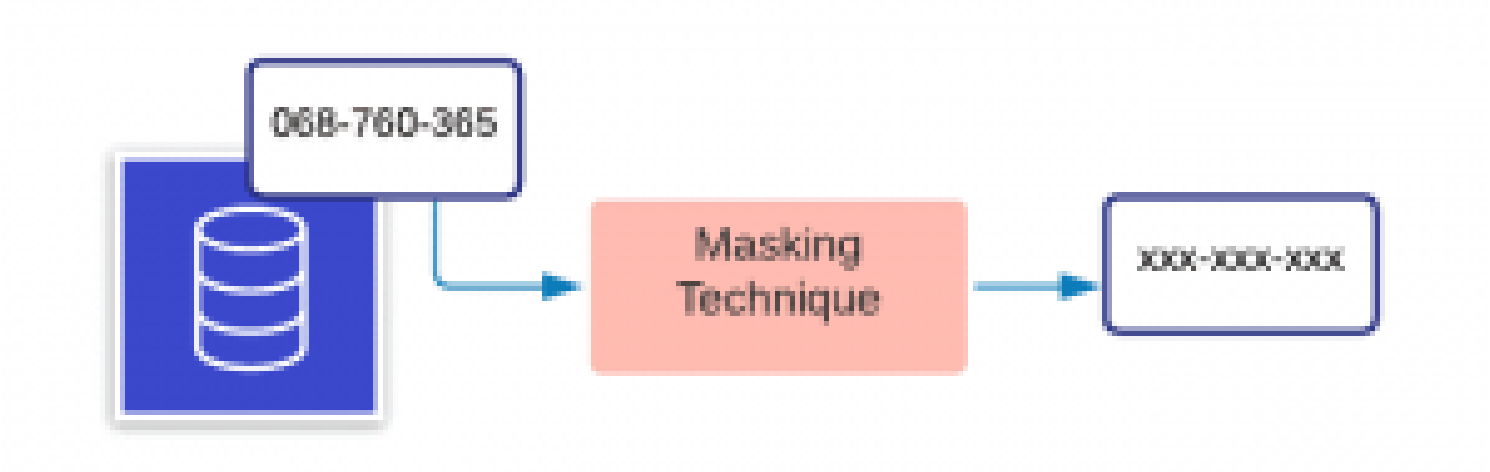
WHAT'S DATA MASKING? TYPES, TECHNIQUES AND EXAMPLES



Data breaches worldwide expose millions of people's sensitive data each year, causing many business organizations to lose millions. In fact, in 2024, the average cost of a data breach was $4.88 million. Personally Identifiable Information (PII) is the costliest type of data among all the compromised data types.

Consequently, data protection has become the top priority of many organizations. That's why data masking has become an essential technique many businesses need to protect their sensitive data.



What is data masking?

Data masking, also known as data obfuscation, hides the actual data using modified content like characters or numbers.

The main objective of data masking is creating an alternate version of data that cannot be easily identifiable or reverse engineered, protecting data classified as sensitive. Importantly, the data will be consistent across multiple databases, and the usability will remain unchanged.

There are many types of data that you can protect using masking, but common data types ripe for data masking include:

- PII: Personally identifiable information
- PHI: Protected health information
- PCI-DSS: Payment card information
- ITAR: Intellectual property

Data masking generally applies to non-production environments, such as software development and testing, user training, etc.—areas that do not need actual data. You can use various techniques to mask which we will discuss in the following sections of this article.

(Check out our big data & data security explainers.)

Importance of data masking

Data masking is important to companies in several ways:

- Helps companies to stay compliant with General Data Protection Regulation (GDPR) by eliminating the risk of sensitive data exposure. Because of this, data masking offers a competitive advantage for many organizations.
- Makes data useless for cyberattackers while preserving its usability and consistency.
- Reduces risks associated with sharing the data with integrated third-party applications and cloud migrations.
- Avoids risks associated with outsourcing any project. Because most organizations merely rely on trust when dealing with outsourced persons, masking prevents data from being misused or stolen.

Types of data masking

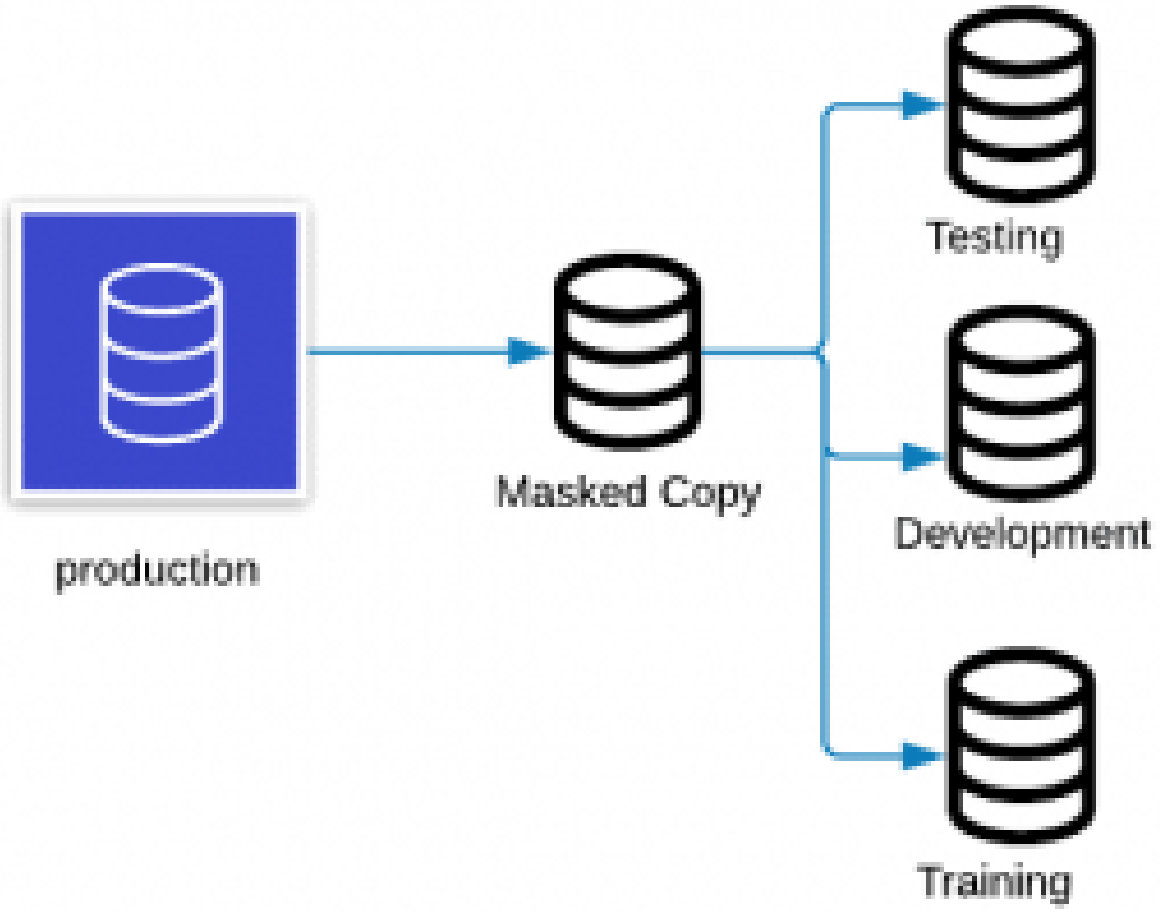
There are several types of data masking types you can use depending on your use case. Of the many, static and on-the-fly data masking are the most common.

Static data masking (SDM)s

Static data masking generally works on a copy of a production database. SDM changes data to look accurate in order to develop, test, and train accurately—without revealing the actual data. The process goes like this:

1. Take a backup or a golden copy of the production database to a different environment.
2. Remove any unnecessary data, and mask it while in stasis.

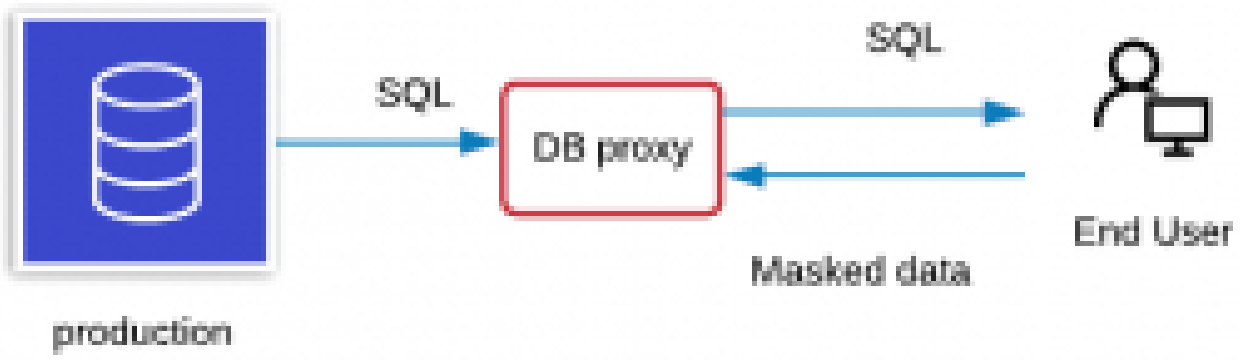
3. Save the masked copy to the desired location.



Dynamic data masking (DDM)

DDM happens dynamically at run time and streams data directly from a production system so that masked data will not need to be saved in another database. It is primarily used for processing role-based security for applications, such as processing customer inquiries and handling medical records. Thus, DDM applies to read-only scenarios to prevent writing the masked data back to the production system.

You can implement DDM using a database proxy which modifies the queries that come to the original database and passes the masked data to the requesting party. With DDM, you do not have to prepare a masked database in advance, but the application can have performance hindrances.



Deterministic data masking

Deterministic data masking involves replacing column data with the same value.

For example, if there is a first name column in your databases that consists of multiple tables, there could be many tables with the first name. If you mask 'Adam' to 'James,' it should show you as 'James' not only in the masked table but also in all associated tables. Whenever you run the masking, it will give you the same result.

On-the-fly data masking

On-the-fly data masking occurs when data transfers from production environments to another environment, like test or development. On-the-fly data masking is ideal for organizations that:

- Deploy software continuously
- Have heavy integrations

Because it is challenging to keep a backup copy of masked data continuously, this process will send only a subset of masked data when needed.

Statistical data obfuscation

The production data can hold different statistical information, which statistical data obscuration techniques can masquerade. Differential privacy is one technique where you can share information about patterns in a data set without revealing information about the actual individuals in the data set.

Data masking techniques

Now let's look at techniques for data masking.



Encryption

Encryption is the most complex—and most secure—type of data masking. Here you use an encryption algorithm that masks the data and requires a key (encryption key) to decrypt the data.

Encryption is more suitable for production data that needs to return to its original state. However, the data will be safe as long as only authorized users have the key. If any unauthorized party compromises, the keys can decrypt the data and view the actual data. Thus proper management of the encryption key is crucial.

Scrambling

Scrambling is a basic masking technique that jumbles the characters and numbers into a random order hiding the original content. Although this is a simple technique to implement, you can only apply it to certain types of data, and it does not make sensitive data as secure as you might expect.

For example, when an employee with ID number 934587 in a production environment goes through character scrambling, it will read 489357 in another environment. Anyone who remembers the original order may still be able to decipher its original value, however.

Nulling out

Nulling out masks the data by applying a null value to a data column so that any unauthorized user does not see the actual data in it. This is another simple technique, but the main problems are that it:

- Reduces data integrity
- Makes testing and development with such data harder

Substitution

Substitution is masking the data by substituting it with another value. This is one of the most effective data masking methods that preserve the original look like the feel of the data.

The substitution technique can apply to several types of data. For example, masking customer names with a random lookup file. This can be pretty difficult to execute, but it is a very effective way of protecting data from breaches.

Shuffling

Shuffling is similar to substitution, but it uses the same individual masking data column for shuffling in a randomized fashion.

For instance, shuffling employee names columns across multiple employee records. The output data looks like accurate data but doesn't reveal any actual personal information. However, if anyone gets to know the shuffling algorithm, shuffled data is prone to reverse engineering.

Number & date variance

The number and data variance method is applicable for masking important financial and transaction date information.

For instance, masking the employee salaries column with the employee salary variance will show the salaries between the highest and lowest paid employees. You can ensure the meaningfulness of the data set by applying the variance around +/- 10% to all salaries in the set.

Date aging

This masking technique either increases or decreases a date field based on the defined data masking policy with an acceptable date range. For example, decreasing the date of the birth field by 1000 days would change the date '1-Jan-2021' to '07-Apr-2018.'

Data masking examples

The following examples show how various techniques for masking data work:

1. **Customer data masking in financial services:** You may need to protect customer credit card numbers, Social Security numbers, and other sensitive data when you share it with others. A credit card number 4111-1234-5678-9012 can be masked to look like 4111-XXXX-XXXX-9012.
2. **Patient data privacy in healthcare:** Various regulations like HIPAA and frameworks such as Section 405(d) of the Health Care Cybersecurity and Resiliency Act require protecting patient

information. You can still use the data if you mask it. For example, a patient named "John Winston" born on May 15, 2001, can be masked as "Patient 1234" born on "05/XX/2001."
3. **Employee salary anonymization for HR analytics:** Your organization may want to analyze salary trends, but will want to maintain individual employee privacy. You can mask the employee's name with a number such as "Emp 001" and associate their salary with this identifier.
4. **Data masking in software testing for e-commerce:** It can be helpful to test systems with realistic data, but you will need to keep real customer data safe. You can mask something like a shipping address as "123 Test Street, Test City, TS, 99999."

Data masking best practices

Ready to start masking data? Here are some best practices to follow.

Identify the sensitive data

Before masking any data, identify and catalog the:

- Sensitive data location(s)
- Authorized person(s) who can view them
- Their usage

Every single data element of a company does not need masking. Instead, thoroughly identify the existing sensitive data in both production and non-production environments. Depending on the complexity of data and the organizational structure, this may require a significant amount of time.

Define your stack of data masking techniques

It is not practical for large organizations to use only a single masking tool across the entire enterprise since data varies greatly. Plus, the technique you choose may require you to comply with specific internal security policies or meet budgetary requirements. In some cases, you may have to develop your masking technique.

So, consider all these necessary factors to choose the right set of techniques. Keep them in sync to ensure the same type of data uses the same technique to preserve referential integrity.

Secure your data masking techniques

Masking techniques and associated data are as critical as sensitive data. For example, the substitution technique can use a lookup file for substitution. If this lookup file falls into the wrong hands, they can reveal the original data set.

Organizations should establish the required guidelines to allow only authorized persons to access the masking algorithms.

Make masking repeatable

Over time, changes to an organization or a particular project or product can result in changes to the data. Avoid starting from square one each time. Instead, make masking a process that is repeatable, quick, and automatic, so you can implement them when changes to the sensitive data occur.

Define an end-to-end data masking process

Organizations must have an end-to-end process that includes:

- Identifying sensitive information
- Applying the appropriate data masking technique
- Continuously auditing to ensure data masking is working as expected

Data masking is essential

Data masking is an essential process for many organizations that protect sensitive data by concealing its authenticity.

Related reading

- BMC Machine Learning & Big Data Blog
- Data Ethics for Companies
- AI Cyberattacks & How They Work, Explained
- Salting vs Stretching Passwords for Enterprise Security
- 7 Business-Critical IT Policies & How To Implement Them
- IT Best Practices: The Best Introduction