Contents

List of Contributors xv
About the Editors xix
Preface xxi
Acknowledgments xxv

PART I BIG DATA SCIENCE

CHAPTER 1 BDA = ML + CC 3
- 1.1 Introduction 3
- 1.2 A Historical Review of Big Data 4
  - 1.2.1 The Origin of Big Data 4
  - 1.2.2 Debates of Big Data Implication 5
- 1.3 Historical Interpretation of Big Data 7
  - 1.3.1 Methodology for Defining Big Data 7
  - 1.3.2 Different Attributes of Definitions 7
  - 1.3.3 Summary of 7 Types Definitions of Big Data 10
  - 1.3.4 Motivations Behind the Definitions 10
- 1.4 Defining Big Data From 3Vs to 3^2Vs 11
  - 1.4.1 Data Domain 11
  - 1.4.2 Business Intelligent (BI) Domain 11
  - 1.4.3 Statistics Domain 13
  - 1.4.4 3^2 Vs Definition and Big Data Venn Diagram 13
- 1.5 Big Data Analytics and Machine Learning 14
  - 1.5.1 Big Data Analytics 14
  - 1.5.2 Machine Learning 15
- 1.6 Big Data Analytics and Cloud Computing 18
- 1.7 Hadoop, HDFS, MapReduce, Spark, and Flink 18
  - 1.7.1 Google File System (GFS) and HDFS 20
  - 1.7.2 MapReduce 24
  - 1.7.3 The Origin of the Hadoop Project 25
  - 1.7.4 Spark and Spark Stack 27
  - 1.7.5 Flink and Other Data Process Engines 27
  - 1.7.6 Summary of Hadoop and Its Ecosystems 32
- 1.8 ML + CC → BDA and Guidelines 34
- 1.9 Conclusion 35
- References 35

CHAPTER 2 Real-Time Analytics 39
2.1 Introduction 39
2.2 Computing Abstractions for Real-Time Analytics 40
2.3 Characteristics of Real-Time Systems 41
2.3.1 Low Latency 42
2.3.2 High Availability 42
2.3.3 Horizontal Scalability 43
2.4 Real-Time Processing for Big Data — Concepts and Platforms 43
2.4.1 Event 43
2.4.2 Event Processing 44
2.4.3 Event Stream Processing and Data Stream Processing 44
2.4.4 Complex Event Processing 44
2.4.5 Event Type 45
2.4.6 Event Pattern 45
2.5 Data Stream Processing Platforms 45
2.5.1 Spark 46
2.5.2 Storm 47
2.5.3 Kafka 47
2.5.4 Flume 48
2.5.5 Amazon Kinesis 48
2.6 Data Stream Analytics Platforms 48
2.6.1 Query-Based EPSs 48
2.6.2 Rule-Oriented EPSs 49
2.6.3 Programmatic EPSs 50
2.7 Data Analysis and Analytic Techniques 53
2.7.1 Data Analysis in General 53
2.7.2 Data Analysis for Stream Applications 53
2.8 Finance Domain Requirements and a Case Study 54
2.8.1 Real-Time Analytics in Finance Domain 54
2.8.2 Selected Scenarios 55
2.8.3 CEP Application as a Case Study 55
2.9 Future Research Challenges 58
References 59
CHAPTER 3 Big Data Analytics for Social Media 63
3.1 Introduction 63
3.2 NLP and Its Applications 63
3.2.1 Language Detection 64
3.2.2 Named Entity Recognition 68
3.3 Text Mining 72

3.3.1 Sentiment Analysis 72
3.3.2 Trending Topics 77
3.3.3 Recommender Systems 81
3.4 Anomaly Detection 85
Acknowledgments 88
References 89

CHAPTER 4 Deep Learning and Its Parallelization 95
4.1 Introduction 95
4.1.1 Application Background 95
4.1.2 Performance Demands for Deep Learning 96
4.1.3 Existing Parallel Frameworks of Deep Learning 96
4.2 Concepts and Categories of Deep Learning 96
4.2.1 Deep Learning 96
4.2.2 Mainstream Deep Learning Models 99
4.3 Parallel Optimization for Deep Learning 104
4.3.1 Convolutional Architecture for Fast Feature Embedding 104
4.3.2 DistBelief 111
4.3.3 Deep Learning Based on Multi-GPUs 112
4.4 Discussions 115
4.4.1 Grand Challenges of Deep Learning in Big Data 115
4.4.2 Future Directions 116
References 117

CHAPTER 5 Characterization and Traversal of Large Real-World Networks 119
5.1 Introduction 119
5.2 Background 120
5.3 Characterization and Measurement 121
5.4 Efficient Complex Network Traversal 124
5.4.1 HPC Traversal of Large Networks 124
5.4.2 Algorithms for Accelerating AS-BFS on GPU 125
5.4.3 Performance Study of AS-BFS on GPU's 126
5.5 *k*-Core-Based Partitioning for Heterogeneous Graph Processing 128
5.5.1 Graph Partitioning for Heterogeneous Computing 129
5.5.2 *k*-Core-Based Complex-Network Unbalanced Bisection 129
5.6 Future Directions 133
5.7 Conclusions 133
Acknowledgments 134
References 134

PART II BIG DATA INFRASTRUCTURES AND PLATFORMS

CHAPTER 6 Database Techniques for Big Data 139

6.1 Introduction 139

6.2 Background 139

6.2.1 Navigational Data Models 139

6.2.2 Relational Data Models 140

6.3 NoSQL Movement 143

6.4 NoSQL Solutions for Big Data Management 144

6.5 NoSQL Data Models 150

6.5.1 Key-Value Stores 150

6.5.2 Column-Based Stores 151

6.5.3 Graph-Based Stores 153

6.5.4 Document-Based Stores 154

6.6 Future Directions 156

6.7 Conclusions 157

References 157

CHAPTER 7 Resource Management in Big Data Processing Systems 161

7.1 Introduction 161

7.2 Types of Resource Management 162

7.2.1 CPU and Memory Resource Management 162

7.2.2 Storage Resource Management 163

7.2.3 Network Resource Management 163

7.3 Big Data Processing Systems and Platforms 163

7.3.1 Hadoop 163

7.3.2 Dryad 164

7.3.3 Pregel 164

7.3.4 Storm 164

7.3.5 Spark 165

7.3.6 Summary 165

7.4 Single-Resource Management in the Cloud 166

7.4.1 Desired Resource Allocation Properties 166

7.4.2 Problems for Existing Fairness Policies 167

7.4.3 Long-Term Resource Allocation Policy 168

7.4.4 Experimental Evaluation 170

7.5 Multiresource Management in the Cloud 171

7.5.1 Resource Allocation Model 172

7.5.2 Multiresource Fair Sharing Issues 174

7.5.3 Reciprocal Resource Fairness 175

7.5.4 Experimental Evaluation 179

139
139
139
139
140
143
144
150
150
151
153
154
156
157
157
161
161
162
162
163
163
163
163
164
164
164
165
165
166
166
167
168
170
171
172
174
175
179

7.6 Related Work on Resource Management 182
7.6.1 Resource Utilization Optimization 182
7.6.2 Power and Energy Cost Saving Optimization 182
7.6.3 Monetary Cost Optimization 182
7.6.4 Fairness Optimization 183
7.7 Open Problems 183
7.7.1 SLA Guarantee for Applications 183
7.7.2 Various Computation Models and Systems 183
7.7.3 Exploiting Emerging Hardware 184
7.8 Summary 184
References 184

CHAPTER 8 Local Resource Consumption Shaping: A Case for MapReduce 189
8.1 Introduction 189
8.2 Motivation 191
8.2.1 Pitfalls of Fair Resource Sharing 192
8.3 Local Resource Shaper 194
8.3.1 Design Philosophy 194
8.3.2 Splitter 195
8.3.3 The Interleave MapReduce Scheduler 195
8.4 Evaluation 198
8.4.1 Experiments With Hadoop 1.x 198
8.4.2 Experiments With Hadoop 2.x 204
8.5 Related Work 210
8.6 Conclusions 211
Appendix CPU Utilization With Different Slot Configurations and LRS 212
References 213

CHAPTER 9 System Optimization for Big Data Processing 215
9.1 Introduction 215
9.2 Basic Framework of the Hadoop Ecosystem 217
9.3 Parallel Computation Framework: MapReduce 218
9.3.1 Improvements of MapReduce Framework 218
9.3.2 Optimization for Task Scheduling and Load Balancing of MapReduce 219
9.4 Job Scheduling of Hadoop 220
9.4.1 Built-In Scheduling Algorithms of Hadoop 220
9.4.2 Improvement of the Hadoop Job Scheduling Algorithm 221
9.4.3 Improvement of the Hadoop Job Management Framework 223
9.5 Performance Optimization of HDFS 224
9.5.1 Small File Performance Optimization 224
9.5.2 HDFS Security Optimization 226

9.6 Performance Optimization of HBase 228
9.6.1 HBase Framework, Storage, and Application Optimization 228
9.6.2 Load Balancing of HBase 229
9.6.3 Optimization of HBase Configuration 230
9.7 Performance Enhancement of Hadoop System 230
9.7.1 Efficiency Optimization of Hadoop 231
9.7.2 Availability Optimization of Hadoop 232
9.8 Conclusions and Future Directions 233
References 233

CHAPTER 10 Packing Algorithms for Big Data Replay on Multicore 239
10.1 Introduction 239
10.2 Performance Bottlenecks 241
10.2.1 Hadoop/MapReduce Performance Bottlenecks 241
10.2.2 Performance Bottlenecks Under Parallel Loads 243
10.2.3 Parameter Spaces for Storage and Shared Memory 244
10.2.4 Main Storage Performance 245
10.2.5 Shared Memory Performance 248
10.3 The Big Data Replay Method 250
10.3.1 The Replay Method 250
10.3.2 Jobs as Sketches on a Timeline 251
10.3.3 Performance Bottlenecks Under Replay 252
10.4 Packing Algorithms 253
10.4.1 Shared Memory Performance Tricks 253
10.4.2 Big Data Replay at Scale 255
10.4.3 Practical Packing Models 256
10.5 Performance Analysis 256
10.5.1 Hotspot Distributions 256
10.5.2 Modeling Methodology 258
10.5.3 Processing Overhead Versus Bottlenecks 259
10.5.4 Control Grain for Drop Versus Drag Models 261
10.6 Summary and Future Directions 262
References 264

PART III BIG DATA SECURITY AND PRIVACY

CHAPTER 11 Spatial Privacy Challenges in Social Networks 269
11.1 Introduction 269
11.2 Background 269
11.3 Spatial Aspects of Social Networks 271

..............228
..............228
..............229
..............230
..............230
..............231
..............232
..............233
..............233

239
..............239
..............241
..............241
..............243
..............244
..............245
..............248
..............250
..............250
..............251
..............252
..............253
..............253
..............255
..............256
..............256
..............256
..............258
..............259
..............261
..............262
..............264

269
..............269
..............269
..............271

11.4 Cloud-Based Big Data Infrastructure 273
11.5 Spatial Privacy Case Studies 275
11.6 Conclusions 281
Acknowledgments 282
References 282

CHAPTER 12 Security and Privacy in Big Data **285**
12.1 Introduction 285
12.2 Secure Queries Over Encrypted Big Data 287
12.2.1 System Model 287
12.2.2 Threat Model and Attack Model 288
12.2.3 Secure Query Scheme in Clouds 289
12.2.4 Security Definition of Index-Based Secure Query Techniques 291
12.2.5 Implementations of Index-Based Secure Query Techniques 291
12.3 Other Big Data Security 295
12.3.1 Digital Watermarking 295
12.3.2 Self-Adaptive Risk Access Control 296
12.4 Privacy on Correlated Big Data 296
12.4.1 Correlated Data in Big Data 296
12.4.2 Anonymity 298
12.4.3 Differential Privacy 300
12.5 Future Directions 304
12.6 Conclusions 305
References 305

CHAPTER 13 Location Inferring in Internet of Things and Big Data **309**
13.1 Introduction 309
13.2 Device-Based Sensing Using Big Data 310
13.2.1 Introduction 310
13.2.2 Approach Overview 310
13.2.3 Trajectories Matching 311
13.2.4 Establishing the Mapping Between Floor Plan and RSS Readings 314
13.2.5 User Localization 318
13.2.6 Graph Matching Based Tracking 318
13.2.7 Evaluation 318
13.3 Device-Free Sensing Using Big Data 319
13.3.1 Customer Behavior Identification 319
13.3.2 Human Object Estimation 328
13.4 Conclusion 334
Acknowledgements 334
References 334

PART IV BIG DATA APPLICATIONS

CHAPTER 14 A Framework for Mining Thai Public Opinions 339

14.1 Introduction 339

14.2 XDOM 340

14.2.1 Data Sources 340

14.2.2 DOM System Architecture 341

14.2.3 MapReduce Framework 342

14.2.4 Sentiment Analysis 343

14.2.5 Clustering-Based Summarization Framework 344

14.2.6 Influencer Analysis 349

14.2.7 AskDOM: Mobile Application 350

14.3 Implementation 350

14.3.1 Server 350

14.3.2 Core Service 351

14.3.3 I/O 351

14.4 Validation 352

14.4.1 Validation Parameter 352

14.4.2 Validation method 352

14.4.3 Validation results 352

14.5 Case Studies 353

14.5.1 Political Opinion: #prayforthailand 353

14.5.2 Bangkok Traffic Congestion Ranking 353

14.6 Summary and Conclusions 354

Acknowledgments 354

References 355

CHAPTER 15 A Case Study in Big Data Analytics: Exploring Twitter Sentiment Analysis and the Weather 357

15.1 Background 357

15.2 Big Data System Components 358

15.2.1 System Back-End Architecture 358

15.2.2 System Front-End Architecture 359

15.2.3 Software Stack 360

15.3 Machine-Learning Methodology 360

15.3.1 Tweets Sentiment Analysis 361

15.3.2 Weather and Emotion Correlation Analysis 371

15.4 System Implementation 373

15.4.1 Home Page 373

15.4.2 Sentiment Pages 374

15.4.3 Weather Pages 374

.............339
.............339
.............340
.............340
.............341
.............342
.............343
.............344
.............349
.............350
.............350
.............350
.............351
.............351
.............352
.............352
.............352
.............352
.............353
.............353
.............353
.............354
.............354
.............355

.............357
.............357
.............358
.............358
.............359
.............360
.............360
.............361
.............371
.............373
.............373
.............374
.............374

15.5 Key Findings....................378
15.5.1 Time Series....................378
15.5.2 Analysis with Hourly Weather Data....................378
15.5.3 Analysis with Daily Weather Data....................380
15.5.4 DBSCAN Cluster Algorithm....................382
15.5.5 Straightforward Weather Impact on Emotion....................383
15.6 Summary and Conclusions....................384
Acknowledgments....................387
References....................387

CHAPTER 16 Dynamic Uncertainty-Based Analytics for Caching Performance Improvements in Mobile Broadband Wireless Networks....................**389**
16.1 Introduction....................389
16.1.1 Big Data Concerns....................391
16.1.2 Key Focus Areas....................391
16.2 Background....................392
16.2.1 Cellular Network and VoD....................392
16.2.2 Markov Processes....................393
16.3 Related Work....................395
16.4 VoD Architecture....................396
16.5 Overview....................398
16.6 Data Generation....................399
16.7 Edge and Core Components....................400
16.8 INCA Caching Algorithm....................401
16.9 QoE Estimation....................403
16.10 Theoretical Framework....................403
16.11 Experiments and Results....................404
16.11.1 Cache Hits With $\mathbf{N_U}$, $\mathbf{N_C}$, $\mathbf{N_M}$ and **k**....................405
16.11.2 QoE Impact With Prefetch Bandwidth....................407
16.11.3 User Satisfaction With Prefetch Bandwidth....................409
16.12 Synthetic Dataset....................409
16.12.1 INCA Hit Gain....................410
16.12.2 QoE Performance....................410
16.12.3 Satisfied Users....................412
16.13 Conclusions and Future Directions....................413
References....................414

CHAPTER 17 Big Data Analytics on a Smart Grid: Mining PMU Data for Event and Anomaly Detection....................**417**
17.1 Introduction....................417
17.2 Smart Grid With PMUs and PDCs....................418

17.3 Improving Traditional Workflow....418
17.4 Characterizing Normal Operation....419
17.5 Identifying Unusual Phenomena....420
17.6 Identifying Known Events....423
17.7 Related Efforts....426
17.8 Conclusion and Future Directions....427
Acknowledgments....428
References....428

CHAPTER 18 eScience and Big Data Workflows in Clouds: A Taxonomy and Survey....431
18.1 Introduction....431
18.2 Background....432
18.2.1 History....432
18.2.2 Grid-Based eScience....434
18.2.3 Cloud Computing....435
18.3 Taxonomy and Review of eScience Services in the Cloud....436
18.3.1 Infrastructure....437
18.3.2 Ownership....437
18.3.3 Application....438
18.3.4 Processing Tools....439
18.3.5 Storage....439
18.3.6 Security....440
18.3.7 Service Models....441
18.3.8 Collaboration....441
18.4 Resource Provisioning for eScience Workflows in Clouds....442
18.4.1 Motivation....442
18.4.2 Our Solution....445
18.5 Open Problems....451
18.6 Summary....452
References....452

Index....457

List c

T. Achalak
King Mong

P. Ameri
Karlsruhe I

A. Berry
Deontik, Br

N. Bojja
Machine Zo

R. Buyya
The Univer
Australia

W. Chen
University o

C. Deeroseja
King Mongk

A. Diaz-Pere
Cinvestav-Ta

H. Ding
Xi'an Jiaoton

X. Dong
Huazhong U

H. Duan
The Univers

S. Dutta
Max Planck

A. Garcia-Ro
Cinvestav-Ta

V. Gramoli
University of

X. Gu
Huazhong U

J. Han
Xi'an Jiaotong

B. He
Nanyang Tec