

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
© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

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.

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



Sushil Pramanick, FCA, PMP

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 unequalled 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!

Contents

1	The Brewing Trends and Transformations in the IT Landscape.....	1
1.1	Introduction	1
1.2	The Emerging IT Trends	2
1.3	The Realization and Blossoming of Digitalized Entities	6
1.4	The Internet of Things (IoT)/Internet of Everything (IoE)	8
1.5	The Tremendous Adoption of Social Media Sites.....	10
1.6	The Ensuring Era of Predictive, Prescriptive, and Personalized Analytics	11
1.7	Apache Hadoop for Big Data and Analytics	17
1.8	Big Data into Big Insights and Actions.....	20
1.9	Conclusions	23
1.10	Exercises	23
2	The High-Performance Technologies for Big and Fast Data Analytics	25
2.1	Introduction	25
2.2	The Emergence of Big Data Analytics (BDA) Discipline	27
2.3	The Strategic Implications of Big Data.....	28
2.4	The Big Data Analytics (BDA) Challenges	30
2.5	The High-Performance Computing (HPC) Paradigms	31
2.6	The High-Performance Approaches Through Parallelism	34
2.7	Cluster Computing	35
2.8	Grid Computing	38
2.9	Cloud Computing.....	43
2.10	Heterogeneous Computing.....	46
2.11	Mainframes for High-Performance Computing.....	49
2.12	Supercomputing for Big Data Analytics	51
2.13	Appliances for Big Data Analytics.....	51
2.13.1	Data Warehouse Appliances for Large-Scale Data Analytics	52
2.13.2	In-Memory Big Data Analytics.....	56
2.13.3	In-Database Processing of Big Data	58

- 2.13.4 Hadoop-Based Big Data Appliances..... 59
- 2.13.5 High-Performance Big Data Storage Appliances..... 63
- 2.14 Conclusions 65
- 2.15 Exercises 66
- References..... 66
- 3 Big and Fast Data Analytics Yearning for High-Performance Computing..... 67**
 - 3.1 Introduction 67
 - 3.2 A Relook on the Big Data Analytics (BDA) Paradigm..... 69
 - 3.3 The Implications of Big and Fast Data 72
 - 3.4 The Emerging Data Sources for Precise, Predictive, and Prescriptive Insights 74
 - 3.5 Why Big Data Analytics Is Strategically Sound? 77
 - 3.6 A Case Study for Traditional and New-Generation Data Analytics 79
 - 3.7 Why Cloud-Based Big Data Analytics?..... 84
 - 3.8 The Big Data Analytics: The Prominent Process Steps 87
 - 3.9 Real-Time Analytics..... 90
 - 3.10 Stream Analytics 95
 - 3.11 Sensor Analytics..... 96
 - 3.11.1 The Synchronization Between Big Data Analytics (BDA) and High-Performance Computing (HPC): The Value Additions..... 97
 - 3.12 Conclusions 98
 - 3.13 Exercises 99
- 4 Network Infrastructure for High-Performance Big Data Analytics 101**
 - 4.1 Introduction 101
 - 4.2 Network Infrastructure Limitations of Present-Day Networks 103
 - 4.3 Approaches for the Design of Network Infrastructures for High-Performance Big Data Analytics 105
 - 4.3.1 Network Virtualization 106
 - 4.3.2 Software-Defined Networking (SDN)..... 117
 - 4.3.3 Network Functions Virtualization (NFV)..... 120
 - 4.4 Wide Area Network (WAN) Optimization for Transfer of Big Data 122
 - 4.5 Conclusions 125
 - 4.6 Exercises 126
 - References..... 126
- 5 Storage Infrastructures for High-Performance Big Data Analytics 127**
 - 5.1 Introduction 127
 - 5.2 Getting Started with Storage Area Network..... 128
 - 5.2.1 Shortcomings of DAS 131

- 5.3 Getting Started with Storage Area Networks (SANs)..... 132
 - 5.3.1 Block-Level Access..... 132
 - 5.3.2 File-Level Access..... 132
 - 5.3.3 Object-Level Access..... 133
- 5.4 Storage Infrastructure Requirements for Storing Big Data..... 134
 - 5.4.1 Chapter Organization..... 135
- 5.5 Fiber Channel Storage Area Network (FC SAN)..... 136
- 5.6 Internet Protocol Storage Area Network (IP SAN)..... 137
 - 5.6.1 Fiber Channel Over Ethernet (FCoE)..... 138
- 5.7 Network-Attached Storage (NAS)..... 138
- 5.8 Popular File Systems Used for High-Performance
Big Data Analytics..... 139
 - 5.8.1 Google File System (GFS)..... 139
 - 5.8.2 Hadoop Distributed File System (HDFS)..... 142
 - 5.8.3 Panasas..... 143
 - 5.8.4 Luster File System..... 148
- 5.9 Introduction to Cloud Storage..... 150
 - 5.9.1 Architecture Model of a Cloud Storage System..... 150
 - 5.9.2 Storage Virtualization..... 153
 - 5.9.3 Storage Optimization Techniques Used
in Cloud Storage..... 157
 - 5.9.4 Advantages of Cloud Storage..... 158
- 5.10 Conclusions..... 159
- 5.11 Exercises..... 159
- References..... 159
 - Further Reading..... 159
- 6 Real-Time Analytics Using High-Performance Computing..... 161**
 - 6.1 Introduction..... 161
 - 6.2 Technologies That Support Real-Time Analytics..... 161
 - 6.2.1 Processing in Memory (PIM)..... 161
 - 6.2.2 In-Database Analytics..... 164
 - 6.3 MOA: Massive Online Analysis..... 166
 - 6.4 General Parallel File System (GPFS)..... 167
 - 6.4.1 GPFS: Use Cases..... 168
 - 6.5 GPFS Client Case Studies..... 174
 - 6.5.1 Broadcasting Company: VRT..... 174
 - 6.5.2 Oil Industry Migrates from Lustre to GPFS..... 176
 - 6.6 GPFS: Key Distinctions..... 177
 - 6.6.1 GPFS-Powered Solutions..... 177
 - 6.7 Machine Data Analytics..... 178
 - 6.7.1 Splunk..... 179
 - 6.8 Operational Analytics..... 180
 - 6.8.1 Technology for Operational Analytics..... 180
 - 6.8.2 Use Cases and Operational Analytics Products..... 181
 - 6.8.3 Other IBM Operational Analytics Products..... 183

- 6.9 Conclusions 184
- 6.10 Exercises 184
- 7 High-Performance Computing (HPC) Paradigms 187**
 - 7.1 Introduction 187
 - 7.2 Why Do We Still Need Mainframes???. 188
 - 7.3 How Has HPC Evolved Over the Years in Mainframes? 188
 - 7.3.1 Cost – An Important Factor for HPC 189
 - 7.3.2 Cloud Computing Centralized HPC..... 189
 - 7.3.3 Requirements to Centralized HPC 189
 - 7.4 HPC Remote Simulation..... 189
 - 7.5 Mainframe Solution Using HPC 190
 - 7.5.1 Intelligent Mainframe Grid 190
 - 7.5.2 How Does an IMG Work?..... 191
 - 7.5.3 IMG Architecture 192
 - 7.6 Architecture Models..... 195
 - 7.6.1 Storage Server with Shared Drive..... 195
 - 7.6.2 Storage Server Without Shared Drive 196
 - 7.6.3 Communication Network Without Storage Server..... 196
 - 7.7 SMP (Symmetric Multiprocessing)..... 197
 - 7.7.1 What Is SMP? 197
 - 7.7.2 SMP and Cluster Methods 197
 - 7.7.3 Is SMP Really Important and Why? 198
 - 7.7.4 Thread Model..... 198
 - 7.7.5 NumaConnect Technology 199
 - 7.8 Virtualization for HPC 199
 - 7.9 Innovation for Mainframes..... 200
 - 7.10 FICON Mainframe Interface..... 200
 - 7.11 Mainframe Mobile 201
 - 7.12 Windows High-Performance Computing 202
 - 7.13 Conclusions 204
 - 7.14 Exercises 205
- 8 In-Database Processing and In-Memory Analytics 207**
 - 8.1 Introduction 207
 - 8.1.1 Analytics Workload vs. Transaction Workload 208
 - 8.1.2 Evolution of Analytic Workload..... 209
 - 8.1.3 Traditional Analytic Platform..... 211
 - 8.2 In-Database Analytics 212
 - 8.2.1 Architecture..... 215
 - 8.2.2 Benefits and Limitations 216
 - 8.2.3 Representative Systems..... 217
 - 8.3 In-Memory Analytics 219
 - 8.3.1 Architecture..... 220
 - 8.3.2 Benefits and Limitations 221
 - 8.3.3 Representative Systems..... 222

- 8.4 Analytical Appliances 226
 - 8.4.1 Oracle Exalytics 227
 - 8.4.2 IBM Netezza 227
- 8.5 Conclusions 229
- 8.6 Exercises 229
- References 230
 - Further Reading 231
- 9 High-Performance Integrated Systems, Databases, and Warehouses for Big and Fast Data Analytics 233**
 - 9.1 Introduction 233
 - 9.2 The Key Characteristics of Next-Generation IT Infrastructures and Platforms 234
 - 9.3 Integrated Systems for Big and Fast Data Analytics 235
 - 9.3.1 The Urika-GD Appliance for Big Data Analytics 235
 - 9.3.2 IBM PureData System for Analytics (PDA) 237
 - 9.3.3 The Oracle Exadata Database Machine 238
 - 9.3.4 The Teradata Data Warehouse and Big Data Appliances 239
 - 9.4 Converged Infrastructure (CI) for Big Data Analytics 241
 - 9.5 High-Performance Analytics: Mainframes + Hadoop 243
 - 9.6 In-Memory Platforms for Fast Data Analytics 246
 - 9.7 In-Database Platforms for Big Data Analytics 249
 - 9.8 The Cloud Infrastructures for High-Performance Big and Fast Data Analytics 251
 - 9.9 Big File Systems for the Big Data World 256
 - 9.10 Databases and Warehouses for Big and Fast Data Analytics 258
 - 9.10.1 NoSQL Databases for Big Data Analytics 259
 - 9.10.2 NewSQL Databases for Big and Fast Data Analytics 263
 - 9.10.3 High-Performance Data Warehouses for Big Data Analytics 265
 - 9.11 Streaming Analytics 269
 - 9.12 Conclusions 274
 - 9.13 Exercises 274
- 10 High-Performance Grids and Clusters 275**
 - 10.1 Introduction 275
 - 10.2 Cluster Computing 278
 - 10.2.1 Motivation for Cluster Computing 278
 - 10.2.2 Cluster Computing Architecture 279
 - 10.2.3 Software Libraries and Programming Models 282
 - 10.2.4 Advanced Cluster Computing Systems 292
 - 10.2.5 Difference Between Grid and Cluster 293
 - 10.3 Grid Computing 294
 - 10.3.1 Motivation for Grid Computing 295
 - 10.3.2 Evolution of Grid Computing 297
 - 10.3.3 Design Principles and Goals of a Grid System 298

- 10.3.4 Grid System Architecture 300
- 10.3.5 Benefits and Limitations of a Grid Computing System . 304
- 10.3.6 Grid Systems and Applications 305
- 10.3.7 Future of Grid Computing..... 311
- 10.4 Conclusions 313
- 10.5 Exercises 313
- References..... 314
 - Further Reading..... 315
- 11 High-Performance Peer-to-Peer Systems 317**
 - 11.1 Introduction 317
 - 11.2 Design Principles and Characteristics 319
 - 11.3 Peer-to-Peer System Architectures..... 320
 - 11.3.1 Centralized Peer-to-Peer Systems 320
 - 11.3.2 Decentralized Peer-to-Peer Systems 322
 - 11.3.3 Hybrid Peer-to-Peer Systems 324
 - 11.3.4 Advanced Peer-to-Peer Architecture
Communication Protocols and Frameworks 326
 - 11.4 High-Performance Peer-to-Peer Applications..... 328
 - 11.4.1 Cassandra 329
 - 11.4.2 SETI@Home..... 331
 - 11.4.3 Bitcoin: Peer-to-Peer-Based Digital Currency 333
 - 11.5 Conclusions 334
 - 11.6 Exercises 335
 - References..... 335
 - Further Reading..... 337
- 12 Visualization Dimensions for High-Performance**
- Big Data Analytics 339**
 - 12.1 Introduction 339
 - 12.2 Common Techniques..... 344
 - 12.2.1 Charts 345
 - 12.2.2 Scatter Plot 347
 - 12.2.3 Treemap..... 348
 - 12.2.4 Box Plot..... 349
 - 12.2.5 Infographics..... 349
 - 12.2.6 Heat Maps 350
 - 12.2.7 Network/Graph Visualization..... 352
 - 12.2.8 Word Cloud/Tag Cloud 353
 - 12.3 Data Visualization Tools and Systems 353
 - 12.3.1 Tableau 353
 - 12.3.2 BIRST 355
 - 12.3.3 Roambi 357
 - 12.3.4 QlikView 359
 - 12.3.5 IBM Cognos 360

12.3.6	Google Charts and Fusion Tables.....	361
12.3.7	Data-Driven Documents (D3.js)	361
12.3.8	Sisense.....	362
12.4	Conclusions	362
12.5	Exercises	363
References	364
Further Reading	365
13	Social Media Analytics for Organization Empowerment	367
13.1	Introduction	367
13.1.1	Social Data Collection.....	368
13.1.2	Social Data Analysis	368
13.1.3	Proliferation of Mobile Devices.....	369
13.1.4	Robust Visualization Mechanisms	369
13.1.5	Rapid Change in the Nature of Data	370
13.2	Getting Started with Social Media Analytics	371
13.2.1	Chapter Organization	373
13.3	Building a Framework for the Use of Social Media Analytics for Business.....	373
13.4	Social Media Content Metrics.....	374
13.5	Predictive Analytic Techniques for Social Media Analytics	376
13.6	Architecture for Sentiment Analysis Using Text Mining.....	377
13.7	Network Analysis on Social Media Data	379
13.7.1	Getting Started with Network Analysis of Social Media Data.....	380
13.7.2	Network Analysis Using Twitter	381
13.7.3	Polarized Network Map	381
13.7.4	In-Group Map.....	382
13.7.5	The Twitter Brand Map	382
13.7.6	Bazaar Networks	382
13.7.7	Broadcast Map	383
13.7.8	Support Network Maps	383
13.8	Different Dimensions of Social Media Analytics for Organizations.....	383
13.8.1	Revenue and Sales Lead Generation	386
13.8.2	Customer Relationship and Customer Experience Management.....	387
13.8.3	Innovation.....	387
13.9	Social Media Tools.....	388
13.9.1	Social Media Monitoring Tools.....	388
13.9.2	Social Media Analytics Tools.....	389
13.10	Conclusions	389
13.11	Exercises	390
Reference	390

14 Big Data Analytics for Healthcare 391

14.1 Introduction 391

14.2 Market Factors Affecting Healthcare 393

14.3 Different Shareholders Envisaging Different Objectives 394

14.4 Big Data Benefits to Healthcare 395

 14.4.1 Healthcare Efficiency and Quality 396

 14.4.2 Earlier Disease Detection 397

 14.4.3 Fraud Detection 397

 14.4.4 Population Health Management 398

14.5 Big Data Technology Adoptions: A New Amelioration! 400

 14.5.1 IBM Watson 400

 14.5.2 IBM Watson Architecture 401

14.6 Watson in Healthcare 401

 14.6.1 WellPoint and IBM 401

14.7 EHR Technology 402

 14.7.1 EHR Data Flow 403

 14.7.2 Advantages of EHR 403

14.8 Remote Monitoring and Sensing 404

 14.8.1 Technological Components 404

 14.8.2 Healthcare Areas Where Remote Monitoring Is Applied 405

 14.8.3 Limitations of Remote Monitoring 405

14.9 High-Performance Computing for Healthcare 406

14.10 Real-Time Analysis of Human Brain Networks 406

14.11 Detection of Cancer 407

14.12 3D Medical Image Segmentation 408

14.13 New Medical Therapies 408

14.14 Use Cases for BDA in Healthcare 409

14.15 Population Health Control 409

14.16 Care Process Management 410

 14.16.1 Core IT Functionality 411

14.17 Hadoop Use Cases 412

14.18 Big Data Analytics: Success Stories 414

14.19 Opportunities for BDA in Healthcare 416

14.20 Member 360 416

14.21 Genomics 418

14.22 Clinical Monitoring 419

14.23 Economic Value for BDA in Healthcare 420

14.24 Big Data Challenges for Healthcare 421

14.25 Future of Big Data in Healthcare 421

14.26 Conclusions 421

14.27 Exercises 423

Index 425