DAYANANDA SAGAR COLLEGE OF ENGINEERING

(An Autonomous Institute Affiliated to VTU, Belagavi)

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560111

APPROVED BY AICTE, UGC
& NAAC WITH 'A' GRADE

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

APPLIED BIG DATA AND CLOUD COMPUTING

LABORATORY MANUAL

BE-VI SEMESTER

(2023-2024)

DAYANANDA SAGAR COLLEGE OF ENGINEERING
Accredited by National Assessment & Accreditation Council (NAAC) with 'A' Grade
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi
ISN 9001:2008, ISO 14001:2004 and ISO 22000:2005 Certified)
SHAVIGE MALLESHWARA HILLS, KUMARASWAMY LAYOUT
BENGALURU-560111

Vision of the Institution

To impart quality technical education with a focus on Research and Innovation emphasizing on Development of Sustainable and Inclusive Technology for the benefit of society.

Mission of the Institution

- To provide an environment that enhances creativity and Innovation in pursuit of Excellence.
- To nurture teamwork in order to transform individuals as responsible leaders and entrepreneurs.
- To train the students to the changing technical scenario and make them to understand the importance of Sustainable and Inclusive technologies.

DAYANANDA SAGAR COLLEGE OF ENGINEERING

Accredited by National Assessment & Accreditation Council (NAAC) with 'A' Grade
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi
ISN 9001:2008, ISO 14001:2004 and ISO 22000:2005 Certified)
SHAVIGE MALLESHWARA HILLS, KUMARASWAMY LAYOUT
BENGALURU-560111

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Vision of the Department

To provide progressive education and flourish the student's ingenuity to be successfulprofessionals impacting the society for a smarter and ethical world.

Mission of the Department

- To adopt an engaging teaching learning process with emphasis on problem solving and programming skills.
- To promote additional skill development through enhanced and experiential learning.
- To collaborate with industries and professional bodies and make the students industry ready.
- To encourage innovation through multi-disciplinary research and development activities.
- To imbibe human values and ethics in students to make them impact the society's as responsible professionals.

DAYANANDA SAGAR COLLEGE OF ENGINEERING

Accredited by National Assessment & Accreditation Council (NAAC) with 'A' Grade (An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
ISN 9001:2008, ISO 14001:2004 and ISO 22000:2005
Certified)SHAVIGE MALLESHWARA HILLS, KUMARASWAMY
LAYOUT BENGALURU-560111

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DOs and DON'Ts in Laboratory:

1. Make entry in the Log Book as soon as you enter the Laboratory.

2. All the students should sit according to their roll numbers starting from their Left to right.

3. All the students are supposed to enter the terminal number in the log book.

4. Do not change the terminal on which you are working.

5. All the students are expected to get at least the algorithm of the Program/concept to be implemented.

6. Strictly observe the instructions given by the teacher/Lab Instructor.

7. Do not disturb machine Hardware / Software Setup

Instruction for Laboratory Teachers:

1. Submission related to whatever lab work has been completed should be done during the next lab session along with signing the index.

2. The promptness of submission should be encouraged by way of marking and evaluation patterns that will benefit the sincere students.

3. Continuous assessment in the prescribed format must be followed.

DAYANANDA SAGAR COLLEGE OF ENGINEERING

Accredited by National Assessment & Accreditation Council (NAAC) with 'A' Grade
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi
ISN 9001:2008, ISO 14001:2004 and ISO 22000:2005 Certified)
SHAVIGE MALLESHWARA HILLS, KUMARASWAMY LAYOUT
BENGALURU-560111

Course objectives:

1. To understand the need of Big Data, challenges and different analytical architectures
2. Installation and understanding of Hadoop Architecture and its eco systems
3. Processing of Big Data with Advanced architectures like Spark.
4. Describe graphs and streaming data in Spark

Course Outcomes: At the end of the course, Student will be able to:

CO1	Demonstrate a solid understanding of the principles, characteristics, and challenges of big data and cloud computing.
CO2	Apply cloud computing concepts to effectively deploy, manage, and scale big data infrastructure and resources.
CO3	Examine practical scenarios and real-life examples where the integration of Big Data analytics and cloud computing technologies has been applied to address various challenges.
CO4	Develop analytical skills to systematically analyze and select optimal tools based on problem analysis.
CO5	Implementing, and deploying a big data analytics solution on a cloud computing platform
CO6	Create simple software programs and web applications, as well as effectively utilize programming tools and environments

Sl. No.	Topics	Course Outcome
1	Install Virtual box/VMware Workstation with different flavors of Linux or windows OS on topofwindows7or8.	**CO2**
2	Install a C compiler in the virtual machine created using virtual box • Write a C program for storing data in a simulated cloud storage environment using a local file	**CO1,CO6**
3	Write a C-Program for CPU usage Monitoring and Logging on Cloud-Based-Ubuntu Servers.	**CO6**
4	Use Google App Engine Launcher to launch the web applications	**CO4,CO6**
5	Installation of Single Node Hadoop Cluster on Ubuntu 22.04 LTS.	**CO4,CO5**
6	Hadoop Programming: Word Count Map Reduce Program Using Eclipse	**CO6**
7	File Management tasks in Hadoop using HDFS commands a. Adding files to HDFS b. Retrieving files from HDFS c. Deleting files from HDFS	**CO5,CO6**
8	Creation of Data Frames using Create Data Frame working with an example related to FIFA dataset.	**CO5**
9	Simulator program: Analyze the average, minimum and maximum response time and datacenter Processing time and write the findings.	**CO4**
10	Conduct experimentation by changing different service broker policies to analyze the average, minimum and maximum response time and total datacenter Processing time. Also, analyze the total cost required to run the application.	**CO4**

1. Install Virtual box/VMware Workstation with different flavors of Linux or windows OS on topofwindows7or8.

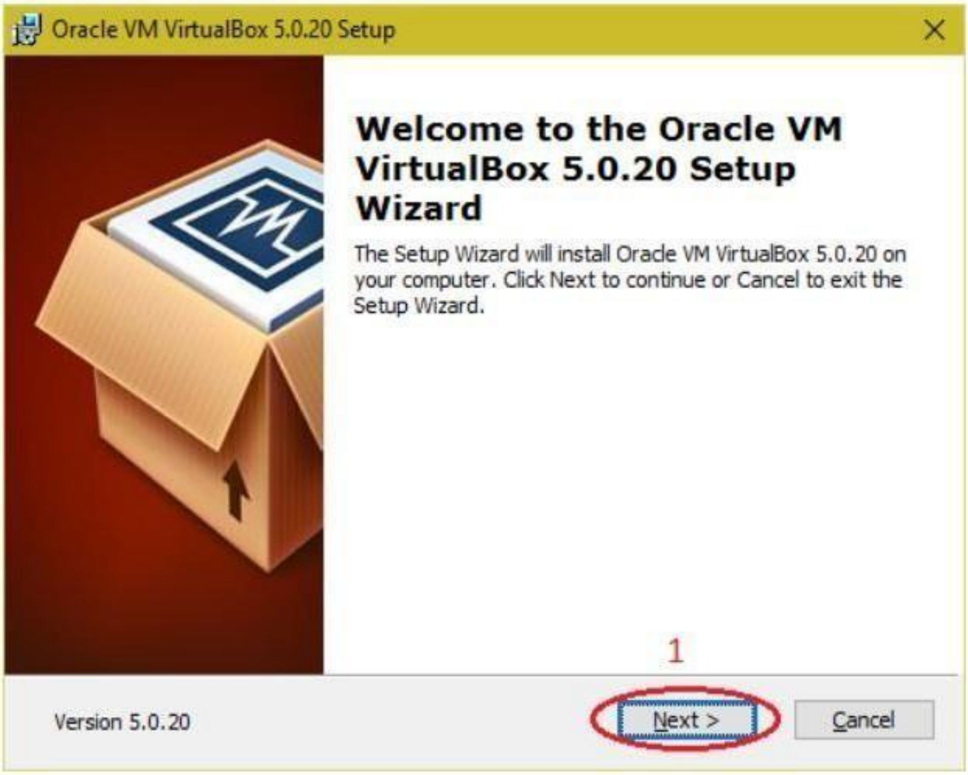
Aim: Find procedure to Install Virtual box/VMware Workstation with different flavors of Linux or windows OS on top of windows7or8.

This experiment is to be performed through portal.

PROCEDURE TO INSTALL

1. Download and Install VirtualBox. Using the link-
2. Download and Install Ubuntu. Using the link-

Run the virtual box setup and click on "Next" Button.

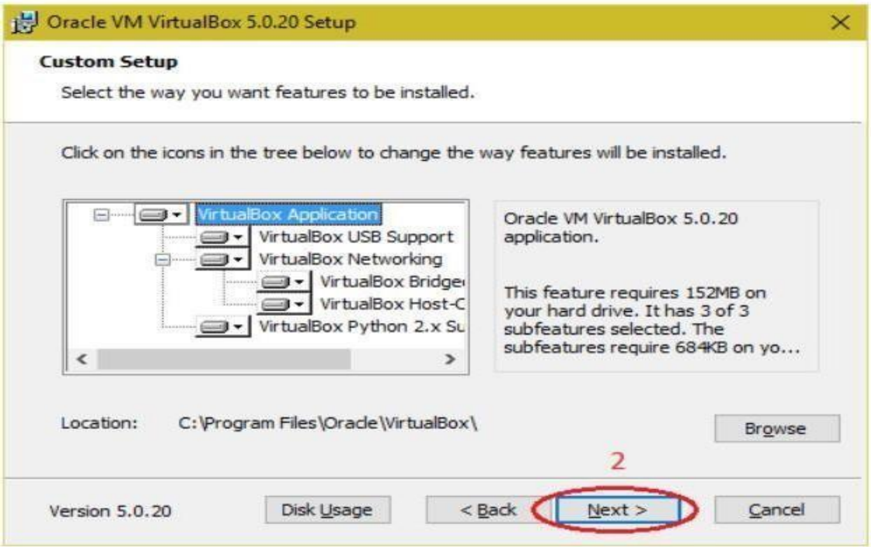


Oracle VM VirtualBox 5.0.20 Setup

Welcome to the Oracle VM VirtualBox 5.0.20 Setup Wizard

The Setup Wizard will install Oracle VM VirtualBox 5.0.20 on your computer. Click Next to continue or Cancel to exit the Setup Wizard.

Click on “Next” button.



Oracle VM VirtualBox 5.0.20 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 Bridge
    - VirtualBox Host-C
  - VirtualBox Python 2.x Su

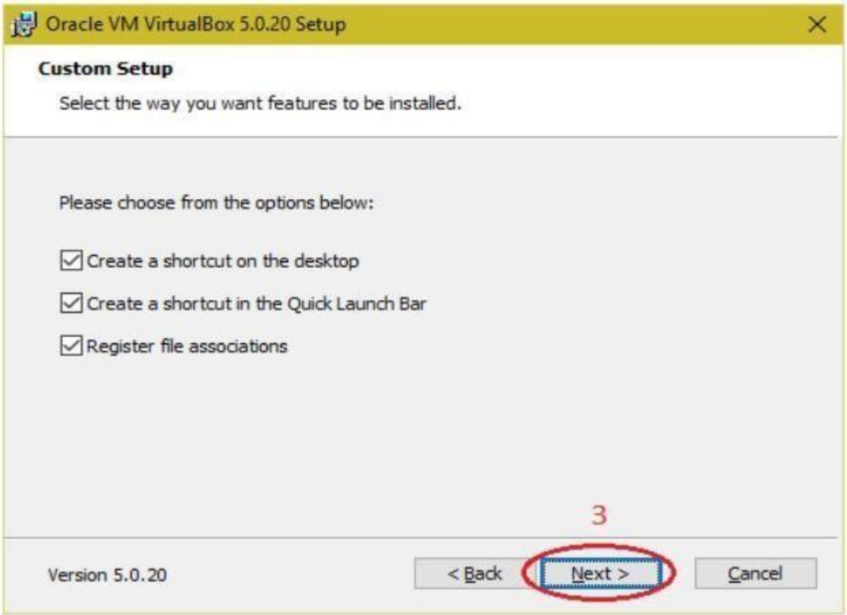
Oracle VM VirtualBox 5.0.20 application.

This feature requires 152MB on your hard drive. It has 3 of 3 subfeatures selected. The subfeatures require 684KB on yo...

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

Version 5.0.20

Click on “Next” button.



Oracle VM VirtualBox 5.0.20 Setup

Custom Setup

Select the way you want features to be installed.

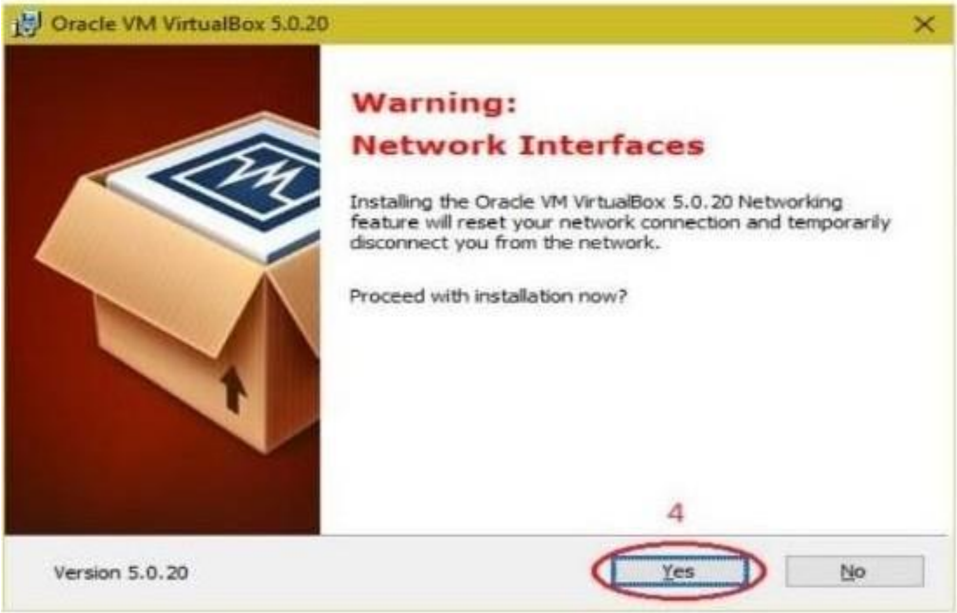
Please choose from the options below:

☑ Create a shortcut on the desktop

☑ Create a shortcut in the Quick Launch Bar

☑ Register file associations

Version 5.0.20



Oracle VM VirtualBox 5.0.20

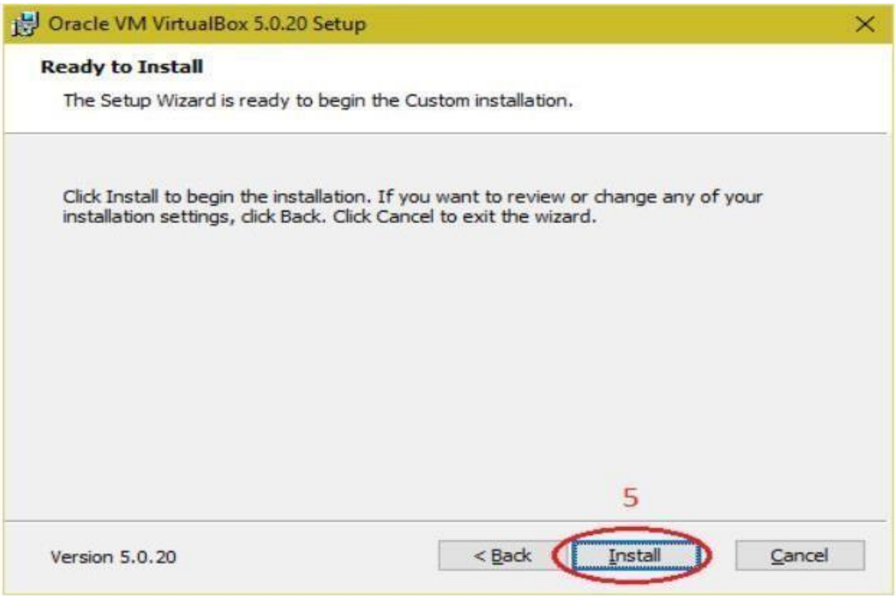
Warning:
Network Interfaces

Installing the Oracle VM VirtualBox 5.0.20 Networking feature will reset your network connection and temporarily disconnect you from the network.

Proceed with installation now?

Version 5.0.20

Click on “Yes” button



Oracle VM VirtualBox 5.0.20 Setup

Ready to Install

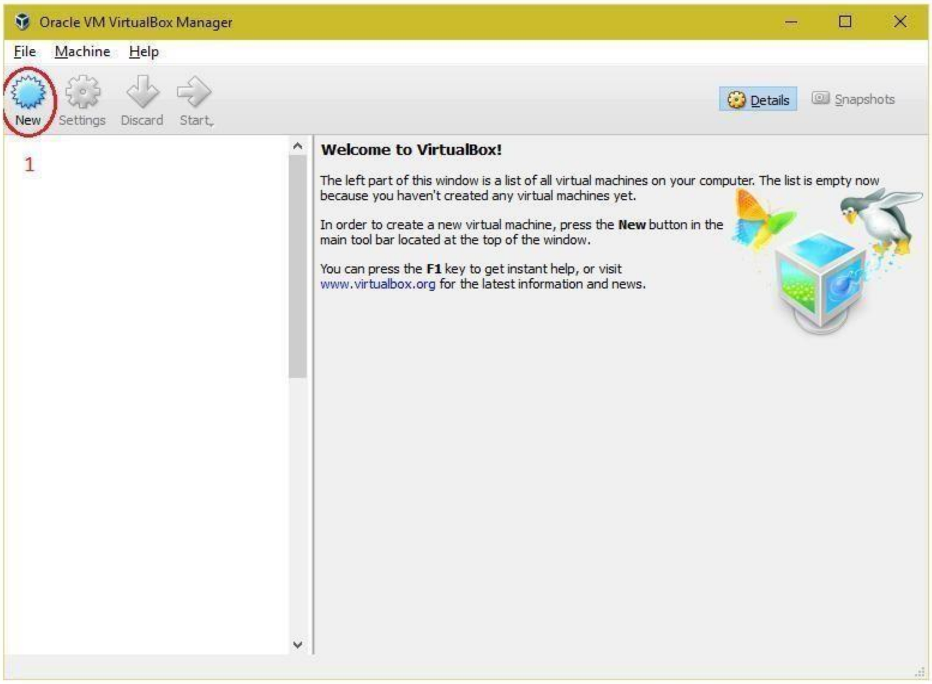
The Setup Wizard is ready to begin the Custom installation.

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.

Version 5.0.20

Installing “Ubuntu”as Virtual machine in “Oracle VM VirtualBox”.

1. Open “Oracle VM VirtualBoxManager”.



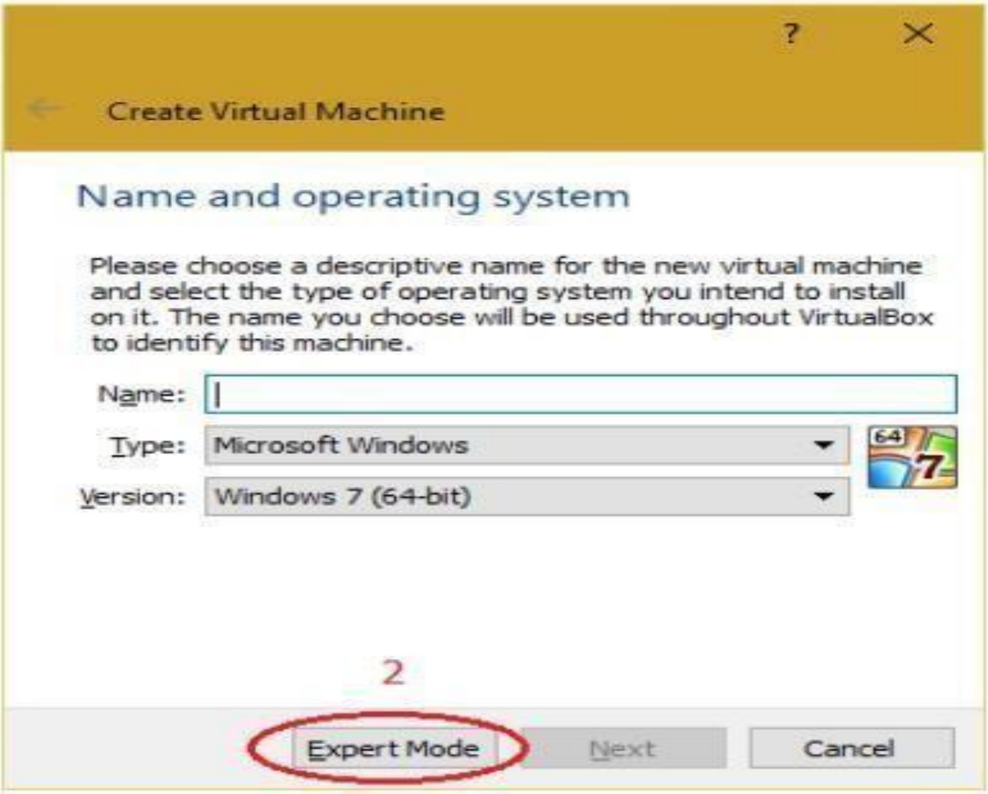
Welcome to VirtualBox!

The left part of this window is a list of all virtual machines on your computer. The list is empty now because you haven't created any virtual machines yet.

In order to create a new virtual machine, press the **New** button in the main tool bar located at the top of the window.

You can press the **F1** key to get instant help, or visit www.virtualbox.org for the latest information and news.

1. Click on “New “button and select “Expert Mode”.



Name and operating system

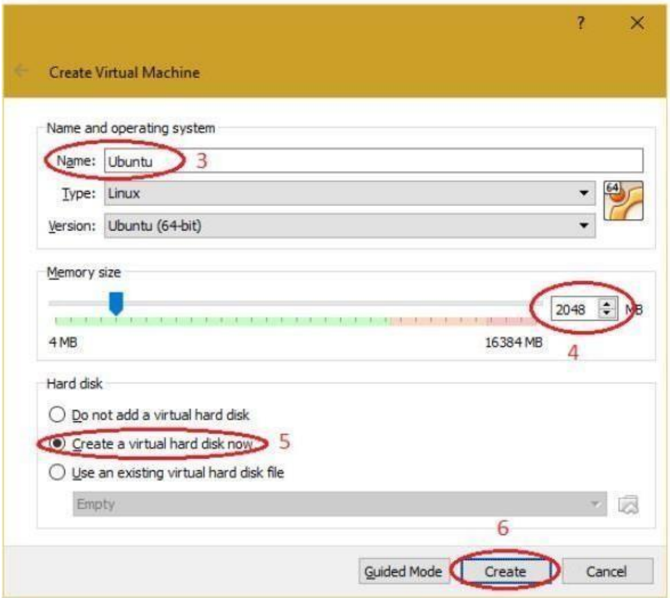
Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.

Name:

Type: Microsoft Windows

Version: Windows 7 (64-bit)

2. Provide the name and operating system information for virtual machine.



Name and operating system

Name: Ubuntu

Type: Linux

Version: Ubuntu (64-bit)

Memory size

4 MB 16384 MB

2048 MB

Hard disk

○ Do not add a virtual hard disk

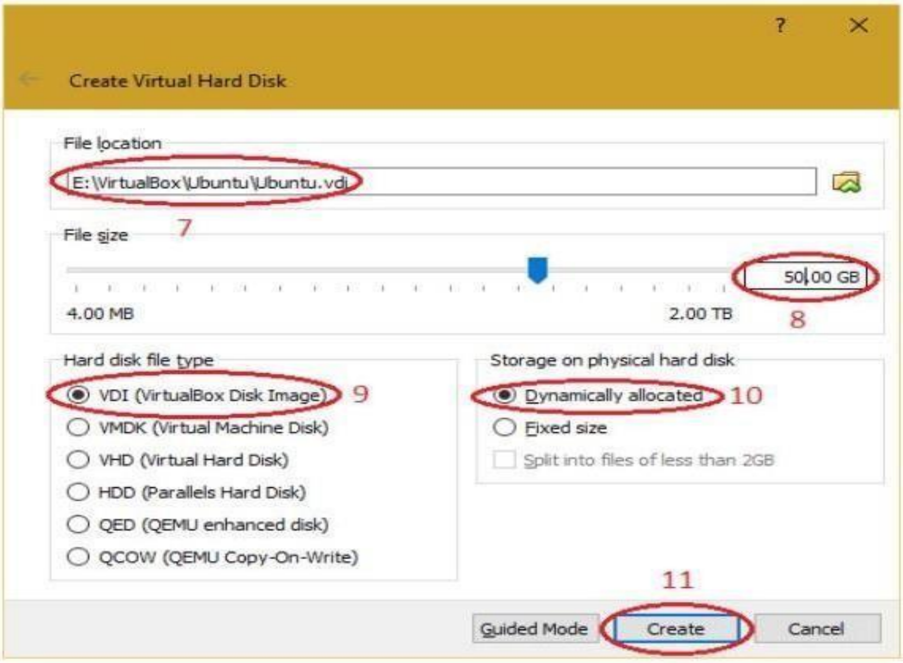
● Create a virtual hard disk now

○ Use an existing virtual hard disk file

Empty

Note: Before installing 64-bit operating system, Intel VT-x/AMD-V must be enabled in "BIOS" on the system. To enable Intel VT-x/AMD-V, open BIOS and search for "Intel Virtualization Technology" or "AMD-V", save the BIOS and boot the PC/Laptop.

3. Select the path for the virtual hard disk and Click on “create “button.



File location

E:\VirtualBox\Ubuntu\Ubuntu.vdi

File size

4.00 MB 2.00 TB

50,00 GB

Hard disk file type

● VDI (VirtualBox Disk Image)

○ VMDK (Virtual Machine Disk)

○ VHD (Virtual Hard Disk)

○ HDD (Parallels Hard Disk)

○ QED (QEMU enhanced disk)

○ QCOW (QEMU Copy-On-Write)

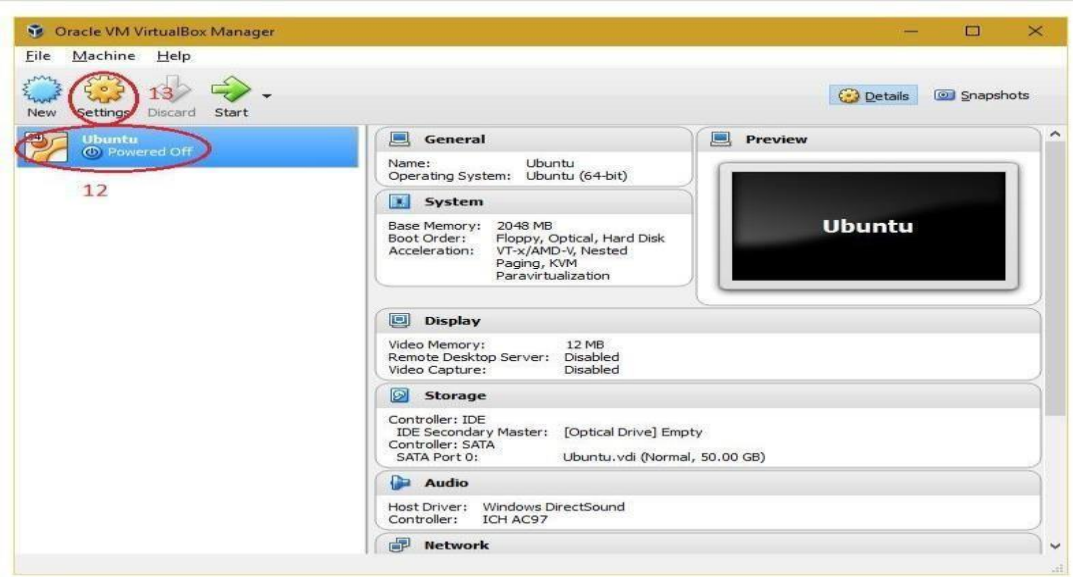
Storage on physical hard disk

● Dynamically allocated

○ Fixed size

□ Split into files of less than 2GB

4. **Select the virtual machine from the Virtual box manager and click on“settings “button.**



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

System
Base Memory: 2048 MB
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization

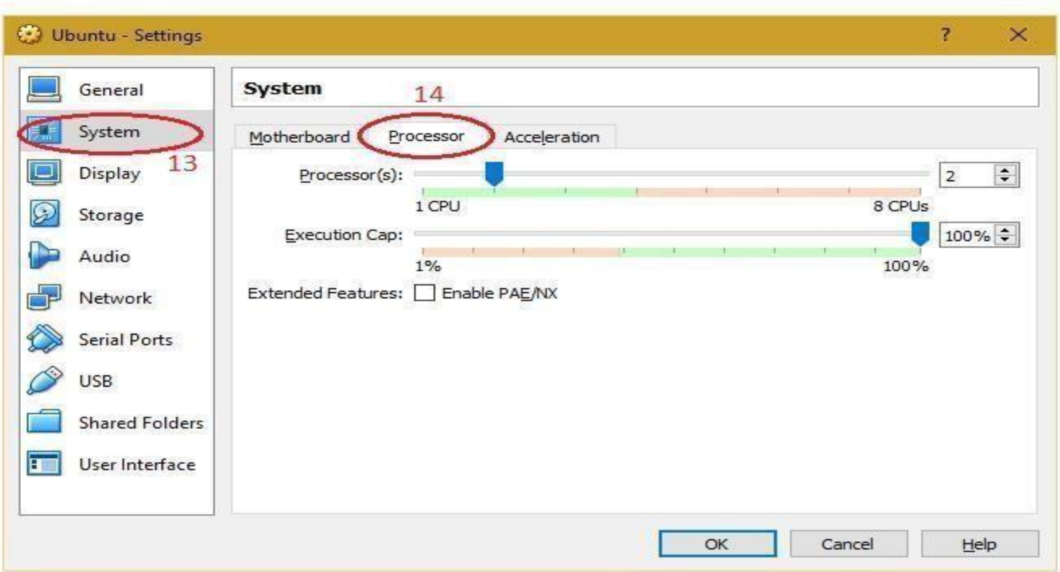
Display
Video Memory: 12 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

Storage
Controller: IDE
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: Ubuntu.vdi (Normal, 50.00 GB)

Audio
Host Driver: Windows DirectSound
Controller: ICH AC97

Network

5. **Select “System” and navigate to “processor” tab to adjust number of processor virtual machine for better performance.**



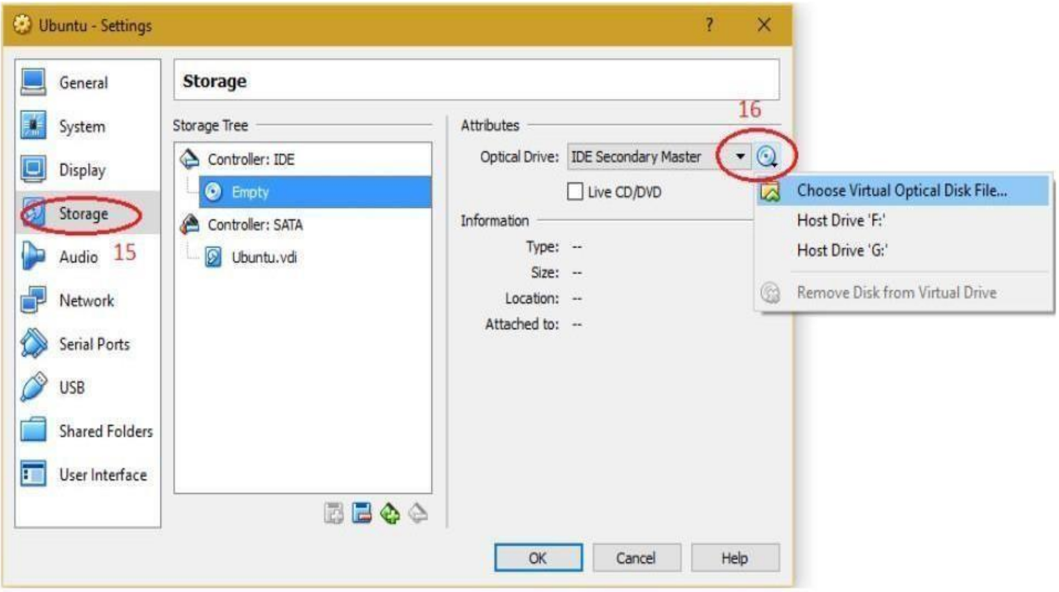
System

Processor(s): 2

Execution Cap: 100%

Extended Features: □ Enable PAE/NX

6. **Select “storage” and choose the installation media of Operation System (ISO/CD/DVD). Preferred Linux “.iso” can be download from CC ftp site. Also many different flavours of Linux are available on the internet- Fedora ,CentOS , Ubuntu, Debian, Mageia, openSUSE, Arch Linux, Slackware Linux, etc.**



Storage Tree

- Controller: IDE
  - Empty
- Controller: SATA
  - Ubuntu.vdi

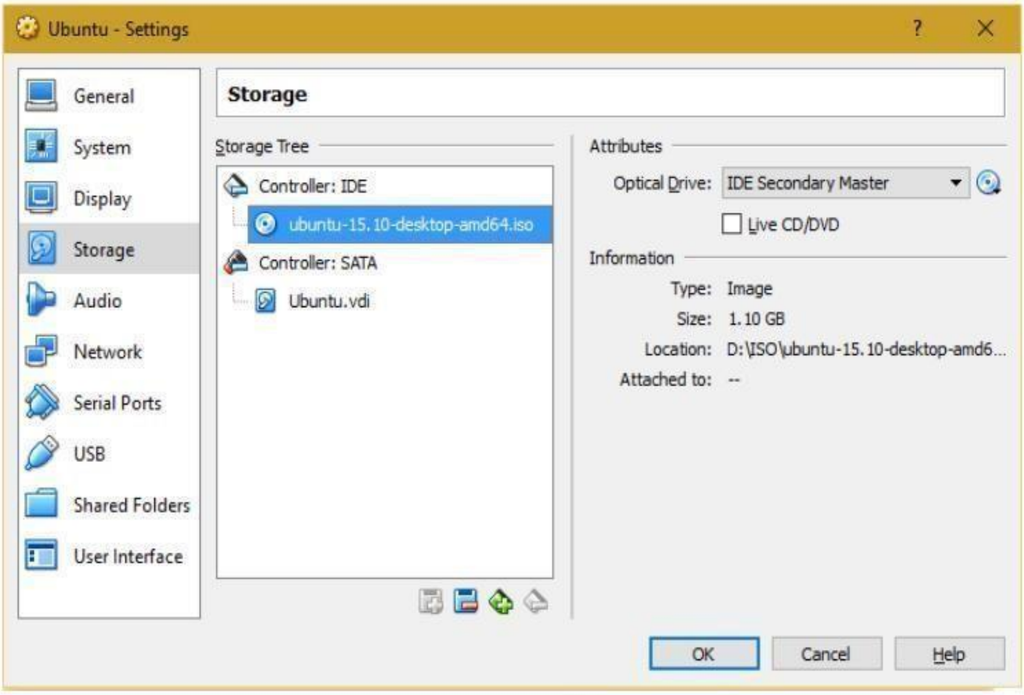
Attributes

Optical Drive: IDE Secondary Master

□ Live CD/DVD

Information

Type: --
Size: --
Location: --
Attached to: --



Storage Tree

- Controller: IDE
  - ubuntu-15.10-desktop-amd64.iso
- Controller: SATA
  - Ubuntu.vdi

Attributes

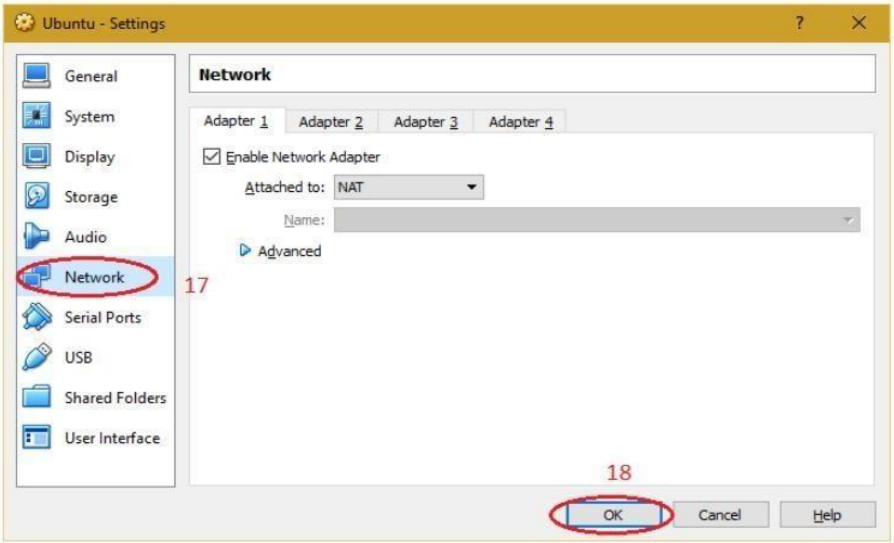
Optical Drive: IDE Secondary Master

□ Live CD/DVD

Information

Type: Image
Size: 1.10 GB
Location: D:\ISO\ubuntu-15.10-desktop-amd6...
Attached to: --

7. **Select “Network” to make changes required for network setting of virtual machine and click on “OK”.**

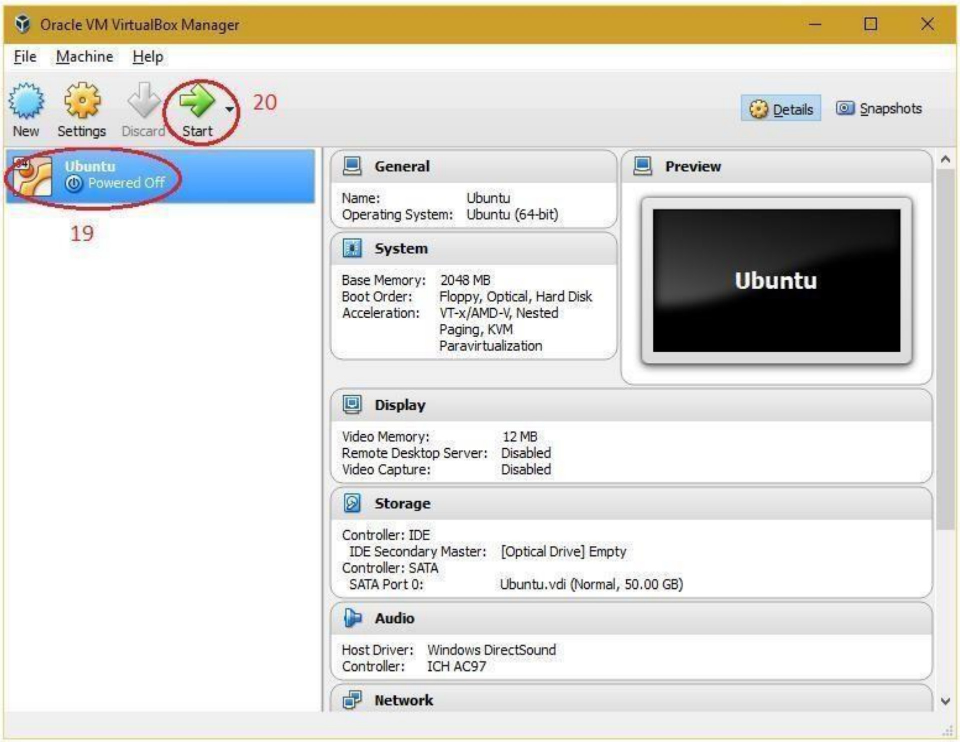


☑ Enable Network Adapter

Attached to: NAT

Name:

8. **Select the created Virtual machine and click on “Start” button.**



General

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

System

Base Memory: 2048 MB
Boot Order: Floppy, Optical, Hard Disk
Acceleration: VT-x/AMD-V, Nested Paging, KVM Paravirtualization

Display

Video Memory: 12 MB
Remote Desktop Server: Disabled
Video Capture: Disabled

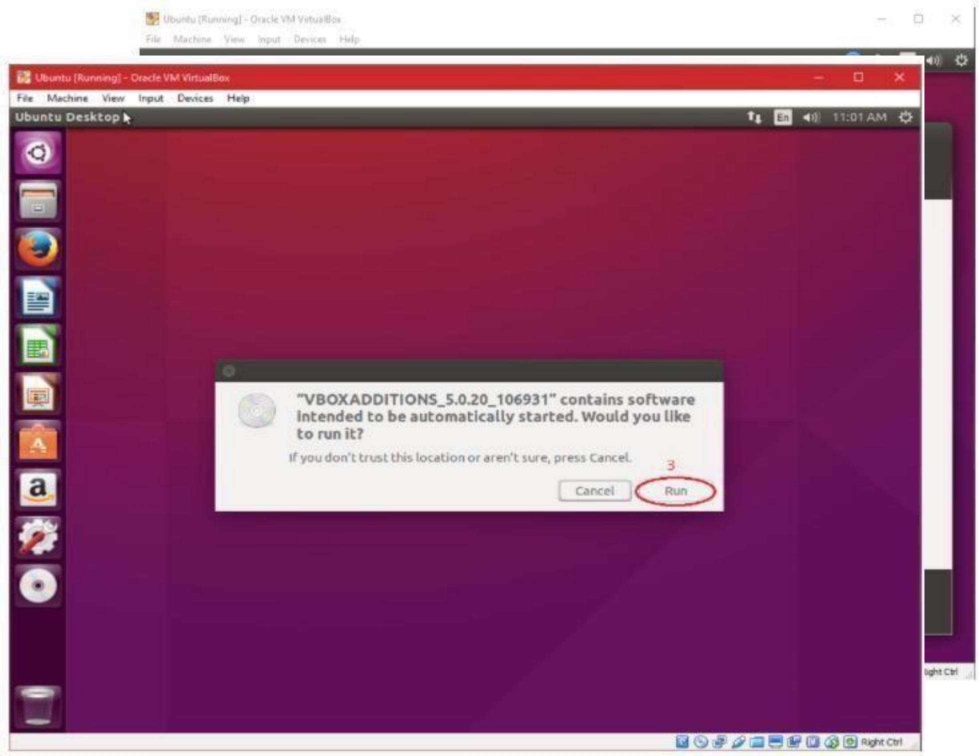
Storage

Controller: IDE
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: Ubuntu.vdi (Normal, 50.00 GB)

Audio

Host Driver: Windows DirectSound
Controller: ICH AC97

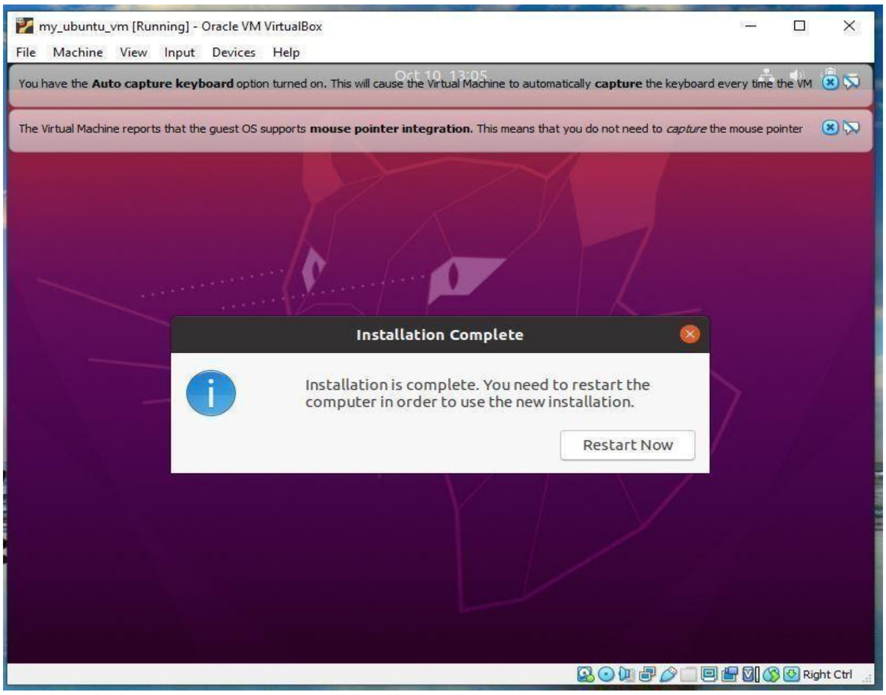
9. Proceed with the installation of operating system in virtualmachine.


Ubuntu [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Ubuntu Desktop
11:01 AM
Cancel
Run
3
Right Ctrl

“VBOXADDITIONS_5.0.20_106931” contains software intended to be automatically started. Would you like to run it?

If you don't trust this location or aren't sure, press Cancel.

“Run “to install on Ubuntu virtual machine. Once the installation, restart the virtual system.

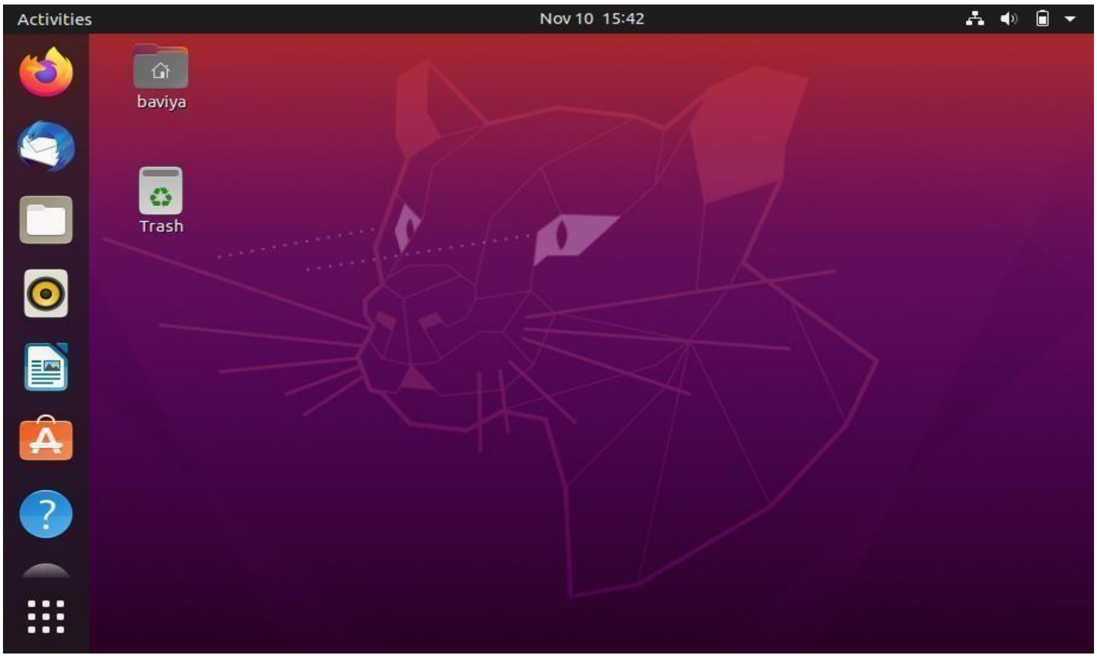

my_ubuntu_vm [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Oct 10 13:05
Installation Complete
Restart Now
Right Ctrl

You have the **Auto capture keyboard** option turned on. This will cause the Virtual Machine to automatically **capture** the keyboard every time the VM

The Virtual Machine reports that the guest OS supports **mouse pointer integration.** This means that you do not need to *capture* the mouse pointer

Installation is complete. You need to restart the computer in order to use the new installation.

Complete the installation process.



RESULT:

Thus, the Virtual Box was installed with Ubuntu OS over Windows Successfully.

2. Install a C compiler in the virtual machine created using virtual box

- **Write a C program for storing data in a simulated cloud storage environment using a local file**

```
#include <stdio.h>
#include <stdlib.h>
int main() {
    char data[1024];
    // Open a file in write mode. Simulating a cloud storage by using a local file.
    FILE *fptr = fopen("cloud_storage.txt", "w");
    if (fptr == NULL) {
        printf("Error opening the file!\n");
        exit(1);
    }

    // Get user input
    printf("Enter text to store in the cloud: ");
    fgets(data, sizeof(data), stdin);

    // Write data to the file
    fprintf(fptr, "%s", data);

    // Close the file
    fclose(fptr);

    printf("Data successfully saved to 'cloud_storage.txt'\n");

    return 0;
}
```

OUTPUT:

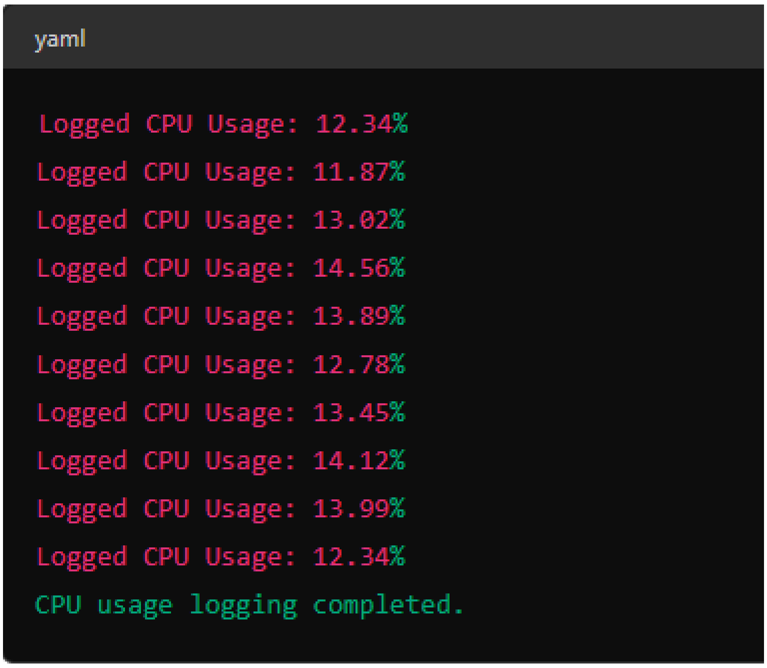
```
Data successfully saved to 'cloud_storage.txt
```

3. Write a C-Program for CPU usage Monitoring and Logging on Cloud-Based-Ubuntu Servers.

```
#include <stdio.h>
#include <unistd.h>
#include <stdlib.h>
#include <string.h>

// Function to read the current CPU usage from the /proc/stat file
float get_cpu_usage() {
    long double a[4], b[4], loadavg;
    FILE *fp;

    fp = fopen("/proc/stat","r");
    fscanf(fp, "%*s %Lf %Lf %Lf %Lf", &a[0], &a[1], &a[2], &a[3]);
    fclose(fp);
    sleep(1);

    fp = fopen("/proc/stat","r");
    fscanf(fp, "%*s %Lf %Lf %Lf %Lf", &b[0], &b[1], &b[2], &b[3]);
    fclose(fp);

    loadavg = ((b[0]+b[1]+b[2]) - (a[0]+a[1]+a[2])) / ((b[0]+b[1]+b[2]+b[3]) - (a[0]+a[1]+a[2]+a[3]));
    return loadavg;
}

int main() {
    FILE *log_file;
    char *filename = "cpu_usage.log";

    // Open log file for writing
    log_file = fopen(filename, "w");
    if (log_file == NULL) {
        perror("Failed to open log file");
        return EXIT_FAILURE;
    }

    // Loop to record CPU usage
    for (int i = 0; i < 10; ++i) {
        float usage = get_cpu_usage();
        fprintf(log_file, "CPU Usage: %.2f%%\n", usage * 100);
        printf("Logged CPU Usage: %.2f%%\n", usage * 100);
```

```
     sleep(5); // Sleep for 5 seconds
    }

    // Close log file
    fclose(log_file);
    printf("CPU usage logging completed.\n");

    return 0;
}
```

OUTPUT:



```
Logged CPU Usage: 12.34%
Logged CPU Usage: 11.87%
Logged CPU Usage: 13.02%
Logged CPU Usage: 14.56%
Logged CPU Usage: 13.89%
Logged CPU Usage: 12.78%
Logged CPU Usage: 13.45%
Logged CPU Usage: 14.12%
Logged CPU Usage: 13.99%
Logged CPU Usage: 12.34%
CPU usage logging completed.
```

4. Use Google App Engine Launcher to launch the web applications.

Aim: To use GAE launcher to launch the web applications

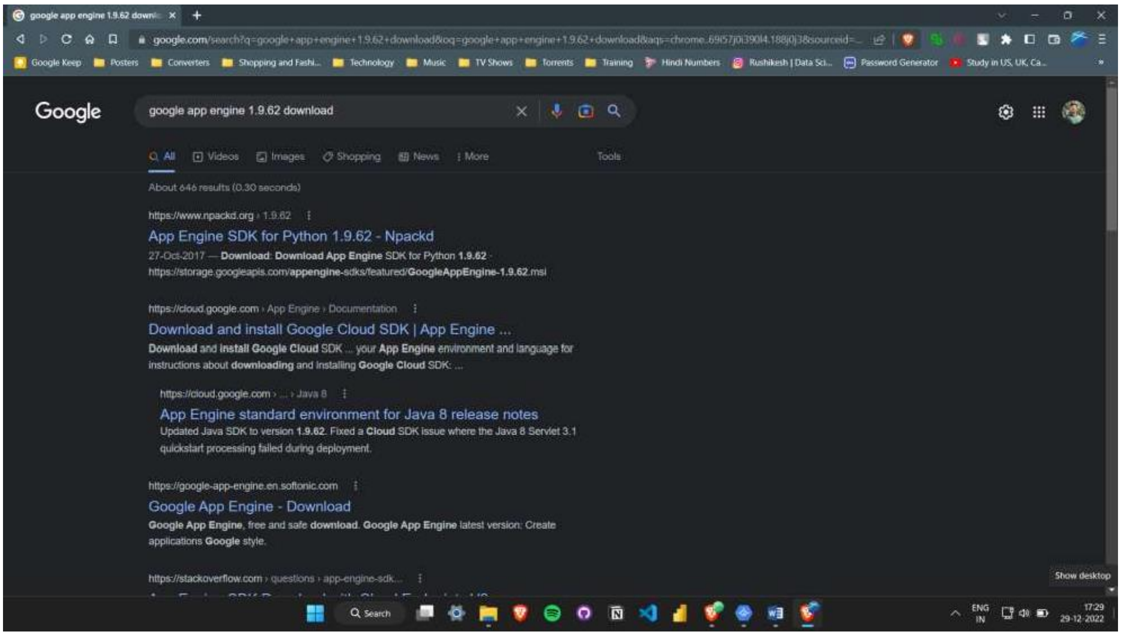
Procedure:

1. Install Google App Engine:

a. Install Google App Engine 1.9.62

i. Go To: https://www.npackd.org/p/com.google.AppEnginePythonSDK/1. 9.62 and install the 1.9.62 version

Make sure Python 2.7 version is installed in your system, if notdownload it on python.org



About 646 results (0.30 seconds)

https://www.npackd.org › 1.9.62

App Engine SDK for Python 1.9.62 - Npackd

27-Oct-2017 — **Download**: **Download App Engine** SDK for Python **1.9.62** · https://storage.googleapis.com/**appengine**-sdks/featured/**GoogleAppEngine-1.9.62**.msi

https://cloud.google.com › App Engine › Documentation

Download and install Google Cloud SDK | App Engine ...

Download and **install Google Cloud** SDK ... your **App Engine** environment and language for instructions about **downloading** and installing **Google Cloud** SDK: ...

https://cloud.google.com › ... › Java 8

App Engine standard environment for Java 8 release notes

Updated Java SDK to version **1.9.62**. Fixed a **Cloud** SDK issue where the Java 8 Servlet 3.1 quickstart processing failed during deployment.

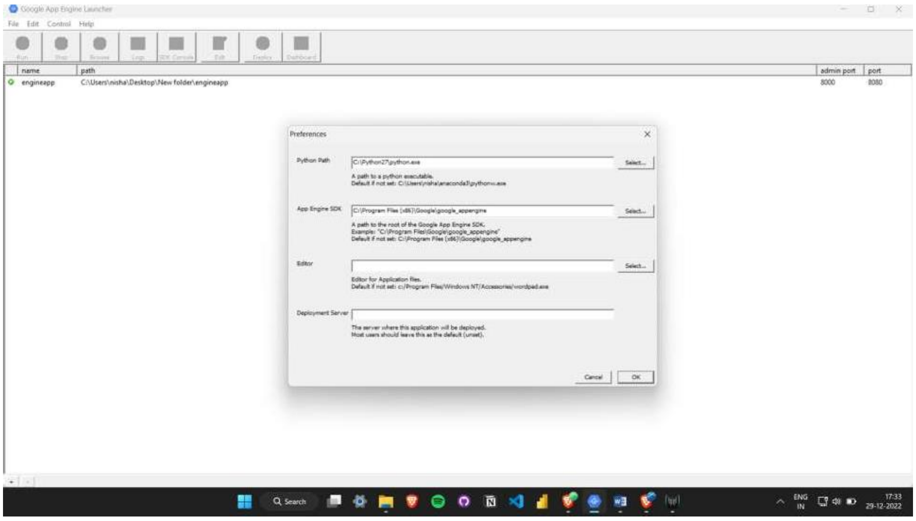
https://google-app-engine.en.softonic.com

Google App Engine - Download

Google App Engine, free and safe **download**. **Google App Engine** latest version: Create applications **Google** style.

https://stackoverflow.com › questions › app-engine-sdk...

a. Create a new folder on Desktop
b. Go to Google App Engine and click on Edit-> Preferences-> Set the path of Google App Engine and Python present in Program Files



Google App Engine Launcher
File Edit Control Help
Run Stop Browse Logs SDK Console Edit Deploy Dashboard

name	path	admin port	port
engineapp	C:\Users\nisha\Desktop\New folder\engineapp	8000	8080

Preferences

Python Path: C:\Python27\python.exe

A path to a python executable.
Default if not set: C:\Users\nisha\anaconda3\pythonw.exe

App Engine SDK: C:\Program Files (x86)\Google\google_appengine

A path to the root of the Google App Engine SDK.
Example: "C:\Program Files\Google\google_appengine"
Default if not set: C:\Program Files (x86)\Google\google_appengine

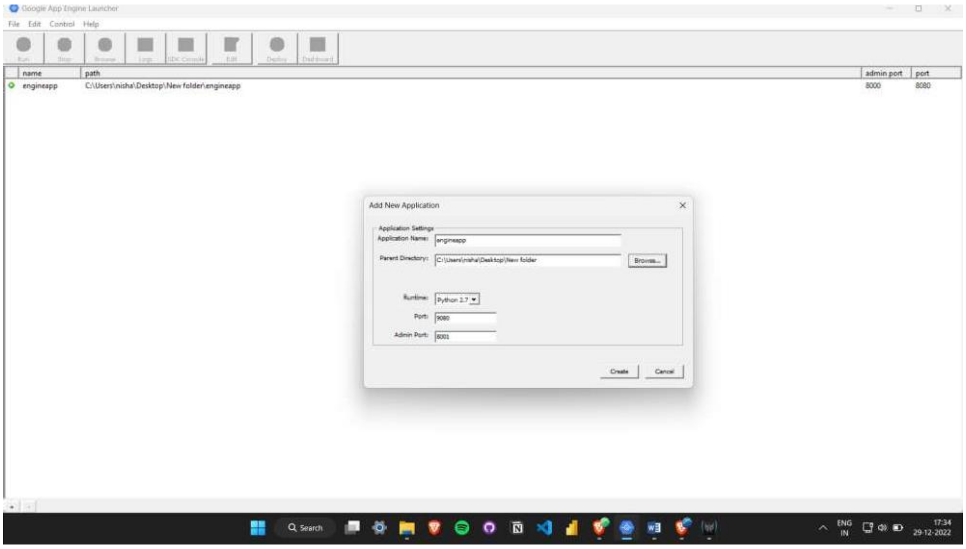
Editor:

Editor for Application files.
Default if not set: c:/Program Files/Windows NT/Accessories/wordpad.exe

Deployment Server:

The server where this application will be deployed.
Most users should leave this as the default (unset).

c. Click on File->Create new Application and select the path of the folder createdearlier.



Google App Engine Launcher
File Edit Control Help
Run Stop Browse Logs SDK Console Edit Deploy Dashboard

name	path	admin port	port
engineapp	C:\Users\nisha\Desktop\New folder\engineapp	8000	8080

Add New Application

Application Settings

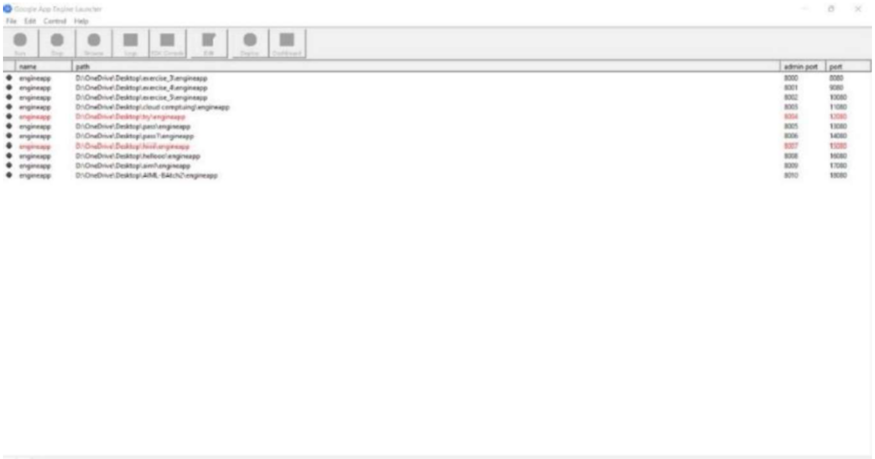
Application Name: engineapp

Parent Directory: C:\Users\nisha\Desktop\New folder

Runtime: Python 2.7

Port: 9080

Admin Port: 8001



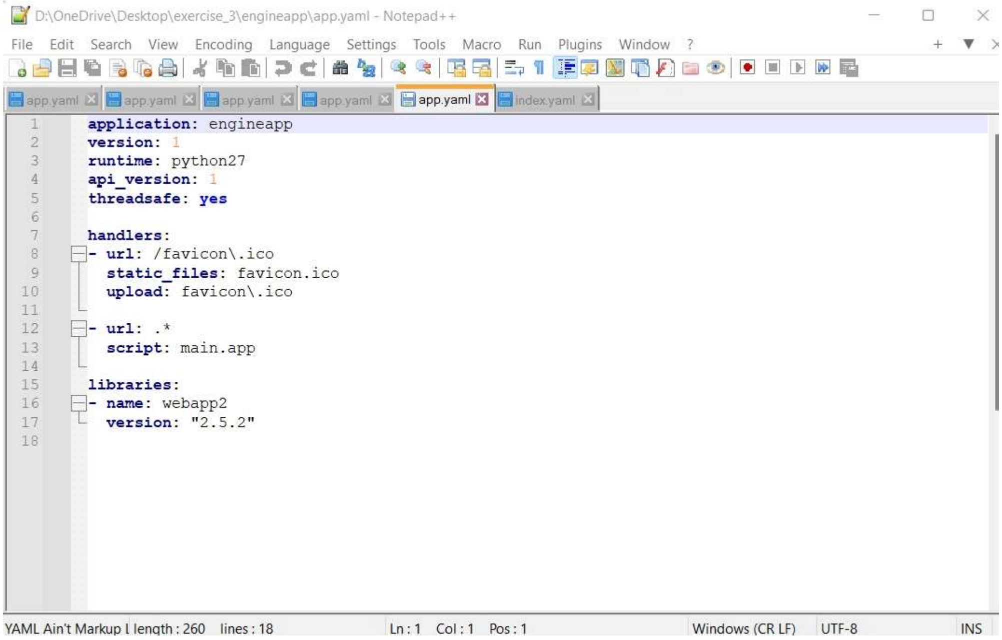
Google App Engine Launcher
File Edit Control Help
Run Stop Browse Logs SDK Console Edit Deploy Dashboard

name	path	admin port	port
engineapp	D:\OneDrive\Desktop\exercise_3\engineapp	8000	8080
engineapp	D:\OneDrive\Desktop\exercise_4\engineapp	8001	9080
engineapp	D:\OneDrive\Desktop\exercise_5\engineapp	8002	10080
engineapp	D:\OneDrive\Desktop\cloud computing\engineapp	8003	11080
engineapp	D:\OneDrive\Desktop\try\engineapp	8004	12080
engineapp	D:\OneDrive\Desktop\pass\engineapp	8005	13080
engineapp	D:\OneDrive\Desktop\pass1\engineapp	8006	14080
engineapp	D:\OneDrive\Desktop\hiiii\engineapp	8007	15080
engineapp	D:\OneDrive\Desktop\hellooo\engineapp	8008	16080
engineapp	D:\OneDrive\Desktop\aim1\engineapp	8009	17080
engineapp	D:\OneDrive\Desktop\AIML-BAtch2\engineapp	8010	18080

d. Open the created folder :

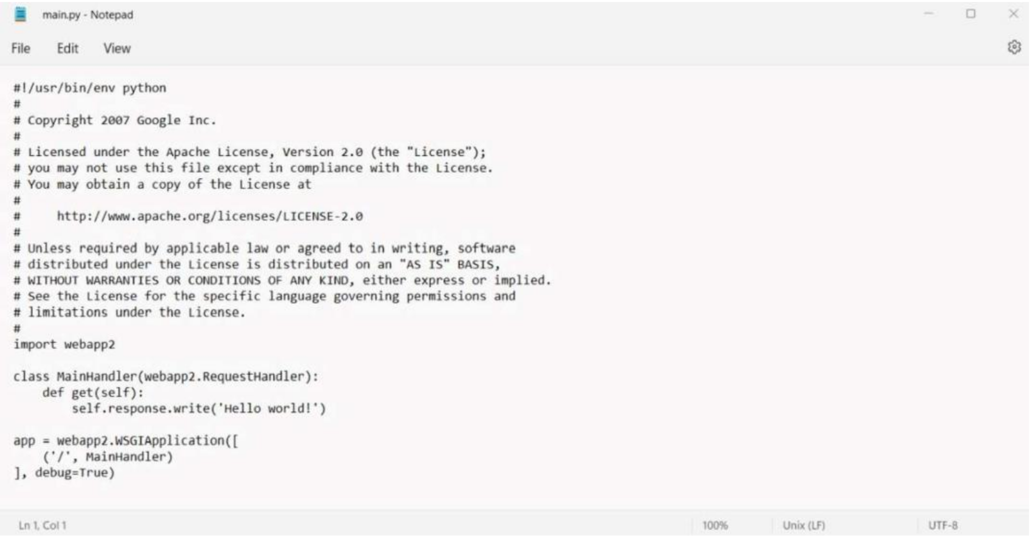
i. Engineapp created automatically

ii. Folder or Project directory must contain `app.yaml` file which contains instruction for GoogleApp Engine To provision the resources for your app.


D:\OneDrive\Desktop\exercise_3\engineapp\app.yaml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
app.yaml
app.yaml
app.yaml
app.yaml
app.yaml
index.yaml
YAML Ain't Markup L length : 260 lines : 18
Ln : 1 Col : 1 Pos : 1
Windows (CR LF)
UTF-8
INS

```yaml
application: engineapp
version: 1
runtime: python27
api_version: 1
threadsafe: yes

handlers:
- url: /favicon\.ico
  static_files: favicon.ico
  upload: favicon\.ico

- url: .*
  script: main.app

libraries:
- name: webapp2
  version: "2.5.2"

```

iii. Contain `main.py` or any other file will handle all the logic.



```python
#!/usr/bin/env python
#
# Copyright 2007 Google Inc.
#
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# 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.
#
import webapp2

class MainHandler(webapp2.RequestHandler):
    def get(self):
        self.response.write('Hello world!')

app = webapp2.WSGIApplication([
    ('/', MainHandler)
], debug=True)
```

Deploying web service:

Run this app in Google App Engine Launcher Launch your browser and access your web service in localhost:port_no of your app

Result:

A web application has been successfully deployed to Google App Engine

5. Installation of Single Node Hadoop Cluster on Ubuntu 22.04 LTS.

Aim: Installation of Single Node Hadoop Cluster on Ubuntu

Prerequisite Test
==============================

```
sudo apt update
sudo apt install openjdk-8-jdk -y

java -version; javac -version
sudo apt install openssh-server openssh-client -y sudo adduser hdoop
su - hdoop
ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys chmod 0600 ~/.ssh/authorized_keys
ssh localhost
```

Downloading Hadoop (Check the path where Hadoop has installed and set HADOOP_HOME to this path)
================================

```
wget https://downloads.apache.org/hadoop/common/hadoop-3.2.3/hadoop-3.2.3.tar.gz tar xzf
hadoop-3.2.3.tar.gz
```

Editng 6 important files
==================================

1st file
============================

sudo nano .bashrc - here you might face issue saying hdoop is not sudo user if this issue comes then

```
su - aman
sudo adduser hdoop sudo

sudo nano .bashrc
#Add below lines in this file

#Hadoop Related Options
export HADOOP_HOME=/home/hdoop/hadoop-3.2.3 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=$HADOOP_HOME/lib/native export
PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS"-Djava.library.path=$HADOOP_HOME/lib/nativ"

source ~/.bashrc 2nd File
============================
sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh #Add below line in this file in the
end
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

3rd File
===============================
sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml

#Add below lines in this file(between "<configuration>" and "<"/configuration>")
<property>
<name>hadoop.tmp.dir</name>
<value>/home/hdoop/tmpdata</value>
<description>A base for other temporary directories.</description>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
<description>The name of the default file system></description>
</property>

4th File
====================================
sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml

#Add below lines in this file(between "<configuration>" and "<"/configuration>")

<property>
<name>dfs.data.dir</name>
<value>/home/hdoop/dfsdata/namenode</value>

</property>
<property>
<name>dfs.data.dir</name>
<value>/home/hdoop/dfsdata/datanode</value>
</property>
<property>
```

```
<name>dfs.replication</name>
<value>1</value>
</property>
```

5th File
===

```
sudo nano $HADOOP_HOME/etc/hadoop/mapred-site.xml
```

#Add below lines in this file(between "<configuration>" and "<"/configuration>")

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

6th File
===

```
sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
```

#Add below lines in this file(between "<configuration>" and "<"/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_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
```

Launching Hadoop (Change to sbin path and run)
=================================

```
hdfs namenode -format

./start-dfs.sh
./start-yarn.sh
```

To check all the daemons of Hadoop s running give JPS and check

6. Hadoop Programming: Word Count Map Reduce Program Using Eclipse

AIM: Hadoop Programming: Word Count MapReduce Program Using Eclipse The entire MapReduce program can be fundamentally divided into three parts:
Driver Code

Mapper Phase Code

Reducer Phase Code

1. Driver Code

```
import java.io.IOException;
import org.apache.hadoop.conf.Configured; import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable; import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat; import
org.apache.hadoop.mapred.FileOutputFormat; import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf; import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;
public class WCDriver extends Configured implements Tool {

public int run(String args[]) throws IOException
{
if (args.length < 2)
{
System.out.println("Please give valid inputs"); return -1;
}
JobConf conf = new JobConf(WCDriver.class); FileInputFormat.setInputPaths(conf, new
Path(args[0])); FileOutputFormat.setOutputPath(conf, new Path(args[1]));
conf.setMapperClass(WCMapper.class); conf.setReducerClass(WCReducer.class);


conf.setMapOutputKeyClass(Text.class); conf.setMapOutputValueClass(IntWritable.class);
conf.setOutputKeyClass(Text.class); conf.setOutputValueClass(IntWritable.class);
JobClient.runJob(conf);
return 0;
}
public static void main(String args[]) throws Exception
{
```

```
int exitCode = ToolRunner.run(new WCDriver(), args); System.out.println(exitCode);
}
}
```

2. Mapper Code

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable; import org.apache.hadoop.io.LongWritable; import
org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase; import
org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector; import
org.apache.hadoop.mapred.Reporter;
public class WCMapper extends MapReduceBase implements Mapper<LongWritable, Text,
Text, IntWritable>
{
public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output,
Reporter rep) throws IOException
{
String line = value.toString();

// Splitting the line on spaces for (String word : line.split(" "))
{
if (word.length() > 0)
{
output.collect(new Text(word), new IntWritable(1));
}
}
}
}
```

3. Reducer Code

```
import java.io.IOException; import java.util.Iterator;
import org.apache.hadoop.io.IntWritable; import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase; import
```

```
org.apache.hadoop.mapred.OutputCollector; import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;
public class WCReducer extends MapReduceBase implements Reducer<Text, IntWritable,
Text, IntWritable>

{
public void reduce(Text key, Iterator<IntWritable> value, OutputCollector<Text, IntWritable>
output,
Reporter rep) throws IOException
{

int count = 0;

// Counting the frequency of each words while (value.hasNext())
{
IntWritable i = value.next(); count += i.get();
}

output.collect(key, new IntWritable(count));
}
}
```

Run the MapReduce code:

The command for running a MapReduce code is:

```
Hadoop jar hadoop-mapreduce-example.jar WordCount /sample/input /sample/output
```

7. File Management tasks in Hadoop using HDFS commands
a. Adding files to HDFS
b. Retrieving files from HDFS
c. Deleting files from HDFS

AIM: File Management tasks in Hadoop using HDFS commands
a. Adding files to HDFS
b. Retrieving files from HDFS
c. Deleting files from HDFS

1. Create a directory in HDFS atgiven path(s).

Usage:

hadoop fs -mkdir <paths> Example:

hadoop fs -mkdir /user/saurzcode/dir1 /user/saurzcode/dir2

2. List the contents of adirectory.

Usage :

hadoop fs -ls <args> Example:
hadoop fs -ls /user/saurzcode

3. Upload and download a file in HDFS.

Upload: hadoop fs put:

Copy single src file, or multiple src files from local file systemto the Hadoop data file system

Usage:

hadoop fs -put <localsrc> ... <HDFS_dest_Path>

Example: hadoop fs -put /home/saurzcode/Samplefile.txt /user/saurzcode/dir3/

Download: hadoop fs -get: Copies/Downloads files to the local file system

Usage:

hadoop fs -get <hdfs_src> <localdst>

Example: hadoop fs get/user/saurzcode/dir3/Samplefile.txt /home/

4. See contents of a file

Same as unix cat command:

Usage:

hadoop fs -cat <path[filename]>Example:

hadoop fs -cat /user/saurzcode/dir1/abc.txt

5. Remove a file or directory inHDFS.

Remove files specified as argument. Deletes directory only when it is empty

Usage : hadoop fs -rm <arg>

Example: hadoop fs -rm /user/saurzcode/dir1/abc.txt

6. Display last few lines of a file.

Similar to tail command in Unix.

Usage :

hadoop fs -tail <path[filename]>

Example: hadoop fs tail/user/saurzcode/dir/ ab.txt

8. Creation of Data Frames using Create Data Frame working with an example related to FIFA dataset.

AIM: Creation of DataFrames using CreateDataFrame working with an example related to FIFA dataset.

Creation of DataFrame in Spark

Let us use the following code to create a new DataFrame.
Here, we shall create a new DataFrame using the createDataFrame method. we shall design the schema for the data that we will read from fifa csv file. Finally, let us use the createDataFrame method to create our DataFrame Hence, we create DataFrame and display it by using the .show method.

Before we read the data from a CSV file, we need to import certain libraries which we need for processing the DataFrames in Spark.

```
import org.apache.spark._ import org.apache.spark.sql._
import org.apache.spark.sql.types._

import org.apache.spark.storage.StorageLevel import scala.collection.mutable.HashMap
import java.io.File
import org.apache.spark.sql.Row import org.apache.spark.util.IntParam
import scala.collection.mutable.ListBuffer

import org.apache.spark.sql.types.{StructType, StructField, StringType, IntegerType};
```

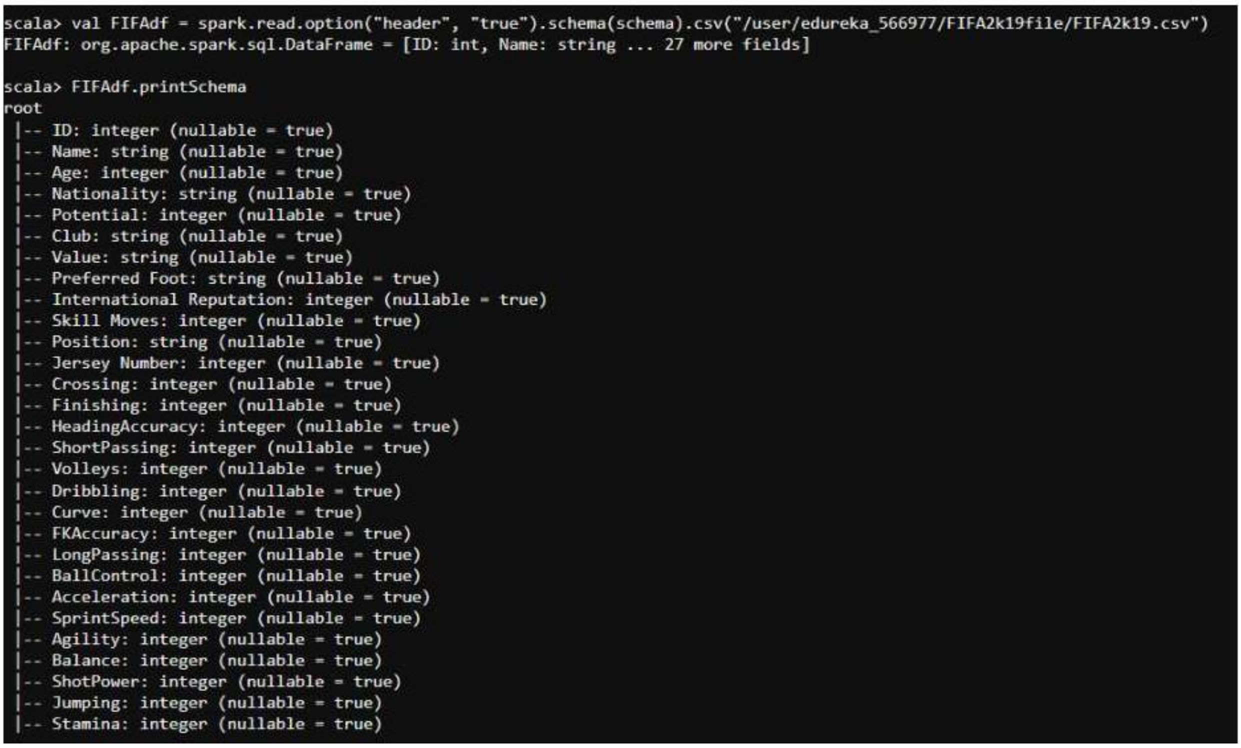
We design the schema for our CSV file once we import libraries, val schema =

```
StructType(Array(StructField("ID",IntegerType,true),StructField("Name",StringType,true),Struc
tField("Age",IntegerType,true),StructField("Nationality",StringType,true),StructField("Potential
",IntegerType,true),StructField("Club",StringType,true),StructField("Value",StringType,true),Str
uctField("PreferredFoot",StringType,true),StructField("InternationalReputation",IntegerType,tru
e),StructField("SkillMoves",IntegerType,true),StructField("Position",StringType,true),StructFiel
d("JerseyNumber",IntegerType,true),StructField("Crossing",IntegerType,true),StructField("Finis
hing",IntegerType,true),StructField("HeadingAccuracy",IntegerType,true),StructField("ShortPas
sing",IntegerType,true),StructField("Volleys",IntegerType,true),StructField("Dribbling",Integer
Type,true),StructField("Curve",IntegerType,true),StructField("FKAccuracy",IntegerType,true),S
tructField("LongPassing",IntegerType,true),StructField("BallControl",IntegerType,true),StructFi
eld("Acceleration",IntegerType,true),StructField("SprintSpeed",IntegerType,true),StructField("A
gility",IntegerType,true),StructField("Balance",IntegerType,true),StructField("ShotPower",Integ
erType,true),StructField("Jumping",IntegerType,true),StructField("Stamina",IntegerType,true)))
```

- Let us load the Fifa data from a CSV file from the HDFS as shown below. We are firstgoing to use Spark.read.format(“csv”) method for reading our CSV file from our HDFS.

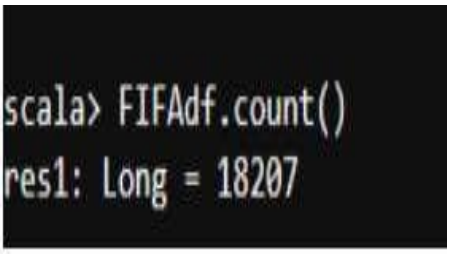
ValFifAdfspark.read.format("csv").option("header",true).load("/home/rash/Downloads/players_16.c sv")

- Let us use .print Schema() method to see the schema of our CSV file.



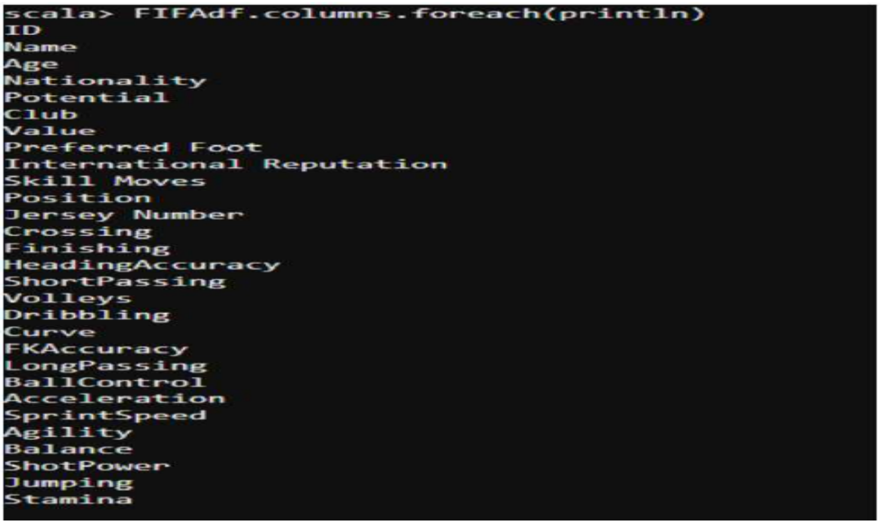
```
scala> val FIFAdf = spark.read.option("header", "true").schema(schema).csv("/user/edureka_566977/FIFA2k19file/FIFA2k19.csv")
FIFAdf: org.apache.spark.sql.DataFrame = [ID: int, Name: string ... 27 more fields]

scala> FIFAdf.printSchema
root
 |-- ID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Nationality: string (nullable = true)
 |-- Potential: integer (nullable = true)
 |-- Club: string (nullable = true)
 |-- Value: string (nullable = true)
 |-- Preferred Foot: string (nullable = true)
 |-- International Reputation: integer (nullable = true)
 |-- Skill Moves: integer (nullable = true)
 |-- Position: string (nullable = true)
 |-- Jersey Number: integer (nullable = true)
 |-- Crossing: integer (nullable = true)
 |-- Finishing: integer (nullable = true)
 |-- HeadingAccuracy: integer (nullable = true)
 |-- ShortPassing: integer (nullable = true)
 |-- Volleys: integer (nullable = true)
 |-- Dribbling: integer (nullable = true)
 |-- Curve: integer (nullable = true)
 |-- FKAccuracy: integer (nullable = true)
 |-- LongPassing: integer (nullable = true)
 |-- BallControl: integer (nullable = true)
 |-- Acceleration: integer (nullable = true)
 |-- SprintSpeed: integer (nullable = true)
 |-- Agility: integer (nullable = true)
 |-- Balance: integer (nullable = true)
 |-- ShotPower: integer (nullable = true)
 |-- Jumping: integer (nullable = true)
 |-- Stamina: integer (nullable = true)
```

Let us find out the total number of rows we have using the following code.
FifaAdf.count()



```
scala> FIFAdf.count()
res1: Long = 18207
```

Let us now find the columns we have in our CSV file. We shall use the following code.
FifaAdf.columns.foreach(println)

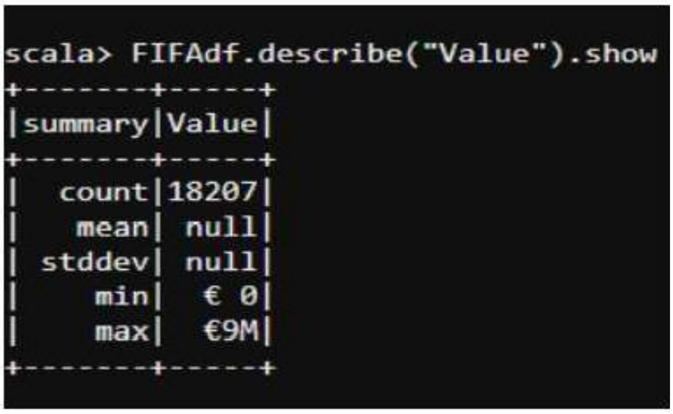


```
scala> FIFAdf.columns.foreach(println)
ID
Name
Age
Nationality
Potential
Club
Value
Preferred Foot
International Reputation
Skill Moves
Position
Jersey Number
Crossing
Finishing
HeadingAccuracy
ShortPassing
Volleys
Dribbling
Curve
FKAccuracy
LongPassing
BallControl
Acceleration
SprintSpeed
Agility
Balance
ShotPower
Jumping
Stamina
```

- If you wish to look at the summary of a particular column in a DataFrame, we can apply to describe command. This command will give us the statistical summary of a particular selected column if nothing is specified, and then it provides the statistical information of the DataFrame.

- Let us find out the description of the Value column to know the minimum and maximum values present in it.

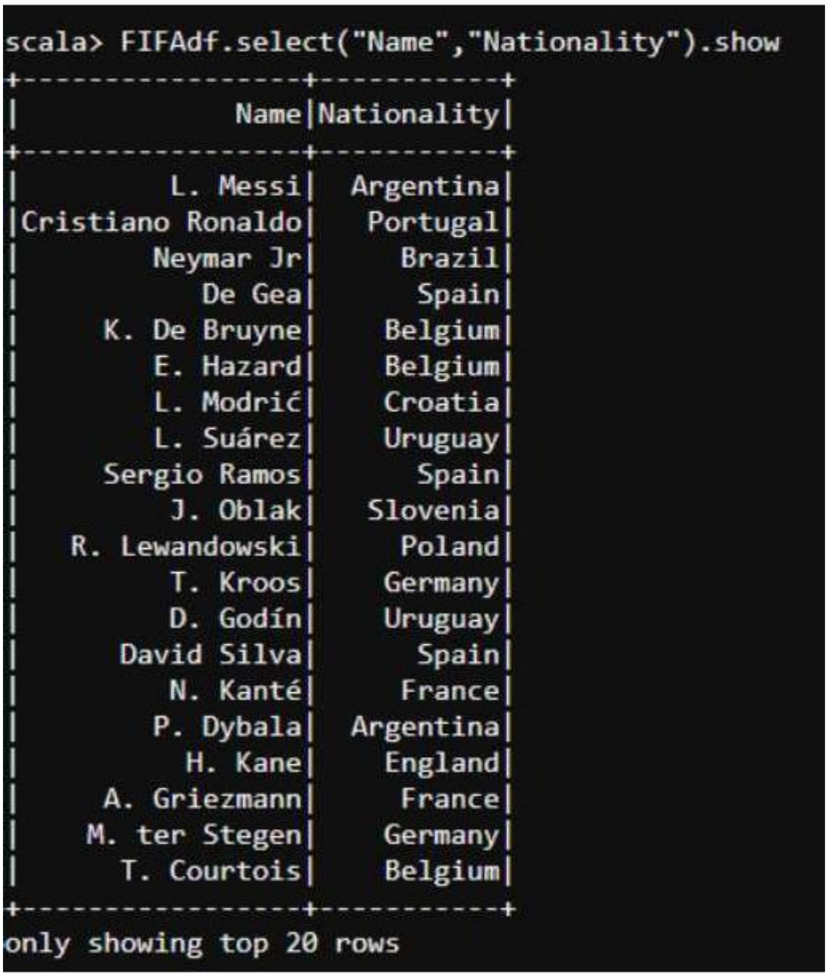
FifaAdf.describe("Value").show



```
scala> FIFAdf.describe("Value").show
+-------+-----+
|summary|Value|
+-------+-----+
|  count|18207|
|   mean| null|
| stddev| null|
|    min|  € 0|
|    max|  €9M|
+-------+-----+
```

- We shall find out the Nationality of a particular player by using the select command.

FifaAdf.select("name","Nationality").show



```
scala> FIFAdf.select("Name","Nationality").show
+-----------------+-----------+
|             Name|Nationality|
+-----------------+-----------+
|         L. Messi|  Argentina|
|Cristiano Ronaldo|   Portugal|
|        Neymar Jr|     Brazil|
|           De Gea|      Spain|
|     K. De Bruyne|    Belgium|
|        E. Hazard|    Belgium|
|        L. Modrić|    Croatia|
|        L. Suárez|    Uruguay|
|     Sergio Ramos|      Spain|
|         J. Oblak|   Slovenia|
|   R. Lewandowski|     Poland|
|         T. Kroos|    Germany|
|         D. Godín|    Uruguay|
|      David Silva|      Spain|
|         N. Kanté|     France|
|        P. Dybala|  Argentina|
|          H. Kane|    England|
|     A. Griezmann|     France|
|     M. ter Stegen|    Germany|
|       T. Courtois|    Belgium|
+-----------------+-----------+
only showing top 20 rows
```

Introduction to Cloud Analyst

"Cloud Analyst" is a simulator used for simulating large scaled applications along with a novel approach for such research oriented studies. There are several toolkits that can be used to model a simulated environment to study the behavior of a large scaled application on the Internet. But having an easy to use tool with a level of visualisation capability is even better than just a toolkit. Such a tool separates the simulation experiment set up exercise from a programming exercise and enables a modeler to concentrate on the simulation parameters rather than the technicalities of programming. A graphical output of the simulation results enables the results to be analyzed more easily and more efficiently and it may also help in quickly highlighting any problems with the performance and accuracy of the simulation logic.

Features of the Simulator

2.2.1 Ease of use

Ease of setting up and executing a simulation experiment is the main point of having a simulation tool. The simulator needs to provide an easy to use graphical user interface which is intuitive yet comprehensive.

2.2.2 Ability to define a simulation with a high degree of configurability and flexibility. Perhaps the most important feature is the level of configurability the tool can provide. A simulation, especially of the nature of modelling something as complex as an Internet Application depends on many parameters and most of the time the values for those parameters need to be assumed. Therefore it is important to be able to enter and change those parameters quickly and easily and repeat simulations.

2.2.3 Graphical output

A picture is said to be worth a thousand words. Graphical output in the form of tables and charts is highly desirable to summarize the potentially large amount of statistics that is collected during the simulation. Such effective presentation helps in identifying the important patterns of the output parameters and helps in comparisons between related parameters.

2.2.4 Repeatability

Repeatability of experiments is a very important requirement of a simulator. The same experiment with the same parameters should produce similar results each time the simulation is executed. Otherwise the simulation becomes just a random sequence of events rather than a controlled experiment. It is also helpful to be able to save an experiment (the set of input parameters) as a file and also be able to save the results of an experiment as a file.

2.2.5 Ease of extension

As already mentioned simulating something like the Internet is a complex task and it is 9 unlikely a 100% realistic simulation framework and a set of input parameters can be achieved in a few attempts. Therefore the simulator is expected to evolve continuously rather than a program that is written once and for all and then used continuously. Therefore the simulator architecture should support extensions with minimal effort with suitable frameworks.

2.3 Simulation Output / what is being measured

Following are the statistical measures produced as output of the simulation in the initial version of the simulator.

• Response time of the simulated application

- Overall average, minimum and maximum response time of all user requests simulated

- The response time broken down by user groups, located within geographical regions
- The response time further broken down by the time showing the pattern of change over the duration of a day

- The usage patterns of the application
    - How many users use the application at what time from different regions of the world, and the overall effect of that usage on the data centers hosting the application
- The time taken by data centers to service a user request
    - The overall request processing time for the entire simulation
    - The average, minimum and maximum request processing time by each data center
    - The response time variation pattern during the day as the load changes

2.4 Technologies Used

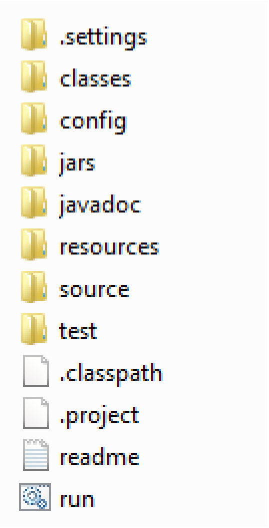
- Java – The simulator is developed 100% on Java platform, using Java SE 1.6.
- Java Swing – The GUI component is built using Swing components.
- CloudSim – CloudSim features for modelling data centers is used in CloudAnalyst.
- SimJava – Sim Java is the underlying simulation framework of CloudSim and some features of SimJava are used directly in CloudAnalyst.

CloudAnalyst Design

The CloudAnalyst is built on top of CloudSim tool kit, by extending CloudSim functionality with the introduction of concepts that model Internet and Internet Application behaviors.

Steps tp install and Run Cloud Analyst

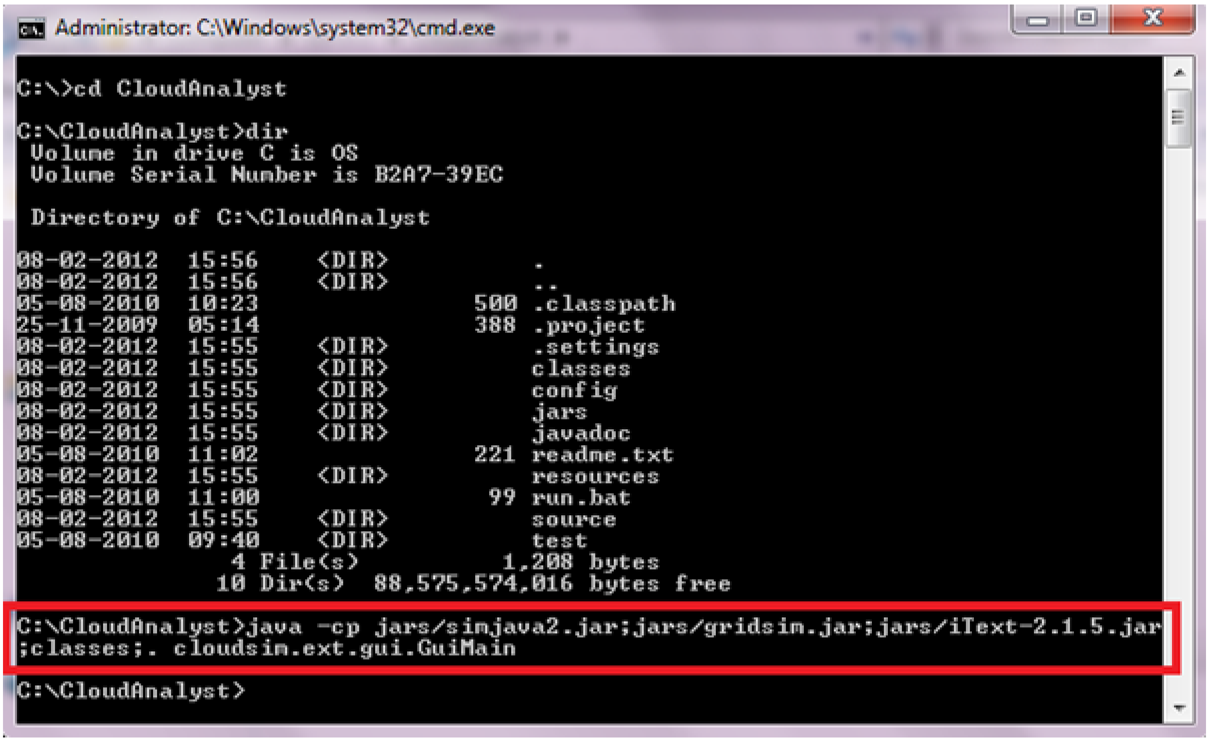
1. Download installer package from http://www.cloudbus.org/cloudsim/CloudAnalyst.zi
2. Extract Files from the compressed package file which will give following folder structure



```
.settings
classes
config
jars
javadoc
resources
source
test
.classpath
.project
readme
run
```

3. To start the simulator, you can got Command line then use the following command on the command prompt as shown in the screenshot below C:\CloudAnalyst>java – cp jars/simjava2.jar;jars/gridsim.jar;jars/iText-

2.1.5.jar;classes;. Cloudsim.ext.gui.GuiMain

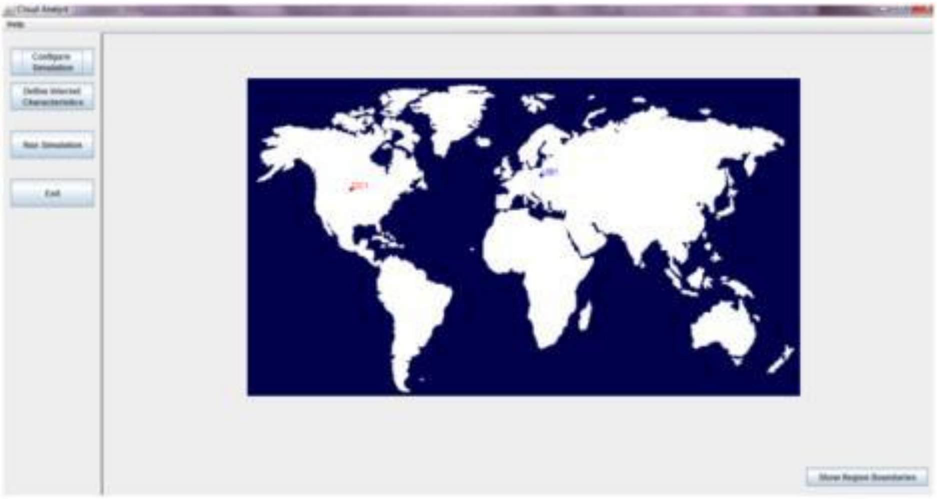


Administrator: C:\Windows\system32\cmd.exe

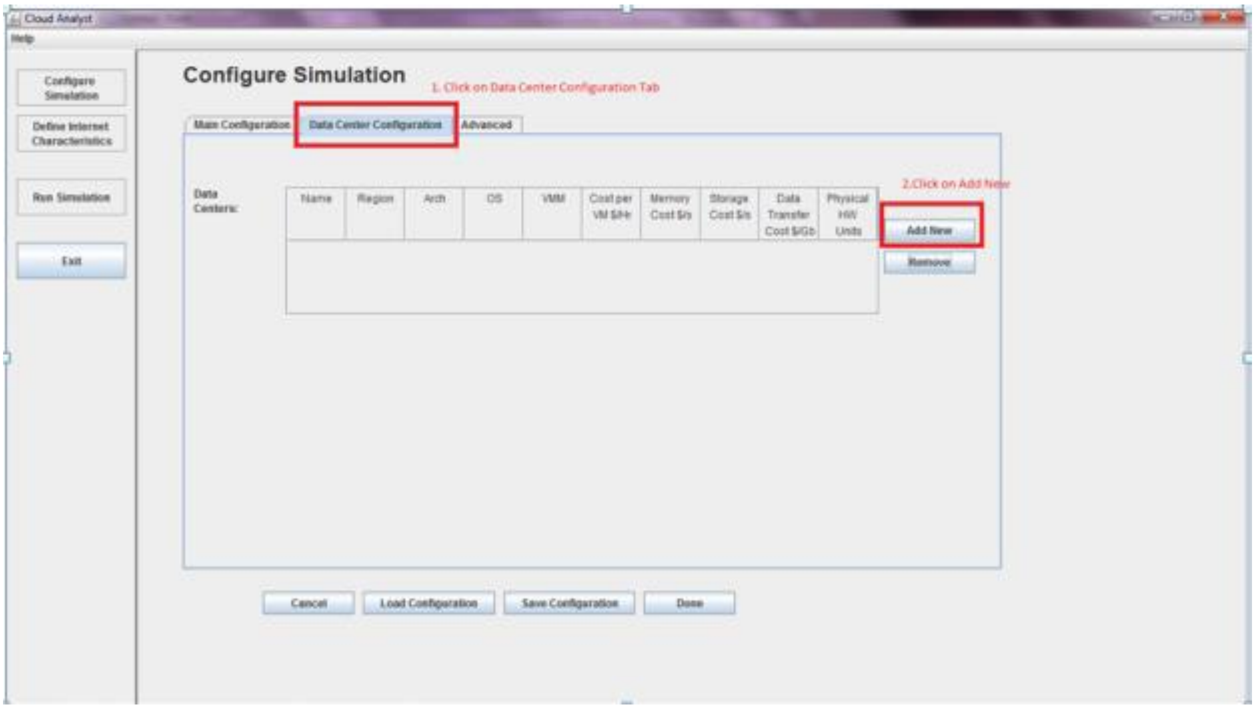
```
C:\>cd CloudAnalyst

C:\CloudAnalyst>dir
 Volume in drive C is OS
 Volume Serial Number is B2A7-39EC

 Directory of C:\CloudAnalyst

08-02-2012  15:56    <DIR>          .
08-02-2012  15:56    <DIR>          ..
05-08-2010  10:23               500 .classpath
25-11-2009  05:14               388 .project
08-02-2012  15:55    <DIR>          .settings
08-02-2012  15:55    <DIR>          classes
08-02-2012  15:55    <DIR>          config
08-02-2012  15:55    <DIR>          jars
08-02-2012  15:55    <DIR>          javadoc
05-08-2010  11:02               221 readme.txt
08-02-2012  15:55    <DIR>          resources
05-08-2010  11:00                99 run.bat
08-02-2012  15:55    <DIR>          source
05-08-2010  09:40    <DIR>          test
               4 File(s)          1,208 bytes
              10 Dir(s)  88,575,574,016 bytes free

C:\CloudAnalyst>java -cp jars/simjava2.jar;jars/gridsim.jar;jars/iText-2.1.5.jar
;classes;. cloudsim.ext.gui.GuiMain

C:\CloudAnalyst>
```

Alternatively you can also click on run.bat file and click done!!

4. The welcome screen of the simulator looks like the following image



5. To configure the parameters for experimentation click on “Show Region Boundaries”. Here you can datacenters and other physical hardware details of the datacenter as shown in the following screenshots.



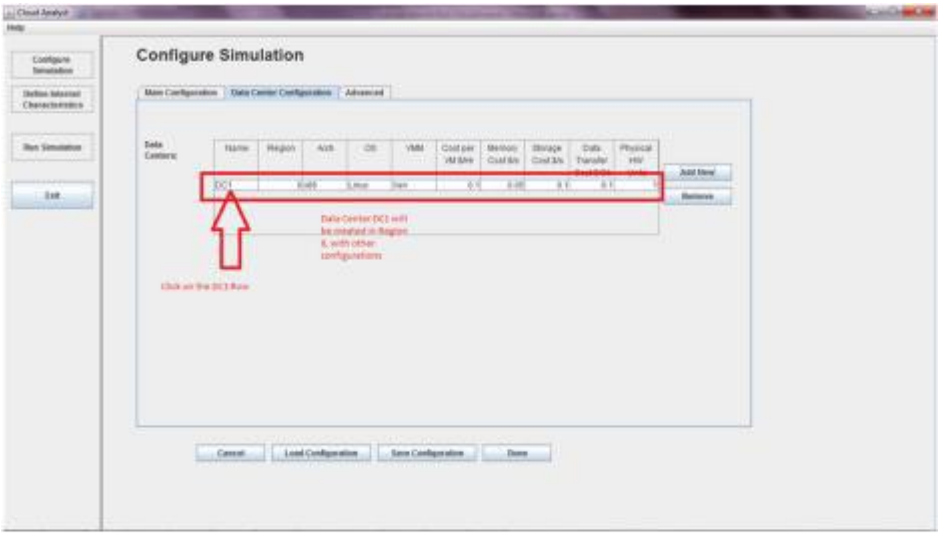
Configure Simulation

1. Click on Data Center Configuration Tab

2.Click on Add New

Data Centers:

Name	Region	Arch	OS	VMM	Cost per VM $/Hr	Memory Cost $/s	Storage Cost $/s	Data Transfer Cost $/Gb	Physical HW Units



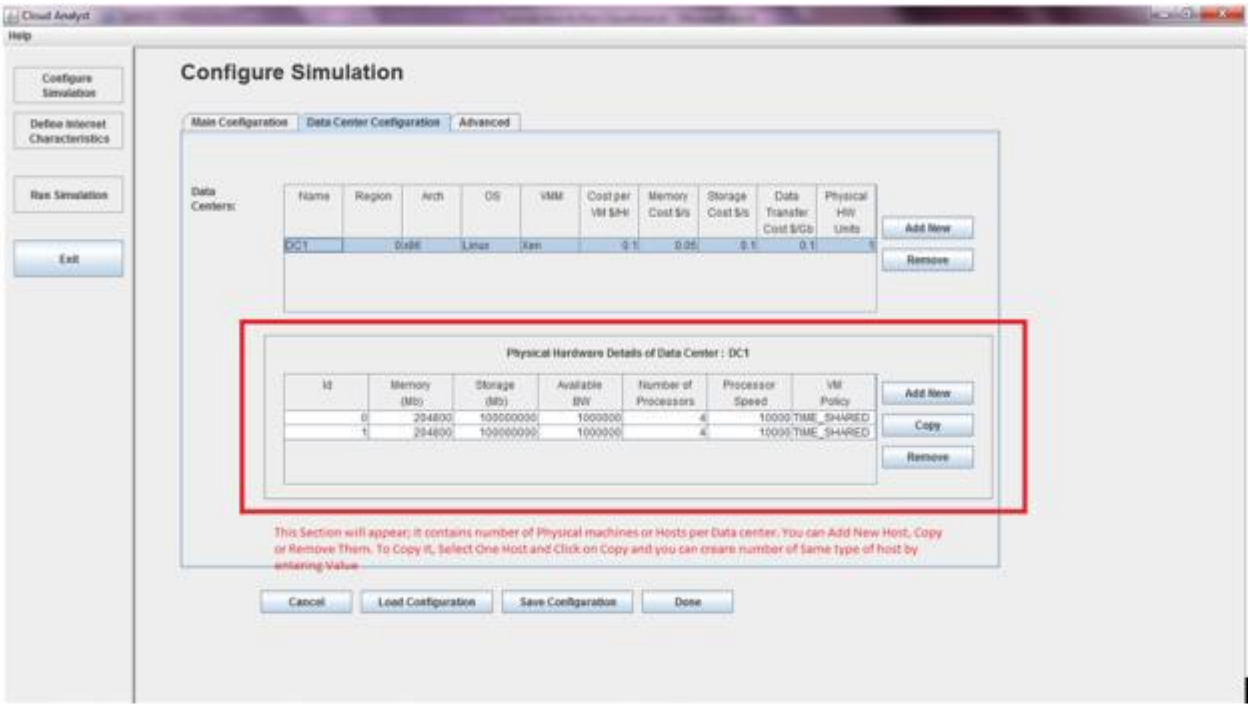
Configure Simulation

Data Centers:

Name	Region	Arch	OS	VMM	Cost per VM $/Hr	Memory Cost $/s	Storage Cost $/s	Data Transfer Cost $/Gb	Physical HW Units
DC1	0	x86	Linux	Xen	0.1	0.05	0.1	0.1	1

Data Center DC1 will be created in Region 0, with other configurations

Click on the DC1 Row



Configure Simulation

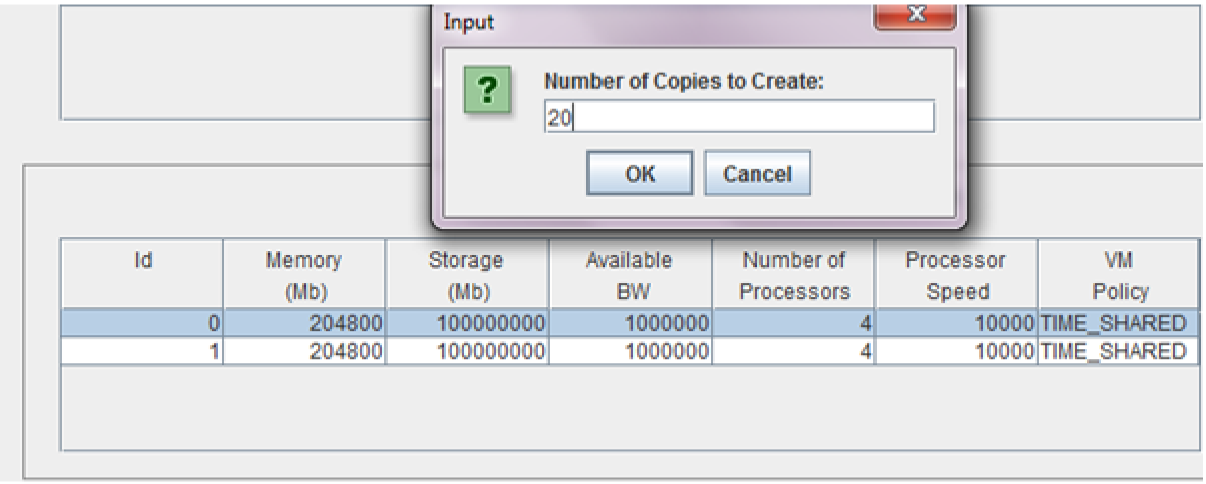
Data Centers:

Name	Region	Arch	OS	VMM	Cost per VM $/Hr	Memory Cost $/s	Storage Cost $/s	Data Transfer Cost $/Gb	Physical HW Units
DC1	0	x86	Linux	Xen	0.1	0.05	0.1	0.1	1

Physical Hardware Details of Data Center : DC1

Id	Memory (Mb)	Storage (Mb)	Available BW	Number of Processors	Processor Speed	VM Policy
0	204800	100000000	1000000	4	10000	TIME_SHARED
1	204800	100000000	1000000	4	10000	TIME_SHARED

This Section will appear; it contains number of Physical machines or Hosts per Data center. You can Add New Host, Copy or Remove Them. To Copy it, Select One Host and Click on Copy and you can creare number of Same type of host by entering Value

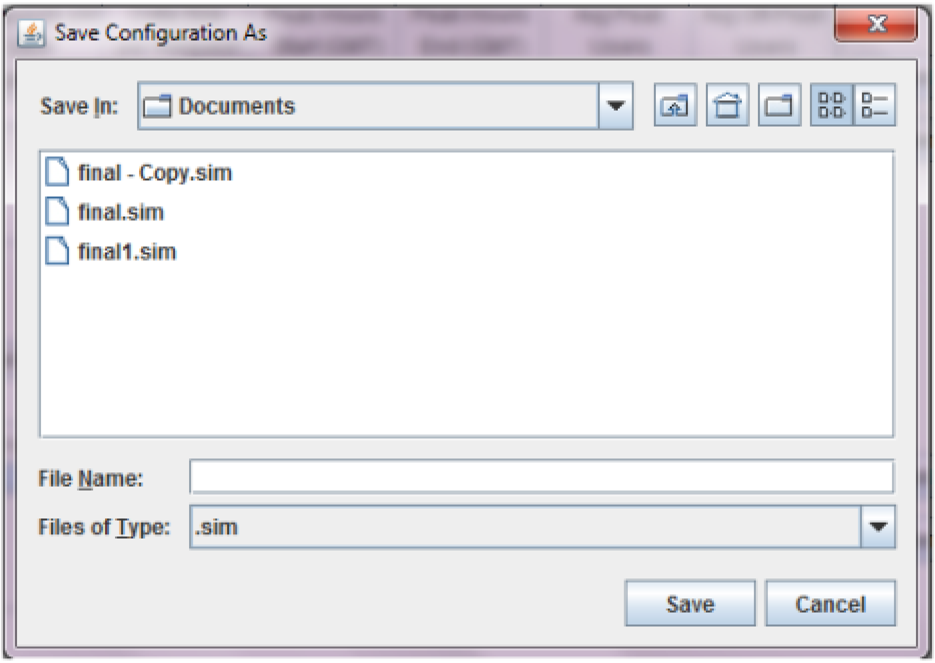


Number of Copies to Create:

20

Id	Memory (Mb)	Storage (Mb)	Available BW	Number of Processors	Processor Speed	VM Policy
0	204800	100000000	1000000	4	10000	TIME_SHARED
1	204800	100000000	1000000	4	10000	TIME_SHARED

6. You can save this Configuration as well in case you want to use it later. It is stored as .sim file. XML data is generated and saved as Sim file.



Save Configuration As

Save In: Documents

final - Copy.sim

final.sim

final1.sim

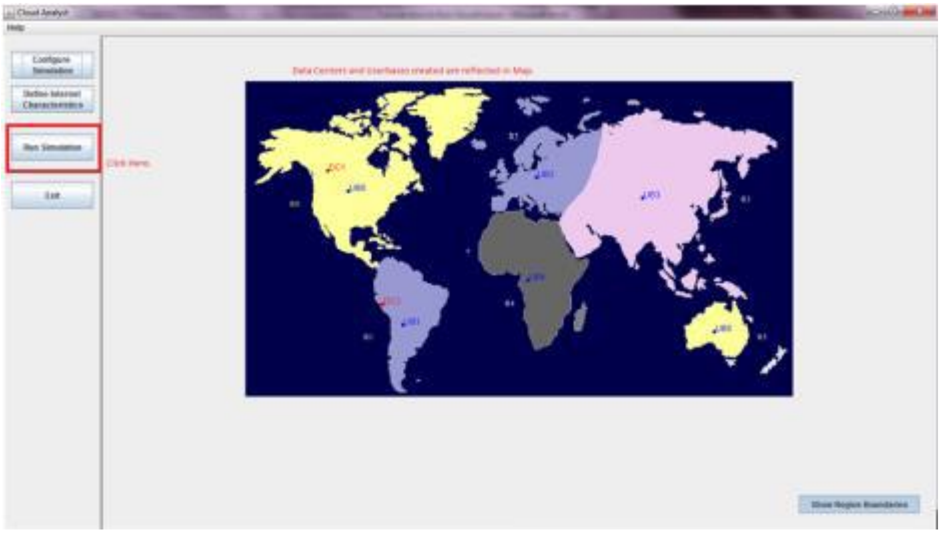
File Name:

Files of Type: .sim

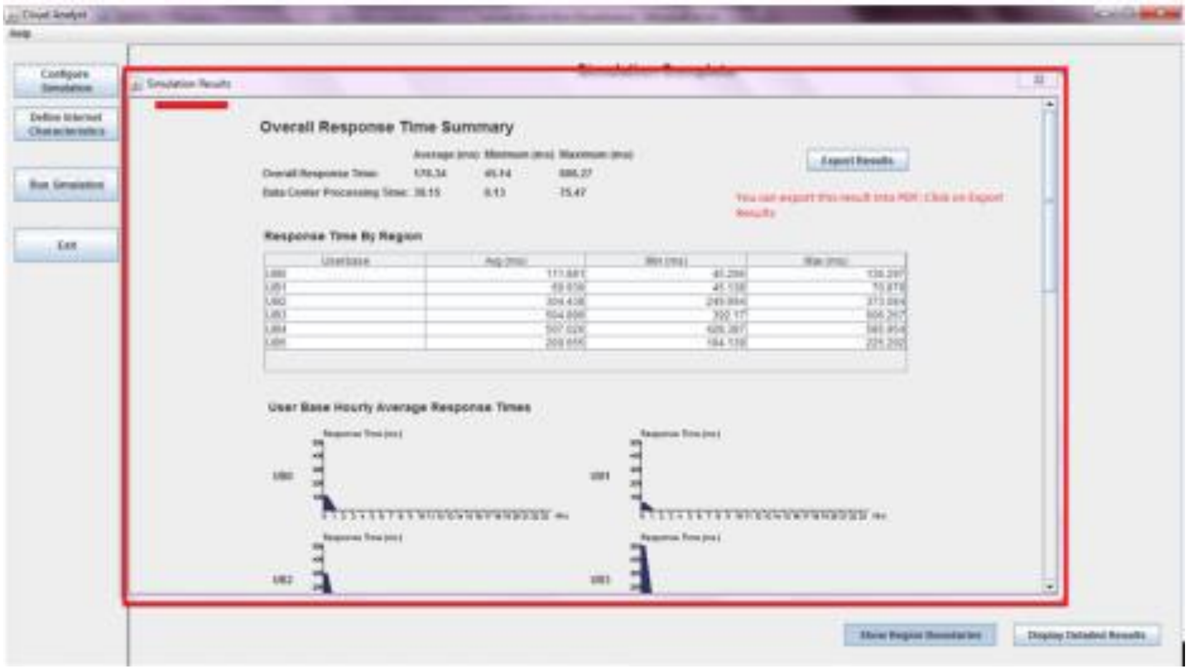
Saved configurations can be loaded anytime easily into Cloud Analyst.

7. Once you are done with Configuration; click on done.

8. To run the simulation click on Run Simulation



9. Simulation Results Window will open



Overall Response Time Summary

	Average (ms)	Minimum (ms)	Maximum (ms)
Overall Response Time:	170.34	45.14	606.27
Data Center Processing Time:	38.15	0.13	75.47

You can export this result into PDF. Click on Export Results

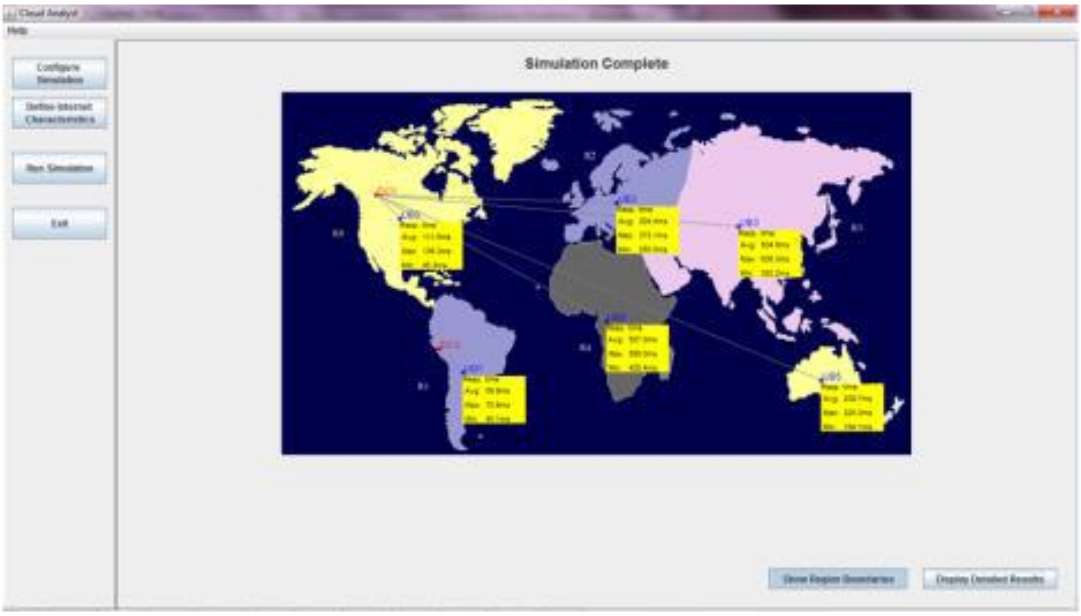
Response Time By Region

Userbase	Avg (ms)	Min (ms)	Max (ms)
UB0	111.881	45.299	136.297
UB1	49.939	45.138	70.878
UB2	304.438	249.884	373.064
UB3	504.898	392.17	606.267
UB4	507.026	426.387	585.854
UB5	209.656	184.139	225.292

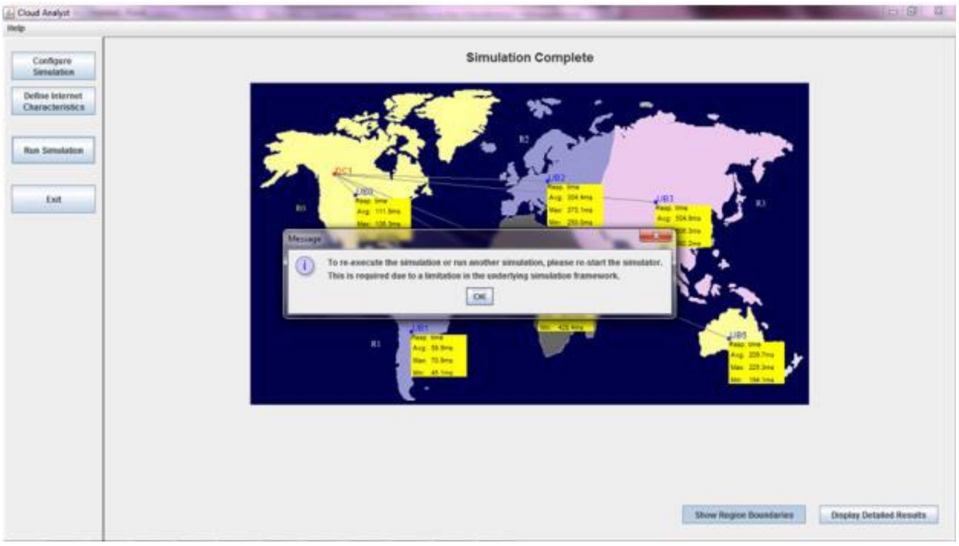
User Base Hourly Average Response Times

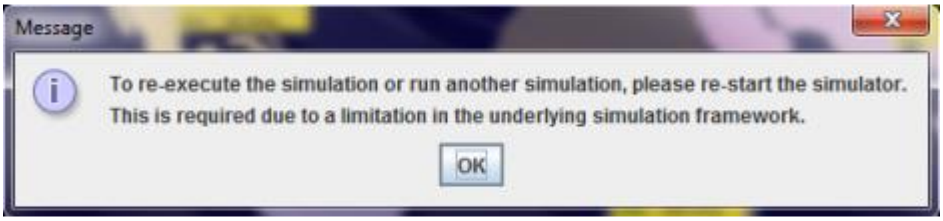
Then click on “Close it.”

10.Main Window will give all statistics



11.If you will try to run Simulation again then it will give Error





To re-execute the simulation or run another simulation, please re-start the simulator.
This is required due to a limitation in the underlying simulation framework.

12. Click on Exit & restart the simulator.

10. Simulator program: Analyze the average, minimum and maximum response time and datacenter Processing time and write the findings.

In given cloud simulator to host a simple web application with the below given configurations

i. Six user bases and one datacenter with 50 virtual machines each with 1024 Mb of memory and processor speed as 100 MIPS.

ii. Four user bases and one datacenter with 25 virtual machines each with 1024 Mb of memory and processor speed as 100 MIPS.

Analyze the average, minimum and maximum response time and datacenter processing time and write the findings.

Procedure for (i):

Step1: Start the Cloud Analyst simulator by clicking on run.bat file.

Step 2: Once the simulator starts navigate to Configure Simulation tab.

Step 3: In Main Configuration tab, add six user bases by clicking add new button in the User Bases table.

Step4: Set the memory to 1024 Mb in the Application Deployment Configuration table.

Step 5: Set the number of VM's to 50 in the Application Deployment Configuration.

Step 6: Navigate to Data Center Configuration tab and click on the Datacenter name ie DC1.

Step 7: The physical hardware details of Data Center table is displayed where you have to set the processor speed to 100 MIPS.

Step 8: Click on Save Configuration to save the simulation by naming it.

Step 9: Click Done.

Step 10: Click on Run Simulation to run the experiment.

Step 11: Wait for the results to be loaded.

Step 12: Note down the average, maximum and minimum response time for the user bases & datacenter processing times.

Step 13: Draw the corresponding graphs.

Procedure for (ii):

Step1: Start the Cloud Analyst simulator by clicking on run.bat file.

Step 2: Once the simulator starts navigate to Configure Simulation tab.

Step 3: In Main Configuration tab, add four user bases by clicking add new button in the User Bases table.

Step4: Set the memory to 1024 Mb in the Application Deployment Configuration table.

Step 5: Set the number of VM's to 25 in the Application Deployment Configuration.

Step 6: Navigate to Data Center Configuration tab and click on the Datacenter name ie DC1.

Step 7: The physical hardware details of Data Center table is displayed where you have to set the processor speed to 100 MIPS.

Step 8: Click on Save Configuration to save the simulation by naming it.

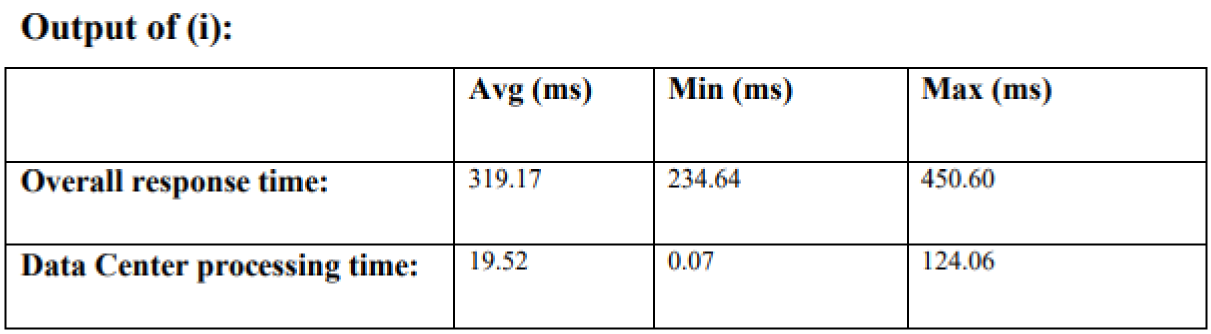
Step 9: Click Done.

Step 10: Click on Run Simulation to run the experiment.

Step 11: Wait for the results to be loaded.

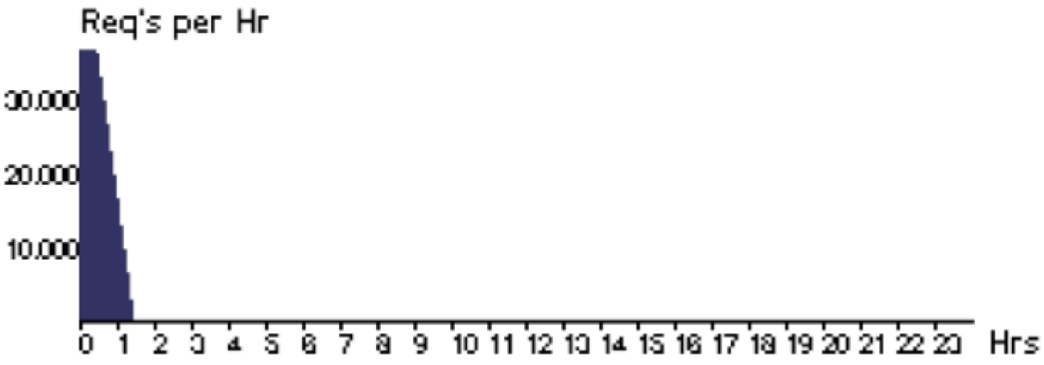
Step 12: Note down the average, maximum and minimum response time for the user bases & datacenter processing times.

Step 13: Observe & note the corresponding graphs.



Output of (i):

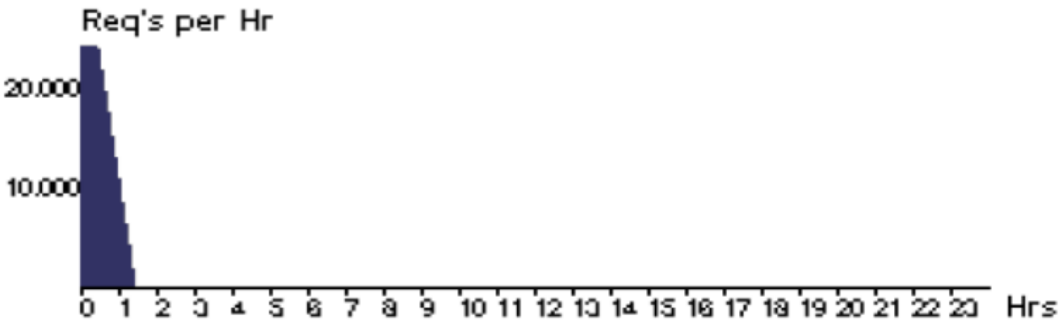
	Avg (ms)	Min (ms)	Max (ms)
Overall response time:	319.17	234.64	450.60
Data Center processing time:	19.52	0.07	124.06



Output of (ii)

	Avg (ms)	Min (ms)	Max (ms)
Overall response time:	300.59	232.81	375.33
Data Center processing time:	0.55	0.04	0.86

DC1



Analysis of the experimentation:

The response times & datacenter processing time in (i) is more than (ii) as we have more number of user bases in (i).

Outcome:

We learnt that the user base, which is collection of users, plays a vital role in generating requests. As the number of user bases increase the time taken to process the user request also increases. The datacenter which is used to process the user requests generated by the user base, needs more processing time. Hence the response time of the datacenter also increases.

10. In given cloud simulator to host a simple web application on cloud with the configurations given below

i. Two datacenters with 50 VM’s each with 1024 Mb memory and processor speed as 100 MIPS. Set the number of user bases to 6.

Conduct experimentation by changing different service broker policies to analyze the average, minimum and maximum response time and total datacenter processing time. Also, analyze the total cost required to run the application.

Procedure for (i):

Step1: Start the Cloud Analyst simulator by clicking on run.bat file.

Step 2: Once the simulator starts navigate to Configure Simulation tab.

Step 3: In Main Configuration tab, add six user bases by clicking add new button in the User Bases table.

Step 4: Navigate to Data Center Configuration tab and click on the Add New to add second datacenter.

Step5: Navigate back to Main Configuration tab, in the Application Deployment Configuration table click on Add New to configure the datacenter.

Step 6: Configure the datacenters DC1 & DC2 with 50 VM’s each and Memory of 1024 Mb each.

Step 7: Configure the Service Broker Policy to Closest Datacenter for first run. This is the default policy.

Step 8: Click on Save Configuration to save the simulation by naming it.

Step 9: Click Done.

Step 11: Click on Run Simulation to run the experiment.

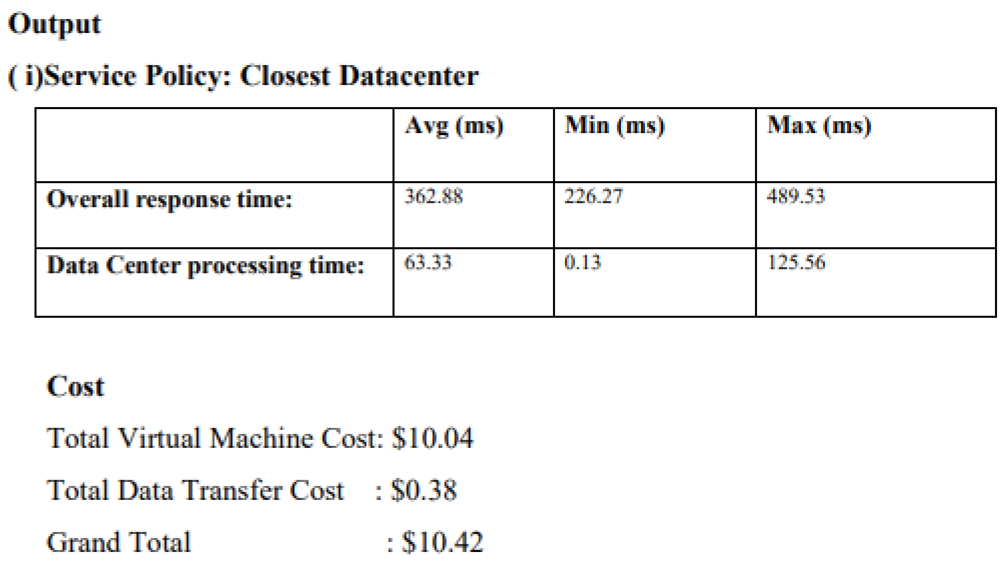
Step 12: Wait for the results to be loaded.

Step 13: Note down the average, maximum and minimum response time for the user bases & datacenter processing times.

Step 14: Note down the cost computed to run the simulation.

Note:

- for second run all the above mentioned steps remain the same except Step 8, where you have to change the service broker policy to Optimize Response time
- for third run all the above mentioned steps remain the same except Step 8, where you have to change the service broker policy to Reconfigure Dynamically with Load.



Output

(i)Service Policy: Closest Datacenter

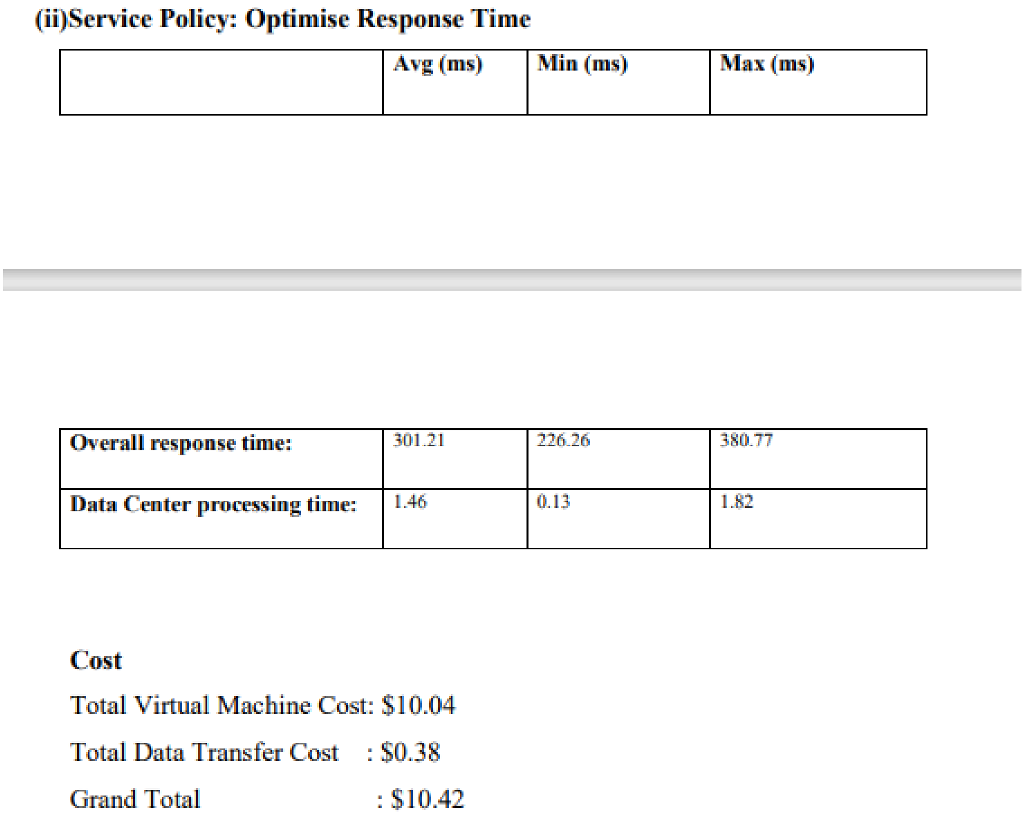
	Avg (ms)	Min (ms)	Max (ms)
Overall response time:	362.88	226.27	489.53
Data Center processing time:	63.33	0.13	125.56

Cost

Total Virtual Machine Cost: $10.04

Total Data Transfer Cost : $0.38

Grand Total : $10.42



(ii)Service Policy: Optimise Response Time

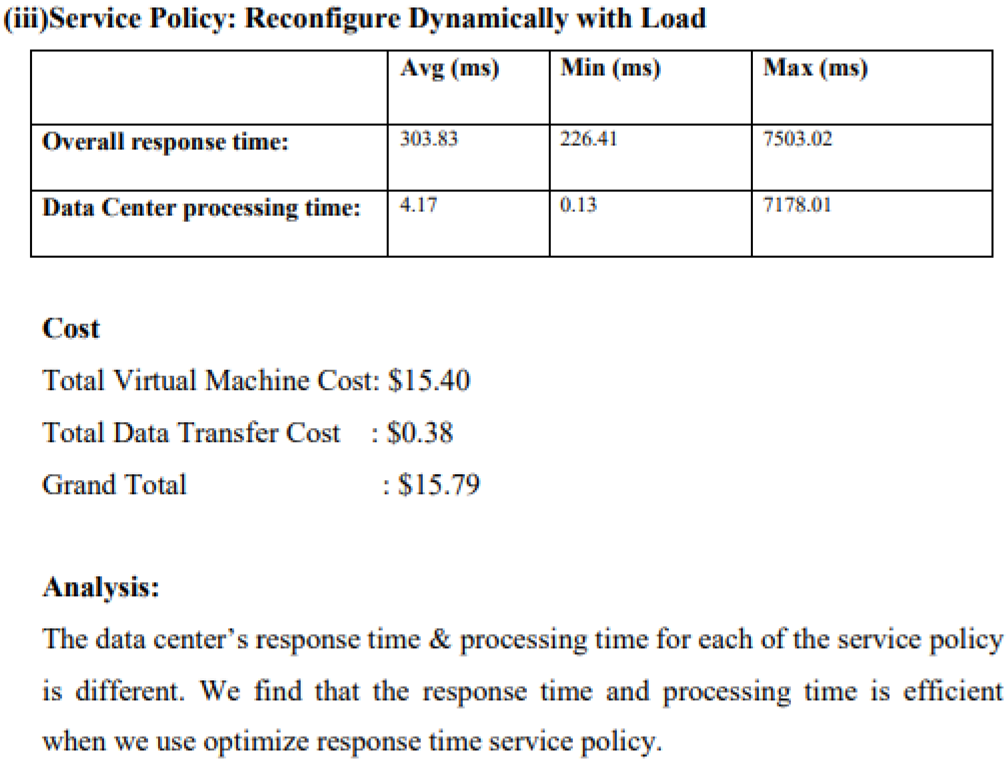
	Avg (ms)	Min (ms)	Max (ms)
Overall response time:	301.21	226.26	380.77
Data Center processing time:	1.46	0.13	1.82

Cost

Total Virtual Machine Cost: $10.04

Total Data Transfer Cost : $0.38

Grand Total : $10.42



(iii)Service Policy: Reconfigure Dynamically with Load

	Avg (ms)	Min (ms)	Max (ms)
Overall response time:	303.83	226.41	7503.02
Data Center processing time:	4.17	0.13	7178.01

Cost

Total Virtual Machine Cost: $15.40

Total Data Transfer Cost : $0.38

Grand Total : $15.79

Analysis:

The data center's response time & processing time for each of the service policy is different. We find that the response time and processing time is efficient when we use optimize response time service policy.

Analysis:

The data center's response time & processing time for each of the service policy is different. We find that the response time and processing time is efficient when we use optimize response time service policy.

Outcome of experimentation:

In this experiment we noted the response time and processing time of two datacenters by varying the service policies. It was noted that when we use optimize response time service policy, the response time and processing time of the datacenter is better compared to the other two service policies.

VIVA QUESTIONS AND ANSWERS

1. Define Cloud computing with example.

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

2. What is the working principle of Cloud Computing?

The cloud is a collection of computers and servers that are publicly accessible via the Internet. This hardware is typically owned and operated by a third party on a consolidated basis in one or more data center locations. The machines can run any combination of operating systems.

3. What are the advantages and disadvantages of Cloud Computing?

Advantages

Lower-Cost Computers for Users

Improved Performance

Lower IT Infrastructure Costs

Fewer Maintenance Issues

Lower Software Costs

Instant Software Updates

Increased Computing Power

Unlimited Storage Capacity

Increased Data Safety

Improved Compatibility between Operating Systems

Improved Document Format Compatibility

Easier Group Collaboration

Universal Access to Documents

Latest Version Availability

Removes the Tether to Specific Devices

Disadvantages

Requires a Constant Internet Connection

Doesn’t Work Well with Low-Speed Connections

Can Be Slow

Features Might Be Limited

Stored Data Might Not Be Secure

If the Cloud Loses Your Data, You’re screwed

4. What is distributed system?

A distributed system is a software system in which components located on networked computers communicate and coordinate their actions by passing messages. The components interact with each other in order to achieve a common goal.

Three significant characteristics of distributed systems are:

- Concurrency of components
- Lack of a global clock
- Independent failure of components

5. What is grid computing?

Grid Computing enables virtual organizations to share geographically distributed resources as they pursue common goals, assuming the absence of central location, central control, omniscience, and an existing trust relationship.

6. What are the business areas needs in Grid computing?

- Life Sciences
- Financial services
- Higher Education
- Engineering Services
- Government
- Collaborative games

7. List out the Grid Applications:

- Application partitioning that involves breaking the problem into discrete pieces
- Discovery and scheduling of tasks and workflow
- Data communications distributing the problem data where and when it is required
- Provisioning and distributing application codes to specific system nodes
- Autonomic features such as self-configuration, self-optimization, self-recovery and self-management.

8. List some grid computing toolkits and frameworks?

- Globus Toolkit Globus Resource Allocation Manager (GRAM)
- Grid Security Infrastructure (GSI)

- Information Services
- Legion, Condor and Condor-G
- NIMROD, UNICORE, NMI.

9. What are Desktop Grids?

These are grids that leverage the compute resources of desktop computers. Because of the true (but unfortunate) ubiquity of Microsoft® Windows® operating system in corporations, desktop grids are assumed to apply to the Windows environment. The Mac OS™ environment is supported by a limited number of vendors.

10. What are Server Grids?

- Some corporations, while adopting Grid Computing, keep it limited to server resources that are within the purview of the IT department.
- Special servers, in some cases, are bought solely for the purpose of creating an internal “utility grid” with resources made available to various departments.
- No desktops are included in server grids. These usually run some flavor of the Unix/Linux operating system.

11. Define Open nebula.

OpenNebula is an open source management tool that helps virtualized data centers oversee private clouds, public clouds and hybrid clouds. ... Open Nebula is vendor neutral, as well as platform- and API-agnostic. It can use KVM, Xen or VMware hypervisors.

12. Define Eclipse.

Eclipse is an integrated development environment (IDE) used in computer programming, and is the most widely used Java IDE. It contains a base workspace and an extensible plug-in system for customizing the environment.

13. Define Net beans.

Net Beans is an open-source integrated development environment (IDE) for developing with Java, PHP, C++, and other programming languages. NetBeans is also referred to as a platform of modular components used for developing Java desktop applications.

14. Define Apache Tomcat.

Apache Tomcat (or Jakarta Tomcat or simply Tomcat) is an open source servlet container developed by the Apache Software Foundation (ASF). Tomcat implements the Java Servlet and the Java Server Pages (JSP) specifications from Sun Microsystems, and provides a "pure Java" HTTP web server environment for Java code to run."

15. What is private cloud?

The private cloud is built within the domain of an intranet owned by a single organization. Therefore, they are client owned and managed. Their access is limited to the owning clients and their partners. Their deployment was not meant to sell capacity over the Internet through publicly accessible interfaces. Private clouds give local users a flexible and agile private infrastructure to run service workloads within their administrative domains.

16. What is public cloud?

A public cloud is built over the Internet, which can be accessed by any user who has paid for the service. Public clouds are owned by service providers. They are accessed by subscription. Many companies have built public clouds, namely Google App Engine, Amazon AWS, Microsoft Azure, IBM Blue Cloud, and Salesforce Force.com. These are commercial providers that offer a publicly accessible remote interface for creating and managing VM instances within their proprietary infrastructure.

17. What is hybrid cloud?

A hybrid cloud is built with both public and private clouds, Private clouds can also support hybrid cloud model by supplementing local infrastructure with computing capacity from an external public cloud. For example, the research compute cloud (RC2) is a private cloud built by IBM.

18. What is a Community Cloud?

A community cloud in computing is a collaborative effort in which infrastructure is shared between several organizations from a specific community with common concerns (security, compliance, jurisdiction, etc.), whether managed internally or by a third-party and hosted internally or externally. This is controlled and used by a group of organizations that have shared interest. The costs are spread over fewer users than a public cloud (but more than a private cloud

19. Define IaaS?

The IaaS layer offers storage and infrastructure resources that is needed to deliver the Cloud services. It only comprises of the infrastructure or physical resource. Top IaaS Cloud Computing Companies:

Amazon (EC2), Rackspace, Go Grid, Microsoft, Terre mark and Google.

20. Define PaaS?

PaaS provides the combination of both, infrastructure and application. Hence, organisations using PaaS don’t have to worry for infrastructure nor for services. Top PaaS Cloud Computing Companies: Salesforce.com, Google, Concur Technologies, Ariba, Unisys and Cisco..

21. Define SaaS?

In the SaaS layer, the Cloud service provider hosts the software upon their servers. It can be defined as a in model in which applications and softwares are hosted upon the server and made available to customers over a network. Top SaaS Cloud Computing Companies: Amazon Web Services, AppScale, CA Technologies, Engine Yard, Salesforce and Windows Azure.

22. What is meant by virtualization?

Virtualizationisacomputerarchitecturetechnologybywhichmultiplevirtualmachines (VMs) are multiple exuding the same Hardwar machine. The ideaof VMs can be dated back to the 1960s.The purpose of a VM is to enhance resource sharing by many users and improve computer performance in terms of resource utilization and application flexibility.

23. What are the implementation levels of virtualization?

The virtualization types are following

1. OS-level virtualization
2. ISA level virtualization
3. User-ApplicationLevel virtualization
4. Hardware level virtualization
5. Library level virtualization

24. List the requirements of VMM?

There are three requirements for a VMM.

First, a VMM should provide an environment for programs which is essentially identical to the original machine. Second, programs run in this environment should show, at worst, only minor decreases in speed. Third, a VMM should be in complete control of the system resources.

25. Explain Host OS and Guest OS?

A comparison of the differences between a host system, a guest system, and a virtual machine within a virtual infrastructure. A host system (host operating system) would be the primary & first installed operating system. If you are using a bare metal Virtualization platform like Hyper-V or ESX, there really isn't a host operating system besides the Hypervisor. If you are using a Type-2 Hypervisor like VMware Server or

Virtual Server, the host operating system is whatever operating system those applications are installed into. A guest system (guest operating system) is a virtual guest or virtual machine (VM) that is installed under the host operating system. The guests are the VMs that you run in your virtualization platform.

26. Write the steps for live VM migration?

The five steps for live VM migration is

Stage 0: Pre-Migration

Active VM on Host A

Alternate physical host may be preselected for migration

Block devices mirrored and free resources maintained

Stage 1: Reservation

Initialize a container on the target host

Stage 2: Iterative pre-copy

Enable shadow paging

Copy dirty pages in successive rounds.

Stage 3: Stop and copy

Suspend VM on host A

Generate ARP to redirect traffic to Host B

Synchronize all remaining VM state to Host B

Stage 4: Commitment

VM state on Host A is released

Stage 5: Activation

VM starts on Host B

Connects to local devices

Resumes normal operation

27. Define Globus Toolkit: Grid Computing Middleware

□ Globus is open source grid software that addresses the most challenging problmes in distributed resources sharing.

□ The Globus Toolkit includes software services and libraries for distributed security, resource management, monitoring and discovery, and data management.

28. Define Blocks in HDFS

- a disk has a block size, which is the minimum amount of data that it can read or write. File systems for a single disk build on this by dealing with data in blocks, which are an integral multiple of the disk block size. Filesystem blocks are typically a few kilobytes in size, while disk blocks are normally 512 bytes. This is generally transparent to the filesystem user who is simply reading or writing a file—of whatever length.

29. Define Name nodes and Data nodes

- An HDFS cluster has two types of node operating in a master-worker pattern:
- a namenode (the master) and
- a number of datanodes(workers).
- The namenode manages the filesystem namespace. It maintains the filesystem tree and the metadata for all the files and directories in the tree. This information is stored persistently on the local disk in the form of two files: the namespace image and the edit log.
- The namenode also knows the datanodes on which all the blocks for a given file are located, however, it does not store block locations persistently, since this information is reconstructed from datanodes when the system starts.

30. Define HADOOP.

Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.

31. Define HDFS.

Hadoop Distributed File System (HDFS) is a Java-based file system that provides scalable and reliable data storage that is designed to span large clusters of commodity servers. HDFS, MapReduce, and YARN form the core of Apache™ Hadoop®.

32. Write about HADOOP.

Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Cutting, who was working at Yahoo! At the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project.

33. Definition of Grid Portal:

A Grid Portal provides an efficient infrastructure to put Grid-empowered applications on corporate Intranet/Internet.

34. Define GAE.

Google App Engine (often referred to as GAE or simply App Engine) is a Platform as a Service and cloud computing platform for developing and hosting web applications in Google-managed data centers. Applications are sandboxed and run across multiple servers. App Engine offers automatic scaling for web applications—as the number of requests increases for an application, App Engine automatically allocates more resources for the web application to handle the additional demand.

35. What is Cloudsim?

CloudSim is a simulation toolkit that supports the modeling and simulation of the core functionality of cloud, like job/task queue, processing of events, creation of cloud entities(datacenter, datacenter brokers, etc), communication between different entities, implementation of broker policies, etc. This toolkit allows to:

- Test application services in a repeatable and controllable environment.
- Tune the system bottlenecks before deploying apps in an actual cloud.
- Experiment with different workload mix and resource performance scenarios on simulated infrastructure for developing and testing adaptive application provisioning techniques.

36. Core features of CloudSim are:

- The Support of modeling and simulation of large scale computing environment as federated cloud data centers, virtualized server hosts, with customizable policies for provisioning host resources to virtual machines and energy-aware computational resources
- It is a self-contained platform for modeling cloud's service brokers, provisioning, and allocation policies.
- It supports the simulation of network connections among simulated system elements.
- Support for simulation of federated cloud environment that inter-networks resources from both private and public domains.
- Availability of a virtualization engine that aids in the creation and management of multiple independent and co-hosted virtual services on a data center node.
- Flexibility to switch between spaces shared and time shared allocation of processing cores to virtualized services.

37. Uses of Cloudsim.

- Load Balancing of resources and tasks
- Task scheduling and its migrations
- optimizing the Virtual machine allocation and placement policies
- Energy-aware Consolidations or Migrations of virtual machines
- optimizing schemes for Network latencies for various cloud scenarios

38. Define OpenStack.

OpenStack is a cloud operating system that controls large pools of compute, storage, and networking resources throughout a datacenter, all managed and provisioned through APIs with common authentication mechanisms. A dashboard is also available, giving administrators control while empowering their users to provision resources through a web interface.

39. Define Trystack.

TryStack is a great way to take OpenStack for a spin without having to commit to a full deployment. This free service lets you test what the cloud can do for you, offering networking, storage and compute instances, without having to go all in with your own hardware.

It’s a labor of love spearheaded by three Red Hat OpenStack experts Will Foster, Kambiz Aghaiepour and Dan Radez. TryStack’s set-up must bear the load of anyone who wants to use it, but instead of an equally boundless budget and paid staff, it was originally powered by donated equipment and volunteers from Cisco, Dell, Equinix, NetApp, Rackspace and Red Hat who pulled together for this OpenStack Foundation project.

40. Define Hadoop.

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.