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. [Article copies are available for purchase from InfoSci-on-Demand.com]

Keywords: Business Intelligence; Database Management; Data Integration; Enterprise IS; Unstructured Data; XML

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,
integration of information from, large numbers of information sources (in some cases up to as many as 50 different 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.
Related Content

**Persuasion, Learning and Context Adaptation**
[www.igi-global.com/article/persuasion-learning-and-context-adaptation/100451?camid=4v1a](www.igi-global.com/article/persuasion-learning-and-context-adaptation/100451?camid=4v1a)

**Agents, Availability Awareness, and Decision Making**
[www.igi-global.com/chapter/agents-availability-awareness-decision-making/54429?camid=4v1a](www.igi-global.com/chapter/agents-availability-awareness-decision-making/54429?camid=4v1a)

**Intelligent Control of Vehicles' Number Plates on Toll Gates in Developing Nations**
[www.igi-global.com/chapter/intelligent-control-of-vehicles-number-plates-on-toll-gates-in-developing-nations/205838?camid=4v1a](www.igi-global.com/chapter/intelligent-control-of-vehicles-number-plates-on-toll-gates-in-developing-nations/205838?camid=4v1a)
Integrating Modified Delphi with Fuzzy AHP for Concrete Production Facility Location Selection

www.igi-global.com/article/integrating-modified-delphi-with-fuzzy-ahp-for-concrete-production-facility-location-selection/94620?camid=4v1a