Open DMIX - Data Integration and Exploration Services for Data Grids, Data Web and Knowledge Grid Applications

Robert L. Grossman*, Yunhong Gu,
Dave Hanley, Xinwei Hong and Gokulnath Rao
Laboratory for Advanced Computing
University of Illinois at Chicago

September, 2003

This is a draft of a paper from the Proceedings of the IEEE 2003 Workshop on Knowledge Grid and Grid Intelligence (KGGI 03), October 13, 2003, Halifax, Canada.

1 Introduction

The analysis and mining of remote and distributed data is critical for many applications being deployed on grid-based and web-based computing platforms. Broadly speaking, the data mining community divides data mining into three broad phases:

1. Data preprocessing — the extraction, transformation, normalization, and aggregation of data from disparate sources to produce a learning set.

2. Data mining — the application of data mining and statistical algorithms to learning sets.

3. Deployment — the deployment of data mining in operational and decision support systems.

Although the bulk of the research focus has been on the second phase, most data mining projects spend the majority of the time and effort on the first and third phases. This will certainly continue to be the case as data mining is applied to distributed data using grid based and web based computing platforms.

It is convenient to think of the output of the data mining phase as a statistical or data mining model, which can be expressed using the XML Predictive Model Markup Language or PMML [11]. The PMML model can then be used in the deployment phase. From this perspective, the PMML model captures an important part of what is sometimes called the "grid intelligence" of a distributed data grid, data web, or knowledge grid application.

In this paper, we introduce data integration and data exploration services designed to facilitate the Phase 1 preprocessing of data. We de-

*Robert L. Grossman is also with Open Data Partners

fine data integration to be the seamless integration of data from disparate sources as a foundation for its exploration, analysis and mining [8]. We define data exploration to be the process of using statistical tools to investigate data sets in order to understand their important characteristics [19]. These data integration and data exploration services are designed to operate in a layered environment of OGSA compliant web services and to interoperate with OGSA compliant transport, access and data mining services.

We have implemented a preliminary version of these services in an open source project called Open DMIX. DMIX is an abbreviation for data mining, integration, and exploration services.

Version 2 of PMML supports the definition of data transformations, normalizations, and aggregations. Other data integration and data exploration operations can be defined using PMML extensions [14]. In this way an XML document based upon PMML and PMML extensions can be used to encode the parts of grid intelligence related to data exploration and integration, complementing the PMML information related to the data mining and deployment phases of the data mining process.

We believe that the Open DMIX services and applications we describe in this paper are novel from the following perspective:

First, data integration has not generally been singled out and supported explicitly by distributed data mining systems and services. Instead, either data integration has been done as part of a general grid computation or, more usually, data integration has been tacitly assumed to have been performed prior to the analysis of the data by the distributed data mining system.

Second, the majority of prior work on wide area data services has been based either on files, such as GridFTP, or on SQL, such as ODBC, JDBC, and OGSA DAIS [15]. In contrast, the Open DMIX services described in this paper are based upon certain predefined data mining operations, for example, those defined by the Predictive Model Markup Language (PMML) [11]. These types of data mining operations are intermediate between SQL selects on remote databases and arbitrary computations on files supported by data grids.

Section 2 contains background and related work. Section 3 provides an overview and taxonomy of data services for distributed grid, web, and knowledge based computing platforms. Section 4 describes the Open DMIX architecture. Section 5 contains some experimental studies. Section 6 is the summary and conclusion.

2 Background and Related Work

To date, there have been a number of different approaches used to build data intensive, distributed applications. One way to make sense of these different technologies is to place them along two axes. The horizontal axis is used to differentiate the action taken with the data, such as viewing it, mining it, or computing with it. The vertical axis is used to differentiate the object of the action, which may be a file, data attributes, or higher order concepts such as the ontologies and related concepts underlying a knowledge management system. This viewpoint is summarized in the Table 1.

Four of these approaches, which are summarized below, are closely connected to the work described in this paper.

Data Grids. Data grid services combine authentication, authorization and access (AAA)

	View	**Discover/Mine**	**Compute**
Knowledge	Digital Libraries	Knowledge Grids	Semantic Webs
Data Attributes	Web Databases	Data Webs	Data Grids
Files	Persistent Archives	Dist. Data Mining	Grids

Table 1: A number of different technologies have been used to build applications for working with remote and distributed data. Data webs provide direct access to data attributes. Data grids enable large scale resource sharing of computational and data resources. Semantic webs provide knowledge based access to data using ontologies, RDF and agent based architectures.

controls with resource managers so that arbitrary computations can be done using distributed computational and data resources belonging to a virtual organization [7]. Good examples of data grids are the data grids developed by physicists [17] or astronomers [10] to process the data collected by the collaboration's instruments. Data in data grids is stored in files and transported using GridFTP [3].

SOAP/XML-based Web Services. Web services based upon SOAP and XML are a rapidly maturing infrastructure for accessing XML-based data [18]. SOAP enables the serialization of XML-data so that it may be transported using TCP or HTTP. SOAP-based services can be described using the Web Services Description Language or WSDL, while the Universal Description, Discovery and Integration (UDDI) provides a simple mechanism for the discovery of web services. SOAP/XML based web services are designed to deal with general XML based data.

Semantic Web. The Semantic Web extends the web's HTML infrastructure to include semantic information defined by XML and the Resource Description Framework (RDF) [18]. RDF views information as a directed labeled graph and serializes it in XML. Less formally, RDF codes information using subject-verb-object triples. As a very simple hypothetical example, the triple (www.ncar.ucar.edu/ccm/1/1, Temperature, 45.5) is a subject-verb-object triple giving the Temperature for a particular data record specified by the URL. RDF can be used to encode much more complicated assertions about data, metadata and relationships defined from them. The semantic web also supports ontologies so that data taxonomies can be used, which is very important for many data analysis and data mining applications.

Data Web. By a *data web* we mean a web based infrastructure for accessing, analyzing and mining remote and distributed data [11]. Data webs can be implemented using general web services and protocols such as XML and SOAP or more specialized protocols and services [13]. It is important to note that data webs are data and its attributes and not with semantic information associated with data. Conceptually, data webs provide much less functionality than semantic webs. Data webs have direct support for just those few concepts required in order to work with remote and distributed data.

Remote Access to Databases. SQL is the standard query language for databases and ODBC and JDBC are widely deployed protocols

for accessing remote data resident in databases. In contrast to data grids, the ODBC and JDBC protocols support the full power of SQL and enable attribute level access to records, as well as a rich range of different types of SQL-based selections.

Data grids are file based infrastructures which support general computation; in particular, they do not provide attribute based access to remote data nor do they provide specific functionality directly related to data and its attributes. Databases were not designed to scale in the fashion of data grids nor to work with remote and distributed data. Finally, web services today do not have the scalability of data grid applications nor do they provide specific data related functionality, although they do provide (using XPath for example) access to attributes.

There are also efforts to combine these approaches. The Open Grid Services Architecture is a standards effort which integrates grids with web services [16]. The OGSA Database Access and Integration Services (OGSA DAIS) [15] combines grid services with databases.

OGSA DAIS is a recent effort that like Open DMIX uses web services to manage some of the metadata and bindings required for remote data access. Both Open DMIX and DAIS support multiple protocols and access to remote databases. DAIS is based upon a grid model and requires an underlying AAA infrastructure such as Globus, while Open DMIX is based upon a web model. In addition, Open DMIX supports templated data analysis and data mining operations, unlike DAIS which is limited to data access and integration operations. Finally, DAIS does not provide a simple mechanism to support distributed joins. Open DMIX Services do: these are called universal correlation keys and are described below.

There are a variety of current efforts to develop data middleware for grids [8]. Perhaps the most relevant are the Storage Resource Broker [2]. The storage resource broker provides transparent access to distributed files based upon metadata, but does not provide the attribute level access to data that is one of the goals of Open DMIX Services. Both Open DMIX Services and the SRB manage relational data using underlying relational databases.

Another related project is Chimera [9], which is much more ambitious than Open DMIX Services. Chimera supports arbitrary data transformations and integration services using the general compuational power of grids. In contrast, Open DMIX supports the templated, predefined transformations and data mining operations of PMML together with simple distributed joins enabled by the universal keys supported by data webs [13] and [11].

Finally, there are a variety of domain specific efforts, such as the DODS, which is developing data access methods for the oceangraphic community [5].

3 A Taxonomy of Data Services

Tables 2, 3 4 summarize some of the data services being used in applications for data grids, data webs and knowledge grids. There are quite a few different data services and there are several different ways of dividing them up.

One basic grouping includes: network services to move bits; data access services to move, create, update and query data records; data exploration and integration services to explore data and to create learning sets for further study; data analysis and data mining services to further explore learning sets; data deployment services to

score data records and process data; and discovery services to discover relevant data and services. See Tables 2 and 3.

Another group of services, such as data replication services, security services, etc. play a supporting role. See Table 4.

Finally, services may be divided into off-line or non-real time services and on-line or real time services. Tables 5 and 6 divide the services with distinction in mind.

4 The Architecture of Open DMIX

The Open DMIX Services that we have prototyped are designed around the layered architecture described in Tables 5 and 6. The analysis and deployment stacks are very similar: layer 4 of the analysis stack is concerned with data analysis and data mining, which can be thought of very concretely as a process whose input is a learning set and whose output is a PMML statistical or data mining model. On the other hand, layer 4 of the deployment stack is concerned with scoring and decision support, which can also be thought of very concretely as a process which once given a fixed PMML model, takes data records as inputs and produces scores as outputs. In general these scores are then used as an input to a decision support process.

Also note that layer 3 of the analysis services includes both data integration and data exploration services, while layer 3 of the deployment services includes just data integration services.

Our layer 1 and 2 design of Open DMIX Services are based upon the Data Space Transfer Protocol or DSTP [13]. This is a protocol specifically designed to access and transport data and metadata and has support for keys, attributes, missing values, and selects. Broadly speaking the Open DMIX architecture extends the architecture used in the DSTP servers by adding layer 3 data exploration and integration services and layer 4 data analysis and data mining services.

Security and policy are not part of the architecture per se, but rather added as required for each application. The Open DMIX Services are based upon web services, and any web service compliant security mechanism, such as TLS, can be used. In addition, since Open DMIX is OGSA compliant, OGSA compliant security infrastructures, such as the Globus Security Infrastructure (GSI) can also be used.

There are a wide variety of different data integration, data exploration, and data analysis services one could imagine. Our approach has been to focus on those services which are supported by Predictive Model Markup Language (PMML) [11].

5 Open DMIX Experimental Studies

We have prototyped some of the DMIX services described in the section above in a preliminary release of an open source server. The DMIX server is based in part upon the DataSpace Transfer Protocol (DSTP) server we have previously developed [13] and [6]. As mentioned above, the Open DMIX server also includes the layer 3 data integration and exploration functionality and the layer 4 data analysis and data mining functionality, while the DSTP server does not.

The DSTP server was designed to provide efficient data access to remote and distributed data and was developed prior to the emergence of web services [13]. It is based upon a protocol called

Service	Status	Future Directions and Challenges
Data Discovery		
Data discovery	Web Service Description Language (WSDL) and Universal Description, Discovery & Integration (UDDI) available	WSDL and UDDI seem adequate for the near term
Data Deployment Services		
Scoring and processing services	PMML-based scoring services are emerging; data processing can be done using data integration.	Scoring and processing services are generally less resource intensive than data analysis services.
Data Analysis and Data Mining		
Data analysis and data mining	PMML-based services; XML-A based web services; grid services can be used for arbitrary computations.	Scaling data analysis and data exploration to large distributed data sets is an open problem.
Data Integration and Exploration		
See accompanying table below.		
Data Access Services		
Data access	At least three approaches being used: data mining based DSTP; data grid based GridFTP; and grid service based OGSA-DAI. The current efforts may be combined.	Scaling data access to large data sets requires integrating data access and scalable network protocols, an open research problem.

Table 2: A Taxonomy of data services for data grids, data webs, and knowledge grids —- The Data Service Stack. Part 1 of 3.

Service	Status	Future Directions and Challenges
Data Exploration and Integration		
Data exploration	Open DMIX supports simple summary and exploration of data	Scaling data exploration to large remote data sets and distributed data sets is an open problem.
Data transformations	PMML Version 2.1 supports templated transformations; grid services can be used to define arbitrary transformations.	Integrating data transformations with databases, web services, and grid services presents several interesting research problems, as does scaling transformations to large data sets.
Distributed data joins	DSTP supports UCK based joins; semantic webs support ontology based merges.	Understanding how to scale distributed data joins to large distributed data sets is still an open problem.
Data workflow	Several preliminary efforts underway.	Virtual data services are being investigated, such as Chimera; data workflow are planned into future versions of PMML; business process workflow markup languages can also be used.

Table 3: A Taxonomy of data services for data grids, data webs, and knowledge grids —- Data Integration and Exploration Services. Part 2 of 3.

Service	**Status**	**Future Directions and Challenges**
Supporting Services		
Resource integration & management	Grid Resource Allocation Management (GRAM).	The challenge for data intensive applications is to use resource management to provide access not only to computational resources but also to data resources.
Security & policy	Grid Security Infrastructure (GSI).	Always a challenge, especially data privacy issues raised as more data sources become available and data integration services grow more common and powerful.
Data replication	Globus Replica Management	Emerging high bandwidth networks are changing the tradeoffs for replica managmenet.

Table 4: A Taxonomy of data services for data grids, data webs, and knowledge grids —- Supporting Services. Part 3 of 3.

Layer	Functionality	Open DMIX Implementation
5. Discovery services	locate relevant data	WSDL and UDDI used
4. Data analysis and mining services	templated data analysis & data mining services	based upon R Project, custom code
3. Data integration & exploration services	transform, aggregate, normalize & integrate data; summarize and extract basic statistical properties of data;	integration services currently supported; PMML-based transformation services being prototyped; statistical summary, clustering, & basic regression supported;
2. Data access services	access and query data and metadata	suported by the DSTP server
1. Network transport services	transport bits	multiple implementations supported by DSTP servers, including TCP, SABUL, GridFTP, etc.

Table 5: Open DMIX Analysis Services are based upon a simple layered architecture.

Layer	Functionality	Open DMIX Implementation
5. Discovery services	locate relevant services and data feeds	WSDL and UDDI used
4. Scoring services	scoring, decision support, etc.	basic scoring engine available
3. Data integration	transform, aggregate, normalize & integrate data;	integration services currently supported; PMML-based transformation services being prototyped;
2. Data access services	access and query data and metadata	suported by the DSTP server
1. Network transport services	transport bits	multiple implementations supported by DSTP servers, including TCP, SABUL, GridFTP, etc.

Table 6: Open DMIX Deployment Services are based upon a simple layered architecture.

the DataSpace Transfer Protocol or DSTP [13]. DSTP supports attribute-based data, attribute metadata, data set metadata, and distributed keys. DSTP binds globally unique IDs or GUIDS to distributed keys. The assumption in DataSpace is that data attributes with the same distributed key, identified through its GUID, can be meaningfully compared. Although very simple, this idea can be used to develop applications which integrate data from distributed data sets in a meaningful way. DSTP also provides support for specifying sampling, mechanisms for working with missing values, and using different network protocols when transporting data and metadata.

From the layered viewpoint described in the previous section, the DSTP server integrates network transport (Layer 1) and data access services (Layer 2). For the work described in this paper, a W3C and OGSA compliant web service interface was added to the DSTP server. Previously, requests to the DSTP server were sent via the DSTP protocol [13], while results were returned via XML using the DSTP protocol or via specialized streaming formats using the DSTP protocol. Currently, requests to the DSTP server are via SOAP/XML [18], while data may be returned via SOAP/XML or via other mechanisms supported by the DSTP protocol, which for many applications is more efficient. For example, using DSTP data can be streamed back in specialized binary formats over separate data channels using network transports designed for high speed data transfers [4].

Table 1 contains some performance measurements which compares Open DMIX's two modes of data access. The data used for this is a simple ten column dataset of random integers. As the table makes clear, the Open DMIX Server's SOAP/XML mode does not scale linearly with query size, and in fact breaks with sufficiently large queries. The 1 million row soap query consumed 99% of the CPU and much of the RAM on the client and server for its marshalling and demarshalling. We emphasize that Open DMIX supports W3C and OGSA compliant web services, due to the prevalence of these standards, but also supports more specialized data access and network protocols in order to work with large remote and distributed data sets.

In addition, for the work described in this paper, we added data exploration and data integration (Layer 3) services and data analysis and data mining (Layer 4) services to the Open DMIX server. The data integration services being prototyped are implemented as a library which can be used by Open DMIX clients and supports the data transformations, aggregations, and normalizations defined by Version 2.0 of the DMG Predictive Model Markup Language (PMML) [14]. In particular, we implemented the following transformations and aggregations for our Open DMIX library:

1. Aggregation Functions
    - Sum
    - Count
    - Max
    - Min
    - Average
2. Normalization Functions - Normalize
3. Discretization Functions - Discretize
4. Value Mapping

Although certainly a small set, there are sufficient to for a surprising variety of different applications.

In addition, since Open DMIX supports DSTP, Open DMIX clients and servers can also use the distributed keys supported directly by the DSTP protocol for data integration.

Finally, we integrated the data analysis and data mining functions provided by the open source R Project into our Open DMIX Server [1].

As an example, Figure 2 is a snapshot of a distributed DMIX application involving protein data and chemical compound data. Chemical compound data from multiple DMIX severs may be combined using a distributed key we have defined for small chemical compounds [12]. A DMIX client application can, for example, pull protein data from one DMIX server, chemical compound data from several other DMIX servers, cluster the chemical compound data based upon its properties, use the cluster to single one chemical compound of interest, and then dock chemical compound in one or more of the proteins retrieved. To support this type of analysis, the DMIX clients and servers use Layer 3 integration and transformation services and Layer 4 analysis services.

6 Summary and Conclusion

In this paper, we described some of the data services required by data intensive applications for data grids, knowledge girds, and data webs. In particular, we have highlighted the importance of data exploration and data integration services, which form an important part of the data preparation process so important in data mining. We have implemented an integrated stack of data mining, data integration and data exploration or DMIX services in an open source server called open DMIX.

In our experimental studies using the DMIX server, we have observed some scalability problems using web services on large remote and distributed data sets. For this reason, the open DMIX server also employs an alternative protocol based upon the DataSpace Transfer Protocol or DSTP which can be used to return the results of queries.

We have also developed two bioinformatics applications which require the integration of data from different DMIX servers and found the integration mechanism based upon universal keys to be quite sufficient for these types of applications.

References

[1] R project. Retrieved from http://www.r-project.org, January 10, 2003.

[2] Chaitanya Baru, Reagan Moore, Arcot Rajasekar, and Michael Wan. The sdsc storage resource broker. *Proceedings of CASCON '98*, 1998.

[3] A. Chervenak, I. Foster, C. Kesselman, and S. Tuecke. Protocols and services for distributed data-intensive science. In *ACAT2000 Proceedings*, pages 161–163, 2000.

[4] A. Chien, T. Faber, A. Falk, J. Bannister, R. Grossman, and J. Leigh.

[5] DODS. Distributed oceangraphic data system. http://www.unidata.ucar.edu/packages/dods/, retrieved on March 30, 2003.

[6] Source Forge. Project dataspace. http://www.sourceforge.net/projects/dataspace, June, 2003.

number records	SOAP/XML Mode (min:sec)	Streaming Mode (sec)
10,000	0.65	0.21
50,000	2.57	0.72
150,000	11.13	2.05
375,000	51.18	5.01
1,000,000	5:52.10	13.43

Figure 1: Open DMIX servers have two modes for data access. One mode uses SOAP/XML, which works for small data sets, and (small) metadata. The other mode uses a streaming protocol and scales much better for large and complex data sets.

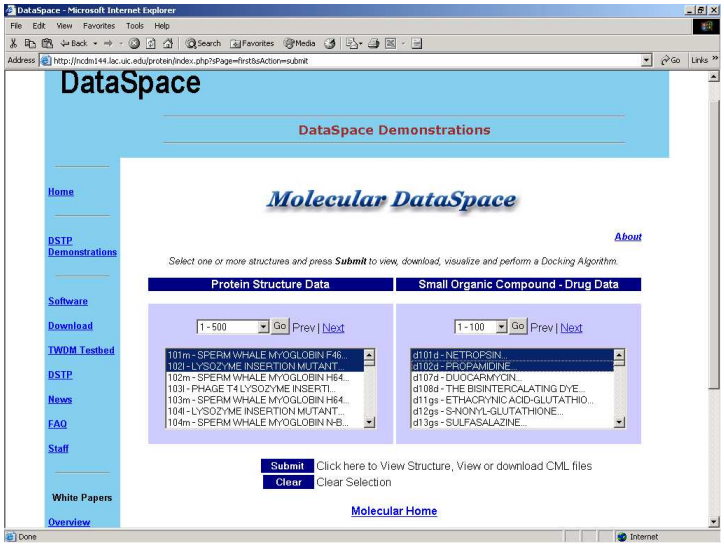

Figure 2: Biowebs.

[7] I. Foster and C. Kesselman. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Francisco, California, 1999.

[8] Ian Foster and Robert L. Grossman. Petascale data integration on terascale networks. *Communications of the ACM*, 2003, to appear.

[9] Ian Foster, Jens Vockler, Michael Wilde, and Yong Zhao. Chimera: A virtual data system for representing, querying, and automating data derivation. In *14th International Conference on Scientific and Statistical Database Management*, 2002.

[10] Jim Gray and Alexander S. Szalay. The world-wide telescope. *Science*, 293:2037–2040, 2001.

[11] R. L. Grossman. Standards and infrastructures for data mining. *Communications of the ACM*, 45(8):45–48, 2002.

[12] Robert Grossman, Donald Hamelberg, Pavan Kasturi, and Bing Liu. Experimental studies of the universal chemical key (uck) algorithm on the nci database of chemical compounds. In *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB 2003)*. IEEE, 2003.

[13] Robert Grossman and Marco Mazzucco. Dataspace - a web infrastructure for the exploratory analysis and mining of data. *IEEE Computing in Science and Engineering*, pages 44–51, July/August, 2002.

[14] Data Mining Group. Predictive model markup language (pmml). http://www.dmg.org, January 10 2002.

[15] Norman W. Paton, Malcolm P Atkinson, Vijay Dialani, Dave Pearson, Tony Storey, and Paul Watson. Database access and integration services on the grid. 2002.

[16] The Globus Project. Towards globus toolkit 3.0: Open grid services architecture. http://www.globus.org/ogsa/, retrieved on January 10, 2003.

[17] The GriPhyN Project. Griphyn - grid physics network. http://www.griphyn.org, retrieved on April 1, 2003.

[18] W3c semantic web. Retrieved from www.w3.org/2001/sw/, September 2, 2002.

[19] J. W. Tukey. *Exploratory Data Analysis*. Addisson Wesley, 1977.