

First Data Investigation on the Grid : FirstDIG

T. M. Sloan^{1,3}, A. Carter^{1,3}, P. J. Graham^{1,3}, D. Unwin², I. Gregory²

¹EPCC, The University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh, EH9 3JZ, UK

²First South Yorkshire, Midland Road, Rotherham, S61 1TF, UK

³National e-Science Centre, e-Science Institute, 15 South College Street, Edinburgh, EH8 9AA, UK

Abstract

The FirstDIG [1] project is a collaboration between First plc [2] and the National e-Science Centre (NeSC) [3] as represented by EPCC [4]. The project aims to deploy an early implementation of the OGSA Data Access and Integration services (OGSA-DAI) [5] within the First South Yorkshire bus operational environment. The project has two central goals. The first is to demonstrate the deployment of OGSA-DAI services in a commercial environment, and learn from this process. The second goal is to answer specific business questions posed by First through a short data mining analysis using the OGSA-DAI service enabled data sources. The project started in May 2003 and will run through to January 2004. This paper describes the project and its current status as of August 2003.

Introduction

As stated in [6], the world wide web provides seamless access to information that is stored in many millions of different geographical locations. Grid technology takes this concept one stage further by allowing seamless access and use of computing resources as well as information. The transformation of the Grid from an enabling technology in the scientific domain to a widely used business tool is a key requirement if the success expected by the UK Government is to be realised [7]. To date the UK e-Science Programme [8] has involved a large number of companies in collaborative projects. Many of these companies are either IT related or classic early adopters. Few projects involve service companies such as First plc. However, the company has a clear business problem the Grid can help them solve and for this reason they are collaborating in the FirstDIG project [7].

Since its formation in 1995, First plc has striven to improve the opportunities for bus and rail travel in cities and towns across the UK. Operating world-wide in many different transport sectors, the company runs over 10,000 vehicles in the UK and has 23% of the market, making it the UK's largest operator. First is represented in this project by First South Yorkshire buses [7].

Many businesses turn to data mining to better inform their decision-making. In the transport industry, the huge range of fragmented data sources has hindered its adoption. In the FirstDIG project we intend to demonstrate how

OGSA-DAI services can provide cost-effective access to disparate data sources and hence enable data mining of these sources to answer specific business questions [7].

This paper provides a brief description of OGSA and OGSA-DAI. It then goes on to describe how OGSA-DAI will be used to provide access to the various data sources to be mined in the FirstDIG project. Finally the paper outlines the current status of the project

OGSA and OGSA-DAI

A Web Service is an interface that describes a collection of operations that are network accessible [9]. Open Grid Architectures facilitate the sharing of data and computing resources amongst collaborating organisations. Such collaborating organisations may be geographically distributed with heterogeneous platforms.

The Open Grid Services Architecture (OGSA) [10], an extension of XML-based Web services, pulls together the concept of an Open Grid Architecture with Web Services to define a set of implementation and platform independent protocols and standards based around the concept of creating, managing and exchanging information among entities called Grid Services. A Grid Service is a stateful Web Service with an associated lifetime that conforms to a set of interfaces and behaviours with which a client may interact [9].

The purpose of OGSA-DAI is to provide uniform service interfaces for data access and

integration via the Grid. Through OGSA-DAI interfaces, disparate, heterogeneous data sources and resources can be treated as a single logical resource. Moreover, OGSA-DAI will allow the same data source and resource to be incorporated into an OGSA-compliant architecture. The OGSA-DAI Grid services will themselves then provide the basic operations that can be used to perform sophisticated operations such as data federation and distributed queries, hiding concerns such as database driver technology, data formatting techniques and delivery mechanisms[9].

FirstDIG Data Sources

The data sources to be used in the FirstDIG project are from the following systems within First South Yorkshire.

- Customer Contact – this records correspondence with customers including commendations and complaints.
- Vehicle Mileage – this records the daily vehicle mileage for bus services.
- Ticket Revenue – this contains the daily tickets sold and the money taken for the bus services.
- Schedule Adherence – a satellite tracking system that records whether a bus is arriving and departing on time from a bus stop.

These systems are located at various company sites, on differing platforms in different databases. The databases range from SQL sources to ODBC sources to COBOL files. It is precisely these issues that any technology must address in order to be applicable and useful to business.

OGSA-DAI will be deployed on the relevant systems at First South Yorkshire and grid services written to enable access to the relevant data. A series of data filters, converters and query tools will be developed to provide access to the necessary data for these resulting OGSA-DAI data services [7].

Data Mining and OGSA-DAI

Through data mining and statistical analyses the FirstDIG project aims to answer specific business questions posed by First. This will require the consolidation of data from the customer contact system, the satellite tracking schedule adherence system, the mileage records system and the revenue system. The project will determine if the OGSA-DAI-enabled data

services can extract and consolidate the necessary data from these systems to answer these business questions. Figure 1 illustrates how OGSA-DAI will be used to extract data from the various systems and make it available for data mining and analysis.

Current Status

As previously mentioned the various data sources to be used in the project have been identified. In addition, First South Yorkshire has formulated a set of business questions whose answers require consolidation of data from these various sources. The questions cover topics such as:

- the effect of lost mileage on revenue, where lost mileage is due to activities such as road works and breakdowns;
- the effect of lost mileage on the number of complaints;
- the effect of reliability on revenue;
- the effect of reliability on complaints received.

In addition, these questions will be answered at deeper levels such as frequency of services and passenger ticket type.

Thus far the data sources have been examined in detail to establish the relevant tables and fields necessary to answer the questions. This examination has also been necessary to determine what the various data filters, converters and query tools will need to do in order to make the necessary data available via the OGSA-DAI services.

The next stages of the project are to design and implement the necessary filters, converters and query tools, deploy these with OGSA-DAI and so build the data services. These data services can then be used to help answer the business questions posed by First South Yorkshire.

Further Information

For more information on the project and its deliverables, please visit the project web site at <http://www.epcc.ed.ac.uk/firstdig/>.

References

- [1] FirstDIG project home page, <http://www.epcc.ed.ac.uk/firstdig/>
- [2] First plc home page, <http://www.firstgroup.com/>
- [3] National e-Science Centre home page, <http://www.nesc.ac.uk/>
- [4] EPCC home page, <http://www.epcc.ac.uk/>

- [5] OGSA-DAI home page,
<http://www.epcc.ed.ac.uk/gridserve/>
- [6] "GRID technology explained...",
<http://www.escience-grid.org.uk/docs/gridtech/busbrief.htm>
- [7] M.Parsons, B. Bedford, D. Unwin "First Data Investigation on the Grid – FirstDIG", National e-Science Centre project proposal, Doc Id: FirstDIG-GCP-PR-V1.0.doc, 31st October 2002
- [8] The UK e-Science Programme home page,
<http://www.escience-grid.org.uk/>
- [9] M. Antonioletti and M. Jackson, "OGSA-DAI Product Overview", June 2003,
<http://www.ogsa-dai.org.uk/docs/current/OGSA-DAI-USER-UG-PRODUCT-OVERVIEW.pdf>
- [10] Open Grid Services Architecture (OGSA),
<http://www.globus.org/ogsa/>

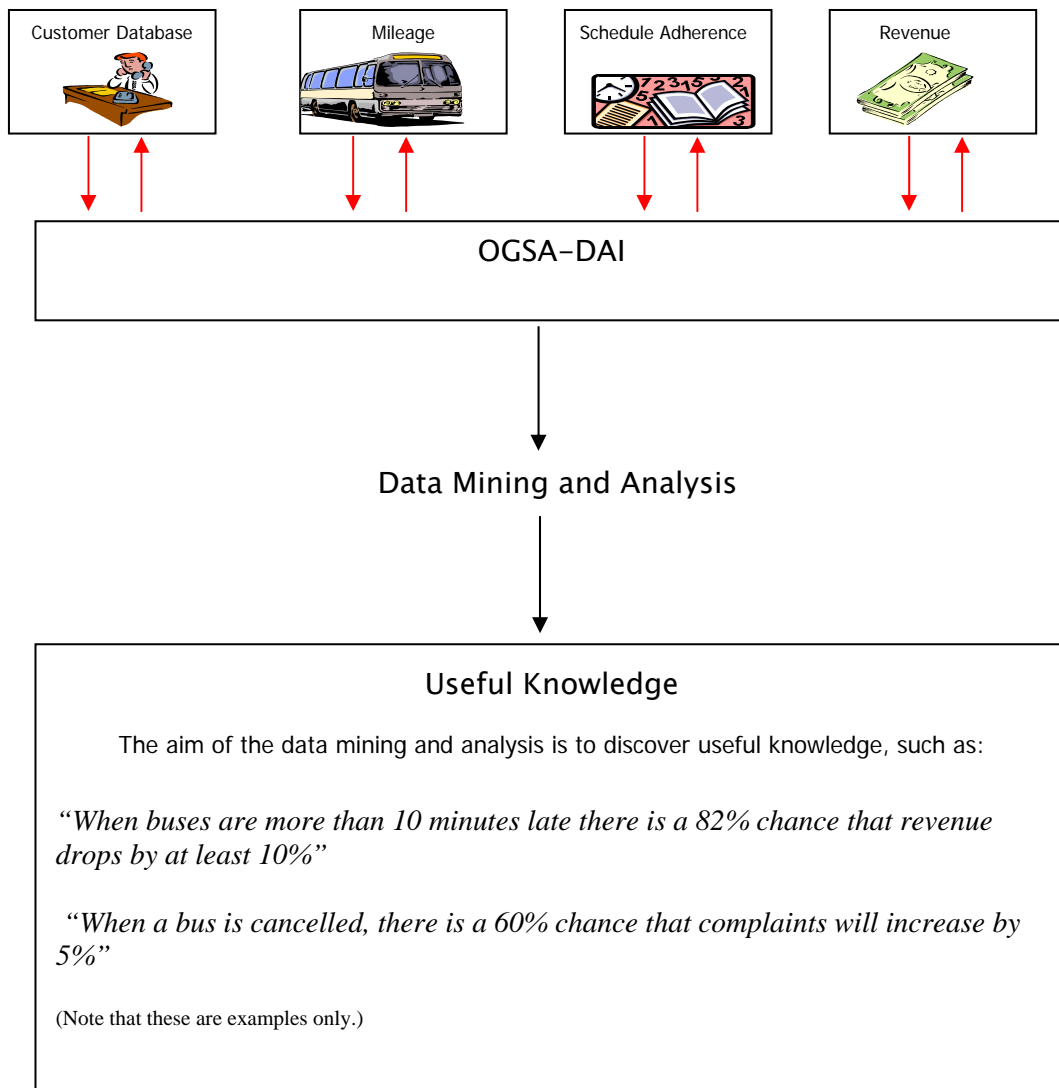


Figure 1: Using OGSA-DAI to extract and consolidate data from various sources for subsequent data mining and analysis.

