Optimizing the Amount of Data to Evaluate the Events of Cyber Security

Martin Dvorak, Libor Dostalek, Zora Rihova

Abstract—**In the Czech Republic, the law on cyber security has been in force since the beginning of this year, which is based on the "Directive of the European Parliament and of the Council concerning measures to ensure a high common level of network and information security across the Union", already approved by the European Parliament. This legislation on cyber security orders to the operators of critical infrastructure and critical information systems to monitor and evaluate the events associated with their operation. The problem is that the IT systems generate huge amounts of events, which makes the implementation of this regulation very costly. The vast majority of data associated with the events is however irrelevant, from the perspective of cyber security. The essay therefore solves the problems of how to select and optimize the data that are processed by the systems evaluating the events.**

The aim of this work is to design new methods for optimizing data processed in the systems evaluating the events of information security. Methods used in this work are analysis, synthesis, and statistical methods for empirical verification of the proposed method.

***Index Terms*—Cyber security, Data flow optimization, Information Security, Security event,**

I. Introduction

Statistics [1] confirm that over the last ten years, the number of transactions carried out electronically via the Internet is constantly increasing, and the number of users of IT (information technologies) is also increasing. Simultaneously, there are growing the transactions, which may be at risk from the point of view of information security. On the other hand, there occurs an increasing number of security incidents with a view of financial gain or theft of sensitive information of organizations. Attackers perform more sophisticated and more ingenious attacks for the purpose of getting rich not only with the help of information technologies, but also with the help of social engineering techniques, etc. Government institutions are aware of this growing trend, and they are taking actions to help increasing the information security of the state. Proof of this is the "Directive of the European Parliament and of the Council concerning measures to ensure a high common level of network and information security across the Union", recently adopted by the European Parliament, and on the basis of this Directive, the law on cyber security adopted by many countries of European Union, such as Czech Republic [2], [3], [4] and [5]. The purpose of this essay is not to go into details of the law and its requirements, and supplementary decrees associated with it.

But it is not only the law that requires monitoring of events in order to ensure a level of information security, as defined by the organization. The requirements for monitoring and assessment of security events can also be found in the International Standard ISO / IEC 27001 [6], which defines the system of information security management.

The subject of this research work is the information security of the organization, which uses a system of evaluation of information security events. Specifically, the authors deal with advanced tools of protection of information security. Unlike the basic tools of information security, advanced tools are a practical application of the knowledge from the scientific areas of artificial intelligence, machine learning and semantic analysis.

The researched problem is that the application generate too much data in the critical infrastructure, while not all the data are relevant to protection of information security in order the system of evaluation of events worked properly. If all the data entered into the system of evaluation event, its slowdown threatens due to "oversaturation" or complete "congestion" and subsequent malfunction. Those consequences mean impossibility of assessment of events in real time, which is a key feature of these systems. The system of evaluating event would have required a lot of time for their evaluation, and that would bring too much delay in the work of users and user comfort.

The underlying assumption of this essay is the fact that the organization has already defined a system of information security management (e.g. according to best practice, which is enshrined in the International Standard ISO / IEC 27001), and the organization is planning or has already implemented a system of evaluating events of information security. The reason for this assumption is the fact that at its fulfilment, the organization has a defined security perimeter, classified data, and the organization has also defined the policies for information security management, data handling; and the applications in the organization are mapped, including the determination of whether they are within the scope of management of information security or not.

Highlight a section that you want to designate with a certain style, and then select the appropriate name on the style menu.

Manuscript received August 01, 2015.

Martin Dvorak, Department of System Analysis, Faculty of Informatics and Statistics, University of Economics, Winston Churchill sq. 3, Prague, Czech Republic. xdvom31@vse.cz

Libor Dostalek, Institute of Applied informatics, University of South Bohemia, Branisovska 31a, Budweis Czech Republic. dostalek@prf.jcu.cz

Zora Rihova, Institute of Applied informatics, University of South Bohemia, Branisovska 31a, Budweis Czech Republic. zora.rihova@prf.jcu.cz

The style will adjust your fonts and line spacing. **Do not change the font sizes or line spacing to squeeze more text into a limited number of pages.** Use italics for emphasis; do not underline.

II. Definitions of Basic Concepts

The objective of this subchapter is to define the concepts that are most used in the essay, and thus achieve understanding of the researched topic.

Information security

The following definition applies in the scientific community for the concept of information security: Information security is maintenance of confidentiality, integrity and availability of information and other properties such as authenticity, accountability, non-repudiation and reliability. International Standard ISO / IEC 27001 also took over this definition.

Event

For the purposes of this essay, the term of event will mean any action in the environment of an organization that is so important for the information security of the organization, that it is worth to be recorded; it is limited in time, and it is possible to clearly attribute to the event the additional attributes - notably, where and when it took place, the subject of the event, type of the event action (read, edit, delete) and who initiated it.

Security incident

For the purposes of this essay, this term will be defined as follows: a security incident is an event that involves a breach of security (availability, integrity or confidentiality) of the information. A security incident can occur without being directly harmful, or without any damage would be apparent at the moment of its occurrence. Everything may occur later.

III. Optimization of the Amount of Data

To solve the above defined problem, it is necessary to limit the amount of data entering into the system evaluating the event. For this purpose, the authors of the essay propose a method described below. With regard to the fact that the storage of data in databases is nowadays a standard in the IT field, the method focuses on SQL (Structured Query Language) queries to the database. Method of optimization of the data amount for the system evaluating the event has the following steps:

1) Analysis of sensitive data
2) Analysis of the applications processing sensitive data
3) Compilation of rules of the data filtration
4) Implementation of the data filtration
5) Verification and possible adjustment of rules of the filtration

Steps 3-5 can be repeated iteratively until the optimal amount of data is reached, which enters a system of evaluation of the event.

A. *Analysis of sensitive data*

The aim of this step is to perform a data analysis and to find not only sensitive data, but also identify the data that may (under certain combinations) bring a potential attacker to sensitive data.

The authors specified the mentioned theory on an organization which works with sensitive data in the form of personal data. In this context therefore, the sensitive data are for example the personal identification number or any other identifier, which leads to personal data.

The SQL query only on the surname is uninteresting in terms of protection of personal data, but it can lead to a search for personal data. Combination of more “uninteresting” searching data has already determined a specific person, or determines such a small group of persons in which it is easy to find specific persons. As soon as a SQL query contains a combination of the surname and street in the city district, it is a combination of data, which in itself is not sensitive, but can already lead to a specific person; this means that the answer to this SQL query will already probably contain sensitive data. Other SQL queries that do not contain the fields with sensitive data or fields that lead to find sensitive data are not relevant in terms of access to sensitive data.

The output of this step is a matrix of identifiers of sensitive data (data items that contain sensitive data) and data items that lead to sensitive data, at suitable combinations. Example of a completed matrix is shown in Table I.

B. *Analysis of the applications processing sensitive data*

The objective of this process step is to identify applications in which the sensitive data are processed and how the sensitive data are processed in these applications. The first part basically means identifying the flows of sensitive data through the organization applications. The second part of the analysis is primarily concerned with finding the following method: according what items the data are searched, what data items can be displayed in the given applications and what options the users have for copying, editing or deleting specific data items.

The analysis is performed always over applications which fall within the defined security perimeter of security policy of the organization. However, in order to control, it is also possible to start from the schema of architecture or data model of the organization if these materials are available.

The output of this step is an updated matrix from the previous step by adding the applications that process the sensitive data or make them available to the user.

Example of a completed matrix is shown in the following Table I. NOTE: For security reasons, the authors could not publish a complete filled matrix and they could not use the real names of the applications in the organization.

Table I: Example of a matrix of identifiers of sensitive data; Source: [authors].

Example of a matrix of identifiers of sensitive data											
	Identifiers of person		Identifiers of an organization				Data leading to the personal data				
Application	Personal identification number	ID of a person in the DB	Company ID	Bank account number	Tax identification number	Name	Surname	Maiden name	Birthdate	Reference number	⋮
SAP	X	X	X	X	X	X	X	X	X	X	…
Siebel		X					X	X	X	X	…
EXC		X									…
…	…	…	…	…	…	…	…	…	…	…	…

C. *Compilation of rules of the data filtration*

The aim of this step is to define a rule or a matrix of rules that determine such SQL queries that will continue to enter the system evaluating the events of information security. This means defining such SQL queries that lead to the provision / modification of sensitive data and are thus important for the information security of the organization.

In this step, there is application of logic function *OR* or *AND*, if necessary, for the data items, which (at appropriate combination) might bring the unauthorized user / attacker to the sensitive data of the organization.

In the example we chose, this means the definition of the following rules:

If the SQL query contains an identifier of the person, the function verifying the presence of these identifiers will be equal to TRUE. From the above matrix it is clear that there are two direct identifiers of the person, but to find the given person it is sufficient to use only one of these identifiers; we will use the OR logic function:
Personal identification number *or* ID of a person in the DB = *TRUE*

In compliance with this rule, the SQL query will enter the system of evaluating the events of information security, and it will be further evaluated.

Analogously, it is possible to define a rule of filtration for the data that lead to the personal data at their combination:
Surname *or* Maiden name *or* Birthdate *or* Reference number / File number = *TRUE*

In compliance with this rule, the SQL query will again enter the system of evaluating the events of information security.

D. *Implementation of the data filtration*

The aim of this step is to create a functional prototype of a filter in the form of a separate application or a database probe module of the evaluating system of the events of information security.

Application functionality in the operation consists of verifying the achievement of the defined filter rules, and then at achieving the rule, it will send the SQL queries for further processing in the system of evaluating the events of information security. In the event of non-fulfilment of the condition, the application will not let the SQL query further into the system of evaluating the events of information security.

The output of this process step is the module / application filtering the SQL queries, so that only those SQL queries entered into the system of evaluating the events that are relevant for the information security.

E. *Verification and possible adjustment of rules of the filtration*

For the verification, it is possible to use the compiled matrix of sensitive data from the process step No. 2, and the created filtering rules. Both of these documents serve as the test scenarios. The tester enters the SQL queries according to this matrix and then verifies whether, if necessary, the specific queries SQL were reflected or were not reflected in the system of evaluating the events of information security. If the results match the defined filtering rules, the application can be put into actual operation.

The authors deal with the verification itself of the proposed method in the chapter below.

F. *Key roles and responsibilities in the process*

Based on the experience from implementation in practice, the team of authors compiled a matrix of key roles and their responsibilities in the process of optimization - the so-called RACI matrix (Responsible, Accountable, Consulted, and Informed). The following table shows this matrix. For completeness, the key output and input is added to each step to enable it to be implemented.

Table II: Definitions of key responsibilities and roles in the process; Source: [authors].

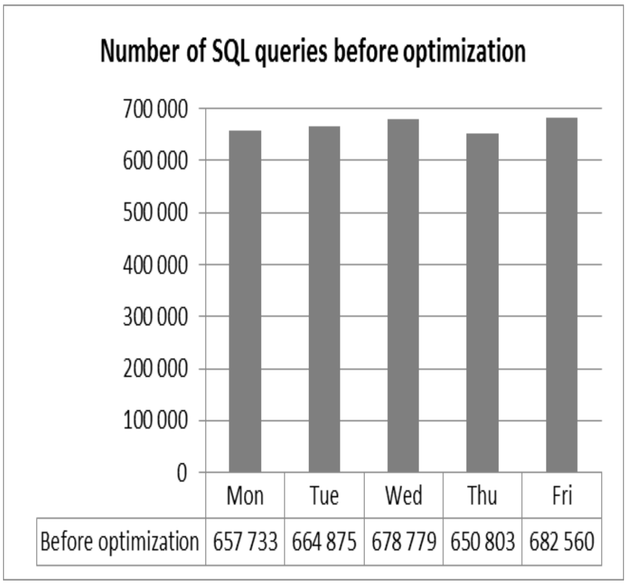
Definitions of key responsibilities and roles in the process							
Step	Input	Output	Security manager	Application specialist	Database specialist	Programmer	Tester
Analysis of sensitive data	Security policy of the organization Classification of data in the organization Data model of the organization	Matrix of sensitive data	A	C	R		
Analysis of the processing of sensitive data in applications	Matrix of sensitive data Range of the system of the information security management of the organization Diagram of the application architecture Data model of the organization	Updated matrix of sensitive data	A	R	C	I	I
Compilation of rules of the data filtration	Matrix of sensitive data	Rules of the data filtration	A	C	R	C	I
Implementation of the data filtration	Rules of the data filtration Matrix of sensitive data	Module filtering the SQL queries	A	C	C	R	I
Verification and possible adjustment of rules of the filtration	Matrix of sensitive data Rules of the data filtration	Confirmation of compliance of the rules of filtration with reality	A	I	I	C	R

IV. Verification of the Proposed Method

The authors verified the above proposed method empirically. It was a practical application in an organization processing personal data. Organization is a state owned company which deals with social security administration. It has over 10 00 employees and has implemented system for evaluating user behavior. The organization has detected in average around 4 information security incidents per day regarding attempts to unauthorized access to sensitive data (Authors were not allowed to publish more detailed graph of security incidents per day).

Due to lack of time, the measurements lasted a week before use of the above proposed method, and then a week after use from 6th April 2015. It is thus available a week-long series of measurements. On the other hand, such a short weekly cycle of measurement has an advantage in exclusion of seasonal effects. It was a measurement in two consecutive weeks when there was no holiday or significant vacation. Influences of the surrounding IT environment on measured values are also minimal due to the fact that no changes have taken place in the organization at the time of measurement, nor the outputs of any of the projects were released.

Before using the proposed filtering rules, averaged 666,950 queries SQL entered daily into the system of evaluating the events of information security for evaluation. The figure below shows the detailed number of queries.



	Mon	Tue	Wed	Thu	Fri
Before optimization	657 733	664 875	678 779	650 803	682 560

Fig. 1: Number of SQL queries before the optimization; Source: [authors].

The figure below shows the number of SQL queries after the use of filtering rules. The graph shows a significant drop in the number of queries that need to be released for processing by the system of evaluating the events of information security.

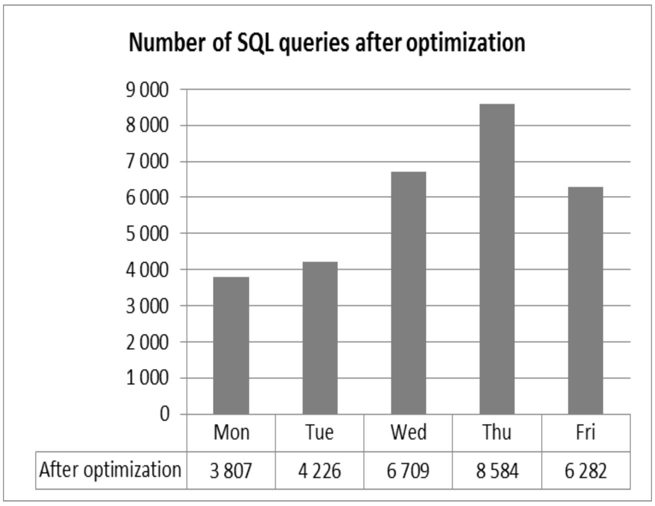


Fig. 2: Number of SQL queries after optimization; Source: [authors].

After the use of filtering rules, an average 666,950 SQL queries per day decreased to 5,992 SQL queries per day. Average number of queries after the use of filtering rules is 123 times lower, which means a drop in the SQL queries by two orders.

The SQL queries which were not released for further processing can be divided into several categories. These categories are chronologically sorted according to their share in the group of these queries:

1) So-called "Dummy queries" or "Dummy selects" e.g. queries *Select "1" from Dual* or *Select sysdate() from Dual*. This type of queries could serve to application for connection verification to database, but from the view of the data loss protection are irrelevant. Into this category are also grouped queries which store log information of application: "*Insert into log…*" These queries also do not lead to disclosure of sensitive data.
2) Queries for enumerator – these queries typically return values of some specific enumerator and thus could be also filtered out.
3) Procedures in the database – e.g.: *proc delete_logfile(day)*. Some application can initiate the execution of these predefined procedures. It has to be verified in the analysis (step number 2) whether some of these procedures do not work with sensitive data. Above mentioned procedure deletes log file for specified day and thus could be also filtered out.
4) Last category consists of those SQL queries which didn't led to disclosure of sensitive data. These are those queries which after parsing to so-called SQL construct do not work with predefined fields which contain sensitive data (parsed structure of SQL query is: Command, Object, Field).

If we compare both figures (before and after application of filtering rules), we could notice higher variability in the number of SQL queries per day. Higher variability was also analysed.

Traffic of SQL queries from applications to database consists from "parts":

1) So called "fixed traffic" – this traffic occurs every day and does not relate to user activity in the applications. E.g. applications execute daily closing report, statistics calculation and its storage in the database for management reporting purposes or application servers renew their cache in regular intervals to achieve better response times of applications.
2) SQL queries from surrounding systems – e.g. inputs from "real world" such as daily batch imports. Amount of data in each batch varies very significantly and thus this kind of traffic contributes mostly to higher variability of the traffic.
3) SQL queries caused by user activity – this depends on the working hours per day and type of user activity. Some activities do not necessarily result in SQL query for sensitive data. Typical example of this activity could be update of enumerators or other administrative activity. Real working hours also affect the amount of generated SQL queries. In order to calculate real working hours must be nominal working hours adjusted to absence, trainings, breaks, meetings and scheduled and non-scheduled downtimes. In the above mentioned organization where the proposed methodology was verified are regular team meetings in the beginnings of the week, which can explain lower number of SQL queries on Monday and Tuesday. Significant drop in the number of SQL queries on Friday can be explained by the fact that there is a shorter working day. Therefore users leave their workplace 2 hours earlier.

Verification also consisted of measuring the impact on the implemented system which evaluates user behaviour. The number of information security incidents remained the same as before application of filtering rules, i.e. in average 4 per day. The authors were not given the permission to publish detailed graph of incidents per day.

Benefits of the proposed methods were: lower number of events processed by the system evaluating user behaviour which led to decreased costs for licence fees and lower data traffic in the network of the company.

V. Conclusion

The aim of this work was to design and verify a new method that reduces the amount of data, which are processed by the system evaluating the events of information security. The authors proposed a method that is based on the analysis of sensitive data and how they are processed in the IT environment of the organization, and the subsequent application of logic functions on the basis of which the filtering rules are established that limit the amount of data entering into the system evaluating the events of information security.

Verification of the method was carried out empirically based on the weekly measurement of the amount of SQL queries. Measurements confirmed that the proposed method allows a significant reduction in the amount of SQL queries that would enter into the system evaluating the events of information security. Average number of queries after the use of filtering rules is 123 times lower, which means a drop in the SQL queries by two orders.

During the verification of the above proposed method, the authors also identified possible improvements of the optimization, using the monitoring and filtering of SQL queries according to their initiator. The authors wish to verify this hypothesis in a further research. In particular, it is about the implementation of metrics for each search data, which would enable to carry out scoring of the individual SQL queries - similarly as is used in deciding whether the e-mail message is a SPAM.

VI. References

[1] CZECH STATISTICAL OFFICE. *Internet and communication*. https://www.czso.cz/documents/10180/23180875/5_cinnosti_provadene_pomoci_internetu.xls/d340bcc3-9acd-4778-a3b9-ed797c419c07?version=1.0

Laws and standards:

[2] Government regulation No. 315/2014 governmental order which changes order no. 432/2010 on criterions defining element of critical infrastructure.

[3] Notice No. 315/2014 on security measures, cyber security incidents, reactive measures and on requirements on administration of cyber security (Notice on cyber security)

[4] Notice No. 317/2014 on significant information systems and their defining criteria.

[5] Act No. 181/2014 Sb. on cyber security

[6] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 27001 *Information technology – Security techniques – Information security management systems – Requirements.*