The Many Roles of Meta Data in Data Integration

Philip A. Bernstein
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6099 U.S.A.
philbe@microsoft.com

ABSTRACT
This paper is a short introduction to an industrial session on the use of meta data to address data integration problems in large enterprises. The main topics are data discovery, version and configuration management, and mapping development.

1. USER’S VIEWPOINT
From a database system engine viewpoint, data integration is the problem of processing queries against mediated schemas, where each mediated schema is mapped to a set of data sources. From a user’s viewpoint, data integration problems focus more on the development and management of the mappings that feed the query processing view.

Data integration is a very large fraction of the work in information technology (IT) departments in large enterprises. Based on anecdotal information from users and vendors, data integration appears to comprise at least half of the work done by software engineers in IT departments today. Some of the main data integration activities are the following:

- Data warehousing – developing extraction, transformation and loading (ETL) programs to load a data warehouse.
- Lineage tracing – integrating ETL tools and data transformation applications to show the sequence of transformations used to populate a field in a database or data warehouse.
- Data integration – developing mappings from a mediated schema to data source schemas.
- Enterprise application integration – wrapping applications with message interfaces and developing mappings between the input and output messages of different applications.
- Business-to-Business E-commerce – developing software to translate messages to and from business partners’ format.
- Information resource management – maintaining an inventory of information assets, such as databases, applications, forms, and message types.
- Information portals – displaying information from a variety of applications and databases in a common user interface.
- Document management – Attaching meta data to documents, placing them in a searchable store, and integrating them with workflow processes.
- Domain model development – To facilitate the above integration tasks, develop a domain model (e.g. in UML), to serve as intermediary between two databases or applications to be integrated.

2. SESSION OVERVIEW
The integration activities of the previous section are quite diverse. However, there are many common meta data problems that are generic in the sense that they appear in most of these activities. Three of them are covered in this session, namely:

- Data discovery – finding relevant data sources to be integrated and capturing their schemas to facilitate access. Then integrating them into a form that is comprehensible to users and mapped to sources for subsequent querying.
- Version and configuration management – maintaining versions of schemas and of mappings between them, to manage their evolution and enable accurate lineage tracing.
- Mapping development – designing mappings between schemas, based on schema definitions, schema constraints, and sample data instances, and then generating and optimizing code to execute the mappings.

Commercial solutions to these common problems are presented in this session by MetaMatrix [3], Meta Integration Technologies [1], and IBM [2], respectively. The presentations also touch on related meta data problems, such as impact analysis, mapping visualization, mapping composition, and schema and mapping evolution.

3. REFERENCES

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGMOD 2005 June 14-16, 2005, Baltimore, Maryland, USA
Copyright 2005 ACM 1-59593-060-4/05/06 $5.00