Contents

Foreword by Inderpal Bhandari **xvii**

Foreword by Aaron Zornes **xix**

Preface **xxi**

PART ONE: Getting Started **1**

Chapter 1: An Introduction to Big Data Governance **3**

Chapter 2: The Big Data Governance Framework **9**
2.1 Big Data Types 10
2.2 Information Governance Disciplines 12
2.3 Industry and Functional Scenarios for Big Data Governance 15
Summary 27

Chapter 3: The Maturity Assessment **29**
3.1 The IBM Information Governance Council Maturity Model 29
3.2 Sample Questions to Assess Maturity 31
Summary 36

Chapter 4: The Business Case **37**
4.1 Improve On-Time Performance and Passenger Safety Through Big Data Governance 37
4.2 Quantify the Financial Impact of Big Data Governance on Customer Privacy 39
4.3 Reduce IT Costs by Governing the Lifecycle of Big Data 40
4.4 Estimate the Impact of Data Quality and Master Data on Big Data Initiatives 41
Summary 42

Chapter 5: The Roadmap **43**
5.1 The Roadmap Case Studies 43
Summary 46

PART TWO: Big Data Governance Disciplines **47**

Chapter 6: Organizing for Big Data Governance **49**
6.1 Map Key Processes and Establish a RACI Matrix to Identify Stakeholders in Big Data Governance 49
6.2 Determine the Appropriate Mix of New and Existing Roles for Information Governance 54
6.3 Appoint Big Data Stewards as Appropriate 55

6.4 Add Big Data Responsibilities to Traditional Information Governance Roles as Appropriate 60
6.5 Establish a Merged Information Governance Organization with Responsibilities That Include Big Data 63
Summary 65

Chapter 7: Metadata **67**

7.1 Establish a Glossary That Represents the Business Definitions for Key Big Data Terms 68
7.2 Understand the Ongoing Support for Metadata Within Apache Hadoop 71
7.3 Tag Sensitive Big Data Within the Business Glossary 73
7.4 Import Technical Metadata from the Relevant Big Data Stores 74
7.5 Link the Relevant Data Sources to the Terms in the Business Glossary 74
7.6 Leverage Operational Metadata to Monitor the Movement of Big Data 75
7.7 Maintain Technical Metadata to Support Data Lineage and Impact Analysis 75
7.8 Gather Metadata from Unstructured Documents to Support Enterprise Search 77
7.9 Extend Existing Metadata Roles to Include Big Data 77
Summary 78

Chapter 8: Big Data Privacy **79**

8.1 Identify Sensitive Big Data 84
8.2 Flag Sensitive Big Data Within the Metadata Repository 86
8.3 Address Privacy Laws and Regulations by Country, State, and Province 86
8.4 Manage Situations Where Personal Data Crosses International Boundaries 96
8.5 Monitor Access to Sensitive Big Data by Privileged Users 98
Summary 99

Chapter 9: Big Data Quality **101**

9.1 Work with Business Stakeholders to Establish and Measure Confidence Intervals for the Quality of Big Data 102
9.2 Leverage Semi-Structured and Unstructured Data to Improve the Quality of Sparsely Populated Structured Data 107
9.3 Use Streaming Analytics to Address Data Quality Issues In-Memory Without Landing Interim Results to Disk 107
9.4 Appoint Data Stewards Accountable to the Information Governance Council for Improving the Metrics Over Time 111
Summary 112

Chapter 10: Business Process Integration **113**

10.1 Identify the Key Processes That Will Be Impacted by Big Data Governance 114
10.2 Build a Process Map with Key Activities 115

10.3 Map Big Data Governance Policies to the Key Steps in the Process 116
Summary 116

Chapter 11: Master Data Integration 117
11.1 Improve the Quality of Master Data to Support Big Data Analytics 119
11.2 Leverage Big Data to Improve the Quality of Master Data 121
11.3 Improve the Quality and Consistency of Key Reference Data to Support the Big Data Governance Program 124
11.4 Consider Social Media Platform Policies to Determine the Level of Integration with Master Data Management 125
11.5 Extract Meaning from Unstructured Text to Enrich Master Data 126
Summary 131

Chapter 12: Managing the Lifecycle of Big Data 133
12.1 Expand the Retention Schedule to Include Big Data Based on Local Regulations and Business Needs 134
12.2 Document Legal Holds and Support eDiscovery Requests 136
12.3 Compress and Archive Big Data to Reduce IT Costs and Improve Application Performance 137
12.4 Manage the Lifecycle of Real-Time, Streaming Data 138
12.5 Retain Social Media Records to Comply with Regulations and Support eDiscovery Requests 139
12.6 Defensibly Dispose of Big Data No Longer Required Based on Regulations and Business Needs 140
Summary 140

PART THREE: The Governance of Big Data Types 141

Chapter 13: Web and Social Media 143
13.1 Consider Evolving Regulations and Customs When Establishing Policies Regarding the Acceptable Use of Social Media Data About Customers 145
13.2 Set Up Policies Regarding the Acceptable Use of Social Media Data About Employees and Job Candidates 150
13.3 Leverage Confidence Intervals to Assess the Quality of Social Media Data 152
13.4 Establish Policies Regarding the Acceptable Use of Cookies and Other Web Tracking Devices 154
13.5 Define Policies to Link Online and Offline Data in a Way That Does Not Violate Privacy Concerns and Regulations 162
13.6 Ensure the Consistency of Web Metrics 165
Summary 167

Chapter 14: Machine-to-Machine Data **169**
14.1 Assess the Types of Geolocation Data Currently Available 170
14.2 Establish Policies Regarding the Acceptable Use of Geolocation Data Pertaining to Customers 172
14.3 Establish Policies Regarding the Acceptable Use of Geolocation Data Pertaining to Employees 175
14.4 Ensure the Privacy of RFID Data 176
14.5 Define Policies Relating to the Privacy of Other Types of M2M Data 179
14.6 Address the Metadata and Quality of M2M Data 181
14.7 Establish Policies Regarding the Retention Period for M2M Data 184
14.8 Improve the Quality of Master Data to Support M2M Initiatives 184
14.9 Secure the SCADA Infrastructure from Vulnerability to Cyber Attacks 187
Summary 192

Chapter 15: Big Transaction Data **193**
Summary 198

Chapter 16: Biometrics **199**
16.1 Assess the Privacy Implications Relating to the Acceptable Use of Biometric Data 200
16.2 Work with Legal Counsel to Determine the Impact of Evolving Regulations on the Use of Biometric Data for Customers and Employees 202
Summary 204

Chapter 17: Human-Generated Data **205**
17.1 Establish Policies to Mask Sensitive Human-Generated Data 206
17.2 Use Unstructured Human-Generated Data to Improve the Quality of Structured Data 207
17.3 Manage the Lifecycle of Human-Generated Data to Reduce Costs and Comply with Regulations 208
17.4 Extract Insights from Unstructured Human-Generated Data to Enrich MDM 208
Summary 209

PART FOUR: Industry Perspectives **211**

Chapter 18: Healthcare **213**
18.1 Leverage Unstructured Data to Improve the Quality of Sparsely Populated Structured Data 214
18.2 Extract Additional Relevant Clinical Factors Not Available Within Structured Data 215
18.3 Define Consistent Definitions for Key Business Terms 216
18.4 Ensure Consistency in Patient Master Data Across Facilities 216

18.5 Adhere to Privacy Requirements for Protected Health Information in Accordance with HIPAA 216
18.6 Creatively Manage Reference Data to Yield Effective Clinical Insights 217
Summary 217

Chapter 19: Utilities 219
19.1 Duplicate Meter Readings 222
19.2 Referential Integrity of the Primary Key 222
19.3 Anomalous Meter Readings 222
19.4 Data Quality for Customer Addresses 223
19.5 Information Lifecycle Management 223
19.6 Database Monitoring 224
19.7 Technical Architecture 224
Summary 226

Chapter 20: Communications Service Providers 227
20.1 Big Data Types 228
20.2 Integrating Big Data with Master Data 229
20.3 Big Data Privacy 231
20.4 Big Data Quality 232
20.5 Big Data Lifecycle Management 233
Summary 233

PART FIVE: Big Data Technology 235

Chapter 21: Big Data Reference Architecture 237
21.1 Big Data Sources 239
21.2 Open Source Foundational Components 239
21.3 Hadoop Distributions 241
21.4 Streaming Analytics 242
21.5 Databases 243
21.6 Big Data Integration 244
21.7 Text Analytics 246
21.8 Big Data Discovery 247
21.9 Big Data Quality 248
21.10 Metadata for Big Data 249
21.11 Information Policy Management 249
21.12 Master Data Management 250
21.13 Data Warehouses and Data Marts 251
21.14 Big Data Analytics and Reporting 252
21.15 Big Data Security and Policy 254
21.16 Big Data Lifecycle Management 255

21.17 The Cloud 258
Summary 259

Chapter 22: Big Data Platforms 261
22.1 IBM 262
22.2 Oracle 268
22.3 SAP 272
22.4 The Microsoft Big Data Platform 276
22.5 HP 278
22.6 Informatica 279
22.7 SAS 282
22.8 Teradata 283
22.9 EMC 284
22.10 Amazon 284
22.11 Google 285
22.12 Pentaho 285
22.13 Talend 286
Summary 286

Appendix A: List of Acronyms 287

Appendix B: Glossary 291

Appendix C: Reviewer Profiles 313

Appendix D: Contributor Profiles 317

Index 333