Abstract

In computing world, data is referred as the factual information in a digital form used as a basis for various qualitative and quantitative analysis. Over the past decades, a large volume of limitless data has been growing rapidly from various sources in diverse formats with many folds beyond terabytes or petabytes. This immense volumes of data are continuously processed, transformed, and migrated due to their life-cycle operations including data generation, storage, transmission, and deletion. In addition, social media and web-based communication are also continuously generating a very large size of unstructured data related to the various fields of our day to day life. With the rapid evolvement of social media platforms, everyone has become more enthusiastic about sharing their thoughts, ideas, opinions, and other content through a social media platform, causing an exponential growth in the size of social media data. This proliferation of data requires an efficient data management in these application domains. Although, the amount of data processed in these application domains are collected from various sources, but the major part of this data is generated from different transformations applied on the existing data. To know about the data quality, reliability, and its trustworthiness for different intentions such as audit trail, fact investigations, error tracing, and rumour detection etc., an extensive set of metadata, i.e., "description about data" is needed. In addition to these metadata, some other information such as various sources and origin of data, derivations history and derivation process of data is also required.

Provenance data is a sort of metadata which gives the following information about a data element such as owner of a data, its direct/indirect sources, history of a data and transformations applied on data etc. Alternatively, prove-

nance can be described by answering the questions *how* and *where* the data is produced and *when* and by *whom*. Provenance information is obligatory to support in understanding the implication, and credibility of a piece of information. It serves different purposes in a database system such as audit trail, data discovery, update propagation, incremental maintenance, rumour identification and justification of a query result etc. In recent years, a piece of information published in an article on social media is also facing a critical challenge to determine its *social provenance* or *social data provenance* which involves following three dimensions viz., "*What*", "*Who*", and "*When*". *What*, provides the description about the social media posts, *Who* describes the correlations among social media users, and *When* characterizes the evolution of users' behaviour over time. Like data provenance, social provenance also describes the ownership and origin of such information. The continuously growing social media data is the major source of big data that is characterised by 7 V's viz., Volume, Velocity, Veracity, Variety, Variability, Visualization, and Value. Veracity of big data is directly linked with data provenance. Provenance for big social data is referred as *Big Social Data Provenance*. In *Social Data Analytics*, the credibility of an analysis is generally depends upon the quality and truthfulness of input data which is assured by the *Social Data Provenance*. In this way, social data provenance plays a major role in clarifying opinions to avoid rumors, investigations and explaining how and when this information is created and by whom. But distillation of provenance information from such a huge amount of complex data, however, is an extremely tedious task, due to its diverse formats.

For applications those are related to auditing, security, and accountability, there is a need to restore all the operations performed on a database to produce the same result as of their previous executions. This leads to the requirement of managing all the updates (i.e., insert, delete, and update operations) without any loss of information as a provenance data. But the conventional/snapshot database systems does not maintain the history of all the data objects and

stores only the current snapshot of data, as a result they are ill-suited for such applications. Zero-Information Loss Database (ZILD) which is a special type of database based on temporal database maintains temporal data as a history of all the updates along with the complete information of operational activities performed in that database. Therefore, it is well-suited for designing the provenance framework especially for capturing provenance for update, insert, delete, and historical queries.

Nowadays, the necessity to capture and query the provenance information about this mammoth data has raised a growing interest in the era of data analytics with several remarkable challenges. Therefore, an effective data model along with an efficient *Data Provenance Framework* is required to distill the provenance information from this enormous size of data.

To address the above issues, we present the following three provenance frameworks on the top of a *Zero-Information Loss Database (ZILD)* in this thesis. First, *Data Provenance for Historical Queries (DPHQ)* Framework for relational database. Second, *Social Data Provenance (SDP)* Framework for social data in graph database. Third, *Big Social Data Provenance (BSDP)* Framework for big social data in key-value pair database. Some of the major contributions of this research work are as follows:

1. We designed and implemented specialized *Zero-Information Loss Databases (ZILD)*, i.e., *ZILRDB*, *ZILGDB*, and *ZILKVD* in Relational, Graph, and Key-Value Pair Database to develop provenance frameworks on top of these. These enable data versioning to maintain the history of all the data updates as provenance information and also support provenance generation for historical queries.
2. We presented a *Data Provenance for Historical Queries (DPHQ)* framework for relational database on top of *Zero-Information Loss Relational Database (ZILRDB)* that supports to capture provenance for all update, insert, and delete operations. It supports multi-layer provenance generation and multi-depth provenance querying for tracing out the sources of a data

element upto a certain depth, and justifying a query result. The captured provenance information is stored in both relational and graph database for efficient provenance querying.

3. We proposed the *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 database.
4. We presented a *Social Data Provenance (SDP)* framework for social data, based on *Zero-Information Loss Graph Database (ZILGDB)*. ZILGDB supports historical data queries, querying through time, and capturing provenance information for historical queries. The proposed framework has the capability to capture provenance information for a query set including select, aggregate, and range queries with timeline.
5. We conducted a real life use case study to evaluate the usefulness of our SDP framework in terrorist attack investigation, to identify the suspicious persons, and their linked communities in a social media platform.
6. We 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. We presented a *Big Social Data Provenance (BSDP)* framework for KVP Database that is capable of capturing fine-grained 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)* architecture.

In addition, we propose various provenance query templates for querying provenance information in a way which is independent of underlying database & application, and to facilitate users to pose useful & common provenance queries using the templates. The notations of these query templates are simple, unambiguous, and easy to understand.