Chapter 13
Intelligent Information Integration:
Reclaiming the Intelligence

Naveen Ashish
University of California, Irvine, USA

David A. Maluf
NASA Ames Research Center, USA

ABSTRACT

The authors present their work in the conceptualization, design, implementation, and application of "lean" information integration systems. They present a new data integration approach based on a schema-less data management and integration paradigm, which enables developing cost-effective large scale integration applications. They have designed and developed a highly scalable, information-on-demand system called NETMARK, which facilitates information access and integration based on a theory of articulation management and a context sensitive paradigm. NETMARK has been widely deployed for managing, storing, and searching unstructured or semi-structured arbitrary XML and HTML information at the National Aeronautics Space Administration (NASA). In this paper the authors describe the theory, design and implementation of our system, present experimental benchmark evaluations, and validate our approach through real-world applications in the NASA enterprise.

INTRODUCTION

This article describes an approach to achieving scalable and cost-effective information integration for large-scale enterprise information management applications. Our work is motivated by requirements in the United States National Aeronautics and Space Administration (NASA) enterprise, where many information and process management applications demand access to, and integration of information from, large numbers of information sources (in some cases up to as many as 50 differ-

DOI: 10.4018/978-1-60960-595-7.ch013

ent sources), across multiple divisions, and with information of different kinds in different formats. An example is the application of assembling an agency level annual report that requires information such as project status, division updates, budget information, personnel progress, etc., from different data sources in different departments, divisions, and centers within NASA. By the early 2000s, when we had initiated this work, intelligent information integration research projects such as SIMS, TSIMMIS, HERMES, InfoMaster, Information Manifold (Halevy, Rajaraman, & Ordille, 2006; Halevy, 2003) to name a few, that were concerned with building data integration systems based on a *mediator* architecture had reached considerable maturity. We had solutions to challenging problems such as providing efficient query processing over multiple distributed data sources, schema mapping and integration tools, wrapper technology for legacy data sources and also Internet data sources, and technologies for entity resolution and matching across multiple sources. There were also data integration start-ups such as Nimble (Draper, Halevy, & Weld, 2001), Junglee, Mergent, and Fetch, and bigger companies such as IBM touting off-the-shelf data integration technology that could address the required information integration needs. While functionally meeting the requirements, none of these technologies could provide scalable and cost-effective information integration solutions for large scale applications. The basic problem was that such middleware based technology being offered became rather "heavy-weight" in the face of large-scale applications. A significant amount of investment was required in assembling new integration applications. Particularly the effort in managing models and meta-data i.e., in describing the many sources being integrated and also in providing an integrated view over the various sources, became formidable - to the extent that this became one of the key impediments to the widespread adoption of "Enterprise Information Integration" (EII) technology in general. A testament to this is articulated in a review of EII technology (Halevy et al., 2005) where a CTO of (a then prominent) EII start-up observes *"A connected thread to this (key impediments for EII) is to address modeling and metadata management, which is the highest cost item in the first place"*.

The above problems carried over to the area of the "Semantic-Web" (Berneres-Lee, Hendler, & Lasilla, 2001) where most applications demand a heavy investment in creating various *ontologies* and further providing semantic linkages across such ontologies. The substantial effort and complexity in ontology creation and maintenance continues to be a major impediment in realizing practical semantic-web applications.

The lack of scalable and cost-efficient data integration technologies was however not because this was something that could not be achieved, but rather because the original vision of *intelligent information integration* had gone awry. The original vision of Intelligent Information Integration (or I^3) [1] research sponsors such as DARPA [2] was a nimble and flexible approach where clients could at will select and integrate information from different sources in a manner suited to their particular applications and the complexity of each new application was confined to the application itself (Figure 1(a)). In practice however this degenerated to a situation where the complexity of *all* applications was added on to the mediation layer (Figure 1(b)). The reason this happened was due to some flawed assumptions about how enterprise data should be managed and integrated. These assumptions, along with our alternative solutions are presented below, namely:

- *"Data must always be stored and managed in DBMS systems"* Actually, requirements of applications vary greatly ranging from data that can well be stored in spreadsheets, to data that does indeed require DBMS storage.
- *"The database must always provide for and manage the structure and seman-*