PROTEGRITY

METHODS OF DATA PROTECTION

Reference Guide

TABLE OF CONTENTS

Data Gathering Requires Trust

CONSUMERS ARE EMPOWERED, MORE THAN EVER, to spend money however and wherever they want. With information literally at their fingertips, thanks to mobile devices, they can freely move from one brand to the next. That doesn't mean loyalty is a thing of the past. They still look for quality and reasonable prices, but, more than anything, consumers also expect to trust brands.

As Harvard Business Review observed, "If you're selling a product, you're now selling trust."[1]

A big reason why trust matters is that the flow of information goes two ways. The price of shopping digitally is that customers hand over a wealth of their own personal information. While this data empowers brands—they can use customer information to personalize sales and marketing, and better anticipate consumers' needs and wants—businesses are also expected to properly handle it, protect it. Neglecting that obligation will tarnish a brand's reputation and erode consumer trust.

> *Privacy is such a vital ingredient to organizational success, both to protect data and foster innovation.*
>
> **John N. Steward**
> Senior Vice-President
> Chief Security and Trust Officer

To gain and hold that trust, without having to abandon the spirit of innovation and without losing business agility, businesses are turning to data-security experts to protect their customers' data. A data-security provider that understands an enterprise's unique business processes and how it manages and uses data is best positioned to keep the data safe.

Sensitive data must be secure whether it's in use, in transit, or at rest. No matter where the data lives and no matter what it does, it has to be fully protected in case the source is breached and data is stolen or compromised.

Data protection comes in several shapes and sizes; security vendors rely on a variety of methods and technologies. Before your brand makes good on your promise to customers to fully protect their data, it's best to understand the different methods of data classification and protection, and the technologies behind it all. You will then be able to choose the most appropriate data-security technology based on what your business does. This reference guide helps you take that first step.

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

The Many Methods of Data Protection

Many of the terms used to describe data protection methods are misused, often creating confusion in the marketplace. Here is a lexicon of sorts for the various data-protection technologies.

Pseudonymization

Pseudonymization is a means of hiding and protecting the identity of the data subject and is one of many ways to de-identify sensitive data. Pseudonymization hides elements of data by replacing information fields with artificial identifiers or pseudonyms. Encryption and tokenization are two common ways to pseudonymize data.

Pseudonymization hides the identity of the subject, unlike anonymization, which removes the identity of the subject entirely. Pseudonymizing data doesn't mean it's gone forever; the process is reversible and allows authorized users to view and manage the protected data afterwards.

Think of an author using a pseudonym to author a book, such as Mark Twain and Sam Clemmons or Theodor Geisel and Dr. Suess. There is still an identity, but one that is obscured by false information.

Pseudonymization is an ideal way to protect operational and transactional data and comes in handy when only parts of the data need to be protected, typically for lines of business, where some employees can have complete or mostly complete access to data, while many others only need limited access.

In healthcare, for instance, the doctors and nurses in a medical office need full access to a patient's health records but not the billing data, whereas the business staff only need to see the billing data and should not have access to a patient's health information.

Dynamic Data Masking (DDM)

Dynamic data masking (DDM) is used to protect data on the move. It does not change cleartext data at rest. Agents are created to mask all or parts of the data when displayed to unauthorized users who see the information before it reaches authorized users.

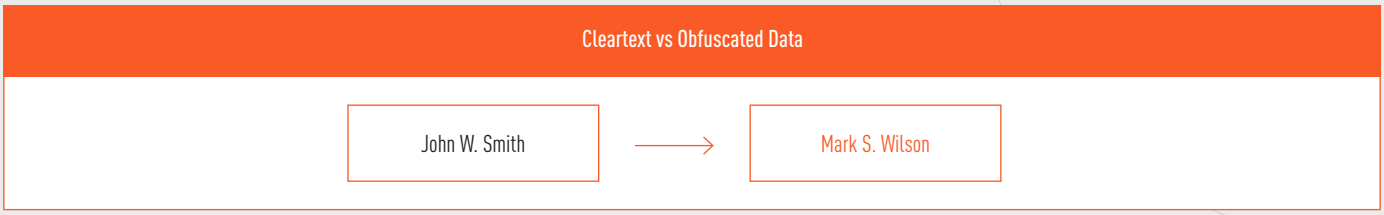
Cleartext vs Dynamically Masked Data

Authorized User	381-58-6294	→	381-58-6294		
Un-authorized User	381-58-6294	→	XXX-XX-6294	OR	290-46-2289

DDM is a data protection method often used in production environments. Vendors that specialize in masking use algorithms to create masked data. Still, DDM brings considerable risk, because the data at rest remains clear and unprotected.

Static Data Masking (SDM)

Static data masking (SDM) is used to protect data in test and development environments, also known as non-production or lower environments. An SDM solution pulls data from a production environment and masks the data so that it looks like authentic production data, but it really isn't.



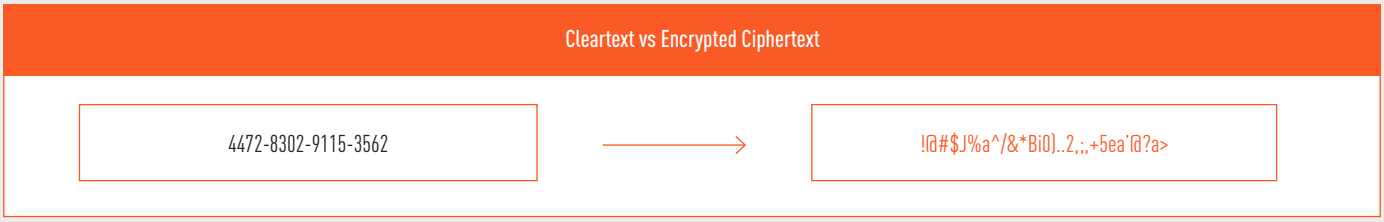
One approach to masking is to shuffle the production data so that what appears to be identifying information is not linked to an actual individual. Names and other identifiers are swapped across records.

There are other techniques to protect data in non-production environments. However, these techniques are not reversible, meaning that data cannot be restored to the original production state. That's not ideal for transactional business systems where data must be returned to its original form.

Vendors of this type of data protection offer what is known as test data-management solutions.

Data Encryption

Encryption technology uses mathematical algorithms and cryptographic keys to alter data into binary ciphertext. A key with an authorized algorithm will unlock the data, reversing the process. There are many forms of data encryption and various key strengths. Encrypted output in the form of ciphertext is binary data and looks nothing like the original cleartext.



Encrypting 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, instead of individual fields, because of these challenges. This method is known as coarse-grained protection.

Format-preserving encryption (FPE) combines some of the benefits of encryption (using a standards-based mathematical algorithm) and tokenization (preserving the same datatype as original cleartext data). This method comes at a cost, however.

FPE requires the same initial central processing unit (CPU) cycles to encrypt. Then it requires additional processing to convert the binary ciphertext into the original data type and avoid "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 another drawback of FPE. What's more: the National Institute of Standards and Technologies has been looking at tightening FPE specifications.

Data Tokenization

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

The token data type and length usually remain the same as the cleartext value, and the token lookup table becomes a key, allowing the cleartext value to be retrieved from the token. Tokenization is reversible and 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 or connected to the security of cardholder data as "in scope." In 2011, PCI DSS guidelines determined that using tokenization to replace credit card data with tokens—when there was no need to reverse the token back to original credit card data—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 datasets grow and IT infrastructures becomes increasingly complex, these dynamic, vault-based token-lookup tables quickly become unmanageable.

A more sophisticated form of tokenization, Protegrity Vaultless Tokenization (PVT), solves this challenge. 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. Instead, users benefit from a highly scalable, flexible and powerful protection method for structured and semi-structured data.

Cleartext vs Tokenized Data

jsmith@protegrity.com	→	uycodi@strekajvp.com

Vaultless tokenization is a "go-to" method for protecting confidential and private information, but in certain circumstances—typically, when data is unstructured, such as binary files, images, or biometrics—it makes sense to use encryption rather than tokenization.

The lack of standardization for tokenization might 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.

De-Identification

With this method, a patient or customer can be identified by information such as name, address, Social Security number (SSN) or National Insurance (NI) number. As the name implies, de-identification disassociates those identifiers from PII.

The methods of de-identification fall into two distinct classes: anonymization and pseudonymization. In simple terms, anonymization de-identifies data without providing the ability to re-identify it. This is a big shortcoming because most users want to reestablish the link between an individual's identifiers and the information about the individual.

Anonymization

Anonymization is considered a form of "differential privacy" and includes k-anonymity, l-diversity, and t-closeness. Anonymization removes the direct identifiers in data, which results in quasi-identifiers. Quasi-identifiers are pieces of data that are true about a specific data subject and other data subjects (e.g., general information like race, age, date of birth, and gender).

With anonymization, you are either adding fake data to a data set or generalizing the real data so it is still true, but not as precise. The key is to generalize the data until it is "fuzzy" enough that you can no longer identify a person with a given record. Anonymization is an irreversible process.

Anonymization is rising in popularity, because of recent data privacy legislation such as GDPR and CCPA. These laws state that if data has been properly anonymized, it does not fall under these privacy regulations. Anonymized data can be used for long-term analytics initiatives, archiving, or shared with third parties anywhere in the world.

Anonymization is extremely important for secure machine learning (ML) and artificial intelligence (AI) initiatives. There is little statistical difference between training ML machines with open data versus anonymized data. The main difference between using real data and anonymized data to train ML models is privacy – using anonymized data achieves the same results with no privacy risk.

A real-world analogy is to envision an author using a pseudonym to byline a book – such as Mark Twain and Sam Clemmons or Theodor Geisel and Dr. Suess. There is still an identity there, it is just obfuscated with fake information. Anonymization, on the other hand, would be if someone wrote a book and bylined it "Anonymous." We would not be able to tell if the book was written by one person or ten people.

K-Anonymity, L-Diversity, T-Closeness

Within anonymization are k-anonymity, l-diversity, and t-closeness. These techniques are used when a company storing sensitive data does not want to use fake data in their data set. K-anonymity removes direct identifiers so the data is left with quasi-identifiers – extracting away the specifics in a data set. This means that for every record that has a unique set of quasi-identified data, there must be at least 'k' other records that have the same quasi-identifiers.

L-diversity and t-closeness are techniques that are applied on top of k-anonymity. Once a data set is de-identified via k-anonymity, the next step is to examine the data set and ensure values are sufficiently diverse (using l-diversity models to further anonymize sensitive values within a group that exhibits a level of homogeneity) and that the distribution of these values is close enough to the distribution in the entire data set (using t-closeness models).

When looking at a data set, if there is a record for a male that lives in the 43081 ZIP code and is 45 years old on September 2, 1975, there needs to be at least k other records that are also males living in 43081 zip code that are 45 years old on September 2, 1975. If the k record is four, then the data set must have four other records that match those quasi-identifiers.

If the data set does not return four records that meet this criteria, then the values must be made more general. The quasi-identifiers will be changed to looking for males living in an area code that contains 4308 and born in September 1975. If this level of generalization still does not deliver the level of k-anonymity required, the data must be generalized again. This process continues until the data set has been sufficiently anonymized, with sufficient diversity of responses and an overall closeness to the mean of the entire population.

Pseudonymization can also be used to de-identify data, but, here, any de-identification can be reversed. Unlike anonymization, the link between an individual's identifiers and the data about the individual can be re-established.

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	2/28/35	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

Data Classification and Data Security Policy

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

These mandates require that a variety of sensitive data—identifiers including name, address, SSN, photos, license number, and credit card number (CCN), as well as quasi-identifiers, such as date of birth (even though, this will not, on its own, identify an individual) must be protected.

An organization's data security policy must define and codify which information or fields should be protected, as well as how it is all protected, following these mandates. Dataclassification techniques are key to finding and identifying such sensitive information within an organization's systems.

Protection Data at Rest

At some point in its lifecycle, data will reside in a storage device. It could rest in a file, a relational database or a modern datastore such as Hadoop or NoSQL, either on-premises or in the cloud.

Protecting data at rest can be accomplished by using simple access controls, file-level encryption, or fine-grained methods such as data de-identification or tokenization, where the data itself is protected. Another option is transparent database encryption (TDE), a technology used by Microsoft, IBM, and Oracle to encrypt database files. TDE offers encryption at the file level and protects data at rest by encrypting databases on a hard drive and backup media.

While each method brings different levels of security risks, it has been established that protecting data itself offers the best protection, while also carrying the lowest level of risk.

Protection Data in Transit

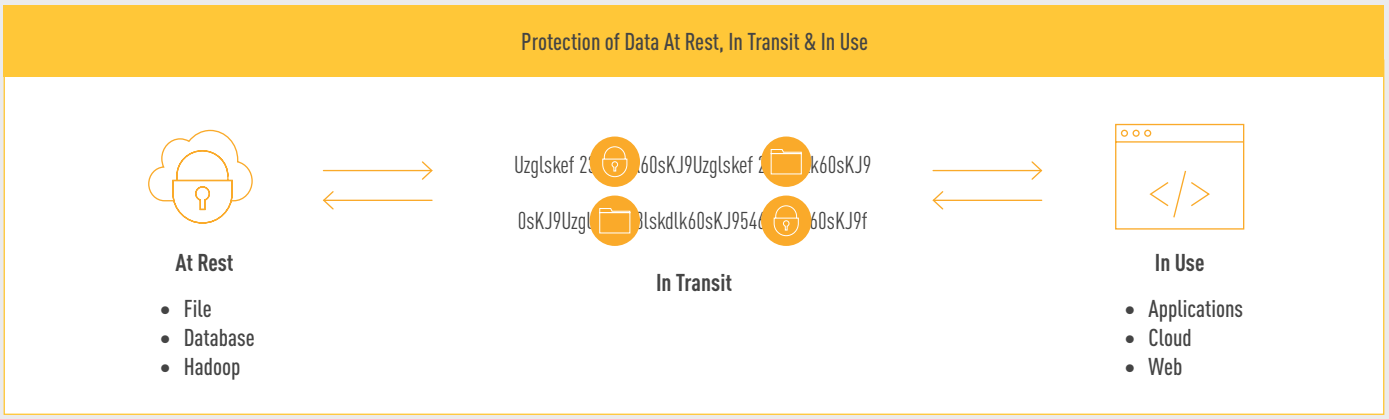
Protecting data in transit is most often the domain of network-traffic encryption protocols, such as SFTP, HTTPS, SS and TLS. Field-level protection, such as tokenization or encryption, can also be used to add a layer of security as data flows between systems. Field-level protection also serves as protection in transit when encryption protocols are not used, such as when data moves within a company's perimeter. For example, a credit card can be protected with tokenization and with TLS protocol. If TLS is terminated by a web application firewall, for instance, data is still protected in its tokenized form.

Protection Data in Use

Organizations also need to protect data that's in use, say, for business transactions or analytics. Safeguarding data in use means following the NIST "Least Privileged" principle that suggests only a minimal amount of data needed to perform a business function is delivered to:

- Authorized users who should be able to access sensitive data in the clear, or
- Privileged users, such as system administrators, database administrators (DBAs), and IT staff members who 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; instead, they should be created by a security officer. This is a critical NIST principle known as separation of duties.



Consistent Data Security Policy

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

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

The Best Data Security Method for Your Organization

In most cases, choosing the best method of protection is determined by the way your organization uses the data. Most of the time, you're probably handling data in a non-production environment, and that calls for SDM.

Transactional business systems such as claims processing and on-line banking transactions require a protection method that can be reversed. Because encryption lacks transparency and requires changes to business systems, it's best to avoid that approach with transactional systems. Instead, consider using pseudonymization or tokenization, which best meet the requirements of transactional systems.

For archiving, data sharing, or analytics, the preferred method will depend on whether the use case requires re-identification or un-protection of data. If neither is called for, anonymization— redacting all identifiers and masking of quasi-identifiers—fits the bill. If re-identification or unprotection is, in fact, needed, pseudonymization or tokenization 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	High	Not Recommended	High	Low
Applicability for PHI	High	Highest	Not Recommended	High	Low

Security in the Cloud

The limitations of traditional IT architecture are prompting organizations to shift workloads to hybrid- and multi-cloud platforms. Cloud use means that data is no longer static in isolated siloes that are easy to protect. Data is constantly moving between digital business processes and applications that are on-premises and in the cloud.

As organizations increasingly move applications and data to cloud environments such as Snowflake, Yellowbrick, AWS, Microsoft Azure, Google, and IBM, it is critical to understand that cloud vendors advocate what is known as a "shared responsibility model." Essentially, while vendors offer highly secure cloud environments, the onus of data security falls on customers. Users are responsible for protecting their own cloud data.

That's why it's even more critical for organizations to turn to data-centric security to ensure their always-moving information is protected from end to end. Data-centric security safeguards the data itself, so that you can protect or de-identify sensitive information on-premises, and then move it to the cloud and un-protect or re-identify it on site.

To answer what is undoubtedly a burning question: Data-centric security integrates with object stores, auto-scaling containers, serverless functions, cloud HSM, and other cloud offerings.

Your Data-centric Security Solution Checklist

Protecting sensitive data throughout its lifecycle, on-premises, and in the cloud delivers the following functionality:

- A single, centralized solution that 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 de-identification
- Tamper-proof monitoring, logging, reporting, and auditing
- Comprehensive data-type support
- Proven success in production environments
- Flexible, frictionless integration with on-premises and cloud systems.

Get Ahead of Security Expectations

It's clear that industry, local government, and cross-border regulatory requirements for dataprotection will continue to evolve, creating new security expectations and introducing new complexities. Still, companies shouldn't fully protect data because it is mandated. Protecting customer data has become the defining element of customer trust. "Good privacy is good business," said Alastair Mactaggart, the force behind California's landmark data-privacy initiative.[2]

An organization's most valuable assets must be protected throughout the enterprise and customer data privacy must be ensured. Of course, breaches will happen, malware will invade, and authorized users will make mistakes with security controls—but if sensitive information is itself protected, it will remain safe.

Choosing the right protection technology for your organization is a priority because of the exponential rise in data, data sources, and platforms available on-premises and in the cloud. Remember, Protegrity Vaultless Tokenization (PVT) solves the many challenges of data security. PVT is a highly scalable, flexible, and powerful protection method for your structured and semistructured data, no matter where it is and what it is doing. Organizations and businesses can't make full use of data if it isn't properly managed and sorted. And you definitely can't if the data isn't secure. Hopefully, you're on your way to protecting your and your customers' data and strengthening the bonds of trust.

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

PROTEGRITY

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 Salt Lake City, Utah, 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.

1165 E Wilmington Ave., Suite 200
Salt Lake City, Utah 84106
Phone: +1.203.326.7200

Protegrity EMEA

1 St. Katherine's Way
London, E1W 1UN
United Kingdom
Phone: +44 20 7113 3730

Protegrity Asia Pacific

Level 6 Republic Plaza 1
9 Raffles Place
Singapore 048619
Phone: +65 9130 9618