

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