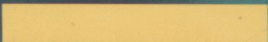


PRASHANTH SOUTHEKAL



DATA QUALITY



**EMPOWERING BUSINESSES WITH
ANALYTICS AND AI**

WILEY

Praise for Data Quality

"The promise of the 'data economy' (i.e., data is the new oil), combined with the naive belief that AI can turn a company's data into gold, is leading many enterprises to experiment with adopting AI. But early adopters are learning that one of the primary causes of AI failure is poor-quality data being fed into AI models (garbage in, garbage out). This highlights that raw data is insufficient and that it must be refined to create value. Consequently, implementing data-centric AI is rapidly becoming a best practice at both start-ups and established technology companies. This book provides a pragmatic approach for enterprises to acquire and manage good-quality data based on proven best practices. If you are a C-level executive or an AI practitioner seeking to deploy AI at scale to drive value creation, I highly recommend this book."

—Anik Bose

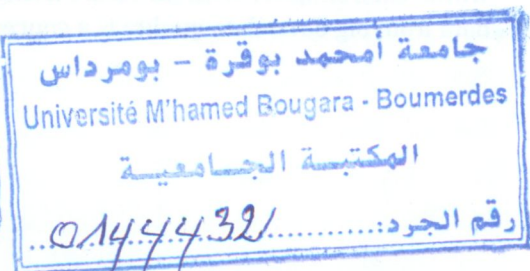
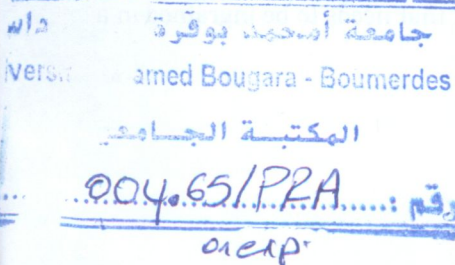
BGV managing general partner and founder, Ethical AI Governance Group
(EAIGG) (United States)

"Prashanth Southekal has nailed the 'why' and the 'how' about data quality's relation to analytics in this highly readable book. To be an economically viable company in today's transparent, global, and competitive world, business leaders must champion the data quality and analytics journey and embed them in decision support systems as an operational core competency. The companies that advocate data quality and analytics integrate them in their DNA to outsmart their competitors in strategic and tactical decision making that yields sustainable success."

—Gary Cokins

President, Analytics-Based Performance Management LLC;
co-author, *Predictive Business Analytics* (United States)

"Good data is a source of myriad opportunities, while bad data is a tremendous burden. Data is now exposed at a much more strategic level (e.g., through business intelligence systems), increasing manifold the stakes involved for individuals and corporations, as well as government agencies. There, the lack of knowledge about data accuracy, currency, or completeness can have erroneous and even catastrophic results. In this book, Dr. Southekal



Contents

Foreword	xvii
Preface	xix
About the Book	xix
Quality Principles Applied in This Book	xx
Organization of the Book	xxi
Who Should Read This Book?	xxiii
References	xxiii
Acknowledgments	xxv
PART I: DEFINE PHASE	1
Chapter 1: Introduction	3
Introduction	3
Data, Analytics, AI, and Business Performance	5
Data as a Business Asset or Liability	6
Data Governance, Data Management, and Data Quality	7
Leadership Commitment to Data Quality	10
Key Takeaways	12
Conclusion	13
References	13
Chapter 2: Business Data	17
Introduction	17
Data in Business	18
Telemetry Data	21
Purpose of Data in Business	22
Business Data Views	24
Key Characteristics of Business Data	31
Critical Data Elements (CDEs)	32
	xiii

Key Takeaways	34
Conclusion	35
References	35

Chapter 3: Data Quality in Business 37

Introduction	37
Data Quality Dimensions	39
Context in Data Quality	51
Consequences and Costs of Poor Data Quality	52
Data Depreciation and Its Factors	54
Data in IT Systems	56
Data Quality and Trusted Information	59
Key Takeaways	60
Conclusion	61
References	62

PART II: ANALYZE PHASE 63

Chapter 4: Causes for Poor Data Quality 65

Introduction	65
Data Quality RCA Techniques	66
Typical Causes of Poor Data Quality	71
Key Takeaways	78
Conclusion	79
References	80

Chapter 5: Data Lifecycle and Lineage 81

Introduction	81
Business-Enabled DLC Stages	82
IT Business-Enabled DLC Stages	86
Data Lineage	88
Key Takeaways	90
Conclusion	90
References	91

Chapter 6: Profiling for Data Quality 93

Introduction	93
Criteria for Data Profiling	95
Data Profiling Techniques for Measures of Centrality	98
Data Profiling Techniques for Measures of Variation	100
Integrating Centrality and Variation KPIs	109
Key Takeaways	112

Conclusion	112
References	112

PART III: REALIZE PHASE **113**

Chapter 7: Reference Architecture for Data Quality **115**

Introduction	115
Options to Remediate Data Quality	116
DataOps	118
Data Product	120
Data Fabric and Data Mesh	123
Data Enrichment	126
Key Takeaways	131
Conclusion	132
References	132

Chapter 8: Best Practices to Realize Data Quality **133**

Introduction	133
Overview of Best Practices	134
BP 1: Identify the Business KPIs and the Ownership of These KPIs and the Pertinent Data	136
BP 2: Build and Improve the Data Culture and Literacy in the Organization	138
BP 3: Define the Current and Desired State of Data Quality	142
BP 4: Follow the Minimalistic Approach to Data Capture	145
BP 5: Select and Define the Data Attributes for Data Quality	148
BP 6: Capture and Manage Critical Data with Data Standards in MDM Systems	152
Key Takeaways	155
Conclusion	158
References	158

Chapter 9: Best Practices to Realize Data Quality **161**

Introduction	161
BP 7: Rationalize and Automate the Integration of Critical Data Elements	162
BP 8: Define the SoR and Securely Capture Transactional Data in the SoR/OLTP System	168
BP 9: Build and Manage Robust Data Integration Capabilities	173
BP 10: Distribute Data Sourcing and Insight Consumption	181
Key Takeaways	186
Conclusion	189
References	189

PART IV: SUSTAIN PHASE	191
Chapter 10: Data Governance	193
Introduction	193
Data Governance Principles	195
Data Governance Design Components	197
Implementing the Data Governance Program	202
Data Observability	203
Data Compliance – ISO 27001, SOC1, and SOC2	205
Key Takeaways	206
Conclusion	208
References	209
Chapter 11: Protecting Data	211
Introduction	211
Data Classification	212
Data Safety	216
Data Security	218
Key Takeaways	220
Conclusion	220
References	221
Chapter 12: Data Ethics	223
Introduction	223
Data Ethics	224
Importance of Data Ethics	224
Principles of Data Ethics	225
Model Drift in Data Ethics	226
Data Privacy	228
Managing Data Ethically	230
Key Takeaways	235
Conclusion	235
References	236
Appendix 1: Abbreviations and Acronyms	237
Appendix 2: Glossary	241
Appendix 3: Data Literacy Competencies	245
About the Author	249
Index	251

PRAISE FOR **DATA QUALITY**

"Incredible! Get ready to be inspired, motivated, and to accelerate on improving your data quality landscape! Regardless of your industry, experience, or role, this book provides the invaluable guidance and framework to start tackling the most critical, common roadblocks encountered with data quality. Spanning the entire data lifecycle, the author has provided trusted, foundational tools with approaches for understandable, everyday business scenarios to demonstrate how organizations can realistically improve their data quality. Incorporating both business performance considerations and technical fundamental best practices together, this timely book is for anyone passionate about truly unlocking the value of their data, engaging and guiding both technical and nontechnical business leaders alike on a critical topic that we should all be passionate about."

—**Alena Godin**, Chief Data and Analytics Officer, Best Buy, Canada

"You may have the best people, processes, technology, analytics, and AI practice in your organization. But if the underlying data that is used is poor in terms of quality, the outcome achieved will be poor. The quality of data driven business decision making/value creation is dependent on the quality of data used. I have been a data quality practitioner for over 30 years. Undoubtedly, this is one of the best books on data quality I have ever read. It provides comprehensive coverage of the end-to-end data quality lifecycle. What is impressive about this book is the language used is simple and understandable by anyone at all levels in an organization and without any technology jargon or tools. A definite book for those who want to be literate on data quality and for those who want to develop and execute a successful data quality program in their organization to derive business benefits."

—**Ram Kumar**, Chief Data and Analytics Officer, Cigna International Markets, Singapore

"In his impeccably well-researched book, *Data Quality: Empowering Businesses with Analytics and AI*, Prashanth Southekal covers just about all the bases—tying data quality to business outcomes, and laying out the many dimensions of data quality, where data quality fits into the overall data and business lifecycle, and a variety of techniques and best practices for identifying and mitigating data quality issues. We'll be seeing this book regularly on the desks of business and data executives and practitioners."

—**Douglas Laney**, Innovation Fellow, Data and Analytics Strategy, West Monroe, USA

"The problem is known: leveraging the value of available data ... but how? Those promising great insights from analytics often start where this book ends: good data quality. Prashanth Southekal takes up the challenge of addressing what is needed to achieve this. Ensuring good data quality at the source is the only viable approach, knowing that fixing bad data (if possible, at all) is way more expensive and cumbersome. Introducing the Define-Assess-Realize-Sustain Model, the book offers the reader a very structured approach in finding the right, fit-to-purpose, level of data quality and sustaining it. It achieves this by describing the main root causes for poor data quality, offering techniques to identify them, and sharing best practices in data capturing and organization around data governance. This book is a valuable read for every organization that is willing to make the effort to unleash the endless opportunities good data quality entails."

—**Astrid von Perbandt**, Vice President Group Controlling, LPKF Laser & Electronics AG, Germany

COMPUTERS/Databases/Data Mining
\$40.00 USA/\$47.99 CAN

Cover Design: Wiley
Cover Image: © amiak/Shutterstock

www.wiley.com

WILEY

 Also available
as an e-book

ISBN 978-1-394-16523-0



9 781394 165230