VIEW POINT

NEED FOR DATA MASKING IN A DATA-CENTRIC WORLD

Paromita Shome, *Senior Project Manager, Infosys Limited*

Abstract

With data gaining increasing prominence as the foundation of organizational operations and business, ensuring data security is emerging as a main priority. It is critical to safeguard sensitive data and customer privacy, the lack of which can lead to financial and reputational losses. Thus, there is a rising demand to protect personally identifiable information during transfer within organizations as well as across the external ecosystem. This paper highlights the need for data masking solutions. It also explains how customized data masking solutions can be used in today's data centric world.

Infosys®
Navigate your next

Introduction

The key differentiator for today's businesses is how they leverage data. Thus, ensuring data security is of utmost importance, particularly for organizations that deal with sensitive data. However, this can be challenging because data that is marked critical and sensitive often needs to be accessed by different departments within an organization. Without a well-defined enterprise-wide data access management strategy, securing data transfer can be difficult. The failure to properly control handling of sensitive information can lead to dangerous data breaches with far-reaching negative effects. For instance, a 2017 report by the Ponemon Institute titled 'Cost of a Data Breach Study, 2017'1 found that:

- *The average consolidated total cost of a data breach is US $3.62 million*
- *The average size of a data breach (number of records lost or stolen) increased by 1.8% in the past year*
- *The average cost of a data breach is US $141 per record*
- *Any incident – either in-house, through a third party or a combination of both – can attract penalties of US$19.30 per record. Thus, for a mere 100,000 records, the cost of a data breach can be as high as US $1.9 million*

These statistics indicate that the consequences of data breaches go beyond financial losses. They also affect the organization's reputation, leading to loss of customer and stakeholder trust. Thus, it is imperative for organizations to adopt robust solutions that manage sensitive data to avert reputational damage and financial losses.

Data masking as a solution

Data masking refers to hiding data such that sensitive information is not revealed. It can be used for various testing or development activities. The most common use cases for data masking are:

- ***Ensuring compliance with stringent data regulations*** *– Nowadays, there are many emerging protocols that mandate strict security compliance such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR). These norms do not allow organizations to transfer personal information such as personally identifiable information (PII), payment card information (PCI) and personal health information (PHI)*
- ***Securely transferring data between project teams*** *– With the increasing popularity of offshore models, project management teams are concerned about how data is shared for execution. For instance, sharing production data raises concerns about the risk of data being misused/mishandled during transition. Thus, project teams need to build an environment that closely mimics production environments and can be used for functionality validation. This requires hiding sensitive information when converting and executing production data*

It is important to note that data sensitivity varies across regions. Organizations with global operations are often governed by different laws. Hence, the demand for data security and the potential impact of any breach differ based on the operating regions. Thus, having an overarching data privacy strategy is paramount to ensure that sensitive data remains protected. This calls for a joint data protection strategy that includes vendors in offshore and near-shore models as well.

Types of data masking

There are various masking models or algorithms that can be leveraged to address the above use cases. These ensure data integrity while adhering to masking demands. The most common types are:

- *Substitution or random replacement of data with substitute data*
- *Shuffling or randomizing existing values vertically across a data set/column*
- *Data encryption by replacing sensitive values with arithmetically formulated data and using an encryption key to view the data*
- *Deleting the input data for sensitive fields and replacing with a null value to prevent visibility of the data element*
- *Replacing the input value with another value in the lookup table*

While the above models enable straight-forward masking, they cannot be applied to all cases, thus creating the need for customized data masking. Customized data masking uses an indirect masking technique where certain business rules must be adhered to along with encryption as shown in Fig 1.

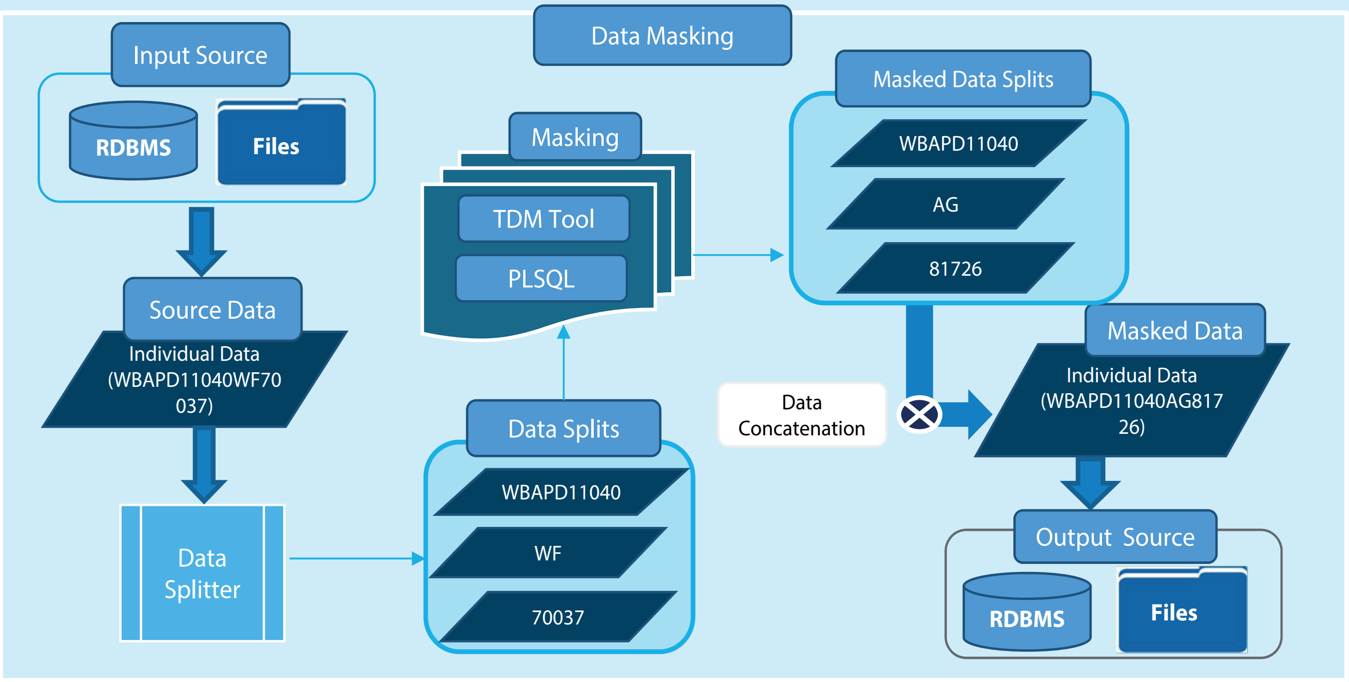

Fig 1: *A technical approach to customized data masking*

In the figure, the source data – **WBAPD11040WF70037** – is received from any source system like RDBMS/Flat files. The business rules state that:

1. There should be no change in the first 10 characters post masking
2. The next 2 letters should be substituted with letters post masking
3. The last 5 numerals should be substituted with numerals only post masking

As part of customized masking, the source data is passed through a data splitter. The single source data is then split into individual source data based on the business rules as shown in Fig 1.

After this, the individually-split data is run through the data masking tool and the masking algorithm defined in the tool is executed for each item, yielding an output of masked data. In each section, the masking type is selected based on the business rule and then the encryption is applied. The three individual sections of masked data are finally concatenated before being published at the output, which can be a database or a flat file. The masked data for the reference source data now reads as WBAPD11040AG81726. This output data still holds the validity of the source input data, but it is substituted with values that do not exist in line with the business rules. Hence, this can be utilized in any non-production environment.

Customized masking can be used in various other scenarios such as:

- *To randomly generate a number to check Luhn's algorithm where masked data ensures that the source data lies within the range of Luhn's algorithum*
- *To check number variance in a range between 'x' and 'y' where the input values will be replaced with a random value between the border values, and the decimal points are changed*
- *To check number variance of around +/- * % where a random percentage value between defined borders will be added to the input value*

Conclusion

As the demand for safeguarding sensitive data increases, organizations need effective solutions that support data masking capabilities. Two key areas where data masking is of prime importance are ensuring compliance with data regulations and protecting data while it is transferred to different environments during testing. While there are several readily-available tools for data masking, some datasets require specialized solutions. Customized data masking tools can help organizations hide source data using encryption and business rules, allowing safe transfer while adhering to various global regulatory norms. This not only saves manual effort during testing but averts huge losses through financial penalties and reputational damages arising from data breaches.

References

1. *https://securityintelligence.com*

For more information, contact askus@infosys.com

Infosys®
Navigate your next