How To Satisfy EU Directive On Data Protection In a Data Warehouse?

Izidor Golob, Boštjan Brumen, Tatjana Welzer
University of Maribor
Faculty of Electrical Engineering and Computer Science
Smetanova 17, 2000 Maribor, Slovenia
e-mail: izidor.golob@uni-mb.si

ABSTRACT

The paper deals with meeting the required level of measures to satisfy the EU Directive on data protection in a data warehouse system. A data warehouse is an integrated and a time-varying collection of data from many diverse and heterogeneous sources, used primarily by business users.

1. Introduction

The development of the information age came about without conscious directions. As technology developed, the transition from industrial age came along as a handmaiden. Having a technological imperative – what can be developed is developed and implemented – business and society face today new business and ethical issues, such as information privacy. While contributing to economic and social progress, trade expansion and the well being of individuals, data processing systems were ultimately being designed to serve man. Therefore they must respect the fundamental rights and freedoms that include respect the right to privacy.

Privacy is one of the fundamental human rights and is therefore often discussed as though it were a moral or constitutional right. Clarke treats privacy as an interest, rather than as a right: *Privacy* is the interest that individuals have in sustaining a 'personal space', free from interference by other people and organizations [Clarke, 2000].

In 1995 the European Union passed Directive 95/46/EC "On the protection of individuals with regard to the processing of personal data and on the free movement of such data" (Directive) which is now implemented by all EU member states. The Directive is intended to protect individual privacy by prohibiting the improper collection, use and transfer of data relating to individuals.

Many aspects of the Directive appear, including free flaw of data within the EU and limitations in third-party data transfer. Having in mind global cooperation, the Directive is important for global enterprises for at least three reasons [Blackmer, 2002]. First, companies must comply with national laws based on the Directive in each country in which they have operations. Second, the Directive should make it easier to consolidate data processing operations and provide remote access to data within Europe. Third, and more ominously, the Directive raises the prospect of data protection authorities blocking or restricting data transfers to the non-EU countries not deemed to provide "adequate protection" for personal data.

In spite of treating privacy as a universal human right, different approaches how to deal with privacy problems appear worldwide. While EU approach favors the privacy protection using legal instruments, the US government on the other hand, for example, advocates industry self-regulation approach. Walczuch gives a possible explanation for these differences [Walczuch et al, 2001] in major cultural differences between national cultures within Europe and worldwide. In general, it seems there is no sense

for the negotiation of an international, technology-neutral, certifiable, management standard for the implementation of the information privacy principles that may be implemented by any public or private organization that collects, uses, processes and discloses personal information via the Internet, or through any other public or private network. However, according to Bernett, it does make sense to establish a common standard on privacy [Bennett, 2002].

Looking at technical world, today's information technology has improved very much in last decades. Large databases are easily affordable, using modern database technology. Data warehouses are integrated, time varying, subject oriented, non-volatile, integrated large collection of data [Inmon, 1996]. Data warehousing is today one of the hot topics in the research community and in the industry. Data warehouses with their concepts and tools provide an area for efficient storage and access of huge amounts of data that can be further explored, primarily by business users. The data warehousing and business intelligence market continues to outperform many other technology markets. In a typical data warehouse, data from a variety of systems is extracted, transformed, and cleansed, and business rules are developed to help clarify and standardize the data. Coupled with other tools, one is easily able to collect and integrate personal information into a data warehouse, from operational sources or Web.

In this paper, we focus on impacts of the Directive in a data warehouse while conforming the adequate level as the Directive requests. We identify problems at technical (data administration) level and give solution to those.

2. Problem

Data warehouses, especially coupled with data mining, open a new dimension in managing large collections of data. Using the concepts of data warehouses and available tools it is possible to build and manage large quantity of data from various sources. This implies that separate data (observations, facts) are integrated into one information.

While each EU member has it's own data protection law, the Directive serves as a basis for local laws. Our research will therefore enlighten the information privacy problem from the Directive perspective. First of all, there is a question what problems could arouse using data warehouse technology on personal data. For this reason, an identification of possible data privacy problem in usage of data warehouses with personal data is an important part of this research.

By linking and matching the identifiers of object's owners, one is able to identify an individual. This was, for example, shown in [Samarati et al, 2002], where data from commercial information providers were collected. The research showed that it is possible to identify an individual out of previous non-related data, even if they are de-identified.

Article 3 of the Directive describes what privacy data-processing mean. Personal data is defined as any information, which identifies or relates to a specific individual. Undoubtedly, personal data are data where names, addresses or day-of-birth are given. If attributes of that kind are missing, one can identify an individual by using additional data, by special tools and algorithms. The decision if one can identify persons out of a specific data set depends on the forecast of identification possibility.

A data warehouse is a special kind of a database and thus inherits all the properties of a typical database. In general, a data warehouse offers more and more precise data with broader time-horizon as it is intended to provide the owner with relevant data of high quality. As of today, data warehouses are the most important source of (business) information for decision-makers in the organizations.

If we omit the personal data for the database, we may miss a number of data warehouse objectives. In an extreme case, we miss all the objectives. Removing personal data has also an impact to information quality, as information quality is clearly radically degraded as information is not so complete anymore.

3. How to make a data warehouse the EU Directive friendly

Independently of the chosen architecture (centralized, distributed, federated) of a data warehouse, it is the best way to introduce a new layer just after ETL tools to do their job, regardless of a source of data (legacy operational systems, WWW, information service provider, digital library). Bellow we identify and analyse some operations that can be performed in this new layer:

a) DELETING personal data: the easiest approach, can neglect the data warehouse existence.

b) DE-IDENTIFYING. We define de-identifying as removing identifiers from data in such a way that specific data set cannot be identified again.

   By de-identifying data before loading them into a data warehouse, data loose their "personal" note and as such are not subject to the Directive anymore. As electronic medical records and medical data repositories get more common and widespread, the issue of making sensitive data anonymous becomes increasingly important. However, the possibility to identify a person still does not disappear. In spite of de-identified data it is theoretically possible to identify a group with a relatively small number of people, as we mentioned in the introduction, especially if people possess a specific, such as triples or similar.

c) PSEUDONYMIZING. We define pseudonymizing as replacing identifiers from data in such a way that specific relations among data sets, previously known, cannot be restored again later.
   This approach has some advantage over the de-identifying. It has an advantage to use the personal data by an authorized user, which can make use of identification. If the mappings being used reside in a data warehouse, it is possible to use them when pseudonymizing new coming data. This option does not decrease data quality in a way de-identifying does. Pseudonymized data can be compared to non-pseudonymized personal data when measuring "output power". Using this option, we keep the possibility to use personal data, but we limit the number of people that can actually achieve this. By revealing mappings and de-pseudomyzing data became personal data, which is subject to the Directive. As such, we need to use all the measures, required by law to protect the data.

d) LIMITING THE USE OF DATA. Mahler describes two ways to achieve this: manually or automatically. [Mahler, 2000]. Manually approach is the most effecting by use of Meta-Database Systems. Having it integrated in a data warehouse systems, a meta systems contains a catalog where users are given permissions to use specific data elements and thus limits users to access unauthorized data via browser. Using an automatic approach, a data warehouse system should be able to limit the use of personal data itself by checking the constraints.

e) LIMITING THE TIME HORIZON. A large time-horizon can be problematic if personal data still reside in a data warehouse. We have options to use meta-database system or a data warehouse-database as a limiting instrument, as described above in d). The operation (deleting, de-identifying) is performed on a front-end layer, not in a back-stage area.

f) LIMITING THE TOOLS FOR ADDITIONAL ANALYSIS. To protect privacy data one could also limit the usage of tools, as there is usually a number of tools that are provided and associated with data-warehouse management system. In most cases these are OLAP tools that provide a multidimensional view of data. According to the Directive, any processing of personal data is also subject to the Directive. Any processing or analytic operations are thus performed only, if they conform the law.

4. Conclusions and Future Work

People desire anonymity for a variety of reasons. Governments worldwide acknowledge this right, but use different instruments to guarantee the certain level of anonymity for their citizens.

The objective of this paper was to focus on problems in a data warehouse while meeting the requirements of the EU Directive on data privacy. We have identified the problems, discussed them and gave solutions to the problems by describing six different approaches in data administration.

In this article, we focused only on pure technical solutions within a context of a data warehouse privacy protection. Future work includes analysis the power of applying more sophisticated tools such as data mining into a data warehouse-

References

[1] Bennett Collin J. An International Standard for Privacy Protection: Objections to the Objections. 2000.

[2] Blackmer Scott. Briefing Report for the Privacy & American Business Meeting on Model Data Protection Contracts and Laws. Washington, D.C. 1998.

[3] Clarke Roger. Beyond the OECD Guidelines: Privacy Protection for the 21st Century, 2000.

[4] Inmon William H., Building the Data Warehouse (Second Edition) New York: Wiley, 1996.

[5] Mahler Thomas, "Die Zulässigkeit von Data Warehousing und Data Mining nach der EG Datenschutzrichtlinie und dem norwegischen Datenschutz-gesetz", Mjur Thesis, 2001.

[6] Samarati Pierangela and Sweeney Latanya. Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression. 1998.

[7] Walczuch Rita M. and Steeghs Lizette, "Implications of the new EU Directive on data protection for multinational corporations," Information Technology & People, Vol 14, No. 2, pp. 142-162, 2001.