comforte

Securing Big Data Analytics

On-premises, in the Cloud, or Both – How to Find the Right Strategy

© 2019 comforte AG. All rights reserved.

Contents

Securing Big Data Analytics

Security as a Catalyst for Big Data Analytics

Big data delivers big insights, with seemingly endless use cases and benefits – from improving how companies market to customers to stopping attempted financial fraud, from tailoring cancer treatment for better outcomes to reducing the number of traffic accidents.

Not only are the uses and benefits expanding, the environments where big data is collected and analyzed are expanding too. In fact, big data is literally everywhere today – on-premises, in the cloud, streaming from sensors and devices, and moving across the internet. Increasingly, some of that data, including sensitive personal information, is ending up in the hands of cybercriminals who sell it to other bad actors for use in myriad malicious activities.

As enterprises use sophisticated analytics to power greater profitability, effectiveness, competitive advantage and shareholder value, the burning question becomes:

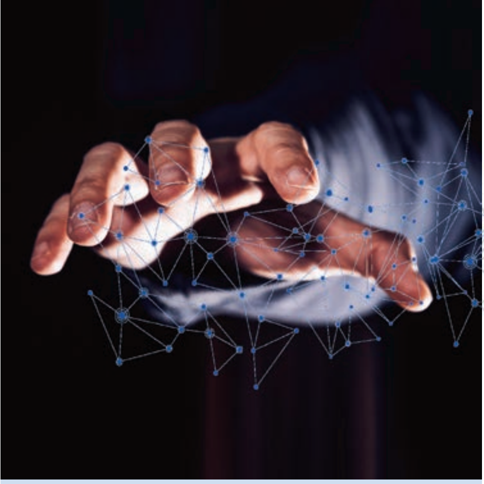
"How can our enterprise make data available to data scientists and other business users while complying with privacy regulations and protecting sensitive, confidential information from prying eyes?"

While perimeter, network and storage security solutions such as identity and access management, vulnerability management, data leakage prevention, and web application security can successfully prevent most cyberattacks, no solution is infallible. The reality is that these solutions cannot stop every attempted breach, which ultimately puts sensitive data at risk.

There's only one way to prevent cybercriminals – and insiders – who have gained access behind the firewall from accessing and exfiltrating sensitive data: adopt a data-centric security approach. This white paper explains how data-centric security using technologies such as tokenization can be the catalyst for big data analytics by eliminating the risks and disadvantages of typical security solutions.

78% of organizations experienced a successful cyberattack in the last 12 months.

Source:
"2019 Cyberthreat
Defense Report,"
CyberEdge Group, LLC,
March 2019



Big Data is Everywhere, But Security Isn't

There's no disputing that the cloud is one of the best things that ever happened in terms of advancing and accelerating innovation across nearly every industry. For big data analytics, cloud computing and cloud data storage give organizations a flexible, scalable and cost-effective way to store and analyze massive volumes of data.

While cloud computing has been a boon for big data analytics, it has also expanded the enterprise's attack surface and increased the risk of data leakage. Despite the effort to inform cloud users of the shared responsibility model common across the industry between cloud service/application providers and their customers, many enterprises don't realize that they ultimately must take steps to protect their data in the cloud beyond turning on the minimal viable security provided by cloud vendors and other third parties. Not only is data no longer solely stored and analyzed on-premises, it's also constantly moving between different environments, databases, and applications, which can be on-premises, in the cloud or a combination of both. Often, data in motion is not well-protected and too frequently it's not protected at all.

No matter where the data goes and resides, its security represents one of the biggest obstacles to companies looking to reap the benefits of big data analytics. When enterprises cannot reliably and consistently secure sensitive data used in big data analysis, it often leads to one or more of the following negative outcomes:

- **Access is refused:**
  The part of the organization that needs the data is denied access because it contains sensitive elements that the company is unable to protect.

- **The company risks non-compliance:**
  Should the enterprise allow access to the data without proper protection, it may face penalties for non-compliance with data privacy regulations, including the Payment Card Industry Data Security Standard (PCI DSS), the Health Insurance Portability and Accountability Act of 1996 (HIPAA), and the European Union's General Data Protection Regulation (GDPR).

- **A breach occurs:**
  Should a breach happen, it can cause extensive damage to the enterprise in the form of mitigation costs, penalties, revenue loss, loss of share-holder value, and customer churn.

- 1.7 MB of data will be created every second for every human being by 2020
- 84% of enterprises have launched advanced analytics and big data initiatives to improve accuracy and accelerate decision making
- 73% of global businesses performed big data processing in the cloud in 2018
- 10 million Internet of Things devices will be in use by 2020
- $3.86 million is the global average cost of a data breach

*Sources:
Domo, NewVantage Venture Partners, Qubole, IOT Analytics, IBM Security and Ponemon Institute*

Why Current Security Solutions Aren't Protecting the Data

As they rightly should, popular security solutions found across most enterprises focus on keeping intruders out. However, even with the addition of artificial intelligence, machine learning, next-generation capabilities and other advances, these solutions can't stop all attacks.

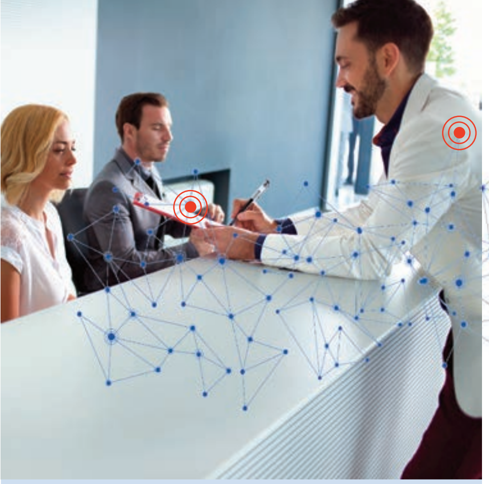
Since it's inevitable that a breach will occur at some point, enterprises have relied on solutions for discovering cyberattacks and mitigating the damage as quickly as possible. Despite these efforts, an attacker often has the luxury of weeks or months to access and exfiltrate data before an organization even discovers the breach – let alone begin to mitigate it. In its annual report, Verizon states that the time it takes cybercriminals to compromise a system is minutes or even seconds, while 68 percent of breaches take months or longer to discover.

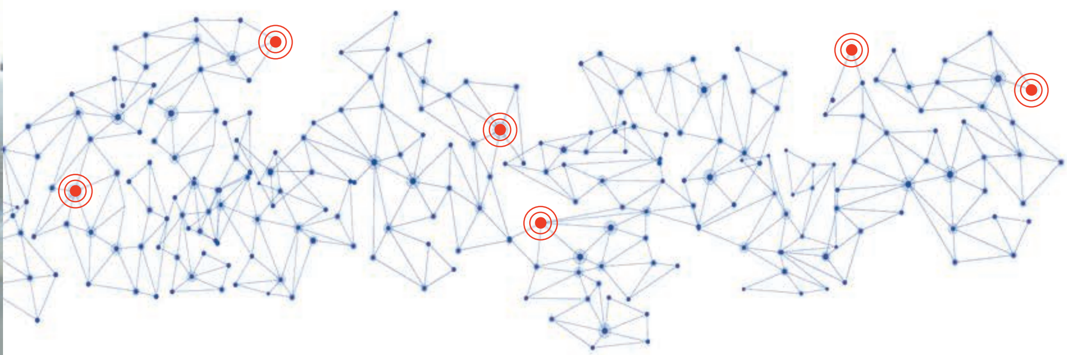
When detection and mitigation fail, the next line of protection has typically been classic data encryption. The idea was to encrypt data stored in databases so that any attacker gaining access to it wouldn't be able to use it. However, there are two major gaps in this approach that cybercriminals readily use to their advantage:

Encryption keys can be stolen: Many companies make it all too easy for attackers to find and access encryption keys by neglecting to secure them properly, using the same key for all of the company's data, not changing the key periodically, or making many other mistakes common with key management.

It only protects data at rest: Applications and data live and travel everywhere, across on-premises systems and the cloud. What happens to data that is unencrypted as it travels between applications or when it's being used by an application? Encryption of data at rest does not protect it when it is unencrypted and transmitted somewhere else.

While not a security gap per se, another disadvantage of classic encryption that keeps organizations from using it effectively is that it typically adds an unacceptable level of complexity to the IT environment. Often, implementing classic encryption requires significant modifications to existing applications and systems. In light of the ever-increasing number of data breaches and the shortcomings of classic encryption, it's clear that popular security solutions aren't the answer to effectively protecting sensitive data used in big data analytics.





Classic encryption fails to protect hotel's data

In the Marriott-Starwood big data breach, sensitive information on 500 million guests was stolen over a period of four years. While the data was encrypted, the company couldn't rule out that the attackers didn't steal the encryption keys as well as the data. The cybercriminals even employed their own encryption of the data from the hacked database to avoid being detected by data loss prevention tools during exfiltration.

Source: "Marriott: Data on 500 Million Guests Stolen in 4-Year Breach," KrebsonSecurity, November 30, 2018

A Data-Centric Security Strategy for Big Data

Given the obvious inadequacy of perimeter defenses, breach detection, and classic encryption, companies need a new approach to security that protects sensitive data when all else fails. That new approach is data-centric security, which protects the data throughout the entire data lifecycle, going wherever the data goes to provide strong protection without affecting data usability.

Instead of focusing on protecting the perimeter, network, endpoints or application, data-centric security prioritizes datasets to protect the data itself. It protects big data everywhere – for example, while it's in use for analytics, when it's in motion between on-premises data stores and the cloud, and when it's at rest.

It also protects the individual data elements wherever possible. That means if a dataset contains a mix of sensitive data such as personally identifiable information (PII) along with other data that is not sensitive or regulated, a data-centric security strategy protects the data at the individual element level.

This fundamentally different approach to protecting sensitive data is based on three pillars of capabilities:

- Data discovery: Discovering sensitive data without the need to make source code changes to business applications or big data analytics environments

- Data classification: Classifying sensitive data to determine the appropriate protection strategy required

- Data protection: Using mechanisms such as tokenization to render sensitive data useless to cybercriminals while maintaining usability of the data

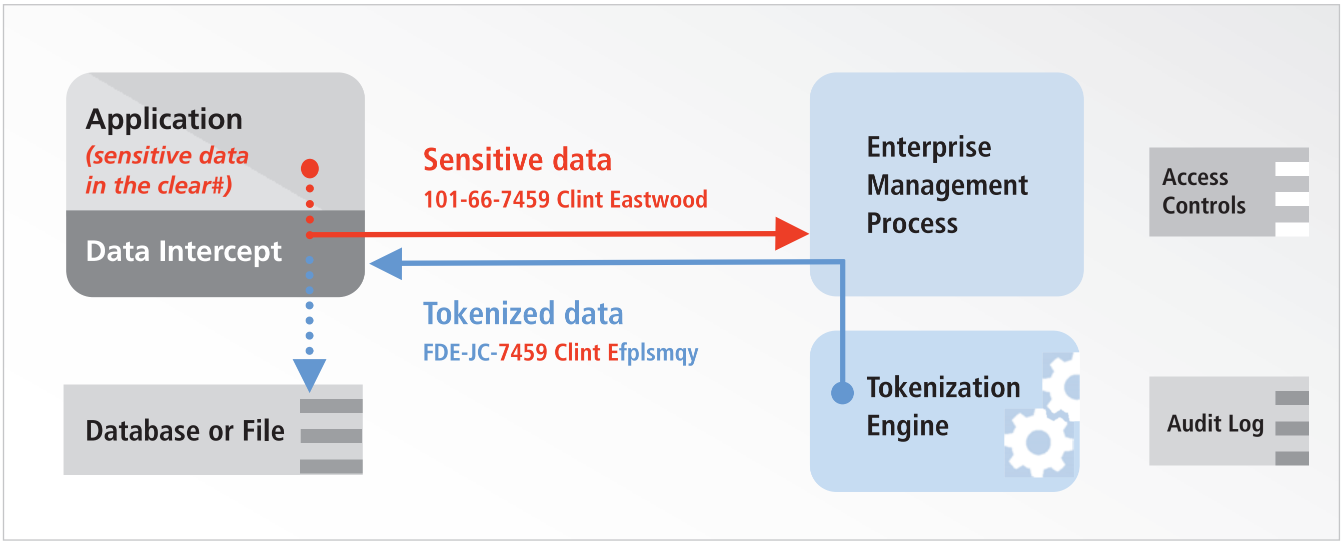
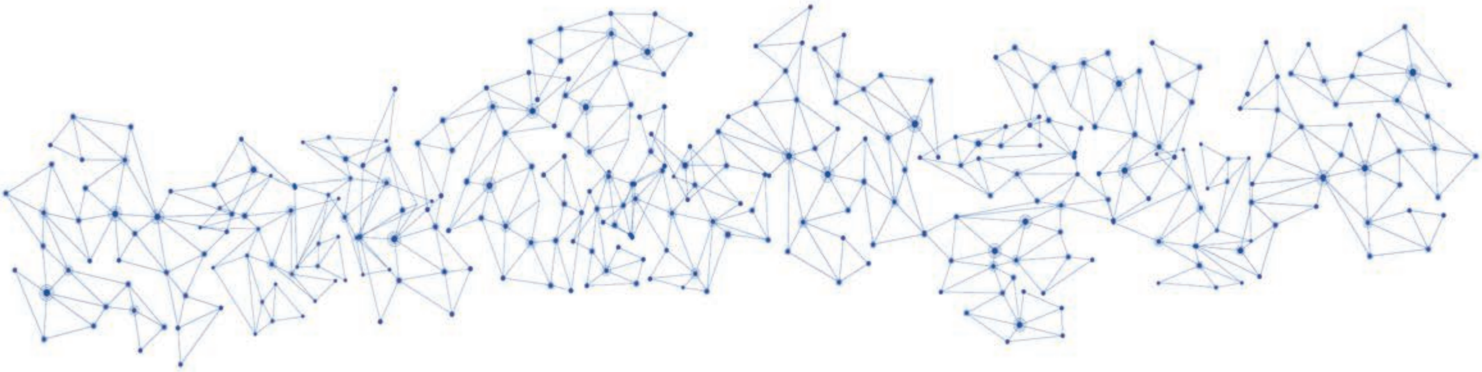


Figure 1. Data-Centric Security Approach

Essential Data Protection Mechanisms for Data-Centric Security

Protection mechanisms such as tokenization address the shortcomings of classic security solutions and are essential components of a data-centric strategy. These data protection mechanisms protect sensitive data while preserving its original format, giving it referential integrity and resulting in a dataset that is the same size as the original. The de-sensitized data has the identical statistical distribution as the original data to ensure that all the characteristics and properties of the dataset are preserved. This eliminates the dilemma of having to choose between either security or analytics as data scientists are able to perform analytics and produce reports on the data while it's still protected.

Data Protection Mechanisms	How They Work
Tokenization	Tokenization replaces the original data with a randomly generated, unique placeholder. There is no mathematical relationship between the token and the original data, so hackers cannot reverse-engineer it.
Format Preserving Encryption (FPE)	Similar to tokenization and unlike classic encryption, format-preserving encryption (FPE) encrypts the data in such a way that it maintains the same format as the original data.
Masking	Data masking anonymizes sensitive data by creating a structurally similar but inauthentic version of the data. Unlike tokenization and FPE, masking is permanent; that is, it's impossible to reverse it to obtain the original values.



Essential Data Protection Mechanisms for Data-Centric Security

Figure 2 shows an example of how tokenization removes confidential data from internal systems and big data environments by replacing it with randomly generated data of no exploitable value to cybercriminal

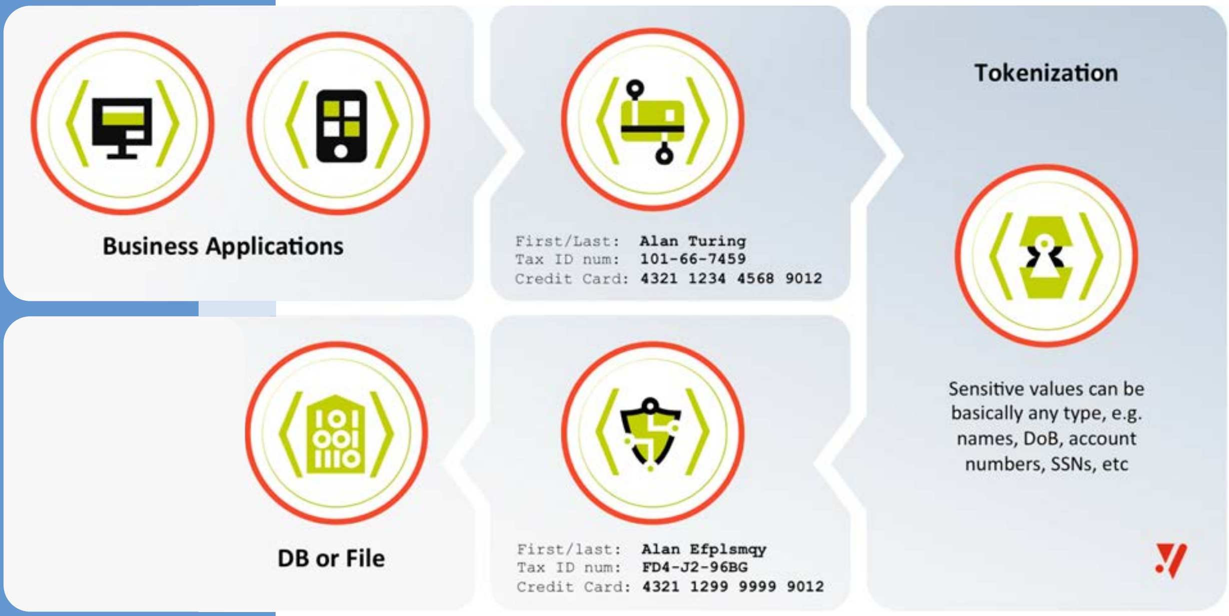

Business Applications
First/Last: Alan Turing
Tax ID num: 101-66-7459
Credit Card: 4321 1234 4568 9012
Tokenization
Sensitive values can be basically any type, e.g. names, DoB, account numbers, SSNs, etc
DB or File
First/last: Alan Efplsmqy
Tax ID num: FD4-J2-96BG
Credit Card: 4321 1299 9999 9012

Figure 2. Tokenization Replaces Sensitive Values with Non-Sensitive Ones

By adopting a data-centric security strategy, enterprises can:

- Protect sensitive information within big data analytics environments without impacting the ability to use the data in existing applications and systems
- Comply with regulatory mandates without prohibiting or restricting access to certain datasets containing sensitive information
- Prevent costly and reputation-damaging data breaches

A Checklist for Choosing the Right Solution for Big Data Environments

As data-centric security gains mainstream recognition for its ability to protect sensitive big data from theft and malicious use, enterprises are faced with choosing a solution from an increasing array of options. However, not all solutions that claim to be data-centric are designed for the demands of a big data analytics environment, one where scalability, performance and availability are crucial to support the throughput requirements of today's – and tomorrow's – analytical workloads.
To protect big data effectively, insist on a solution that offers:

Linear scalability:
As big data environments increasingly include analysis of streaming, real-time data, you need a security solution that can keep pace. To make the most of your data, whether it's real-time or historical, your data-centric security solution should easily scale as your workloads scale, with no performance impact.

High performance:
More business intelligence solutions are now using artificial intelligence capabilities and taking advantage of in-memory data processing. To keep up with the speed and performance requirements of these systems, look for a data-centric security solution that delivers high performance with features such as intelligent streaming and load distribution.

High availability:
Security shouldn't impact the availability of your big data environment. Choose a solution with built-in fault tolerance so that any unexpected failures are resolved automatically without interruption of service.

Furthermore, big data analytic environments are evolving, making it essential to choose a high-performance solution that is not only data-centric and goes everywhere the data goes, but one that can easily adapt to new technology and infrastructure as it emerges. Look for a data-centric security solution that provides:

Flexibility:
As mainstream, big data technologies such as Hadoop and Spark become legacy systems and new solutions and infrastructure emerge, your data-centric security solution must adapt easily to these changes.

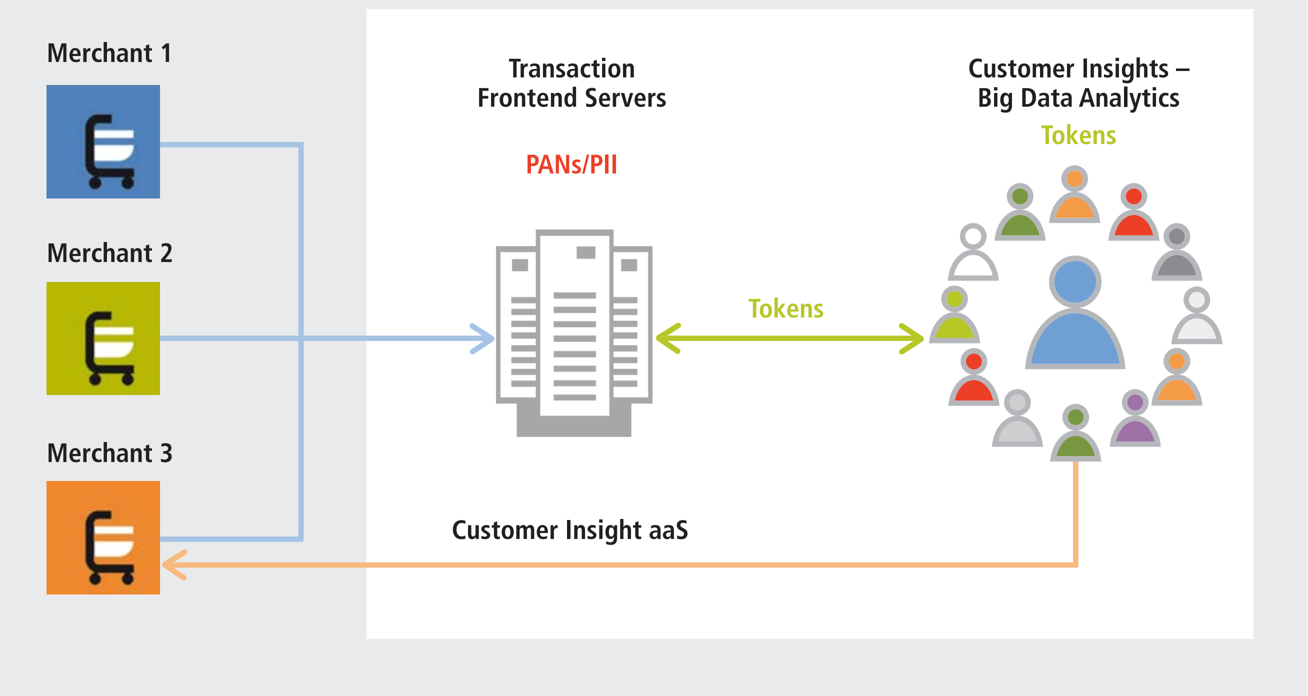
Support for multicloud and hybrid environments:
You need a solution that supports your current and future big data analytics environment, whether that's on-premises, in the cloud, or a hybrid of both.

Both native and API-based integration:
To minimize implementation effort and avoid the cost and time to change existing applications, look for a solution that offers flexible integration with big data analytics tools and platforms.

Tokenization at Work in Big Data

In the retail industry, big data can yield insights that help companies better understand their customers, improve customer loyalty, increase promotional effectiveness, acquire more customers, increase supply chain efficiency and much more. Using data-centric security, tokenizing sensitive data lets analysts extract insights without the risk of exposing personal, confidential customer data.

CUSTOMER INSIGHT WITH SECURE BIG DATA ANALYTICS



A Final Word About Protecting Sensitive Data Wherever It Goes

Don't let the threat of a data breach stop your big data analytics efforts before they can deliver transformative insights for your enterprise. Data-centric security does what classic security solutions cannot: protect sensitive data wherever it goes.

With its proven data-centric security solution, comforte can help you make security a catalyst for big data success. comforte's data protection suite, SecurDPS, was built from the ground up to best address data security in a big data world.

Learn more about data-centric security for big data analytics environments. Get more details at: https://www.comforte.com/enterprise-data-protection/big-data-security/

About comforte AG

With more than 25 years of experience in data protection on truly mission-critical systems, comforte is the perfect partner for organizations who want to protect their most valuable asset: data. comforte's Data Protection Suite, SecurDPS, has been built from the ground up to best address data security in a world that is driven by digital business innovations, empowered customers and continuous technology disruptions.

We are here to enable your success by providing expertise, an innovative technology suite and local support. To learn more, talk to your comforte representative today and visit www.comforte.com.

© 2019 comforte AG. All rights reserved.