Abstract iii
Resum v
Résumé viii
Acknowledgements xi
List of Figures xvi
List of Tables xviii
Thesis Details xix
1 Introduction 1
1 Background and Motivation 1
2 The NoSQL Systems and Document Stores 3
3 Data Design for Document Stores 5
3.1 State of the Art and Challenges 8
4 Structure of the Thesis 12
5 Thesis Overview 13
5.1 On the Performance Impact of Using JSON, Beyond Impedance Mismatch 13
5.2 Managing Polyglot System Metadata with Hypergraphs 14
5.3 A Cost Model for Random Access Queries in Document Stores 16
5.4 Automated Database Design for Document Stores 17
6 Contributions 18
2 On the Performance Impact of Using JSON, Beyond Impedance Mismatch 21
1 Introduction 22
2 Related Work 23
3 Representational Differences 24
3.1 Schema variability 24
3.2 Schema declaration 25
3.3 Structure complexity 26
4 Experimental evaluation 28
4.1 Schema variability 29
4.2 Schema declaration 31
4.3 Structure complexity 32
5 Discussion 34
3 Managing Polyglot Systems Metadata with Hypergraphs 36
1 Introduction 37
2 Preliminaries 38
2.1 Resource Description Framework (RDF) 38
2.2 SOS Model 39
3 Formalization 40
4 Metadata Management 44
4.1 Query Representation 46
4.2 Constraints and Transformation Rules on Data Stores . 47
5 Calculating Statistical and Storage Metadata 50
5.1 Storage size estimation 51
5.2 Physical access patterns for workloads 53
6 Use Case 55
7 Related Work 57
4 A cost model for random access queries in document stores 60
1 Introduction 61
2 Background and Related Work 63
3 Formalization of the Cost Model 66
3.1 Generic Component 67

3.2 Specific Component 71
4 Applying the cost model 76
4.1 Couchbase Server (THP) 76
4.2 MongoDB (TDSL) 78
5 Experiments 80
5.1 Couchbase Server 81
5.2 MongoDB 82
5.3 Accuracy of Prediction 87
5.4 Comparison to Other Approaches 88
5 Automated Database Design for Document Stores with Multi-criteria Optimization 89
1 Introduction 90
2 Related Work 92
3 Overview 94
3.1 User Inputs 95
3.2 Design Processes 96
3.3 Loss Function 96
3.4 Search Algorithm 97
4 Canonical Model 99
4.1 Immutable Graph 100
4.2 Storage-Agnostic Constructs 102
4.3 Document Store-Specific Constructs 104
5 Design Processes Over the Canonical Model 105
5.1 Random Design Generation 105
5.2 Design transformations 108
6 Experiments 112
6.1 Quality of the Design 112
6.2 Scalability of the Approach 114
6.3 Threats to Validity 116
6 Conclusions and Future Directions 118
1 Conclusions 118
2 Future Directions 121
Appendices 122
A DocDesign: Cost-Based Database Design for Document Stores 123
1 Introduction 124
2 DocDesign 126
2.1 Design Alternatives 127
2.2 Canonical Representation 127
2.3 Query Workload 128
2.4 Estimating the Runtime 128
3 Demonstration Overview 129
4 Conclusion 130
B DocDesign 2.0: Automated Database Design for Document Stores with Multi-criteria Optimization 132
1 Introduction 133
2 DocDesign 2.0 in a nutshell 136
2.1 User Inputs 136
2.2 Design Operations 138
2.3 Optimization 139
3 Demonstration Overview 140
C Calculating Internal B-tree Blocks 142

D Cost Calculation Examples for MongoDB 143
1 Single Collection with Primary Index 143
2 Multiple Collections 145
E Algorithm to build hyperedges from connected components 148
F Formalized transformations 150
G Validation of operations against MongoDB Design Patterns 152
Bibliography 155
References 156