Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Master Thesis

Query-Execution Equivalence of Anonymized Data

Today, many applications, such as query recommendation or identification of hot spots of user interest, are based on analyzing SQL logs. To conduct research on this topic, scientists rely on the availability of large real-world query logs. For the organizations owning the logs, the insights of log analysis would give way to innovations. However, most existing research on SQL logs is academic in nature. The main reason is that many companies, while they cannot do such analyses themselves, are afraid of sharing their SQL logs. These concerns are natural, as the data may contain confidential information, which must not be revealed (including privacy concerns).

In a current research project, we intend to address this issue by anonymization/encryption of the attribute values in the log. The core idea is to use an encryption function that provides sufficient anonymization. The encryption function shall also keep certain properties of the data, allowing us to conduct meaningful research on the encrypted logs. In particular, the function shall encrypt attribute values in an order-preserving fashion. In this way, we intend to achieve what we call *execution equivalence* for a certain sub-set of the relational operators. In a nutshell, query execution equivalence for two queries is as follows. The first query is one found in the log. The second query is generated by encrypting the attribute values in the first query. Two such queries are execution equivalent if their execution always has the same structure.

As we only want to preserve the order of attribute values, we can rely on simple compression schemes, such as Dictionary Encoding, instead of complex encryption schemes. In this thesis, the candidate shall extend our idea of query-execution equivalence for order-preserving encryption schemes and conduct meaningful experiments on real-world case studies. This includes:

- Conceptually improve the notion of query execution equivalence,
- Come up with a concept and implement an approach how to test *SQL statements* for query execution equivalence, including meaningful error reporting in case of non-equivalence,
- Extend the approach by considering also *query plans*,
- Design and conduct meaningful experiments on real-world databases and SQL logs (e.g., SkyServer) answering questions such as:
    - What fraction of the statements in the log are execution equivalent?
    - What other (external) factors may influence execution equivalence (e.g., current load on the database)?

In this thesis, you will gain in depth knowledge on principles of modern database systems as well as on state-of-the-art knowledge-discovery techniques and their deployment. An ideal candidate should be familiar with concepts from an introductory course on database systems and should be able to do conceptual work with clear results. This thesis is best suited for a Master Thesis.

Contact

Dr. Martin Schäler | martin.schaeler@kit.edu | +49 721 608-47351 | Room: 365

Am Fasanengarten 5 | 76131 Karlsruhe | Building: 50.34

Master Thesis