

Big Data Analytics and Visualization Lab

A Project Report Submitted in Fulfillment

of the Degree of

MASTER

In

COMPUTER APPLICATION

Year 2022-2023

By

Mr. GUPTA SUDHIR PRAHLAD SABITREE

(Seat No-806061)

(Application Id-171010)

Under the Guidance of

Prof. Bharati



Institute of Distance and Open Learning

Vidya Nagari, Kalina, Santacruz East – 400098.

University of Mumbai

PCP Center

Satish Pradhan Dnyanasadhana College,

Thane.



Institute of Distance and Open Learning

Vidya Nagari, Kalina, Santacruz East – 400098.

CERTIFICATE

This is to certify that, this project report entitled "**Big Data Analytics and Visualization Lab**" is a record of work carried out by **Mr. GUPTA SUDHIR PRAHLAD SABITREE** (Seat no-**806061**), student of **MCA semester-III** class and is submitted to University of Mumbai, in partial fulfilment of the requirement for the award of the degree of **Master in Computer Application**. The project report has been approved.

Guide

External Examiner

Coordinator – M.C.A

Declaration

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)

Mr. GUPTA SUDHIR PRAHLAD SABITREE

Seat No-806061

Date:

Place:

ACKNOWLEDGMENT

After the completion of this work, words are not enough to express my feelings about all those who helped me to reach my goal; feeling above this is my indebtedness to the almighty for providing me this moment in my life.

It's a great pleasure and moment of immense satisfaction for me to express my profound gratitude to my project guide, **Prof. Bharati** whose constant encouragement enabled me to work enthusiastically. His perpetual motivation, patience and excellent expertise in discussion during progress of dissertation work have benefited me to an extent, which is beyond expression. His depth and breadth of knowledge of Engineering field made me realize that theoretical knowledge always help to develop efficient operational software, which is a blend of all core subjects of the field. The completion of this project would not have been possible without his encouragement, patient guidance and constant support.

I would like to thank all staff members for their valuable cooperation and permitting me to work in the computer labs.

Special thanks to my colleagues and friends for providing me useful comments, suggestions and continuous encouragement.

Finally, I thanks my family members, for their support and endurance during this work.

Mr. Gupta Sudhir Prahlad Sabitree

(Seat No:806061)

Index

Practical No	Details
1	Install, configure and run Hadoop and HDFS ad explore HDFS.
2	Implement word count / frequency programs using MapReduce
3	Mongo DB: Installation and Creation of database and Collection CRUD Document: Insert, Query, update and Delete Document
4	Hive: Introduction Creation of Database and Table, Hive Partition, Hive Built in Function and Operators, Hive View and Index.
5	Visualization: Connect to data, Build Charts and Analyze Data, Create Dashboard, Create Stories using Tableau.

Practical No. 1

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

Step 1: Download and install VirtualBox

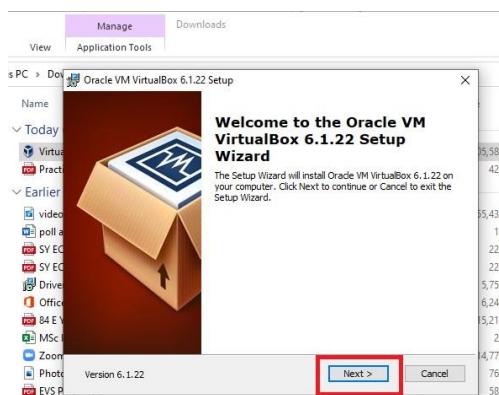
Go to the website of Oracle VirtualBox and get the latest stable version from the following site

<https://www.virtualbox.org/>

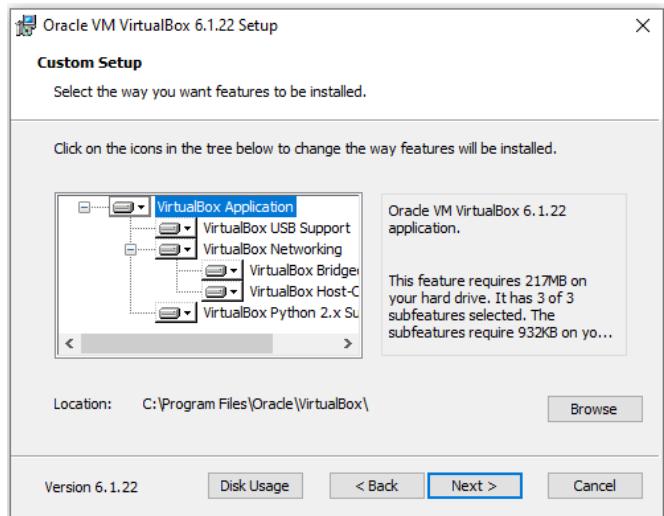
click on ‘Download’



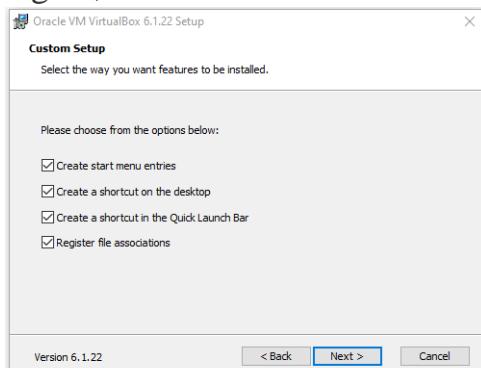
You will get VirtualBox-6.1.22-144080-Win.exe file downloaded. Double click and run it. Click on next.



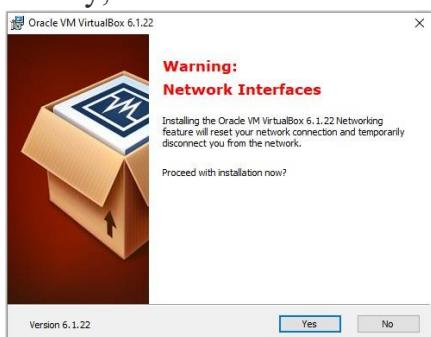
Click on ‘next’ without changing the default folder as shown below:



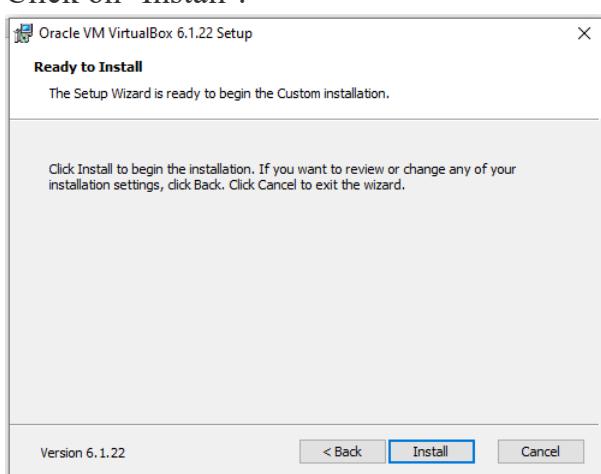
Again, click on next as shown below:



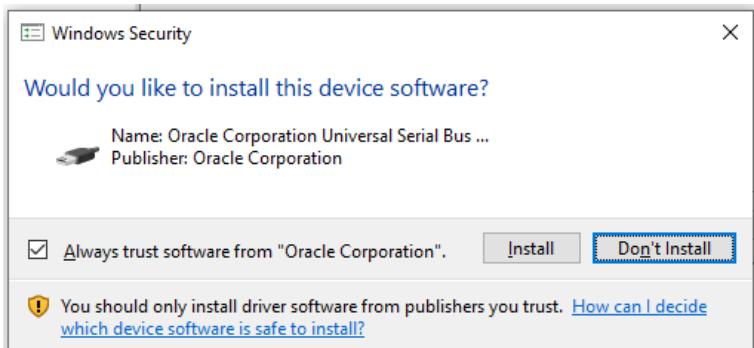
Finally, click on 'Yes'.



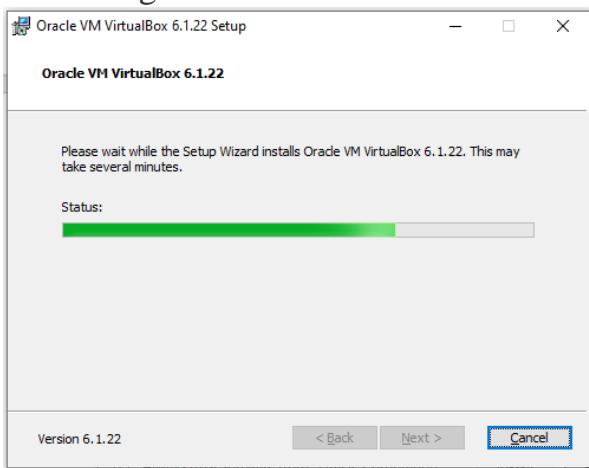
Click on 'Install'.



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



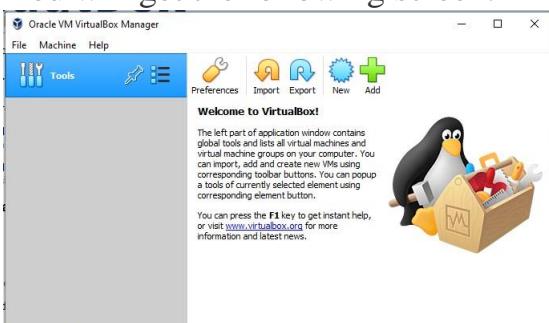
You will get the screen as shown below:



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



You will get the following screen:



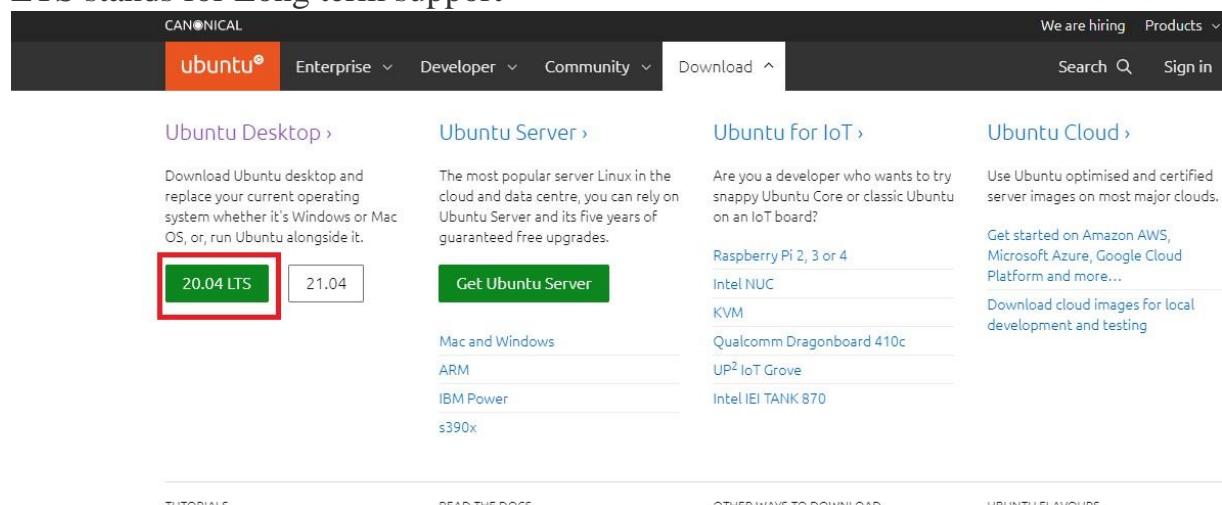
Step 2: download Ubuntu

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:

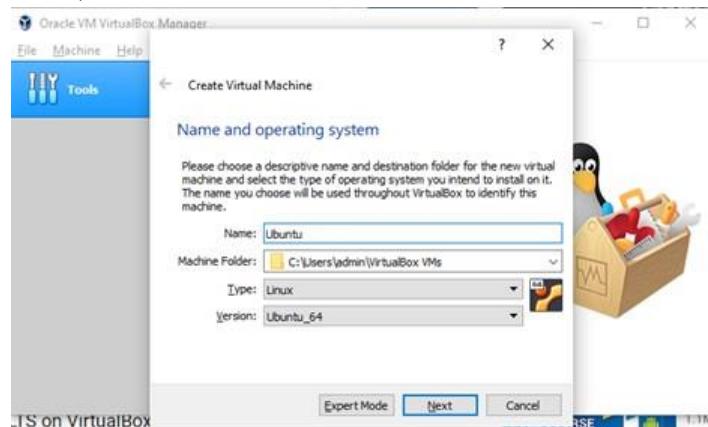
LTS stands for Long term support



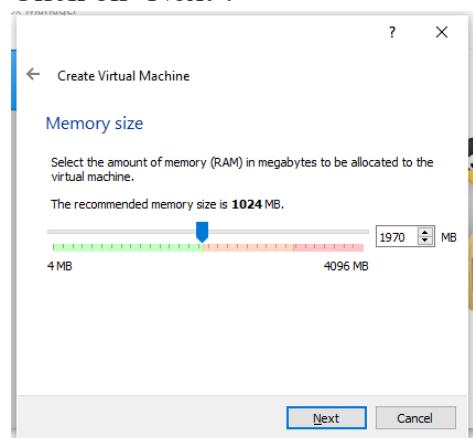
The screenshot shows the Canonical Ubuntu website. The navigation bar includes links for CANONICAL, ubuntu®, Enterprise, Developer, Community, Download, We are hiring, Products, Search, and Sign in. Below the navigation, there are four main sections: Ubuntu Desktop, Ubuntu Server, Ubuntu for IoT, and Ubuntu Cloud. The Ubuntu Desktop section features a '20.04 LTS' button, which is highlighted with a red box. Other options in this section include '21.04', 'Ubuntu Server', 'Ubuntu for IoT', and 'Ubuntu Cloud'. The Ubuntu Server section includes links for Mac and Windows, ARM, IBM Power, and s390x. The Ubuntu for IoT section includes links for Raspberry Pi 2, 3 or 4, Intel NUC, KVM, Qualcomm Dragonboard 410c, UP2 IoT Grove, and Intel IEI TANK 870. The Ubuntu Cloud section includes links for Amazon AWS, Microsoft Azure, Google Cloud Platform, and more, along with a link to download cloud images for local development and testing.

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

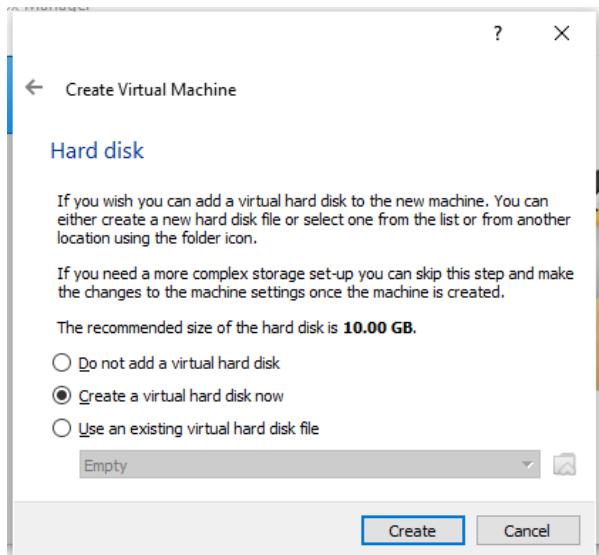
Now, click on 'New' to virtualbox and write Name as 'Ubuntu' as shown below:



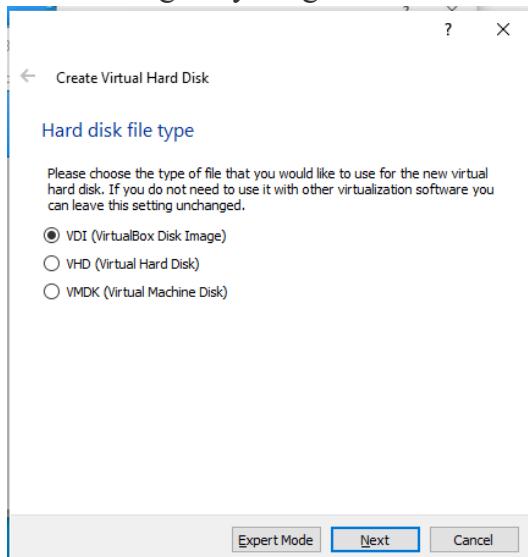
Click on 'Next'.



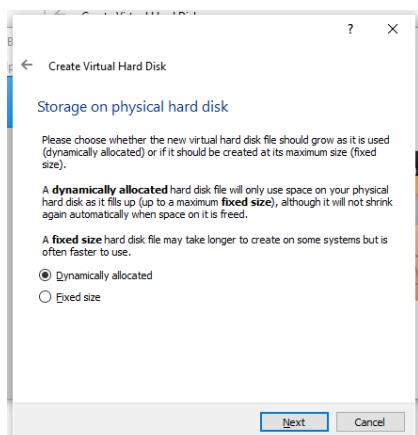
Here, you allow memory size up to green indicator (1970 MB).
Click on ‘Next’.



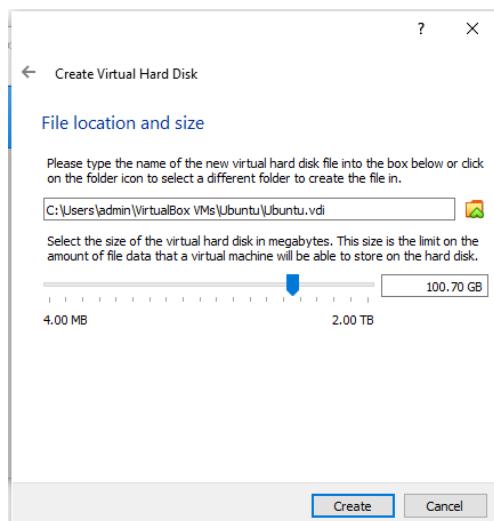
Don’t change anything in this screen and click on ‘Create’.



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

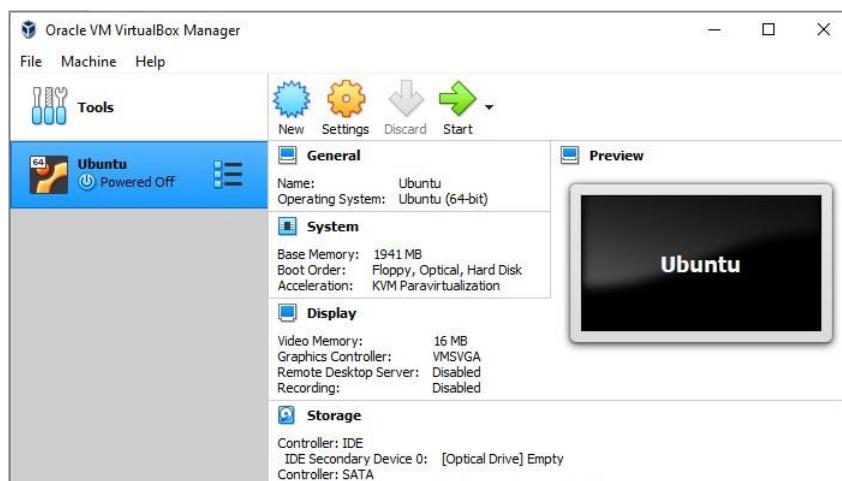


Keep this screen also as it is and click on ‘Next’.



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

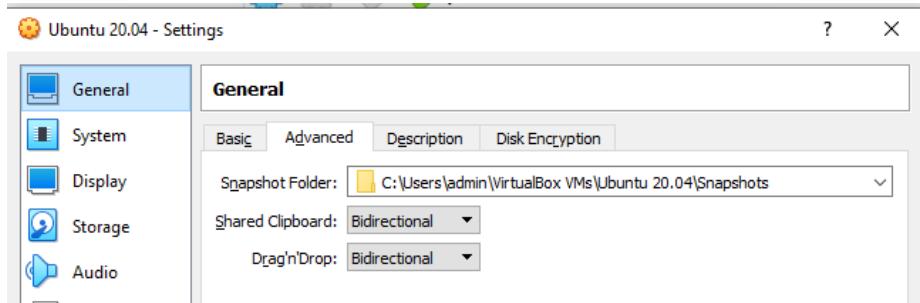
You may see the following screen having Ubuntu on Virtual Machine.



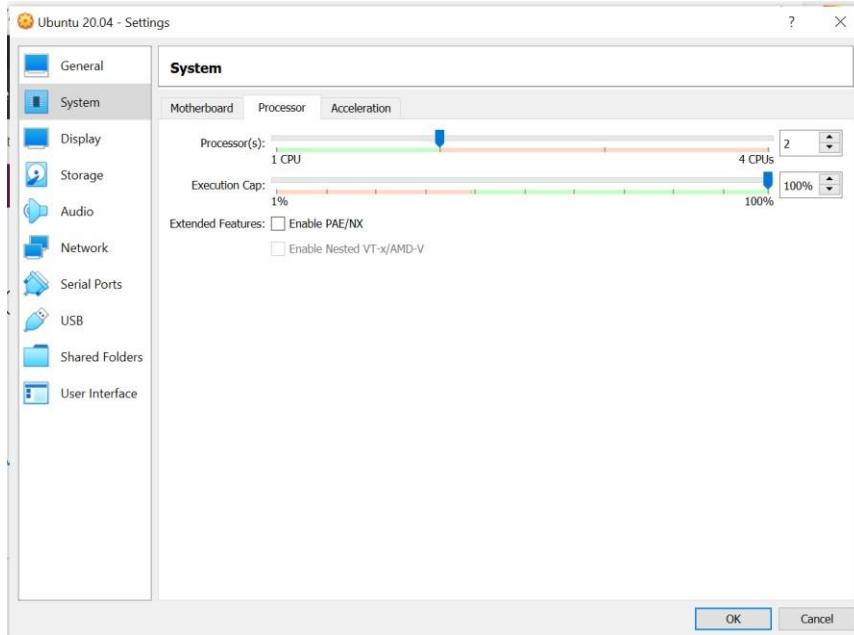
Select ‘General’ -> ‘Basic’ as shown below:

You may change the name from Ubuntu to Ubuntu 20.04

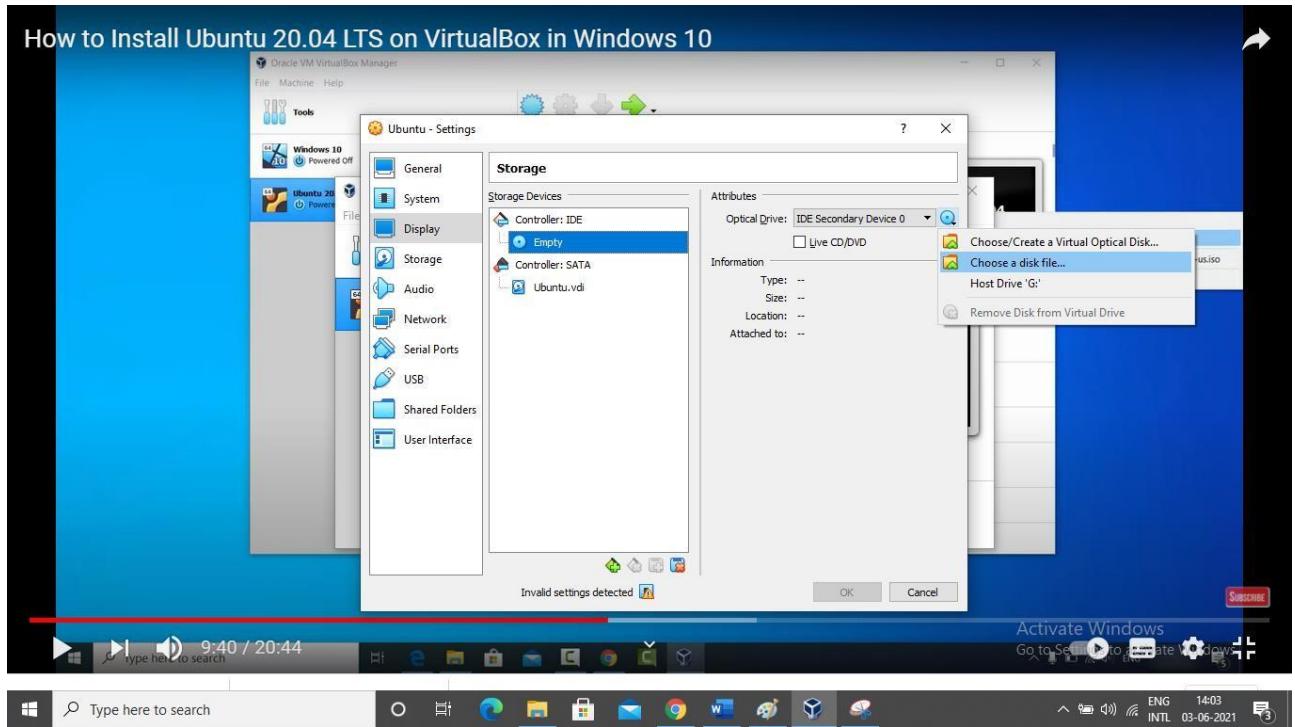
Click on Ubuntu and then click on settings option as shown below: Select bidirectional in ‘General’->‘Advanced’ as shown below:



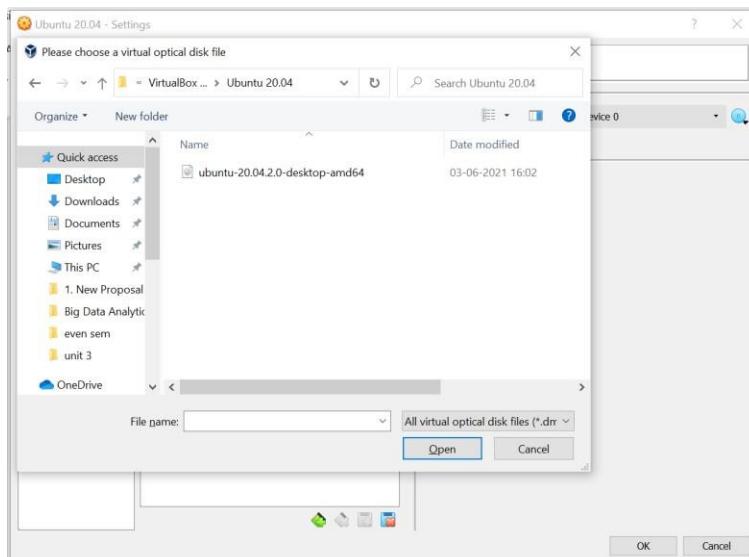
Goto 'System' option and change the processor up to green bar, usually 4. (if it allows)



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 diskfile' as shown below:

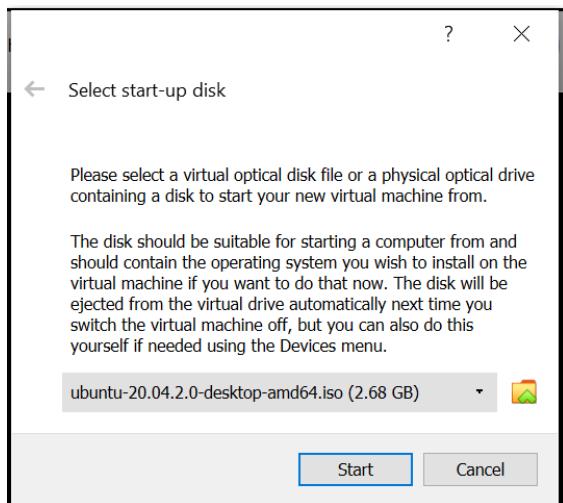


Browse the folder where you have selected ubuntu iso file.

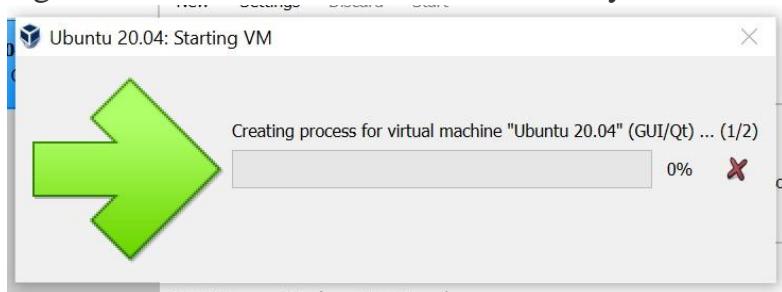


Click on Ubuntu....iso file and click on open and then click on ok.

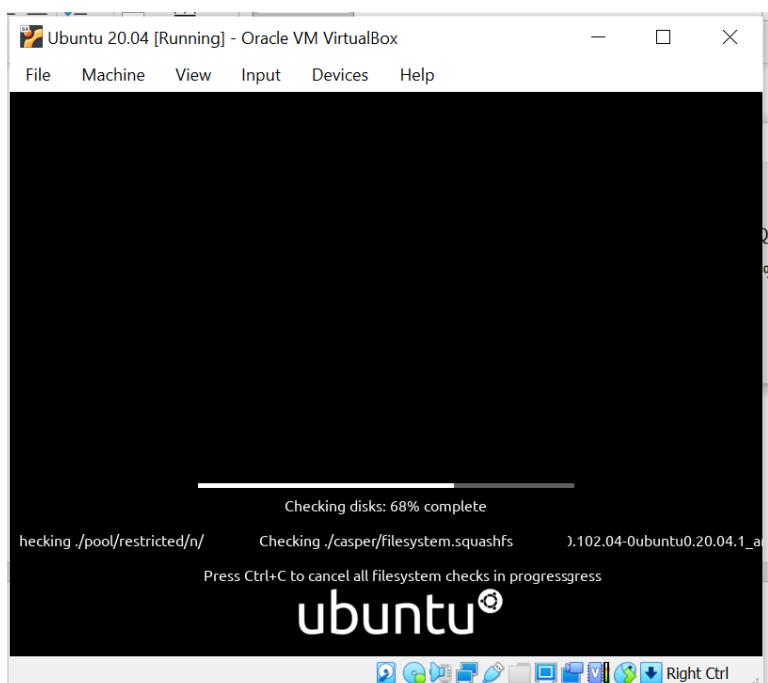
Click on Ubuntu -> start button.



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

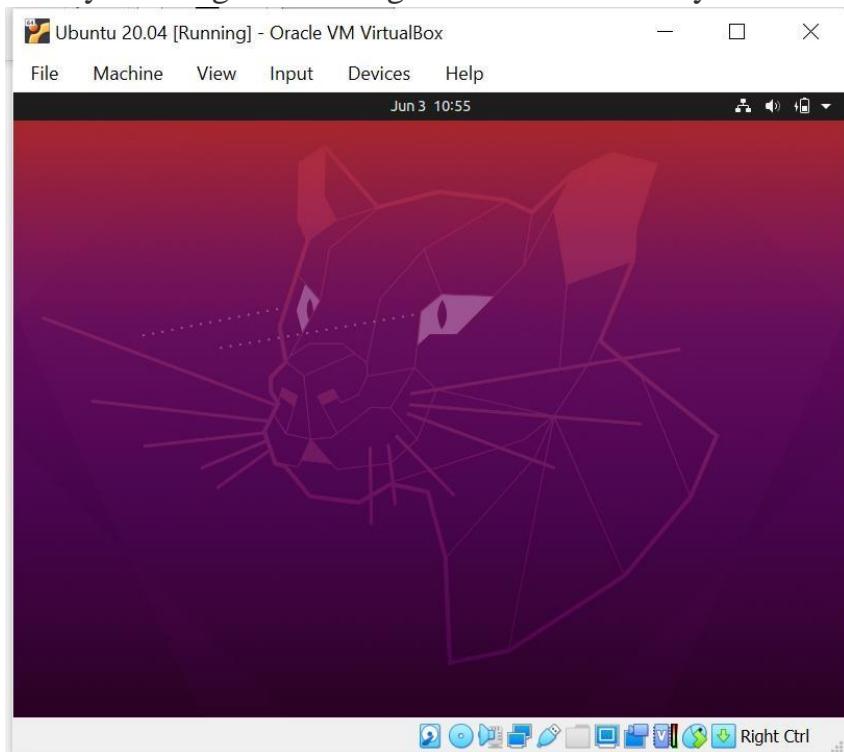


And simultaneously one more screen as follows:



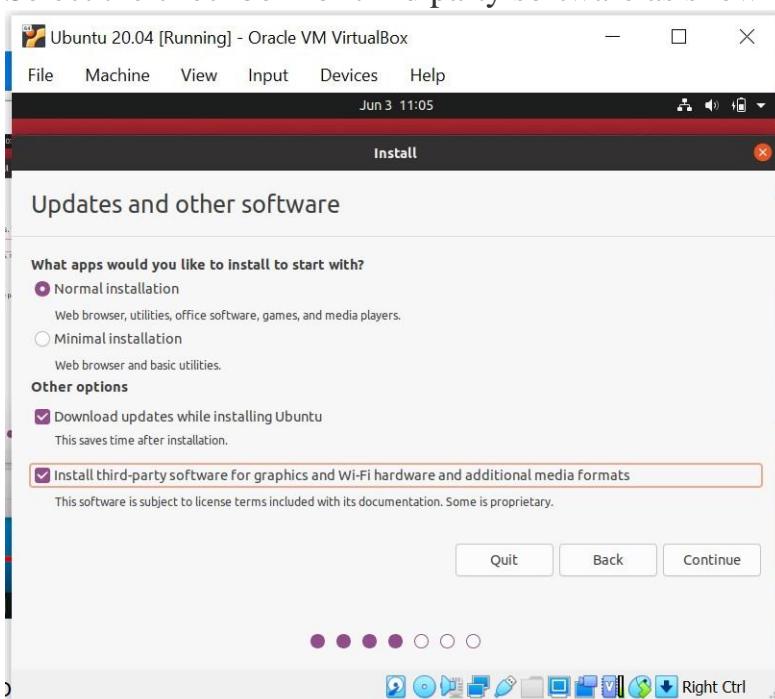
Keep on closing all warnings.

Next you will get following screen automatically.

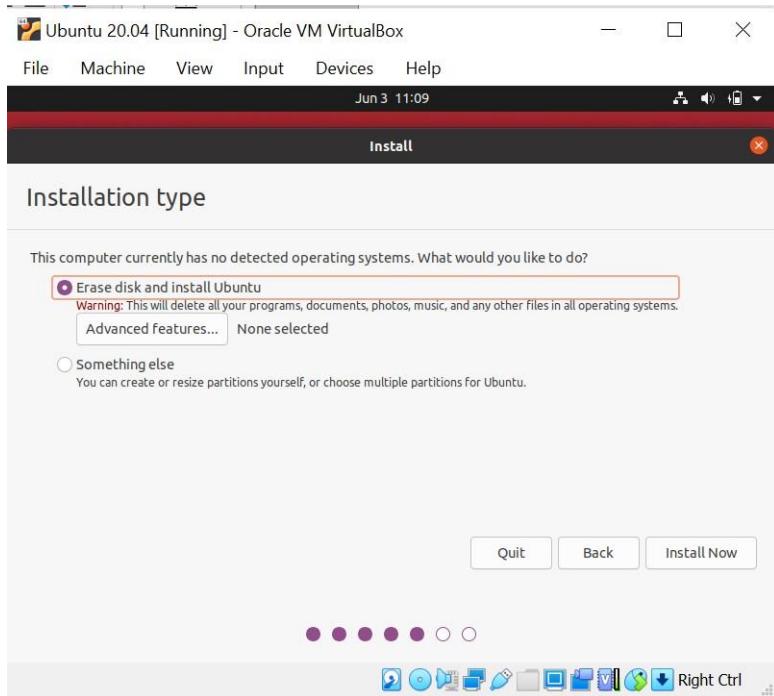


Select language -> English and click on 'Install Ubuntu' in 'KeyboardLayout' screen, select 'English UK'. Click on 'Continue'.

Select the checkbox for third party software as shown below:

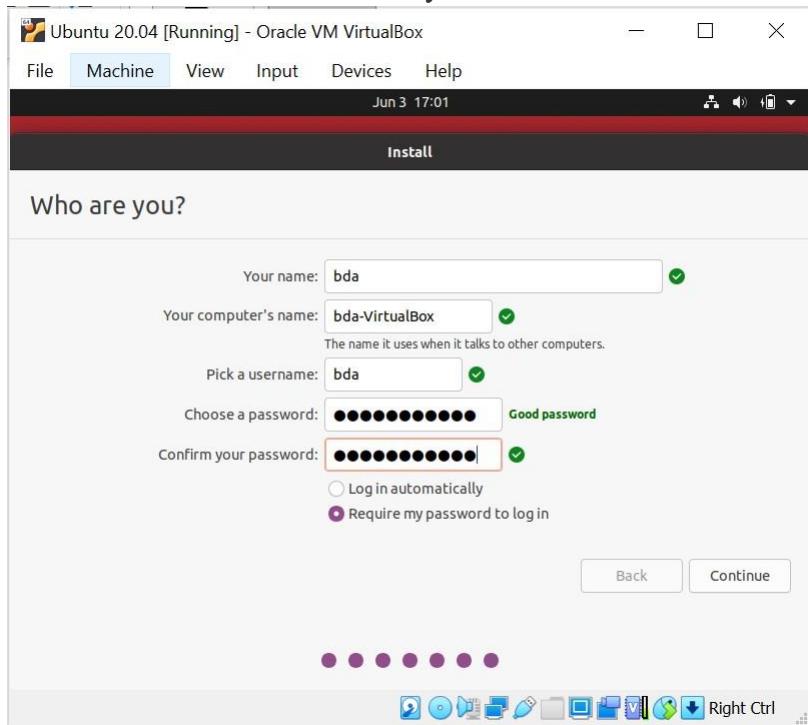


Click on 'continue'.

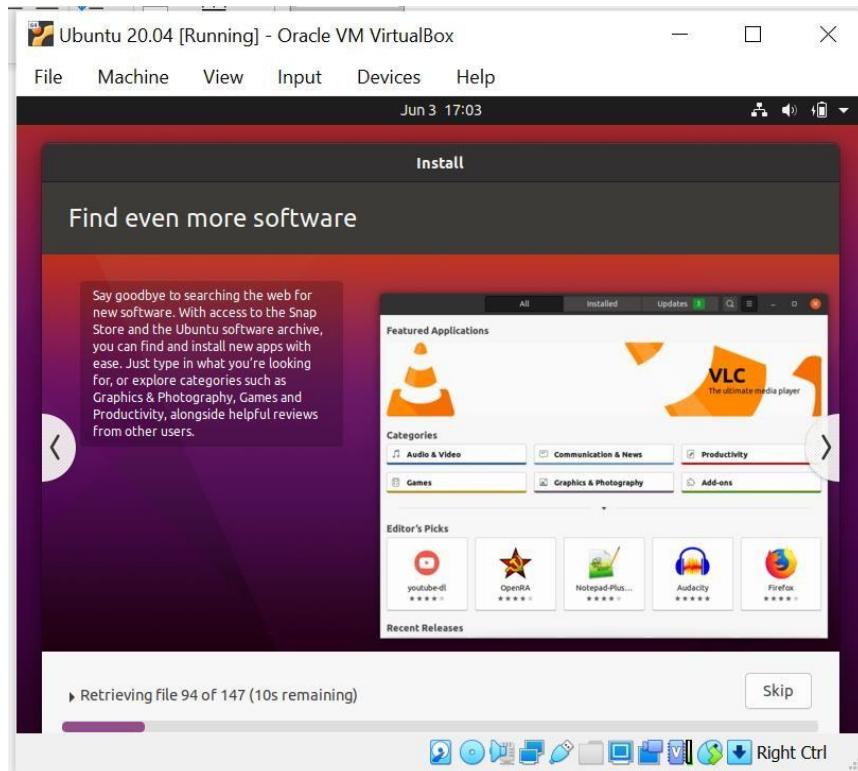


Select Erase disk and Install Ubuntu and click on 'Install Now'. Click on 'Continue' on the next screen.

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



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



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

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

```
$ sudo apt update  
$ sudo apt install default-jdk  
$ ava -version'
```

```
$ wget https://hadoop.apache.org/release/3.2.2.html/hadoop-3.2.2.tar.gz
$ tar xzvf hadoop-3.2.2.tar.gz
$ sudo mv hadoop-3.2.2 /usr/local/hadoop
$ readlink -f /usr/bin/java | sed "s:bin/java::"
```

: Configuring Hadoop's Java Home; To begin, open hadoop-env.sh

```
$ sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh
```

File will be opened. Add the following line at the end of .sh file
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/

to save the changes in the file, press ctrl and x
together. then press Y

then press Enter key

then apply following commands:

```
$ /usr/lib/jvm/java-11-openjdk-amd64/
```

Step 4: Running Hadoop

```
$ /usr/local/hadoop/bin/hadoop
$ mkdir ~/input
$ cp /usr/local/hadoop/etc/hadoop/*.xml ~/input
```

We can use the following command to run the MapReduce hadoop mapreduce-examples program, a Java archive with several options. We'll invoke its grep program, one of the many examples included in hadoop-mapreduce-examples, followed by the input directory, input and the output directory grep_example. The MapReduce grep program will count the matches of a literal word or regular expression. Finally, we'll supply the regular expression allowed[.]^{*} to find occurrences of the word allowed within or at the end of a declarative sentence. The expression is case-sensitive, so we wouldn't find the word if it were capitalized at the beginning of a sentence:

```
$ /usr/local/hadoop/bin/hadoop jar  
/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-  
3.2.2.jar grep ~/input ~/grep_example 'allowed[.]*'<br/>$ cat ~/grep_example/*
```

Practical No. 2

Aim: Implement word count / frequency programs using MapReduce.

We have to write the splitting parameter, Map function logic and Reduce function logic. The rest of the remaining steps will execute automatically.

Prerequisites:

1. Hadoop-3.3.1
2. JDK 8
3. Eclipse

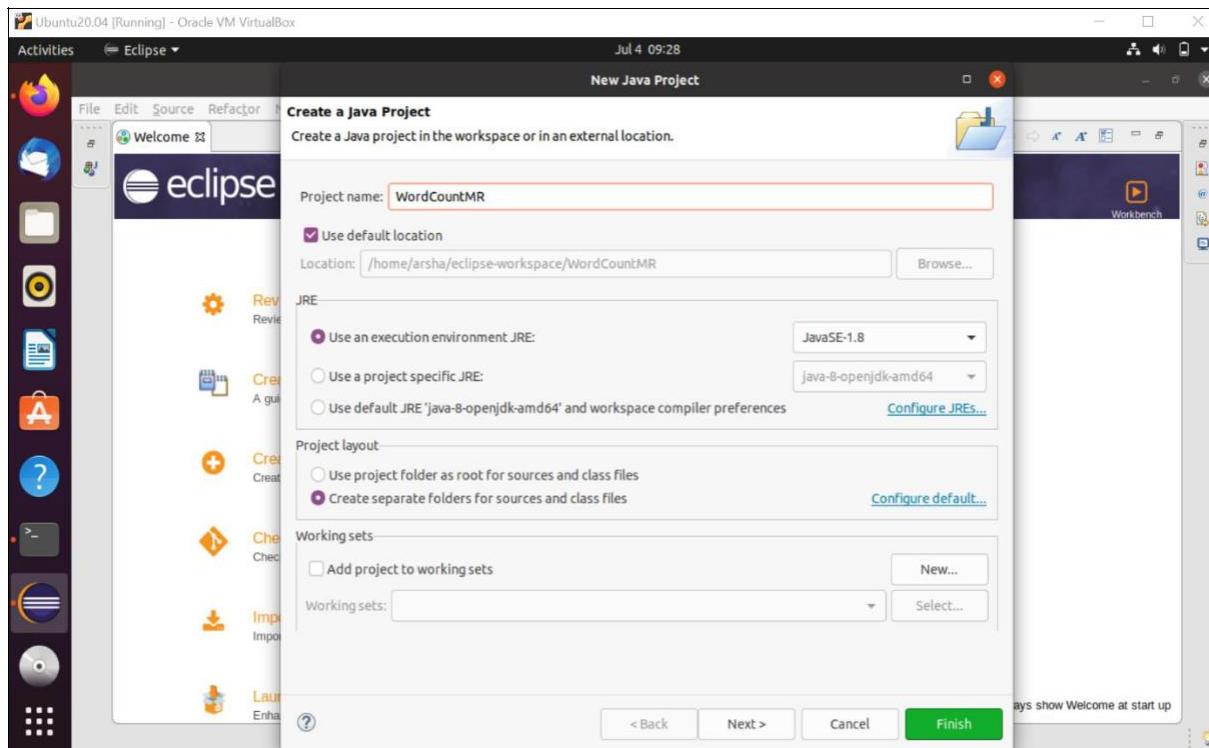
We have already installed Hadoop and Java successfully in the first practical.

To install Eclipse on ubuntu 20.04, go to the terminal and run command

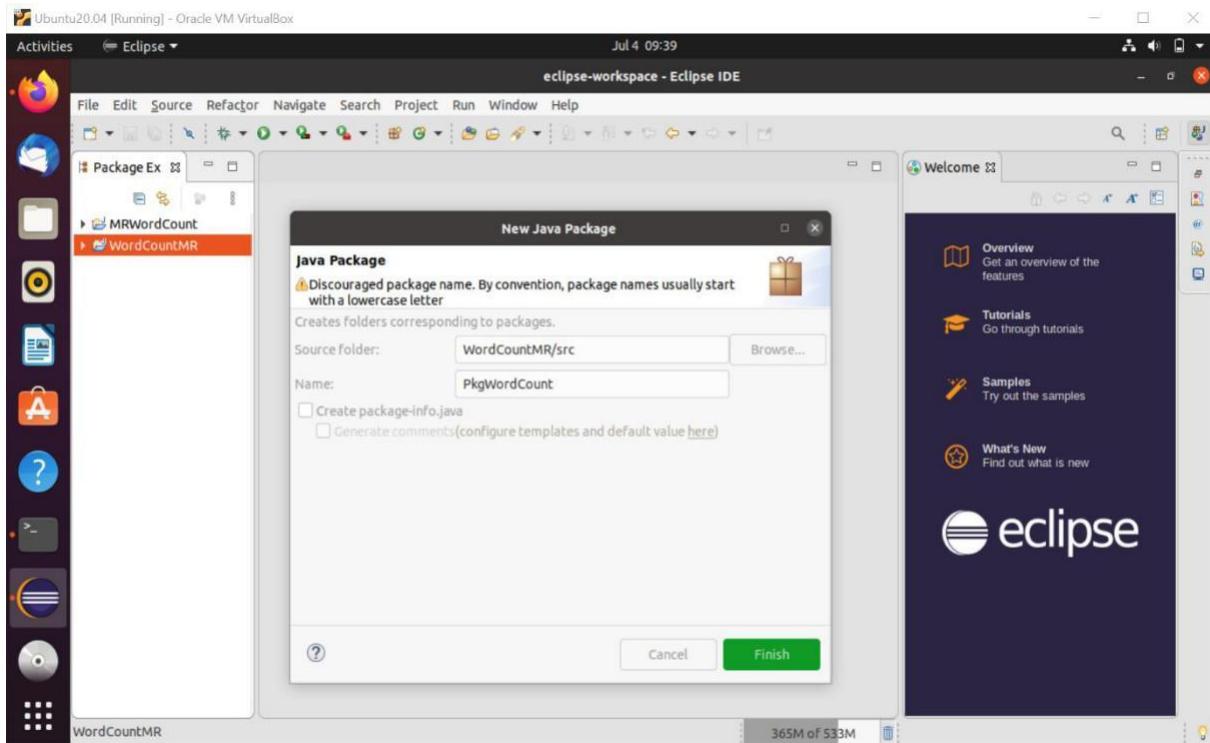
```
$ sudo snap install --classic eclipse
```

Steps

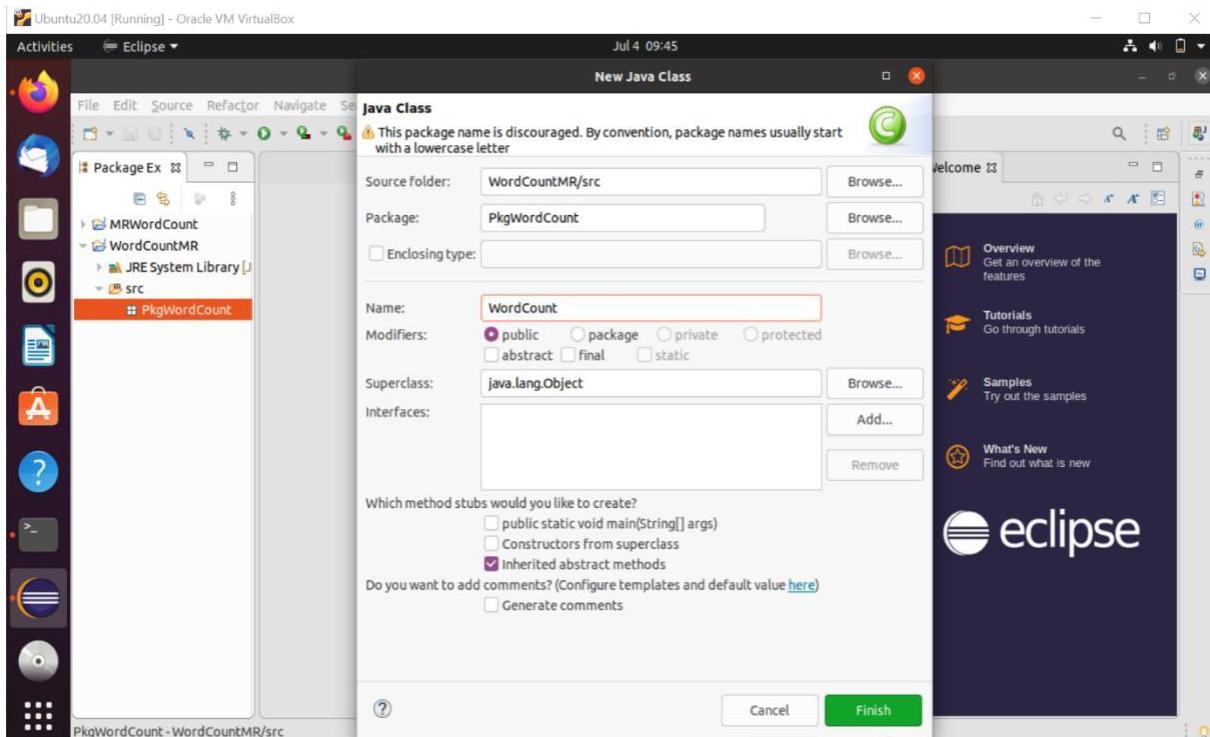
1. Open Eclipse > File > New > Java Project > (Name it - WordCountMR) > Finish.



2. Right Click on project WordCountMR > New > Package (Name it - PkgWordCount) > Finish.



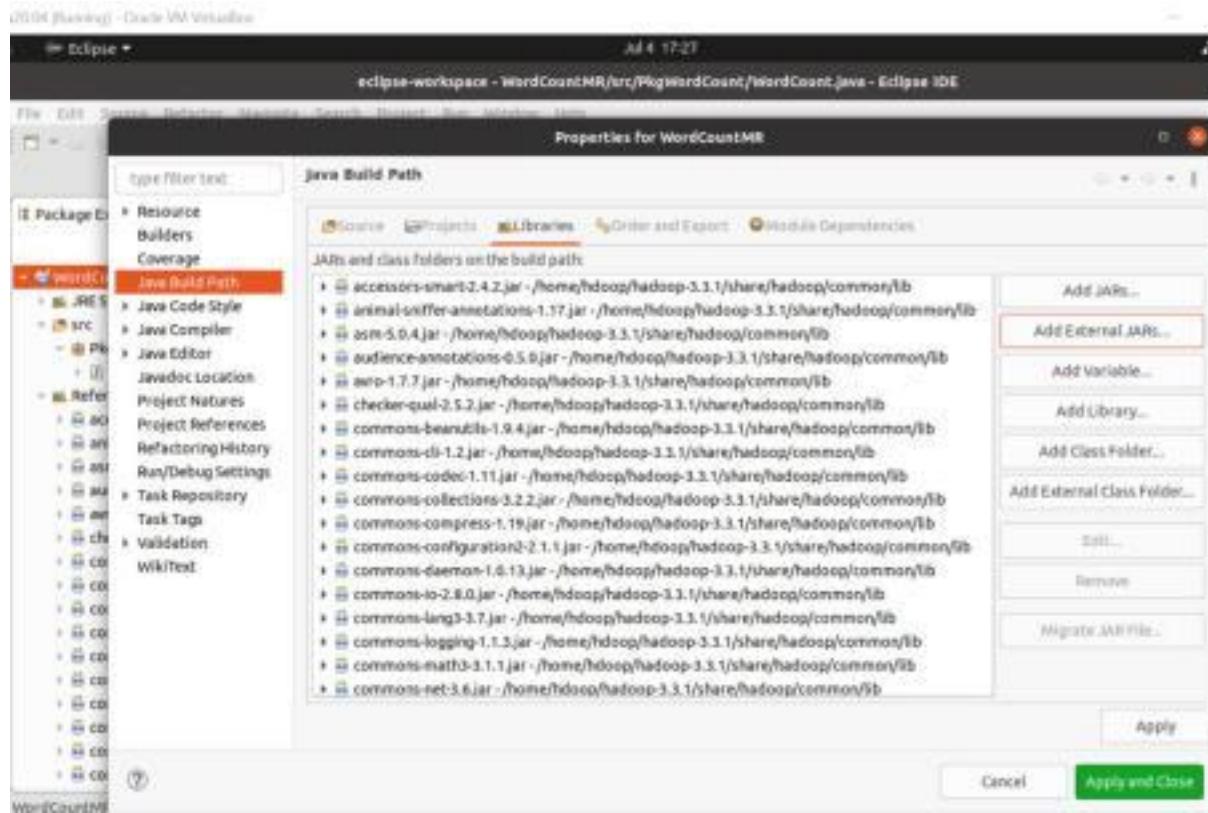
3. Right Click on Package PkgWordCount > New > Class (Name it - WordCount).



4. Add Following Reference Libraries:

Right Click on Project > Configure Build Path> Add External JARs

- All jar under /hadoop-3.3.1/share/hadoop/mapreduce
- All jar files under /hadoop-3.3.1/share/hadoop/common/lib
- /hadoop-3.3.1/share/hadoop/common/hadoop-common-3.3.1.jar> Apply and Close.



5. Type the following code:

```
package PkgWordCount;

import java.io.IOException;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
```

```
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text; import
org.apache.hadoop.mapreduce.Job; import
org.apache.hadoop.mapreduce.Mapper; import
org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat; import
org.apache.hadoop.mapreduce.lib.output.FileOutputFormat; import
org.apache.hadoop.util.GenericOptionsParser;

public class WordCount{

    @SuppressWarnings("deprecation")

    public static void main(String [] args) throws Exception{
        Configuration c=new Configuration();

        String[] files=new
GenericOptionsParser(c,args).getRemainingArgs();

        Path input=new Path(files[0]); Path output=new
Path(files[1]); Job j=new Job(c,"wordcount");
j.setJarByClass(WordCount.class);
j.setMapperClass(MapForWordCount.class);
j.setReducerClass(ReduceForWordCount.class);
j.setOutputKeyClass(Text.class);
j.setOutputValueClass(IntWritable.class);
FileInputFormat.addInputPath(j, input);
FileOutputFormat.setOutputPath(j, output);
System.exit(j.waitForCompletion(true)?0:1);

    }

    public static class MapForWordCount extends Mapper<LongWritable,
Text, Text, IntWritable>{

        public void map(LongWritable key, Text value, Context con)
throws IOException, InterruptedException{
            String line = value.toString();
            String[] words=line.split(" ");
            for(String word: words ){

```

```

        Text outputKey = new Text(word.trim());
        IntWritable outputValue = new IntWritable(1);
        con.write(outputKey, outputValue);

    }

}

}

public static class ReduceForWordCount extends Reducer<Text,
IntWritable, Text, IntWritable>{

    public void reduce(Text word, Iterable<IntWritable> values,
Context con) throws IOException, InterruptedException{

        int sum = 0;
        for(IntWritable value : values){
            sum += value.get();

        }
        con.write(word, new IntWritable(sum));
    }

}

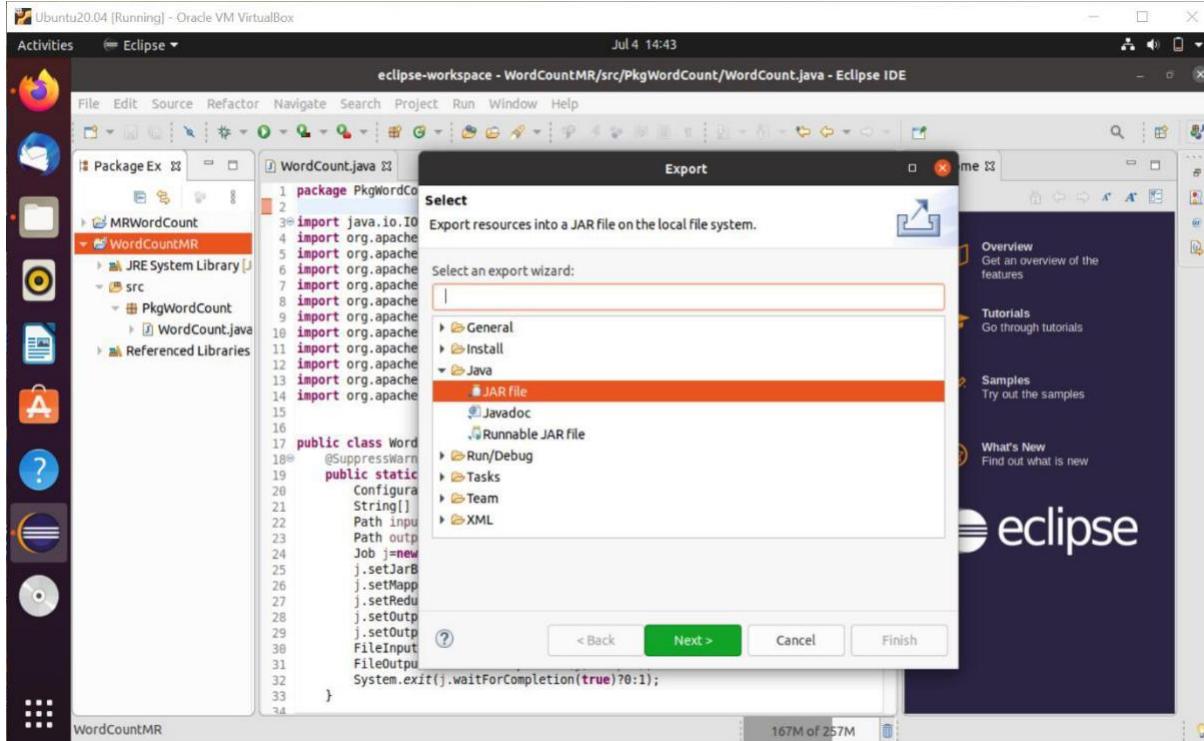
```

The above program consists of three classes:

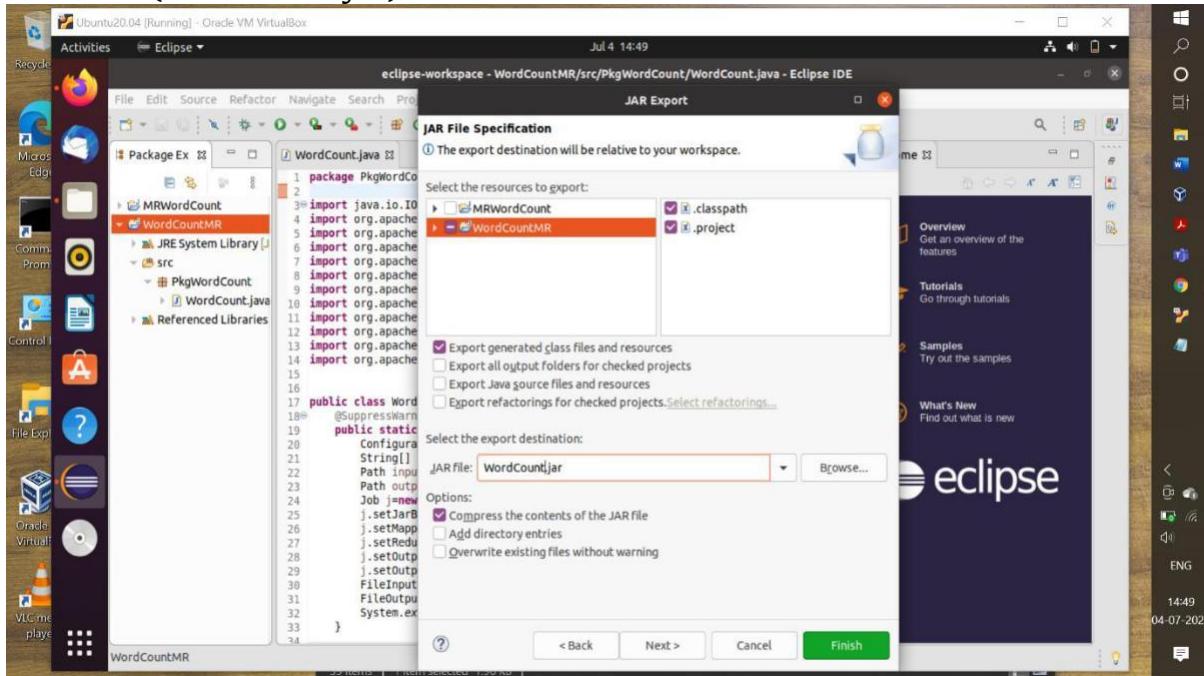
- Driver class (Public, void, static, or main; this is the entry point).
- The **Map** class which extends the public class **Mapper<KEYIN,VALUEIN,KEYOUT,VALUEOUT>** and implements the **Map** function.
- The **Reduce** class which extends the public class **Reducer<KEYIN,VALUEIN,KEYOUT,VALUEOUT>** and implements the **Reduce** function.

6. Make a jar file

- Right Click on Project > Export
- Expand Java tab > JAR file > Next
-



- Select the resources to export WordCountMR > Name the JAR file (WordCount.jar) > Finish



7. Create a text file (data.txt)

- By default, when we right-click inside anywhere in Ubuntu Nautilus file manager, it will not give us the “New document” option. Thus, to get this missing option, we need to run a command.
- Open Ubuntu command terminal. Run command- touch ~/Templates/Text\ document
- Now, go to your Desktop > Right-click > New document > Text document (Name it – data.txt) and Add some words in it.

8. Now put this data.txt in HDFS.

- Open terminal and login with your hadoop user and change directory to hadoop-3.3.1/sbin. Run command-

```
cd hadoop-3.3.1/sbin
```

- Now start DFS and yarn services. Run commands –

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

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

- Now to check all your daemons are running properly or not run

```
command –  
jps
```

The screenshot shows a terminal window titled "hadoop@arsha-VirtualBox: ~/hadoop-3.3.1/sbin". The terminal output is as follows:

```
hadoop@arsha-VirtualBox:~$ cd hadoop-3.3.1/sbin
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ ./start-dfs.sh
Starting namenodes on [0.0.0.0]
Starting datanodes
Starting secondary namenodes [arsha-VirtualBox]
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ jps
4147 NodeManager
4500 Jps
3556 DataNode
3414 NameNode
4011 ResourceManager
3758 SecondaryNameNode
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$
```

- Now create input directory in HDFS. Run command -
hdfs dfs -mkdir /input
- Now copy your data.txt to this input directory. Run command -
hdfs dfs -put /home/arsha/Desktop/data.txt /input
- Now to check the content of this data.txt, run command -
hdfs dfs -cat /input/data.txt

```
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ hdfs dfs -mkdir /input
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ hdfs dfs -put /home/arsha/Desktop/da
ta.txt /input
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ hdfs dfs -cat /input/data.txt
Hello wordcount MapReduce Hadoop program.
This is my first MapReduce program.
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$
```

9. Now Run WordCount.jar that we have created earlier to see the output.
For this command format will be

'hadoop jar <your jar file location> <packagename.classname> /<text
file location in HDFS> /<HDFS output directory name>'

There is no need to create output directory in HDFS

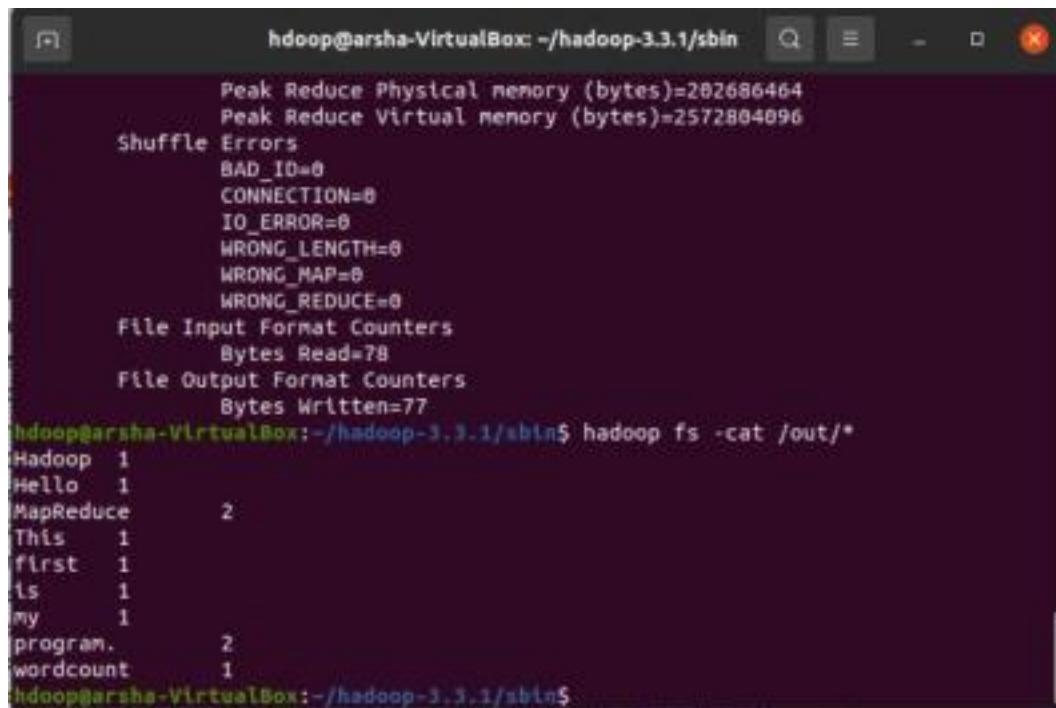
Run command -

Hadoop jar /home/arsha/eclipse-workspace/WordCount.jar
PkgWordCount.WordCount /input /out

```
[Running] - Oracle VM VirtualBox
Terminal Jul 4 17:10
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ hadoop jar /home/arsha/eclipse-workspace/WordCount.jar PkgWordCount.WordCount
/input /out
2021-07-04 17:08:51,502 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-07-04 17:08:52,231 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hd
oop/.staging/job_1625391968675_0001
2021-07-04 17:08:52,561 INFO input.FileInputFormat: Total input files to process : 1
2021-07-04 17:08:53,048 INFO mapreduce.JobSubmitter: number of splits:1
2021-07-04 17:08:53,218 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1625391968675_0001
2021-07-04 17:08:53,218 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-07-04 17:08:53,478 INFO conf.Configuration: resource-types.xml not found
2021-07-04 17:08:53,478 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-07-04 17:08:53,808 INFO impl.YarnClientImpl: Submitted application application_1625391968675_0001
2021-07-04 17:08:53,882 INFO mapreduce.Job: The url to track the job: http://arsha-VirtualBox:8088/proxy/application_16253
91968675_0001/
2021-07-04 17:08:53,883 INFO mapreduce.Job: Running job: job_1625391968675_0001
2021-07-04 17:09:05,092 INFO mapreduce.Job: Job job_1625391968675_0001 running in uber mode : false
2021-07-04 17:09:05,096 INFO mapreduce.Job: map 0% reduce 0%
2021-07-04 17:09:11,217 INFO mapreduce.Job: map 100% reduce 0%
2021-07-04 17:09:17,281 INFO mapreduce.Job: map 100% reduce 100%
2021-07-04 17:09:18,316 INFO mapreduce.Job: Job job_1625391968675_0001 completed successfully
2021-07-04 17:09:18,550 INFO mapreduce.Job: Counters: 54
File System Counters
FILE: Number of bytes read=150
FILE: Number of bytes written=544925
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=177
HDFS: Number of bytes written=77
HDFS: Number of read operations=8
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
HDFS: Number of bytes read erasure-coded=0
Job Counters
Launched map tasks=1
Launched reduce tasks=1
```

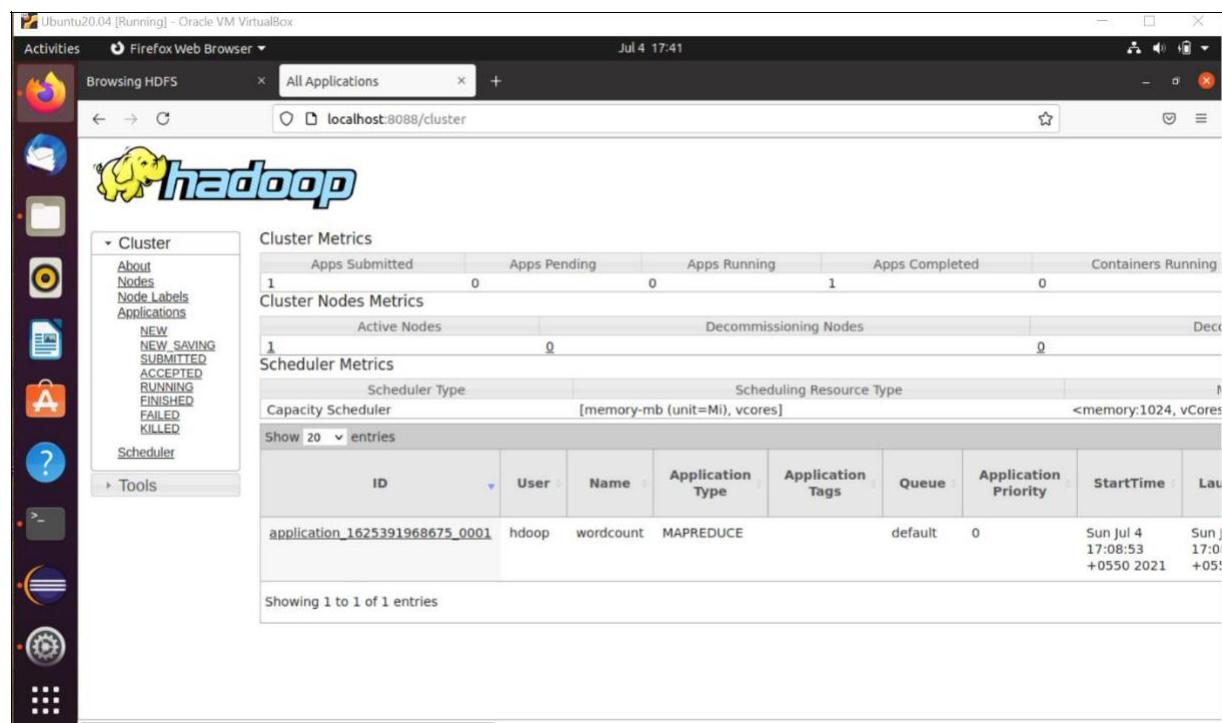
As you can see mapping and reducing job is completed successfully.

10. Now to see the output run command -
hadoop fs -cat /out/*



```
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ hadoop fs -cat /out/*
Peak Reduce Physical memory (bytes)=282686464
Peak Reduce Virtual memory (bytes)=2572804096
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=78
File Output Format Counters
  Bytes Written=77
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ hadoop fs -cat /out/*
Hadoop 1
Hello 1
MapReduce 2
This 1
first 1
is 1
my 1
program. 2
wordcount 1
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$
```

11. See the output on Browser.
• Open localhost:8088 to open Resource Manager.



The screenshot shows a Firefox browser window titled "Browsing HDFS" with the URL "localhost:8088/cluster". The page displays cluster metrics and scheduler metrics for a Hadoop cluster. On the left, there is a sidebar with navigation links for Cluster (About, Nodes, Node Labels, Applications), Scheduler (Scheduler Type: Capacity Scheduler, Scheduling Resource Type: [memory-mb (unit=Mi), vcores], Queue: default, Application Priority: 0), and Tools.

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
1	0	0	1	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Queue	Application Priority	StartTime	LastUpdate
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	default	0	Sun Jul 4 17:08:53 +0550 2021

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LastUpdate
application_1625391968675_0001	hadoop	wordcount	MAPREDUCE		default	0	Sun Jul 4 17:08:53 +0550 2021	Sun Jul 4 17:08:53 +0550 2021

Showing 1 to 1 of 1 entries

You can see 1 active node and 1 app submitted which is our wordcount mapreduce app.

- Now Open localhost:9870 and click on Browse the file system under Utilities. You can see HDFS as follows

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hdoop	supergroup	0 B	Jul 04 15:32	0	0 B	input
drwxr-xr-x	hdoop	supergroup	0 B	Jul 04 17:09	0	0 B	out
drwx-----	hdoop	supergroup	0 B	Jul 03 13:56	0	0 B	tmp

- Click input directory and you can see the data.txt

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hdoop	supergroup	78 B	Jul 04 15:32	1	128 MB	data.txt

- To see the content of file click on data.txt and then click Head the file

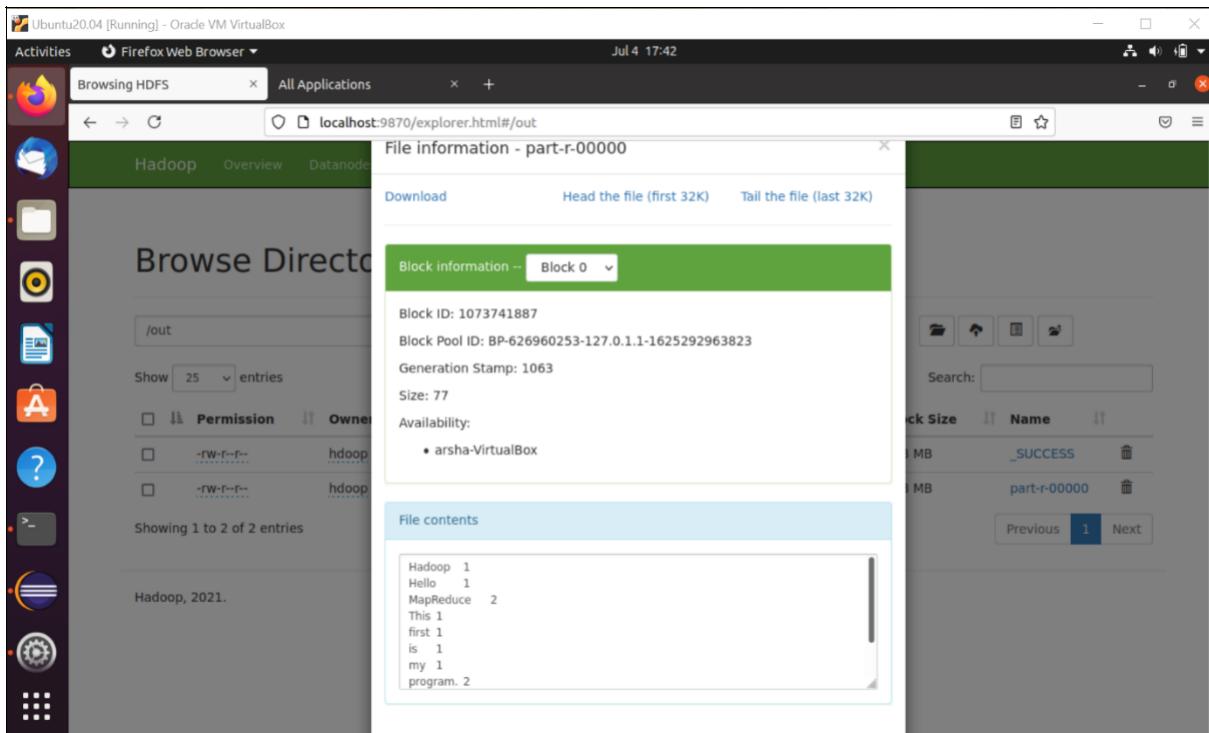
The screenshot shows a Firefox browser window titled "Browsing HDFS" at the URL "localhost:9870/explorer.html#/input". A modal dialog box is open over the page, titled "File information - data.txt". The dialog contains sections for "Download", "Head the file (first 32K)", and "Tail the file (last 32K)". Under "Block information", it shows "Block 0" with details: Block ID: 1073741880, Block Pool ID: BP-626960253-127.0.1.1-1625292963823, Generation Stamp: 1056, Size: 78, Availability: arsha-VirtualBox. The "File contents" section displays the text "Hello wordcount MapReduce program. This is my first MapReduce program.", which is also highlighted with a red box.

- To see output of wordcount program, go to out directory and click part-r-00000

The screenshot shows a Firefox browser window titled "Browsing HDFS" at the URL "localhost:9870/explorer.html#/out". The main content area is titled "Browse Directory" and shows a list of files under "/out". The table includes columns for Name, Size, Last Modified, and Block Size. Two files are listed: "_SUCCESS" (Size: 0 B) and "part-r-00000" (Size: 77 B). The "part-r-00000" file is highlighted with a red box.

Name	Size	Last Modified	Block Size
_SUCCESS	0 B	Jul 04 17:09	128 MB
part-r-00000	77 B	Jul 04 17:09	128 MB

- Click Head the file.



12. Stop all the daemons. Run command-

`./stop-all.sh`

```
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$ ./stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [0.0.0.0]
Stopping datanodes
Stopping secondary namenodes [arsha-VirtualBox]
Stopping nodemanagers
localhost: WARNING: nodemanager did not stop gracefully after 5 seconds: Trying
to kill with kill -9
Stopping resourcemanager
hadoop@arsha-VirtualBox:~/hadoop-3.3.1/sbin$
```

Practical No. 3

Aim: Mongo DB: Installation and Creation of database and Collection
CRUD Document: Insert, Query, update and Delete Document.

Requirement

- a. PyMongo
- b. Mongo Database

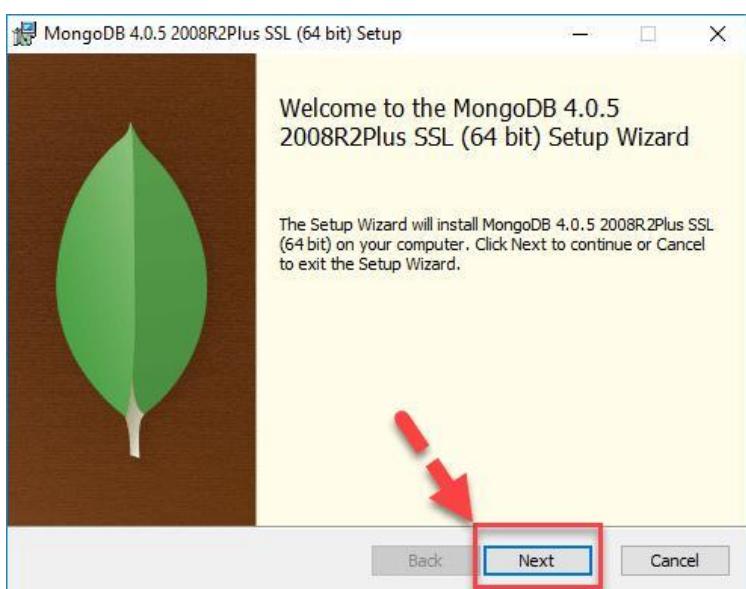
Step A: Install Mongo database

Step 1) Go to (<https://www.mongodb.com/download-center/community>) and Download MongoDB Community Server. We will install the 64-bit version for Windows.

Select the server you would like to run:

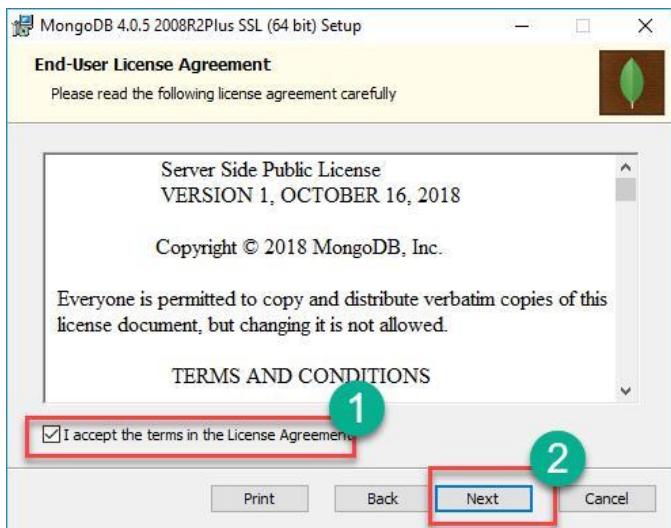


Step 2) Once download is complete open the msi file. Click Next in the start up screen

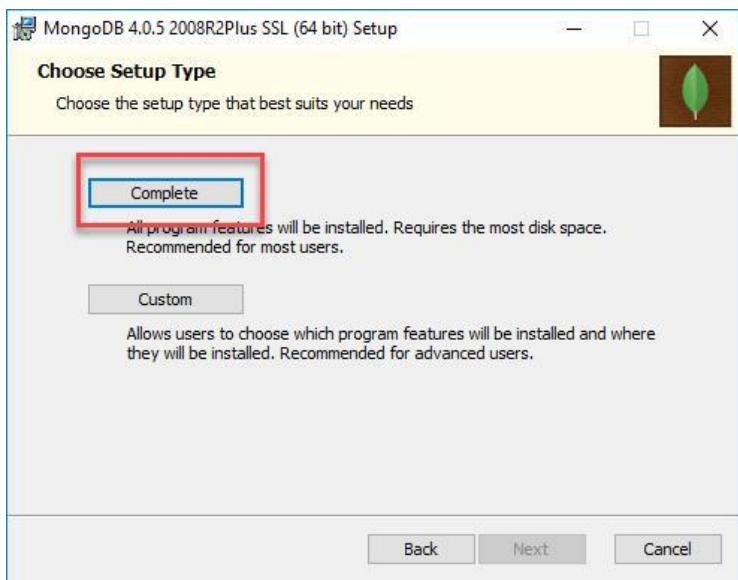


Step 3)

1. Accept the End-User License Agreement
2. Click Next

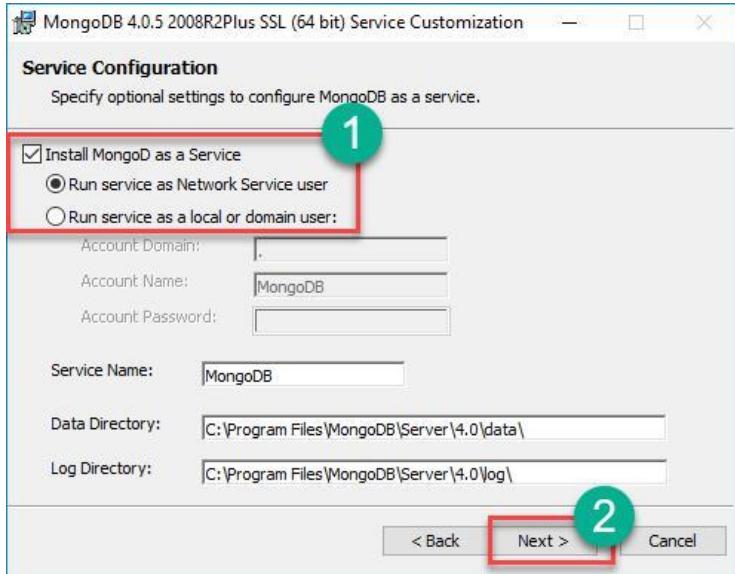


Step 4) Click on the "complete" button to install all of the components. The custom option can be used to install selective components or if you want to change the location of the installation.

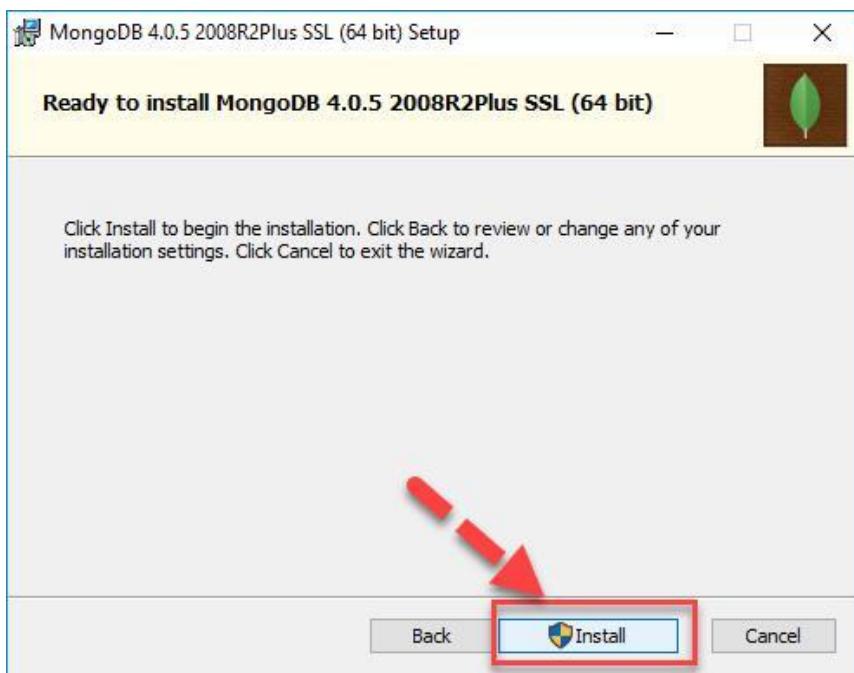


Step 5)

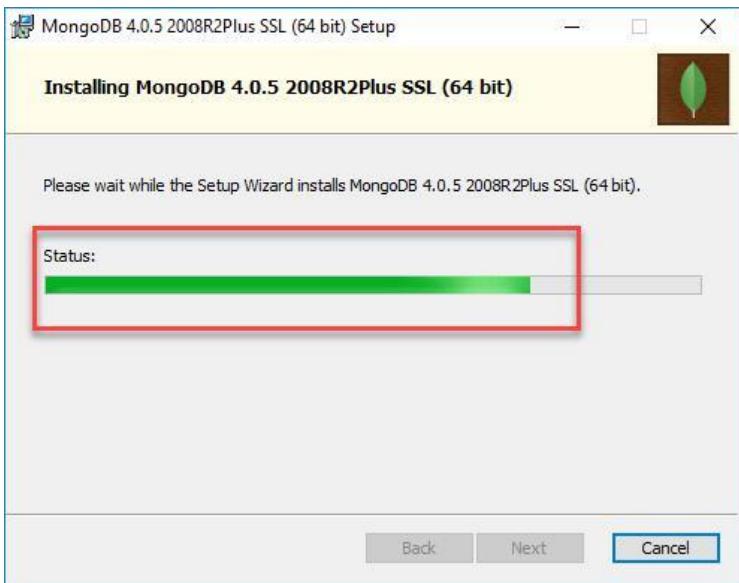
1. Select “Run service as Network Service user”. make a note of the data directory, we’ll need this later.
2. Click Next



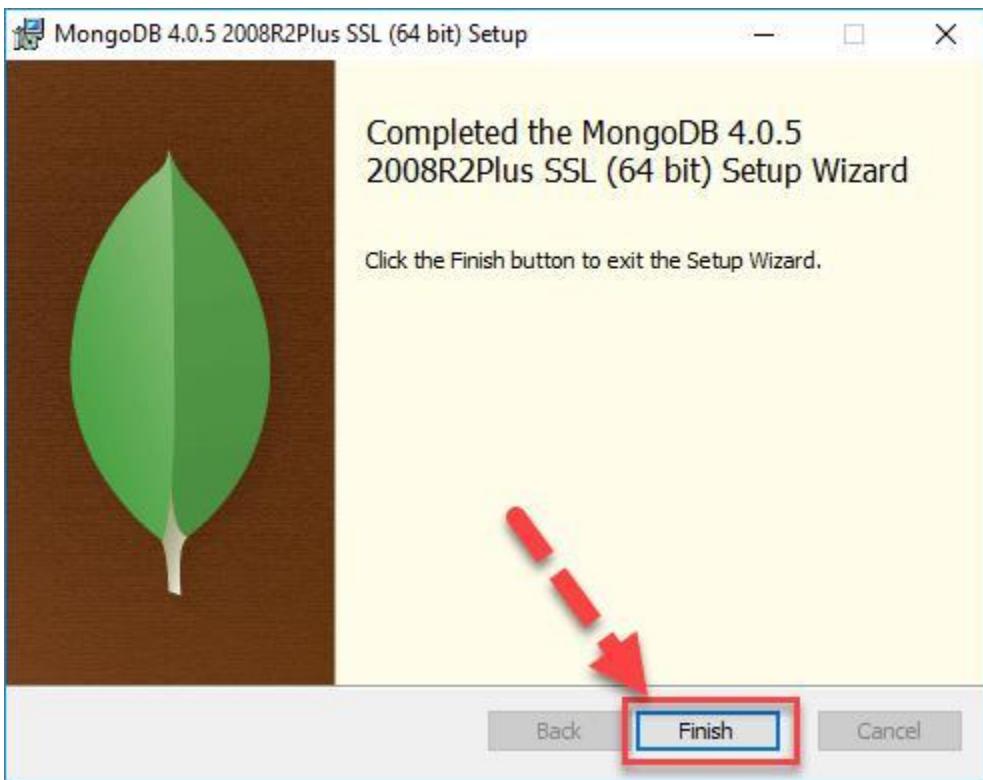
Step 6) Click on the Install button to start the installation.



Step 7) Installation begins. Click Next once completed



Step 8) Click on the Finish button to complete the installation



Test Mongodb

Step 1) Go to " C:\Program Files\MongoDB\Server\4.0\bin" and double click on mongo.exe. Alternatively, you can also click on the MongoDB desktop item

Create the directory where MongoDB will store it's files. From the command prompt run `md \data\db`. This is the default location. However, other locations can be specified using the `--dbpath` parameter. See [the Mongo docs](#) for more information.

- C:\>md data ○
C:\md data\db
- C:\Program Files\MongoDB\Server\4.05\bin>mongod.exe --dbpath
"C:\data"

Start the mongodb daemon by running `C:\mongodb\bin\mongod.exe` in the Command Prompt. Or by running, `C:\path\to\mongodb\bin\mongod.exe`

Connect to MongoDB using the Mongo shell While the MongoDB daemon is running, from a different Command prompt window

run `C:\mongodb\bin\mongo.exe`

`C:\Program Files\MongoDB\Server\4.05\bin>mongod.exe --dbpath "C:\data"`

`C:\Program Files\MongoDB\Server\4.05\bin>mongo.exe`

Step B: Install PyMongo

`C:\Users\ Your Name\AppData\Local\Programs\Python\Python36-32\Scripts>python -m pip install pymongo`

Now you have downloaded and installed a mongoDB driver.

Test PyMongo

demo_mongodb_test.py:

```
import pymongo
```

Program 1: Creating a Database

```
import pymongo
```

```
myclient = pymongo.MongoClient("mongodb://localhost:27017/")
mydb = myclient["mybigdata"]
print(myclient.list_database_names())
```

Progam 2: Creating a Collection

```
import pymongo
```

```
myclient = pymongo.MongoClient("mongodb://localhost:27017/")
mydb = myclient["mybigdata"]
mycol=mydb["student"]
print(mydb.list_collection_names())
```

Progam 3: Insert into Collection

```
import pymongo
```

```
myclient = pymongo.MongoClient("mongodb://localhost:27017/")
mydb = myclient["mybigdata"]
mycol=mydb["student"]
```

```
mydict={"name":"Kaushal", "address":"Mumbai"}  
x=mycol.insert_one(mydict) # insert_one(containing the name(s) and value(s) of each  
field
```

Program 4: Insert Multiple data into Collection

```
import pymongo  
  
myclient = pymongo.MongoClient("mongodb://localhost:27017/")  
mydb = myclient["mybigdata"]  
  
mycol=mydb["student"]  
  
mylist=[{"name":"Kaushal", "address":"Mumbai"}, {"name":"A", "address":"Mumbai"},  
{"name":"B", "address":"Pune"}, {"name":"C", "address":"Pune"},]  
x=mycol.insert_many(mylist)
```

Test in Mongodb to check database and data inserted in collection

- a. If you want to check your database list, use the command **show dbs** in mongo command prompt
- b. If you want to use a database with name mybigdata, then use database statement would be as follow: **use mybigdata**
- c. If you want to check collection in mongodb use the command **show collections**
- d. If you want to display all the data from collection: **db.collection_name.find()** or **db.collection_name.find().pretty()**

Practical No. 4

Aim: Hive: Introduction Creation of Database and Table, Hive Partition, Hive Built in Function and Operators, Hive View and Index.

Steps for hive installation

- ⑩ Download and Unzip Hive
- ⑩ Edit **.bashrc** file
- ⑩ Edit **hive-config.sh** file
- ⑩ Create **Hive directories** in HDFS
- ⑩ Initiate **Derby database**
- ⑩ Configure **hive-site.xml** file

download and unzip Hive

wget

<https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz>

tar xzf apache-hive-3.1.2-bin.tar.gz

Edit **.bashrc** file

```
sudo nano .bashrc

export HIVE_HOME= /home/hadoop/apache-hive-
3.1.2-bin

export PATH=$PATH:$HIVE_HOME/bin

source ~/.bashrc
```

Edit **hive-config.sh** file

```
sudo nano $HIVE_HOME/bin/hive-config.sh

export HADOOP_HOME=/home/hadoop/hadoop-
3.2.1
```

Create **Hive directories** in HDFS

```
hdfs dfs -mkdir /tmp

hdfs dfs -chmod g+w /tmp

hdfs dfs -mkdir -p /user/hive/warehouse

hdfs dfs -chmod g+w /user/hive/warehouse
```

Fixing guava problem – Additional step

```
rm $HIVE_HOME/lib/guava-19.0.jar  
cp  
$HADOOP_HOME/share/hadoop/hdfs/lib/guava-  
27.0-jre.jar $HIVE_HOME/lib/
```

Initialize Derby and hive

```
schematool -initSchema -dbType derby
```

hive

optional Step – Edit hive-site.xml

```
cd $HIVE_HOME/conf
```

```
cp hive-default.xml.template hive-site.xml
```

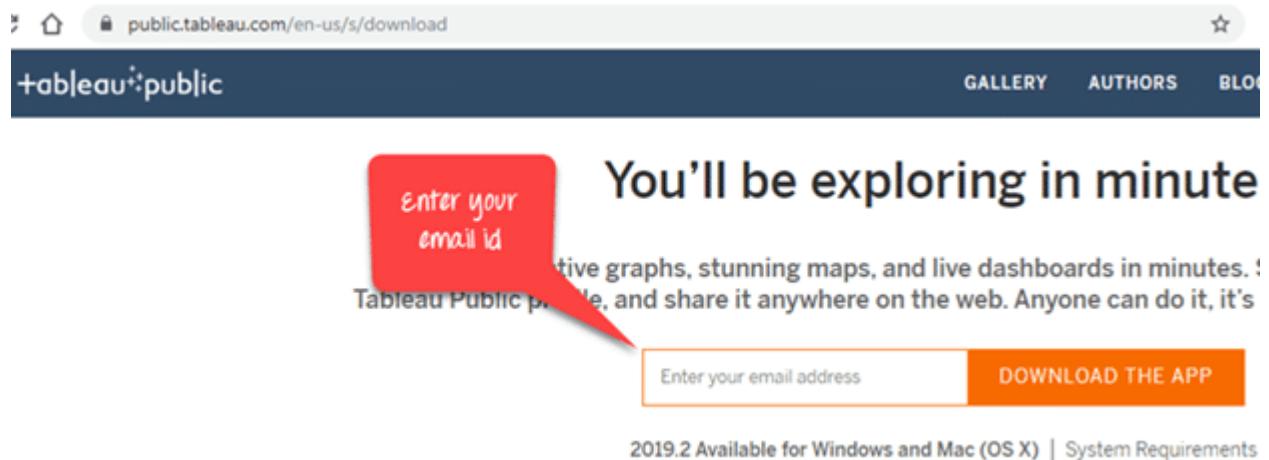
```
sudo nano hive-site.xml - change metastore location to above created hdfs  
path(/user/hive/warehouse)
```

Practical No. 5

Aim: Visualization: Connect to data, Build Charts and Analyze Data, Create Dashboard, Create Stories using Tableau.

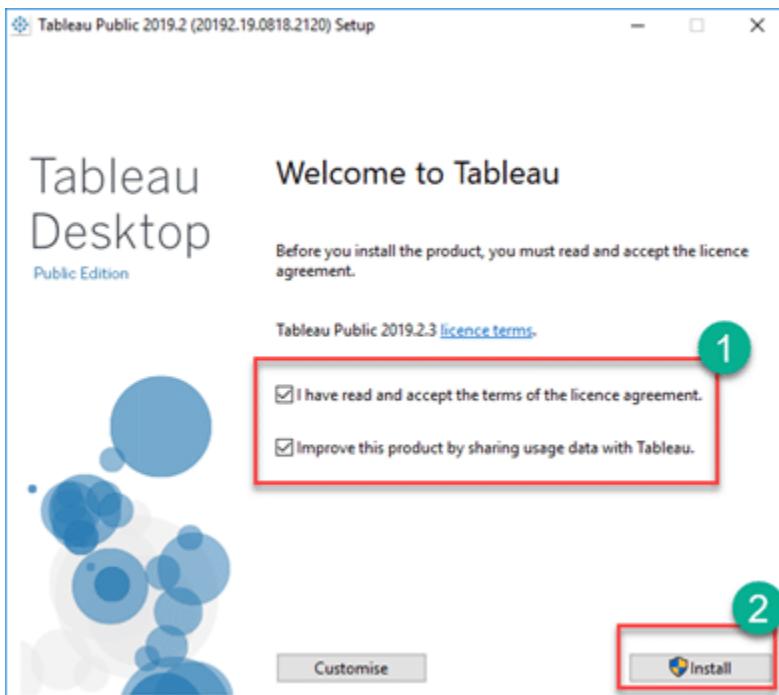
Downloading and Installing Tableau Public

1- Visit the URL <https://public.tableau.com/en-us/s/download> on your web browser. Once the window opens, enter your email id when asked, and click on the “Download the App” button.



2- The file will start downloading in “.exe” format. You can view the download progress on the bottom-left corner of the tab.

3- Once the progress is 100 percent, open the file. Accept the terms and conditions by selecting the checklist boxes and click on the “Install” button.



4- Once the installation is complete, open Tableau and start the screen of Tableau Public as shown below.



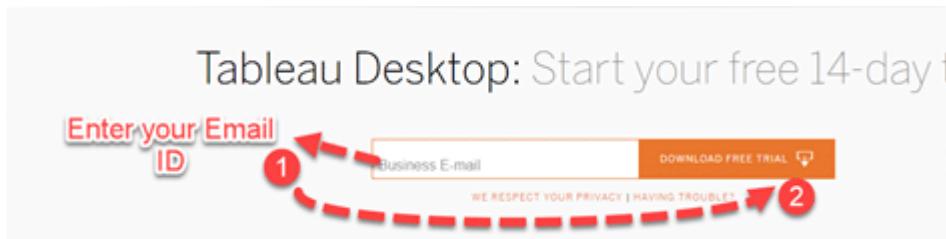
Downloading and Installing Tableau Desktop

1- Enter this URL <https://www.tableau.com/products/desktop> on your web browser.

2- Click on the “TRY NOW” button in the top-right corner of the website as shown below.

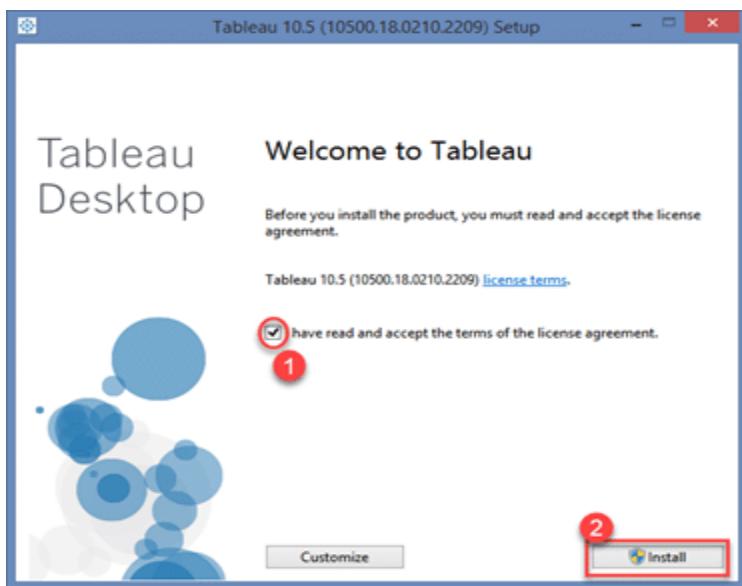


3- Once you click on the “TRY NOW” button, you will be redirected to a page that will ask you to feed in your official email address. After filling in the email address, click on the “DOWNLOAD FREE TRIAL” button.



4- The latest version of Tableau Desktop will start downloading, and you will be able to view the download progress in the bottom-left corner of the screen.

5- Once downloaded, open the file. Accept the terms and conditions, and click on the “Install” button.



6- A pop-up option will appear asking for the approval of the administrator to install the software. Click on “YES” to approve and move further.

7- On approval, the installation will start. On the completion of the installation, open Tableau.

8- This is the final stage that asks for registration. Click on “Activate Tableau” and enter your license details or credentials.

9- Click on “Start Trial Now” and wait for the registration process to complete.

10- Once it is completed, open the Tableau screen as shown below.



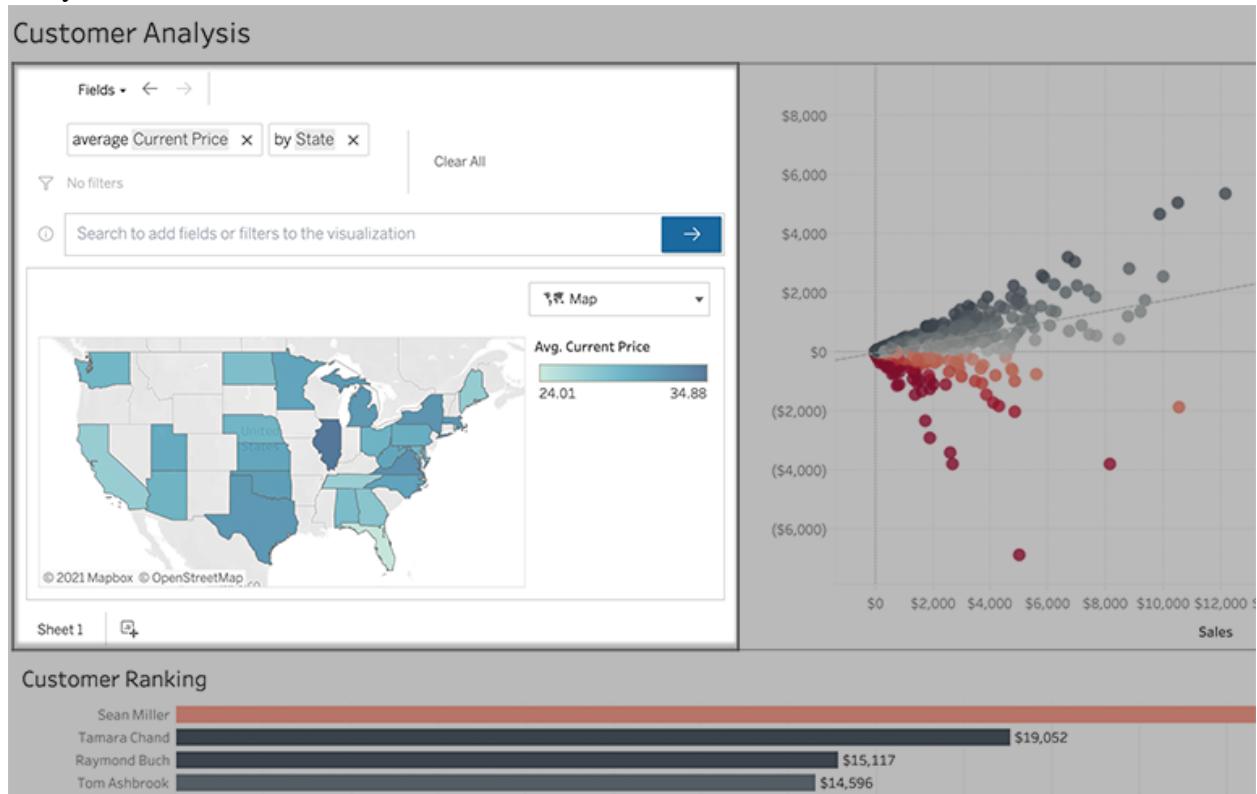
Connect from Tableau Desktop

Connect to Data:

- a. Open Tableau Desktop and click on the "Connect to Data" button.
- b. Choose the type of data source you want to connect to (e.g., Excel, CSV, SQL Server, etc.) and select the file or database.
- c. If necessary, enter any login credentials or other connection details required to access the data.
- d. Tableau will automatically import the data into a new worksheet.

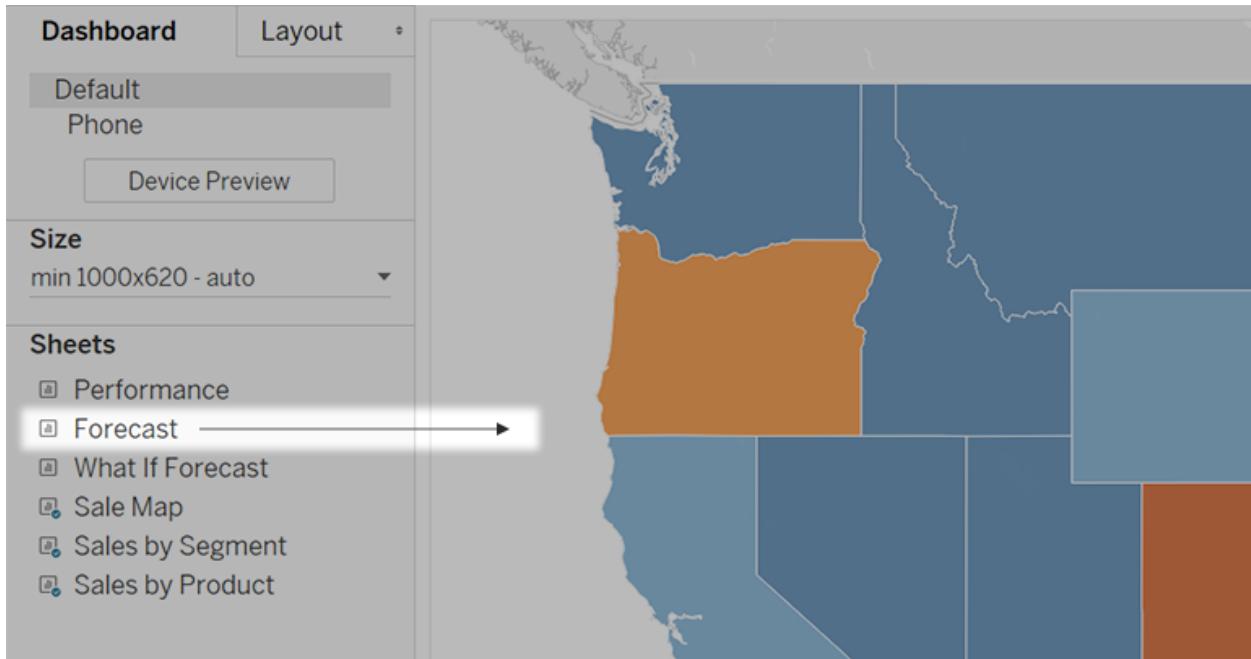
Build Charts and Analyze Data:

- a. In the worksheet, select the data you want to visualize and drag it onto the "Columns" and "Rows" shelves to create a chart.
- b. Choose the type of chart you want to create from the "Show Me" panel (e.g., bar chart, line chart, scatter plot, etc.).
- c. Use the "Marks" card to customize the appearance of the chart (e.g., color, size, shape, etc.).
- d. Use the "Filters" pane to refine the data displayed in the chart by selecting or excluding specific values or ranges.
- e. Use the "Analytics" pane to add statistical functions or trend lines to the chart for deeper analysis.

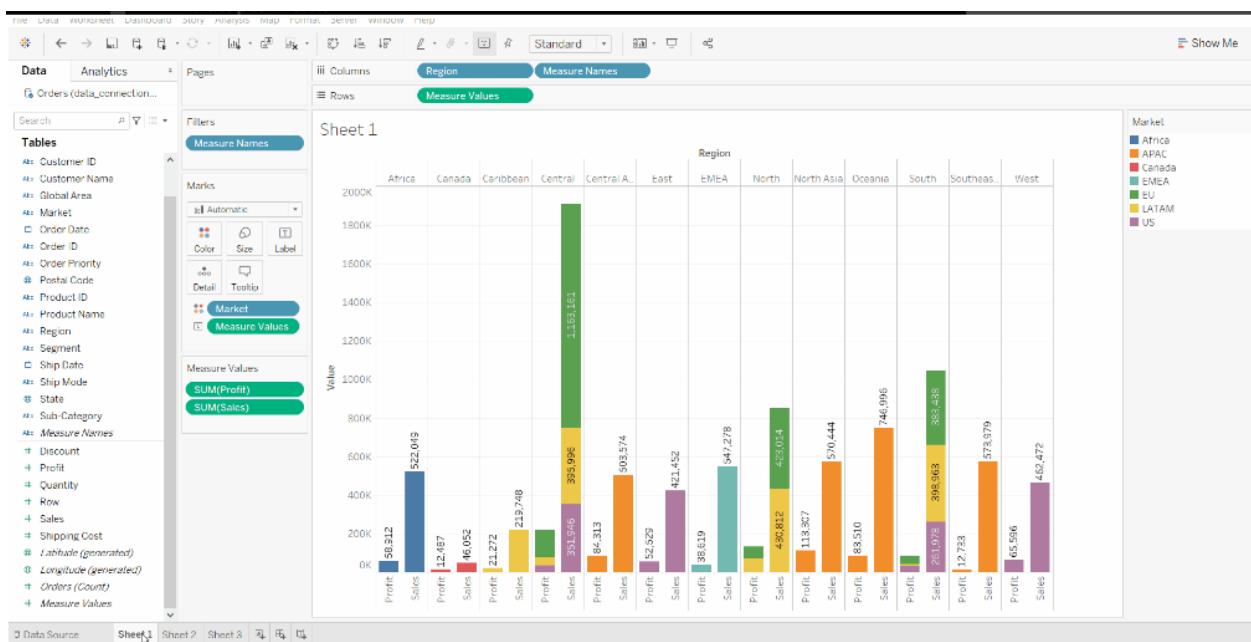


Create Dashboard:

- a. Click on the "New Dashboard" button to create a new dashboard.
- b. Drag the charts or worksheets you want to include on the dashboard onto the canvas.



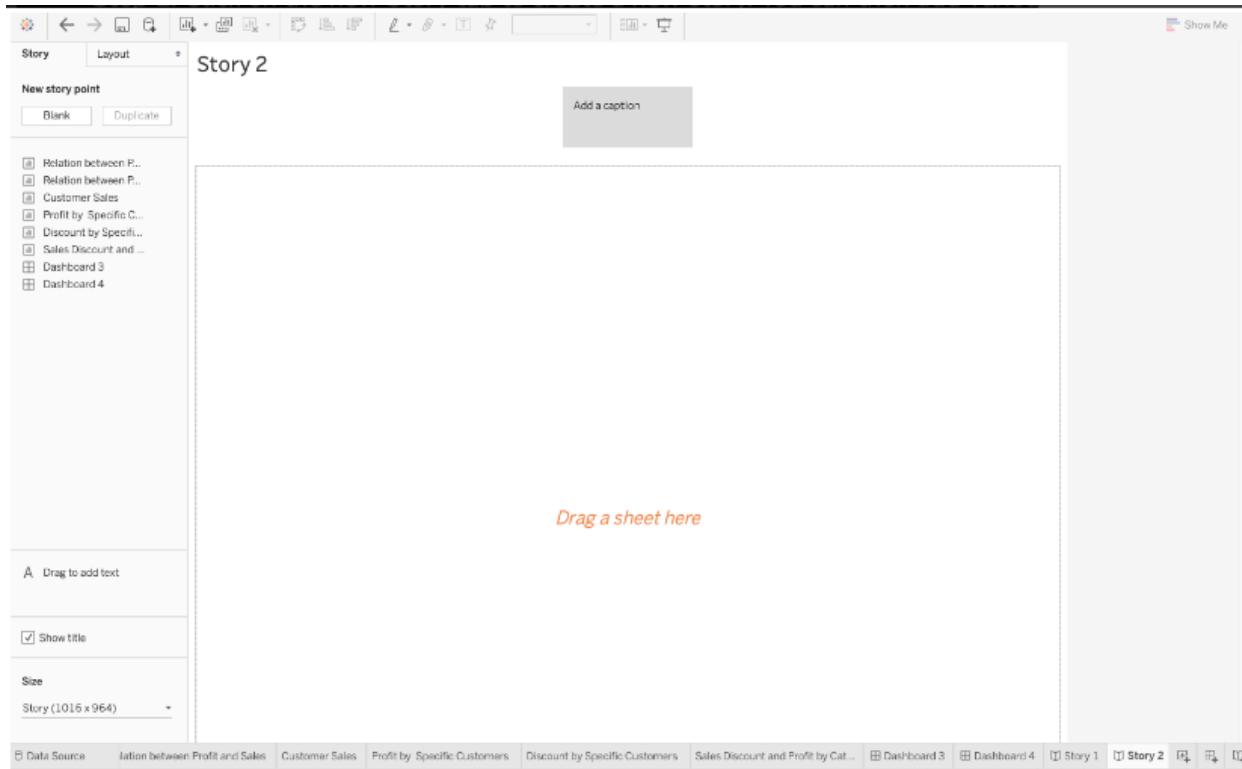
- Use the "Dashboard" menu to customize the layout and formatting of the dashboard (e.g., adding titles, adjusting the size and position of the components, etc.).
- Use the "Actions" pane to create interactivity between the components on the dashboard (e.g., selecting a value in one chart filters the data displayed in another chart).



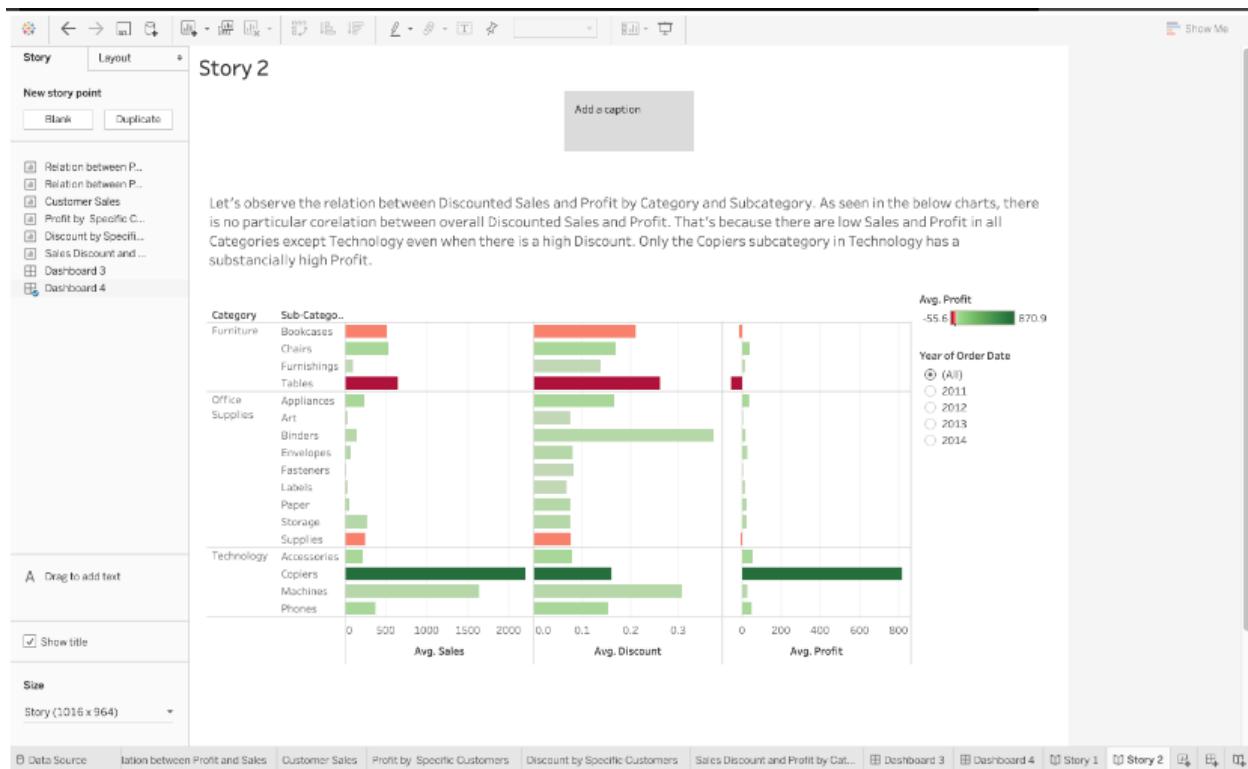
Create Stories:

- Click on the "New Story" button to create a new story.

- b. Use the "Story" menu to add slides to the story and customize the layout and formatting of each slide (e.g., adding titles, adjusting the size and position of the components, etc.).



- c. Use the "Sheets" pane to select the worksheets or dashboards you want to include on each slide.



d. Use the "Annotations" pane to add annotations, text boxes, or other visual aids to explain the data and insights on each slide.

Story 2

< Relation between Discounted Sales and Profit by Category Add a caption >

Drag a sheet here

A Drag to add text

Show title

Size
Story (1016 x 964)

Data Source Ilation between Profit and Sales Customer Sales Profit by Specific Customers Discount by Specific Customers Sales Discount and Profit by Cat... Dashboard 3 Dashboard 4 Story 1 Story 2

- f. Use the "Navigation" pane to create a sequence of slides and add transitions between them for a more engaging storytelling experience.

