
Contents

Foreword	xi
Anne LAURENT and Dominique LAURENT	
Preface	xiii
Olivier PIVERT	
Chapter 1. NoSQL Languages and Systems	1
Kim NGUYỄN	
1.1. Introduction	1
1.1.1. The rise of NoSQL systems and languages	1
1.1.2. Overview of NoSQL concepts	4
1.1.3. Current trends of French research in NoSQL languages	6
1.2. Join implementations on top of MapReduce	7
1.3. Models for NoSQL languages and systems	12
1.4. New challenges for database research	16
1.5. Bibliography	18
Chapter 2. Distributed SPARQL Query Processing: a Case Study with Apache Spark	21
Bernd AMANN, Olivier CURÉ and Hubert NAACKE	
2.1. Introduction	21
2.2. RDF and SPARQL	22
2.2.1. RDF framework and data model	22
2.2.2. SPARQL query language	25

2.3. SPARQL query processing	29
2.3.1. SPARQL with and without RDF/S entailment	29
2.3.2. Query optimization	30
2.3.3. Triple store systems	33
2.4. SPARQL and MapReduce	34
2.4.1. MapReduce-based SPARQL processing	35
2.4.2. Related work	39
2.5. SPARQL on Apache Spark	41
2.5.1. Apache Spark	41
2.5.2. SPARQL on Spark	42
2.5.3. Experimental evaluation	48
2.6. Bibliography	53

Chapter 3. Doing Web Data: from Dataset Recommendation to Data Linking

.	57
-----------	----

Manel ACHICHI, Mohamed BEN ELLEFI, Zohra BELLAHSENE and
Konstantin TODOROV

3.1. Introduction	57
3.1.1. The Semantic Web vision	57
3.1.2. Linked data life cycles	58
3.1.3. Chapter overview	61
3.2. Datasets recommendation for data linking	62
3.2.1. Process definition	63
3.2.2. Dataset recommendation for data linking based on a Semantic Web index	64
3.2.3. Dataset recommendation for data linking based on social networks	64
3.2.4. Dataset recommendation for data linking based on domain-specific keywords	65
3.2.5. Dataset recommendation for data linking based on topic modeling	65
3.2.6. Dataset recommendation for data linking based on topic profiles	66
3.2.7. Dataset recommendation for data linking based on intensional profiling	67
3.2.8. Discussion on dataset recommendation approaches	68
3.3. Challenges of linking data	69
3.3.1. Value dimension	70

3.3.2. Ontological dimension	74
3.3.3. Logical dimension	77
3.4. Techniques applied to the data linking process	78
3.4.1. Data linking techniques	79
3.4.2. Discussion	83
3.5. Conclusion	86
3.6. Bibliography	87

Chapter 4. Big Data Integration in Cloud Environments: Requirements, Solutions and Challenges 93

Rami SELLAMI and Bruno DEFUDE

4.1. Introduction	93
4.2. Big Data integration requirements in Cloud environments	96
4.3. Automatic data store selection and discovery	99
4.3.1. Introduction	99
4.3.2. Model-based approaches	99
4.3.3. Matching-oriented approaches	100
4.3.4. Comparison	102
4.4. Unique access for all data stores	103
4.4.1. Introduction	103
4.4.2. ODBAPI: a unified REST API for relational and NoSQL data stores	104
4.4.3. Other works	105
4.4.4. Comparison	107
4.5. Unified data model and query languages	108
4.5.1. Introduction	108
4.5.2. Data models of classical data integration approaches	109
4.5.3. A global schema to unify the view over relational and NoSQL data stores	110
4.5.4. Other works	113
4.5.5. Comparison	117
4.6. Query processing and optimization	118
4.6.1. Introduction	118
4.6.2. Federated query language approaches	118
4.6.3. Integrated query language approaches	121
4.6.4. Comparison	124

4.7. Summary and open issues	125
4.7.1. Summary	125
4.7.2. Open issues	127
4.8. Conclusion	129
4.9. Bibliography	129

Chapter 5. Querying RDF Data: a Multigraph-based Approach 135

Vijay INGALALLI, Dino IENCO and Pascal PONCELET

5.1. Introduction	135
5.2. Related work	137
5.3. Background and preliminaries	137
5.3.1. RDF data	138
5.3.2. SPARQL query	140
5.3.3. SPARQL querying by adopting multigraph homomorphism	142
5.4. AMBER: a SPARQL querying engine	143
5.5. Index construction	144
5.5.1. Attribute index	144
5.5.2. Vertex signature index	145
5.5.3. Vertex neighborhood index	148
5.6. Query matching procedure	149
5.6.1. Vertex-level processing	151
5.6.2. Processing satellite vertices	152
5.6.3. Arbitrary query processing	154
5.7. Experimental analysis	159
5.7.1. Experimental setup	159
5.7.2. Workload generation	160
5.7.3. Comparison with RDF engines	161
5.8. Conclusion	164
5.9. Acknowledgment	164
5.10. Bibliography	164

Chapter 6. Fuzzy Preference Queries to NoSQL Graph Databases 167

Arnaud CASTELLTORT, Anne LAURENT, Olivier PIVERT,
Olfa SLAMA and Virginie THION

6.1. Introduction	167
6.2. Preliminary statements	168

6.2.1. Graph databases	168
6.2.2. Fuzzy set theory	174
6.3. Fuzzy preference queries over graph databases	176
6.3.1. Fuzzy preference queries over crisp graph databases	176
6.3.2. Fuzzy preference queries over fuzzy graph databases	182
6.4. Implementation challenges	193
6.4.1. Modeling fuzzy databases	193
6.4.2. Evaluation of queries with fuzzy preferences	193
6.4.3. Scalability	195
6.5. Related work	197
6.6. Conclusion and perspectives	198
6.7. Acknowledgment	199
6.8. Bibliography	199
Chapter 7. Relevant Filtering in a Distributed Content-based Publish/Subscribe System	203
Cédric DU MOUZA and Nicolas TRAVERS	
7.1. Introduction	203
7.2. Related work: novelty and diversity filtering	205
7.3. A Publish/Subscribe data model	206
7.3.1. Data model	206
7.3.2. Weighting terms in textual data flows	207
7.4. Publish/Subscribe relevance	208
7.4.1. Items and histories	208
7.4.2. Novelty	209
7.4.3. Diversity	209
7.4.4. An overview of the filtering process	210
7.4.5. Choices of relevance	210
7.5. Real-time integration of novelty and diversity	212
7.5.1. Centralized implementation	212
7.5.2. Distributed filtering	216
7.6. TDV updates	221
7.6.1. TDV computation techniques	221
7.6.2. Incremental approach	223
7.6.3. TDV in a distributed environment	225

7.7. Experiments	228
7.7.1. Implementation and description of datasets	229
7.7.2. TDV updates	229
7.7.3. Filtering rate	230
7.7.4. Performance evaluation in the centralized environment	234
7.7.5. Performance evaluation in a distributed environment	238
7.7.6. Quality of filtering	240
7.8. Conclusion	241
7.9. Bibliography	242
List of Authors	245
Index	247