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 VISUALZATION

BIG DATA AND VISUALZATION

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. Introduction**
- 2. HDFS and MapReduce
 - 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

3. No SQL

- **3.1. Introduction**
- **3.2. Hardware Required**
- **3.3. Software Required**
- **3.4. Installation and Configuration**
- 3.5. Practice
- 4. Hadoop Eco-system (HIVE and PIG)
 - 4.1. Introduction
 - 4.2. Hardware Required
 - 4.3. Software Required
 - 4.4. Installation and Configuration
 - 4.5. Practice

5. Data Visualization

- 5.1. Introduction
- 5.2. Hardware Required

5.3. Software Required

- 5.4. Installation and Configuration
- 5.5. Practice

3

47

55

81

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

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.
 - a) cloudera-quickstart-vm-5.4.2-0-virtualbox
 - b) 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/



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



Step 5: Click on Next

Calact the way you u	ant features to be installed	
Select the way you w	ant reatures to be installed.	
Click on the icons in t	he tree below to change the	way features will be installed.
Uirtu	alBox Application	Oracle VM VirtualBox 6.1.26
		application.
	VirtualBox Bridge	This feature requires 217MP on
	VirtualBox Host-C	your hard drive. It has 3 of 3
	VirtualBox Python 2.x Su	subfeatures selected. The subfeatures require 932KB on yo
<	>	
Location: C: Prog	gram Files (Oracle (VirtualBox	Brow

Oracle VIVI VirtualBox 0.1.20 S	etup	×
Custom Setup		
Select the way you want featu	ures to be installed.	
Please choose from the options	s below:	
Create start menu entries		
Create a shortcut on the de	esktop	
Create a shortcut in the Qu	iidk Launch Bar	
Register file associations		
Version 6.1.26	< Back Next >	Cancel

Step 7: Click on Next

Pracle VM VirtualBox 6.1.26	Warning: Network Interfaces Installing the Oracle VM VirtualBox 6.1.26 Networking feature will reset your network connection and temporarily disconnect you from the network. Proceed with installation now?	×
Version 6.1.26	Yes No	

Oracle VM VirtualBox 6.1.26 Setu	up	\times
Ready to Install		
The Setup Wizard is ready to beg	in the Custom installation.	
Click Install to begin the installation	on. If you want to review or change any of your	
installation settings, dick Back. Cli	lick Cancel to exit the wizard.	

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.

Oracle VM VirtualBox 6.1.26 Se	etup		-		×
Oracle VM VirtualBox 6.1.26	5				
Please wait while the Setup \ take several minutes	Wizard installs Oracle VM	VirtualBox 6	1.26. Tl	his may	
Status: Copying new files					



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

Coudera-quickstart-vm-5.4.2-0-virtualbox	Welcome to VirtualBox! The left part of application window contains global tools and lists all virtual machines and virtual machine groups on your computer. You can moort, add and create new VMs using corresponding toolbar buttoms. You can popula a tools of currently selected element using corresponding toolbar buttoms. You can press the F1 key to get instant help, or visit <u>www.virtualbox.org</u> for more information and latest news.
	VirtualBox - Information X VirtualBox - Information A new version of WirualBox has been released! Version D.6 a available at <u>strubulance.co</u> : Voc can download this version using the link: <u>https://download.virtualbox.co/wirualbox/70.0.6/</u> VirtualBox-7.0.6-155128-Wirualbox-7.0.6/ OK OK

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

Oracle VM VirtualBox Manager File Machine Help		o x
Control of the second	Preference: Import Appliance (Crti-I) Woom to Virtualitation: We will be the set of spokation window contains global books and lists all virbal machines and virbal machine groups on your computer. You can import, add and create new We using corresponding toobar buttom. You can popup a tools of currently selected element using corresponding import, add and create new We using corresponding toobar buttom. You can popup a tools of currently selected element using corresponding and create new We using corresponding toobar buttom. You can popup a tools of currently selected element using corresponding to use an press the F1 key to get instant help, or visit <u>www.virbabox.org</u> for more information and latest news.	

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

	?	2	×
Import Virtual Appliance			
Appliance to import			
Please choose the source to import appliance from. This can be a local file system to import OVF archive or one of known cloud se VM from.	rvice providers to imp	ort do	ud
Source: Local File System			Ŧ
Please choose a file to import the virtual appliance from. VirtualBox currently supports importing appliances saved in the Open Vir continue, select the file to import below.	tualization Format (O)	VF). To	
File:		[2
			Ch

Please choose a virtual appliance file to import			×
\leftarrow \rightarrow \checkmark \uparrow \blacksquare $<$ Dow $>$ cloudera-quickstart-vm \sim	ē	, Search cl	oudera-quickstart
Organize 🔻 New folder			•••
 Downloads Albmsetl BDA Software- Classwork for F Classwork for F Cloudera-quick Criteria-2Teach d3 data-integratio fwdclasstestdo groupdiscussic groupinginpig 	Da 09-	te modified -06-2015 10:54	Type Open Virtualizatio
File name:	~	Open Virtualiza	> ation Format (*.o ~ Cancel

Step 14: Click on Next in below screen.

Source:	Local File System			
Please cho continue, s	ose a file to import the virtual appliance select the file to import below.	e from. VirtualBox currently su	pports importing appliances saved in the Open Virtualization Form	at (OVF). To
File:	C:\Users\Admin\Downloads\doudera-c	aidstart-vm-5.4.2-0-virtualb	ox\doudera-quickstart-vm-5.4.2-0-virtualbox.ovf	ľ

Step 15: Click on import in below screen.

	System 1		
3	Name	cloudera-quickstart-vm-5.4.2-0-virtualbox	
	Guest OS Type	Red Hat (64-bit)	
	CPU	1	
	RAM	4096 MB	
\odot	DVD		
F	Network Adapter	Intel PRO/1000 MT Desktop (82540EM)	
0	Storage Controller (IDE)	PIIX4	
× 0	Storage Controller (IDE)	PIIX4	
	Virtual Disk Image	cloudera-quickstart-vm-5.4.2-0-virtualbox-disk1.vmdk	
	Base Folder	C:\Users\Admin\VirtualBox VMs	
dia la	Primary Group	1	
6.65		,	
		,	
lachine	Base Folder: C:\Users	, VAdmin\VirtualBox VMs	
lachine I	Base Folder: C:\Users Idress Polcy: Include only 1	, Vadmin\VirtualBox VMs NAT network adapter MAC addresses	
lachine I MAC Ad Additio	Base Folder: C:\Users Idress Polcy: Include only 1 onal Options: Import har	, Vadmin\VirtualBox VMs NAT network adapter MAC addresses rd drives as VDI	

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

Guest C CPU RAM	Cloud DS Type 2 R 1 4096	led Hat (64-bit)
Guest C CPU RAM O DVD	DS Type Marked R 1 4096	ted Hat (64-bit)
CPU RAM OVD	1 4096	
RAM DVD	4096	
OVD		MB
_	\checkmark	
Networ 🖒 Storage	Manter In	tel DRO/1000 MT Deckton (82540EM) nce: Importing appliance "\\192.168.10.145\mca 2022\Sadhana madam so X
Storage		nporting virtual disk image 'doudera-quickstart-vm-5.4.2-0-virtualbox-disk1.vmdk' (2/3) 3%
Primary		

Step 17: After importing file below screen will appear

Ø Oracle VM Vietnalliere Manager File Machine Help		- a ×
Tools	Nor acting part let	
timires quickstart em 5.42 8 virtusibar Present CB	E Secretal New: dealtrequidstart m 5.1.20-vitable:	Proven
	Syntem Soc Peters Soc Peters Vouses Process Vouses Vouses Vouses Vouses Vouses Vouses	cloudera-quickstart- vm-5.4.2-0-virtualbox
	Trephy Yello Manony Copier Contain Kanono Cintain Kanono Cintain Rocking Stateful Statef	
	Storage Controller: DE Controller DE Finers (Perce C. doubter-calcostationers 4.2,-o-astuation-dat.Luck (terms, e4.01.09) DE Seconder Denie C. logical Direct Empty	
	Deabled	
	Ø INS Disabled	
	Shared folders	