

---

# Table of Contents

<b>Preface.....</b>	<b>xiii</b>
---------------------	-------------

---

## **Part I. Foundation and Building Blocks**

<b>1. Data Engineering Described.....</b>	<b>3</b>
What Is Data Engineering?	3
Data Engineering Defined	4
The Data Engineering Lifecycle	5
Evolution of the Data Engineer	6
Data Engineering and Data Science	11
Data Engineering Skills and Activities	13
Data Maturity and the Data Engineer	13
The Background and Skills of a Data Engineer	17
Business Responsibilities	18
Technical Responsibilities	19
The Continuum of Data Engineering Roles, from A to B	21
Data Engineers Inside an Organization	22
Internal-Facing Versus External-Facing Data Engineers	23
Data Engineers and Other Technical Roles	24
Data Engineers and Business Leadership	28
Conclusion	31
Additional Resources	32
<b>2. The Data Engineering Lifecycle.....</b>	<b>35</b>
What Is the Data Engineering Lifecycle?	35
The Data Lifecycle Versus the Data Engineering Lifecycle	36
Generation: Source Systems	37

Storage	40
Ingestion	41
Transformation	45
Serving Data	46
Major Undercurrents Across the Data Engineering Lifecycle	50
Security	51
Data Management	52
DataOps	61
Data Architecture	66
Orchestration	66
Software Engineering	68
Conclusion	70
Additional Resources	71
<b>3. Designing Good Data Architecture.....</b>	<b>73</b>
What Is Data Architecture?	73
Enterprise Architecture Defined	74
Data Architecture Defined	77
“Good” Data Architecture	78
Principles of Good Data Architecture	79
Principle 1: Choose Common Components Wisely	80
Principle 2: Plan for Failure	81
Principle 3: Architect for Scalability	82
Principle 4: Architecture Is Leadership	82
Principle 5: Always Be Architecting	83
Principle 6: Build Loosely Coupled Systems	83
Principle 7: Make Reversible Decisions	85
Principle 8: Prioritize Security	86
Principle 9: Embrace FinOps	87
Major Architecture Concepts	89
Domains and Services	89
Distributed Systems, Scalability, and Designing for Failure	90
Tight Versus Loose Coupling: Tiers, Monoliths, and Microservices	92
User Access: Single Versus Multitenant	96
Event-Driven Architecture	97
Brownfield Versus Greenfield Projects	98
Examples and Types of Data Architecture	100
Data Warehouse	100
Data Lake	103
Convergence, Next-Generation Data Lakes, and the Data Platform	104
Modern Data Stack	105
Lambda Architecture	106

Kappa Architecture	107
The Dataflow Model and Unified Batch and Streaming	107
Architecture for IoT	108
Data Mesh	111
Other Data Architecture Examples	112
Who's Involved with Designing a Data Architecture?	113
Conclusion	113
Additional Resources	113
<b>4. Choosing Technologies Across the Data Engineering Lifecycle.....</b>	<b>119</b>
Team Size and Capabilities	120
Speed to Market	121
Interoperability	121
Cost Optimization and Business Value	122
Total Cost of Ownership	122
Total Opportunity Cost of Ownership	123
FinOps	124
Today Versus the Future: Immutable Versus Transitory Technologies	124
Our Advice	126
Location	127
On Premises	127
Cloud	128
Hybrid Cloud	131
Multicloud	132
Decentralized: Blockchain and the Edge	133
Our Advice	133
Cloud Repatriation Arguments	134
Build Versus Buy	136
Open Source Software	137
Proprietary Walled Gardens	141
Our Advice	142
Monolith Versus Modular	143
Monolith	143
Modularity	144
The Distributed Monolith Pattern	146
Our Advice	146
Serverless Versus Servers	147
Serverless	147
Containers	148
How to Evaluate Server Versus Serverless	149
Our Advice	150
Optimization, Performance, and the Benchmark Wars	151

Big Data...for the 1990s	152
Nonsensical Cost Comparisons	152
Asymmetric Optimization	152
Caveat Emptor	153
Undercurrents and Their Impacts on Choosing Technologies	153
Data Management	153
DataOps	153
Data Architecture	154
Orchestration Example: Airflow	154
Software Engineering	155
Conclusion	155
Additional Resources	155

---

## Part II. The Data Engineering Lifecycle in Depth

<b>5. Data Generation in Source Systems.....</b>	<b>159</b>
Sources of Data: How Is Data Created?	160
Source Systems: Main Ideas	160
Files and Unstructured Data	160
APIs	161
Application Databases (OLTP Systems)	161
Online Analytical Processing System	163
Change Data Capture	163
Logs	164
Database Logs	165
CRUD	166
Insert-Only	166
Messages and Streams	167
Types of Time	168
Source System Practical Details	169
Databases	170
APIs	178
Data Sharing	180
Third-Party Data Sources	181
Message Queues and Event-Streaming Platforms	181
Whom You'll Work With	185
Undercurrents and Their Impact on Source Systems	187
Security	187
Data Management	188
DataOps	188
Data Architecture	189

Orchestration	190
Software Engineering	191
Conclusion	191
Additional Resources	192
<b>6. Storage.....</b>	<b>193</b>
Raw Ingredients of Data Storage	195
Magnetic Disk Drive	195
Solid-State Drive	197
Random Access Memory	198
Networking and CPU	199
Serialization	199
Compression	200
Caching	201
Data Storage Systems	201
Single Machine Versus Distributed Storage	202
Eventual Versus Strong Consistency	202
File Storage	203
Block Storage	206
Object Storage	209
Cache and Memory-Based Storage Systems	215
The Hadoop Distributed File System	215
Streaming Storage	216
Indexes, Partitioning, and Clustering	217
Data Engineering Storage Abstractions	219
The Data Warehouse	219
The Data Lake	220
The Data Lakehouse	220
Data Platforms	221
Stream-to-Batch Storage Architecture	221
Big Ideas and Trends in Storage	222
Data Catalog	222
Data Sharing	223
Schema	223
Separation of Compute from Storage	224
Data Storage Lifecycle and Data Retention	227
Single-Tenant Versus Multitenant Storage	230
Whom You'll Work With	231
Undercurrents	232
Security	232
Data Management	232
DataOps	233

Data Architecture	234
Orchestration	234
Software Engineering	234
Conclusion	234
Additional Resources	235
<b>7. Ingestion.....</b>	<b>237</b>
What Is Data Ingestion?	238
Key Engineering Considerations for the Ingestion Phase	239
Bounded Versus Unbounded Data	240
Frequency	241
Synchronous Versus Asynchronous Ingestion	242
Serialization and Deserialization	243
Throughput and Scalability	243
Reliability and Durability	244
Payload	245
Push Versus Pull Versus Poll Patterns	248
Batch Ingestion Considerations	248
Snapshot or Differential Extraction	250
File-Based Export and Ingestion	250
ETL Versus ELT	250
Inserts, Updates, and Batch Size	251
Data Migration	251
Message and Stream Ingestion Considerations	252
Schema Evolution	252
Late-Arriving Data	252
Ordering and Multiple Delivery	252
Replay	253
Time to Live	253
Message Size	253
Error Handling and Dead-Letter Queues	253
Consumer Pull and Push	254
Location	254
Ways to Ingest Data	254
Direct Database Connection	255
Change Data Capture	256
APIs	258
Message Queues and Event-Streaming Platforms	259
Managed Data Connectors	260
Moving Data with Object Storage	261
EDI	261
Databases and File Export	261

Practical Issues with Common File Formats	262
Shell	262
SSH	263
SFTP and SCP	263
Webhooks	263
Web Interface	264
Web Scraping	264
Transfer Appliances for Data Migration	265
Data Sharing	266
Whom You'll Work With	266
Upstream Stakeholders	266
Downstream Stakeholders	267
Undercurrents	267
Security	268
Data Management	268
DataOps	270
Orchestration	272
Software Engineering	272
Conclusion	272
Additional Resources	273
<b>8. Queries, Modeling, and Transformation.....</b>	<b>275</b>
Queries	276
What Is a Query?	277
The Life of a Query	278
The Query Optimizer	279
Improving Query Performance	279
Queries on Streaming Data	285
Data Modeling	291
What Is a Data Model?	292
Conceptual, Logical, and Physical Data Models	293
Normalization	294
Techniques for Modeling Batch Analytical Data	298
Modeling Streaming Data	311
Transformations	313
Batch Transformations	314
Materialized Views, Federation, and Query Virtualization	327
Streaming Transformations and Processing	330
Whom You'll Work With	333
Upstream Stakeholders	333
Downstream Stakeholders	334
Undercurrents	334

Security	334
Data Management	335
DataOps	336
Data Architecture	337
Orchestration	337
Software Engineering	337
Conclusion	338
Additional Resources	339
<b>9. Serving Data for Analytics, Machine Learning, and Reverse ETL.....</b>	<b>341</b>
General Considerations for Serving Data	342
Trust	342
What's the Use Case, and Who's the User?	343
Data Products	344
Self-Service or Not?	345
Data Definitions and Logic	346
Data Mesh	347
Analytics	348
Business Analytics	348
Operational Analytics	350
Embedded Analytics	352
Machine Learning	353
What a Data Engineer Should Know About ML	354
Ways to Serve Data for Analytics and ML	355
File Exchange	355
Databases	356
Streaming Systems	358
Query Federation	358
Data Sharing	359
Semantic and Metrics Layers	359
Serving Data in Notebooks	360
Reverse ETL	362
Whom You'll Work With	364
Undercurrents	364
Security	365
Data Management	366
DataOps	366
Data Architecture	367
Orchestration	367
Software Engineering	368
Conclusion	369
Additional Resources	369



---

## Part III. Security, Privacy, and the Future of Data Engineering

<b>10. Security and Privacy.....</b>	<b>373</b>
People.....	374
The Power of Negative Thinking.....	374
Always Be Paranoid.....	374
Processes.....	375
Security Theater Versus Security Habit.....	375
Active Security.....	375
The Principle of Least Privilege.....	376
Shared Responsibility in the Cloud.....	376
Always Back Up Your Data.....	376
An Example Security Policy.....	377
Technology.....	378
Patch and Update Systems.....	378
Encryption.....	379
Logging, Monitoring, and Alerting.....	379
Network Access.....	380
Security for Low-Level Data Engineering.....	381
Conclusion.....	382
Additional Resources.....	382
<b>11. The Future of Data Engineering.....</b>	<b>383</b>
The Data Engineering Lifecycle Isn't Going Away.....	384
The Decline of Complexity and the Rise of Easy-to-Use Data Tools.....	384
The Cloud-Scale Data OS and Improved Interoperability.....	385
“Enterprisey” Data Engineering.....	387
Titles and Responsibilities Will Morph.....	388
Moving Beyond the Modern Data Stack, Toward the Live Data Stack.....	389
The Live Data Stack.....	389
Streaming Pipelines and Real-Time Analytical Databases.....	390
The Fusion of Data with Applications.....	391
The Tight Feedback Between Applications and ML.....	392
Dark Matter Data and the Rise of...Spreadsheets?!.....	392
Conclusion.....	393
<b>A. Serialization and Compression Technical Details.....</b>	<b>395</b>
<b>B. Cloud Networking.....</b>	<b>403</b>
<b>Index.....</b>	<b>407</b>

