Chapter 7

Conclusions and Future Work

The thesis focused on design and development of provenance frameworks to support provenance capturing and querying for different types of databases and applications. We successfully implemented the following three provenance frameworks on the top of Zero-Information Loss Database (ZILD); First, *Data Provenance for Historical Queries (DPHQ)* framework for relational database. Second, *Social Data Provenance (SDP)* framework for social data in graph databases. Third, *Big Social Data Provenance (BSDP)* framework for big social data in KVP databases. In addition, we proposed various provenance query templates as a step towards designing a provenance query language which allows users to query provenance data seamlessly across database models and applications.

7.1 Conclusions

The proposed provenance frameworks are designed and implemented by using the concept of Zero-Information Loss Database for capturing and querying provenance data for select, aggregate, historical queries along with insert, delete, and update operations. It enables data versioning to maintain the history of all updates as provenance information, and supports provenance generation for historical queries. The work proposed in thesis revolves around the concept of Zero-Information Loss Database (ZILD). In a ZILD, nothing is ever lost. Details of ZILD are presented in Chapter 2.

In Chapter 3, we proposed *Data Provenance for Historical Queries (DPHQ)* framework for provenance capturing and querying in relational database. The proposed framework is built upon *Zero-Information Loss Relational Database (ZILRDB)* using the concept of nested relations that maintains all insert, delete, and update operations efficiently. We proposed *Provenance Relational Algebra (PRA)* as an extension of traditional relational algebra to capture the provenance for *ASPJU (Aggregate, Select, Project, Join, Union)* queries in relational databases. Framework supports multi-layer provenance capturing. We stored the provenance data for RDBMS in RDBMS & graph database, & compared provenance storage overhead & provenance querying efficiently. By storing provenance in graph database, we can perform almost all types of queries on provenance without requiring any data from relational database. We found that graph databases offer more significant performance gains over relational databases for executing multi-depth queries on provenance.

A systematic *Social Data Provenance (SDP)* framework for social data based on *Zero-information Loss Graph Database (ZILGDB)* is presented in Chapter 4. It supports historical data queries, and querying through time using updates management in *ZILGDB*. The proposed data model for social data in graph database is proven to be very efficient for range queries using timeline approach. The proposed provenance framework has the capability to capture provenance information for select, aggregate, and historical queries. However, a small overhead is associated with provenance capturing of aggregate queries, and queries having large number of result tuples as compared to queries having less number of result tuples. It also provides support to a detailed provenance analysis through visualization along with efficient multi-depth querying to know about the direct/indirect sources of any information. Generating social provenance in a detailed visual form can help in analyzing a huge social network for critical decision making and in strategical planning. Our proposed framework and provenance algorithm prove to be very promising in dealing with increasingly challenging issue of trust in social media. We conducted a real life use case study to evaluate the usefulness of such detailed social provenance generated by our framework in terrorist attack investigation, to identify suspicious persons and their linked communities using a social media platform, particularly in Twitter's network.

We proposed a *Big Social Data Provenance (BSDP)* framework for *Key-Value Pair (KVP)* Databases using the concept of *Zero-Information Loss Database (ZILD)* in Chapter 5. In the proposed framework, a large volume of real life social data are fetched from the Twitter's network through live streaming and modelled in a KVP database by applying an efficient query driven approach. The proposed framework has the capability to capture provenance information for select and aggregate queries. It also supports to capture provenance information for insert, delete, update, and historical queries using *Zero-Information Loss Key-Value Pair Database (ZILKVD)*. We evaluated the performance of proposed framework in terms of provenance capturing overhead and provenance query execution time for different query sets. We found that the proposed framework is efficient in terms of execution time for provenance queries. However, a small execution overhead is measured for some aggregate queries where aggregation is performed on a large number of input tuples as compared to other aggregate queries where aggregation is performed on a few input tuples only. Framework also provides supports for querying provenance information for historical data queries as well as in tracing out the source of result tuples of data retrieval queries.

A set of provenance query templates were identified in Chapter 6 which allows users to seamlessly query provenance data captured for different database models. We propose to continue this work to come up with a full-fledged PQL.

7.1.1 Salient Features of Thesis

We believe that the main goal of this thesis is successfully achieved through our research contributions. Some of the salient features of our research work are given below:

1. *Zero-Information Loss Database Design:* Zero-Information Loss Databases is a special kind of database based on temporal database, and is designed on the top of relational, graph, and key-value pair databases that supports data versioning to maintain history of all updates as provenance information along with the provenance of insert and delete operations. It also supports to capture provenance for historical queries.

2. *Provenance Relational Algebra (PRA):* Provenance Relational Algebra, an extension of traditional relational algebra, supports to capture provenance for ASPJU queries in relational database.

3. *DPHQ Framework for Relational Database:* DPHQ is implemented on top of ZILRDB

for capturing provenance for ASPJU (Aggregate, Select, Project, Join, Union) along with provenance for update, insert, and delete queries in relational database. It is capable to capture Multi-Layer provenance and efficient Multi-Depth querying on provenance using graph databases.

4. *Social Data Provenance Framework for Graph Databases:* Social data provenance framework is developed on top of Zero-Information Loss Graph Database (ZILGDB) that supports to capture provenance information for select, aggregate, historical, insert, update, and delete queries. It also provides support to a detailed provenance analysis through visualization along with efficient multi-depth querying of provenance data to know about the direct/indirect sources of any information along with historical data queries.

5. *Applicability of Social Data Provenance Framework:* We conduct a real life use case study to evaluate the usefulness of such detailed social provenance generated by social data provenance framework for graph databases in terrorist attack investigation, to identify suspicious persons and their linked communities using a social media platform, particularly in Twitter's network.

6. *Query Driven Key-Value Pair (KVP) Data Model:* Designed and implemented an efficient Key-Value Pair (KVP) data model based upon a query driven approach to correlate a huge volume of live streamed real life social data through relationships and dependencies.

7. *Big Social Data Provenance Framework for key-value pair databases:* Big Social Data Provenance (BSDP) framework for Key-Value Pair Databases is developed on top of Zero-Information Loss Key-Value Pair Database (ZILKVD). BSDP is suitable to capture provenance for almost all type of queries including historical queries, and also supports for querying provenance information for historical data queries, as well as to trace the source of result tuples of data retrieval queries.

8. *Multi-Layer Provenance Capture:* Both DPHQ and social data provenance framework for graph databases are capable to capture a detailed multi-layer provenance.

9. *Multi-Depth Provenance Querying:* Proposed frameworks also support multi-depth provenance querying to know about direct and indirect sources of a piece of information.

10. *Provenance Query Templates:* Various provenance query templates are proposed to facilitate users to express provenance queries on provenance data.

7.2 Future Work

We have identified several directions in which the work presented in the thesis can be taken forward, those are outlined below:

1. Provenance information was captured for select, aggregate, historical queries along with the insert, delete and update operations. We plan to extend the proposed framework to capture the provenance information for complex queries such as nested queries, sub-queries, difference etc.

2. As the social data is growing rapidly at an unprecedented scale, we plan to further extend our social data provenance framework for distributed graph database on different nodes in a cluster using Neo4j Aura.

3. Currently BSDP is implemented on a single node of Apache Cassandra cluster; our plan is to further extend this framework for a large volume of data, distributed on different nodes in a cluster.

4. We also believe that the captured provenance information can be used as the basis for a trust model to evaluate the trust value of information. We plan to design such a trust model in future.

5. This research is in progress with the aim of designing a new provenance query language which will abstract the underlying database models, and will allow expressing any query on provenance data. This work is, of course, an interesting area for research, therefore, remains as a suggestion for future work.