Computer Communications and Networks

Pethuru Raj
Anupama Raman
Dhivya Nagaraj
Siddhartha Duggirala

High-Performance Big-Data Analytics

Computing Systems and Approaches

Springer

Computer Communications and Networks

Series editor
A.J. Sammes,
Centre for Forensic Computing
Cranfield University, Shrivenham Campus
Swindon, UK

The **Computer Communications and Networks** series is a range of textbooks, monographs and handbooks. It sets out to provide students, researchers, and non-specialists alike with a sure grounding in current knowledge, together with comprehensible access to the latest developments in computer communications and networking.

Emphasis is placed on clear and explanatory styles that support a tutorial approach, so that even the most complex of topics is presented in a lucid and intelligible manner.

More information about this series at http://www.springer.com/series/4198

Pethuru Raj • Anupama Raman
Dhivya Nagaraj • Siddhartha Duggirala

High-Performance Big-Data Analytics

Computing Systems and Approaches

Springer

Pethuru Raj
IBM India
Bangalore, India

Anupama Raman
IBM India
Bangalore, India

Dhivya Nagaraj
IBM India
Bangalore, India

Siddhartha Duggirala
Indian Institute of Technology
Indore, MP, India

ISSN 1617-7975 ISSN 2197-8433 (electronic)
Computer Communications and Networks
ISBN 978-3-319-20743-8 ISBN 978-3-319-20744-5 (eBook)
DOI 10.1007/978-3-319-20744-5

Library of Congress Control Number: 2015951624

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Foreword

In the recent past, the data growth has been unbelievably phenomenal due to a plethora of converging technologies (digitization, connectivity, integration, perception, miniaturization, consumerization, commoditization, orchestration of knowledge discovery and dissemination, etc.). In short, every common and casual thing in our personal as well as professional environments gets connected with one another and service-enabled to innately enable them to participate seamlessly and sagaciously in the mainstream computing. The device landscape is also going through a variety of innovations and improvisations in the recent past. Precisely speaking, the deeper and extreme connectivity among all kinds of digitized artifacts and devices is primarily responsible for the massive amounts of data.

This brewing trend and transformation brings forth a variety of challenges as well as opportunities not only for IT professionals but also for data scientists. The discipline of data analytics is bound to grow in multiple dimensions and directions. Newer types of data analytics (generic as well as specific) are bound to emerge and evolve fast. The compute, storage, and network challenges are also destined to become severe. With the ever-increasing data size, structure, scope, speed, and value, the biggest challenge for any enterprise and its IT team is how to flawlessly capture and process data and extract actionable insights in time. Each type of data emanating internally as well as externally provides hidden insights in the form of usable patterns, fruitful associations, timely alerts, fresh possibilities, and opportunities.

In this book, the authors have passionately explained why big and fast data analytics need high-performance infrastructures (server machines, storage appliances, and network connectivity solutions) to do justice for the next-generation data analytic solutions. With the sole aim of conveying the nitty-gritty of the state-of-the-art technologies and tools which are emerging and evolving in the high-performance big data analytics domain, authors have consciously focused on the various types of enabling IT infrastructures and platforms for the same.

Data Analytics: The Process Steps It is a well-known fact that typically there are three major phases/stages in any data analytics task:

1. *Data capture* through data virtualization platforms
2. *Data processing/interpretation* platforms for knowledge discovery
3. *Knowledge dissemination* through a host of data visualization platforms

The Emerging Analytics Types With the ever-increasing data volume, velocity, variety, variability, viscosity, and veracity, a bevy of powerful analytics (generic as well as specific) use cases is getting unearthed. All kinds of business verticals and industry segments are employing different types of analytics to squeeze out actionable insights from their data. The generic/horizontal analytic types include:

- Sensor analytics
- Machine analytics
- Operational analytics
- Real-time analytics
- High-performance analytics

The domain-specific analytic types include social media and network analytics, customer sentiment analytics, brand optimization analytics, financial trading and trending analytics, retail analytics, energy analytics, medical analytics, and utility analytics.

Emerging IT Infrastructures and Platforms IT infrastructures are increasingly becoming converged, centralized, federated, pre-integrated, optimized, and organized in an attempt to evolve into a right and relevant option for futuristic businesses. Analytical platforms are hitting the market like never before. Having understood the significance of carefully and cognitively doing analytics on every form of data to be competitive in their business operations, enterprises across the globe are eagerly looking out for high-performing IT infrastructures and platforms to run big and fast data analytical applications in an effective and efficient manner.

The authors have extensively written about the existing and emerging high-performance IT infrastructures and platforms for efficiently accomplishing flexible data analytics. The prominent IT infrastructures which are discussed in this book include:

1. Mainframes
2. Parallel and supercomputing systems
3. Peer-to-peer, cluster, and grid computing systems
4. Appliances (data warehouse and Hadoop-specific)
5. Expertly integrated and specially engineered systems
6. Real-time systems
7. Cloud infrastructures

The authors have consciously brought forth all the value-adding information on next-generation IT infrastructures and platforms for big and fast data analytics in this book. This book will be highly beneficial for technical experts, consultants,

evangelists, and exponents apart from business executives, decision-makers, and other stakeholders. I am doubly sure that software engineers, solution architects, cloud professionals, and big data scientists across the world will find this book informative, interesting, and inspiring to deeply understand how data analytics will emerge as a common service and play an indispensable role in shaping up the world towards becoming rewardingly smart.

Sushil Pramanick, FCA, PMP

IBM Analytics, Orange County, CA, USA
TBDI, Orange County, CA, USA

Preface

Several industry trends and a host of powerful technologies and tools in a synchronized fashion undoubtedly lead to the massive data explosion. Incidentally data is overwhelmingly acquiring the status of a strategic asset across industry verticals. The prominent foundations for the unprecedented data include the following noteworthy transitions: the device ecosystem expands continuously as per the widely changing peoples' imaginations, machines are becoming intelligent with smart instrumentation and interconnectivity generating data in the range of petabytes and exabytes, personal and professional applications are meticulously service-enabled to be interoperable for beneficial data sharing, everyday social networking sites are producing terabytes of data, ordinary objects in and around us are minutely digitized generating a lot of multi-structured data at different velocities, etc. On the other hand, ICT infrastructures and platforms are highly optimized and organized for effective data storage and processing and analytics, adaptive WAN technologies are being formulated to accelerate data transfer securely, newer architectural patterns are assimilated, processes are systematically made nimble, etc. to make sense out of data.

This data when analyzed carefully can provide a wealth of information which can revolutionize all facets of our life. This idea has evolved to be a game changer in the present day IT domain and is widely referred to as big data analytics. Considering the data size, speed, scope, and structure, the compute, storage, and networking infrastructures need to be immaculately efficient. Primarily the big data world brings forth three crucial challenges for the IT world: big data storage and management, big data analytics, and producing sophisticated applications leveraging analytics. Precisely speaking big data analytics (BDA) is quickly turning out to be the next-generation high-performance computing discipline for our students, scholars, and scientists to unearth competent algorithms, patterns, methodologies, best practices, key guidelines, evaluation metrics, etc.

This book is a humble attempt to provide a bird's eye view of those techniques. A sincere and serious attempt is made to analyze the network and storage infrastructure optimizations which are the need of the hour in order to capture, ingest, crunch,

and handle big data efficiently for knowledge discovery and dissemination. Some use cases of big data analytics in various industry sectors are also included in this book to let you know the heightening importance of data analytics in a simplified and streamlined manner.

Chapter 1: *The Brewing Trends and Transformations in the IT Landscape*—This chapter, being the first chapter for the book, lists out the emerging transitions in the IT arena especially in the context for the big and fast data world. The promising and potential technologies and tools in the ICT domain are being given prime importance in this chapter in order to give an indication of what is in store for the readers in this book.

Chapter 2: *The High-Performance Technologies for Big and Fast Data Analytics*—This chapter has categorized the most visible technologies for high-performance big and fast data analytics.

Chapter 3: *Big and Fast Data Analytics Yearning for High-Performance Computing*—We have explained the nitty-gritty of big and fast data analytics here as a precursor for conveying the significance of high-performance computing needs for productively extracting actionable insights out of data heaps.

Chapter 4: *Network Infrastructure for High-Performance Big Data Analytics*—This chapter summarizes the network infrastructure requirements for effective transfer of big data. In order to transfer big data efficiently through the networks, it is necessary to perform some modifications to the existing network infrastructure. Some of the techniques which could be used are network virtualization, software-defined networks (SDN), two-tier leaf spine architecture, and network functions virtualization. Each of these techniques is explained in detail in this chapter. It is also necessary to optimize the existing Wide Area Network infrastructure so that it can transfer big data efficiently. A novel approach for transfer of big data efficiently using TCP/IP protocol using a technology called FASP is also discussed in this chapter. Some of the implementation aspects of FASP are also included in the discussion of this chapter.

Chapter 5: *Storage Infrastructures for High-Performance Big Data Analytics*—This chapter summarizes the storage infrastructure requirements of applications which generate big data. Present-day storage infrastructures are not optimized to store and handle big data. The main concern with the existing storage techniques is their lack of scalability. Hence it is the need of the day to devise new storage techniques which can handle big data efficiently. In this chapter, we are adopting a multifaceted approach: at first we discuss the existing storage infrastructure and their suitability to handle big data. Later on we discuss some platforms and file systems which are designed only for handling big data like PANASAS file system, Lustre file system, GFS, and HDFS.

Chapter 6: *Real-Time Analytics Using High-Performance Computing*—This chapter talks about analytics in the real-time environment. It covers all recent real-time analytics solutions like machine data analytics and operational analytics. This is an eye opener on how data can be handled in real time and the value that it adds in changing our lives for a better tomorrow.

Chapter 7: *High-Performance Computing (HPC) Paradigms*—This chapter covers in detail the reasons behind the evolution of high-performance computing over the years in mainframe. A few years ago, the world came to a conclusion that mainframes are going to be extinct with the evolving technology. But organizations like IBM have proved that mainframes are not going to be extinct but have come back with a bang providing solutions that were once assumed completely impossible.

Chapter 8: *In-Database Processing and In-Memory Analytics*—This chapter is for elucidating the in-database analytics techniques and in-memory analytics techniques. When the business systems run at large scale, moving data in and out of the data store can be really daunting and expensive. While moving the processing near to the data, the data processing is done in the data store itself; by doing this we can reduce the data movement costs and use much larger data sets to mine the data. While the businesses are moving, the speed has become crucial. This is where real-time databases come into the picture. This chapter covers all aspects pertaining to in-database and in-memory analytics techniques with appropriate examples.

Chapter 9: *High-Performance Integrated Systems, Databases, and Warehouses for Big and Fast Data Analytics*—For the ensuing big data era, there is a distinctive need for newer kinds of data management systems. We have clearly catalogued and written about the emerging clustered SQL, NoSQL, and NewSQL databases. There is an explanation on big data-specific data warehouses.

Chapter 10: *High-Performance Grids and Clusters*—This chapter is for elucidating the techniques and software tools available to enable big data analytics and high data-intensive processing. Businesses around the globe are under pressure to reduce the TCO of analytical platforms and yet quite ably perform at the required level. Using these versatile high-performance systems, businesses are able to meet the performance demands that are put forward to them. This chapter explains the different use cases about the usage of cluster and grid computing systems in the realm of big data analytics.

Chapter 11: *High-Performance Peer-to-Peer Systems*—This chapter is for elucidating the peer-to-peer techniques and tools that are used in big data analytics domain. Due to the large-scale nature of the data stores or analytical systems typically have master-slave relationship between servers. This helps in parallelizing the application, but a problem arises when the master fails. Then no request will be replied back. In these scenarios, if the software structure is decentralized, meaning no master servers, then there would be no single point of failure. Hence the entire request will be answered. This chapter explains the different use cases about the usage of high-performance peer-to-peer systems.

Chapter 12: *Visualization Dimensions for High-Performance Big Data Analytics—This chapter* is for elucidating the visualization techniques and tools. As the data size increases and complexity of the data increases, it becomes difficult to comprehend the meaning of the data and the analysis which is uncovered. If the data or analytical output is displayed in some visual format instead as simple numbers, users can easily grab the meaning and work on it accordingly. This chapter explains the different use cases about the usage of information visualization techniques which are prominently used in the big data analytics field.

Chapter 13: *Social Media Analytics for Organization Empowerment*—This chapter highlights social media analytics which is one of the prominent technology use cases of big data analytics. One of the major drivers for big data is the huge amount of unstructured data which is generated by the social media networks. This has led to the evolution of a new stream of analytics which is called social media analytics. This chapter discusses the various drivers for the evolution of social media analytics. The various use cases which depict the usage of social media analytics for the transformation of organizations are discussed at length in this chapter. The content metrics which are used to track the impact of social media for organizations are also discussed at length in this chapter. Some key predictive analytic techniques which are used for social media analytics are network analysis and sentiment analysis using text mining. These two techniques are discussed in this chapter. Some of the tools which are used for social media analytics are also discussed in this chapter.

Chapter 14: *Big Data Analytics for Healthcare*—This chapter explains the primary importance of analytics in the healthcare sector. It is rightly said that the future of healthcare is the future for all of us. This chapter covers important driving factors for analytics in healthcare and the use cases for big data analytics in healthcare. The chapter sets an example on how data that got unnoticed in the past has proven to be a ray of hope to deliver quality care to patients in a cost-effective manner.

Disclaimer: The book and its contents are intended to convey the views of the authors and not their organizations.

Acknowledgment

The Acknowledgment (Pethuru Raj)

I express my sincere gratitude to Mr. **Simon Rees**, Springer, Associate Editor, Computer Science, for immensely helping us from the conceptualization to the completion of this book. I wholeheartedly acknowledge the fruitful suggestions and pragmatic contributions of my esteemed colleagues Anupama, Siddardh, and Dhivya. I need to remember my supervisors Prof. Ponnammal Natarajan, Anna University, Chennai; the late Prof. Priti Shankar, Computer Science and Automation (CSA) Department, Indian Institute of Science (IISc), Bangalore; Prof. Naohiro Ishii, Department of Intelligence and Computer Science, Nagoya Institute of Technology; and Prof. Kazuo Iwama, School of Informatics, Kyoto University, Japan, for shaping my research life. I wholeheartedly thank my managers at IBM in extending their moral support all along this book journey.

I, at this point of time, recollect and reflect on the selfless sacrifices made by my parents in shaping me up to this level. I would expressly like to thank my wife (Sweetlin Reena) and sons (Darren Samuel and Darresh Bernie) for their perseverance. I give all the glory and honor to my Lord and Savior Jesus Christ for His grace and guidance.

The Acknowledgment (Anupama Raman)

At the outset, I would like to express my heartfelt thanks to Mr. Simon Rees, Wayne Wheeler, and the Springer publishing team for their wholehearted support. My special thanks to Dr. Pethuru Raj for his constant support, guidance, and insights which helped me in crafting various chapters of this book. I would like to thank IBM management for their wholehearted support in the successful completion of this book project. At this stage I would also like to sincerely acknowledge the sacrifice of my

parents which made me what I am today. A special note of thanks to my husband (R. Murali Krishnan) and daughter (Aparna) for their constant support and motivation. I would also like to acknowledge the support given to me by my parents-in-law, my sisters, and their family. I thank all my friends who have constantly helped and supported me to complete the book successfully.

The Acknowledgment (Dhivya Nagaraj)

I would like to express my thanks to the editors and the publishing team of Springer Publications for their support. I cannot express enough my sincere appreciation to my mentor Anupama Murali for her majestic professional help in shaping the contents of this book and also to Dr. Pethuru Raj for all the encouragement and moral support.

I would like to extend my heartfelt thanks to my managers Sandip Singha and Prasenjit Bandhopadhyay and also the IBM management for being with me throughout this journey and helping me to aid the successful completion of the book. It is difficult to meet success without the support from our dear ones. This would be the right opportunity to thank my dear husband (Dr. Sudeep Gowda) and my parents for the unequaled and exceptional support in this travel. A special thanks to my daughter without whose support this wouldn't have been achievable and to all my dear friends for all their prayers.

The Acknowledgment (Siddhartha Duggirala)

On this note, I would like to express sincere thanks to each and everyone in the whole team for letting me be a part of this book and helping me out. My unfeigned thanks to Dr. Pethuru Raj and Anupama Raman for their guidance and support throughout the execution of this book.

I would like to thank each and every one from IIT Indore for being a part of my life. Especially Dr. Abhishek Srivatsava, Dr. Aruna Tiwari, and Dr. Siddharth Malu for believing in me and encouraging me to pursue whatever my heart bleeds for. I would like also to thank Sarfraz Qureshi my senior in IIT Indore and Dr. Kaushik Pavani my mentor who introduced me to the world of data.

I'd take this opportunity to thank my parents, brother, and cousins for helping me out, keeping me motivated, and showering their love. Without these great people, their support, and motivation, none of this would've been achievable. I am really grateful to all my friends for helping me and supporting me till the completion of the book. So, thanks everyone!