PRIVITAR

Managing the Lifecycle of Sensitive Data

with the Privitar Data Privacy Platform

www.privitar.com

There was a time when data collected about individuals was limited, was spread across multiple systems and was difficult to connect.

Regulations were limited or non-existent. And data privacy protections were rudimentary. Not anymore.

The explosion in the availability of personal data and tools to process it at scale creates remarkable opportunities for organizations to improve growth, profitability and otucomes. As a result, most organizations have imperatives to create a data marketplace that makes data quickly and safely available to its data scientists and business lines. However, regulations and internal data privacy and security policies often stand in the way.

The Privitar Data Privacy Platform provides a modern data architecture that enables organizations to realize the promise of safe, usable data across all of their data sources and environments by:

- supporting the diverse combinations of data, purpose and use
- applying consistent, auditable privacy policies
- streamlining and automating data provisioning
- scaling to enterprise data volumes
- preserving compatibility with applications and tools
- managing regulatory, reputational, and ultimately financial risk.

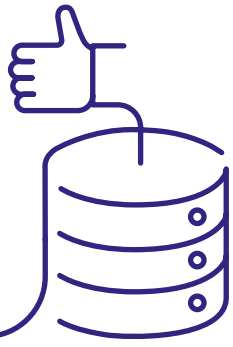
Data privacy is contextual

Each data analysis represents a unique combination of data, purpose and method(s), which in combination define its context.

To unlock its value, the right data must be provided to the right analysts and systems for processing. But when it includes personal data, privacy also must be safeguarded, since it is during analysis that the risk of misuse is greatest, regardless whether it is intentional or accidental.

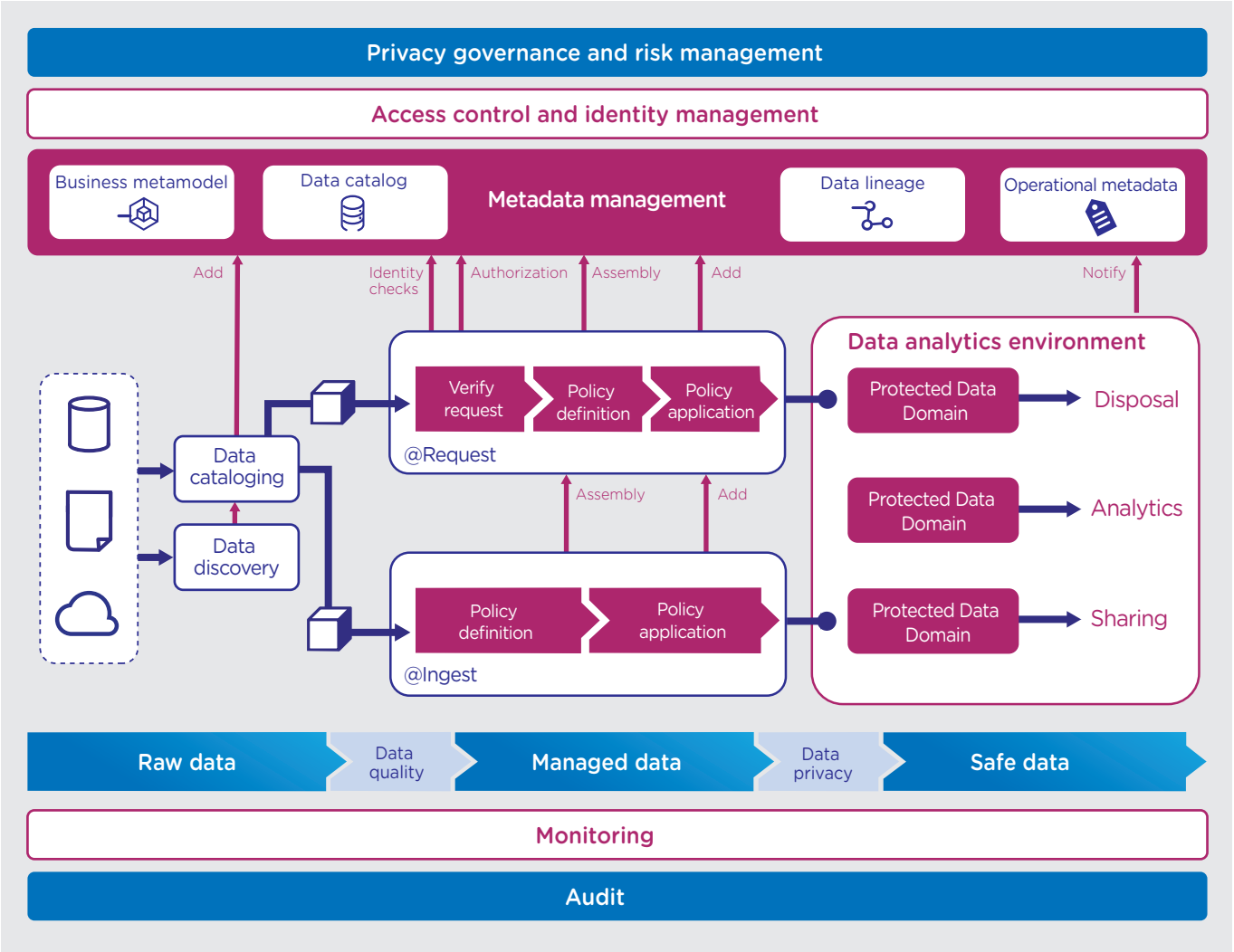
Providing access to datasets that are larger or richer than strictly required exacerbates the consequences from misuse. When too much data is provided, more individuals will suffer harm from misuse. When too much detail for individuals is available, misuse can be more specific, and thereby more intrusive and damaging to each individual. Moreover, richer information heightens the potential for linkage attacks, as discussed later.

When you prepare a dataset for use in a particular context, you can also optimize the dataset for that context. This allows you to maximize its value to data consumers. Data resolution can be retained only where it is critical for an analysis, and coarser protections can be applied elsewhere. In so doing, data patterns that play an important role in the analysis are preserved and traded off against reduced resolution in other dimensions.



Raw > managed > safe data pipeline

To enable enterprises to realize context-specific data, Privitar has productized a three-stage data privacy pipeline that enables organizations to design data flows that automate privacy best practices. In successive stages of the pipeline, raw data becomes cataloged, managed, protected and, ultimately, made available in a Protected Data Domain. Data moves through the three stages of raw, managed and safe data as follows.



1. **Raw Data.** Data, whether newly ingested or legacy, that has little metadata is raw data. This data should be treated as high risk until the scope of any personal information it contains is understood. It is critical to maintain tight access controls on this data.
2. **Managed Data.** Raw data becomes managed data through the processes of data discovery and data cataloging. Data discovery is the process by which raw data is analyzed and attributes within the data are identified. Data cataloging produces an index of what data is available and enforces a standardized tagging system for attributes across all data.
3. **Safe Data.** Applying privacy transformations to managed data creates a Protected Data Domain containing safe data to be used by specific analysts for a defined purpose. Protected Data Domains break privacy risk into manageable units. We will explore the components required to optimally manage risk for these units throughout this whitepaper.

Metadata-based entitlements

Each Protected Data Domain contains the minimum amount of data for a particular use. It therefore represents the ideal unit of data for managing access controls. Importantly, each one also includes immutable metadata for properties such as requester, approver, user(s), purpose and duration of use. This metadata can be used in conjunction with your data catalog, identity and access management and other data pipeline systems to automate the application of access entitlements and data expiration.

Full range of de-identification techniques

In addition to access controls, enterprises must be able to apply the right combination of privacy transformations at the data level. The following techniques for producing context-specific Protected Data Domains can be applied in any combination as part of Privitar Privacy Policies.

- Pseudonymization. Replace direct identifying information with random pseudonyms, while preserving the format, structure and referential integrity of the data
- Generalization, perturbation or noise addition. Reduce the resolution or accuracy of the data to limit the knowledge that data consumers can obtain
- Data minimization. Remove information not required for the specified purpose from the dataset

Consistent Privacy Policies

When dealing with data at scale, it is imperative that reusable, modular privacy definitions are created and managed centrally. This allows you to apply protections consistently across your data, whether new or legacy, regardless of the data source or execution environment.

Access control is necessary but insufficient

Access control is an important defense, preventing the use of data by unauthorized actors, but it simply cannot accommodate all contexts within an enterprise.

It does not prevent authorized data consumers from learning more about the data subjects than their analyses require. Moreover, these authorized users may inadvertently or maliciously misuse or improperly distribute data, whether of their own accord or through coercion. And at some point, malicious actors will find a weakness in your access controls or perimeter defenses and gain access to some of your data.

Privacy Policies describe the privacy transformations to apply to data in order to de-identify it. When you apply these Privacy Policies to the data, the de-identified data is added to a Protected Data Domain for use by defined users. This means that you don't need to move your data from its secured environments such as a cloud or Hadoop cluster or build out a separate privacy processing infrastructure. Further, Privacy Policies can be applied across data sources and platforms, regardless of where they exist.

For example, the same Privacy Policy can be embedded in a Kafka streaming architecture and a big data processing task in Spark whether on premise, cloud or hybrid. This policy acts identically in every case, producing safe data that is based on the same privacy requirements.

Regardless of how heterogeneous your data architecture is – whether you have a data lake, streaming, cloud or mixed environment; whether you are modernizing and migrating from one technology to another – you can apply the same policies across your data landscape.

Automation is essential at scale

Manual processes are slow, prone to errors, easily circumvented and subject to unnecessary exceptions. Conversely, automation enables enterprises to apply consistent processes and controls across their data landscape, however heterogeneous or ever-changing they may be. By integrating privacy controls into your data pipeline, you can automatically apply your privacy controls to each dataset based on your catalog metadata.

Privitar is designed with a complete set of REST APIs – the same APIs the user interface and server processes call. These APIs enable you to integrate Privitar into your data processing and provisioning pipeline so that you can:

> Systematically apply consistent Privacy Policies regardless of the data sources and environments you have now or in the future

> Apply business rules according to the context inherent in the data and metadata

> Integrate with other systems that populate or consume catalog metadata

Embedded data provenance

Recall that Protected Data Domain metadata includes the requester, approver, user(s), purpose and duration of use. It also incorporates the data location, schema, and the Privacy Policy and de-identification techniques applied to the source data. This metadata is encoded into an undetectable and unique watermark that can be applied to a Protected Data Domain.

By tailoring specific de-identification techniques, Privitar can embed digital fingerprints into your data without any additional information loss. These watermarks enable detection and attribution of unauthorized copies of data. They are a powerful deterrent against insider threats and can accelerate forensic investigation in the event of a breach.

Controlled data linkage

Data Linkage is the ability to enrich a dataset by combining it with one or more other datasets on the basis of common attributes present in those datasets. For example, a retail analysis might join customer profiles with a database of transactions to analyze purchasing behavior taking into account demographic segmentation. Joining the two is possible due to a common customer identifier present in both.

There are clear benefits in being able to expand datasets containing personal information by linking and aggregating more information; indeed, it's something that analysts do every day. However, without proper controls the potential for unintended or unexpected linking is a genuine source of risk:

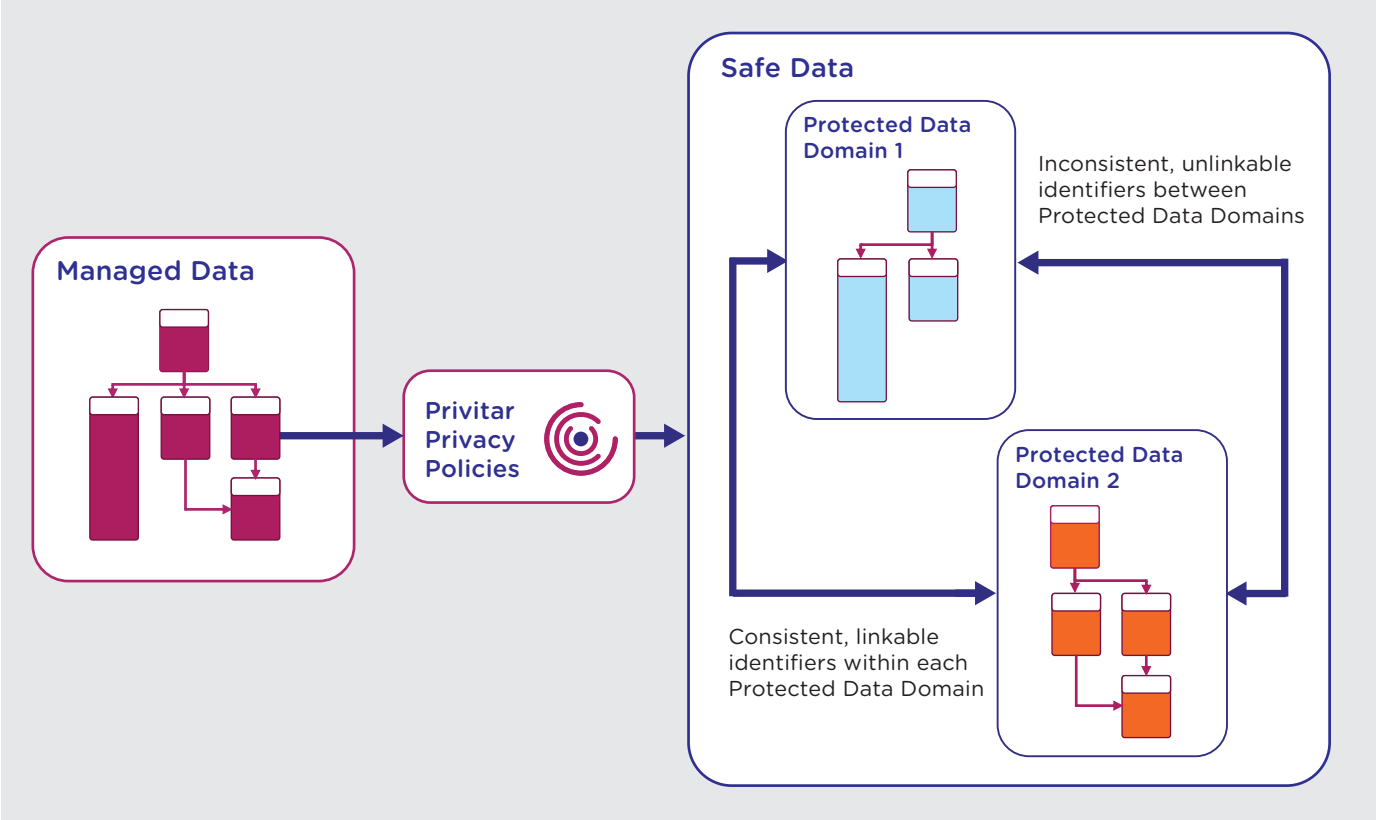
> **Unrelated analyses may overlap within an organization**. Analysts working on unrelated projects may link their datasets, either accidentally or deliberately without approval or an understanding of the risks.

> **Linkage with public datasets**. Public datasets contain a wealth of sensitive information. If analytic datasets can be linked to public datasets, there is a significant risk of reidentification or sensitive attribute disclosure.

> **Data sharing may reveal personal information.** When datasets are shared with a 3rd party, you lose control and visibility of how that dataset may be linked to other datasets.

In each of these cases, linkage allows anyone with access to the data to learn more about individuals than is required for a task and provides opportunities for abuse.

Protected Data Domains eliminate common identifying attributes between datasets. Among various techniques, random pseudonyms are substituted for identifiers to prevent linkage to public datasets. Referential integrity within a single Protected Data Domain can be preserved by using the same pseudonyms for all data inside, but linkages with other PDDs are prevented by using different pseudonyms for each one.



Audit and reporting

Organizations must demonstrate to auditors and regulators, both internal and external, the precise measures they are taking to protect sensitive data. When assessing breaches, regulators have demonstrated that they will take into account whether reasonable and responsible measures have been taken.

Privitar records the Privacy Policy that was applied, and which data privacy protections were used to create each Protected Data Domain across the entire enterprise. These reports can be used to demonstrate to regulators that consistent, comprehensive privacy protections have been applied.

Application and tool compatibility

Protected Data Domains are real datasets. After you apply privacy transformations, safe data in a Protected Data Domain retains the structure of the original, including its format, storage location and metadata. This has several advantages:

- **Preserve application compatibility.** No changes to applications are required as a result of privacy protections, because they preserve the data schema and format. Safe data can be consumed by all applications in exactly the same way as the raw data.
- **Seamlessly support analytics and BI tools.** Cognos, Microsoft, Python, Qlik, R, SAS, Tableau, TIBCO, or whatever your tool of choice continues to work simply because data formats are preserved. Accessing safe data in these tools is no different from accessing raw data.

Raw Data				
Wealth Id	Name	Id	Date of birth	Net worth
519802	Irene Hill	329-92-1196	15 August 1969	19,306,886
230120	Daniel Cohen	720-30-2518	16 July 1969	24.284.491
824482	Abigail Baker	931-09-4452	8 May 1945	18,919,593
384901	Charles Knox	274-24-5568	7 April 1969	21,266,545
190338	Audrey Poole	710-71-5618	6 August 1945	16,870,355

Safe Data				
Wealth Id	Name	Id	Date of birth	Net worth
907908	Oscar Hooper	999-30-5482	15 August 1969	19,000,000
934449	Jasmine Howell	999-05-8556	16 July 1969	24.000.000
985313	Molly Flynn	999-17-1545	7 April 1945	21,000,000
938114	Harrison Reed	999-79-3904	8 May 1945	19,000,000
985647	Donimic Banks	999-70-4263	6 August 1945	17,000,000

Enrich & analyze data across boundaries & borders

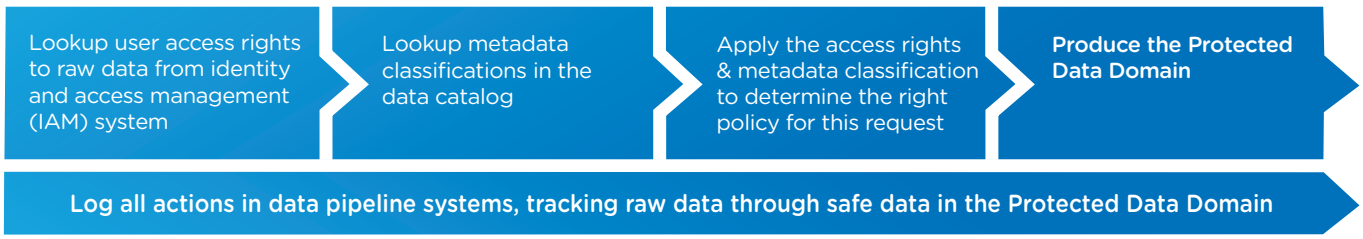
Analyses using data collected from different divisions, geographies or even organizations can benefit from the increased richness that the additional data affords. This can generate insights that otherwise would not be possible. All parties who contribute sensitive private information to such a combined dataset must be confident that the linking process is performed such that individuals' identifying information and their commercially sensitive information is protected end-to-end.

Privitar SecureLink™ is a multiparty solution that uses blind matching and homomorphic encryption to enable collection of data from multiple parties to produce richer datasets while addressing privacy concerns. Identifying fields are encrypted at their sources and are never decrypted, which ensures protection of identifying data in flight, in memory and at rest at the data recipient. It uses encrypted fields to link datasets with no ability to decrypt or discover the common identifier used to link them. There is therefore no risk of re-identifying individuals in the combined dataset.

Privacy @Request

Analysts and business users can make individual requests for data, which include a specific purpose and result in the creation of a context-specific dataset. Requests are subject to an approval process that assesses risk factors, such as the sensitivity of the data, the recipient(s) and their entitlements, the purpose of the analysis and how the resulting Protected Data Domain will be delivered. A privacy orchestration system can resolve this systematically, automating the provisioning of safe data in a Protected Data Domain.

The benefit of this approach is that the information provisioned respects the data consumer's entitlements and is tailored precisely for the use case.



Privacy @Ingest

When data at rest is protected upon data ingest, such as landing in a data lake or reservoir, defining an appropriate Privacy Policy requires you to understand the classes of analyses in advance. The Protected Data Domain produced is typically cataloged as a safe library that can be used by many data consumers on an ad-hoc basis.

The advantage of this approach is that Privacy Policies are defined and understood in advance, and the resulting Protected Data Domains have wider applicability. This means that fewer are needed, which reduces system and process overhead. The disadvantage is that the resulting Protected Data Domains are higher risk due to the less restrictive Privacy Policies required to service less specific use cases and the larger number of data consumers that can access them.

It is important to understand that since a Protected Data Domain is a real dataset, it can be used as a data source to apply additional controls in a second Privacy Policy to produce a second Protected Data Domain from the original. This is especially useful in response to a request for a sensitive analysis or sharing data with a 3rd party, as it lets you daisy-chain privacy protections with finer grain controls for more specific use cases.

Privacy @Borders

A special case of applying protections on ingest involves moving data between administrative zones where different controls on sensitive data must be applied. Examples of this include moving personal data from on premise to a public cloud or across borders for countries or jurisdictions. In these cases, required Privacy Policies are determined by the regulations and company policies that are in effect between the zones.

Conclusion

As data collection and analytics technology has grown, so too have data privacy risks. Established protections are now inadequate. A new and rigorous solution, embracing both technology and process, is required.

Organizations must recognize that privacy is contextual. A solution must deliver tailored, context-relevant datasets that optimize utility, implement data minimization and allow centralized management of risk. The solution must be built on a foundation of robust data discovery and data catalogs that power automated data privacy processes, respecting both the unique nature of each dataset and the processing being performed on it.

The Privitar platform achieves these requirements producing privacy-preserving datasets, known as Protected Data Domains, based on metadata and user-level entitlements. Protected Data Domains provide complete control of data linkage within datasets and prohibit linkage between them.

Centralized, reusable Privacy Policies are applied across your data landscape, with broad support for streaming, on-premise cluster and cloud platforms. Privitar's architecture brings the workload to these environments, preserving data structure and thereby ensuring application and tool compatibility.

Privitar tracks Protected Data Domains and their lineage to give data owners visibility of data use and risk across the organization. Protected Data Domains have a managed lifecycle, which can include safe disposal once an analysis is concluded.

About Privitar

Organizations worldwide rely on Privitar to protect their customers' sensitive personal data and to deliver uncompromising data privacy that frees them to extract maximum value from the data they collect and manage.

With the powerful Privitar Data Privacy Platform, businesses can safely use data to gain valuable insights that support data driven decisions over intuition to innovate, identify market opportunities, accelerate time to market, acquire and retain customers, improve customer experience, and identify inefficiencies that ultimately grow revenues, reduce costs and increase profitability.

Founded in 2014, Privitar is headquartered in London and has offices in New York, Boston, Munich, Paris and Singapore.

Contact us:

e: info@privitar.com
t: +44 203 282 7136
w: www.privitar.com

PRIVITAR

@PrivitarGlobal

www.privitar.com