Methods of Data Protection

REFERENCE GUIDE

CONTENTS

Methods of Data Protection

1. INTRODUCTION

Consumers are increasingly making trust a primary criteria for purchase decisions. A recent Harvard Business Review article declared, "If you're selling a product, you're now selling trust."[1] That wealth of information customers are handing over creates an organizational obligation to protect their data; neglecting that obligation will negatively impact a brand's reputation and customer trust.

To merit trust while fostering innovation and driving business agility, enterprises must protect their data every step of the way. Every business needs a nuanced data protection strategy that ensures their customers' sensitive information is kept safe — in use, in transit and at rest.

> "Privacy is such a vital ingredient to organizational success, both to protect data and foster innovation."
>
> **John N. Stewart**,
> Senior Vice-President,
> Chief Security and Trust Officer,
> Cisco 2019 Data Privacy Benchmark Study

Data-centric protection focuses on the securing data itself, protecting sensitive enterprise data throughout its lifecycle. This data-first approach to protection can be realized using a variety of methods and technologies.

This reference guide describes data-centric security, and explains the different technologies available, the importance of data classification and security policies, as well as how to choose the most appropriate technology based on use cases.

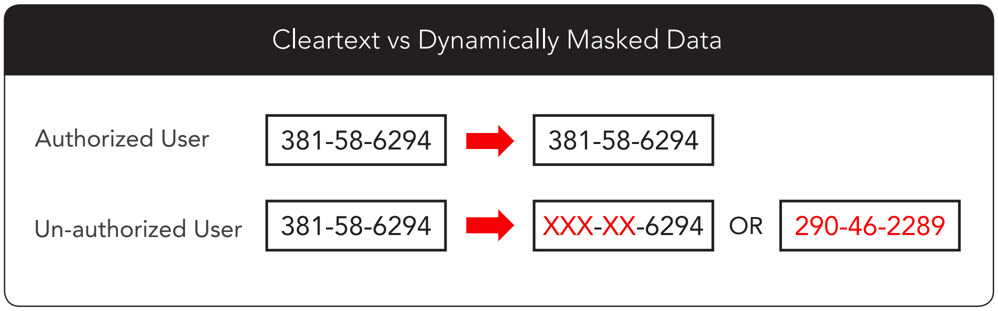
[1] https://hbr.org/2019/03/cybersecurity-is-putting-customer-trust-at-the-center-of-competition

2. DEFINITIONS

Many of the terms used to describe data protection methods are misused, creating confusion in the marketplace. This section provides a lexicon for various data protection technologies which are covered in this document.

2.1. Dynamic Data Masking (DDM)

DDM is used to protect data on the move – it does not alter cleartext data at rest. Agents are created that mask all or part(s) of the data when displayed to unauthorized users while passing through cleartext to reach authorized users.

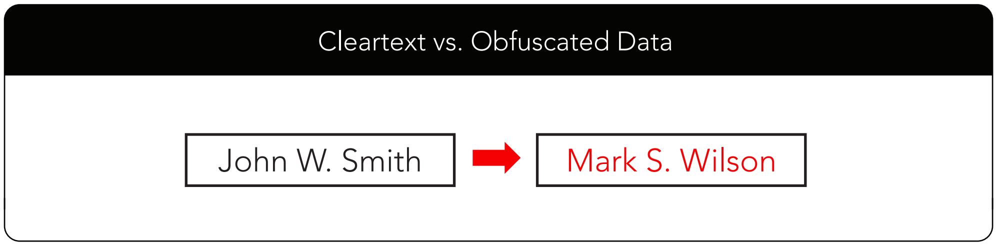

Cleartext vs Dynamically Masked Data

DDM is a data protection method delivered by many vendors to provide data protection solutions in production environments. Masking vendors, use their algorithms to deliver masked data to unauthorized users.

From a security risk point of view, there remains considerable risk to the company that is using DDM to protect their data, since the data at rest remains clear and unprotected.

2.2. Static Data Masking (SDM)

SDM is used to protect data in test and development environments. These are also known as non-production or lower environments. The SDM solution pulls data from a production environment and applies masking to the data so that it looks like real production data but actually, the data is masked.


Cleartext vs. Obfuscated Data

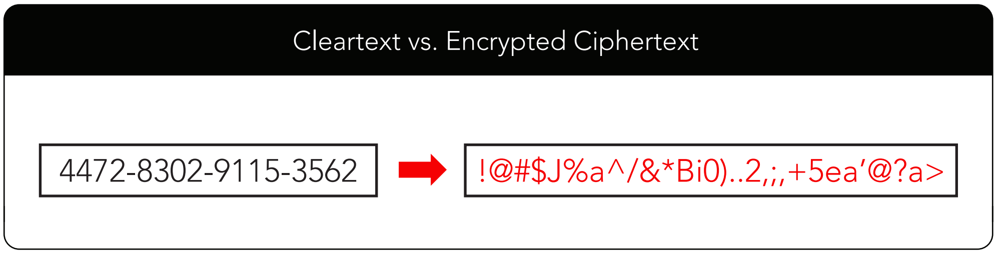
One approach to masking is to shuffle the production data so that data about an individual is not linked to the original individual. Names and other identifiers are swapped across records. There are many techniques that can be used to protect data in non-production environments.

These techniques are not reversible, so data cannot be restored back to the original production state. That means this technique is not usable in transactional business systems where data must be restored.

Vendors of this type of data protection often fall into the category of Test Data Management solutions.

2.3. Data Encryption

Encryption technology utilizes mathematical algorithms and cryptographic keys to alter data into binary ciphertext. Using the correct key with the algorithm will unlock the data, reversing the process. There are many forms of data encryption, various key strengths and other options. Encrypted (ciphertext) output is binary data and looks nothing like the original cleartext.


Cleartext vs. Encrypted Ciphertext

Using encryption to protect individual fields in databases is a challenge. Modifications need to be made to the database schema and applications to accommodate changes to the data type and length of the field.

Encryption is typically used to encrypt files versus individual fields because of these challenges. Often this is called coarse grained protection. Format-Preserving Encryption (FPE)

Format Preserving Encryption provides some benefits of both encryption (using a standards-based mathematical algorithm) and tokenization (preserving the same datatype as original cleartext data).

These combined benefits, however, come at a cost. FPE requires the same initial central processing unit (CPU) cycles to encrypt, then additional processing to convert the binary ciphertext into the same data type as the original and avoid the “collisions” (the same output for two different input values) that result when converting a larger binary field into a smaller alpha, numeric, or alphanumeric data type.

Limited support for extreme field lengths and various data types are alo common issues in the use of FPE and the usefulness of FPE is about to become a lot restrictive due to NIST’s revised FPE specifications in response to negative cryptanalysis.

2.4. Data Tokenization

In its most basic form, tokenization simply substitutes a randomly generated value (token) for a cleartext value. A lookup table (token vault) is kept in a secure place, mapping the cleartext value to the corresponding token.

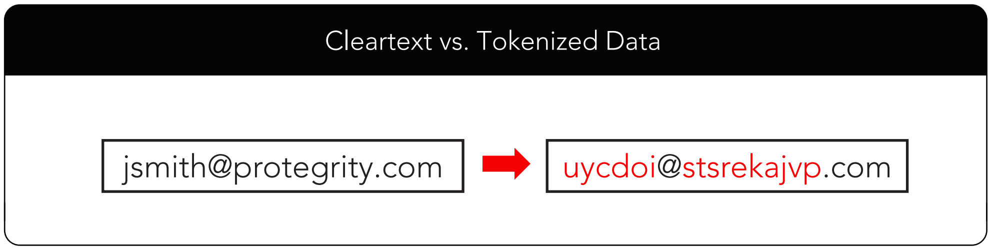
The token data type and length typically remain the same as the cleartext value, and the token lookup table becomes the ‘key’, allowing the cleartext value to be retrieved from the token.

Tokenization can be reversible and is an excellent method for protecting individual fields of data in transactional or analytical systems because the data type and length do not change

The Payment Card Industry Data Security Standard (PCI DSS) defines systems directly involved with, connected to or impacting the security of cardholder data as ‘in scope’. In 2011, PCI DSS guidelines determined that using tokenization to replace credit cards with tokens, when there was no need to reverse the token back to a credit card, is now considered out-of-scope.

Tokenization can also be used to de-identify other types of sensitive data, from Personally Identifiable Information (PII), to business-related intellectual property (IP). However, as tokenized data sets grow and IT infrastructures becomes increasingly complex, these dynamic, vault-based token lookup tables quickly become unmanageable.

A more sophisticated form of tokenization created by Protegrity, is Vaultless Tokenization (PVT). This method of data protection uses small, static token tables to create unique, random token values without the need for a dynamic, vaulted token lookup table. The result is a highly scalable, flexible and powerful protection method for structured and semi-structured data.



Cleartext vs. Tokenized Data

jsmith@protegrity.com ➡ uycdoi@stsrekajvp.com

Vaultless Tokenization is used by many organizations as the 'go to' method for protecting any information that is considered to be confidential and private but in certain circumstances it makes sense to use encryption rather than tokenization, typically when data is unstructured, such as binary files, images, biometrics, etc.

The lack of standardization in relation to tokenization may also make encryption seem like a more obvious choice but it is worth noting that PVT has been validated by some of the most distinguished names in cryptography and data security.

2.5. De-Identification

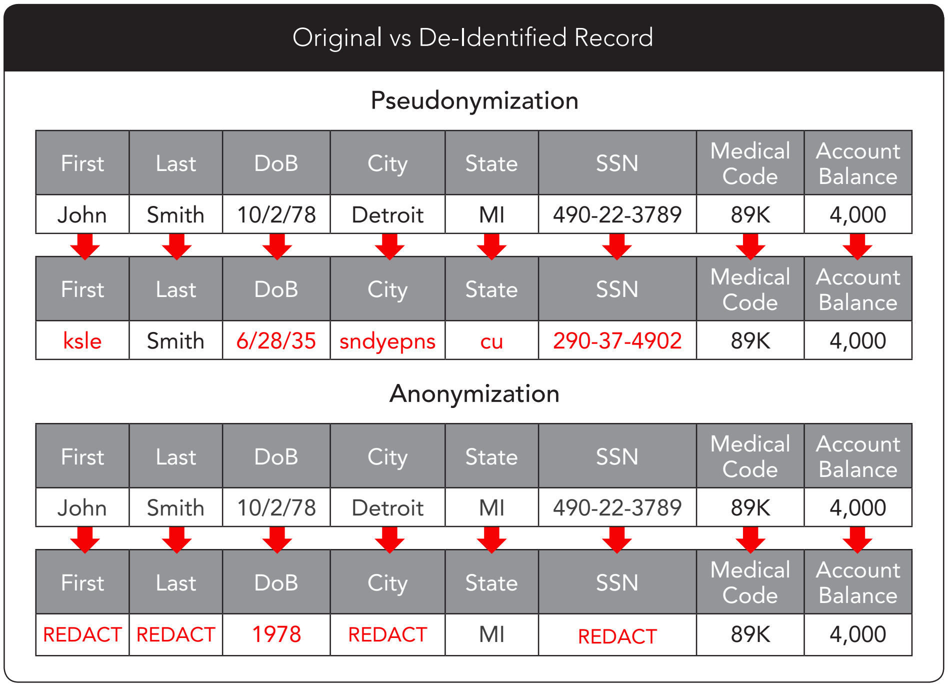
A patient or customer can be identified by information such as Name, Address and Social Security Number (SSN) or National Insurance (NI) number.

As the name implies, de-identification is used to describe the the process of disassociating a data subject's identifiers from their PII.

Different methods of de-identification exist and fall into two distinct classes; Anonymization and Pseudonymization.

In simple terms, anonymization de-identifies data without providing the ability to re-identify the data. Re-identification is the term used to re-establish a link between an individual's identifiers and the information about the individual.

Pseudonymization can also be used to de-identify data but the de-identification can be reversed. In other words, the link between an individual's identifiers and the data about the individual can be re-established.



Original vs De-Identified Record

Pseudonymization

First	Last	DoB	City	State	SSN	Medical Code	Account Balance
John	Smith	10/2/78	Detroit	MI	490-22-3789	89K	4,000

First	Last	DoB	City	State	SSN	Medical Code	Account Balance
ksle	Smith	6/28/35	sndyepns	cu	290-37-4902	89K	4,000

Anonymization

First	Last	DoB	City	State	SSN	Medical Code	Account Balance
John	Smith	10/2/78	Detroit	MI	490-22-3789	89K	4,000

First	Last	DoB	City	State	SSN	Medical Code	Account Balance
REDACT	REDACT	1978	REDACT	MI	REDACT	89K	4,000

3. DATA CLASSIFICATION AND DATA SECURITY POLICY

Regulations like the PCI DSS, Health Insurance Portability and Accountability Act (HIPAA), General Data Protection Regulation (GDPR) and US State laws such as the California Consumer Privacy Act (CCPA) mandate the protection of sensitive and personal information.

Based on these mandates, a complex variety of sensitive data and identifiers – including Name, Address, SSN, Picture, License Number, Credit Card Number (CCN), etc. – and quasi-identifiers – such as date of birth (where a date of birth will not, on its own identify an individual) – must be protected.

As a result, an organization's data security policy must define and codify which information or fields should be protected, as well as how. Data classification techniques are key to finding and identifying that all sensitive information within an organization's systems so it can be properly protected per industry and regulatory compliance mandates.

3.1. Protecting Data at Rest

At some point in its lifecycle, data will reside in a storage device. This could be in the form of a file or a relational database or in some modern data store such as Hadoop or a NoSQL database, on-premise and/or in the cloud.

Protecting data at rest can be accomplished in several ways including simple access controls, file level encryption, or fine-grained protection such as data de-identification or tokenization where data itself is protected. Another option is Transparent Database Encryption (TDE) which is a technology employed by Microsoft, IBM and Oracle to encrypt database files. TDE offers encryption at file level. And solves the problem of protecting data at rest, encrypting databases both on the hard drive and consequently on backup media.

While each method has different levels of breach risk, it has been well established that protecting data itself offers the best protection with the lowest risk of a breach.

3.2. Protecting Data in Transit

Protecting data in transit is most often the domain of network traffic encryption protocols such as SFTP, HTTPS, SS and Transport Level Security (TLS). Field level protection such as tokenization or encryption can also be used to add an additional layer of security for data as it flows between systems. Field level protection also serves as protection in transit when encryption protocols are not used, such as when moving data within a company's perimeter.

For example, a credit card can be protected with tokenization and also with a protocol like TLS. If TLS is terminated with some device like a Web Application Firewall, data is still protected in it's tokenized form.

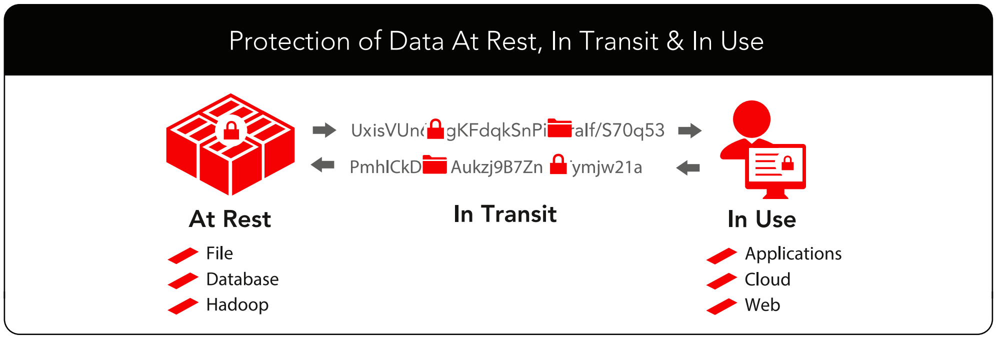
3.3. Protecting Data in Use

Best practices for minimizing risk of a data breach include protecting sensitive data at rest but organization's also need to protectect data utilized for business transactions or in analytics.

Protecting data in use means following the NIST Least Privilege principle where only the minimum data a user needs to perform their business function is delivered:

- Authorized users should be able to access sensitive data in the clear.
- Privileged users such as System Administrators, Database Administrators (DBAs) and IT should not be able to access sensitive data in the clear.

Policies should be created to enforce the delivery of sensitive data in the clear to authorized users; these policies should not be created by IT, they should be created by a different role such as a Security Officer. This is a critical principle NIST calls Separation of Duties (SoD).


Protection of Data At Rest, In Transit & In Use

3.4. Consistent Data Security Policy

Another important consideration for a data-centric approach is the ability to apply data security policies consistently across all environments within an organization. Protection method(s), limited access rights, accountability and tamper-proof audit trail should be applied consistently, throughout the entire data life-cycle from creation to deletion or archive.

A SSN first entered into a web form should remain protected according to policy at all times, wherever it can be found – on the web server, the application server, enterprise data warehouse, data lake, or archive media – regardless of the platform used for processing, analytics or storage.

4. DECIDING WHICH METHOD IS MOST APPROPRIATE

In most cases, the method of protection is determined by the specific use case for which the data is being applied. The most appropriate approach to protect data in non-production environments is Static Data Masking (SDM).

Transactional business systems such as claims processing or on-line banking transactions require a protection method that can be reversed. Pseudonymization, tokenization and encryption are reversible methods of protection but given the lack of transparency or changes that a business system must undergo when using encryption, pseudonymization or tokenization best meet the requirements of transactional systems.

In the case of Archive, Data Sharing or Analytics, the method to use will depend on whether the use case requires the re-identification or un-protection of the data. If no un-protection or re-identification is needed, then anonymization by redacting all identifiers and masking of quasi identifiers potentially fits the bill. If re-identification or un-protection is needed, then pseudonymization or tokenization better meet the requirement.

FACTORS TO CONSIDER IN SELECTING AND COMPARING PROTECTION METHODS

Algorithm ➔ Properties ↓	Encryption (AES/TDES)	Vaultless Tokenization	Vault-based Tokenization	Format Preserving Encryption	Masking / Obfuscation
Strength	Strong	Strong	Strong	Strong	Strong-Medium
Where Used	Production	Production / Non-Production	Production / Non-Production	Production / Non-Production	Non-Production
Performance	Fastest	Fast	Slowest	Medium	Medium-N/A
Transparency	Poor	High	High	High	High
Reversibility	Reversible	Reversible	Reversible	Reversible	Not Reversible
Standards Based	NIST, FIPS & Others	None	None	NIST	None
Usability with Analytics	Medium	High	Medium	High	Medium
Deployment Choices	Cluster or In-Process	Cluster or In-Process	Cluster	Cluster or In-Process	N/A
Applicability for PCI DSS	Medium	Highest	High	Medium	Not Recommended
Applicability for PII	High	Highest	Not Recommended	High	Low
Applicability for PHI	High	Highest	Not Recommended	High	Low

5. CLOUD DATA SECURITY

The traditional IT architecture is becoming obsolete with the high adoption of hybrid and multi-cloud technology. Today, 21% of data in the cloud is sensitive information, with an astounding growth rate of 53% per annum.1

Data is no longer static in isolated siloes which are easy to protect, it is constantly in flight moving between digital business processes and applications, on-premise and in the cloud.

As organizations increasingly move applications and data to cloud environments like AWS, Azure and Google, it is critical to understand that cloud vendors advocate a Shared Responsibility Model. This means that while cloud vendors offer highly secure environments, their customers are responsible for the data stored within them.

Data-centric security protects the data itself, so organizations can protect or de-identify sensitive information on-premise, then move it to the cloud and un-protect or re-identify in situ.

Data-centric security in the cloud should integrate with technology such as object stores, auto scaling containers, server-less functions, cloud HSM and other new offerings uniquely provided in cloud environments.

6. DATA-CENTRIC SECURITY SOLUTION CHECKLIST

A truly data-first approach to protecting sensitive data throughout its lifecycle, on-premise and in the cloud will deliver all of the following functionality:

- A single, centralized solution which works consistently across all platforms
- Data and security governance to specify sensitive fields and identify risk
- Automated discovery of sensitive data
- Security policy creation and management
- Role based access controls with Least Privilege and Separation of Duties
- Scalable, flexible protection and deidentification
- Tamper-proof monitoring, logging, reporting and auditing
- Comprehensive data type support
- Proven success in production environments
- Flexible, frictionless integration, on-premise and in the cloud

7. SUMMARY

It is clear that industry, government and cross-border regulatory requirements for data protection will continue to evolve, creating new and shifting cybersecurity complexities for global companies. Protecting data, however, is not something that should be done just because it is mandated by formal regulations.

Protecting customer data has become THE key element of customer trust. "Good privacy is good business" as Alastair Mactaggart, the force behind California's landmark data-privacy initiative, declared in November 2018.[2]

With a data-first approach, an organization's most valuable asset – and customer privacy – is protected, throughout the enterprise. Breaches will happen, malware will invade, and authorized users will do unauthorized things, but if sensitive information is protected using data-first security, it will remain safe regardless.

Choosing the right protection technologies for each unique data use cases is critical due to the exponential rise in data sources and platforms available on-premise and in the cloud. Flexibility should be the most important aspect of any data security solution.

Careful review of data use cases and the security methods and technologies described in this guide will help organizations identify a data-first security solution that organizations and their customers can trust.

[2] https://iapp.org/news/a/video-mactaggart-on-the-genesis-of-the-ccpa

PROVEN EXPERTS IN DATA SECURITY

Protegrity is the only data-first security solutions provider, trusted by enterprise security and data leaders in data-centric industries around the world. For more than 15 years, Protegrity's laser-like focus on data security has set the standard, and its innovative approach is unmatched in its depth and breadth, protecting sensitive data in motion, in use and at rest.

Protegrity partners with customers to secure the ever-changing data landscape through continuous innovation. Its proven approach to scalable, data-first security allows customers to optimize their use of data for greater business impact throughout the enterprise, while ensuring complete data privacy and regulatory compliance.

With Protegrity, enterprises can embrace a data-first security posture that enables a customer-first approach to innovation, service, and leadership.

Protegrity is headquartered in Stamford, Connecticut USA, with regional offices around the world. For additional information visit www.protegrity.com or call 1.203.326.7200.

Corporate Headquarters
Protegrity USA, Inc.
333 Ludlow Street, South Tower, 8th Floor
Stamford, CT 06902, USA
Phone: +1.203.326.7200

Protegrity Europe
Suite 2, First Floor, Braywick House West
Windsor Road, Maidenhead,
Berkshire SL6 1DN, United Kingdom
Phone: +44 1494 857762

Protegrity Asia Pacific
Level 6 Republic Plaza 1
9 Raffles Place, Singapore 048619
Phone: +65 9130 9618
Phone: +61 283 808 829 (AUS)

www.protegrity.com
info@protegrity.com