Dataspaces: Co-existence with Heterogeneity

David Maier
Portland State University

Alon Halevy
Google Inc. and U. Washington

Michael Franklin
University of California at Berkeley

Most information management scenarios today rarely have a situation in which all the data that needs to be managed can fit nicely into a single management system, such as a relational database or a knowledge base. Instead, we need to manage a set of loosely connected data sources, and typically face the following recurring challenges:

- Users want be able to search the entire collection without having knowledge of individual sources. In some cases, they merely want to know *where* the information exists as a starting point to further exploration.
- An organization may want to enforce certain rules, integrity constraints, or conventions (e.g., on naming entities) across the entire collection, or track flow and lineage between systems. Furthermore, the organization needs to create a coherent external view of the data.
- The administrators may want to impose a single "support system" in terms of recovery, availability, and redundancy, as well as uniform security and access controls.
- Users and administrators need to manage the evolution of the data, both in terms of content and schemas, in particular as new data sources get added (e.g., as a result of mergers or new partnerships).
- The organization would like to globally capture human interactions with the sources (e.g., querying, excerpting, annotating, co-use) to provide value-added information about semantics, linkages and quality.

The aforementioned data management challenges are ubiquitous – they arise in enterprises (large or small), coordination within and across government agencies, data analysis in large science-related research or development projects, management of libraries (digital or otherwise), information collection and dissemination in the battlefield, search on one's PC desktop or other personal devices, coordination between devices in a "smart" home, and in search for structured objects on the web. In these scenarios, there is some well-understood scope and control across the data and systems within these organizations, and hence one can identify a space of data, which, if managed in a principled way, will offer significant benefits to the organization.

Traditionally, the problem of querying multiple sources has been addressed by research on information integration systems. In such a system, we let the data reside in the individual systems, and create *semantic mappings* to a logical mediated schema. Query processing proceeds by reformulating queries over a mediated schema onto appropriate queries on the individual data sources. Creating these semantic mappings can be a major bottleneck in building these systems because they require significant upfront effort.

We recently introduced *dataspaces* (Franklin, Halevy, & Maier 2005) as a new abstraction for data management for the aforementioned scenarios, and proposed the development of DataSpace Support Platforms (DSSPs). In a nutshell, a DSSP offers a suite of interrelated services and guarantees that enables an application developer to focus on the specific challenges of an application, rather than the recurring challenges involved in dealing consistently and efficiently with large amounts of interrelated but disparately managed data.

Dataspace management is not an information integration approach; rather, it is more of an *information co-existence* approach. The goal of DSSPs is to provide base functionality over all information sources, regardless of how integrated they are. For example, a DSSP can provide keyword search over all of the data sources it contains, similar to the way that existing desktop search systems. When more sophisticated operations are required, such as database-style query processing, data mining or sophisticated reasoning over certain sources, then additional effort can be applied to more closely integrate those sources, in an incremental, "pay-as-you-go" fashion. Furthermore, as we perform more integration tasks, we expect the cost of integration to decrease.

We will describe how recent work from both the AI and DB communities can be applied to building DSSPs and some of the challenges we face. We will also argue that such a bottom-up pay-as-you-go approach should also be applied to knowledge-rich settings.

References

Franklin, M.; Halevy, A.; and Maier, D. 2005. From databases to dataspaces: A new abstraction for information management. *Sigmod Record* 34(4):27–33.