Data Provenance in ETL Scenarios

Timos Sellis, Dimitrios Skoutas
National Technical University of Athens
{timos,dskoutas}@dblab.ece.ntua.gr

Alkis Simitsis
IBM Almaden Research Center
asimits@us.ibm.com

Panos Vassiliadis
University of Ioannina
pvassil@cs.uoi.gr

Extract – Transform – Load Processes

Data in large organizations are typically distributed in several heterogeneous sources, organized and stored under different naming conventions, structures, and formats. For supporting the functionality of On-Line Analytical Processing (OLAP) applications and Decision Support Systems (DSS), Data Warehouses (DW) are employed to integrate the data and provide a uniform infrastructure for querying, reporting, mining, and other advanced analysis techniques. The process of populating the DW with data stemming from the operational sources, in a way that the schema and business requirements of the DW are met, is referred to as *Extract-Transform-Load* (ETL) process. Typically, such processes are handled by ETL tools, which are pieces of software responsible for the extraction of data from several sources, their cleansing, customization, and insertion into a DW. However, even though the term ETL is traditionally related to data warehousing, it may be used in a wider sense to refer to any process of exchanging and transforming data between data stores. For example, ETL is the core functionality of a recently emerging type of web applications, called *mashups*, where information is extracted from various web sites, and it is appropriately transformed and integrated before presented to the final user.

Data provenance in ETL

In this context, *provenance* refers essentially to capturing, representing, and managing metadata for tracing the origin of data elements that take part in various ETL operations or populate the DW. It can be viewed at different levels of detail, such as: which sources contribute to a given target element; which particular elements within these sources; which transformations have the target element undergone. At the finest level of granularity, provenance can be considered as the inverse of the ETL process: whereas ETL processes facilitate the transferring of data from the sources to the warehouse, provenance refers to the explanation of why a warehouse tuple is found there and through which process. In other words, the formal characterization of a tuple's provenance involves the description – via a certain formalism – of the inverse process for an ETL scenario.

Clearly, since it is the ETL process per se which is of primary interest in a data warehouse setting, the semi-automatic computation of the inverse process in a formal way is a clear problem per se. Moreover, this is a technically challenging task, since typically, there are several non-invertible operators (e.g., aggregation) involved in an ETL scenario.

Provenance plays a significant role in several aspects, most notably in the following.

- Explanation of analysis results: provenance information can be used to determine why a specific data value exists in the DW, for instance, whether a recorded profit loss is due to an actual decrease on sales or due to changes on currency exchange rates.
- Dealing with incomplete information: for example, default values may be applied to missing values of various data elements, based on their source of origin, creation time, and so forth.
- Handling inconsistencies: rules for conflict resolution can be defined and enforced using provenance information to determine the origin of the conflicting data elements, and then examining trust and policies on the corresponding sources.
- What-if analysis: knowledge of the source elements that affect the value of a DW element can facilitate the execution of alternative analysis scenarios, e.g., determining whether a change of the values of specific source elements would affect a decision, and how.

- Quality of the analysis results: provenance can help in the qualitative evaluation of the results by contributing in the determination of the accuracy, timeliness, and confidence of the information presented to the user; that could be very useful, for example, in the case of mashups applications.
- Optimization of ETL workflows: the whole process should be optimized due to specific constraints in the available time window; e.g., determining the frequency or execution order of specific ETL operations.

Our Objectives

Nevertheless, provenance has not been the primary focus of research in this area, and typically, it is not addressed explicitly. (However, some research efforts have already presented results on the related topic of data lineage, e.g., [1], which is considered as one aspect of the problem.) Therefore, we would like to bridge our work on a framework towards the modeling of ETL processes and the optimization of ETL workflows with data provenance. The goal of our research in this topic was to facilitate, manage, and optimize the design and implementation of ETL processes both during the initial design and deployment stage as well as during the continuous evolution of a DW [2-5]. In particular, our results include:

- The provision of a novel conceptual model for the tracing of inter-attribute relationships and the respective ETL transformations in the early stages of a DW project along with a methodology for its construction.
- The provision of a novel logical model for the representation of ETL processes with two main characteristics: genericity and customization.
- The presentation of a methodology for the semi-automatic transition from the conceptual to the logical model for ETL processes.
- An attempt to use ontology-based mechanisms to semi-automatically capture the semantics and the relationships among the various sources.
- The tuning of an ETL scenario through several algorithms for the optimization of the execution order of the transformations.

In this presentation, we will present the above results and discuss their close relationship with data provenance, along with open issues towards establishing a formal, end-to-end model for provenance in the context of ETL scenarios considered in a broader setting not necessarily related to the traditional DW environment.

References

[1]. Y. Cui, J. Widom. Lineage Tracing for General Data Warehouse Transformations. In VLDB Journal 12(1), 2003.

[2]. A. Simitsis, P. Vassiliadis, T. K. Sellis. Optimizing ETL Processes in Data Warehouses. In IEEE International Conference on Data Engineering (ICDE), 2005.

[3]. A. Simitsis, P. Vassiliadis, T. K. Sellis. State-Space Optimization of ETL Workflows. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 17(10), 2005.

[4]. A. Simitsis, P. Vassiliadis. A Method for the Mapping of Conceptual Designs to Logical Blueprints for ETL Processes. In Decision Support Systems (to appear.)

[5]. P. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis, S. Skiadopoulos. A Generic and Customizable Framework for the Design of ETL Scenarios. In Information Systems, 30(7), 2005.