SmartCities

IST Project Number 12252

Schlumberger Systèmes
Sema UK Ltd
Schlumberger Industries Ltd
Southampton City Council
University of Southampton
MasterCard Europe
Technolution
Crid
City of Goteborg
IT Innovation
Black Sea Consulting

D7.4
Best practice for smart city card information analysis report.

Abdul-Rashid Abdi, IT Innovation
Steve Taylor, IT Innovation

Version 1
15 May 2003

Contents

1 Introduction

In the SmartCities IST project, a Multi-owner Data Analysis System (hereafter referred to as "the Data Analysis System") that consists of data collection, storage and analysis compnents is introduced. SmartCities smart card transactions are captured and securely stored together with card and cardholder related information. SmartCities Data Analysis System users are given access to a range of analysis tools that enable them to analyse stored data over the web.

This report describes the approach taken in SmartCities to build a secure, data protection and privacy compliant Data Analysis System. Lessons learned in setting up such a system are summarised and implementation tips are introduced. The report is intended to give advice to those who are about to embark on a similar undertaking.

2 System design

SmartCities experience showed that setting-up a multi-owner data analysis system has many advantages including:

- giving user organisations' analysts powerful analysis tools that can be accessed over the web from anywhere using low band-width connections;
- clearly defining the organisation hosting the Data Analysis System as a Data Processor – thus establishing lines of accountability;
- reducing the cost of setting up the Data Analysis System as setup and ongoing maintenance costs can be spread amongst participating organisations/departments;
- increasing collaboration between different department in the same organisation by allowing them to cross reference and cross-analyse their data; and
- centralized management of security and access controls, thus enhancing privacy and data protection in general.

In setting up the Data Analysis System, three distinct components are identified:

- data collection;
- data storage and
- analysis tools.

The data collection module should allow for a seamless data collection from various systems and deliver collected data to the data storage modules securely and safely. The storage module should allow for the safe and secure storage of the data. Analysis tools should allow access to complex analysis functionalities over a low bandwidth connection in order to enable information workers to access the data from anywhere through a basic web connection.

The data collection module will in most cases require development of bespoke tools integrated with available modules (such as ftp, encryption and cryptographic tools), as there are no stand-alone off-the-shelf data collection system that will perform exactly the function required. In the case of SmartCities Data Analysis System, in-house development resulted in an on-line/off-line set of data collection tools.

In selecting data storage and analysis tools, existing data storage and analysis products are likely to address most Data Analysis System's requirements. Oracle database and Oracle OLAP offerings were selected to store data and to enable web-based analysis for SmartCities but alternatives should not be ruled out.

In order to illustrate the technological possibility of the SmartCities Data Analysis System a Dissemination System was set up. This was populated with realistic-looking dummy data and made available to users to explore proposed solution features and comment as necessary.

The Dissemination System was enormously beneficial in giving potential users a good idea of what the Data Analysis System was capable of, and in generating interest in general. Through a number of seminars, the Dissemination System was introduced to users and comments gathered, which were fed into design process. Furthermore, the Dissemination System was used to construct case study scenarios to further demonstrate business benefits of a Multi-owner Data Analysis System.

2.1 Data collection

The SmartCities experience showed that designing and deploying a universal data collection system is likely to take the bulk of the development effort.

Early on in the development phase, it is recommended that the data collection system should be viewed as four distinct components that delimit the four distinct phases of data movements from source to the data storage unit/s. The distinct phases can be described as:

- data preparation;
- data transmission;
- data transformation and
- data loading.

Data preparation is the process of transforming the source data into formats that can be exported to the outside world. Data preparation in SmartCities involves two fundamental processes;

- de-personalisation in accordance with the relevant privacy and data protection guidelines and
- encrypting data files.

In all cases, the de-personalisation and encryption operations must take place within the data Owner's organisation. Data preparation software will run on the data owner's platforms and therefore should be designed with target systems capabilities in mind. Data preparation software should be institutive and easy to install and deploy as it will, in most cases, be handed over to target systems' administrator/s.

At the data transmission stage, files that contain 'prepared source data' are transmitted to the Data Analysis System. 'Prepared source data' can be pulled from or pushed by the Data Owner's system. In SmartCities, 'prepared source data' is pulled from the Data Owner's system.

Once 'transmitted prepared source data' reaches the data storage environment it needs to be decrypted and then translated into a format that can be uploaded into the data storage unit/s. In SmartCities, the transformation processes is accomplished in a data staging area.

Data loading software consumes the transformation stage output and loads the data into the data storage unit/s.

Data collection should be designed so that collecting data is viewed as a background task that runs with little visibility and minimum impact on resources. Once set-up, collecting data should require little human intervention. Other features that should be considered in data collection systems are; ability to recover from error gracefully, good error notification routines and adequate traceability features.

2.2 Data storage solutions and data analysis tools

In SmartCities, Oracle database solutions were selected to store data as Oracle provides mature, powerful and popular data storage solutions. Unlike data collection systems, databases systems are well understood and it is likely that most, if not, all popular data storage offering will address

multi-owner data analysis requirements.

In terms of storing data, data owned by different SmartCities organisations are uploaded into different secure schemas and any cross-analysis is achieved by permission of the data owners and through the introduction of new schemas. This form of data organisation achieves the privacy requirement that application provider's data should be kept separate and any cross-analysis must be achieved through the introduction of new schemas.

In SmartCities, Oracle Discoverer tools were selected to give access to simple reporting and complex analysis functionalities. Discoverer provides a friendly graphical user interface to help business users to conduct complex queries locally or remotely over the web.

3 Security and Privacy

In any data handling system, security and privacy are naturally of paramount importance. SmartCities Data Analysis System design has taken a number of steps to ensure that data is secure and that privacy is not infringed.

3.1 Security

A major lesson learned here is that Data Owners need to be convinced of the security of the data warehouse and that their data will be protected safe from unauthorised access. In SmartSities, a number of security features are built into the architecture of the Data Analysis System to safe guard data from unwarranted access;

- Each Data Owner has its own secure area within the Data Analysis System, in which their data is stored. Access to a Data Owner's secure area is restricted to members of that organisation.
- The Data Analysis System is located in a De-Militarized Zone (DMZ). This is a secure area that is guarded from the Internet by a firewall. Furthermore, the DMZ is cordoned off from the internal network of its hosting organisation by another firewall. This means that it is protected from intrusion from both outside and within its hosting organisation.
- Access to the Data Analysis System is granted on an IP address basis. Only recognised IP addresses are allowed to access the DMZ by its internet-side firewall.
- All transmissions to and from the Data Analysis System are secured using HTTPS. This ensures that the data cannot be observed or altered during transmission, and that the endpoints are authenticated.
- All access to the Data Analysis System is logged. Thus, a complete audit trail exists so that any misuse of the system will be recorded.

3.2 Privacy

There must be no data in Data Analysis System that may identify an individual. This is important not only for legal reasons, but also from a user perception point of view. Individuals must not get the impression that, without their consent and knowledge, their behaviour is monitored and decisions are being made based on their activities.

In SmartCities, data collected from each data owner is anonymised at source before leaving its owning organisation. The practical implementation is that all personal data is either removed (name, address, etc) or scrambled (customer ID, account number, etc). The resultant effect is that using the data from the Data Analysis System alone, an individual cannot be identified.

To enable cross-analysis without exposing individuals, Data Owners should be given scrambling keys that are only accessible to them. These keys are to be used to scramble IDs (Card IDs, customer number etc.). Cross-analysis of data is made possible as one key is used to scramble corresponding IDs in different data sets.

3.3 Data Privacy and Security vs multi-owner data analysis

Multi-owner data analysis involves using more than one organisation/department data. This results in a conflict of interest between the desired ability to analyse multi-owned data and the need to ensure that an organisation/department data is made secure. Naturally privacy and

security take precedence over the multi-owner data analysis requirement, and as such a solution to the problem that does not compromise security and privacy is needed.

The chief lesson learned here is that data owners within the project were not willing to share data with outside agencies. This is understandable, since some data is likely to contain information that is proprietary and thus concerned organisations/departments will not want to share it. However, it is perfectly possible that there are some organisations/departments willing to share data. In SmartCities, legal and technical solution (as described below) in order to facilitate data sharing amongst those are willing to do so were investigated and implemented in the form of analysing data that belongs to different departments within the same organisation.

The data share solution takes the form of a legally binding agreement between two parties that wish to share data. Sharing data is one-way, in which there is a data Owner, and a second party to which the Owner grants access (the Grantee). Should there be a need to two-way data sharing, there should be two agreements. The terms of the access agreement (i.e. to which sets of data the Grantee is permitted access, and for how long, etc) are determined per agreement.

The technical implementation of the above legal agreement is as fellows;

- a new secure area is to be created in the Data Analysis System, to which the Owner has write access and the Grantee has read access only;
- the data Owner to generate a new scrambling key for the purpose of enabling data share with the Grantee
- the data Owner to scramble the IDs in the source data to be shared using the above key
- the scrambled shared source data should then be placed in the shared area in Data Analysis System an the Grantee given the scramble key
- the Grantee may then use the shared scramble key to scramble his own source data, thus producing the same scrambled value for a given ID in the corresponding Owner's data set
- the Grantee may then cross analyse his data with the relevant owner's data set

If the relationship is long-term, the anonymisation and upload to the multi-owner data analysis system's shared area may be automated in a similar manner to that of single organisation data collection and upload into own private area.

The above described data share approach has a number of advantages;

- The Owner controls exactly what is available in the shared area;
- No information exists in the multi-owner data analysis system that may identify a real person;
- Neither the Owner's nor the Grantee's private scrambling key is divulged and
- Should the shared scrambling key be compromised, all that is required is for the data to be deleted from the shared area, rather than the Owner's private area in the Data Analysis System.

The major disadvantage with this method is that it requires a separate data collection for each sharing relationship a data Owner has in addition to their usual data collection. However, the security and privacy advantages outweigh this disadvantage, as security and privacy must take precedence over all.

3.4 Legal Issues

The use of techniques to ensure anonymity and the wider issues surrounding the holding and processing of data should be investigated and technical solutions proposed to satisfy the concerns of the Directive 95/46/EC of the European Parliament and of the Council. Those who are embarking on a Multi-owner Data Analysis System should be aware of the provision of the Directive as well as National laws. Most notable and important points of the Directive as observed by SmartCities are listed below.

- The Directive refers to the concept that data can be processed whenever the Data Controller (organisations that owns the data) has a legitimate interest in doing so and this interest is not overridden by the interest of protecting the fundamental rights of the Data Subject, particularly the right to privacy. Deciding what is a reasonable balance between the business interest and the need for privacy of Data Subjects is the responsibility of the Data Controller.
- Data Controllers and their nominated Data Processors (an organisation that does not own the data but processes data on behalf of a Data Controller) are required to observe certain data processing rules, including:
    - data processing should not be excessive in relation to the purpose to which is to be processed; and
    - data is to be kept up to date.
- Data Subjects must be given information of the purposes of the processing.
- As the Data Processor has no legal right to choose the purpose of the processing of the data, they should not be able to identify individuals. The identification of an individual is thought to be possible if the IDs or combinations of personal data that collectively can identify an individual are made available to the Data Processor.

The cross profiling of a Data Subject on many fields of activity that where previously unconnected, but can be connected in the presence of a Data Analysis System, is not "per se" unlawful. There are, however, perception issues, and chief amongst these is the so-called the “Big Brother” issue. Steps must be taken to avoid citizens getting the impression that their activities are monitored.

In SmartCities, a Data Controller’s legitimacy to collect and analyse data was addressed by the Data Controller itself. A data processing contract between Data Controllers and the Data Processor was found to be necessary. Data Subjects were informed of Data Controllers’ intent and the Data Processor was prevented from identifying Data Subjects, as all records were de-personalized at source. Cross-analysis work was restricted to cover data belonging to different departments within the same organisation to allay citizens’ “Big Brother” concerns.

4 Conclusions

Multi-owner Data Analysis systems have many advantages including sharing the cost of setting up and maintenance amongst users, centralised management of security and privacy rules and increased collaborations amongst participants to name few.

During the design phase, prototyping was found useful. A Dissemination System was found valuable in soliciting users, educating them, and improving their understanding of how to utilise Multi-owner Data Analysis System to gain measurable business advantages. Multi-owner Data Analysis Systems should be viewed as consisting of three distinct but interconnected components; collection, storage and analysis.

The data collection component will in most cases require bespoke development. Storage and analysis components can be implemented using existing commercial products.

Multi-owner Data Analysis systems must address relevant data protection and privacy laws, as well as public perception issues. Data Controllers and Processors responsibilities must be fed into the design processes and implemented solutions must respect Data Subject privacy rights. Data Subjects' concerns must be tackled and allayed at the outset.