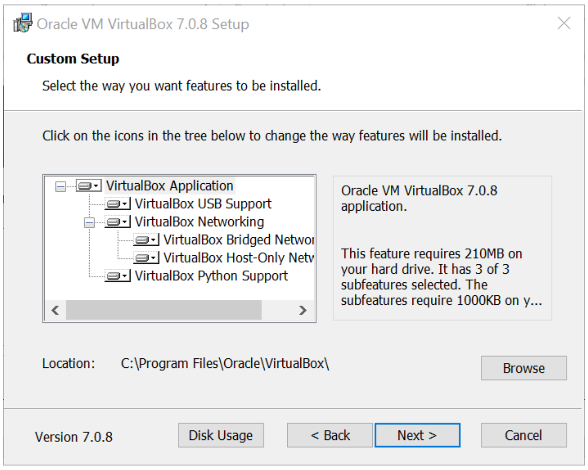
Step 2: Run downloaded installer with default selection



Oracle VM VirtualBox 7.0.8 Setup

Welcome to the Oracle VM VirtualBox 7.0.8 Setup Wizard

The Setup Wizard will install Oracle VM VirtualBox 7.0.8 on your computer. Click Next to continue or Cancel to exit the Setup Wizard.



Oracle VM VirtualBox 7.0.8 Setup

Custom Setup

Select the way you want features to be installed.

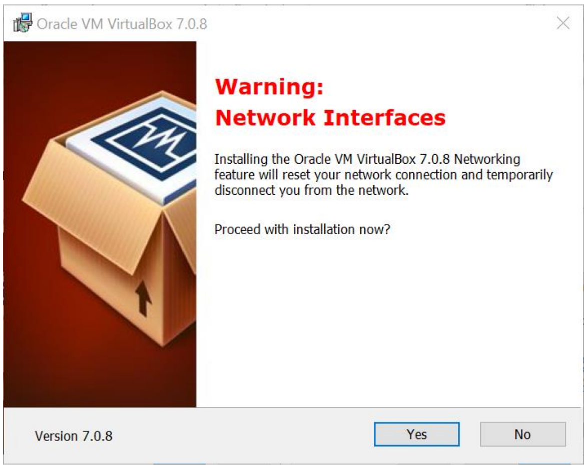
Click on the icons in the tree below to change the way features will be installed.

- VirtualBox Application
  - VirtualBox USB Support
  - VirtualBox Networking
    - VirtualBox Bridged Networ
    - VirtualBox Host-Only Netv
  - VirtualBox Python Support

Oracle VM VirtualBox 7.0.8 application.

This feature requires 210MB on your hard drive. It has 3 of 3 subfeatures selected. The subfeatures require 1000KB on y...

Location: C:\Program Files\Oracle\VirtualBox\

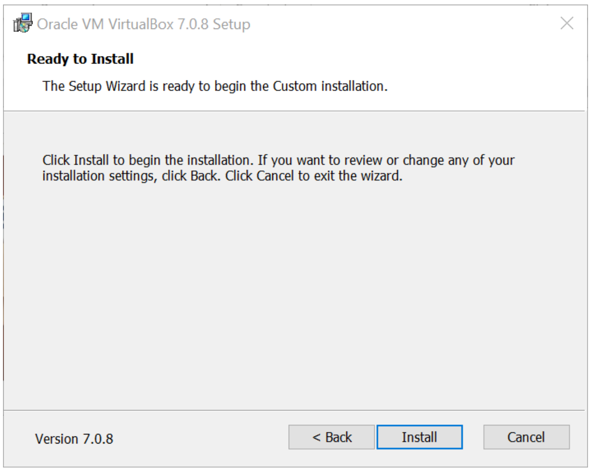


Oracle VM VirtualBox 7.0.8

Warning: Network Interfaces

Installing the Oracle VM VirtualBox 7.0.8 Networking feature will reset your network connection and temporarily disconnect you from the network.

Proceed with installation now?

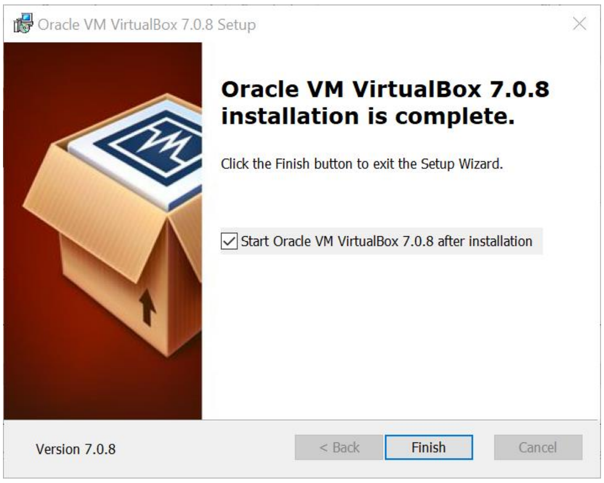


Oracle VM VirtualBox 7.0.8 Setup

Ready to Install

The Setup Wizard is ready to begin the Custom installation.

Click Install to begin the installation. If you want to review or change any of your installation settings, click Back. Click Cancel to exit the wizard.



Oracle VM VirtualBox 7.0.8 Setup

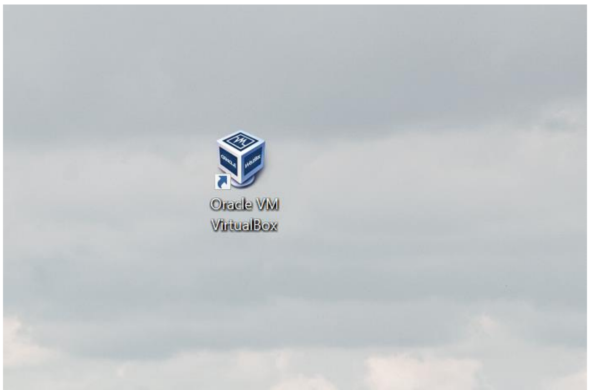
Oracle VM VirtualBox 7.0.8 installation is complete.

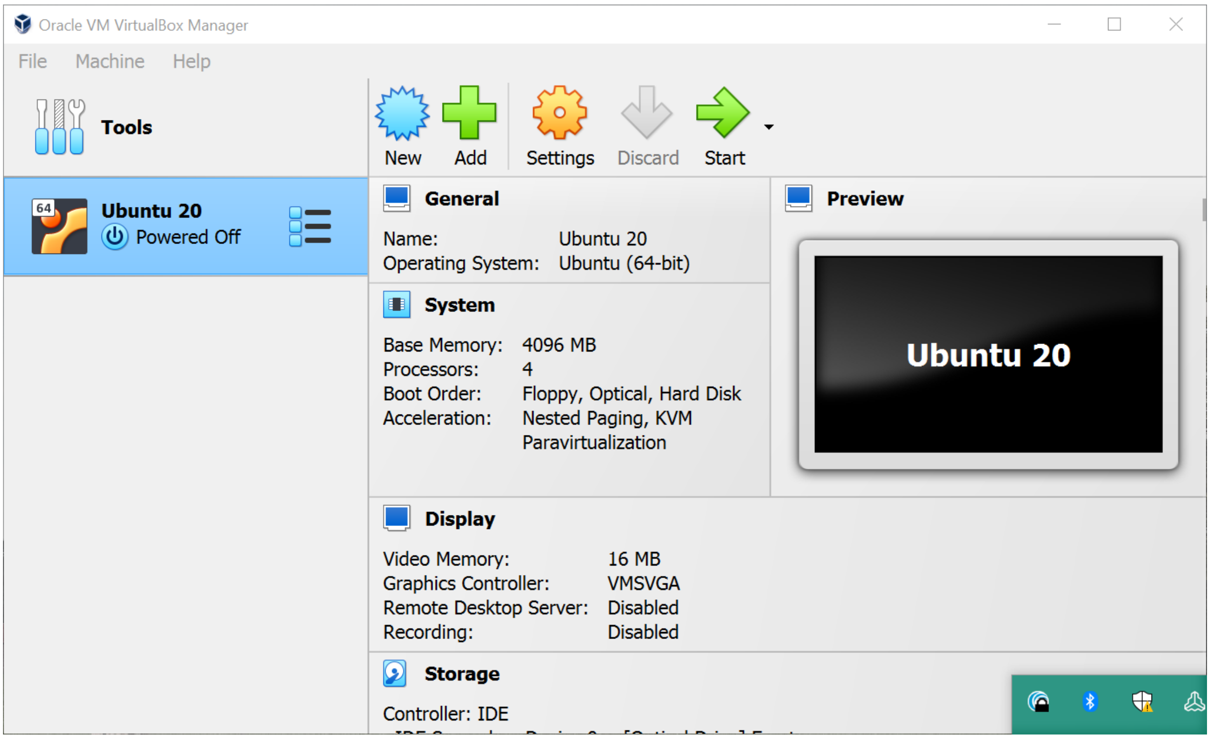
Click the Finish button to exit the Setup Wizard.

☑ Start Oracle VM VirtualBox 7.0.8 after installation

Version 7.0.8

Step 4: Once setup is complete, Virtual Box shortcut is created on desktop.





Oracle VM VirtualBox Manager

Ubuntu 20
Powered Off

General

Name: Ubuntu 20
Operating System: Ubuntu (64-bit)

System

Base Memory: 4096 MB
Processors: 4
Boot Order: Floppy, Optical, Hard Disk
Acceleration: Nested Paging, KVM Paravirtualization

Preview

Display

Video Memory: 16 MB
Graphics Controller: VMSVGA
Remote Desktop Server: Disabled
Recording: Disabled

Storage

Controller: IDE

Virtual Box installation is complete now.

Practical No: 8

Aim: Ubuntu Installation

Description:

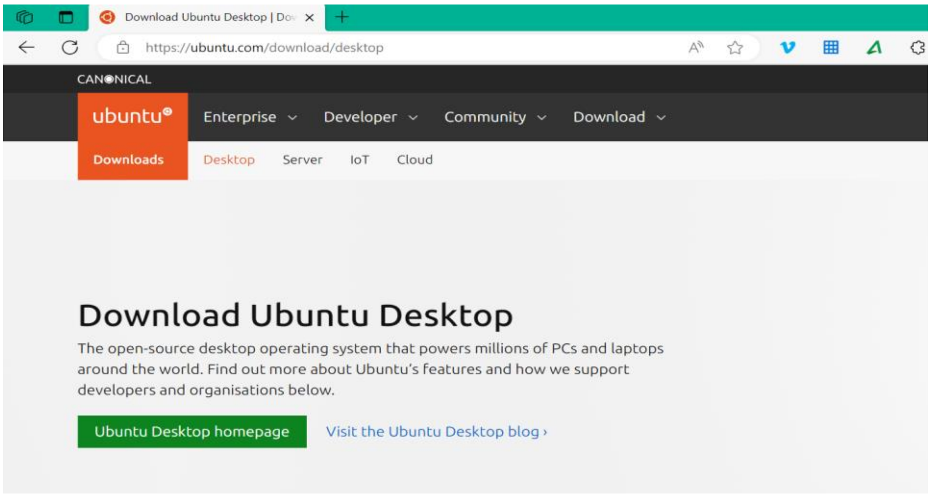
Practical No: 8

Aim: Ubuntu Installation

Installation Steps:

We are going to install ubuntu on virtual box, for this we need ubuntu image.

Step 1: Download ubuntu image from https://ubuntu.com/download/desktop

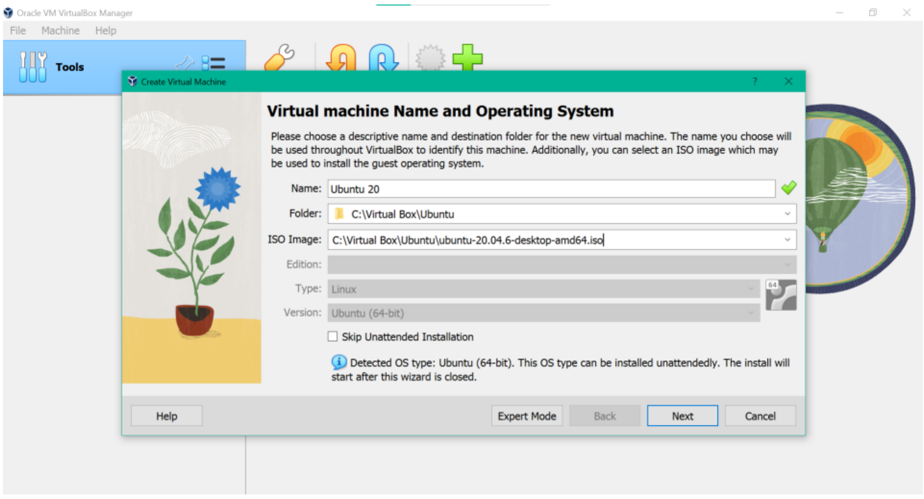


Download Ubuntu Desktop

The open-source desktop operating system that powers millions of PCs and laptops around the world. Find out more about Ubuntu's features and how we support developers and organisations below.

Ubuntu Desktop homepage   Visit the Ubuntu Desktop blog ›

Step 2: Selected downloaded iso image for installation



Virtual machine Name and Operating System

Please choose a descriptive name and destination folder for the new virtual machine. The name you choose will be used throughout VirtualBox to identify this machine. Additionally, you can select an ISO image which may be used to install the guest operating system.

Name: Ubuntu 20

Folder: C:\Virtual Box\Ubuntu

ISO Image: C:\Virtual Box\Ubuntu\ubuntu-20.04.6-desktop-amd64.iso

Edition:

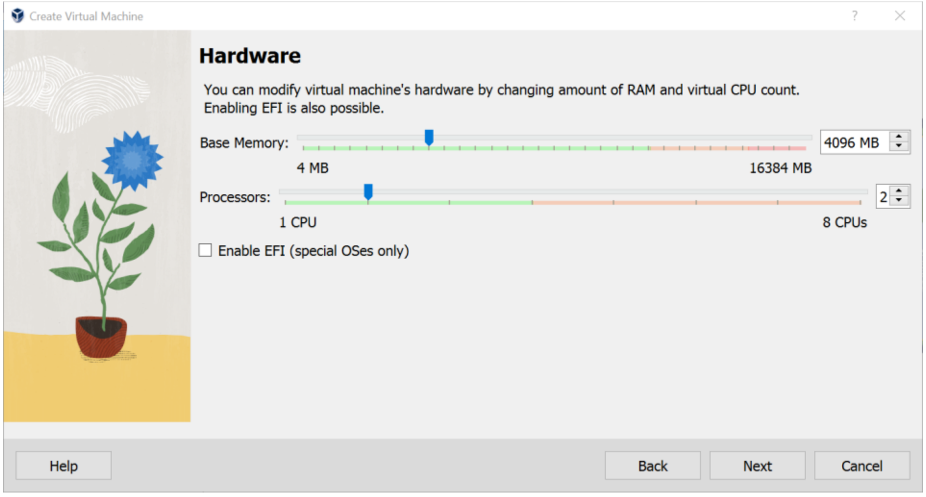
Type: Linux

Version: Ubuntu (64-bit)

☐ Skip Unattended Installation

Detected OS type: Ubuntu (64-bit). This OS type can be installed unattendedly. The install will start after this wizard is closed.

Step 3: Configure RAM & Processor



Hardware

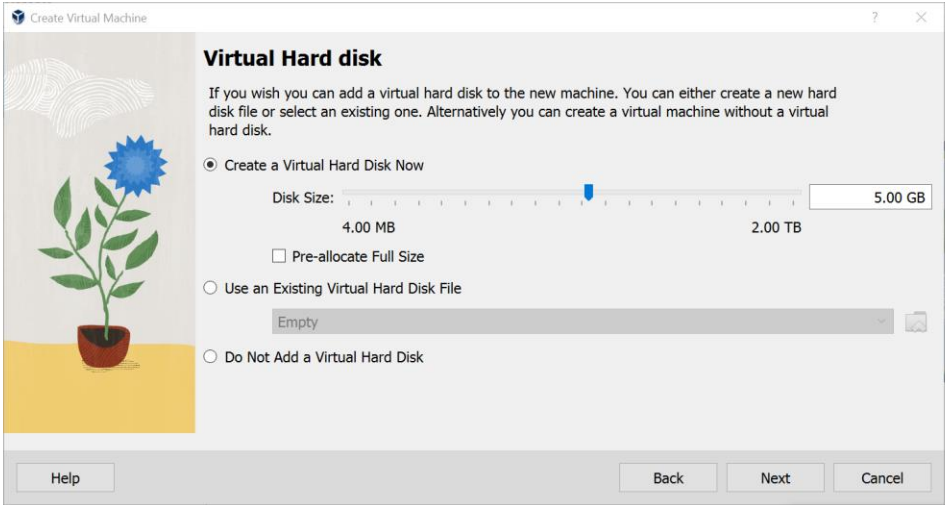
You can modify virtual machine's hardware by changing amount of RAM and virtual CPU count. Enabling EFI is also possible.

Base Memory: 4096 MB (4 MB – 16384 MB)

Processors: 2 (1 CPU – 8 CPUs)

☐ Enable EFI (special OSes only)

Step 4: Allocate disk size



Virtual Hard disk

If you wish you can add a virtual hard disk to the new machine. You can either create a new hard disk file or select an existing one. Alternatively you can create a virtual machine without a virtual hard disk.

● Create a Virtual Hard Disk Now

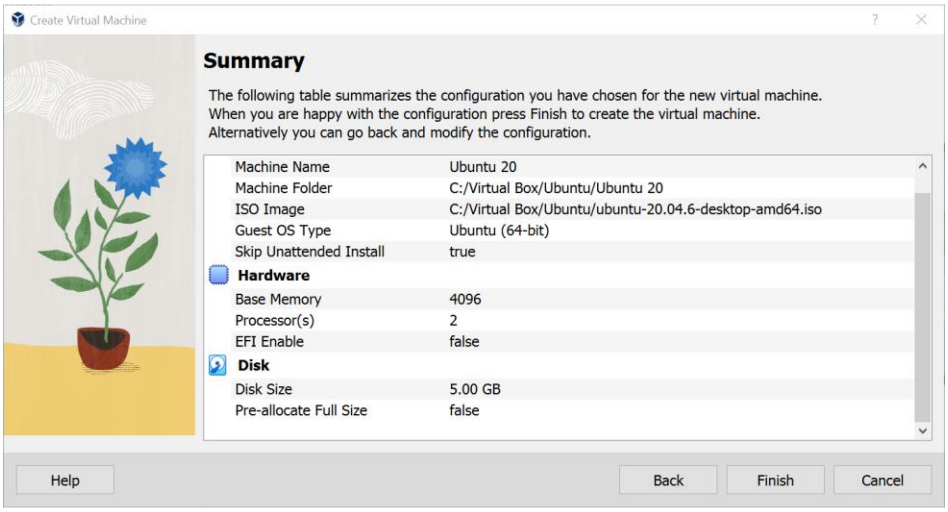
Disk Size: 5.00 GB (4.00 MB – 2.00 TB)

□ Pre-allocate Full Size

○ Use an Existing Virtual Hard Disk File

Empty

○ Do Not Add a Virtual Hard Disk

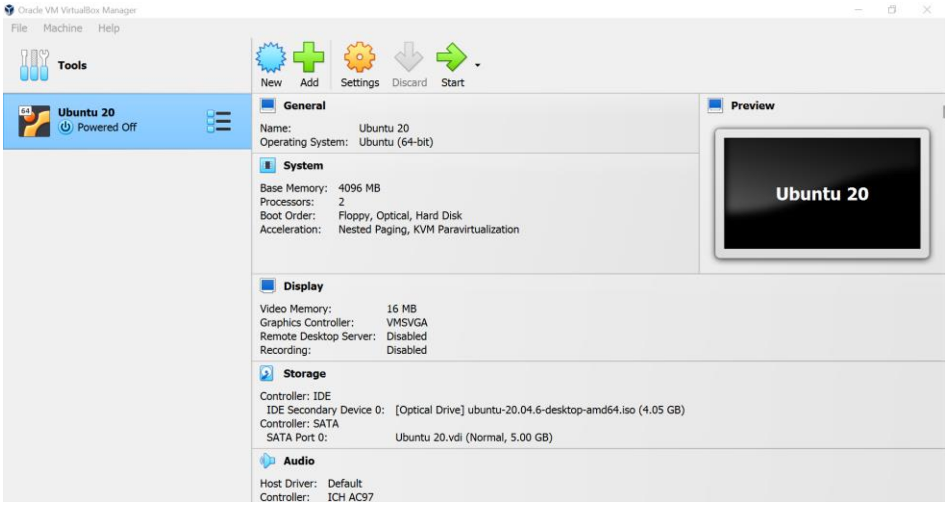


Summary

The following table summarizes the configuration you have chosen for the new virtual machine. When you are happy with the configuration press Finish to create the virtual machine. Alternatively you can go back and modify the configuration.

Machine Name	Ubuntu 20
Machine Folder	C:/Virtual Box/Ubuntu/Ubuntu 20
ISO Image	C:/Virtual Box/Ubuntu/ubuntu-20.04.6-desktop-amd64.iso
Guest OS Type	Ubuntu (64-bit)
Skip Unattended Install	true
Hardware	
Base Memory	4096
Processor(s)	2
EFI Enable	false
Disk	
Disk Size	5.00 GB
Pre-allocate Full Size	false

Step 5: Start virtual box for initial installation of Ubuntu



General

```
Name:              Ubuntu 20
Operating System:  Ubuntu (64-bit)
```

System

```
Base Memory:   4096 MB
Processors:    2
Boot Order:    Floppy, Optical, Hard Disk
Acceleration:  Nested Paging, KVM Paravirtualization
```

Display

```
Video Memory:            16 MB
Graphics Controller:     VMSVGA
Remote Desktop Server:   Disabled
Recording:               Disabled
```

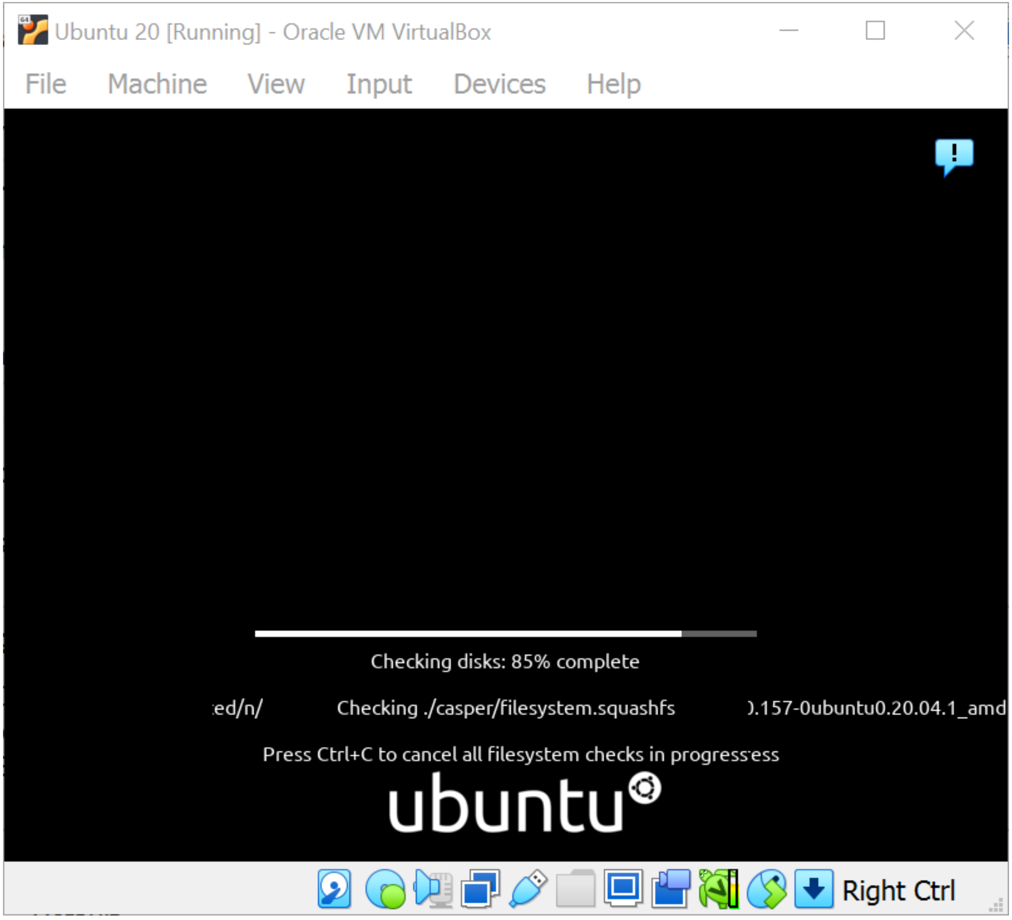
Storage

```
Controller: IDE
  IDE Secondary Device 0:  [Optical Drive] ubuntu-20.04.6-desktop-amd64.iso (4.05 GB)
Controller: SATA
  SATA Port 0:             Ubuntu 20.vdi (Normal, 5.00 GB)
```

Audio

```
Host Driver:  Default
Controller:   ICH AC97
```

Step 6: Ubuntu installation progress

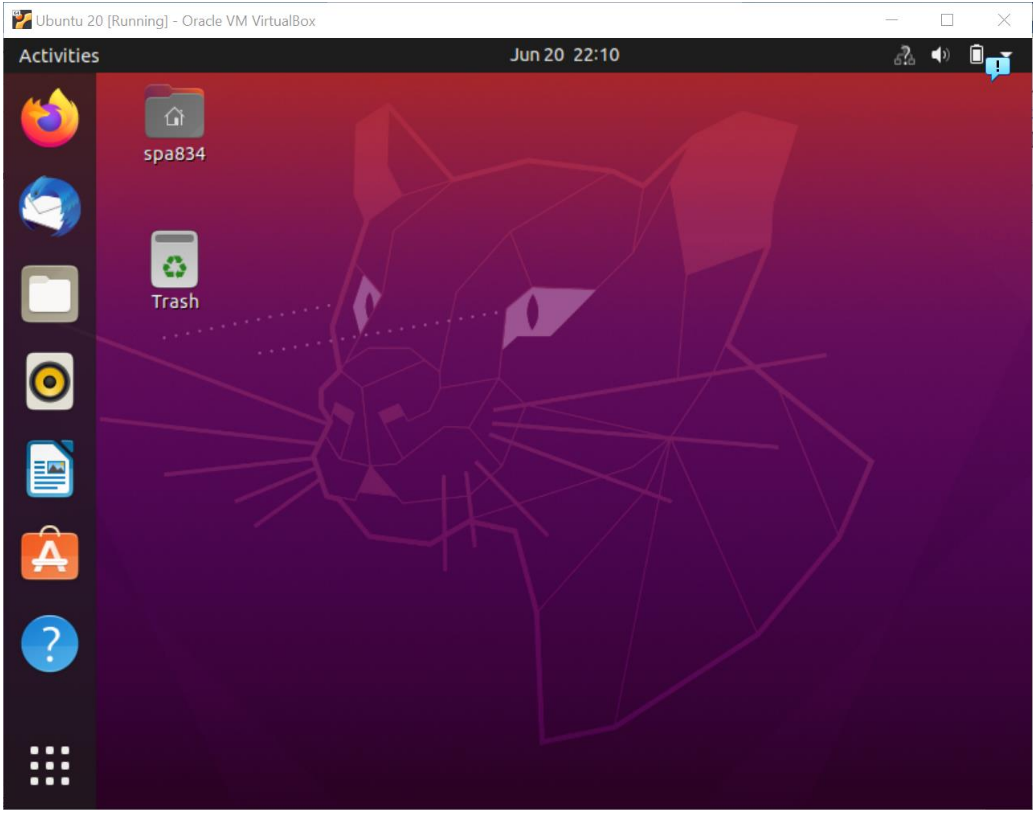

Ubuntu 20 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Right Ctrl

Checking disks: 85% complete

:ed/n/ Checking ./casper/filesystem.squashfs).157-0ubuntu0.20.04.1_amd

Press Ctrl+C to cancel all filesystem checks in progressess

Step 7: Again launch ubuntu virtual box, this time it starts Ubuntu operating system.



Ubuntu installed successfully.

Practical No: 9

Aim: Hadoop Installation

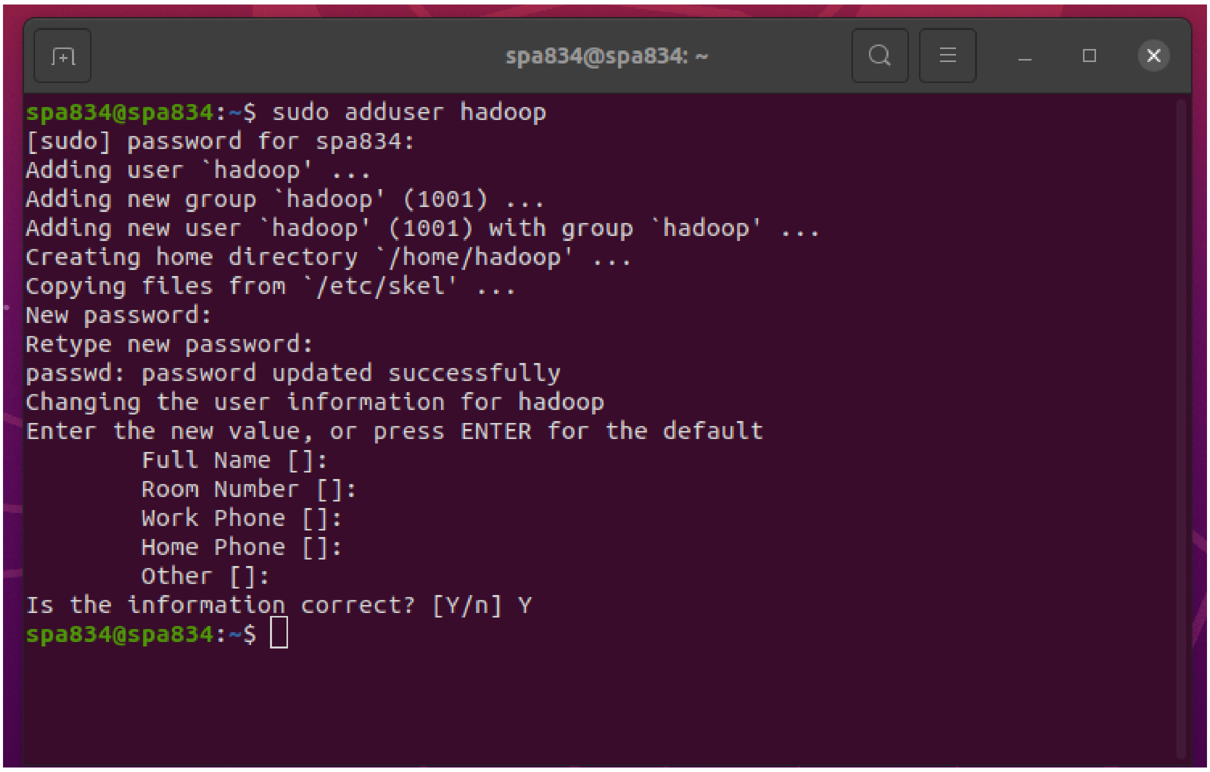
Description:

Practical No: 9

Aim: Hadoop Installation

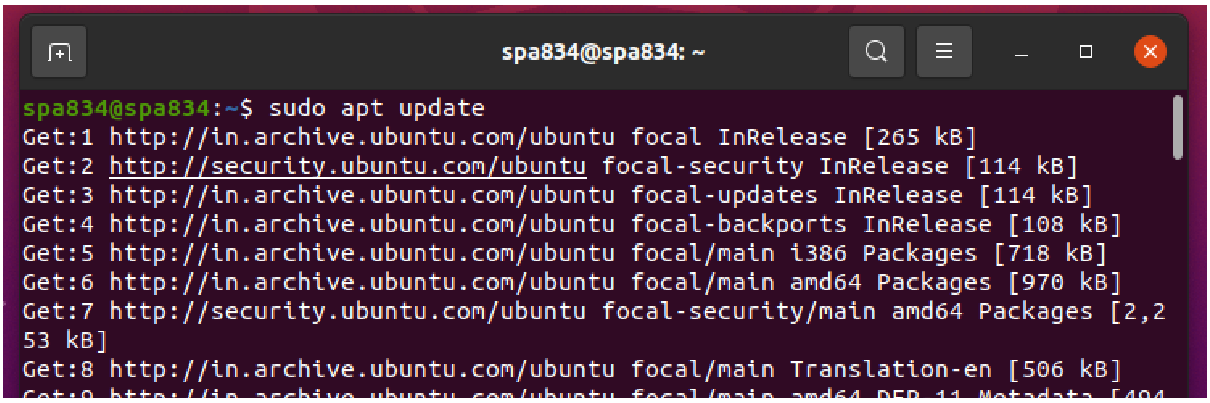
Installation Steps:

Step 1: Create user for Hadoop environment

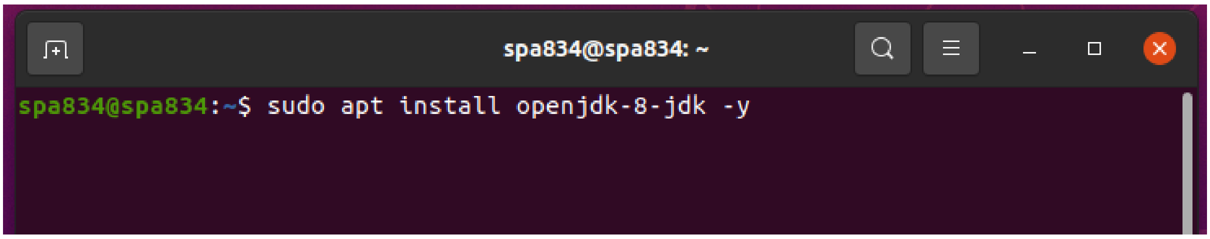


```
spa834@spa834:~$ sudo adduser hadoop
[sudo] password for spa834:
Adding user `hadoop' ...
Adding new group `hadoop' (1001) ...
Adding new user `hadoop' (1001) with group `hadoop' ...
Creating home directory `/home/hadoop' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hadoop
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
spa834@spa834:~$
```

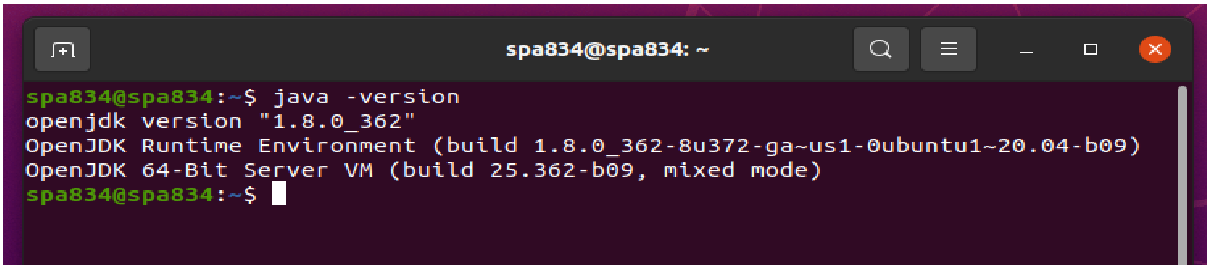
Step 2: Install java



```
spa834@spa834:~$ sudo apt update
Get:1 http://in.archive.ubuntu.com/ubuntu focal InRelease [265 kB]
Get:2 http://security.ubuntu.com/ubuntu focal-security InRelease [114 kB]
Get:3 http://in.archive.ubuntu.com/ubuntu focal-updates InRelease [114 kB]
Get:4 http://in.archive.ubuntu.com/ubuntu focal-backports InRelease [108 kB]
Get:5 http://in.archive.ubuntu.com/ubuntu focal/main i386 Packages [718 kB]
Get:6 http://in.archive.ubuntu.com/ubuntu focal/main amd64 Packages [970 kB]
Get:7 http://security.ubuntu.com/ubuntu focal-security/main amd64 Packages [2,2
53 kB]
Get:8 http://in.archive.ubuntu.com/ubuntu focal/main Translation-en [506 kB]
Get:9 http://in.archive.ubuntu.com/ubuntu focal/main amd64 DEP-11 Metadata [494
```

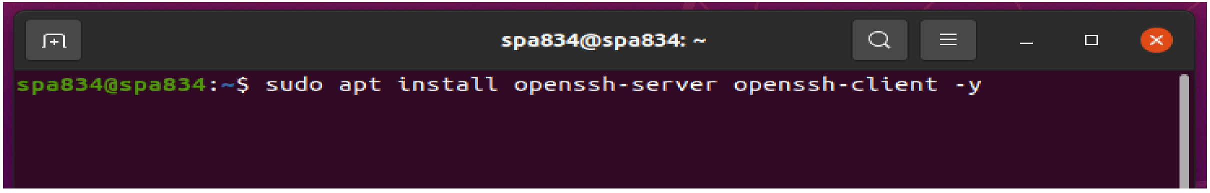


```
spa834@spa834:~$ sudo apt install openjdk-8-jdk -y
```

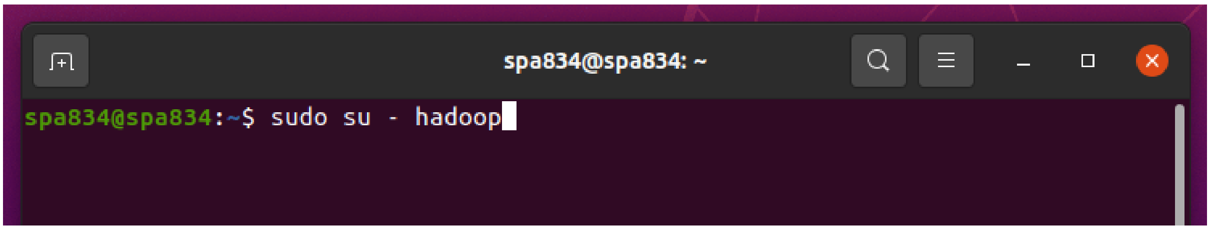


```
spa834@spa834:~$ java -version
openjdk version "1.8.0_362"
OpenJDK Runtime Environment (build 1.8.0_362-8u372-ga~us1-0ubuntu1~20.04-b09)
OpenJDK 64-Bit Server VM (build 25.362-b09, mixed mode)
spa834@spa834:~$
```

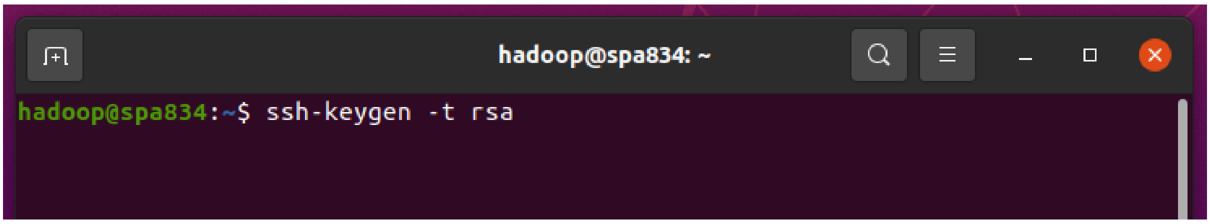
Step 3: Install OpenSSH



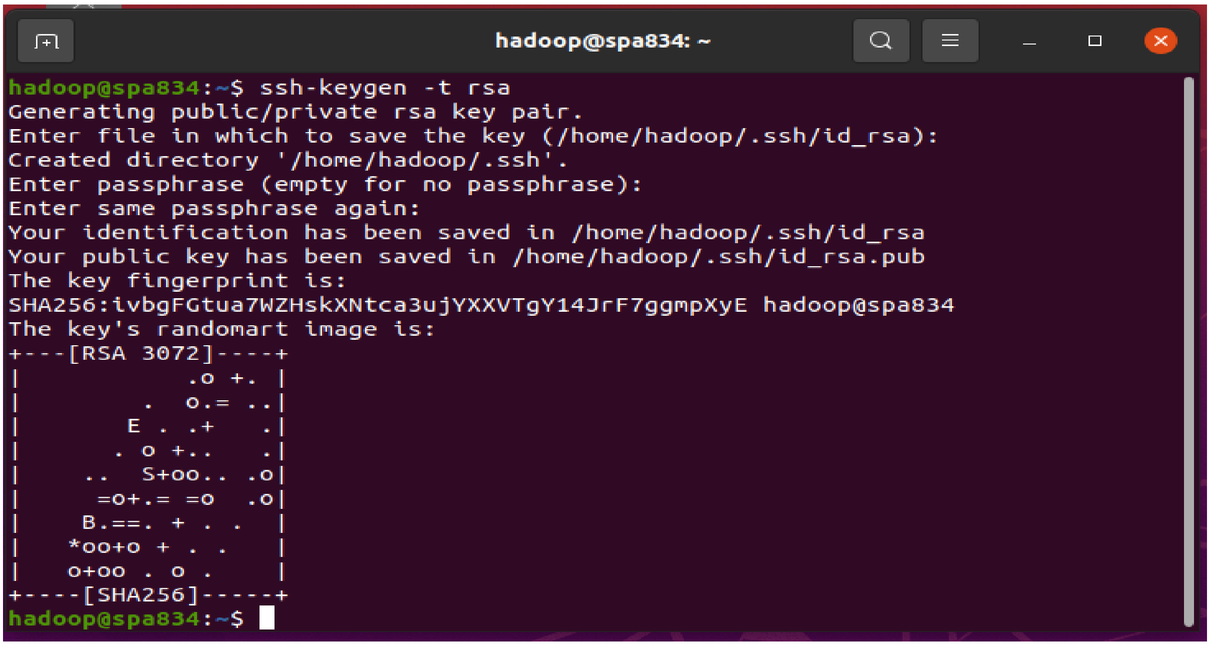
```
spa834@spa834:~$ sudo apt install openssh-server openssh-client -y
```



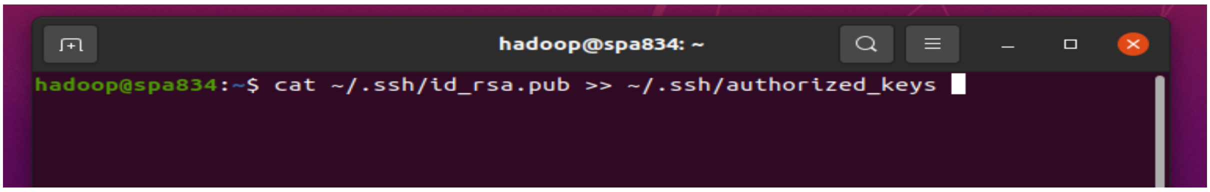
```
spa834@spa834:~$ sudo su - hadoop
```



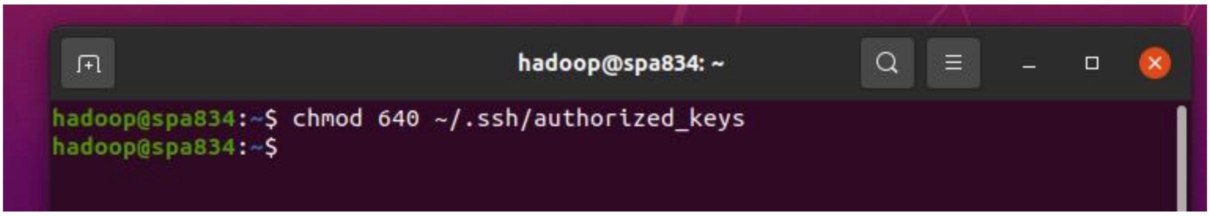
```
hadoop@spa834:~$ ssh-keygen -t rsa
```



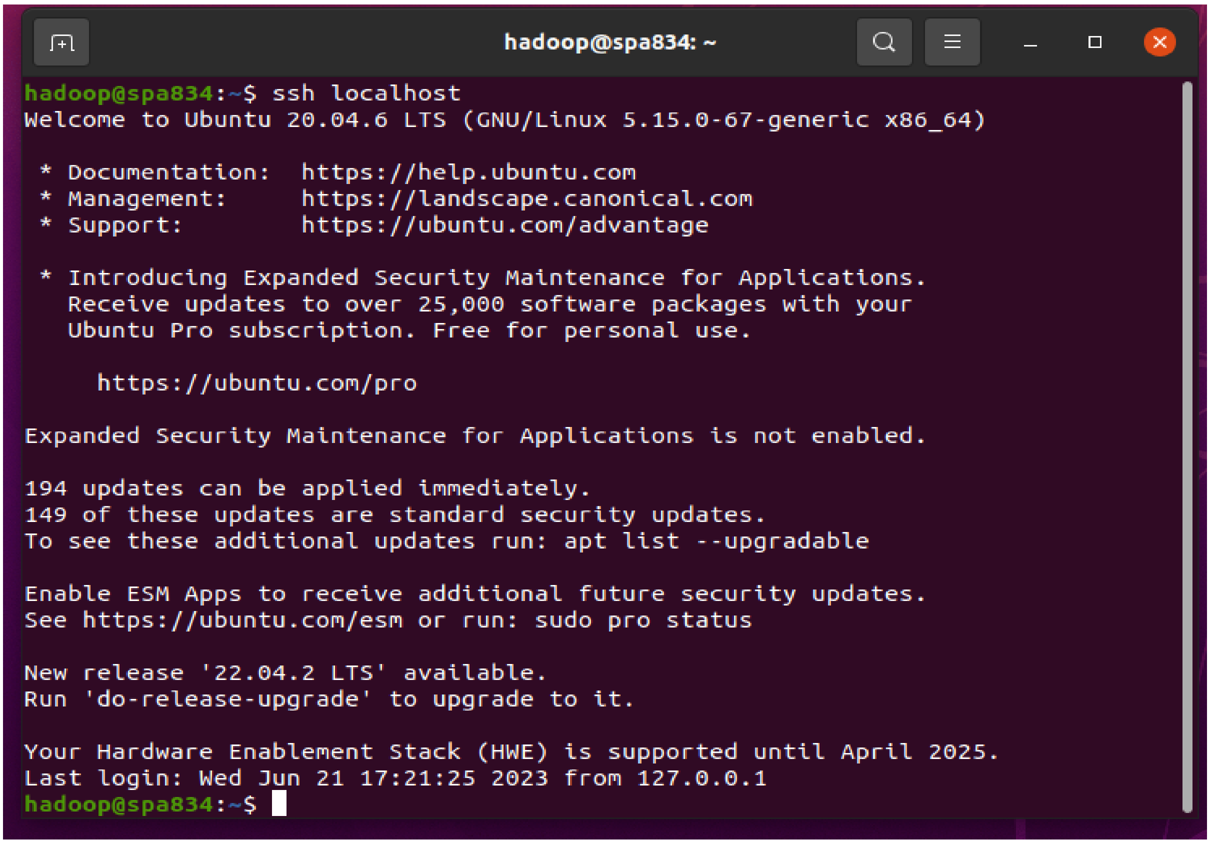
```
hadoop@spa834:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:ivbgFGtua7WZHskXNtca3ujYXXVTgY14JrF7ggmpXyE hadoop@spa834
The key's randomart image is:
+---[RSA 3072]----+
|           .o +. |
|        .  o.= ..|
|       E . .+    .|
|      . o +..    .|
|    ..  S+oo.. .o|
|     =o+.= =o  .o|
|    B.==. + . .  |
|   *oo+o + . .   |
|   o+oo . o .    |
+----[SHA256]-----+
hadoop@spa834:~$
```



```
hadoop@spa834:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

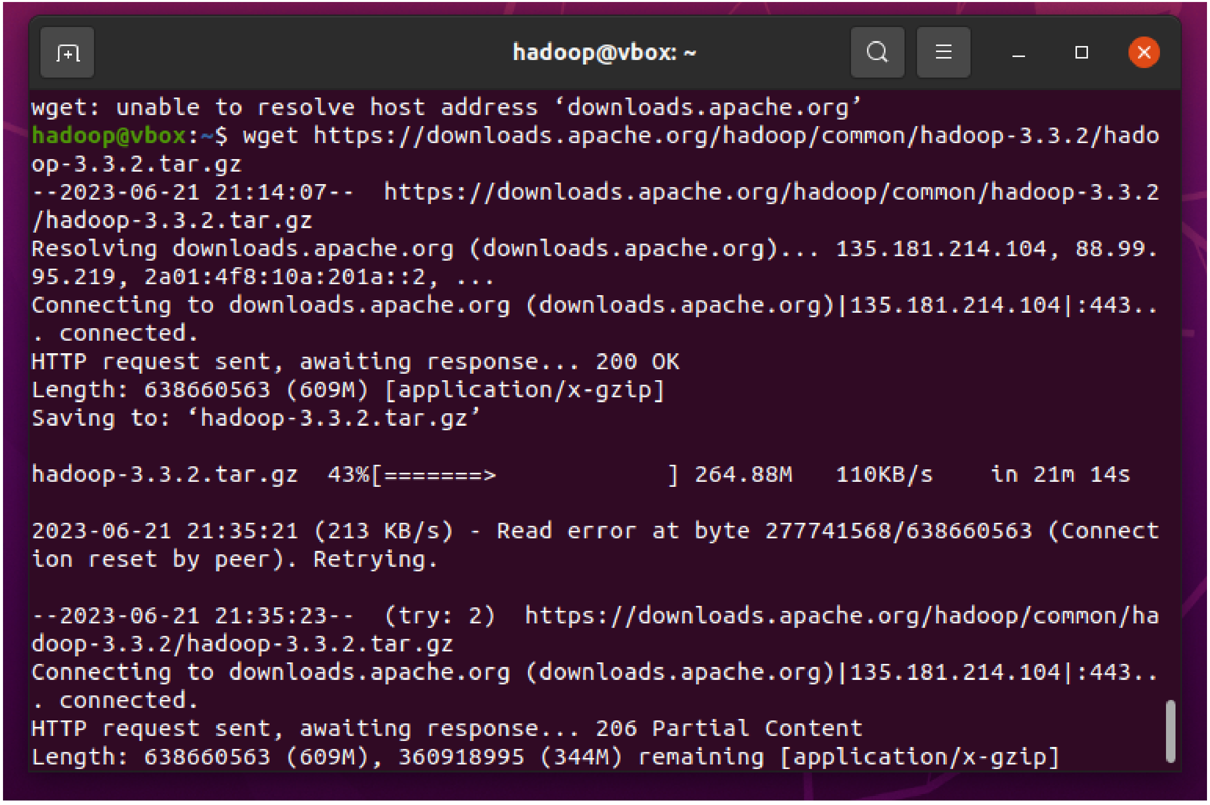


```
hadoop@spa834:~$ chmod 640 ~/.ssh/authorized_keys
hadoop@spa834:~$
```

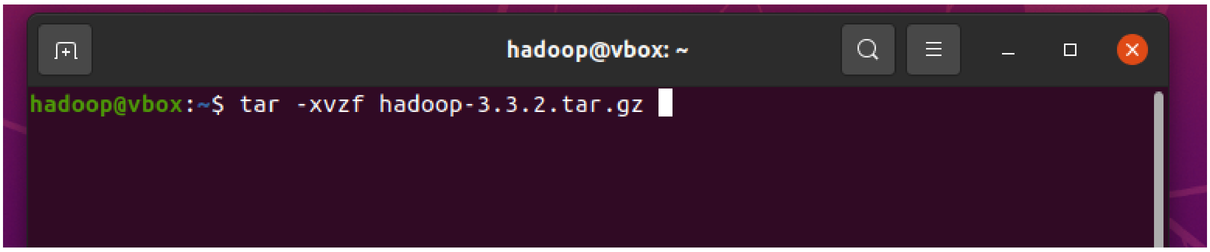


```
hadoop@spa834:~$ ssh localhost
Welcome to Ubuntu 20.04.6 LTS (GNU/Linux 5.15.0-67-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Introducing Expanded Security Maintenance for Applications.
   Receive updates to over 25,000 software packages with your
   Ubuntu Pro subscription. Free for personal use.

     https://ubuntu.com/pro

Expanded Security Maintenance for Applications is not enabled.

194 updates can be applied immediately.
149 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

New release '22.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Your Hardware Enablement Stack (HWE) is supported until April 2025.
Last login: Wed Jun 21 17:21:25 2023 from 127.0.0.1
hadoop@spa834:~$
```

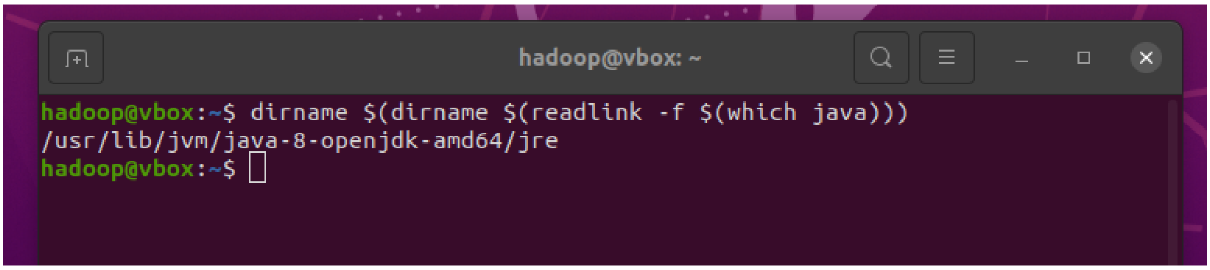
Step 4: Install Apache Hadoop



```
wget: unable to resolve host address ‘downloads.apache.org’
hadoop@vbox:~$ wget https://downloads.apache.org/hadoop/common/hadoop-3.3.2/hado
op-3.3.2.tar.gz
--2023-06-21 21:14:07--  https://downloads.apache.org/hadoop/common/hadoop-3.3.2
/hadoop-3.3.2.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 135.181.214.104, 88.99.
95.219, 2a01:4f8:10a:201a::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443..
. connected.
HTTP request sent, awaiting response... 200 OK
Length: 638660563 (609M) [application/x-gzip]
Saving to: ‘hadoop-3.3.2.tar.gz’

hadoop-3.3.2.tar.gz  43%[=======>            ] 264.88M   110KB/s    in 21m 14s

2023-06-21 21:35:21 (213 KB/s) - Read error at byte 277741568/638660563 (Connect
ion reset by peer). Retrying.

--2023-06-21 21:35:23--  (try: 2)  https://downloads.apache.org/hadoop/common/ha
doop-3.3.2/hadoop-3.3.2.tar.gz
Connecting to downloads.apache.org (downloads.apache.org)|135.181.214.104|:443..
. connected.
HTTP request sent, awaiting response... 206 Partial Content
Length: 638660563 (609M), 360918995 (344M) remaining [application/x-gzip]
```

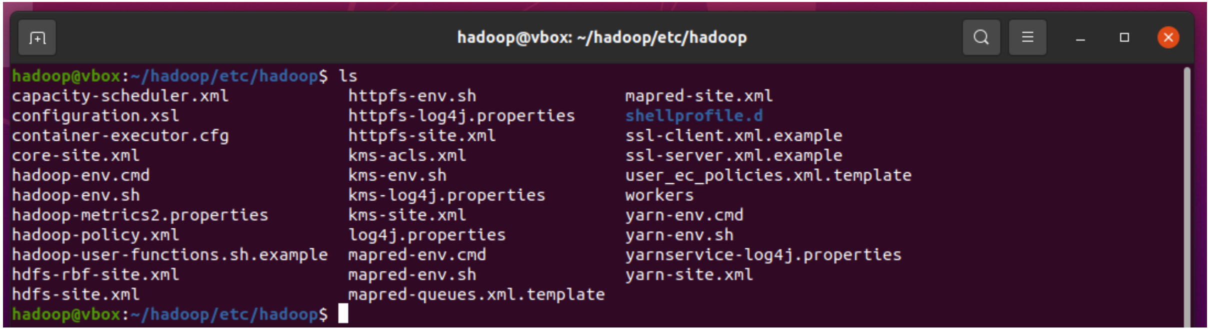


```
hadoop@vbox:~$ tar -xvzf hadoop-3.3.2.tar.gz
```



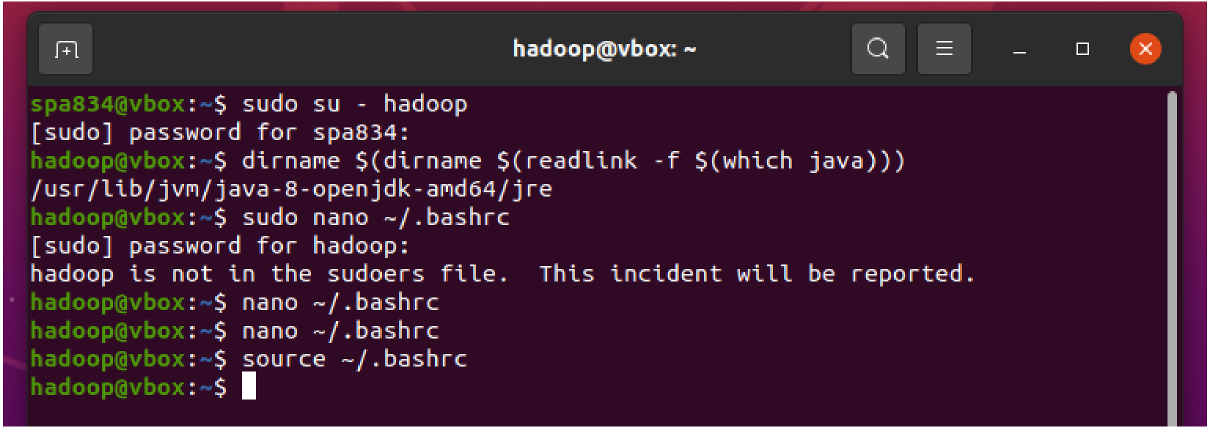
```
hadoop@vbox:~$ dirname $(dirname $(readlink -f $(which java)))
/usr/lib/jvm/java-8-openjdk-amd64/jre
hadoop@vbox:~$
```

Step 5: Configure Hadoop



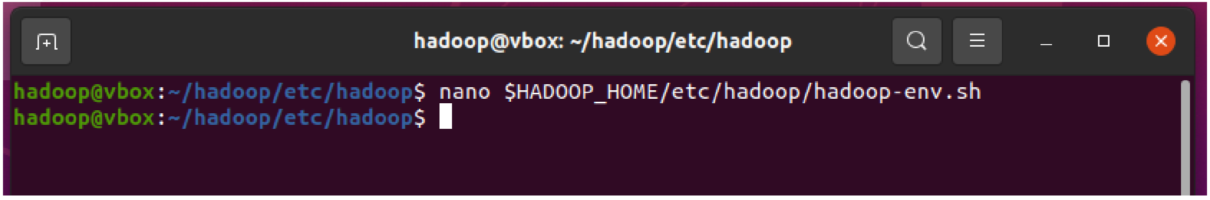
```
hadoop@vbox:~/hadoop/etc/hadoop$ ls
capacity-scheduler.xml            httpfs-env.sh              mapred-site.xml
configuration.xsl                 httpfs-log4j.properties    shellprofile.d
container-executor.cfg            httpfs-site.xml            ssl-client.xml.example
core-site.xml                     kms-acls.xml               ssl-server.xml.example
hadoop-env.cmd                    kms-env.sh                 user_ec_policies.xml.template
hadoop-env.sh                     kms-log4j.properties       workers
hadoop-metrics2.properties        kms-site.xml               yarn-env.cmd
hadoop-policy.xml                 log4j.properties           yarn-env.sh
hadoop-user-functions.sh.example  mapred-env.cmd             yarnservice-log4j.properties
hdfs-rbf-site.xml                 mapred-env.sh              yarn-site.xml
hdfs-site.xml                     mapred-queues.xml.template
hadoop@vbox:~/hadoop/etc/hadoop$
```

Step 5a: Configure Hadoop Environment Variables (bashrc)

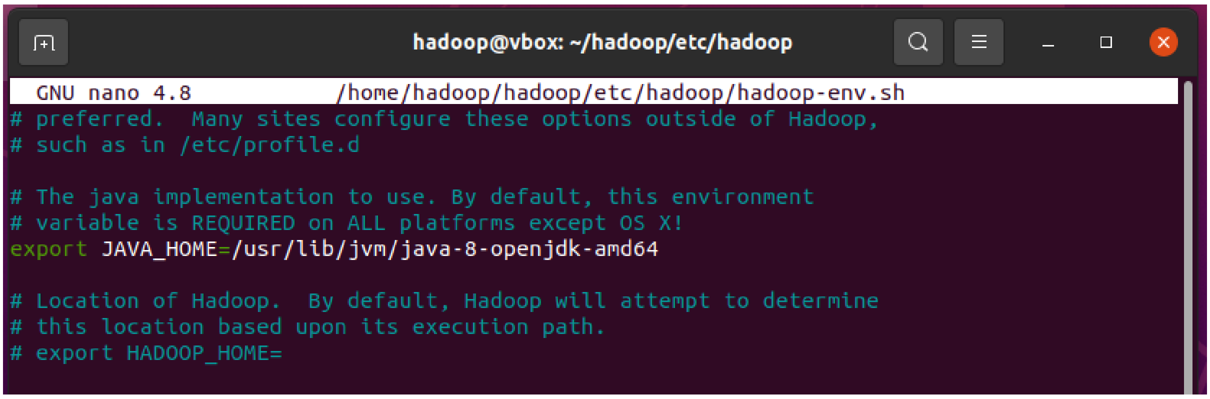


```
spa834@vbox:~$ sudo su - hadoop
[sudo] password for spa834:
hadoop@vbox:~$ dirname $(dirname $(readlink -f $(which java)))
/usr/lib/jvm/java-8-openjdk-amd64/jre
hadoop@vbox:~$ sudo nano ~/.bashrc
[sudo] password for hadoop:
hadoop is not in the sudoers file.  This incident will be reported.
hadoop@vbox:~$ nano ~/.bashrc
hadoop@vbox:~$ nano ~/.bashrc
hadoop@vbox:~$ source ~/.bashrc
hadoop@vbox:~$
```

Step 5b: Edit hadoop-env.sh file

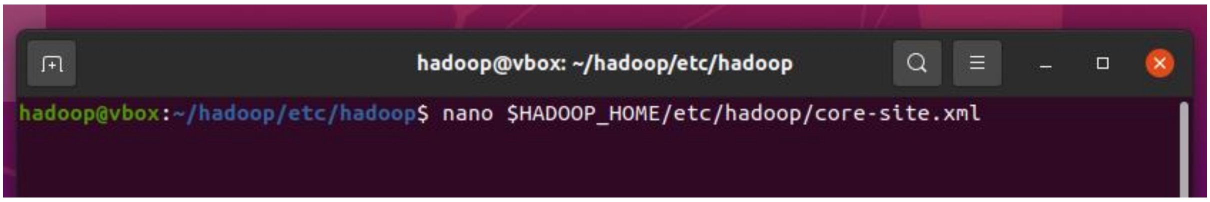


```
hadoop@vbox:~/hadoop/etc/hadoop$ nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
hadoop@vbox:~/hadoop/etc/hadoop$
```

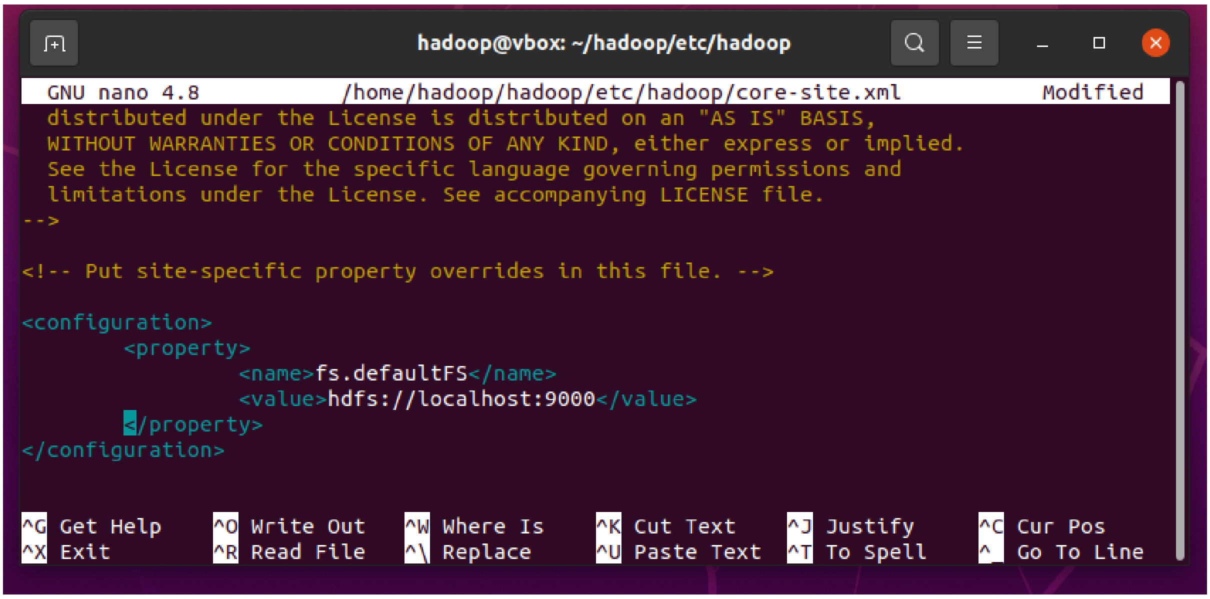


```
  GNU nano 4.8              /home/hadoop/hadoop/etc/hadoop/hadoop-env.sh
# preferred.  Many sites configure these options outside of Hadoop,
# such as in /etc/profile.d

# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# Location of Hadoop.  By default, Hadoop will attempt to determine
# this location based upon its execution path.
# export HADOOP_HOME=
```

Step 5c: Edit core-site.xml file

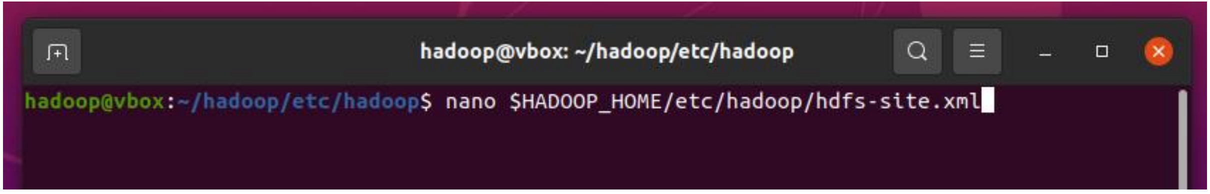


```
hadoop@vbox:~/hadoop/etc/hadoop$ nano $HADOOP_HOME/etc/hadoop/core-site.xml
```

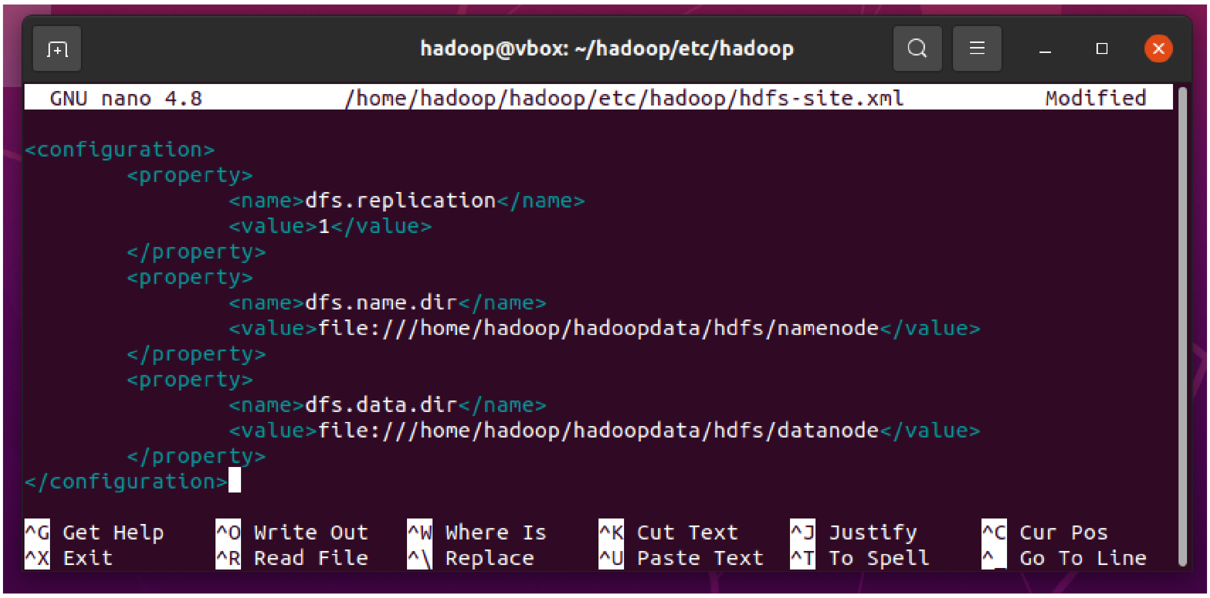


```
  GNU nano 4.8             /home/hadoop/hadoop/etc/hadoop/core-site.xml            Modified
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>fs.defaultFS</name>
                <value>hdfs://localhost:9000</value>
        </property>
</configuration>

^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line
```

Step 5d: Edit hdfs-site.xml file

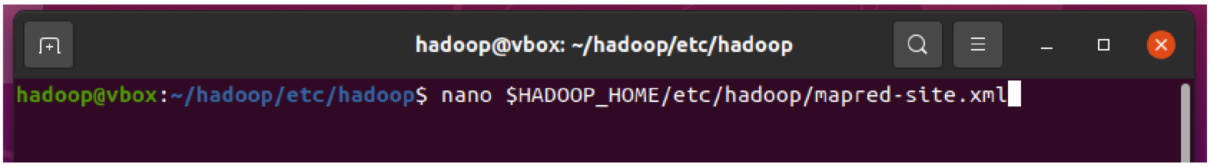


```
hadoop@vbox:~/hadoop/etc/hadoop$ nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

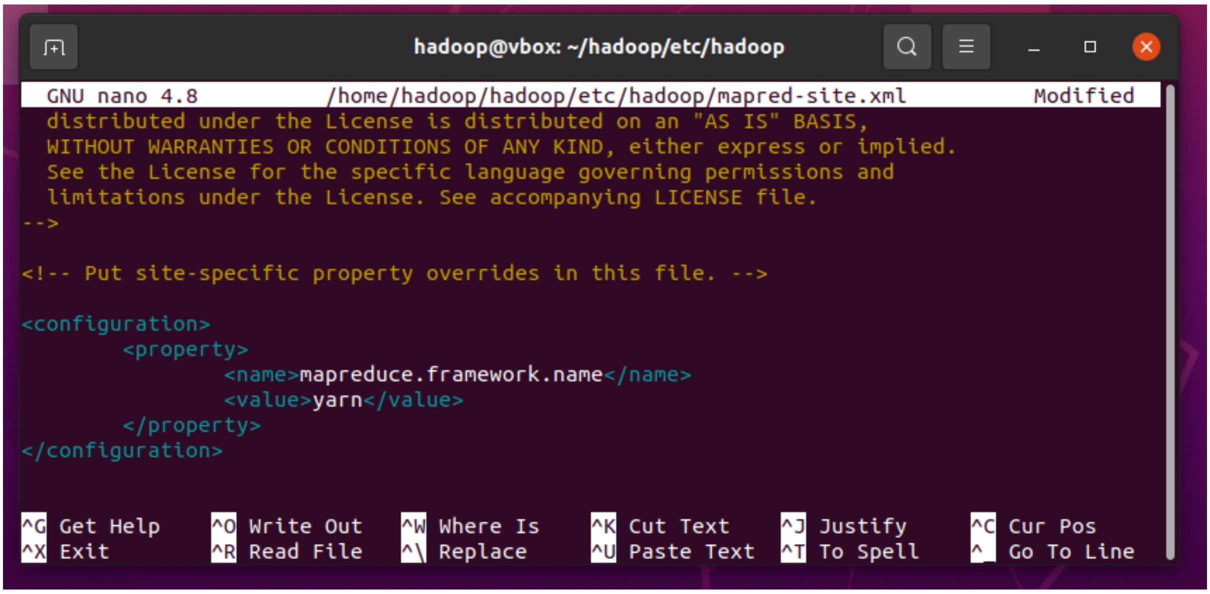


```
  GNU nano 4.8          /home/hadoop/hadoop/etc/hadoop/hdfs-site.xml          Modified

<configuration>
        <property>
                <name>dfs.replication</name>
                <value>1</value>
        </property>
        <property>
                <name>dfs.name.dir</name>
                <value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>
        </property>
        <property>
                <name>dfs.data.dir</name>
                <value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>
        </property>
</configuration>

^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line
```

Step 5e: Edit mapred-site.xml file

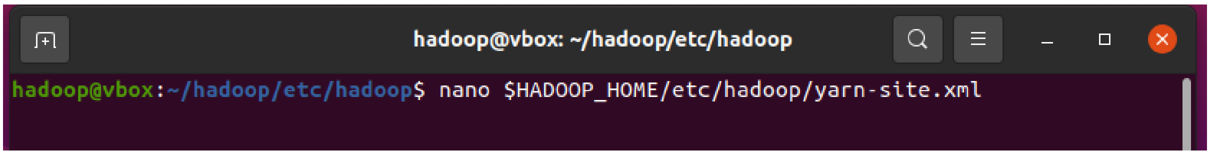


```
hadoop@vbox:~/hadoop/etc/hadoop$ nano $HADOOP_HOME/etc/hadoop/mapred-site.xml
```

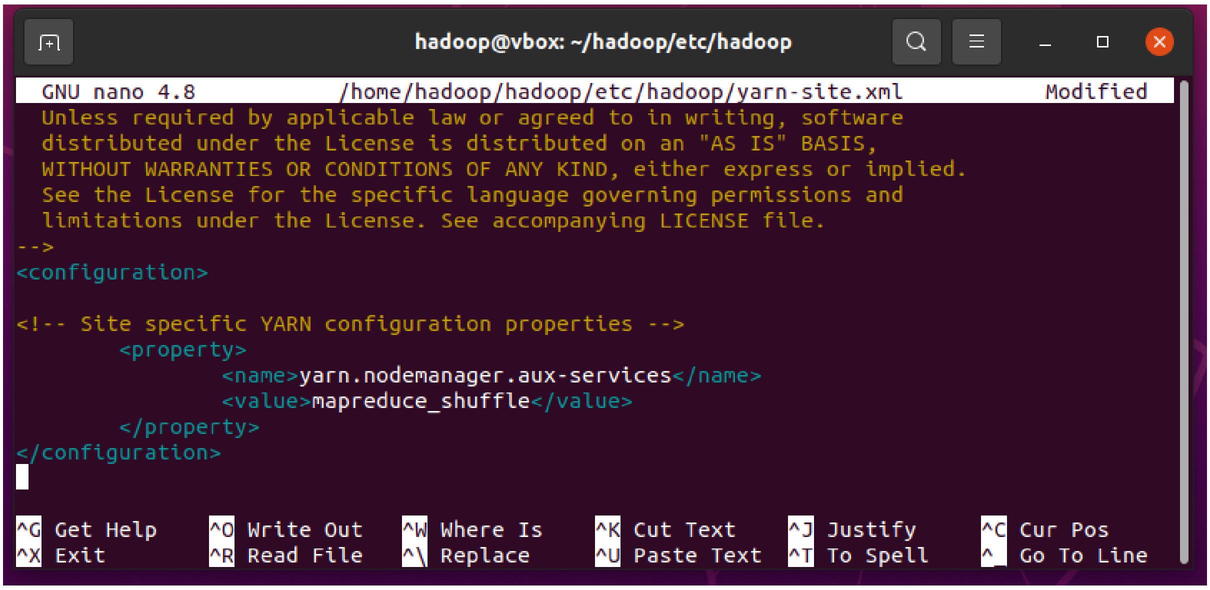


```
  GNU nano 4.8         /home/hadoop/hadoop/etc/hadoop/mapred-site.xml         Modified
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
        <property>
                <name>mapreduce.framework.name</name>
                <value>yarn</value>
        </property>
</configuration>

^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line
```

Step 5f: Edit yarn-site.xml file

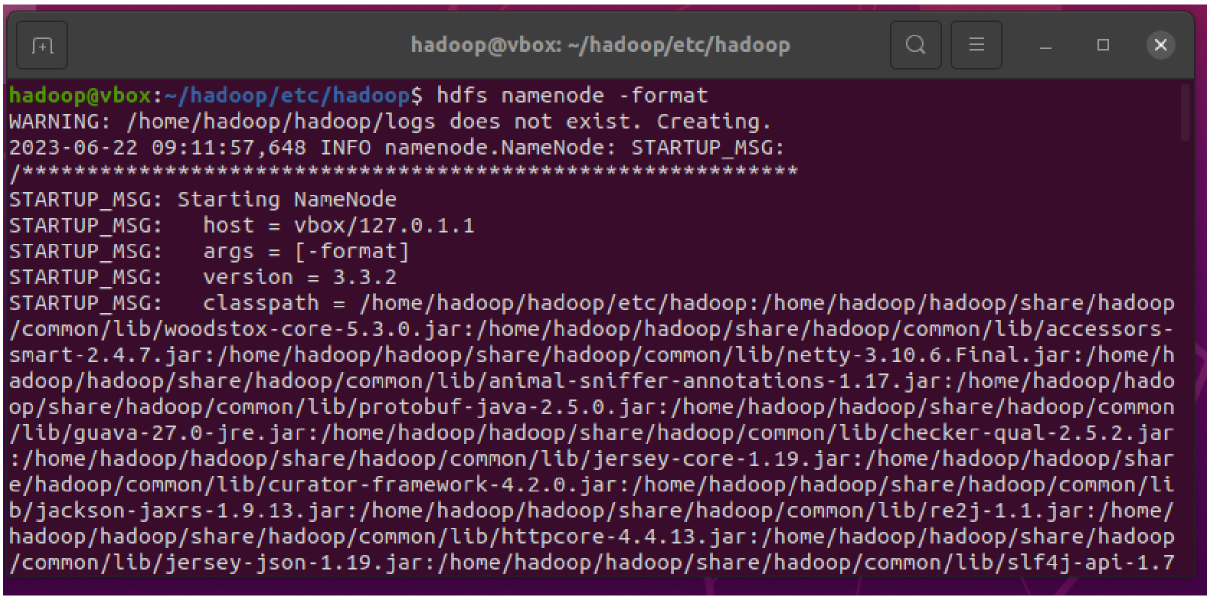


```
hadoop@vbox:~/hadoop/etc/hadoop$ nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
```

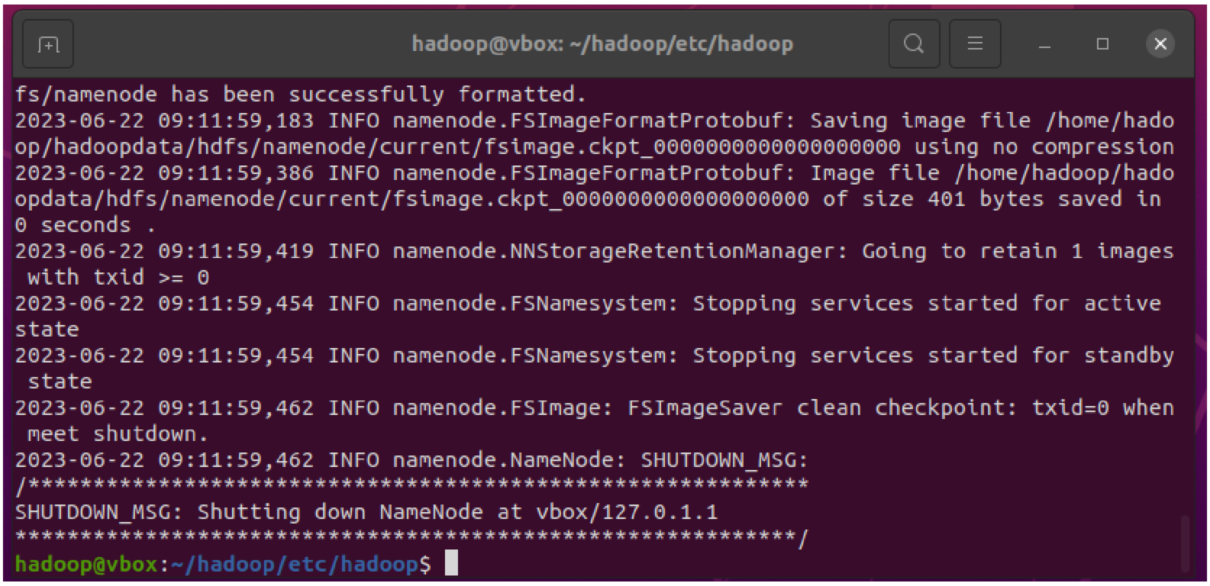


```
  GNU nano 4.8             /home/hadoop/hadoop/etc/hadoop/yarn-site.xml           Modified
  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
        <property>
                <name>yarn.nodemanager.aux-services</name>
                <value>mapreduce_shuffle</value>
        </property>
</configuration>


^G Get Help    ^O Write Out   ^W Where Is    ^K Cut Text    ^J Justify     ^C Cur Pos
^X Exit        ^R Read File   ^\ Replace     ^U Paste Text  ^T To Spell    ^_ Go To Line
```

Step 5g: Format HDFS NameNode

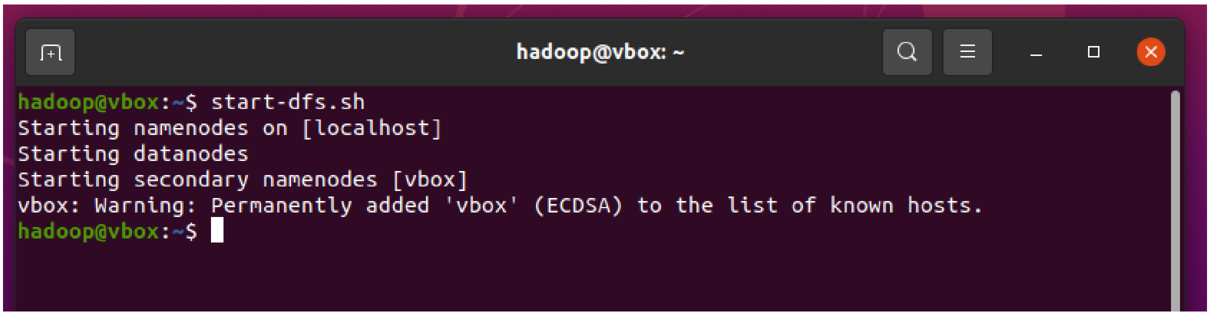


```
hadoop@vbox:~/hadoop/etc/hadoop$ hdfs namenode -format
WARNING: /home/hadoop/hadoop/logs does not exist. Creating.
2023-06-22 09:11:57,648 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = vbox/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.2
STARTUP_MSG:   classpath = /home/hadoop/hadoop/etc/hadoop:/home/hadoop/hadoop/share/hadoop
/common/lib/woodstox-core-5.3.0.jar:/home/hadoop/hadoop/share/hadoop/common/lib/accessors-
smart-2.4.7.jar:/home/hadoop/hadoop/share/hadoop/common/lib/netty-3.10.6.Final.jar:/home/h
adoop/hadoop/share/hadoop/common/lib/animal-sniffer-annotations-1.17.jar:/home/hadoop/hado
op/share/hadoop/common/lib/protobuf-java-2.5.0.jar:/home/hadoop/hadoop/share/hadoop/common
/lib/guava-27.0-jre.jar:/home/hadoop/hadoop/share/hadoop/common/lib/checker-qual-2.5.2.jar
:/home/hadoop/hadoop/share/hadoop/common/lib/jersey-core-1.19.jar:/home/hadoop/hadoop/shar
e/hadoop/common/lib/curator-framework-4.2.0.jar:/home/hadoop/hadoop/share/hadoop/common/li
b/jackson-jaxrs-1.9.13.jar:/home/hadoop/hadoop/share/hadoop/common/lib/re2j-1.1.jar:/home/
hadoop/hadoop/share/hadoop/common/lib/httpcore-4.4.13.jar:/home/hadoop/hadoop/share/hadoop
/common/lib/jersey-json-1.19.jar:/home/hadoop/hadoop/share/hadoop/common/lib/slf4j-api-1.7
```

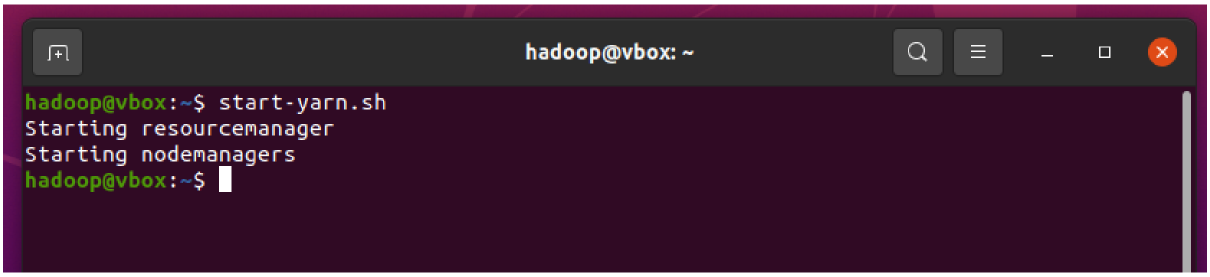


```
fs/namenode has been successfully formatted.
2023-06-22 09:11:59,183 INFO namenode.FSImageFormatProtobuf: Saving image file /home/hado
op/hadoopdata/hdfs/namenode/current/fsimage.ckpt_0000000000000000000 using no compression
2023-06-22 09:11:59,386 INFO namenode.FSImageFormatProtobuf: Image file /home/hadoop/hado
opdata/hdfs/namenode/current/fsimage.ckpt_0000000000000000000 of size 401 bytes saved in
0 seconds .
2023-06-22 09:11:59,419 INFO namenode.NNStorageRetentionManager: Going to retain 1 images
 with txid >= 0
2023-06-22 09:11:59,454 INFO namenode.FSNamesystem: Stopping services started for active
state
2023-06-22 09:11:59,454 INFO namenode.FSNamesystem: Stopping services started for standby
 state
2023-06-22 09:11:59,462 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when
 meet shutdown.
2023-06-22 09:11:59,462 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at vbox/127.0.1.1
************************************************************/
hadoop@vbox:~/hadoop/etc/hadoop$
```

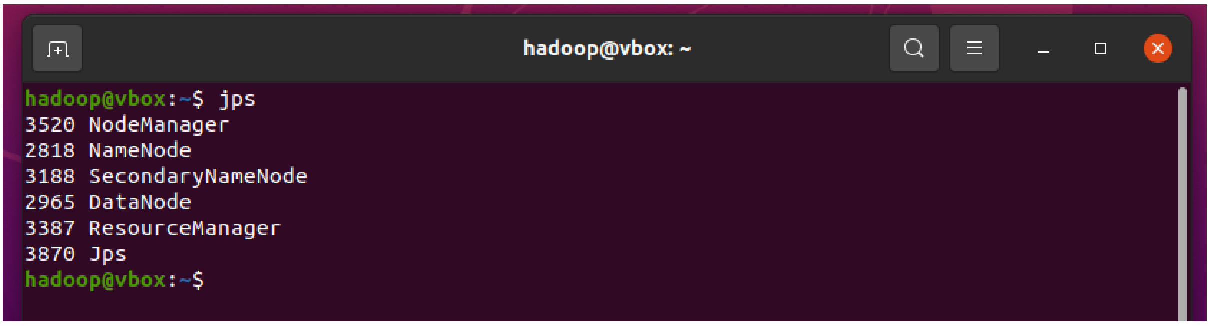
Step 6: Start Hadoop Cluster



```
hadoop@vbox:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [vbox]
vbox: Warning: Permanently added 'vbox' (ECDSA) to the list of known hosts.
hadoop@vbox:~$
```

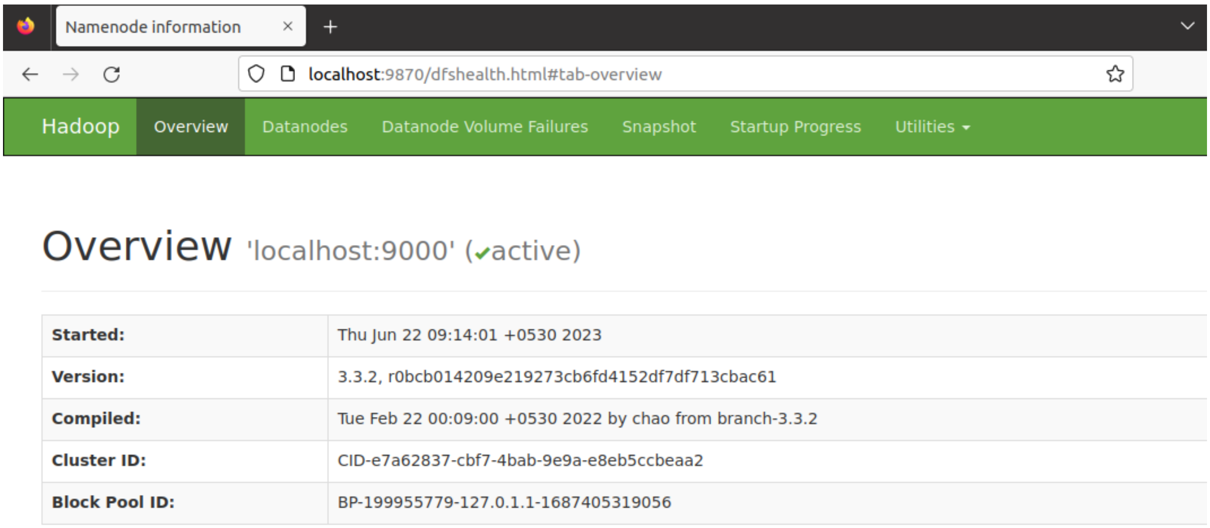


```
hadoop@vbox:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@vbox:~$
```



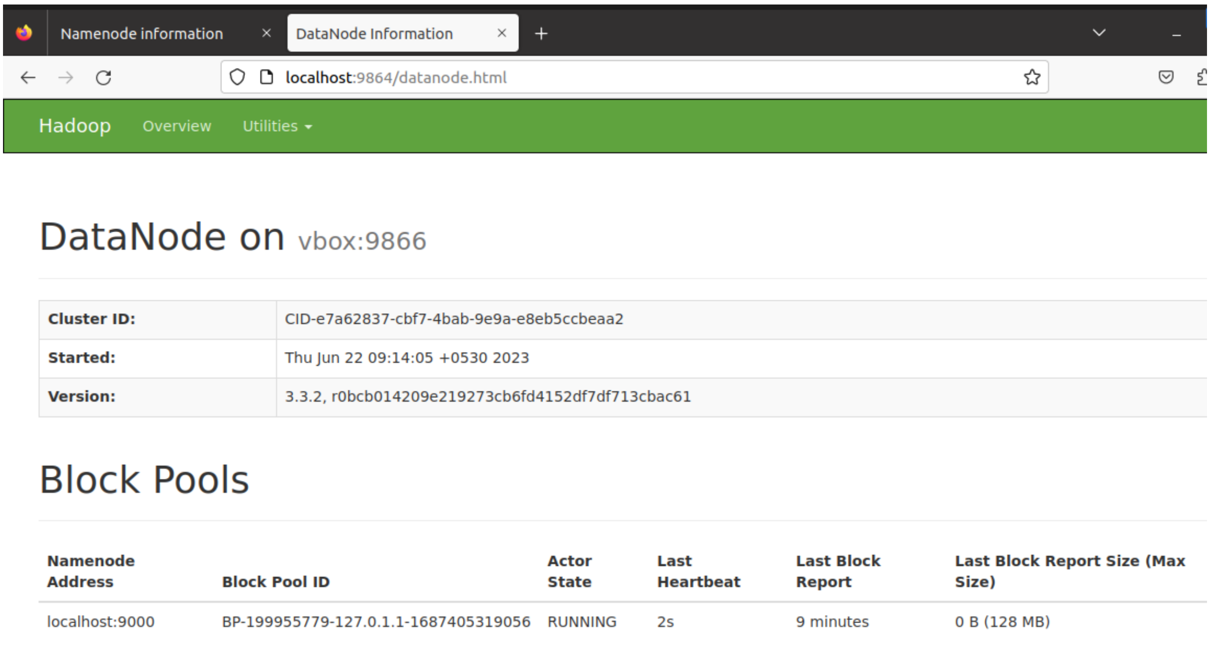
```
hadoop@vbox:~$ jps
3520 NodeManager
2818 NameNode
3188 SecondaryNameNode
2965 DataNode
3387 ResourceManager
3870 Jps
hadoop@vbox:~$
```

Step 7: Access Hadopp UI from Browser



Overview 'localhost:9000' (✓active)

Started:	Thu Jun 22 09:14:01 +0530 2023
Version:	3.3.2, r0bcb014209e219273cb6fd4152df7df713cbac61
Compiled:	Tue Feb 22 00:09:00 +0530 2022 by chao from branch-3.3.2
Cluster ID:	CID-e7a62837-cbf7-4bab-9e9a-e8eb5ccbeaa2
Block Pool ID:	BP-199955779-127.0.1.1-1687405319056

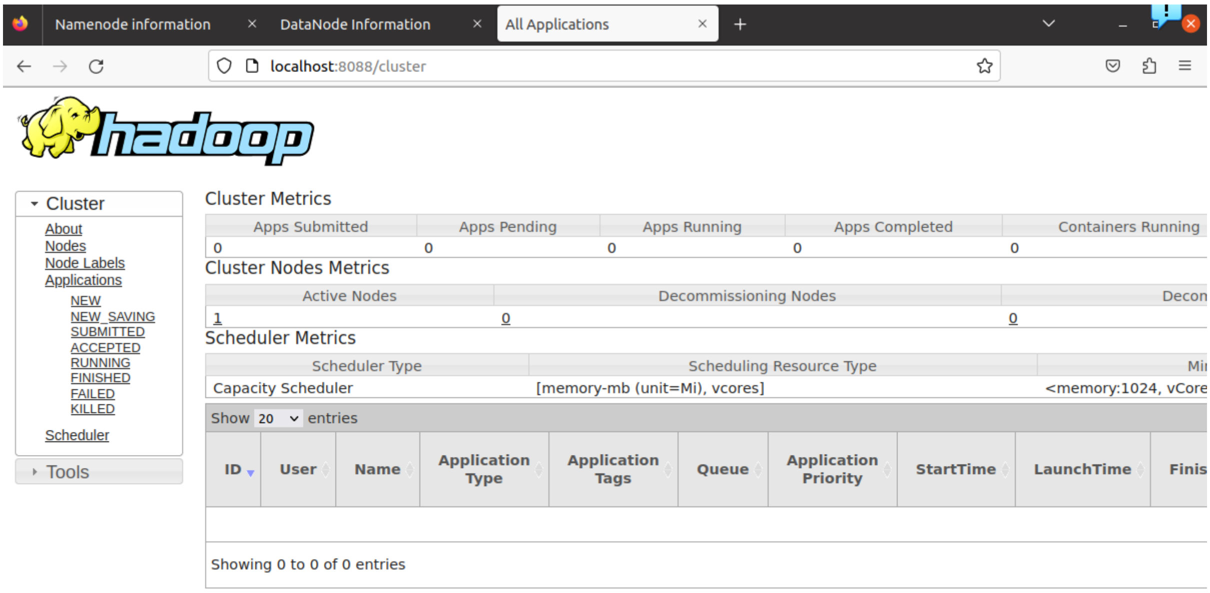


DataNode on vbox:9866

Cluster ID:	CID-e7a62837-cbf7-4bab-9e9a-e8eb5ccbeaa2
Started:	Thu Jun 22 09:14:05 +0530 2023
Version:	3.3.2, r0bcb014209e219273cb6fd4152df7df713cbac61

Block Pools

Namenode Address	Block Pool ID	Actor State	Last Heartbeat	Last Block Report	Last Block Report Size (Max Size)
localhost:9000	BP-199955779-127.0.1.1-1687405319056	RUNNING	2s	9 minutes	0 B (128 MB)



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
0	0	0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decon
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Mi
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCore

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	Finis

Showing 0 to 0 of 0 entries