electronic Journal of Health Informatics
http://www.ejhi.net
2009; Vol 4(1): e6

Incorporating Privacy Support into Clinical Data Warehouses

Anders H. Landberg[1], Heather Grain[2],

J. Wenny Rahayu[1], Eric Pardede[1]

[1] Department of Computer Science, La Trobe University, Melbourne, Australia

[2] Llewelyn Grain Informatics, Melbourne, Australia

Abstract

This paper presents an analysis and implementation of a clinical data warehouse. It focuses on the nature of health data and points out implications that arise when warehousing this data. Especially concerns in regards to data privacy and authentication, data completeness and quality are addressed. First, we explore privacy preserving methods and propose a query-time validation scheme that protects against privacy disclosure caused by combining data attributes. To enforce the access control, we propose a novel concept of composite security levels.

Second, we introduce techniques and methods to overcome these issues, and suggest strategies for practical implementation. Finally, we introduce the system prototype that was developed during this project, and explain and illustrate, how these techniques and methods were applied in practice with emergency data.

Keywords: Data privacy, data linkage, Electronic Health Records, data analysis

1. Introduction

In recent years, the collection of health information in electronic form, such as patient hospitalisation data and diagnostics information has experienced large growth [1]. Consequently, the need for integrating this collected data has arisen, so that health information from different sources across the country can be conveniently compared and summarised. The benefits of this approach are to enable healthcare professionals, medical staff, and researchers, to discover new knowledge from the data, and to improve existing procedures in healthcare. Database-driven applications and spreadsheets with extended functionality have offered means to view and manipulate the data sources individually, but they do not specifically address analysis-heavy and integrated application contexts. With this issue of integrating the data into a centralised component, data warehouses have made their way in to the health information systems sector.

1.1. Motivation

In health informatics, we witness an ever increasing volume of data that is being collected. Currently, the Department of Human Services, Victoria, maintains several collections of health information. These are (i) the Victorian Admitted Episode Data Set (VAED), (ii) the Victorian Emergency Minimum Data Set (VEMD), (iii) the Elective Surgery Information System, and (iv) the Victorian Perinatal Data Collection [2]. Due to the increasing number and type of data, the need has arisen to warehouse these collections and to extract new knowledge and valuable information from it.

The *electronic Journal of Health Informatics* is an international journal committed to scholarly excellence and dedicated to the advancement of Health Informatics and information technology in healthcare. ISSN: 1446-4381
© Copyright of articles is retained by authors; originally published in the *electronic Journal of Health Informatics* (http://www.ejhi.net). This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License (http://creativecommons.org/licenses/by-nc-sa/2.5/au).

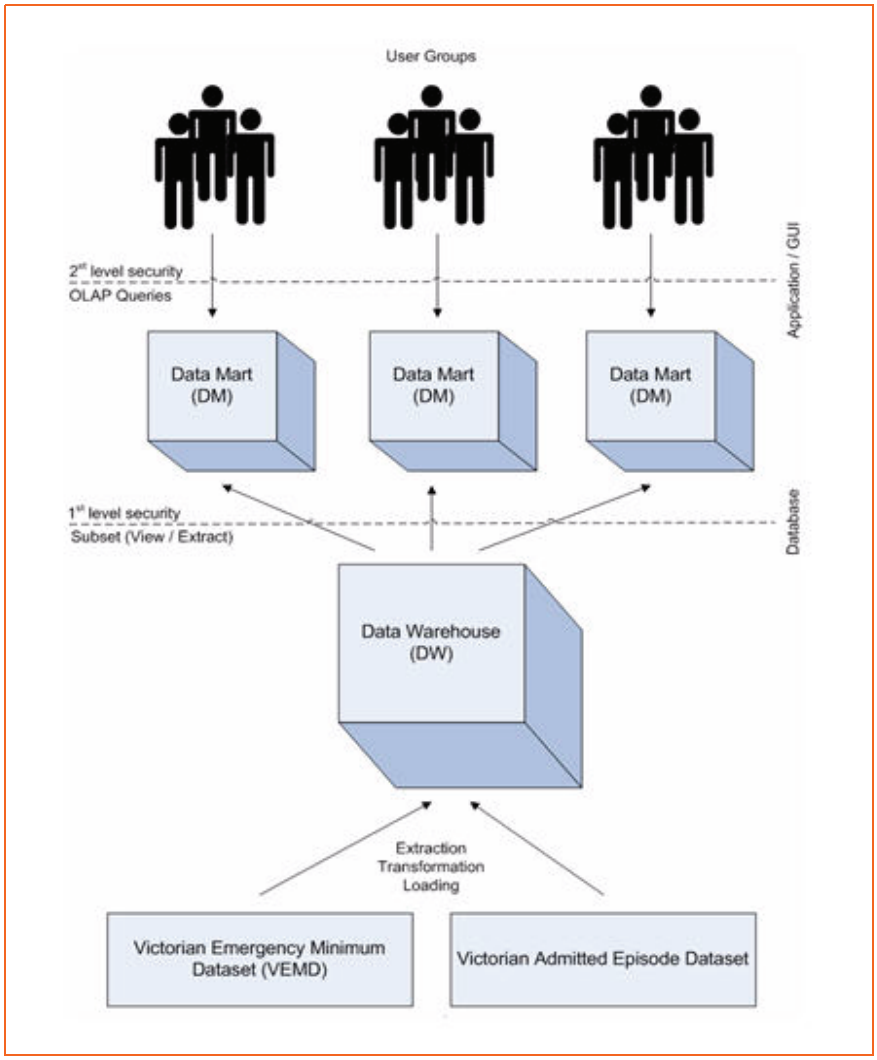

Figure 1. Data Warehouse and Virtual Data Marts, Multiple Users

For the purpose of this project, the VAED and the VEMD data sets were joined, such that various time measurements arising from common time-stamp attributes in both data sets were attained. One of the goals was to compare inpatient admission times in the VAED with the time-stamps that record the patients' treatment times while in the emergency room (VEMD). Valuable information could now be extracted, such as patient waiting times in the emergency room as well as during the transmission from ER to a ward. We would like to note that the data sets used for this project are de-identified, which means that any attributes that could lead to a disclosure of individual patients' information, were removed.

The first challenge is in the design and development of the clinical data warehouse. Data Warehousing provides solutions to integrate multiple data sources into a central point of storage, offering various views and summarisations at different levels of data granularity, and enabling users to perform complex data analysis queries on the data (see Figure 1)[3]. In this way, health information can be quickly analysed in respect to different patient patterns, symptoms, and demographic locations, to name a few. But also efficiency concerns of health institutions, such as hospitals and care centres, can be analysed. Valuable measures such as patients' waiting times, or ratios that denote the effectiveness of specific wards within a hospital can be calculated, and subsequent decisions can be made upon the results.

The second challenge is to ensure that the warehoused data is adequately protected against unauthorised access and privacy attacks. An example of this is as follows. An adversary issues specific queries against the data warehouse that will possibly reveal deterministic results about individuals, or increase the possibility of doing so. Such an attack can easily be performed by combining certain attributes in queries and analysing the diversity of the resulting data. As such attacks are subject to individual queries, protection mechanisms must be in place that prevent such attacks at query-time. The

reason for this is because we don't know what queries a possible adversary will issue on run-time, or what their intentions are.

The rest of this paper is organised as follows. Section 2 discusses characteristics and issues in storing and publishing health information. Section 3 explains how data warehousing contributes as a core component in clinical decision support systems and focuses on access control and end user profiles of the clinical data warehouse. Section 4 explains how our query-time access control scheme is applied. Section 5 introduces the clinical data warehouse decision support system prototype. Finally, section 6 provides an analysis of our implementation and methodologies.

2. Characteristics and issues in storing and publishing health information

Health data has the following main characteristics, (i) different levels of granularity, (ii) different levels of quality (accuracy, correctness, currency, completeness and relevance), (iii) contains sensitive information.

2.1. Granularity

The notion of data granularity is closely related to the way that the data was collected at the respective sources. Healthcare institutions with advanced information systems technologies are able to record measures such as time much more accurately than if the data is manually entered into a file or basic administrative system. Analogously, some healthcare professionals and medical staff may choose to record patient diagnostics in a very detailed manner, whereas others document only the most significant information. When these data sources that contain different levels of detail (granularity) are to be summarised, it is evident that there will be issues as what to do about the difference in data detail.

2.1.1. Multi-dimensionality of data warehouses

A feature of data warehousing that addresses the issue of granularity, is the concept of dimensions (see Figure 2). Dimensions can be thought of as the sides or the edges of a multidimensional data cube, representing different categories, such as injury type, ward type, or demographic location. This enables the decision support system to provide manifold views and query-options to the user, so that the underlying data can be analysed from different perspectives and angles.

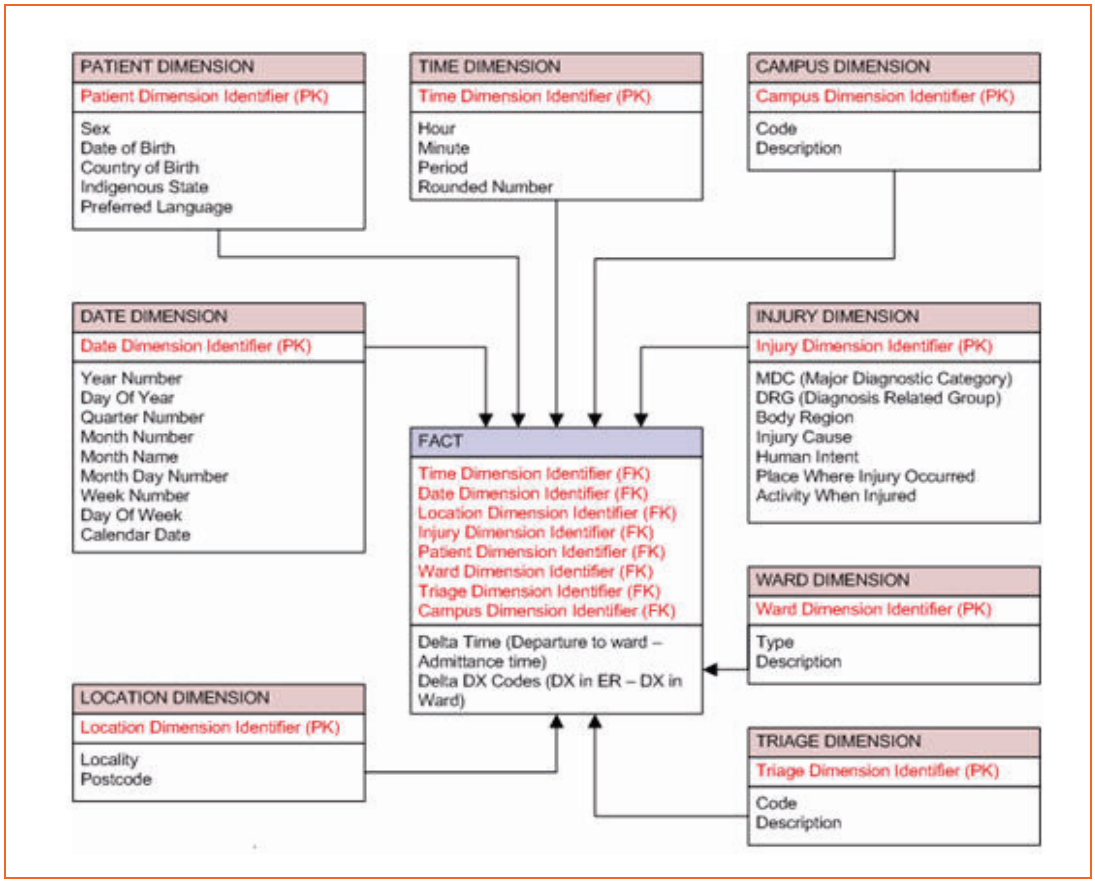

Figure 2. Star Schema, Data Warehouse Design Process

A data warehouse can have an arbitrary number of dimensions, some of which store the same kind of information, but for different levels of detail, or granularity. Varying granularity levels within the data (e.g. time) can also be stored in one and the same dimension. In this case, however, missing detail in some of the data sources may lead to incomplete query results.

A single dimension can hold more than one level of granularity. This is achieved by storing the same data at different degrees of aggregation in the same dimension. For example, the age group dimension has a finest grain of detail where an individual age (e.g. 50) is represented by one record. An aggregate thereof is the age group 45-50, which the previous, finer, granularity falls into. This approach of storing aggregates of data is repeated until a sufficient level of aggregation in a dimension is reached. The benefits of several granularity levels in one dimension is that the user can drill-down into the data from a coarse grain of detail (e.g. age group 20-65) or roll-up from fine grained aggregated data till more summarized details.

2.2. Data Quality

Differences in data quality generally occur for similar reasons as the previous paragraph stated, but it also includes the possibility of human error in the data entry. If modern information systems are used, certain measures such as timestamps are automatically recorded by information systems. Other information such as patient age, location of accident, or main injury, however, is often subject to error when entered into a system.

A good example for this is an emergency room. At busy times, healthcare staff have more important things to do rather than dealing with a computer system and the correct entry of data values. Caring for a seriously injured patient will always have higher priority than making sure that the patient's correct age is entered into the system. It is for this reason that errors are naturally introduced into the data, and it is the task of the developer to prepare the system how to deal with these errors, not the other way around.

Numerous data quality measures exist that are to be taken into consideration when integrating and analysing healthcare data. A difficult issue in the area of health information is to measure data quality, and decide which impact on the data analysis the errors will make [4][5]. Although consequences of this issue are addressed in this section, it is beyond the scope of this paper, and will not be elaborated on further.

2.2.1. Data cleansing

Data cleansing is the process of correcting the raw data that comes from the initial data sources, before it is loaded into the data warehouse. This step is so important in regards to health information, because it will have a major impact on the quality and accuracy of the resulting queries, and consequent decisions upon the results. Hereby, it is important to rectify the data in such a way that it does not lose its detail, but follows standards in terms of data format and value. Particularly missing or incomplete data must be treated with care, because a missing data record should not simply be discarded, but also taken into consideration by data analysis. Such data records can be marked as incomplete, so that the query engine identifies them as a record, but treats their values and significance in a separate way.

2.3. Sensitive information

Health information systems store personal information about real persons. From the data, a person can be uniquely identified by so-called quasi-identifiers such as name, date of birth, gender, and postcode. Other quasi-identifiers can be used to determine a person if the above mentioned ones do not suffice. Together with this personal information, health information systems also store medical diagnostics, such as disease codes and descriptions, and medical treatments. These parts of the stored information are called sensitive values, as they reveal additional information about a person that is confidential, and is not supposed to be known by anyone other than the person themselves and a trusted healthcare professional, or medical staff [6].

Deterministic person attributes such as name, and health insurance number are generally removed from the data records, such that a person cannot be re-identified directly from looking at the data. However, the quasi-identifier attributes that remain in the data, such as age, gender, postcode, to name a few, can be used to uniquely re-identify a person in some cases. In addition to that, the sensitive information of that person is then known as well. With most health decision support systems being multi-user systems, there are a number of different users, or user-groups, that have access to the warehoused information. In order to prevent certain groups from accessing data that may lead to the re-identification of individual persons, access control techniques are required that protect the sensitive information of individuals from unauthorised access. Although certain attributes can be removed or modified prior to the warehousing process, quasi-identifiers remain, and it can only be determined on query-time, whether a certain set of attributes can be accessed by a particular system user. This is an issue in the area of privacy protection and access control in databases and data warehouses, and will be addressed in detail later in this paper.

2.3.1. Privacy protection

Protecting sensitive information in a clinical data warehouse-driven decision support system is paramount [7]. Access control models have been developed that restrict unauthorised access to data, by defining privacy levels to portions of the data as well as to user groups. More recent works

have focused on query-time validation of privacy rules, such as mentioned above when discussing quasi-identifiers. Not only does health information contain several attributes that can be used to re-identify a person, it also stores several attributes that give insight in to the medical condition of the person. Treating the sensitive information as an atomic value is one solution to the problem of privacy protection. However, it ignores the fact that diagnostics codes and descriptions can be subsets of each other and identify diseases to different levels of detail. This paper addresses issues involving the access of sensitive information, and how the re-identification of patients can be prevented for certain user groups depending on the nature of the query. The second part, which concentrates on more advanced privacy protection of sensitive information, is beyond the scope of this paper and is suggested as future work.

3. Data Warehouse (DW) as component of a clinical decision support system

As the paper has explained in previous sections, a clinical data warehouse stores integrated domain oriented health data in a multidimensional structure. This allows for quick and complex querying, and for the extraction of valuable information.

Equipped with these features, a clinical data warehouse offers the predestined back-end, or core component, of a clinical decision support system. Acting as the interface between the underlying data and the end-user, the decision support system must offer easy means for the user to access the data source, to design and execute queries, and to provide ways to view and read the result data for analysis. An important consideration is that the requirements for a data warehouse and decision support system vary, whether they are being used on the state level, or on the local level. For this project, we have centred upon the state level, however, the approach also functions appropriately at a more specific hospital level.

In healthcare, there are several important parts of information that are of interest when analysing and mining patient records. These are quality measures in regards to the healthcare processes that are performed on the patients, such as waiting times for the patient to be seen by a nurse or doctor or measures that capture overall efficiency and effectiveness of hospitals and health care centres. For example, for the purpose of determining time differences between the emergency room visit and the admittance to a hospital ward, data sets are linked, and the resulting deltas are calculated.

For these types of information to be discovered, it is necessary to equip the decision support system with the correct functions. Further, it is of high concern to protect the privacy of the persons whose data is stored in the data warehouse. The following sections will now give an overview over user groups for clinical decision support systems, and present the developed prototype system.

3.1. Data

Health data is generally available at different levels of granularity, completeness, and accuracy. Moreover, new data is continuously being added to the data sets. The procedures of data cleansing have been discussed in previous sections, therefore, this section shall now elaborate on the data representation. We hereby refer to two prominent datasets that are maintained by the Victorian department of human services. These are the VAED (Victorian Admitted Episode Dataset), and the VEMD (Victorian Emergency Minimum Dataset), which store patient information when admitted to a ward at a hospital or care centre, and when visiting an emergency room.

Nearly all of the information that is stored in the patient records in these datasets is encoded using alphanumeric codes, and the corresponding descriptors and additional information are stored in lookup tables. Timestamps are stored in a very precise format, but this does not guarantee that the time-stamp has been recorded precisely. Some attributes, such as diagnosis and procedures, appear up to 8 times each within a patient record. This means that such attributes must be either stored in an array or multi-value attributes, or the record must be duplicated for each different attribute value.

3.2. End users

This section addresses the identification of end-users for clinical data warehouse systems [8]. In health care, there are four major information end users: (i) clinicians, (ii) financial personnel, (iii) researchers, and (iv) educators. Strictly speaking, financial personnel is further to be sub-classified into the groups of healthcare service planners, and government funders. Clinicians perform clinical evaluation and management of patients and are therefore mainly interested in the actual diagnostics and procedures that are associated with patients of different ages, origins, etc. Financial personnel deal with the evaluation and management of the economics of health care. State wide performance records of hospitals and care centres are of interest for this user group, as it aims to identify profitable and under-performing institutions, and to discover unused resources. Researchers are interested in the discovery, integration, and application of new knowledge and to propose new practices and policies. Finally, educators provide new knowledge to experts and participants and must therefore have access to a broad part of the data warehouse.

In addition to these end users, there is another user group that has interest in accessing health data: the public. Particularly for this user group, proper enforcement of privacy constraints and patient re-identification control are necessary.

3.3. Privacy Policies

The above described user groups have respective interest in parts of the

data warehouse. Moreover, they have different levels of access authorization for particular attributes. A simple yet effective way to ensure such authentication constraints is to map the user groups' goals and interests in the data against the model, i.e. against the dimensions and attributes. In this way it can be guaranteed that users of a certain user group only will have access to their respective view of the data.

While separate views of the data allow for static access control for user groups, this method does not suffice in order to protect the data from attacks that are performed once the adversary has successfully been granted access to the data warehouse. This means that a further specification of access levels is necessary, both for the user groups and for the data itself. The underlying idea of this concept is to associate each user group with a user security level, and also associate model security levels with dimensions, facts, and attributes. Also, combinations of model security levels are now possible, which can be used to validate access at the time when a query is issued.

4. Sensitive information and privacy protection

Although most of the identifying information such as patient name, date of birth, medicare number, has been removed from the data records, there still remain a number of quasi-identifying attributes[1]. For the VEMD these are country of birth, age, postcode, and gender. Research has shown that a majority of patients can be re-identified by the three quasi-identifiers age, postcode, and gender [9]. Associated to these data items, there is a number of diagnostics information stored in each data row. These are primary and other diagnostics codes, as well as procedures, and type of visit. For our case study, we warehoused the most relevant attributes, and ended up with a total of 16 dimensions.

The concept of privacy protection, and particularly anonymisation [11], has been the one of the main research focuses of this project. For the reason of protecting re-identification of patient details, our system allows for custom specification of security levels that apply to individual attributes, as well as our novel concept of composite security levels, which protect the data against queries containing particular quasi-identifiers. Composite security levels (CSL) are constraints that apply to a certain set of attributes, and prevent users from accessing these attributes together. CSLs also have an associated security level and apply on query-time, as the queries that the user performs are not known before hand. For this reason, users and user groups are associated with user profiles, such that their security levels (the degree to which they can access the data) can be matched against security levels of attributes, and against composite security levels.

The privacy model that we apply in this project separates static access control levels from dynamic ones, whereby the former is achieved by creating user views of the data, and the latter one is realized by combined security levels (CSL). A CSL applies when the user issues a query that contains dimensions, attributes, and/or facts that all belong to a pre-defined CSL. In other words, if the user attempts to access parts of the data warehouse that all belong to a CSL, then this query must be validated before proceeding to the database engine. As every CSL must be associated with a level of access (a numerical value), it can quickly be compared against the security level associated with the user who is issuing the query. If the user's security level is greater than or equal to the CSL's security level, then the query is successfully validated, and can be processed.

5. Clinical DW Decision Support System

This section describes and illustrates how the developed system addresses issues of storing, warehousing and publishing health information. The following features of the system are hereby addressed: user profiles configuration, security levels, composite security levels, query generator, filters, graphing.

The prototype described in this paper has been developed by Anders H. Landberg at the Department of Computer Science at LaTrobe University as part of an on-going collaboration in health informatics between the Department of Computer Science and the Department of Health Information Management. One of the goals of the project is to provide means of access to health emergency data, while considering implications and constraints during the process. Some screen shots that illustrate the user interface of the prototype system can be found in Figure 3.

1. attributes that can be combined, and this combination leads to re-identification of a person in a data record, are often referred to as quasi-identifying (QI) attributes in the literature

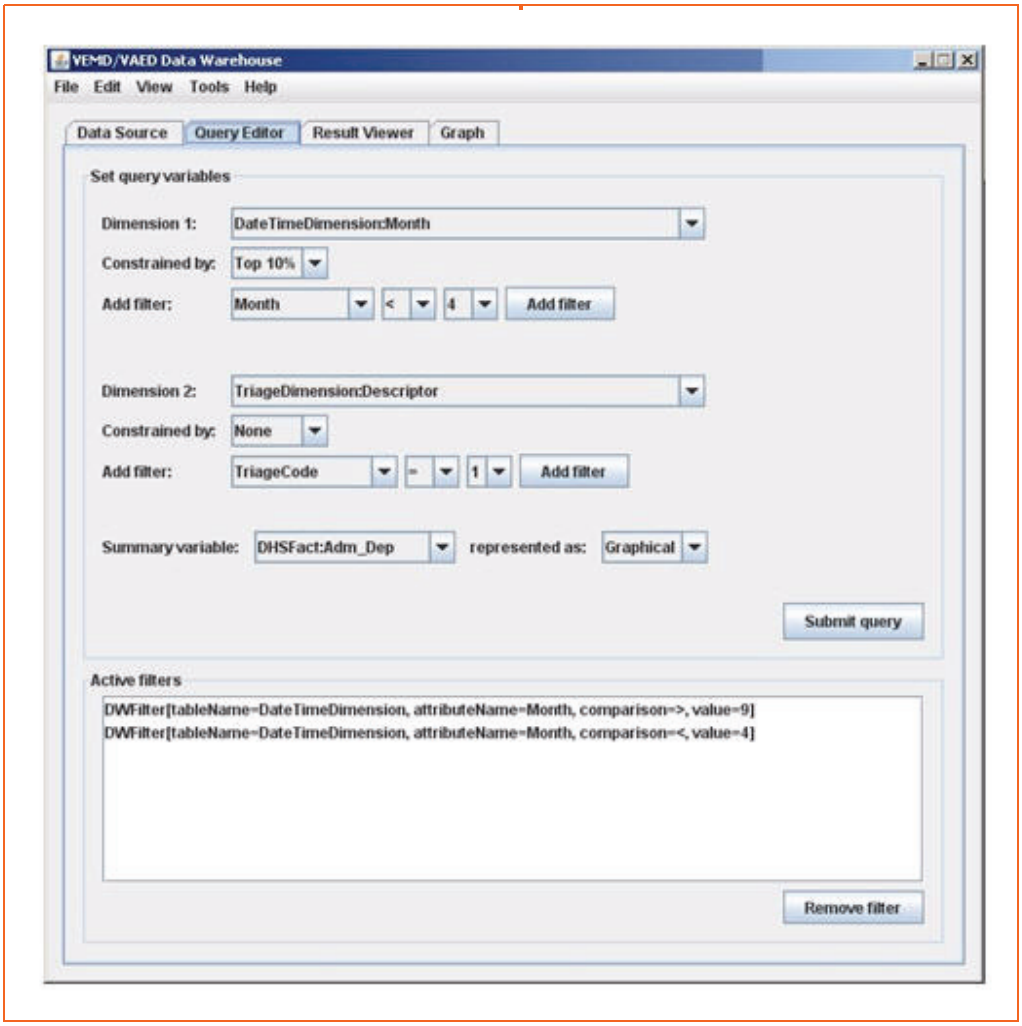

Figure 3. Query Editor

5.1. Granularity and multi-dimensionality

The time dimension of the data warehouse represents the most fine-grained dimension, as it is used to evaluate various timestamps that are recorded at critical points, for example when a seriously injured patient that visits an emergency room is first seen by a doctor. Other levels of granularity that are of interest are periods of the day, such as morning, afternoon, and evening. These time intervals are so important as they denote parts of the day when people are more or less active, and when it may be more or less necessary to allow more resources. An emergency room will usually be most busy on weekends and in particular in the evenings and nights, whereas a post-surgery ward will experience its peak time during mornings.

Diagnostic information is represented by a classification that groups disease and treatment provided called the DRG Diagnosis related group. These groups can be aggregated together to represent Major Diagnostic Categories (MDCs). Diagnostic related groups and major diagnostic categories are warehoused in such a way that drill-down on MDCs is possible in order to view more fine-grained diagnostics values that are generally stored using DRGs.

To address the concept of different levels of granularity, the system allows for an arbitrary number of dimensions and the definition of drill-down and roll-up dimensions. Granularity levels can be defined as finer or larger grain of each other, and in this way form a vertical hierarchy that allows the user to drill down on a particular attribute within a dimension. In the case study conducted during this project, we implemented the time dimension as a drill-down and roll-up dimension. OLAP queries are therefore possible for 'year' attribute, and subsequent drill-down options on months, and days within that year. Vice versa, the roll-up function performs the opposite functionality.

Not only does the system support multiple dimensions and levels of granularity, it also caters for multiple

fact tables and cross fact querying. In this way several fact tables can be used to warehouse different parts of the domain, or to capture different levels of granularity. Referring back to the section on end users, we believe that this feature can be very useful, as different user groups will be interested in different parts of the data. The dimensions, however, can still be shared among various fact tables, such that no supplication of look-up tables and dimensions is necessary.

5.2. Data quality and data cleansing

The most frequent data quality issue is incompleteness and inaccuracy of recorded health data. Incomplete data is marked with appropriate 'miss-codes', which enables the user of the data warehouse to analyse the data in regards to these data gaps. It is particularly interesting to find out why such data gaps exist and under which circumstances they occur. Inaccurate data entries can not always be spotted, and in our system they are treated just as any other data values. However, timestamps such as 12:00:00 or 02:30:00 often seem very unlikely, especially if they occur frequently. Therefore, too accurate timestamps can be used to identify inaccurate data entry, and to discover possible lack of resources in particular areas.

5.3. Composite security levels

A composite security level is not the means of identifying sensitive information or to de-identify it in some way. Numerous works in the area of de-identification and privacy protecting techniques have been proposed to solve this problem [9][11][12]. A CSL is rather the concept and mechanism that can be implemented to enforce that sensitive information is exposed to the public in a secure way. In simple terms, our approach accomplishes this by locking certain combinations of data fields and giving these combinations a level of access that can only be used by authorized users. This means that if the combinations are incorrectly identified beforehand, then this negative effect will also be reflected upon the CSL mechanism.

Figure 4 shows a sample star schema with four dimensions and a fact table, and the attributes in each dimension that are protected by the CSL. When a query is executed that accesses all these attributes in combination, then the CSL applies and it is validated against the accessing entity's security level.

Approaches such as Anatomy [12], l-diversity [9] and anonymisation [11] are used to identify and arrange combinations, whereas CSLs are used to ultimately enforce these.

A common misconception of the approach is that separate querying of the data and aggregation of the results will still lead to disclosure of sensitive information and thus not protect privacy. This is incorrect for the case when the data has been insufficiently de-identified and diversified, i.e. when anonymisation approaches as mentioned above have not been applied appropriately.

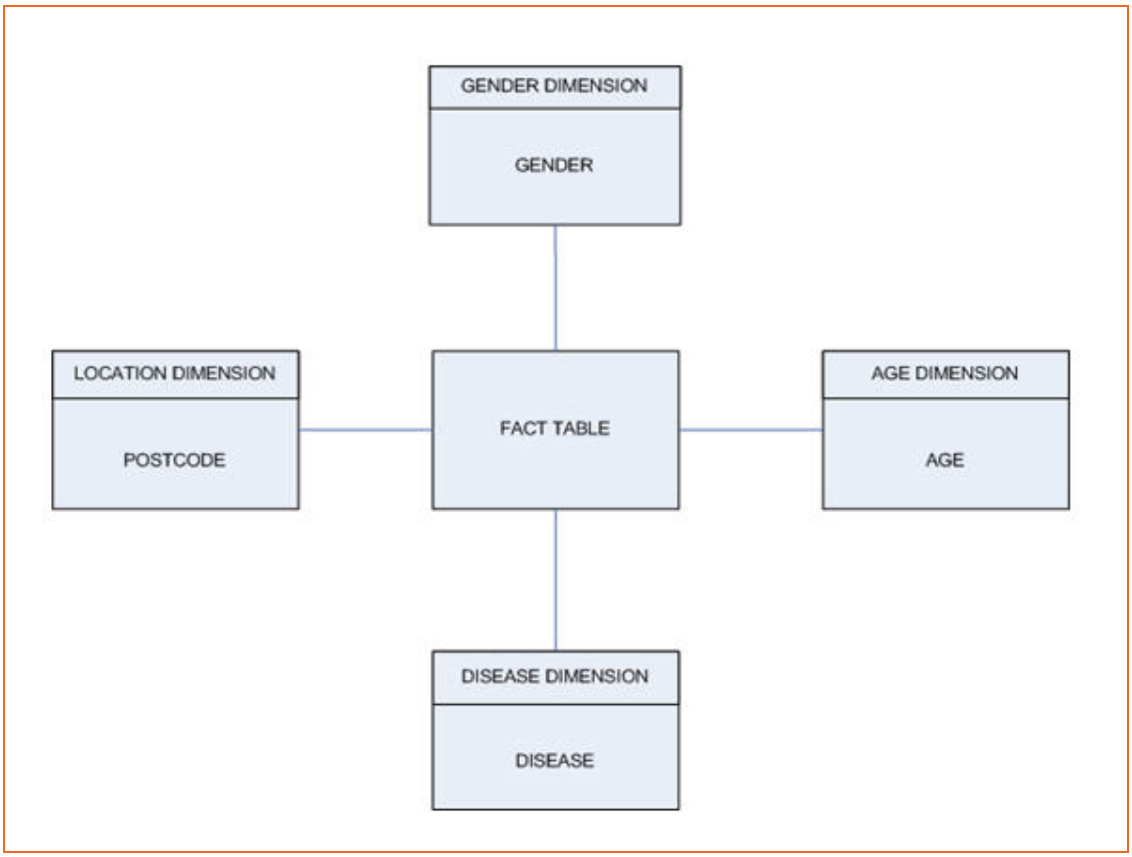

Figure 4. CSL protects combination of attributes against deterministic query

The following example explains how a single query against a combination of secured attributes, and three separate queries against the same data with aggregation will produce different results. The tables below show sample result sets from the partial queries, and Listing 1 shows a sample OLAP query that triggers a CSL as result of querying these attributes. Further, the column disclosure percentage indicates the cardinality in a group/quadrant and thus the possibility that a record in the group is deterministic.

Knowledge of user: Bob, 50 lives has postcode 3000 – disease unknown. This is the information that a possible adversary has, and which they want to use to identify Bob in the database.

CSL participation attributes: Gender, Postcode, Age, Disease(*) sensitive attribute. The goal of the adversary is to determine Bob's disease, by querying the data in such a way that it will produce a deterministic result.

Partial query 1: grouped by Gender and Postcode

Gender	Postcode	(Age)	Disease	Disclosure %
M	3000	(48)	Heart	25
M	3000	(49)	Kidney	25
M	**3000**	**(50)**	**Liver**	**25**
M	3000	(51)	Lung	25

Partial query 2: grouped by Gender and Age

Gender	(Postcode	Age	Disease	Disclosure %
M	(3002)	50	Heart	25
M	(3001)	50	Kidney	25
M	**(3000)**	**50**	**Liver**	**25**
M	(3000)	50	Lung	25

Partial query 3: Postcode grouped by Age

(Gender)	Postcode	Age	Disease	Disclosure %
M	3000	50	Heart	25
M	**3000**	**50**	**Liver**	**25**
F	3000	50	Kidney	25
F	3000	50	Lung	25

Deterministic query: grouped by Gender, Postcode and Age

Gender	Postcode	Age	Disease	Disclosure %
M	3000	48	Heart	0
M	3000	49	Kidney	0
M	**3000**	**50**	**Liver**	**100**
M	3000	51	Lung	0

Partial queries on the CSL participating attributes results in an approximation of matching a possible results row against the known individual. In our example there is a 25% chance that an adversary will disclose the record of Bob, 50. However, when querying a full combination of the attributes, Bob's identity is disclosed and the adversary can identify Bob's record in the database and gain the sensitive information.

Further, results sets generally are returned in sorted order by the selected attributes. In each of the partial queries, one attribute is left out, and hence the result sets are sorted differently every time.

Thus, re-constructing a deterministic data-tuple from partial query results will prove impossible when CSLs are in place that protect from accessing certain sensitive attributes in combination.

5.4. Defining and enforcing composite security levels

Applying the concept of composite security levels is simple. The DW schema is defined in an XML document and allows two special tags, namely <CSL> and <CSLid>. The first tag specifies the level of the CSL that is validated against the accessing entity's security level if the CSL applies during a query. The second tag specifies a unique identifier that

groups all participating attributes that belong to a CSL. All attributes that belong to the same CSL must specify the same values for the CSL and CSLid tags.

A sample OLAP query that triggers CSL_ID1 is given in Listing 1. As can be seen in the attributes selected, all participating attributes of the CSL are contained in the query. By the nature of the user interface of our application, the user can issue any ad-hoc query and hence such combined security levels cannot be determined in advance. This example clearly illustrates the necessity of run-time evaluation of CSLs.

To validate a query as given in Listing 1, three major steps are performed. First, individual security levels that apply to the selected query attributes are verified. If this step fails, the query aborts. Otherwise, all selected attributes are checked for the CSL tags. For all distinct CSLid's that are identified among the selected nodes, it is checked if for each of these CSLid's, all participating nodes are found in the query's selection. If the query contains all participating attributes of any CSL, then the (numerical) security level of the identified CSL(s) is verified against the accessing entities' security level. This step is repeated for all applying CSLs. If this step fails, the query is aborted. Otherwise, the query is successfully validated and forwarded to the database.

```
SELECT * FROM
  (
  SELECT LocationDimension.Postcode
       , AgeGroupDimension.Age
       , GenderDimension.Gender
       , DiseaseDimension.Disease
       , MEDIAN( ClinicalDWFact.WaitingTime1 ) AS MEDIAN_WAITING_TIME
      , CUME_DIST OVER ( PARTITION BY AgeGroupDimension.AgeGroupCode
         ORDER BY MEDIAN( ClinicalDWFact.WaitingTime1 ) DESC ) AS
Constr1,

  FROM LocationDimension
       , AgeGroupDimension
       , GenderDimension
       , DiseaseDimension
       , ClinicalDWFact

  WHERE LocationDimension.LocationCode =
      ClinicalDWFact.LocationCode
  AND AgeGroupDimension.AgeGroupCode =
      ClinicalDWFact.AgeGroupCode
  AND GenderDimension.GenderCode =
      ClinicalDWFact.GenderCode
  AND DiseaseDimension.DiseaseCode =
      ClinicalDWFact.DiseaseCode

  GROUP BY CUBE ( LocationDimension.LocationCode
       , AgeGroupDimension.AgeGroupCode
       , GenderDimension.GenderCode
       , DiseaseDimension.DiseaseCode )

  )
  WHERE Constr1 >= 0.9;
```

Listing 1. OLAP query that will trigger the CSL_ID1

6. Analysis

This section discusses the results of the project and points out methods and techniques that were used during the research and development of the prototype. Also, we have included an illustration that presents the performance analysis of the composite security levels feature.

6.1. Results of implementation

As the original, raw health data was not accessible at the time of the project, synthesized data was used instead. However, the format and notation was kept exactly the same to best replicate the health data for the warehouse. Additionally, the initial system prototype was implemented as a single user system. However, by manually swapping the user profile definition files, a multi-user system was imitated. Of course, concurrent operations were not captured by this, but the general idea of having multiple user roles with different security levels was completely imitated.

During the use of the system and experimental evaluation, it was found that a powerful filtering functionality is of crucial necessity to be able to refine query results against the data warehouse in a quick and easy manner. An initial feature called 'constraint' was slightly modified to suit the purpose of providing these filters. By using filters, certain attribute values or ranges thereof can be included or excluded by the query, or, numerical ranges can be limited with inequality operators (see Figure 3).

It was also found that the CSL feature can be used for extended constraints on the data, which were not subject to privacy protection. For example, the data in the data warehouse can be partitioned by composite security levels, such that each partition has its own CSL. Then, these partitions can be mapped to user profiles by applying security levels to user profile definitions appropriately. Although a simplified version of this feature can be implemented using basic security levels, the CSL concept caters for the access of combined elements, and is therefore much more flexible, because different combinations of attributes can have different CSL security levels.

6.2. Methodologies

During design and implementation of the system prototype, three main aspects were covered with particular focus. These were (i) data warehouse structure, (ii) user profiles, and (iii) data privacy. To address each of these aspects, respective methodologies and concepts were applied. These were (i) the data warehouse architecture according to Inmon, which uses a top-down approach for designing the data warehouse, (ii) user profiles with security levels, and (iii) Composite Security Levels (CSL) for the enforcement of data privacy and access control.

Inmon data warehouse architecture. We chose the architecture proposed by Inmon [10], as it suggests the data warehouse to be a centralised repository for the entire domain with attached dimension tables. For our prototype, we chose a single fact table that stored a number of different timestamp deltas, i.e. time intervals between two timestamps, as well as total number of health data records. Inmon also suggests maintaining data marts that represent the different aspects of the data from an enterprise perspective. In our project however, we chose not to create separate data marts, but rather creating dynamic views of the data according to the user profiles. This means that when a user starts the system under a certain user profile definition, only the parts of the data they are authorised to see are actually available. The advantage of this highly centralised design is that consistency is maintained at all times, regardless of how many 'virtual' data marts or views exist. Also, from a performance point of view, our design choice is justifiable. Re-creating views of data every time a user runs a query is nearly not feasible. But this would be necessary to enforce composite security levels using views. Therefore, our solution maintains a high level of flexibility while performing not worse than if physical data marts were created.

User profiles with security levels. In order to enforce access control on the data in the data warehouse, it is necessary to assign a level of access or, security level to each user. This is done by user profile definition files, which keep information about the user or user group, as well as the security level (SL). This SL is then used to determine which data the user group has access to, and to match against possible composite security levels. By specifying SLs to user groups, virtual views or data marts are created, because certain attributes and values are not available to unauthorised users. If for example user Bob has security level 3, then he can only access data that is labeled with a security level of 3 or below. Also, Bob can only access attributes together in a single query that do not have a composite security level greater than 3.

Composite Security Levels (CSL). Patient re-identification is the process of mapping a set of attribute (values), the so-called quasi-identifiers from a database query against a known person profile, and then identifying, which of the resulting data records (if the result is greater than 1) matches the person. One process of preventing re-identification of personal data is anonymisation, whereby identifying attributes are removed from the data. However, this technique does not always suffice. It is important to prevent certain attributes to be accessed together, as their combination reveals information that leads to the disclosure of patient information.

For this reason, we are using a previously proposed method that is based on composite security levels (CSL) to enforce these extended privacy constraints. CSLs are defined as a group of attributes, i.e. age, postcode, and gender, and a separate security level, say 3. In order to access these three attributes in the same query, the accessing entity, i.e. the user, must satisfy the security level of 3 to execute the query. If the user's security level does not match the CSLs security level, then the user is restricted from the query. Using CSLs, we are able to model dynamic views of the data, complex privacy constraints, and other complex constraints that apply to the data in the data warehouse.

6.3. Future Work

Securely warehousing health data is a challenge that has been addressed in this paper and our proposed concepts and methodologies have been implemented to facilitate privacy protection in a DW that stores the Victorian Health Datasets VAED and VEMD. With the increasing interest in extracting new knowledge from this warehoused data, there is also high motivation in linking this data with other sources, such as education and employment data. Valuable demographic information can thus be extracted, and as a result of that, more sophisticated privacy protecting mechanisms are required.

As a continuation of this research work, we plan further extensions and adjustments to our model, such that linked data sources can also be supported by the privacy scheme.

References

1. T. J. Eggebraaten, J. W. Tenner, and J. C. Dubbels, "A health-care data model based on the hl7 reference information model," IBM Systems Journal, vol. 46, no. 1, pp. 5–18, 2007.

2. V. G. H. Information, "Information about health data standards and systems (hdss) used in victoria's hospitals," 2008. This is an electronic document. Date of publication: [2008]. Date retrieved: May 1, 2008. Date last modified: 29 April, 2008. http://www.health.vic.gov.au/hdss/index.htm.

3. T. R. Sahama and P. R. Croll, "A data warehouse architecture for clinical data warehousing," in ACSW '07: Proceedings of the fifth Australasian symposium on ACSW frontiers, (Darlinghurst, Australia, Australia), pp. 227–232, Australian Computer Society, Inc., 2007.

4. R. L. Leitheiser, "Data quality in health care data warehouse environments," in HICSS, 2001.

5. D. J. Berndt, J. W. Fisher, A. R. Hevner, and J. Studnicki, "Healthcare data warehousing and quality assurance," IEEE Computer, vol. 34, no. 12, pp. 56–65, 2001.

6. X. Zou, Y.-S. Dai, B. N. Doebbeling, and M. Qi, "Dependability and security in medical information system," in HCI (4), pp. 549–558, 2007.

7. J. Bhattacharya, S. K. Gupta, and B. Agrawal, "Protecting privacy of health information through privacy broker," in HICSS, 2006.

8. A. J. J. McLeod and J. G. Clark, "Identifying the user in healthcare information systems research," in HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences, (Washington, DC, USA), p. 141, IEEE Computer Society, 2007.

9. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ldiversity: Privacy beyond k-anonymity," in ICDE, p. 24, 2006.

10. W. H. Inmon, Building the data warehouse. Wellesley, MA, USA: QED Information Sciences, Inc., 1992.

11. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):571–588, 2002.

12. X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In VLDB, pages 139–150, 2006.

Correspondence

Eric Pardede, PhD
Department of Computer Science and Computer Engineering
La Trobe University
Bundoora Vic 3084, Australia

Phone: +61 (0)3 9479 3459
Fax: +61 (0)3 9479 3060

e.pardede@latrobe.edu.au