WHITEPAPER

DATA MASKING IN THE ENTERPRISE

By

Saurabh Gupta
Cloud Compliance

Akarsh Singh
Tsaaro
INTRODUCTION

In the last couple of years, there has been an increase in cyber threats placing personal and sensitive information of individuals at the risk of being breached. In 2019, an IBM security survey showed the average cost of the data breach was more than $3.9 million. The European Union’s General Data Protection Regulations (GDPR) and California Consumer Protection Act (CCPA) in the US, is encouraging a lot of businesses that collect personal data to minimize the usage of customer’s data. This has incentivised businesses to protect and secure the personal data they collect by investing in security solutions.

One solution which could prove to be useful for businesses is the method of Data Masking (also known as data obfuscation). It has proved to be very useful and cost-effective for businesses to protect their customer’s data from any breach.

WHAT IS DATA MASKING?

Data masking is a simple data security technique wherein the personal data sets collected by organizations are masked to create fictitious data. In other words, personal information such as name, address, and age are replaced with inauthentic information so that the person using the data is not able to identify the individual to whom the data belongs. However, by doing so the quality of the information is not depleted and the information remains characteristically intact. An example is as follows:

<table>
<thead>
<tr>
<th>Unmasked information</th>
<th>Masked Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name: Adam Sandler</td>
<td>Name: Jane Doe</td>
</tr>
<tr>
<td>Age: 45</td>
<td>Age: 78</td>
</tr>
<tr>
<td>Phone no.: +44-7911-765984</td>
<td>Phone no.: +44-1234-567890</td>
</tr>
</tbody>
</table>

The purpose of data masking is to protect the personal data of the individuals which the organization collects for their training and research purposes. It also helps in securing the data when it is being shared with third parties.

WHY SHOULD COMPANIES MASK THEIR DATA?

1. In cases of a data breach, the data makes no sense to the hacker

2. Sharing personal data of individuals becomes simpler with external vendors as the real and authentic data is not exposed

3. Useful for data sanitization as any traces that are left even after deleting the file will be redundant.
4. Minimises cloud adoption risks.

5. To be compliant with various country specific data privacy regulations.

THE GROWING NEED FOR DATA MASKING

1. The EU GDPR, CCPA in the US, and APPI in Japan and Australia require organizations collecting data to use methods like encryption, pseudonymization, or anonymisation to mask the actual data of the individual. Under Article 6(4)(e) of the GDPR appropriate safeguards such as encryption or pseudonymization are to be set in place for the protection of data. Further, under Article 32(a) appropriate technical and organizational measures should be adopted to ensure the proper level of security. Similarly, provisions are laid under the CCPA and APPI to de-identify the data so that the data cannot be used to identify an individual. Also, countries like Brazil and India are coming up with their data protection laws which also require the masking of customer data. The chart below shows the country-wise data protection laws.

- Australia - APPI
- Brazil - LGPD*
- Canada - PIPEDA*
- China - Draft*
- Europe - GDPR
- Japan - APPI
- India - PDPB (Draft)*
- USA/ - CCPA/CPRA*
- South Africa – POPIA

*Privacy laws in progress/revision/recently enforced.

2. A survey done by Insider, Data Breach 2019 showed that 61% of employees have put the organization at the risk of data breach maliciously. Hence, it also becomes important that the employees who have access to the personal data that is being collected can be trusted and will not indulge in wrongful or malicious activity. Further, issues with the irregular update to the security mechanism, unpatched systems, and recording errors also constitute a failure of internal working in the organization.
3. A report by Kaspersky showed that more than 726 million cyber-attacks have occurred in the first six months of 2020. Similarly, during the Covid-19 pandemic there were a significant rise in the numbers of malware, ransomware, and phishing attacks that stole both financial data and other sensitive data of individuals. As the data stored in organizations are not generally kept in an anonymised format, this makes data easily accessible to hackers.

4. As individuals, people have also started to fully utilize the digitization process. However, what individuals do not know is that their every action on internet leaves a digital footprint. This footprint is constantly growing and with the advantage of automated technology, it has become far easier to find and access the data of an individual. To protect this data from unauthorised access, data masking can be utilized to de-identify and anonymize the information.

**DATA MASKING TECHNIQUES**

**1. Pseudonymisation**

Under this technique the personal data is no longer attributed to a particular individual without using any additional information i.e., the data is processed in a manner that the person who is viewing the data cannot identify the individual without using any other additional information. The additional information is stored in separate locations and is subject to strict technical and organizational measures.

To get access to the original data, the person accessing needs to have the key to de-identify the data.
Under Pseudonymisation three techniques are used to mask the data:

**A. Custom Seeding**

**B. Encryption of Personally Identifiable Information (PII)**

Encryption of PII requires the use of different complex algorithms that converts the data into a non-readable text and can only be accessed by the person holding the key. ExFor a particular project the name of the person giving their data can be coded as C; C+1; C+2 and further on.

Encryption helps in safeguarding the PII as when the data is being shared with third parties over a network there are chances of hackers accessing such data. In such a case the PII being encrypted will not be of any use to the hacker as he will not be able to make sense out of the data.

**C. Others**

1. **Character Scrambling**
   
   It is a simple method of rearranging and re-ordering the characters in personal data. This once done cannot be undone.

2. **Deletion**
   
   Organizations can sometimes simply delete their data but this reduces their data efficiency in the case of testing or research.
3. Number and Date Variance
   This method is used mostly for financials. It uses variance where new data sets are created and are used in the generation of synthetic data.

4. Substitution
   Original information is replaced (substituted) with some other random data. This helps in disguising the data so that authentic data can be preserved.

5. Shuffling
   Similar to the substitution method, the data from which an individual can be identified is shuffled randomly with other authentic data.

2. Anonymization

   It is the process of protecting personal or sensitive information by erasing or encrypting identifiers that connect an individual to the stored data. For example, you can run PII such as names, social security numbers, and addresses through a data anonymization process that retains the data but keeps the source anonymous.

   1. Attribute suppression
      It refers to the removal of a part of data (also referred to as “column” in databases and spreadsheets) in a dataset when an attribute is not required.

   2. Generalization
      It refers to a deliberate reduction in the precision of data. Data can be modified into a set of ranges or a broad area with appropriate boundaries. The purpose is to eliminate some of the identifiers while retaining a measure of data accuracy.

   3. Aggregation
      It is converting a dataset from a list of records to summarized values when the individual records are not required for the intended purpose.

**Drawbacks of Data Anonymization**

The GDPR stipulates that websites must obtain consent from users to collect personal information such as IP addresses, device IDs, and cookies. Collecting anonymous data and deleting identifiers from the database limit your ability to derive value and insight from your data. For example, anonymized data cannot be used for marketing efforts, or to personalize the user experience.
DATA RETENTION AND STORAGE

The data retention policy is a set of guidelines that helps organizations keep track of how long information must be kept and how to dispose of the information when it is no longer needed. The policy should also outline the purpose of processing personal data. This ensures that you have documented proof that justifies your data retention and disposal periods.

Two options one gets when the deadline for data retention expires: delete it or anonymise it. To delete the data, must ensure that all copies have been discarded. To do this, will need to find out where the data is stored. Is it a digital file, hard copy, or both? It is easy to erase physical data, but digital data often leaves traces or copies in forgotten file servers and databases. To comply with the GDPR, will need to put the data ‘beyond use’. All copies of the data should be removed from live and back-up systems.

The data retention policy will address the types of information covered in the policy. Different types of information will be subject to different rules, so you must keep a record of what data you are processing (names, addresses, contact details, financial records, and so on) and how long you are entitled to keep information.

The length of time you hold particular data is a subjective decision for you to make, based on your reasons for processing the data.

THE DIFFERENCE BETWEEN DATA MASKING AND SYNTHETIC DATA

<table>
<thead>
<tr>
<th>Data Masking</th>
<th>Synthetic Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data masking is the process of replacing confidential data or restricted data with functional fictitious data such as characters</td>
<td>Synthetic data is artificial data that is created by algorithms that mirror the statistical properties of the original data but does not reveal any information regarding individuals</td>
</tr>
<tr>
<td>There is a slight chance somebody can identify individual people based on the trends on the masked data</td>
<td>Synthetic data is a reversible solution to preserve the data</td>
</tr>
</tbody>
</table>
Data masking techniques are very easy to implement such as substitution and shuffling. Business can mask the data with one code. It requires deep learning algorithms to produce better outcomes, it takes time to create a good data set.

Data masking is a quick solution if a company wants to comply with CCPA and GPDR. It is better for testing purpose because it contains statistical attributes of an original data set.

<table>
<thead>
<tr>
<th>DATA MASKING &amp; GDPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data masking is accepted as a technique to protect individuals’ data by GDPR. Here is the related article where GDPR encourages businesses to use pseudonymization:</td>
</tr>
</tbody>
</table>

**Article 6 (4)(e):** “The existence of appropriate safeguards, which may include encryption or pseudonymization.”

**Article 25 (1):** “Taking into account the state of the art, the cost of implementation and the nature, scope, context, and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organizational measures, such as pseudonymization, which are designed to implement data-protection principles, such as data minimization, effectively and to integrate the necessary safeguards into the processing to meet the requirements of this Regulation and protect the rights of data subjects.”

**Article 32 (a):** “The controller and the processor shall implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate: the pseudonymization and encryption of personal data.”

**Article 40 (2):** “Associations and other bodies representing categories of controllers or processors may prepare codes of conduct, or amend or extend such codes, to specify the application of this Regulation, such as about:

- (d) the pseudonymization of personal data

**Article 89 (1):** “Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards including data minimization and pseudonymization.”
SUMMARY

Owing to the rapid digitalization that is taking place globally it has become a challenge to protect the personal data of the individual as it is the source for identification of an individual human.

With the technological advances that are taking place, it becomes important to use those advances in compliance and in harmony with the rules and regulations. One such as explained above is data masking. It can help organizations to better equip themselves to protect and secure the personal data that they have collected for their processing purposes. It also provides safeguards to the organisation against undue liability of non-compliance.

If you are looking to mask your data you can try out this free version of a Sandbox Data Masker by Cloud Compliance.
Sources

https://cloudcompliance.app/
https://appexchange.salesforce.com/appxListingDetail?listingId=a0N3u00000OO2JKEA1
https://research.aimultiple.com/data-masking/
https://dataprivacymanager.net/pseudonymization-according-to-the-gdpr/
https://scoop-cms.s3.amazonaws.com/566e8c75ca2f3a5d5d8b45ae/documents/egress-opinionmatters-insider-threat-research-report-a4-uk-digital.pdf
Company Profile

Tsaaro provides privacy and cybersecurity services to help organizations meet regulatory requirements while maintaining a robust security infrastructure.

Our industry-standard privacy services include: Privacy compliance, DPO-as-a-service, DPIA to name a few, delivered by our expert privacy professionals recognized by IAPP.

We take a pragmatic, risk-based approach to provide our clients with real-world, workable advice, and support, that helps them deal with a wide range of security and privacy challenges.

Contact Us

You can assess risk with respect to personal data and strengthen your data security by contacting Tsaaro.

Call us:

Monday – Friday (10AM-6PM)
+91-0522–3581306
+31-644837150

Email us:

info@tsaaro.com