PRIVITAR

Data Privacy for Analytics — at the Speed of Business

Introduction

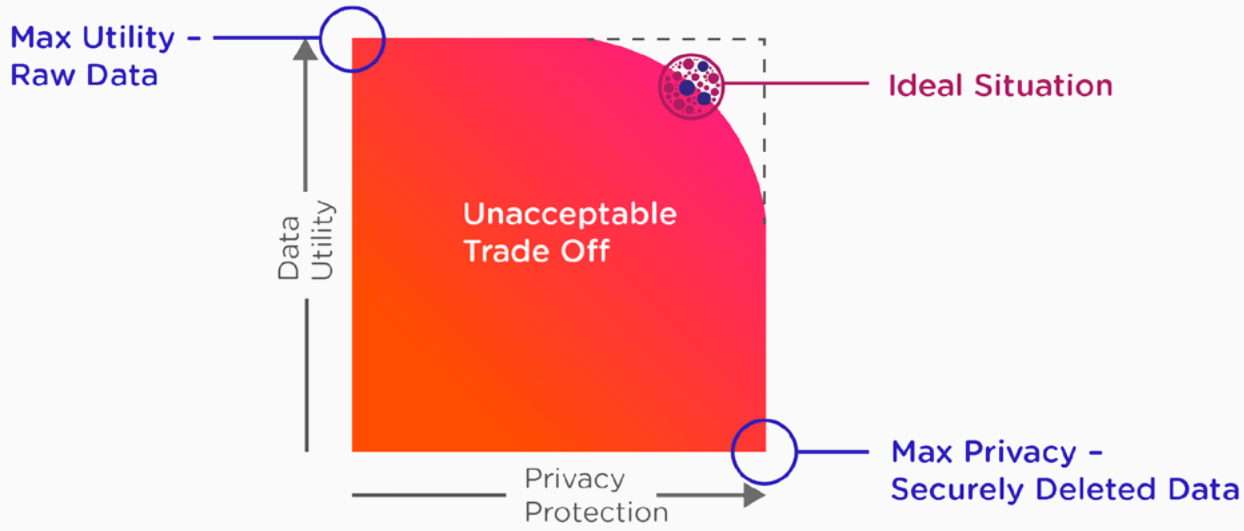
An ongoing challenge for organizations of all sizes has centered around how to accelerate analytics projects and extract analytical value without compromising privacy. Data scientists struggle to get access to sensitive data for analytics rapidly, and using a one-size-fits-all de-identification technique leaves them with data that is insufficient for complex use cases.

The Big Data paradigm assumes that more is better: more data, collected from more sources means that we can make better predictions. This holds true, but leaves out the imperative that organizations have to protect the privacy of the consumers in the data. Technology has created many ways to collect and analyze information, and it holds the key to protecting the data that is the lifeblood of innovation.

Privacy risk in sensitive data

Using raw data in analytics enables data scientists to access the highest utility data, however it also increases risk unnecessarily. The risks are threefold: significant fines due to data protection regulations when data is not properly protected, damage to reputation in case of a data leak or outright breach, and lack of trust in an organization leading to decreased revenues. Given the risks, it's clear that data protection is essential, and it is critical to solve the challenge of how to retain data utility while protecting it.

Historically, most organizations have been well aware of the importance of protecting data in analytics, so they separated it into two large groups: (1) the data you need to use for day-to-day operations, which is the risk you have to accept in order to do business, and (2) data that you don't use because you're simply not willing to accept the risk that comes with using that data. Of course, by not using large portions of data, there is additional risk due to the opportunity cost of not using that data. How, then, to create data that's safe to use and analyze without incurring significant risk or losing data utility?



“

The perceived opportunities in big data provide incentives to collect as much data as possible and to retain this data as long as possible for yet unidentified future purposes.

”

The European Data Protection Supervisor

Five Key Principles to maximize the utility of the data:

There are five key principles that organizations must consider when seeking to maximize the utility of their data while ensuring that it remains protected:

1. To accelerate the provisioning of safe data by defining, managing, and systematically applying consistent data privacy policies across locations and data environments.

2. To optimize data utility and privacy by fine tuning policies applied directly to data according to context.

3. To automate data de-identification for analytics by integrating with leading cloud and on-premise analytics technologies.

4. To deter insider threats and facilitate forensic investigations via watermarks that describe data provenance.

5. To maximize the value of analytics investments by providing data users with access to more types of data in a safe and secure way.

Compliance

Over 100 countries and 30 states in the United States of America have enacted data privacy laws, and that list continues to grow. While it may not feel that way to those struggling to meet these changing regulations, the intent of the regulators is not to prevent organizations from using their data, but rather to require the responsible use of data that protects each individual's sensitive personal information. The regulations require organizations to demonstrate the measures they're taking to protect this data to auditors and regulators. Furthermore, these measures must represent reasonable best practices and be applied consistently. The consequences of non-compliance are steep:

The General Data Protection Regulation (GDPR) affects all companies who do business in European Union with fines up to 4% of worldwide annual group revenues.

The California Consumer Privacy Act (CCPA) went into effect on January 1, 2020, and includes fines of $2,500-$7,500 per person per incident, plus damages of $100-$750 per person per incident, or actual damages, whichever is greater. California also passed the California Privacy Rights Act (CPRA) in November 2020, overlaying the CCPA.

The Health Insurance Portability and Accountability Act, more commonly known as HIPAA, is a federal law in the U.S. that protects sensitive patient health information from being disclosed without the consent or knowledge of the patient. HIPAA violations can result in civil money penalties (fines) and criminal penalties (jail time).

Three significant changes from the CPRA to the CCPA:

The CPRA makes a wide range of changes to the CCPA. Three significant changes business should be aware of are:

1. **Fines and claims have been made a little tougher.** Fines for unintentional violations relating to children's data have been tripled and the 30 day 'cure' provision, whereby an organization had 30 days to try and remedy a breach, has been removed.

2. **Consumers can opt out of all sharing.** Under the CCPA the "opt out" only relates to selling data. While the definition of sale was very broad, the CPRA is even broader, extending it to all data sharing.

3. **Consumers can opt out of secondary uses of sensitive data.** The CPRA introduces a new category of data: sensitive data. This includes things like health information, information about race or sexual orientation, personal messages, and precise geolocation. Under the CPRA, for sensitive data, consumers can opt out not just from sharing from one company to another, but also any secondary uses of the data within a company. That means consumers can request their data not be used for any purposes beyond what is necessary to provide the goods or services.

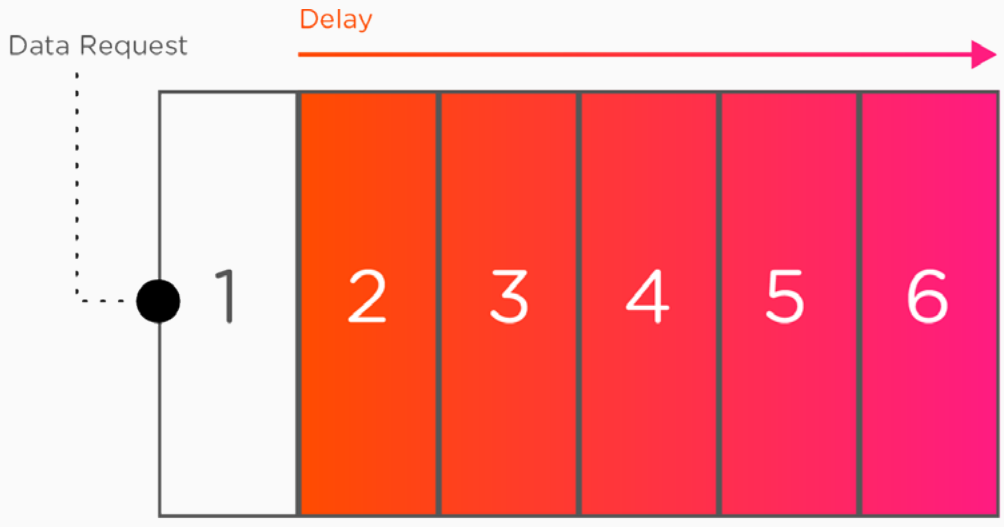
To meet these and other regulatory requirements, many organizations implement policies, processes, and access controls (which are often manual).

This can result in a delay from two to six months from a data request to fulfilment of that original data request. It may even decrease data resolution to the point of it becoming useless to the data users. Regulatory compliance combined with strict corporate policies and processes can prevent organizations from maximizing the value of sensitive data for analytics.

What's more, meeting the requirements of the regulations requires interpretation and creates further challenges. For example, Article 5 of the GDPR defines the data minimisation principle as follows: "Personal data shall be…adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed." So, the next problem comes for real-world data scientists, who must decide how to select which data to include and justify their selections. Using all the available data on the basis that it could be necessary contravenes the data minimization principle, and goes against guidance from the Information Commissioner's Office (ICO) in the United Kingdom that "finding the correlation does not retrospectively justify obtaining the data."

Requirements such as the Right to be Forgotten bring new challenges. Companies must comply with GDPR, among other regulations, and support Right to be Forgotten (RtbF) requests, but

doing so can be quite difficult and expensive for businesses. When an organization receives a request it typically requires that all data for that individual be deleted, so that it can no longer be processed or analyzed, and ensuring that data is removed from all instances. It is quite difficult for most organizations to ensure that all data touch points are accounted for, and deleting data inherently reduces the analytical value of the dataset. Many organizations have struggled to handle RtbF requests quickly and effectively. Technology, particularly in the form of automation and effective workflows, may help to solve these challenges. Workflow



Fulfillment/Time from Data Request to Access (months)

“

Personal data shall be… adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

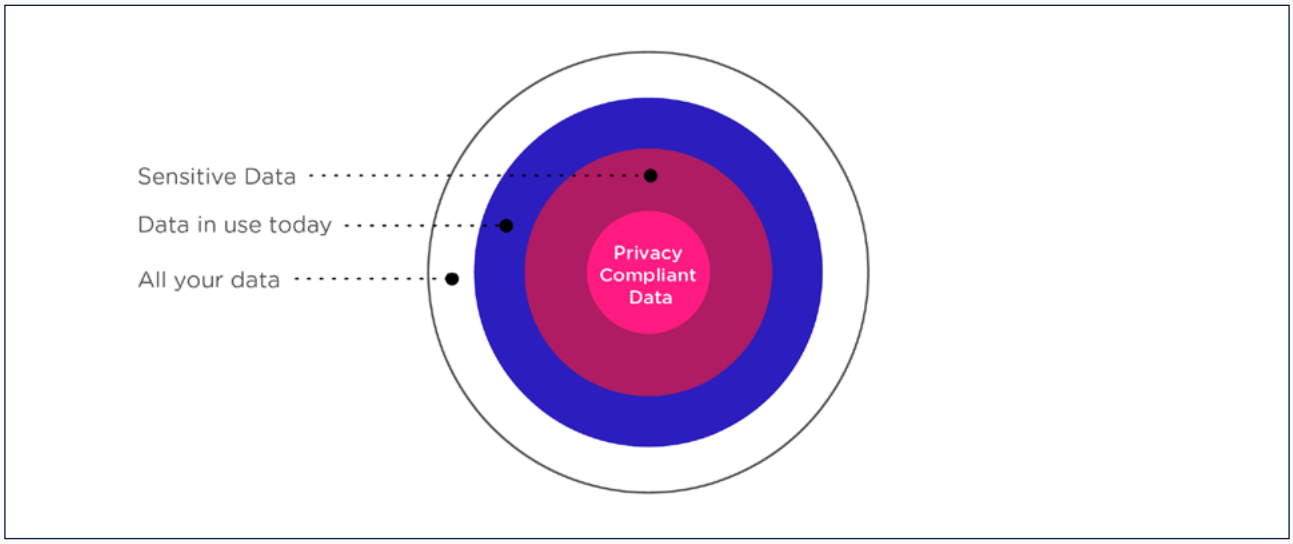
”

GDPR
Article 5

Workflow

Most organizations spend years accumulating vast amounts of data, and while many datasets are stored, they are not available for analysis. And unfortunately, there is a great deal of discrepancy between the data that is officially "available" (because it's considered safe to use) and the data that is actually being used for analysis. Essentially, there is a "black market of data" in many organizations that's consistently being used, and yet that data has not been protected, and should not be available for use.

Regulators would find these uses of data unacceptable, yet that data continues to be used in most organizations. If a data breach were to occur, an investigation would almost certainly result in significant fines as well as damage to reputation. Instead of utilizing this black market data, organizations must adopt automation and workflows that allow them to access 100% of their data — without compromising on data privacy.



The first step is to centralize privacy management, which offers a number of advantages. First, it fosters a consistent approach across an organization because it creates a central forum for decisions about pre-processing data. Data goes through a series of steps during pre-processing, including data cleaning, data integration, data transformation, data reduction, data discretization, and data sampling. In a project-specific approach, these decisions can become slow, inconsistent, and a challenge to audit. Second, centralization allows you to document the transformations (such as tokenization) applied to the data. This can accelerate data preparation for machine learning (ML) projects, because you can make decisions on how to construct a training dataset once and then apply those decisions consistently.

Indeed, documenting transformations supports compliance with the GDPR requirement to record processing (Article 30) and explainability in the context of the ICO's guidance. Using Privacy Enhancing Techniques (PETs) to de-identify data addresses the risk related to using sensitive data, which enables data scientists to minimize the time spent collating data and allows them to spend more time running and analyzing models.

Creating business value from data requires organisations to convert data into actionable insights, typically through the analytics and

artificial intelligence/machine learning (AI/ML) capabilities within the organisation. Many organisations move to data lake architectures in order to ingest, store, and catalog data at scale. Next, organizations must implement and automate a data pipeline that allows for the creation of a de-identified safe dataset for analysis. Data scientists and business analysts can now access that data using their choice of tools and frameworks, including analytics and ML services. This enables data scientists and business lines to access safe, usable data quickly.

There are several critical changes to infrastructure and workflows that organizations must plan for to empower data scientists as they adopt data privacy solutions:

1. Enable discovery of sensitive data sources and personal data across the organization.

2. Create a data catalogue to accelerate and empower the search for useful information.

3. Make advanced privacy enhancing techniques available, in the form of rules to be applied to protect sensitive records.

4. Create privacy policies that map to datasets structures that protect sensitive information.

5. Control the data releases in specific domains and with traceability through the use of data watermarking.

By adopting these critical changes to infrastructure and workflow, security and privacy departments can accelerate data access, protect sensitive data, and allow data scientists to explore and visualize data faster.

Speed to data

Data scientists and machine learning experts spend approximately 80% of their time generating, preparing, and labeling data. These experts spend only 20% of their time building and training models! Obtaining, crunching, and preparing data is part of the job and has significant implications on the performance of the final model. And, naturally, a learning model is only going to be as good as the supporting data. This is why it is crucial to pay attention and maximize efficiency for the time spent in the data preparation stage.

For many data scientists, the access request process is where the slow down occurs as it can take days, weeks, or even months. The time frame typically depends on a number of factors, including: sensitivity of the data, control processes currently in place, technology limitations, and cross-departmental approvals. A data scientist must provide justification for access, and may even need to have specific meetings with security and privacy teams in order to get approval. If the data will be processed in cloud environments, there may be additional processes to ensure that the data is adequately protected from a data breach to minimize risk to the organization.

By adopting the right infrastructure, workflows, and processes, these data science and machine learning teams can analyze more data, faster. Faster access to more data enables organizations to make better decisions, improve customer engagement, enhance innovation, and meet customer needs in a competitive business environment.

Value of data

The analytical value of data protected with both security and privacy is significant, helping organizations make better business decisions, innovate faster, and engage more effectively with customers. Yet the significant barriers due to privacy risks, compliance concerns, workflows, and processes that slow access to data down have historically been a challenge when it comes to maximizing the analytical value of sensitive data and making it more easily and broadly available.

If data protection and privacy concerns were not constraining organizations, data scientists could feed as much data as possible into a training dataset. Data science is exploratory; therefore the data scientist doesn't know in advance what correlations the model may unearth. Training datasets must be as large as possible to maximize the possibilities of innovative insight. That insight isn't limited to a market disrupting solution, it may be a crucial medical treatment or an innovation that preserves the environment. The data itself is valuable, but the data processing is what transforms that data into actionable information. Of course, we do not live in a world free of data protection and privacy concerns.

Organizations today have access to the raw computing power necessary to perform advanced analytics and deploy machine learning and artificial intelligence (AI) projects. This enables these organizations to innovate quickly as they process and learn from an ever-growing cache of data.

The opportunities offered by the ability to gather vast amounts of information and perform complex computations are enormous, from detecting and preventing financial crimes to finding new and groundbreaking health treatments. Protecting sensitive data opens up these possibilities.

Data democratization

By applying workflows and simplifying processes to de-identify data, your organization can make sensitive data far more available, essentially democratizing data science. This enables organizations to distribute insight production to larger groups within the company and empower more of your workforce to perform high-leverage data science work. However, data democratization can't be achieved unless you have the appropriate technical and non-technical guidance, with policies and automation to help manage access levels and maintain consistency to maximize data utility.

As digital transformations and big data initiatives begin to scale, enterprises require a streamlined provisioning process that can meet the volume and breadth of data usage in their organizations while also standing up to regulations and audit. Many organizations have tried manual approval processes and custom scripting solutions in small-scale pilots or departmental stages of deployment. Unfortunately, these approaches are slow, unreproducible, and ultimately break under enterprise load. They must be replaced with a systematic and automated approach that removes the friction between data users and sources while enforcing data privacy.

Conclusion

Every day, organizations gather vast amounts of information, generated by a population of trillions. Before COVID-19, organizations were

already aware of the tremendous value gained by using, analyzing, and exchanging data for organizations worldwide. This global pandemic has only accelerated existing trends towards digital transformation.

In 2020, work meetings, social interactions, education, exercise classes and shopping shifted to online environments, and many of them are likely to remain there long past the necessity of interacting online. This shift served as a change catalyst for everyone as businesses adjusted to new market needs. These new market trends and needs have pushed many organizations to accelerate their migration to and usage of the cloud, adopting a blend of on premises, hybrid, public, and multi-cloud environments to meet the various needs of their customers.

Along with this acceleration, organizations are scaling data-driven services, embracing integrated analytics, data engineering, and machine learning technologies, which in turn dramatically accelerate data-driven insights and innovations. What remains an ethical and business imperative is to be prepared for the challenges of keeping data protected, whether it is secured data at rest or data in use that has been de-identified to ensure privacy.

How Privitar can help?

Data-driven organizations rely on Privitar to realize the promise of one of their most valuable assets – safe, usable data. The Privitar Data Privacy Platform™ integrates with data platforms to protect and manage sensitive data while optimizing its utility for analytic applications. This enables organizations to use all of their data for analytics to gain timely insights and support data-driven decisions that lead to better products, services, and customer experiences. Plus, with native and API integrations to the full range of cloud and on-prem big data standards, and a commitment to remain vendor neutral, you can evolve your data pipeline and analytics environments without risking loss of support or vendor lock in.

Privitar has the most flexible data privacy engine in the world, capable of supporting batch, data flow, and on-demand processing of data across cloud, hybrid cloud, multi-cloud and on-premise environments. So whether you want to protect data as part of your data flow prior to ingesting into your cloud data lake or produce purpose-limited datasets in batch, Privitar empowers you to protect data privacy while maximizing analytic capabilities.

Together this protects organizations from the fallout of any data breach, enables faster and better business decisions, better customer engagement, enhanced innovation, and faster time to market. Only Privitar can help achieve privacy, utility, and speed of sensitive data together.

Privitar protects sensitive data for analytics:

Reduce privacy risk and meet compliance

Maximize analytical value of sensitive data

Increase the agility of sensitive data by accelerating and democratising its use

PRIVITAR.COM