Evaluation of Data Anonymization Tools

Sergey Vinogradov
Corporate Technology
Siemens LLC
Saint-Petersburg, Russia
sergey.vinogradov@siemens.com

Alexander Pastsyak
Corporate Technology
Siemens LLC
Saint-Petersburg, Russia
alexander.pastsyak@siemens.com

***Abstract* — This survey became possible due to coming request of one of Siemens Business Units to look for data anonymization solutions being presented in the market today. The customer plans to implement and deploy it within software development projects to provide offshore team with a fully functional environment without any critical data in it. Critical data are, for instance, Personal Identifiable Information (PII), which is related to the nature of business application to be developed. In this survey paper, the introduction to data anonymization topic is given, the major challenges in data privacy an IT company may face during outsourcing of software development are considered, and the results of evaluation of data anonymization tools are provided.**

Keywords-data anonymization; test data generation; pseudonymization; data masking, de-identification.

I. INTRODUCTION

The outsourcing of software development is becoming a common practice in Siemens due to global structure of the company. Either the whole development process or its particular phases like design, implementation or testing can be outsourced. Transfer of the testing process only to a remote location is still a rare practice since it requires special consideration regarding intellectual property, security and privacy [1]. In this paper, we consider the problem of data privacy during outsourcing of the testing process for those business domains where data management in applications is especially important [2]. Usually, such applications deal with a lot of private information such as names, addresses, phone numbers, customer names, bank accounts, transactions, so, it is very important to hide this information from off-shore test team [3]. On the other hand, the environment for the test team has to be as identical to the production as possible. In this situation, data anonymization solution may help.

In this work, we provide an introduction to data anonymization topic (Section II), review of major challenges related to data privacy that an IT company may face during outsourcing of software development (Section II), and evaluation results of data anonymization tools from vendors in SAP and non-SAP domains (Sections III and IV).

II. DATA ANONYMIZATION

A. Definitions

Data Anonymization (also referred as data obfuscation, data masking, de-sensitization, de-identification or data scrubbing) is the process that helps to conceal private data. It protects sensitive information in production data base so it can be transferred to a test team. Data anonymization can be classified to pure anonymization [2] and pseudo-anonymization [4]. Pure anonymization does not provide any possibility to reconstruct the initial data, while pseudo-anonymization indeed provides such possibility through special algorithms. The former approach is the most reliable when the highest security is required, while the latter one might be interesting in the situation when the issue found by the test team has to be reproduced with production data values.

Let us consider the following example: a database with personal information (names, birth dates, bank accounts) need to be transferred to the offshore test team. The following typical data anonymization approaches can be applied to hide sensitive information:

- Data generation. Completely new data are generated. Special cases for dates and bank accounts need to be properly handled.
- Data encryption. The data is simply encrypted. Can be restored if the key is saved.
- Shuffling. The data is shuffled in one column. In this case the combination (name, bank account) will not be real.

Also pseudo anonymization approach can be implemented with almost any anonymization technique in the following way: there is a special database, which keeps track of the changes during application of the anonymization algorithms. If the reverse operation is required the lookup over this database returns the old value.

In the real situation, the combination of the approaches applied for different data fields can be considered: data generation for card numbers, data encryption for some description field and shuffling for personal names and addresses.

In our tools evaluation, we consider different anonymization algorithms as well as other criteria described in sub-section B.

B. Criteria

The following criteria have been used to evaluate data anonymization tools. We must note that the criteria described below belong to purely technical features of tools, while

 ISBN: 978-1-61208-185-4

business-related requirements (like licensing, maintenance costs, etc.) of the tools overview are out of scope of this paper.

SAP and non-SAP solution – this criterion points to a tool's ability to cope with data management applications from SAP (SAP is a market leader in enterprise application software. SAP stands for Systems, Applications and Products in Data Processing) and non-SAP domains. Significant number of Siemens business divisions uses SAP products in their projects.

According to [5], businesses that run SAP face a common challenge: how to get real SAP data into non-production systems for testing, training and Production support. The key issue is that, while new data reflecting the latest business activity is constantly being added to Production, business users cannot easily access this data. Non-production updates are essential for testing newly-developed features, production-support issues, and support packs. However, using live SAP data for testing is becoming increasingly difficult. Client or system copies disrupt the landscape, require large amounts of disk space, take a long time to prepare, and increase technical-support overheads. These challenges have to be taken into account as far as the testing activities are planned to outsource where data privacy become critical. So it becomes obvious that data anonymization solution for SAP applications cannot be standalone and has to be integrated in overall data-copy solution.

Out of the box SAP schemas support – a tool provides out-of-the-box the solution for different SAP types of the system, i.e., SAP ERP (Enterprise Resources Planning System including such parts as HCM – Human Capital Management, FI – Financing. LO – Logistics) and SAP CRM & SRM (Custom Relationship Management System and Supplier Relationship Management System, accordingly). Important issue here is a specific set of data within each SAP scheme and a tool's ability to mask such data.

Such variety of types of application data makes us searching for vendors which provide support on data copy and data anonymization for the most of SAP system types. Especially it concerns the availability of pre-defined data transformation (conversion) rules in a solution. The examples of such conversions for HR data implemented in Data Sync Manager for HCM tool are shown in Fig. 1.

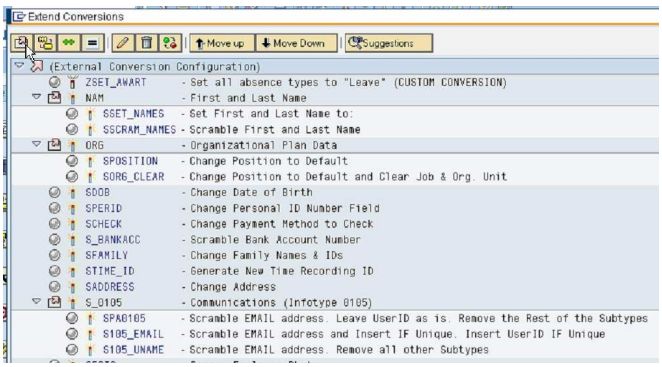

Figure 1. Conversions for HR data.

Multi Database – underlying database used in a project puts certain constraints on a data anonymization solution. Ability to work with different databases or independence of underlying database widens the application scope of a tool. This criterion is not relevant and would not be considered for SAP-dedicated data anonymization solutions.

Resulting security – since we want to protect our sensitive data the level of security needs to be assured. Usually test data generation tools have the highest security level since the data in non-production environment will be completely different from the source system.

In general, data anonymization algorithms can be grouped in three categories: data generation algorithms, algorithms which are dealing with already existing data and the combination of the above.

The first group includes all algorithms, which create completely new entries in the data base; they are used for anonymization of bank accounts, credit card numbers, social security numbers, generating random numbers, dates, etc.

The second group of algorithms operates with already existing data in the data base. The typical examples in this group are: shuffling of the fields in one or several columns, encryption, scrambling the letters in the string or figures in numbers and so on.

Also it is possible to make a combination of the algorithms from the above groups, e.g., shuffle the fields in one of the data base columns and add a random number as a string literal at the end of the field.

Preserving application and data integrity – data anonymization technologies should satisfy a simple, yet strict rule: the application that runs against masked data performs as if masked data is real [6].

This is a MUST requirement for every data anonymization tool since without conforming to this criterion the resulting database will be useless. The main focus during evaluation has been done on the tool ability to preserve data relations automatically (without user assistance).

Support of roles assignment – tool offers different roles to operate and use tool's functions. There might be system administrator role, super user role, user role with different sets of access rights to project data.

According to [6], there are four phases for data masking lifecycle identified: Data discovery and analysis, Data planning and modeling, Developing and Implementation.

The goal of Data Analysis phase is to identify the data that needs to be masked in order to sufficiently protect the data without compromising data utility. At this phase, the highest level of access a tool's user has to be provided assuming the work with data mining and analysis of customer data.

The Planning and Modeling phase is designed to set in place the criteria that will be used within your environment to mask the data and create context around the info that was discovered in first phase. The work to be done at this phase is not supposed to deal with critical data itself, rather than is related to selection of data anonymization rules and data anonymization strategy at all.

 ISBN: 978-1-61208-185-4

The Develop phase is designed to build data masking configuration suites based upon customer specific Functional Masking Needs. Again, the preparation of data anonymization scripts and the proper configuration of data masking rules within a tool do not assume the direct access to customer data.

The Implementation and Execution phase is designed to put in place a plan for integrating data masking into the overall production-to-non-production business process. This work mostly should be done in place (creating test database(s), moving masking scripts to source code implementation libraries, etc.); thus, requiring assignment of special persons to perform it and granting temporal access to the production database.

Batch operation support – due to possibly huge amount of sensitive data to be transferred to non-production environment, due to time-consuming tasks of creating non-production copies itself it would be useful to be able to plan and run such tasks from scripts.

Algorithms for anonymization – here, the different alternatives have been considered: test data generation, encryption, data masking within several masking rules, pseudo-anonymization, dynamic data masking, etc.

It has to be noted that most of the evaluated data anonymization tools provide possibility for customization of data masking algorithms. Such possibility allows for implementing almost every imaginable algorithm, however it requires significant time to study programming techniques (from simple Java™ subroutine in Camouflage Enterprise to ABAP routine in Data Sync Manager or even a pluggable C++ library in IBM Optim™.). That is why this particular criterion is focused on the algorithms which are provided by the tool out-of-the-box. The default list of data transformers available in the Camouflage Enterprise is presented on Fig. 2.

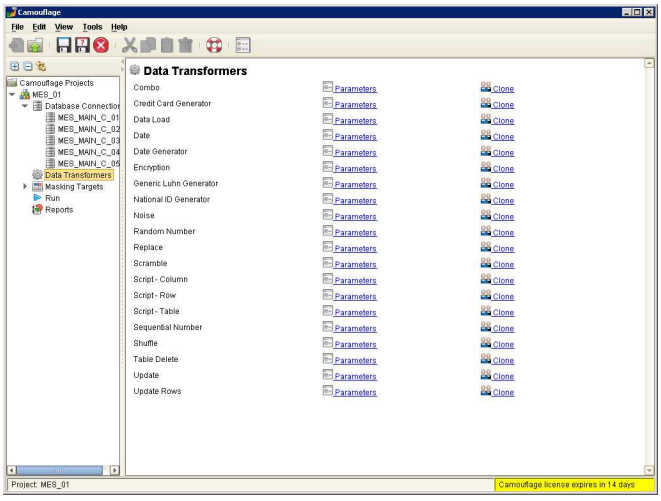

Figure 2. Default data transformers in Camouflage Enterprise

III. TOOLS ANALYSIS

There was a wide range of solutions found for data anonymization task in SAP and non-SAP domains. Moreover, in the market, we faced with several vendors who usually provide products in non-SAP domain, nevertheless they did not refuse to elaborate a data anonymization solution for SAP business applications (e.g., Grid Tools Ltd, UK). Other vendors, like IBM (US), are able to keep a separate product line to cope with data privacy projects in SAP domain. The tools found for each domain are summarized below.

There have been also found the test data generation solutions as well as non-commercial software.

Let us shortly introduce the found tools structured in a three domains: SAP, non-SAP and universal.

A. SAP tools

SAP® Test Data Migration Server (TDMS) [7] software from the market leader allows for extracting data from the production system and creating test landscapes of lower volume of data. Despite of tight integration of this solution with SAP systems, such issues like unknown anonymization algorithm (sales department could not provide such information) and time-consuming installation and setup procedure to be performed only by SAP consultants compromised the perspectives of this tool.

Accenture Clone and Test HCM (Accenture Software for SAP HCM) [8] enables the simple configuration of reliable and realistic test environments using actual data from the current SAP ERP HCM system. Within copying data and creating clone the tool provides rule-based data scrambling. The solution deserves the serious attention due to no specific customization during setup and relatively fast copying procedure. The clear disadvantage of such tool is its limit of support of SAP schemas.

The product from **GASPARIN Software Solutions, hr.dat.copy** [9] is a pure transport of personnel data (HR data) across systems and/or clients. It allows for modification or anonymization of personnel numbers or other sensitive data; however the application scope of such solution is rather limited.

BCV5™ solution from **(Enterprise Systems Associates, Inc, IBM affiliate)** [10] stands for fast, reliable copying and refresh / replication of DB2 data. It is rather fast solution customized for IBM mainframe.

Data Sync Manager for HCM from **EPI-USE** [5] pretends to be the most complete data anonymization solution for SAP systems (support can be extended to other schemas of SAP ERP as well as to SAP SRM and CRM). The solution is realized as a transport to both SAP source and target systems, thus no additional hardware and middleware required. Along with seamless integration with SAP system the setup of the tool requires no specific customization, and graphical representation of workflow for creating clones along with data masking procedures proved to be very clear and comprehensible.

B. non-SAP tools

With **Oracle Enterprise Manager Data Masking Pack** [11] from **Oracle** sensitive information such as credit card or social security numbers can be replaced with realistic values, allowing production data to be safely used in development and

 ISBN: 978-1-61208-185-4

testing or shared with out-source or off-shore partners for other non-production purposes. The clear advantage of this solution is a library of templates and format rules used, constantly transforming data in order to maintain referential integrity for applications. However, the possibility to use the tool with other DBs needs to be investigated.

Camouflage Enterprise from **Camouflage Software Inc., US** [6] provides complete tool-chain for data management tasks from data sub setting and data masking to creating copies of production system. It also delivers the highest level of customizable masking with Masking Engine, Scripting Engine and database specific transformers. Extendable and scalable for large organizations, with highly user-friendly interface this solution proves to be a key player in the market of non-SAP data anonymization solutions.

Jumble DB from **Orbium Software** [12] is a complete data scrambling solution for Oracle and SQL Server databases. Despite of moderate scalability and standard data masking features (scrambling), the vendor declares the tool's capability to keep referential integrity intact.

FieldShield product from **Innovative Routines International (IRI), Inc.** [13] masks private data at the field level with obfuscation and encryption functions which are applied according to your business rules. It works with data in the format of sequential flat files extracted from database (Oracle and DB2 are supported). Despite of multiplatform support and wide set of data anonymization technologies (from encryption till masking via custom functions) this solution offers non-trivial workflow which threatens to be very time-consuming.

DataVantage Global ® [14] solution from **Direct Computer Resources, Inc.,** protects confidential health, financial, personnel and other data and uses data obfuscation and encryption methods for this. It declares multi-database support and provides only graphical user interface to perform data anonymization tasks.

Data Masker [15] software from **Net 2000 Ltd** removes sensitive data from test databases and replaces it with realistic looking false information. Along with existing data scrambling rules a user is allowed to define her ones. Tools editions are defined on a per DB type base (Oracle, SQL, DB2).

C. Tools providing universal solutions

Datamaker™ [3] solution from **Grid Tools Ltd** creates data from scratch, creates subset databases, de-identifies (mask or obfuscate) existing data and bulks up data for performance testing. The vendor declares the support of all known databases across multiple platforms. It is worth to be mentioned that based on existing products the solution for data anonymization in SAP domain can be found and developed as well.

Tricryption® [16] solution from **Eruces, Inc.** provides secure replacement of sensitive identifiable data with anonymous but unique alias/pseudonym labels. Pros and cons of pseudo anonymization were discussed in Section II. Nevertheless, the encryption technology used in this product might be rather powerful to gain the resulting security of data anonymization while preserving the possibility to recover data. Anyway it might be useful for so called "anonymization in place" and not relevant in the case of offshore development and testing.

dgmasker™ [1] solution from **dataguise Inc.** masks data to help enterprises meet various compliance requirements such as PCI, HIPAA, GLBA, PII and SOX. The vendor declares the support of Oracle, DB2 and MS SQL Server databases, multiple advanced masking algorithms and tool's capability to perform data anonymization of SAP applications.

ActiveBase Security™ [17] product **ActiveBase Ltd** is another novel solution which offers a new approach to database security. It protects production environment by adding a security layer within and around business applications, masking or scrambling sensitive information in real time with no changes to applications or databases. Due to this independence of the solution from underlying data model in database SAP ERP, CRM applications are supported.

Optim™ [18] from **IBM** presents the complete data management solution for all known databases and significant number of application types like SAP Apps, PeopleSoft, Jd Edwards EnterpriseOne, Siebel Apps etc. It delivers powerful data transformation capabilities to mask personal information within a structured workflow of extracting data from production and sending it to development, test and training systems on demand.

IV. EVALUATION PROCEDURE

The evaluation procedure for the above tools has been separated into three different phases:

Market Evaluation – Comprised high level market analysis to choose most promising tools with regard to specified criteria for further technical evaluation.

This study has been performed using available marketing materials and calls to vendor representatives. 16 different tools described in Section III were found and screened according to the specified criteria. As the result of this phase 8 different tools proposed for further studies had been presented to the customer, who selected three of them for technical evaluation.

Solution Design – Comprised the preparation for technical evaluation of 4 tools-candidates selected at the previous project phase.

The technical evaluation of selected tools required the representative database, which has been defined and prepared during this phase. Two systems were used: Oracle database with MES data for evaluation of tools' capabilities in non-SAP domain and system with SAP HR data for evaluation of tools in SAP domain.

Technical Evaluation – Comprised evaluation of the tools-candidates which was intended to perform on test system with real data. The study included application of selected solutions and clarification of its capabilities on sample datasets.

This phase has been carried out by applying the tools in the real data anonymization scenarios. Special use cases have been designed to verify tools capabilities: 10 use cases for non-SAP

 ISBN: 978-1-61208-185-4

TABLE I. CROSS COMPARISON OF EVALUATION CRITERIA

Criteria	SAP and Non-SAP	SAP Scheme support	Multi Database	Multi Platform	Technical Features	Licensing Costs	Maintenance Costs	Setup Costs	Service / Support	Training costs	Customization costs	Required user profile	Points per Criterion	Relative Benefit Factor
SAP and Non-SAP	-	1	0	0	0	1	1	1	1	1	1	1	8	0,06
SAP Scheme support	1	-	1	1	1	2	2	2	2	2	2	2	18	0,14
Multi Database	2	1	-	1	2	2	2	2	2	2	2	2	20	0,15
Multi Platform	2	1	1	-	2	2	2	2	2	2	2	2	20	0,15
Technical Features	2	1	0	0	-	2	2	2	2	2	2	2	17	0,13
Licensing Costs	1	0	0	0	0	-	2	1	2	2	1	2	11	0,08
Maintenance Costs	1	0	0	0	0	0	-	0	1	0	0	2	4	0,03
Setup Costs	1	0	0	0	0	1	2	-	1	0	0	2	7	0,05
Service / Support	1	0	0	0	0	0	1	1	-	1	0	2	6	0,05
Training costs	1	0	0	0	0	0	2	2	1	-	0	2	8	0,06
Customization costs	1	0	0	0	0	1	2	2	2	2	-	2	12	0,09
Required user profile	1	0	0	0	0	0	0	0	0	0	0	-	1	0,01

data and 12 use cases for SAP data. These use cases covered common difficulties in data anonymization process: foreign-primary key relationships, triggers, dependency between different columns (e.g., dates dependency), user-defined objects and fields in SAP system, bank accounts, application of custom anonymization algorithms, etc.

Quantitative assessment of capabilities of 4 tools-candidates turned to be possible at this phase. Firstly, the customer provided the cross comparison of the evaluation criteria in order to define the "relative importance" of each criterion. The results of such cross comparison are presented in Table I. Secondly, there was performed the benefit value analysis of the tools-candidates which received as an input the list of weighted criteria and the criteria values from our technical evaluation. The results of the benefit value analysis are summarized in Table II.

V. SUMMARY

The results of data anonymization tools market evaluation have shown that there is a wide range of solutions available for data anonymization task in SAP and non-SAP domains. The solutions usually offer the longer list of functionality for creating non-production environment in addition to data anonymization, e.g., identification of sensitive data, data subsetting, application templates, and data copying. However in most cases practical applicability of the promising data anonymization solutions has to be confirmed in the course of more detailed study with involvement of trial versions of evaluated solutions.

Camouflage DLM Suite (Camouflage Software Inc.) and **DataVantage Global® tools** in non-SAP domain and **Data Sync Manager (EPI-USE)** for SAP domain have been preliminary selected as the most promising data solutions due to the following distinctive characteristics:

- User-friendly GUI and relatively short learning curve
- Comprehensive workflow and intuitive interface
- Available predefined conversions/masking rules for the most types of sensitive data
- Sensitive data identification can be provided by the vendor.

In this work data anonymization solutions only have been studied. Nevertheless, Camouflage DLM Suite aims at full functional support of creating non-production environment (sub setting, sensitive data identification, data masking) and provides the average learning curve for future users. Camouflage allows for masking the data in non SAP applications working with databases.

DSM covers SAP application domain providing transparent workflow for creating non-production environment and large set of pre-defined conversions for SAP ERP business objects.

 ISBN: 978-1-61208-185-4

TABLE II. BENEFIT VALUE ANALYSIS

Tool Comparison		**Camouflage Enterprise (DLM Suite)**		**IBM Optim™ Data Privacy 7.2**		**IBM InfoSphere Optim TDM for SAP**		**Data Sync Manager for HCM**	
Criteria	***Relative Benefit Factor***	***Value for tool 1 (0-10)***	***Relative Value Tool 1***	***Value for tool 2 (0-10)***	***Relative Value Tool 2***	***Value for tool 3 (0-10)***	***Relative Value Tool 3***	***Value for tool 4 (0-10)***	***Relative Value Tool 4***
SAP and Non-SAP	0,06	0	0,00	0	0,00	0	0,00	0	0,00
SAP Scheme support	0,14	0	0,00	0	0,00	3	0,41	9	1,23
Multi Database	0,15	10	1,52	10	1,52	10	1,52	10	1,52
Multi Platform	0,15	10	1,52	10	1,52	10	1,52	10	1,52
Technical Features	0,13	10	1,29	8	1,03	3	0,39	10	1,29
Licensing Costs	0,08	9	0,75	7	0,58	4	0,33	5	0,42
Maintenance Costs	0,03	5	0,15	5	0,15	5	0,15	6	0,18
Setup Costs	0,05	9	0,48	7	0,37	7	0,37	5	0,27
Service / Support	0,05	7	0,32	6	0,27	4	0,18	9	0,41
Training costs	0,06	8	0,48	4	0,24	4	0,24	7	0,42
Customization costs	0,09	8	0,73	6	0,55	1	0,09	6	0,55
Required user profile	0,01	8	0,06	5	0,04	4	0,03	6	0,05
Total Benefit Value	1,00		7,29		6,27		5,23		7,83

REFERENCES

[1] Why Add Data Masking to Your Best Practices for Securing Sensitive Data, Dataguise Inc., Whitepaper, 2009. www.dataguise.com <retrieved: October, 2011>

[2] www.datactics.com <retrieved: October, 2011>

[3] www.grid-tools.com <retrieved: August, 2011>

[4] www.sapior.com <retrieved: October, 2011>

[5] www.epiuse.com <retrieved: November, 2011>

[6] Data Masking Best Practices, Camouflage Inc., Whitepaper, March 2010. www.datamasking.com <retrieved: November, 2011>

[7] www.sap.com <retrieved: October, 2011>

[8] www.ehr-solutions.de <retrieved: October, 2011>

[9] www.gasparin.at <retrieved: August, 2011>

[10] www.esaigroup.com, http://ubs-hainer.com/en/db2-products/bcv5-fast-cloning-of-db2-data <retrieved: October, 2011>

[11] www.oracle.com<retrieved: October, 2011>

[12] www.orbiumsoftware.com <retrieved: October, 2011>

[13] www.iri.com <retrieved: October, 2011>

[14] Data Solutions for Data Privacy, Direct Computer Resources Inc., Whitepaper, June 2010. www.datavantage.com <retrieved: October, 2011>

[15] www.datamasker.com/index.html <retrieved: August, 2011>

[16] www.eruces.com <retrieved: August, 2011>

[17] www.active-base.com <retrieved: October, 2011>

[18] http://www-01.ibm.com/software/data/data-management/optim-solutions <retrieved: October 2011>

 ISBN: 978-1-61208-185-4