

LATE BHAUSAHEB HIRAY S.S. TRUST'S INSTITUTE
OF COMPUTER APPLICATION, MUMBAI

Big Data and Visualization

**DR. SADHANA OJHA,
ASST. PROF. AQUILA SHAIKH**

Faculty, Master of Computer Application (M.C.A.)

Late Bhausaheb Hiray S.S. Trust's Institute of Computer Application



BIG DATA AND VISUALIZATION

BIG DATA AND VISUALIZATION

LATE BHAUSAHEB HIRAY S.S. TRUST'S INSTITUTE OF
COMPUTER APPLICATION

Dr. Sadhana Ojha

Co Author: Asst. Prof. Aquila Shaikh

Faculty, Master of Computer Application (M.C.A.)

Late Bhausaheb Hiray S.S. Trust's Institute of Computer Application



First Edition, 2023

Copyright © Late Bhausaheb Hiray S.S. Trust's Institute Of Computer Application, Bandra (E), Mumbai-51, 2023

All rights reserved. No part of this publication may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, without the prior written permission of the author, except in the case of brief quotations embodied in critical reviews and certain other non-commercial uses permitted by copyright law. For permission requests, write to the publisher at the address below.

This book can be exported from India only by the publishers or by the authorized suppliers. Infringement of this condition of sale will lead to Civil and Criminal prosecution.

Paperback ISBN: 978-81-19221-26-4

eBook ISBN: 978-81-19221-15-8

WebPDF ISBN: 978-81-19221-03-5

Note: Due care and diligence has been taken while editing and printing the book; neither the author nor the publishers of the book hold any responsibility for any mistake that may have inadvertently crept in.

The publishers shall not be liable for any direct, consequential, or incidental damages arising out of the use of the book. In case of binding mistakes, misprints, missing pages, etc., the publishers' entire liability, and your exclusive remedy, is replacement of the book within one month of purchase by similar edition/reprint of the book.

Printed and bound in India by

16Leaves

2/579, Singaravelan Street

Chinna Neelankarai

Chennai – 600 041, India

info@16leaves.com

www.16Leaves.com

Call: 91-9940638999

Contents

1. Introduction of Big Data	1
1.1. Introduction	
<hr/>	
2. HDFS and MapReduce	3
2.1. Introduction	
2.2. Hardware Required	
2.3. Software Required	
2.4. Installation and Configuration	
2.5. Practice	
2.6. MapReduce	
2.7. Practice	
<hr/>	
3. No SQL	47
3.1. Introduction	
3.2. Hardware Required	
3.3. Software Required	
3.4. Installation and Configuration	
3.5. Practice	
<hr/>	
4. Hadoop Eco-system (HIVE and PIG)	55
4.1. Introduction	
4.2. Hardware Required	
4.3. Software Required	
4.4. Installation and Configuration	
4.5. Practice	
<hr/>	
5. Data Visualization	81
5.1. Introduction	
5.2. Hardware Required	
5.3. Software Required	
5.4. Installation and Configuration	
5.5. Practice	

Chapter 1 Introduction of Big Data

1.1 Introduction

Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions. These processes use familiar statistical analysis techniques—like clustering and regression—and apply them to more extensive datasets with the help of newer tools.

Big data is a relative term. If big data is referred by “volume” of transactions and transaction history, then hundreds of terabytes (10¹² bytes) may be considered “big data” for a pharmaceutical company and volume of transactions in petabytes (10¹⁵ bytes).

Big Data Analytics as shown in Fig. 1.1 is the result of three major trends in computing: Mobile Computing using hand-held devices, such as smartphone and tablets; Social Networking, such as Facebook and Pinterest; and Cloud Computing by which one can rent or lease the hardware setup for storing and computing.

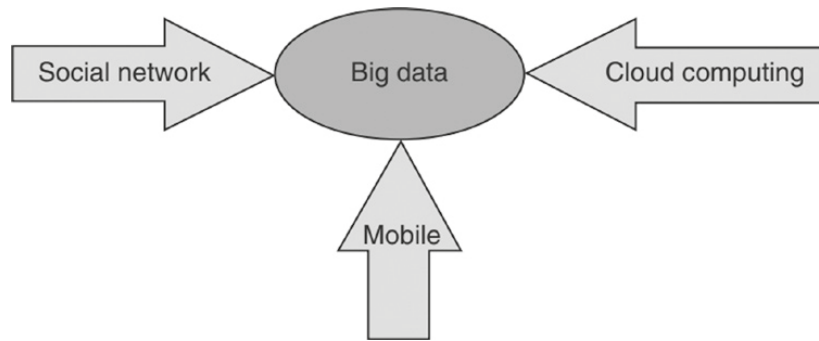


Figure 1.1 Big Data: Result of three computing trends.

Big data analytics is the use of advanced analytic techniques against very large, diverse big data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zettabytes.

It can be defined as data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency. Characteristics of big data include high volume, high velocity and high variety. Sources of data are becoming more complex than those for traditional data because they are being driven by artificial intelligence (AI), mobile devices, social media and the Internet of Things (IoT). For example, the different types of data originate from sensors, devices, video/audio, networks, log files, transactional applications, web and social media – much of it generated in real time and at a very large scale.

Chapter 2 HDFS and MapReduce

2.1 Introduction

HDFS is a distributed file system that provides a limited interface for managing the file system to allow it to scale and provide high throughput. HDFS creates multiple replicas of each data block and distributes them on computers throughout a cluster to enable reliable and rapid access. When a file is loaded into HDFS, it is replicated and fragmented into “blocks” of data, which are stored across the cluster nodes; the cluster nodes are also called the DataNodes. The NameNode is responsible for storage and management of metadata, so that when MapReduce or another execution framework calls for the data, the NameNode informs it where the data that is needed resides. Figure 2.1 shows the NameNode and DataNode block replication in HDFS architecture.

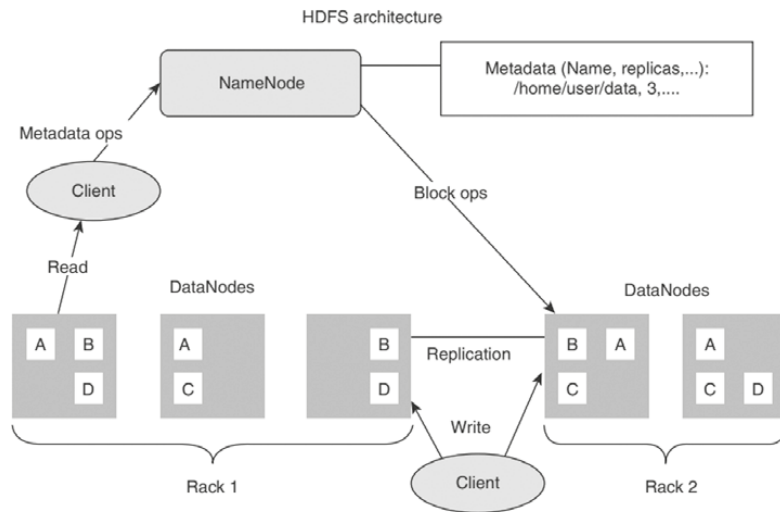
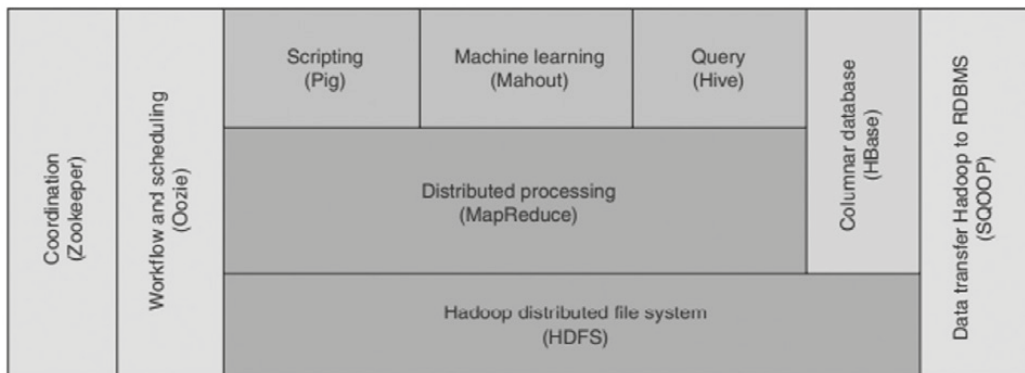


Figure 2.1 NameNode and DataNode block replication.

Hadoop Ecosystem



2.2 Hardware Required

- » Windows 10,11,etc.
- » 8 GB of RAM
- » i3 Processor
- » 64-bit operating system, x64-based processor

2.3 Software Required

- » Oracle VM virtualBox
- » cloudera-quickstart-vm-5.4.2-0-virtualbox

2.4 Installation and Configuration

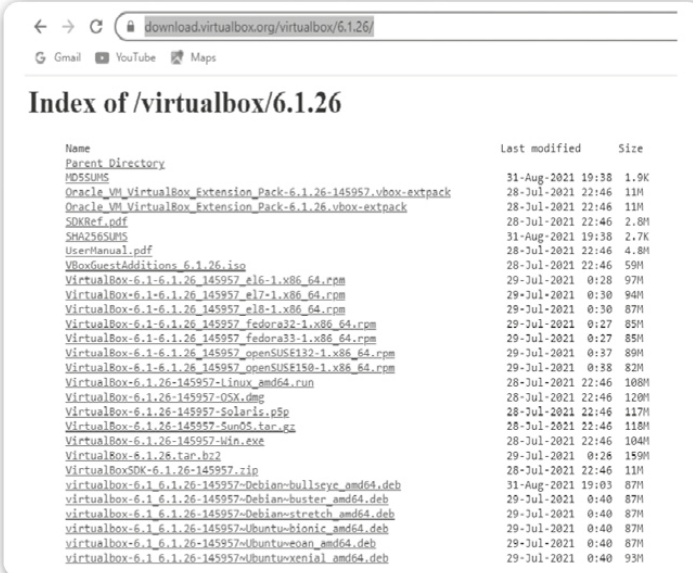
Step 1: download cloudera image using below link https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.4.2-0-virtualbox.zip

Step 2: Unzip the downloaded Zipped file. After unzipping we will get a folder named cloudera-quickstart-vm-5.4.2-0-virtualbox. Inside this folder two files will be there.

- cloudera-quickstart-vm-5.4.2-0-virtualbox
- cloudera-quickstart-vm-5.4.2-0-virtualbox-disk1

Now next step is to download Oracle VM virtual box by following below steps.

Step 3: Download Oracle VM virtualBox for WINDOWS using this link <https://download.virtualbox.org/virtualbox/6.1.26/>

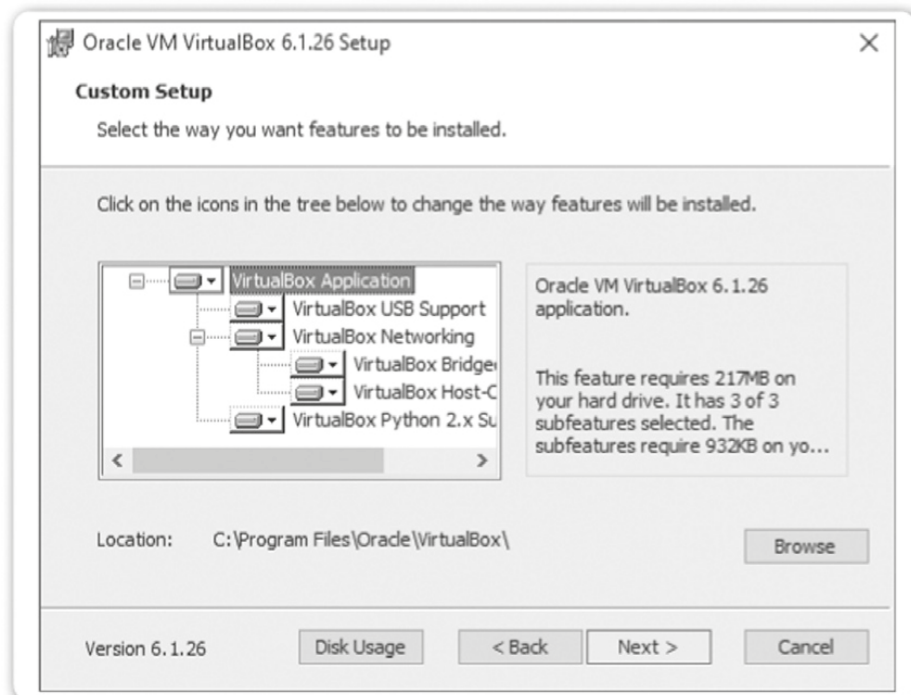


Name	Last modified	Size
Parent Directory		
HPSSUSPS	31-Aug-2021 19:38	1.9K
Oracle_VM_VirtualBox_Extension_Pack-6.1.26-145957.vbox-extpack	28-Jul-2021 22:46	11M
Oracle_VM_VirtualBox_Extension_Pack-6.1.26.vbox-extpack	28-Jul-2021 22:46	11M
SDKRef.pdf	28-Jul-2021 22:46	2.8M
SHA256SUMS	31-Aug-2021 19:38	2.7K
UserManual.pdf	28-Jul-2021 22:46	4.8M
VBoxGuestAdditions_6.1.26.iso	28-Jul-2021 22:46	59M
VirtualBox-6.1.6.1.26-145957-el6-1.x86_64.rpm	29-Jul-2021 0:28	97M
VirtualBox-6.1.6.1.26-145957-el7-1.x86_64.rpm	29-Jul-2021 0:30	94M
VirtualBox-6.1.6.1.26-145957-el8-1.x86_64.rpm	29-Jul-2021 0:30	87M
VirtualBox-6.1.6.1.26-145957-fedora32-1.x86_64.rpm	29-Jul-2021 0:27	85M
VirtualBox-6.1.6.1.26-145957-fedora33-1.x86_64.rpm	29-Jul-2021 0:27	85M
VirtualBox-6.1.6.1.26-145957-openSUSE132-1.x86_64.rpm	29-Jul-2021 0:37	89M
VirtualBox-6.1.6.1.26-145957-openSUSE150-1.x86_64.rpm	29-Jul-2021 0:38	82M
VirtualBox-6.1.26-145957-linux_amd64.run	28-Jul-2021 22:46	108M
VirtualBox-6.1.26-145957-OSX.dmg	28-Jul-2021 22:46	129M
VirtualBox-6.1.26-145957-Solaris.p5p	28-Jul-2021 22:46	117M
VirtualBox-6.1.26-145957-SunOS.tar.gz	28-Jul-2021 22:46	118M
VirtualBox-6.1.26-145957-win.exe	28-Jul-2021 22:46	104M
VirtualBox-6.1.26.tar.bz2	29-Jul-2021 0:28	159M
VirtualBoxSDK-6.1.26-145957.zip	28-Jul-2021 22:46	11M
virtualbox-6.1.6.1.26-145957-debian-bullseye_amd64.deb	31-Aug-2021 19:03	87M
virtualbox-6.1.6.1.26-145957-debian-buster_amd64.deb	29-Jul-2021 0:40	87M
virtualbox-6.1.6.1.26-145957-debian-stretch_amd64.deb	29-Jul-2021 0:40	87M
virtualbox-6.1.6.1.26-145957-ubuntu-bionic_amd64.deb	29-Jul-2021 0:40	87M
virtualbox-6.1.6.1.26-145957-ubuntueven_amd64.deb	29-Jul-2021 0:40	87M
virtualbox-6.1.6.1.26-145957-ubuntuxenial_amd64.deb	29-Jul-2021 0:40	93M

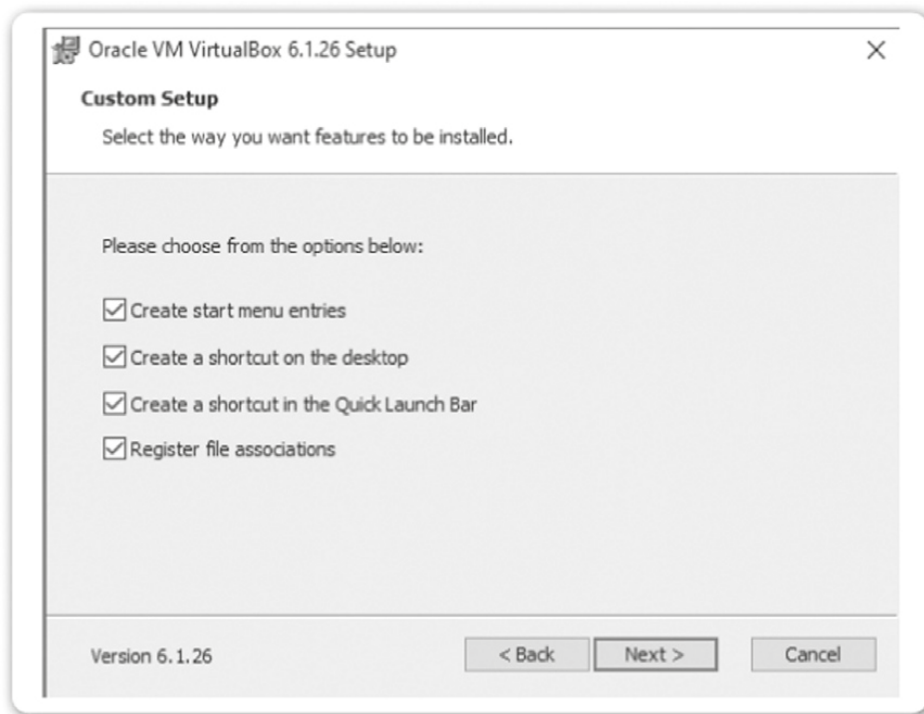
Step 4: After downloading this .exe file simply double click on it and install it.



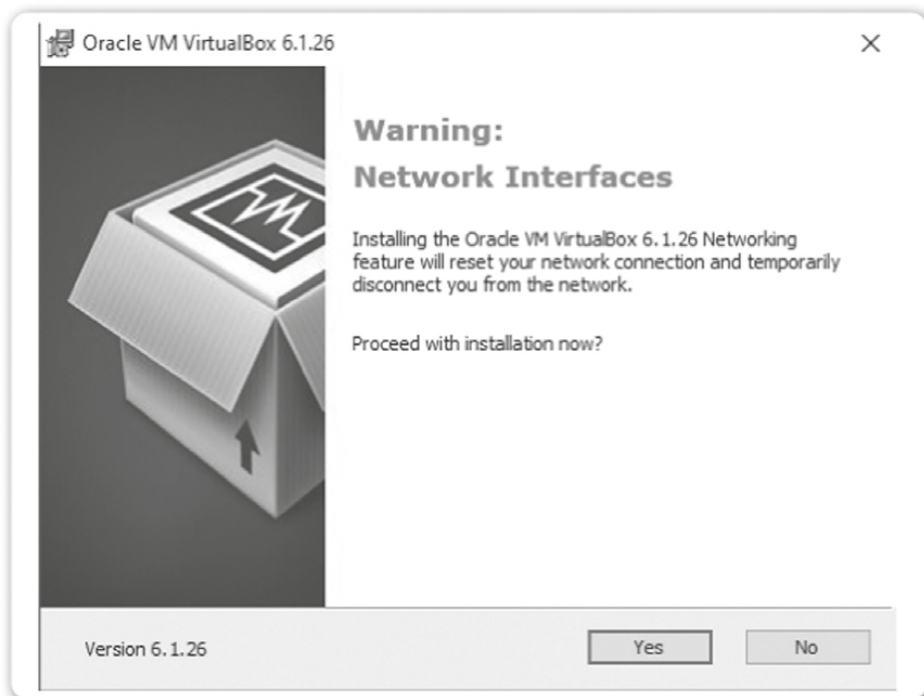
Step 5: Click on Next



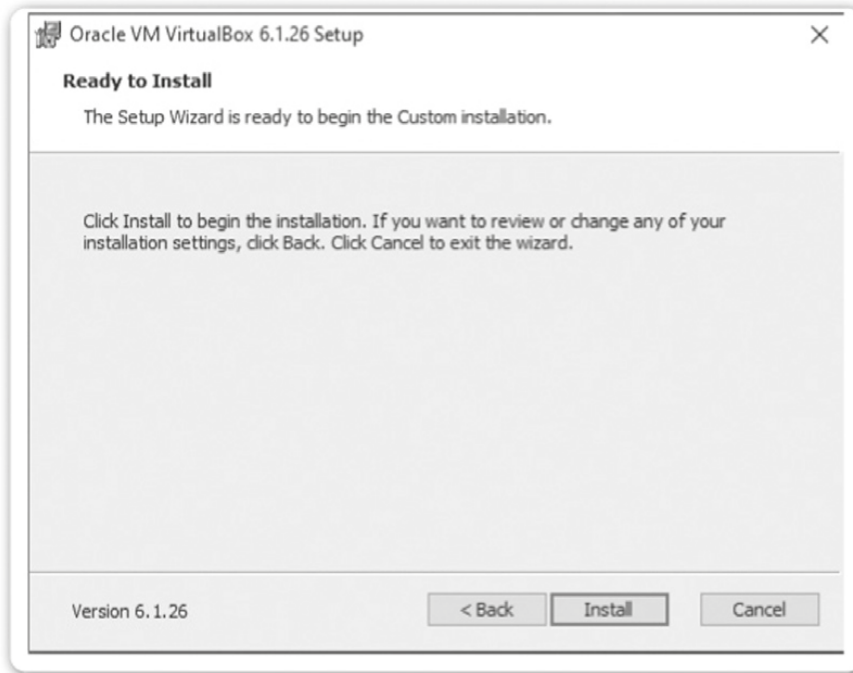
Step 6: Click on Next



Step 7: Click on Next

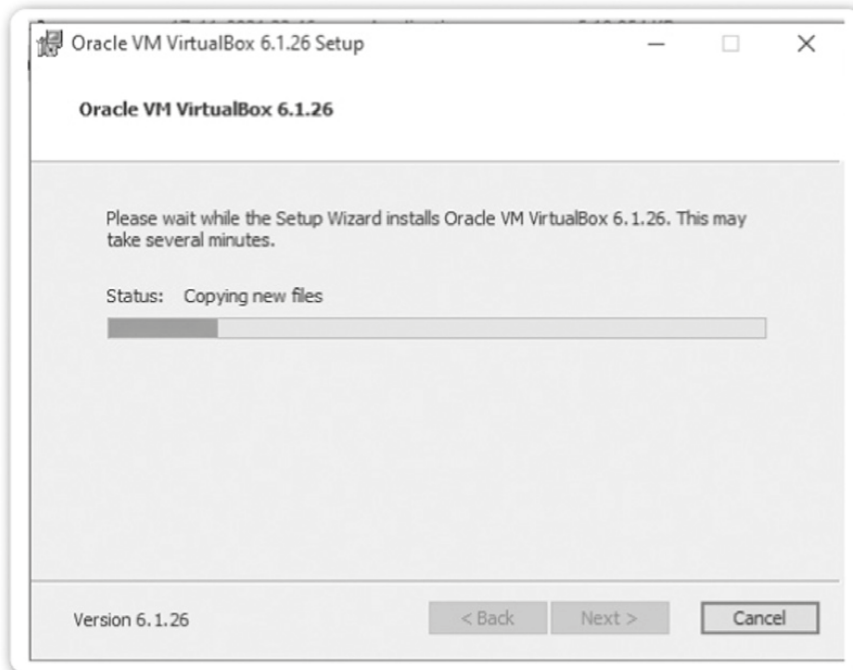


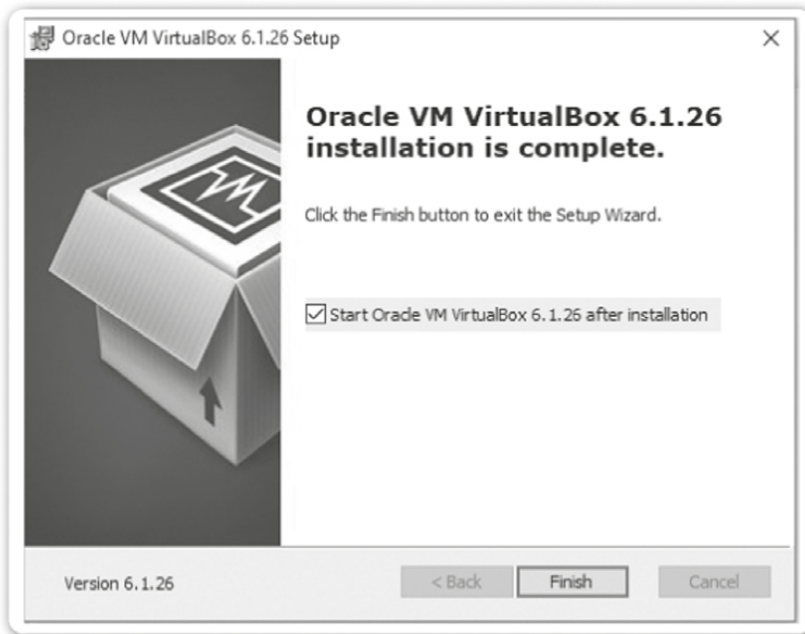
Step 8: Click on Yes



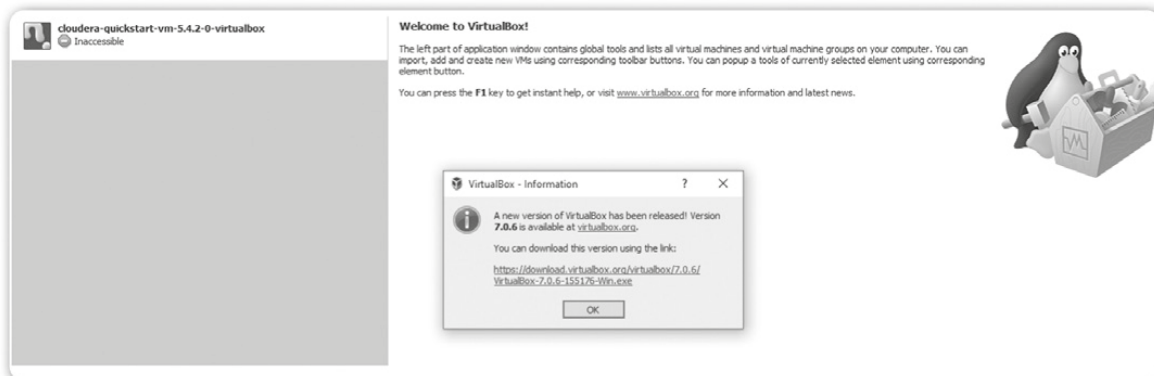
Step 9: Click on Install

Step 10: After clicking install it will ask to allow changes in device then click on yes. After clicking on Yes it will start installing.





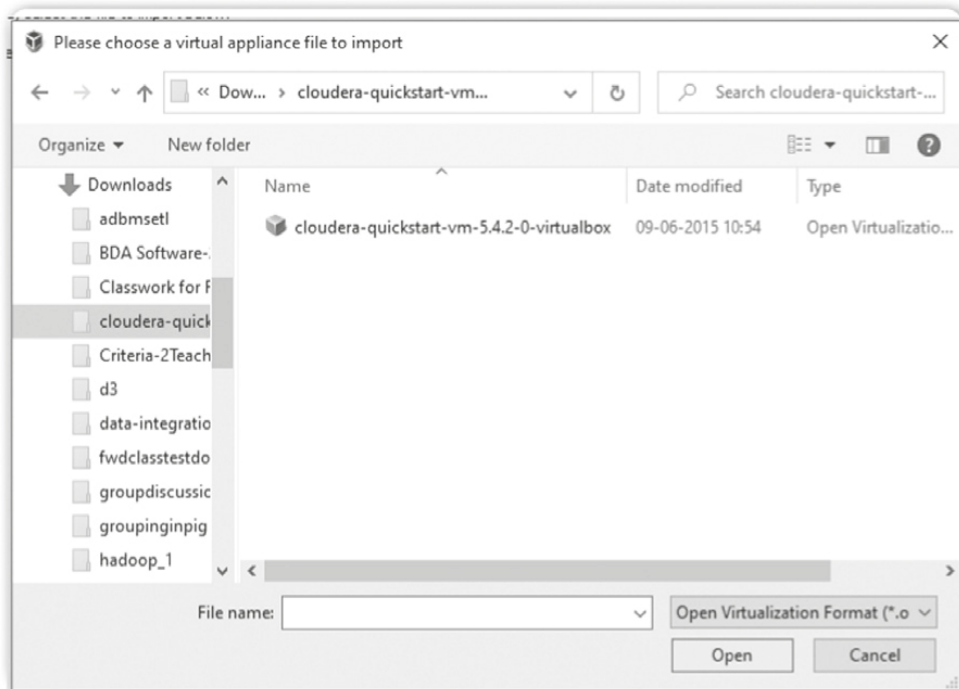
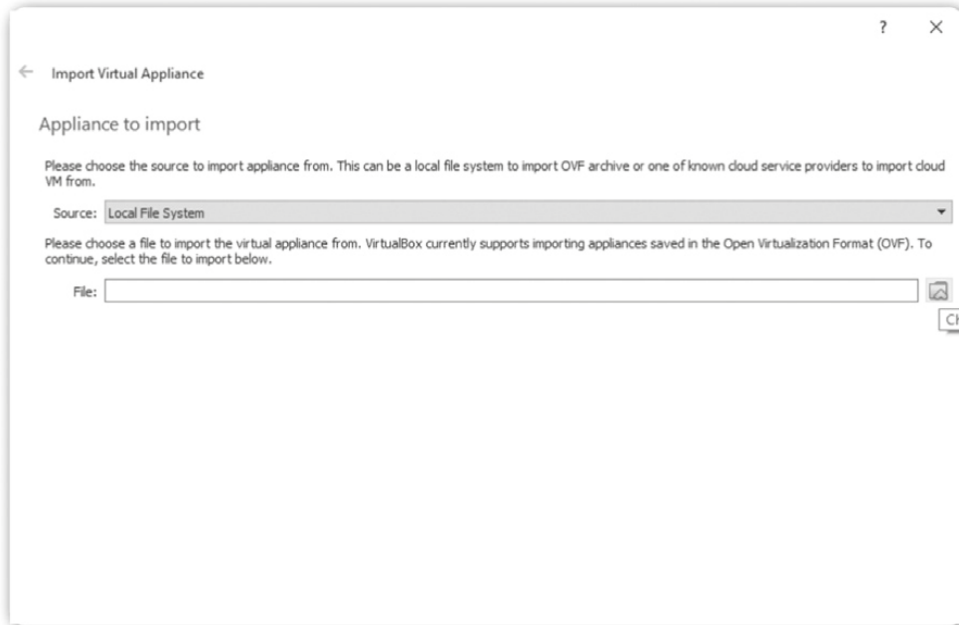
Step 11: Click on Finish. Below screen will appear. Click on OK.



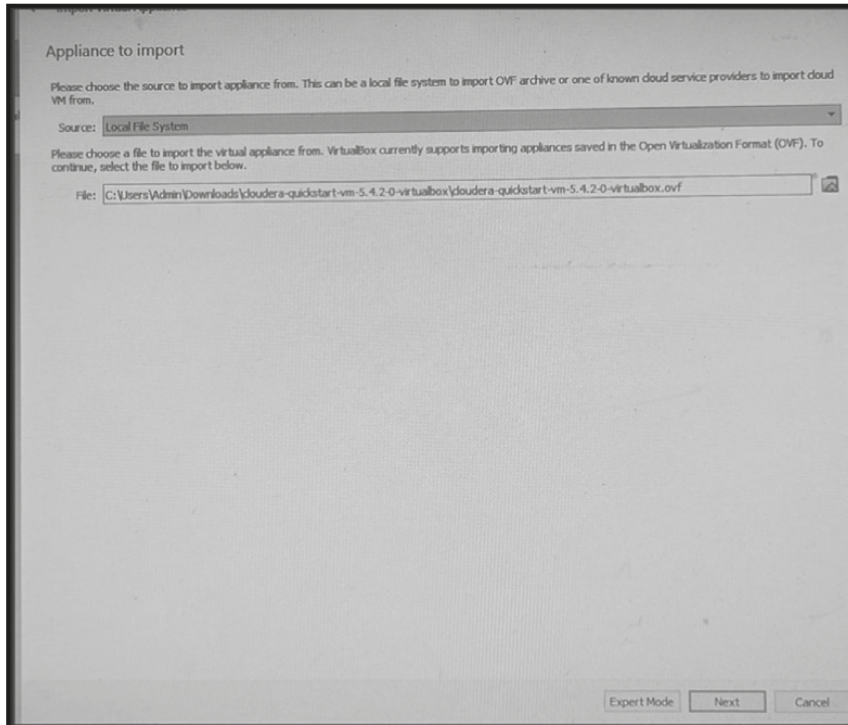
Step 12: After clicking OK in step 11 below screen will appear then click on import icon.



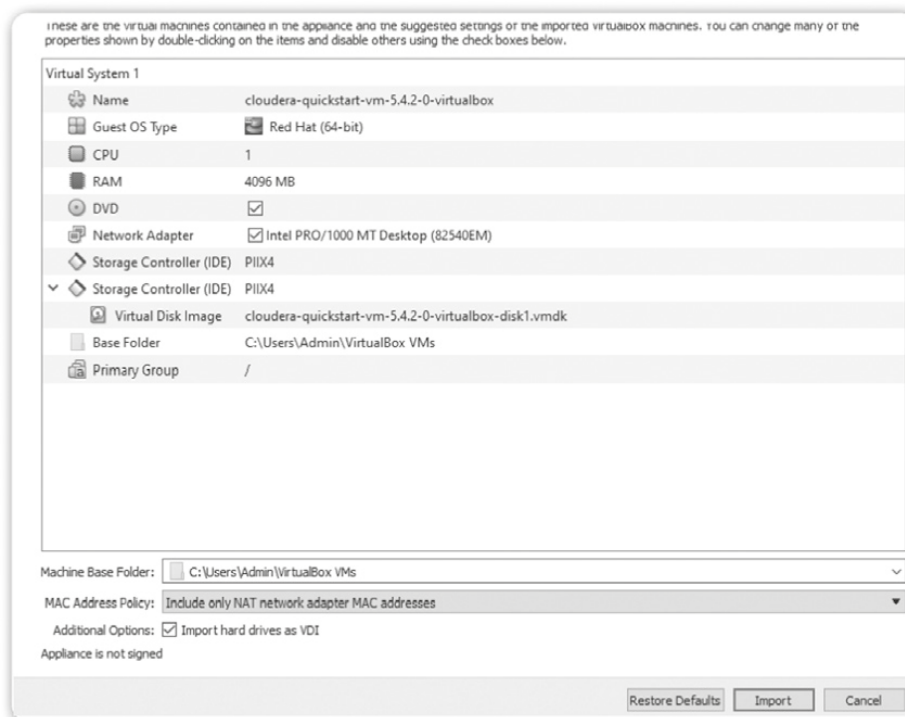
Step 13: Import the cloudera file mentioned in step 2 (a) and click on Open.



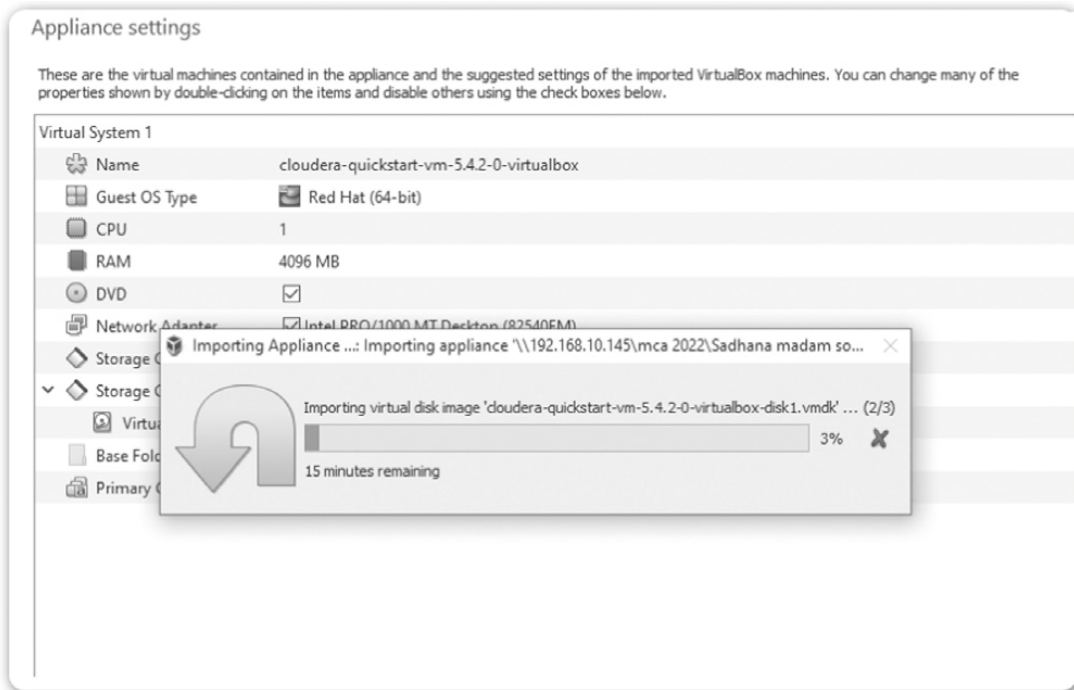
Step 14: Click on Next in below screen.



Step 15: Click on import in below screen.



Step 16: After clicking on import it will start importing file.



Step 17: After importing file below screen will appear

