

## **UNIT 4: BIG DATA ESSENTIALS**

### **4.0 OVERVIEW**

In this Unit, you will get an in-depth knowledge about Big Data landscape. You will become conversant with the terminology and core concepts behind Big Data such as evolution of Big Data, the characteristics of Big Data and different challenges that have cropped up the era of Big Data and the growing volume of data. The unit also covers the different application domains where Big Data can be applied such as healthcare, banking and finance, retail, hospitality and transportation, government and security. Furthermore, the unit will include steps to set up the Cloudera Big Data platform to get acquainted with the Big Data environment.

### **4.1 LEARNING OUTCOMES**

Upon completion of this unit, you will be able to:

- Explain Big data concepts.
- Learn the characteristics of Big Data.
- Recognise the challenges of Big Data.
- Be acquainted with the application domains for Big Data.
- Know how to set up a Big Data environment.

## 4.2 BIG DATA OVERVIEW

This section provides the definitions for big data and shows how Big Data has evolved over the past years.

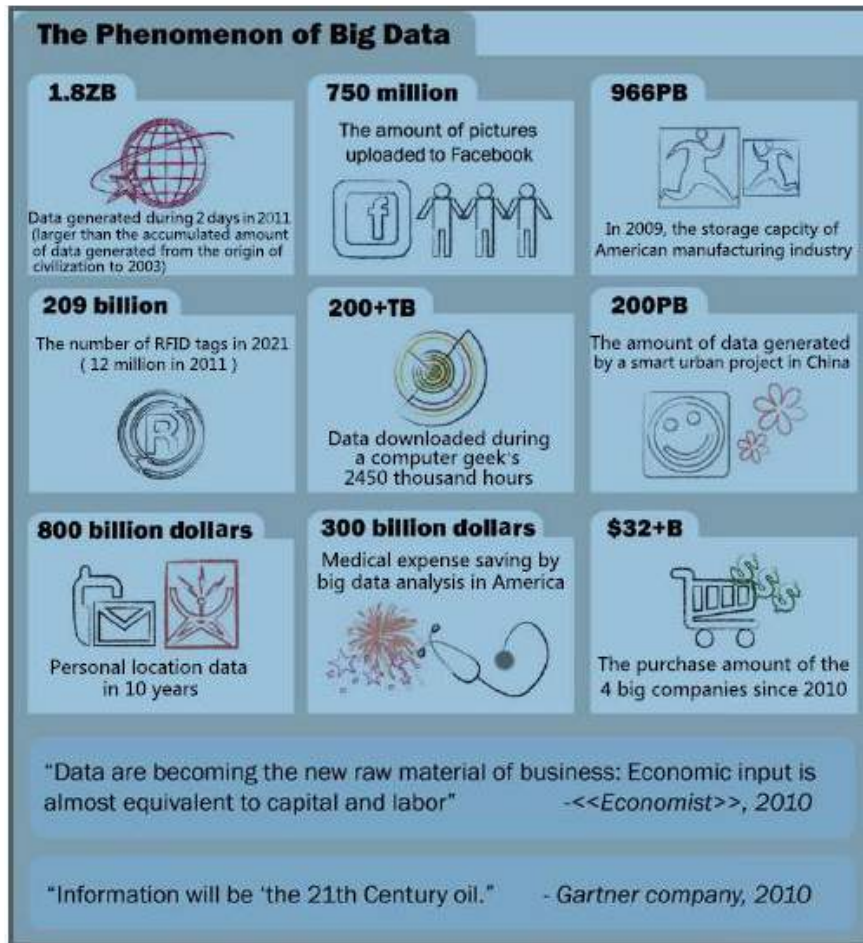
### Definitions

Big data is a term used for describing large data sets that were beyond storage capabilities. Sagioglu and Sinanc (2013) defines Big Data as a term “for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results”. Manyika *et al.* (2011) defines Big Data as “the volume of data whose size is beyond the ability of typical databases to store, manage, and analyze”. Internet Data Center (IDC) defines Big Data as “a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and analysis” (Vesset *et al.*, 2012).

### Evolution of Big Data

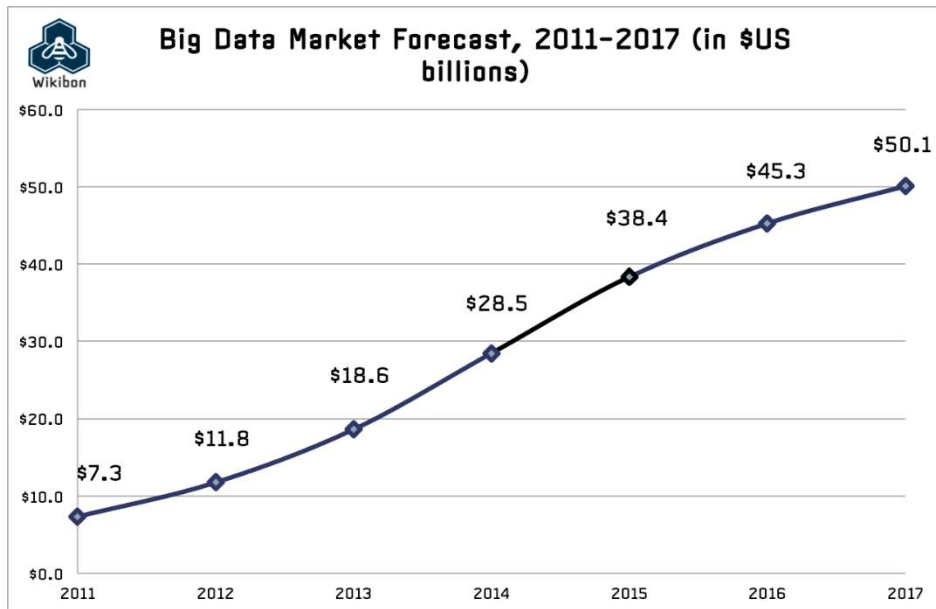
The explosion of the Internet, social media, technologies such as mobile devices, sensors and applications have led to the creation of massive data sets. According to McAfee *et al.* (2012), as of 2012, about 2.5 exabytes of data were created each day and that number is doubling every 40 months and so. As of 2014, Google processes data of hundreds of Petabyte (PB) and Facebook generates log data of over 10 PB per month (Chen *et al.*, 2014).

**Figure 4.1** shows how the phenomenon of Big Data has changed things for the past years. The Big Data market is measured “by vendor revenue derived from sales of related hardware, software and services” (Wikibon, 2014). As it can be seen from **Figure 4.2**, it was forecasted in 2014 by Wikibon that Big Data market will reach \$50.1 billion in 2017 but the actual figure is \$57 billion (PR Newswire, 2017), showing the rapid growth of Big Data technologies.



**Figure 4.1: The Phenomenon of Big Data**

(Source: Chen *et al.*, 2014).



**Figure 4.2: Big Data Market Forecast**

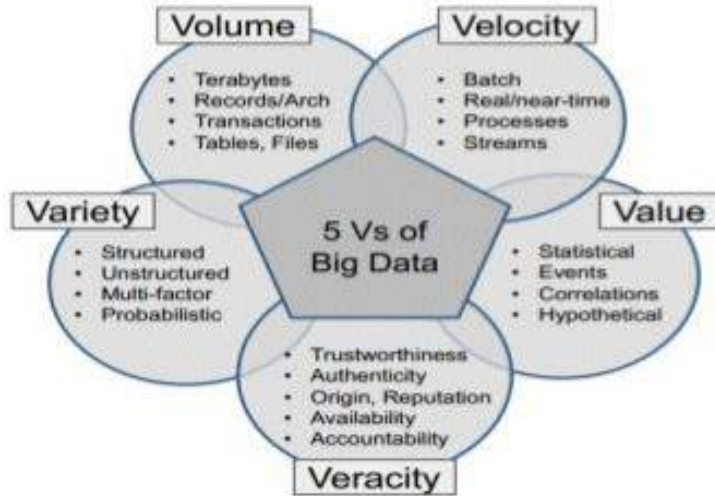
(Source: <http://wikibon.org>).

**Activity 1**

1. Identify sources of data that have contributed to the phenomenon of Big Data.

**4.3 CHARACTERISTICS OF BIG DATA**

Big Data is characterised by the multi-V model as shown in **Figure 4.3**. These Vs are further explained.



**Figure 4.3: 5 V's of Big Data**

(Source: Anuradha, 2015).

- **Volume:** The volume of data refers to large amount of data with size varying from terabytes to zettabyte. Analysing and manipulating such a large amount of data require substantial resources and represent a major challenge (Tole, 2013).
- **Velocity:** The velocity refers to the speed at which data is created (McAcfee *et al.*, 2012). It can be measured using data volume per time.
- **Variety:** Variety refers to different types of data: structured, semi-structured and unstructured data that are being stored and analysed. Semi- structured data consist of a combination of structured and unstructured data. The types of data can include text, audio, video, images, sensor data, emails, log files, social media posts amongst others (Kune *et al.*, 2016). **Figure 4.4** provides some examples to differentiate between traditional data and Big Data.

<i>Traditional Data</i>	<i>Big Data</i>
Documents	Photos
Finances	Audio and Video
Stock Records	3D Models
Personnel files	Simulations
	Location data

**Figure 4.4: Difference between traditional data and Big Data**

(Source: Tole, 2013).

- **Veracity:** Veracity refers to trustworthiness of the data. It includes other data quality attributes such as authenticity, reputation, availability, consistency and accountability of data (Tole, 2013; Anuradha, 2015; Kune *et al.*, 2016).
- **Value:** Raw data is of no value. Big data has to be transformed into smart data to add value to a business or even generate revenue. Additionally, it has to be properly managed to offer better insights (Tole, 2013).

### **Activity 2**

1. Differentiate between structured and unstructured data, by providing examples of both types.
2. How is Big Data different from traditional data?

## **4.4 CHALLENGES**

With the era of big data and the growing volume of data, a number of challenges have cropped up. The traditional RDBMS cannot handle the large volume of data and cope with heterogeneity of big data (McAfee *et al.*, 2012; Chen and Zhang, 2014). This section highlights a number of challenges that have to be tackled to enable development of big data applications:

- **Data Representation**

The evolution of Big Data has led to creation of large amount of heterogeneous data with variations in type, structure, semantics, organisation, granularity and accessibility. Thus representing data so that it is meaningful and efficient, is a major challenge (Chen *et al.*, 2014).

- **Data Analysis**

Analysing Big Data is a challenging task due to the incompleteness and inconsistencies of semi-structured and unstructured data (Ammu and Irfanuddin, 2013; Chen and Zhang, 2014). Additionally, according to Ammu and Irfanuddin (2013), data volume is scaling faster than compute resources. Moreover, the analysis of large data sets is very time consuming (Ammu

and Irfanuddin, 2013; Chen and Zhang, 2014). It is of utmost importance to address these challenges to realize the full potential of Big Data analysis. Furthermore, in order to get much benefit and insights from Big Data analysis, Big Data has to be pre-processed, cleaned and transformed properly.

- Data Acquisition

Data acquisition consists of data collection, data transmission and data pre-processing (Chen et al., 2014). Big data collected from various sources such as log files and sensors often consist of large amount of redundant data. It is therefore a major challenge to remove this high redundancy. According to Chen *et al.* (2014), appropriate compression algorithms have to be applied.

- Data Storage

Traditional RDBMS are found to be ill-suited for storing and processing Big data (McAfee *et al.*, 2012; Chen *et al.*, 2014). NoSQL databases, also referred to as non-traditional databases are becoming increasingly popular for Big Data storage. Some examples of Big Data databases include Dynamo, Voldemort, BigTable, Cassandra, MongoDB, SimpleDB and CouchDB (Chen *et al.*, 2014). It is a major challenge to offer information storage service with reliable storage space as well as powerful access interface for query and analysis of a large amount of data (Chen *et al.*, 2014).

- Data Management

According to Kaisler et al. (2013), managing Big Data is the most difficult problem. A number of issues still have to be resolved such as “access, metadata, utilization, updating, governance, and reference (in publications)”. The authors also emphasise on the need for new approaches to qualify and validate data as they find it impractical to perform validation on every data item in large datasets.

### **Activity 3**

1. Describe some more challenges that have cropped up with the evolution of Big Data.
2. There are three types of Big Data databases namely key-value databases, column-oriented databases, and document-oriented databases.
  - (i) Differentiate between these three types of databases.
  - (ii) Categorise the existing Big Data databases into these three groups.

## **4.5 APPLICATION DOMAINS**

Big Data can be applied in various domains such as healthcare, banking and finance, retail, hospitality and transportation, government and security. Some of these examples are described below:

### **▪ Healthcare**

Big Data is currently being used in healthcare for the prediction and surveillance of diseases (Kune *et al.*, 2016). Analysing disease patterns can prevent the spreading of the disease. Furthermore, analysing large data sets of patients' information can help identification of patients who are likely to suffer from a particular disease such as diabetes (Raghupathi and Raghupathi, 2014 ; Joaheer and Nagowah, 2017).

### **▪ Banking and Finance**

Financial institutions are often faced with difficulties to retain their customers. By utilizing Big Data and applying sentiment analysis and predictive analysis techniques, it is possible to predict who are the potential customers and additionally offer targeted products to the customers (Kune *et al.*, 2016).

### **▪ Retail and Manufacturing**

According to Sagiroglu and Sinanc (2013), analysing Big Data can lead to various benefits in the retail and manufacturing sector such as store behavior analysis, variety and price optimization,



product placement design, labor inputs optimization, distribution and logistics optimization and demand forecasting amongst others.

- **Government**

Big Data has huge potential of transforming governments by fostering collaboration and providing real-time solutions to enhance decision making (Bertot and Choi, 2013). Analysing Big Data in the government sector can lead to safer communities, smarter decisions, better served citizens, improved fiscal performance greater innovation and citizen engagement amongst others (Rajagopalan and Vellaipandiyam, 2013). Furthermore, to ensure security of general public, relevant departments of the government can analyse images from aerial cameras, news feeds, and social networks or any other items of interest (Kune *et al.*, 2016).

#### **Activity 4**

1. Suggest other application domains where Big Data could be applied.
2. Investigate and describe other applications of Big Data in the healthcare sector.
3. Discuss the trends of Big Data in shaping the financial sector.

## **4.6 BIG DATA TOOLS**

This section gives an overview of different Big Data tools currently being used. Chen and Zhang (2014) gives an overview of Big Data Tools based on batch processing namely Apache Hadoop, Dryad, Apache Mahout, Jaspersoft BI Suite, Pentaho Business Analytics, Skytree Server, Tableau, Karmasphere Studio and Analyst and Talend Open Studio. The use of the different tools is shown in Figure 4.5 and the advantages are presented in **Figure 4.6**.

Name	Specified Use
Apache Hadoop	Infrastructure and platform
Dryad	Infrastructure and platform
Apache Mahout	Machine learning algorithms in business
Jaspersoft BI Suite	Business intelligence software
Pentaho Business Analytics	Business analytics platform
Skytree Server	Machine learning and advanced analytics
Tableau	Data visualization, Business analytics,
Karmasphere Studio and Analyst	Big Data Workspace
Talend Open Studio	Data management and application integration

**Figure 4.5: Big Data Tools based on batch processing and their uses**

(Source: Chen and Zhang, 2014).

Name	Advantage
Apache Hadoop	High scalability, reliability, completeness
Dryad	High performance distributed execution engine, good programmability
Apache Mahout	Good maturity
Jaspersoft BI Suite	Cost-effective, self-service BI at scale
Pentaho Business Analytics	Robustness, scalability, flexibility in knowledge discovery
Skytree Server	Process massive datasets accurately at high speeds
Tableau	Faster, smart, fit, beautiful and ease of use dashboards
Karmasphere Studio and Analyst	Collaborative and standards-based unconstrained analytics and self service
Talend Open Studio	Easy-to-use, eclipse-based graphical environment

**Figure 4.6: Big Data Tools based on batch processing and their advantages**

(Source: Chen and Zhang, 2014).

There are additional tools and platform that distribute open-source Hadoop platforms namely AWS, Cloudera, Hortonworks, and MapR Technologies (Raghupathi and Raghupathi, 2014). Proprietary options such as IBM's BigInsights are also available.

## 4.7 SETTING UP THE ENVIRONMENT FOR BIG DATA

In this section, installation details of the basic platforms and setup required to use R on Big Data platforms are given. The three main setups include:

1. The Virtual Machine
2. Hadoop Environment
3. R on Hadoop

### Prerequisites

The minimum requirements to run the following setups include minimum 16 GB Ram.

The Software Packages used in this tutorial are as follows:

- Guest OS: Windows 10 Pro with 16 GB RAM
- Virtualization software: Oracle VirtualBox 5.2.12
- Cloudera Hadoop VM: CDH 5.13
- R: 3.4.4
- RStudio server: 1.2.637

### 4.7.1 Installing the Virtual Machine

There are different virtualization products that exist on the market. The following steps provide installation details for the **Oracle VirtualBox**.

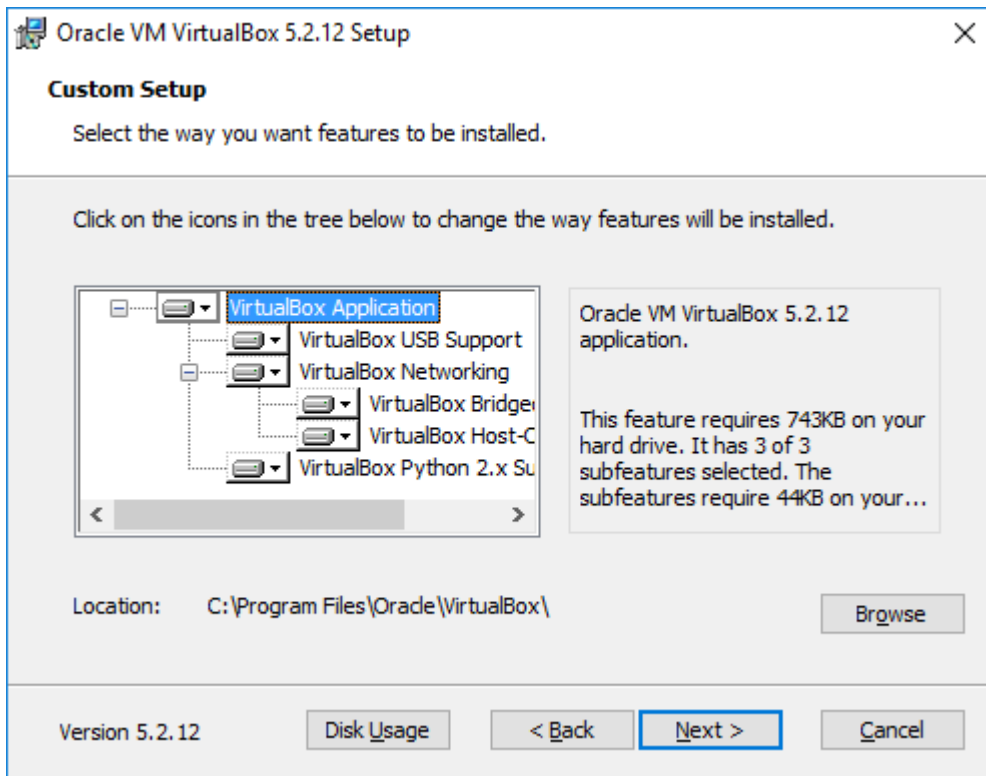
1. Download the software from the following URL:

<https://download.virtualbox.org/virtualbox/5.2.12/VirtualBox-5.2.12-122591-Win.exe>

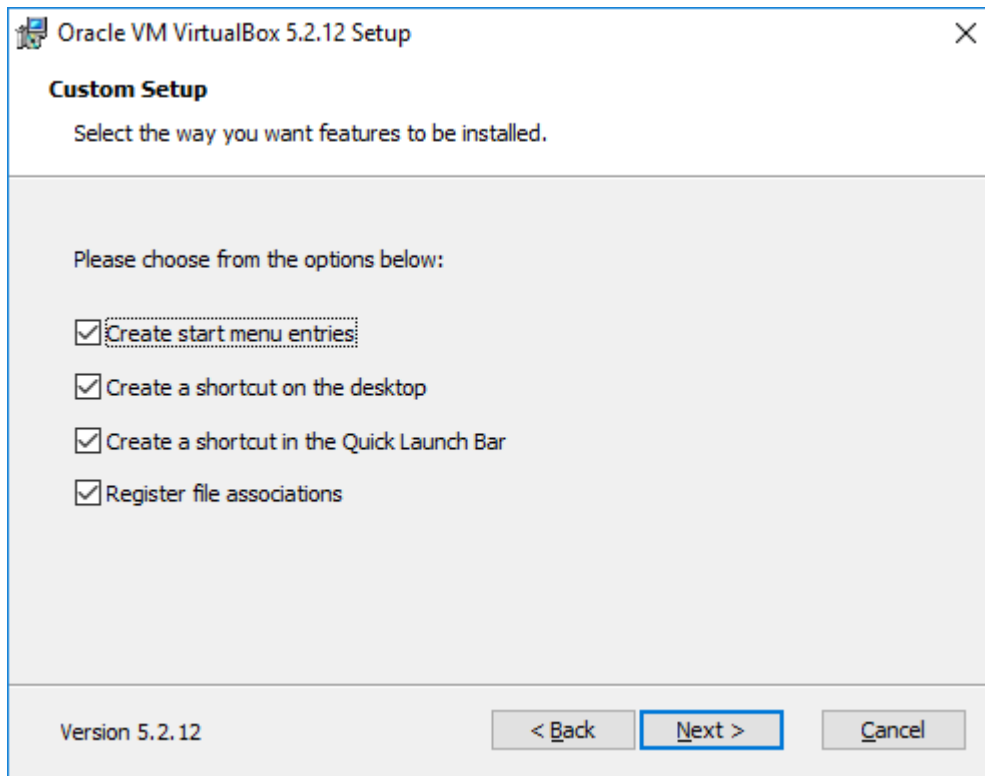
2. Install the Oracle VM VirtualBox and follow the steps outline below.



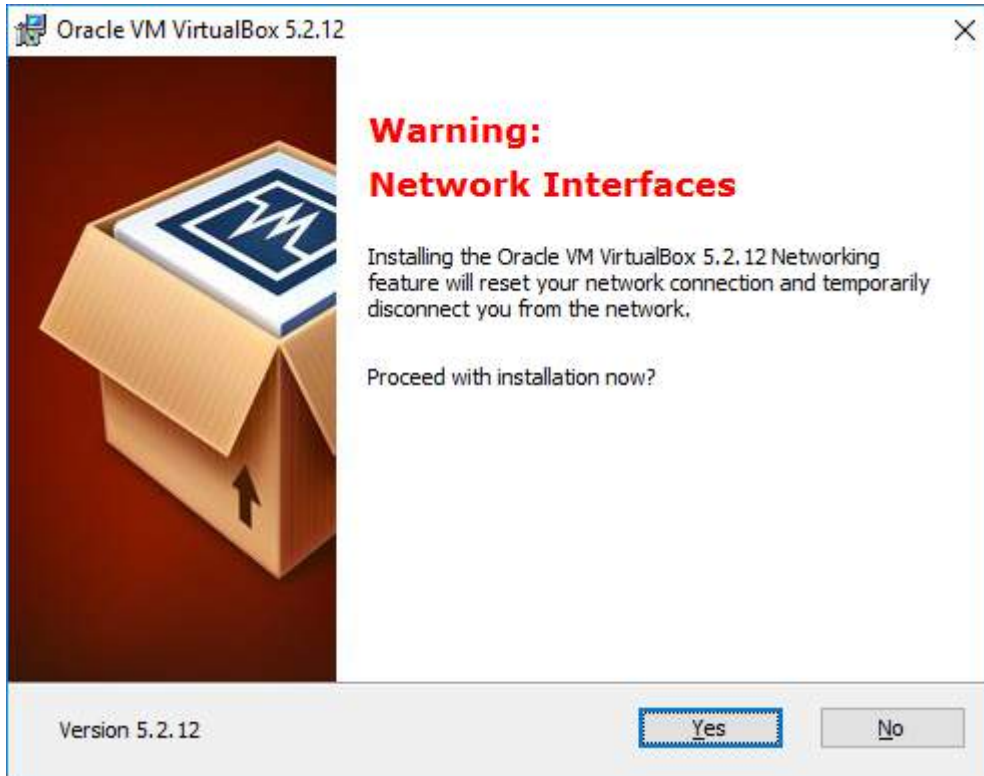
3. Click Next to start the setup wizard.



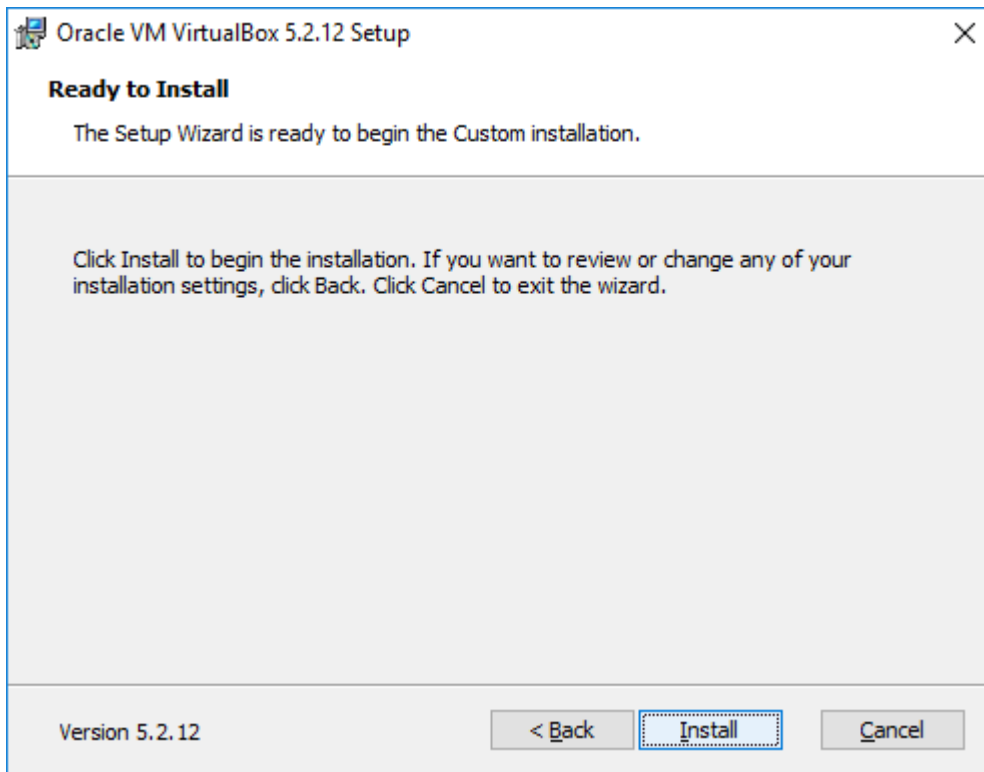
4. Click Next to accept the default settings.



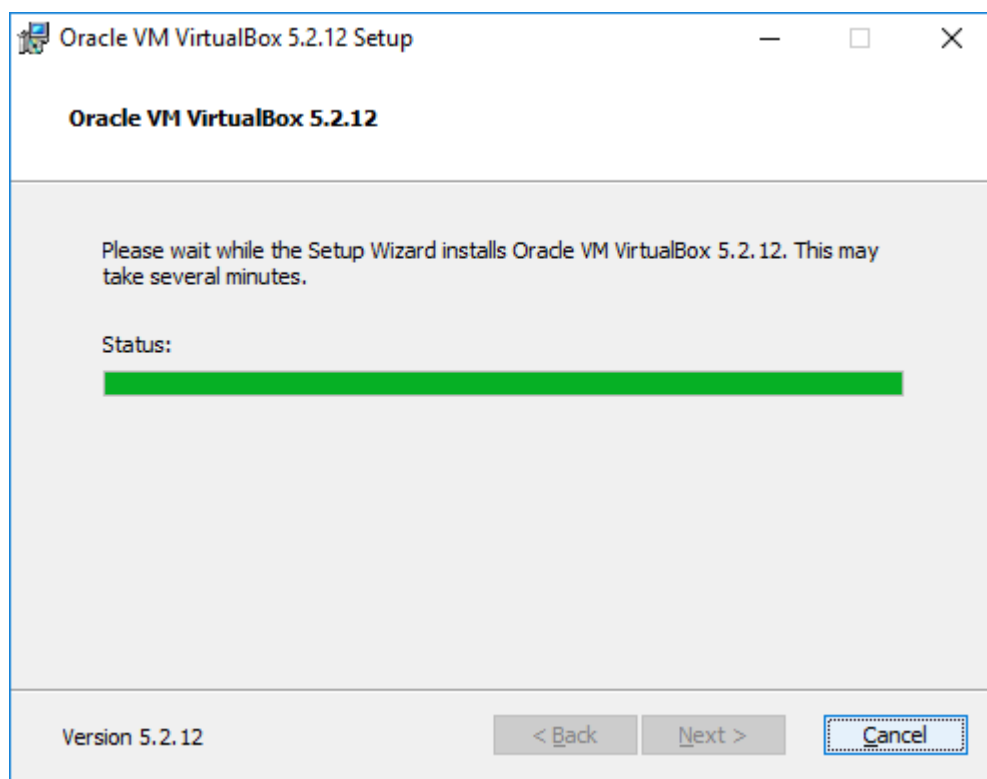
5. Click Next.



6. Note that you will be temporarily disconnected from your network. Click Yes.



7. Click Install to start the installation of the Oracle VM VirtualBox which might take several minutes.



8. Once the Oracle VM VirtualBox has been installed, the following screen will appear.



9. You will now proceed to the next section to download the Cloudera's Hadoop environment and will come back to the Oracle VM VirtualBox to create a new virtual machine.

#### 4.7.2 Installing the Cloudera Hadoop Environment

This section provide installation details pertaining to Cloudera (<https://www.cloudera.com/>).

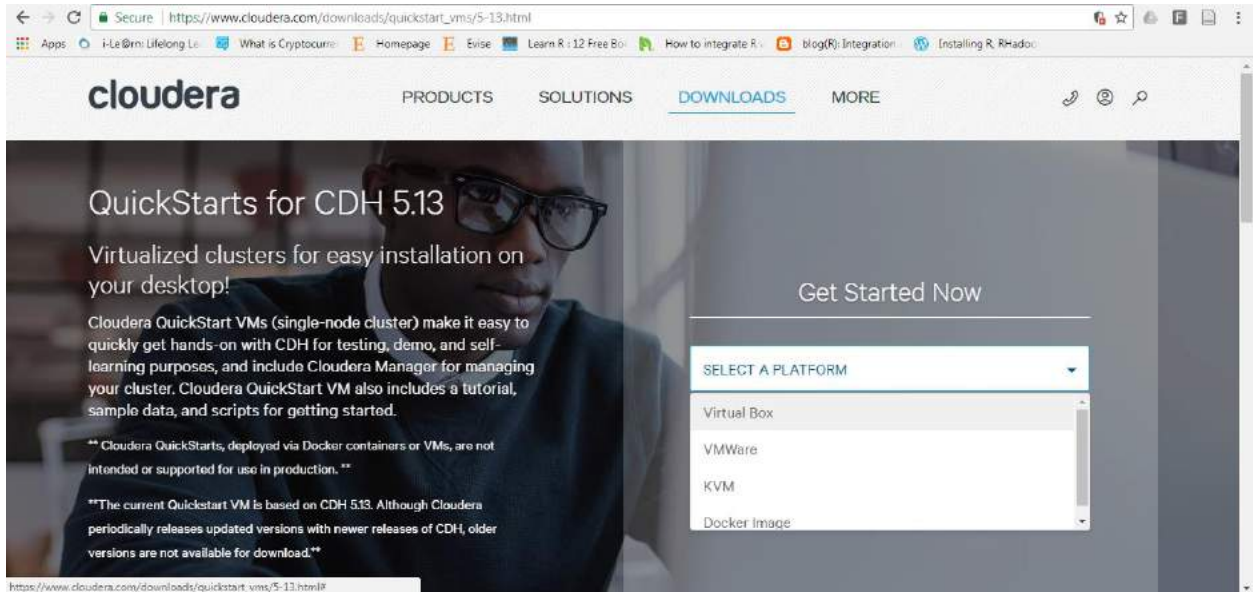
Cloudera is an open source platform and is the most popular distribution of Hadoop.


The following steps consist of installation details of the **Cloudera Platform on Oracle VM VirtualBox (Ver 5.2.12)**:

1. Download the latest version of 'Cloudera's Hadoop Quickstart VM' for your Guest OS. CDH 5.13 which runs CentOS 6.7 has been used in this tutorial and has been obtained from the following link:


[https://www.cloudera.com/downloads/quickstart\\_vms/5-13.html](https://www.cloudera.com/downloads/quickstart_vms/5-13.html)



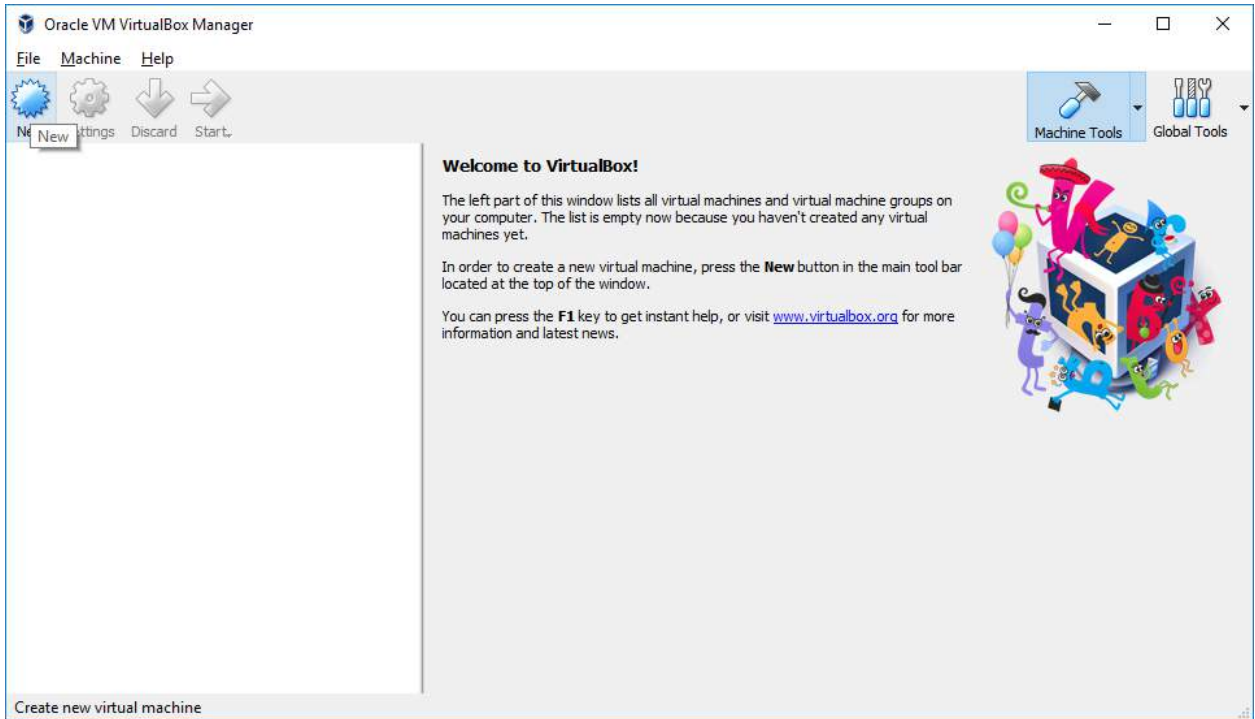


 <b>Note it!</b>	<ul style="list-style-type: none"> <li>• <i>The CDH should be downloaded for the correct platform. In this tutorial, the <b>Virtual Box</b> virtualization product has been used.</i></li> </ul>
--	--

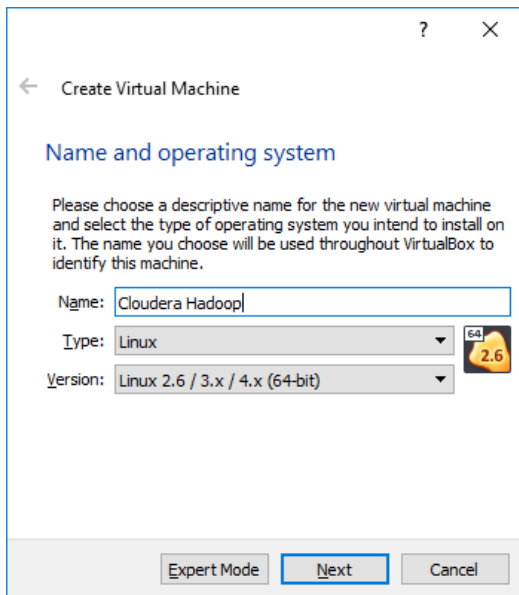
2. Extract 'Cloudera's QuickStart VM' compressed (zip) file. It extracts Virtual Machine Disk Format (VMDF) file: '**cloudera-quickstart-vm-5.13.0-0-virtualbox.vmdk**'
3. Copy this virtual machine image to a desired folder (e.g. a folder named 'Cloudera Hadoop').

 <b>Note it!</b>	<ul style="list-style-type: none"> <li>• <i>This folder and image file has to be the permanent location of your Hadoop installation.</i></li> <li>• <i>Do NOT delete this folder</i></li> </ul>
--	---

4. You will now create a virtual machine on VirtualBox. Open the **Oracle VM VirtualBox** application.



5. Click on 'New' to create a new virtual machine.

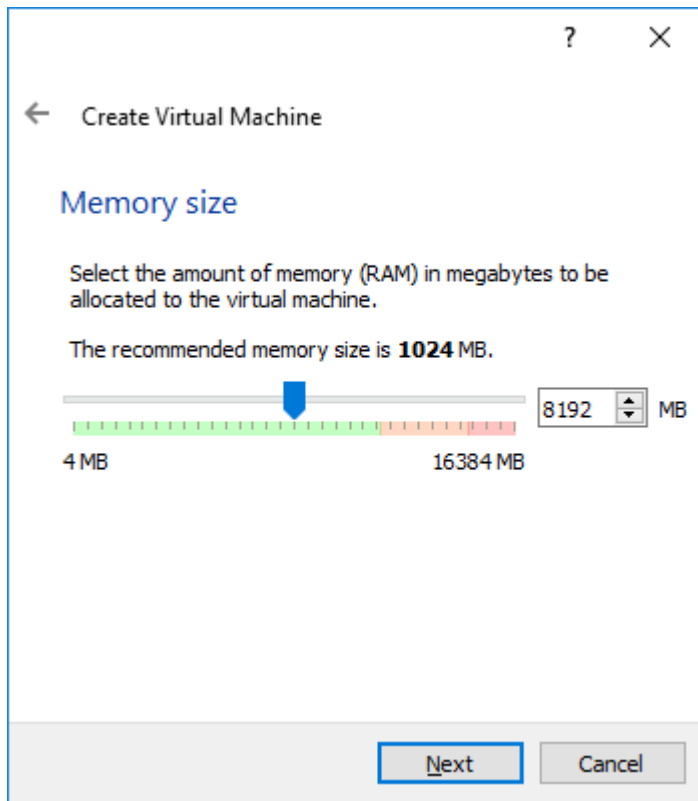


6. Give a name to the VM. Here 'Cloudera Hadoop' has been used.

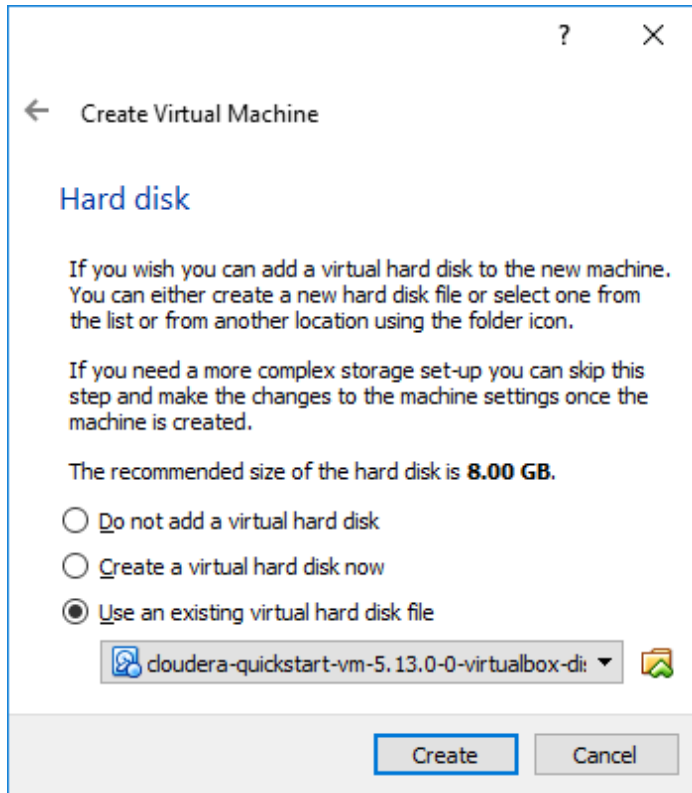
Pick the type: **Linux**

Choose version: **Linux 2.6/3.x/4.x (64 bit)**

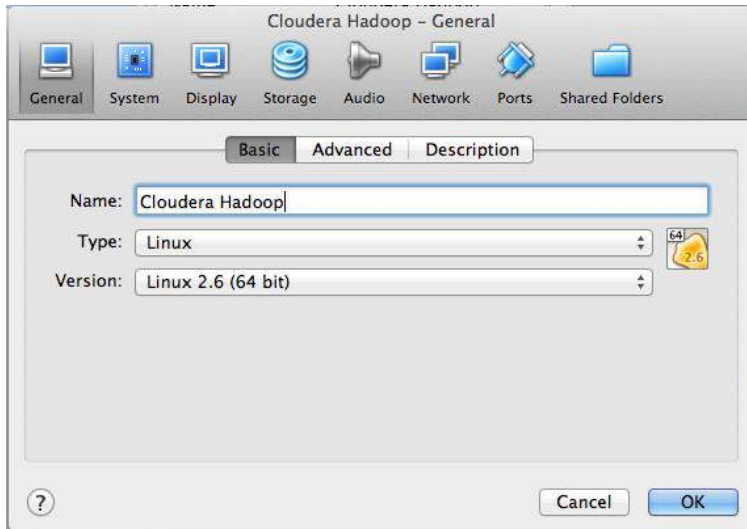
Click 'Next'.



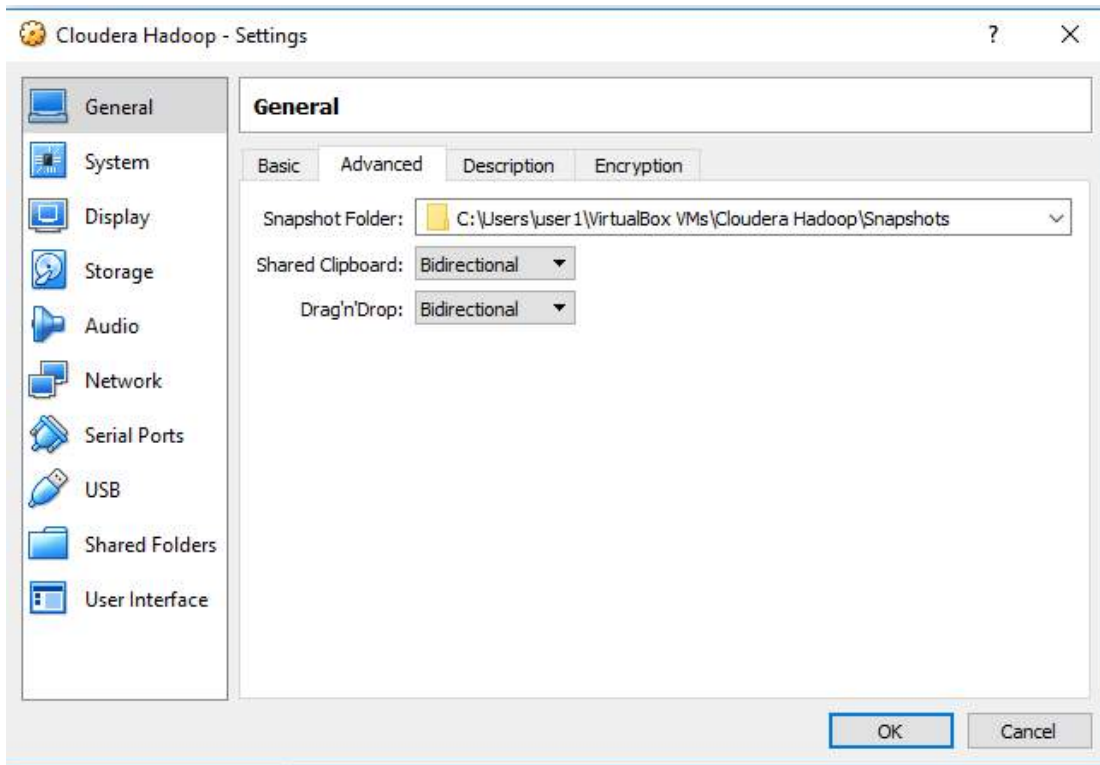
7. It is recommended that you use at least **8192** MB of RAM. Click 'Next'.



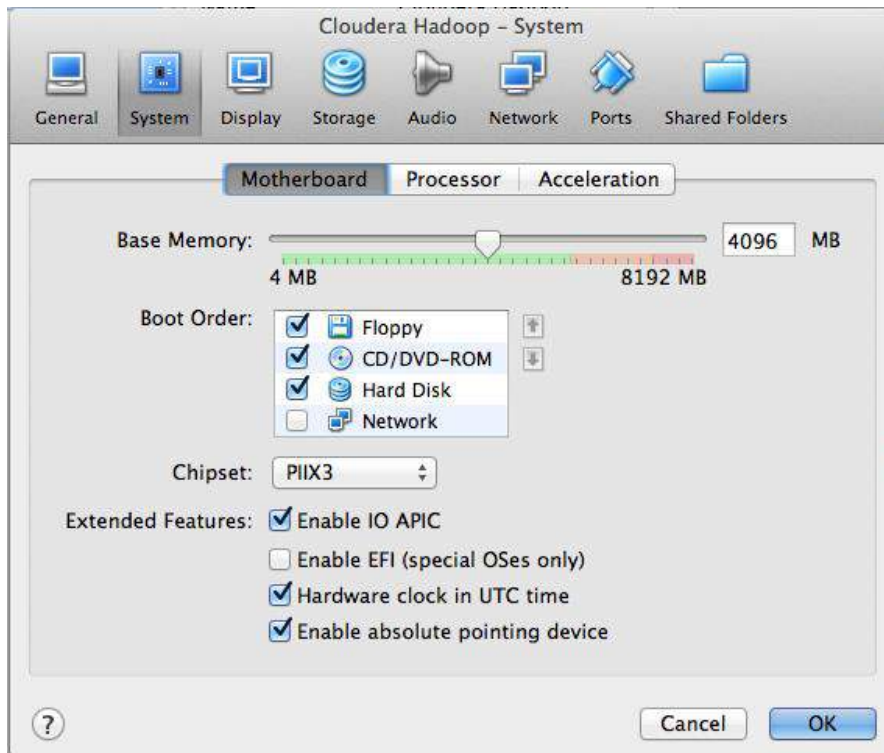
8. Choose the option 'Use an existing virtual hard drive file' and select the virtual image file ('**cloudera-quickstart-vm-5.13.0-0-virtualboxdemo-vm.vmdk**' ) saved in folder 'Cloudera Hadoop' (refer to step 2).
9. Click 'Create'.
10. Click 'Settings' to make a few recommended changes.



11. Click on tab 'Advanced' under 'General' category.



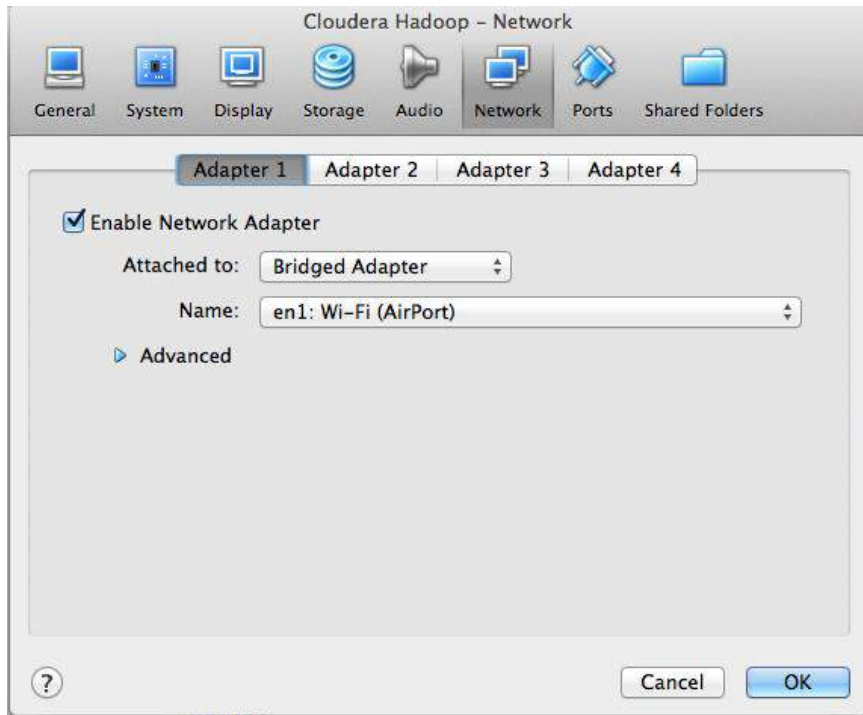
12. Pick 'Bidirectional' option for items: '**Shared Clipboard**' and '**Drag'n'Drop**'



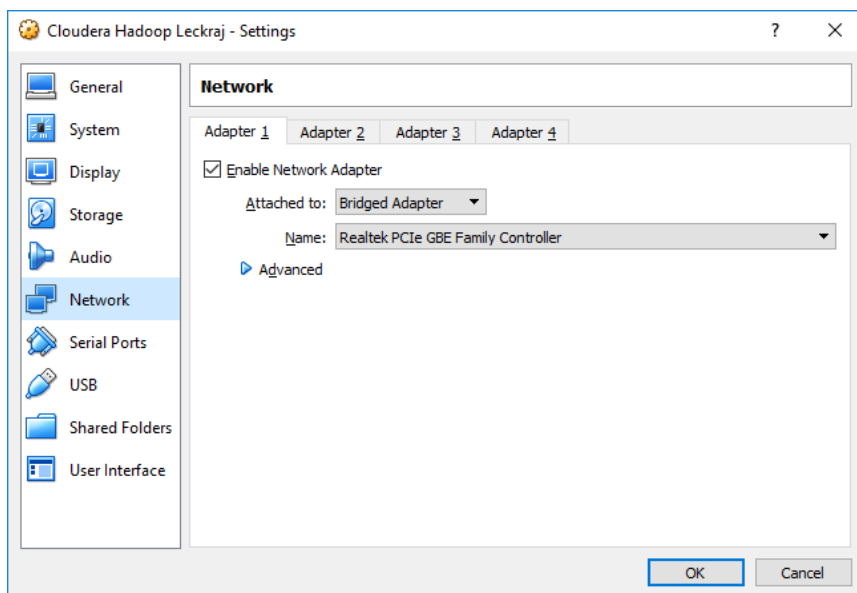
13. Click on 'System' category. Ensure option '**Enable IO APIC**' under 'Extended Features' is checked on (This is default!)

14. Click on 'Network' category. Choose Adapter 1 option '**Attached to:**' as '**Bridged Adapter**'. This gives you access to the physical Wi-Fi or Default Adapter: “Realtek PCIe GBE Family Controller”)

If you connected to a Wi-Fi network, you will have something similar to the following:

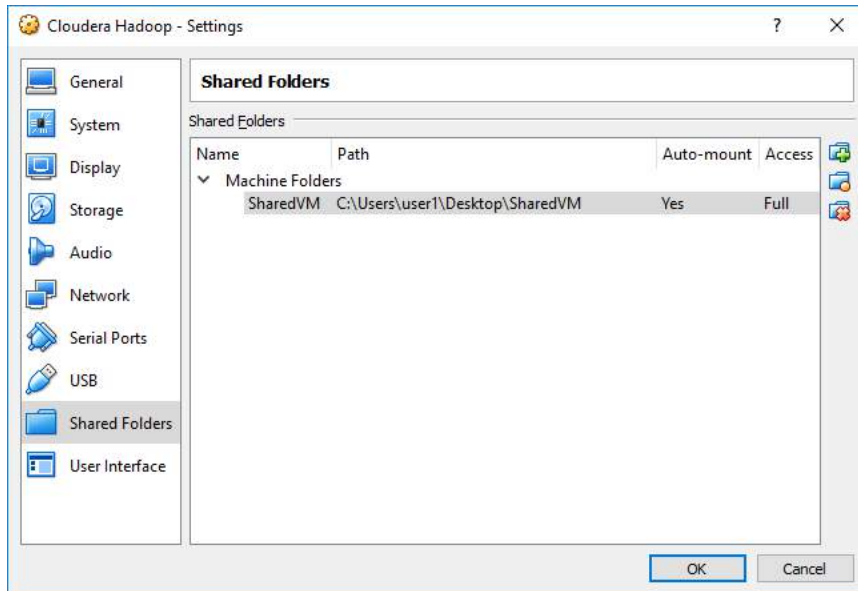


Alternatively, if you are using a wired connection, you might have the following screen.

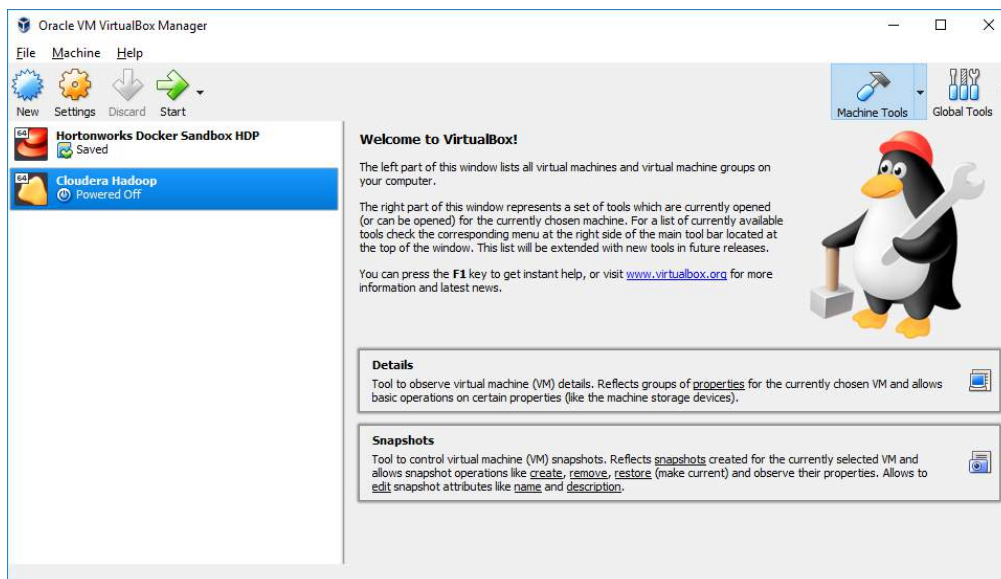


15. Click on '**Shared Folders**' category. Create a shared folder on desktop “**SharedVM**” and select the path accordingly. In this case, the path “C:\Users\user1\Desktop\SharedVM” has been used. Also, check **Auto-Mount**.

16. Click Ok and the following screen will appear.

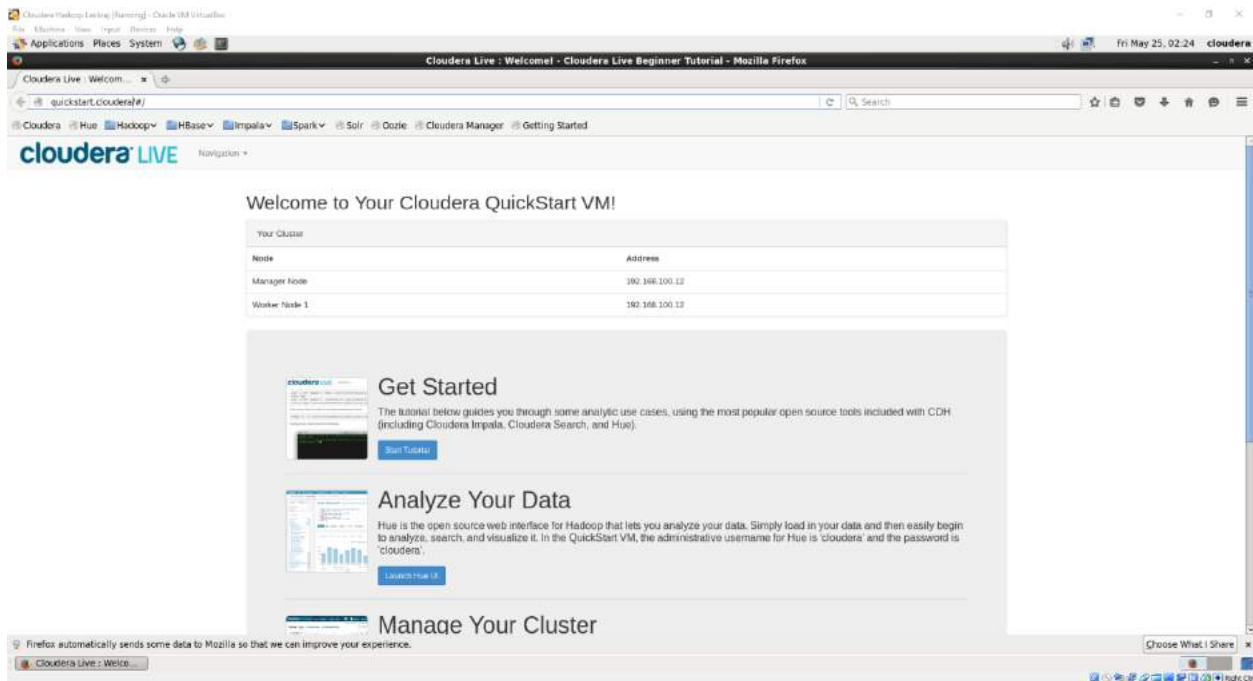


17. Click OK. The following screen appears:





18. You are now ready to start the virtual machine which includes the Hadoop Environment. Click **'Start'** to initiate the virtual machine. You will see several pages of output on a black screen until you finally see the desktop of the virtual machine.



19. Once you launch the VM, you are automatically logged in as the cloudera user.
20. If details about username and password is being requested, use the following:

The account details are:

username: cloudera

password: cloudera

The cloudera account has sudo privileges in the VM.

21. Open a Terminal on the Cloudera Hadoop Environment by clicking on Applications -> System Tools -> Terminal.

22. You will now update your kernel.

Switch to root user by typing the following:

```
sudo bash
```

Update the linux kernel:

```
yum install kernel -y
```

```
reboot
```



**Note it!**

- *Note that installation might take some time and will depend on internet connection.*

### 4.8.3 Getting Cloudera ready to run R

To be able to run R on the Big Data platform, the related R packages and IDE have to be installed. For this tutorial, the following have been used:

1. R
  2. R studio
1. To install R, first add the EPEL repository, then install *git*, *wget* and *R*. You need to find the latest release of the EPEL repository (<http://fedoraproject.org/wiki/EPEL>) and update the URL accordingly.

```
# yum install https://dl.fedoraproject.org/pub/epel/epel-release-latest-6.noarch.rpm  
sudo yum -y install git wget R
```

2. Download the latest version of RStudio server from the following link and for your specific platform. In this tutorial the CentOS Linux platform has been used.  
<https://www.rstudio.com/products/rstudio/download-server/>

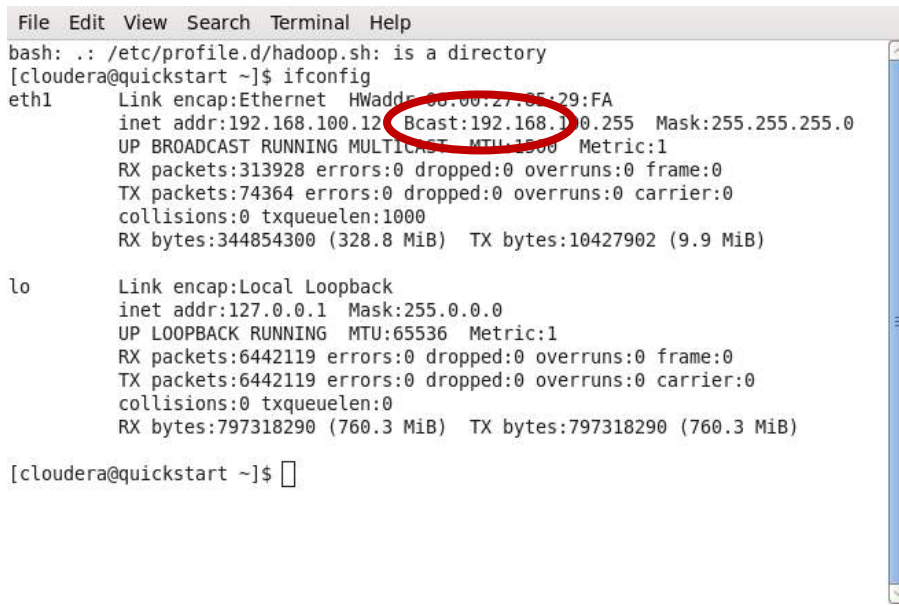
```
wget https://download2.rstudio.org/rstudio-server-rhel-1.1.453-x86_64.rpm
sudo yum install rstudio-server-rhel-1.1.453-x86_64.rpm
```

3. RStudio uses the port 8787. You can access RStudio from the browser of the **Virtual Machine** by either typing the following as the URL:  
<http://quickstart.cloudera:8787>

OR by using the IP address of the machine.

You can check the IP address of your virtual machine by running the following command on a terminal:

```
ifconfig
```



```
File Edit View Search Terminal Help
bash: ./etc/profile.d/hadoop.sh: is a directory
[cloudera@quickstart ~]$ ifconfig
eth1      Link encap:Ethernet  HWaddr 08:00:27:05:29:FA
          inet addr:192.168.100.12 Bcast:192.168.100.255 Mask:255.255.255.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500 Metric:1
          RX packets:313928 errors:0 dropped:0 overruns:0 frame:0
          TX packets:74364 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:344854300 (328.8 MiB)  TX bytes:10427902 (9.9 MiB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1 Mask:255.0.0.0
          UP LOOPBACK RUNNING  MTU:65536 Metric:1
          RX packets:6442119 errors:0 dropped:0 overruns:0 frame:0
          TX packets:6442119 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:797318290 (760.3 MiB)  TX bytes:797318290 (760.3 MiB)

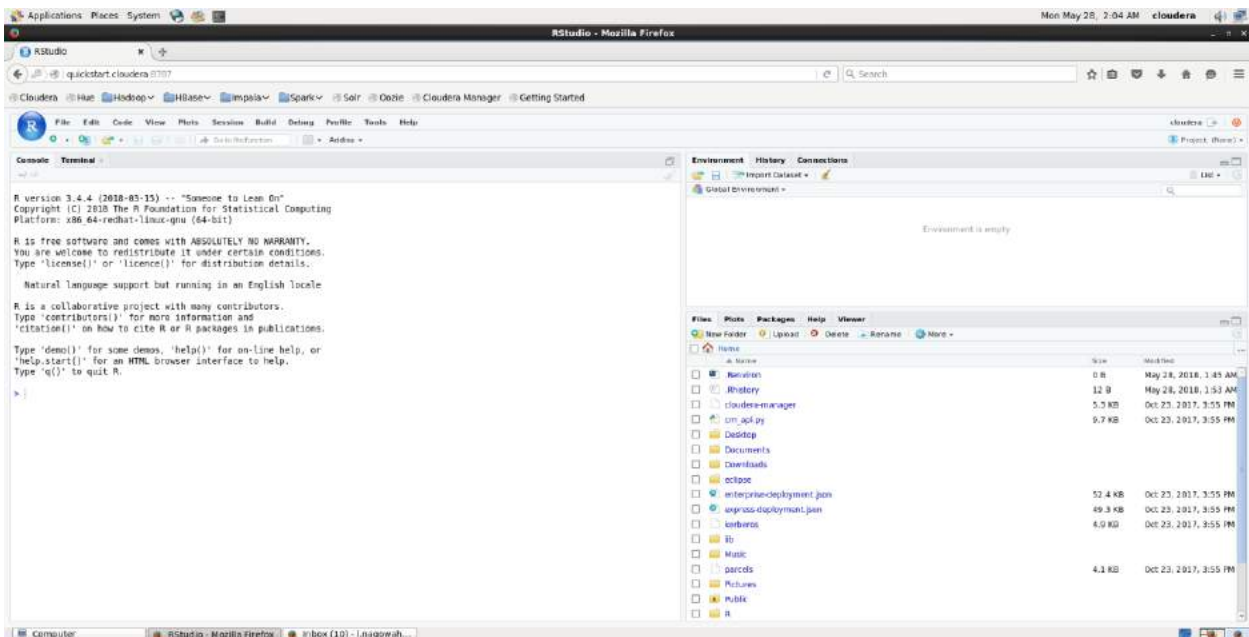
[cloudera@quickstart ~]$
```

From the above snapshot, RStudio can be accessed via the browser by typing the address <http://192.168.100.12:8787/>

4. Log in into RStudio by using the following username and password

Username: cloudera

Password: cloudera



## 4.8 SUMMARY

In this Unit, you learned the concepts and characteristics of Big Data along with the challenges that have cropped up with the high volume of data. You became familiar with the different application domains for Big Data. Additionally, you have learned how to set up a Big Data environment. Installation details pertaining to Cloudera Big Data platform has been included in the unit.

## 4.9 ADDITIONAL READINGS

1. John Walker, S., 2014. Big data: A revolution that will transform how we live, work, and think.
2. Zikopoulos, P. and Eaton, C., 2011. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.
3. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J. and Barton, D., 2012. Big data: the management revolution. *Harvard business review*, 90(10), pp.60-68.
4. White, T., 2012. *Hadoop: The definitive guide*. "O'Reilly Media, Inc.".
5. Cloudera, <https://www.cloudera.com/>

#### 4.10 REFERENCES

- Ammu, N. and Irfanuddin, M., 2013. Big data challenges. *International Journal of Advanced Trends in Computer Science and Engineering*, 2(1), pp.613-615.
- Anuradha, J., 2015. A brief introduction on Big data 5Vs characteristics and Hadoop Technology. *Procedia computer science*, 48, pp.319-324.
- Bertot, J.C. and Choi, H., 2013, June. Big data and e-government: issues, policies, and recommendations. In *Proceedings of the 14th Annual International Conference on Digital Government Research* (pp. 1-10). ACM.
- Chen, C.P. and Zhang, C.Y., 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, pp.314-347.
- Chen, M., Mao, S. and Liu, Y., 2014. Big data: A survey. *Mobile Networks and Applications*, 19(2), pp.171-209.
- Joaher, R. and Nagowah, S., 2017. A Big Data Framework for Diabetes in Mauritius. In *Infocom Technologies and Unmanned Systems (ICTUS' 2017)*.
- Kaisler, S., Armour, F., Espinosa, J.A. and Money, W., 2013, January. Big data: Issues and challenges moving forward. In *46th Hawaii international conference on System sciences (HICSS), 2013* (pp. 995-1004). IEEE.
- Kune, R., Konugurthi, P.K., Agarwal, A., Chillarige, R.R. and Buyya, R., 2016. The anatomy of big data computing. *Software: Practice and Experience*, 46(1), pp.79-105.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H., 2011. Big data: The next frontier for innovation, competition, and productivity.
- McAfee, A., Brynjolfsson, E. and Davenport, T.H., 2012. Big data: the management revolution. *Harvard business review*, 90(10), pp.60-68.
- Raghupathi, W. and Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1), p.3.
- Rajagopalan, M.R. and Vellaipandiyam, S., 2013, November. Big data framework for national e-governance plan. In *ICT and Knowledge Engineering (ICT&KE), 2013 11th International Conference on* (pp. 1-5). IEEE.
- Sagiroglu, S. and Sinanc, D., 2013, May. Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.
- Tole, A.A., 2013. Big data challenges. *Database Systems Journal*, 4(3), pp.31-40.

Vesset, D., Woo, B., Morris, H.D., Villars, R.L., Little, G., Bozman, J.S., Borovick, L., Olofson, C.W., Feldman, S., Conway, S. and Eastwood, M., 2012. Market Analysis–Worldwide Big Data Technology and Services 2012-2015 Forecast. *IDC Analyze the Future, 1*, pp.1-34.

Wikibon, 2014, <http://wikibon.org>