

CONTENTS

About the Author	iv
Forewords	xv
<i>by Aditya Kongara</i>	xv
<i>by John R. Talburt</i>	xvi
<i>by Aaron Zornes</i>	xviii
Preface	xxi
PART I—INTRODUCTION	1
1: An Introduction to Data Governance	3
Definition	3
Case Study	5
The Pillars of Data Governance	5
Summary	6
2: Enterprise Data Management Reference Architecture	7
EDM Categories	8
Big Data	13
Data Governance Tools	14
Summary	14
PART II—CATEGORIES OF DATA GOVERNANCE TOOLS	15
3: The Business Glossary	17
Bulk-Load Business Terms in Excel, CSV, or XML Format	17
Create Categories of Business Terms	20

Facilitate Social Collaboration	20
Automatically Hyperlink Embedded Business Terms	21
Add Custom Attributes to Business Terms and Other Data Artifacts	22
Add Custom Relationships to Business Terms and Other Data Artifacts	23
Add Custom Roles to Business Terms and Other Data Artifacts	23
Link Business Terms and Column Names to the Associated Reference Data	24
Link Business Terms to Technical Metadata	25
Support the Creation of Custom Asset Types	26
Flag Critical Data Elements	28
Provide OOTB and Custom Workflows to Manage Business Terms and Other Data Artifacts	28
Review the History of Changes to Business Terms and Other Data Artifacts	32
Allow Business Users to Link to the Glossary Directly from Reporting Tools	33
Search for Business Terms	34
Integrate Business Terms with Associated Unstructured Data	35
Summary	36
4: Metadata Management	37
Pull Logical Models from Data Modeling Tools	37
Pull Physical Models from Data Modeling Tools	38
Ingest Metadata from Relational Databases	40
Pull in Metadata from Data Warehouse Appliances	41
Integrate Metadata from Legacy Data Sources	42
Pull Metadata from ETL Tools	43
Pull Metadata from Reporting Tools	44
Reflect Custom Code in the Metadata Tool	45
Pull Metadata from Analytics Tools	47
Link Business Terms with Column Names	48
Pull Metadata from Data Quality Tools	48
Pull Metadata from Big Data Sources	50
Provide Detailed Views on Data Lineage	51
Customize Data Lineage Reporting	52
Manage Permissions in the Metadata Repository	55
Support the Search for Assets in the Metadata Repository	57
Summary	58
5: Data Profiling	59
Conduct Column Analysis	59

Discover the Values Distribution of a Column	61
Discover the Patterns Distribution of a Column	62
Discover the Length Frequencies of a Column	63
Discover Hidden Sensitive Data	64
Discover Values with Similar Sounds in a Column	65
Agree on the Data Quality Dimensions for the Data Governance Program	66
Develop Business Rules Relating to the Data Quality Dimensions	67
Profile Data Relating to the Completeness Dimension of Data Quality	69
Profile Data Relating to the Conformity Dimension of Data Quality	69
Profile Data Relating to the Consistency Dimension of Data Quality	71
Profile Data Relating to the Synchronization Dimension of Data Quality	71
Profile Data Relating to the Uniqueness Dimension of Data Quality	73
Profile Data Relating to the Timeliness Dimension of Data Quality	74
Profile Data Relating to the Accuracy Dimension of Data Quality	75
Discover Data Overlaps Across Columns	76
Discover Hidden Relationships Between Columns	80
Discover Dependencies	81
Discover Data Transformations	84
Create Virtual Joins or Logical Data Objects That Can Be Profiled	86
Summary	88
6: Data Quality Management	89
Transform Data into a Standardized Format	89
Improve the Quality of Address Data	93
Match and Merge Duplicate Records	95
Create a Data Quality Scorecard	98
<i>Select the Data Domain or Entity</i>	98
<i>Define the Acceptable Thresholds of Data Quality</i>	98
<i>Select the Data Quality Dimensions to Be Measured for the Specific Data Domain or Entity</i>	99
<i>Select the Weights for Each Data Quality Dimension</i>	99
<i>Select the Business Rules for Each Data Quality Dimension</i>	100
<i>Assign Weights to Each Business Rule in a Given Data Quality Dimension</i>	101
<i>Bind the Business Rules to the Relevant Columns</i>	102
View the Data Quality Scorecard	103
Highlight the Financial Impact Associated with Poor Data Quality	104
Conduct Time Series Analysis	104

Manage Data Quality Exceptions	106
Summary	108
7: Master Data Management	109
Define Business Terms Consumed by the MDM Hub	109
Manage Entity Relationships	111
Manage Master Data Enrichment Rules	112
Manage Master Data Validation Rules	113
Manage Record Matching Rules	114
Manage Record Consolidation Rules	116
View a List of Outstanding Data Stewardship Tasks	117
Manage Duplicates	119
View the Data Stewardship Dashboard	121
Manage Hierarchies	122
Improve the Quality of Master Data	122
Integrate Social Media with MDM	125
Manage Master Data Workflows	126
Compare Snapshots of Master Data	127
Provide a History of Changes to Master Data	128
Offload MDM Tasks to Hadoop for Faster Processing	129
Summary	131
8: Reference Data Management	133
Build an Inventory of Code Tables	134
Agree on the Master List of Values for Each Code Table	135
Build Simple Mappings Between Master Values and Related Code Tables	137
Build Complex Mappings Between Code Values	137
Manage Hierarchies of Code Values	139
Build and Compare Snapshots of Reference Data	140
Visualize Inter-Temporal Crosswalks Between Reference Data Snapshots	141
Summary	143
9: Information Policy Management	145
Manage Information Policies, Standards, and Processes Within the Business Glossary	147
Manage Business Rules	147
Leverage Data Governance Tools to Monitor and Report on Compliance	149
Manage Data Issues	149
Summary	157

PART III—THE INTEGRATION BETWEEN ENTERPRISE DATA MANAGEMENT AND DATA GOVERNANCE TOOLS	159
10: Data Modeling	161
Integrate the Logical and Physical Data Models with the Metadata Repository	162
Expose Ontologies in the Metadata Repository	163
Prototype a Unified Schema Across Data Domains Using Data Discovery Tools	163
Establish a Data Model to Support Master Data Management	166
Summary	167
11: Data Integration	169
Deploy Data Quality Jobs in an Integrated Manner with Data Integration	170
Move Data Between the MDM or Reference Data Hub and the Source Systems	172
Leverage Reference Data for Use by the Data Integration Tool	173
Integrate Data Integration Tools into the Metadata Repository	174
Automate the Production of Data Integration Jobs by Leveraging the Metadata Repository	174
Summary	175
12: Analytics and Reporting	177
Export Data Profiling Results to a Reporting Tool for Further Visual Analysis	177
Export Data Artifacts to a Reporting Tool for the Visualization of Data Governance Metrics	178
Integrate Analytics and Reporting Tools with the Business Glossary for Semantic Context	179
Summary	180
13: Business Process Management	181
Data Governance Workflows Should Leverage BPM Capabilities	181
Master Data Workflows Should Leverage BPM Capabilities	186
Data Governance Tools Should Map to BPM Tools	187
Summary	188
14: Data Security and Privacy	189
Determine Privacy Obligations	190
Discover Sensitive Data Using Data Discovery Tools	190
Flag Sensitive Data in the Metadata Repository	191
Mask Sensitive Data in Production Environments	193
Mask Sensitive Data in Non-Production Environments	193
Monitor Database Access by Privileged Users	194

Document Information Policies Implemented by Data Masking and Database Monitoring Tools	195
Create a Complete Business Object Using Data Discovery Tools That Can Be Acted Upon by Data Masking Tools	199
Summary	200
15: Information Lifecycle Management	201
Document Information Policies in the Business Glossary That Are Implemented by ILM Tools	204
Discover Complete Business Objects That Can Be Acted on Efficiently by ILM Tools	204
Summary	205
PART IV—BIG DATA GOVERNANCE TOOLS	207
16: Hadoop and NoSQL	209
Conduct an Inventory of Data in Hadoop	211
Assign Ownership for Data in Hadoop	212
Provision a Semantic Layer for Analytics in Hadoop	213
View the Lineage of Data In and Out of Hadoop	216
Manage Reference Data for Hadoop	219
Profile Data Natively in Hadoop	221
Discover Data Natively in Hadoop	223
Execute Data Quality Rules Natively in Hadoop	224
Integrate Hadoop with Master Data Management	226
Port Data Governance Tools to Hadoop for Improved Performance	229
Govern Data in NoSQL Databases	231
Mask Sensitive Data in Hadoop	232
Summary	232
17: Stream Computing	233
Use Data Profiling Tools to Understand a Sample Set of Input Data	234
Govern Reference Data to Be Used by the Stream Computing Application	235
Govern Business Terms to Be Used by the Stream Computing Application	236
Summary	237
18: Text Analytics	239
Big Data Governance to Reduce the Readmission Rate for Patients with Congestive Heart Failure	240

Leverage Unstructured Data to Improve the Quality of Sparsely Populated Structured Data	241
Extract Additional Relevant Predictive Variables Not Available in Structured Data	242
Define Consistent Definitions for Key Business Terms	242
Ensure Consistency in Patient Master Data Across Facilities	242
Adhere to Privacy Requirements	243
Manage Reference Data	243
Summary	243
PART V—EVALUATION CRITERIA AND THE VENDOR LANDSCAPE	245
19: The Evaluation Criteria for Data Governance Platforms	247
The Total Cost of Ownership	247
Data Stewardship	248
Approval Workflows	248
The Hierarchy of Data Artifacts	249
Data Governance Metrics	256
The Cloud	258
Summary	258
20: ASG	259
ASG-metaGlossary	259
ASG-Rochade	260
ASG-becubic	260
21: Collibra	263
Business Glossary	263
Reference Data Management	264
Data Stewardship	265
Workflows	265
Metadata	265
Data Profiling	265
22: Global IDs	267
Data Profiling	268
Data Quality	269
Metadata	270

23: IBM	271
Metadata	272
Information Integration	272
Data Quality	272
Master Data Management	273
Data Lifecycle Management	273
Privacy and Security	274
24: Informatica	275
Data Profiling and Data Quality	275
Metadata and Business Glossary	277
Master Data Management	277
Information Lifecycle Management	277
Security and Privacy	278
Cloud	278
25: Orchestra Networks	279
Workflows	279
Data Modeling	280
Master Data Management	282
Reference Data Management	283
Business Glossary	283
26: SAP	285
An In-Memory Database	286
Data Quality and Metadata Management	286
Master Data Management	287
Content Management	287
Information Lifecycle Management	287
Enterprise Modeling	287
Data Integration	287
27: Talend	289
The Extended Ecosystem	290
Big Data	291
Data Integration	291
Data Quality	291
Master Data Management	291

Enterprise Service Bus (ESB)	292
Business Process Management (BPM)	292
28: Notable Vendors	293
Adaptive	293
BackOffice Associates	293
Data Advantage Group	293
Diaku	294
Embarcadero Technologies	294
Global Data Excellence	295
Harte-Hanks Trillium	295
Oracle	296
SAS	296
Appendix A: List of Acronyms	299
Appendix B: Glossary	301
Appendix C: Potential Data Governance Tasks to Be Automated with Tools	319
Business Glossary	319
Metadata Management	320
Data Profiling	320
Data Quality Management	321
Master Data Management	321
Reference Data Management	322
Information Policy Management	322
Data Modeling	322
Data Integration	323
Analytics and Reporting	323
Business Process Management	323
Data Security and Privacy	323
Information Lifecycle Management	324
Hadoop and NoSQL	324
Stream Computing	324
Text Analytics	324
Index	325