

Big Data Analytics and Visualization Lab

Subject Code: MCAL31

A Practical Journal Submitted in Fulfilment of the Degree of

MASTER

In

COMPUTER APPLICATION

Year 2023-2024

By

Mr. Ravishankar Jaiswal

(Seat Number:- 5000058)

(Application No:129211)

Under the Guidance of

Ms Bharti Mam



Institute of Distance and Open Learning

Vidya Nagari, Kalina, Santacruz East – 400098.

University of Mumbai

PCP Center

[Satish Pradhan Dyanasadhana College, Thane]



Institute of Distance and Open Learning
Vidyanagari, Kalina, Santacruz (E) -400098

CERTIFICATE

This to certify that, **Ravishankar Jaiswal** appearing **Master in Computer Application (Semester III) Seat Number: 5000058** has satisfactory completed the prescribed Practical of **MCAL31- Big Data Analytics and Visualization Lab** as laid down by the University of Mumbai for the academic year **2023- 24**

Teacher in charge

Examiners

Coordinator

IDOL, MCA

University of Mumbai

Date: -

Place: -

Index

| Practical No | Aim | Page No. | Sign. |
|-------------------------|--|---------------------|--------------|
| 1. | Install, configure, and run Hadoop and HDFS ad explore HDFS. | 1 - 21 | |
| 2. | Word Count in Hadoop | 22-29 | |
| 3. | MongoDB installation & CRUD Operation | 30- 43 | |
| 4 | Data Visualization | 44-56 | |
| 5 | Data Visualization using tableau | 57-62 | |

Practical No: 01

Aim: Install, configure, and run Hadoop and HDFS and explore HDFS.

Step 1: Download and install VirtualBox

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

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

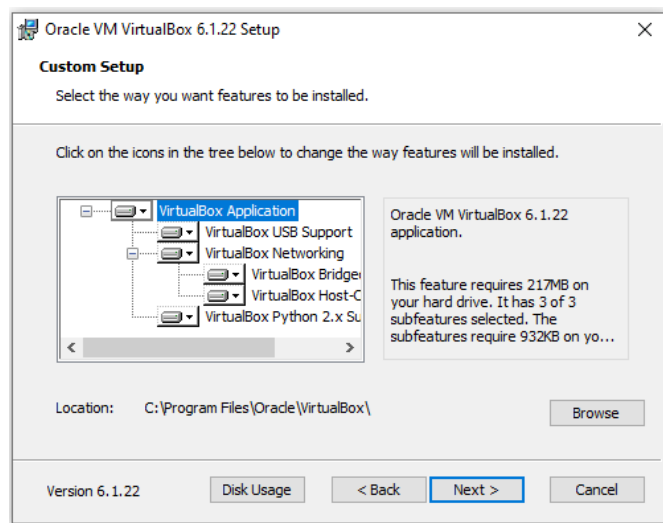


You will get VirtualBox-6.1.22-144080-Win.exe file downloaded.
Double click and run it. Click on next.

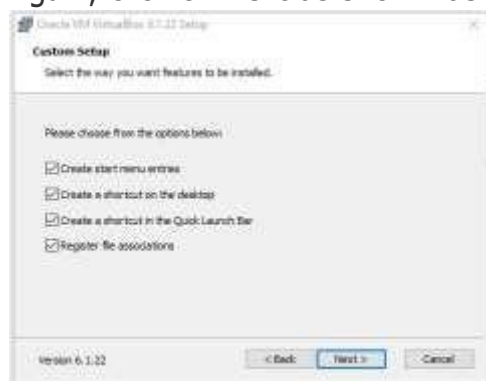


Click on 'next' without changing the default folder as shown below:

Big Data Analytics and Visualization Lab



Again, click on next as shown below:

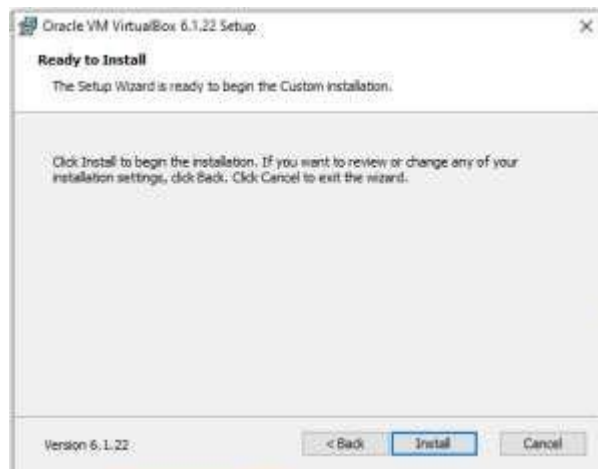


Finally, click on 'Yes'.

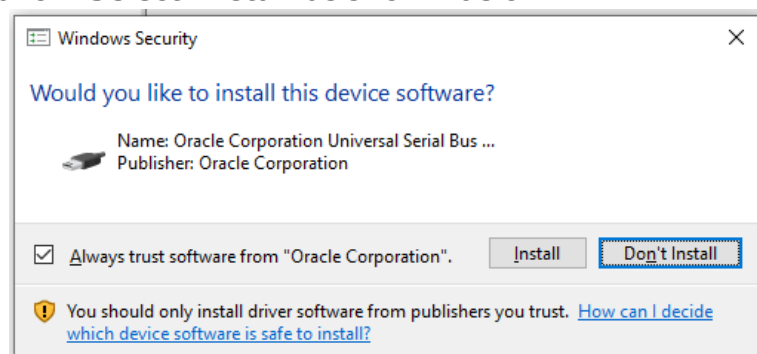


Click on 'Install'.

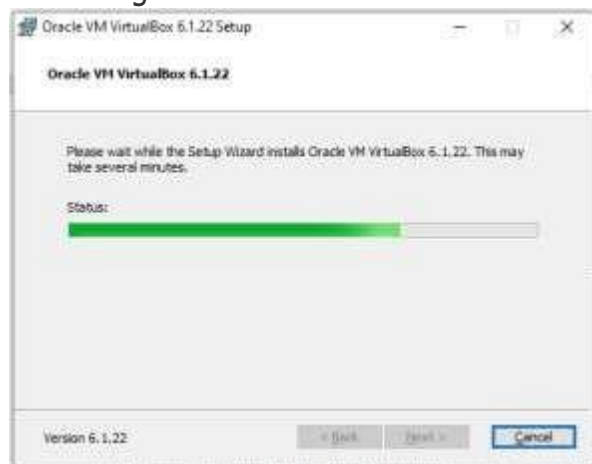
Big Data Analytics and Visualization Lab



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



You will get the screen as shown below:



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

Big Data Analytics and Visualization Lab



You will get the following screen:



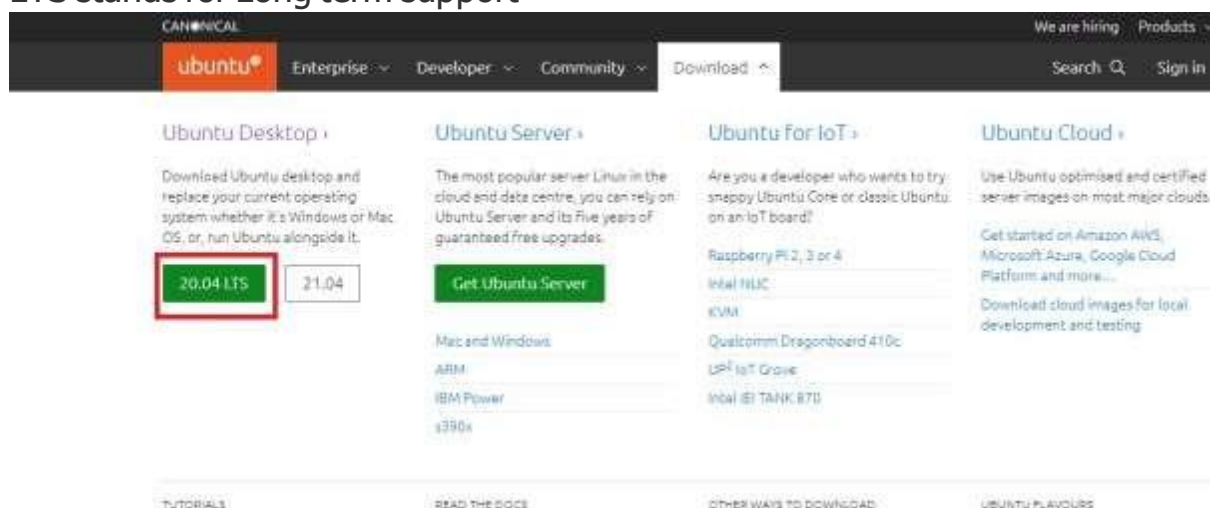
Step 2: download Ubuntu

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

Browse ubuntu.com

Click on download and 20.04 LTS as shown below:

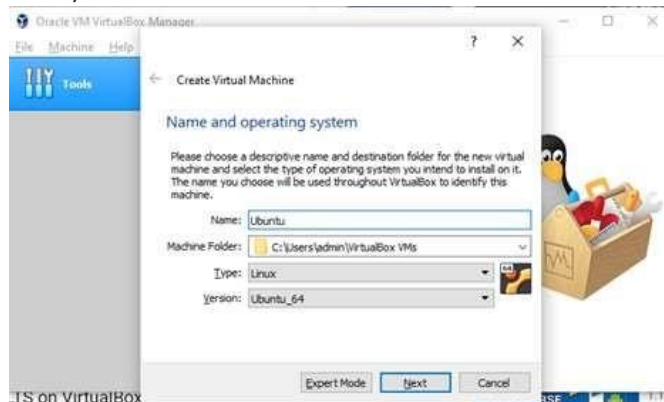
LTS stands for Long term support



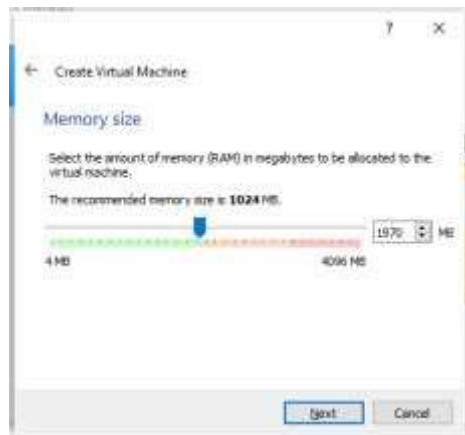
Big Data Analytics and Visualization Lab

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

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

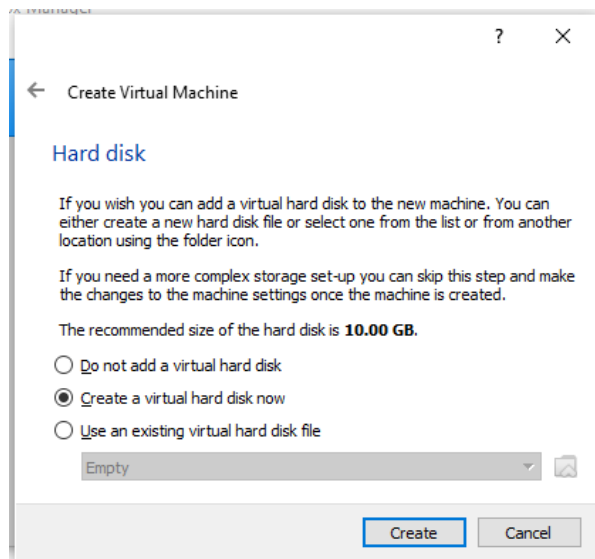


Click on 'Next'.



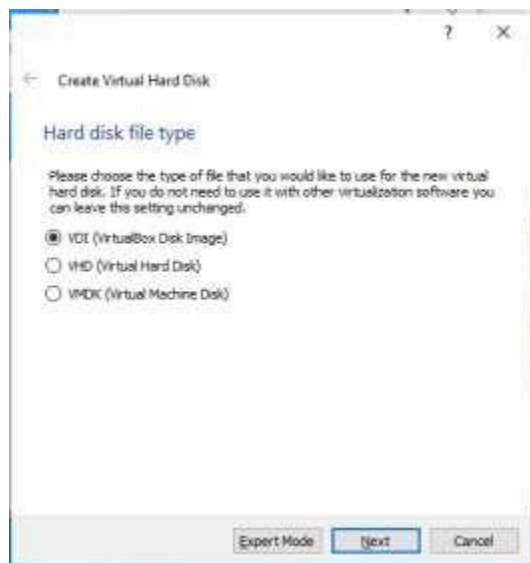
Here, you allow memory size up to green indicator (1970 MB).

Click on 'Next'.



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

Big Data Analytics and Visualization Lab

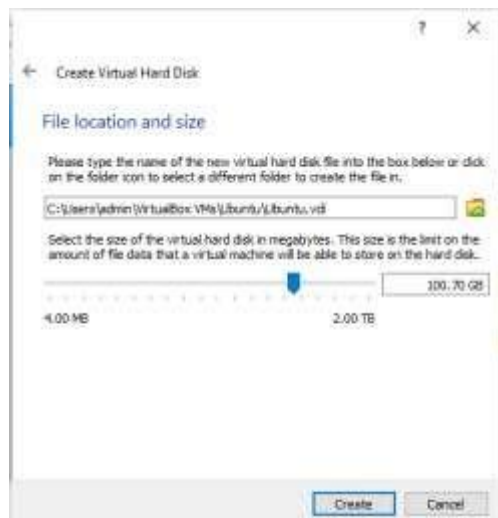


Click on 'Next', keeping the selection as it is (on VDI).'



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

Big Data Analytics and Visualization Lab



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

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

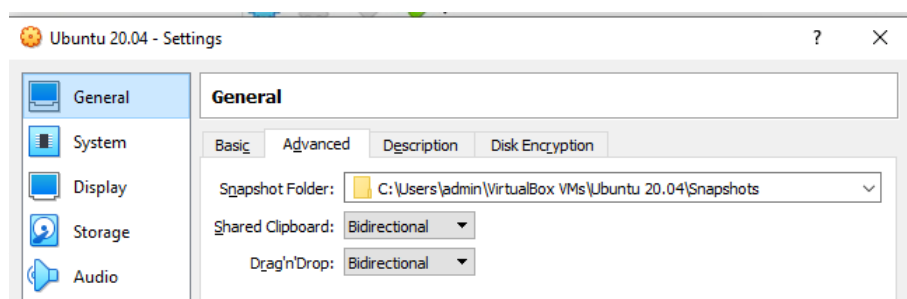


Select 'settings'

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

You may change the name from Ubuntu to Ubuntu 20.04

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



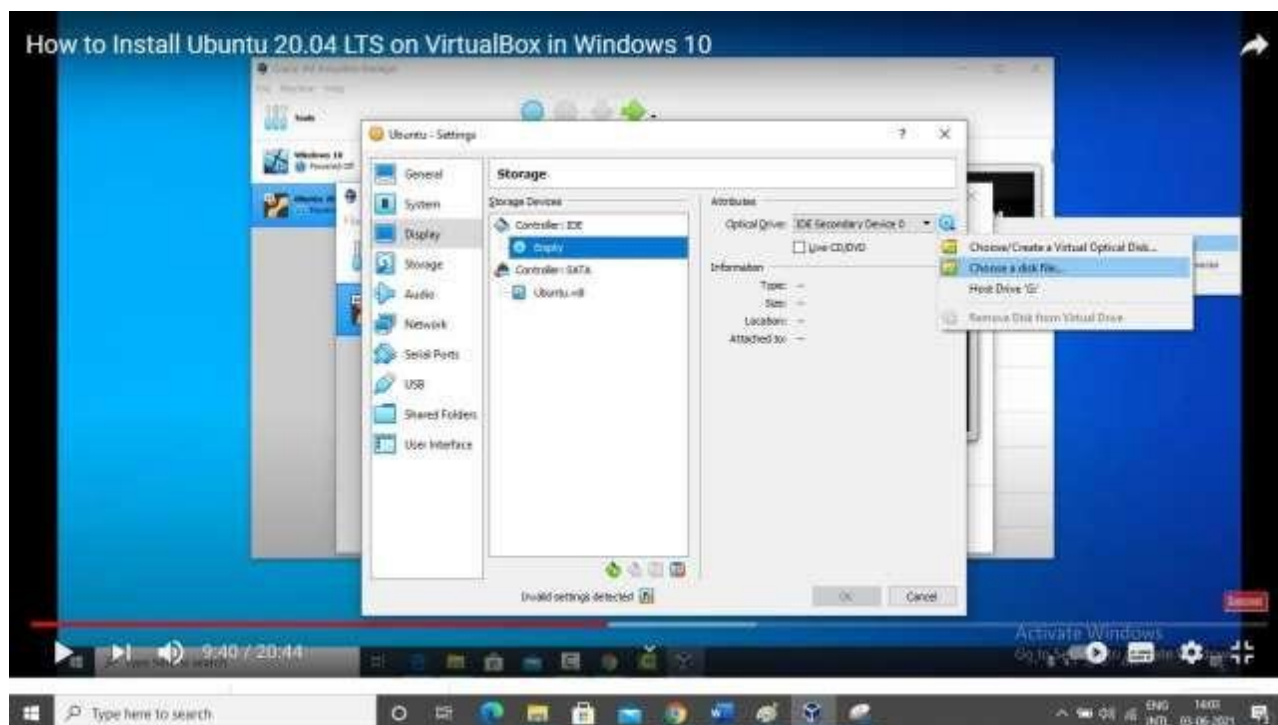
Big Data Analytics and Visualization Lab

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



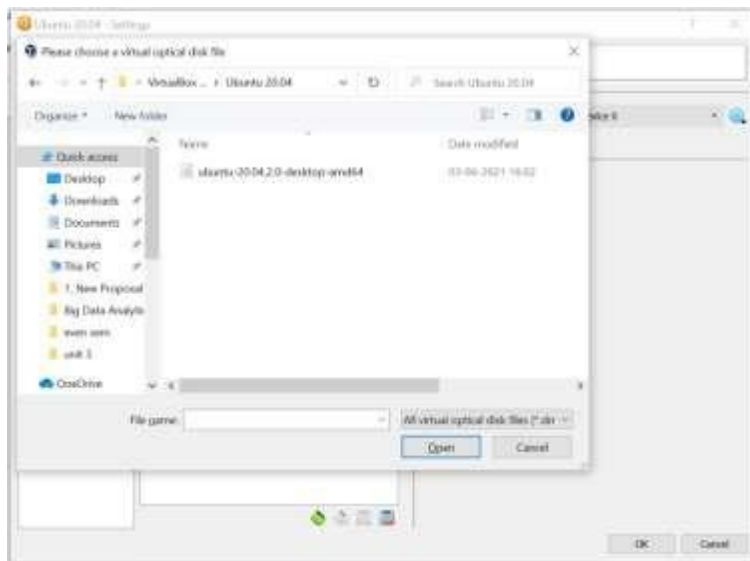
Cut and paste your ubuntu .iso file from current folder to

C:\Users\ADMIN\VirtualBox VMs\Ubuntu 20.04 folder.
Click on 'Storage' and click on 'Empty' followed by 'Choose a disk file' as shown below:



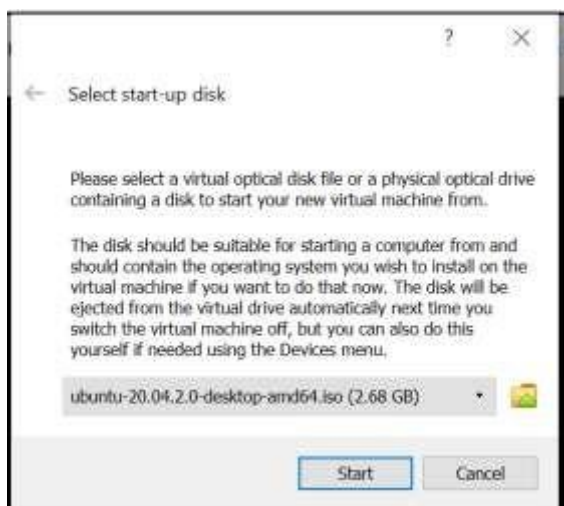
Browse the folder where you have selected ubuntu iso file.

Big Data Analytics and Visualization Lab



Click on Ubuntu....iso file and click on open and then click on ok.

Click on Ubuntu -> start button.

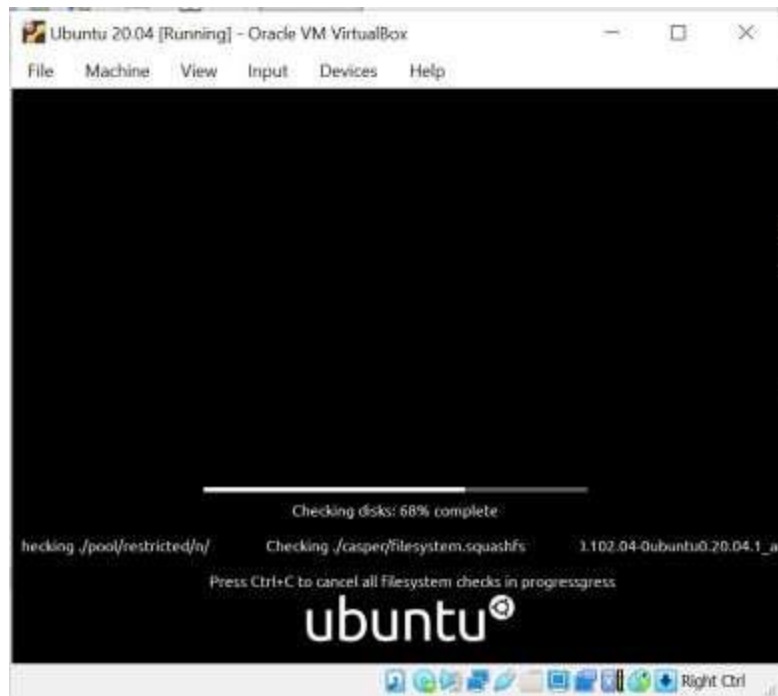


Again, click on 'Start' button. It will show you the following screen.



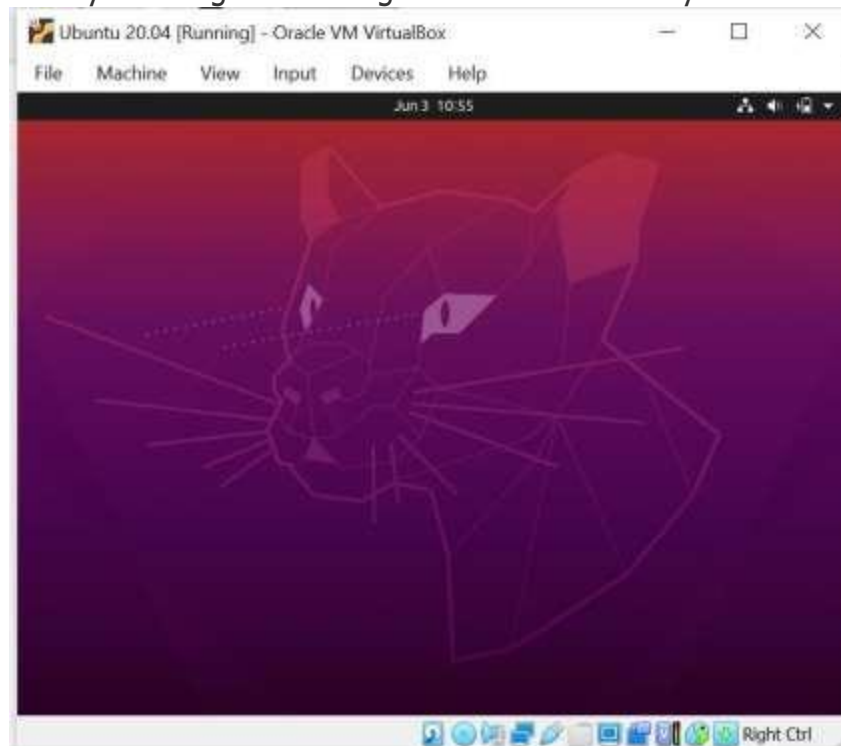
Big Data Analytics and Visualization Lab

And simultaneously one more screen as follows:



Keep on closing all warnings.

Next you will get following screen automatically.



Big Data Analytics and Visualization Lab

Select language -> English and click on 'Install Ubuntu'.in 'Keyboard Layout' screen,
select 'English US. Click on 'Continue'. Click on 'Continue'.

(if you will select 'English UK', then some key will be changed as follows:

****Note:**

Some Keys for Ubuntu under UK keyboard layout

“ -> @

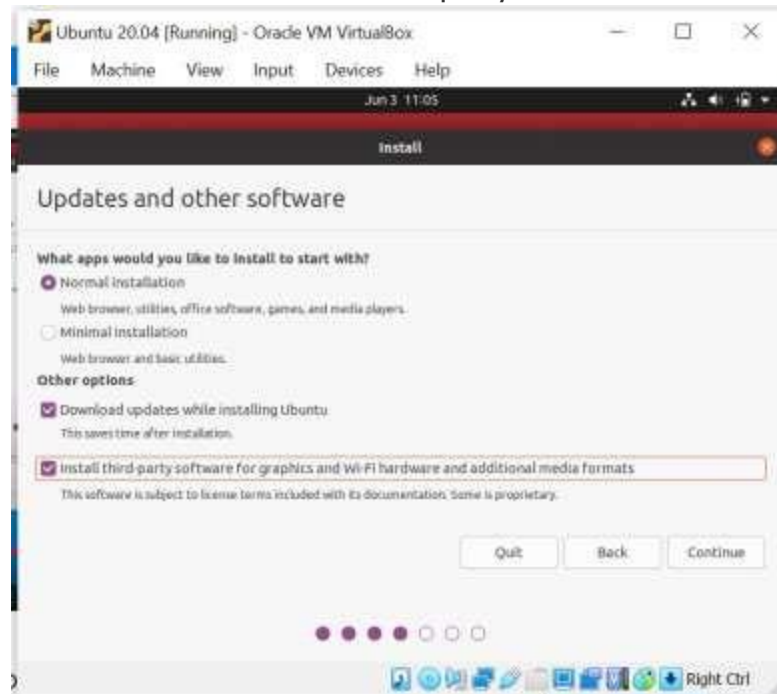
@ -> “

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

~ -> pipe

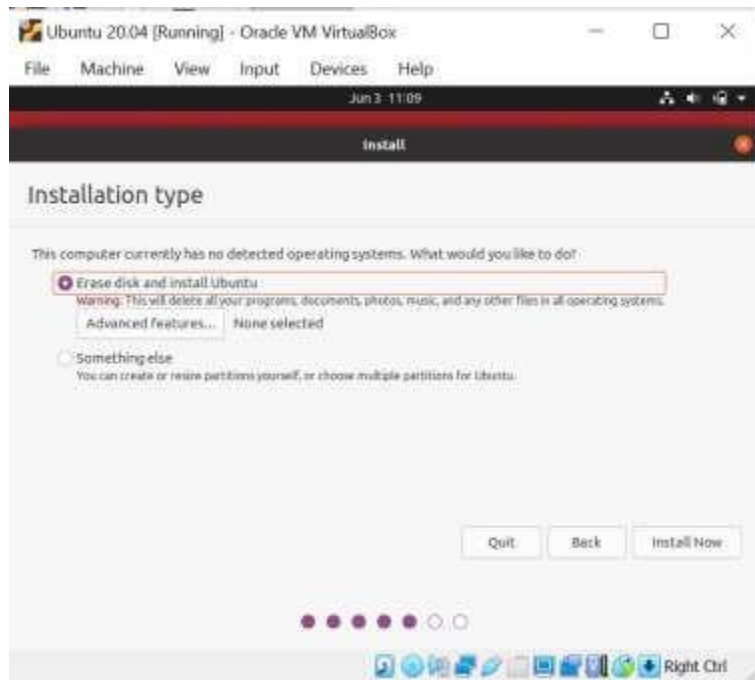
)

Select the checkbox for third party software as shown below:

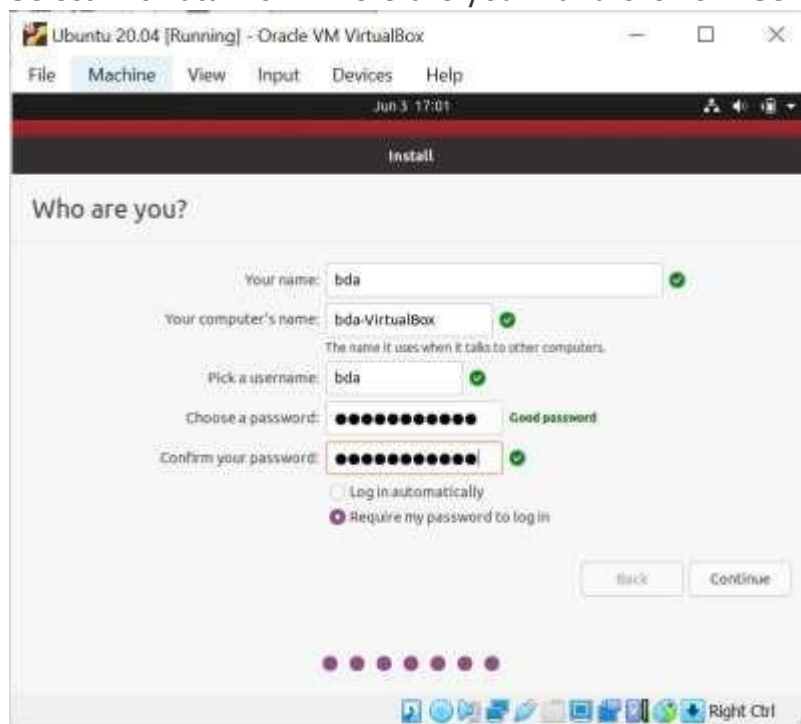


Click on 'continue'.

Big Data Analytics and Visualization Lab



Select Erase disk and Install Ubuntu and click on '**Install Now**'. Click on '**Continue**' on the next screen. Select "Kolkata" for "where are you?" and click on '**Continue**'.



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

Big Data Analytics and Visualization Lab



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

Step 3 Install Hadoop

Login to ubuntu

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

** please refer to note - Some Keys for Ubuntu under UK keyboard layout – at the

end. Search for Ubuntu terminal on search bar, after login done.

Apply following commands from ubuntu terminal

Prerequisite

```
bda@bda-VirtualBox:~$ sudo apt update
```

Name: Yadav Uday Sagindra

Seat Number: 5000088

Big Data Analytics and Visualization Lab

```
Ign:1 cdrom://Ubuntu 20.04.2.0 LTS _Focal Fossa_ - Release amd64 (20210209.1)
focalInRelease
Hit:2 cdrom://Ubuntu 20.04.2.0 LTS _Focal Fossa_ - Release amd64 (20210209.1)
focalRelease
Hit:4 http://archive.ubuntu.com/ubuntu focal InRelease
Hit:5 http://archive.ubuntu.com/ubuntu focal-updates
InRelease Hit:6 http://security.ubuntu.com/ubuntu focal-
security InReleaseReading package lists... Done
Building dependency tree
Reading state information...
Done
291 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

bda@bda-VirtualBox:~\$ sudo apt install default-jdk

```
Reading package lists...
DoneBuilding
dependency tree
:
Setting up default-jdk (2:1.11-72) ...
Setting up libxt-dev:amd64 (1:1.1.5-1)
...
```

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

```
openjdk version "11.0.11" 2021-04-20
OpenJDK Runtime Environment (build 11.0.11+9-Ubuntu-0ubuntu2.20.04)
OpenJDK 64-Bit Server VM (build 11.0.11+9-Ubuntu-0ubuntu2.20.04, mixed mode, sharing)
```

open ssh server

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

```
Reading package lists...
DoneBuilding
dependency tree
:
Processing triggers for ufw (0.36-6) ...
```

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

```
Adding user `hdoop' ...
Adding new group `hdoop' (1000) ...
Adding new user `hdoop' (1000) with group
`hdoop' ...Creating home directory `/home/hdoop'
...
Copying files from
```

Big Data Analytics and Visualization Lab

```
`/etc/skel' ...New password:
hadoop
Retype new password:
passwd: password updated
successfully Changing the user
information for hadoop
Enter the new value, or press ENTER for the
  defaultFull Name []:
    Room Number []:
    Work Phone []:
    Home Phone []:
    Other []:
Is the information correct? [Y/n] y
```

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

Password: hadoop

hadoop@bda-VirtualBox:~\$ ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa

```
Generating public/private rsa key pair.
Created directory '/home/hadoop/.ssh'.
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in
/home/hadoop/.ssh/id_rsa.pubThe key fingerprint is:
SHA256:EDxiHTL1r3LUCdKFWc0moPHUh1D8tU6Y0b2rnXuWUtQ
hadoop@bda-VirtualBox
The key's randomart image is:
```

```
+---[RSA 3072]---+
| o+=.+X++ . . |
| oo+Oo.= * + .|
| . .+. = * E.|
|   o + . = o. |
|   S + = .|
|   . . . + .|
|   . o.....|
|   o   .. o|
|           .+.|
+----[SHA256]----+
```

hadoop@bda-VirtualBox:~\$ cat ~/.ssh/id_rsa.pub >>

~/.ssh/authorized_keys**hadoop@bda-VirtualBox:~\$ chmod 0600**

~/.ssh/authorized_keys

hadoop@bda-VirtualBox:~\$ ssh localhost

Big Data Analytics and Visualization Lab

The authenticity of host 'localhost (127.0.0.1)' can't be established.

ECDSA key fingerprint is

SHA256:4TE4DDAv14vhARPWjZcW3C5UM3X94B7wUudPrT+Z
mF0.

Are you sure you want to continue connecting (yes/no/[fingerprint])? **yes**

:

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.

Downloading Hadoop

hadoop@bda-VirtualBox:~\$ wget

**https://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop-
3.3.1.tar.gz**

--2021-06-14 08:52:00--

https://downloads.apache.org/hadoop/common/hadoop-3.3.1/hadoop-
3.3.1.tar.gz

Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 135.181.209.10,
135.181.214.104, ...

Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.

HTTP request sent, awaiting response... 200 OK

Length: 359196911 (343M) [application/x-gzip]

Saving to: 'hadoop-3.3.1.tar.gz'

hadoop-3.3.1.tar.gz 100%[=====>] 342.56M 15.4MB/s in 33s

2021-06-14 08:52:34 (10.2 MB/s) - 'hadoop-3.3.1.tar.gz' saved

[359196911/359196911]

hadoop@bda-VirtualBox:~\$ ls

hadoop-3.3.1.tar.gz

hadoop@bda-VirtualBox:~\$ tar xzf hadoop-

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

hadoop-3.3.1 hadoop-3.3.1.tar.gz

Editing 6 important files for creating a single cluster

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

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

Adding user 'hadoop' to group

`sudo' ...Adding user hadoop to

group sudo Done.

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

Big Data Analytics and Visualization Lab

1.

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

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

```
#Hadoop Related Options
export HADOOP_HOME=/home/hdoop/hadoop-
3.3.1export
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"
```

save this file as ctrl x and y. Press enter.

hdoop@bda-VirtualBox:~\$ source

~/bashrc2.

Edit hadoop-env.sh File

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

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

hdoop@bda-VirtualBox:~\$ sudo nano \$HADOOP_HOME/etc/hadoop/hadoop-env.sh

at the end of the file add the following line

export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/

save it.

3

Edit core-site.xml File

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

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

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

Big Data Analytics and Visualization Lab

hadoop@bda-VirtualBox:~\$ sudo nano \$HADOOP_HOME/etc/hadoop/core-site.xml

```
<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/hadoop/tmpdata</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:9000</value>
</property>
</configuration>
```

4

hadoop@bda-VirtualBox:~\$ sudo nano \$HADOOP_HOME/etc/hadoop/hdfs-site.xml

```
<configuration>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hadoop/dfsdata/namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hadoop/dfsdata/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
</configuration>
```

5

hadoop@bda-VirtualBox:~\$ sudo nano \$HADOOP_HOME/etc/hadoop/mapred-site.xml

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

6

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

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

Big Data Analytics and Visualization Lab

```
<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,HADOO
P_MAPRED_HOME</value>
</property>
</configuration>
```

Format HDFS NameNode

hdoop@bda-VirtualBox:~\$ hdfs namenode -format

:

xid=0 when meet shutdown.

2021-06-18 14:16:33,353 INFO namenode.NameNode: SHUTDOWN_MSG:

/*****

SHUTDOWN_MSG: Shutting down NameNode at bda-VirtualBox/127.0.1.1

*****/

Start Hadoop Cluster (services)

hdoop@bda-VirtualBox:~\$ cd Hadoop-3.3.1

hdoop@bda-VirtualBox:~/Hadoop-3.3.1\$ cd

sbin

hdoop@bda-VirtualBox:~/hadoop-3.3.1/sbin\$./start-dfs.sh

Starting namenodes on

[localhost]Starting datanodes

Starting secondary namenodes [bda-VirtualBox]

bda-VirtualBox: Warning: Permanently added 'bda-virtualbox' (ECDSA) to the list of knownhosts.

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

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

Starting

resourcemanager

Starting

nodemanagers

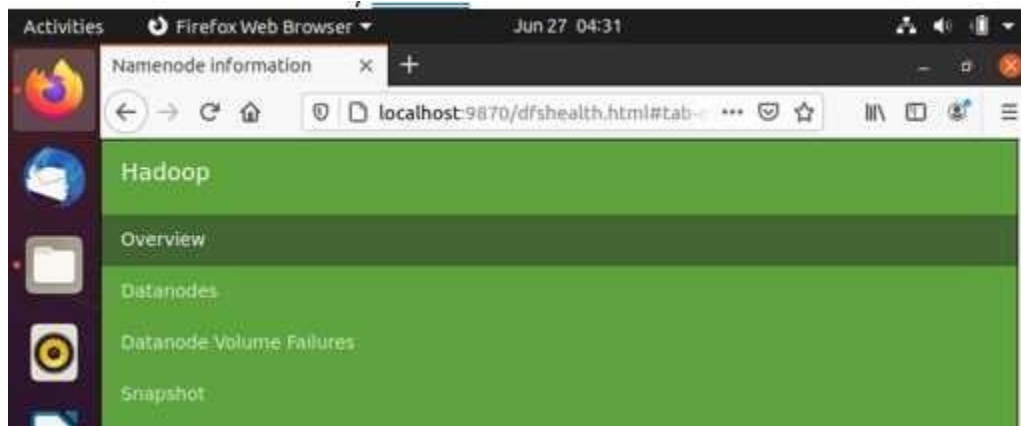
Big Data Analytics and Visualization Lab

To see all components, we use jps command:

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

```
11744 NodeManager
11616 ResourceManager
12192 Jps
11268 SecondaryNameNode
11077 DataNode
10954 NameNode
```

Browse localhost:9870 on any browser:



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

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

```
hdoop@bda-VirtualBox:~/hadoop-3.3.1/sbin$ sudo nano /home/bda/sample.txt
```

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

edit the file by adding some text and save and exit

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

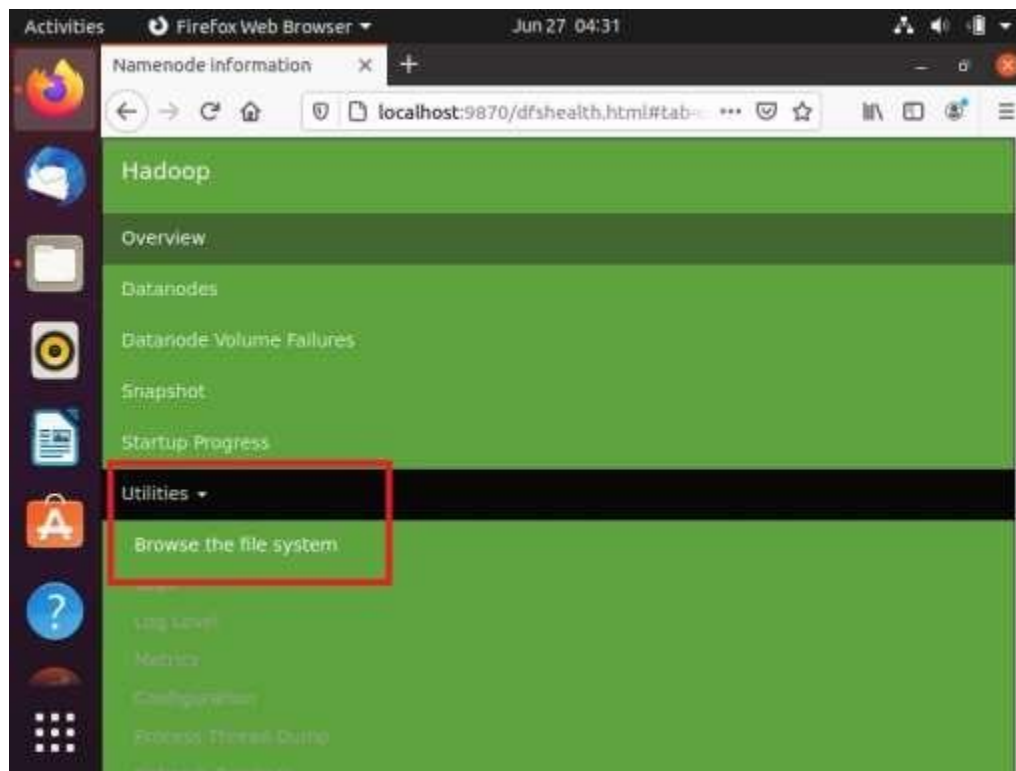
```
Desktop  Downloads  Pictures  sample.txt
Videos  Documents  Music    Public
Templates
```

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

```
/home/bda/sample.txt / 2021-06-18 14:44:24,257 WARN util.NativeCodeLoader:
Unable to load native-hadooplibrary for your platform... using builtin-java classes
where applicable
```

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

Big Data Analytics and Visualization Lab



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

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

```
Found 1 items
```

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


Practical No: 02

Aim: Wordcount in HADOOP.

Login to bda user of Ubuntu

create a folder wordcount under home folder of Ubuntu

create file WordCount.java and store in that folder:

code:

```
import
java.io.IOException;
import
java.util.StringTokenize
r;

import
org.apache.hadoop.conf.Configuration
;import org.apache.hadoop.fs.Path;
import
org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import
org.apache.hadoop.mapreduce.Job;
import
org.apache.hadoop.mapreduce.Mapper
;import
org.apache.hadoop.mapreduce.Reduce
r;
import
org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import
org.apache.hadoop.mapreduce.lib.output.FileOutputForma
t;

public class WordCount {

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

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

Big Data Analytics and Visualization Lab

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

public static class IntSumReducer extends
    Reducer<Text,IntWritable,Text,IntWritable> {private IntWritable result = new
    IntWritable();
    public void reduce(Text key, Iterable<IntWritable> values, Context context) throws
    IOException,InterruptedException
    {
        int sum = 0;
        for (IntWritable val :
            values) {sum +=
            val.get();
        }
        result.set(sum);
        context.write(key
        , result);
    }
}

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

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

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

create a folder wordcount_classes in wordcount folder.

create env variable using hdoop terminal as follows and apply subsequent commands:
bda@bda-VirtualBox:~\$ su hdoop

Password:

```
hdoop@bda-VirtualBox:/home/bda$ export  
HADOOP_CLASSPATH=$(hadoop classpath)hdoop@bda-  
VirtualBox:/home/bda$ echo $HADOOP_CLASSPATH  
/home/hdoop/hadoop-3.3.1/etc/hadoop:/home/hdoop/hadoop-  
3.3.1/share/hadoop/common/lib/*:/home/hdoop/hadoop-  
3.3.1/share/hadoop/common/*:/home/hdoop/hadoop-  
3.3.1/share/hadoop/hdfs:/home/hdoop/hadoop-  
3.3.1/share/hadoop/hdfs/lib/*:/home/hdoop/hadoop-  
3.3.1/share/hadoop/hdfs/*:/home/hdoop/hadoop-  
3.3.1/share/hadoop/mapreduce/*:/home/hdoop/hadoop-  
3.3.1/share/hadoop/yarn:/home/hdoop/hadoop-  
3.3.1/share/hadoop/yarn/lib/*:/home/hdoop/hadoop-3.3.1/share/hadoop/yarn/*
```

hdoop@bda-VirtualBox:/home/bda\$ start-dfs.sh

```
Starting namenodes on  
[localhost]Starting  
datanodes  
Starting secondary namenodes [bda-VirtualBox]  
2021-06-26 13:13:58,694 WARN util.NativeCodeLoader: Unable to load native-hadoop  
library for your platform... using builtin-java classes where applicable
```

hdoop@bda-VirtualBox:/home/bda\$ start-yarn.sh

```
Starting  
resourcemanager  
Starting  
nodemanagers  
hdoop@bda-  
VirtualBox:/home/bda$ jps  
3248 ResourceManager
```

Big Data Analytics and Visualization Lab

3779 Jps

3382 NodeManager

2631 NameNode

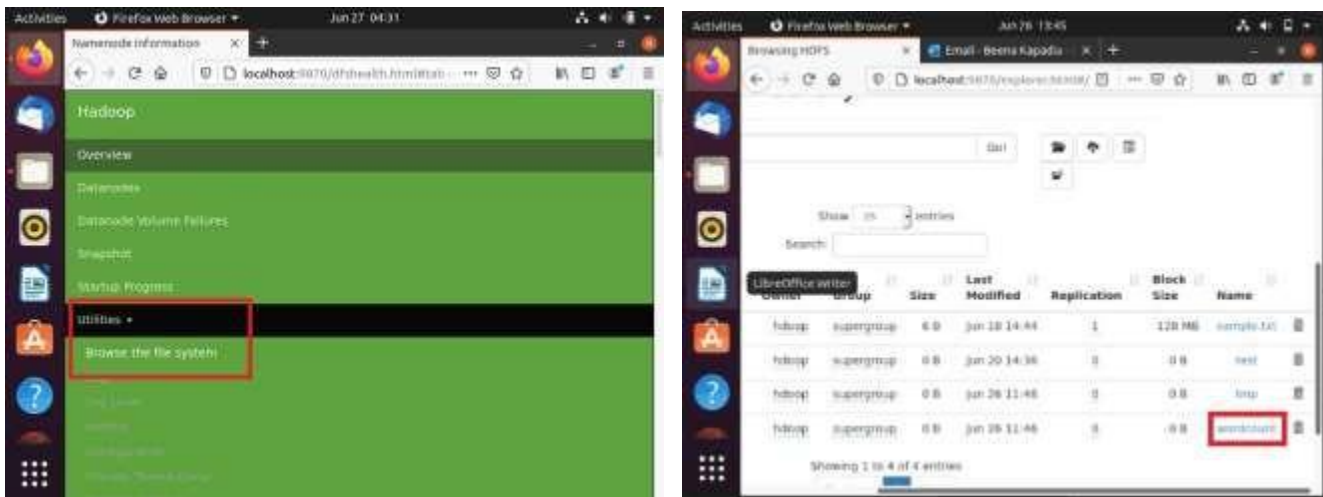
2760 DataNode

2953 SecondaryNameNode

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

(Browse localhost:9870 on any browser and click on **utility** and **select browse the file system**

you can see your folder there)

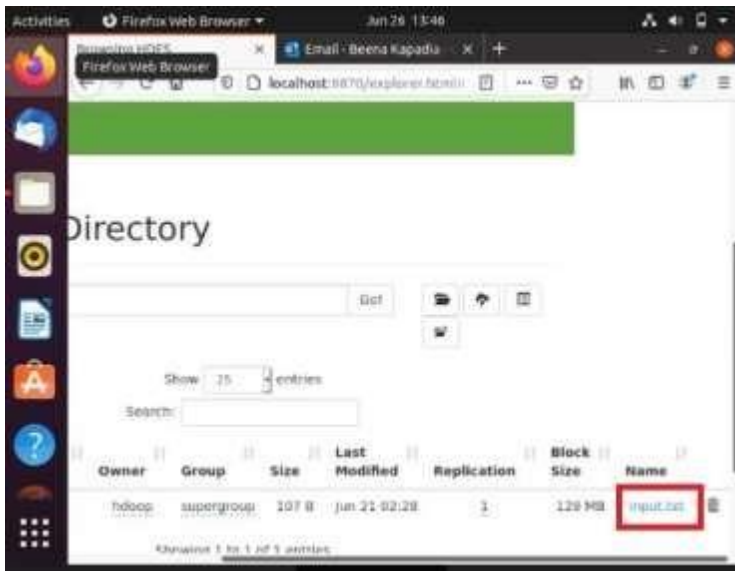


hadoop@bda-VirtualBox:/home/bda\$ hdfs dfs -mkdir

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

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

Big Data Analytics and Visualization Lab



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

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

beena

yashesh

vinod

beena

kruti

nidhi

vinod

yashesh

dhavan

kruti

nidhi

komal

jeenisha

mitish

nidhi

yashesh

hadoop@bda-

VirtualBox:/home/bda\$ su bda

password

bda@bda-VirtualBox:~\$ cd wordcount

bda@bda-VirtualBox:~/wordcount\$ javac -classpath

\${HADOOP_CLASSPATH} -d'/home/bda/wordcount/wordcount_classes'

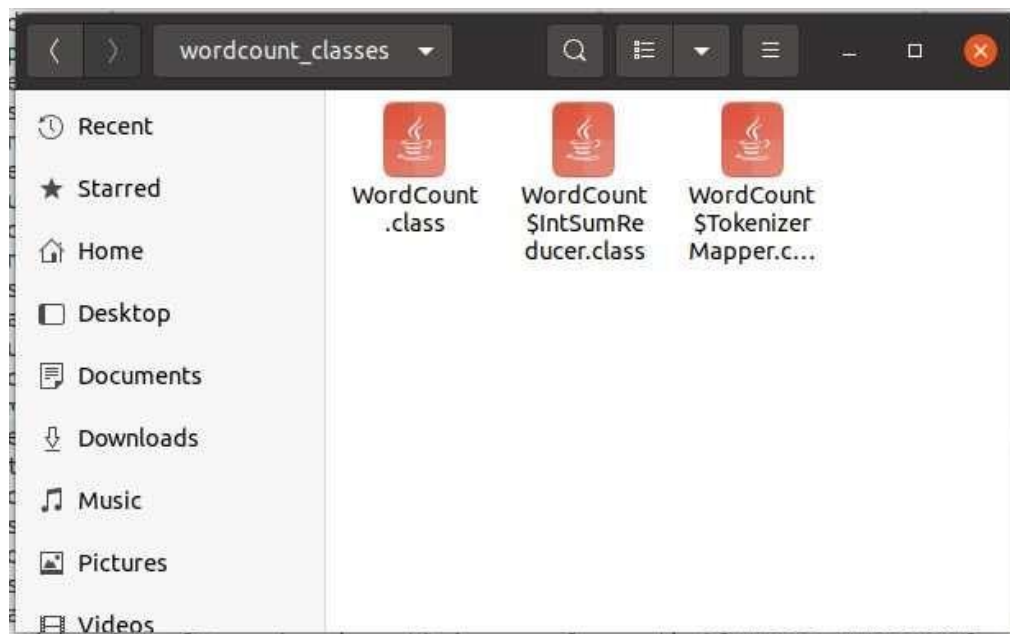
'/home/bda/wordcount/WordCount.java'

Big Data Analytics and Visualization Lab

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

```
javac -classpath ${HADOOP_CLASSPATH} -d  
'/home/bda/wordcount/classes'  
'/home/bda/wordcount/WordCount.java'
```

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

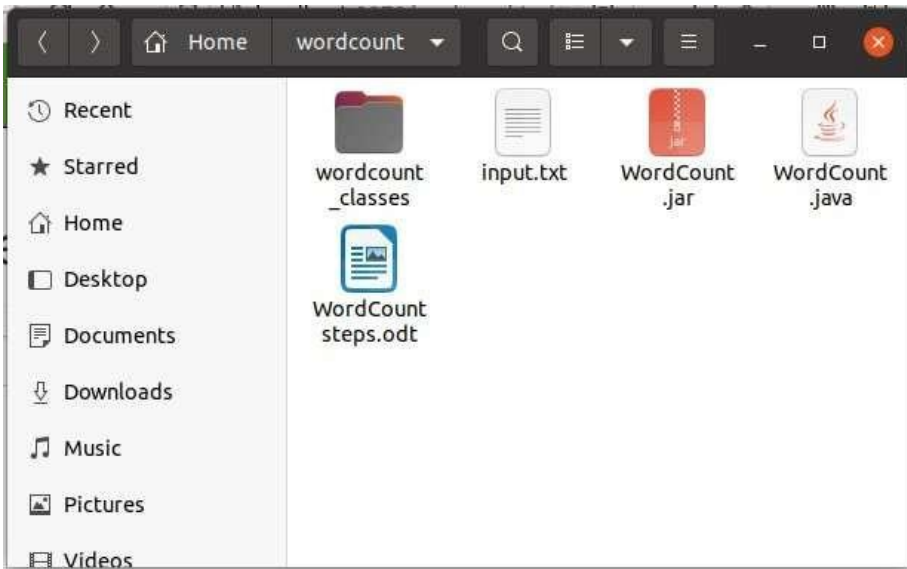


```
bda@bda-VirtualBox:~/wordcount$ jar -cvf WordCount.jar -C  
'/home/bda/wordcount/classes'/. 
```

added manifest

```
adding: WordCount$IntSumReducer.class(in = 1755) (out=  
750)(deflated 57%) adding: WordCount$TokenizerMapper.class(in =  
1752) (out= 761)(deflated 56%) adding: WordCount.class(in = 1511)  
(out= 832)(deflated 44%)
```

Big Data Analytics and Visualization Lab



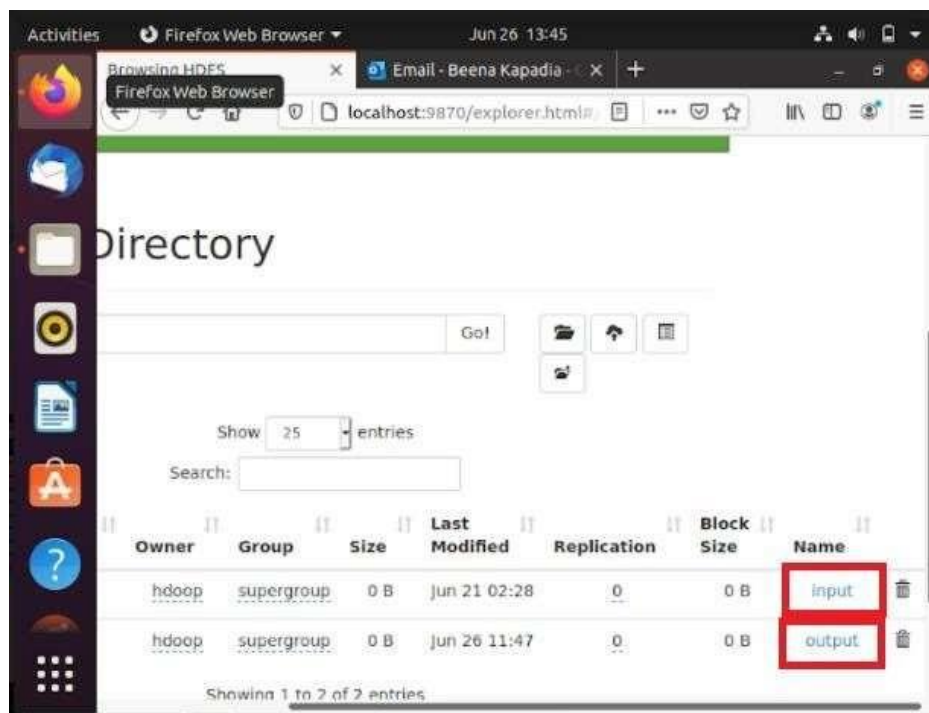
```
bda@bda-VirtualBox:~/wordcount$ jar -cvf
WordCount.jar -C
'/home/bda/wordcount/wordcount_classes'
/.
```

```
added manifest
adding: WordCount$IntSumReducer.class(in = 1755) (out=
750)(deflated 57%) adding: WordCount$TokenizerMapper.class(in =
1752) (out= 761)(deflated 56%) adding: WordCount.class(in = 1511)
(out= 832)(deflated 44%)
```

```
hadoop@bda-VirtualBox:/home/bda/wordcount$ hadoop jar
'/home/bda/wordcount/WordCount.jar' WordCount /wordcount/input/
/wordcount/output
```

output is created.

Big Data Analytics and Visualization Lab



Get the output:

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

```
2021-06-26 13:33:57,781 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
```

```
beena 2
```

```
dhavan
```

```
1
```

```
jeenish
```

```
a
```

```
1
```

```
komal 1
```

```
kruti 2
```

```
mitish 1
```

```
nidhi 3
```

```
vinod 2
```

```
yashesh 3
```

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


Practical No: 03

Aim: MongoDB installation & CRUD Operation.

MONGO DB ENVIRONMENT

To get started with MongoDB, you have to install it in your system. You need to find and download the latest version of MongoDB, which will be compatible with your computer system. You can use this (<http://www.mongodb.org/downloads>) link and follow the instruction to install MongoDB in your PC.

The process of setting up MongoDB in different operating systems is also different, here various installation steps have been mentioned and according to your convenience, you can select it and follow it.

Install Mongo DB in Windows:

The website of MongoDB provides all the installation instructions, and MongoDB is supported by Windows, Linux as well as Mac OS.

It is to be noted that, MongoDB will not run in Windows XP; so you need to install higher versions of windows to use this database.

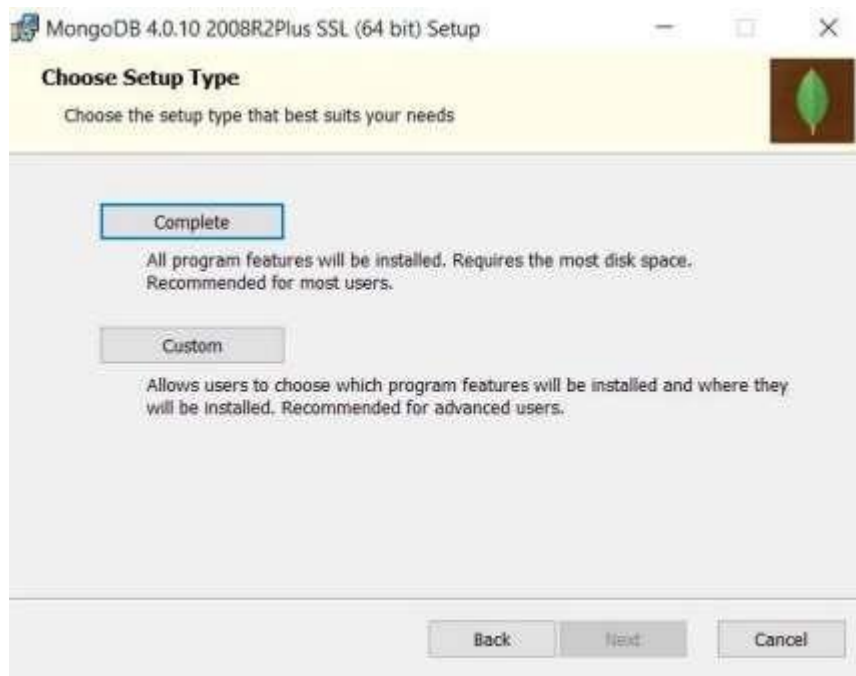
Once you visit the link (<http://www.mongodb.org/downloads>), click the download button.

1. Once the download is complete, double click this setup file to install it. Follow the steps:
- 2.

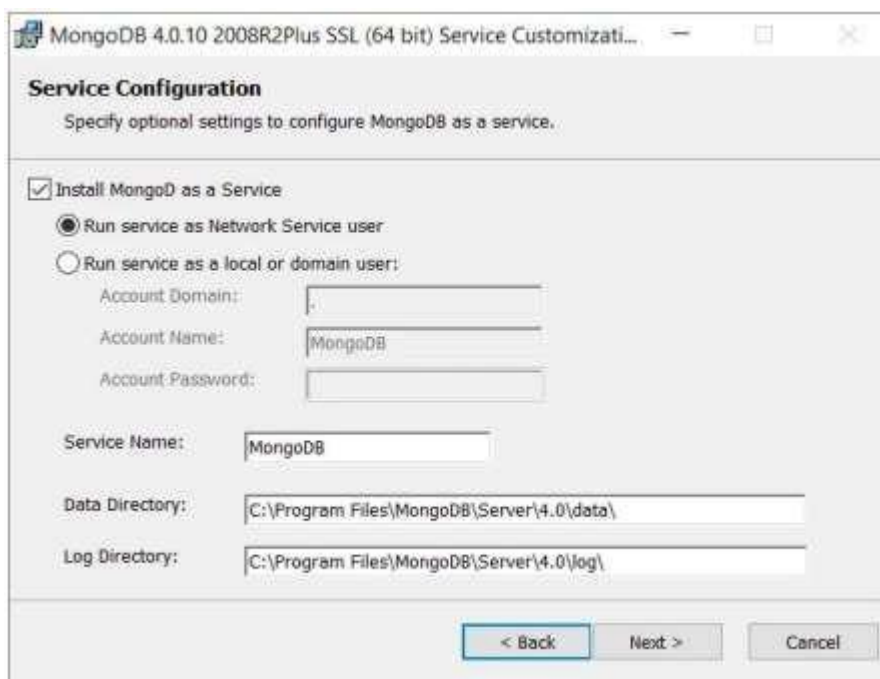
3. N



Big Data Analytics and Visualization Lab

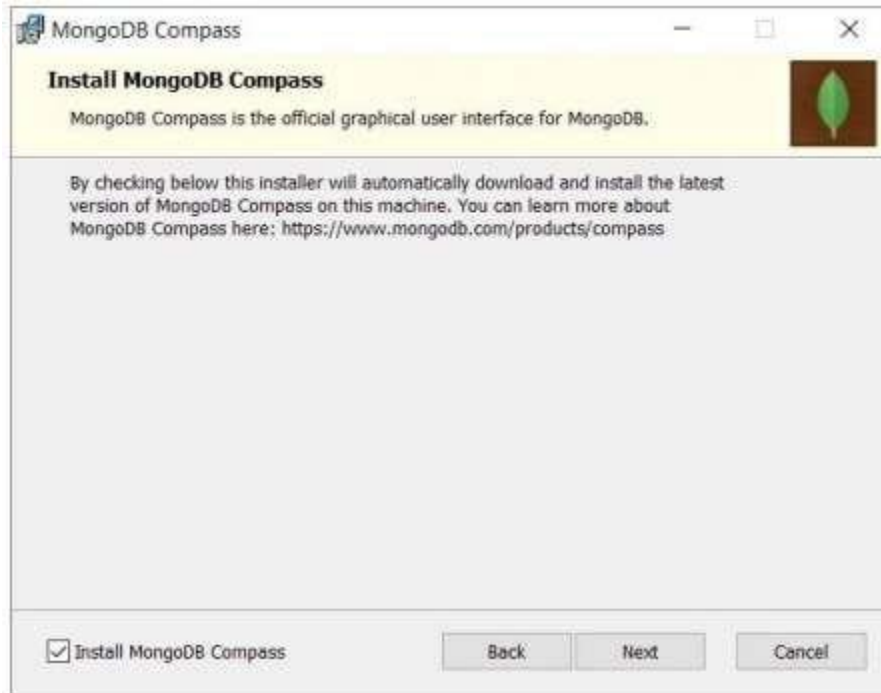


4. Then, select the radio button "Run services as Network service user."



5. The setup system will also prompt you to install MongoDB Compass, which is MongoDB official graphical user interface (GUI). You can tick the checkbox to install that as well

Big Data Analytics and Visualization Lab



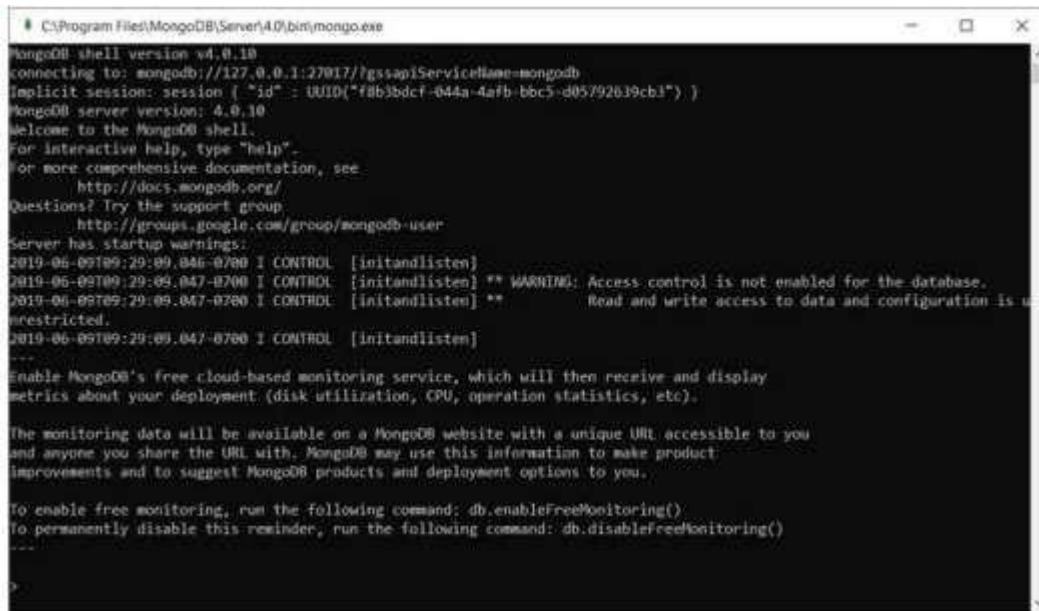
Once the installation is done completely, you need to start MongoDB and to do so follow the process:

1. Open Command Prompt.
2. Type: `C:\Program Files\MongoDB\Server\4.0\bin`
3. Now type the command simply: **mongod** to run the server.

In this way, you can start your MongoDB database. Now, for running MongoDB primary client system, you have to use the command:

`C:\Program Files\MongoDB\Server\4.0\bin>mongo.exe`

Big Data Analytics and Visualization Lab

A screenshot of a Windows command prompt window titled "C:\Program Files\MongoDB\Server\4.0\bin\mongo.exe". The window displays the MongoDB shell version 4.0.10 and shows the connection to a local MongoDB instance. It includes startup warnings about access control and free monitoring service. The text in the window is as follows:

```
MongoDB shell version v4.0.10
connecting to: mongodb://127.0.0.1:27017/?gssapiServiceName=mongodb
Implicit session: session { "id" : UUID("f8b3bdcf-044a-4afb-bbc5-d05792639cb1") }
MongoDB server version: 4.0.10
Welcome to the MongoDB shell.
For interactive help, type "help".
For more comprehensive documentation, see
  http://docs.mongodb.org/
Questions? Try the support group
  http://groups.google.com/group/mongodb-user
Server has startup warnings:
2019-06-09T09:29:09.046-0700 I CONTROL [initandlisten]
2019-06-09T09:29:09.047-0700 I CONTROL [initandlisten] ** WARNING: Access control is not enabled for the database.
2019-06-09T09:29:09.047-0700 I CONTROL [initandlisten] **      Read and write access to data and configuration is unrestricted.
2019-06-09T09:29:09.047-0700 I CONTROL [initandlisten]
...
Enable MongoDB's free cloud-based monitoring service, which will then receive and display
metrics about your deployment (disk utilization, CPU, operation statistics, etc).

The monitoring data will be available on a MongoDB website with a unique URL accessible to you
and anyone you share the URL with. MongoDB may use this information to make product
improvements and to suggest MongoDB products and deployment options to you.

To enable free monitoring, run the following command: db.enableFreeMonitoring()
To permanently disable this reminder, run the following command: db.disableFreeMonitoring()
...
>
```

CREATING A DATABASE USING MONGODB

To build a database in MongoDB, first construct a MongoClient object, and then supply a connection URL with the right IP address and the database name. If the database does not already exist, MongoDB will create it and connect to it.

Example: Create a database called “mydb”

```
var MongoClient = require('mongodb').MongoClient; var url =
"mongodb://localhost:27017/mydb";
```

```
MongoClient.connect(url, function(err, db) { if (err) throw err;
console.log("Database created!");
db.close();
});
```

Save the code above in a file called "demo_create_mongo_db.js" and run the file:

Run "demo_create_mongo_db.js"

C:\Users\Your Name>node demo_create_mongo_db.js This will give you this result:

Database created!

Note: MongoDB waits until you have created a collection (table), with at least one document (record) before it actually creates the database (and collection).

Big Data Analytics and Visualization Lab

The use Command

MongoDB use DATABASE_NAME is used to create database. The command will create a new database if it doesn't exist, otherwise it will return the existing database.

Syntax

Basic syntax of use DATABASE statement is as follows –

use DATABASE_NAME

Example

If you want to use a database with name **<mydb>**, then use **DATABASE** statement would be as follows –

```
>use mydb
```

switched to db mydb

To check your currently selected database, use the command **db**

```
>db Mydb
```

If you want to check your databases list, use the command **show dbs**.

```
>show dbs
```

```
local    0.78125GB
```

```
test     0.23012GB
```

Your created database (mydb) is not present in list. To display database, you need to insert at least one document into it.

```
>db.movie.insert({"name":"tutorials point"})
```

```
>show dbs
```

```
local    0.78125GB
```

```
mydb     0.23012GB
```

```
test     0.23012GB
```

In MongoDB default database is test. If you didn't create any database, then collections will be stored in test database.

DROP DATABASE

The **dropDatabase()** Method

MongoDB db.dropDatabase() command is used to drop a existing database.

Big Data Analytics and Visualization Lab

Syntax

Basic syntax of dropDatabase() command is as follows –

```
db.dropDatabase()
```

This will delete the selected database. If you have not selected any database, then it will delete default 'test' database.

Example

First, check the list of available databases by using the command, showdbs.

```
>show dbs
```

```
local    0.78125GB
```

```
mydb     0.23012GB
```

```
test     0.23012GB
```

```
>
```

If you want to delete new database <mydb>, then dropDatabase() Command would be as follows

```
>use mydb
```

```
switched to db mydb
```

```
>db.dropDatabase()
```

```
>{ "dropped" : "mydb", "ok" : 1 }
```

```
>
```

Now check list of databases.

```
>show dbs
```

```
local 0.78125GB
```

```
test 0.23012GB
```

CREATING COLLECTIONS IN MONGO DB

The createCollection() Method

MongoDB db.createCollection(name, options) is used to create collection.

Syntax

Basic syntax of createCollection() command is as follows –

```
db.createCollection(name, options)
```

In the command, **name** is name of collection to be created. **Options** is a document and is used to specify configuration of collection.

Name: Yadav Uday Sagindra

Seat Number: 5000088

Big Data Analytics and Visualization Lab

| Parameter | Type | Description |
|-----------|----------|---|
| Name | String | Name of the collection to be created |
| Options | Document | (Optional) Specify options about memory size and indexing |

Options parameter is optional, so you need to specify only the name of the collection. Following is the list of options you can use –

| Field | Type | Description |
|-------------|---------|---|
| capped | Boolean | (Optional) If true, enables a capped collection. Capped collection is a fixed size collection that automatically overwrites its oldest entries when it reaches its maximum size. If you specify true, you need to specify size parameter also. |
| autoIndexId | Boolean | (Optional) If true, automatically create index on _id field.s Default value is false. |
| size | number | (Optional) Specifies a maximum size in bytes for a capped collection. If capped is true, then you need to specify this field also. |
| max | number | (Optional) Specifies the maximum number of documents allowed in the capped collection. |

While inserting the document, MongoDB first checks size field of capped collection, then it checks max field.

Examples

Name: Yadav Uday Sagindra
Seat Number: 5000088

Big Data Analytics and Visualization Lab

Basic syntax of **createCollection()** method without options is as follows –

```
>use test
switched to db test
>db.createCollection("mycollection")
{ "ok" : 1 }
>
```

You can check the created collection by using the command show collections.

```
>show collections
mycollection
system.indexes
```

The following example shows the syntax of **createCollection()** method with few important options –

```
> db.createCollection("mycol", { capped : true, autoIndexID : true, size :
6142800, max : 10000 } ){
"ok" : 0,
"errmsg" : "BSON field 'create.autoIndexID' is an unknown field.", "code" : 40415,
"codeName" : "Location40415"
}
>
```

In MongoDB, you don't need to create collection. MongoDB creates collection automatically, when you insert some document.

```
>db.tutorialspoint.insert({"name" : "tutorialspoint"}),WriteResult({
"nInserted" : 1 })
>show collectionsmycol
mycollection system.indexes
tutorialspoint
>
```

The drop() Method

MongoDB's **db.collection.drop()** is used to drop a collection from the database.

Syntax

Basic syntax of **drop()** command is as follows –

```
db.COLLECTION_NAME.drop()
```

Example

First, check the available collections into your database **mydb**.

Big Data Analytics and Visualization Lab

```
>use mydb  
switched to db mydb  
>show collectionsmycol  
mycollection system.indexes  
tutorialspoint  
>
```

Now drop the collection with the name **mycollection**.

```
>db.mycollection.drop()true  
>
```

Again check the list of collections into database.

```
>show collections  
mycol  
system.indexes  
tutorialspoint  
>
```

drop() method will return true, if the selected collection is dropped successfully, otherwise it will return false.

CRUD DOCUMENT

As we know, we can use MongoDB for a variety of purposes such as building an application (including web and mobile), data analysis, or as an administrator of a MongoDB database. In all of these cases, we must interact with the MongoDB server to perform specific operations such as entering new data into the application, updating data in the application, deleting data from the application, and reading the data of the application.

MongoDB provides a set of simple yet fundamental operations known as CRUD operations that will allow you to quickly interact with the MongoDB server.



Big Data Analytics and Visualization Lab

3.4.1 Create Operations — these operations are used to insert or add new documents to the collection. If a collection does not exist, a new collection will be created in the database. MongoDB provides the following methods for performing and creating operations:

| Method | Description |
|-----------------------------------|--|
| db.collection.insertOne() | It is used to insert a single document in the collection. |
| db.collection.insertMany() | It is used to insert multiple documents in the collection. |

insertOne()

As the namesake, insertOne() allows you to insert one document into the collection. For this example, we're going to work with a collection called RecordsDB. We can insert a single entry into our collection by calling the insertOne() method on RecordsDB. We then provide the information we want to insert in the form of key-value pairs, establishing the Schema.

Example

```
db.RecordsDB.insertOne({
  name: "Marsh",
  age: "6 years",
  species: "Dog",
  ownerAddress: "380 W. Fir Ave",
  chipped: true
})
```

If the create operation is successful, a new document is created. The function will return an object where "acknowledged" is "true" and "insertID" is the newly created "ObjectId."

```
> db.RecordsDB.insertOne({
  name: "Marsh",
  age: "6 years",
  species: "Dog",
  ownerAddress: "380 W. Fir Ave",
  chipped: true
})
```

Big Data Analytics and Visualization Lab

```
{  
  "acknowledged" : true,  
  "insertedId" : ObjectId("5fd989674e6b9ceb8665c57d")  
}
```

insertMany()

It's possible to insert multiple items at one time by calling the insertMany() method on the desired collection. In this case, we pass multiple items into our chosen collection (RecordsDB) and separate them by commas. Within the parentheses, we use brackets to indicate that we are passing in a list of multiple entries. This is commonly referred to as a nested method.

Example:

```
db.RecordsDB.insertMany([ {name: "Marsh",  
  age: "6 years", species: "Dog",  
  ownerAddress: "380 W. Fir Ave",chipped: true},  
  { name: "Kitana",  
    age: "4 years",species: "Cat",  
    ownerAddress: "521 E. Cortland",  
    chipped: true} ])
```

```
db.RecordsDB.insertMany([ { name: "Marsh", age: "6 years", species:  
  "Dog",  
  ownerAddress: "380 W. Fir Ave", chipped: true}, {name: "Kitana", age: "4  
  years",  
  species: "Cat", ownerAddress: "521 E. Cortland", chipped: true} ])  
  
{  
  "acknowledged" : true,  
  "insertedIds" : [  
    ObjectId("5fd98ea9ce6e8850d88270b4"),  
    ObjectId("5fd98ea9ce6e8850d88270b5")  
  ]  
}
```

Read Operations:

You can give specific query filters and criteria to the read operations to indicate the documents you desire. More information on the possible query filters can be found in the MongoDB manual. Query modifiers can also be used to vary the number of results returned.

MongoDB offers two ways to read documents from a collection:

- `db.collection.find()`
- `db.collection.findOne()`

find()

In order to get all the documents from a collection, we can simply use the `find()` method on our chosen collection. Executing just the `find()` method with no arguments will return all records currently in the collection.

db.RecordsDB.find()

In order to get one document that satisfies the search criteria, we can simply use the `findOne()` method on our chosen collection. If multiple documents satisfy the query, this method returns the first document according to the natural order which reflects the order of documents on the disk. If no documents satisfy the search criteria, the function returns null. The function takes the following form of syntax.

`db.{collection}.findOne({query}, {projection})`

Update Operations

Update operations, like create operations, work on a single collection and are atomic at the document level. Filters and criteria are used to choose the documents to be updated during an update procedure.

You should be cautious while altering documents since alterations are permanent and cannot be reversed. This also applies to remove operations.

There are three techniques for updating documents in MongoDB CRUD:

- `db.collection.updateOne()`
- `db.collection.updateMany()`
- `db.collection.replaceOne()`

updateOne()

With an update procedure, we may edit a single document and update an existing record. To do this, we use the `updateOne()` function on a specified collection, in this case "RecordsDB." To update a document, we pass two arguments to the method: an update filter and an update

Big Data Analytics and Visualization Lab

action.

The update filter specifies which items should be updated, and the update action specifies how those items should be updated. We start with the `updatefilter`. Then we utilise the "\$set" key and supply the values for the fields we wish to change. This function updates the first record that matches the specified filter.

updateMany()

`updateMany()` allows us to update multiple items by passing in a list of items, just as we did when inserting multiple items. This update operation uses the same syntax for updating a single document.

replaceOne()

The `replaceOne()` method is used to replace a single document in the specified collection. `replaceOne()` replaces the entire document, meaning fields in the old document not contained in the new will be lost.

Delete Operations

Delete operations, like update and create operations, work on a single collection. For a single document, delete actions are similarly atomic. You can provide delete actions with filters and criteria to indicate which documents from a collection you want to delete. The filter options use the same syntax as the read operations.

MongoDB provides two ways for removing records from a collection:

- `db.collection.deleteOne()`
- `db.collection.deleteMany()`

deleteOne()

`deleteOne()` is used to remove a document from a specified collection on the MongoDB server. A filter criterion is used to specify the item to delete. It deletes the first record that matches the provided filter.

deleteMany()

`deleteMany()` is a method used to delete multiple documents from a desired collection with a single delete operation. A list is passed into the method and the individual items are defined with filter criteria as in `deleteOne()`

Let us sum up

- MongoDB is an open-source database that uses a document-oriented

Big Data Analytics and Visualization Lab

data model and a non-structured query language

- It is one of the most powerful NoSQL systems and databases around, today.
- The data model that MongoDB follows is a highly elastic one that lets you combine and store data of multivariate types without having to compromise on the powerful indexing options, data access, and validation rules.
- A group of database documents can be called a collection. The RDBMS equivalent to a collection is a table. The entire collection exists within a single database. There are no schemas when it comes

Practical No: 04

Aim: Data Visualization

- Microsoft Power BI
- Oracle Visual Analyzer
- SAP Lumira
- SAS Visual Analytics
- Tibco Spotfire
- Zoho Analytics
- D3.js
- Jupyter
- MicroStrategy
- Google Charts

DATA VISUALIZATION BEST PRACTICES

- Set the context:
- Know your audience(s)
- Choose an effective visual
- Keep it simple



Fig 1: Good data visualization

BASIC DATA VISUALIZATION PRINCIPLES

Three basic principles seem especially useful to guide your creation of better, more effective visualizations.

1. Show the Data

People read will read the graphs in your report, article, or blog post to better understand your argument.

2. Reduce the Clutter

Chart clutter, the use of unnecessary or distracting visual elements, tends to reduce effectiveness of the graph. Clutter comes in many forms: dark or heavy gridlines; unnecessary tick marks, labels, or text; unnecessary icons or pictures; ornamental shading and gradients; and unnecessary dimensions.

3. Integrate the Text and the Graph

Legends define or explain a series on a graph are often placed faraway from the content—off to the right or below the graph.

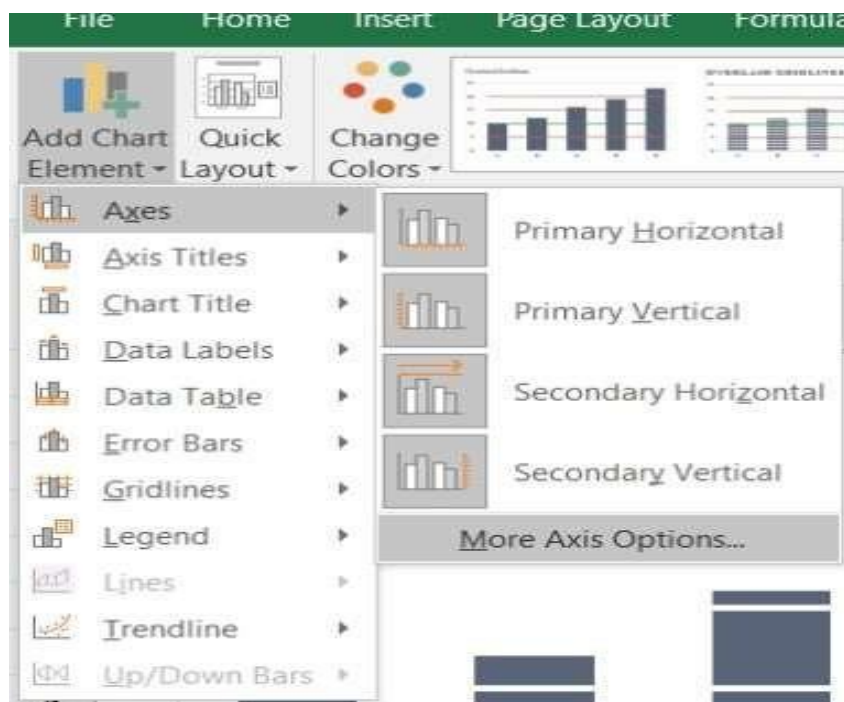
Design Tab

Chart Tools tab will appear at the top of your ribbon consisting of two tabs: Design and Format. The Design tab contains options that allow you to apply different default 'Chart Layouts' and 'Chart Styles'. The options available under the 'Add Chart Element' button replaces the Layout tab on previous versions of Excel and allows you to modify the appearance of axes, titles, gridlines, and more.

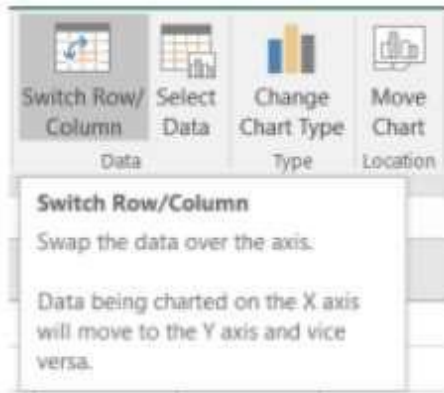


Each of the options in the 'Add Chart Element' menu allows you to choose from a set of pre-populated options, or to open a menu with more options.

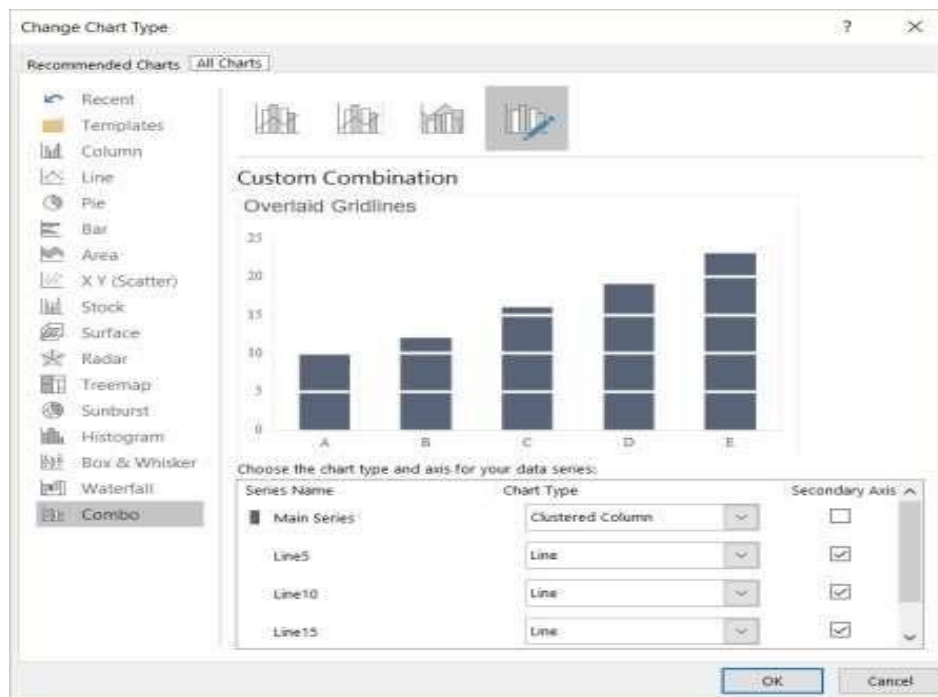
Big Data Analytics and Visualization Lab



The 'Change Chart Type' button will allow you to change the type of chart for all the data on the chart, or a selected series.



Big Data Analytics and Visualization Lab



Format Tab

The Format tab contains the standard outline and fill color options



In the very top-left section of the Format tab is the 'Chart Elements' drop-down menu. The list in this drop-down menu consists of everything in your chart including titles, axes, error bars, and every series. If you have a lot of objects on your chart, this drop-down menu will help you to easily find and select what you need.

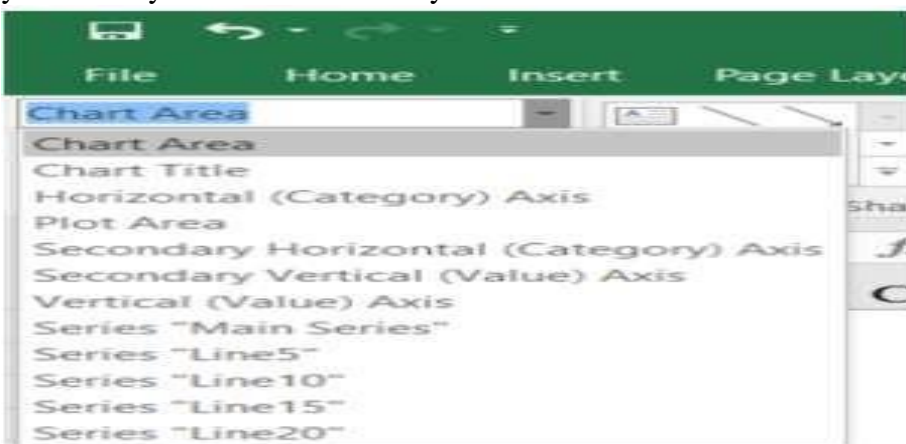
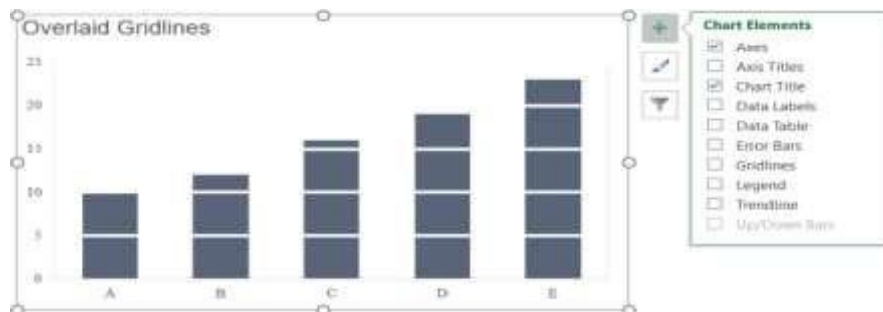
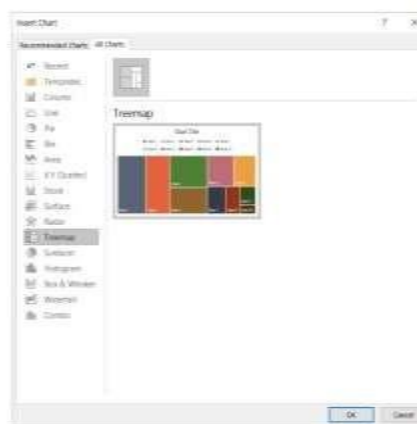
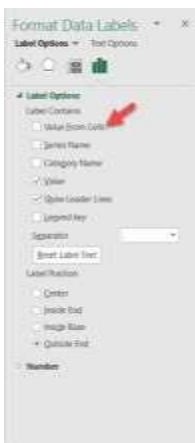


Chart Elements Menu

‘Chart Elements Menu’ that appears just outside the top-right part of the chart when you select it. Appearing as a ‘plus’ symbol, the menu is identical to the ‘Add Chart Elements’ button in the Design tab.



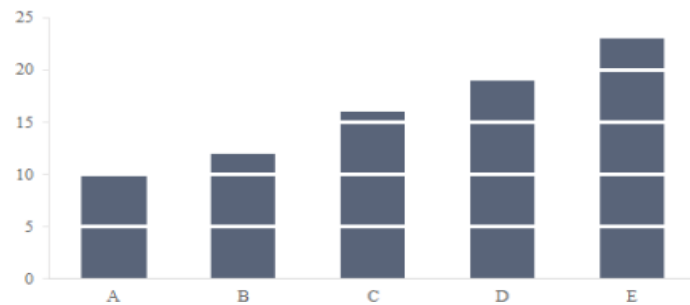
Select a specific data range to use as labels in your chart. This comes in quite handy when, for example, you want to add custom labels to a scatterplot. Instead of having to do the labeling manually, you can select the data labels series in the spreadsheet. Among the new chart types is a Treemap, Histogram, Box & Whisker, and Waterfall chart.



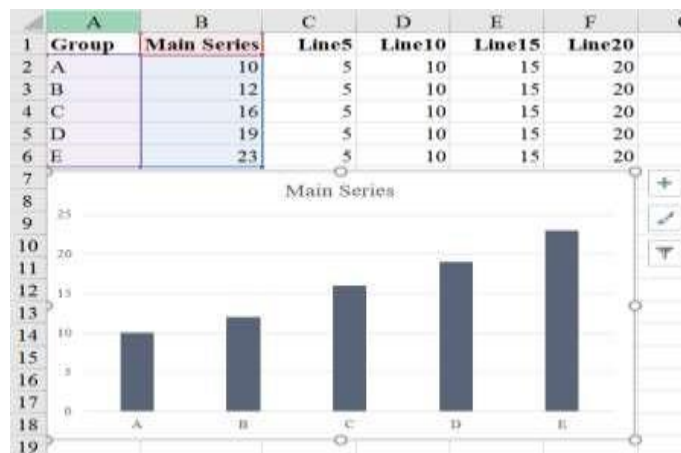
Big Data Analytics and Visualization Lab

Overlaid Gridlines

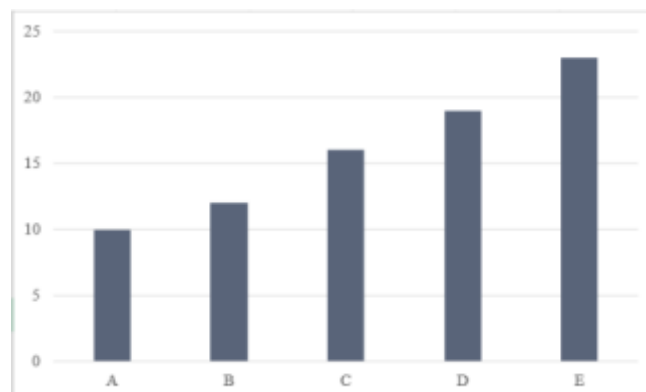
The Overlaid Gridline chart is a column chart with gridlines on top of the columns. This type of chart allows viewers to absorb the column data as segments rather than single columns. Use the OverlaidGridline tab in the Advanced Data Visualizations with Excel 2016 Hands-On.xlsx spreadsheet to create the chart.



1. Begin by creating a column chart from columns A (“Group”) and B (“Main Series”).

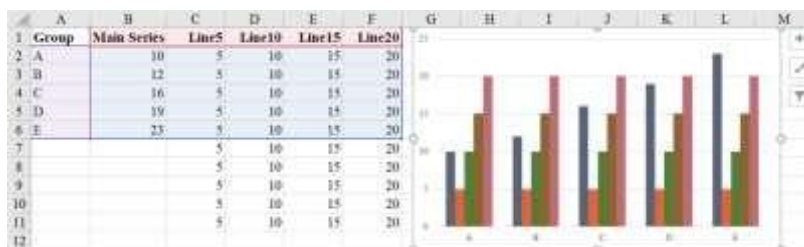


2. Remove the title.

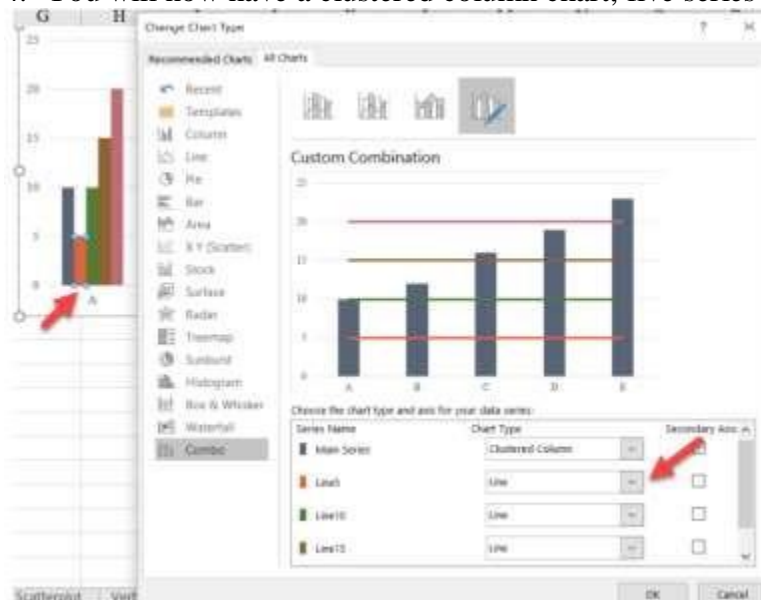


Big Data Analytics and Visualization Lab

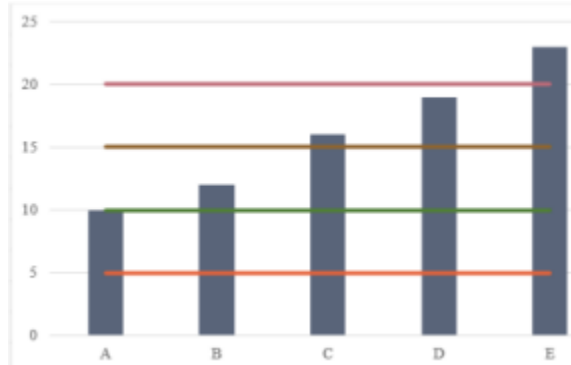
3. We're now going to add the four "Line" series to the chart.



4. You will now have a clustered column chart, five series for each group.

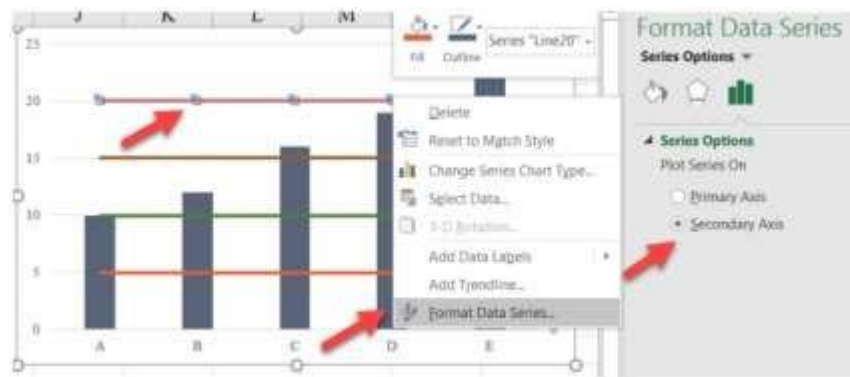


Each series except the "Main Series" now become lines.

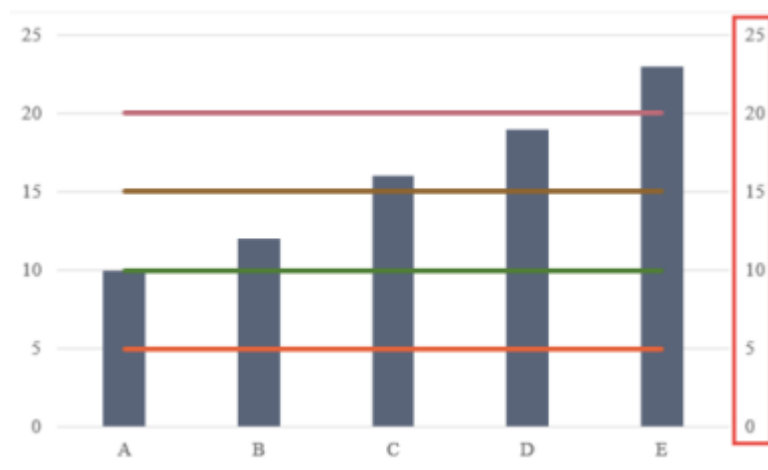


5. If we were to simply change the lines to white, they would end in the middle of the bars of the A and E groups. We now move each of those four lines to the "secondary axis" so we can get them to stretch through the bars. To do so, first select a line, right-click, and select "Format Data Series". Go to the "Series Options" tab and select the "Secondary Axis" option.

Big Data Analytics and Visualization Lab



6. You'll notice that a new y-axis has appeared on the right side of the graph. When you're done moving all four series to the secondary axis, this new y-axis should go from 0 to 25.



7. There is also now a secondary x-axis, but we need to turn it on. To do so, select the "Axes" option in the "Chart Elements" menu by pressing the "plus" button that will appear when you select the chart. By hovering over the "Axes" menu, three of the boxes will have checkmarks next to them. Turn on the "Secondary Horizontal" axis by selecting the checkbox



8. Change the colors of the lines to white using the "Format" tab option.

Big Data Analytics and Visualization Lab



9. We fix that by changing how the data points line up with the tick marks. In a default line graph in Excel, the data markers line up between the tick marks; notice how the line begins in the middle of the A bar, between the y-axis and the tick mark between the A and B groups. By placing the data markers on the tick marks, we can extend the lines through the bars. To do so, we'll format the secondary x-axis (by rightclicking and navigating to the "Axis position" options under "Axis Options" in the "Format axis" menu.

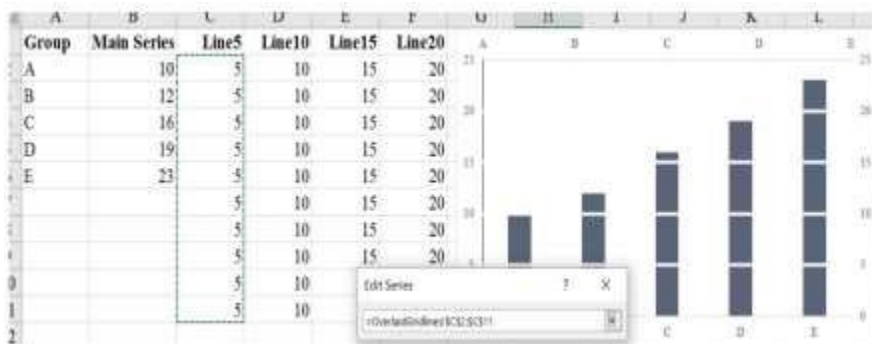


10. Add your vertical primary axis line, select the axis and add the line under the "Format Axis" menu.

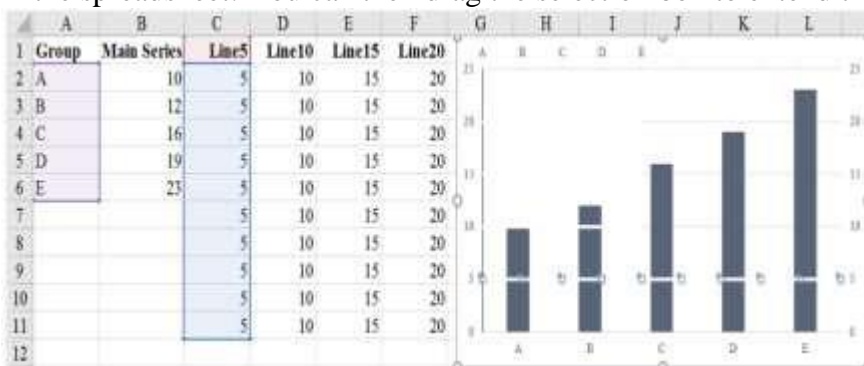


Big Data Analytics and Visualization Lab

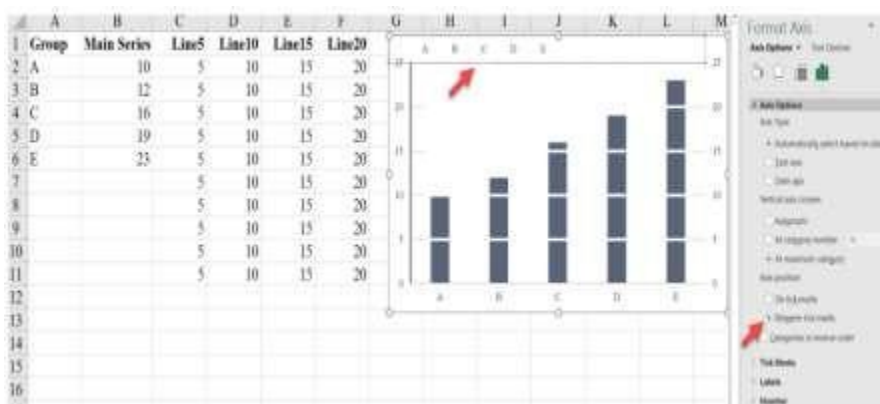
11. We want to extend the data series for each “Line” series through row One way to do this is to right-click on the graph, select the “Select Data” option, and edit each of the 4 “Line” series to extend the data series.



Alternatively, you can select the line on the chart and you’ll notice that your data are selected in the spreadsheet. You can then drag the selection box to extend the data series.



12. We need to now change where the data markers line up with the tick marks. Once again, format the secondary x-axis and change the “Position Axis” back to “Between tick marks”.

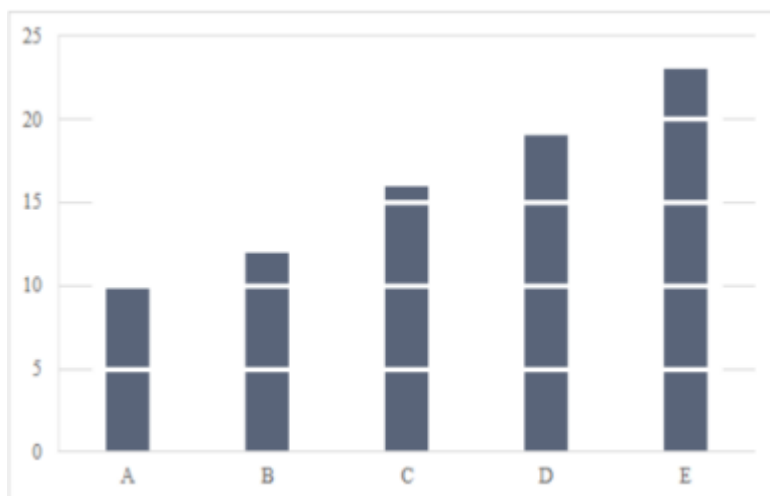


13. We also want to turn off the secondary horizontal axis and set the “Line Color” to “No line”.

Big Data Analytics and Visualization Lab

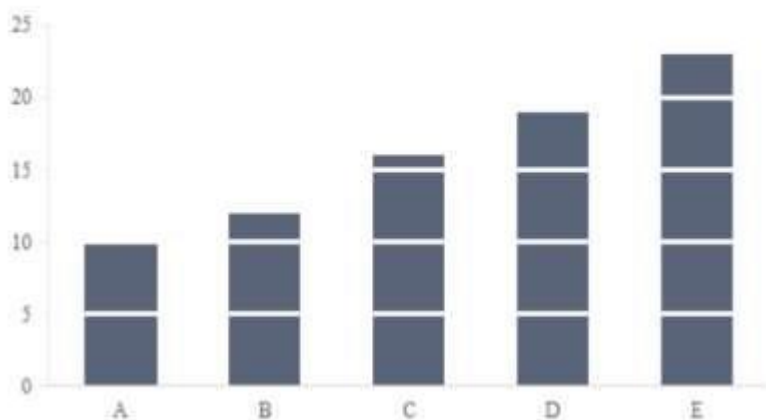


14. Repeat the process in Step 13 for the secondary vertical axis, remove the gridlines and style the rest as you see fit.



Final Version with Styling

Overlaid Gridlines

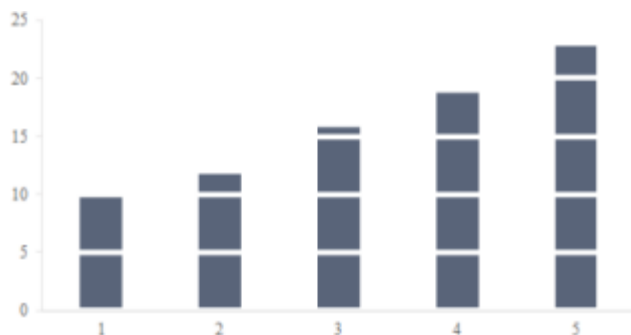


Overlaid Gridlines with a Formula

In this version of the Overlaid Gridlines chart, I create a stacked column chart. Each section of

Big Data Analytics and Visualization Lab

the chart is given a white outline so that it appears like there are gridlines. Use the Overlaid Gridlines_Formula tab in the Advanced Data Visualizations with Excel 2016 Hands-On.xlsx spreadsheet to create the chart.



Create a stacked column chart from cells C16:M20. These are the cells that contain the formula.



To plot the rows, select the chart and the “Switch Row/Column” button in the “Design” tab of the ribbon.

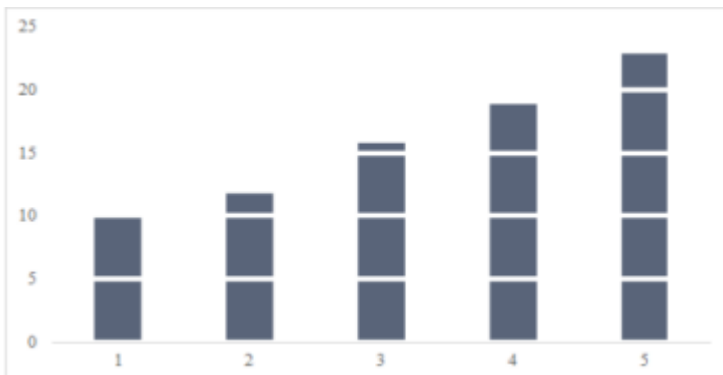


Change the fill of each shape under the “Shape Fill” dropdown in the “Format” menu to the same color. Similarly, change the color of the “Shape Outline” to white and increase the thickness to your desired weight. Of course, delete the existing (default) gridlines, legend, etc.

Big Data Analytics and Visualization Lab

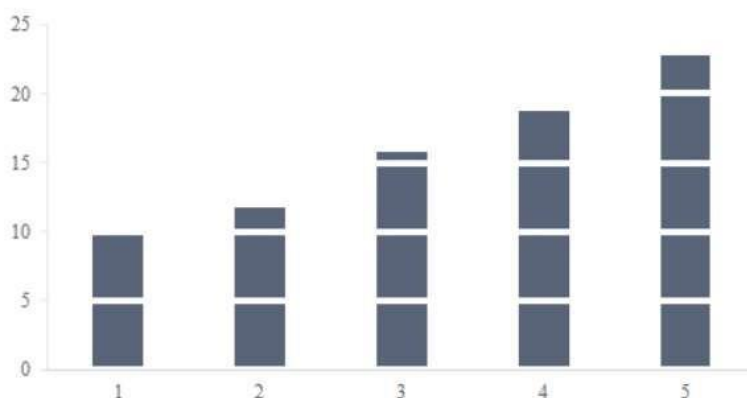


Repeat for all 5 series



Final Version with Styling

Overlaid Gridlines with a Formula



Practical No: 05

Aim: Data Visualization using tableau.

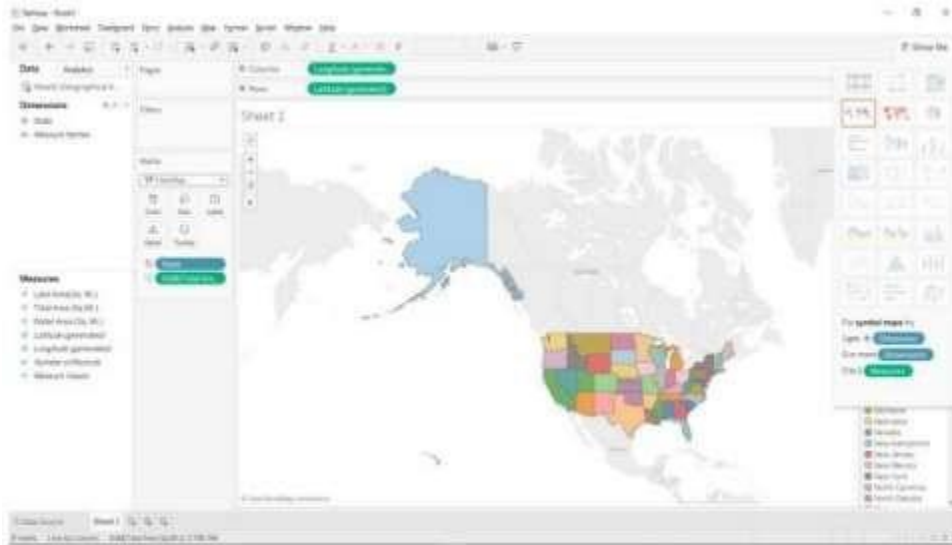


Geographical map visualization using Tableau:

Steps to visualize Filled Maps in Tableau:

1. Open Tableau and connect excel sheet.
2. The data with states and their geographical data is shown in Tableau.
3. Drag and drop states onto columns and Total Area onto rows.
4. Drag and drop state onto color and total area onto tool tip.
5. Now select the filled map. Then the map is represented with the geographical data of each state as shown below.

Big Data Analytics and Visualization Lab



Visualization of Combo and Gauge Charts using QlikView

1. For this we have taken two different datasets.

- Dataset -1 (Cost_Student). Contains data which representing amount spend by the countries, on a student per year for primary to secondary education and postsecondary education.
- Dataset - 2 (Sales_Rep). Contains data, which represents the sales made by the representatives of a stationary chain around USA over particular dates

Big Data Analytics and Visualization Lab

| OrderDa | Regi | Rep | Item | Sali | Unit C | Reveni | Targe |
|----------|---------|----------|---------|------|--------|----------|-------|
| 1/6/14 | East | Jones | Pencil | 95 | 1.99 | 189.05 | 115 |
| 1/23/14 | Central | Kivell | Binder | 50 | 19.99 | 999.50 | 115 |
| 2/9/14 | Central | Jardine | Pencil | 36 | 4.99 | 179.64 | 115 |
| 2/26/14 | Central | Gill | Pen | 27 | 19.99 | 539.73 | 115 |
| 3/15/14 | West | Sorvino | Pencil | 56 | 2.99 | 167.44 | 115 |
| 4/1/14 | East | Jones | Binder | 60 | 4.99 | 299.40 | 115 |
| 4/18/14 | Central | Andrews | Pencil | 75 | 1.99 | 149.25 | 115 |
| 5/5/14 | Central | Jardine | Pencil | 90 | 4.99 | 449.10 | 115 |
| 5/22/14 | West | Thompson | Pencil | 32 | 1.99 | 63.68 | 115 |
| 6/8/14 | East | Jones | Binder | 60 | 8.99 | 539.40 | 115 |
| 6/25/14 | Central | Morgan | Pencil | 90 | 4.99 | 449.10 | 115 |
| 7/12/14 | East | Howard | Binder | 29 | 1.99 | 57.71 | 115 |
| 7/29/14 | East | Parent | Binder | 81 | 19.99 | 1,619.19 | 115 |
| 8/15/14 | East | Jones | Pencil | 35 | 4.99 | 174.65 | 115 |
| 9/1/14 | Central | Smith | Desk | 2 | 125.00 | 250.00 | 15 |
| 9/18/14 | East | Jones | Pen Set | 16 | 15.99 | 255.84 | 115 |
| 10/5/14 | Central | Morgan | Binder | 28 | 8.99 | 251.72 | 115 |
| 10/22/14 | East | Jones | Pen | 64 | 8.99 | 575.36 | 115 |
| 11/8/14 | East | Parent | Pen | 15 | 19.99 | 299.85 | 115 |
| 11/25/14 | Central | Kivell | Pen Set | 96 | 4.99 | 479.04 | 115 |
| 12/12/14 | Central | Smith | Pencil | 67 | 1.29 | 86.43 | 115 |
| 12/29/14 | East | Parent | Pen Set | 74 | 15.99 | 1,183.26 | 115 |
| 1/15/15 | Central | Gill | Binder | 46 | 8.99 | 413.54 | 115 |
| 2/1/15 | Central | Smith | Binder | 87 | 15.00 | 1,305.00 | 115 |
| 2/18/15 | East | Jones | Binder | 4 | 4.99 | 19.96 | 115 |
| 3/7/15 | West | Sorvino | Binder | 7 | 19.99 | 139.93 | 115 |
| 3/24/15 | Central | Jardine | Pen Set | 50 | 4.99 | 249.50 | 115 |
| 4/10/15 | Central | Andrews | Pencil | 66 | 1.99 | 131.34 | 115 |
| 4/27/15 | East | Howard | Pen | 96 | 4.99 | 479.04 | 115 |
| 5/14/15 | Central | Gill | Pencil | 53 | 1.29 | 68.37 | 115 |
| 5/31/15 | Central | Gill | Binder | 80 | 8.99 | 719.20 | 115 |
| 6/17/15 | Central | Kivell | Desk | 5 | 125.00 | 625.00 | 15 |
| 7/4/15 | East | Jones | Pen Set | 62 | 4.99 | 309.38 | 115 |
| 7/21/15 | Central | Morgan | Pen Set | 55 | 12.49 | 686.95 | 115 |
| 8/7/15 | Central | Kivell | Pen Set | 42 | 23.95 | 1,005.90 | 115 |
| 8/24/15 | West | Sorvino | Desk | 3 | 275.00 | 825.00 | 15 |
| 9/10/15 | Central | Gill | Pencil | 7 | 1.29 | 9.03 | 115 |
| 9/27/15 | West | Sorvino | Pen | 76 | 1.99 | 151.24 | 115 |
| 10/14/15 | West | Thompson | Binder | 57 | 19.99 | 1,139.43 | 115 |
| 10/31/15 | Central | Andrews | Pencil | 14 | 1.29 | 18.06 | 115 |
| 11/17/15 | Central | Jardine | Binder | 11 | 4.99 | 54.89 | 115 |
| 12/4/15 | Central | Jardine | Binder | 94 | 19.99 | 1,879.06 | 115 |
| 12/21/15 | Central | Andrews | Binder | 28 | 4.99 | 139.72 | 115 |

2. Open QlikView. A QlikView getting started screen opens, now click on the New Document button on right below corner of the Wizard.
3. A new Window opens up (Kind of pop up) close the pop up window. (Since it is Auto chart generator, we do not want that wizard).
4. On closing the pop up, user will be able to see a Main blank sheet.
Now have quick look on to the toolbar.
5. Over the tool bar below to the selections label1, click on the Edit Script icon (Icon looks like pen on a paper).
6. A new Edit Script window opens up (We use this window to load an Excel file), bottom of the window under the Data from files Click Table Files button, and choose the saved Excel file, then click open.
7. A new File Wizard Type window opens up. Now select the Tables drop down and choose Cost_Student if it has not defaulted and click the Labels drop down and choose Embedded Labels. Click next until, next button disables (If it disables you are on final screen). Select

Big Data Analytics and Visualization Lab

check box load all and click finish.

8. On Finish you will be able to see Edit Script screen , with script added to load excel file (Script shown below) On Edit Script window tool bar below file , click the reload icon. It asks you save the file. `LOAD * FROM [C:\Manusha\DataViz\DataViz-Qlik\QlikView_Dataset.xls] (biff, embedded labels, table is Cost_Per_Student$);`

9. Now a sheet property window opens, please choose Country (Which works as a filter).

10. Right Click anywhere on the sheet, a window opens choose New Sheet Object and select Table box. New window open again, add all the available fields and click apply. You will be able to see a table box with content same as to excel sheet.

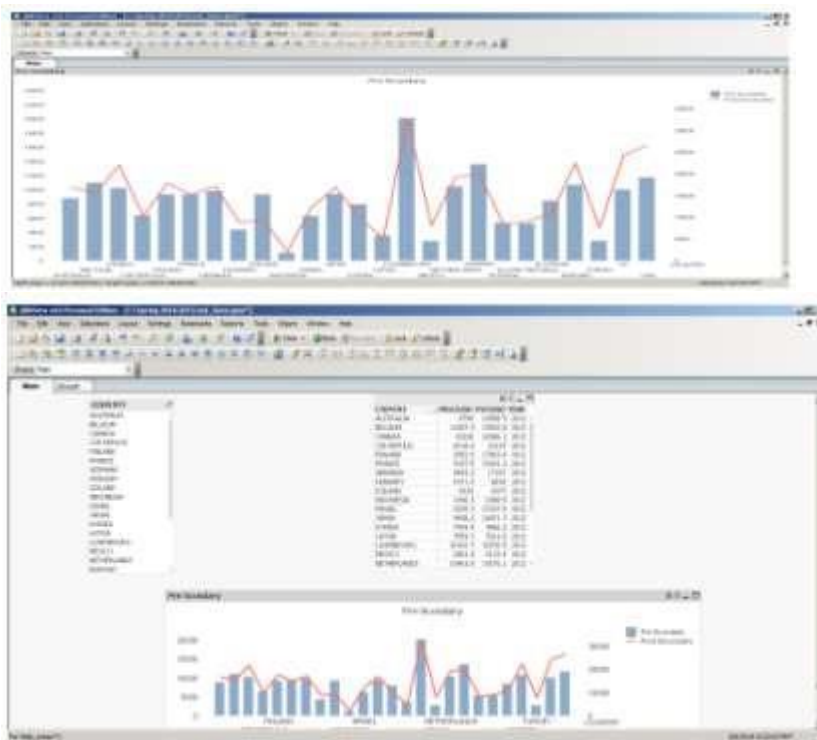
11. Right Click anywhere on the sheet, a window opens choose New Sheet Object and select Chart. New window open again, Select Combo Chart from chart type and click next.

12. Choose Country as dimension and click next now choose the expressions.

- Expression 1 for bar chart Sum (PRSCIUSD), now click Add button to add one more expression Sum (PSCIUSD) and click finish.

13. Now we have seen the Combo Chart.

14. If you observe both pre-Secondary and post-secondary are on same axis. Let us split them. Right click on the chart and choose the properties, and navigate to the Axes tab, Select Post-Secondary Expression and choose the position Right (Top) click apply and ok. You will see charts similar to below.



15. Now click on layout shown in the tool bar and add new sheet. A new sheet will open.

16. Repeat step 7 to step 12, in step 9 choose sales_Rep sheet and in step 11 Choose Rep to the field display in list boxes.

Big Data Analytics and Visualization Lab

17. Right Click anywhere on the sheet, a window opens choose New Sheet Object and select Chart. New window open again, Select Gauge from chart type and click next.
18. Gauge charts have no dimensions, so we do not select any dimension, will be moving directly to expression.
19. Since we are calculating the performance, we add a division between sales and targets (Sum (sales)/sum (targets)), add label for expression (Sales Achieved) finish you will see a chart like below.



20. Let us make it more readable, right click, choose the properties, navigate to presentation add a segment in segment setup (we can go with 2 segments as well for clear user readability we choose 3) and change the colors of each segment by choosing color band (Red, Yellow, Green).
21. Now look for Show scale section left –below to segment setup add margin units a 4 and show labels on every major entry margin units 1 (Select the check box Show scale if it is not selected).
22. Now Switch to Number tab since we are representing it in percentage select the radio button Fixed to and select the check box Show in percentage and click apply.



Big Data Analytics and Visualization Lab

23. In order to see the needle value (Speedometer current value). Right click on the chart and select properties move to presentation tab and select Add button next to Text in Chart (located below right corner of the window) and add this formula($=\text{num}(\text{Sum}(\text{Sales})/\text{sum}(\text{Target}), '###\%')$), and click apply, value appears on the top left corner of the chart. Press CTL +SHIFT to drag the text anywhere in the chart. Final Gauge chart with data along with chart looks like below.



CREATING A STORY WITH TABLEAU PUBLIC

With Tableau public, you are able to organize your data in order to tell a meaningful story. This is beneficial when you are doing a presentation, creating an article, or uploading to a website, as it helps your audience understand your data. Stories are created through assembling the different worksheets and dashboards. We can highlight important data points, add text box and pictures to help convey our story. We will use our health expenditure worksheets to create a tailoring in story and illustrate the changes in Canada's spending in a meaningful way. To begin, select "New Story" at the bottom right of your screen