Metadata Propagation in Large, Multi-Layer Database Systems

Arnon Rosenthal
The MITRE Corporation
arnie@mitre.org

Edward Sciore
Boston College and MITRE
sciore@bc.edu

1. The problem

Large database systems (e.g., federations, warehouses) have several virtual or physical databases, each derived from other layers. Today, metadata tend to be locked into the schema where it is supplied, and are rarely available through other schemas. Propagating metadata to other schemas is difficult for three reasons.

First, syntactically similar metadata values can propagate very differently. For example, the access rights associated with a join view may be the intersection of the source access rights, whereas the completeness of the view may depend on the join domain’s cardinality.

Second, a derived table’s metadata may be an order of magnitude larger than its view derivation. This scale precludes any solution that requires humans to extend view definitions to specify metadata transformations.

Third, when updates to metadata can be propagated to another schema, it may be necessary to negotiate about whether (or how) the update should be performed. Several administrator roles may be involved (at least one per schema), and each needs to see metadata relative to their schema of interest.

2. Our solution

We address all three difficulties by means of a semi-automatic inference of meta-attribute values. Humans need only supply a relatively small number of basic inference rules [5], one for each type of metadata (e.g., units, accuracy) and query operator (e.g., Select, multiply). Rules are applied to produce bounds on view meta-attribute values.

Semi-automatic inference addresses the diversity of inference semantics by allowing the administrator to assign a pre-specified behavior (via an inference rule) to each meta-attribute. It promotes reuse – different meta-attributes can use the same inference rule, and a meta-attribute can use the same rule wherever it appears in the schema. By providing metadata in terms of the recipient’s schema, it eases communication [3].

We combine this direct inference with the principle that one can infer correct metadata bounds using any query that computes view V (not just V’s definition). This approach has several desirable properties. First, metadata can be asserted on any table, source or derived. Second, metadata obtained through different derivations can be combined. By reusing query processing theory and (ideally) software, our approach handles non-Boolean properties and reduces the cost of metadata wizards [1].

Our html mockup of an administration interface [2] illustrates how non-programmers can effectively manage rules’ applicability. Our formal research focuses on a particular kind of metadata – access permissions, to extend the “single copy” semantics of SQL query/update to security metadata in a federation [3]. We have been quite successful in upward inference; downward inference (like view update) is ambiguous and difficult.

3. Open problems

Globalizing the metadata is a long-term challenge, which raises many research problems, such as:

- Identify inference and “view update” rules for all types of metadata and operations [4].
- Design a framework system into which rules can be plugged. Design the administrator’s interface.
- Services for negotiating about metadata values (or data values) among autonomous administrators, e.g., what-if analyses, suggesting changes
- Algorithmics: wizards to implicitly enumerate a large set of queries equivalent to a view definition.

References (URLs are in www.cs.bc.edu/~sciore/papers)