Contents

About the Author iv

Forewords xv
  by Aditya Kongara xv
  by John R. Talburt xvi
  by Aaron Zornes 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

Select the Data Domain or Entity 98
Define the Acceptable Thresholds of Data Quality 98
Select the Data Quality Dimensions to Be Measured for the Specific Data Domain or Entity 99
Select the Weights for Each Data Quality Dimension 99
Select the Business Rules for Each Data Quality Dimension 100
Assign Weights to Each Business Rule in a Given Data Quality Dimension 101
Bind the Business Rules to the Relevant Columns 102
View the Data Quality Scorecard 103
Highlight the Financial Impact Associated with Poor Data Quality 104
Conduct Time Series Analysis 104
7: **Master Data Management**
- Define Business Terms Consumed by the MDM Hub
- Manage Entity Relationships
- Manage Master Data Enrichment Rules
- Manage Master Data Validation Rules
- Manage Record Matching Rules
- Manage Record Consolidation Rules
- View a List of Outstanding Data Stewardship Tasks
- Manage Duplicates
- View the Data Stewardship Dashboard
- Manage Hierarchies
- Improve the Quality of Master Data
- Integrate Social Media with MDM
- Manage Master Data Workflows
- Compare Snapshots of Master Data
- Provide a History of Changes to Master Data
- Offload MDM Tasks to Hadoop for Faster Processing

8: **Reference Data Management**
- Build an Inventory of Code Tables
- Agree on the Master List of Values for Each Code Table
- Build Simple Mappings Between Master Values and Related Code Tables
- Build Complex Mappings Between Code Values
- Manage Hierarchies of Code Values
- Build and Compare Snapshots of Reference Data
- Visualize Inter-Temporal Crosswalks Between Reference Data Snapshots

9: **Information Policy Management**
- Manage Information Policies, Standards, and Processes Within the Business Glossary
- Manage Business Rules
- Leverage Data Governance Tools to Monitor and Report on Compliance
- Manage Data Issues

Summary

109
109
111
112
113
114
116
117
119
121
122
122
125
126
127
128
129
131
133
134
135
137
137
139
140
141
143
145
147
147
149
149
157
PART III—THE INTEGRATION BETWEEN ENTERPRISE DATA MANAGEMENT AND DATA GOVERNANCE TOOLS

10: Data Modeling
- Integrate the Logical and Physical Data Models with the Metadata Repository
- Expose Ontologies in the Metadata Repository
- Prototype a Unified Schema Across Data Domains Using Data Discovery Tools
- Establish a Data Model to Support Master Data Management
- Summary

11: Data Integration
- Deploy Data Quality Jobs in an Integrated Manner with Data Integration
- Move Data Between the MDM or Reference Data Hub and the Source Systems
- Leverage Reference Data for Use by the Data Integration Tool
- Integrate Data Integration Tools into the Metadata Repository
- Automate the Production of Data Integration Jobs by Leveraging the Metadata Repository
- Summary

12: Analytics and Reporting
- Export Data Profiling Results to a Reporting Tool for Further Visual Analysis
- Export Data Artifacts to a Reporting Tool for the Visualization of Data Governance Metrics
- Integrate Analytics and Reporting Tools with the Business Glossary for Semantic Context
- Summary

13: Business Process Management
- Data Governance Workflows Should Leverage BPM Capabilities
- Master Data Workflows Should Leverage BPM Capabilities
- Data Governance Tools Should Map to BPM Tools
- Summary

14: Data Security and Privacy
- Determine Privacy Obligations
- Discover Sensitive Data Using Data Discovery Tools
- Flag Sensitive Data in the Metadata Repository
- Mask Sensitive Data in Production Environments
- Mask Sensitive Data in Non-Production Environments
- Monitor Database Access by Privileged Users
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
Metadata
Information Integration
Data Quality
Master Data Management
Data Lifecycle Management
Privacy and Security

24: Informatica
Data Profiling and Data Quality
Metadata and Business Glossary
Master Data Management
Information Lifecycle Management
Security and Privacy
Cloud

25: Orchestra Networks
Workflows
Data Modeling
Master Data Management
Reference Data Management
Business Glossary

26: SAP
An In-Memory Database
Data Quality and Metadata Management
Master Data Management
Content Management
Information Lifecycle Management
Enterprise Modeling
Data Integration

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