

Contents

Part I Big Data Technologies

1	Introduction to Big Data	3
	Borko Furht and Flavio Villanustre	
	Concept of Big Data	3
	Big Data Workflow	4
	Big Data Technologies	5
	Big Data Layered Architecture	5
	Big Data Software	6
	Splunk	6
	LexisNexis' High-Performance Computer Cluster (HPCC)	6
	Big Data Analytics Techniques	7
	Clustering Algorithms for Big Data	8
	Big Data Growth	9
	Big Data Industries	9
	Challenges and Opportunities with Big Data	10
	References	11
2	Big Data Analytics	13
	Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos	
	Introduction	14
	Data Analytics	16
	Data Input	17
	Data Analysis	17
	Output the Result	19
	Summary	22
	Big Data Analytics	24
	Big Data Input	25
	Big Data Analysis Frameworks and Platforms	26
	Researches in Frameworks and Platforms	27
	Comparison Between the Frameworks/Platforms of Big Data	30

Big Data Analysis Algorithms	31
Mining Algorithms for Specific Problem	31
Machine Learning for Big Data Mining	33
Output the Result of Big Data Analysis	36
Summary of Process of Big Data Analytics	37
The Open Issues	40
Platform and Framework Perspective	40
Input and Output Ratio of Platform	40
Communication Between Systems	40
Bottlenecks on Data Analytics System	41
Security Issues	41
Data Mining Perspective	42
Data Mining Algorithm for Map-Reduce Solution	42
Noise, Outliers, Incomplete and Inconsistent Data	42
Bottlenecks on Data Mining Algorithm	43
Privacy Issues	43
Conclusions	44
References	45
3 Transfer Learning Techniques	53
Karl Weiss, Taghi M. Khoshgoftaar and DingDing Wang	
Introduction	53
Definitions of Transfer Learning	55
Homogeneous Transfer Learning	59
Instance-Based Transfer Learning	60
Asymmetric Feature-Based Transfer Learning	61
Symmetric Feature-Based Transfer Learning	64
Parameter-Based Transfer Learning	68
Relational-Based Transfer Learning	70
Hybrid-Based (Instance and Parameter) Transfer Learning	71
Discussion of Homogeneous Transfer Learning	72
Heterogeneous Transfer Learning	73
Symmetric Feature-Based Transfer Learning	74
Asymmetric Feature-Based Transfer Learning	79
Improvements to Heterogeneous Solutions	82
Experiment Results	83
Discussion of Heterogeneous Solutions	83
Negative Transfer	85
Transfer Learning Applications	88
Conclusion and Discussion	90
Appendix	92
References	93

4 Visualizing Big Data 101
Ekaterina Olshannikova, Aleksandr Ometov, Yevgeni Koucheryav
and Thomas Olsson
Introduction 101
Big Data: An Overview 103
Big Data Processing Methods 104
Big Data Challenges 107
Visualization Methods 109
Integration with Augmented and Virtual Reality 119
Future Research Agenda and Data Visualization Challenges 121
Conclusion 123
References 124

5 Deep Learning Techniques in Big Data Analytics 133
Maryam M. Najafabadi, Flavio Villanustre, Taghi M. Khoshgoftaar,
Naeem Seliya, Randall Wald and Edin Muharemagc
Introduction 133
Deep Learning in Data Mining and Machine Learning 136
Big Data Analytics 138
Applications of Deep Learning in Big Data Analytics 140
Semantic Indexing 141
Discriminative Tasks and Semantic Tagging 144
Deep Learning Challenges in Big Data Analytics 147
 Incremental Learning for Non-stationary Data 147
 High-Dimensional Data 148
 Large-Scale Models 149
Future Work on Deep Learning in Big Data Analytics 150
Conclusion 152
References 153

Part II LexisNexis Risk Solution to Big Data

6 The HPCC/ECL Platform for Big Data 159
Anthony M. Middleton, David Alan Bayliss,
Gavin Halliday, Arjuna Chala and Borko Furht
Introduction 159
 Data-Intensive Computing Applications 160
 Data-Parallelism 161
 The “Big Data” Problem 161
Data-Intensive Computing Platforms 162
 Cluster Configurations 162
 Common Platform Characteristics 163
HPCC Platform 164
 HPCC System Architecture 164
 HPCC Thor System Cluster 167
 HPCC Roxie System Cluster 169

- ECL Programming Language 170
 - ECL Features and Capabilities 171
 - ECL Compilation, Optimization, and Execution 173
 - ECL Development Tools and User Interfaces 177
 - ECL Advantages and Key Benefits 177
- HPCC High Reliability and High Availability Features 179
- Conclusion 180
- References 182

- 7 Scalable Automated Linking Technology for Big Data**
- Computing** 185
 - Anthony M. Middleton, David Bayliss and Bob Foreman
 - Introduction 185
 - SALT—Basic Concepts 186
 - SALT Process 187
 - Specification File Language 188
 - SALT—Applications 195
 - Data Profiling 196
 - Data Hygiene 199
 - Data Source Consistency Checking 201
 - Delta File Comparison 202
 - Data Ingest 202
 - Record Linkage—Process 205
 - Record Matching Field Weight Computation 206
 - Generating Specificities 208
 - Internal Linking 209
 - External Linking 213
 - Base File Searching 218
 - Remote Linking 219
 - Attribute Files 220
 - Summary and Conclusions 220
 - References 222

- 8 Aggregated Data Analysis in HPCC Systems** 225
 - David Bayliss
 - Introduction 225
 - The RDBMS Paradigm 226
 - The Reality of SQL 227
 - Normalizing an Abnormal World 228
 - A Data Centric Approach 230
 - Data Analysis 232
 - Case Study: Fuzzy Matching 233
 - Case Study: Non-obvious Relationship Discovery 234
 - Conclusion 235

- 9 Models for Big Data** 237
 - David Bayliss
 - Structures Data. 237
 - Text (and HTML) 241
 - Semi-structures Data. 242
 - Bridging the Gap—The Key-Value Pair 243
 - XML—Structured Text 244
 - RDF. 246
 - Data Model Summary. 247
 - Data Abstraction—An Alternative Approach 247
 - Structured Data 248
 - Text 249
 - Semi-structured Data. 249
 - Key-Value Pairs. 250
 - XML 251
 - RDF. 252
 - Model Flexibility in Practice 253
 - Conclusion 255

- 10 Data Intensive Supercomputing Solutions.** 257
 - Anthony M. Middleton
 - Introduction. 257
 - Data-Intensive Computing Applications. 259
 - Data-Parallelism. 260
 - The “Data Gap”. 260
 - Characteristics of Data-Intensive Computing Systems 261
 - Processing Approach 262
 - Common Characteristics 263
 - Grid Computing 264
 - Data-Intensive System Architectures 265
 - Google MapReduce 265
 - Hadoop. 269
 - LexisNexis HPCC 273
 - Programming Language ECL. 279
 - Hadoop Versus HPCC Comparison 282
 - Terabyte Sort Benchmark 283
 - Pig Versus ECL. 285
 - Architecture Comparison. 287
 - Conclusion 303
 - References. 305

- 11 Graph Processing with Massive Datasets: A Kel Primer** 307
 - David Bayliss and Flavio Villanustre
 - Introduction. 307
 - Motivation. 308

Background	309
The Open Source HPCC Systems Platform Architecture	309
KEL—Knowledge Engineering Language for Graph Problems.	309
KEL—A Primer	310
Proposed Solution	313
Data Primitives with Graph Primitive Extensions	313
Generated Code and Graph Libraries	315
KEL Compiler	316
KEL Language—Principles	316
KEL Language—Syntax	318
KEL—The Summary	323
KEL Present and Future	328
References.	328
Part III Big Data Applications	
12 HPCC Systems for Cyber Security Analytics	331
Flavio Villanustre and Mauricio Renzi	
The Advanced Persistent Threat	332
LexisNexis HPPS Systems for Deep Forensic Analysis	335
Pre-computed Analytics for Cyber Security	335
The Benefits of Pre-computed Analytics	337
Deep Forensics Analysis	338
Conclusion	339
13 Social Network Analytics: Hidden and Complex	
Fraud Schemes	341
Flavio Villanustre and Borko Furht	
Introduction	341
Case Study: Insurance Fraud	341
Case Study: Fraud in Prescription Drugs	341
Case Study: Fraud in Medicaid	342
Case Study: Network Traffic Analysis.	343
Case Study: Property Transaction Risk	346
14 Modeling Ebola Spread and Using HPCC/KEL System	347
Jesse Shaw, Flavio Villanustre, Borko Furht, Ankur Agarwal and Abhishek Jain	
Introduction	347
Survey of Ebola Modeling Techniques	349
Basic Reproduction Number (R_0)	349
Case Fatality Rate (CFR)	350
SIR Model	351
Improved SIR (ISIR) Model	352
SIS Model.	353
SEIZ Model	353

- Agent-Based Model 355
- A Contact Tracing Model 357
- Spatiotemporal Spread of 2014 Outbreak
of Ebola Virus Disease 360
- Quarantine Model 361
- Global Epidemic and Mobility Model 362
- Other Critical Issues in Ebola Study 364
 - Delays in Outbreak Detection 364
 - Lack of Public Health Infrastructure 365
 - Health Worker Infections 366
 - Misinformation Propagation in Social Media 367
- Risk Score Approach in Modeling and Predicting Ebola Spread 368
 - Beyond Compartmental Modeling 368
 - Physical and Social Graphs 369
 - Graph Knowledge Extraction 369
 - Graph Propagation 370
- Mobile Applications Related to Ebola Virus Disease 373
 - ITU Ebola—Info—Sharing 373
 - Ebola Prevention App 373
 - Ebola Guidelines 373
 - About Ebola 374
 - Stop Ebola WHO Official 374
 - HealthMap 374
 - #ISurvivedEbola 374
 - Ebola Report Center 374
 - What is Ebola 375
 - Ebola 375
 - Stop Ebola 375
 - Virus Tracker 375
 - Ebola Virus News Alert 376
 - Sierra Leone Ebola Trends 376
 - The Virus Ebola 376
 - MSF Guidance 376
 - Novarum Reader 376
 - Work Done by Government 378
- Innovative Mobile Application for Ebola Spread 378
 - Registering a New User 379
 - Login the Application 380
 - Basic Information 380
 - Geofencing 380
 - Web Service Through ECL 382
- Conclusion 383
- References 384

- 15 Unsupervised Learning and Image Classification in High Performance Computing Cluster 387**
 - I. Itauma, M.S. Aslan, X.W. Chen and Flavio Villanustre
 - Introduction 387
 - Background and Advantages of HPCC Systems^R 388
 - Contributions 389
 - Methods 390
 - Image Reading in HPCC Systems Platform 390
 - Feature Learning 391
 - Feature Extraction 393
 - Classification 393
 - Experiments and Results 393
 - Discussion 398
 - Conclusion 398
 - References 399



<http://www.springer.com/978-3-319-44548-9>

Big Data Technologies and Applications

Furht, B.; Villanustre, F.

2016, XVIII, 400 p. 118 illus., Hardcover

ISBN: 978-3-319-44548-9