The IU-IBM DiscoveryLink Project Final Report

Executive Summary

This highly fruitful collaborative effort between IU and IBM life sciences/DB2 teams has resulted in the institution at IU of a Centralized Life Science Data (CLSD) service. CLSD provides to IU researchers a unified, SQL interface to diverse, public life sciences data.

The CLSD uses IBM's DiscoveryLink/Information Integrator and DB2 products on IU's IBM Regatta SP node to federate data from BLAST searches with SQL searches against LocusLink, BIND, Enzyme, UniGene, ePCR, Nucleotide, Homologene, PubMed, SGD, and dbSNP public life sciences databases.

During the course of the project:

- An IBM life sciences team visited IU
- IU IT personnel attended DB2 and DiscoveryLink training at IBM sites in San Jose and Dallas
- Software was shared between IU and IBM

A number of presentations about the CLSD were made at IU and elsewhere, including a joint IU-IBM press release in July 2003. A publicly accessible web site for DiscoveryLink resources (such as parsers, scripts, etc.), called the "IBM Data Federation User Group," was also created at IU.

IU engaged with IBM in beta testing version 8 of DB2 and Information Integrator (DiscoveryLink's successor) products. Following the test, a production instance of CLSD employing DB2 v8 and Information Integrator was created on the Regatta node.

IU also acted as a reference site for IBM and showcased CLSD for a number of interested academic institutions.

Future projects currently being considered include the use of the SRS wrapper with Information Integrator and the possibility of making Information Integrator/DB2 grid-aware.

Phase 1

- IBM life sciences team visited IU from June 10-12, 2002 and met with the IU core IT team and the Center for Medical Genomics (CMG; Howard Edenberg's lab) team.
- During the visit, a test instance of DB2 along with the DiscoveryLink components relational and life sciences connect were installed on the IU IBM SP. The life sciences connect component was tested.
- The IU core IT team gained considerable insight and understanding of DiscoveryLink by attending an IBM-delivered DiscoveryLink class in California in May 2002 and by direct contact with the IBM life sciences team during their visit and via weekly teleconferences.

- The precise goals of the CMG continue to evolve during phase 1, making a production query involving DiscoveryLink difficult within the 12-week period of phase 1. Nevertheless, the utility of DiscoveryLink to include data accessible via BLAST searches in a laboratory production pipeline was demonstrated to the CMG, which plans to use DiscoveryLink to implement a system for annotating data collected at its gene expression microarray facilities.

Phase 2

- The first goal in phase 2 was to unite far-flung information about individual genes into a single queriable source for the CMG. Data to be united include: variants of the genes, locations on chromosomes, reactions catalyzed, metabolic pathways in which gene products operate, etc. To do this, DiscoveryLink was used to federate data not only from BLAST searches, LocusLink, BIND, and Enzyme (identified in phase 1) but also data from UniGene, ePCR, Nucleotide, Homologene, PubMed, SGD, and dbSNP.
- IBM delivered parsers for LocusLink and UniGene.
- DiscoveryLink was moved from a test to a production instance, giving the CMG an SQL interface to various data sources. The IU core IT team, with the IBM Life Science team's help, installed a production instance of DB2 on the INGEN-funded IBM SP Regatta node at IUPUI. A facility (clsd-update) was developed to automatically transfer various public data sources to local disks at IUPUI, to parse these data, and to load them into DB2. A model was developed for providing user accounts and support for users who need SQL access to the databases. A production instance of BLAST daemon and public sequence databases was installed and maintained on IU's Sun E10k.
- IBM provided DB2 help to make DB2 administration easier for IU. In addition, an IU core IT team member attended DB2 training in Dallas in late October 2002 (administered under the IBM Scholars program) to increase his knowledge of DB2.
- The scope of the project was expanded to include other, non-CMG groups within the School of Medicine. These now include: Medical Genetics (Tatiana Foroud), Psychiatry (Eric Meyer), Informatics (Narayanan Perumal), Computer Science (Kamal Kumar), Medicine (Michael Econs), Informatics (Stuart Young), and Hematology/Oncology (Bob Hicky).
- The IU core IT team worked with the IBM life sciences team and local IU users to publicize work resulting from the use of DiscoveryLink. These efforts include submission of papers to Bio-IT awards, IBM's CASCON 2003 conference, and the ACM Symposium on Applied Computing 2004. A review of Information Integrator and the CLSD has been accepted by Briefings in Bioinformatics and will be published in December 2003.
- The CLSD was demonstrated at the I-Light workshop at IUPUI in December 2002, at the IBM Institute of Innovation event at IUB in September 2003, and is scheduled at the BioSensor conference in Indianapolis in October 2003.

- CLSD overview presentations have been given at numerous occasions at IUB and IUPUI and are available at: http://storage.iu.edu/presentations/
- IU and IBM worked on a joint press release on CLSD which came out on July 15, 2003. Following the press release, CLSD was featured at a number of online life sciences web sites.
- IU created a publicly accessible website for DiscoveryLink resources (such as parsers, scripts, etc.) called the "IBM Data Federation User Group": http://www.indiana.edu/~clsd/forum.html
- IU engaged with IBM in beta testing DB2 V8 and Information Integrator (DiscoveryLink's new name). In the process (which ended in Oct. 2003), the following components were tested and feedback provided to IBM:
    - MSSQL wrapper (for the Microsoft SQL server)
    - DRDA wrapper (for DB2 v7)
    - Entrez wrapper (for the Entrez API)
    - HMMER (for the HMMER program)
    - XML wrapper
    - BLAST wrapper (for the BLAST program)
    - Flat (for flat files)
- A production instance of DB2 v8 and Information Integrator was installed on the Regatta node and the CLSD service upgraded to this new version. As a result, dbSNP data are now being provided via the CLSD using II's relational connect wrapper from a native instance of dbSNP hosted in Microsoft SQL server by the School of Medicine's Information Systems and Technology Management (ISTM) group.
- Also, the DRDA wrapper is currently being used to provide access to dbSNP release 114 which is being maintained in the DB2 v7 instance of CLSD.
- Upon IBM's request, IU acted as a DiscoveryLink reference site for Princeton Univ., the Univ. of Arizona, and McGill University.

Beyond Phase 2

Currently active projects include:

- Using the SRS wrapper to provide queries to the SRS version 7 installation at the Center for Genomics and Bioinformatics at IUB. Since the SRS wrapper runs under DB2 v7, this would also require the DRDA wrapper.
- Making Information Integrator/DB2 grid-aware. The idea is to link multiple DB2/II instances running at geographically different sites together somehow and allow grid users to query disparate life sciences databases via a single SQL query. An exploratory teleconference occurred between IU IT core team and IBM's DB2/life sciences teams in June 2003.