Model Management and Schema Mappings: Theory and Practice

Philip A. Bernstein
Microsoft Research
Redmond, WA, U.S.A.
philbe@microsoft.com

Howard Ho
IBM Almaden Research Center
San Jose, CA, U.S.A.
ho@almaden.ibm.com

ABSTRACT

We present an overview of a tutorial on model management—an approach to solving data integration problems, such as data warehousing, e-commerce, object-to-relational mapping, schema evolution and enterprise information integration. Model management defines a small set of operations for manipulating schemas and mappings, such as Match, Compose, Inverse, and Merge. The long-term goal is to build generic implementations of the operations that can be applied to a wide variety of data integration problems.

1. TUTORIAL OVERVIEW

Data integration is a large fraction of the database business. Solving a data integration problem requires the manipulation of metadata: database schemas that describe the sources and targets of the integration and mappings between those schemas. Work on metadata problems goes back to at least the early 1970's, when data translation was a hot database research topic. However, until recently there was no widely-accepted conceptual framework for this field, as there is for most other database topics.

In recent years, work in this field has coalesced around a small set of operations for manipulating metadata, with the goal of applying them to a wide variety of data integration problems. They include operations to match schemas, compose mappings, diff schemas, invert mappings, merge schemas, and translate schemas and mappings between languages. This conceptual framework is called model management in [1].

While model management has been broadly focused on the collection of such metadata operations as a framework, a key component of this framework has been developed independently as part of the Clio project [3][7]. Clio started at about the same time that the model management vision was proposed, but it had a particular focus on the problem of generating schema mappings and the associated run-time artifacts (e.g., data translation programs) [6]. The Clio research pioneered several algorithms and languages for generating, manipulating and representing schema mappings. Later, it moved towards the vision of model management by addressing ways to combine individual schema mappings into more general flows or data integration applications. Concurrently, model management research moved towards the types of mapping generation and compilation problems introduced by the Clio project. Today, the two lines of work are following similar research directions.

Much is now known about model management operations. There are theoretical results that explain their semantics and basic properties. There are practical algorithms for their implementation. There are prototypes of many of those algorithms. And there are products that use the operations.

There is a big research literature on these operations. Most of these papers predate the introduction of model management. Some of them are more recent. There is now a steady stream of new papers appearing regularly, many of them from the tutorial speakers' research groups. Recent overviews of this area cite over a hundred papers [2][4]. An on-line bibliography about schema evolution lists over three hundred references [8].

This tutorial offers an introduction to model management and related schema mapping problems and a synopsis of the state of the art. It covers the following topics:

- What are schema mapping problems, why are they important, and why are they hard to solve?
- A design pattern for solving schema mapping problems.
- A short history of model management—the initial proposal and how it has changed.
- Semantics of the main operations, algorithms to implement them, and open problems.
- A short history of the Clio project and its mapping-related technologies.
- Two applications: object-to-relational mapping in Microsoft's ADO.NET Entity Framework [5]; and schema matching, schema mapping and query generation in IBM Rational Data Architect [9].

This tutorial is intended for both engineers and researchers. The former will learn about solutions to use and hard problems to avoid. The latter will get a snapshot of the research field and problems that are worth tackling next.

2. SPEAKERS

Philip A. Bernstein is a Principal Researcher at Microsoft and Affiliate Professor at University of Washington. For the past fifteen years his research focus has been metadata management. He is an ACM Fellow, a winner of the ACM SIGMOD Innovations Award, and a member of the U.S. National Academy of Engineering.

Howard Ho is a research staff member at the IBM Almaden Research Center. He received his Ph.D. in parallel processing

VLDB '07, September 23-28, 2007, Vienna, Austria.

research from Yale in 1990. He has been managing the Clio project for the last six years. Before that, he was active in data mining and parallel processing. He is currently Editor-in-Chief of Journal of Interconnection Networks.

3. REFERENCES

[1] P.A. Bernstein, A.Y. Halevy, R. Pottinger: A vision of management of complex models. SIGMOD Record 29(4): 55-63.

[2] P.A. Bernstein, S. Melnik: Model management 2.0: manipulating richer mappings. SIGMOD 2007: 1-12.

[3] L.M. Haas, M.A. Hernández, C.T.H. Ho, L. Popa. M. Roth: Clio grows up: from research prototype to industrial tool. SIGMOD 2005: 805-810.

[4] P. G. Kolaitis: Schema mappings, data exchange, and metadata management. PODS 2005: 61-75.

[5] S. Melnik, A. Adya, P.A. Bernstein: Compiling mappings to bridge applications and databases. SIGMOD 2007: 461-472.

[6] R.J. Miller, L.M. Haas, and M.A. Hernández: Schema Mapping as Query Discovery. *VLDB 2000*: 77-88.

[7] R.J. Miller, M.A. Hernández, L.M. Haas, L. Yan, H. Ho, R. Fagin, L. Popa: The Clio project: managing heterogeneity. SIGMOD Record 30(1): 78-83 (2001).

[8] E. Rahm, P.A. Bernstein: An online bibliography on schema evolution. SIGMOD Record 35(4): 30-31 (2006).

[9] M. Roth, M.A. Hernández, P. Coulthard, L.-L. Yan, L. Popa, C.T.H. Ho, C.C. Salter: XML mapping technology: Making connections in an XML-centric world. IBM Systems Journal 45(2): 389-410 (2006).