Contents

Foreword by David Corrigan vii
Foreword by Inderpal Bhandari ix

PART I: Big Data Integration and Governance with IBM InfoSphere
Chapter 1: An Introduction to Big Data Governance 1

Chapter 2: The Big Data Governance Framework 5
2.1 Big Data Types 6
2.2 Information Governance Disciplines 8
2.3 Industry and Functional Scenarios for Big Data Governance 11

Chapter 3: The IBM Big Data Platform 15
3.1 IBM Big Data Products 16
3.2 IBM Big Data Platform Differentiators 20

Chapter 4: Big Data Integration 23
4.1 Bulk Data Movement 23
4.2 Data Replication 27
4.3 Data Virtualization 28

Chapter 5: Metadata 31
5.1 Establish a Glossary That Represents the Business Definitions for Key Big Data Terms 32
5.2 Tag Sensitive Big Data Within the Business Glossary 34
5.3 Maintain Technical Metadata to Support Data Lineage and Impact Analysis 34
5.4 Gather Metadata from Unstructured Documents to Support Enterprise Search 37

Chapter 6: Big Data Security and Privacy 39
6.1 Identify Sensitive Big Data 40
6.2 Flag Sensitive Big Data Within the Metadata Repository 41
6.3 Mask Sensitive Big Data in Production and Non-Production Environments 42
6.4 Monitor Access to Sensitive Big Data by Privileged Users 44

Chapter 7: Big Data Quality 49
7.1 Leverage Semi-Structured and Unstructured Data to Improve the Quality of Sparsely Populated Structured Data 50

7.2 Use Streaming Analytics to Address Data Quality Issues In-Memory Without Landing Interim Results to Disk 52
7.3 Cleanse Big Data Before or After Processing in Hadoop 55

Chapter 8: Master Data Integration 57
8.1 Improve the Quality of Master Data to Support Big Data Analytics 59
8.2 Leverage Big Data to Improve the Quality of Master Data 60
8.3 Improve the Quality and Consistency of Key Reference Data to Support the Big Data Governance Program 62
8.4 Extract Meaning from Unstructured Text to Enrich Master Data 62
8.5 Enrich Customer Master Data with Insights from Social Media to Create Social MDM 67
8.6 Turbo-Charge MDM with Hadoop Technologies 69

Chapter 9: Managing the Lifecycle of Big Data 71
9.1 Expand the Retention Schedule to Include Big Data Based on Local Regulations and Business Needs 72
9.2 Document Legal Holds and Support eDiscovery Requests 72
9.3 Compress and Archive Big Data on Hadoop to Reduce Storage Costs 73
9.4 Archive Big Data in Immutable Format with Seamless Access to Hadoop for Analytics 74
9.5 Manage the Lifecycle of Real-Time, Streaming Data 74
9.6 Defensibly Dispose of Big Data No Longer Required Based on Regulations and Business Needs 75

PART II: Process Data Governance with IBM InfoSphere
Chapter 10: An Introduction to Process Data Governance 77

Chapter 11: Retail Case Study: Process Data Governance of Social Media 79

Chapter 12: Oil and Gas Case Study: Process Data Governance of Sensor Data 81

Chapter 13: Healthcare Case Study: Process Data Governance of Big Claims Transaction Data 87
13.1 A Primer on Claim Codes Used by Health Plans 88

Notes 93
Appendix: Reviewer and Contributor Profiles 95