

Table of Contents

Foreword.....	ix
Preface.....	xi
1. The Data Quality Imperative.....	1
High-Quality Data Is the New Gold	3
Data-Driven Companies Are Today's Disrupters	3
Data Analytics Is Democratized	4
AI and Machine Learning Are Differentiators	4
Companies Are Investing in a Modern Data Stack	5
More Data, More Problems	7
Issues Inside the Data Factory	7
Data Migrations	9
Third-Party Data Sources	11
Company Growth and Change	11
Exogenous Factors	13
Why We Need Data Quality Monitoring	15
Data Scars	15
Data Shocks	16
Automating Data Quality Monitoring: The New Frontier	18
2. Data Quality Monitoring Strategies and the Role of Automation.....	19
Monitoring Requirements	19
Data Observability: Necessary, but Not Sufficient	21
Traditional Approaches to Data Quality	23
Manual Data Quality Detection	23
Rule-Based Testing	25
Metrics Monitoring	29

Automating Data Quality Monitoring with Unsupervised Machine Learning	32
What Is Unsupervised Machine Learning?	32
An Analogy: Lane Departure Warnings	35
The Limits of Automation	36
A Four-Pillar Approach to Data Quality Monitoring	39
3. Assessing the Business Impact of Automated Data Quality Monitoring.....	41
Assessing Your Data	42
Volume	42
Variety	42
Velocity	47
Veracity	48
Special Cases	49
Assessing Your Industry	50
Regulatory Pressure	50
AI/ML Risks	51
Data as a Product	54
Assessing Your Data Maturity	54
Assessing Benefits to Stakeholders	57
Engineers	57
Data Leadership	58
Scientists	59
Consumers	59
Conducting an ROI Analysis	60
Quantitative Measures	60
Qualitative Measures	63
Conclusion	64
4. Automating Data Quality Monitoring with Machine Learning.....	65
Requirements	65
Sensitivity	66
Specificity	66
Transparency	67
Scalability	67
Nonrequirements	67
Data Quality Monitoring Is Not Outlier Detection	68
ML Approach and Algorithm	69
Data Sampling	71
Feature Encoding	76
Model Development	77
Model Explainability	81
Putting It Together with Pseudocode	85

Other Applications	87
Conclusion	88
5. Building a Model That Works on Real-World Data.....	91
Data Challenges and Mitigations	91
Seasonality	91
Time-Based Features	92
Chaotic Tables	93
Updated-in-Place Tables	94
Column Correlations	97
Model Testing	98
Injecting Synthetic Issues	99
Benchmarking	102
Improving the Model	110
Conclusion	111
6. Implementing Notifications While Avoiding Alert Fatigue.....	113
How Notifications Facilitate Data Issue Response	113
Triage	114
Routing	114
Resolution	115
Documentation	115
Taking Action Without Notifications	116
Anatomy of a Notification	117
Visualization	118
Actions	118
Text Description	120
Who Created/Last Edited the Check	120
Delivering Notifications	120
Notification Audience	121
Notification Channels	122
Notification Timing	124
Avoiding Alert Fatigue	124
Scheduling Checks in the Right Order	125
Clustering Alerts Using Machine Learning	126
Suppressing Notifications	127
Automating the Root Cause Analysis	131
Conclusion	134
7. Integrating Monitoring with Data Tools and Systems.....	135
Monitoring Your Data Stack	136
Data Warehouses	137

Integrating with Data Warehouses	137
Security	141
Reconciling Data Across Multiple Warehouses	141
Data Orchestrators	143
Integrating with Orchestrators	143
Data Catalogs	146
Integrating with Catalogs	148
Data Consumers	150
Analytics and BI Tools	150
MLOps	150
Conclusion	152
8. Operating Your Solution at Scale.....	153
Build Versus Buy	153
Vendor Deployment Models	155
Configuration	157
Determining Which Tables Are Most Important	157
Deciding What Data in a Table to Monitor	158
Configuration at Scale	158
Enablement	159
User Roles and Permissions	159
Onboarding, Training, and Support	160
Improving Data Quality Over Time	161
Initiatives	161
Metrics	162
From Chaos to Clarity	164
Appendix: Types of Data Quality Issues.....	167
Index.....	191