Toward Automated Large-Scale Information Integration and Discovery

Paul Brown, Peter Haas, Jussi Myllymaki,
Hamid Pirahesh, Berthold Reinwald, Yannis Sismanis

IBM Almaden Research Center
{pbrown1, phaas, jussi, pirahesh, reinwald, syannis}@us.ibm.com

Abstract. The high cost of data consolidation is the key market inhibitor to the adoption of traditional information integration and data warehousing solutions. In this paper, we outline a next-generation integrated database management system that takes traditional information integration, content management, and data warehouse techniques to the next level: the system will be able to integrate a very large number of information sources and automatically construct a global business view in terms of "Universal Business Objects". We describe techniques for discovering, unifying, and aggregating data from a large number of disparate data sources. Enabling technologies for our solution are XML, web services, caching, messaging, and portals for real-time dashboarding and reporting.

1 Introduction

Efficient business management requires a flexible, unified view of all data in the enterprise. Many mid-sized and large companies are now facing the challenging problem of bringing together all of their data in a unified form. Enterprise business data is typically spread across hundreds or thousands of sources: inside applications such as SAP, Siebel, and PeopleSoft, on web sites, as syndicated data feeds, in content-management warehouses and marts, in email archives, in spreadsheets and other office documents, and inside a tremendous variety of custom applications. Data collected over a long period of time is juxtaposed with ready-made data sources that come as part of turnkey systems.

Disparate data sources are brought together through business acquisitions and mergers. The various operational systems within an enterprise are usually isolated, have data that is inconsistent both in format and semantics with other operational systems, and offer different end-user tools for data retrieval and reporting. There is a marked lack of data integrity; for example, a given entity can have different names in different operational systems, distinct entities can have the same name, and an entity may exist in some systems but not in others. Adding to the problem, most traditional databases were

T. Härder and W. Lehner (Eds.): Data Management (Wedekind Festschrift), LNCS 3551, pp. 161-180, 2005.
© Springer-Verlag Berlin Heidelberg 2005

not designed with future data integration in mind. These databases, which place a premium on data integrity and consistency, are usually encased in a very large body of difficult-to-change procedural code that is inextricably entwined with the schema semantics.

The major costs of an integration project are usually incurred while trying to "understand the data," i.e., trying to figure out the schema of each input data source, the constraint rules that govern the data in each schema, and, perhaps most importantly, the rules that describe the relationships between data from different sources. These assertions are borne out in the traditional setting for data integration, namely data warehousing. Traditional warehouses try to provide a unified view of the data while hiding the diversity of the underlying operational systems. Experience with data-warehousing projects suggests that the capital cost of integration is very high, and an average design time of six months to two years causes many projects to fail. Manual schema integration constitutes the bulk of the expense. At least one vendor [11] estimates that labor costs for large warehousing projects currently comprise 70% of the total costs, and we believe that the relative labor costs for traditional warehousing solutions will only increase in the future. The initial per-document cost of enterprise content management systems is less than that of data warehouses, but the deep document-analytic functionality provided by content management systems is frequently inferior to the BI capabilities of data warehouses.

In this paper, we outline a next-generation integrated database management system (DBMS) that brings together information integration, content management, and data warehousing with business intelligence. The integrated DBMS (1) taps into the data flow between operational systems and captures raw data, (2) automatically constructs a global business view in terms of Universal Business Objects (UBOs), (3) provides warehousing functionality for the UBOs such as cube queries, advanced analytics, and efficient propagation of data changes, and (4) provides rich interfaces for browsing and searching UBOs that hide the details and variety of the underlying operational systems. We take a data-centric integration approach in that we capture raw data only, and do not rely on any schema information. To ensure that the system scales to very large numbers of information sources, we employ asynchronous messaging and crawling techniques similar to those found in web-scale applications such as Google. We apply machine-learning techniques to map information across disparate sources, and develop new algorithms to map information sources into higher-level business artifacts; these artifacts can either be discovered by the system or introduced into the system through enterprise data dictionaries, taxonomies, and ontologies. We introduce a new query paradigm that takes queries over UBOs and converts them into queries over specific information sources based on discovered relationships between their sources. The enabling core technologies for this next-generation integrated DBMS are XML, web services, tools for information dissemination and dashboards, frameworks for unstructured information management (such as IBM's UIMA architecture) and advanced search capabilities.