

Integrating Data and Services: Products and Challenges at BEA

Michael J. Carey
BEA Systems, Inc.
2315 North First Street
San Jose, CA 95131
mcarey@bea.com

II FOR SOA AT BEA

Data needed for today's enterprise applications lives in a variety of information sources, including relational databases, packaged applications, various homegrown applications, external Web services, and files. An application developer who wishes to integrate or even simply access data in such a wide range of sources must cope with at least three kinds of heterogeneity. First, each source has its own associated data model or data format (tables, WSDLs with XML Schemas, XML documents, files, and so on); this is *model heterogeneity*. Second, each type of source has its own programming interface (JDBC/SQL, SOAP, file I/O calls, RPC, and custom APIs like BAPI for SAP); this is *API heterogeneity*. Finally, a given piece of information such as a name or an address may be represented differently in different sources; this is *schema heterogeneity*. The BEA AquaLogic Data Services Platform (ALDSP) is aimed at alleviating these problems in the context of today's emerging service-oriented architectures (SOA).

ALDSP [BEA05] was first released in mid-2005, as ALDSP 2.0, and it has been followed up with two subsequent releases (2.1 in March 2006 and 2.5 in September 2006). The purpose of ALDSP is to make it easy for SOA developers to design, develop, deploy, and maintain a *data services layer* in the new world of service-oriented architectures. Just as relational databases dramatically changed the daily lives of developers of data-centric applications in the last millennium, replacing tedious (to develop) and brittle (to maintain and tune) procedural programs with declarative SQL queries and updates, ALDSP aims to simplify the lives of developers of data-centric SOA applications by giving them declarative tools to do their jobs. Put simply, ALDSP is BEA's answer to the question: *Data + SOA = ?* In order to answer this question, perhaps the most important capability of ALDSP is its support for integrating data from services with data from traditional databases.

ALDSP is based on the *data service modeling* methodology described in [BCMMPT05]. Briefly, a universe of enterprise data sources of interest -- regardless of their nature -- can be modeled as a set of interrelated, business-meaningful *data services*. Each data service has a shape (an XML schema) and a set of service methods for accessing its data in various ways, for modifying its data, as well as for navigating to data from other, related data services. Technically, ALDSP can be viewed as something of a reincarnation of the Functional Data Model [S81] and the functional approach to integration [LR82] updated to use widely accepted modern and open standards. XML-related standards provide the technical foundation for this methodology. The relevant W3C standards include XML, Web services, XML Schema, and (last but not least) XQuery. A technical overview of ALDSP is presented in [C+06], and more details of its approach to federated query processing are available in [BCLWE06].

SOA II CHALLENGES

Although the field of Information Integration (II) is nearly twenty-five years old [LR82], numerous challenges and opportunities remain [H+05]. The integration of services with traditional data, particularly given the lack of readily available, database-style semantic information and quality of service guarantees, poses a number of challenges. Some we are working on today at BEA. Others we just wish we were, and would thus love to see the academic community tackle effectively. Some of the open challenges include:

Query Costing and Statistics: In a heterogeneous world involving non-database data, varying system loads, and varying network latencies, database-style statistics and cost information are simply not available. Dynamic, observational approaches are needed.

Semantic Query Optimization: In an environment where a given datum may be available from alternative functions (e.g., `getCustomersByName`, `getCustomersByZipcode`, `getCustomerByCustId`), query optimization requires selecting the best logical access path that answers a given query. A framework for annotating service methods with semantic information and rewriting queries based on the captured information is needed.

Transactions and Sagas: Many enterprise data sources are not transaction-aware, or if they are, are more often than not unwilling to participate in traditional distributed transactions [ACKM04]. Techniques are needed for providing approximate transaction-like guarantees when updating non-transactional sources and their limitations must be understood and handled effectively.

Practical Schema Integration: Effective techniques to aid enterprise data architects in coping with schema heterogeneity, even after twenty-five years of work in the field, remain elusive. Recently, there has been a flurry of activity on AI-based approaches to resolving schema differences and automatically inferring mappings. One property of any practical approach to this problem, at least in the enterprise, is that mappings will never be left to chance. Approaches are needed that assist (but do not replace) human data analysts in dealing with schema heterogeneity, especially for large schemas.

REFERENCES

- [ACKM04] Alonso, G., Casati, F., Kuno, H., and Machiraju, V. *Web Services: Concepts, Architectures, and Applications*. Springer-Verlag, Berlin/Heidelberg, 2004.
- [BEA05] *BEA AquaLogic Data Services Platform*, <http://www.bea.com/dataservices>, 2005.
- [BCMMPT05] Borkar, V., Carey, M., Mangtani, N., McKinney, D., Patel, R., and Thatte, S., XML Data Services. *Int'l. Journal of Web Service Research* 3(1), January-March 2006.
- [BCLWEO06] Borkar, V., Carey, M., Lychagin, D., Westmann, T., Engovatov, D., and Onose, N. Query Processing in the AquaLogic Data Services Platform. *Proc. VLDB Conf.*, Seoul, Korea, September 2006.
- [C+06] Carey, M., et al. Data Delivery in a Service-Oriented World: The BEA AquaLogic Data Services Platform. *Proc. ACM SIGMOD Conf.*, Chicago, IL, June 2006.
- [H+05] Halevy, A., et al. Enterprise Information Integration: Successes, Challenges, and Controversies. *Proc. ACM SIGMOD Conf.*, Baltimore, MD, June 2005.
- [LR82] Landers, T., and Rosenberg, R. An Overview of MULTIBASE. *Proc. 2nd Int'l. Symp. on Distributed Data Bases*. Berlin, Germany.
- [S81] Shipman, D. The Functional Data Model and the Data Language DAPLEX. *ACM Trans. on Database Systems* 6(1), March 1981.