

# **GEDDM: Grid Management Framework to Support Computationally Intensive On Demand Data Mining Services in a Business Environment**

**Mark Prentice  
Karen Loughran  
Paul Donachy  
Terry Harmer  
Ron H. Perrott**

Belfast e-Science Centre  
[www.qub.ac.uk/escience](http://www.qub.ac.uk/escience)

**Jens Rasch  
Sarah Bearder**

Datactics Ltd  
[www.datactics.co.uk](http://www.datactics.co.uk)

## **Abstract**

With the explosion in size of data warehouses and the proliferation of databases, managing and mining large volumes of unstructured data is still the most critical element currently affecting companies attempting to control their data assets. At present, it can not be stressed enough how poorly developed many of the current practices are and how as the size of datasets is only going to increase massively in the near future, it is of significant commercial importance to develop scalable affordable techniques for handling these issues. A grid enabled environment has the potential to solve this problem by providing the core processing capabilities with secure, reliable and scaleable high bandwidth access to the various distributed data sources and formats across various administrative domains.

When running computationally intensive processes such as data mining operations in a dynamic grid environment, it is advantageous to have an accurate representation of the available resources and their current status. This resource information can then be used for optimal on-demand allocation of distributed resources at runtime, and can significantly improve both performance and end user quality of service. This is extremely important when you wish to maximize the efficient utilization of a set of the available resources within an organization.

GEDDM is a collaborative industrial e-Science project, run by the Belfast e-Science Centre in conjunction with industrial partner Datactics Ltd. This paper describes mechanisms employed by GEDDM that expose the backend capabilities of the core data mining engine via grid services and allows these core services to be dynamically allocated to on-demand resources based upon the availability of the underlying infrastructure. The grid management framework dynamically collects, monitors and manages information from all of the resources in the system through the use of a innovative lightweight grid management agent that is deployed on each available resource. This lightweight agent continually monitors the status of that resource and periodically reports the status information to the grid management server.

The information stored in the grid management service is then used by a job allocation service when allocating data mining operations to particular resources. This approach enables resulting job allocation to take into account resource availability (such as resources that are not responding or resources that currently have a high CPU load). It is achieved through the use of OGSA based grid services developed using Globus Toolkit 3.

The benefits of this approach have been proven in a real world system using the grid services and the data mining capabilities developed by Datactics. The solution described takes full advantage of the underlying security mechanisms provided by the Globus Toolkit and is designed to be used by Datactics clients to mine data in a real world environment.

This management framework is being used in conjunction with data mining grid services, where large datasets are preloaded into memory across parallelised clusters for fast, real-time computations. These grid services provide on demand access to remote datasets and large scale computational resources through a simple grid based API. The solution uses Globus based grid services accessed through a web portal created using Gridsphere.

This paper presents the background and architecture used in developing grid services which exposes both the backend capabilities of the data mining engine and defines the grid management framework. Two commercial use cases are outlined in this paper. The first deals with sending large scale data mining operations to distributed computational grid resources for load balanced jobs, which otherwise would take days or weeks without the use of grid resources. The second use case deals with large numbers of small queries against a single large dataset that is preloaded into memory.

In addition, this paper will detail the preliminary results of using the system in a business environment and the issues encountered while developing grid services using the Globus toolkit. Experiences of using and deploying GT3 grid solutions will also be discussed as well as possible areas for improvement. Finally, results of large scale end user tests will be presented which will outline how this grid solution has been applied to a real world business environment.