Security and Challenges in Privacy Preservation of unstructured data using Pseudonymization and Data masking techniques

Dr. M.Suresh Babu[1], Professor, Department of CSE, K.L.University – off Campus – Hyderabad.
Suneetha. V[2], Research Scholar, Rayalaseema University, Kurnool.
P. Neelakanteswara[3] , Assistant Professor, Department of CSE, SVIT,Anantapur,India

Abstract

Protecting sensitive data is a challenge. And, the historic digital transformation has made this challenge even greater by the exponential increase in data. The amount of sensitive data to be protected increases at almost unbelievable rates, the data comes in numerous forms, and while data needs to be safe from cybercriminals, it must also be available to use in an ever increasing number of applications as enterprises pursue their digital transformation. And all of this needs to be done on a budget! With malicious actors becoming increasingly sophisticated, the answer is not to layer on additional endpoint and network security, or even to put all of one's stock in traditional encryption. Rather, it is to protect specific sets of particularly vulnerable data using the most appropriate solutions available for that data. One such solution is tokenization, which can be combined with dynamic data masking.

Key words : Data masking, tokenization, pseudonymization

1.0 Introduction

Pseudonymization Techniques: How to Protect Your Data

In an increasingly digital age, business changes are more than just necessary, they're inevitable. Businesses all over the world now have access to more data than ever before thought possible. This certainly brings about a whole host of opportunities, but many, many challenges as well.

For instance, companies rely on customer data to market to new and existing customers. They may collect customer email addresses and passwords to create a database of accounts. There's also customer data that a business collects at a point of sale, like credit card and address information. Businesses also have all their employee and vendor information stored somewhere. The digitization of the global economy has given businesses access to a wealth of information and given organizations a way to store and centralize it, but it has also put them at risk in a lot of ways. As technology advances and grows, so do the threats against it.

What is pseudonymization?

Data protection and privacy continue to be a priority for businesses. One technique that GDPR compliant businesses are utilizing in their commitment to data privacy is pseudonymization. This is what that looks like.

Pseudonymization takes personal data and ensures that it can't be linked back to one source or single user without additional data. For instance, if a company has your name, email address, age, nationality, and workplace name, pseudonymization takes the data that's identifiable about you specifically (your name, age, etc.) and makes it inaccessible and separate from nonidentifying data, like your nationality. Pseudonymous data can be put back together at some point so that all information can be taken together and linked back to a specific source or person.

Can pseudonymization help protect your business's data?

Here are some techniques that pseudonymization uses.

1.1. Scrambling

This technique mixes or randomizes letters in identifiable information. "Thomas" can become "Msaoht" for instance. The data is still there, it's just mixed up and harder to understand at face value.

1.2. Encryption

Encryption brings up thoughts of old spy movies where data can't be accessed without a specific code to render the data both accessible and intelligible. This is a fairly accurate description of how encryption works. Encrypting data makes it unreadable and can't be revealed or reversed to its original form without the decryption code. GDPR policies state that the decryption key is to be kept separate from the encrypted data.

1.3. Masking

This technique is found on credit card statements or documents with someone's social security number on them. The first series of numbers is usually represented by an X and the last few digits are shown as the true digits. A social security number would be rendered as XXX-XXX-4567. It's also the same technique that shows your password as a series of asterisks when being typed into a password field.

1.4. Tokenization

This pseudonymization method protects data by replacing sensitive data with non-sensitive data, referred to as tokens. The tokens have no meaning or value. It doesn't alter the length or type of data, so it can later be processed by a system that's sensitive to length and type characteristics.

1.5. Data blurring

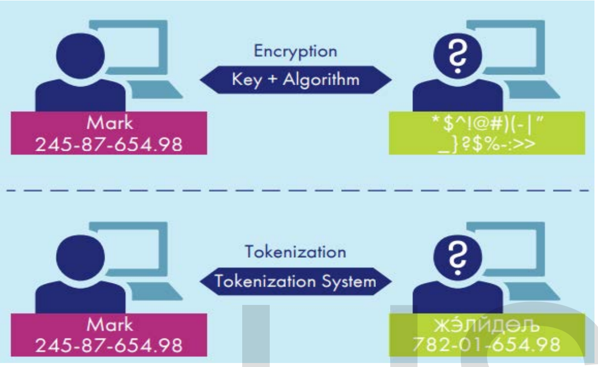
This technique uses literal blurring to create an image that is impossible to identify. A typical example is the pixelated blurring of faces you might see on the news.

Pseudonymization helps reduce the possibility that even if data is stolen or accessed, no one will easily be able to trace the sensitive, personal data back to one person or source, even if other non-sensitive information is accessible. It also allows for those who may need that sensitive information later, and individuals who have given permission to have and use it – banks, for instance – to be able to store the information safely with a way of rendering the sensitive information later as readable. This two-way process is important in protecting the data from unauthorized access and also allowing certain levels of accessibility. It has proven to be an important part of protecting privacy and data in the ongoing quest for reduced data exposure and privacy breaches.

2.0 Encryption vs. Tokenization

ENCRYPTION

One advantage of encryption, in general, is it hides the original format (e.g. size and character set) of the cleartext data, in the sense that its “raw” use will always create an output in multiples of the cipher block size. For Data Encryption Standard (DES) it is 8 bytes, for Advanced Encryption Standard (AES) it is 16 bytes. This makes it very difficult to determine what type of data was encrypted. For example, the AES encryption of both a nine-digit social security number and a 15-digit credit card number would be ciphertext of at least 16 bytes in binary data. Just by looking at this ciphertext, you would have no idea what the original data type was, because the schema is destroyed. However, this same attribute becomes a problem when encrypting clear-text data that resides in a fixed-length database field. Using the same example, if a field is designed to only contain a nine-digit social security number, the resulting 16-byte ciphertext would break the database schema, application APIs and even protocols.



2.1 TOKENIZATION WITH DYNAMIC DATA MASKING

Tokenization solves this problem. Tokenization protects sensitive data by substituting random data. It creates an unrecognizable tokenized form of the data that maintains the format of the source data. For example, a credit card number (1234-5678-1234-5678) when tokenized (2754-7529-6654-1987) looks similar to the original number and can be used in many operations that call for data in that format without the risk of linking it to the cardholder’s personal information. The tokenized data can also be stored in the same size and format as the original data. So storing the tokenized data requires no changes in database schema or process and maintains referential integrity. This makes tokenization an ideal way to store such personally identifiable information (PII) and protected health information (PHI)as:

Name National IDs, such as India’s Aadhar card number

Data of birth

Address

Telephone number

eMail address

Social Security Number

Payment card number

Passport number

Driver’s License number Etc.

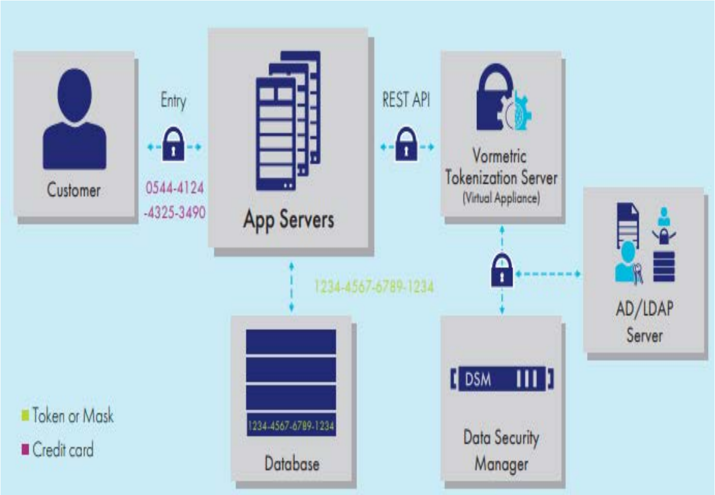


Fig1 : E Security Vormetric tokenization process.

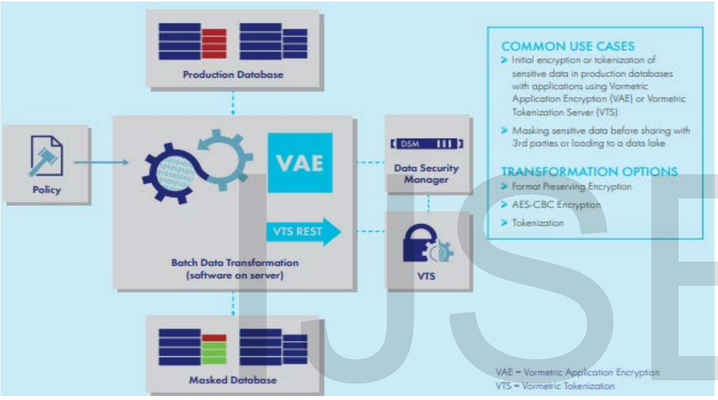

Production Database
Policy
VAE
VTS REST
Batch Data Transformation
(software on server)
DSM
Data Security
Manager
VTS
Masked Database
COMMON USE CASES
Initial encryption or tokenization of sensitive data in production databases with applications using Vormetric Application Encryption (VAE) or Vormetric Tokenization Server (VTS)
Masking sensitive data before sharing with 3rd parties or loading to a data lake
TRANSFORMATION OPTIONS
Format Preserving Encryption
AES-CBC Encryption
Tokenization
VAE = Vormetric Application Encryption
VTS = Vormetric Tokenization

Fig 2: Batch data transformation

Higher Security at Field Level Tokenization is a good security solution for an enterprise's most sensitive data, because tokenization occurs at the field level. This means the data is tokenized before it goes into the data base, which reduces the danger of insider access and credential-theft breaches.

2.2 Dynamic Data Masking

Dynamic Data Masking is a technology that protects data by dynamically masking parts of a data field, and eSecurity's Vormetric Token Server can set up rules regarding what information gets returned to specific users and roles. So, for example, a security team could establish policies so that a user with customer service representative credentials would only receive a credit card number with the last four digits visible, while a customer service supervisor could access the full credit card number in the clear. And these rules can easily be set up by non-developers, which increases ease of use.

2.3 Two-Way vs. One-Way Tokenization

Tokenization can also be one-way or two-way. In most cases users would choose two-way tokenization, because this allows retrieval of the original data. But in situations where an enterprise might want to keep information associated with a record or account, but destroy any PII associated with that account, such as using production data in testing or training for big data, the organization could use one way tokenization. This effectively destroys the PII aspect of the record and puts in its place a token in the same format. Such tokenization enables the use of numerous applications, including some used in big data analysis, without jeopardizing PII.

2.4 Optional Batch Data Transformation

Speaking of big data, another challenge enterprises face when using big data, whether on premises or in the cloud, is the time necessary to pseudonymize the data. eSecurity offers a Batch Data Transformation utility to solve this problem. With this utility, you can tokenize high volumes of sensitive records without lengthy maintenance windows and downtime.

2.5 RESTful APIs and the Cloud A security challenge

we've seen with our customers who use the cloud is the inability to put agents in the cloud. eSecurity's Vormetric Tokenization with Dynamic Masking can overcome this obstacle by employing RESTful APIs to receive and authenticate requests. RESTful APIs are text-based and are available on any operating system that supports Web services. Because they're network- and text-based, you can use them in command lines (for example, using CURL on Linux and Windows) for testing, tuck them into scripts, or embed them in application programming language. This makes tokenization simple to integrate into applications, including those in the cloud. It also enables users to send data to the cloud while maintaining control of the keys necessary to detokenize the data. Consequently, it is a viable approach to securing PII in big data in the cloud while maintaining local control over data pseudonymization, in this case, by tokenization. Cost-Effective In addition, tokenization is cost effective. According to Securosis: The most common reason organizations select tokenization over alternatives is cost reduction: reduced costs for application changes, followed by reduced audit/assessment scope. We also see organizations select tokenization when they need to update security for large enterprise applications — as long as you have to make a lot of changes, you might as well reduce potential data exposure and minimize the need for encryption at the same time.

3.0 Sectors for which Tokenization is a good fit

Among the target markets for tokenization and data masking are those that must follow strict compliance regulations, such as payments. But, PII is everywhere, and we're seeing clients moving to protect it wherever it exists in their organizations, regardless of how they acquire it. As you'll see below, there are many industries and applications for which tokenization is an excellent fit.

PAYMENTS The payments industry must abide by the Payment Card Industry Data Security Standard (PCI DSS). Using tokenization and data masking, enterprises can save money and time by reducing sensitive information sprawl. For example, if credit card numbers are stored in a database, they are governed by PCI DSS regulations. Tokenizing credit card numbers – thus making them not valuable to hackers – removes this information from the PCI DSS scope for audits. By tokenizing or masking the sensitive information that needs to be protected, the PCI DSS audit automatically becomes smaller in scope and consequently less expensive to the enterprise.

4.0 DATA ANALYTICS USING PANS

Primary account numbers (PANs) for credit cards are personally identifiable information (PII), which must be protected under mandates and regulations. However, PANs frequently are used as identifiers in customer and other kinds of analytics, because they are linked with transactions and when, where and how these transactions were made. For example a merchant might store the

PANs with their associated transactional data to do analyses of customer loyalty, heavy vs. light users, etc. If the PANs are tokenized, they are protected and out of scope. If not, the organization using them is at high risk.

HEALTHCARE

The healthcare industry is subject to numerous data privacy regulations including HIPAA3.in the U.S., the General Data Protection Regulation (GDPR)4 in the EU and similar regulations in other countries. For healthcare, tokenization and data masking help ensure that critical healthcare information is viewed only by those who need to see it. This helps the organization in question meet ePHI disclosure rules.

4.1 HUMAN RESOURCES

Tokenization is an excellent data security approach for applications that require PII, such as national identification numbers, which are regularly required by accounting, benefits and other HR services. Similarly, automated batch jobs, such as payroll deposits and 401K contributions are performed by HR applications. In all these cases, these social security numbers need to be protected. And, just as with payment cards, it may be useful to display the last four digits of the number to data users. ESecurity's Vormetric Tokenization with Dyanamic Masking is particularly appropriate for cases like this

4.2 STATE AND FEDERAL GOVERNMENT

Similar to HR, state and federal government agencies frequently handle sensitive formatted information, such as social security, driver's license and passport numbers from citizens and employees. This PII, too, needs protection, and tokenization with dynamic masking is a great option for this application.

4.3 BIG DATA

Enterprises leveraging big data environments are also finding tokenization to be an excellent solution. Given the myriad platforms used to support big data frameworks, it is important that enterprises have a number of data security options at their disposal. One immediate challenge enterprises face in this endeavor is protecting PII. In addition, at the enterprise level, the analytics applications that help businesses optimize their operations by examining all this data must also protect sensitive data from competitive or malicious threats and the consequent risk of competitive exposure and loss of customer privacy. By tokenizing PII enterprises not only protect the data from these threats by making it indecipherable, they bring their organizations into compliance with industry mandates and government regulations.How eSecurity can help eSecurity has a number of products and services that can help your organization choose the best data protection approach for your enterprise, including tokenization with dynamic data masking.

5.0 VORMETRIC TOKENIZATION WITH DYNAMIC DATA MASKING

Vormetric Vaultless Tokenization with Dynamic Data Masking dramatically reduces the cost and effort required to comply with security policies and regulatory mandates like PCI DSS, HIPAA and GDPR. The solution delivers capabilities for database tokenization and dynamic display security. It enables your organization to efficiently address its objectives for securing and anonymizing sensitive assets—whether they reside in data center, big data, container or cloud environments. Foster Innovation Without Introducing Risk: Tokenize data and maintain control and compliance when moving to the cloud, big data, and outsourced environments.

Scale Globally: Deploy the solution globally without concerns about token synchronization, performance or uncontrolled costs. The vaultless tokenization approach and pricing model enables easy to manage and affordable scale.

Increase Security for More Sensitive Data: Protect sensitive information in database columns quickly with minimal disruption, effort, and cost.

Simplify Training and Operations: Centrally manage data security policies and keys.

Streamline Key Management: Provision and manage keys for all TeS products as well as manage keys for third-party devices.

Mask Data Dynamically: Administrators can establish policies to return an entire field tokenized or dynamically mask parts of a field. For example, a security team could establish policies so that a user with customer service representative credentials would only receive a credit card number with the last four digits visible, while a customer service supervisor could access the full credit card number in the clear.

Efficiently Reduce PCI DSS Compliance Scope: Remove card holder data from PCI DSS scope with minimal cost and effort and save on complying with the industry standard with Vormetric Vaultless Tokenization with Dynamic Data Masking.

Implement without Disruption: With the solution’s format-preserving tokenization capabilities, you can restrict access to sensitive assets without changing the existing database schema. The solution’s REST API implementation makes it fast, simple, and efficient for application developers to institute sophisticated tokenization capabilities.

5.1 OPTIONAL BATCH DATA TRANSFORMATION

e-Security Tokenization customers can also deploy the Batch Data Transformation utility from e-Security. The Vormetric Batch Data Transformation utility is a high speed batching tool for encryption and tokenization. It works in conjunction with the Vormetric Application Encryption or Vormetric Tokenization Server products to facilitate the encryption or tokenization of high volumes of sensitive records without lengthy maintenance windows and downtime. You can also tokenize or mask sensitive columns in production databases and in copies of databases before they are shared with third-party developers and big data environments. No changes to applications, network systems or storage architectures are necessary.

Accelerate Transformation of Existing Sensitive Data: Protect sensitive information in database columns quickly and efficiently using encryption or tokenization with minimal disruption, effort and cost.

Refresh Cryptographic Keys Efficiently: Avoid taking your systems offline by rotating your database encryption keys in the background to ensure compliance with data protection regulations does not become a burden that affects your system availability.

Reduce Risk when Sharing Data: Leverage static data masking to remove the sensitive information before sharing with third-party developers and big data environments while maintaining your data integrity but still supporting the critical testing and analytical activities.

Flexible Tokenization: The utility can be used in conjunction with the Vormetric Tokenization Server to tokenize selected columns in the database using a policy-based approach for the number of records specified in the batch. This avoids the need for any application changes and helps ensure that sensitive information such as credit card numbers are not stored in the clear. The reverse de-tokenize process is supported so that your applications can access the clear data again when required.

Static Data Masking: In situations where sharing a representative database with a third party is required, sensitive data needs to be removed in advance because of compliance and security concerns. Static data masking is an effective method supported by the Batch Data Transformation utility that keeps the data accurate, consistent and safe. It even supports tokenizing dates within your defined date range.

6.0 Conclusion

Tokenization is an excellent pseudonymization approach for many kinds of data, but it is particularly appropriate when you need to maintain the data format to maintain data base schemas and referential integrity. It is also appropriate for particularly sensitive data, because the data is tokenized before it is stored in the database. This leaves control of the pseudonymization process in the hands of the enterprise, not those of third party storage providers. In addition, with a RESTful API interface and batch data transformation, tokenization is easy to implement and use. And it is cost effective. While tokenization was initially adopted for protection of PANS in payment applications, its utility is much broader, because: PII is increasingly ubiquitous Regulations and mandates are becoming more and more rigorous and challenging to meet The digital transformation is continuing to add to the stores of sensitive data and this data must be protected Cybercriminals are inventing new ways to steal your data around the clock and around the world Tokenization is cost competitive with other forms of pseudonymization.

REFERENCES

[1] Lee Chung, H.; Cranage David, A. 2010. *Personalisation-privacy paradox: The effects of personalisation and privacy assurance on customer responses to travel websites.* Elsevier. http://www.elsevier.com/locate/tourman

[2] Yanying Gu, Anthony Lo, 2009. *A Survey of Indoor Positioning Systems for Wireless Personal Networks.* IEEE Communications Surveys & Tutorials, Vol. 11, №1, First Quarter.

[3] Manyika, J., et. al. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Online:
http://www.mckinsey.com/Insights/MGI/Research/Technology_andInnovation/Big_data_The_next_frontier_for_innovation.

[4] Tene, O., and Polonetsky J. (2012). Privacy in the age of big data: A time for big decisions. Stanford Law Review 64, 63.

[5] J. Camenisch and E. van Herreweghen. Design and implementation of the Idemix anonymous credential system. In CCS'02: Proceedings of the 9th ACM conference on Computer and communications security, pages 21–30. ACM, 2002.

[6] W. Diffie and M. Hellman. New directions in cryptography. IEEE Trans. on Information Theory, 22(6):644–654, 1976.

[7] J. Domingo-Ferrer, editor. Inference Control in Statistical Databases, number 2316 in Lect. Notes Comp. Sci. Springer, Berlin, 2002.
[8] https://www.elastic.co/blog/gdpr-personal-data-pseudonymization-part-1
[9] Dalianis, H. and Velupillai, S. (2010). De-identifying Swedish Clinical Text - Refinement of a Gold Standard and Experiments with Conditional Random Fields, Journal of Biomedical Semantics, 1:6 (12 April 2010).
[10] Douglass, M., Clifford, G., Reisner, A., Moody, G. and Mark, R. (2004). Computer-assisted de-identification of free text in the MIMIC II database, Computers in Cardiology 31: 341–344. Internet: http://mimic.mit.edu/Archive/Publications/Douglass04 .pdf.
[11] Haverinen, K., Ginter, F., Laippala, V., Viljanen, T. and Salakoski, T., (2010). Dependency-based PropBanking of clinical Finnish, In Proceedings of The Fourth Linguistic Annotation Workshop (LAW held at ACL2010, Uppsala, Sweden. Internet: http://bionlp.utu.fi/clinicalcorpus.html HIPAA (2003).
[12] Health insurance portability and accountability (HIPAA), privacy rule and public health guidance, From CDC and the U.S. Department of Health and Human Services IV) held at ACL2010, Uppsala, Sweden. Internet: http://bionlp.utu.fi/clinicalcorpus.html HIPAA (2003). Health insurance portability and accountability (HIPAA), privacy rule and public health guidance, From CDC and the U.S. Department of Health and Human Services.