DBMSs with Native XML Support: Towards Faster, Richer, and Smarter Data Management

Min Wang

IBM T.J. Watson Research Center
Hawthorne, NY 10532, USA

Abstract. XML provides a natural mechanism for representing semi-structured and unstructured data. It becomes the basis for encoding a large variety of information, for example, the ontology. To exploit the full potential of XML in supporting advanced applications, we must solve two issues. First, the integration of structured (relational) data and unstructured or semi-structured data, and on a higher level, the integration of data and knowledge. In this talk, we will address these two issues by introducing a solution that leverages the power of pure XML support in DB2 9.

The semistructured and structured data models represent two seemingly conflicting philosophies: one focuses on being flexible and self-describing, and the other focuses on leveraging the rigid data schema for a wide range of benefits in traditional data management. For many applications such as e-commerce that depend heavily on semistructured data, the relational model, with its rigid schema requirements, fails to support them in an effective way; on the other hand, the flexibility of XML in modeling semistructured data comes with a big cost in terms of storage and query efficiency, which to a large extent has impeded the deployment of pure XML databases to handle such data. We introduce a new approach called eXtricate that taps on the advantages of both philosophies. We argue that semistructured documents, such as data in an E-catalog, often share a considerable amount of information, and by regarding each document as consisting of a shared framework and a small diff script, we can leverage the strengths of relational and XML databases at the same time to handle such data effectively. We also show that our approach can be seamlessly integrated into the emerging support of native XML data in commercial DBMSs (e.g., IBM's recent DB2 9 release with Native XML Support). Our experiments validate the amount of redundancy in real e-catalog data and show the effectiveness of our method.

The database community is on a constant quest for better integration of data management and knowledge management. Recently, with the increasing use of ontology in various applications, the quest has become more concrete and urgent. However, manipulating knowledge along with relational data in DBMSs is not a trivial undertaking. In this paper, we introduce a novel, unified framework for managing data and domain knowledge. We provide the user with a virtual view that unifies the data, the domain knowledge and the knowledge inferable from the data using the domain knowledge. Because the virtual view is in the relational

format, users can query the data and the knowledge in a seamlessly integrated manner. To facilitate knowledge representation and inferencing within the database engine, our approach leverages native XML support in hybrid relational-XML DBMSs. We provide a query rewriting mechanism to bridge the difference between logical and physical data modeling, so that queries on the virtual view can be automatically transformed to components that execute on the hybrid relational-XML engine in a way that is transparent to the user.