Big Data and Analytics

The key concepts and practical applications of big data analytics

Dr. Jugnesh Kumar

Dr. Anubhav Kumar

Dr. Rinku Kumar

www.bpbonline.com

First Edition 2024

Copyright © BPB Publications, India

ISBN: 978-93-55516-176

All Rights Reserved. No part of this publication may be reproduced, distributed or transmitted in any form or by any means or stored in a database or retrieval system, without the prior written permission of the publisher with the exception to the program listings which may be entered, stored and executed in a computer system, but they can not be reproduced by the means of publication, photocopy, recording, or by any electronic and mechanical means.

LIMITS OF LIABILITY AND DISCLAIMER OF WARRANTY

The information contained in this book is true to correct and the best of author's and publisher's knowledge. The author has made every effort to ensure the accuracy of these publications, but publisher cannot be held responsible for any loss or damage arising from any information in this book.

All trademarks referred to in the book are acknowledged as properties of their respective owners but BPB Publications cannot guarantee the accuracy of this information.

To View Complete
BPB Publications Catalogue
Scan the QR Code:

www.bpbonline.com

Dedicated to

My beloved wife, ***Varsha*** *and*
my son ***Nikhil Kumar***
-Dr. Jugnesh Kumar

My beloved wife, ***Rekha*** *and*
my son ***Krishna Sharma***
-Dr. Anubhav Kumar

My father, ***Mr. Omprakash Singh*** *and*
my mother, ***Mrs. Bala Devi***
-Dr. Rinku Kumar

About the Authors

- **Dr. Jugnesh Kumar** is working as a professor in the Computer Science and Engineering department at Echelon Institute of Technology, Faridabad. He also has more than 18 years of teaching experience. He holds an M.Tech. and PhD in computer science and engineering. He had successfully organized international and national-level conferences. He has published more than 35 articles in Scopus-indexed, SCI-high-impact journals and international conferences.

- **Dr. Anubhav Kumar** is working as a Professor at SCSE Galgotias University, Greater Noida. He also has more than 17 years of teaching experience. He holds an M.Tech and PhD in computer science and engineering. His doctoral work in computer science and engineering focused on NLP. He has published more than 65 Articles in Scopus Indexed/SCI/High Impact Journals and International Conferences.

- **Dr. Rinku Kumar** is working as an Assistant Professor in the Computer Science and Engineering Department at St. Andrews Institute of Technology and Management, Gurugram. He also has more than ten years of working experience. He has published more than 12 Articles in Scopus and High Impact Journals and Conferences.

About the Reviewer

Gaurav Gangwar is a Cloud Data Engineer, Certified in Azure and AWS Cloud Data Platforms. He is currently focused on and working with cloud data engineering platform solutions and tools, including Azure Databricks, Snowflake, Azure Synapse, AWS Redshift, and Big Query. He also focuses on other ETL/ELT tools like Azure Data Factory, DBT, Matillion, AWS Glue, and Airflow. Some reporting tools like PowerBI and ThoughtSpot are passionate about automating the data and AI jobs, avoiding doing anything that resembles manual work. He is an active reader of books related to data and AI/ML and a technical reviewer for various books about big data analytics, data engineering, data warehouse/database, and Azure Databricks/data engineering.

Acknowledgements

- I want to express my deepest gratitude to my family and friends, especially my friends Dr. Anubhav Kumar and Dr. Rinku Kumar, for their unwavering support and encouragement throughout this book's writing.

  I am also grateful to BPB Publications for their guidance and expertise in bringing this book to fruition. It was a long journey of revising this book, with valuable participation and collaboration of reviewers, technical experts, and editors.

  I would also like to acknowledge the valuable contributions of my colleagues and co-workers during many years working in the tech industry, who have taught me so much and provided valuable feedback on my work.

  Finally, I would like to thank all the readers who have taken an interest in my book and for their support in making it a reality. Your encouragement has been invaluable.

  –Dr. Jugnesh Kumar

- I want to express my deepest gratitude to my family and friends for their unwavering support and encouragement throughout this book's writing, especially my friends Dr. Jugnesh Kumar and Dr. Rinku Kumar, my parents Mr. Avnish Kumar and Smt. Kusum Lata for their love and support.

  I am also grateful to BPB Publications for their guidance and expertise in bringing this book to fruition. It was a long journey of revising this book, with valuable participation and collaboration of reviewers, technical experts, and editors.

  I would also like to acknowledge the valuable contributions of my colleagues and co-workers during many years working in the tech industry, who have taught me so much and provided valuable feedback on my work.

Finally, I would like to thank all the readers who have taken an interest in my book and for their support in making it a reality. Your encouragement has been invaluable.

–Dr. Anubhav Kumar

- I want to express my deepest gratitude to my family and friends for their unwavering support and encouragement throughout this book's writing, especially my friend Dr. Jugnesh Kumar and family.

  I am also grateful to BPB Publications for their guidance and expertise in bringing this book to fruition. It was a long journey of revising this book, with valuable participation and collaboration of reviewers, technical experts, and editors.

  I would also like to acknowledge the valuable contributions of my colleagues and co-workers during many years working in the tech industry, who have taught me so much and provided valuable feedback on my work.

  Finally, I would like to thank all the readers who have taken an interest in my book and for their support in making it a reality. Your encouragement has been invaluable.

  –Dr. Rinku Kumar

Preface

Welcome to the dynamic world of Big Data and Analytics. In an era where information is revered as the new gold, the ability to harness, manage, and extract insights from vast repositories of data has become indispensable.

This book is an exploration into the realms of Big Data and Analytics, designed to be a comprehensive guide for both beginners and seasoned professionals. It delves into the fundamental principles, methodologies, and advanced techniques essential for understanding, building, and leveraging robust data infrastructure and extracting valuable knowledge from it.

Big Data and Analytics introduces the foundational concepts behind the creation and management of centralized repositories, offering a blueprint for designing efficient data storage systems. Big Data and Analytics, on the other hand, navigate the terrain of extracting meaningful patterns, trends, and associations from extensive datasets.

Chapter 1: Introduction to Big Data - In this chapter, we learned that big data has fundamentally changed how we approach data analysis and decision-making in the digital age. Massive amounts of data are being produced as a result of the spread of digital technologies and connected devices, including data from social media, sensors, transactions, and machine-generated data

Chapter 2: Big Data Analytics - In this chapter, we delved into big data analytics, a pivotal discipline in the modern data-driven world, enabling organizations to extract valuable insights from vast and complex datasets. It involves the systematic analysis of data using various techniques and tools to uncover patterns, trends, and relationships that can drive informed decision-making. Classification of analytics categorizes analytics into descriptive, diagnostic, predictive, and prescriptive, each serving specific purposes in data analysis.

Chapter 3: Introduction of NoSQL - This chapter tells us that big data analytics has become a transformative force in the digital era, revolutionizing how we analyze and draw conclusions from enormous and complex datasets. Organizations can use various analytical techniques depending on the needs

of their respective businesses thanks to the classification of analytics into descriptive, diagnostic, predictive, and prescriptive categories

Chapter 4: Introduction to Hadoop - Hadoop's key characteristics, including distributed processing and storage, fault tolerance, and scalability, have made it the perfect solution for managing large and varied datasets. Hadoop has emerged as a top option for many data-driven applications due to benefits like affordability, compatibility with commodity hardware, and the capacity to process unstructured data. Its strong ecosystem, which includes different Hadoop distributions and Apache Pig for data processing, further strengthens its capabilities.

Chapter 5: Map Reduce - The MapReduce method involves a number of steps, such as splitting, mapping, shuffling, sorting, and reducing, which significantly increases the speed and scalability of data processing. Combiners are a useful addition to the MapReduce workflow because they reduce data transfer and increase productivity. Combiners, however, might not always be appropriate because they have the potential to make the code more complex and demand more resources.

Chapter 6: Introduction to MongoDB - The evolution of MongoDB into a potent document-oriented database, with features like schema flexibility, horizontal scalability, and support for various data types, can be seen in the database's history and development. Data manipulation is made simple by the flexible query language and CRUD operations, and data retrieval effectiveness is improved by features like count, sort, limit, and skip. Additionally, complex data processing and analysis tasks are made possible by MongoDB's strong MapReduce and aggregation capabilities.

Code Bundle and Coloured Images

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

https://rebrand.ly/9ukhu2d

The code bundle for the book is also hosted on GitHub at
https://github.com/bpbpublications/Big-Data-and-Analytics.
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.

Did you know that BPB offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.bpbonline.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at :

business@bpbonline.com for more details.

At **www.bpbonline.com**, you can also read a collection of free technical articles, sign up for a range of free newsletters, and receive exclusive discounts and offers on BPB books and eBooks.

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

1. Introduction to Big Data 1

Introduction 1

Structure 2

Diverse facets of big data 2

Digital data and its types 3

Characteristics of big data 4

Types of big data 5

Structured data 6

Unstructured data 8

Semi-structured data 9

Evolution of big data 10

Applications and challenges of big data 12

Challenges of big data in retail 15

Manufacturing and supply chain 16

Transportation and logistics 18

Challenges for the transportation and logistics industry 19

Energy and utilities 20

Challenges of big data in utility and energy 21

Telecommunications 23

Challenges of big data in telecommunications 25

Government and public sector 26

Challenges of big data in the public sector and government 28

Marketing and advertising 29

Challenges of big data in marketing and advertising 31

Sports analytics 33

Challenges of big data in sports analytics 34

3 Vs of big data 36

Example of 3Vs of big data 39

Non-definitional traits of big data 39
Big data workflow management 41
Business intelligence versus big data 44
Definition and scope *46*
Data volume and variety *46*
Data processing and analysis *46*
Time sensitivity *47*
Decision-making scope *47*
The future of business intelligence is big data analytics *48*
Data science process steps 48
Foundations for big data systems and programming 51
Distributed systems *52*
Data storage and management *52*
Data processing and analytics *52*
Programming languages and tools *53*
Data streaming and real-time processing *53*
Cloud computing *53*
Data visualization and reporting *53*
Distributed file systems 54
Definition and purpose *54*
Characteristics and key features *55*
Architecture and data flow *55*
Use cases and benefits *56*
Data warehouse and Hadoop environment 56
Data warehouse *56*
Hadoop *56*
Difference between data warehouse and Hadoop environment *57*
Coexistence 59
Conclusion 61
Questions 61

2. Big Data Analytics 63
Introduction 63
Structure 63
Classification of analytics 64
Data science 67
Difference between data science and big data 69
Data characteristics 69
Tools and technologies 69
Terminologies in big data 70
CAP theorem 72
Example of the CAP theorem in big data 73
BASE concept 74
Conclusion 75
Questions 76

3. Introduction of NoSQL 77
Introduction 77
Structure 78
Introduction to NoSQL 78
NoSQL databases creation history 79
NoSQL categories 81
Document databases 81
Key-value stores 83
Column-family stores 85
NoSQL advantages 86
Graph databases 88
Graph database in NoSQL 88
NewSQL 89
SQL versus NoSQL versus NewSQL 91
Conclusion 93
Questions 93

4. Introduction to Hadoop 95
Introduction 95
Structure 96
History of Hadoop 96
Features of Hadoop 98
Distributed storage 98
Scalability 100
Fault tolerance 101
Parallel processing 102
Flexibility 103
Ecosystem 105
Cost-effectiveness 106
Advantages of Hadoop 108
Versions of Hadoop 109
Hadoop ecosystems 111
Hadoop distributed file system 112
MapReduce 113
Apache Hive 115
Apache Pig 117
Apache Spark 118
Apache HBase 119
Apache Kafka 121
Apache Sqoop 122
Apache Zeppelin 123
Apache Oozie 124
Hadoop distributions 126
HQL versus SQL 128
Relational database management system versus Hadoop 129
Relational database management system 130
Hadoop 130
Hadoop architecture 132

Conclusion....134
Questions....135

5. Map Reduce....137
Introduction....137
Structure....138
MapReduce architecture....138
Steps in MapReduce....140
Working of MapReduce....141
Example of MapReduce....143
Limitations of MapReduce....145
Mapper....145
Working of Mapper....147
Reducer....148
Combiner....150
Working of MapReduce combiner....151
Advantages of combiners....154
Disadvantages of combiners....155
Partitioner....156
Need of MapReduce partitioner....157
Advantages and disadvantages of partitioner....157
Searching....158
Sorting....160
The sorting algorithm....162
Compression....163
Hadoop 2....165
Architecture of Hadoop 2....166
Hadoop YARN architecture....168
Interacting with Hadoop ecosystems....169
Conclusion....171
Questions....171

6. Introduction to MongoDB 173
Introduction 173
Structure 173
Not only SQL databases 174
Mongo DB 175
Advantages of MongoDB 176
Data modeling in MongoDB 177
History of MongoDB 177
MongoDB key features 178
Document model 179
Scalability 180
High performance 182
Replication and high availability 183
Replication 183
High availability 184
Flexible querying 184
Integration with programming languages 186
Data types 187
MongoDB query language 191
Create, read, update, and delete operations 193
Arrays 194
Functions 195
Count 195
Sort 196
Limit 197
Skip 198
Aggregate 198
MapReduce 199
Cursors in MongoDB 200
Indexes in MongoDB 201
mongoimport and mongoexport 202

mongoimport *202*

mongoexport *203*

Conclusion 204

Questions 205

Index 207-214

Chapter 1
Introduction to Big Data

Introduction

The amount of data produced by humanity is increasing exponentially because of the rapid development of technology, the proliferation of devices, and the widespread use of social networking sites. To put things in perspective, humankind produced 5 billion gigabytes of data between the beginning of time and 2003, which could cover an entire football pitch if represented as physical discs.

Amazingly, however, the same amount of data was generated every ten minutes in 2013, up from every two days in 2011, and it has continued to increase significantly. Even though this vast amount of information has many valuable insights and the potential to be helpful when processed, it is frequently underutilized and ignored. The enormous volume of data being produced at an unprecedented rate worldwide is called **Big Data**. Both structured and unstructured data types are possible. Businesses heavily rely on data in today's knowledge-based economy to fuel their success. So, it becomes crucial and enormously rewarding to make sense of this data, identify patterns, and expose hidden connections within this vast ocean of information. The urgent need is to turn big data into easily usable, actionable

business intelligence for enterprises. Businesses of all sizes, locations, market shares, and customer segments can develop successful strategies by accessing and analyzing high-quality data. This is where Hadoop, the go-to platform for processing enormous volumes of data, comes into play.

Structure

In this chapter, we will discuss the following topics:

- Diverse facets of big data
- Digital data and its types
- Characteristics of big data
- Types of big data
- Evolution of big data
- Applications and challenges of big data
- 3Vs of big data
- Non-definitional traits of big data
- Big data work flow management
- Business intelligence versus big data
- Data science process steps
- Foundations for big data systems and programming
- Distributed filesystems
- Data warehouse and Hadoop environment
- Coexistence

Diverse facets of big data

Alternatively, we can define big data as a collection of sizable datasets processed faster than traditional computing techniques. It has developed into a broad discipline that includes various tools, techniques, and frameworks, not just a single technique or tool. The data consists of the enormous amount produced by different devices and applications. The following industries fall under the umbrella of big data, as shown in *Table 1.1*:

Big data involves	Description
Black box data	It is a crucial component of helicopters, aircraft, and jets, among others, recording audio from the flight crew, audio from microphones, and audio from earphones, as well as information about the aircraft's performance.
Social media data	Social media platforms like Facebook and Twitter store vast amounts of information and opinions shared by millions of users worldwide.
Stock exchange data	Stock exchange data contains valuable insights into customers' buying and selling decisions regarding various company shares.
Power grid data	Power grid data details the energy consumption of specific nodes about a central station.
Transport data	Transport data encompasses vehicle models, capacities, distances, and availability.
Search engine data	Search engines gather extensive data from diverse databases.

***Table 1.1**: Shows the involvement of big data in various organizations*

Digital data and its types

Digital data can be classified into several types based on their characteristics and formats. Here are some common types of digital data given below:

- **Textual data:** This type includes written or typed text, such as documents, emails, webpages, and social media posts. Textual data is typically represented as a sequence of characters.
- **Numeric data:** Numeric data consists of numbers and mathematical values. It can be discrete (whole numbers) or continuous (decimal numbers). Examples of numeric data include measurements, financial data, and statistical records.
- **Image data:** Image data represents visual information through pictures or graphical content. It consists of a grid of pixels, where each pixel contains color or grayscale information. Image data is commonly used in photography, digital art, and computer vision applications.
- **Audio data:** Audio data represents sound or audio signals. It can be in the form of speech, music, or other audio recordings. Audio data

is typically stored as waveform samples, capturing variations in air pressure over time.

- **Video data:** Video data consists of an order of images (frames) presented in rapid succession. It combines image and audio data to represent moving visual content. Video data is commonly used in movies, television, surveillance systems, and video streaming platforms.
- **Geospatial data:** Geospatial data refers to data with geographical or spatial information. It includes coordinates, maps, satellite imagery, and location-based data. Geospatial data is widely used in navigation, urban planning, mapping, and environmental analysis.
- **Time series data:** Time series data capture measurements or observations taken at different points in time. It includes data points recorded at regular intervals, such as stock prices, weather data, sensor readings, and device logs.
- **Structured data:** This type of data follows a predefined format and schema. It is organized in a tabular or relational form, with well-defined rows and columns. Structured data is stored in databases and spreadsheets and can be easily queried and analyzed.
- **Unstructured data:** Unstructured data refers to data that does not have a predefined format or structure. It includes free-form text, multimedia content, social media posts, emails, and documents. Unstructured data requires advanced techniques like machine learning and natural language processing to extract meaningful insights.
- **Metadata:** Metadata provides descriptive information about other types of data. It includes file names, creation dates, author information, data sources, and formats. Metadata helps in organizing, managing, and understanding other data types.

Characteristics of big data

Data can possess several characteristics that impact its management, analysis, and interpretation. Some important features of data include:

- **Volume:** Volume denotes the amount or size of data. It can range from small-scale data sets to massive volumes of data from various sources.

- **Velocity:** Velocity denotes the speed at which data is created, processed, collected, and analyzed. Real-time data requires fast processing capabilities to extract timely insights.
- **Variety:** The diversity of data types and formats is called variety. Text, images, audio, video, and other data types can exist in structured, unstructured, or semi-structured forms.
- **Semi-structured**: The data shows some organization but lacks a strict structure, in contrast to structured data, which is prearranged in a tabular format with a predetermined schema. A certain level of hierarchy or relationship is possible because this kind of data frequently contains elements like tags, keys, or attributes.
- **Veracity:** Veracity refers to the quality and reliability of data. Data may contain errors, inconsistencies, or inaccuracies that must be addressed to ensure data veracity.
- **Value:** Value refers to data's usefulness, relevance, and potential insights. Extracting value from data involves analysis, interpretation, and decision-making based on the obtained insights.
- **Variability:** Variability refers to the dynamic landscape of data. Data can exhibit variations in volume, velocity, and variety over time. Handling data variability requires adaptability and flexibility in data.

Types of big data

Big data can be classified into three main types based on the nature of the data and its characteristics. These types are mentioned in *Table1.2*:

Based on	Structured data	Unstructured data
Technology	It is built on a relational database.	It is based on binary along with character data.
Flexibility	Less flexible and schema-dependent, structured data.	Schema is not present, making it more flexible.
Scalability	Scaling a database schema is challenging.	It can scale up more.
Robustness	It has great strength.	It has less strength.