Adopting privacy regulations in a data warehouse:
A case of the anonimity versus utility dilemma

Chaïm van Toledo[1] and Marco Spruit[1]
[1]*Institute of Information and Computing Sciences, Utrecht University, Princetonplein 5,Utrecht, the Netherlands*
Contact: m.r.spruit@uu.nl

Keywords: Privacy, k-anonimity, p-sensitivity, data warehouse, privacy enhancing technologies, ETL, having clause.

Abstract: This paper ìnvestigates how privacy can be protected in a data warehouse while, at the same time, an organisation tries to be as open as possible. First, we perform a literature review on relevant techniques and methods to preserve privacy and show that k-anonimity can be applied to comply with an organisation's requirements. Then, we propose two design strategies to adopt privacy regulations within a data warehouse. The first proposal shows that during the ETL process a data transformation can be performed to effectively realise anonimised records in a business intelligence environment. The second proposal shows that with views and a having clause, anonimisation can be arranged as well.

1 INTRODUCTION

National governments and international organisations like the European Union insist more and more that the privacy of individuals must be protected. Organisations store more and more large amounts of sensitive data of individuals, such as their income, medical conditions and so on in data warehouses (DWHs; *e.g.* (Spruit and Sacu, 2015)). These types of sensitive data may not be tracked back to individuals for the sake of privacy. On the other hand, organisations often want to analyse these data or even want to share it with the public, especially organisations financed from public funds (Kim *et al.*, 2014). This leads to the anonimity versus utility dilemma, the conflict of wanting to be as open as possible on the one hand and wanting the maximum protection of privacy on the other (Bezzi and Pazzaglia, 2009). Privacy and security solutions are adapted slowly by the DWH community (Kimball and Ross, 2011), while data usage and ownership have also become key topics in the emerging field of master data management (Spruit and Pietzka, 2015).

The Oxford English Dictionary has several definitions about what privacy is, this paper uses the very short definition: "protection from public knowledge or availability" (OED, 2015). For a DWH, this paper uses the definition of Inmon (2002, p. 31) that a DWH is "a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management's decisions". So in a DWH the individual's data must be protected from public knowledge or availability.

This paper explores the field of privacy in data warehousing with a case study. The following article tries to give an answer to the following research question: *How can privacy be protected in a DWH while the organisation tries to be as open as possible?* We investigate this main question through the following sub questions:

(1) What are the criteria to protect the privacy of individuals? (2) What methods and techniques are available to protect privacy in DWHs? (3) How to implement these techniques and methods in a DWH?

The paper is structured as follows. In section 2 we outline our design science research method to explain how the research was conducted. Section 3 summarise our extensive literature review on relevant methods and solutions for privacy preservation in DWHs, with their pros and cons. Section 4 test the different methods and solutions in our case study organisation. We conclude in section 5 by answering our research questions, summarising our findings and directions for further research.

2 METHOD

This case study tests different solutions for Utrecht University (UU). The UU is an educational and research institute in the Netherlands, according to the

Academic Ranking of World Universities (2015) the number 1 university in the Netherlands and the 56th in the world. The organisation runs their business intelligence environment on an Oracle database with SAP Business Objects to extract data, as shown in Figure 1. Their problem or dilemma is the ambition to be an open and transparent organisation, but with the guarantee to protect the privacy of its employees, students and other stakeholders. Some of the users of the DWH are authorised to see all the data, unauthorised users must only see the anonimised data. UU requires that only groups larger than ten are shown in the results to unauthorised users. This means that results or groups within results less than ten have to be anonimised, somehow. The threshold of ten is arbitrary, by the way, and was decided by UU based on a general rule of thumb.

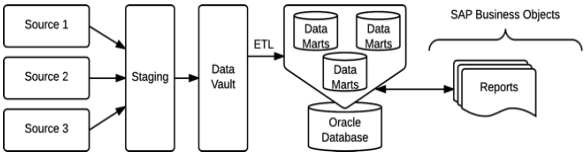

Figure 1: DWH situation at UU

This paper follows an action research method. Action research is defined as:

> "A participatory, democratic process concerned with developing practical knowing in the pursuit of worthwhile human purposes, grounded in a participatory worldview which we believe is emerging at this historical moment. It seeks to bring together action and reflection, theory and practice, in participation with others, in the pursuit of practical solutions to issues of pressing concern to people, and more generally the flourishing of individual persons and their communities" (Reason and Bradbury, 2001:2).

This means that this research is conducted in collaboration with the participants, namely the research systems and data management department of UU. It follows concepts of F, M and A by Checkland (Oates, 2005), where F stands for the framework of ideas, M for a problem-solving methodology and A for the area of application. The framework of ideas is elaborated within the literature review. Here the existing methods and techniques about privacy are scrutinised, derived from existing literature, but also with help from the ideas of the participants.

The problem-solving methodology is the part where the framework of ideas will be transported to the area of application, the organisation. Together with the participants, for each solution derived from the literature there will be assessed if, maybe in combination with other solutions, it is the right solution.

The literature is gathered with a systematic literature review. Hereby privacy in DWHs was examined. Techniques and methods that came out as a result were examined further.

3 LITERATURE REVIEW

This section elaborates on the current state of the art of privacy in DWHs. At first it defines what privacy is and what kind of problems concerning anonimity and privacy in data warehousing can be identified. The second part tells which kind of technologies and methods can solve these problems.

3.1 Privacy in DWH

Looking to statistical databases with data of the population, three types of data can be categorised: the identifier (like full names or social security numbers), quasi-identifier (like postal codes, age or ethnic background) and sensitive data (someone's diseases or income). The quasi-identifiers combined can be seen as an identifier (Sweeney, 2000).

The consequences when an individual's data are not protected is that "once information is released, it may be impossible to prevent misuse" (Clifton *et al.*, 2002:192). This release of an individual's sensitive data can happen in three possible ways (Article 29 Data Protection Working Party, 2014):

- *Singling out*, which means that individuals can be identified out of the data;
- *Linkability*, when two or more records linked with each other can identify an individual;
- *Inference*, when a dataset is sensitive to deductions, data can be traced to individuals because of the logical conclusions derived out of the deductions.

The last two, linkability and inference, can also be described as collective privacy, whereby multiple sources can identify the individuals out of data sets.

3.2 The solutions

Handling privacy in big DWHs can be very difficult and there are different solutions proposed which can be stored under the common name of privacy enhancing technologies (PETs) (Bezzi and Pazzaglia, 2009). It includes privacy management systems, privacy measurement anonimisation techniques, privacy preserving data mining, privacy-preserving

authentication and also protections directly for regular computer users, like the TOR-browser to getanonymously on the internet or adblockers to prevent being tracked (Federrath, 2005; Clifton *et al.*, 2002; Bezzi and Pazzaglia, 2009). Put simply, a lot of technologies focus on reducing personal information in data collections, it can anonimise, pseudonymise or hide data. Privacy measurements check to which degree the data are anonimised. Privacy measurements are metrics where the degree of privacy can be measured, this can be the degree of anonimity or how much diversity there is in the data set. Another classification came from Agrawal and Srikant (2000). They classify two techniques to protect sensitive information of individuals, namely query restriction and data perturbation. To compare these classifications with the methods of PETs, described Bezzi and Pazzaglia (2009), it is hard to put all the methods under the two classifications, for example, the measurement methods are not suited to be presented in query restriction nor in data perturbation. Figure 2 shows how the categories, metrics and methods are classified under the PETs.

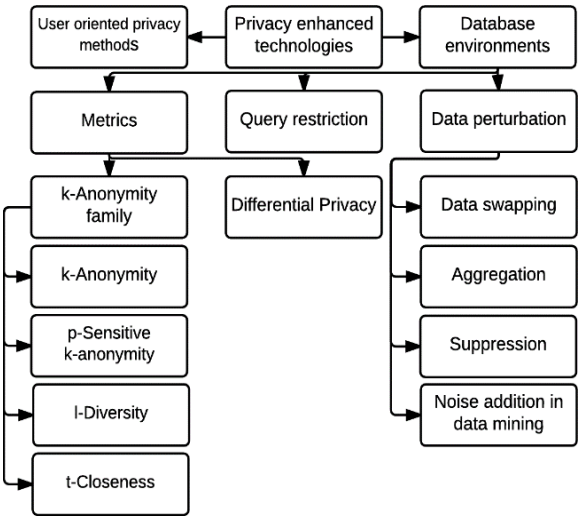

Figure 2: PETs with database solutions for privacy

3.2.1 Query restriction

Query restriction can be "restricting the size of a query result, controlling overlap amongst successful queries, keeping an audit trail of all answered queries and constantly checking for possible compromises, suppression of data cells of small sizes, and clustering entities into mutually exclusive atomic population" (Agrawal and Srikant, 2000:440). In other words, the system keeps track of the user queries and analyses them. Side effects are that there is a computational burden to the tracked queries and second with collusion attacks the query restriction can be bypassed (Domingo-Ferrer and Soria-Comas, 2014).

3.2.2 Data perturbation

Data perturbation can hold "swapping values between records, replacing the original database by a sample from the same distribution, adding noise to the values in the database, adding noise to the results of a query, and sampling the result of a query" (Agrawal and Srikant, 2000:440). Hence, data perturbation techniques anonimise the data. One problem with data perturbation is that information can be lost. Methods for data perturbation are data swapping, aggregation, suppression and noise addition in data mining (Sharma *et al.*, 2013):

- *Data swapping* "replace the original data set by another one where some original values belonging to a sensitive attribute are exchanged between them" (Sharma *et al.*, 2013:44). In this sense it is not sure if the data is part of the row or not. This technique can be helpful in development stages of the DWH.
- *Aggregation* generalises the data. For example, an exact birth day will become only a birth year.
- *Suppression* will delete or suppress certain records. This can be for example deleting or replacing the individual name with an asterix/*.
- *Noise addition in data mining* adds random numbers to numerical attributes: "Noise is added in a controlled way so as to maintain variance, co-variance and means of the attributes of a data set" (Sharma *et al.*, 2013:45). Likewise, the data swapping technique, this technique can also be helpful in the development process.

3.2.3 Privacy measurements

In this section four metrics will be elaborated. The first four are part of the k-anonimity family, the last one falls under the differential privacy category.

3.2.3.1 k-Anonimity

k-Anonimity is one of the most popular privacy protecting methods (Dankar and Al Ali, 2005). The definition of k-anonimity is: "A protected data set is said to satisfy k-anonimity for k > 1 if, for each combination of key attributes, at least k records exist in the data set sharing that combination" (Domingo-Ferrer and Torra, 2008:991). "The goal of k-anonimity is to only release data where for all possible queries, at least k results will be returned"

(Clifton *et al.*, 2002:201). To reach this goal, some results need to be generalised and suppressed for at least k records: "The aim is to hide every individual in a crowd of k look-alikes" (Dankar and Al Ali, 2005:572). The k stands for the minimum number of rows in a column.

Although k-anonimity is a popular method, it is not completely privacy attackable proof. One of the critiques is that "k-anonimity does not model sensitive information and attacker background knowledge" (Machanavajjhala *et al.*, 2008). Another critique is that it may fail to protect against attribute disclosure, by combining several characteristics the individual is still traceable (Domingo-Ferrer and Torra, 2008).

3.2.3.2 p-Sensitive k-anonimity

p-Sensitive is an adjustment of k-anonimity and its "purpose is to protect against attribute disclosure by requiring that there will be at least p different values for each confidential attribute within the records sharing a combination of key attributes" (Domingo-Ferrer and Torra, 2008:991). It works in combination with k-anonimity, whereby k has to be bigger than p. p-Sensitive ask for every sensitive attribute that there will be also other sensitive attributes with the same combination. Limitations are that there can be a loss of information, because to fulfil p-sensitive, sometimes the data rows needs to be fuzzier and coarsened.

3.2.3.3 l-Diversity

l-Diversity can be seen as an extension of k-anonimity. It tackles two problems of k-anonimity, first the discovery of sensitive data when there is little diversity and second the problem of background knowledge of an attacker (Machanavajjhala *et al.*, 2007). "The main idea behind l-diversity is the requirement that the values of the sensitive attributes are well-represented in each group" (Machanavajjhala *et al.*, 2008:5).

l-Diversity also has its limitations. Domingo-Ferrer and Torra (2008) identified two critique points. The first is that l-diversity can be difficult and also unnecessary to achieve in the current databases. The second is that it is insufficient to prevent disclosure of attributes, this is possible because of two attacks, namely the skewness attack and the similarity attack. The skewness attack means that with a certain set of records the chances can be estimated if some records can be applied to a person. Of course this means that an attacker must have certain background knowledge about the dataset. The similarity attack occurs when some records are semantically similar and therefore attribute disclosure can occur.

3.2.3.4 t-Closeness

Definition of t-closeness: "A data set is said to satisfy t-closeness if, for each group of records sharing a combination of key attributes, the distribution of the confidential attribute in the group and the distribution of the attribute in the whole data set is no more than a threshold t" (ENISA, 2015:31; Domingo-Ferrer and Torra, 2008:992). t-Closeness was created out of the shortcomings of l-diversity. It "requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table" (Li *et al.*, 2007:106). t-Closeness tries to solve the earlier presented skewness and similarity attacks of l-diversity. The skewness attacks can be solved because the "within-group distribution of confidential attributes is the same as the distribution of those attributes for the entire dataset" (Domingo-Ferrer and Torra, 2008:992). The similarity attack can be solved by "the within-group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset" (Domingo-Ferrer and Torra, 2008:992).

Although t-closeness has several ways to be checked, there is no computational procedure available to enforce this. Domingo-Ferrer and Torra (2008) state that if such a procedure was available, it would damage the utility of data.

3.2.3.5 Differential Privacy

Differential privacy "guarantees that the difference in the probability of an output between two data sets differing in just one element is at most a factor of e^{ε}" (Kerschbaum *et al.*, 2011). It falls under the randomisation techniques whereby noise is added beforehand.

4 PROPOSALS

This part will elaborate on a conceptual level how to implement privacy regulations in a DWH. To reflect back on the literature review, the metric of k-anonimity meets the requirements of the case study organisation, representing a group larger than 10, but only on categorical values. For ordinal data, such as an individual's income, the p-sensitive k-anonimity metric must be applied.

Then, how to implement anonimisation in a DWH? To implement this, two solution strategies are proposed. The first proposition is to anonimise the data in the ETL (Extract, Transform and Load) processes. The data can be stored in the same tables as the original, with an additional column to

distinguish the anonimised data with the real data, whereby with an authorisation method the right data can be queried.

The second method is simpler to implement and that is working with views in the Oracle database. The downside of this technique is that a lot of data can be 'lost' for the unauthorised user. This especially happens when the table grows horizontally, after all it is likely that the having clause detects less than the same rows when there are more quasi-identifiers. The following code below shows a view with three unions, each with a `SELECT` command. It is an example of people with a disease, stored with their postal code and their ethnicity. The result is that there is a 2-anonimity when the person is not allowed to see detailed results (`SELECT * FROM XXX WHERE allowed = 0`) and all the detailed results when the person is allowed (`SELECT * FROM XXX WHERE allowed = 1`).

The first `SELECT` command selects every row which complies to the 2-anonimity requirement. The second `SELECT` command selects the rest of the rows, but changes the values to an asterix (*). The third and last `SELECT` command selects just the detailed rows:

```
CREATE OR REPLACE FORCE
NONEDITIONABLE VIEW
"SYSTEM"."PEOPLE_DESEASES_VIEW"
("ID", "NAME", "POSTAL_CODE",
"ETHNICITY", "DESEASE", "ALLOWED")
AS SELECT ID, '*' as NAME,
   m1.POSTAL_CODE, m1. ETHNICITY,
   m1.DESEASE, 0 as allowed
FROM PEOPLE_DESEASES m1 JOIN (
  SELECT POSTAL_CODE, ETHNICITY
  FROM PEOPLE_DESEASES
  group by POSTAL_CODE, ETHNICITY
  having count(*) >= 2 ) m2
ON m2.POSTAL_CODE = m1.POSTAL_CODE
   AND m2.ETHNICITY = m1.ETHNICITY
UNION SELECT ID, '*' as NAME,'*' as
   POSTAL_CODE, '*' as ETHNICITY,
   m1.DESEASE, 0 AS allowed
FROM PEOPLE_DESEASES m1 JOIN (
  SELECT POSTAL_CODE, ETHNICITY
  FROM PEOPLE_DESEASES
  GROUP BY POSTAL_CODE, ETHNICITY
  having count(*) < 2 ) m2
ON m2.POSTAL_CODE = m1.POSTAL_CODE
   AND m2.ETHNICITY = m1.ETHNICITY
UNION SELECT ID, NAME, POSTAL_CODE,
   ETHNICITY, DESEASE, 1 as allowed
FROM PEOPLE_DESEASES;
```

With underlying views date ranges can be created for the sake of showing more information. But still, there is a disadvantage compared to the ETL proposal.

5 CONCLUSIONS

This paper investigated the following question: *"How can privacy be protected in a DWH while the organisation tries to be as open as possible?"* We addressed this main question by formulating and answering three sub questions.

The first sub question was: *'What are the criteria to protect the privacy of individuals?'* The literature review showed that three important features spring out to protect the privacy of inference, namely singling out, linkability and inference.

The second sub question was: *'What kind of methods and techniques are available to protect privacy in DWHs?'* Thereby the literature review showed the PETs, with three categories for the database environment: metrics, query restriction and data perturbation. At the metrics, we see that k-anonimity and sometimes k-anonimity p-sensitive is sufficient to comply with an organisation's requirements. With the data perturbation techniques, the aggregation and suppression are sufficient to modify the data.

The third sub question was: *'How to implement these techniques and methods in a DWH?'* This paper shows two proposals to comply with the organisation's need. The first is to adapt an ETL-process whereby the data will be modified to eventually comply to the metrics. The second proposal is to work with views in the database. Working with the having clause will also meet the k-anonimity requirement, but the problem with this method is that a lot of data can be lost for the unauthorised user. This can be solved by adding extra views whereby the data will be modified.

Back to the research question, this paper showed different kinds of technologies followed by two proposals on how privacy can be protected in the DWH at our case study organisation. Further research is now being prepared to investigate the application of algorithms for the ETL processes to transform data into anonimised data. Also further refinements are needed concerning the privacy regulations of organisations, to clarify what is necessary to protect the individual's privacy and what needs to be examined to accomplish that goal. Finally, to comply against linkability and inference, new features need to be added to the DWH.

REFERENCES

Academic Ranking of World Universities. (2015). *Academic Ranking of World Universities 2015*. Retrieved from ARWU World University Rankings 2015: http://www.shanghairanking.com/World-University-Rankings-2015/Netherlands.html

Agrawal, R., and Srikant, R. (2000). Privacy-preserving data mining. *ACM Sigmod Record, 2*(29), 439-450.

Article 29 Data Protection Working Party. (2014). *Opinion 05/2014 on Anonymization Techniques*. Retrieved February 23, 2016, from European Commission: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

Bezzi, M., and Pazzaglia, J. (2009). The anonymity vs. utility dilemma. ISSE 2008 Securing Electronic Business Processes: Highlights of the Information Security Solutions Europe 2008 Conference, 99-107.

Clifton, C., Kantarcioglu, M., and Vaidya, J. (2002). Defining privacy for data mining. *National Science Foundation Workshop on Next Generation Data Mining*, 199-207.

Dankar, F. K., and Al Ali, R. (2005). A theoretical multi-Level privacy protection framework for biomedical data warehouses. *Procedia Computer Science, 63*, 569-574.

Domingo-Ferrer, J., and Soria-Comas, J. (2014). Data anonymization. *Risks and Security of Internet and Systems*, 267-271.

Domingo-Ferrer, J., and Torra, V. (2008). A critique of k-anonymity and some of its enhancements. *Third International Conference on Availability, Reliability and Security*, 990-993.

ENISA. (2015, 12 17). *Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics*. Retrieved from European Union Agency for Network and Information Security: https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport

Federrath, H. (2005). Privacy enhanced technologies: methods–markets–misuse. *Trust, Privacy, and Security in Digital Business*, 1-9.

Inmon, W. (2002). *Building the data warehouse: Third edition.* New York: John Wiley & Sons.

Kerschbaum, F., Strüker, J., and Koslowski, T. (2011). Confidential information-sharing for automated sustainability benchmarks. *Thirty Second International Conference on Information Systems*, 1-17.

Kim, G., Trimi, S., and Chung, J. (2014). Big-data applications in the government sector. *Communications of the ACM, 57*, 78-85.

Kimball, R., and Ross, M. (2011). The data warehouse toolkit: the complete guide to dimensional modeling. New York: John Wiley & Sons.

Li, N., Li, T., and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. *IEEE 23rd International Conference on Data Engineering*, 106-115.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. *IEEE 24th International Conference on Data Engineering*, 277-286.

Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1-53.

Oates, B. J. (2005). Researching information systems and computing. London: Sage.

OED. (2015). *privacy, n.* Retrieved from Oxford English Dictionary: http://www.oed.com/view/Entry/151596?redirectedFrom=privacy&

Reason, P., and Bradbury, H. (. (2001). Handbook of action research: Participative inquiry and practice. London: Sage.

Sharma, M., Chaudhary, A., Mathuria, M., and Chaudhary, S. (2013). A review study on the privacy preserving data mining techniques and approaches. *International Journal of Computer Science and Telecommunications, 4*(9), 42-46.

Spruit, M., and Catalina, S. (2015). DWCMM: The Data Warehouse Capability Maturity Model. *Journal of Universal Computer Science, 21*(11), 1508-1534.

Spruit, M., and Pietzka, K. (2015). MD3M: The master data management maturity model. *Computers in Human Behavior, 51(B), 1068-1076*.

Sweeney, L. (2000). *Simple demographics often identify people uniquely.* Retrieved May 5, 2016, from Data Privacy Lab: http://dataprivacylab.org/projects/identifiability/paper1.pdf