Virtual Database Technology, XML, and the Evolution of the Web

STS Prasad and Anand Rajaraman*

Junglee Corporation
1250 Oakmead Parkway, Suite 310
Sunnyvale, CA 94086-4027
http://www.junglee.com

Abstract

We describe Junglee's Virtual Database (VDB) technology, which makes the World Wide Web and other external data sources behave as an extension of an enterprise's relational database (RDBMS) system. We provide examples of powerful applications enabled by the technology. We then consider XML, a new markup language standard; we conjecture how XML will transform the web, and the role that will be played by Virtual Database technology in this transformation.

1 Introduction

Virtual database (VDB) technology makes the World Wide Web and other external data sources behave as an extension of an enterprise's relational database (RDBMS) system. According to some estimates, as much as 90% of the world's data is outside of relational database systems. Vital data is scattered across web sites, file systems, database systems, and legacy applications. These data sources differ in the way they organize the data, in the vocabulary they use, and in their data-access mechanisms. Many of them do not even support native query operations. Writing applications that combine data from these sources is a complex, often impossible, task because of the heterogeneity involved.

Junglee's patent-pending VDB technology can fundamentally transform enterprise computing and the World-Wide Web by providing a solution to this data scatter problem. VDB technology lets applications ask powerful queries of data that is scattered over a variety of data sources. The VDB gathers, structures and integrates the data from these disparate data sources and provides the application programmer with the appearance of a single, unified relational database system. VDB technology enables the development of an exciting new breed of applications that use all the data.

As an illustration of the applications enabled by VDB technology, consider job hunting on the Web. In order to make a meaningful career choice, a job seeker needs information on available opportunities as well as related data — such as information on housing, school districts, and crime statistics in the job area. Information on job openings is scattered across thousands of different web sites — company home pages and several aggregate sites,

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*Phone (408) 522-9494; Fax (408) 522-9470. For more information contact Anand Rajaraman at anand@junglee.com

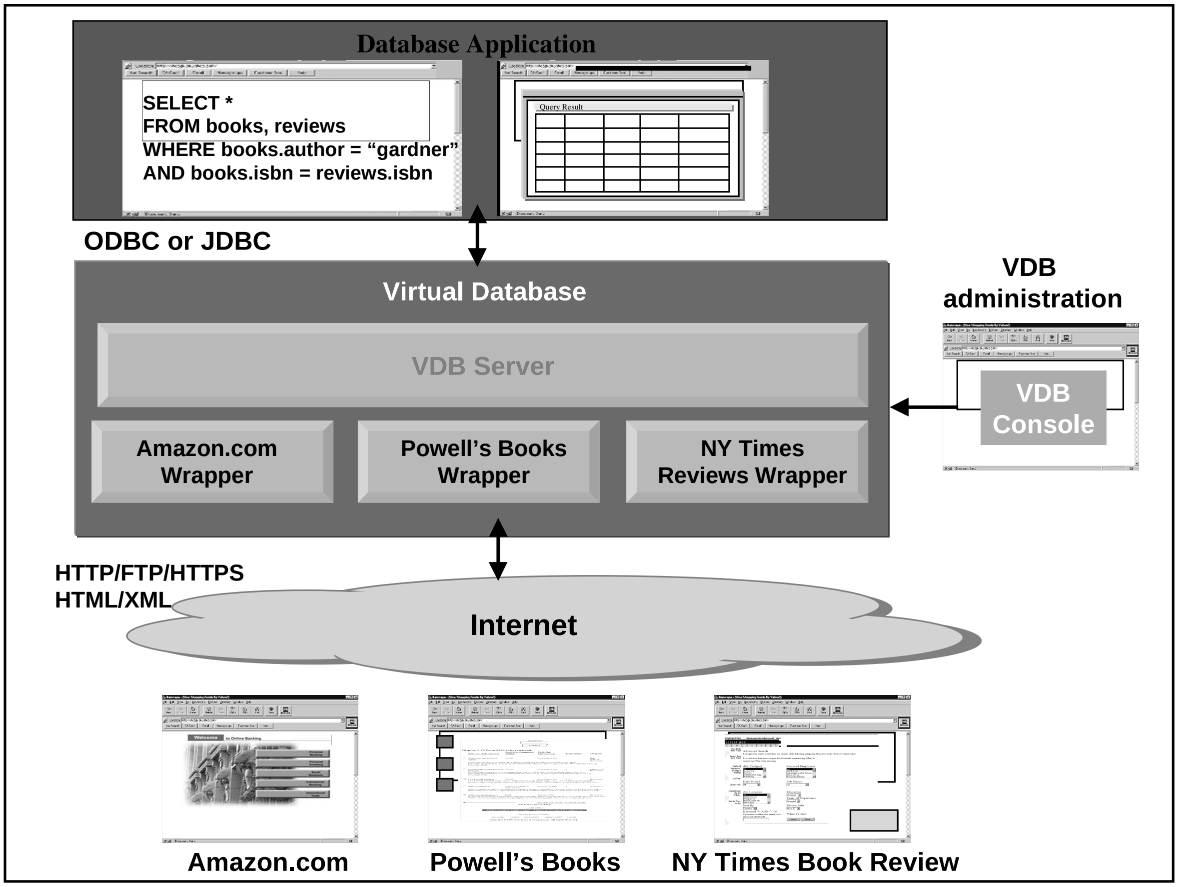

Figure 1: The Books Virtual Database

such as newspaper classifieds sites. Keyword search capabilities on words appearing in the job listing are the only available search choice.

VDB technology converts all these data sources into a single virtual relational database. Using an application based on VDB technology, the job seeker can now obtain answers to the following query posed to the Web, "find marketing manager positions in a company that is within 15 miles of San Francisco and whose stock price has been growing at a rate of at least 25% per year over the last three years." This single query would span the Web employment listings of many corporations, in addition to web sites that have geographical mapping information and websites that contain historical records of corporate equity prices. The query would also return, for each position, related information including statistics on housing prices, school districts, and crime statistics. Section 3 provides details on this and other VDB applications that are deployed on several high-traffic web sites, including those of Yahoo!, The Wall Street Journal, The Washington Post, and San Jose Mercury News.

2 Technology Architecture

Figure 1 is a run-time view of a simple Virtual Database (VDB), which we'll call the Books VDB for future reference. This VDB integrates the contents of two bookstores (Amazon.com and Powell's Books) and the New York Times Book Reviews and presents a unified schema with two tables, books and reviews. The database application operates on this unified schema, issuing SQL queries through the JDBC or ODBC API; the application itelf can be built using standard RAD tools such as Delphi, PowerBuilder, Visual Basic, or similar Java toolkits.

The VDB is accessed through the VDB Server, and is administered through the browser-based VDB Console. The VDB also contains, for each external data source, a wrapper that interfaces the data source to the VDB server. A wrapper makes an arbitrary external data source, such as a web site, behave like an RDBMS, while the VDB Server integrates these separate relational databases into a unified Virtual Database (VDB).

A wrapper interfaces with a web site, typically using HTTP and HTML or XML. It handles HTTP protocol-related issues such as forms, cookies, and authentication. The wrapper is accessed via the JDBC API, through which clients can issue SQL queries. A SQL query issued to the wrapper might result in the wrapper filling out a HTML form on the Amazon.com web site, navigating and parsing the resulting HTML pages, and transforming the data into rows in a relational table. The wrapper uses extraction rules to apply sophisticated linguistic processing to extract attributes from the web pages, uses data transformation rules to transform and format the data to fit the schema, and uses the data validation rules to ensure data integrity.

Lightweight Java applications that interact with one (or a few) data sources can interface directly with wrappers. The application sees each data source as a separate JDBC source with its own schema, and must connect to each source separately and combine the data as needed.

Sophisticated applications that use more than a few data sources use the full functionality of the VDBMS, as shown in Figure 1. The VDBMS exposes tables in multiple data sources as virtual tables in a single Virtual Database (VDB), and supports full RDBMS functionality over virtual tables including view definitions and query processing across sources. In the example of Figure 3, the VDB defines the view books as the union of the amazon and powell's virtual tables. When the VDBMS receives the query shown in the figure, the query processor component decomposes the query, determines the fragments to be sent down to the individual data sources, and combines their results. The query result cache caches results from data sources for performance. In addition, the publishing system can be set up to periodically create physical snapshots of virtual tables in a local relational data store in order to speed up data access.

3 The VDB At Work: Real-World Applications

Junglee has applied VDB technology in several key domains: Employment Classifieds, Consumer Shopping, Real Estate, and Apartment listings. We describe below two such applications.

3.1 Online Recruitment

The JobCanopy VDB application integrates job listings from over 700 data sources, including employer web sites, flat files, and legacy data feeds. The schema for this VDB includes 31 attributes of interest to employers and jobseekers, including job title, job category, job location, and contact information. These data sources are scoured each week to ensure that the information is always fresh. Listings from different employers are normalized to have the same set of fields and the same vocabulary. The JobCanopy product is accessible from the web sites of several major newspapers and online media companies, including The Wall Street Journal Interactive Edition, The Washington Post, The San Jose Mercury News, Classifieds2000, and Westech Virtual Job Fair.

3.2 Web Commerce

The ShopCanopy VDB application allows comparison shopping over 40 merchants in 8 categories, including Books, Music, Computer Hardware, and Consumer Electronics. ShopCanopy is deployed on the Yahoo! Visa Shopping Guide web site at http://shopguide.yahoo.com.

The ShopCanopy application brings together buyers and sellers online to create marketplaces on the Web. ShopCanopy allows consumers to easily access and compare product and pricing information from merchants simultaneously, and then link to a specific merchant's site to make a purchase. VDB technology reduces the time spent looking for specific items by searching through affiliated online merchants and compiling a single list of all the vendors that offer the specified item, plus availability, shipping, pricing and other information helpful for making product choices.

4 The Three Phases of XML

The ever-increasing reach of the Web is based to a large extent on the simplicity of HTML, which enables authors to distribute documents at low cost and with ease. Most documents on the Web today are stored and transmitted in HTML. HTML is adequate for handling the presentation aspects of small and simple documents. Going beyond presentation of data and documents, HTML is used today to represent the form-based query capabilities and transactional functionality of Web sites. For example, a bookstore would allow shoppers to search for books by author or title. This query capability is presented through an HTML form. A transaction, e.g., the ability to buy a book and order its shipment, is also presented through HTML.

Virtual database technology transforms the Web into a database, using adapters (called *wrappers*) that analyze the HTML from Web sites to present them as relational data sources. The limited descriptive capability of HTML necessitates manual analysis for the creation of adapters. XML will add the next level of automation. As data sources become self-describing, VDB technology will automatically create the adapters. Virtual databases of several hundred thousand data sources will occur. In fact, the entire Web could become one unified database, fulfilling Junglees vision. With the increasing complexity of documents and their inter-relationships, the Web is evolving from a collection of hyperlinked documents to dynamic content that is generated from relational databases, document libraries and other forms of organized content. The limitations of HTML — stemming from the fact that structure, content and presentation are intermingled in HTML — is becoming increasingly evident to Web developers. The need for extensibility, structure and validation is the basis for the evolution of the Web towards XML. Both HTML and XML are derived from SGML (Standard Generalized Markup Language, ISO 8879). SGML allows documents to be self-describing, through the specification of tag sets and the structural relationships between the tags. This specification is referred to as the Document Type Definition, or DTD. HTML is a small hard-wired set of about 70 tags and 50 attributes, which allows HTML users to skip the self-describing aspect from a document. XML, on the other hand, retains the key SGML advantage of self-description through DTDs, while avoiding the complexity of full-blown SGML

The XML specification was adopted as a standard by the W3C in February 1998. Since then, there has been a groundswell of support for XML from the development community. We believe that XML will become the dominant data interchange format on the web, and that this transition will happen in three phases outlined below.

4.1 Phase 1: Data in XML

The delivery of XML content to browsers from Web sites enables the distribution of a significant proportion of the processing load from the Web server to the Web client. Sorting, grouping and pagination can be performed locally, avoiding round-trips to the Web server. The highly structured delivery of data enables clients to present different views of the same data to different users through style sheets.

Phase 1 represents a gigantic leap forward towards structured data interchange between Web servers and browsers, and potentially between Web servers themselves.

4.2 Phase 2: Data and queries in XML

The growth in the number of query-based Web data sources presents a challenge and a business opportunity for Web search engines, Web portals and Web content aggregators that will push XML to the next level. In Phase 2, Web sites will describe their query capabilities through XML to facilitate integration of individual Web sites into structured search engines and content syndicates. For example, a bookstore's web site may state that it allows searches by author or by title, but not a search that will enumerate all the books in the store.

A key benefit from describing query capabilities in XML is the ability to present the query based on a style sheet, enabling custom search forms for individual users. In addition, it would enable Web clients to mediate between small collections of heterogeneous Web data sources.

In Phase 2, VDB technology will be the enabling technology for the next generation of search engines. This next generation will go beyond mere keyword searches, allowing domain-specific attribute-based searches. For example, a corporate procurement search engine would allow procurement analysts to search the entire collection of supplier catalogs on the Web by supplier location, product category, category-specific attributes and price range. Cross-domain searches will present new business opportunities. VDB technology will normalize the query capabilities across Web sites, presenting uniform query capabilities across vast collections of autonomous Web sites conceivably the entire Web.

4.3 Phase 3: Data, queries and transactions in XML

Phase 3 will accelerate the pace of virtualization of organizations, where each organization focuses on its core competencies, and increasingly use Web-based or extranet-based outsourcing of ancillary services. However, to the consumer, there is an illusion of a complete service-provider. There are signs of this trend even today in financial services, health care, retail, business-to-business suppliers and electronic marketplaces. VDB technology will form the basis for the virtual organization, leveraging its ability to pull together disparate transaction capabilities into a common model. Web-based transaction integration follows naturally from the Web data integration capabilities of the preceding phases. In other words, the VDB search engines of Phase 2 will naturally gravitate to full-service transaction facilities, to cement their position in the value chain to the consumer.

5 Conclusion

VDB technology enables rapid deployment of applications with at least one of the following characteristics:

- Large numbers of data sources
- Data sources are autonomous, there is no centralized control
- Data sources can have a mixture of structured and unstructured data

The World Wide Web, and most Intranets, have all of these characteristics. The emergence of XML and related standards, such as RDF, will accelerate the deployment of VDB technology since they have the potential to lower drastically the cost of incorporating a web site into a virtual database.

Acknowledgements

We thank Ashish Gupta for helpful discussions, and the entire Junglee team who have implemented everything described in this paper.