The Queen's University of Belfast
BeSC
Belfast e-Science Centre

# GEDDM - Commercial Data Mining Using Distributed Resources

Mark Prentice

1st December 2004

datactics

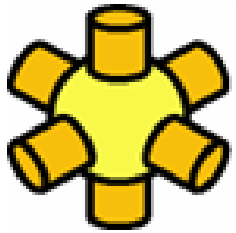
# Introduction



- Industrial partner
- Overview of GEDDM project
- Application areas
- Grid enabled implementation
- Current status

# Industrial Partner





datactics

- Northern Ireland based company (formed 1999)
- Provide data mining services using custom engines
  - Engines already parallelisable over an internal cluster
- Data mining software being used in the real world
- Improve data quality by applying fuzzy matching and parallel processing to achieve greater depth and accuracy

# Commercial Business Drivers



- Data sources
  - numerous structures, formats, locations administrative domains…
- Few customers want to buy their own hardware
- Example: Bank with say 10,000 branches
  - Each branch could send in a request for a query against a large scale in house datasets
  - Need to be able to handle these requests efficiently and securely
- Example: US County Court litigation case
  - Datactics asked to mine 45TB of data
  - Spread over thousands of PCs
  - Extract and process highly distributed data

datactics

# Fuzzy logic - How many errors can you spot?

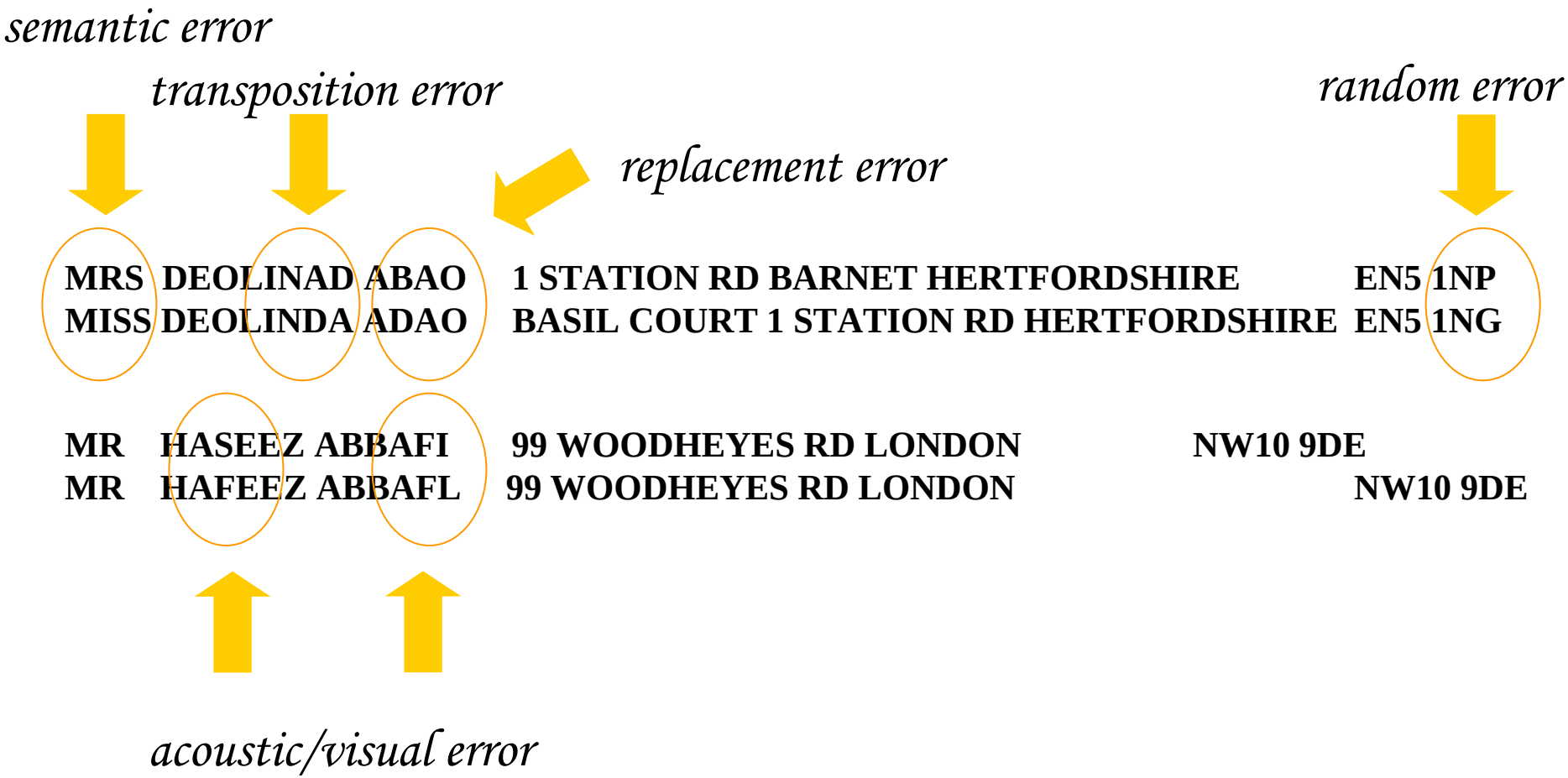
Belfast e-Science Centre

| | | | |
|---|---|---|---|
| MRS | DEOLINAD ABAO | 1 STATION RD BARNET HERTFORDSHIRE | EN5 1NP |
| MISS | DEOLINDA ADAO | BASIL COURT 1 STATION RD HERTFORDSHIRE | EN5 1NG |
| MR | HASEEZ ABBAFI | 99 WOODHEYES RD LONDON | NW10 9DE |
| MR | HAFEEZ ABBAFL | 99 WOODHEYES RD LONDON | NW10 9DE |

datactics

# Typical errors






datactics

# GEDDM Overview

Belfast e-Science Centre

- GEDDM - **G**rid **E**nabled **D**istributed **D**ata **M**ining
- 2 year project - started August 2003
- Uses Datactics fuzzy parallelised data-matching and transformation engine to perform data-mining operations
  - Deals with large volumes of data, currently anything from a few MB to around 100 GB
  - Existing engine and GUI are platform independent
- Computationally intensive - need to compare every record with every other record ($n^2$ process)

datactics

# Objectives



- Use Grid Technology to expose core engine as Grid Services using Globus Toolkit
- Provide secure remote access to data mining engines through grid mechanisms
- Provide secure file transportation between remote clients and data mining hardware
- Provide basic node management of underlying hardware
- Use basic load balancing when allocating data mining jobs

# Objectives (continued)

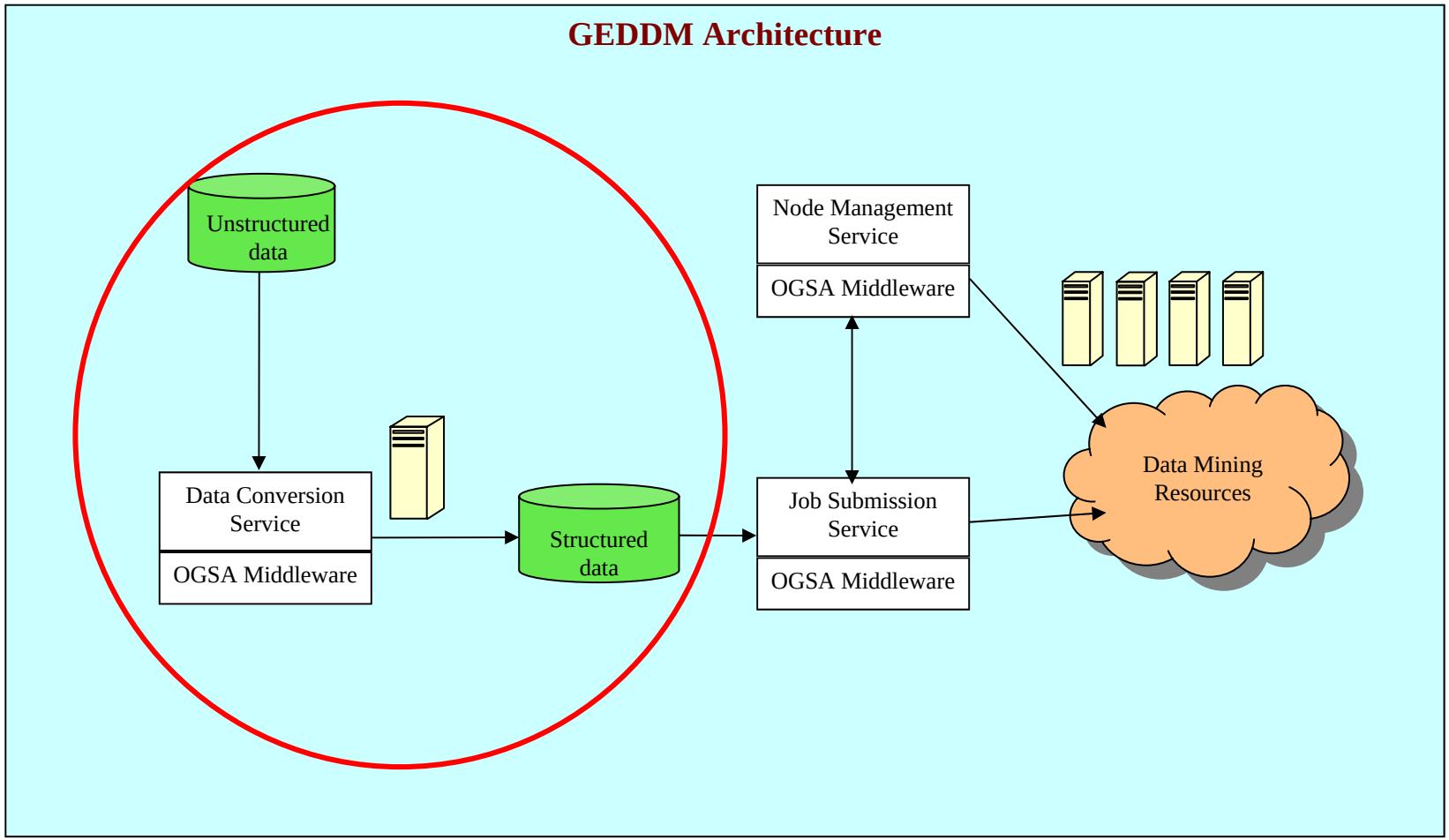
- Provide services to convert unstructured data sources into common structured data format
- Allow conversion of web logs, email, pdf, RDBMS, Word documents, etc
- Minimal dependencies

# Applications



- **Watch List Compliance -** checking bank account lists or passenger lists with lists of suspected criminals
- **Forensic accounting –** e.g. checking databases for fraudulent billing
- **Financial/Telco/Government/Direct Marketing –** checking for duplication of customer data
- **Structural analysis -** examining documents for common phrases e.g. insurance claims
- **Astro-physical image analysis –** using catalogue data

# GEDDM Architecture



# Grid Enabled Solution – Unstructured Data Conversion



- Provides GT3.2 grid services to convert unstructured data into a common structured format
- Provides XML templates to describe common unstructured formats (e.g. web logs)
- Common output format files can be passed to Data Mining Services for data matching operations

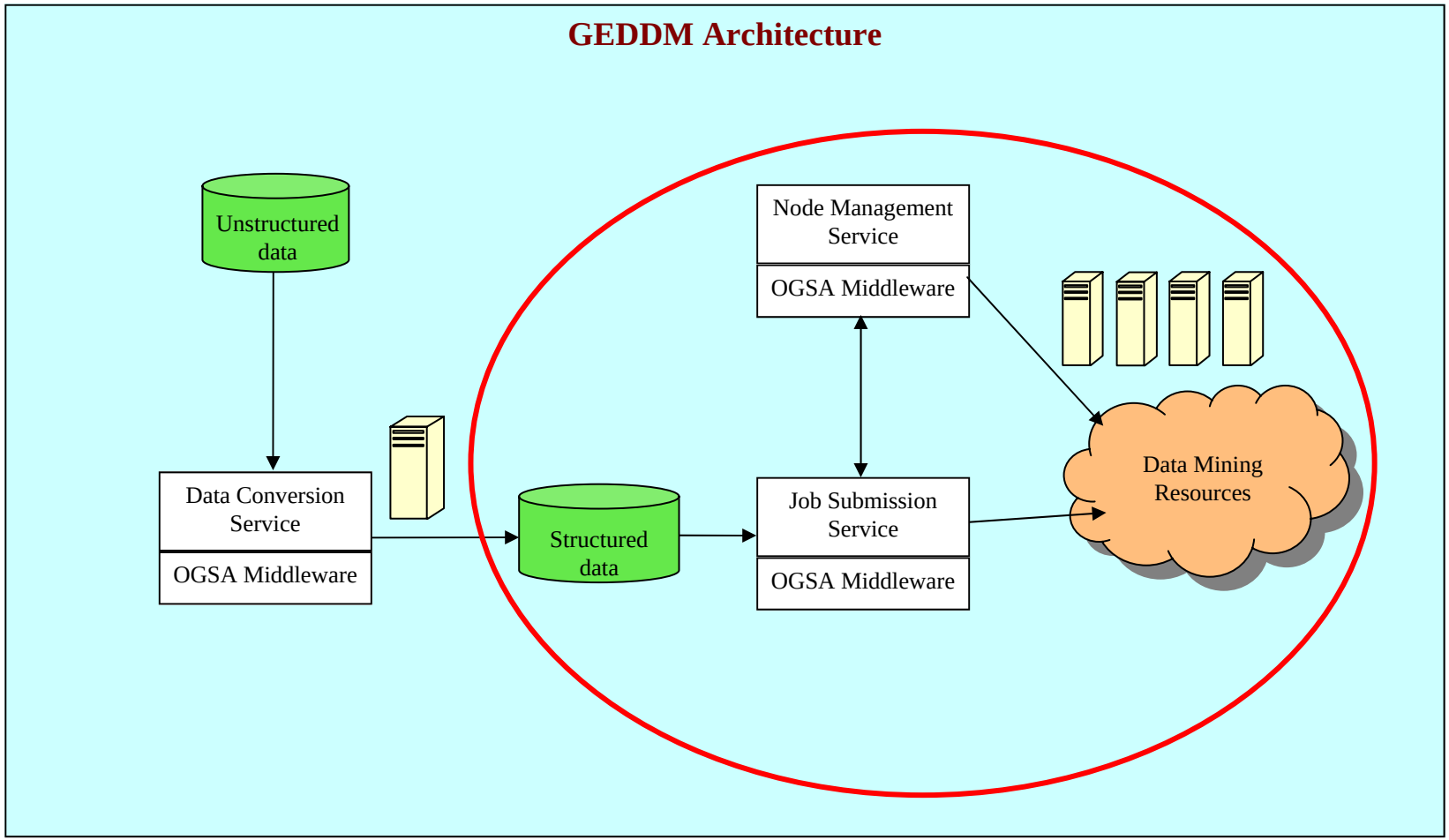
# Unstructured Data Format Supported

The Queen's University of Belfast
BeSC
Belfast e-Science Centre

- Emails
- Web Logs
- PDF's, Word Documents
- RDBMS
- Reports

datactics

# GEDDM Architecture



# Data Mining Services



- Node Management Adaptor
  - C++ & gSOAP application (small footprint)
  - Used to register nodes on a cluster
- Node registry service
  - GT3 service
  - Used by job submission service for load balancing when allocating jobs
- Job submission service
  - Secure GT3 service
  - Creates job management service instance per job
- Job management service
  - Secure GT3 service
  - Starts data mining engines and monitors job progress

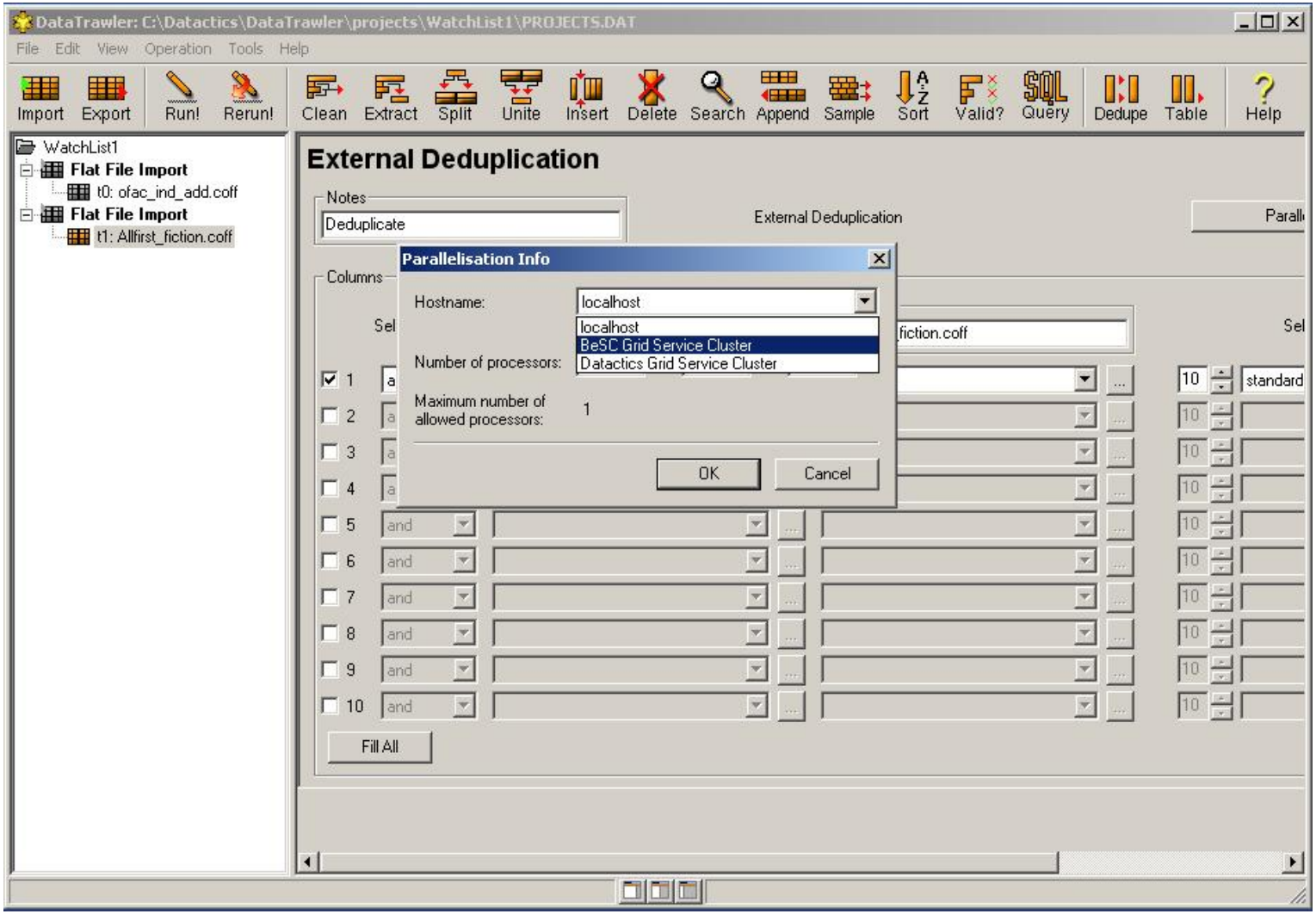
# Commercial Software Integration



- Automated file transfer between distributed resources (currently using scp)
- All communication with the remote grid services uses GSI message level security
- Changes to existing Data Mining software application minimal
- Client side dependencies minimal
- User selects parallelization of a job on a grid cluster and the rest is transparent

# Selecting Grid Environment






datactics

# Benefits of Grid Enabled Solution



- Remote job submission
- Status of jobs can be monitored remotely
- Status of cluster(s) can be monitored remotely
- Specification of cluster can be viewed
- Secure, reliable and scaleable
- Decoupling of GUI from data mining engine
- Extends range of data sources that can be queried by data mining engine

# Current Status



- Beta testing stage of data mining services
- Client side integration working under Windows and Linux
- Demoed software at AHM04
- Data Conversion services currently being developed
- OnDemand services starting development
  - Embed data mining engine

- Email: m.prentice@qub.ac.uk
- Project Webpage : www.qub.ac.uk/escience/geddm
- Demo available for viewing