MOVIE RECOMMENDATION SYSTEM

A PROJECT REPORT

Submitted By

Ojasvi Tyagi
(University Roll No- 2000290140081)
Tanu Sharma
(University Roll No-2000290140125)
Rajat Deol
(University Roll No-2000290140096)
Shivani Chauhan
(University Roll No-2000290140114)

Submitted in partial fulfilment of the Requirements for the Degree of

MASTER OF COMPUTER APPLICATIONS

Under the Supervision of
Mrs. Vidushi Mishra
Assistant Professor,
KIET Group of Institutions

Submitted to

DEPARTMENT OF COMPUTER APPLICATIONS
KIET Group of Institutions, Ghaziabad
Uttar Pradesh-201206

CERTIFICATE

Certified that **Ojasvi Tyagi (Enrollment No-20002901401400), Rajat Deol (Enrollment No-200029014005781), Tanu Sharma** (**Enrollment No-200029014014005810**), **Shivani Chauhan (Enrollment No-200029014014005799**) have carried out the project work having "**Title of Report – Movie Recommendation**" for Master of Computer Applications from Dr. A.P.J. Abdul Kalam Technical University (AKTU**)** (formerly UPTU), Technical University, Lucknow under my supervision. The project report embodies original work, and studies are carried out by the student himself / herself and the contents of the project report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Date:

Ojasvi Tyagi (2000290140081)
Rajat Deol (2000290140096)
Tanu Sharma (2000290140125)
Shivani Chauhan (2000290140114)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date:

Mrs.Vidushi Misra
Assistant Professor
Department of Computer Applications
KIET Group of Institutions, Ghaziabad

Signature of Internal Examiner **Signature of External Examiner**

Dr. Ajay Shrivastava
Head, Department of Computer Applications
KIET Group of Institutions, Ghaziabad

ABSTRACT

This paper discusses about recommendations of the movies. A movie recommendation is important in our social life due to its strength in providing enhanced entertainment. Such a system can suggest a set of movies to users based on their interest, or the popularities of the movies. A recommendation system is used for the purpose of suggesting items to purchase or to see. They direct users towards those items which can meet their needs through cutting down large database of Information. A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. They are primarily used in commercial applications. MOVREC also help users to find the movies of their choices based on the movie experience of other users in efficient and effective manner without wasting much time in useless browsing.
Keywords: Filtering, Recommendation System, Recommender.

ACKNOWLEDGEMENT

Success in life is never attained single handedly. My deepest gratitude goes to my thesis supervisor, **Mrs. Vidushi Misra** for her guidance, help and encouragement throughout my research work. Their enlightening ideas, comments, and suggestions.

Words are not enough to express my gratitude to Dr. Ajay Kumar Shrivastava, Professor and Head, Department of Computer Applications, for his insightful comments and administrative help at various occasions.

Fortunately, I have many understanding friends, who have helped me a lot on many critical conditions.

Finally, my sincere thanks go to my family members and all those who have directly and indirectly provided me moral support and other kind of help. Without their support, completion of this work would not have been possible in time. They keep my life filled with enjoyment and happiness.

Ojasvi Tyagi

Rajat Deol

Tanu Sharma

Shivani Chauhan

TABLE OF CONTENTS

Chapter 1 - Introduction

Project description:

A movie recommendation is important in our social life due to its strength in providing enhanced entertainment. Such a system can suggest a set of movies to users based on their interest, or the popularities of the movies. A recommendation system is used for the purpose of suggesting items to purchase or to see. They direct users towards those items which can meet their needs through cutting down large database of Information. A recommender system, or a recommendation system (sometimes replacing 'system' with a synonym such as platform or engine), is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. They are primarily used in commercial applications. MOVREC also help users to find the movies of their choices based on the movie experience of other users in efficient and effective manner without wasting much time in useless browsing. Keywords: Filtering, Recommendation System, Recommender.

Project Scope:

A recommendation system has become an indispensable component in various e-commerce applications. Recommender systems collect information about the user's preferences of different items (e.g., movies, shopping, tourism, TV, taxi) by two ways, either implicitly or explicitly. An implicit acquisition of user information typically involves observing the user's behavior such as watched movies, purchased products, downloaded applications. On the other hand, a direct procurement of information typically involves collecting the user's previous ratings or history. Collaborative filtering (CF) is the way of filtering or calculating items through the sentiments of other people. It first gathers the movie ratings given by individuals and then recommends movies to the target user based on like-minded people with similar tastes and interests in the past.
Additional impression on which some recommender systems are based is clustering. Clustering is a popular unsupervised data mining tool that is used for partitioning a given dataset into homogeneous groups based on some similarity or dissimilarity metric .

Collaborative filtering and clustering have been discussed in detail in the next section. Hybrid cluster and optimization approach is applied to improve movie prediction accuracy. Such a hybrid approach has been used to overcome the limitations of typical content-based and collaborative recommender systems. For clustering, k-means algorithm is applied and for optimization, cuckoo search optimization is implemented. K-means algorithm is an enormously greater clustering algorithm when compared toother clustering methods in relations of time, complexity, or effectiveness for a particular number of clusters. Clustering algorithm with a bio-inspired algorithm such as cuckoo search delivers optimize results. The cuckoo search has shown best performance when compared with other algorithms such as genetic algorithms and particle swarm optimization. Simulations and comparison of the cuckoo search were greater to these existing algorithms for multimodal objective functions. To find the best results we must find the most suitable weight among all possible ones.

Hardware/Software used in Project

Hardware Used in Project

- Window 10
- 8 GB RAM
- i7 10th processor
- With 512 GB SSD

Software Used in Project

- Hadoop
- Oracle VirtualBox
- Centos
- Putty

Chapter 2 - Feasibility Study:

Feasibility Study is a study to evaluate feasibility of proposed project or system. Feasibility study is one of stage among important four stages of Software Project Management Process.

As name suggests feasibility study is the feasibility analysis or it is a measure of the software product in terms of how much beneficial product development will be for the organization in a practical point of view.

Feasibility study is carried out based on many purposes to analyze whether software product will be right in terms of development, implantation, contribution of project to the organization etc.

Technical feasibility:

The objective of the technical feasibility step is to confirm that the product will perform and to verify that there are no production barriers. Product: The product of this activity is a working model.

In Technical Feasibility current resources both hardware software along with required technology are analyzed/assessed to develop project. This technical feasibility study gives report whether there exists correct required resources and technologies which will be used for project development. Along with this, feasibility study also analyzes technical skills and capabilities of technical team, existing technology can be used or not, maintenance and up-gradation is easy or not for chosen technology etc.

Operational Feasibility:

The operational feasibility to help users find items that they deem of interest to them. They can be seen as an application of data mining process. In this paper, a new recommender system based on multi-features is introduced. Demographic and psychographic features are used to asses' similarities between users.

In Operational Feasibility degree of providing service to requirements is analyzed along with how much easy product will be to operate and maintenance after deployment. Along with these other operational scopes are determining usability of product, determining suggested solution by software development team is acceptable or not etc.

Economic Feasibility:

In Economic Feasibility study cost and benefit of the project is analyzed. Means under this feasibility study a detail analysis is carried out what will be cost of the project for development which includes all required cost for final development like hardware and software resource required, design and development cost and operational cost and so on. After that it is analyzed whether project will be beneficial in terms of finance for organization or not.

Behavioral Feasibility:

It evaluates and estimates the user attitude or behavior towards the development of new system. It helps in determining if the system requires special effort to educate, retrain, transfer, and changes in employee's job status on new ways of conducting business.

Chapter 3 - Technologies Used:

Virtual Box:

VirtualBox is a software that is provided by Oracle to install virtual machines onto your system. It was introduced in the year 2007 by Innotek Gmbh (a former German software company which created VirtualBox) and later was developed by Oracle. It is also called a software virtualization package that is capable to load multiple operating systems.

Some features of VirtualBox :

- It offers a variety of operating systems such as Windows XP, Linux, Ubuntu, macOS.
- It is a very rich, robust, and extremely high-performance product
- It is used by professionals in the form of Open Source Software.
- It is open-source software that provides virtualization which is capable to offer services on almost every type of host operating system.

How to install Ubuntu on VirtualBox?

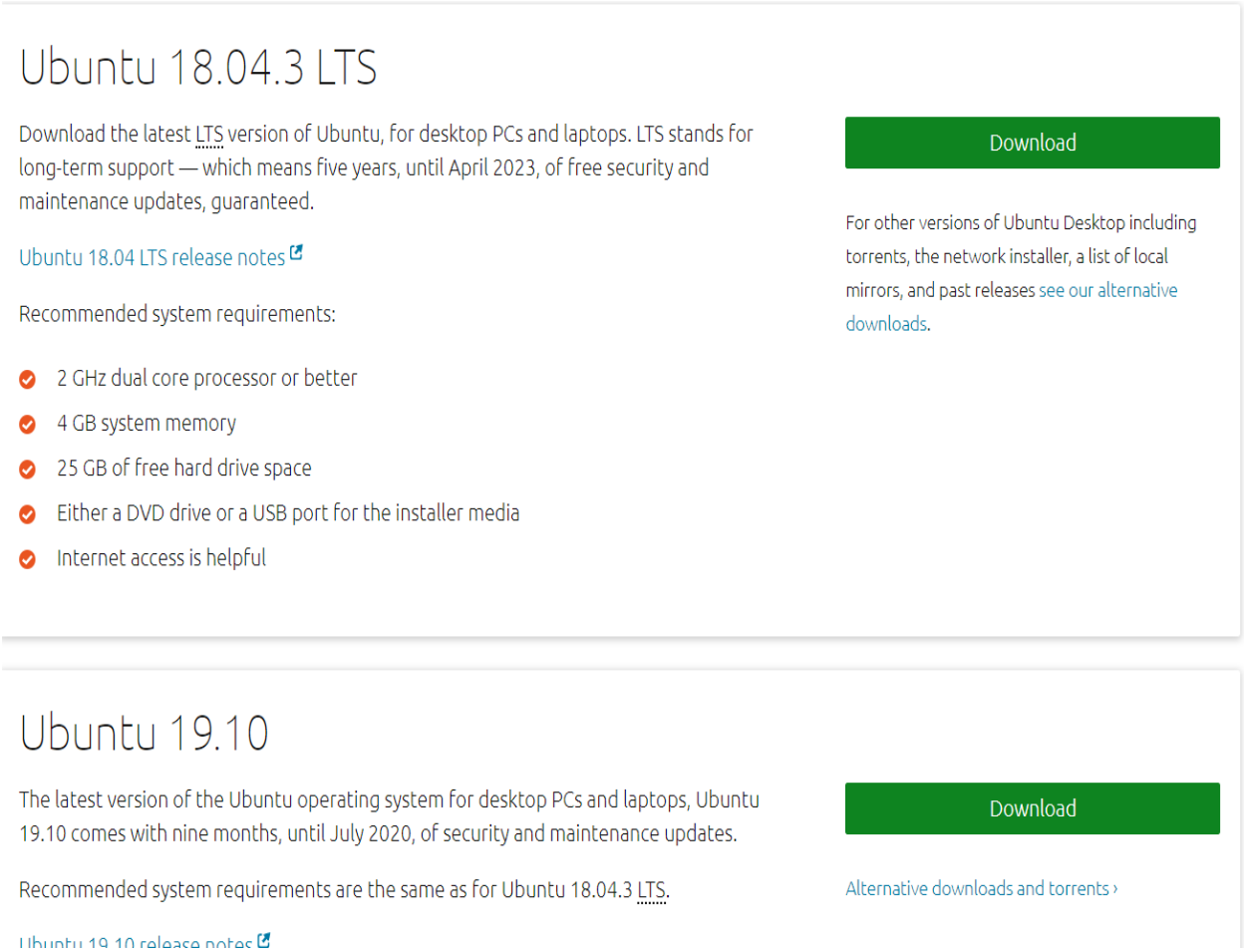
Virtual Machine abstracts the hardware of our personal computers such as CPU, disk drives, memory, NIC (Network Interface Card), etc, into many different execution environments as per our requirements, hence giving us a feeling that each execution environment is a single computer. For example, Virtual Box.

We can create a virtual machine for several reasons, all of which are fundamentally related to the ability to share the same basic hardware yet can also support different execution environments, i.e., different operating systems simultaneously.

To use Ubuntu along with Windows, one must have VirtualBox installed in their machine.

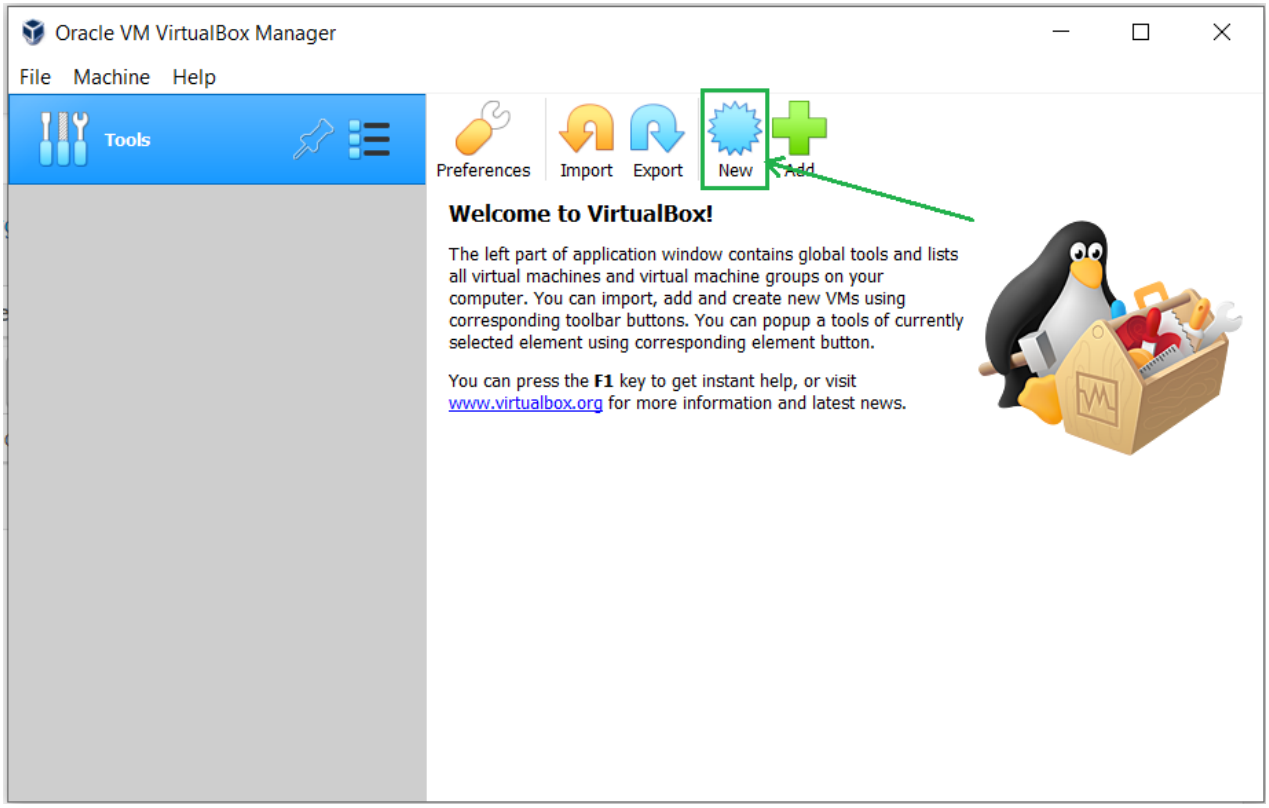
Downloading and Installing Ubuntu

Before, we begin with the installation process, we need to download ISO file for Ubuntu. For that, all the versions of Ubuntu are available on the official site ubuntu.com

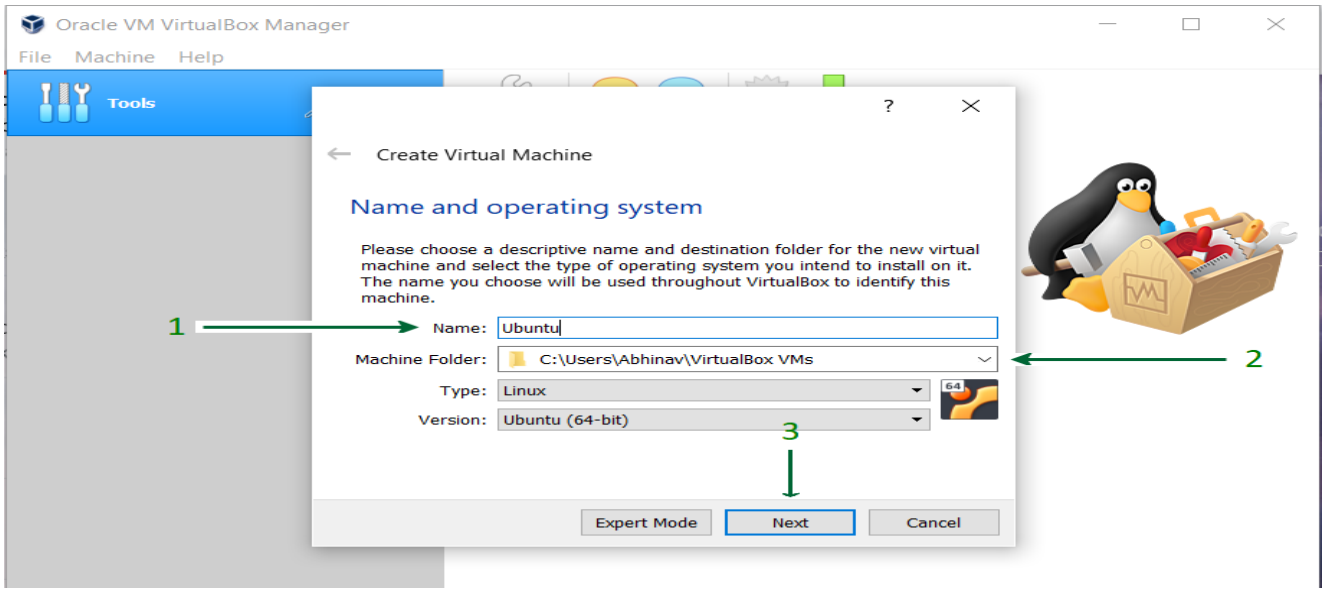


After the downloading is over, you can install Ubuntu on VirtualBox with the help of following instructions:

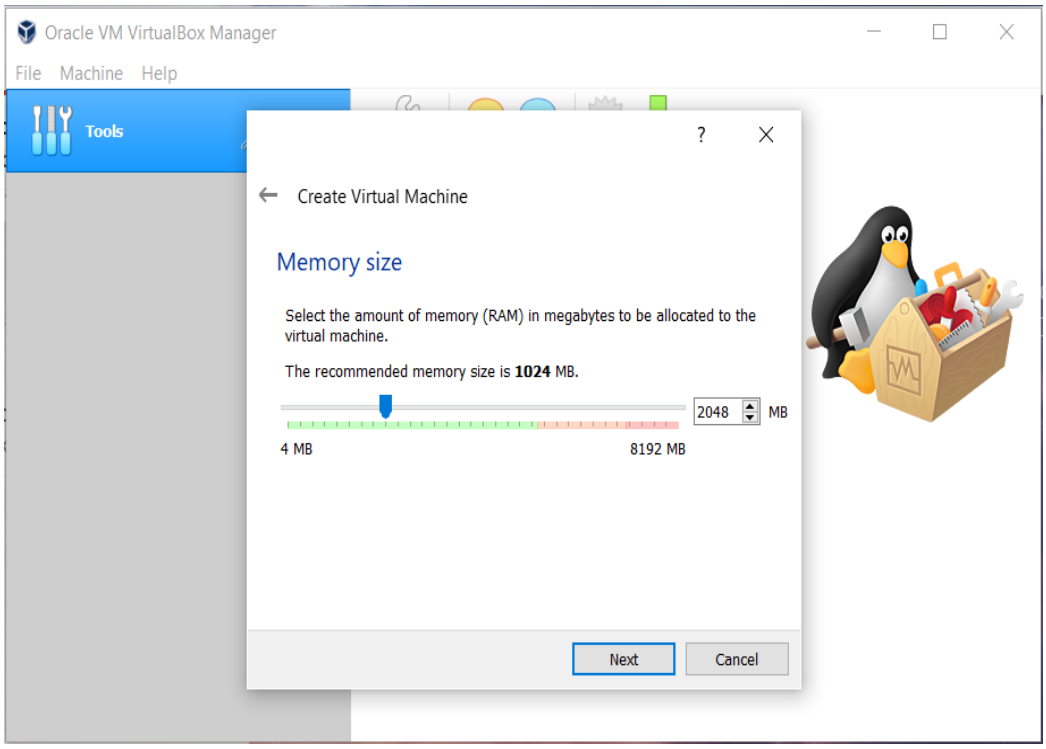
Step 1: Open **VirtualBox** and click on the **New** button.



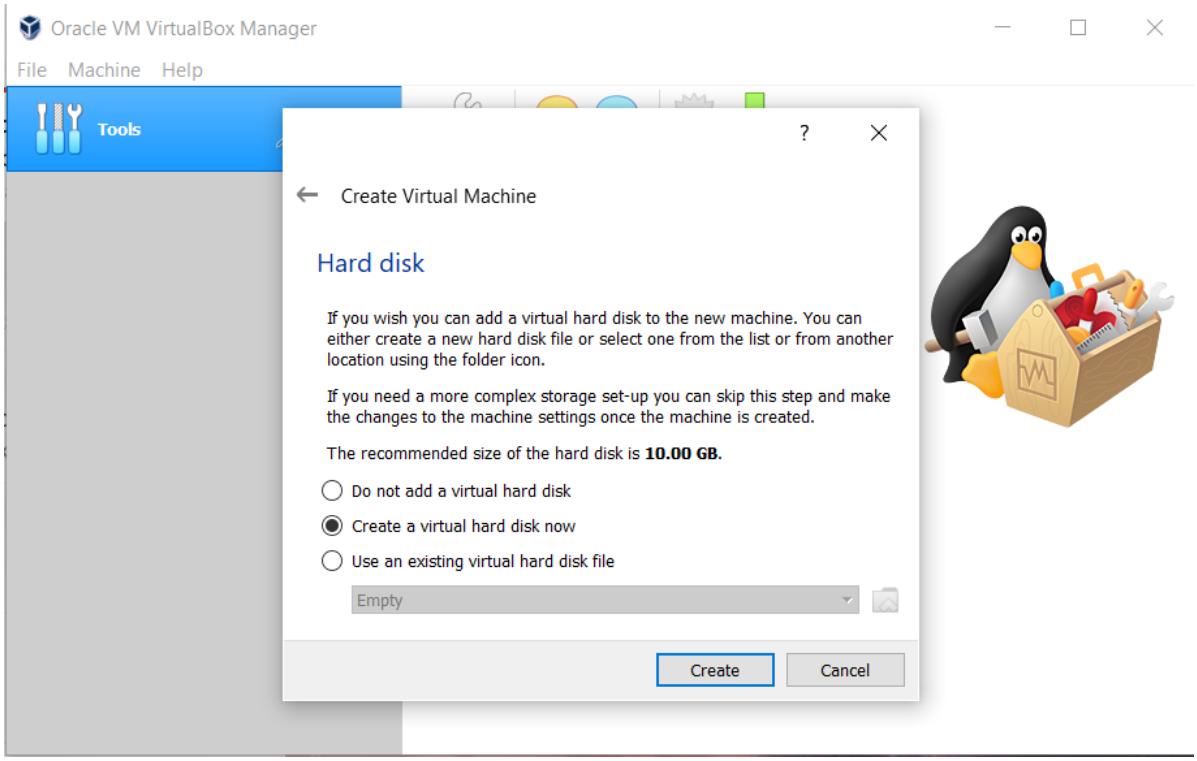
Step 2: Give a name to your Virtual Machine and select the location for it to install.



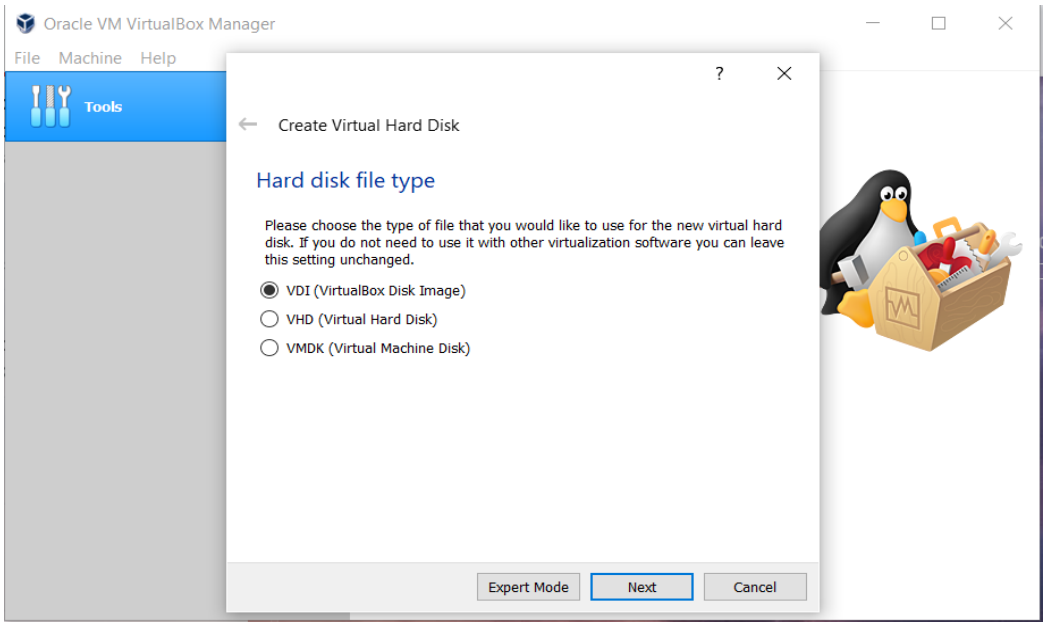
Step 3: Assign RAM size to your Virtual Machine.



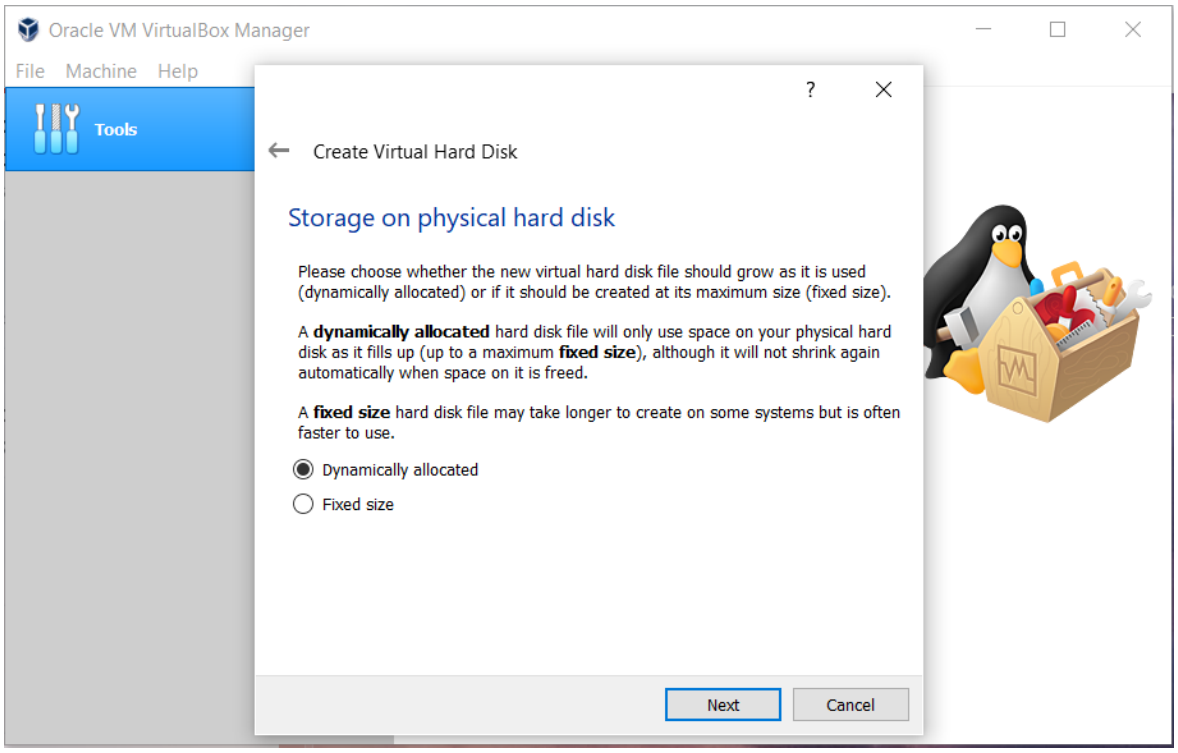
Step 4: Create a Virtual Hard disk for the machine to store files.



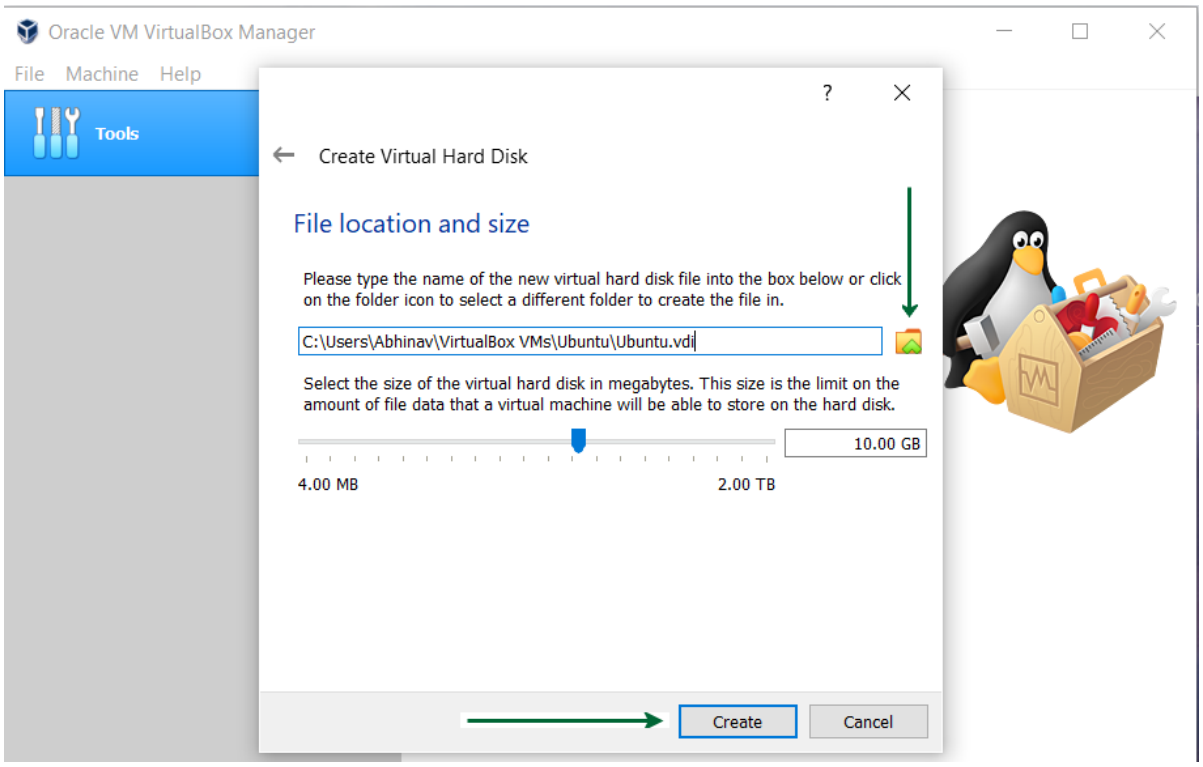
Step 5: Select the type of Hard disk. Using **VDI** type is recommended.



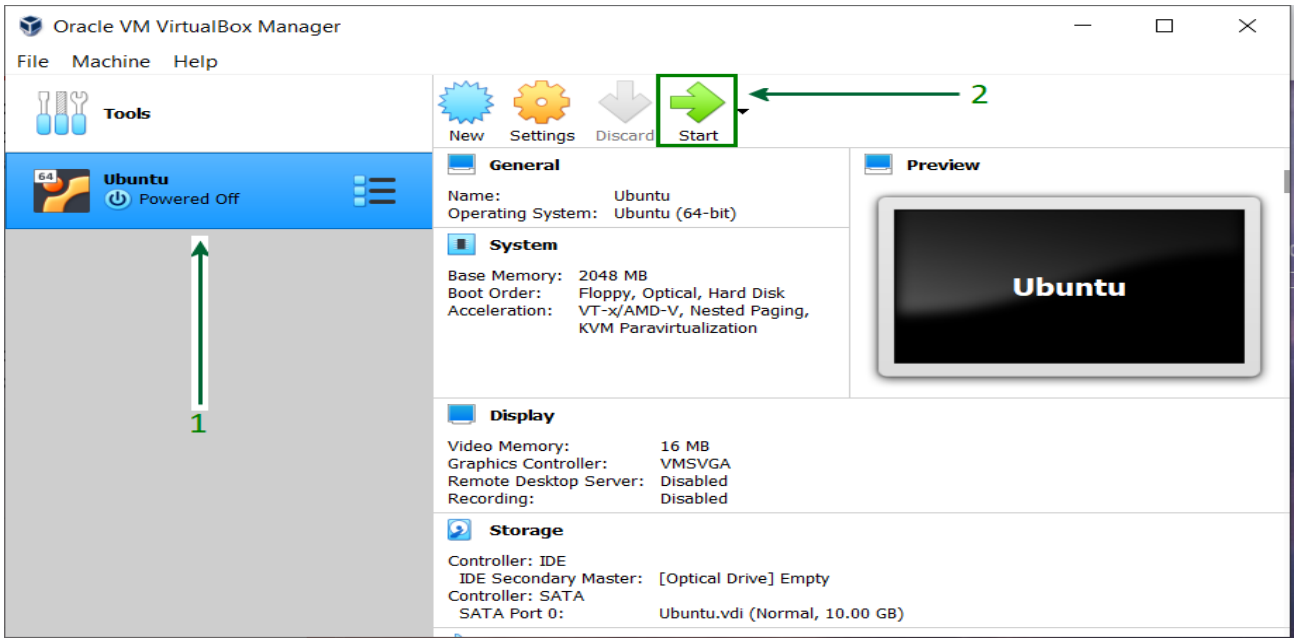
Step 6: Either of the physical storage type can be selected. Using dynamically allocated disk is by default recommended.



Step 7: Select disk size and provide the destination folder to install.

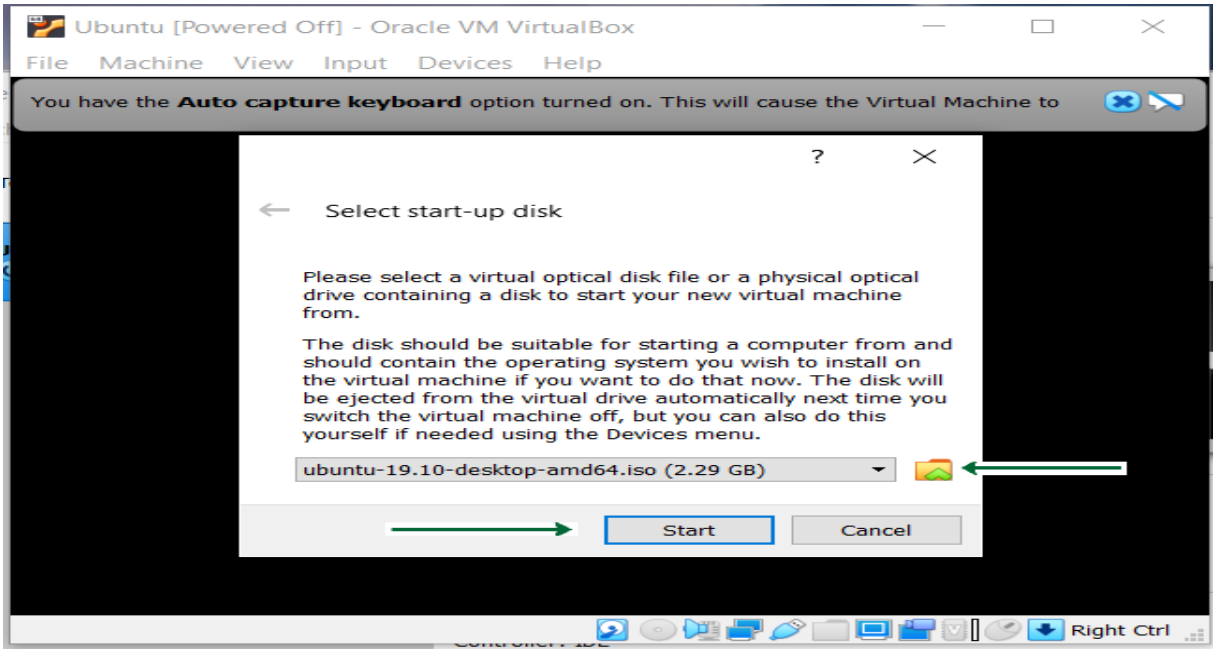


Step 8: After the Disk creation is done, boot the Virtual Machine and begin installing Ubuntu.

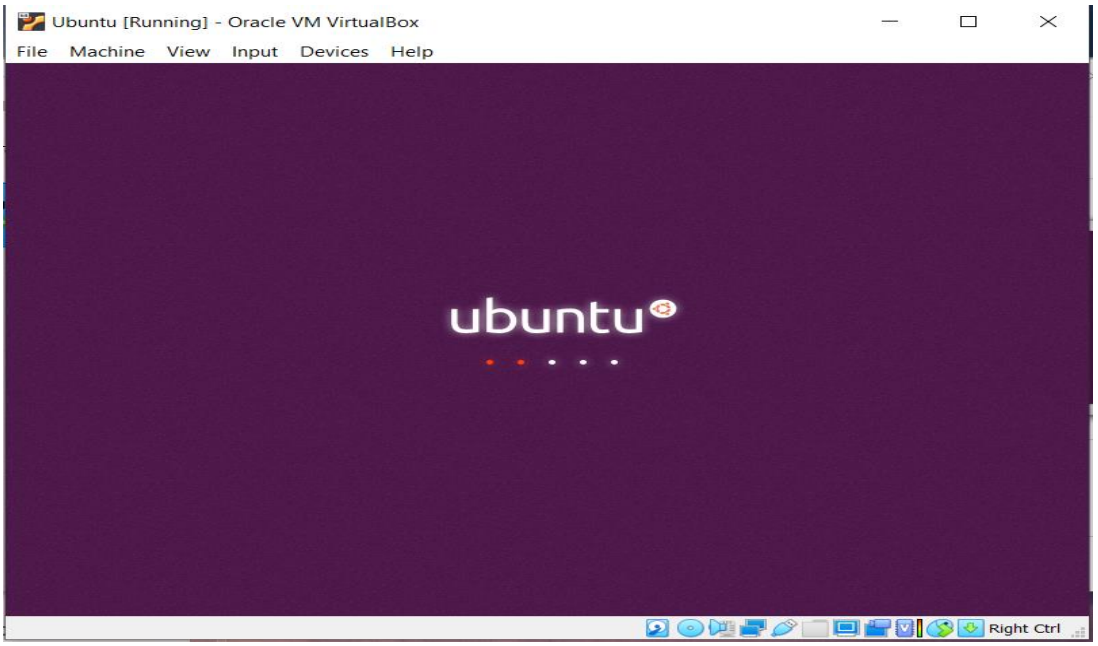


Step 9: If the installation disk is not automatically detected. Browse the file location

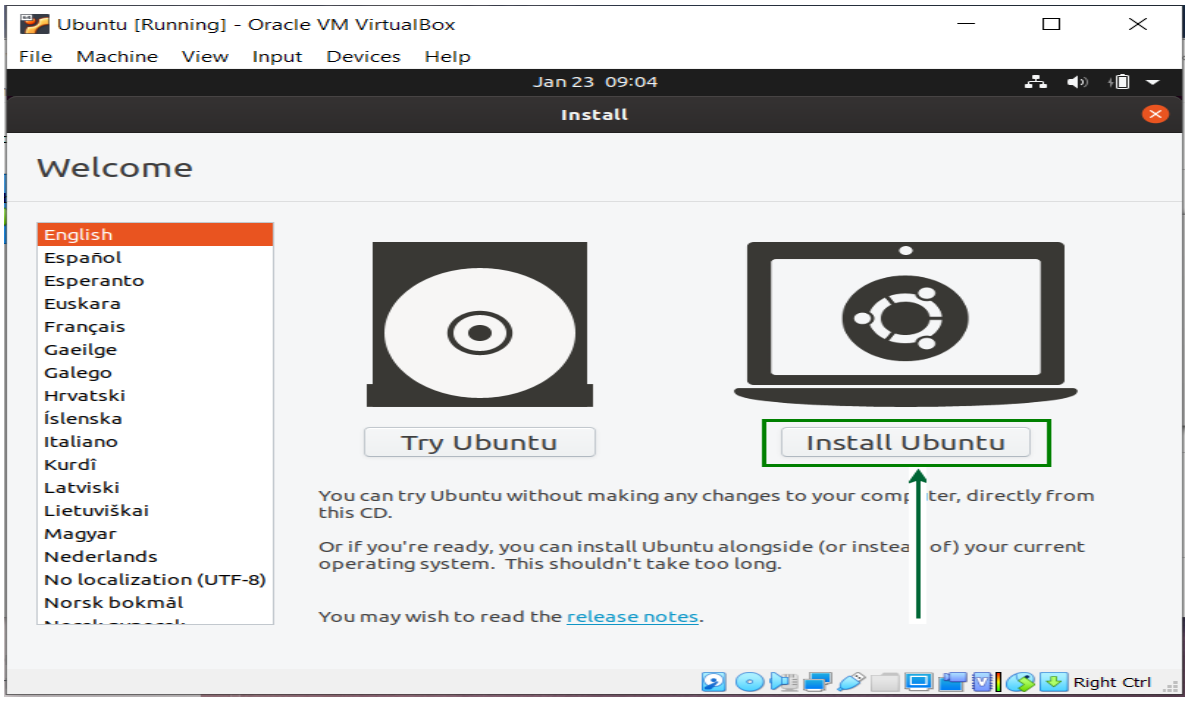
and select the ISO file for Ubuntu.



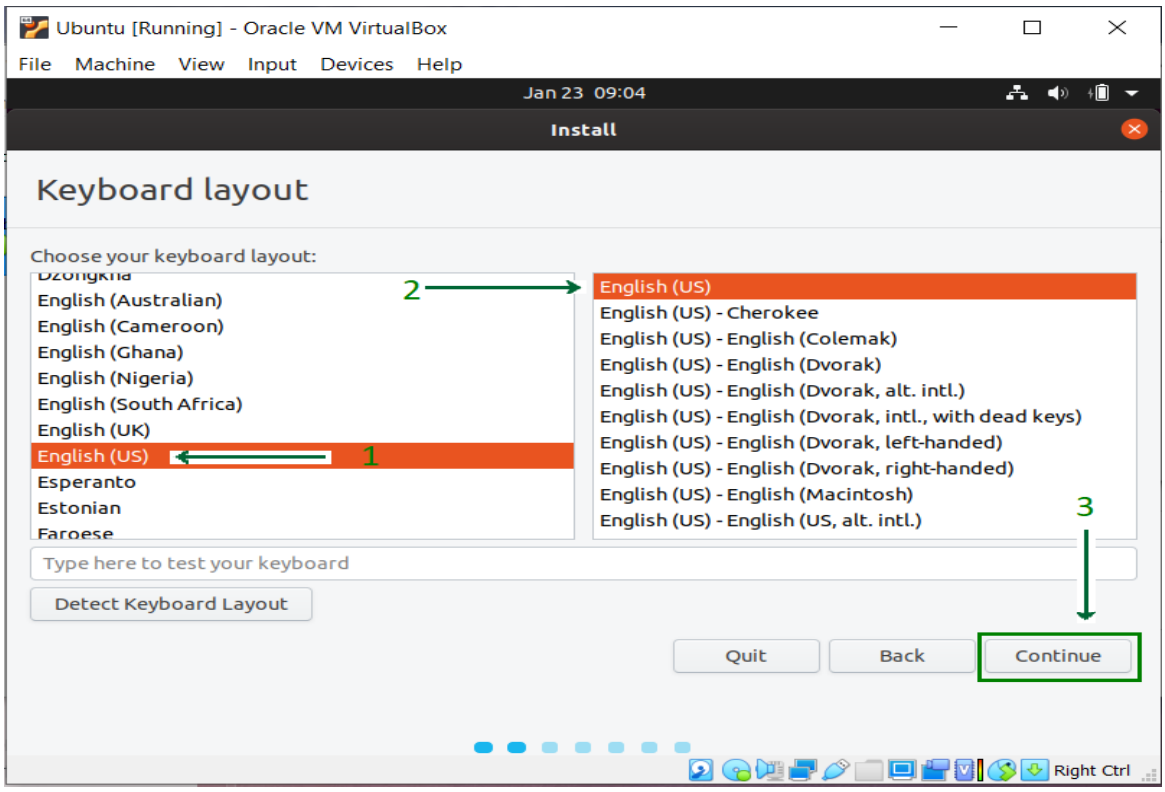
Step 10: Proceed with the installation file and wait for further options.



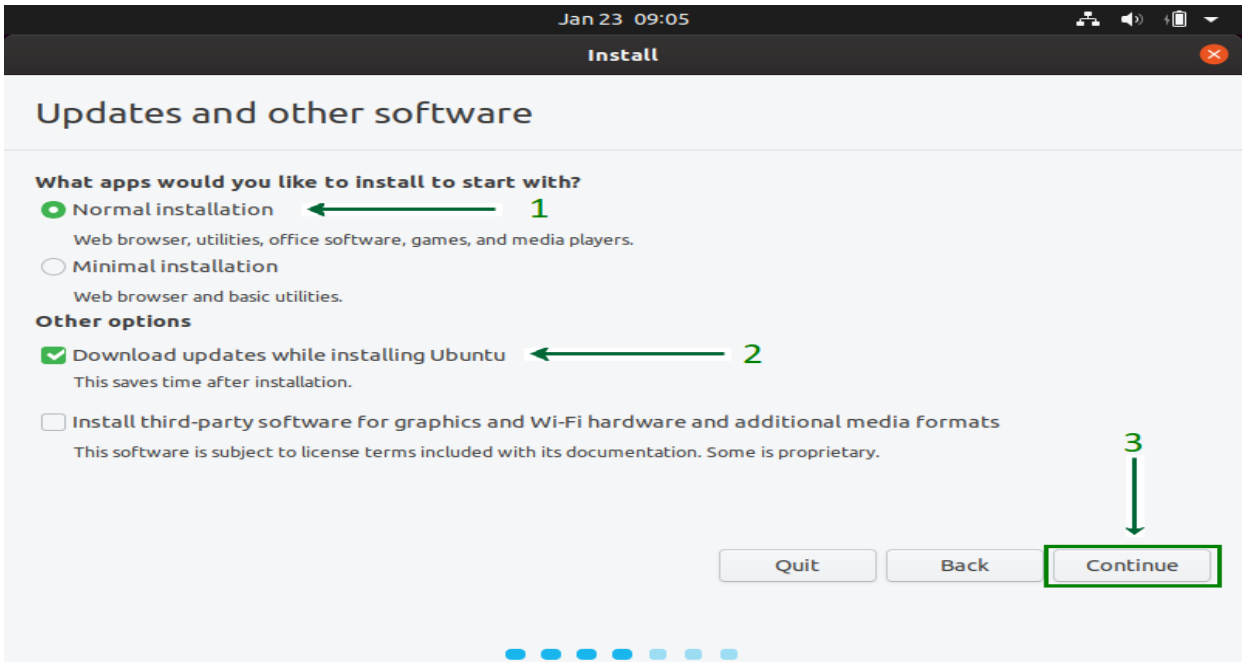
Step 11: Click on the Install Ubuntu option, this might look different for other Ubuntu versions.



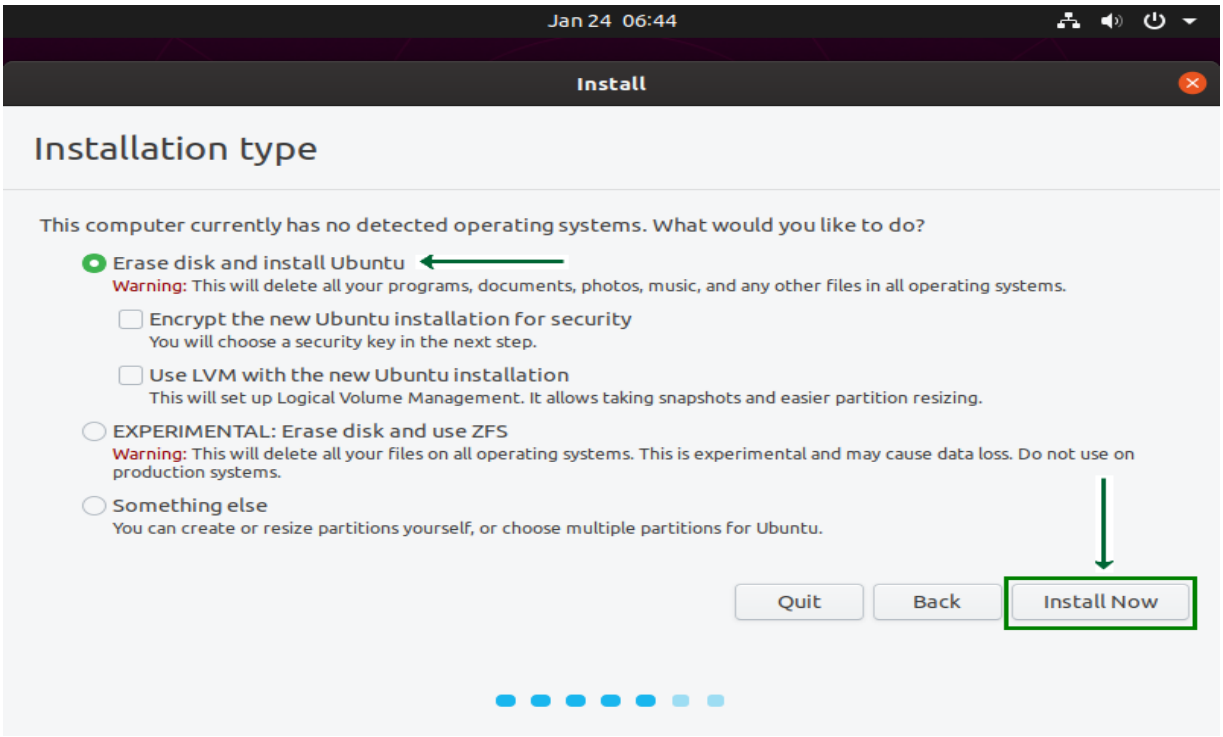
Step 12: Select Keyboard layout, if the defaults are compatible, just click on the **continue** button and proceed.



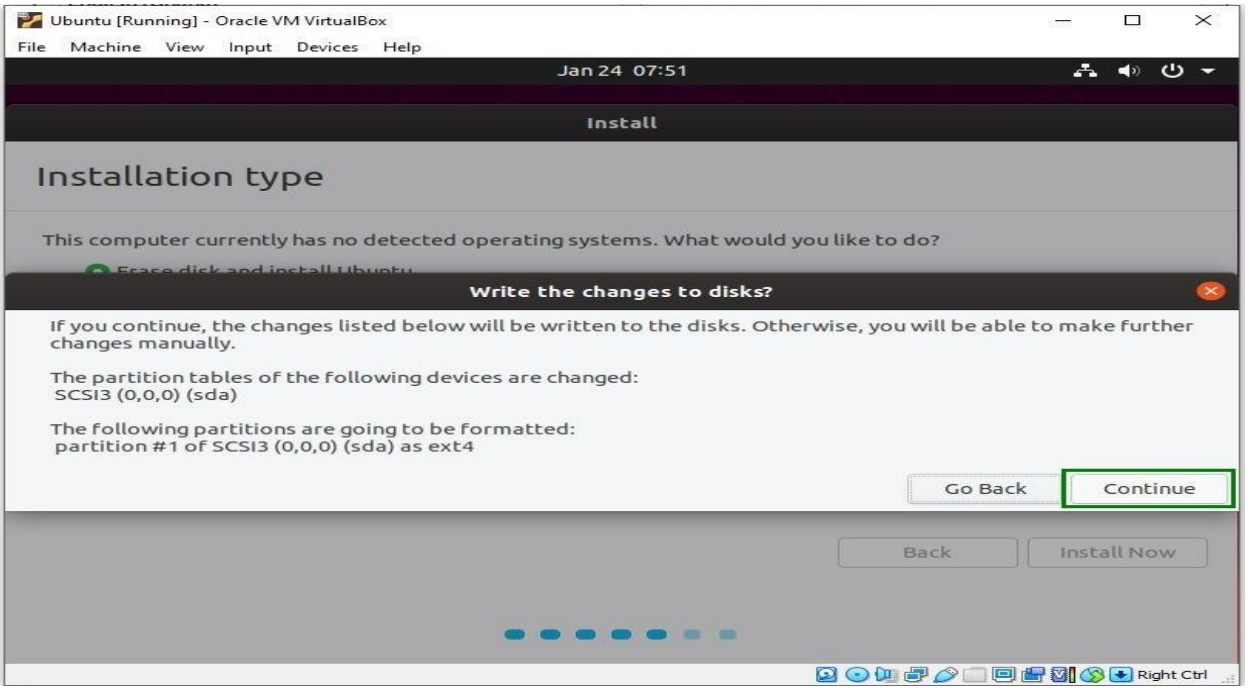
Step 13: Select installation type. By default, it is set to Normal installation, which is recommended, but it can also be changed to Minimal installation if there is no need for all Ubuntu features.



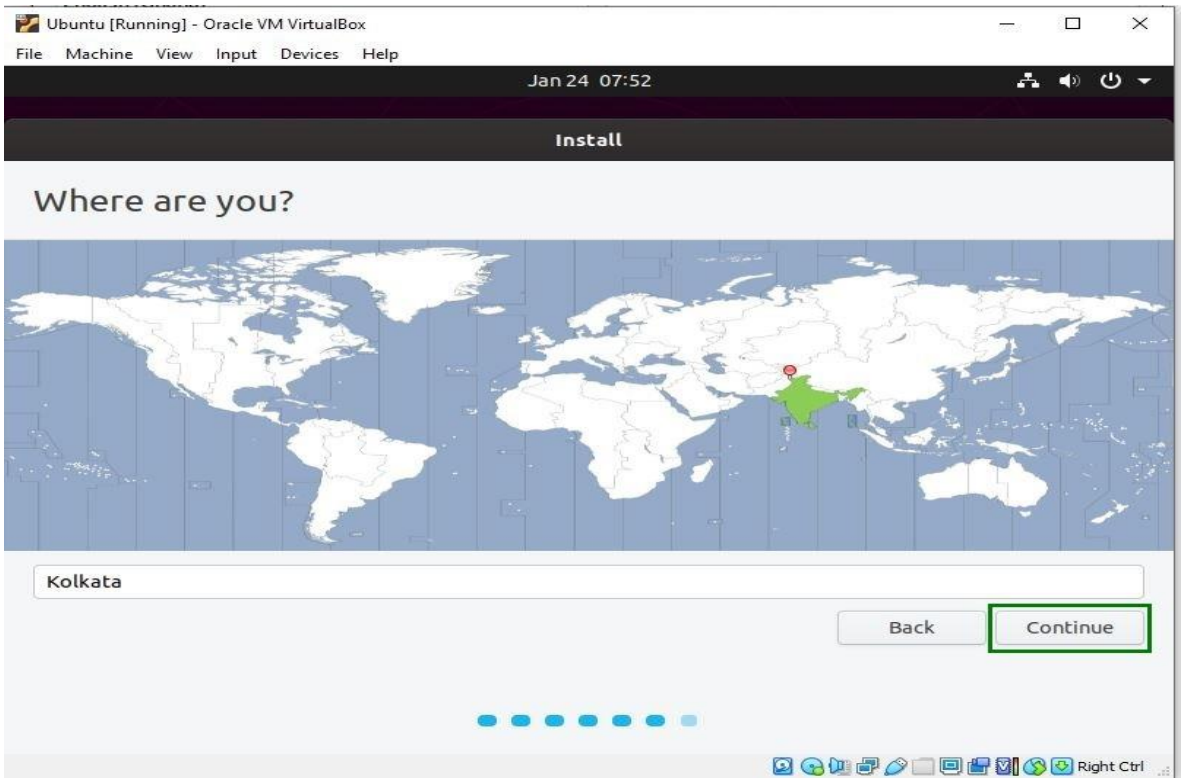
Step 14: Click on the **Install Now** button and carry on with the installation. Do not get worried with the **Erase disk** option, it will only be effective inside the virtual machine, other system files outside the VirtualBox remain intact.



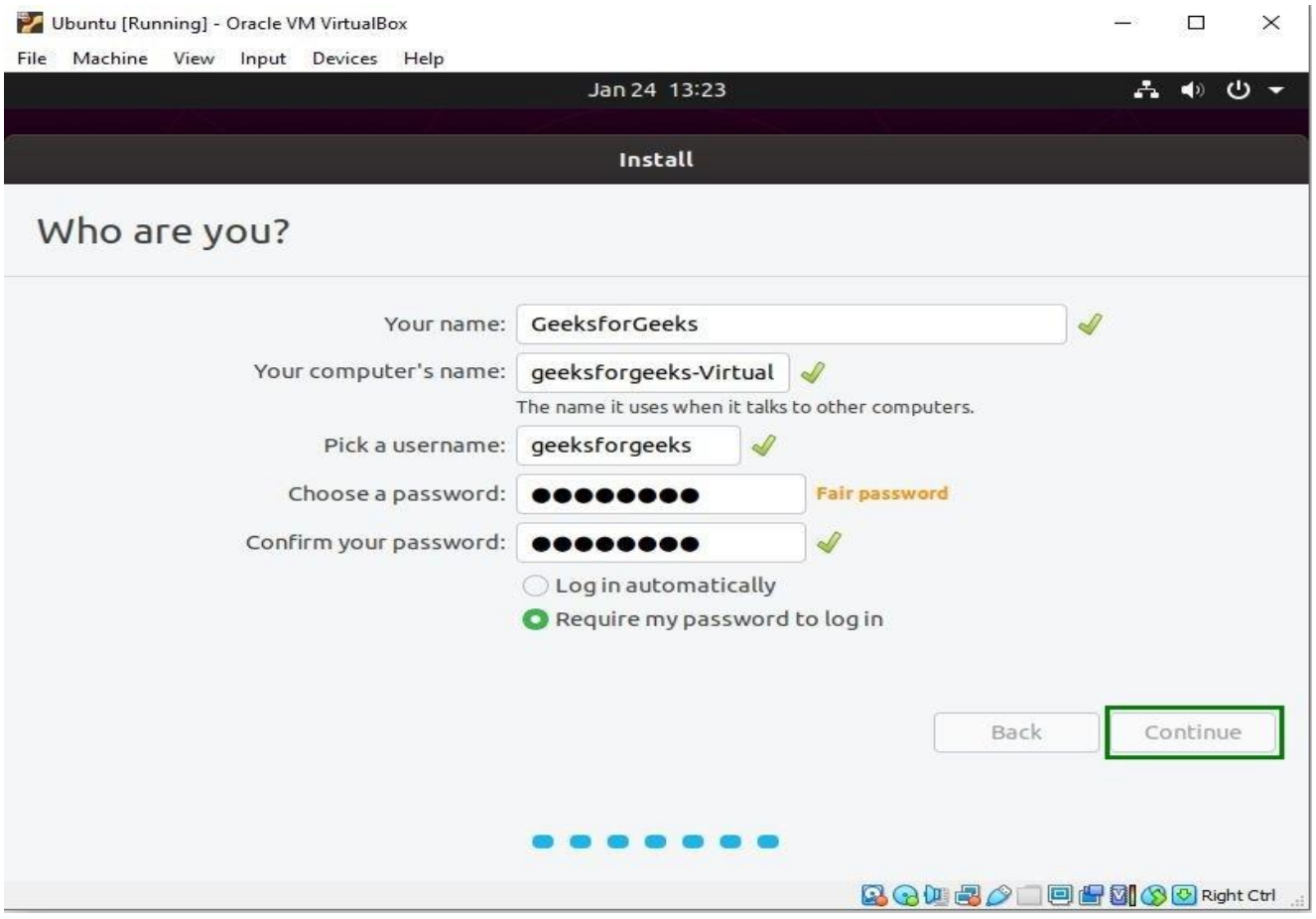
Step 15: Click on the **continue** button, and proceed with writing changes on the disk.



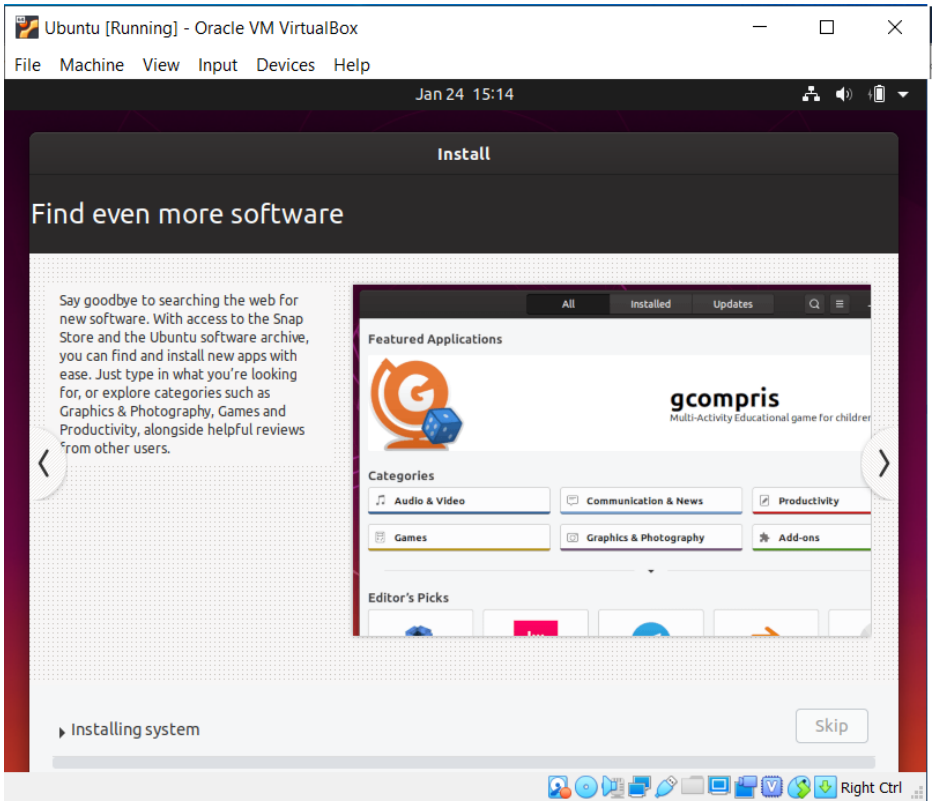
Step 16: Select your location to set the Time Zone.



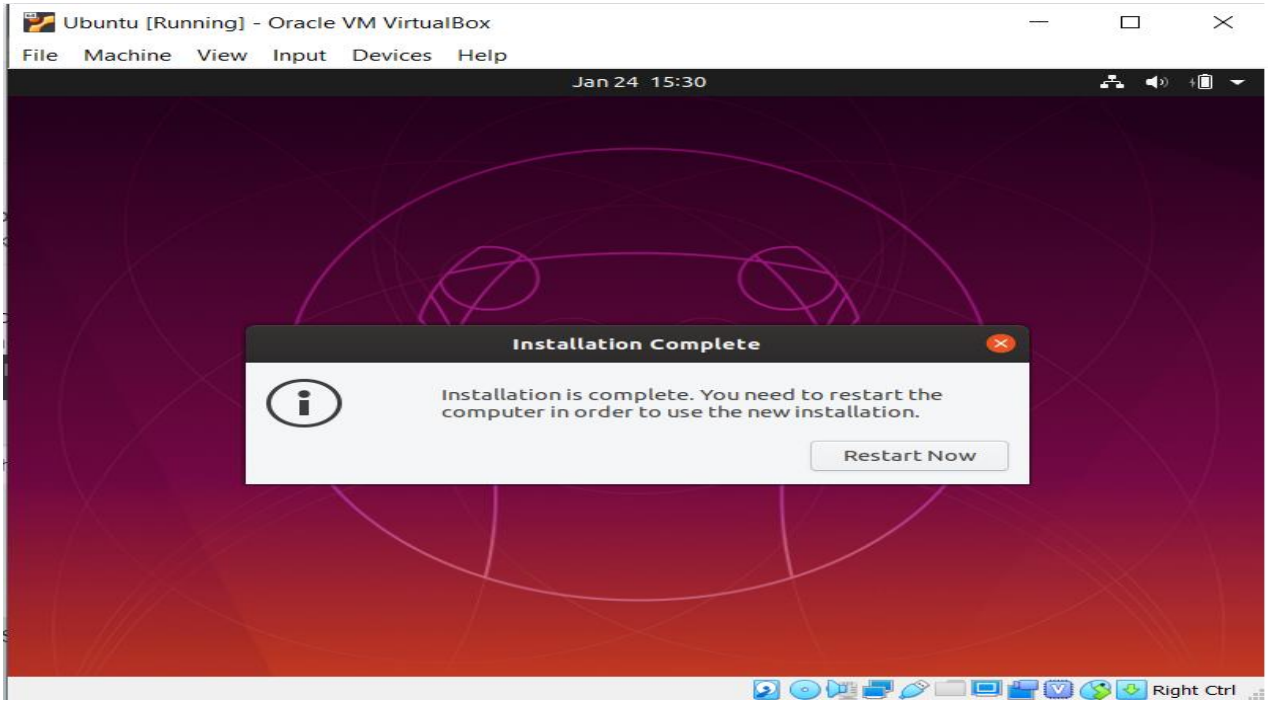
Step 17: Choose a name for your computer and set a password to secure login info.



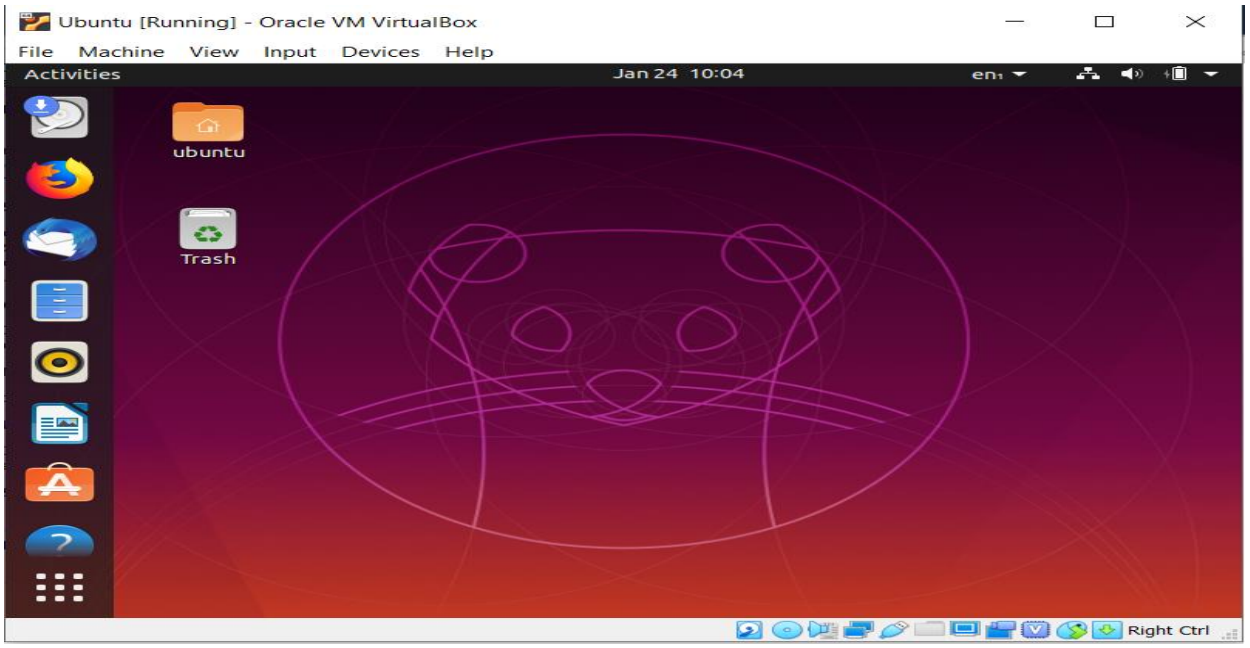
Step 18: Wait for the installation process to complete.



Step 19: Once the installation process is over, reboot your Virtual Machine.



Step 20: You're finished with the installation process. Now you can use Ubuntu along with the Windows, without creating a dual boot.

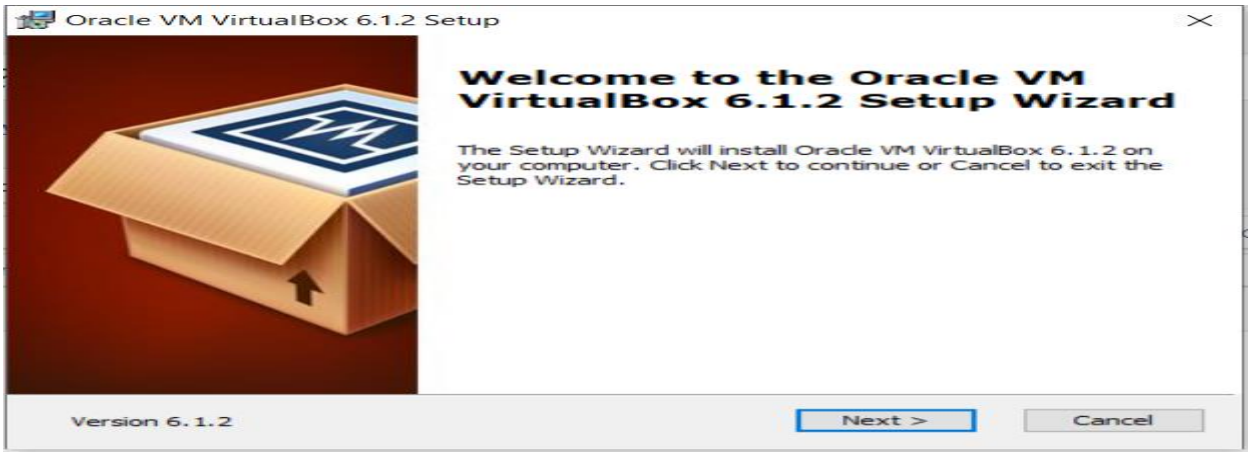


Downloading and Installing VirtualBox on window

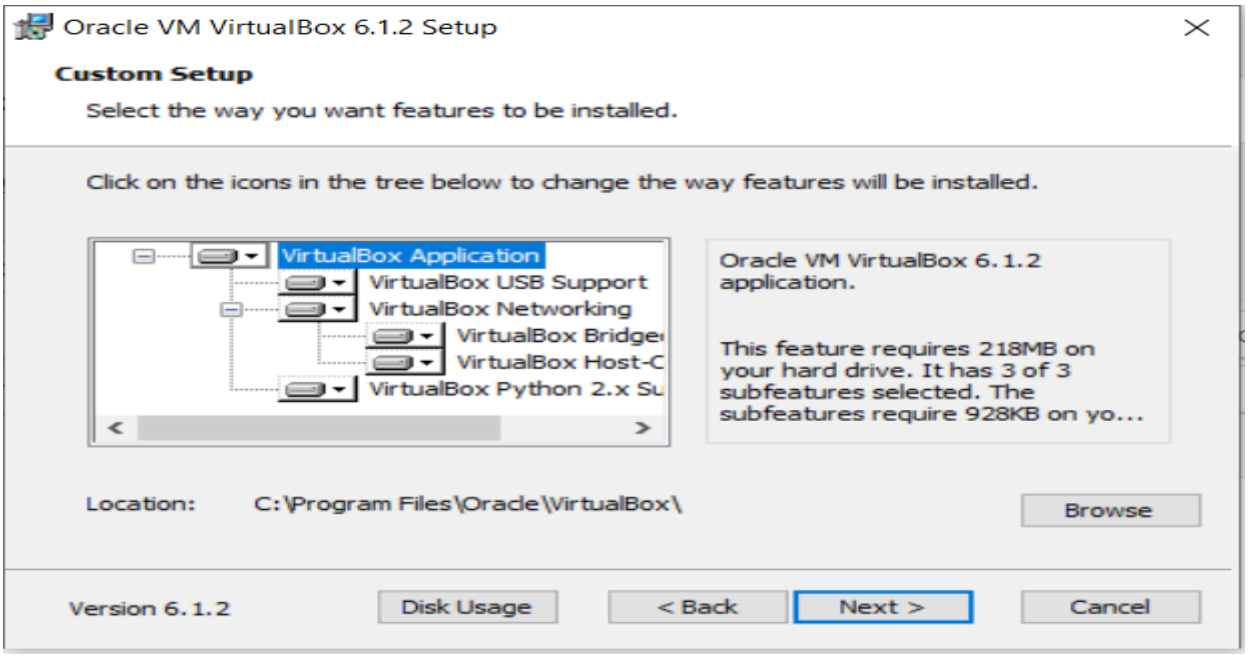
To download Virtual Box, go to the official site **virtualbox.org** and download the latest version for windows.

Beginning with the Installation:

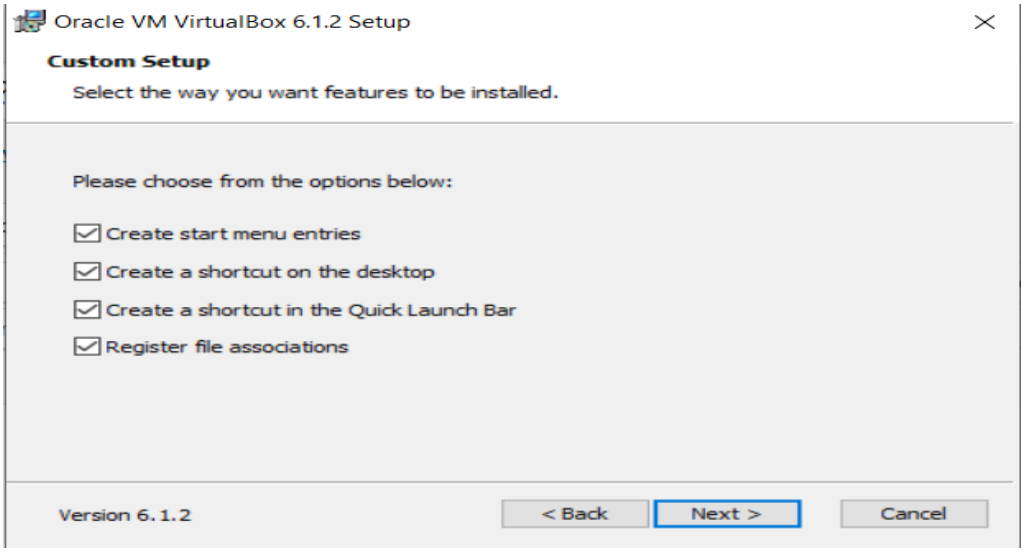
Getting Started:



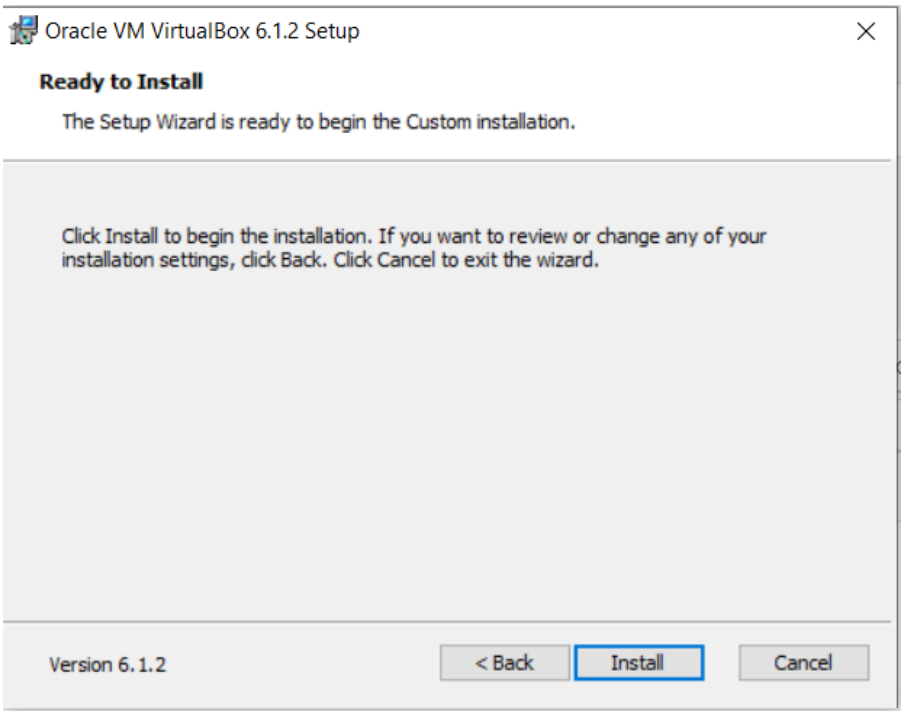
- **Select Installation Location:**



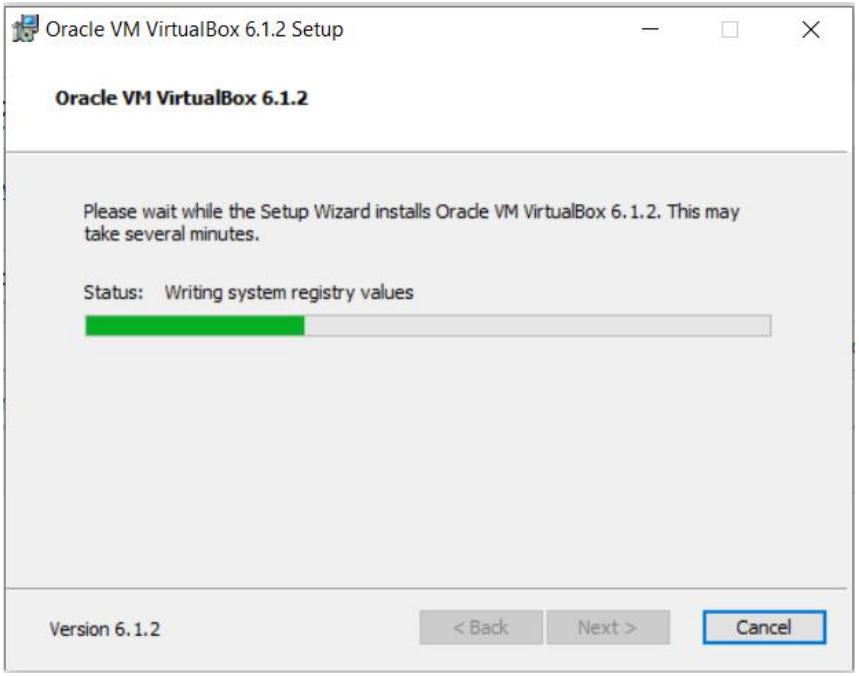
- **Creating Entries and Shortcuts:**



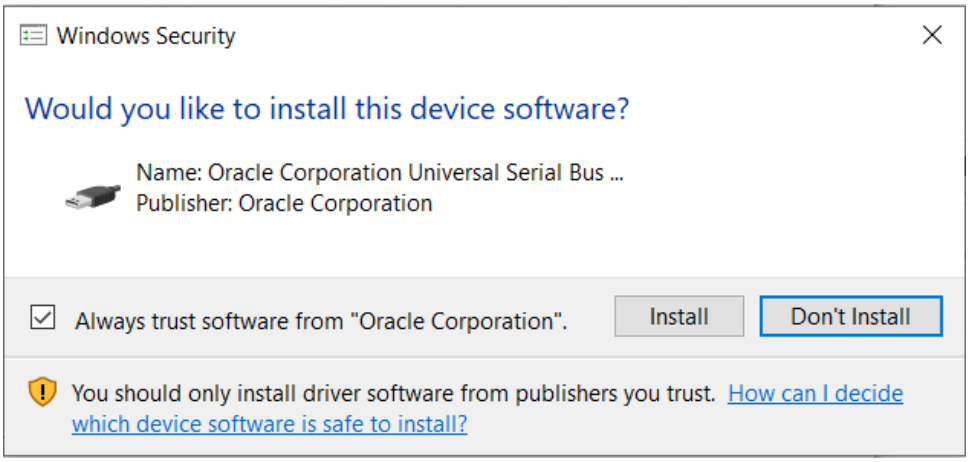
- **Ready to Install:**



- **Installing Files and packages:**

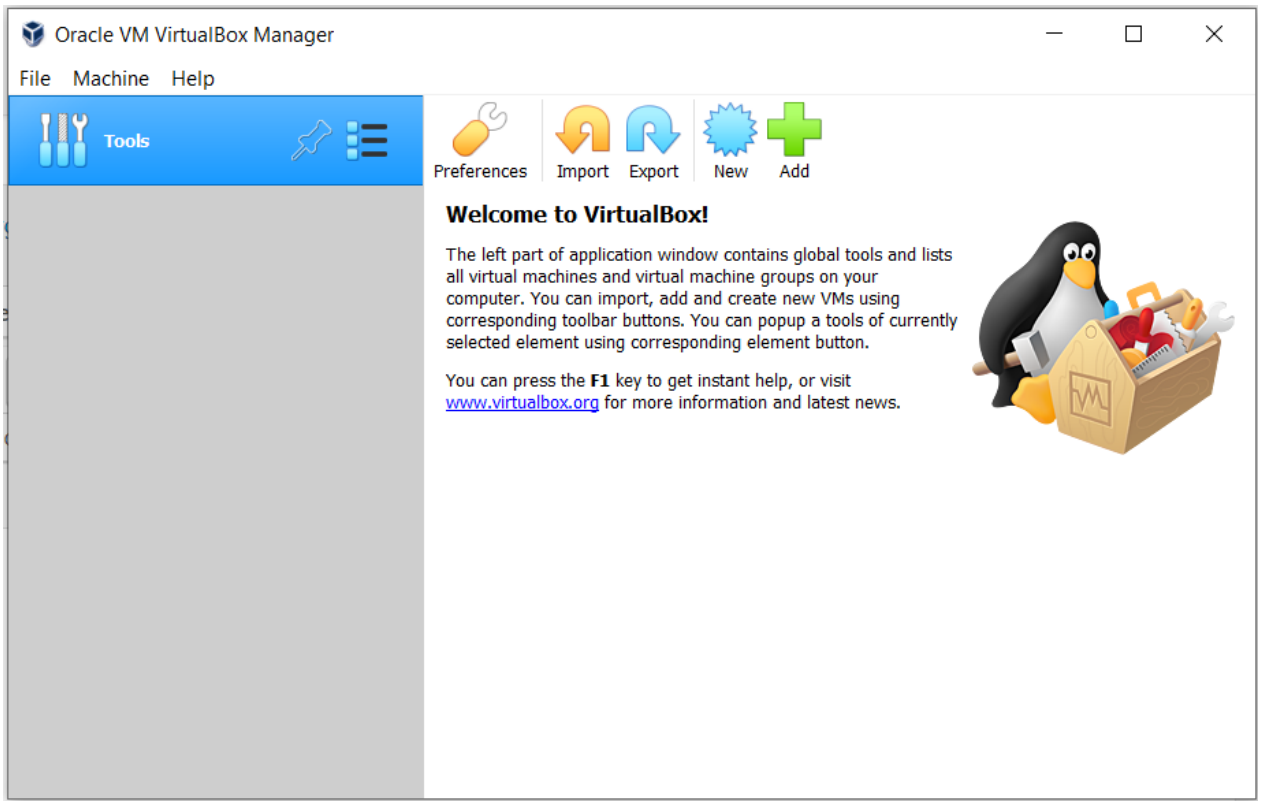


- **Installing Certificates:**



- **Finished Installation:**

When you will open virtualbox it will look like as shown below:



Hadoop:

The definition of a powerful person has changed in this world. A powerful is one who has access to the data. This is because data is increasing at a tremendous rate. Suppose we are living in 100% data world. Then 90% of the data is produced in the last 2 to 4 years. This is because now when a child is born, before her mother, she first faces the flash of the camera. All these pictures and videos are nothing but data. Similarly, there is data of emails, various smartphone applications, statistical data, etc. All this data has the enormous power to affect various incidents and trends. This data is not only used by companies to affect their consumers but also by politicians to affect elections. This huge data is referred to as **Big Data**. In such a world, where data is being produced at such an exponential rate, it needs to maintain, analyzed, and tackled. This is where Hadoop creeps in.

Hadoop is an Apache open-source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Big Data was defined by the "***3Vs***" but now there are "***5Vs***" of Big Data which are also termed as the characteristics of Big Data.

1. **Volume:** With increasing dependence on technology, data is producing at a large volume. Common examples are data being produced by various social networking sites, sensors, scanners, airlines, and other organizations.

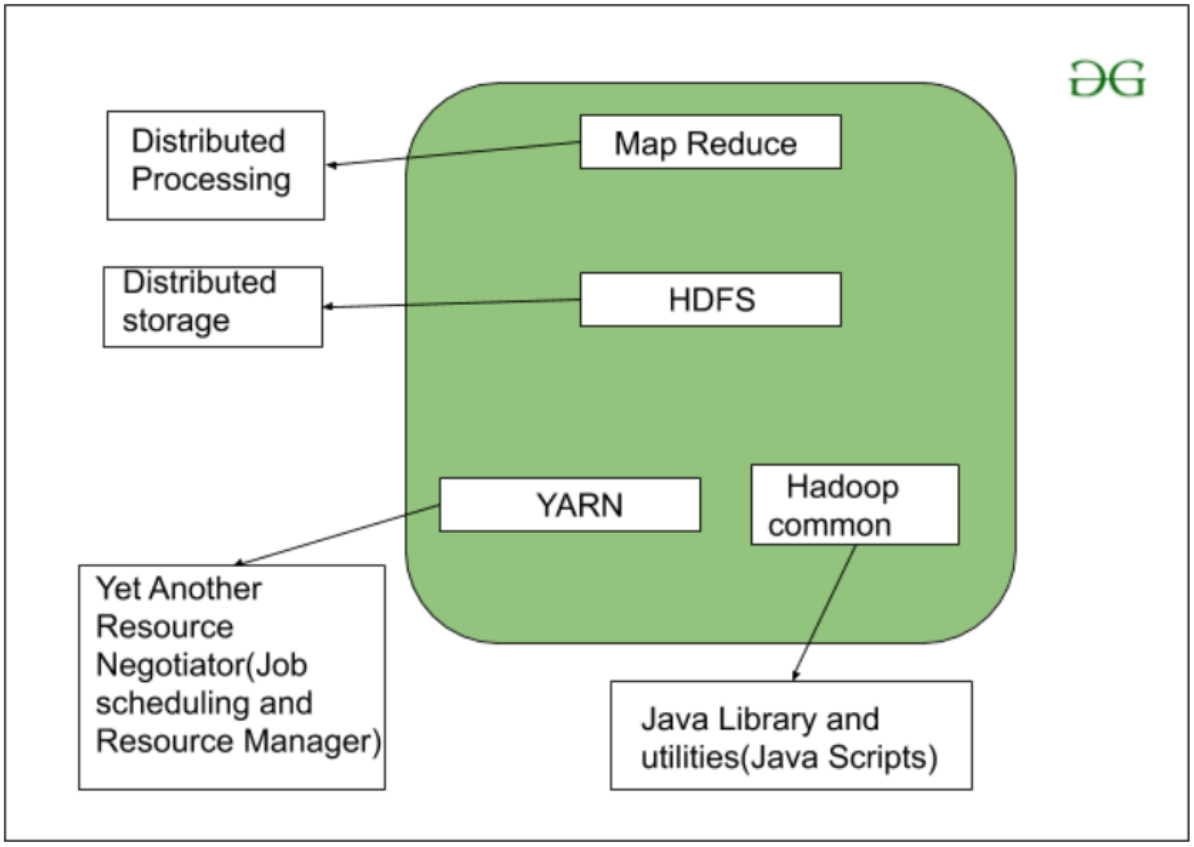
2. **Velocity:** Huge amount of data is generated per second. It is estimated that by the end of 2020, every individual will produce 3mb data per second. This large volume of data is being generated with a great velocity.

3. **Variety:** The data being produced by different means is of three types:
    - **Structured Data:** It is the relational data which is stored in the form of rows and columns.
    - **Unstructured Data:** Texts, pictures, videos etc. are the examples of unstructured data which can't be stored in the form of rows and columns.
    - **Semi Structured Data:** Log files are the examples of this type of data.

4. **Veracity:** The term Veracity is coined for the inconsistent or incomplete data which results in the generation of doubtful or uncertain Information. Often data inconsistency arises because of the volume or amount of data e.g., data in bulk could create confusion whereas less amount of data could convey half or incomplete Information.

5. **Value:** After having the 4 V's into account there comes one more V which stands for Value! Bulk of Data having no Value is of no good to the company, unless you turn it into something useful. Data in itself is of no use or importance, but it needs to be converted into something valuable to extract Information. Hence, you can state that Value! is the most important V of all the 5V's.

Architecture Of Hadoop:



Install Hadoop: Setting up a Single Node Hadoop Cluster

You must have got a theoretical idea about Hadoop, HDFS and its architecture. But to get Hadoop Certified you need good hands-on knowledge. I hope you would have liked our previous blog on *HDFS Architecture*, now I will take you through the practical knowledge about Hadoop and HDFS. The first step forward is to install Hadoop.
There are two ways to install Hadoop, i.e. Single node and Multi-node.

A single node cluster means only one DataNode running and setting up all the NameNode, DataNode, ResourceManager, and NodeManager on a single machine. This is used for studying and testing purposes. For example, let us consider a sample data set inside the healthcare industry. So, for testing whether the Oozie jobs have scheduled all the processes like collecting, aggregating, storing, and processing the data in a proper sequence, we use a single node cluster. It can easily and efficiently test the sequential workflow in a smaller environment as compared to large environments which contain terabytes of data distributed across hundreds of machines.

While in a Multi-node cluster, there are more than one DataNode running and each DataNode is running on different machines. The multi-node cluster is practically used in organizations for analyzing Big Data. Considering the above example, in real-time when we deal with petabytes of data, it needs to be distributed across hundreds of machines to be processed.
Thus, here we use a multi-node cluster.
In this blog, I will show you how to install Hadoop on a single node cluster. You can get a better understanding from the Hadoop Admin Training in Bangalore.

Prerequisites

- *VIRTUAL BOX*: it is used for installing the operating system on it.
- *OPERATING SYSTEM*: You can install Hadoop on Linux-based operating systems. Ubuntu and CentOS are very commonly used. In this tutorial, we are using CentOS.
- *JAVA*: You need to install the Java 8 package on your system.
- *HADOOP*: You require Hadoop 2.7.3 package.

Install Hadoop

Step1: https://drive.google.com/file/d/0BwIBPXSoIx_FTk1RN2IzMU9zMm8/view?usp=sharingutm_source%3Dblog&utm_medium=blog-internal-search-box&utm_term=install-hadoop to download the Java 8 Package. Save this file in your home directory.

Step 2: Extract the Java Tar File.
***Command*:** tar -xvf jdk-8u101-linux-i586.tar.gz

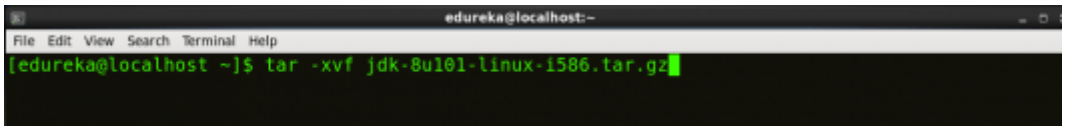


Fig: Hadoop Installation – Extracting Java Files

Step 3: Download the Hadoop 2.7.3 Package.
***Command*:** wget https://archive.apache.org/dist/hadoop/core/hadoop-2.7.3/hadoop-2.7.3.tar.gz

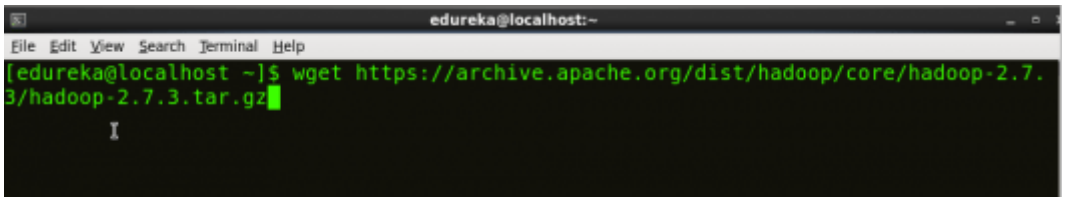


Fig: Hadoop Installation – Downloading Hadoop

Step 4: Extract the Hadoop tar File.
***Command*:** tar -xvf hadoop-2.7.3.tar.gz

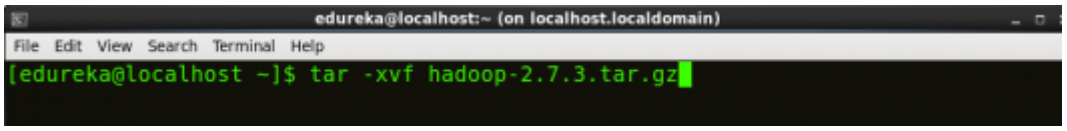
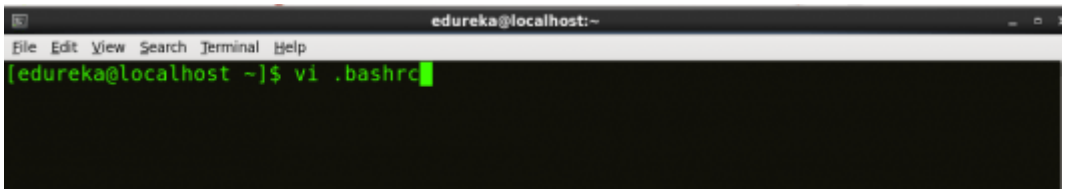


Fig: Hadoop Installation – Extracting Hadoop Files

Step 5: Add the Hadoop and Java paths in the bash file (.bashrc).
Open**. bashrc** file. Now, add Hadoop and Java Path as shown below. Learn more about the Hadoop Ecosystem and its tools with the Hadoop Certification.

***Command*:** vi .bashrc



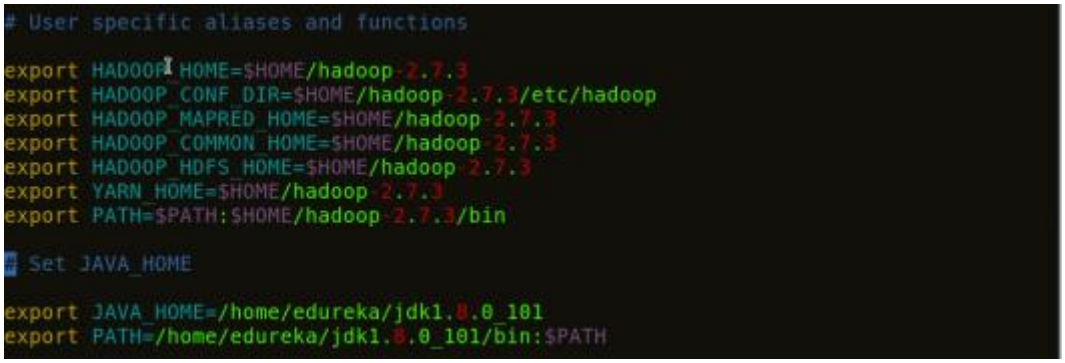


Fig: Hadoop Installation – Setting Environment Variable

Then, save the bash file and close it.

For applying all these changes to the current Terminal, execute the source command.

***Command*:** source. Bashrc

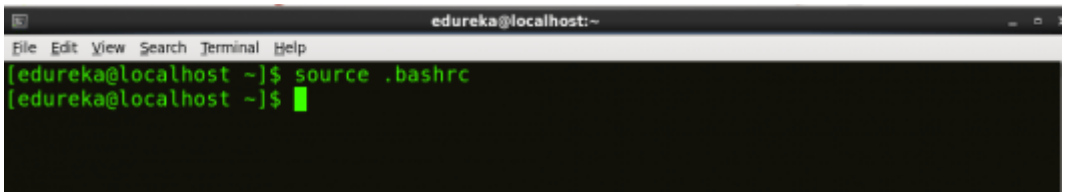


Fig: Hadoop Installation – Refreshing environment variables

To make sure that Java and Hadoop have been properly installed on your system and can be accessed through the Terminal, execute the java -version and Hadoop version commands.

***Command*:** java -version

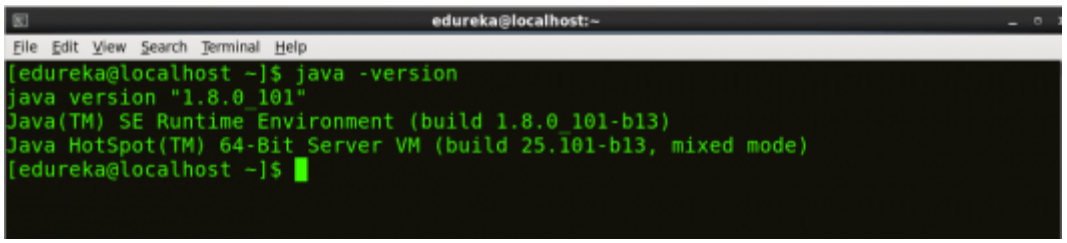


Fig: Hadoop Installation – Checking Java Version

***Command*:** hadoop version

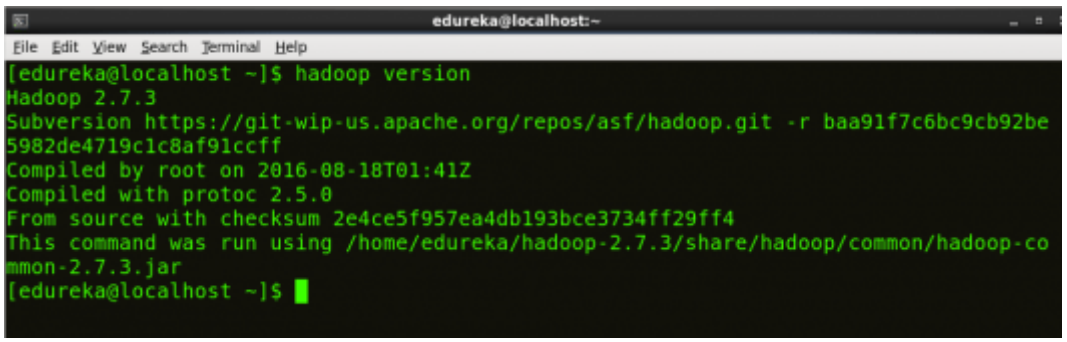


Fig: Hadoop Installation – Checking Hadoop Version

Step 6: Edit the **Hadoop Configuration files**.
Command: cd hadoop-2.7.3/etc/hadoop/

Command: ls All the Hadoop configuration files are located in **hadoop-2.7.3/etc/Hadoop** directory as you can see in the snapshot below:

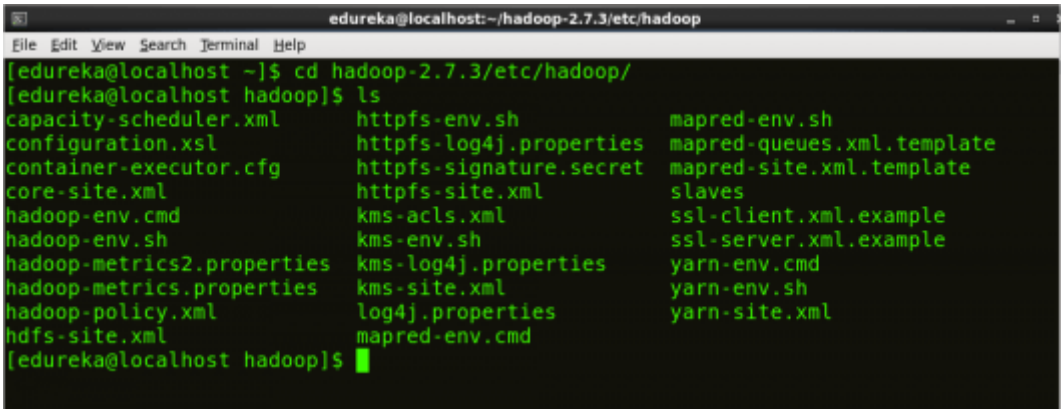


Fig: Hadoop Installation – Hadoop Configuration Files

Step 7: Open *core-site.xml* and edit the property mentioned below inside configuration tag:

core-site.xml informs Hadoop daemon where NameNode runs in the cluster. It contains configuration settings of Hadoop core such as I/O settings that are common to HDFS & MapReduce.

***Command*:** vi core-site.xml

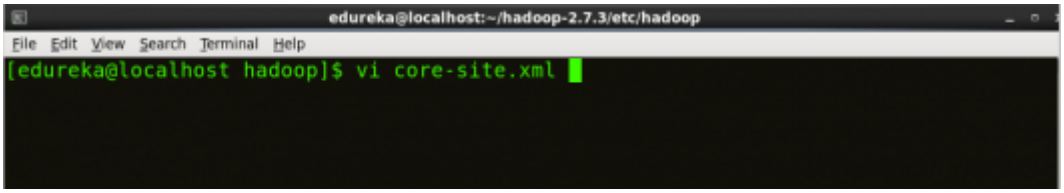


Fig: Hadoop Installation – Configuring core-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

Step 8: Edit *hdfs-site.xml* and edit the property mentioned below inside configuration tag:

hdfs-site.xml contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, Secondary NameNode). It also includes the replication factor and block size of HDFS.

***Command*:** vi hdfs-site.xml

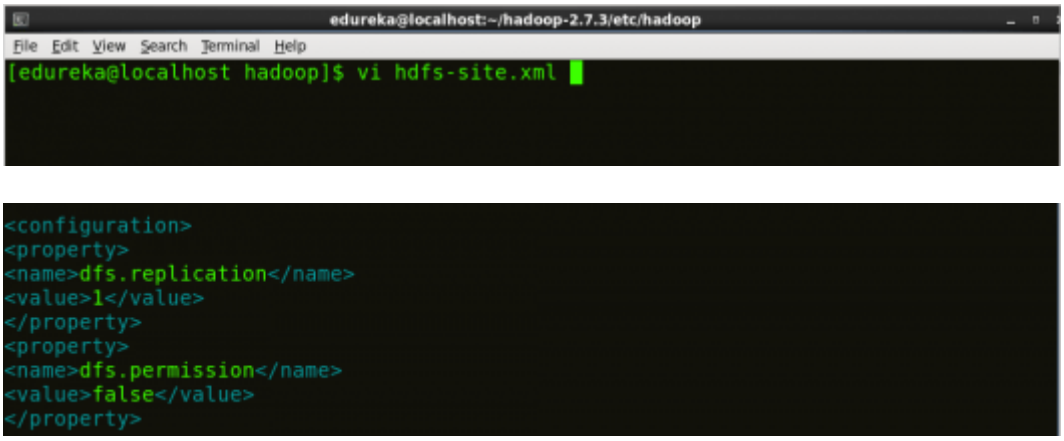


Fig: Hadoop Installation – Configuring hdfs-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.permission</name>
<value>false</value>
</property>
</configuration>
```

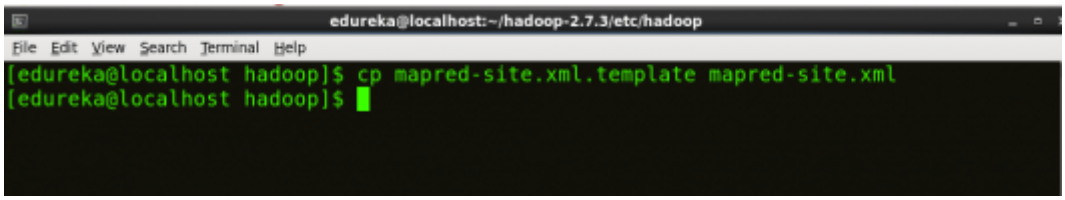
Step 9: Edit the *mapred-site.xml* file and edit the property mentioned below inside configuration tag:

mapred-site.xml contains configuration settings of MapReduce application like number of JVM that can run in parallel, the size of the mapper and the reducer process, CPU cores available for a process, etc.

In some cases, mapred-site.xml file is not available. So, we have to create the mapred-site.xml file using mapred-site.xml template.

***Command*:** cp mapred-site.xml.template mapred-site.xml

***Command*:** vi mapred-site.xml.



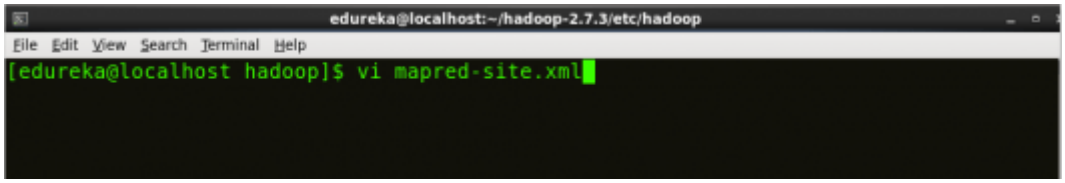


Fig: Hadoop Installation – Configuring mapred-site.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

Step 10: Edit *yarn-site.xml* and edit the property mentioned below inside configuration tag:

yarn-site.xml contains configuration settings of ResourceManager and NodeManager like application memory management size, the operation needed on program & algorithm, etc.

***Command*:** vi yarn-site.xml

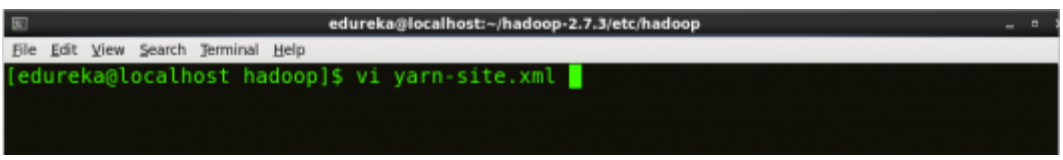


Fig: Hadoop Installation – Configuring yarn-site.xml

```
<?xml version="1.0">
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.auxservices.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
</configuration>
```

Step 11: Edit *hadoop-env.sh* and add the Java Path as mentioned below:
hadoop-env.sh contains the environment variables that are used in the script to run Hadoop like Java home path, etc.

***Command*:** vi hadoop–env.sh

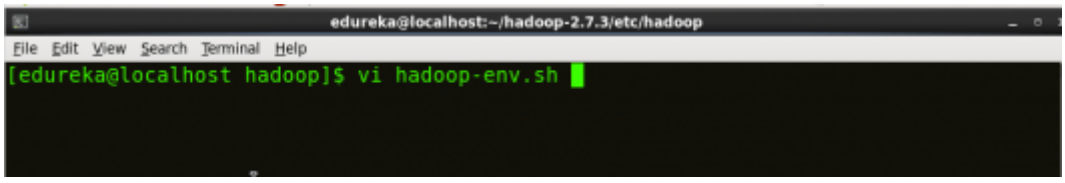


Fig: Hadoop Installation – Configuring hadoop-env.sh

Step 12: Go to Hadoop home directory and format the NameNode.
***Command*:** cd

***Command*:** cd hadoop-2.7.3

***Command*:** bin/hadoop namenode -format

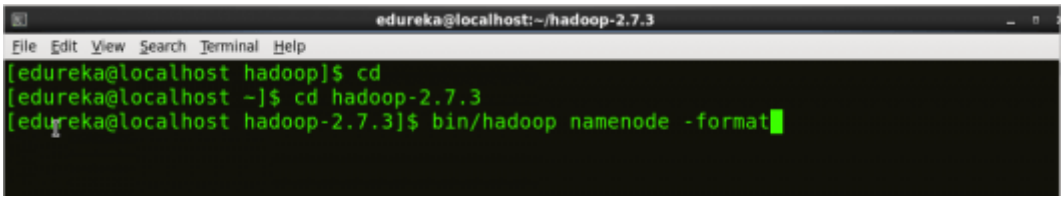


Fig: Hadoop Installation – Formatting NameNode

This formats the HDFS via NameNode. This command is only executed for the first time. Formatting the file system means initializing the directory specified by the dfs.name.dir variable.

Never format, up and running Hadoop filesystem. You will lose all your data stored in the HDFS.

Step 13: Once the NameNode is formatted, go to hadoop-2.7.3/sbin directory and start all the daemons.

Command: cd hadoop-2.7.3/sbin

Either you can start all daemons with a single command or do it individually.

Command: ./start-all.sh

The above command is a combination of ***start-dfs.sh, start-yarn.sh*** & ***mr-jobhistory-daemon.sh***

Or you can run all the services individually as below:

Start NameNode:

The NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files stored in the HDFS and tracks all the file stored across the cluster.

Command: ./hadoop-daemon.sh start namenode

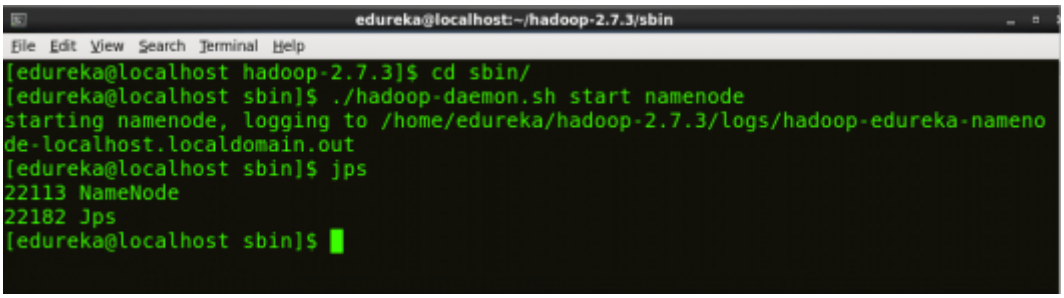


Fig: Hadoop Installation – Starting NameNode

Start DataNode:

On startup, a DataNode connects to the Namenode and it responds to the requests from the Namenode for different operations.

Command: ./hadoop-daemon.sh start datanode

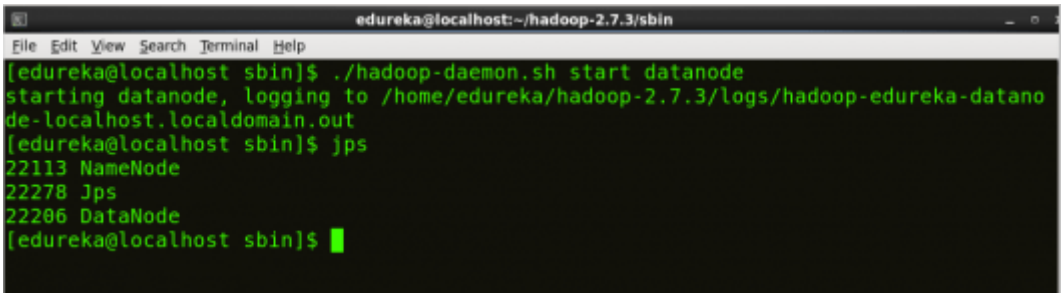


Fig: Hadoop Installation – Starting DataNode

Start ResourceManager:

ResourceManager is the master that arbitrates all the available cluster resources and thus helps in managing the distributed applications running on the YARN system. Its work is to manage each NodeManagers and the each application's ApplicationMaster.

Command: ./yarn-daemon.sh start resourcemanager

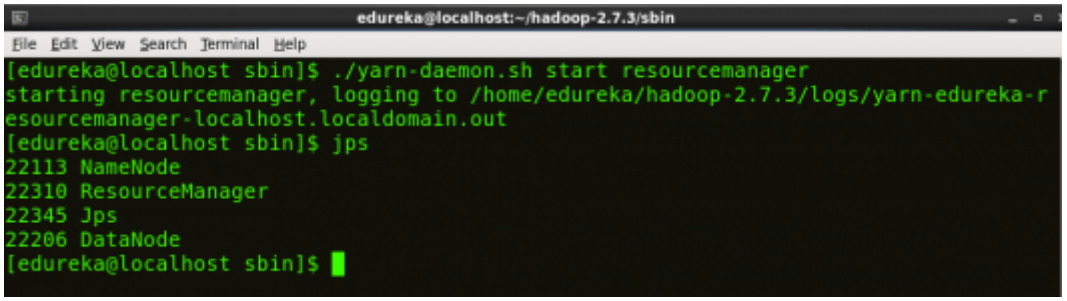
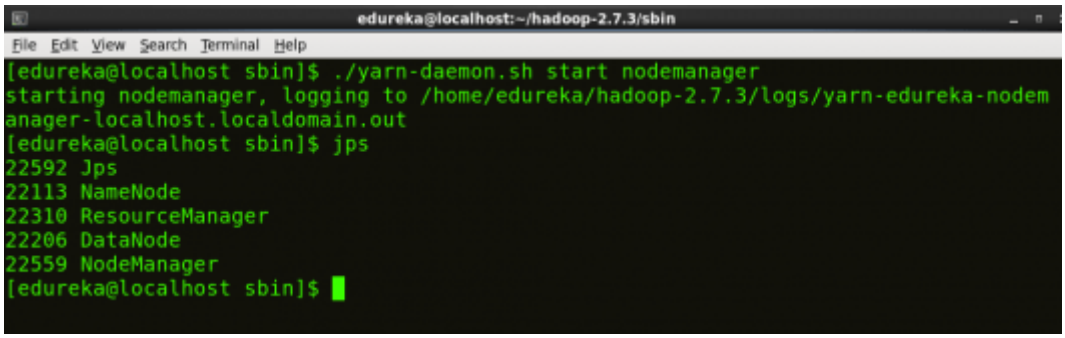


Fig: Hadoop Installation – Starting ResourceManager

Start NodeManager:

The NodeManager in each machine framework is the agent which is responsible for managing containers, monitoring their resource usage and reporting the same to the ResourceManager.

Command: ./yarn-daemon.sh start nodemanager



Start JobHistoryServer:

JobHistoryServer is responsible for servicing all job history related requests from client.

Command: ./mr-jobhistory-daemon.sh start historyserver

Step 14: To check that all the Hadoop services are up and running, run the below command.

Command: jps

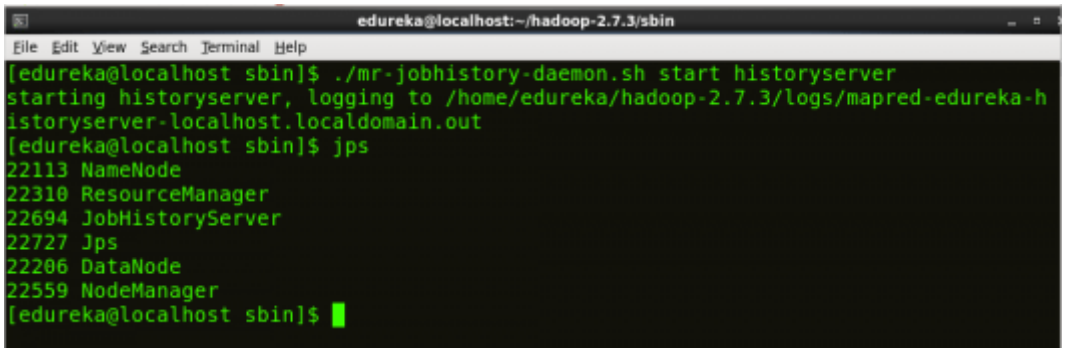


Fig: Hadoop Installation – Checking Daemons

Step 15: Now open the Mozilla browser and go to **localhost**:**50070/dfshealth.html** to check the NameNode interface.

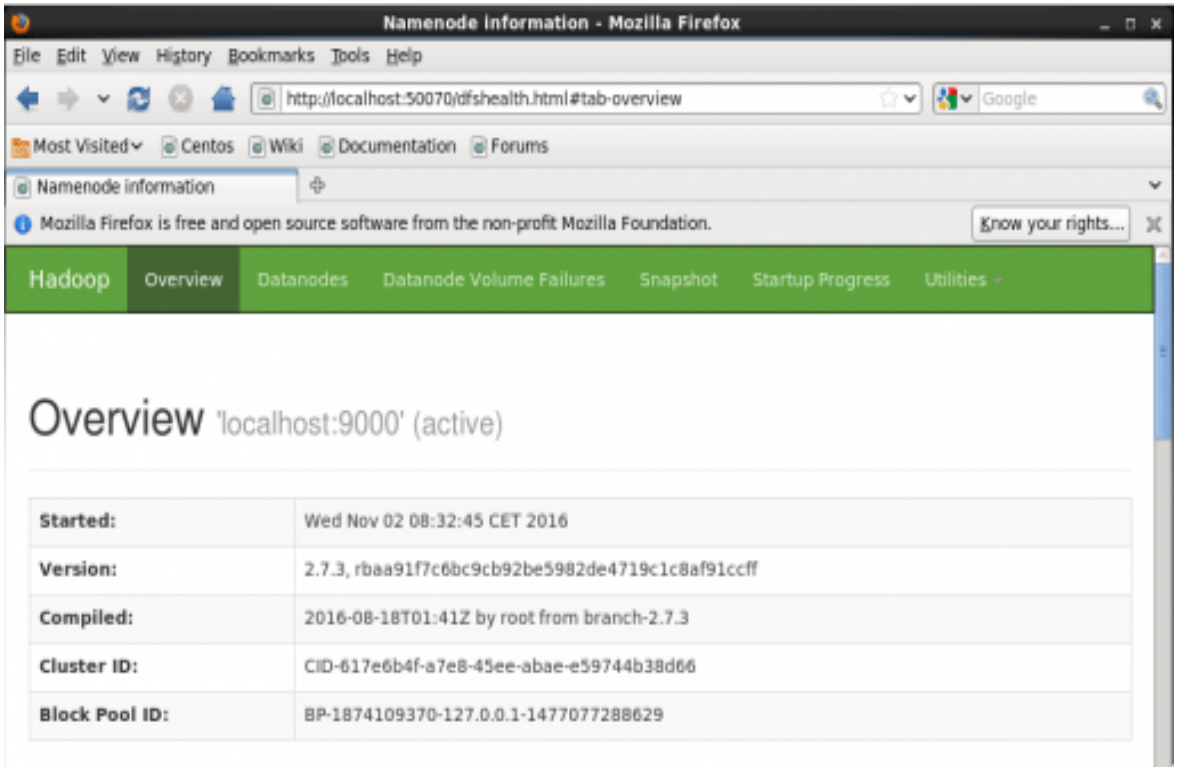


Fig: Hadoop Installation – Starting WebUI

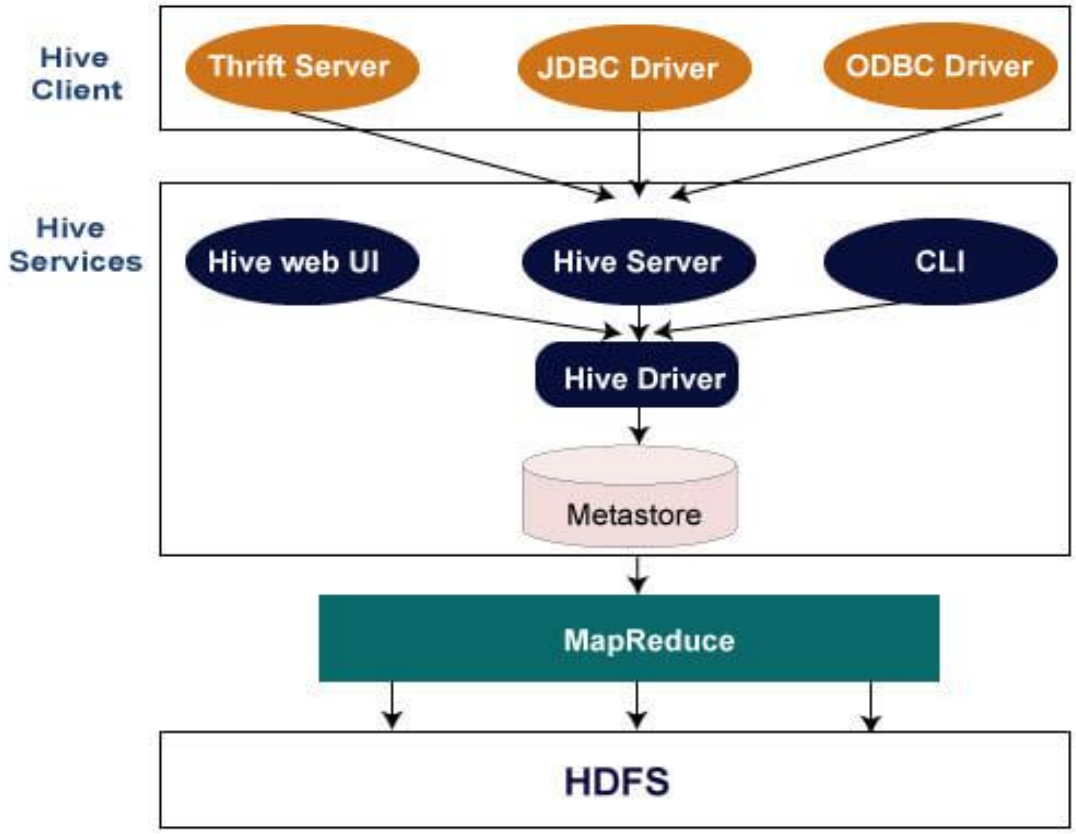
Congratulations, you have successfully installed a single-node Hadoop cluster in one go.

HIVE

Apache Hive is a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale. A data warehouse provides a central store of information that can easily be analyzed to make informed, data driven decisions. Hive allows users to read, write, and manage petabytes of data using SQL.

Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data. What makes Hive unique is the ability to query large datasets, leveraging Apache Tez or MapReduce, with a SQL-like interface.

Architecture Of Hive:



Map Reduce

MapReduce is a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster. As the processing component, MapReduce is the heart of Apache Hadoop.

MapReduce is a programming model that is used for processing and generating large data sets on clusters of computers. It was introduced by Google. MapReduce is a concept or a method for large scale parallelization. It is inspired by functional programming's map () and reduce () functions.
MapReduce program is executed in three stages they are:

- **Mapping:** Mapper's job is to process input data. Each node applies the map function to the local data.

- **Shuffle:** Here nodes are redistributed where data is based on the output keys. (Output keys are produced by map function).

- **Reduce:** Nodes are now processed into each group of output data, per key in parallel.

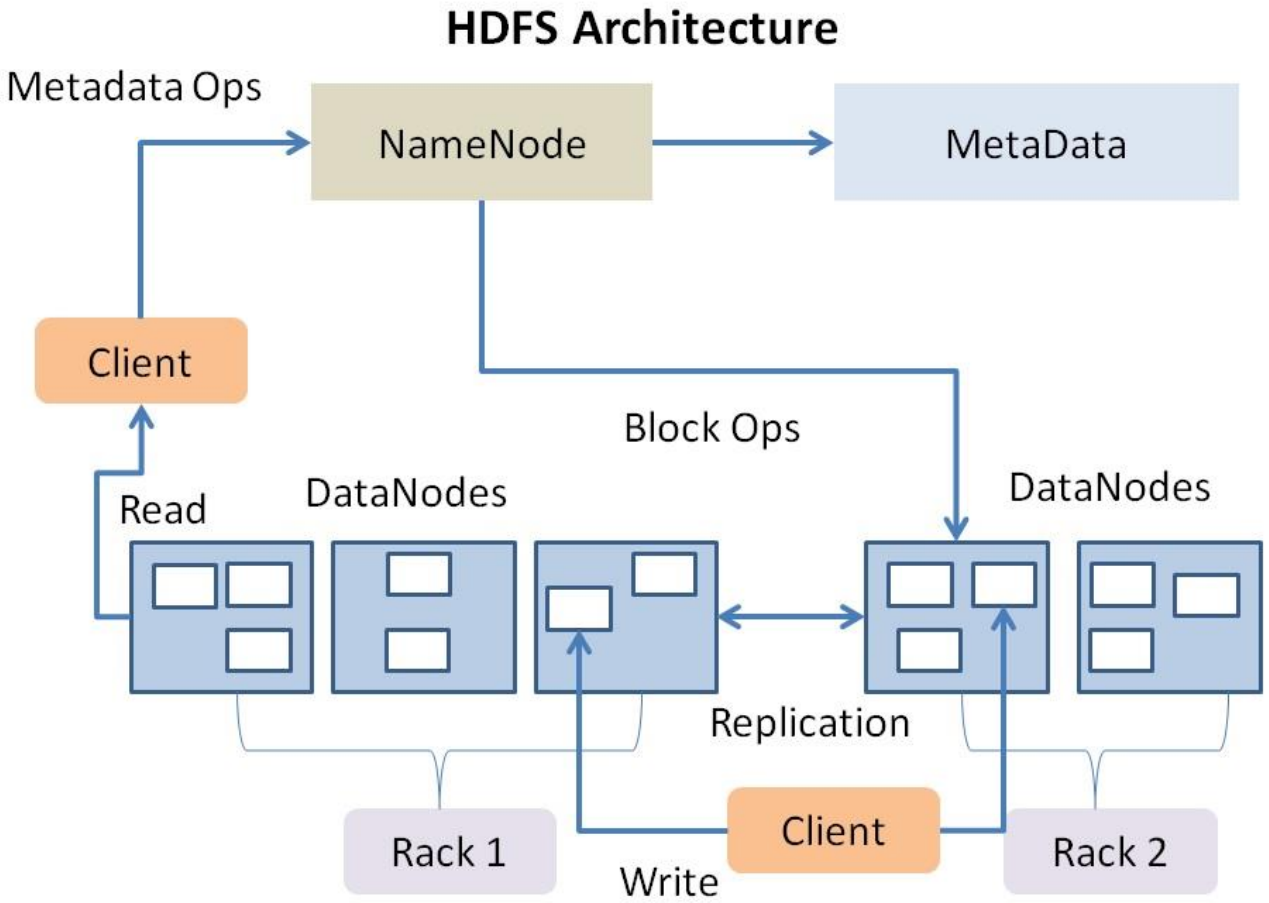
HDFS

HDFS is a distributed file system that handles large data sets running on commodity hardware. It is used to scale a single Apache Hadoop cluster to hundreds (and even thousands) of nodes. HDFS is one of the major components of Apache Hadoop, the others being map reduce and YARN.

Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly faulttolerant and designed using low-cost hardware.
HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

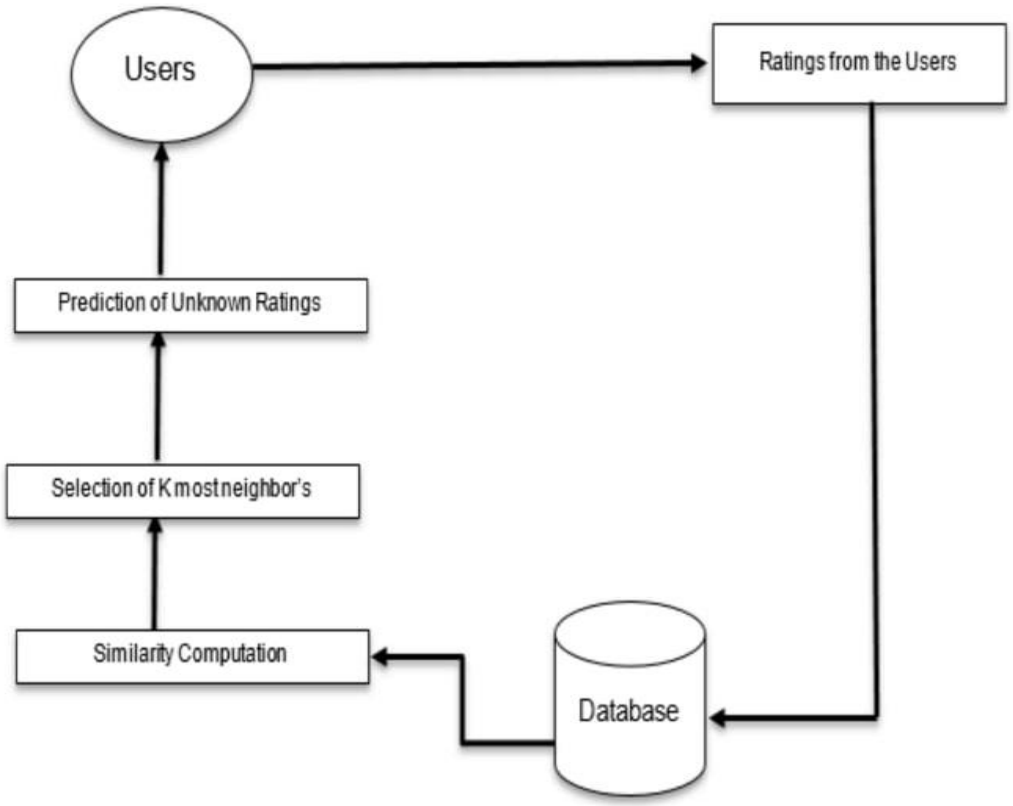
Features of HDFS:

- It is suitable for the distributed storage and processing.
- Hadoop provides a command interface to interact with HDFS.
- The built-in servers of namenode and datanode help users to easily check the status of cluster.
- Streaming access to file system data.
- HDFS provides file permissions and authentication.


HDFS Architecture
Metadata Ops
NameNode
MetaData
Client
Block Ops
Read
DataNodes
DataNodes
Replication
Rack 1
Client
Write
Rack 2

Chapter 4 - Database Design:

The Architecture of Movie Recommendation System:

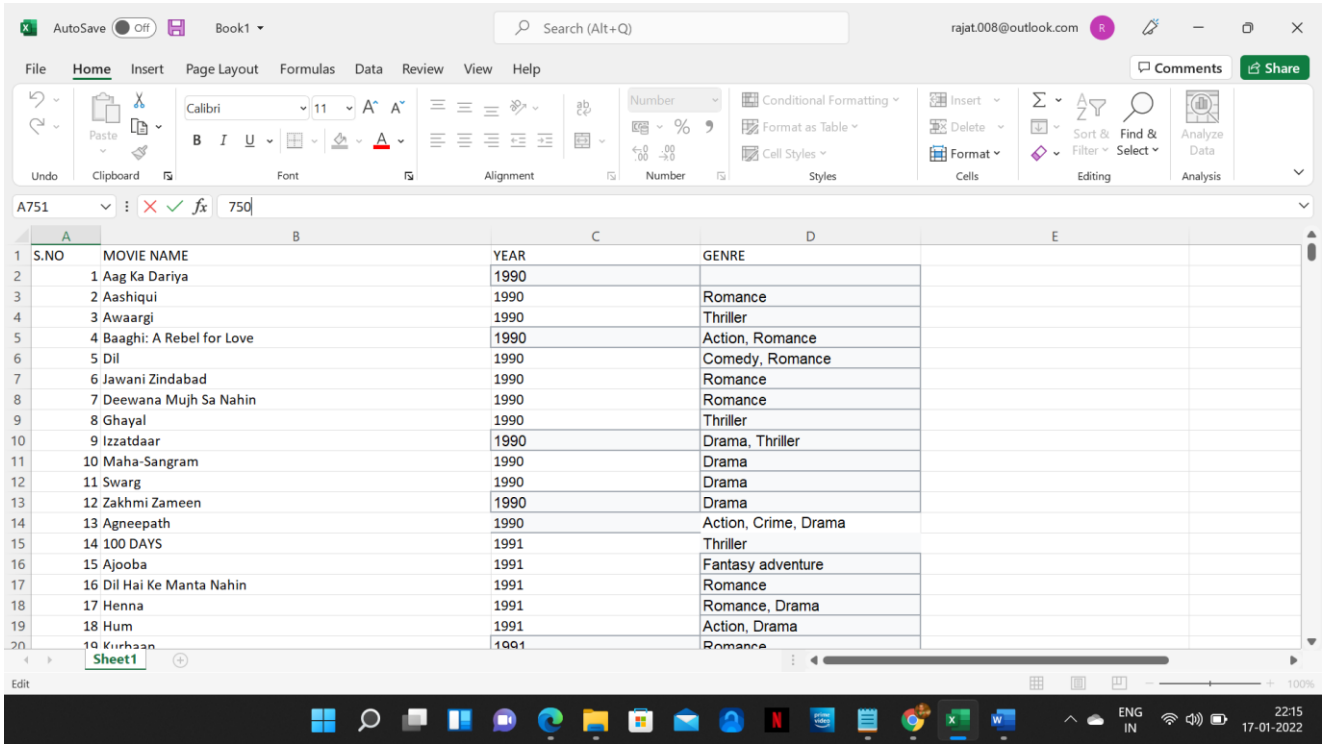


Dataset:

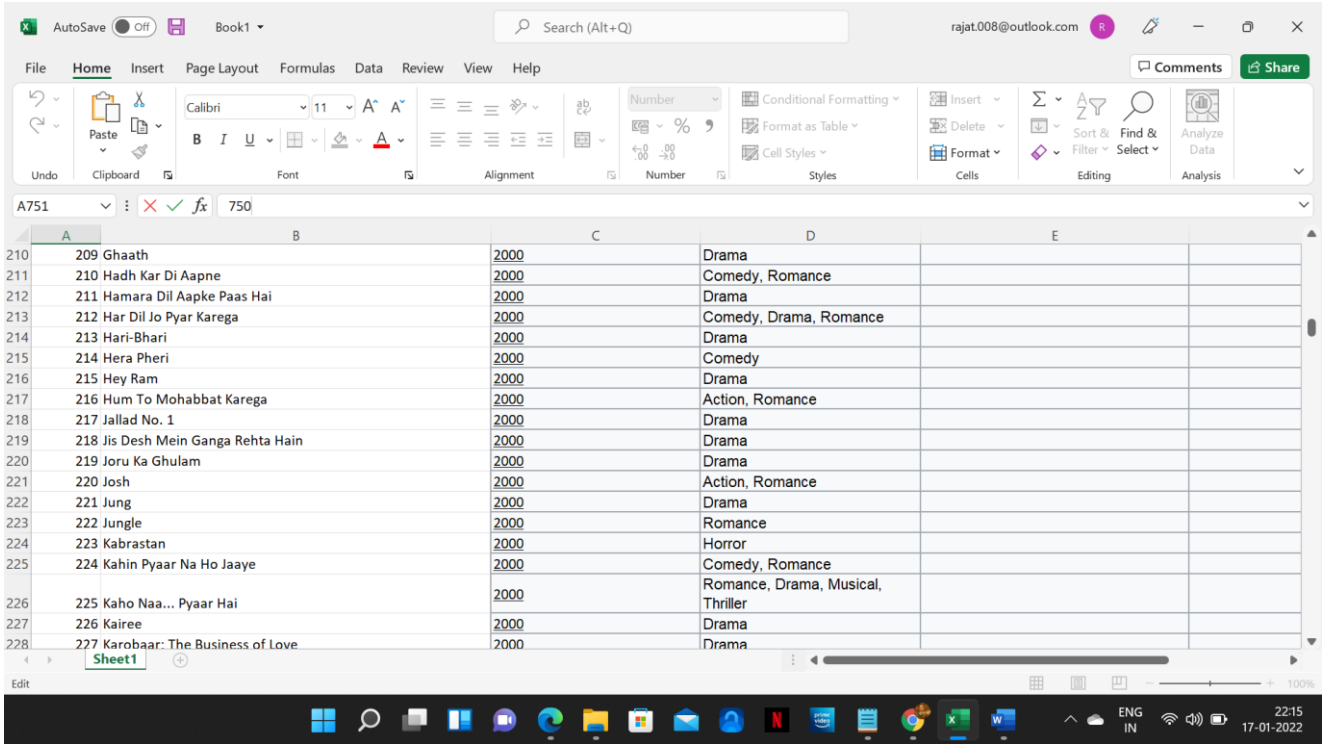
A dataset is a collection of records, like a relational database table. Records are like table rows, but the columns can contain not only strings or numbers, but also nested data structures such as lists, maps, and other records.
Basically, we have stored our Dataset with the name Dataset with csv file format. The Dataset contains many movies with their releasing year (which movie has been released in which year) as well as Genre such as Comedy, Action, Drama, Family, Horror, Science Thriller, Romance, Biography.

Database Tables:

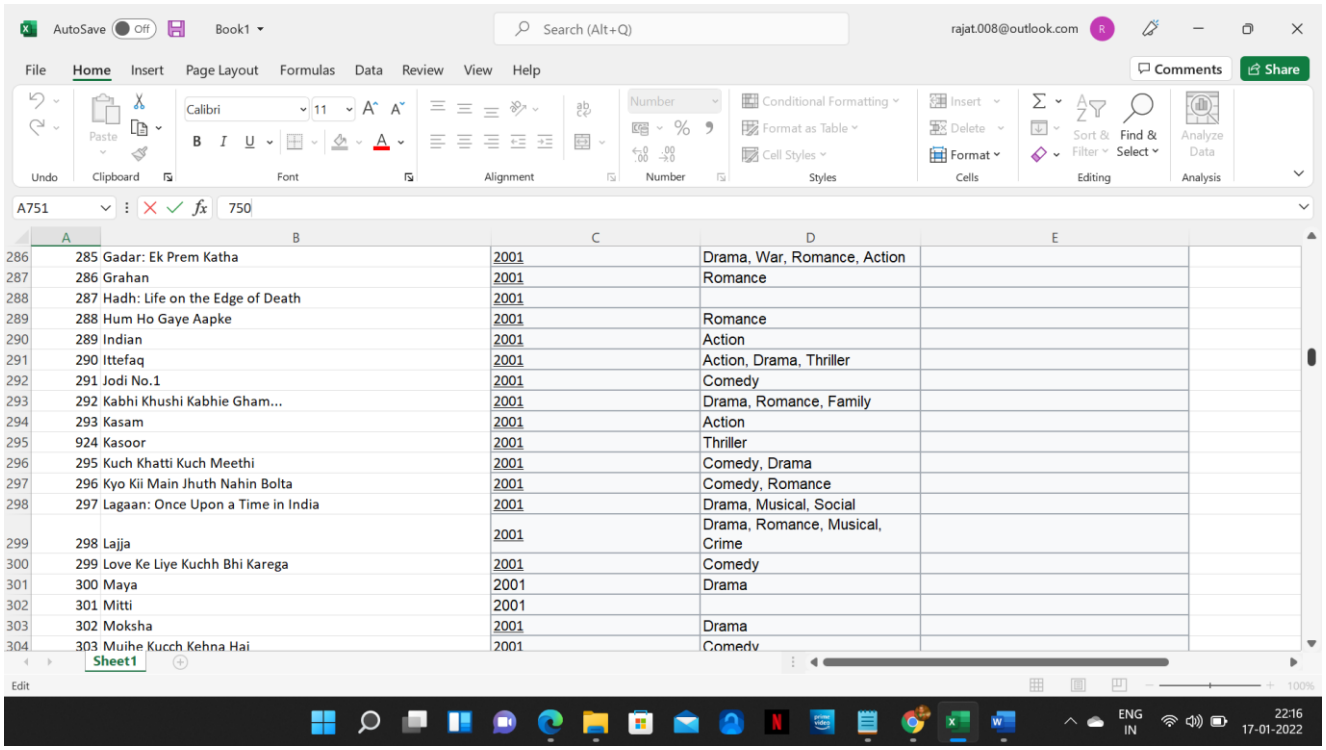
For the database portion, we use a table of dataset which include large number details of the movies. Screenshot of such Dataset is:



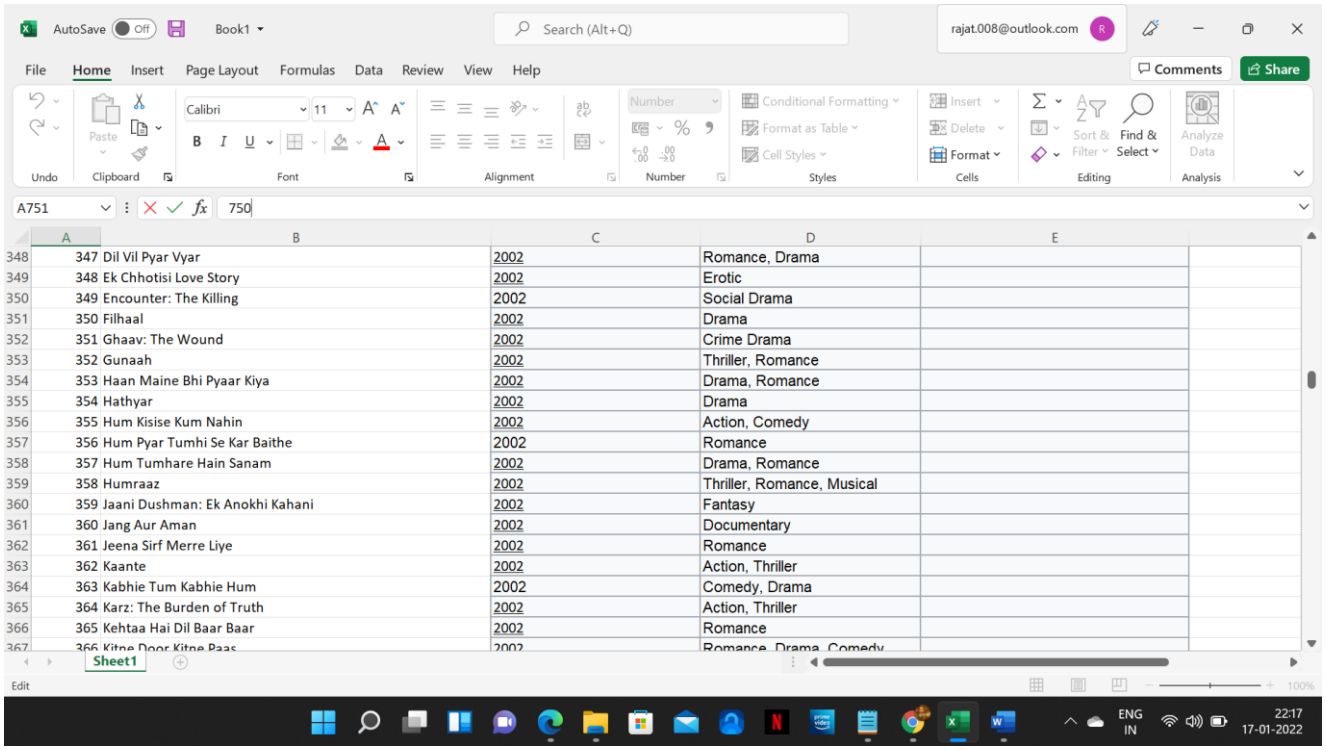
S.NO	MOVIE NAME	YEAR	GENRE
1	Aag Ka Dariya	1990	
2	Aashiqui	1990	Romance
3	Awaargi	1990	Thriller
4	Baaghi: A Rebel for Love	1990	Action, Romance
5	Dil	1990	Comedy, Romance
6	Jawani Zindabad	1990	Romance
7	Deewana Mujh Sa Nahin	1990	Romance
8	Ghayal	1990	Thriller
9	Izzatdaar	1990	Drama, Thriller
10	Maha-Sangram	1990	Drama
11	Swarg	1990	Drama
12	Zakhmi Zameen	1990	Drama
13	Agneepath	1990	Action, Crime, Drama
14	100 DAYS	1991	Thriller
15	Ajooba	1991	Fantasy adventure
16	Dil Hai Ke Manta Nahin	1991	Romance
17	Henna	1991	Romance, Drama
18	Hum	1991	Action, Drama
19	Kurbaan	1991	Romance



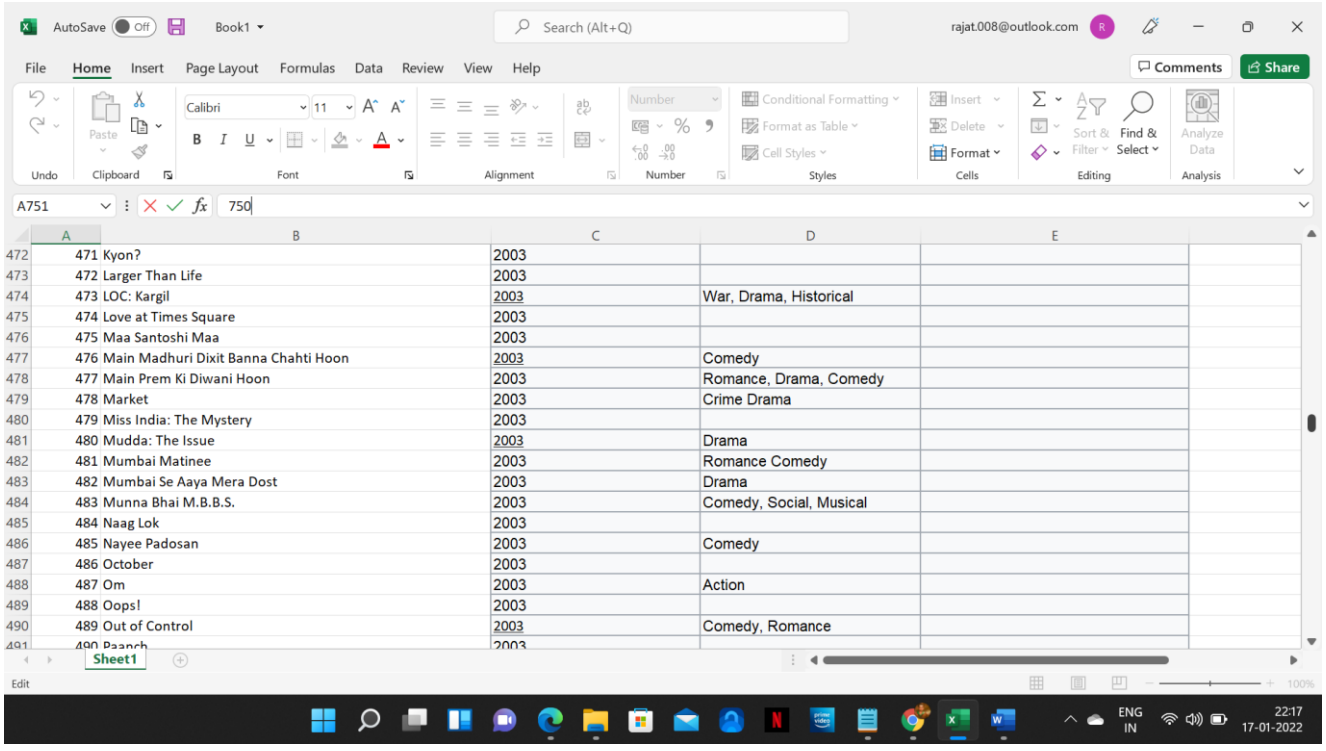
	A	B	C	D	E
210	209	Ghaath	2000	Drama	
211	210	Hadh Kar Di Aapne	2000	Comedy, Romance	
212	211	Hamara Dil Aapke Paas Hai	2000	Drama	
213	212	Har Dil Jo Pyar Karega	2000	Comedy, Drama, Romance	
214	213	Hari-Bhari	2000	Drama	
215	214	Hera Pheri	2000	Comedy	
216	215	Hey Ram	2000	Drama	
217	216	Hum To Mohabbat Karega	2000	Action, Romance	
218	217	Jallad No. 1	2000	Drama	
219	218	Jis Desh Mein Ganga Rehta Hain	2000	Drama	
220	219	Joru Ka Ghulam	2000	Drama	
221	220	Josh	2000	Action, Romance	
222	221	Jung	2000	Drama	
223	222	Jungle	2000	Romance	
224	223	Kabrastan	2000	Horror	
225	224	Kahin Pyaar Na Ho Jaaye	2000	Comedy, Romance	
226	225	Kaho Naa... Pyaar Hai	2000	Romance, Drama, Musical, Thriller	
227	226	Kairee	2000	Drama	
228	227	Karobaar: The Business of Love	2000	Drama	



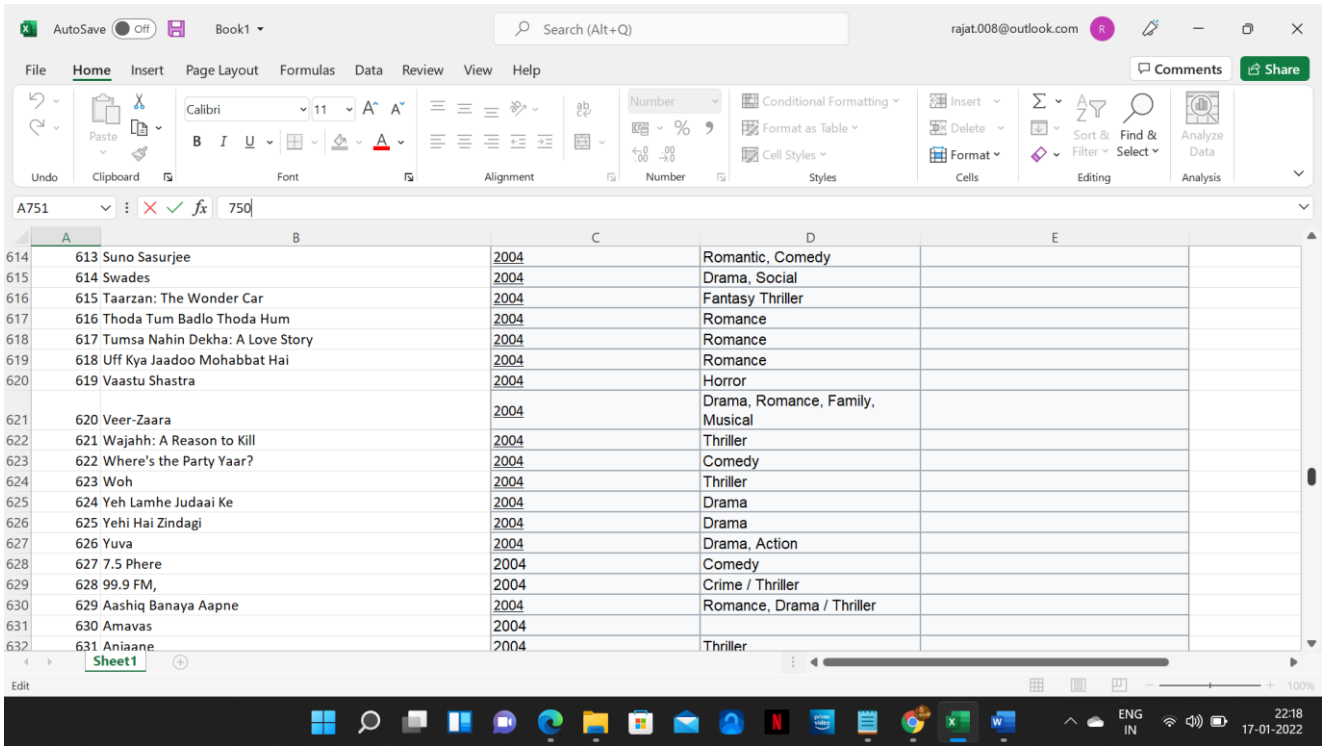
	A	B	C	D	E
286	285	Gadar: Ek Prem Katha	2001	Drama, War, Romance, Action	
287	286	Grahan	2001	Romance	
288	287	Hadh: Life on the Edge of Death	2001		
289	288	Hum Ho Gaye Aapke	2001	Romance	
290	289	Indian	2001	Action	
291	290	Ittefaq	2001	Action, Drama, Thriller	
292	291	Jodi No.1	2001	Comedy	
293	292	Kabhi Khushi Kabhie Gham...	2001	Drama, Romance, Family	
294	293	Kasam	2001	Action	
295	924	Kasoor	2001	Thriller	
296	295	Kuch Khatti Kuch Meethi	2001	Comedy, Drama	
297	296	Kyo Kii Main Jhuth Nahin Bolta	2001	Comedy, Romance	
298	297	Lagaan: Once Upon a Time in India	2001	Drama, Musical, Social	
299	298	Lajja	2001	Drama, Romance, Musical, Crime	
300	299	Love Ke Liye Kuchh Bhi Karega	2001	Comedy	
301	300	Maya	2001	Drama	
302	301	Mitti	2001		
303	302	Moksha	2001	Drama	
304	303	Mujhe Kucch Kehna Hai	2001	Comedy	

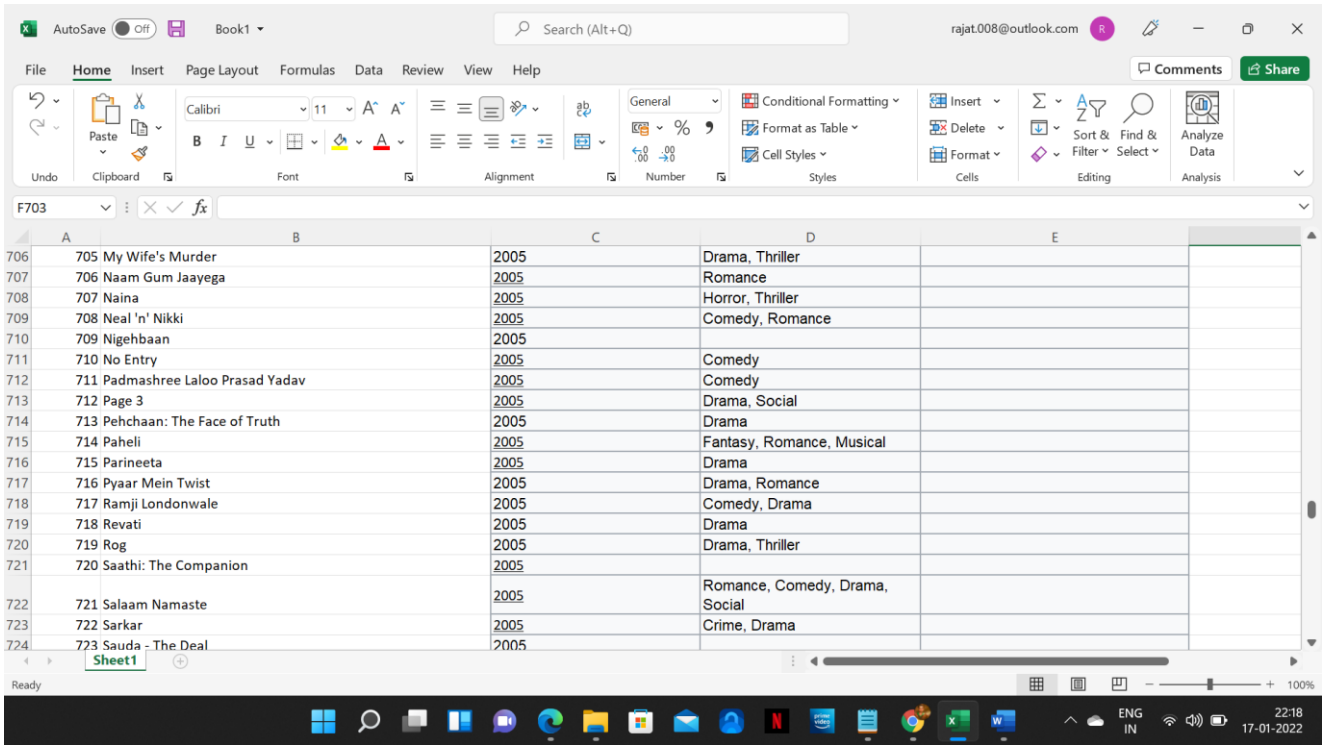


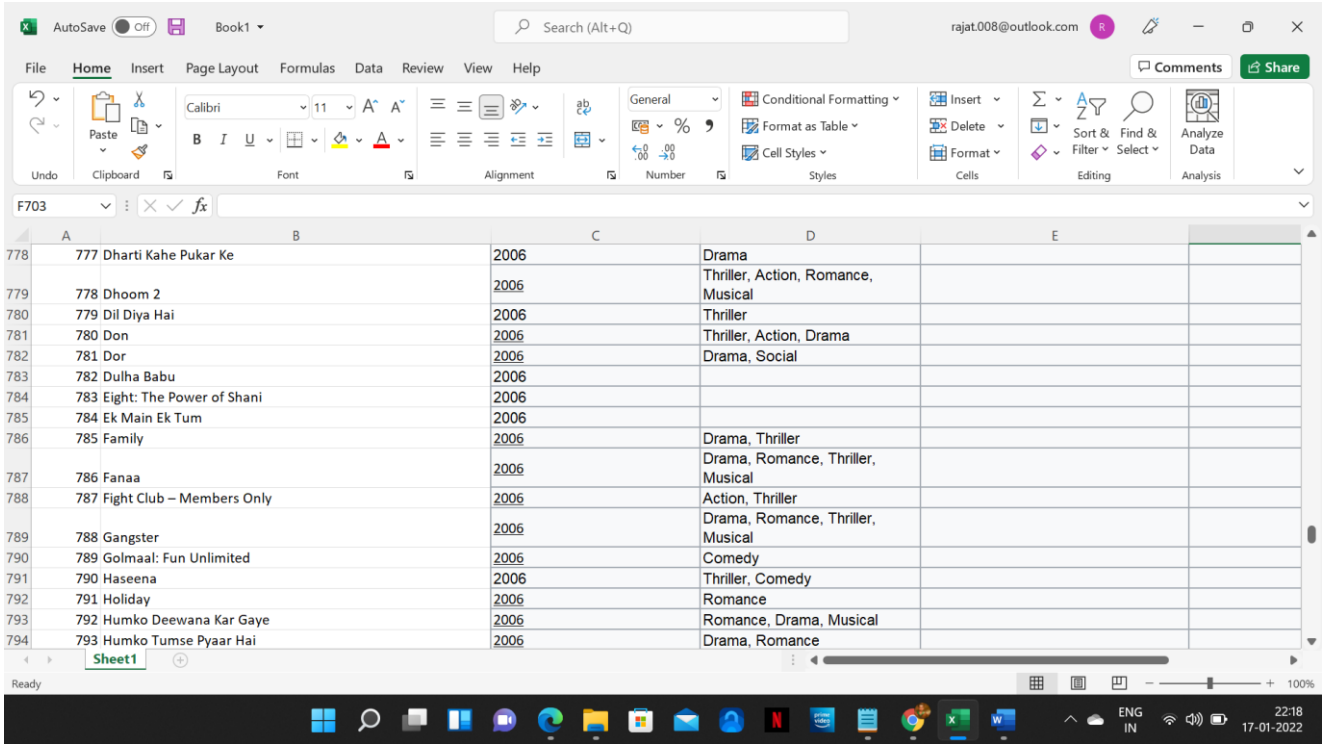
	A	B	C	D	E
348	347	Dil Vil Pyar Vyar	2002	Romance, Drama	
349	348	Ek Chhotisi Love Story	2002	Erotic	
350	349	Encounter: The Killing	2002	Social Drama	
351	350	Filhaal	2002	Drama	
352	351	Ghaav: The Wound	2002	Crime Drama	
353	352	Gunaah	2002	Thriller, Romance	
354	353	Haan Maine Bhi Pyaar Kiya	2002	Drama, Romance	
355	354	Hathyar	2002	Drama	
356	355	Hum Kisise Kum Nahin	2002	Action, Comedy	
357	356	Hum Pyar Tumhi Se Kar Baithe	2002	Romance	
358	357	Hum Tumhare Hain Sanam	2002	Drama, Romance	
359	358	Humraaz	2002	Thriller, Romance, Musical	
360	359	Jaani Dushman: Ek Anokhi Kahani	2002	Fantasy	
361	360	Jang Aur Aman	2002	Documentary	
362	361	Jeena Sirf Merre Liye	2002	Romance	
363	362	Kaante	2002	Action, Thriller	
364	363	Kabhie Tum Kabhie Hum	2002	Comedy, Drama	
365	364	Karz: The Burden of Truth	2002	Action, Thriller	
366	365	Kehtaa Hai Dil Baar Baar	2002	Romance	
367	366	Kitne Door Kitne Paas	2002	Romance, Drama, Comedy	



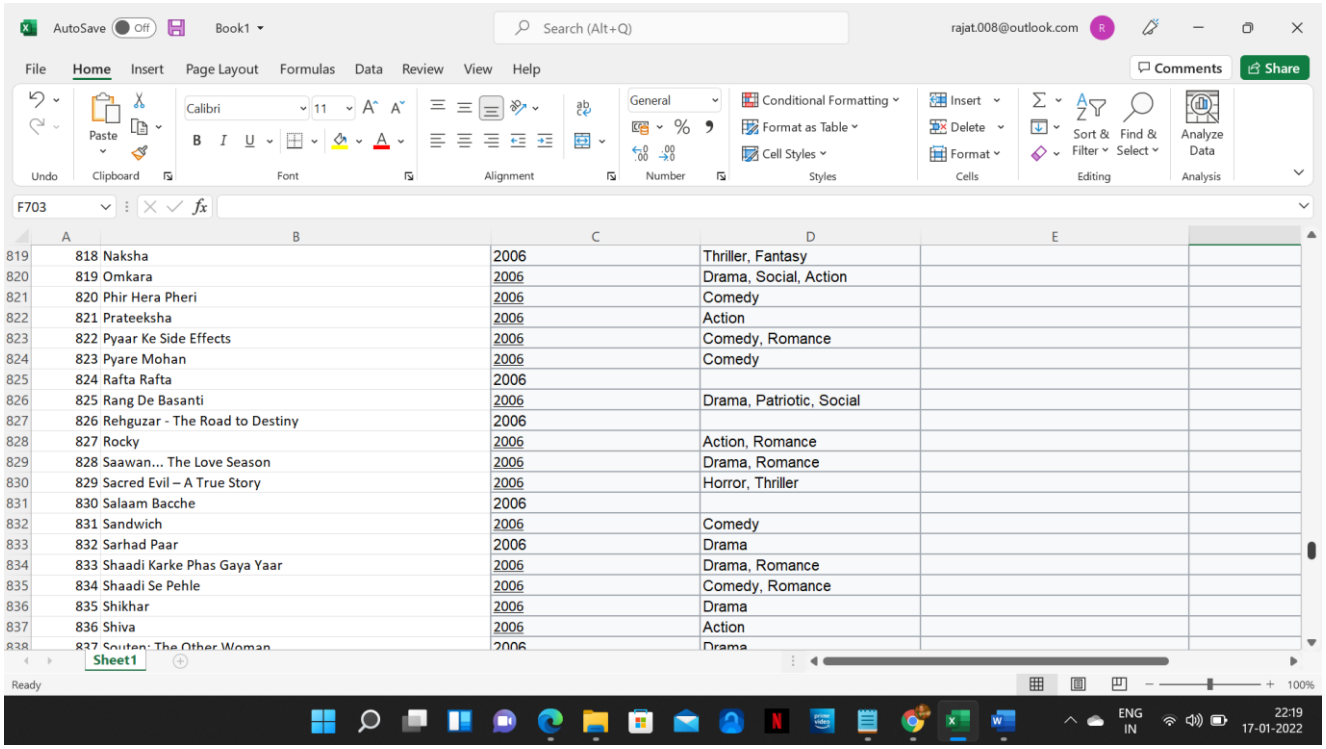
	A	B	C	D	E
472	471	Kyon?	2003		
473	472	Larger Than Life	2003		
474	473	LOC: Kargil	2003	War, Drama, Historical	
475	474	Love at Times Square	2003		
476	475	Maa Santoshi Maa	2003		
477	476	Main Madhuri Dixit Banna Chahti Hoon	2003	Comedy	
478	477	Main Prem Ki Diwani Hoon	2003	Romance, Drama, Comedy	
479	478	Market	2003	Crime Drama	
480	479	Miss India: The Mystery	2003		
481	480	Mudda: The Issue	2003	Drama	
482	481	Mumbai Matinee	2003	Romance Comedy	
483	482	Mumbai Se Aaya Mera Dost	2003	Drama	
484	483	Munna Bhai M.B.B.S.	2003	Comedy, Social, Musical	
485	484	Naag Lok	2003		
486	485	Nayee Padosan	2003	Comedy	
487	486	October	2003		
488	487	Om	2003	Action	
489	488	Oops!	2003		
490	489	Out of Control	2003	Comedy, Romance	
491	490	Paanch	2003		


AutoSave Off
Book1
Search (Alt+Q)
rajat.008@outlook.com
File Home Insert Page Layout Formulas Data Review View Help
Comments Share
Undo Clipboard Font Alignment Number Styles Cells Editing Analysis
A751
750
613 Suno Sasurjee 2004 Romantic, Comedy
614 Swades 2004 Drama, Social
615 Taarzan: The Wonder Car 2004 Fantasy Thriller
616 Thoda Tum Badlo Thoda Hum 2004 Romance
617 Tumsa Nahin Dekha: A Love Story 2004 Romance
618 Uff Kya Jaadoo Mohabbat Hai 2004 Romance
619 Vaastu Shastra 2004 Horror
620 Veer-Zaara 2004 Drama, Romance, Family, Musical
621 Wajahh: A Reason to Kill 2004 Thriller
622 Where's the Party Yaar? 2004 Comedy
623 Woh 2004 Thriller
624 Yeh Lamhe Judaai Ke 2004 Drama
625 Yehi Hai Zindagi 2004 Drama
626 Yuva 2004 Drama, Action
627 7.5 Phere 2004 Comedy
628 99.9 FM, 2004 Crime / Thriller
629 Aashiq Banaya Aapne 2004 Romance, Drama / Thriller
630 Amavas 2004
631 Anjaane 2004 Thriller
Sheet1
Edit
ENG IN 22:18 17-01-2022

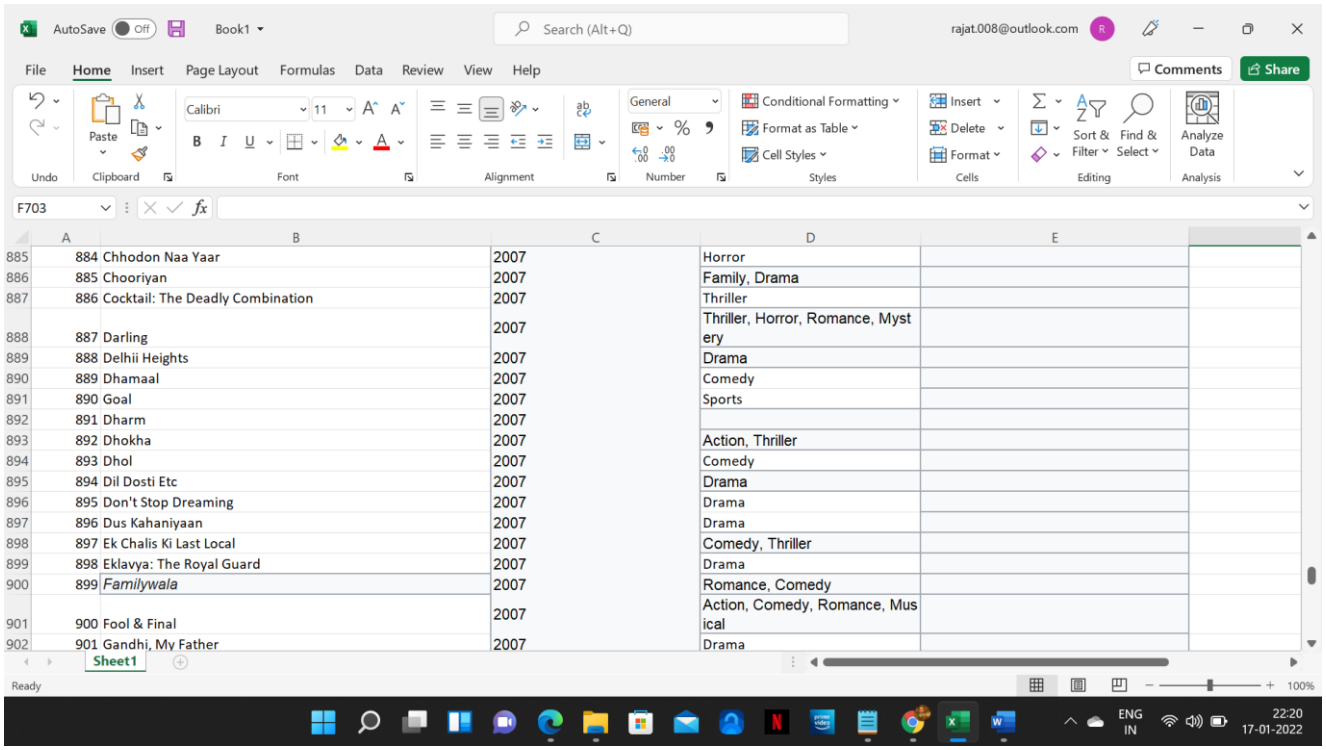

AutoSave Off
Book1
Search (Alt+Q)
rajat.008@outlook.com
File Home Insert Page Layout Formulas Data Review View Help
Comments Share
Undo Clipboard Font Alignment Number Styles Cells Editing Analysis
F703
705 My Wife's Murder 2005 Drama, Thriller
706 Naam Gum Jaayega 2005 Romance
707 Naina 2005 Horror, Thriller
708 Neal 'n' Nikki 2005 Comedy, Romance
709 Nigehbaan 2005
710 No Entry 2005 Comedy
711 Padmashree Laloo Prasad Yadav 2005 Comedy
712 Page 3 2005 Drama, Social
713 Pehchaan: The Face of Truth 2005 Drama
714 Paheli 2005 Fantasy, Romance, Musical
715 Parineeta 2005 Drama
716 Pyaar Mein Twist 2005 Drama, Romance
717 Ramji Londonwale 2005 Comedy, Drama
718 Revati 2005 Drama
719 Rog 2005 Drama, Thriller
720 Saathi: The Companion 2005
721 Salaam Namaste 2005 Romance, Comedy, Drama, Social
722 Sarkar 2005 Crime, Drama
723 Sauda - The Deal 2005
Sheet1
Ready
ENG IN 22:18 17-01-2022



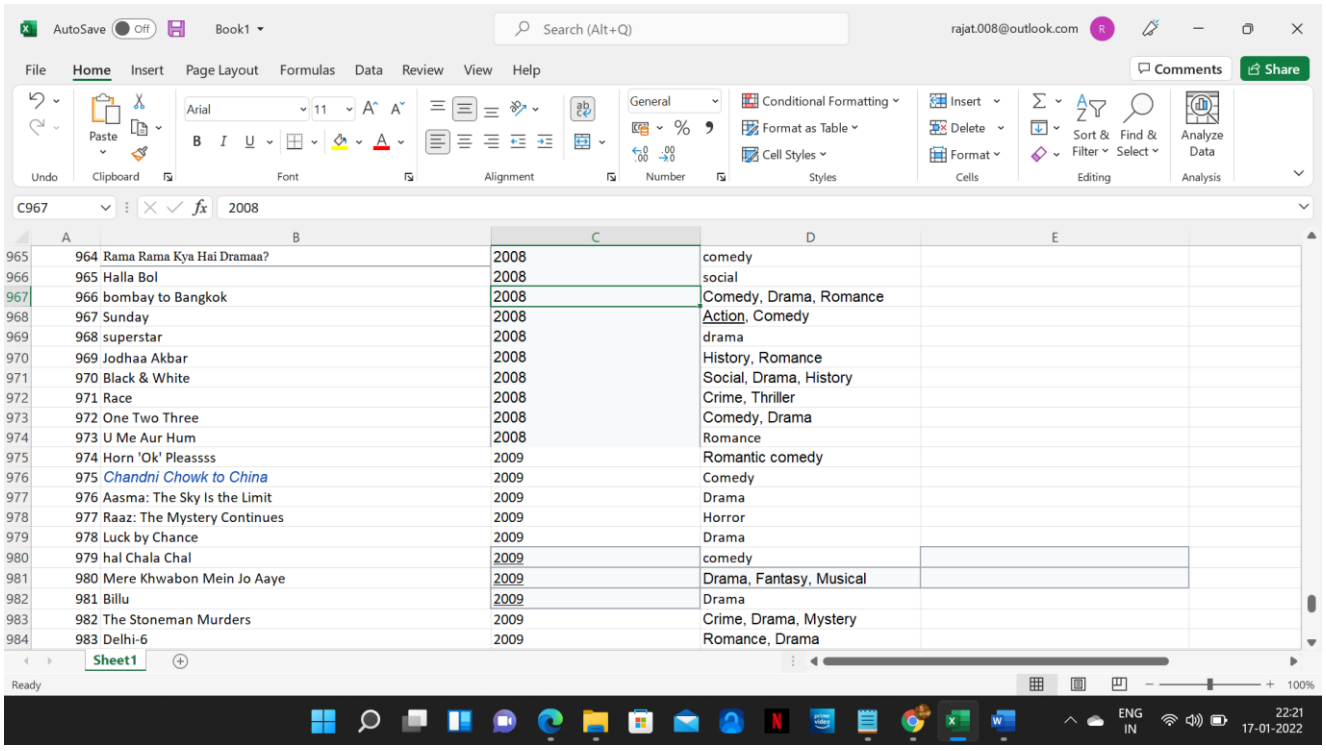
	A	B	C	D	E
778	777	Dharti Kahe Pukar Ke	2006	Drama	
779	778	Dhoom 2	2006	Thriller, Action, Romance, Musical	
780	779	Dil Diya Hai	2006	Thriller	
781	780	Don	2006	Thriller, Action, Drama	
782	781	Dor	2006	Drama, Social	
783	782	Dulha Babu	2006		
784	783	Eight: The Power of Shani	2006		
785	784	Ek Main Ek Tum	2006		
786	785	Family	2006	Drama, Thriller	
787	786	Fanaa	2006	Drama, Romance, Thriller, Musical	
788	787	Fight Club – Members Only	2006	Action, Thriller	
789	788	Gangster	2006	Drama, Romance, Thriller, Musical	
790	789	Golmaal: Fun Unlimited	2006	Comedy	
791	790	Haseena	2006	Thriller, Comedy	
792	791	Holiday	2006	Romance	
793	792	Humko Deewana Kar Gaye	2006	Romance, Drama, Musical	
794	793	Humko Tumse Pyaar Hai	2006	Drama, Romance	



	A	B	C	D	E
819	818	Naksha	2006	Thriller, Fantasy	
820	819	Omkara	2006	Drama, Social, Action	
821	820	Phir Hera Pheri	2006	Comedy	
822	821	Prateeksha	2006	Action	
823	822	Pyaar Ke Side Effects	2006	Comedy, Romance	
824	823	Pyare Mohan	2006	Comedy	
825	824	Rafta Rafta	2006		
826	825	Rang De Basanti	2006	Drama, Patriotic, Social	
827	826	Rehguzar - The Road to Destiny	2006		
828	827	Rocky	2006	Action, Romance	
829	828	Saawan... The Love Season	2006	Drama, Romance	
830	829	Sacred Evil – A True Story	2006	Horror, Thriller	
831	830	Salaam Bacche	2006		
832	831	Sandwich	2006	Comedy	
833	832	Sarhad Paar	2006	Drama	
834	833	Shaadi Karke Phas Gaya Yaar	2006	Drama, Romance	
835	834	Shaadi Se Pehle	2006	Comedy, Romance	
836	835	Shikhar	2006	Drama	
837	836	Shiva	2006	Action	
838	837	Souten: The Other Woman	2006	Drama	

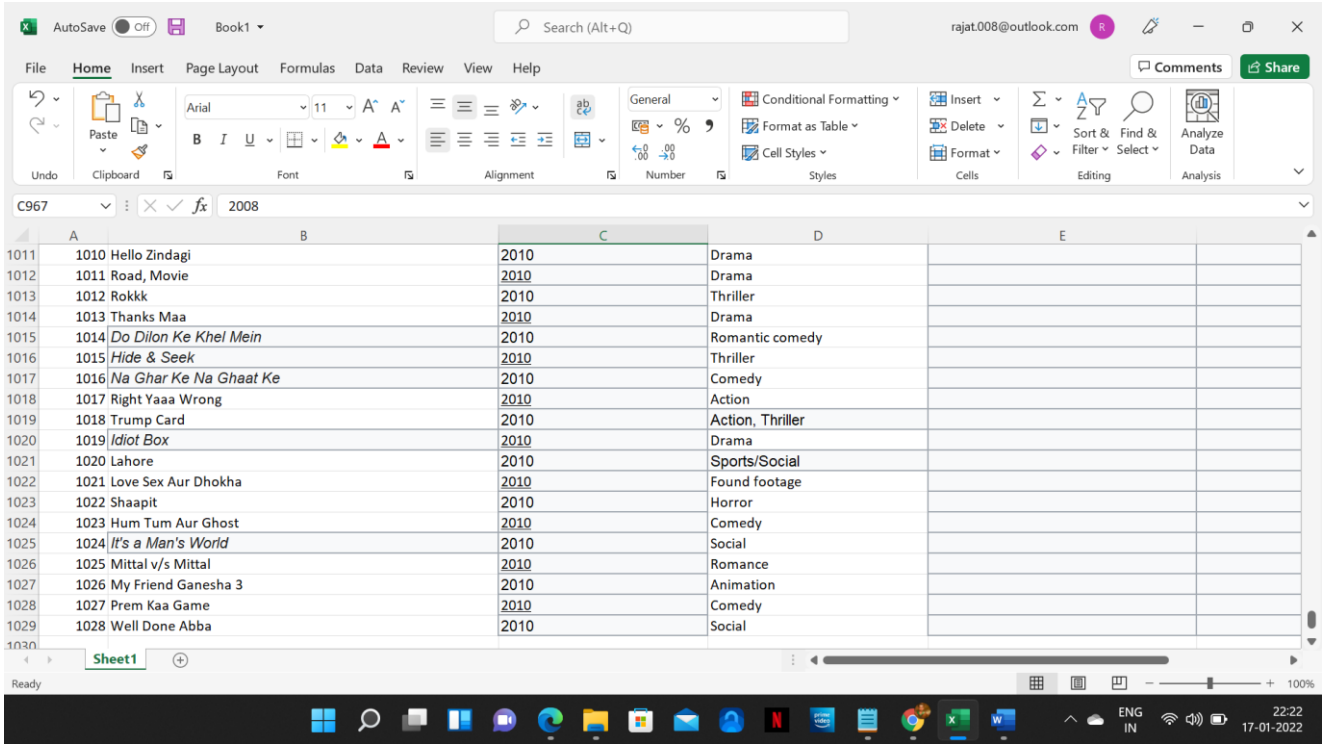


	A	B	C	D	E
885	884	Chhodon Naa Yaar	2007	Horror	
886	885	Chooriyan	2007	Family, Drama	
887	886	Cocktail: The Deadly Combination	2007	Thriller	
888	887	Darling	2007	Thriller, Horror, Romance, Mystery	
889	888	Delhii Heights	2007	Drama	
890	889	Dhamaal	2007	Comedy	
891	890	Goal	2007	Sports	
892	891	Dharm	2007		
893	892	Dhokha	2007	Action, Thriller	
894	893	Dhol	2007	Comedy	
895	894	Dil Dosti Etc	2007	Drama	
896	895	Don't Stop Dreaming	2007	Drama	
897	896	Dus Kahaniyaan	2007	Drama	
898	897	Ek Chalis Ki Last Local	2007	Comedy, Thriller	
899	898	Eklavya: The Royal Guard	2007	Drama	
900	899	*Familywala*	2007	Romance, Comedy	
901	900	Fool & Final	2007	Action, Comedy, Romance, Musical	
902	901	Gandhi, My Father	2007	Drama	



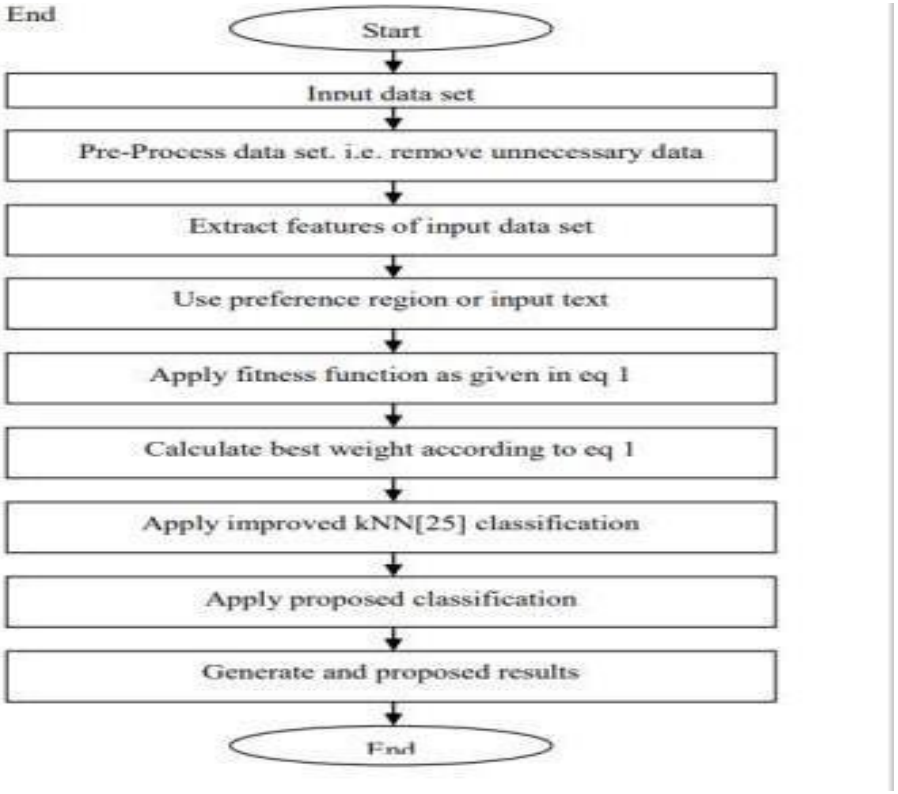
	A	B	C	D	E
965	964	Rama Rama Kya Hai Dramaa?	2008	comedy	
966	965	Halla Bol	2008	social	
967	966	bombay to Bangkok	2008	Comedy, Drama, Romance	
968	967	Sunday	2008	Action, Comedy	
969	968	superstar	2008	drama	
970	969	Jodhaa Akbar	2008	History, Romance	
971	970	Black & White	2008	Social, Drama, History	
972	971	Race	2008	Crime, Thriller	
973	972	One Two Three	2008	Comedy, Drama	
974	973	U Me Aur Hum	2008	Romance	
975	974	Horn 'Ok' Pleassss	2009	Romantic comedy	
976	975	*Chandni Chowk to China*	2009	Comedy	
977	976	Aasma: The Sky Is the Limit	2009	Drama	
978	977	Raaz: The Mystery Continues	2009	Horror	
979	978	Luck by Chance	2009	Drama	
980	979	hal Chala Chal	2009	comedy	
981	980	Mere Khwabon Mein Jo Aaye	2009	Drama, Fantasy, Musical	
982	981	Billu	2009	Drama	
983	982	The Stoneman Murders	2009	Crime, Drama, Mystery	
984	983	Delhi-6	2009	Romance, Drama	

li



	A	B	C	D	E
1011	1010	Hello Zindagi	2010	Drama	
1012	1011	Road, Movie	2010	Drama	
1013	1012	Rokkk	2010	Thriller	
1014	1013	Thanks Maa	2010	Drama	
1015	1014	*Do Dilon Ke Khel Mein*	2010	Romantic comedy	
1016	1015	*Hide & Seek*	2010	Thriller	
1017	1016	*Na Ghar Ke Na Ghaat Ke*	2010	Comedy	
1018	1017	Right Yaaa Wrong	2010	Action	
1019	1018	Trump Card	2010	Action, Thriller	
1020	1019	*Idiot Box*	2010	Drama	
1021	1020	Lahore	2010	Sports/Social	
1022	1021	Love Sex Aur Dhokha	2010	Found footage	
1023	1022	Shaapit	2010	Horror	
1024	1023	Hum Tum Aur Ghost	2010	Comedy	
1025	1024	*It's a Man's World*	2010	Social	
1026	1025	Mittal v/s Mittal	2010	Romance	
1027	1026	My Friend Ganesha 3	2010	Animation	
1028	1027	Prem Kaa Game	2010	Comedy	
1029	1028	Well Done Abba	2010	Social	
1030					

Flow Chart


End
Start
Input data set
Pre-Process data set. i.e. remove unnecessary data
Extract features of input data set
Use preference region or input text
Apply fitness function as given in eq 1
Calculate best weight according to eq 1
Apply improved kNN[25] classification
Apply proposed classification
Generate and proposed results
End

Chapter 5 - Form Design:

Install Putty

After installation of putty, enter the ip addrss of virtual machnine and then you can work on putty for

your virtual machine using theses commands:

cd

cd hadoop-2.7.3

start-dfs.sh - Starts the Hadoop DFS daemons, the namenode and datanodes. Use this before

start-mapred.shstop-dfs.sh - Stops the Hadoop DFS daemons.

start-mapred.sh - Starts the Hadoop Map/Reduce daemons, the jobtracker and tasktrackers.

stop-mapred.sh - Stops the Hadoop Map/Reduce daemons.

start-all.sh - Starts all Hadoop daemons, the namenode, datanodes, the jobtracker and tasktrackers.

Deprecated; use start-dfs.sh then start-mapred.sh

stop-all.sh - Stops all Hadoop daemons. Deprecated; use stop-mapred.sh then stop-dfs.sh

enter jsp

hive

select * from Dataset;(to print data of the table)

show tables;

drop table table_name;(to deleate the table)

To create a table in hive the command is:

CREATE EXTERNAL TABLE IF NOT EXISTS Dataset

(SNO INT,

MOVIENAME STRING,

YEAR INT,

GENRE STRING

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

LOCATION '/dataset';

Qeries to run:

Select GENRE from Dataset where Genre like’%Comedy’;

Select * from Dataset where YEAR between 1990 AND 1995;

Select * from Dataset where YEAR between 1995 AND 2000;

Select * from Dataset where YEAR between 2000 AND 2005;

Select * from Dataset where YEAR between 2005 AND 2010;

Select GENRE from Dataset where Genre like’%Romance’;

Select GENRE from Dataset where Genre like’%Action’;

Select GENRE from Dataset where Genre like’%Social’;

Select GENRE from Dataset where Genre like’%Thriller’;

Select GENRE from Dataset where Genre like’%Family’;

Select GENRE from Dataset where Genre like’%Horror’;

CREATE TABLE One(MOVIENAME STRING);

INSERT INTO TABLE One Select MOVIENAME from Dataset where YEAR between 1990 AND

1995;

select * from One;

CREATE TABLE One(MOVIENAME STRING);

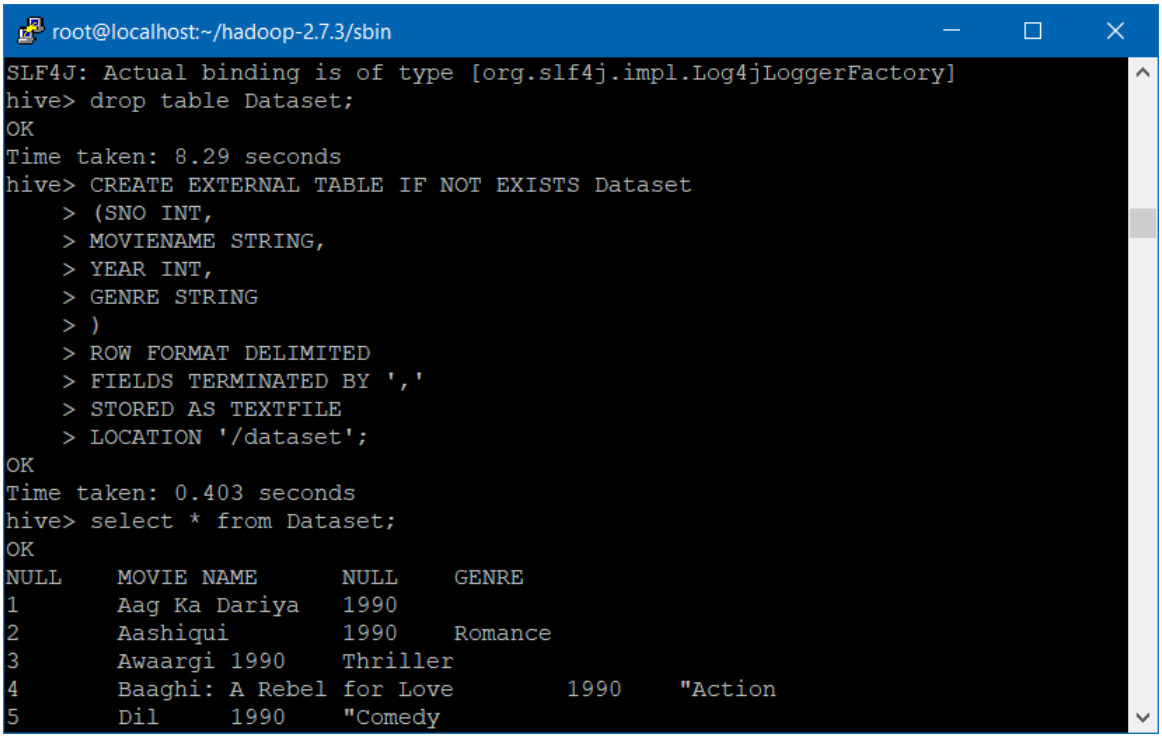
```
INSERT INTO TABLE Two Select MOVIENAME from Dataset where YEAR between 1995 AND

2000;

select * from Two;

CREATE TABLE Three(MOVIENAME STRING);

INSERT INTO TABLE Three Select MOVIENAME from Dataset where YEAR between 2000 AND

2005;

select * from Three;

CREATE TABLE Four(MOVIENAME STRING);

INSERT INTO TABLE Four Select MOVIENAME from Dataset where YEAR between 2005 AND

2010;

select * from Five;

CREATE TABLE Countcomedy(MOVIENAME STRING);

INSERT INTO TABLE Countcomedy Select COUNT(MOVIENAME) from Dataset where GENRE

like '%Comedy';

select * from Countcomedy;

CREATE TABLE Countaction(MOVIENAME STRING);

INSERT INTO TABLE Countaction Select COUNT(MOVIENAME) from Dataset where GENRE

like '%Action';

select * from Countaction;

CREATE TABLE Countfamily(MOVIENAME STRING);

INSERT INTO TABLE Countfamily Select COUNT(MOVIENAME) from Dataset where GENRE

like '%Family';
```

```
select * from Countfamily;

CREATE TABLE Countsocial(MOVIENAME STRING);

INSERT INTO TABLE Countsocial Select COUNT(MOVIENAME) from Dataset where GENRE

like '%Social';

select * from Countsocial;

CREATE TABLE Countromance(MOVIENAME STRING);

INSERT INTO TABLE Countromance Select COUNT(MOVIENAME) from Dataset where
GENRE

like '%Romance';

select * from CountRomance;

CREATE TABLE Counthorror(MOVIENAME STRING);

INSERT INTO TABLE Counthorror Select COUNT(MOVIENAME) from Dataset where
GENRE

like '%Horror';

select * from Counthorror;

CREATE TABLE Countthriller(MOVIENAME STRING);

INSERT INTO TABLE Countthriller Select COUNT(MOVIENAME) from Dataset where
GENRE

like '%Thriller';

select * from Countthriller
```

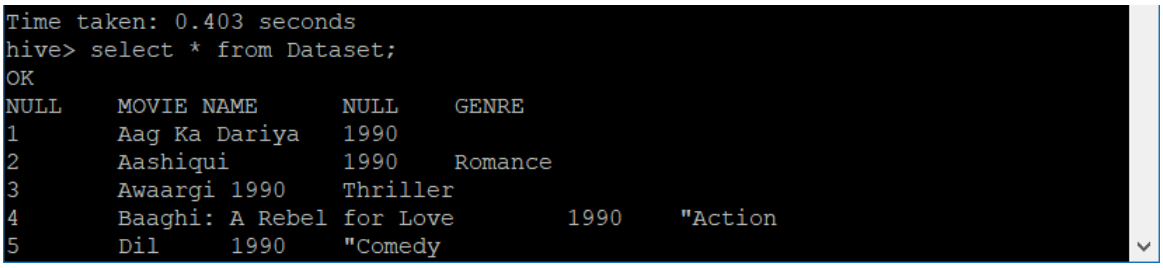
Input / Output Form (Screenshot)

- **Input Screenshot**

Table creation

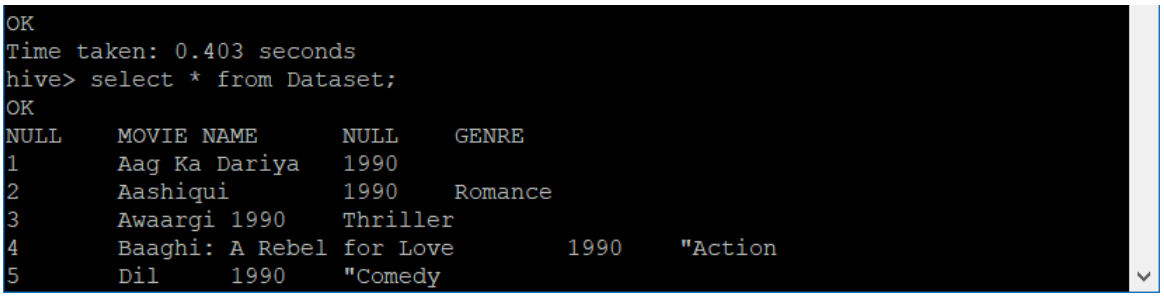


Query for fetching the data:



Output Screenshot

Output of the queries

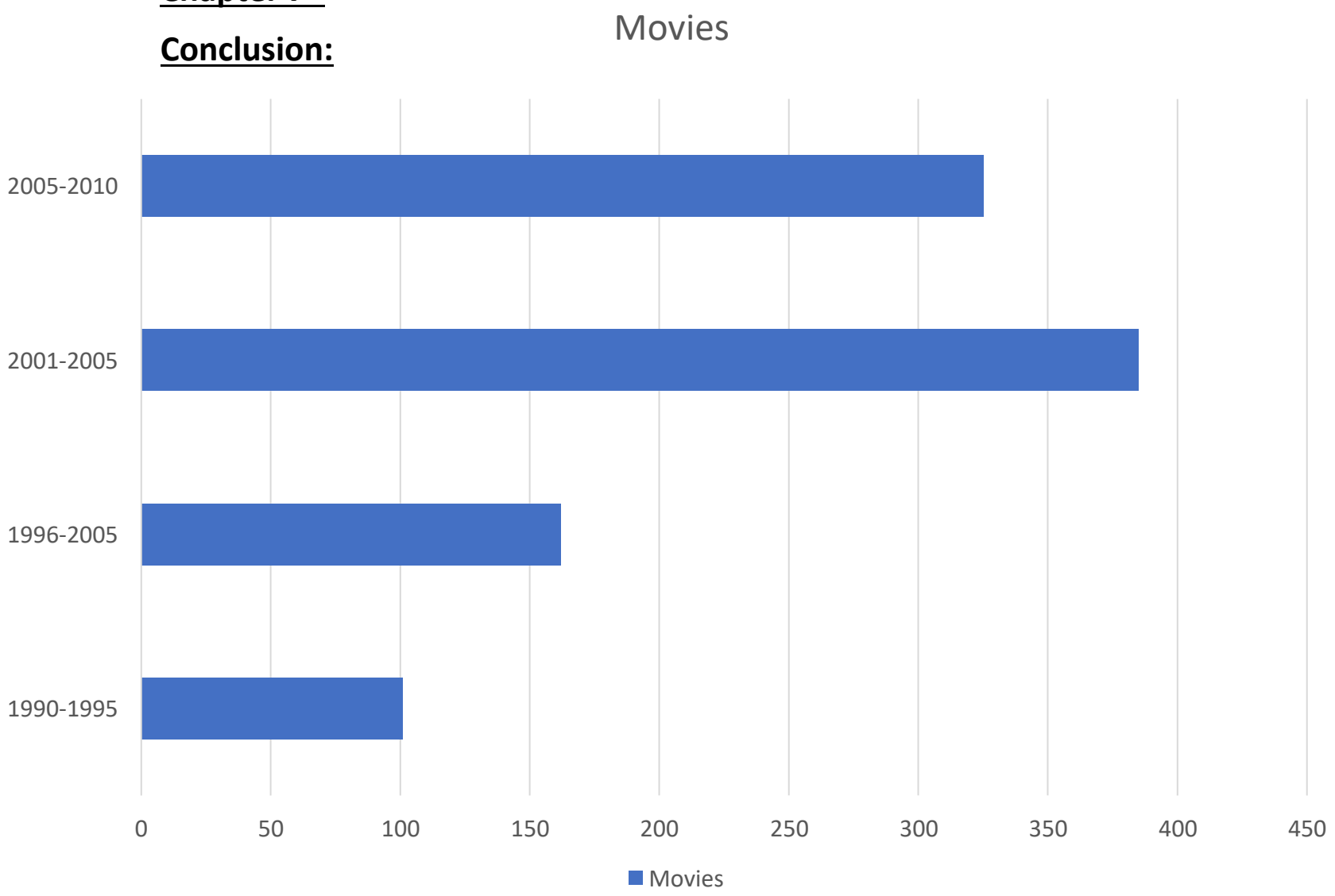


Chapter 6 – Testing

- Check whether the machine working properly on oracle virtual box
- Check is there any issue in the implementation of virtual machine
- Check the entering valid credentials like Password of local host should be correct
- Check the dataset we are using should contain more than 500 entries (Minimum)

Chapter 7 -

Conclusion:



Chapter 8 – BIBLIOGRAPHY:

I have done this project with the help of my supervisor Mrs. Vidushi Mishra & alumni mentor & taking references from the following:

www.edurekha.com

www.javatpoint.com

I used:

- Oracle
- Virtual box
- CentOS-7
- Putty-64bit-0.7installer
- WinSCP
- Internet Explorer
- Chrome