The INtersect Concept for Multidatabase System Integration in the Pharmaceutical Industry

Tony Schaller
Molecular Design Ltd.
2132 Farallon Drive
San Leandro, CA 94577
tonys@molecular.com

The industry trends facing information systems in the 1990's involve a matrix of complex requirements. The migration from centralized mainframe computing to desktop personal computing has created opportunities and challenges in moving access to and control of information closer to the end-user. Within this new wave of computing, there has been a drive to move the access control for information to the desktop of the user. Although, graphical user standards were introduced in order to ease of use and provide a consistent look and feel to desktop applications. This provided some assistance in shielding the complexity of the various systems, however it did not address the difficulties presented in accessing heterogeneous databae systems on various platforms.

These same trends are driving specific changes in the scientific computing environment. Terminals linked to mainframe computer systems are giving way to dedicated desktop workstations or personal computers. Research scientests are using more of the various desktop productivity tools in order to complete the necessary preparation of documents, statistical modeling, etc. This need for cross-application compatibility resulted in the objective to share text and chemical structure representation.

In addition to the increased use of diverse desktop applications, evolutionary changes were taking place with respect to the database management systems that were in use. Traditionally, scientests accessed chemical structure information for searching particular representations of molecules or reactions via mainframe based systems. These chemical database applications utilized multi-dimensional search algorithms to streamline the search of chemical structures and reactions and quickly provide hit lists of candidates to the user. In addition, more and more data were being collected in relational systems for repetitive sets of information (eg. test results from laboratory screening). Other forms of data as well was being stored in other systems; text databases used for composing research documentation and new drug submission proposals; image databases containing results from spectral analysis.

Effective integration of these diverse data sources was seen as necessary for the comprehension of all of this data, especially when considering options when synthesizing new drug compounds, or presentation to other agencies (eg. FDA submission).

To respond to these dramatic changes in the computing environment, Molecular Design Limited (MDL) launched a project in 1987 to build a distributed heterogeneous multidatabase system to access chemical, relational, text, object databases in a

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
SIGMOD /5/93/Washington, DC,USA
© 1993 ACM 0-89791-592-5/93/0005/0502...$1.50

manner that would shield the complexity of the underlying dbms from the end user. The resulting work was a client-server system that shipped in September 1990.

The client consists of a local chemical database combined with a graphical user interface to enable access to the data in a consistent manner with the ease-of use that is characteristic of a "point and click" interface. The server consists of a multidatabase technology, known as INtersect, hat provides complex data integration from heterogeneous database sources transparent of location of system.

INtersect's architecture is an object-oriented structure that comprises multiple internal layers (communications, integration, and gateway interfaces). It operates as a loosely-coupled system that utilizes an external schema known as an Hview (Heterogeneous View) to define the virtual database for the user, application or project. Links are established to the underlying database systems dynamically when the actual view is opened. Hviews may utilize alternate access routes or inherit other hview definitions that exist previously.

The "nested relational" model is used to support heterogeneous data models. It integrates relational, hierarchical, flat, and object-oriented systems while maintaining and using the structure of the varying data models. It does not require a single model to be imposed on all of the database sources (eg. relational). Cursor objects are provided to the client to permit explicit manipulation and browsing of record from the underlying sources if so desired.

The gateway interface to INtersect supports an object paradigm as new data types may be added, along with their own supporting methods for object creation, deletion and modification. Query operators may be registered and supported on a gateway specific basis allowing the support of heterogeneous access languages (eg. SQL, Text, etc.). Navigation of data within the gateway can be extended to be sensitive to the semantics of the data.

Although the conventional method for merging data from multiple sources is via a "merging" or join step, INtersect integrates in a pipelined fashion. Results from each data source search are coupled with other search predicates and used in combination to drive query against the next data source. This provides for an optimized response time model in which incremental results can be passed to the client as soon as they are available.

The paper that will be presented illustrates the foundations that called for the product. An architectural description of INtersect will be detailed along with its ability to support access and data model heterogeneity. As INtersect is currently used within the ISIS product at over 70 pharmaceutical customer sites today, the application of the technology will be reviewed to discuss its implementation in the discovery and chemical industry.