Contents

Acknowledgments xv
About the Author xvii
Introduction xix

PART 1 BIG DATA

CHAPTER 1 Introduction to Big Data 3

Introduction 3
Big Data 3
Defining Big Data 5
Why Big Data and why now? 5
Big Data example 6
  Social Media posts 6
  Survey data analysis 7
  Survey data 8
  Weather data 11
  Twitter data 11
  Integration and analysis 11
  Additional data types 13
Summary 14
Further reading 14

CHAPTER 2 Working with Big Data 15

Introduction 15
Data explosion 15
Data volume 17
  Machine data 17
  Application log 18
  Clickstream logs 18
  External or third-party data 18
  Emails 18
  Contracts 19
  Geographic information systems and geo-spatial data 19
  Example: Funshots, Inc 21
Data velocity 23
  Amazon, Facebook, Yahoo, and Google 24
  Sensor data 24
  Mobile networks 24
  Social media 24

Data variety 25

Summary 27

CHAPTER 3 Big Data Processing Architectures 29

Introduction 29

Data processing revisited 29

Data processing techniques 30

Data processing infrastructure challenges 31

Storage 31

Transportation 32

Processing 32

Speed or throughput 33

Shared-everything and shared-nothing architectures 33

Shared-everything architecture 34

Shared-nothing architecture 34

OLTP versus data warehousing 35

Big Data processing 36

Infrastructure explained 39

Data processing explained 40

Telco Big Data study 40

Infrastructure 42

Data processing 42

CHAPTER 4 Introducing Big Data Technologies 45

Introduction 45

Distributed data processing 46

Big Data processing requirements 49

Technologies for Big Data processing 50

Google file system 51

Hadoop 53

Hadoop core components 54

Hadoop summary 85

NoSQL 86

CAP theorem 87

Key-value pair: Voldemort 88

Column family store: Cassandra 88

Document database: Riak 96

Graph databases 97

NoSQL summary 97

Textual ETL processing 97

Further reading 99

CHAPTER 5 Big Data Driving Business Value 101

Introduction 101

Case study 1: Sensor data 102

- Summary 102
- Vestas 102
- Overview 102
- Producing electricity from wind 102
- Turning climate into capital 104
- Tackling Big Data challenges 104
- Maintaining energy efficiency in its data center 105

Case study 2: Streaming data 105

- Summary 105
- Surveillance and security: TerraEchos 105
- The need 106
- The solution 106
- The benefit 106
- Advanced fiber optics combine with real-time streaming data 107
- Solution components 107
- Extending the security perimeter creates a strategic advantage 107
- Correlating sensor data delivers a zero false-positive rate 107

Case study 3: The right prescription: improving patient outcomes with Big Data analytics 108

- Summary 108
- Business objective 108
- Challenges 108
- Overview: giving practitioners new insights to guide patient care 109
- Challenges: blending traditional data warehouse ecosystems with Big Data 109
- Solution: getting ready for Big Data analytics 109
- Results: eliminating the "Data Trap" 110
- Why aster? 110
- About aurora 111

Case study 4: University of Ontario, institute of technology: leveraging key data to provide proactive patient care 111

- Summary 111
- Overview 111
- Business benefits 112
- Making better use of the data resource 112
- Smarter healthcare 113
- Solution components 113
- Merging human knowledge and technology 114

Broadening the impact of artemis 115
Case study 5: Microsoft SQL server customer solution 115
Customer profile 115
Solution spotlight 115
Business needs 116
Solution 116
Benefits 117
Case study 6: Customer-centric data integration 118
Overview 118
Solution design 121
Enabling a better cross-sell and upsell opportunity 121
Summary 123

PART 2 THE DATA WAREHOUSING

CHAPTER 6 Data Warehousing Revisited 127
Introduction 127
Traditional data warehousing, or data warehousing 1.0 128
Data architecture 129
Infrastructure 130
Pitfalls of data warehousing 131
Architecture approaches to building a data warehouse 137
Data warehouse 2.0 140
Overview of Inmon's DW 2.0 141
Overview of DSS 2.0 141
Summary 144
Further reading 144
CHAPTER 7 Reengineering the Data Warehouse 147
Introduction 147
Enterprise data warehouse platform 148
Transactional systems 149
Operational data store 149
Staging area 149
Data warehouse 149
Datamarts 150
Analytical databases 150
Issues with the data warehouse 150
Choices for reengineering the data warehouse 152
Replatforming 152
Platform engineering 153
Data engineering 154

Modernizing the data warehouse 155
Case study of data warehouse modernization 157
Current-state analysis 157
Recommendations 158
Business benefits of modernization 158
The appliance selection process 159
Summary 162
CHAPTER 8 Workload Management in the Data Warehouse 163
Introduction 163
Current state 163
Defining workloads 164
Understanding workloads 165
Data warehouse outbound 167
Data warehouse inbound 169
Query classification 170
Wide/Wide 170
Wide/Narrow 171
Narrow/Wide 171
Narrow/Narrow 172
Unstructured/semi-structured data 172
ETL and CDC workloads 172
Measurement 174
Current system design limitations 175
New workloads and Big Data 176
Big Data workloads 176
Technology choices 177
Summary 178
CHAPTER 9 New Technologies Applied to Data Warehousing 179
Introduction 179
Data warehouse challenges revisited 179
Data loading 180
Availability 180
Data volumes 180
Storage performance 181
Query performance 181
Data transport 181
Data warehouse appliance 182
Appliance architecture 183
Data distribution in the appliance 184
Key best practices for deploying a data warehouse appliance 186

Big Data appliances 187
Cloud computing .... 187.
Infrastructure as a service 188
Platform as a service 188
Software as a service 189
Cloud infrastructure 189
Benefits of cloud computing for data warehouse 190
Issues facing cloud computing for data warehouse 190
Data virtualization 191
What is data virtualization? 191
Increasing business intelligence performance 193
Workload distribution 193
Implementing a data virtualization program 193
Pitfalls to avoid when using data virtualization 194
In-memory technologies 194
Benefits of in-memory architectures 195
Summary 195
Further reading 195

PART 3 BUILDING THE BIG DATA – DATA WAREHOUSE

CHAPTER 10 Integration of Big Data and Data Warehousing 199
Introduction 199
Components of the new data warehouse 200
Data layer 200
Algorithms 202
Technology layer 203
Integration strategies 204
Data-driven integration 204
Physical component integration and architecture 207
External data integration 209
Hadoop & RDBMS 211
Big Data appliances 212
Data virtualization 214
Semantic framework 215
Lexical processing 216
Clustering 216
Semantic knowledge processing 217
Information extraction 217
Visualization 217
Summary 217

CHAPTER 11 Data-Driven Architecture for Big Data 219
Introduction..219
Metadata 219
Technical metadata...221
Business metadata..221
Contextual metadata..221
Process design–level metadata...221
Program-level metadata 222
Infrastructure metadata 222
Core business metadata...222
Operational metadata 223
Business intelligence metadata 223
Master data management...223
Processing data in the data warehouse...225
Processing complexity of Big Data 228
Processing limitations 229
Processing Big Data...229
Machine learning 235
Summary...240

CHAPTER 12 Information Management and Life Cycle for Big Data 241
Introduction...241
Information life-cycle management...241
Goals 242
Information management policies...243
Governance 243
Benefits of information life-cycle management..247
Information life-cycle management for Big Data..247
Example: information life-cycle management and social media data 248
Measuring the impact of information life-cycle management............................250
Summary..250

CHAPTER 13 Big Data Analytics, Visualization, and Data Scientists 251
Introduction..251
Big Data analytics...251
Data discovery 253
Visualization 254
The evolving role of data scientists 255
Summary...255

CHAPTER 14 Implementing the Big Data – Data Warehouse – Real-Life Situations 257
Introduction: building the Big Data – Data Warehouse 257
Customer-centric business transformation 257
Outcomes 260
Hadoop and MySQL drives innovation 261
Benefits 263
Integrating Big Data into the data warehouse 264
Empowering decision making 264
Outcomes 265
Summary 265

Appendix A: Customer Case Studies 267
Appendix B: Building the Healthcare Information Factory 289
Summary 333
Index 335