Privacy By Design: From Principles to Requirements

Mark Settle

About the Author

Mark Settle is a seven time CIO, two time book author and three time CIO 100 honoree. He has led IT teams in public and private companies that supported global business operations, software development and online Ecommerce. His most recent book is *Truth from the Valley, A Practical Primer on IT Management for the Next Decade*.

Acknowledgments

Many thanks to Adam Gavish, Gal Helemski, Lonne Jaffe, Declan Morris, Joanne Peca, Seth Pierrepont, Amit Shaked and Priyanka Somrah for reviewing an earlier version of this report.

Disclaimer

References to the products and services of commercial vendors that appear within this report are included solely for illustration purposes. They do not represent personal endorsements on the part of the author. Many of these vendors are early stage startup firms. Their product visions may exceed their current capabilities and the utility of their offerings may not be widely tested by companies of varying size operating within different industries.

Table of Contents

Introduction

The term *Privacy by Design* can be traced to a collaborative project performed by the Information and Privacy Commission of Ontario, Canada, the Dutch Data Protection Authority and the Netherlands Organization for Applied Scientific Research in 1995. This concept was popularized by Ann Cavoukian, the Assistant Information and Privacy Commissioner of Ontario at the time. Cavoukian proposed a set of Foundational Principles that should govern the construction and operation of IT systems employing privacy data. These Principles were officially endorsed as an essential component of privacy protection at the 2010 Assembly of International Data Protection and Privacy Commissioners.

Cavoukian's Privacy by Design Principles are a manifesto or call to arms, highlighting the importance of privacy data protection and underscoring the responsibility of commercial firms to safeguard the handling and processing of Personally Identifiable Information (PII). Cavoukian's Principles have gained traction in global regulatory agencies. They are directly reflected in the European Union's 2016 General Data Protection Regulation (GDPR), specifically in Article 25 which is entitled Data Protection by Design and Default. They are also selectively referenced in the 2017 ISO 29100 standard dealing with information technology, security techniques and privacy.

Cavoukian's Principles were formulated in a very different era. The Privacy Act regulating the use of PII data within U.S. government agencies was passed in 1974. It was followed by HIPAA (Health Insurance Portability and Accountability Act) in 1996 and the European Union's Data Protection Directive (forerunner of GDPR) in 1995. The Payment Card Industry Data Security Standard (PCI DSS) was one of the first control frameworks designed to regulate the use of PII within private industry. It did not go into effect until 2006. From a technology perspective, SaaS applications, APIs, cloud service providers and mobile devices were virtually nonexistent in the late 1990s and early 2000s. The quantity and sensitivity of digital PII data employed by commercial firms was far less and widely publicized data breaches were far less common.

Privacy By Design - Foundational Principles
Defined By Ann Cavoukian

1. Proactive, Not Reactive; Preventative, Not Remedial
2. Privacy as the Default
3. Privacy Embedded Into Design
4. Full Functionality – Positive Sum, Not Zero Sum
5. End-to-End Security – Lifecycle Protection
6. Visibility and Transparency
7. Respect for User Privacy

Note that the Users referenced in these Principles are equivalent to the Data Subjects defined in GDPR. The term Data Subject is adopted throughout this report to refer to individuals who have contributed PII to commercial firms.

Revisiting Privacy by Design Principles in 2021

Regulatory requirements, technological capabilities and human concerns regarding PII management have clearly changed during the 25 years since the introduction of the original PbD Principles. However, the way we think about privacy data has changed in some fundamental ways as well.

Human beings possess legitimate rights regarding the use of their personal information. There is no reference to citizen data rights in the U.S. Constitution but data rights have been explicitly recognized in a series of legislative acts such as the Federal Educational Rights and Privacy Act, the Fair Credit Reporting Act, HIPAA, the Driver's Privacy Protection Act and the Children's Online Privacy Protection Act. GDPR explicitly defines the rights possessed by individuals who contribute personal information to commercial enterprises (aka data subjects). The rights to access contributed information, correct it, erase it and control/approve its usage have been widely copied in other regulations such as the California Consumer

Protection Act (CCPA), the Virginia Consumer Data Protection Act, Brazil's General Law for the Protection of Personal Data and the New Zealand Privacy Act. Whether explicitly defined or implicitly assumed, a standard set of data subject rights has emerged in our corporate consciousness that impacts commercial operations almost everywhere. Evangelism concerning the importance of PII protection is no longer necessary. PII protection has either been mandated by law or become a standard expectation of individuals doing business with commercial firms.

PII protection is a shared responsibility between commercial enterprises and data subjects. The Foundational Principles emphasize the responsibilities of commercial firms to proactively embed safeguards into every aspect of their privacy data operations. There's an implication that after a data subject has voluntarily surrendered certain forms of PII, its protection is solely the responsibility of its commercial recipient. In practice, it's difficult for corporations to anticipate all the ways they might potentially use PII. It's equally difficult for data subjects to anticipate all the concerns they may have regarding the ways their PII is being used. For example, a data subject may object to the sharing of her PII with another corporation for marketing purposes but may eagerly agree to such sharing if it results in exclusive product access, free shipping or pricing discounts. Consumer concerns regarding PII usage are neither uniform nor predictable. They're contextual in nature. For all of these reasons, the contribution of PII to a commercial firm is the initiation of an ongoing relationship – not a one-time event. PII can only be protected if commercial firms provide alerts and updates on its intended usage and data subjects exercise their rights accordingly. PII protection is not solely the responsibility of commercial firms.

Collection limitation is a thing of the past. One of the key implications of the Privacy as the Default Principle is that PII data should be minimized upon collection and usage. Simply put, the scope and nature of information collected from a data subject should be defined as narrowly as possible to address a specific purpose. The subsequent usage of such information should proceed on a highly selective basis, employing only those subsets of the acquired data that are actually needed to perform a specific task. In reality, commercial B2C businesses have spent the past 10 years trying to acquire insights into the interests, backgrounds, living circumstances, preferences and behaviors of their paying customers to enrich and personalize their multichannel buying experiences. The corporate compulsion to know as much as possible about paying customers will only intensify in the immediate future as steps are taken to discontinue the use of third party tracking cookies on commercial websites. It's important to note that many consumers actually welcome the opportunity to share personal information with retail firms in the hopes that such information may provide them with early access, stock alerts, pricing discounts or premium support on future purchases. Minimization upon usage remains an important principle which will be highlighted later in this report but minimization upon collection is no longer realistic within today's multichannel marketplace.

Privacy protection is becoming a standard functional requirement for IT systems handling PII data. The Third Principle dictates that Privacy is to be Embedded Into Design and not added as a bolt-on capability or afterthought to a PII processing system. Article 35 of GDPR specifically requires commercial firms controlling such processing to "carry out an assessment of the envisaged processing operations on the protection of personal data….where a type of processing….is likely to result in a high risk to the rights and freedoms of natural persons". Firms implementing new systems that pose such risks must prepare Data Protection Impact Assessments (DPIAs) and consult with the appropriate EU supervisory authority to determine if privacy risks have been fully identified and properly remediated. PII impact assessments are not limited to in-scope GDPR systems. Compliance with the ISO 29100 standard requires the conduct of Privacy Impact Assessments. DPIAs are also required by the California Privacy Rights Act that will go into effect in 2023. In reality, privacy considerations are rapidly becoming a formal forethought in system design and are increasingly considered to be a formal functional requirement for any PII handling system.

Personal information (PI), personally identifiable information (PII) and personal health information (PHI) are defined in different ways by different pieces of legislation. PII is commonly interpreted to be information that can be used to uniquely identify an individual either directly or when used in conjunction with other forms of information. Cell phone locations, web surfing behaviors, credit card usage and even residential power consumption all constitute various forms of privacy data that can potentially be linked to a data subject's identity. For the purposes of this report, the terms PII and privacy data will be used interchangeably to refer to any form of data that can be linked to an individual human identity.

It's time to translate Design Principles into Design Requirements

The Agile Manifesto was published in 2001. It was a truly revolutionary document that called into question traditional waterfall methods of software development. It recommended a much more collaborative and incremental approach to software engineering that was designed to improve the productivity and accountability of development teams. The Manifesto has had a profound impact on how IT professionals go about designing and implementing new systems. Twenty years later, many if not most IT teams are still refining and extending their use of Agile principles.

The PbD Principles were published at roughly the same time but they've had considerably less impact on IT thinking or system design practices. There are several reasons why the Principles failed to galvanize the attention of the IT industry not the least of which is that they were stated in vague, evangelical terms that weren't readily comprehensible or compelling to IT practitioners. It's equally true that the Manifesto addressed deficiencies in software development practices that impacted a broad cross section of the IT community. Data security concerns were less pervasive at that time. Consequently, during the past twenty years the Principles have found a broader and more receptive audience within governmental regulatory agencies than the IT industry.

The purpose of this report is to translate the original Principles into a set of actionable Requirements that can ensure the protection of PII data in future IT systems. These Requirements are grounded in the business and technology realities of the 2020s.

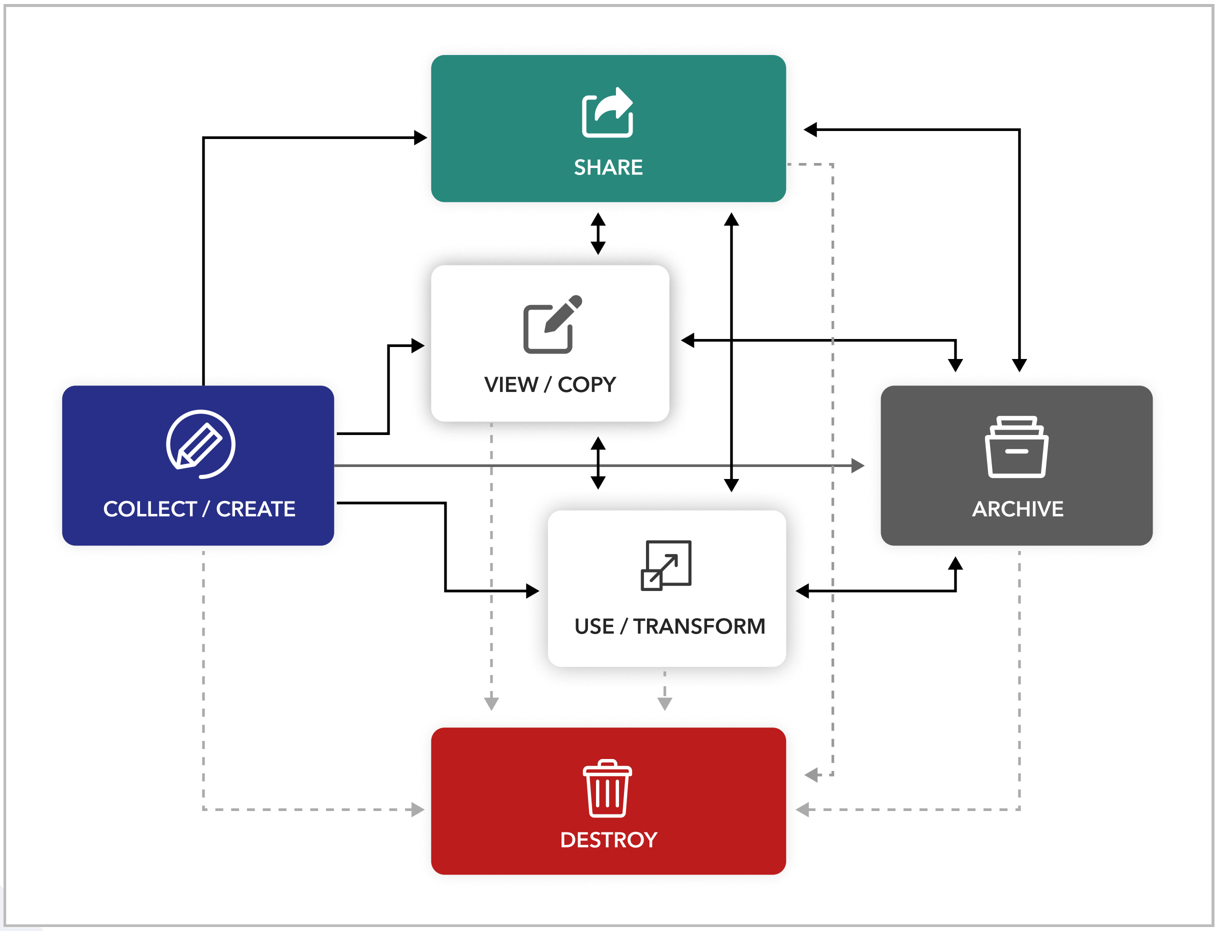
The following discussion includes references to new tools possessing capabilities that can address the functionality of each Requirement in whole or in part. However, this report does not prescribe a reference technology stack for building privacy-preserving systems because no such stack currently exists.

Note that this report builds upon several concepts discussed in NextGen DLP: Data Misuse Protection, a companion report authored by Mark Settle and Sid Trivedi.

In many places, the following discussion revises or restates portions of the original Principles. This is not an exercise in plagiarism but rather an attempt to present PbD concepts in a terminology and context that IT professionals can easily understand, and more importantly, act upon.

Requirement 1. Full Lifecycle Protection

Data assets can have very complex histories. After PII data has been initially collected it can be used to construct stores, databases and files that can be viewed, copied, used or transformed in various ways. Derivative assets resulting from any of these events can be shared with others or they can be archived or destroyed. Many of the critical assets used to support routine business operations are continuously modified by one or more of these events performed in a sequential manner that's rarely standardized and devilishly difficult to reconstruct after the fact.



Data Lifecycle Events

Privacy protection begins at the point of data collection and is instantiated in the construction of primary data assets. Primary assets need to contain a rich set of metadata that can be used to protect the PII they contain. This metadata should include the following:

- Information describing how, when, where and why the PII was originally collected as well as who is responsible for the stewardship of the primary asset.
- Information referencing the policies that govern the use of the PII whether they are based on regulatory requirements or discretionary obligations that a company may have voluntarily assumed at the point of data collection.
- Data sensitivity classifications – companies may choose to establish their own sensitivity classification schemes or adopt those in common use by others.
- Data usage restrictions which may be specified in terms of business purpose, retention time, geography, etc. It's likely that usage restrictions will become more granular in the future as data subjects are offered a broader range of opt-in and opt-out choices at the point of data collection.

These forms of metadata should be inherited by all derivative assets. They provide a basis for managing and protecting all subsequent data products. As noted above, it's extremely difficult if not impossible to reliably reconstruct these forms of metadata in assets that are separated from primary data sources by dozens of intermediary data products.

Derivative assets should be accompanied by similar information regarding the timing, purpose and method of their construction as well as their ownership, compliance requirements, sensitivity and usage restrictions.

Full lifecycle protection requires the continual maintenance of metadata regarding an asset's chain of custody. Chain of custody is a broad concept that should include the following:

- Business custody - the business departments and business leaders who assumed responsibility for the proper management and handling of derivative assets.

- Usage custody - the individuals who have been granted access permissions and authorization privileges to derivative assets. This information is usually maintained in access and authorization systems or various types of logs. Asset metadata merely needs to provide pointers to the systems or logs containing usage history.

- Environmental custody – the infrastructure environments that have hosted successive assets derived from a primary source.

PII lifecycle protection ends at the point of data destruction or data anonymization. Anonymization can be accomplished in several different ways but its net effect is to irreparably destroy the connection between information supplied by a data subject and the data subject's human identity. In privacy parlance, a data subject's identity is no longer discoverable following anonymization.

Privacy data management exposes commercial firms to a variety of risks and burdens. Improper handling or usage of PII data may result in financial penalties, brand damage or an irreversible loss of customer confidence. Company employees may experience significant friction in performing their jobs due to the safeguards that have been put in place to protect PII. These risks and burdens provide powerful incentives for anonymizing and destroying PII at a file/database/store level or a field/record level at the earliest possible opportunity.

Differential privacy techniques involving the injection of noise into assets containing PII data provide a further measure of protection that may be warranted in certain circumstances, especially those involving public disclosures of partially de-identified PII. However, standard anonymization techniques are sufficient to protect PII employed in the vast majority of internal business operations

Realistically, it will be some time before the enriched metadata described above is routinely specified in all PII source systems and consistently passed down to all derivative assets. In the meantime, data discovery tools will continue to play a critical role in identifying the location of existing PII and inferring its heritage as accurately as possible.

A variety of vendors are developing new capabilities for mapping data flows, inferring data lineage, enriching metadata and instituting usage controls that are applicable to the full lifecycle protection of PII assets. These capabilities are illustrated by the following companies.

- **Octopai** automates the discovery of metadata information within assets constructed by a wide variety of database, ETL, BI and reporting systems. This metadata provides a basis for reconstructing asset lineage. Octopai can also be used to establish a consistent glossary of business terminology through the inspection of data expressions within the physical, semantic and presentation layers of common reporting systems.

- **Solidatus** provides an ML-assisted means of mapping the flow of specific data fields across multiple assets enabling the reconstruction of data lineage at a highly granular level. It also can be used to facilitate the construction of enterprise-wide data dictionaries, catalogs and business glossaries. Its discovery and cataloging capabilities can be applied to assets created by both modern and legacy (mainframe) systems.

- **Alex Solutions** harvests metadata from existing data assets and employs pre-defined rules and ML-based inference techniques to contextualize such metadata in a consistent fashion. In many instances lineage relationships can be readily reconstructed from internally consistent metadata catalogs.

- **Cyberhaven** provides a data tracing solution that continuously tracks file movement and ownership through multiple channels such as email, Box, Zoom, MS Teams and Slack, without employing any classification or tagging procedures. File lineage can be determined retroactively and monitored prospectively.

- **Manta** captures lineage information by continuously scanning software algorithms that act upon data, not the data itself. Many conventional solutions infer lineage by matching identical data fields detected in multiple assets. Manta establishes lineage relationships by monitoring the code being used to construct derivative assets.

- **TigerGraph** employs graph database technology to discover and map lineage relationships at multiple levels including data stores, tables, files, fields and cells.

- **Privitar** controls the usage of data within protected domains by requiring metadata documentation of individual usage requests including information concerning the requester, business purpose, additional users, duration of usage and the request approver. Digital watermarks can be embedded in assets to document data provenance and monitor data movements.

- **DoControl** monitors data lifecycle events such as create, view, share and edit occurring within SaaS applications, collaboration tools and file sharing services. It can provide notifications of such events, suspend their execution pending management approval or block them altogether on the basis of contextual parameters such as user identity, organizational hierarchy, file history, usage time, etc.

Amundsen and **Datahub** are two open source tools developed by Lyft and LinkedIn that can be used to discover and catalog metadata within existing assets. Existing metadata can play a key role in reconstructing lineage relationships. It can also be used to restore critical information regarding sensitivity classifications or usage restrictions that may have been lost in the construction of derivative assets.

Requirement 2. Maximum Use of Embedded Safeguards

The 'shift left' movement in software engineering originally referred to testing newly constructed code as early and often as possible in the development lifecycle. Early detection and correction of software bugs plays a key role in enabling the continuous integration and deployment of new software capabilities. Security as Code is a natural extension of the shift left movement. It refers to the application of security tests and vulnerability scans to newly constructed code as early and often as well. More recently, the term Policy as Code has become popular. It refers to the insertion of controls within newly developed software systems that can be used to regulate user authentication, API authorization, container resource utilization or infrastructure configurations in ways that are consistent with company policies.

Shift left concepts can be readily extended to establish *Privacy as Code* engineering practices. Many of the technical safeguards discussed in this report can be directly incorporated in new PII processing systems. Access permissions, authorization privileges, exposure restrictions, encryption requirements, infrastructure configurations and many other safeguards can be managed as inherent features of such systems. Although these capabilities may be supplied by third parties as API-enabled services, they no longer need to be implemented with a completely independent set of tools after a software system has been placed in production.

Privacy as Code relieves privacy teams of the cost and burden of implementing selective safeguards after a system has gone live. These practices also enable developers to extend the functionality, customize the usability and improve the resiliency of new systems, as well as accelerate their time to market.

Privacy as Code toolkits and practices are in early stages of development. They're certainly not complete and many lack the sophistication of existing tools developed for similar purposes. The following vendors illustrate some of the Privacy as Code capabilities that are currently available.

- **Skyflow** provides a cloud-based vault for sensitive data that can be accessed via an API, eliminating the need to host PII in any form. Data within the vault is encrypted in multiple ways, providing an additional measure of protection.

- **Evervault** is a proxy service that can encrypt data being ingested by an application or data service and decrypt it on its outbound delivery to third party APIs or users. If desired, encrypted data can be processed within serverless execution environments secured and hosted by Evervault.

- **Auth0** is a developer toolkit containing APIs and widgets that can be used to build customized end user authentication procedures for accessing most common databases. (Okta purchased Auth0 in 2021.)

- **Oso** and **Authzed** are developer toolkits that can be used to build customized procedures for granting fine-grained authorization privileges to end users and enforcing authorization policies.

- **Cyral** is a stateless interception service for data endpoints that can monitor usage and enforce access policies on a real time, in-line basis, eliminating reliance on traditional monitoring agents and host-based policy management procedures. It's designed to be integrated into existing DevOps/SecOps workflows and deployed by tools such as Terraform (HashiCorp) and CloudFormation (AWS).

- **PlainID** is a policy engine that can be inserted between the business service and data layers of an application to regulate data access at a finer scale, independently of the coarse grained rules governing application access.

- **Transcend** provides developers with APIs that can be used to automate the implementation of data access/erasure requests and consent agreement changes via customized workflows, eliminating the need for human intervention in the administration of data subject rights.

- **Laminar** provides a platform that can be integrated into CI/CD pipelines, providing development teams with immediate feedback on the implementation of security policies before code is placed in production. Policies are encoded in configuration files that can be automatically tested in dev, staging and production environments.

- **Terratrue** flags privacy protection issues and needs throughout the software development process through deep integration with common development tools such as Github, Jira and Slack. It enables safeguards for data subject access and Record of Processing Activity (ROPA) updates to be inserted in new systems before the fact, eliminating the need to satisfy such requirements post-production. Terratrue's knowledge base is continually updated to reflect global regulatory requirements.

Requirement 3. Exposure Based Upon Least Need

As indicated earlier, attempts to limit the collection of PII are becoming increasingly futile as companies seek deeper insight into their customers' interests, preferences and behaviors. This makes the minimization of PII data upon subsequent viewing, copying, use, transformation or sharing doubly important.

The *principle of least privilege* is familiar to all IT professionals. It's commonly used to restrict access to application modules, data assets and infrastructure entities. It's used to place further restrictions on the actions or privileges a user is authorized to perform after access has been gained. And finally it can be used to restrict a user's entitlement to perform those actions on specific data sets or configuration controls. For example, a member of the HR compensation team may be permitted to *access* Workday's compensation module, he may be *authorized* to modify data within that module, but he may not be *entitled* to apply this privilege to the records within the executive compensation database. In practice, the terms permission, privilege and entitlement are used interchangeably but in reality there are important distinctions between access permissions, authorization privileges and data/configuration entitlements.

The principle of least privilege is relevant to the handling and use of PII data but it needs to be restated for purposes of clarity and utility. *Need to know* is a similar principle that's been used by military organizations for centuries to restrict the exposure of information concerning troop movements and battle plans. This well-known concept is directly applicable to the usage and handling of PII data and can be succinctly restated as the *principle of least need*.

The principle of least need should be routinely used to restrict the nature, quantity and precision of PII data required to perform a particular task or activity, as well as the length of time such data remains accessible to a team or user. Data that is not absolutely required should be generalized, redacted, anonymized or eliminated altogether. Data that is not in active use should be archived or destroyed. In reality, there is far too much gratuitous duplication, sharing and retention of PII data in most enterprises that serves no business purpose.

There are obviously situations in which the nature and quantity of data required for a particular analysis or modeling exercise cannot be specified in advance. Experimentation may be required to cull an assortment of variables into a useful dataset. Lean manufacturing principles need to be enforced throughout this experimentation process. In lean manufacturing, anything that doesn't create customer value is considered to be waste. In building a lean data pipeline, any form of PII data that isn't required to substantiate analytical conclusions or maximize forecast accuracy should be treated as a potential liability that should be eliminated from further use or handling at the earliest opportunity.

The heightened interest in AI/ML technology has increased the size and sophistication of DataOps (Data Operations) teams in many companies. DataOps groups are actively combating the gratuitous replication and sharing of unnecessary PII data by establishing feature stores that serve as data brokerages, allowing AI/ML modelers to select only those features (variables) that are relevant for a particular task or project. DataOps groups have performed the diligence required to certify the quality and integrity of data exposed in such stores, eliminating duplicative data inspection and clean-up efforts on primary data sources by multiple modeling teams.

A wide variety of technologies are available to minimize data exposure at an increasingly granular level. Masking, encryption, redaction and anonymization techniques can be applied at a field, record and cell level within structured databases. Permissions, privileges and entitlements can be managed at these increasingly granular levels as well. These emerging capabilities are illustrated by the services offered by the following vendors.

- **Satori** employs a cloud-based proxy service that functions as a data access controller to any type of data store. Data user identities can be managed across multiple stores, independently of the user identities being used to control application access. Data access requests can be approved or rejected on a contextual basis, employing end user attributes at the time of a request. Satori can be configured to mask, redact or time bound data delivered to end users at a table, field or record level.

- **Privacera** employs Apache Ranger technology to enforce consistent access policies across a wide variety of cloud-native data services and warehouses. It also provides encryption capabilities based upon the Apache Ranger Key Management Service supporting advanced encryption and format-preserving encryption standards.

- **Immuta** provides the ability to establish purpose-based access controls for PII assets. It can be used to dynamically hash, mask, round or replace field values or k-anonymize records at query time without copying or moving the data it is shielding. Immuta also offers differential privacy (noise injection) techniques that can be applied at query time.

- **Okera** provides a platform for federated data access management that gives business departments and work teams the ability to establish customized privacy controls for individual data assets at a granular cell level.

- **Privitar** offers a wide selection of de-identification techniques including tokenization, encryption, generalization, masking, perturbation and substitution. Protected data domains established for different purposes can employ one or more of these techniques to minimize data exposure, greatly reducing the ability to link PII data across multiple domains.

- **Gretel** provides a variety of API-based services that can be used to anonymize, encrypt or replace PII fields or records in real time for application testing or model development purposes. It can also generate anonymized synthetic data with statistically equivalent properties to source PII data sets.

- **Skyflow** applies multiple de-identification techniques to data within individual fields as well as sub-elements of individual fields (such as the month/day/year elements that designate an individual's Date of Birth). These transformations are performed when data is ingested into Skyflow's data vault, not at query time. All data is served to users through a single API in a form dictated by policy (e.g. tokenized, masked, redacted or encrypted).

- **Baffle** offers a cloud-based service that can apply tokenization, format preserving encryption (FPE) and AES-256 file encryption to all stages of a data pipeline, including data in memory. Baffle's service is capable of handling both structured and unstructured data.

- **TripleBlind** provides a means of encrypting data assets and the applications that act upon them, allowing data to remain permanently encrypted throughout its lifecycle.

- **Iguazio's** data transformation tools can be applied to real time streaming data or batched data to build feature stores for ML modeling purposes. These tools can also be used to monitor feature drift over time.

Requirement 4. Shared Management of Personally Identifiable Information

Visibility and Transparency is one of the Foundational PbD Principles. These concepts require additional clarification to become actionable design Requirements. *Visibility* refers to the ability of a data subject to exercise their rights to access, modify and delete personal information that they have voluntarily submitted to a commercial enterprise. *Transparency* refers to the ability of a data subject to remain informed about the ways in which their personal information is being employed and proactive assurance that any restrictions or obligations agreed upon at the time of submission are being enforced.

As noted earlier, the solicitation and submission of personal information is the beginning of an ongoing relationship between a commercial enterprise and a data subject. Enterprises will invariably find potential uses for customer data that were never explicitly envisioned at the time of solicitation and data subjects are likely to develop concerns about the use of their information that they failed to anticipate at the time of submission. IT systems need to provide a mechanism for enabling the joint management of personal information by the enterprise and the data subject until such data has been permanently destroyed.

Individuals of differing age, gender, wealth, health, race, nationality or political persuasion can have highly idiosyncratic concerns about the use of their PII. Furthermore, these concerns change over time as an individual's circumstances change. The broadly stated usage restrictions referenced in current corporate privacy statements are unlikely to be a viable mechanism for obtaining widespread PII in the future. Consent agreements will likely require additional specificity regarding the context in which PII will be used. The greater the sensitivity of the information an individual submits to a commercial firm, the greater the burden on that firm to describe the ways in which such information will be used.

Broad statements regarding the intended use of PII by a commercial firm need to be translated into generic use cases that are meaningful to a data subject, with reference to such activities as personal marketing, product development or customer support. Personal marketing might involve future offers of early access to new products, free shipping or product discounts. Product development might involve the use of PII to evaluate the efficacy of personal healthcare products or detect trends in clothing preferences. Customer support could involve offers to join affinity groups of like-minded consumers who share common cooking interests or pet care concerns. Future systems need to provide data subjects with the flexibility to opt in or out of generic use cases that are implied but not formally specified within conventional privacy statements.

Progressive enterprises may also offer data subjects the opportunity to opt in to certain types of alerts or notifications regarding the use of their PII. In much the same way that a bank may notify its customers about an account login event, password change or significant cash withdrawal, companies may elect to inform customers about the use of their PII in designing the next version of the Apple iPhone or developing a joint marketing plan offering exclusive discounts on Nike and Lululemon products over the holiday shopping season. Privacy-related alerts and notifications represent a form of active transparency that is likely to become more highly desired in the future and potentially required in certain circumstances.

It's doubtful that every system employing PII data within an enterprise would offer data subjects a unique interface to address the visibility and transparency requirements referenced above. System-unique dashboards would be difficult to maintain and confusing to data subjects. It's far more likely that companies will develop a single corporate interface that provides data subjects with an enterprise-wide view of a company's PII holdings and a corresponding ability to exercise their rights and monitor the usage of their data on an enterprise-wide basis. Under these circumstances, it's incumbent on processing systems to establish APIs that expose the nature and use of the PII they are manipulating.

Data subjects require access to self-service tools that enable them to inspect their data, exercise their rights, opt out of specific usage scenarios and opt in to specific notifications without recourse to lengthy, cumbersome request processes facilitated by human agents. European Union supervisory authorities have received multiple complaints from data subjects regarding the administrative difficulties that they've encountered in trying to exercise their GDPR rights. Automation technologies are far too sophisticated to justify the continuation of form-based privacy management processes requiring extensive human interaction.

In many ways the shared management capabilities outlined above exceed current regulatory requirements. They also exceed the current business practices of many firms. They're likely to become more prevalent in the future as consumer privacy concerns increase.

These visibility and transparency capabilities may initially appear to impose additional burdens on commercial firms with little or no apparent business benefits. Progressive firms have already realized, however, that investments in these capabilities foster a deeper sense of customer trust that pays long term financial dividends. It also makes it easier to obtain additional PII regarding customer interests and behaviors that can be translated into future merchandising opportunities.

A variety of vendors provide platforms for managing data subject consent agreements and administering data subject rights. **OneTrust**, **TrustArc**, **Wirewheel**, **Ethyca** and **DataGrail** are leaders in this space which is commonly referred to as PrivacyOps (Privacy Operations). **Securiti**, **Ketch**, **Soveren** and **Manetu** are more recent entrants. The success of these firms to date is largely attributable to privacy controls required by GDPR and CCPA regulations.

Mine takes a different approach. It provides individuals with a digital assistant that can discover firms holding their PII by scanning email subject lines. Users can delete PII from companies or services they no longer use. Mine's capabilities can also be used by companies.

All of these vendors are forerunners of the more comprehensive shared management platforms envisioned above.

Requirement 5. Proactive Privacy Assurance

PII protection cannot be achieved solely through the implementation of technical safeguards. Operational processes are also required to ensure that IT systems are managed in ways that minimize privacy-related risks. Several of the processes discussed below exceed the rigor of contemporary practices but are likely to become more common in the future. Note that several of these processes make selective use of one or more of the technologies discussed earlier in this report.

Privacy controls enforcement. IT groups within U.S. public companies are intimately familiar with the SOX controls that govern the use of financial systems. Similar control frameworks need to be established for privacy protection. Controls should ensure compliance with any regulations governing a company's operations, any obligations it has assumed in consent agreements established with data subjects and any operating standards it has freely elected to impose upon itself. Although IT groups frequently view controls as a bureaucratic imposition on their agility and efficiency, controls are essential in ensuring that a commercial firm is doing in practice what it claims to be doing in principle. Privacy controls should be engineered and audited in the same fashion that financial, safety or security controls are managed today.

Ancestry testing. To ensure that metadata is being continuously enriched with lineage information, assets containing sensitive PII data should be randomly inspected on a selective basis to determine if chains of business, usage and environmental custody can be reconstructed.   Note that this requirement may actually be satisfied in practice by MLOps (Machine Learning Operations) teams that devote considerable effort to understanding the genealogy of data employed within ML models.

Archiving and retention policies. Assets containing sensitive PII may be subject to more stringent archiving and retention policies than other forms of data. Assets that are only used sporadically should be archived. Assets that have been unused for some predetermined period of time should be destroyed. Asset lineage may also be used in defining effective archiving and retention policies. Primary assets and their immediate derivatives may be retained for use for longer periods of time than 10th, 20th or 30th generation derivative products that could be reconstructed if necessary.

Periodic rollback of unused data rights. Unused PII access permissions, authorization privileges and entitlements should be suspended or cancelled after prespecified periods of time. Note that modern security tools enable these rights to be managed at a highly granular level. Permissions, privileges and entitlements can be rescinded at a file, table, field, record or cell level.

Continuous expansion of time-denominated or attribute-based data rights. A complementary means of minimizing PII exposure is to grant selective data rights to users on a qualified basis, limiting the duration of such rights or the circumstances under which they can be exercised (e.g. user location, user device, time of day, etc.).

Proactive validation of PII data currency, accuracy and completeness. Progressive companies may invite data subjects to periodically review PII that is being maintained within their systems. Firms may also elect to publish periodic privacy statements providing such information.

Discovery testing. PII data shared with business partners or externally exposed in some fashion should be screened to determine if it can be used to infer the human identity of any data subject. In some instances the ability to link certain forms of PII (e.g. travel loyalty program memberships) to unique human identities may be permissible whereas in other cases (e.g. prescription drug usage) it may not.

The business criticality of proper PII handling

Privacy awareness has grown considerably since the PbD Principles were published 25 years ago. Growing awareness has been fostered by government agencies that have imposed regulatory restrictions upon the commercial use of PII. It's been reinforced by well-publicized PII breaches that have grown in size and frequency as well.

2021 B2C Breaches

Kroger	FATFACE UNITED KINGDOM	facebook	GEICO	PELOTON
2/22/21	3/22/21	4/5/21	4/19/21	5/31/21

The IT industry has been slow to respond to expanding awareness of privacy concerns. Its response to date has been largely focused upon breach prevention and regulatory compliance. Privacy safeguards employed in today's IT systems are frankly insufficient to satisfy the broad, unpredictable and idiosyncratic concerns of individuals who have contributed their PII to commercial organizations. Corporations that employ PII in ways that are deemed to be unsanctioned or unethical by such individuals are exposing themselves to considerable risk even if they have successfully prevented breaches and are fully compliant with existing privacy regulations.

PII is essential to the success of almost every B2C business and many B2B2C businesses. The possession of PII by such businesses should be considered a conditional privilege that can be easily revoked by their customers and partners. Companies that go to extra lengths to ensure the proper handling of PII are not only more likely to retain their current customers but also more likely to obtain additional PII that will enable them to anticipate their customers' needs and desires in the future. The investments in PII systems engineering and operations envisioned in this report are a small price to pay to obtain such business-critical information. The future success of many businesses may literally depend upon it.

Suggested Reading

Privacy by Design, The 7 Foundational Principles, Anne Cavoukian

Information privacy law, Wikipedia

CCPA and GDPR Comparison Sheet, Laura Jehl and Alan Friel, Baker Hostetler LLP, 2018

U.S. Federal Trade Commission Privacy Impact Statements, through 2021

The Cost of Privacy, Okta, 2020

The Top 8 Benefits of Data Lineage, David Loshin, Erwin, August 2019

Magic Quadrant for Metadata Management Solutions, Gartner Research, November 2020

Magic Quadrant for Data Quality Solutions, Gartner Research, July 2020

The Metadata Revolution, Priyanka Somrah, Work-Bench, January 2021

Data Security in the SaaS Age: Focus on What You Can Control, Mike Rothman, Securosis, June 2020

Emerging Architectures in Modern Data Infrastructure, Matt Bornstein, Martin Casado and Jennifer Li, Andreessen Horowitz, October 2020

Real Time Feature Engineering with a Feature Store, Adi Hirschtein, Iguazio, December 2020

Glossary

AI – Artificial Intelligence

API – Application Programming Interface

B2C – Business to Consumer

B2B2C – Business to Business to Consumer

BI – Business Intelligence

CCPA – California Consumer Protection Act

CI/CD – Continuous Integration/Continuous Delivery

DataOps – Data Operations

DPIA - Data Protection Impact Assessment

ETL – Extract Transfer and Load

FPE – Format Preserving Encryption

GDPR – General Data Protection Regulation (European Union)

HIPAA – Health Insurance Portabiity and Accountability Act

ML – Machine Learning

MLOps – Machine Learning Operations

PbD – Privacy by Design

PCI DSS – Payment Card Industry Data Security Standard

PHI – Personal Health Information

PI – Personal Information

PII – Personally Identifiable Information

ROPA – Record of Processing Activity

SaaS – Software as a Service