

Data Integration Via Universal Keys

Robert Grossman
University of Illinois at Chicago

September 3, 2006

1 Introduction

We describe an infrastructure for integrating distributed data called DataSpace [3]¹ In addition to supporting data and metadata, the infrastructure also supports globally unique keys for integrating data that we call universal keys. In contrast to some of the standard approaches to data integration, DataSpace does not try to achieve the full semantic integration of distributed data, but instead provides the minimum infrastructure necessary to integrate distributed data that is attached to universal keys. We also describe some applications that have been built with this infrastructure in astronomy, bioinformatics, and earth science.

TCP-based web and grid services, as usually deployed, have been shown to have problems for integrating very large data sets over wide area, high performance networks [1]. Recently, we have implemented a peer-to-peer version of DataSpace called Sector that is designed for working with large data sets over wide area, high performance networks [7]. Sector has been used to distributed the terabyte size catalog data for the Sloan Digital Sky Survey (SDSS) from Chicago to locations in the U.S., Europe and Asia.

This paper is based in part on [2]

2 Key Concepts

In the DataSpace approach to data integration, we assume that data is distributed, has attached metadata, and is associated with globally unique identifiers called universal keys.

Data. DataSpace is designed to work with several different types of data, including relational data, distributed columns of data, semi-structured data and blobs of data. We refer to the components of data as elements. For example, elements of relational data are fields, elements of columns of data are columns, elements of semi-structured XML data are XML elements, etc.

Metadata. We assume that there is descriptive information about data sets, and components of data sets, such as columns or elements. Examples of metadata can include data types, associated schemas and taxonomies, and provenance information.

Universal keys. We assume that data, metadata, and their components may have, but are not required to have, globally unique identifiers (GUIDs) associated with them. We refer to these GUIDs as universal keys. Universal keys are used to define the integration of data.

Derived data. DataSpace allows new data to be derived from existing data in two ways, as described below.

DataSpace supports the following operations on data:

¹In 2006, Halevy, Franklin and Maier [8] introduced an approach to data integration also called dataspace that also does not require the full semantic integration of information. In contrast, the dataspace that was introduced in 2002 in [3] is less ambitious and does not require support for constraints, consistency, or recovery.

Append, delete and update. Using universal keys, we can define the operations of appending data, deleting data, and updating data. For example, for blob data, given a data set of blobs with universal keys, if a new data blob shares a universal key with an existing data blob, the old blob is updated with the new blob; otherwise, the new blob is appended to the data set of blobs.

Integration. Using universal keys, we can define the integration of new data elements by concatenating data elements and identifying two or more data records with the same universal keys.

Derivation. Derivation is the process of taking data as input and producing new data as output. There are two common approaches that are used in practice: derived data may be defined by executing a procedural program or by executing a declarative program.

3 Examples

We have used the DataSpace infrastructure to build several different applications, including the three described below.

Integrating Astronomical Data Sets. We have performed various experimental studies integrating distributed astronomical data sets. In one of these experiments, we put a copy of the Sloan Digital Sky Survey (SDSS) Data on a computer cluster in Tokyo and a copy of the Two-Micron All-Sky Survey (2MASS) on a computer cluster in Chicago. The SDSS consisted of about 82 million stars in the visible spectrum, while the 2MASS data consisted of about 208 million stars in the infrared spectrum. We streamed both terabyte size data sets to Seattle, where we joined the two data sets using a common vector-valued universal key consisting of (right ascension, declination) pairs, measured in degrees. We used an algorithm supplied by Alex Szalay to identify candidate brown dwarfs by looking at data from both spectrums. In this fashion, we identified over 250,000 candidate brown dwarfs that would have been difficult to identify from just one of the two data sets. We note that universal keys are all the data integration infrastructure required to perform the type of analysis required to identify candidate brown dwarfs from these two data sets.

Proteomics Data Sets. We built several other applications that used a DataSpace infrastructure to integrate proteomics data. In particular, we built several applications that required integrating databases of chemical compounds. It turns out that the standard mechanisms for identifying chemical compounds, such as CAS identifiers or UNIQUE SMILES strings are not necessarily unique. For this reason, we developed unique identifiers for chemical compounds [5] and [6] called Universal Chemical Keys or UCKs. We used these universal keys to integrate local proteomics repositories with data from the NIH NCI database [5]. We have also used DataSpace services and UCKs to integrate metabolic pathways from the KEGG and MetaCyc databases [6].

Integrating Geo-Science Data Sets. In work described in [3], we placed approximately 100 Gigabytes of Community Climate Model (CCM3) data from the National Center for Atmospheric Research (NCAR) on a DataSpace Server located at NCAR. CCM3 data is used by scientists to study CO2 warming and climate change, climate prediction and predictability, atmospheric chemistry, paleoclimate, biosphere-atmosphere transfer and nuclear winter. The data consists of monthly satellite measurements of global surface temperatures, precipitation, ozone levels and vegetation index. There are three universal keys for this application: latitude, longitude and time. The DataSpace client we developed for this application supports queries by universal key, by attribute ranges, by attribute, and by data set. The client can view the data, download the data, graph the data, and do simple exploratory operations on the data. The DataSpace client can also compare the CCM3 data and overlay the CCM3 data with other data sets sharing one or more of the same universal keys.

References

- [1] Ian Foster and Robert L. Grossman, Data Integration in a Bandwidth Rich World, Communications ACM, Volume 46, Issue 11, November, 2003, pages 50-57.
- [2] Ian Foster and Robert L. Grossman, Middleware for Data Integration, in preparation.
- [3] Robert Grossman, and Marco Mazzucco, DataSpace - A Web Infrastructure for the Exploratory Analysis and Mining of Data, IEEE Computing in Science and Engineering, July/August, 2002, pages 44-51.
- [4] Asvin Ananthanarayan, Rajiv Balachandran, Yunhong Gu, Robert Grossman, Xinwei Hong, Jorge Levera, Marco Mazzucco, Data Webs for Earth Science Data, Parallel Computing, Volume 29, 2003, pages 1363-1379.
- [5] Robert L. Grossman, Pavan Kasturi, Donald Hamelberg, Bing Liu, An Empirical Study of the Universal Chemical Key Algorithm for Assigning Unique Keys to Chemical Compounds, Journal of Bioinformatics and Computational Biology, 2004, Volume 2, Number 1, 2004, pages 155-171.
- [6] Greeshma Neglur and Robert L. Grossman, Assigning Unique Keys to Chemical Compounds for Data Integration: Some Interesting Counter Examples, 2nd International Workshop on Data Integration in the Life Sciences (DILS 2005), La Jolla, July 20-22, 2005.
- [7] Yunhong Gu, Robert L. Grossman, Alex Szalay and Ani Thakar, Distributing the Sloan Digital Sky Survey Using UDT and Sector, Second IEEE International Conference on E-Science and Grid Computing, 2006, to appear.
- [8] Alon Halevy, Michael Franklin and David Maier, Principles of Dataspace Systems, Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 1 - 9, 2006.
- [9] NCI Open Database Compounds, retrieved from <http://cactus.nci.nih.gov/> on August 10, 2006.