Ümit Demirbaga
Gagangeet Singh Aujla
Anish Jindal
Oğuzhan Kalyon

Big Data Analytics

Theory, Techniques, Platforms, and Applications

Springer

Big Data Analytics

Ümit Demirbaga · Gagangeet Singh Aujla ·
Anish Jindal · Oğuzhan Kalyon

Big Data Analytics

Theory, Techniques, Platforms, and Applications

Springer

Ümit Demirbaga
Department of Medicine
University of Cambridge
Cambridge, UK

Department of Computer Engineering
Faculty of Engineering, Architecture, and Design
Bartin University
Bartin, Türkiye

Anish Jindal
Department of Computer Science
Durham University
Durham, UK

Gagangeet Singh Aujla
Department of Computer Science
Durham University
Durham, UK

Oğuzhan Kalyon
Faculty of Medical Sciences
Newcastle University
Newcastle Upon Tyne, UK

ISBN 978-3-031-55638-8 ISBN 978-3-031-55639-5 (eBook)
https://doi.org/10.1007/978-3-031-55639-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

To my beloved wife, Dr. Kübra Kırca Demirbaga, and my cherished son, Asaf Aziz Demirbaga. They are not just coauthors of this book, but the coauthors of my life.

Dr. Ümit Demirbaga

To my daughter, Imanat Kaur Aujla, my son, Avitaj Singh Aujla, my love, Navneet Mann Aujla, my parents (Surjit Kaur and Prof. Kulwant Singh Aujla), and my in-laws (Kuljit Kaur and Datar Singh Mann).

Dr. Gagangeet Singh Aujla

To my late grandparents, lovely parents (Dr. Ashok Jindal and Dr. Anita Jindal) and beloved wife, Ankita.

Dr. Anish Jindal

To my mom, Zeynep, my dad, Mehmet, my sister, Arzu, and my love, Feyza Gül.

Oğuzhan Kalyon

Foreword

The world has changed in the information space over the last two decades due to three main factors: firstly there has been the routine deployment of high content measurement devices, from personal photos to satellite images to DNA sequencing to social media feeds. Secondly we have had the network and disk to store information at scale, often storing information that we don't know the value of. Thirdly increasingly sophisticated computational techniques, given labels such as "data science," "machine learning," and "AI," have been developed. All these phenomena can be collected under the heading of "Big Data."

This book provides an overview of these trends and the practical ways to handle this. Much of the complexity of dealing with data at this scale is about engineering—the practicalities about whether one can manage data flows robustly and cheaply—as well as the more statistically and algorithmically sophisticated analysis schemes. Here the reader can learn about both, and see this from a generic perspective of how to transmit, store, and organise data through to more subject-specific topics such as an introduction to Big Data approaches in bioinformatics. The book is designed for a broad audience, applicable to seasoned computational and data scientists as well as people at the start of their careers. The authors have provided both overviews and practical examples.

The world has already been changed by the advent of big data, and there is no doubt this will be part of this century. I recommend this book to everyone who wants to be part of this future.

December 2023

Prof. Ewan Birney, CBE, FRS, FMedSci
Deputy Director General
of the European Molecular Biology Laboratory (EMBL)
Director of European Bioinformatics Institute (EMBL-EBI)
Nonexecutive Director of Genomics England
Chair of the Global Alliance for Genomics and Health
Honorary Professor of Bioinformatics
University of Cambridge
Cambridge, UK

Preface

The deep significance of big data analytics is a beacon that helps enterprises navigate the challenges of making data-driven decisions in the ever-changing and quickly evolving field of information technology. Businesses now have a strategic imperative: to harness, evaluate, and draw useful insights from an unprecedented influx of data from varied sources. This book opens up as a thorough manual by exploring the fundamental ideas and methods that make up the formidable core of big data analytics and shedding light on its transformative potential. The importance of big data analytics stems from its ability to handle enormous amounts of data and its potential to reveal hidden connections, patterns, and trends that are missed by more conventional analytical techniques. Organisations that fully utilise big data have a competitive advantage in a world where data is being generated at a rate never seen before. Effective big data analytics has far-reaching consequences for various industries, from improving operational efficiency and resource allocation to facilitating data-driven innovation.

The revolutionary potential of big data analytics has become a key component of strategic efforts for organisations globally in the ever-expanding digital landscape. Opportunities and difficulties arise from the sheer amount and diversity of data collected as the globe grows more linked. Big data analytics emerges as the compass that leads decision-makers through the complexities of this data-rich terrain. Big data analytics stimulates innovation and advances artificial intelligence, machine learning, and predictive modelling in addition to its function in revealing insights. This book acts as a bridge to this ever-changing world, guiding readers through the fundamental ideas and innovative uses that shape the field of big data analytics.

This collaborative endeavour, authored by experts in the field, serves as your comprehensive guide by offering a multifaceted exploration of big data analytics. Tailored to the reader's unique role, expertise, and aspirations in the dynamic landscape of information technology, each chapter is crafted by a specialised contributor, which provides in-depth insights and expertise on specific aspects of the subject matter. This book is divided into 12 chapters. Chapter 1 establishes the groundwork by elucidating the essential properties of big data analytics, which explores the diverse range of techniques and provides an overview of the subsequent chapters. Chapter 2 provides a comprehensive guide to understanding big

data, unravelling its definition, characteristics, the renowned 5 Vs, challenges, and future directions. Transitioning seamlessly, Chap. 3 introduces the realm of big data analytics that delves into the pivotal role big data analytics plays in risk management, cost reduction, data-driven decision-making, and product development. As the narrative unfolds, Chap. 4 extends the discussion to the intersection of big data and cloud computing by offering a historical backdrop and elucidating cloud computing units. Chapter 5 immerses the reader in the expansive landscape of big data analytics platforms. The chapter dissects the components of systems by delving into the main characteristics and desired properties. It also provides practical case studies through real-world applications. Navigating further, Chap. 6 addresses the critical aspect of big data storage solutions, which explores traditional systems, such as relational databases and data warehouses, and presents contemporary solutions and cloud storage. Chapter 7 focuses on the pivotal realm of big data monitoring. Understanding the nuances of proactive and reactive monitoring, readers explore the monitoring components. The chapter concludes by acquainting readers with a spectrum of monitoring tools. As the reader ventures into Chap. 8, the book unfolds the intricate world of debugging big data systems. Acknowledging real-world performance problems, the chapter outlines systematic debugging steps and addresses common issues by providing a comprehensive guide to root cause analysis. Closing the narrative, Chap. 9 explores the synergy between machine learning and big data analytics. Diving into supervised and unsupervised machine learning, readers gain insights into challenges, preprocessing techniques, and popular algorithms. The chapter paints a holistic picture of the role machine learning plays in extracting meaningful patterns and predictions from massive datasets. Exploring the diverse applications of big data analytics across various sectors unveils a tapestry of innovation and strategic advancements. Chapter 10 delves into real-world case studies and applications that underscore the transformative impact of big data analytics, where each case study unveils the intricate interplay between big data analytics and sector-specific challenges. The examination of big data analytics for smart grids is expanded in Chap. 11, which also explains the intricacies of smart grids and illustrates how big data analytics is essential to improving their functionality. Finally, Chap. 12 delves into bioinformatics, which explores the intersection of big data and genomics. From understanding the challenges posed by big data in bioinformatics to examining frameworks for handling big genomic data, this section provides a holistic view of big data analytics in bioinformatics and a detailed case study in genomic medicine.

Whether you are an undergraduate student embarking on your academic journey, a master's student exploring advanced concepts, or a Ph.D. candidate or postdoctoral researcher delving into the nuanced intersections of big data analytics, this book serves as your comprehensive guide that offers a multifaceted exploration of big data analytics that is tailored to your unique role, expertise, and aspirations in the dynamic landscape of information technology. As a lecturer or educator, you will find valuable resources to support your teaching endeavours by incorporating real-world case studies and practical insights into your curriculum.

Moreover, this book offers a practical handbook for software engineers navigating the evolving IT landscape, which provides in-depth insights into platforms, storage solutions, and monitoring tools. Whether steering established companies or launching startups, business professionals will discover strategic guidance in risk management, cost reduction, and product development. The diverse range of applications is also explored, spanning government, health care, entertainment, banking, retail, and energy, which ensures relevance across industries. No matter your role or expertise, this book equips you with the knowledge to harness the transformative power of big data analytics for innovation, efficiency, and strategic decision-making.

Cambridge, UK — Ümit Demirbaga
Durham, UK — Gagangeet Singh Aujla
Durham, UK — Anish Jindal
Newcastle Upon Tyne, UK — Oğuzhan Kalyon
December 2023

Acknowledgements Umit extends his heartfelt gratitude to the Republic of Turkey and the Turkish Ministry of National Education for their unwavering financial and emotional support, which was pivotal in facilitating the successful stages of his academic journey.

Contents

List of Figures

Introduction

1

The world is being overrun by an unprecedented amount of data in the twenty-first century. This data comes from various sources, ranging from the subtle clicks of a mouse to the complicated data streams obtained via satellite technologies. Big data analytics is a discipline positioned to unearth priceless insights, spur innovation, and revolutionise decision-making paradigms due to the exponential growth of data. This book thoroughly introduces the complex field of big data analytics.

Big data analytics is fundamentally distinguished by its innate ability to uncover hidden possibilities inside the enormous data reservoirs inherent to our digital era [1]. It goes beyond merely managing huge datasets; instead, it explores the worlds of data interpretation, pattern identification, and predictive analysis, all of which lead to the support of crucial judgements. Big data analytics has permeated numerous industries, from the healthcare sector's pursuit of better diagnostics to the finance sector's search for data-driven strategies, offering competitive advantages, operational efficiency, and concrete benefits to various stakeholders.

This journey embarks upon an exploration of the quintessential attributes that delineate big data analytics. It reveals the complex methods and necessary equipment data scientists employ to decipher complex information. The journey continues into machine learning, Artificial Intelligence (AI), and data mining, where models and algorithms are the keys to revealing significant discoveries. A detailed summary of this book's contents is essential to our journey since it aims to give the reader a thorough understanding of big data analytics so they may successfully navigate the complex terrain of this diverse field.

Thus, this odyssey traversing the annals of big data analytics extends an invitation to all, irrespective of their status as seasoned data professionals, driven by the pursuit of honing their expertise or enthusiastic novices, harbouring an intrinsic curiosity regarding the uncharted territories of data exploration. In unison, this sojourn pledges not only to unveil the enigmatic intricacies inherent in the realm of big data analytics but also to scrutinise its expansive potential meticulously. In doing so, it endeavours to endow its passengers with the comprehensive skill set and profound knowledge

Ü. Demirbaga et al., *Big Data Analytics*,
https://doi.org/10.1007/978-3-031-55639-5_1

for harnessing the boundless power latent within the ever-pervasive data domain in our contemporary, data-centric milieu.

1.1 Essential Big Data Analytics Properties

Big data analytics is a multidisciplinary field that uses large, complex datasets to promote innovation across industries, extract useful insights, and influence decision-making. Six key components support its efficacy: it allows for secure parameter modifications, streamlined data integration, sophisticated data exploration, scalable data analysis, strong identity management, and extensive reporting features [2]. These features collectively constitute the fundamental components of a data-driven world. Combining these traits gives individuals and businesses the means to succeed in the data-centric landscape of the twenty-first century.

- **Scalability**: A fundamental quality in big data analytics is scalability, which concerns how well an analytics model can handle enormous amounts of data while maintaining controllable prices for hardware and cloud services [3]. Data scientists often face the difficulty of scaling their models to handle much larger, more complex data because they usually start their analytical journey with smaller datasets. Scalability is a basic design feature of an ideal big data analytics platform, making shifting from small-scale data analysis to large-scale operations easier. In today's data-intensive world, scalability plays a critical role in facilitating the effective management and extraction of relevant insights from the large amounts of data collected.
- **Version Control**: In big data analytics, version control plays an important role in managing iterative modifications to analytics models. Version control enables safe and reversible software alterations by systematically monitoring and documenting different versions [4]. In the event of unanticipated problems or system failures, this feature allows data scientists to quickly return to an earlier iteration of the analytics model, minimising project delays and guaranteeing budgetary compliance. However, fine-tuning model parameters is common for data scientists and comes with inherent hazards. Modest changes have the potential to seriously upset the system and cause major delays and cost overruns in projects.
- **Simple Integration Process**: Big data analytics technologies usually collect information from various sources, including business systems, cloud apps, and data warehouses. Intricate modifications are frequently required for these integrations to guarantee flawless communication and data processing. Data scientists work much more efficiently when analytics tools provide an easy-to-use integration method [5]. These solutions take up valuable time by simplifying the complexities involved in data source integration, enabling data scientists to concentrate on other important activities, such as improving analytics models. Simplifying the integration process makes preparing data smoothly and effectively easier, enabling data scientists to extract valuable insights more quickly.

- **Better Data Exploration**: Another important step in big data analytics is data exploration, in which data scientists thoroughly examine the gathered information to find previously undiscovered relationships, comprehend the context of business problems better, and create relevant analytical questions [6]. Significantly, faster data exploration is achieved with analytics tools that support it. They make it easier to test hypotheses quickly, spot weak or ambiguous data points rapidly, and offer tools for data visualisation. These kinds of capabilities enable data scientists to find important insights faster, which enhances decision-making in general.
- **Identity Management**: Identity management is essential to the overall data protection and cybersecurity strategy, an extensive database containing data about certain computer systems, software, and hardware. Identity management, which carefully regulates access, is essential to data security [7]. Identity management systems play a major role in data security by controlling and limiting access to particular data for systems or individuals. Access must be restricted to authorised individuals or devices to protect sensitive information's confidentiality and integrity and improve an organisation's overall cybersecurity posture.
- **Reporting Features**: The picture of big data analytics is incomplete without reporting features, which include real-time reporting, dashboard management, and location-based insights. These attributes facilitate enterprises to uphold a watchful and knowledgeable posture concerning their data assets [8]. Businesses are notified when significant data linkages or practical insights are discovered. Thanks to this real-time access, organisational leaders can effectively respond to important events, enabling them to make prompt and informed decisions. Reporting features would allow firms to stay flexible, adaptable, and responsive to new possibilities and trends in a data-driven environment.

1.2 Big Data Analytics Techniques

Big data analytics techniques include a broad range of complex approaches and instruments carefully designed to negotiate the complex terrain of large and complex datasets. They act as the cornerstone for businesses and organisations that offer a crucial way to seize important insights and establish a competitive edge in the fast-paced world of modern data-driven ecosystems. These methodologies are invaluable to enterprises and institutions, allowing them to identify trends, deconstruct intricate systems, and extract significant understandings from vast data.

Some of the key big data analytics techniques include:

- **Data Mining**: One of the most important methods in big data analytics is data mining, which uses sophisticated algorithms to find hidden relationships and patterns in enormous datasets [9]. Its main job is to break down and analyse complicated datasets to extract priceless information and spot new trends. Using this approach, data must first be gathered and prepared. Then, specific algorithms must be applied

to search for latent links and recurring patterns within the data environment. Due to its adaptability, data mining is used in many disciplines, including scientific research, financial analysis, market research, and health care.

- **Machine Learning**: Machine learning is the cutting edge of advanced analytics. It is a set of methods that allow computers to learn independently and improve their performance via experience [10]. This capacity is extremely useful in many areas, such as recommendation systems, categorization problems, and predictive analytics. With machine learning algorithms, computers can now identify complex patterns, anticipate outcomes based on data, and adjust to changing conditions. Using historical data, models are trained in this autonomous learning process, improving their capacity for precise prediction or decision-making. Essentially, machine learning elevates the potential of data-driven decision-making and gives businesses the tools to improve consumer experiences, automate processes, and streamline workflows. It is revolutionising how businesses function and interact with data in the modern day, with applications ranging from health care and banking to e-commerce and autonomous driving.
- **Natural Language Processing (NLP)**: NLP is a vital component within big data analytics that analyses and interprets unstructured textual data by exploring the intricacies of human language [11]. By bridging the gap between computational analysis and human communication, NLP techniques allow useful insights to be extracted from large amounts of textual data. NLP is a broad field that includes tasks such as sentiment analysis, which measures emotional tone; language translation, which fills in linguistic gaps; and the interpretation of linguistic nuances, such as idioms, slang, or contextual signals [12]. Organisations can automate language-based processes, generate a deeper knowledge of the content in documents, social media, or customer interactions, analyse textual information, and identify trends in consumer feedback by implementing NLP approaches. The increasing presence of digital material has increased the importance of this sector, making it a vital resource for fields like content suggestion, customer service automation, and social media analytics. Ultimately, these fields will influence how companies use and interact with textual data.
- **Data Visualisation**: Data visualisation plays a crucial role in big data analytics by bridging complex datasets and human understanding. Using interactive dashboards, graphs, and charts, this art form goes beyond simple data display and turns raw data into forms that are easy to understand [13]. When data is unprocessed, it can be confusing and overwhelming. Data visualisation addresses this by offering an understandable and simple way to work with complicated datasets. Decision-making becomes faster and more precise when patterns, trends, and anomalies concealed in the data become instantly visible through visualisation. It democratises access to insights and allows analysts, data scientists, and decision-makers to successfully convey their results to a larger audience. This leads to a more educated, data-driven approach to addressing business challenges.
- **Predictive Analytics**: A key strategy in big data analytics is predictive analytics, which uses historical data to forecast future patterns and results. This method uses advanced predictive modelling and analysis to provide priceless insights that

facilitate proactive strategy development and well-informed decision-making [14]. Using past data patterns and correlations, predictive analytics gives businesses a powerful tool to foresee future events, spot possible hazards, and grab opportunities. Doing this greatly improves their ability to adjust and react strategically to a business environment that is changing quickly. Whether used in marketing, finance, health care, or other industries, predictive analytics is essential for streamlining processes, allocating resources, and controlling risk. All of these factors contribute to more effective and profitable company outcomes.

- **Statistical Analysis**: In big data analytics, statistical analysis is a rock-solid foundation of scientific rigour that provides a systematic framework for hypothesis testing, data inference, and probabilistic inference-based decision-making [15]. Using this technique, data scientists and analysts can make significant findings by revealing hidden patterns, correlations, and trends in large datasets by applying mathematical and statistical methodologies. Statistical analysis is a lighthouse of impartiality that helps practitioners make evidence-based judgements by carefully analysing data's inherent uncertainties and variabilities [16]. It supports the validity and trustworthiness of findings in fields including experimental research, risk assessment, and quality control.
- **Clustering and Segmentation**: Clustering and segmentation techniques are indispensable for organising and extracting insights from complex big datasets that enable grouping similar data points into clusters or segments, illuminating inherent patterns and structures within the data [17]. With the unsupervised learning approach, clustering identifies hidden relationships and categorises data points based on their similarities, shedding light on intricate connections that might otherwise remain concealed [18]. In market analysis and targeted marketing, segmentation techniques divide datasets into distinct groups, which allows businesses to tailor their strategies and offers to specific customer segments [19]. These techniques enhance decision-making and streamline marketing efforts, leading to more effective and personalised approaches.
- **Real-time Analytics**: Real-time analytics enable organisations to extract immediate value from large-scale complex streaming data [20]. Fast, data-driven decision-making is essential in today's fast-paced, constantly changing digital environment, and this approach is motivated by its necessity. Organisations can make quick decisions responding to events, trends, or new information by processing data in real time. This enables them to optimise short-term prospects, reduce possible hazards, and improve overall operational effectiveness. When it comes to financial trading, cybersecurity, and e-commerce, for example, real-time analytics is very helpful when responding quickly to threats [21].
- **Distributed Computing**: When processing large datasets presents a barrier, distributed computing is a powerful paradigm in big data analytics. This method uses distributed systems' capabilities, which are best demonstrated by Hadoop and Spark, to enable the smooth transfer of data and computing workloads over a network of linked nodes or clusters [22]. The main benefit is that it may divide large, complicated data analysis jobs into smaller, more manageable chunks that can be performed simultaneously on several nodes, thereby reducing the time

needed for analysis [23]. From batch processing to real-time data streaming, distributed computing provides the computational power and scalability to handle complex calculations. It speeds up data analysis and improves fault tolerance by ensuring that the workload may be moved to another node without causing a lot of interruption in the case of a hardware breakdown.

1.3 Overview of This Book

This section provides an insightful glimpse into the content and structure of the book, offering a comprehensive framework of the core themes and subjects to be elucidated in subsequent chapters.

Foundations of Big Data Analytics: A fundamental investigation of big data analytics is at the core of this work. Data now serves as the foundation for innovation and decision-making across industries in the twenty-first century. It is essential to comprehend where data comes from and what it means in the modern world. This section examines the historical development of data analytics, tracing its origins and displaying its profound social impact. Readers will understand deeply how we arrived at today's data-centric society by exploring the process from data gathering to analysis.

Core Concepts and Technologies: Understanding big data analytics's fundamental ideas and tools is necessary for navigating the field. This section provides a thorough overview of data analytics's essential components. It introduces readers to the basic elements, tools, and frameworks necessary for successfully handling and analysing huge datasets. The foundation of contemporary data-driven decision-making, scalable data architecture, and fundamental analytics methodologies are highlighted. Readers will thoroughly understand the technology underlying big data analytics by the end of this part.

Practical Applications: Big data analytics is a dynamic force actively transforming industries; it is not merely a theoretical endeavour. This section illustrates practical applications from various fields, putting theory into action. Readers will see how data analytics fosters creativity, process optimisation, and priceless insights through interesting case studies and examples from health care, finance, and marketing. These real-world examples highlight the transformative potential of big data analytics and help readers imagine how it will affect their particular fields of interest.

Fostering Expertise Development: The journey through this book offers readers a way to develop competence in big data analytics. This book meets a range of learning needs, whether the reader is a beginner in the field or an experienced practitioner looking for significant insights. It offers a disciplined process for obtaining the knowledge and skills necessary for successfully navigating our modern, data-driven landscape. With the help of foundational concepts, technological know-how, practical application, and interactive activities, readers will leave the book prepared to take advantage of data's pervasive effect in the twenty-first century.

References

1. G.S. Aujla, N. Kumar, A.Y. Zomaya, R. Ranjan, Optimal decision making for big data processing at edge-cloud environment: An sdn perspective. IEEE Trans. Ind. Inf. **14**(2), 778–789 (2018)
2. 6 features that make big data analytics vital for businesses. Selerity. [Online]. Available: https://seleritysas.com/2019/05/06/6-features-that-make-big-data-analytics-vital-for-businesses/
3. H. Hu, Y. Wen, T.-S. Chua, X. Li, Toward scalable systems for big data analytics: A technology tutorial. IEEE Access **2**, 652–687 (2014)
4. D.L. Atkins, T. Ball, T.L. Graves, A. Mockus, Using version control data to evaluate the impact of software tools: A case study of the version editor. IEEE Trans. Soft. Eng. **28**(7), 625–637 (2002)
5. X.L. Dong, D. Srivastava, Big data integration, in *IEEE 29th International Conference on Data Engineering (ICDE)*. (IEEE, 2013), pp. 1245–1248
6. A. Wasay, M. Athanassoulis, S. Idreos, Queriosity: Automated data exploration, in *IEEE International Congress on Big Data*. (IEEE, 2015), pp. 716–719
7. P. Jain, M. Gyanchandani, N. Khare, Big data privacy: a technological perspective and review. J. Big Data **3**, 1–25 (2016)
8. P. Mikalef, M. Boura, G. Lekakos, J. Krogstie, Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment. British J. Manage. **30**(2), 272–298 (2019)
9. X. Wu, X. Zhu, G.-Q. Wu, W. Ding, Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2013)
10. Challenges and approaches, A. L'heureux, K. Grolinger, H.F. Elyamany, M.A. Capretz, Machine learning with big data. IEEE Access **5**, 7776–7797 (2017)
11. R. Sharma, P. Agarwal, A. Arya, Natural language processing and big data: a strapping combination, in *New Trends and Applications in Internet of Things (IoT) and Big Data Analytics*. (Springer, 2022), pp. 255–271
12. J.C. Eichstaedt, M.L. Kern, D.B. Yaden, H.A. Schwartz, S. Giorgi, G. Park, C.A. Hagan, V.A. Tobolsky, L.K. Smith, A. Buffone et al., Closed-and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. Psychol. Methods **26**(4), 398 (2021)
13. D. Keim, H. Qu, K.-L. Ma, Big-data visualization. IEEE Comput. Graph. Appl. **33**(4), 20–21 (2013)
14. A. Shi-Nash, D.R. Hardoon, Data analytics and predictive analytics in the era of big data, in *Internet of Things and Data Analytics Handbook* (2017), pp. 329–345
15. C. Wang, M.-H. Chen, E. Schifano, J. Wu, J. Yan, Statistical methods and computing for big data. Stat. Interface **9**(4), 399 (2016)
16. A. Der Kiureghian, Analysis of structural reliability under parameter uncertainties. Probab. Eng. Mech. **23**(4), 351–358 (2008)
17. F. Yoseph, N.H. Ahamed Hassain Malim, M. Heikkilä, A. Brezulianu, O. Geman, N.A. Paskhal Rostam, The impact of big data market segmentation using data mining and clustering techniques. J. Intell. Fuzzy Syst. **38**(5), 6159–6173 (2020)
18. O. Nasraoui, C.-E.B. N'Cir, Clustering methods for big data analytics. Tech. Toolbox. Appl. **1**, 91–113 (2019)
19. P. Monil, P. Darshan, R. Jecky, C. Vimarsh, B. Bhatt, Customer segmentation using machine learning. Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET) **8**(6), 2104–2108 (2020)
20. R. Ranjan, Streaming big data processing in datacenter clouds. IEEE Cloud Comput. **1**(01), 78–83 (2014)
21. M. Cao, R. Chychyla, T. Stewart, Big data analytics in financial statement audits. Account. Horizons **29**(2), 423–429 (2015)

22. K. Kambatla, G. Kollias, V. Kumar, A. Grama, Trends in big data analytics. J. Parall. Distrib. Comput. **74**(7), 2561–2573 (2014)
23. U. Demirbaga, G.S. Aujla, Federated-ann based critical path analysis and health recommendations for mapreduce workflows in consumer electronics applications. IEEE Trans. Consum. Electron. (2023)

Big Data

2

Machine learning will automate jobs that most people thought could only be done by people.

—Dave Waters

This chapter introduces the fundamental concepts of big data, offering a comprehensive understanding of its definition, key characteristics, and the widely recognised 5 Vs. The multifaceted challenges associated with realising the enormous potential of big data are explored, encompassing issues related to data collection, storage, privacy, security, and the complexities of deriving value from this extensive resource. In addition, avenues for harnessing the power of big data are investigated, including applying advanced analytics and machine learning, utilising data visualisation techniques, and implementing communication strategies. Lastly, a glimpse into the future of big data is provided, shedding light on emerging trends and directions that will shape its ongoing evolution and influence across various domains.

2.1 Definition of Big Data

Big data is a term that emerged in astronomy and genetics but has now been used on the Internet and has become a part of our everyday lives without our awareness or actively contributing to it. Much data has been stored, processed, and managed due to the computer's effectiveness in every part of our lives. The growing Internet use by businesses, corporations, and individuals has resulted in the circulation, processing, and dissemination of these data in electronic media [1]. The data we have described comprises information entered and stored as a condition of service, as well as a large

Ü. Demirbaga et al., *Big Data Analytics*,
https://doi.org/10.1007/978-3-031-55639-5_2