# Applying Grid Technologies to Distributed Data Mining

**A. C. Hume**[1], A. D.Lloyd[2,3], T. M. Sloan[1], A. C. Carter[1]

[1]EPCC, The University of Edinburgh, James Clerk Maxwell Building, Mayfield Road, Edinburgh, EH9 3JZ, UK

[2]The University of Edinburgh Management School, The University of Edinburgh, 7 Bristo Square, Edinburgh, EH8 9AL, UK

[3]Curtin Business School, Curtin University of Technology, GPO Box U1987, Perth WA 6845, Australia

## Abstract

The Grid promises improvements in the effectiveness with which global businesses are managed, if it enables distributed expertise to be efficiently applied to the analysis of distributed data. We report an ESRC-funded collaboration between EPCC in Edinburgh and Curtin University of Technology in Perth, Australia, that is applying public-domain Grid technologies to secure data mining within a commercial environment. We describe this Grid infrastructure and discuss its strengths and weaknesses.

## 1. Introduction

Data mining projects often require distributed analysts to submit jobs to distributed compute resources that process data from distributed data resources. These requirements, along with others such as secure communications and access control, make data mining an ideal application of Grid technologies.

The INWA project [1] has investigated the suitability of existing grid technologies for secure commercial data mining. This project has been funded under the Pilot Projects in E-Social Science programme [2] of the UK's Economic and Social Research Council (ESRC). The full title of the project is 'Informing Business and Regional Policy: Grid-enabled fusion of global data and local knowledge' but for ease of communication this has been abbreviated to INWA.

The project is a collaboration between various academic and commercial organisations from the UK and Australia. EPCC [3], the University of Edinburgh Management School (UEMS) [4] and Lancaster University Management School (LUMS) [5] are the academic partners in the UK with Curtin Business School [6] from the Curtin University of Technology the academic partner in Perth, Australia. The various commercial partners are from the UK and Australia

Financial, telecommunications and property data have been provided by the commercial partners. These partners have also helped formulate requirements for mining of this data. The major requirement being to ensure that any data supplied can only be accessed by trusted parties. The data from UK partners is sited at EPCC with the Australian data sited at Curtin. Sun Microsystems in Australia provided the project with the compute servers for the data located at Curtin.

Such a collaboration between multiple data services in multiple jurisdictions tests acceptance of the grid – a pre-requisite for anyone to adopt this technology.

## 2. The INWA Grid Infrastructure

The project has designed and implemented a Grid Infrastructure using existing freely available Grid technology. This allows analysts at Edinburgh or Perth to submit batch jobs securely that are run on a compute resource local to the data being processed. The results from the batch jobs are automatically transferred back to the user. The Infrastructure also allows analysts to interact with the relational data sources via SQL queries.

To submit and transfer batch jobs and their results between local and remote sites the Infrastructure uses Grid Engine V5.3[7] as the compute resource manager, and Transfer-queue Over Globus (TOG) [8] with Globus Toolkit V2 [9] for the grid middleware.

Grid Engine is an open source distributed resource management system that allows the

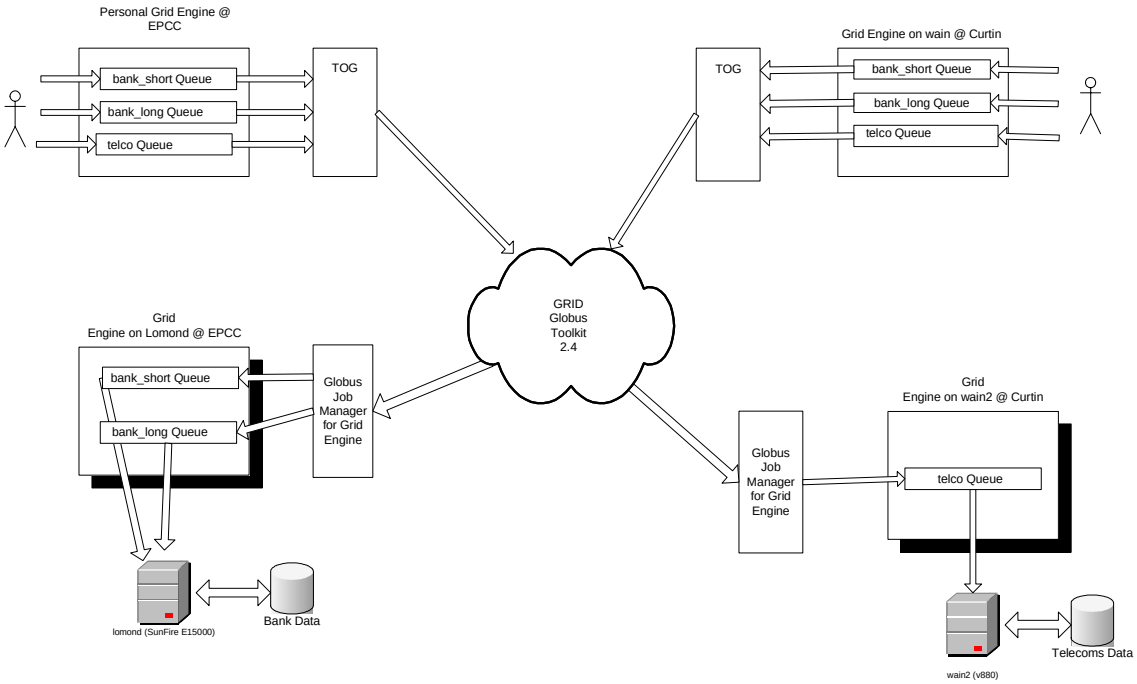


**Figure 1: Job submission and execution using the INWA Grid Infrastructure**

efficient use of compute resources within an organisation. The Globus Toolkit is essentially a Grid API for connecting distributed compute and instrument resources. TOG integrates Grid Engine V5.3 and Globus Toolkit V2 to allow access to remote resources. This allows an enterprise to access remote compute resources at any collaborating enterprises [10].

The Infrastructure uses OGSA-DAI Release 3.1 [11] and the FirstDIG browser [12] to provide interactive access to the relational data resources via SQL queries. This allows Grid technologies to be applied at the crucial data cleaning and preparation stages that precede any data mining.

OGSA-DAI (Open Grid Services Architecture-Data Access and Integration) is grid middleware software designed to assist with the access and integration of data from separate sources via Grid Services. OGSA-DAI Release 3.1 uses the Globus Toolkit V3.0.2. The FirstDIG browser allows a user to interact with OGSA-DAI grid service enabled data sources using SQL queries via a GUI.

This paper describes the INWA Grid Infrastructure and discusses its strengths and weaknesses.

## 3. Batch Job Submission and Execution

The Infrastructure is used to run data cleaning and mining batch jobs on the data located at EPCC and Curtin. The data cleaning batch jobs generally use GritBot [13] and Perl scripts whilst the data mining batch jobs use C5.0 [14].

Batch job submission and execution using the Infrastructure are illustrated in Figure 1. The Infrastructure has two compute resources, one at EPCC and the other at Curtin. Both these compute resources are running Grid Engine. Grid Engine provides job queues into which all jobs to be executed on the compute resource must be placed. Grid Engine will cause the jobs to be executed when the required compute resource is available. The analysts do not interact with the compute resources directly. Analysts submit jobs to personal versions of Grid Engine running on their local system.

The queues in these personal Grid Engine installations are configured to use TOG to forward the job to a Grid Engine queue on a remote compute resource. TOG is configured to ensure that the Globus Toolkit securely encrypts a job and its results during transfer to and from the remote compute resource.

This Infrastructure has several advantages for the analyst over the non-Grid approach of

simply having a secure shell connection to each of the compute resources:

- The Globus toolkit provides a certificate-based authentication mechanism that allows the jobs to be sent to the remote resources without the need to remember a username and password for each resource.
- Analysts can use multiple compute resources without any need to know where the resources are located. This information is held within the personal Grid Engine configuration that maps queue names (which are based on data set names) to queues on remote resources.
- All input and output files associated with jobs are stored on the analyst's account on his local system.

There are also a few potential disadvantages of the infrastructure:

- The analyst has to submit all jobs via Grid Engine. If the analyst is not familiar with Grid Engine this will be an additional learning curve.
- There is no simple way for an analyst to determine the state of a job running on a remote compute server. Querying the state of a job submitted to a personal Grid Engine will return the state of the job on that personal Grid Engine rather than the state of the remote job.

## 4. Interactive Data Access

A secure OGSA-DAI configuration and the FirstDIG browser were installed at EPCC to allow interactive access to the relational data resources. The FirstDIG browser allowed the user to interact with the suite of OGSA-DAI Grid data services and send SQL queries to the data resource. The query results are displayed in a result window and can be saved to a CSV file if desired.

This secure OGSA-DAI configuration ensures that only authorised and authenticated users can access the grid data services and that SQL queries and their results are encrypted during transfer. OGSA-DAI Release 3.1 uses Globus Toolkit V3.0.2 to provide the authorisation, authentication and encryption facilities.

Unfortunately, no general-purpose software packages, such as interactive statistics packages, currently support the OGSA-DAI interface and it is likely to require great leaps forward in terms of Grid technology and standardization before they do so. If the Grid community wishes to use existing data processing and visualization tools with Grid-enabled data resources it ought to be possible to develop ODBC and JDBC bridges for OGSA-DAI that would allow applications to access these data resources over the Grid using standard interfaces.

Given that there are so many existing applications that access data resources using ODBC and JDBC drivers it is worth addressing what advantages would be gained by using OGSA-DAI for simple access to remote data resources. OGSA-DAI offers the following advantages:

- Most JDBC and ODBC drivers pass login details in clear text and hence highly compromise security. OGSA-DAI allows the use of the Globus Toolkit's more secure certificate-based authentication mechanism instead.
- There is no need for the user to obtain and remember login details to every data resource they wish to access. All authentication and authorization will be carried out using the Globus Toolkit's certificate based mechanism.
- OGSA-DAI provides an extra level of access control mapping that eases the task of managing user access at the server.
- OGSA-DAI can encapsulate many diverse data resources (e.g. Oracle, MySQL and DB2 databases) behind a single interface. Thus the client application does not require a driver for each data resource type – these drivers are required only at the server.

## 5. Conclusions

The project has been able to construct a Grid infrastructure that has facilitated many of the required data mining activities. Secure batch job submission, execution and automatic file transfer has been very successful. Using Grid technology has offered the analysts several advantages in terms of usability over non-Grid shell based approaches. The inability to easily discover the state of a job while it is at a remote queue was the major usability drawback.

The construction of the batch job submission part of the Grid Infrastructure took longer than expected. The installation of Globus Toolkit V2 is still not a trivial task and although the researcher had installed it several times before, complications still arose that took several days

to overcome. Software bugs also delayed the construction. These clearly indicate much of the Grid Infrastructure software cannot yet be considered a mature product.

OGSA-DAI does provide an excellent mechanism to allow trusted parties access to a data resource in a secure way. The FirstDIG browser makes it easy for users to interact with OGSA-DAI services. It would be useful if other OGSA-DAI functionality such as meta-data and database schema can be exposed though this browser.

This experimentation with OGSA-DAI provided some useful insights. It is unlikely that commercial institutions will trust Grid technology enough to utilise such mechanisms to provide external access to sensitive data at any point in the near future. However, it does seem reasonable to hope that academic institutions will provide read-only access to research data resources using Grid technologies.

OGSA-DAI failed to handle large data sets due to out-of-memory errors. Extremely large data sets are an essential part of the Grid vision and it is important that OGSA-DAI can handle them without crashing the server. In order to achieve this, streaming data transfer must be placed at the heart of OGSA-DAI thinking. All activity developers must think of streaming data transfer as the primary mechanism. In response to this project and others the OGSA-DAI team have revised several key activities that failed to support streaming and as a result OGSA-DAI Release 4 now handles large data sets considerably better.

Throughout this work, bugs or omissions relating to security have been discovered. This seems to indicate that security considerations are not a high priority for grid software developers. It is crucial that a higher importance is placed on security if the Grid is to gain wider acceptance amongst commercial organisations. This is particularly important if the Grid vision of secure easy access to shared data and resources between commercial organisations is to be achieved.

## References


[1] The INWA project, http://www.epcc.ed.ac.uk/projects/inwa

[2] ESRC Pilot Projects in E-Social Science, http://www.esrc.ac.uk/esrccontent/researchfunding/escience.asp

[3] EPCC, http://www.epcc.ed.ac.uk/

[4] UEMS, http://www.ems.ed.ac.uk/

[5] LUMS, http://www.lums.lancs.ac.uk/

[6] Curtin Business School, http://www.cbs.curtin.edu.au/

[7] Grid Engine, http://gridengine.sunsource.net/

[8] Transfer-queue Over Globus (TOG), http://gridengine.sunsource.net/project/gridengine/tog.html

[9] The Globus Alliance, http://www.globus.org/

[10] T.M.Sloan, R.Abrol, G.Cawood, T.Seed, F.Ferstl, "Sun Data and Compute Grids", Proceedings of the 2nd UK e-Science All Hands Meeting, 2-4 September, 2003, Nottingham, UK

[11] OGSA-DAI, http://www.ogsadai.org/

[12] The FirstDIG project, http://www.epcc.ed.ac.uk/firstdig

[13] GritBot, http://www.rulequest.com/gritbot-info.html

[14] C5.0, http://www.rulequest.com/see5-info.html