Big Data and Hadoop

2nd Edition

Fundamentals, tools, and techniques for data-driven success

Mayank Bhushan

www.bpbonline.com

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor BPB Online or its dealers and distributors, will be held liable for any damages caused or alleged to have been caused directly or indirectly by this book.

BPB Online has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, BPB Online cannot guarantee the accuracy of this information.

First published: 2018
Second published: 2024

Published by BPB Online
WeWork
119 Marylebone Road
London NW1 5PU

UK | UAE | INDIA | SINGAPORE

ISBN 978-93-55516-664

www.bpbonline.com

Dedicated to

My beloved Parents
Smt. Neelam Sharma
Sh. Gopal Krishna Sharma
&
My wife ***Apoorva***

About the Author

Mayank Bhushan has a teaching experience of more than 15 years. He holds a B.Tech. degree in Computer Science and Engineering and an M.Tech. degree in the same field from Motilal Nehru National Institute of Technology Allahabad, Prayagraj. In addition to having good grades, he is certified to have global experience in Big Data Analytics and Salesforce-Cloud computing. Besides that, he has a certificate in Computer Networking from IIT Kharagpur, especially in the Linux platform. Along with this book, he has written various books tailored for vocational studies.

Throughout his career, the privilege of sharing knowledge through lectures at both private and government engineering colleges has been experienced. The focus during these lectures has been on the subject of Big Data and Hadoop. Commitment to education is deeply held by him and a self-designed course on Big Data and Cloud Computing has been developed. In this course, not only knowledge is imparted by him, but also valuable project ideas and real-time solutions to address any doubts are provided.

He has written many books in this area and is known for making important contributions to international study. With a lot of experience, he has written a number of important research papers that have been read around the world. Aside from his study, Mayank Bhushan has been an inspiration to many scholars, helping them with their theses and being a valuable mentor.

His knowledge and devotion have not only made academic literature better, but they have also had a huge impact on the academic careers of people who want to become researchers. His commitment to advancing knowledge and nurturing the next generation of scholars is evident in his prolific research output and mentorship roles.

About the Reviewer

Jai is an experienced Data Engineer well-versed in large-scale data processing and the design of critical data models for cybersecurity and data privacy products. His expertise in machine learning and generative AI has played a significant role in advancing data governance in large corporations.

With an impressive 18-year career across diverse sectors such as FinTech, HealTech, AdTech, and Media, Jai's current research focuses on Children's privacy-enhancing technologies about social media and gaming platforms. His work is aimed at helping both small and large enterprises that gather personal data from minors to enhance their privacy practices and secure the personal information of millions of children and teenagers who frequently use these platforms globally. Currently, Jai is working with four startups to implement best practices in data cloud, data engineering, and data privacy.

In addition to his role as a data engineer, Jai is an author and speaker, primarily on topics related to data security, privacy, and children's rights and regulations. He has recently contributed as an editor to O'Reilly and as an author and peer reviewer for journals by IEEE and ACM.

Jai is a recognized figure in the Data Privacy field, having authored multiple articles and presented his work at various conferences. His ultimate vision is to nurture a community of privacy experts to solve real-world problems with data-driven solutions.

Acknowledgement

I would like to express my heartfelt gratitude to the numerous individuals who supported me throughout the creation of this book. To all those who provided guidance, engaged in discussions, read and reviewed the content, shared their insightful comments, allowed me to quote their valuable remarks, and contributed to the editing, proofreading, and design process, I extend my sincere thanks.

Writing this book was a collaborative effort, and I am deeply thankful to the Hadoop community, whose continuous efforts have been a significant source of learning and inspiration for me.

Gratitude is a sentiment that we all share when others extend their helpful hands during challenging phases of life, assisting us in achieving our set goals. While it is impossible to individually thank everyone who played a role, I humbly make an effort to express my appreciation to some of them.

First and foremost, I offer my thanks to the Almighty, who invisibly guides us all and has kept us on the right path throughout this journey.

I owe a profound debt of gratitude to the following individuals:

1. Prof. (Dr.) Munesh Chandra Trivedi, National Institute of Technology, Agartala
2. Prof. (Dr.) Shailesh Tiwari, Director, KIET Ghaziabad (UP)
3. Prof. (Dr.) Mayank Pandey, MNNIT Prayagraj
4. Dr. Nitin Shukla, Asst. Prof. Thapar University
5. Dr. Shaswati Banerjea, Asst. Prof. MNNIT Prayagraj
6. Mr. Suraj Deb Barma, Principal, Govt. Polytechnic College, Agartala
7. Dr. Sumit Yadav, Director, Income Tax Dept.
8. Dr. Shailendra Kumar, Asst. Prof. IIT Bhilai
9. Dr. Saurabh Kumar Rajput, Asst. Prof. Madhav Institute Gwalior (MP)

Their unwavering support has been invaluable, and without their assistance, this book would not have been possible.

I extend my heartfelt thanks to all my Gurus, friends, and colleagues who provided unwavering support and kept me motivated throughout the writing process.

I am also grateful to the Publisher and the entire team at BPB Publications, technical reviewer of this book Mr. Jai and especially Mr. Manish Jain, for their efforts in presenting this text in a polished and presentable form.

In closing, I extend my appreciation to everyone who directly or indirectly contributed to the completion of this authentic work.

Preface

Welcome to the world of Big Data! In today's data-driven landscape, the ability to harness and process vast amounts of information has become not just an asset but a necessity for businesses, researchers, and individuals alike. This book, titled **Big Data and Hadoop**: **Fundamentals, tools, and techniques for data-driven success** is your gateway to understanding and mastering the fascinating realm of Big Data.

Chapter 1: Big Data Introduction and Demand – In this opening chapter, we embark on a journey to explore the foundations of Big Data. We will delve into the very concept of Big Data, its significance in today's world, and the growing demand for solutions that can handle its challenges. We will also examine industry examples of how Big Data is being utilized and the myriad of possibilities it presents.

Chapter 2: NoSQL Data Management – This chapter takes us into the realm of NoSQL databases, offering an introduction to these non-relational data stores. We will compare SQL and NoSQL databases, explore the nuances of data consistency in NoSQL, and take a deep dive into the HBase database. Additionally, we will discuss the MapReduce paradigm and key concepts like partitioning and combining.

Chapter 3: MapReduce Technique – This chapter discusses a paradigm widely employed in the realm of distributed computing, that revolutionizes the processing of vast datasets with efficiency and scalability. Developed by Google, this technique serves as a cornerstone in the field of big data analytics. By harnessing the power of parallel processing and fault tolerance, MapReduce enables the seamless analysis of massive datasets across distributed clusters, making it a pivotal tool in addressing the challenges posed by the ever-expanding volume of data in diverse domains.

Chapter 4: Basics of Hadoop – To lay a solid foundation for your journey into Big Data technologies, this chapter introduces you to the basics of Hadoop. We will cover essential topics like data formats, analyzing data with Hadoop, scaling strategies, and the design of the Hadoop Distributed File System (HDFS). Concepts such as data flow, Hadoop I/O, compression, serialization, and Avro file-based data structures will be explored in detail.

Chapter 5: Hadoop Installation – Getting hands-on with Hadoop is crucial, this chapter guides you through the step-by-step process of installing Hadoop on various platforms. Whether you're using Ubuntu or setting up a fully distributed Hadoop system, this chapter provides detailed instructions to help you get started.

Chapter 6: MapReduce Applications – This chapter is all about MapReduce, a fundamental programming model for processing Big Data. We will help you understand the principles behind MapReduce, walk you through the traditional way of using it, and explain the MapReduce workflow.

Chapter 7: Hadoop Related Tools-I: HBase and Cassandra – This chapter introduces you to two important tools in the Big Data ecosystem: HBase and Cassandra. You will discover how to install HBase, explore its conceptual architecture, and gain practical insights into its implementation. We will also delve into HBase's key differences from traditional relational databases. The chapter then shifts focus to Cassandra, explaining its data model, providing examples, and discussing its integration with Hadoop.

Chapter 8: Hadoop Related Tool-II: PigLatin and HiveQL – It introduces two more essential tools: PigLatin and HiveQL. You will learn how to install PigLatin, understand its execution types, and explore the Pig data model. We will also guide you through the development and testing of PigLatin scripts. Next, we delve into Hive, exploring its data types, file formats, and comparing HiveQL with traditional database querying languages.

Chapter 9: Practical and Research-based Topics – This chapter is dedicated to practical and research-based topics in the world of Big Data. You will explore real-world applications like data analysis with X, the use of Bloom Filters in MapReduce, leveraging Amazon Web Services, analyzing documents archived from The New York Times, mobile data mining, and Hadoop diagnostics.

Chapter 10: Spark – As we conclude our journey through Big Data and related technologies, this chapter introduces Apache Spark, a powerful framework for distributed data processing. We will explore its capabilities and understand how it fits into the Big Data landscape, setting the stage for your next adventure in data processing.

This book is designed to provide you with a comprehensive understanding of Big Data technologies, enabling you to tackle real-world challenges and leverage the opportunities presented by the ever-expanding world of data. Whether you are a student, a professional, or a curious explorer, we hope this book equips you with the knowledge and skills to thrive in the era of Big Data.

It is said "To err is human, to forgive divine". In this light I wish that the shortcomings of the book will be forgiven. At the same I am open to any kind of constructive criticisms and suggestions for further improvement.

Happy reading and happy data processing!

Code Bundle and Coloured Images

Please follow the link to download the
Code Bundle and the ***Coloured Images*** of the book:

https://rebrand.ly/jb064ll

The code bundle for the book is also hosted on GitHub at
https://github.com/bpbpublications/Big-Data-and-Hadoop-2nd-Edition.
In case there's an update to the code, it will be updated on the existing GitHub repository.

We have code bundles from our rich catalogue of books and videos available at **https://github.com/bpbpublications**. Check them out!

Errata

We take immense pride in our work at BPB Publications and follow best practices to ensure the accuracy of our content to provide with an indulging reading experience to our subscribers. Our readers are our mirrors, and we use their inputs to reflect and improve upon human errors, if any, that may have occurred during the publishing processes involved. To let us maintain the quality and help us reach out to any readers who might be having difficulties due to any unforeseen errors, please write to us at :

errata@bpbonline.com

Your support, suggestions and feedbacks are highly appreciated by the BPB Publications' Family.

Piracy

If you come across any illegal copies of our works in any form on the internet, we would be grateful if you would provide us with the location address or website name. Please contact us at **business@bpbonline.com** with a link to the material.

If you are interested in becoming an author

If there is a topic that you have expertise in, and you are interested in either writing or contributing to a book, please visit **www.bpbonline.com**. We have worked with thousands of developers and tech professionals, just like you, to help them share their insights with the global tech community. You can make a general application, apply for a specific hot topic that we are recruiting an author for, or submit your own idea.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions. We at BPB can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about BPB, please visit **www.bpbonline.com**.

Join our book's Discord space

Join the book's Discord Workspace for Latest updates, Offers, Tech happenings around the world, New Release and Sessions with the Authors:

https://discord.bpbonline.com

Table of Contents

CHAPTER 1

Big Data Introduction and Demand

…data is useless without the skill to analyze it

– Jeanne Harris

Taking a hunch you have about the world and pursuing it in a structural, mathematical way to understand something new about the world

– Hilary Mason

Introduction

In today's scenario, we all are surrounded by a bulk of data. We, as humans, are also an example of big data as we are surrounded by devices that generate data every minute.

I spend most of my time assuming the world is not ready for the technology revolution that will be happening to them soon

– Eric Schmidt

Executive Chairman Google

As a matter of fact, if we compare the present situation to the past scenario, we can find that we are creating as much information in just two days as we did up to 2003, which means we are creating five exabytes of data every two days.

The real problem is user-generated data that they are producing continuously. At the time of data analysis, we have challenges in storing and analyzing those data.

Structure

The following are the topics to be discussed in this chapter:

- Big data
- Characteristics of big data
- Why big data is required?
- Introduction of Hadoop
- Convergence of key trends
- Unstructured data
- Industry examples of big data
- Usages of big data

Objectives

This chapter provides an introduction to big data and its need in the current scenario. Every device can generate data in each second, creating problems with its storage and, after that, its analysis for further usage. These kinds of issues will be addressed in this chapter. Here, we also collect introductory knowledge of Hadoop, which is used for processing and storing a large amount of data. It also helps in ending the confusion of different types of data generated by devices. Data has become a critical resource for businesses and organizations of all sizes in recent years. However, with the increasing volume, velocity, and variety of data, traditional data processing techniques and technologies have become insufficient to handle the data deluge. This is where the concept of big data comes into play. Big data refers to large and complex data sets that traditional data processing applications cannot handle effectively. Big data technologies enable the storage, processing, and analysis of these massive data sets to extract valuable insights and create new opportunities. Hadoop, one of the most popular big data technologies, is an open-source framework that facilitates the distributed processing of large data sets across clusters of commodity hardware. With Hadoop, organizations can analyze vast amounts of data in a cost-effective, scalable, and flexible manner, unlocking new insights and business opportunities.

The real issue is user-generated content

— Schmidt

Big data

Big data refers to the vast and complex volumes of structured and unstructured data that exceed the processing capabilities of traditional data management systems. This data is characterized by its sheer size, velocity, variety, and, more recently, its value. Big data typically encompasses information from diverse sources, including social media, sensors, machine-generated data, and traditional databases. It presents unique challenges and opportunities as organizations seek to capture, store, analyze, and extract meaningful insights from this data to make data-driven decisions and gain a competitive edge.

Mostly, it helps Google to analyze the data and sell data analytics to companies who require it. We are producing data only through mobile as we already logged in when we bought the system:

- **Map**: It collects data on our traveling.
- **App**: It gathers information about our daily life activities and records activities in which we are most involved.
- **E-commerce sites**: It collect information about our requirement and shows whatever we are supposed to buy.
- **E-mails**: It is important to note that while Google scans e-mail content, it claims not to read e-mails for personal information or sensitive data. However, this practice has raised privacy concerns, and Google has faced legal challenges related to e-mail scanning in the past.

Indeed, over the past few decades, advances in technology, such as remote sensing, **Geographic Information Systems** (**GIS**), and **Global Positioning Systems** (**GPS**), have revolutionized the way we understand and analyze the distribution of human populations across the world. For that scenario, we need to map those population data to a meaningful survey that is performed by big companies. As a result, spatially careful changes across scales of days, weeks, or months, or maybe year to year, area units are tough to assess and limit the applying of human population maps in things within which timely data is needed, such as disasters, conflicts, or epidemics. Information being collected daily by mobile network suppliers across the planet, the prospect of having the ability to map up-to-date and ever-changing human population distributions over comparatively short intervals exists, paving the approach for brand new applications and a close to period of time understanding of patterns and processes in human science.

Some of the facts related to exponential data production are as follows:

- Currently, over 2 billion people worldwide are connected to the internet, and over 5 billion individuals own mobile phones. By 2030, 150 billion devices are expected to be connected to the internet. At this point, predicted data production will be 44 times greater than that in 2009.

- To provide a rough idea, by 2020, global internet traffic was estimated to be measured in exabytes per month (1 exabyte = 1 billion gigabytes). Cisco's **Visual Networking Index** (**VNI**) forecasted that monthly global IP traffic would reach 260 exabytes per month by 2020. Facebook alone stores, accesses, and analyzes 30 +PB of user-generated data.
- In 2018, Google was processing 20,000 TB+ of data daily.
- Walmart processes over 1 million customer transactions, thus generating data in excess of 2.5 PB as an estimate.
- More than 5 billion people worldwide call, text, tweet, and browse on mobile devices.
- The number of e-mail accounts created worldwide is expected to increase from 3.3 billion in 2012 to over 4.3 billion by late 2016 at an average annual rate of 6% over the four years. In 2012, a total of 89 billion e-mails were sent and received daily, and this value is expected to increase at an average annual rate of 13% over the next four years to exceed 143 billion by the end of 2016; by this rate, it is expected to reach 500 billion + accounts by 2030.
- Boston.com reported that in 2021, approximately 1507 billion e-mails were sent daily. Currently, an e-mail is sent every 3.5 × 10−12 seconds. Thus, the volume of data increases per second as a result of rapid data generation. At this rate, an imaginary figure can be taken for us, which would be beyond our thinking.
- By 2030, enterprise data is expected to total 400 ZB, as per International Data Corporation.
- The New York Stock Exchange generates about one terabyte of data for new trade.

Based on this estimation, **business-to-consumer** (**B2C**) and internet-**business-to-business** (**B2B**) transactions will amount to 450 billion per day.

All are the facts that are sufficient to prove that the world is generating large amounts of data that are not only structured. That case leads to the innovation or thinking that can provide solutions for solving those issues.

Big data is the one which is used to deal with the current scenario. Big data is the concept for handling unstructured and structured data other than the traditional way. *Table 1.1* shows the flow of data from bottom to top. In today's scenario, any type of data is possible to store and process:

Number	Symbol	In Binary
Bit	b	1 bit
Nibble/Nybble	nibble	4 bits
Byte	B	8 bits
KiloByte	KB	1024 B

Number	Symbol	In Binary
MegaByte	MB	1024 KB
GigaByte	GB	1024 MB
TeraByte	TB	1024 GB
PetaByte	PB	1024 TB
HexaByte	HB	1024 PB
ZettaByte	ZB	1024 HB
YottaByte	YB	1024 ZB

***Table 1.1**: Introduction of data*

Characteristics of big data

Big data is data that gives the capacity to think beyond the traditional database system. As data that can be used in big data may be structured or unstructured data with a huge amount of capacity, it requires fast movement, fast storage, and fast processing other than conventional database techniques. These requirements of processing data demand tools that can perform functions fast and meaningful that are difficult by any traditional database tools. Properties of big data provide a next-generation way to handle situations and provide an easy and efficient way to handle data for an organization. As we all see around, there are a lot of devices that are continuously generating data with exponential increments, and all human beings are digging themselves into social networking. These types of unstructured and structured data create challenges in storing and processing data.

Every day, the world creates 2.5 quintillion bytes of data that are 90% of the data in the world today that was created in the last two years alone, and sources of those data from sensors, videos, post, Twitter, WhatsApp, Facebook, and many more digital sites of many users.

The following is the comparison between big data and traditional techniques of databases:

Traditional Database "Schema on Write"	Big data "Schema on Read"
There is a need to create a schema before data is loaded into the database.	Data is firstly copied to HDFS; after that, transformation is needed.
The load operator performs explicitly to transform the database.	Only required columns are extracted to perform operations.
It uses the scale-in property to the enhancement of data on the server side.	There is the use of scale-out property to enrich data at any time.

***Table 1.2**: Traditional database and big data schema*