

DataFoundry Data Warehousing and Integration for Scientific Data Management

*R. Musick, T. Critchlow, M. Ganesh, K. Fidelis, A. Zemla
and T. Slezak*

U.S. Department of Energy

February 29, 2000

Lawrence
Livermore
National
Laboratory

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Work performed under the auspices of the U. S. Department of Energy by the University of California Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

This report has been reproduced
directly from the best available copy.

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information
P.O. Box 62, Oak Ridge, TN 37831
Prices available from (423) 576-8401
<http://apollo.osti.gov/bridge/>

Available to the public from the
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd.,
Springfield, VA 22161
<http://www.ntis.gov/>

OR

Lawrence Livermore National Laboratory
Technical Information Department's Digital Library
<http://www.llnl.gov/tid/Library.html>

An LDRD Final Report

DataFoundry Data Warehousing and Integration for Scientific Data Management

Tracking code: 97-ERD-033

Ron Musick (Principal Investigator)
Terence Critchlow and Madhavan Ganesh

Center for Applied Scientific Computing

Krzysztof Fidelis, Adam Zemla

Biology and Biotechnology Research Program

Tom Slezak

Computing Application Organization / Human Genome Center

Overview

Data warehousing is an approach for managing data from multiple sources by representing them with a single, coherent point of view. Commercial data warehousing products have been produced by companies such as Rebrick, IBM, Brio, Andyne, Ardent, NCR, Information Advantage, Informatica, and others. Other companies have chosen to develop their own in-house data warehousing solution using relational databases, such as those sold by Oracle, IBM, Informix and Sybase. The typical approaches include federated systems, and mediated data warehouses, each of which, to some extent, makes use of a series of source-specific wrapper and mediator layers to integrate the data into a consistent format which is then presented to users as a single virtual data store. These approaches are successful when applied to traditional business data because the data format used by the individual data sources tends to be rather static. Therefore, once a data source has been integrated into a data warehouse, there is relatively little work required to maintain that connection. However, that is not the case for all data sources.

Data sources from scientific domains tend to regularly change their data model, format and interface. This is problematic because each change requires the warehouse administrator to update the wrapper, mediator, and warehouse interfaces to properly read, interpret, and represent the modified data source. Furthermore, the data that scientists require to carry out research is continuously changing as their understanding of a research question develops, or as their research objectives evolve. The difficulty and cost of these updates effectively limits the number of sources that can be integrated into a single data warehouse, or makes an approach based on warehousing too expensive to consider.

Our research has been driven in response to the concerns above. We have developed, implemented and published several papers on a system and set of techniques that significantly reduce the cost (and time) needed to add a new source to the warehouse, or respond to a change in an existing source. In part, these techniques are based on a novel mechanism for automatically generating much of the code required for an operational data warehouse – the *mediator generator*. This code makes use of a minimal collection of declarative information, or

metadata, that the warehouse administrator provides instead of actually writing the mediator code. The core technology that has been developed is directly applicable to any scientific domain that needs efficient access to multiple heterogeneous data sources, and thus has broad applicability in the Lab. Our research results are described in depth in the seven attached reports, each of which is briefly summarized below.

The research concepts have been validated in a series of prototypes in use by biologists at the Lab. The DataFoundry warehouse prototype integrates protein structure data from the Protein Data Bank (PDB) [CCD96], protein sequence data from SWISS-PROT [BA96], protein structural taxonomy data from SCOP [MBH95], and genomics expressed sequence tag data from dbEST [BLT93]. The prototype supports research activities of biologists at LLNL. The prototype's capabilities include: (1) a web-based interface that allows ad-hoc SQL queries, graphical queries, or the use of query templates to pose questions; (2) direct access to the warehouse from C, C++ and Perl; and (3) basic protein homology search functionality. We have developed an additional set of interfaces with Java and applet/servlet code that runs in a browser. With these interfaces, DataFoundry enhances the scientist's ability to search for and describe proteins and ESTs related to target sequences by sequence and structural homology, as well as similarity in physical and biological characteristics.

This is a multidisciplinary effort involving computer scientists and biologists from the Center for Applied Scientific Computing, the Computing Applications Organization, the Joint Genome Institute, and the Biology and Biotechnology Research Program. We have filed a patent application on DataFoundry technology. A software technology company from Texas is negotiating an exclusive license for DataFoundry code with the intent to enhance their future product offerings. We have follow-on funding to carry on research and development in the genomics area from the Joint Genome Institute. We have been funded by the Accelerated Strategic Computing Initiative to bring the technology to bear on large-scale computational data. There are several other potential funding options that are currently pending, through both internal and external research proposals.

The seven attached reports that describe this effort are listed below:

In the first report, *Digital Libraries and Bioinformatics* (UCRL-JC-131329), we discuss the major challenges facing any community of users that need to make use of multiple data sources distributed on the WWW, and provide a glimpse of a digital libraries approach to these challenges. The main source of difficulties lies in the fact that for many scientific communities, the "data-scape" is best described as a cottage-industry with hundreds of small, independent, heterogeneous data collections that are not linked by standard concepts, vocabularies, formats, or interfaces. For all intents and purposes, though publicly available, the difficulties faced in making use of the data makes it inaccessible to the scientist. In 1998 this report was submitted as a position paper to IEEE Transactions on Information Technology in Biomedicine.

In the second report, *DataFoundry: Information Management for Scientific Data* (UCRL-JC-133640), we provide an overview of our approach to integrating heterogeneous data. In particular, DataFoundry's mediated architecture is a unique mix of federated, multidatabase, and the more traditional approaches to data warehousing. DataFoundry's improved data integration process, which is derived from metadata-based code generation, is described as well. This report will appear in the March issue of IEEE Transactions on Information Technology in Biomedicine, 2000.

In the third report, *Experiences using a Meta-data Based Information Infrastructure* (UCRL-JC-134854), we identify some of the practical issues we have faced in building DataFoundry that stem from our usage of the underlying metadata infrastructure. For example, our warehouse

schema and the ingest process was optimized for clarity and efficiency – but that made it difficult to properly delete objects efficiently due to multiple referencing. Also, for scalability reasons, we needed to partition the warehouse into several databases, which introduces difficulties in maintaining a global schema. This report will appear in the Proceedings of the 2nd International Workshop on Biomolecular Informatics, held in Atlantic City, New Jersey in February, 2000.

In the fourth report, *Metadata Based Mediator Generation* (UCRL-JC-130234), we describe the details of our mediator generation, which implements a unique data ingest process. The mediator generator we have developed uses carefully defined metadata to create a collection of classes and libraries that define class hierarchies and abstraction layers. These are used by the mediator (also created by the mediator generator) to carry out tasks such as mapping and transforming source data into a consistent semantics and format, as well as actually loading that data into the warehouse. This report appeared in the Proceedings of the 3rd International Conference on Cooperative Information Systems, held in New York, New York in August 1998.

In the fifth report, *Automatic Generation of Warehouse Mediators Using an Ontology Engine* (UCRL-JC-129930), we describe the details of the metadata structure and content. The metadata is used to describe the data in the DataFoundry prototypes, drive the ingest process, and drive much of the interface as well. The metadata is composed of four types of information: domain-level abstractions (for semantic matching), database abstractions (for describing source data formats), mappings (for mapping source objects to warehouse objects), and transformations (a repository of common transformations between different representations of identical objects). It is important to note that while each instantiation of the metadata is domain-specific, the metadata schema itself is domain-independent, and will apply to any domain with data of any format. This report appeared in the Proceedings of the 5th International Workshop on Knowledge Representation Meets Databases, held in Seattle, Washington in May 1998.

In the sixth report, *Detecting Data and Schema Changes in Scientific Documents* (UCRL-JC-134444), we describe a primarily theoretical approach to both detecting when a data source schema has changed, and identifying the type of change that has occurred. Schema evolution is a continuous process with the semi-structured documents found on the web and in scientific data sources. This approach describes a specialized graph representation of semi-structured documents, classifies and formalizes the types of changes that can occur over the course of schema evolution, and identifies rules for detecting these changes. This is used in an approach for identifying how a schema has evolved by automated analysis (i.e. with genetic algorithms) of a history of the parsing errors for that data source. This work led to a Ph.D. thesis by one of the authors. This report will appear in the Proceedings of the 4th IEEE Conference on Advances in Digital Libraries, to be held in Washington, DC in May 2000.

In the seventh report, *Practical Lessons in Supporting Large-Scale Computational Science* (UCRL-JC-135606) we discuss supporting query and visualization support on large computational science data with techniques from the database community. This report describes a set of experiments that were carried out, as well as the lessons that were learned in applying object-relational DBMS technology. This work paved the way for our follow-on activities within the Accelerated Strategic Computing Initiative. This report is a condensed version of an internal LLNL technical report (UCRL-ID-129903). This report appears in ACM Sigmod Record, V28 #4 in December, 1999.

References

- [BA96] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its new supplement TrEMBL. *Nucleic Acids Res.* 24:21-25(1996)
- [BLT93] M. S. Boguski, T. M. Lowe and C.M. Tolstoshev. dbEST – database for expressed sequence tags. *Nat. Genet.* 4(4):332-3. Aug 1993.
- [CCD96] J. Callaway, M. Cummings, B. Deroski, P. Esposito, A. Forman, P. Langdon, M. Libeson, J. McCarthy, J. Sikora, D. Xue, E. Abola, F. Bernstein, N. Manning, J. Sussman. Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description Version 2.1. Tech Report at www.pdb.bnl.gov. October. 1996
- [CGM98a] T. Critchlow, M. Ganesh, R. Musick, "Automatic Generation of Warehouse Mediators Using an Ontology Engine", Conference on Knowledge Representation Meets Databases, 1998.
- [CGM98b] T. Critchlow, M. Ganesh, R. Musick, "Metadata-Based Mediator Generation", Submitted to Conference on Cooperative Information Systems, 1998.
- [CGM98c] T. Critchlow, M. Ganesh, R. Musick, K. Fidelis, and T. Slezak, "DataFoundry: Warehousing Techniques for Dynamic Environments", Submitted to IEEE Information Technology Applications in Biomedicine, 1998.
- [MBH95] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology.* 247:536-540.