
Contents

Preface	xi
Chapter 1. Introduction to Data Lakes: Definitions and Discussions	1
Anne LAURENT, Dominique LAURENT and Cédrine MADERA	
1.1. Introduction to data lakes	1
1.2. Literature review and discussion	3
1.3. The data lake challenges	7
1.4. Data lakes versus decision-making systems	10
1.5. Urbanization for data lakes	13
1.6. Data lake functionalities	17
1.7. Summary and concluding remarks	20
Chapter 2. Architecture of Data Lakes	21
Housseem CHIHOUB, Cédrine MADERA, Christoph QUIX and Rihan HAI	
2.1. Introduction	21
2.2. State of the art and practice	25
2.2.1. Definition	25
2.2.2. Architecture	25
2.2.3. Metadata	26
2.2.4. Data quality	27
2.2.5. Schema-on-read	27
2.3. System architecture	28
2.3.1. Ingestion layer	29
2.3.2. Storage layer	31

2.3.3. Transformation layer	32
2.3.4. Interaction layer	33
2.4. Use case: the Constance system	33
2.4.1. System overview	33
2.4.2. Ingestion layer	35
2.4.3. Maintenance layer	35
2.4.4. Query layer	37
2.4.5. Data quality control	38
2.4.6. Extensibility and flexibility	38
2.5. Concluding remarks	39
Chapter 3. Exploiting Software Product Lines and Formal Concept Analysis for the Design of Data Lake Architectures	41
Marianne HUCHARD, Anne LAURENT, Thérèse LIBOUREL, Cédrine MADERA and André MIRALLES	
3.1. Our expectations	41
3.2. Modeling data lake functionalities	43
3.3. Building the knowledge base of industrial data lakes	46
3.4. Our formalization approach	49
3.5. Applying our approach	51
3.6. Analysis of our first results	53
3.7. Concluding remarks	55
Chapter 4. Metadata in Data Lake Ecosystems	57
Asma ZGOLLI, Christine COLLET† and Cédrine MADERA	
4.1. Definitions and concepts	57
4.2. Classification of metadata by NISO	58
4.2.1. Metadata schema	59
4.2.2. Knowledge base and catalog	60
4.3. Other categories of metadata	61
4.3.1. Business metadata	61
4.3.2. Navigational integration	63
4.3.3. Operational metadata	63
4.4. Sources of metadata	64
4.5. Metadata classification	65
4.6. Why metadata are needed	70
4.6.1. Selection of information (re)sources	70

4.6.2. Organization of information resources	70
4.6.3. Interoperability and integration	70
4.6.4. Unique digital identification	71
4.6.5. Data archiving and preservation	71
4.7. Business value of metadata	72
4.8. Metadata architecture	75
4.8.1. Architecture scenario 1: point-to-point metadata architecture	75
4.8.2. Architecture scenario 2: hub and spoke metadata architecture	76
4.8.3. Architecture scenario 3: tool of record metadata architecture	78
4.8.4. Architecture scenario 4: hybrid metadata architecture	79
4.8.5. Architecture scenario 5: federated metadata architecture	80
4.9. Metadata management	82
4.10. Metadata and data lakes	86
4.10.1. Application and workload layer	86
4.10.2. Data layer	88
4.10.3. System layer	90
4.10.4. Metadata types	90
4.11. Metadata management in data lakes	92
4.11.1. Metadata directory	93
4.11.2. Metadata storage	93
4.11.3. Metadata discovery	94
4.11.4. Metadata lineage	94
4.11.5. Metadata querying	95
4.11.6. Data source selection	95
4.12. Metadata and master data management	96
4.13. Conclusion	96
Chapter 5. A Use Case of Data Lake Metadata Management	97
Imen MEGDICHE, Franck RAVAT and Yan ZHAO	
5.1. Context	97
5.1.1. Data lake definition	98
5.1.2. Data lake functional architecture	100

5.2. Related work	103
5.2.1. Metadata classification	104
5.2.2. Metadata management	105
5.3. Metadata model	106
5.3.1. Metadata classification	106
5.3.2. Schema of metadata conceptual model	110
5.4. Metadata implementation	111
5.4.1. Relational database	112
5.4.2. Graph database	115
5.4.3. Comparison of the solutions	119
5.5. Concluding remarks	121

Chapter 6. Master Data and Reference Data in Data Lake Ecosystems 123

C drine MADERA

6.1. Introduction to master data management	125
6.1.1. What is master data?	125
6.1.2. Basic definitions	125
6.2. Deciding what to manage	126
6.2.1. Behavior	126
6.2.2. Lifecycle	127
6.2.3. Cardinality	127
6.2.4. Lifetime	128
6.2.5. Complexity	128
6.2.6. Value	128
6.2.7. Volatility	129
6.2.8. Reuse	129
6.3. Why should I manage master data?	130
6.4. What is master data management?	131
6.4.1. How do I create a master list?	136
6.4.2. How do I maintain a master list?	138
6.4.3. Versioning and auditing	139
6.4.4. Hierarchy management	140
6.5. Master data and the data lake	141
6.6. Conclusion	143

Chapter 7. Linked Data Principles for Data Lakes	145
Alessandro ADAMO and Mathieu D’AQUIN	
7.1. Basic principles	145
7.2. Using Linked Data in data lakes	148
7.2.1. Distributed data storage and querying with linked data graphs	151
7.2.2. Describing and profiling data sources	153
7.2.3. Integrating internal and external data	156
7.3. Limitations and issues	159
7.4. The smart cities use case	162
7.4.1. The MK Data Hub	163
7.4.2. Linked data in the MK Data Hub	165
7.5. Take-home message	169
Chapter 8. Fog Computing	171
Arnault IOUALALEN	
8.1. Introduction	171
8.2. A little bit of context	171
8.3. Every machine talks	172
8.4. The volume paradox	173
8.5. The fog, a shift in paradigm	174
8.6. Constraint environment challenges	176
8.7. Calculations and local drift	177
8.7.1. A short memo about computer arithmetic	178
8.7.2. Instability from within	179
8.7.3. Non-determinism from outside	180
8.8. Quality is everything	181
8.9. Fog computing versus cloud computing and edge computing	184
8.10. Concluding remarks: fog computing and data lake	185
Chapter 9. The Gravity Principle in Data Lakes	187
Anne LAURENT, Thérèse LIBOUREL, Cédrine MADERA and André MIRALLES	
9.1. Applying the notion of gravitation to information systems	187
9.1.1. Universal gravitation	187

9.1.2. Gravitation in information systems	189
9.2. Impact of gravitation on the architecture of data lakes	193
9.2.1. The case where data are not moved	195
9.2.2. The case where processes are not moved	197
9.2.3. The case where the environment blocks the move	198
Glossary	201
References	207
List of Authors	217
Index	219