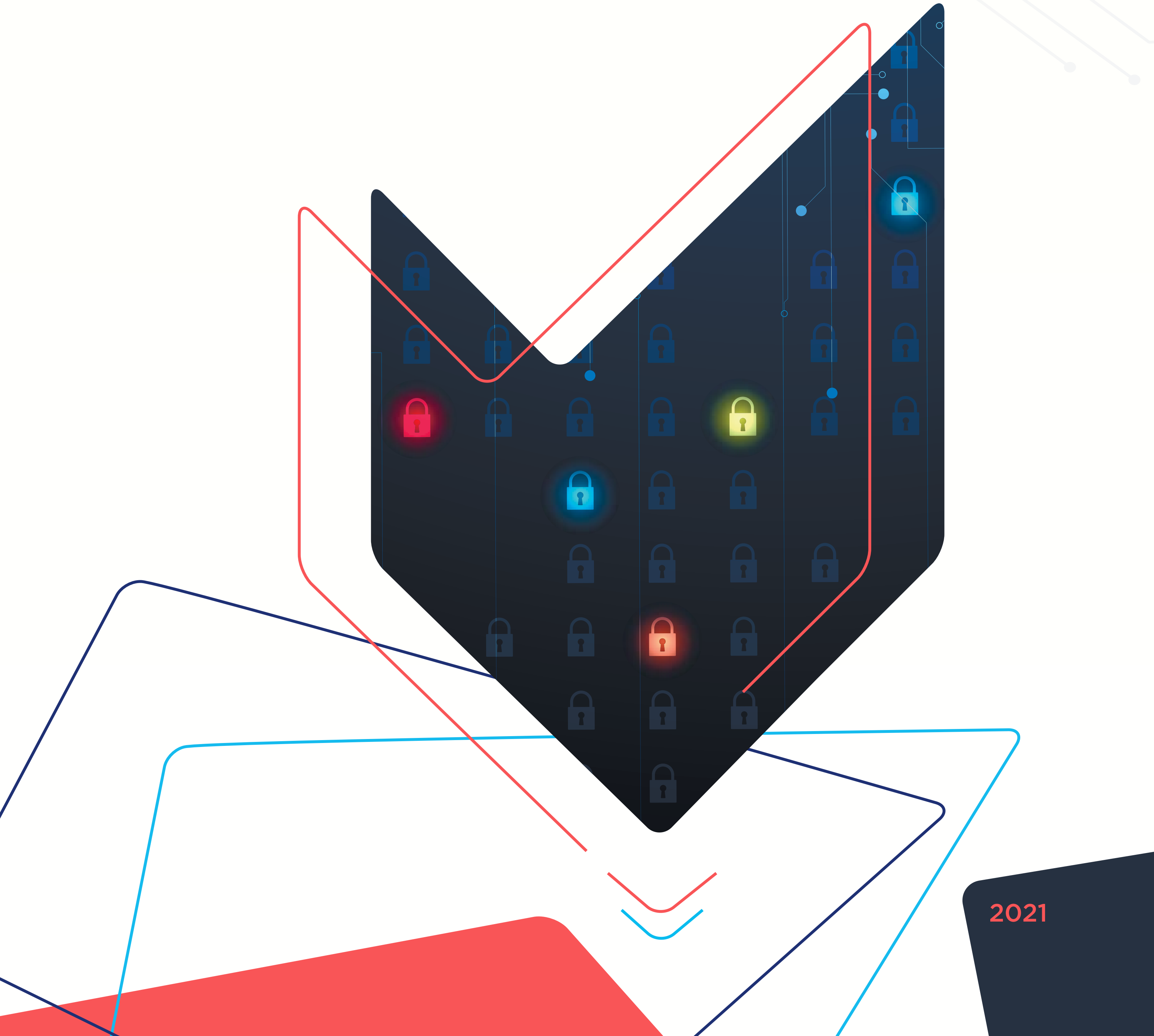
PKWARE

pkware.com

Rethinking Data Management Practices for Privacy Compliance



2021

Data privacy is among the most regulated areas of technology. Unfortunately, the concept is as vague as it is vital when it comes to exactly how to implement data privacy controls for the enterprise. Data privacy regulations remain notoriously difficult to put into practice compared to data security, which is supported by a growing constellation of standards and frameworks and other security best practices.

While security regulations are replete with references to technical controls and automated solutions as part of the compliance effort, data privacy regulations are written as if compliance can be met with paper, policies, contracts, and training. The problem is that while major regulations like the California Consumer Privacy Act (CCPA) and the General Data Protection Regulation (GDPR) were drafted to enforce enterprise compliance, they don't adequately address the crucial support role of automation in data privacy compliance.

That's a big and ironic challenge for technologists. Let's take a closer look at how we can deliver the right technologies to back up the compliance demands of evolving data privacy regulations.

Contents

The Foundation of Privacy: Know Your Data Subjects

One of the requirements in many data privacy regulations is for companies to share the data they maintain about data subjects with them upon request. This basic data privacy right is commonly referred to as a Data Subject Access Request (DSAR). Addressing DSARs is not new; it's been central for decades to data privacy regulations worldwide.

Data privacy regulations provide flexibility around the level of effort a company is expected to assume when searching for information to satisfy a DSAR request. However, companies' current practices stretch the exceptions' boundaries and often limit the DSAR response to searches in a handful of predetermined repositories. These practices are not sustainable for several reasons.

First, data privacy breach notification requirements are prevalent. Following an incident involving their data, individuals who had previously asked for a DSAR may discover in the notification of a breach that the response details they had received were incomplete. For instance, given the right of private action offered under the CCPA, California consumers can claim they would have taken a different action had they known earlier. If they'd known the actual extent of the data the breached company maintained about them, they might have asked for their data to be erased. Now that the personal data details are clear, they can sue for damages based on the incomplete DSAR response they previously received.

Second, breaches lead to audits by regulators. Audits raise scrutiny of a company's overall privacy practices, including how it responds to DSARs.

Finally and importantly, today's data subjects are more knowledgeable about the information companies collect and more educated on how to challenge a DSAR response that appears oddly slim or neatly curated.

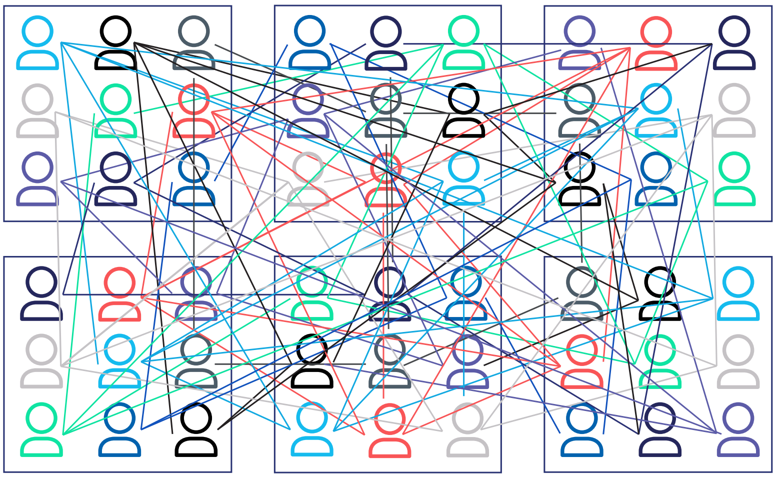
Companies face a constant challenge to manage personal information across what may be very disparate and complex enterprise systems and repositories. Years of insufficient data management practices make this task a technological challenge.

Matching Identities Across Tables

The inconsistent management of primary and foreign keys in databases for matching identities across tables is one example of an area that sorely needs improvement. Each table in a database has a key that allows its data to be matched with other tables' data. For instance, a company may have three tables in a database: one with its customers' contact information, the second with the transaction information of those customers, and the third with the customers' responses to satisfaction surveys. In the three tables, the customers are identified by their Customer ID number (CID).

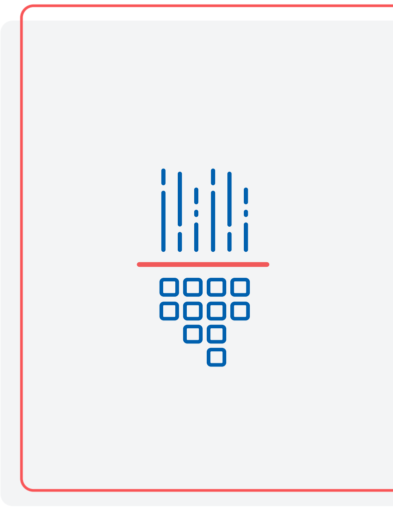
The CID is the primary key that links the three tables and allows the company to correctly connect the same customer's contact information, transaction, and survey response. Management of primary and foreign keys refers to the diligence of tracking how data subjects are identified on the database level—so that when a DSAR needs to be fulfilled, the company can search for the data subject's data efficiently across those various keys.

Many companies are facing the problem that each database could contain thousands of tables with different keys (or foreign keys) for the same person and no mapping that connects the primary with the foreign keys. Different tables may have been created by different database users concerned only with connecting the few tables they needed to use. The result is a tangle of connections and keys—artifacts left behind by untold numbers of business analysts—that make DSARs very hard to satisfy.



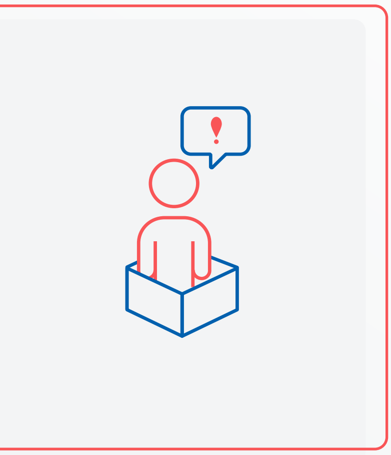
Naming Data Elements

The clear naming of data elements in database tables is another common challenge that needs to be corrected so that data subjects can receive a complete view of the data that companies process about them. This is especially relevant when the data is not self-evident at the outset. Machine learning and artificial intelligence, for instance, can figure out that a string of eight digits in a table's column represents a date. Still, no technology can figure out whether the date stands for account creation, last transaction, last customer service call, or any other possible activity without clear headers in the table. Traditionally, a group named the headers expecting that only a few database administrators would need to understand them. Privacy regulations require us to rethink this practice and establish new requirements to guide our database titles and overall data management.



"I Have 'Orphaned' What?"

When we speak of data subjects, we think of individuals who have transacted with the company in one way or another (employees, applicants, customers, visitors, or others). One of the dirty secrets of data management is that all companies also have many "orphaned identities" in their repositories. Orphaned identities are data subjects to whom a company cannot tie any processing activity or privacy commitment. Orphaned identities might include information added by employees who use company resources for personal use (the resumes of domestic workers, for example), information received in error from partners, tables with no keys, and more.



Orphaned identities remain data subjects, and the inappropriate disclosure of this personal information can still trigger breach notification obligations. The ability to find orphaned identities and correct (or eliminate) them is a data management best practice that will help reduce an organization's risk profile.

Set Standards, Establish Goals, and Reduce Risk

When it comes to unstructured data, such as documents, we need consensus on identifying an identity in personal data accurately. (e.g., is April Green a name of a person, street, company, or just a noun and adjective in proximity). The choice of technique to address such questions calculating the statistical confidence that a specific piece of data represents one thing and not the other.

Unfortunately, we do not have any standards that guide our use of statistical confidence when searching for personal information and responding to DSARs. Such standards are important to create consistency and provide the rationale whenever challenged by a data subject or regulator about existing practices.

The correct identification of data subjects across the enterprise will allow organizations to develop a complete and accurate inventory of its data subjects, a goal most companies have yet to achieve. Such an organized view of data subjects is not only the best way to address DSARs but is also the first stepping stone to address the other data management tasks companies are facing.

> When we became aware of PK Protect and the full capabilities of PK Privacy, what really clicked was the ability to find sensitive data in our structured and unstructured data sources, including Amazon S3 buckets, MySQL, MongoDB databases, and so on.
>
> —Tyrone Mills, CISO, Trōv

Obfuscating and Deleting Data Elements and Subjects

Privacy regulations require companies to limit the use and sharing of personal data and delete data subjects' personal information upon request. The mandate is in the direct call in Article 32 of the GDPR to mask personal information when using it. Its value is reflected in the fact that effective limitations on personal information use can save companies from fallout due to breaches and breach notification requirements.

Unfortunately, obfuscation techniques and the deletion of personal information are easier said than done, creating real data management challenges for companies.

When it comes to obfuscation to limit the personal data authorized users within the organization can access, many systems are just not built with the capacity to accommodate such functionality. Applications often allow some limitations on data views by creating different roles with access rights, but those are not as flexible or aligned with the company's privacy policies.

Obfuscating data in large repositories—whether that data is structured, unstructured, or semi-structured—also involves challenges such as shared drives full of documents and big data lakes.

Meanwhile, the deletion of privacy data, especially data that pertains to a particular data subject, presents a different technical problem, namely replication. Privacy data regularly moves between systems and is validated by comparing it to previous data. That interconnectedness means that when a line item, such as the record of a data subject, is unexpectedly missing, a replication error happens. When such errors occur, financial systems cannot balance their entries, and transaction records do not add up.

The Right Approach to Obfuscation and Deletion

Success involves several different components in addressing these data management challenges. Beyond the obvious need to process personal information in systems that can handle different obfuscation techniques, there is the important question of correctly identifying the privacy data that needs to be protected in the first place.

We can all recite a list of sensitive data elements that should only be made available in limited circumstances, but there are complexities. The reality of data sensitivity is that it's often the combination of several data elements that define the data's risk. The "formula" for effective obfuscation must account for the user's legitimate need for specific data elements, the risk associated with the combination of these data elements, and the minimum exposure of the data elements for the use at hand. Few solutions out there today address such a formula.

As for data deletion, we must engage techniques that address unreconciled data entries' challenges following a data subject's deletion. The most practical way to address this challenge is to use obfuscation to identify privacy data elements while simultaneously preserving necessary data for replication purposes, leaving that data intact.

One obfuscation technique that is especially useful for this purpose is Format Preserving Masking (FPM). FPM replaces real personal information with data elements that look like personal information but don't belong to any data subject; a new different name replaces an actual name, a new different address replaces the actual address, and so forth. FPM has several advantages over more straightforward masking techniques to hide the data, such as replacing characters with asterisks.

Comprehensive Discovery

Irreversible Transformation

Enterprise-Wide Consistency

First, users of the repositories using FPM cannot tell which record was modified and reverse-engineer the deleted identity. Second, FPM can be used to support analytics processes that include modified data elements. The algorithms used to convert personal data in FPM can be designed to keep data in ranges that are meaningful to the company. When this happens, modified data can still be used for analytic purposes, for instance: keeping dates of birth within noted ranges or maintaining an accurate count of the number of data subjects from specified areas. Finally, FPM can be applied consistently across systems so that the same manufactured data elements are applied consistently when the data subject in question is encountered.

As with any technological solution for subjective problems such as identifiability, there are no silver bullets. We need data management standards that guide the use of obfuscation technologies to identify typical versus leading practices. We also need to educate the various stakeholders in the privacy management space, from privacy professionals to regulators. Success requires we employ obfuscation and data deletion processes that hold up in the real world of the extended enterprise and third-party vendor ecosystems.

Considerations for Third-Party Access to Data

Privacy regulations require companies to know what third parties have received data that falls under their scope. This requirement is necessary from a contractual perspective—bringing needed accountability to third parties to provide an accurate notice to data subjects and, in cases involving updates or deletions of data subjects, to ensure third parties follow suit for the relevant data subjects.

As reasonable as these requirements may seem, they touch on one of the most prevalent privacy and data protection weaknesses. Companies have a hard time knowing who all their third parties are and what data they process.

Companies cannot outsource accountability

In the last 15 years, companies outside of the financial services field have started looking more closely at the third parties they share data with. The push was the various state breach notification regulations that started coming up after California first passed stronger data privacy regulations in 2003. Breach notification elucidated the notion that a company cannot outsource accountability, and therefore breaches that happen due to third parties are the company's responsibility.

This evolution in understanding and controlling disclosures was important. Historically, companies did not concern themselves with tracking what data subjects are shared with different third parties.

A Road Map for Third-Party Accountability

From a data management perspective, the good news is that there are plenty of footprints across the enterprise that allow companies to find out how data is being shared, even down to the data subject level. There are four aspects for tracking the disclosures of the specific data subject. Two aspects are easier to solve, and two are more complex from a data management perspective.

Easier to Track: User Access and Email

Let's start with low-hanging fruit for tracking disclosures to third parties: email and user access. Much of the personal information that companies share with third parties is done by providing third-party users with access to systems and emailing that information to them. While many companies are not yet trying to explore disclosure data from these sources, the information is, in fact, readily available. Access rights can be compared with active directory and HR systems to regularly determine when users are not employees of the company and what third party they represent; a company can easily automate this process.

When it comes to personal data that leaves the organization, often via email, many companies already have experience determining whether sensitive identifiers such as social security numbers are shared. There are plenty of Data Loss Prevention (DLP) tools used for detecting such sensitive information, but they're not always suited for identity-level detection. Thankfully, technologies adept at finding identities in unstructured data can do a better job at this task and correlate identities with the third parties' domains.

Tougher to Track: Batch Transfers and Downstream Processors

The more complex disclosure tracking use cases typically have to do with batch transfers and the identification of downstream processors. The term "batch transfers" refers to the large volume of data commonly or regularly sent between companies. The personal data may include an updated list of employees to benefits providers or lists of customers for a marketing vendor. It can be a challenge identifying who in the organization is making disclosures in the first place, not to mention when and how. That's because it's rarely the case these days that such batch transfers are centralized by IT.

The information security tools now commonly available to support the secure transfer of large files allow users across the organization to make disclosures. With that convenience come gaps in data management and an inability to detect such transfers unless resorting to unreliable surveys and questionnaires. Known batch transfers can be scanned for the data subjects they include so that the relevant third party can be associated with them.

Part of the solution involves solving the challenge of identifying how personal information flows from one-third party to the next. A company's third parties have their own third parties that take part in processing the data. While it is common (and often legally required) for third parties to list their downstream processors, these lists are often incomplete or downright inaccurate.

The challenge here is twofold. The first challenge is how to detect such downstream processors. Second, once those processors are identified, the challenge becomes how to track which data subjects they process so that regulatory and contractual requirements can flow down to the relevant third parties and their service providers.

Fortunately, tracking disclosures of identities to third parties rarely needs to happen in real time. For the most part, the purpose is to keep an accurate inventory of the third parties with access to personal information, tracking the identities they receive and the obligations that correspond to each identity (e.g., a CCPA-impacted data subject below the age of 16). Inter-company distrust, having brought about the CCPA in the first place, will lead companies to require this degree of visibility into their disclosures in the near term.

Conclusion

Once you rethink and understand all these necessary privacy compliance and data management practices, you can readily manage your effective deployment and implementation of PKWARE's PK Protect suite. PK Privacy utilizes the key functions of PK Discovery, PK Classification, PK Encryption, and PK Masking, according to your policies, to ensure your company and the third parties whose practices your company is liable for can automate effective compliance with current and emerging privacy regulations around the world.

PKWARE's PK Privacy **automates** privacy compliance and DSAR fulfillment.

Enterprise Wide Policy Management

PK Protect provides a point of control for data protection activity across the entire organization

Simple Workflow

With PKWARE, data protection is automated for end users and easy for administrators to manage

Easy Implementation

PKWARE supports a variety of deployment options, enabling organizations to implement their data protection solution without time-consuming changes to infrastructure and workflows

Protection Without Gaps

PKWARE works on every enterprise operations system and provides persistent protection that remains with data even if it’s copied or shared outside organizations

Integrated Discovery, Classification and Protection

No other solution has the capability to find, classify and protect data in a single automated workflow

Multiple Protection and Remediation Options

Organizations can take a policy-based approach to data protection and choose from action including persistent encryption, quarantine, masking, and deletion.

BOOK A DEMO

PKWARE offers the only data discovery and protection solution that locates and secures sensitive data to minimize organizational risks and costs, regardless of device or environment. Our ultra-efficient, scalable software is simple to use on a broad range of data types and repositories, enabling precise, automated visibility and control of personal data, even in the fastest-moving, most complex IT environments. With more than 1,200 customers, including many of the world's largest financial institutions, retailers, healthcare organizations, and government agencies, PKWARE continues to innovate as an award-winning global leader in data discovery, security, and compliance.
To learn more, visit PKWARE.com.

PKWARE

PKWARE.com

866-583-1795

201 E. Pittsburgh Ave.
Suite 400
Milwaukee, WI 53204