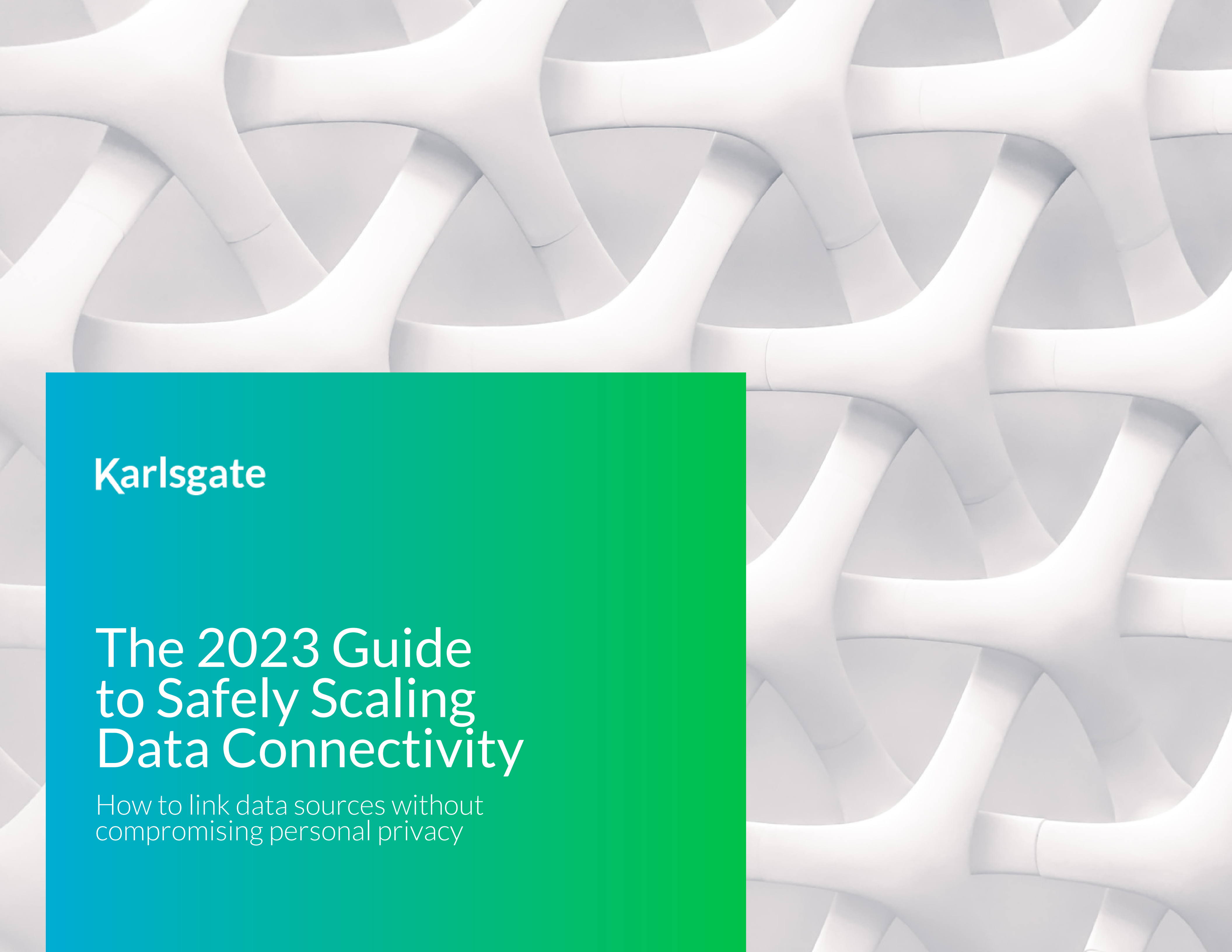


Karlsgate

The 2023 Guide to Safely Scaling Data Connectivity

How to link data sources without compromising personal privacy

Introduction

Data has always been at the heart of why and how organizations do what they do. From healthcare, to finance, to government, to media and beyond, global organizations leverage data to develop solutions and services, drive decisions, and effectively engage with constituents. In fact, recent advancements in technology are rapidly increasing both the amount of data used and the way in which it is used.

Artificial intelligence (AI) and machine learning are being implemented at a record pace, and more and more organizations are finding the need to share data and insights with each other. Data has become an asset, a commodity, and unfortunately, also a potential risk for organizations.

So, how do we improve and enhance the speed and accuracy of using data in a way that truly protects privacy? Now more than ever, there's a need for a paradigm shift.

43%

of data integration projects don't even get completed because there isn't enough IT capacity

Did you know?

44%

of projects don't get done because the security and compliance requirements make the process for connecting data too cumbersome

What's Inside:

- A breakdown of the data-sharing burden
- Can you safely scale data while also simplifying it?
- A look into the current state of secure data exchange
- Privacy regulation deadlines for 2023
- Why traditional data-sharing methods aren't giving you as much protection as you think
- How to step up your data-sharing strategy by adapting a zero-trust approach

Table of Contents

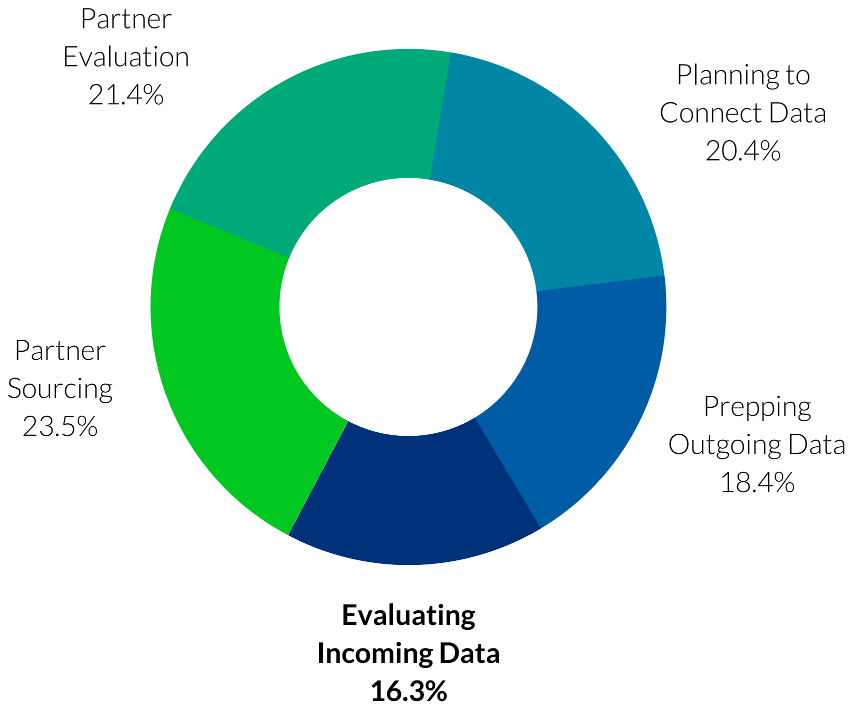
The Hassle of Secure Data Collaboration

From government agencies, to corporations, to healthcare organizations and everyone in between, accessing and exchanging valuable data has always come as a burden. Whether searching for data from an outside vendor, or looking to connect internal systems, the process to connect data from different sources isn't simple. In fact, the average data sharing or data integration project takes over two months (and anecdotally we hear it's even longer).

In addition, lengthy privacy reviews, evolving security regulations, and working to align processes for each new data partner significantly slow down the time and increase the cost of exchanging data safely. And that's all before insights have even begun to be shared.

Thus, projects fail much of the time – 52% of the time to be exact.[1]

Less Than 17% of the Time Spent on Data Integration Projects is Focused on Evaluating Data



Scaling Data Connectivity While Handling Real-World Complexity

As the need for more data from more partners rises, so does the need for methods of data exchange that simplify the process. Finding and vetting data partners can take months, and it's hard to trust data partners not to mishandle or misuse your data. Before you can even evaluate the benefits of connecting data sources, preparing to connect data takes a great deal of manual effort. Many of the steps are lengthy, costly, and cumbersome. Once the data is approved, there's still concern about how to make regular updates, and if the data doesn't pass the evaluation, you must begin the time consuming, lengthy process over from the beginning.

According to IT leaders, the top complexities of data sharing and integration projects are:

- Alignment on security protocol
- Alignment on file formats
- Alignment of data element normalization/standardization
- Downtime associated with data refreshes
- Custom development/coding required for each integration

When we look toward the future of data collaboration, we see a major push toward interoperability and a focus on digital transformation. With this also comes the need for more precise and accurate data from more (and more varied) sources, but real-world complexity has historically been a roadblock to achieving this.

From the processing and storing of data, to data normalization and standardization, to aligning with external partners on file structures, the number of steps required before data can even be evaluated can hold projects back – or keep them from ever getting started. It's more critical than ever to streamline these steps and embrace emerging technologies to easily scale data connectivity and enable collaboration with more partners more quickly, and easily, than ever before.

With all of that said, as mission critical as scaling data connectivity is in today's data landscape, there's still another, major component to consider: Scaling the data collaboration processes are only worthwhile if you can also ensure ultimate protection of personal information. The data sharing process itself is only as secure as the technology and safety measures in place to not only comply with restrictions and legislations put in place to safeguard personally identifiable information (PII), but to also truly safeguard personal privacy by maintaining custody of sensitive information and keeping it out of the wrong hands, whether accidentally or maliciously.

The State of Data Privacy

When it comes to privacy concerns, you only have to look in the news and you'll find an almost constant stream of stories about data breaches, privacy fines, and evolving legislation to protect consumer identities. New data shows that 33% of consumers globally have already become victims of a data breach.[2] In 2022 alone, the U.S. Department of Health and Human Services reported nearly 600 healthcare data breaches affecting more than 40 million individuals. And this is not a new trend. In 2021, Gartner reported that 211.4 million identities in the U.S. were affected by data breaches across all industries.

Why are breaches so common? It's no secret that data-driven decision making is important for businesses across all industries. But with sharing and storing data comes risk, and this risk can mean big business for hackers.

There is already legislation in place at the state, federal, and international levels to ensure companies are compliant in protecting consumer identities, and this legislation continues to evolve year after year.

Some new and updated legislation for 2023 includes:

Jan 1, 2023

California Privacy Rights Act (CPRA)
Virginia Consumer Data Protection Act

July 1, 2023

Connecticut Data Privacy Act (CTDPA)
Colorado Privacy Act

December 31, 2023

Utah Consumer Privacy Act

These strict guidelines aspire to prevent PII from being shared or accessed irresponsibly. However, while most companies are protecting their data to the best of their ability, a study by Positive Technologies reports that with current data-protection methods, cybercriminals can still penetrate 93% of company networks.

Ultimately, unless organizations evolve their data protection methods for all parts of the data lifecycle, data is still being left vulnerable to re-identification and breaches. If we want to better safeguard PII, it’s time to change our approach.

84%

of IT leaders agree that they'll need to make changes to their data practices to comply with changing privacy regulations over the next two years.

79%

of IT leaders say privacy guidelines slow their data strategy plans.

85%

of IT leaders say privacy concerns, and the associated obstacles, have impacted their ability to share data both internally and externally.

The Evolution of PII Protection

Changing the way privacy is managed begins with understanding the different ways in which data has traditionally been protected. From encryption and hashing to the more recent introduction of data clean rooms, the many options for securing data all provide some of level of protection. Yet none of these methods alone fully protects data without processing complications or limiting functionality. So, what's the best way?

The reality is that it depends on your goals. So, let's narrow our focus to the goal of protecting consumer identities while solving the "linkage problem." The critical linkage problem can be defined as:

Two independent entities (public or private) are each managing a dataset about individuals. The understanding of each individual's identity is achieved using various identifiers such as name, postal address, email, and/or social security number. However, these components of personal data are sensitive and are tied to personal privacy rights, regulatory restrictions, and/or ethical handling concerns.

The problem lies in how to enable the 2 independent entities to share the understanding of the individuals in common between the 2 datasets without sharing any personal data and without inadvertently allowing reidentification of those individuals not in common (i.e., outside of the desired intersection).

Now that we have a focus, let's evaluate the evolution of protection methods.

Evolving Data Protection to Keep Up with Modern Needs

There was once a time, before privacy concerns and regulations were prevalent, in which sending clear text was the norm – and you may be surprised, but it does still happen today. Rising concern then led to the introduction of encryption, a method which protects data well while in transit – but requires that you give custody of a copy of your data and supply the decryption key to your partner. This makes it fully re-identifiable once the recipient decrypts it and leaves unnecessary, residual data in your partner's environment; the identities and attributes of individuals that were not common to both you and your partner's file.

Hashing then came into play, scrambling data in a way that's very difficult to reverse. However, this still leaves data changing custody, which can lead to future re-identification attempts against an identity graph. Also, hashing is still considered identifiable data and therefore isn't GDPR-compliant, meaning in today's data ecosystem, it's still not enough.

More recent innovations in privacy-enhancing technology include federated learning, differential privacy, and fully homomorphic encryption. While each is powerful for use in analysis, modeling, and data obfuscation, none of these methods adequately addresses the data linkage problem – a cause for many challenges when it comes to data collaboration. What we're still missing is the ability to adequately protect PII, while also having a meaningful impact on sharing insights on matched identities.

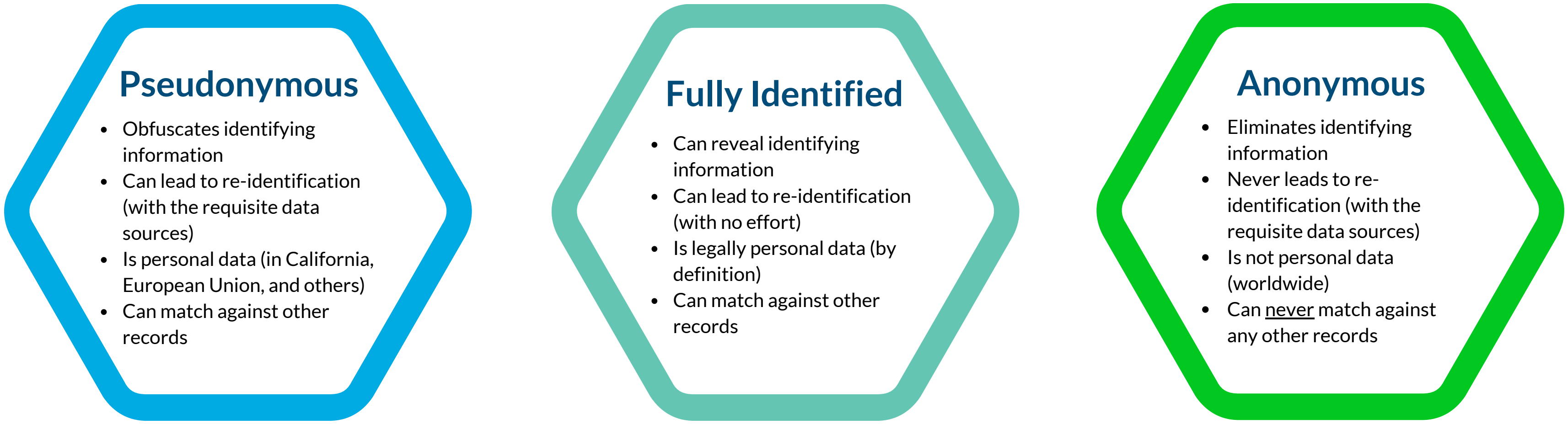
Even data clean rooms, which have become more common today, require that both partners agree on which clean room solution to use and then allow an additional third party to gain custody over all datasets used for matching. It's a full-sharing event with consent and security obligations, subjecting PII to a level of risk.

In addition to the methods described above many industries leverage the use of de-identification, a method of protection intended to protect data at rest, to try to solve linkage problem for data in transit and use. De-identification of data refers to the process used to prevent personal identifiers from being connected or linked to a particular consumer, either directly or indirectly. There are many different de-identification techniques which represent a broad spectrum – from relatively weak to very strong techniques that can effectively eliminate privacy risk for data at rest.

In healthcare, they call it tokenization. In other industries, they create "IDs". Regardless of what it's called, de-identification is a best practice across multiple industries as a privacy-enhancing tool for protecting data at rest. When done well, it can help companies meet their obligations under privacy regulations like CCPA, GDPR, HIPAA, and more, and build trust in the data governance practice. Beyond that, de-identification has other advantages:

- Data is still usable at an individual level by other groups within the business
- Data is usable across multiple verticals from healthcare to retail for both research and marketing purposes
- Demonstrates a privacy-by-design approach to legal entities should a breach occur
- Significantly reduces risk and impact of third-party data breaches

How to Classify Each State of Identity



Pseudonymous

- Obfuscates identifying information
- Can lead to re-identification (with the requisite data sources)
- Is personal data (in California, European Union, and others)
- Can match against other records

Fully Identified

- Can reveal identifying information
- Can lead to re-identification (with no effort)
- Is legally personal data (by definition)
- Can match against other records

Anonymous

- Eliminates identifying information
- Never leads to re-identification (with the requisite data sources)
- Is not personal data (worldwide)
- Can never match against any other records

However, when it comes to the linkage problem, the only way historically to use de-identification is to have persistent pseudonyms to link, which are considered PII under many of the privacy regulations, or send your data through a 3rd party for processing and matching with partners, which relinquishes control of your data and opens up the risk of re-identification.

As we usher in the “Protected Data Age,” a new era in which harnessing the power of data no longer means sacrificing the privacy of individuals, it’s time to acknowledge the current limitations with legacy data protection methods and embrace a better way to safely share data.

How a Zero-Trust Approach to Security Can Improve Your Data Collaboration

The current rate of activity for bad actors means that we need to assume a breach will be attempted, and therefore share data in a way that eliminates the associated risk altogether. This is known as a zero-trust approach.

The concept of zero trust is based on a very real understanding that nearly every digital interaction is risky by nature. Zero-trust frameworks can be applied to nearly every aspect of an organization’s cybersecurity approach. It means that devices and networks are validated, users verified, and access to data and files is strictly controlled, limited to who needs access and when. It also means setting up data sharing in a way that keeps networks safe and prevents the need for identifiable data to leave the secure environment.

Data sharing often means the data is only as secure as the organizations with which we share; zero trust means we can share without exposing our data to someone else’s risks. With many of the existing methods for sharing data, organizations lose control of their data and re-identification is possible. It’s not the strongest protection against data loss or leakage.

In contrast, emerging methods of safeguarding PII, along with new privacy enhancing-technologies, combine a robust process of securing data. One such innovation leverages partitioned knowledge orchestration involving two partners that want to share insights and a third-party facilitator that never accesses any identifiable or usable data. This approach enables partners to match and identify overlaps of their data without ever losing custody of any of their data. It also enables sharing of insights but only non-identifiable attributes and only on matched records. This creates a zero-trust framework for data sharing and blocks re-identification. Layering in best practices associated with salted hashing and encryption makes it so that neither sharing partner nor the blind third-party facilitator can make use of any identifying elements that they did not already have. Further, because data partners only learn new insights about individuals for whom they already had data, there's no residual data left behind.

A matching process like this, without disclosure of PII, whether raw PII or pseudonyms, eliminates the largest and otherwise unavoidable risk vector associated with sharing data: the dissemination of identities. This means that a business risk profile is no longer an aggregation of all of the risk profiles for each and every data partner with whom they work. In a world where there is a tidal wave of demand for data from novel, more extensive, and more comprehensive sources, while, at the same time, hackers are continuing to escalate their activity, it's essential that we continue to be looking for innovative approaches for connecting data safely and securely.

Technology is changing; how we deal with big data must change with it.

About Karlsgate

At our core, we're data scientists and technologists who, after more than two decades in the industry, have come to a major realization: the way companies are handling data simply isn't the best way anymore. It's slow, it's risky, and truthfully – it's outdated. With that in mind, we created Karlsgate to provide Privacy Enhancing Data Processing and Connectivity Tools to protect data at rest, in transit, and in use.

As we at Karlsgate embrace what we call the "Protected Data Age," we're changing the game. Gone are the days of slow data exchanges, vulnerable consumer information, and costly breaches. Data is more powerful than ever and harnessing it to drive business decisions should be quicker, easier, and safer than ever before, too. Founded in 2020 by a team of veteran data technologists and scientists, Karlsgate takes a zero-trust approach to data connectivity with Karlsgate Identity Exchange (KIE™), allowing the free flow of insights while maintaining control of sensitive information.

With Karlsgate, you can:

- Immediately save the cost and time to interoperate your data securely and compliantly across partners, customers, cloud solutions and internal systems.
- Eliminate the data security and compliance burden to effectively host, connect, and manage sensitive data, ensuring compliance with HIPAA, GDPR, CCPA, CPRA and more.
- Empower your business users to access and integrate data sources through a no-code, UI-driven system designed for their complex business needs.
- Automatically clean, correct, and standardize your data for integration with other data sources.

[1] According to data from a 2022 third-party survey conducted on behalf of Karlsgate.
[2] 2022 Thales Consumer Digital Trust Index

Karlsgate

karlsgate.com
contact@karlsgate.com
Follow us on LinkedIn & Twitter @Karlsgate