Studies in Big Data 137

Vinay Rishiwal
Pramod Kumar
Anuradha Tomar
Priyan Malarvizhi Kumar *Editors*

Towards the Integration of IoT, Cloud and Big Data

Services, Applications and Standards

Springer

Studies in Big Data

Volume 137

Series Editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland

The series “Studies in Big Data” (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence including neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

The books of this series are reviewed in a single blind peer review process.

Indexed by SCOPUS, EI Compendex, SCIMAGO and zbMATH.

All books published in the series are submitted for consideration in Web of Science.

Vinay Rishiwal · Pramod Kumar ·
Anuradha Tomar · Priyan Malarvizhi Kumar
Editors

Towards the Integration of IoT, Cloud and Big Data

Services, Applications and Standards

Springer

Editors
Vinay Rishiwal
Department of CSIT
Faculty of Engineering and Technology
M.J.P. Rohilkhand University
Bareilly, India

Anuradha Tomar
Netaji Subhas University of Technology
New Delhi, India

Pramod Kumar
Glocal University
Saharanpur, Uttar Pradesh, India

Priyan Malarvizhi Kumar
Department of Data Science
University of North Texas
Denton, TX, USA

ISSN 2197-6503 ISSN 2197-6511 (electronic)
Studies in Big Data
ISBN 978-981-99-6033-0 ISBN 978-981-99-6034-7 (eBook)
https://doi.org/10.1007/978-981-99-6034-7

Preface

The rapid advancement of technology has led to the emergence of the Internet of Things (IoT), Cloud Computing, and Big Data as transformative forces in various industries. As these technologies continue to evolve, there is a growing need for their integration to unlock their full potential and enable the development of innovative services, applications, and standards. The integration of these three domains presents numerous challenges and opportunities. One of the key challenges is the efficient and secure management of the massive data generated by IoT devices, as well as the seamless integration of IoT devices with cloud-based infrastructure. This requires the development of scalable and robust architectures, protocols, and standards that enable interoperability, data sharing, and resource allocation across heterogeneous systems. Moreover, the integration of IoT, Cloud, and Big Data enables the creation of innovative services and applications. To achieve successful integration, the establishment of common standards is crucial.

To summarise, it is the right time to explore the integration of IoT, Cloud, and Big Data, which holds immense potential to transform industries, enhance services, and enable data-driven decision-making. However, addressing the challenges related to data management, interoperability, and security is vital for successful integration. Moreover, the establishment of standards is crucial to facilitate seamless communication and collaboration between different systems. By leveraging the combined power of IoT, Cloud, and Big Data, organizations can unlock new possibilities and drive digital transformation in the era of interconnected and data-driven ecosystems.

This book consists of eight chapters. The first chapter covers introduction to Big Data analysis and its need, skills required for Big Data analysis, characteristics of Big data analysis, an overview of the Hadoop ecosystem, and some use cases of Big Data analysis. The aim of the second chapter is to study and compare three of the most common classification methods, Support Vector Machines, K-Nearest Neighbours and Artificial Neural Networks, for heart disease prediction using the ensemble of standard Cleveland cardiology data. The objective of the third article is to reduce the energy consumption of the ECG machine. Authors in chapter four, have proposed a system to implement an automatic water supply to the farms based upon their crop, system that measures water level of soil and helps to decide to turn on or off the water

supply. Further, chapter five uses deep convolutional networks algorithms for leaf image classification to provide accurate results. The concept of Blockchain is used in chapter six with the aim to ensure the security of the patient's medical records. Chapter seven offers SHA-PSO, a PSO-based meta-heuristic technique that schedules workloads among Virtual Machines (VM) to minimize energy. Authors in chapter eight have proposed design of field monitoring device using IoT in Agriculture.

Bareilly, India — Vinay Rishiwal
Saharanpur, India — Pramod Kumar
New Delhi, India — Anuradha Tomar
University of North Texas, USA — Priyan Malarvizhi Kumar

Contents

Editors and Contributors

About the Editors

Dr. Vinay Rishiwal Ph.D. is working as a Professor in the Department of Computer Science and Information Technology, Faculty of Engineering and Technology, MJP Rohilkhand University, Bareilly, Uttar Pradesh, India. He obtained B.Tech. degree in Computer Science and Engineering in the year 2000 from M.J.P. Rohilkhand University (SRMSCET), India and received his Ph.D. in Computer Science and Engineering from Gautam Buddha Technical University, Lucknow, India, in the year 2011. He has 23 years of experience into academics. He is a senior member of IEEE, ACM and worked as Convener, Student Activities Committee, IEEE Uttar Pradesh Section, India. He has published more than 90 research papers in various journals and conferences of international repute. He also has 20 patents into his credit. He is a General/Conference chair of four International Conferences namely ICACCA, IoT-SIU, MARC 2020 and ICAREMIT. He has received many awards as best paper/ research/orator at various platforms. Dr. Rishiwal has visited many countries for academic purposes and worked upon many projects of CST, UP Government, MHRD and UGC. His current research interest includes Wireless Sensor Networks, IoT, Cloud Computing, Social networks and Blockchain Technology.

Prof. (Dr.) Pramod Kumar is an accomplished academic leader with over 24 years of experience in the field. He currently serves as the Dean of Academics at Glocal University in Saharanpur, UP, where he has been since September 2022. Prior to this, he held the position of Dean of Computer Science and Engineering at Krishna Engineering College in Ghaziabad and served as the director of Tula's Institute in Dehradun, Uttarakhand. Prof. Pramod Kumar holds a Ph.D. in Computer Science and Engineering, which he earned in 2011, as well as an M.Tech in CSE from 2006. He is a Senior Member of IEEE and an Ex-Joint Secretary of the IEEE U.P. section. Through his research, he has made significant contributions to the fields of Computer Networks, IoT, and Machine Learning. He is the author or co-author of more than 70

research papers and has edited four books. He has also supervised and co-supervised several M.Tech. and Ph.D. students.

Dr. Anuradha Tomar is currently working as an Assistant Professor in the Instrumentation & Control Engineering Division of Netaji Subhas University, Delhi, India. Dr. Tomar has completed her Postdoctoral research in EES, from Eindhoven University of Technology (TU/e), the Netherlands. She received her B.E Degree in Electronics Instrumentation & Control with Honours in the year 2007 from the University of Rajasthan, India. In the year 2009, she completed her M.Tech. Degree with Honours in Power Systems from the National Institute of Technology Hamirpur. She received her Ph.D. in Electrical Engineering, from the Indian Institute of Technology Delhi (IITD), India. Dr. Anuradha Tomar has committed her research work efforts towards the development of sustainable, energy-efficient solutions for the empowerment of society, and humankind. Her areas of research interest are the Operation & Control of Microgrids, Photovoltaic Systems, Renewable Energy based Rural Electrification, Congestion Management in LV Distribution Systems, Artificial Intelligent & Machine Learning Applications in Power Systems, Energy conservation, and Automation.

Dr. Priyan Malarvizhi Kumar is presently employed as an Assistant Professor at the University of North Texas in the United States. Before joining this role, he served as an Assistant Professor in the Department of Computer and Information Science at Gannon University, USA. Prior to his tenure at Gannon University, he held the position of Assistant Professor in the Computer Science and Engineering Department at Kyung Hee University in South Korea. Additionally, he gained valuable experience as a Postdoctoral Research Fellow at Middlesex University in London, UK. Dr. Kumar earned his Ph.D. degree from Vellore Institute of Technology University. His academic journey also includes a Bachelor of Engineering degree from Anna University and a Master of Engineering degree from Vellore Institute of Technology University.

Dr. Kumar's current research focuses on areas such as Big Data Analytics, Internet of Things (IoT), Internet of Everything (IoE), and Internet of Vehicles (IoV) in the context of healthcare. He has authored and co-authored papers published in international journals and conferences, including those indexed by the Science Citation Index (SCI). He maintains a lifetime membership with the International Society for Infectious Disease, the Computer Society of India, and is an active member of the Vellore Institute of Technology Alumni Association.

Contributors

Vikrant Agaskar Vidyavardhani College of Engineering and Technology, Vasai-Virar, Maharashtra, India

Nitin Arora Electronics and Computer Discipline, Indian Institute of Technology, Roorkee, India

Neema Bhandari Birla Institute of Applied Sciences, Bhimtal, Uttarakhand, India

Neeraj Bisht Birla Institute of Applied Sciences, Bhimtal, Uttarakhand, India

Shilpi Bisht Birla Institute of Applied Sciences, Bhimtal, Uttarakhand, India

Sandeep K. Budhani Graphic Era Hill University, Bhimtal, Uttarakhand, India

Rashmi Chaudhry Netaji Subhas University of Technology, Delhi, India

Meetali Chauhan Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India

Goutam Datta School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India

Rakesh Kumar Dwivedi CCSIT, Teerthanker Mahaveer University, Moradabad, UP, India

Sushopti Gawade Pillai College of Engineering, Panvel, India

Aman Kataria Amity Institute of Defence Technology, Amity University, Noida, India

Alex Khang GRITEx and VUST, Ho Chi Minh City, Vietnam

Umesh Kulkarni Vidyalankar Institute of Technology Wadala, Mumbai, Maharashtra, India

Hemant Palivela Manager-AI, Accenture Solutions, Mumbai, Maharashtra, India

Bishwajeet Pandey Gyancity Lab, Guragaon, India

Danilla Parygin Volgograd State Techincal University, Vogograd, Russia

Ravi Prakash CSED, Motilal Nehru National Institute of Technology, Allahabad, India

Sita Rani Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India

Mukesh Rawat CSED, Meerut Institute of Engineering and Technology, Meerut, India

Vinay Rishiwal MJP Rohilkhand University, Bareilly, India

Ashendra Kumar Saxena CCSIT, Teerthanker Mahaveer University, Moradabad, UP, India

Vivek Shahare Department of Computer Science and Engineering, Indian Institute of Technology, Dharwad, India

Swati Sharma IT, Meerut Institute of Engineering and Technology, Meerut, India

Vijay Kumar Sharma CSED, Meerut Institute of Engineering and Technology, Meerut, India

Anupam Singh Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India

Kaustubh Ranjan Singh Delhi Technological University, Delhi, India

Pankaj Singh Birla Institute of Applied Sciences, Bhimtal, Uttarakhand, India

Mano Yadav Bareilly College Bareilly, Bareilly, India

Preeti Yadav MJP Rohilkhand University, Bareilly, India

Introduction to Big Data Analytics

Nitin Arora, Anupam Singh, Vivek Shahare, and Goutam Datta

Abstract Nowadays, social media and networks, scientific instruments, mobile devices, mobile devices, and a high volume of information data (tabular data, text files, images, videos, audio, logos, etc.) is generated at high velocity by social media and networks, scientific instruments, mobile devices, and sensors technology and networks. In these types of data, data quality is usually not guaranteed. This data can be structured or unstructured, necessitating a cost-effective, innovative method of data processing to improve understanding and decision-making. This chapter covers some introduction to Big Data analysis and its need, skills required for Big Data analysis, characteristics of Big data analysis, an overview of the Hadoop ecosystem, and some use cases of Big Data analysis.

Keywords Big data · Hadoop ecosystem · Big data analysis · Business intelligence analysis · Big data domain · Big data quality · Dimensions

N. Arora (✉)
Electronics and Computer Discipline, Indian Institute of Technology, Roorkee, India
e-mail: nitinarora.iitr@gmail.com

A. Singh
Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India
e-mail: anupamsingh.cse@geu.ac.in

V. Shahare
Department of Computer Science and Engineering, Indian Institute of Technology, Dharwad, India
e-mail: vivek.shahare27@gmail.com

G. Datta
School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India
e-mail: gdatta@ddn.upes.ac.in

V. Rishiwal et al. (eds.), *Towards the Integration of IoT, Cloud and Big Data*,
Studies in Big Data 137, https://doi.org/10.1007/978-981-99-6034-7_1

Table 1 Characteristics of data

	Small data	Big data
Volume	Less than 1 TB	Greater than 1 TB
Velocity	Controlled and steady data flow	Enormous data flowing at shorter time frames
Variety	Structured and semi-structured. E.g., Xcel, Table, data, Json	Wide variety of data, i.e., tabular data, text files, images, videos, audios, logs, etc.
Veracity	It contains more quality data	The quality of data is rarely guaranteed
Value	Business Intelligence, analysis, and reporting	Complex data mining, predictions, pattern finding, etc.
Time variance	Data represents the business value for the history of data as well as incremental	At times, history data becomes irrelevant for analyzing business insights
Infrastructure	More defined resources allocation	The Load-on system varies a lot

1 Introduction to Big Data

Big Data is a phrase that relates to a collection of vast and complex data sets that are challenging to store and analyze using standard data processing methods. Big data refers to data assets with a large volume, great velocity, and great diversity that necessitate cost-effective, creative data processing to improve insight and decision-making [1].

2 The Distinction Between Small and Big Data

There are several distinctions between small data and big data. These distinctions include volume, velocity, variety, veracity, value, time variance, and infrastructure [2]. Table 1 summarizes all the differences.

3 Classification of Big Data

Big data is classified as [3]

- **Structured Data**: Structured data has a well-defined format. It can be readily stored in tabular form in relational databases such as MySQL and Oracle.
- **Semi-Structured Data**: Semi-structured data has some structure but can't be recorded in a tabular format in relational databases. XML files, JSON documents, e-mail messages, and so forth are examples.

– **Unstructured Data**: Unstructured data has no structure and cannot be saved in tabular form in relational databases. Examples include video, audio, text, and machine-produced data.

4 Characteristics of Big Data

Big data has many characteristics. Some of them are: [4]

– **Volume**: The term "volume" refers to the "quantity of data," which rapidly increases daily. Humans, technology, and their interactions on social media create enormous amounts of data.
– **Variety**: Because so many sources contribute to Big Data, the types of data they generate are diverse. It might be organized, semi-organized, or unorganized. Many different forms of data can be generated/collected by a single application. All of these forms of data must be connected to extract knowledge.
– **Velocity**: Velocity refers to stream of data that arrives from different social media sites continuously, and the repository gets completed with new data at the same rate. It becomes a challenge to capture this stream of data promptly for further processing.
– **Veracity**: The term "veracity" alludes to the data's unreliability. Data inconsistency and incompleteness create uncertainty in the data supplied. Many extensive data types, such as Twitter postings with hashtags, abbreviations, typos, and colloquial speech, have less controlled quality and accuracy.
– **Value**: It's great to access massive data, but it isn't sensible unless we can transform it into practice.
– **Variability**: Diversity and variation are not the same things. Even though a coffee shop may offer six different coffee blends, diversity only exists when you consistently obtain and enjoy the same blend. The same is true for data; if the meaning changes frequently, it can significantly affect the homogeneity of the data.
– **Visualization**: Using charts and graphs to represent vast quantities of accurate data is far more successful than using spreadsheets.

5 Who's Generating Big Data?

The capacity to acquire data no longer limits development and creativity. However, the capacity to organize, analyze, summarise, display, and find information from acquired data in a timely and scalable manner is critical.

6 Why Is Big Data Important?

Companies obtain a complete knowledge of their company, consumers, products, and rivals if big data is gathered, processed, and analyzed properly and efficiently. It results in enhanced efficiency, more sales, cheaper expenses, better customer service, and better goods and services. Sensors are embedded in manufacturing items to provide a stream of telemetry. Retailers frequently know who buys their products. To determine who didn't buy and why, businesses can leverage social media and blog data from their e-commerce sites, which is the knowledge they don't have today. Using extensive historical call center data more rapidly enhances customer engagement and satisfaction. Use of social media material is encouraged to better and more rapidly assess consumer sentiment about you/your customers and enhance goods, services, and customer interactions [5].

7 Challenges in Big-Data

Big data size is enormous, and this data can be structured or unstructured. There are many challenges in this and are discussed below [6].

- **Volume**: Thanks to new data sources that are developing, the volume of data, particularly machine-generated data, is expanding, as is the rate at which it expands each year. For example, the world's data storage capacity was 800,000 petabytes (PB) in 2000 and is anticipated to reach 35 zettabytes by 2020.
- **Variety and the use of many data sets**: Unstructured data makes up greater than 80% of today's data. Most of this data is so vast for effective management.
- **Velocity**: As organizations realize the benefits of analytics, they face a problem: they want the data sooner, or in other words, they want real-time analytics.
- **Veracity, Data Quality, Data Availability**
- **Data Discovery**: Finding high-quality data from the massive amounts of data available on the Internet is a significant problem.
- **Relevance and Quality**: It's tough to determine data sets' quality and relevance to specific requirements.
- **Personally Identifiable Information**: A lot of this data is about people. This necessitates, in part, efficient industrial processes. "It partly asks for efficient government monitoring. Partly-perhaps even entirely-it necessitates a severe rethinking of what privacy truly entails."
- **Process Challenges**: Finding the appropriate analysis model may take a lot of time and effort; thus, the ability to cycle quickly and 'fail fast' through many (perhaps throwaway) models is crucial.
- **Management Challenges**: Sensitive data, such as personal information, is found in many warehouses. Accessing such data raises legal and ethical problems. As a result, the data must be secured, access restricted, and audited.