AUTOMATING THE DATA SECURITY PROCESS

Florin OGÎGĂU-NEAMȚIU

Head of IT Office, Regional Department of Defense Resources Management Studies, Braşov, România

Contemporary organizations face big data security challenges in the cyber environment due to modern threats and actual business working model which relies heavily on collaboration, data sharing, tool integration, increased mobility, etc. The nowadays data classification and data obfuscation selection processes (encryption, masking or tokenization) suffer because of the human implication in the process. Organizations need to shirk data security domain by classifying information based on its importance, conduct risk assessment plans and use the most cost effective data obfuscation technique. The paper proposes a new model for data protection by using automated machine decision making procedures to classify data and to select the appropriate data obfuscation technique. The proposed system uses natural language processing capabilities to analyze input data and to select the best course of action. The system has capabilities to learn from previous experiences thus improving itself and reducing the risk of wrong data classification.

Key words: *data, obfuscation, encryption, automatization, machine learning, natural language processing.*

1. INTRODUCTION

Modern technological outcomes changed the business environment which now heavily depends on rapid data collection, efficient data manipulation and quick dissemination between involved actors. In many cases technology is deployed as a request for enhancing the organization capabilities to solve a problem without regarding the security of data as a critical capability. Organizations pursue such strategies and do not fully integrate data security in their products because they see security as a systems performance limitation and do not conduct elaborated risk assessment plans to identify the associated impact of data compromising.

Modern organizations have identified the information as being one of the main assets they have

and are building defensive systems to protect it. The classical defensive strategy based on defense in depth poses increased security risks nowadays because of several factors: the need of organization to adopt integrated hardware and software services in order to minimize costs and provide rapid access to information; the need for increased collaboration capabilities due to high mobility requests of modern business model; the increased organization needs to share data with its partners for business advantages; the change of threat actors from script kids to national state actors and the proliferation of advanced persistence threats.

Because of these factors organizations are required to adopt new strategies to protect their data in all of its states: at rest, in use or in motion and apply new models for detecting and mitigating security breaches. Also, because data security imposes constrains to organization business processes and requires investments, sometimes quite considerable and without an evident benefit, organization decision makers are required to conduct extensive risk analysis studies, prioritize the main threats and apply appropriate security measures to ensure the CIA (confidentiality, integrity, availability) data model [1] is preserved while minimizing the costs and the impact on organization business processes.

2. LITERATURE REVIEW

One of the most frequent techniques implemented by modern organization in order to protect sensitive data is the data obfuscation process. The technique is used to alter original data and in this way to hide sensitive data from disclosure. There are multiple data obfuscation techniques but they can be grouped based on data alteration in 3 categories: encryption (all data is affected and the process is reversible), data masking (not all data is affected but the process is irreversible) and tokenization (not all data is affected and the process is reversible).

Data encryption is the process of conducting specific mathematical operations on an original data packet, based on specific cryptographic algorithms, in such a way that the result cannot be deciphered without knowing the secret key. Modern encryption algorithms can be divided in two main categories (symmetric and asymmetric) each of them being used in particular situations based on operational needs and constraints. In both cases one of the most important aspects of encryption is that the function is a bijection [2] and there is an inverse function which provides a way to reconstruct the original message based on a key.

Implementing data encryption in a system comes with great benefits but at the same time comes with

some drawbacks and limitations: requires considerable costs for acquiring dedicated hardware and software capabilities, requires increased resources for system maintenance, increases the difficulty of integrating modern and legacy systems, requires redesigning organization business processes for increased efficiency, high difficulties in providing the technology in large shared environments with increased personnel dynamics, lowers employees efficiency by requiring additional operations, does not provide analytical processes to be conducted on the encrypted data, requires complex key management systems for ensuring encryption key security, prone to malicious insider attacks, etc.

Data masking is another method of providing data security by creating a structurally similar but fake version of a section from a data packet [3]. Many data masking techniques exist today (substitution, shuffling, number and data variance, nulling, deletion, character scrambling, other complex rules) and the selection of which one to be used has to be done based on application integration, business requirements and minimal costs. The main characteristic of data masking technique is that the masked data cannot be converted back to the original data as the data conversion function is not reversible moreover it does not even have to be an injective function.

Data masking techniques is implemented by organizations in processes where operations need to be performed on real data without compromising data security. Such processes need to be performed for many reasons like: testing purposes in research and development environments to increase product's capabilities, enhance user training on special platforms, share data with other internal structures or third party partners in order to conduct data analyses studies to identify customer satisfaction and behavior and ultimately develop better suited products or better marketing strategies, compliance with local or international regulations, etc.

Even if it brings less operational burden on business systems and is best suited in certain environments data masking capabilities have consistent limitations: do not provide recovery of masked data, cannot be used to communicate sensitive data to legitimate applications and humans, does not support authentication mechanisms, cannot be applied to the whole data and selection need to be performed, improper data selection algorithms can lead to accidental exposure, etc.

The third method of data obfuscation is tokenization. Data tokenization is the process of

substituting a sensitive data element within a data packet with a non-sensitive equivalent, referred to as a token, which has no extrinsic or exploitable meaning or value. The token is a reference (identifier) that maps back to the sensitive data through a tokenization system [4]. The system provides capabilities to hide or show sensitive data based on who and what access that user has to information.

In order to provide such services the system keeps a database of tokens called vault linked to the associated sensitive data. Protecting system vault is critical to the system and enhanced procedures need to be in place in order to provide physical security, database integrity, limit access control, provide back-up and restore capabilities, rapid provisioning for peak demands, resiliency, etc.

Implementing tokenization in organizations requires resource investments and even if it solves some problems it comes with certain limitations: if used extensively it generates a rapid database growth, large databases are difficult to maintain and lower systems performance, rely on other obfuscation techniques like encryption to secure the data vault and the communication lines between the data vault and its clients.

None of the data obfuscation techniques is the perfect answer to modern data security challenges organizations face today. Each of the technologies addresses some problems but comes with certain limitations in terms of dedicated investments, required user training, negative influence upon organization business processes performance, allocated resources for system maintenance, etc. In order to protect sensitive data organizations can choose from different alternatives but they need to base their decision on multiple criteria: organizational challenges, technological capabilities, integration efforts within the business, cost of ownership and select the most suitable techniques or a combination of them

Because of the limitations and requirements obfuscation technologies have modern organizations need to conduct a risk assessment plan to minimize data security costs. One of the most important criteria which should guide the efforts is the importance of data for organization.

Organization face today the challenge of being forced to deal with massive amounts of data and trey need to treat data differently based on its importance. Data classification is the process of categorizing data based on some values (importance, domain, location, etc.) [5].

In order to minimize data protection costs in terms of

financial expenditures and impact upon internal business processes organizations need to apply different protection standards for different data classified categories.

The classical process used by organizations to perform data classification is conducted by humans which use their experience, knowledge and a reference system to classify information. After classification they apply an approved system/procedure to manipulate it. This process has considerable limitations like:

- Assign to the whole data object the classification level of the highest classified data packet contained and thus making the whole data unavailable to other systems who need only low level classified information (marketing, research and development, testing environment, etc.);

- Incorrect data classification due to operator inexperience, limited time, high quantity of data or usage of the "better to be safe than sorry" principle;

- Unjustified growth of high level classified data objects;

- Unjustified increase of organizational expenditures for processing higher level classified data;

- Limited flexibility and operational status;

- Increased the impact of data obfuscation upon organizational business processes due to difficulties of integrating different categories of classified systems;

- Increased reaction time of the organization;

- Limits human performance by requesting them to perform side activities;

- Increases operational burden of IT systems maintenance, back-up and data restoration processes.

3. PROPOSAL

This paper proposes a new framework to deal with sensitive data by using the automated machine decision making technology to analyze input data, classify it and then select the appropriate course of action for manipulating it in a secure way. By minimizing the human factor involvement in the whole process the system will benefit in terms of decreasing time reaction, limiting data classification errors, minimizing unjustified organization costs with data protection, increased application integration, etc.

The proposed system mimics human behavior when dealing with a task: observation, comparison to a baseline, create hypothesis, evaluate hypothesis, make decision and finally update knowledge database with special elements encountered. The system consists of 3 main modules:

data analysis, *data learning* (fig. 1) and *data obfuscation* (fig. 2).

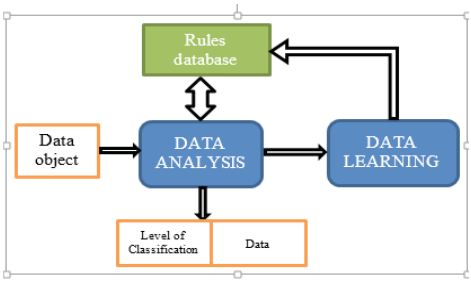

Fig. no. 1. Data analzsis and data learning modules

The systems input will be a data object which needs to be analyzed while the system output will be a data secure object. Depending on the decisions and operations performed by the system the result could be a data object which is encrypted, tokenized, masked or unaltered.

The proposed system has capabilities to analyze input data, create hypothesis based on a rules database, establish the classification level of the information contained by the data object and then apply an algorithm for securing it.

The initial data object consists of information (text, image, audio or video) and associated metadata (who sent the message, who is the recipient, on what communication channel it arrived, what is the quality of the sender and receiver, other miscellaneous indicators).

The first module is the data analysis module. It uses natural language processing algorithms to parse input data objects and decide which level of classification should be allocated to the data.

In order to make decisions the system has a rules database and a data learning module. The rules database consists of an established indicators database composed of specific markers and their corresponding classification level (security numbers, names, financial info, mission info, patterns, etc). The first step is to initialize the database with a baseline indicators pool. For this step a specialized team will develop a list of data manipulated within a specific organization, assign the corresponding classification level and input it in the system. The data learning module (Fig.2) has capabilities to improve the database by adding/removing indicators from the database by using machine-learning algorithms.

The next module (Fig. 2) will process the data and obfuscate it based on the level of classification specified and the implemented obfuscation algorithms. The system has dedicated modules based on organization policies regarding protecting data. The core elements are modules for encryption, tokenization or masking but additional instruments can be established like: provisioning virtual machines to alter data in a secure environment, establish secure

communication links with data storage equipment, configuring data access control, etc.

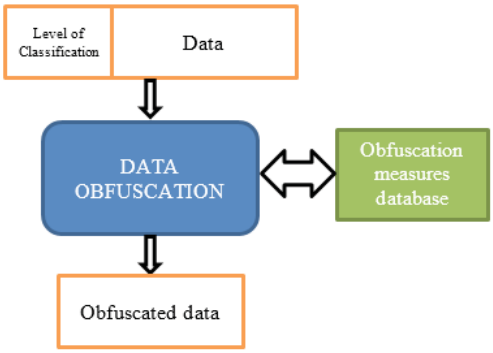

Fig. no. 2. Data analysis and learning modules

The obfuscation measures database is a modularized one and permits updating and modification based on changes appeared in the organization security policy.

The proposed system will be delivered in organization as an application, a service or it can be integrated as a security layer by developing specific APIs. Including such requirements from the design phase of the product will provide flexibility and endurance and will ensure easy integration within the existing IT infrastructure with a minimal negative impact upon organization business processes.

4. IMPLEMENTATION AND TESTING

The proposed system was tested in a cloud environment using the Natural Language Processing (NLP) capability of the Watson computing system offered on IBM Bluemix platform. The NLP is an area from the artificial intelligence domain, highly integrated with computer linguistics and artificial intelligence, aiming to develop a way by which computers can understand the human language. This technology can be applied to organize and structure knowledge, to perform operations like automatic summarization, translation, recognize entities, extract relations, detect patterns, analyze sentiments, recognize speech, etc.

In the conducted experiment an organization was imagined which needed to process data from three classification categories. The proposed system can automatically classify any input data object to one of those three categories without human intervention. The system is able to learn from encountered experiences but in order for it to update its database with new information it requires human validation.

In the first phase three classes were defined (Fig. 3). These classes corresponded to the three data classification categories the hypothetical organization has to manage.

Fig. no. 3. Class definition

Then the system was initialized by building an initial database with "experience". The process consisted of identifying a comprehensive set of words/sentences which could reflect as much as possible the information processed within that organization and then classifying that information based on the three classes created before (Fig. 4). This part is a critical one for the whole process as based on the defined "experience" the system will perform its classification process. The database needs to address the following aspects:

- to correctly define the organization activity as same data could belong to different categories in different organizations;
- to be as comprehensive as possible in order for the system to have an increased reliability. Even if the system has a built-in learning procedure and is learning from previous experiences it requires each human validation. By establishing an initial comprehensive database the validation time can be minimized, classification errors limited and an overall increased operational status;
- to eliminate duplication, one piece of information can belong only to one classification categories.

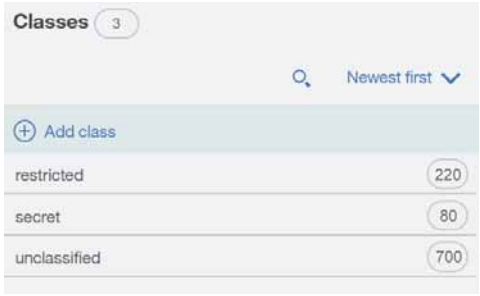

Fig. no. 4. Class population

The next step was to test the system by requesting him to classify an input data object. Multiple data objects from different categories were selected and submitted for analysis to the system (Fig 5) and (Fig 6).

Fig no. 5. Data classification process

The system performed the data classification and returned back the classification category with an error. For the samples from (Fig. 5) for which the system already had a previous "experience" it returned the answer with an error of 0.1 and 0.2 but for the sample from Fig 6, for which the system was not previously trained, it returned an error of 0.35. This output reflects the importance of the "experience" building phase of the system.

Fig. no. 6. Data classification output

The next step was to retrain the system on the unknown data object by assigning it to the classification category specific to the organization. Performing a new classification request returned the following results (Fig. 7):

Fig. no.7. Retrained data classifier

After conducting the classification the data object will be delivered to the obfuscation module where it will be processed based on the established policies and determined classification. Implementing such a module will use standard algorithms for encryption, tokenization or masking.

5. CONCLUSIONS

The proposed model has the following advantages:

- Decreased data classification process time;
- Reduced number of errors within the classification process;
- Increased reliability of the data classification process;
- Increased effectiveness of the data manipulation process within the organization;
- Increased efficiency of the organization human resource by concentrating it on value added activities and releasing from the burden of conducting administrative tasks;
- Increased organizational sharing and collaborative capabilities;
- Rapid adaptability to change;
- Minimized investments in technology by eliminating duplication of resources due to classification reasons - a system can manipulate securely multiple level classified data;
- Minimal operational surveillance after initial release effort the system.

Modern IT environments deal constantly with large quantities of data which humans are unable to process efficiently. The proposed system takes advantage of the technological developments in the natural language processing and minimizes the human intervention in the process. Moreover by taking into considering that the product will be delivered as a service from the designing phase it would have a maximized effectiveness.

Traditional IT security has concentrated on designing secure systems to process raw data rather than securing the actual data. However the cloud computing services combined with the requirements for increased mobility and data exchange has proven this model to be inadequate. Organizations need to ensure that data is processed efficiently and safely, even in untrusted environments, with minimal impact upon organization budget or business processes.

REFERENCES

[1] Panmore Institute, (2017, September 9), *The CIA thiad*, Retrieved from http://panmore .com/ the-cia-triad-confidentiality-integrity-av ailability;

[2] Dasgupta S., Papadimitriou C.H., Vazirani U. V. (2008) *Algorithms*, McGraw-Hill Company, New York, pp 41;

[3] Search security, Data masking (2017, September 18), Retrieved from http://searchse curity.techtarget. com/definition/data-masking;

[4] Wikipedia, Tokenization (data security), (2017, September, 19), Retrieved from https://en.wikipedia. org /wiki/Tokenization;

[5] Aggarwal C.C., An Introduction to Data Classification, Chapman and Hall/CRC, 2015;

[6] Suthaharan S., Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, Springer International Publishing, 2016.