Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-1 (2015) 113–114

Reproducible Database Queries in Privacy Sensitive Applications

Stefan Pröll * **Rudolf Mayer** * **Andreas Rauber** **

* *SBA-Research, Vienna, Austria (e-mail: sproell@sba-research.org, rmayer@sba-research.org)*
** *Institute of Software Technology and Interactive Systems at the Vienna University of Technology, Austria (e-mail: rauber@ifs.tuwien.ac.at)*

Abstract: Research databases are an important building block in eScience and computational science investigations. For enabling reproducible research, an approach is needed which supports the identification and citation of the exact data (sub)sets utilized in experiments. While this itself is a challenge, in many cases the data stored in databases is sensitive and needs to be protected. Due to the increasing complexity of eScience investigations, data is often integrated from different sources, potentially stemming from competing data owners. In order to achieve the research goals, the data needs to be combined and analysed as a whole. As data owners of such sources may have potential conflicts of interest in certain aspects, a mechanism is needed which prevents the retrieval and or recombination of privacy related data while still full access to own data must be granted at all times.

Keywords: Reproducibility, Relational databases, Data handling systems, Data privacy, Data sets

1. INTRODUCTION

Data is a crucial element in computational science. On the one hand, data is often used as input for research, e.g. in statistical machine learning. Also in mathematical models that themselves do not directly learn from data, data observations might still be utilized in an earlier step, to obtain and estimate fitting parameter settings for the models. On the other hand, complex data can also be the output of an experiment or simulation. Important decisions are made on the basis of such data, for example, whether an experiment was considered successful or a model is considered valuable is often settled on the result sets delivered from statistical analysis of data. As data is not only a final product of research investigations, but is transformed, updated or changed in the many processing steps that form an experiment, each intermediate result needs to be traceable and identifiable. Hence it is a fundamental requirement to know which exact data was involved in the research, and its intermediate steps. Data citation tackles the problem of uniquely identifying, referencing and citing datasets and their subsets in order to make them retrievable at a later point in time. The goal of citing data is to attach a persistent identifier to each dataset retrieved from a data source, and use this identifier as a handle which allows to retrieve the exact same dataset again. Thus data citation enables the examination, revision or analysis of data and therefore constitutes evidence for decisions and how they have been made.

In most settings, it is not a feasible approach to export each individual dataset and store it as data dump in an archive, as such a solution does not scale well for large data volumes, caused either by large data itself, or many and repeated processing steps. Further, without additional metadata the management of data files becomes a challenge. Verifying whether a data set is actually the one required can be achieved by applying fingerprinting and checksum mechanisms, but without the knowledge how the data was derived and how a specific subset was selected, the reproducibility of data driven experiments is limited.

In many cases the datasets contain sensitive data and privacy protocols need to be applied. Implementing thorough permission schemes is essential, but the goal of providing a secure eScience environment becomes more challenging if several stakeholders with potentially competing interests need to exchange data. All participating partners obviously need access to the data they contributed, but providing access to aggregated results and compiled result sets is a further requirement. The system needs to support analysis of the data but it needs to prevent the creation of datasets which contain sensitive data or allow the deduction of such information via skilful querying (Dwork and Smith (2010)).

2. DATA CITATION IN VOLATILE SOURCES

In many scenarios, data is not just static but it can also be highly dynamic. New records can be created, some data may get updated whereas older records may be deleted. For this reason the full history of all operations, which either added, altered or deleted any record in the database system, needs to be traced. Still, only versioning of databases itself does not yet allow to retrieve a specific result set from a given point in time. Most modern rela-

⋆ Parts of this work are supported by the project DEXHELPP.

10.1016/j.ifacol.2015.05.202

tional database management systems (RDBMS) support point in time recovery (PITR), which can be used for rolling back the data to a specific date. Although this method allows querying on top of the data as they were at any given moment in time, the approach is not feasible for retrieving historical data in an convenient fashion. This is due to the potentially costly rollback operation to a specific data version, which cannot be reused for other queries. It is to mention that the term version may be misleading in the context of data citation. By a version we understand a specific state of the records in a result set, we do not refer to an export of data which is assigned a version number. The same query for instance will produce different versions of result sets, whenever a single record contained in the result was changed between two executions of a query.

To achieve versioning, all events performed on each record need to be stored with a timestamp in order to create an audit trail of all changes which have been introduced into the database. The system needs to provide information if a record was inserted, updated or a deleted. This data can be stored in a history table and allows retrieving a specific state of the database records at any given time. Additionally to the provenance of any record in the database, the actual query which was used for retrieving the dataset needs to be stored with additional metadata.

For storing the queries with their execution metadata, the concept of the query store was developed by Proell and Rauber (2014). The goal of the query store is to attach persistent identifiers to query results and allow retrieving the same data again by re-executing a query against historical data. A persistent identifier (PID) uniquely identifies a resource for the long term by utilizing a managed infrastructure providing additional services which can be used for accessing the metadata and the object itself in a reliable way. Instead of attaching a PID directly to the exported dataset, the PID references the query which ultimately produces the dataset. The Query Store needs to detect whether a issued query is already stored persistently or if a new query needs to be inserted. In order to validate the re-executed query result for its correctness, a hash key is computed. A result set is only considered correct if and only if all records are included in the same sequence and ordering as in the original query. When querying sensitive data, several security policies need to be applied which need to prevent unauthorized access and impermissible queries. The knowledge necessary for applying such queries needs to be preserved, maintained and re-applied for the re-executed queries retrieving historical data.

3. REQUIREMENTS IN PRIVACY SENSITIVE APPLICATIONS

Many scientific projects are collaborative and involve different organizations working together and exchanging data. Although the stakeholders pursue a common goal, it is not necessarily true that all exchanged data should be available to all project partners without limitations. Research data often contains information which needs to be protected from unauthorized access and privacy needs to be maintained. This is especially true for projects that involve data which can be attributed to individuals and may even contain highly sensitive data such as health records.

Whenever such data is exchanged, the database management handling the data needs to ensure that privacy is maintained within the whole data life cycle. Providing a secure database for several stakeholders with potentially conflicting interests goes beyond the implementation of permission schemes for individual tables. The system must preserve the privacy of the data at all times and may also prevent data leakage through clever queries with could reveal individual details.

Furthermore, a mechanism is needed which allows tracing all executed queries and logs the activities on the system. Researchers and administrators need to be able to reproduce a specific query which was issued against the database and retrieve the very same dataset again. There are several reasons for this requirement. Researchers, as mentioned earlier, need to have the possibility retrieving the data again which was used a model in order to verify and rerun the experiment. Moreover, the project administrators and managers require to verify that the data was retrieved in compliance with the ethical policies defined in a project. Therefore, a mechanism is required which allows auditing the queries and the resulting datasets for their privacy compliance.

To this end, the above introduced data citation approach needs to be adapted to cope with this more complex setting. For example, the re-execution of a query also requires to maintain the permission rights of the records as access to a specific portion of the data may be granted or revoked. This information needs to be captured and stored in the query store in order to reproduce who retrieved sensitive data and prove that access to a specific dataset was granted. Additionally, the query store can also serve as an audit trail and needs to be protected from manipulation.

4. CONCLUSIONS

Reproducibility is a key requirement for computational research and therefore the complete workflow from experimental setup to results needs to be traceable and understandable retrospectively. Data citation enables peers to verify the data used in each step of an experiment, thus it constitutes significantly to reproducibility of experiments. In turn, reproducibility provides documentation which fosters reuse of results and intermediate research products. It is clear that data is a fundamental basis for most experiments, therefore the creation process of a dataset needs to be reproducible as well. For obvious reasons, reproducibility includes privacy preserving methods which hinder disclosure of sensitive data.

REFERENCES

Dwork, C. and Smith, A. (2010). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2).

Proell, S. and Rauber, A. (2014). A Scalable Framework for Dynamic Data Citation of Arbitrary Structured Data. In *3rd International Conference on Data Management Technologies and Applications (DATA2014)*. Vienna, Austria.