

# Contents

<b>Curriculum Vitae</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Resum</b>	<b>vii</b>
<b>Résumé</b>	<b>ix</b>
<b>Acknowledgements</b>	<b>xi</b>
List of Figures . . . . .	xvi
List of Tables . . . . .	xviii
<b>Thesis Details</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1 Background and Motivation . . . . .	1
2 Data Integration . . . . .	2
2.1 Supporting end-to-end data integration . . . . .	4
2.2 Virtual integration . . . . .	5
2.3 Materialized integration . . . . .	6
2.4 Activities in data integration: state of the art and challenges . . . . .	7
3 Structure of the Thesis . . . . .	10
4 Thesis Overview . . . . .	11
4.1 Chapter 2: A software reference architecture for semantic-aware data-intensive systems . . . . .	12
4.2 Chapter 3: An integration-oriented ontology to govern evolution in data-intensive ecosystems . . . . .	14
4.3 Chapter 4: Answering queries using views under semantic heterogeneities and evolution . . . . .	15
4.4 Chapter 5: SLA-driven selection of intermediate results to materialize . . . . .	17

Contents

5	Contributions . . . . .	17
<b>2</b>	<b>A Software Reference Architecture for Semantic-Aware Data-Intensive Systems</b>	<b>20</b>
1	Introduction . . . . .	22
2	Big Data Definition and Dimensions . . . . .	24
2.1	Volume . . . . .	25
2.2	Velocity . . . . .	25
2.3	Variety . . . . .	26
2.4	Variability . . . . .	26
2.5	Veracity . . . . .	26
2.6	Summary . . . . .	27
3	Related Work . . . . .	29
3.1	Selection of papers . . . . .	29
3.2	Analysis . . . . .	29
3.3	Discussion . . . . .	32
4	Bolster: a Semantic Extension for the $\lambda$ -Architecture . . . . .	33
4.1	The design of <i>Bolster</i> . . . . .	33
4.2	Adding semantics to the $\lambda$ -architecture . . . . .	34
4.3	<i>Bolster</i> components . . . . .	36
5	Exemplar Use Case . . . . .	41
5.1	Semantic representation . . . . .	42
5.2	Data ingestion . . . . .	42
5.3	Data processing and analysis . . . . .	43
6	<i>Bolster</i> Instantiation . . . . .	43
6.1	Available tools . . . . .	44
6.2	Component selection . . . . .	47
6.3	Tool evaluation . . . . .	51
7	Industrial Experiences . . . . .	51
7.1	Use cases and instantiation . . . . .	52
7.2	Validation . . . . .	56
8	Conclusions . . . . .	60
<b>3</b>	<b>An Integration-Oriented Ontology to Govern Evolution in Data-Intensive Ecosystems</b>	<b>61</b>
1	Introduction . . . . .	63
2	Overview . . . . .	65
2.1	Running example . . . . .	66
2.2	Notation . . . . .	68
3	Big Data Integration Ontology . . . . .	72
3.1	Global graph . . . . .	72
3.2	Source graph . . . . .	74
3.3	Mapping graph . . . . .	75

Contents

4	Handling Evolution . . . . .	77
4.1	Releases . . . . .	77
4.2	Release-based ontology evolution . . . . .	78
5	Evaluation . . . . .	79
5.1	Functional evaluation . . . . .	79
5.2	Industrial applicability . . . . .	81
5.3	Ontology evolution . . . . .	82
6	Related Work . . . . .	83
7	Conclusions . . . . .	84
<b>4</b>	<b>Answering Queries Using Views Under Semantic Heterogeneities and Evolution</b>	<b>85</b>
1	Introduction . . . . .	86
2	Related Work . . . . .	89
3	Preliminaries . . . . .	90
3.1	Case study . . . . .	90
3.2	Formal background . . . . .	92
3.3	Case study (cont.) . . . . .	97
4	Rewriting Conjunctive Queries . . . . .	99
4.1	Preliminaries . . . . .	99
4.2	Rewriting algorithm . . . . .	100
4.3	Intra-concept generation . . . . .	101
4.4	Inter-concept generation . . . . .	103
4.5	Discussion . . . . .	105
5	Rewriting Conjunctive Aggregate Queries . . . . .	107
5.1	The aggregation graph . . . . .	107
5.2	Generating CAQs . . . . .	109
5.3	Discussion . . . . .	112
6	Experimental evaluation . . . . .	114
6.1	Experimental setting . . . . .	114
6.2	Experimental results . . . . .	115
7	Conclusions . . . . .	117
<b>5</b>	<b>SLA-driven Selection of Intermediate Results to Materialize</b>	<b>118</b>
1	Introduction . . . . .	120
1.1	Motivational example . . . . .	121
2	Formal Building Blocks and Problem Statement . . . . .	123
2.1	Multiquery AND/OR DAGs and data-intensive flows . . . . .	123
2.2	Components . . . . .	124
2.3	Problem statement . . . . .	125
3	Cost Model for Intermediate Results Materialization Selection . . . . .	125
3.1	Data-intensive flow statistics . . . . .	126
3.2	Metrics . . . . .	126

Contents

3.3	Cost functions . . . . .	128
4	State Space Search Algorithm . . . . .	130
4.1	Actions . . . . .	130
4.2	Initial state . . . . .	132
4.3	Heuristic . . . . .	133
4.4	Searching the solution space . . . . .	133
5	Experiments . . . . .	134
5.1	Intermediate results selection evaluation . . . . .	135
6	Related Work . . . . .	140
7	Conclusions . . . . .	141
<b>6</b>	<b>Conclusions and Future Directions</b>	<b>142</b>
1	Conclusions . . . . .	143
2	Future directions . . . . .	145
	<b>Appendices</b>	<b>146</b>
<b>A</b>	<b>Detailed Algorithms for Rewriting Conjunctive Queries</b>	<b>147</b>
1	Preliminaries . . . . .	147
2	Intra-concept generation . . . . .	147
3	Inter-concept generation . . . . .	150
<b>B</b>	<b>Extended Experiments for Rewriting Conjunctive Queries</b>	<b>154</b>
1	Evolution of response time based on wrappers . . . . .	154
2	Evolution of response time based on edges in the query. . . . .	156
<b>C</b>	<b>MDM: Governing Evolution in Big Data Ecosystems</b>	<b>159</b>
1	Introduction . . . . .	161
1.1	Motivational use case . . . . .	162
2	Demonstrable Features . . . . .	164
2.1	Definition of the global graph . . . . .	164
2.2	Registration of new data sources . . . . .	165
2.3	Definition of LAV mappings . . . . .	166
2.4	Querying the global graph . . . . .	167
2.5	Implementation details . . . . .	168
3	Demonstration overview . . . . .	169
	<b>Bibliography</b>	<b>170</b>
	References . . . . .	171