

Table of Contents

Preface.....	xiii
1. A (Book) Case for Eventual Consistency.....	1
<i>Denise Koessler Gosnell, PhD</i>	
2. A/B and How to Be.....	3
<i>Sonia Mehta</i>	
3. About the Storage Layer.....	5
<i>Julien Le Dem</i>	
4. Analytics as the Secret Glue for Microservice Architectures.....	7
<i>Elias Nema</i>	
5. Automate Your Infrastructure.....	9
<i>Christiano Anderson</i>	
6. Automate Your Pipeline Tests.....	11
<i>Tom White</i>	
7. Be Intentional About the Batching Model in Your Data Pipelines.....	13
<i>Raghotham Murthy</i>	
8. Beware of Silver-Bullet Syndrome.....	17
<i>Thomas Nield</i>	

9. Building a Career as a Data Engineer.....	19
<i>Vijay Kiran</i>	
10. Business Dashboards for Data Pipelines.....	21
<i>Valliappa (Lak) Lakshmanan</i>	
11. Caution: Data Science Projects Can Turn into the Emperor's New Clothes.....	23
<i>Shweta Katre</i>	
12. Change Data Capture.....	26
<i>Raghotham Murthy</i>	
13. Column Names as Contracts.....	28
<i>Emily Riederer</i>	
14. Consensual, Privacy-Aware Data Collection.....	30
<i>Katharine Jarmul</i>	
15. Cultivate Good Working Relationships with Data Consumers.....	32
<i>Ido Shlomo</i>	
16. Data Engineering != Spark.....	34
<i>Jesse Anderson</i>	
17. Data Engineering for Autonomy and Rapid Innovation.....	36
<i>Jeff Magnusson</i>	
18. Data Engineering from a Data Scientist's Perspective...	38
<i>Bill Franks</i>	
19. Data Pipeline Design Patterns for Reusability and Extensibility.....	40
<i>Mukul Sood</i>	
20. Data Quality for Data Engineers.....	42
<i>Katharine Jarmul</i>	

21. Data Security for Data Engineers.....	44
<i>Katharine Jarmul</i>	
22. Data Validation Is More Than Summary Statistics.....	46
<i>Emily Riederer</i>	
23. Data Warehouses Are the Past, Present, and Future....	48
<i>James Densmore</i>	
24. Defining and Managing Messages in Log-Centric Architectures.....	50
<i>Boris Lublinsky</i>	
25. Demystify the Source and Illuminate the Data Pipeline.....	52
<i>Meghan Kwartler</i>	
26. Develop Communities, Not Just Code.....	54
<i>Emily Riederer</i>	
27. Effective Data Engineering in the Cloud World.....	56
<i>Dipti Borkar</i>	
28. Embrace the Data Lake Architecture.....	58
<i>Vinoth Chandar</i>	
29. Embracing Data Silos.....	61
<i>Bin Fan and Amelia Wong</i>	
30. Engineering Reproducible Data Science Projects.....	63
<i>Dr. Tianhui Michael Li</i>	
31. Five Best Practices for Stable Data Processing.....	65
<i>Christian Lauer</i>	
32. Focus on Maintainability and Break Up Those ETL Tasks.....	67
<i>Chris Moradi</i>	

33. Friends Don't Let Friends Do Dual-Writes.....	69
<i>Gunnar Morling</i>	
34. Fundamental Knowledge.....	71
<i>Pedro Marcelino</i>	
35. Getting the “Structured” Back into SQL.....	73
<i>Elias Nema</i>	
36. Give Data Products a Frontend with Latent Documentation.....	76
<i>Emily Riederer</i>	
37. How Data Pipelines Evolve.....	78
<i>Chris Heinzmann</i>	
38. How to Build Your Data Platform like a Product.....	80
<i>Barr Moses and Atul Gupte</i>	
39. How to Prevent a Data Mutiny.....	83
<i>Sean Knapp</i>	
40. Know the Value per Byte of Your Data.....	85
<i>Dhruba Borthakur</i>	
41. Know Your Latencies.....	87
<i>Dhruba Borthakur</i>	
42. Learn to Use a NoSQL Database, but Not like an RDBMS.....	89
<i>Kirk Kirkconnell</i>	
43. Let the Robots Enforce the Rules.....	91
<i>Anthony Burdi</i>	
44. Listen to Your Users—but Not Too Much.....	93
<i>Amanda Tomlinson</i>	
45. Low-Cost Sensors and the Quality of Data.....	95
<i>Dr. Shivanand Prabhoolall Guness</i>	

46. Maintain Your Mechanical Sympathy.....	97
<i>Tobias Macey</i>	
47. Metadata \geq Data.....	99
<i>Jonathan Seidman</i>	
48. Metadata Services as a Core Component of the Data Platform.....	101
<i>Lohit VijayaRenu</i>	
49. Mind the Gap: Your Data Lake Provides No ACID Guarantees.....	103
<i>Einat Orr</i>	
50. Modern Metadata for the Modern Data Stack.....	105
<i>Prukalpa Sankar</i>	
51. Most Data Problems Are Not Big Data Problems.....	107
<i>Thomas Nield</i>	
52. Moving from Software Engineering to Data Engineering.....	109
<i>John Salinas</i>	
53. Observability for Data Engineers.....	111
<i>Barr Moses</i>	
54. Perfect Is the Enemy of Good.....	114
<i>Bob Haffner</i>	
55. Pipe Dreams.....	116
<i>Scott Haines</i>	
56. Preventing the Data Lake Abyss.....	118
<i>Scott Haines</i>	
57. Prioritizing User Experience in Messaging Systems....	120
<i>Jowanza Joseph</i>	

58. Privacy Is Your Problem.....	122
<i>Stephen Bailey, PhD</i>	
59. QA and All Its Sexiness.....	124
<i>Sonia Mehta</i>	
60. Seven Things Data Engineers Need to Watch Out for in ML Projects.....	126
<i>Dr. Sandeep Uttamchandani</i>	
61. Six Dimensions for Picking an Analytical Data Warehouse.....	128
<i>Gleb Mezhanskiy</i>	
62. Small Files in a Big Data World.....	131
<i>Adi Polak</i>	
63. Streaming Is Different from Batch.....	134
<i>Dean Wampler, PhD</i>	
64. Tardy Data.....	136
<i>Ariel Shaqed</i>	
65. Tech Should Take a Back Seat for Data Project Success.....	138
<i>Andrew Stevenson</i>	
66. Ten Must-Ask Questions for Data-Engineering Projects.....	140
<i>Haidar Hadi</i>	
67. The Data Pipeline Is Not About Speed.....	143
<i>Rustem Feyzkhanov</i>	
68. The Dos and Don'ts of Data Engineering.....	145
<i>Christopher Bergh</i>	
69. The End of ETL as We Know It.....	148
<i>Paul Singman</i>	

70. The Haiku Approach to Writing Software.....	151
<i>Mitch Seymour</i>	
71. The Hidden Cost of Data Input/Output.....	153
<i>Lohit VijayaRenu</i>	
72. The Holy War Between Proprietary and Open Source Is a Lie.....	155
<i>Paige Roberts</i>	
73. The Implications of the CAP Theorem.....	157
<i>Paul Doran</i>	
74. The Importance of Data Lineage.....	159
<i>Julien Le Dem</i>	
75. The Many Meanings of Missingness.....	161
<i>Emily Riederer</i>	
76. The Six Words That Will Destroy Your Career.....	163
<i>Bartosz Mikulski</i>	
77. The Three Invaluable Benefits of Open Source for Testing Data Quality.....	165
<i>Tom Baeyens</i>	
78. The Three Rs of Data Engineering.....	167
<i>Tobias Macey</i>	
79. The Two Types of Data Engineering and Data Engineers.....	169
<i>Jesse Anderson</i>	
80. The Yin and Yang of Big Data Scalability.....	171
<i>Paul Brebner</i>	
81. Threading and Concurrency in Data Processing.....	173
<i>Matthew Housley, PhD</i>	

82. Three Important Distributed Programming Concepts..	175
<i>Adi Polak</i>	
83. Time (Semantics) Won't Wait.....	177
<i>Marta Paes Moreira and Fabian Hueske</i>	
84. Tools Don't Matter, Patterns and Practices Do.....	179
<i>Bas Geerdink</i>	
85. Total Opportunity Cost of Ownership.....	181
<i>Joe Reis</i>	
86. Understanding the Ways Different Data Domains Solve Problems.....	183
<i>Matthew Seal</i>	
87. What Is a Data Engineer? Clue: We're Data Science Enablers.....	185
<i>Lewis Gavin</i>	
88. What Is a Data Mesh, and How Not to Mesh It Up.....	187
<i>Barr Moses and Lior Gavish</i>	
89. What Is Big Data?.....	189
<i>Ami Levin</i>	
90. What to Do When You Don't Get Any Credit.....	191
<i>Jesse Anderson</i>	
91. When Our Data Science Team Didn't Produce Value..	193
<i>Joel Nantais</i>	
92. When to Avoid the Naive Approach.....	195
<i>Nimrod Parasol</i>	
93. When to Be Cautious About Sharing Data.....	197
<i>Thomas Nield</i>	
94. When to Talk and When to Listen.....	199
<i>Steven Finkelstein</i>	

95. Why Data Science Teams Need Generalists, Not Specialists.....	201
<i>Eric Colson</i>	
96. With Great Data Comes Great Responsibility.....	203
<i>Lohit VijayaRenu</i>	
97. Your Data Tests Failed! Now What?.....	205
<i>Sam Bail, PhD</i>	
Contributors.....	207
Index.....	237