

Table of Contents

List of Figures	xiv
List of Tables	xvii
List of Abbreviations/Symbols	xix
1 Introduction	1
1.1 Importance of Data and Big Data	2
1.2 Journey of Database	2
1.2.1 SQL Database	2
1.2.2 NoSQL Database	3
1.3 Data Provenance	6
1.4 Motivation	8
1.5 Research Foci	11
1.6 Related Work on Data Provenance	13
1.6.1 Relational Databases	18
1.6.2 Graph Databases	23
1.6.3 Key-Value Pair (KVP) Databases	26
1.7 Research Gaps	29
1.8 Research Contributions	31
1.8.1 Zero-Information Loss Database Design	31
1.8.2 Design and Development of Provenance Framework	32
1.8.3 Provenance Query Templates	38
1.9 Organization of Thesis	38

2	Design and Development of Zero-Information Loss Databases	40
2.1	Zero-Information Loss Database (ZILD)	40
2.2	Generalized Architecture for Zero-Information Loss Database (ZILD)	43
2.3	Database Schema Design	47
2.3.1	Relational Database	47
2.3.2	Graph Database	48
2.3.3	Key-Value Pair (KVP) Database	49
2.4	Role of Zero-Information Loss Database in Data Provenance	50
 3	 Provenance Framework for Relational Databases	 54
3.1	Provenance in Relational Databases	55
3.1.1	Multi-Layer/Multi-Depth Provenance	57
3.1.2	DPHQ	60
3.2	ZILD Architecture for Relational Database (ZILRDB)	62
3.3	Data Provenance Framework using ZILRDB	68
3.3.1	Provenance Relational Algebra (PRA)	70
3.3.2	Provenance Generation	77
3.3.3	Provenance Storage	80
3.3.4	Querying Provenance	83
3.4	Experimental Setup and Results	87
3.4.1	Experimental Setup	87
3.4.2	Results and Discussions	88
3.5	Application Scenarios	94
3.6	Conclusions and Future Work	95
 4	 Provenance Framework for Graph Databases	 97
4.1	Introduction to Graph Database	99
4.2	Provenance in Graph Databases	101
4.3	ZILD Architecture for Graph Database (ZILGDB)	104
4.4	Social Data Provenance Framework using ZILGDB: Twitter Case Study	108
4.4.1	Graph Data Model for Twitter	108
4.4.2	Provenance Generation and Storage	112

4.4.3 Querying Provenance	118
4.5 Experimental Setup and Results	121
4.5.1 Experimental Setup	121
4.5.2 Results and Discussions	122
4.6 Application Scenario	128
4.7 Conclusions and Future Work	133
5 Provenance Framework for Key-Value Pair (KVP) Databases	135
5.1 Introduction to Key-Value Pair (KVP) Database	136
5.2 Provenance in KVP Databases	138
5.3 ZILD Architecture for Key-Value Pair Database (ZILKVD)	140
5.4 Big Social Data Provenance Framework using ZILKVD: Twitter Case Study	144
5.4.1 Twitter Data Model using Key-Value Pair Database	144
5.4.2 Provenance Generation	150
5.4.3 Provenance Storage	158
5.4.4 Querying Provenance	159
5.5 Experimental Setup and Results	164
5.5.1 Experimental Setup	164
5.5.2 Results and Discussions	164
5.6 Application Scenarios	169
5.7 Conclusions and Future Work	170
6 Template-Based Provenance Querying : Towards Designing a Provenance Query Language	172
6.1 Introduction	172
6.2 Provenance Query Templates	176
6.2.1 Vocabulary of PQL (Provenance Query Language)	177
6.2.2 Example Provenance Queries in English	179
6.2.3 Provenance Query Template Design	183
6.3 Conclusions and Future Work	198

7	Conclusions and Future Work	200
7.1	Conclusions	200
7.1.1	Salient Features of Thesis	202
7.2	Future Work	204
	Bibliography	205
	Publications	222
	Biographies	224