Addressing the perils of PII with the De-Identified Data Lake
There have been more than 13 billion data records stolen since 2013. That is 72 records lost or stolen every second. Companies across industries, such as Yahoo, Equifax, the Red Cross Blood Service in Australia, and Facebook, have all dealt with the challenge of keeping personally identifiable information (PII) safe.

The EU’s GDPR, the California Consumer Privacy Act (CCPA), Canada’s data privacy laws, and Australia’s Consumer Data Right policy aim to prevent data breaches as much as possible. The goal is to protect the PII, which includes names, addresses, social security numbers, IP addresses, and account numbers—to name just a few—of individuals and consumers. In the cases of General Data Protection Regulation and the California Consumer Privacy Act, a citizen of the EU or resident of California can request, at any time, that their data not be shared, not be used in certain ways, or be “forgotten.”

But these laws carry with them some tough decisions. How can a company remain cost-effectively compliant and still be able to use data for innovations that add value to the business? Storing data is not enough. In many cases, organizations are not always aware of what data they have, where it’s stored, how sensitive and how accurate it is. This can prevent organizations from getting value from their data. It can also create significant risks: from data breaches and misuse of data, to loss of corporate secrets and customer trust.
Innovation vs. privacy

We are currently in what Reinventure calls the "data economy," describing it as “the trade in data between organizations and governments, domestically or internationally, and the derivative data products (algorithms, insights, applications) that arise from that previously unavailable flow of data." In the data economy, innovation is at odds with privacy, with a premium placed on innovation in many cases.

For example, China has fewer privacy regulations or PII usage restrictions than Europe or the U.S., and the Chinese are perceived as dominating areas of innovative technology like AI and driverless cars. A 2017 Forbes article quotes Kai-Fu Lee, Chairman and CEO of Sinovation Ventures, as saying that, in China, there is a “conducive environment for fast launch and fast iterations,” adding “Compared to the U.S., there is less focus on data privacy.” An August 2018 article in Foreign Policy magazine says that Chinese companies have remarkably broad access to the social and purchasing habits of Chinese Internet users. For example, Tencent’s WeChat platform uses broad and deep access to data to deliver AI-driven services to Chinese users.

On the other side of this equation are the European Union and the state of California. They have tilted the scales of innovation in the other direction by emphasizing protections of PII that existing technology cannot currently meet well and, as a result, could impact commerce. The difficult part is complying with all these requirements without adversely impacting the ability to extract value from data and continuing to innovate. A powerful solution from AWS Data and Analytics Competency Partners aims to address that challenge.
The De-Identified Data Lake

AWS and AWS Data and Analytics Competency Partners have a broad approach to data governance based on an architecture called the De-Identified Data Lake (DIDL). A data lake enables organizations to store massive amounts of data in a central location. From this data lake, diverse groups in an organization can access it easily to categorize, process, analyze, and consume company data. Since data can be stored as-is, there is no need to convert it to a predefined schema. In addition, querying this data for analytics no longer requires knowing the questions beforehand.

A de-identified data lake solves the data privacy problem by de-identifying and protecting sensitive information before it even enters a data lake. Because it minimizes the storage and the use of PII, there is less risk of data breaches and misuse of data, and compliance costs can be lower—without losing the ability to understand and use your data for competitive advantage.

Designed by AWS, this solution helps organizations address the underlying challenges related to storing and processing PII. DIDL addresses the perils of PII in the data lake, delivering a reduction in the risks associated with the proliferation of consumer identities in an organization.
Based on a data de-identification process with additional elements for extra security and precaution, DIDL detaches valuable data about habits, behavior, region, and non-identifiable preferences from the actual identity of a person. In other words, the data is no longer associated with a person. This makes the data less attractive to hackers and rogue agents. There are only attributes in the data lake, and they are not tied to personal information.

The components of the solution include a data transformation stage; a data catalog; a de-identification manager and data linking software for de-identified matching; staging; and data analytics, processing, and monitoring.
Data transformation

DIDL includes AWS partner technology and services that extract, de-identify, and transform data from a broad range of sources—databases, CRM, marketing automation, and other software-as-a-service—before loading into the data lake. By transforming data at this stage, organizations can get to the root of any problems at the beginning and remove them. The solution includes a configuration UI, monitoring and alerting services, transformation and modeling, automated ETL operations, and managed infrastructure on AWS.

The specific transformation process used is “tokenization,” which means that the records still exist but the names and other PII are replaced by a token. The goal is to ensure that businesses have access to the behavior or transactions that belong to a person, but not know who the person is. This data transformation process can be used for structured, unstructured and streaming data, and data exchange as follows:

- **Structured data:** For data in relational databases, DIDL identifies the PII fields and then calls a service that replaces the PII data with a token so all the records still exist, but the PII has been replaced.

- **Unstructured and streaming data:** DIDL looks for personal information in images (such as MRIs, audio, video) and other unstructured data, and tokenizes it. It also quarantines images or flat files that might contain PII components like names into a different bucket for manual review. For streaming data, DIDL can scrape any PII, such as an IP address, and replace it with a token.

- **Data exchange:** When two parties need to share data, such as sending information about an employee to a second party for payroll processing, PII matching software uses a dual process of tokenization and hitching. The data being shared is tokenized and then linked, or hitched, to another token.
A data catalog is the central authority for all data surrounding an organization. This single pane of glass is critical to understanding your metadata. It is also used to determine the lineage of data so it is possible to assess how reports, for example, might change or break if a field were removed from the original dataset.

The DIDL data catalog is a solution that monitors every asset in the data lake and provides data stewards the ability to manage access to data assets. It provides the central place for managing data and how it is used. The data catalog sits above storage and compute layers as a virtual private cloud that contains a combination of elastic IPs and public and private subnets. Cluster instances are managed by the partner’s cluster management services, as illustrated here.

A data catalog provides management and governance of data throughout the data supply chain from source to consumer, resulting in a solid foundation for secure, reliable, analytics-ready data. It has controls in place to safeguard data privacy and comply with regulatory requirements.
De-identification manager and de-identified data matching

The DIDL de-identification manager is a policy-based partner solution that extracts and obfuscates or masks sensitive PII data before it ever lands in a data lake. Common methods of obfuscation—encryption, hashing, and removing characters and standardizing lengths—DIDL recommends hashing, which is the transformation of a string of characters into a fixed-length value or key that represents the original string.

For de-identified matching and transfer of matched data between two parties that is regulatory compliant, DIDL has a set of utilities provided by a partner. Each party starts by sending their PII fields into a hitch contributor node, which hashes the data, then slices it into smaller pieces. Portions of those small pieces of what used to be PII are sent to Matcher nodes so they only have a small piece of a hash of what used to be PII. Those systems match the slices, which results in false positives. The true and false matches are sent to the aggregator, which searches for a true match across all of the matching nodes. This allows for de-identified matching for the first time.

As a match table is created between the parties, a partner platform handles the legal and regulatory compliance of transfer of non-PII data between the parties.
Analytics and data processing

After all the data is in the data lake and de-identified, analytics kick in. Analytics services such as Amazon EMR, Athena, QuickSight, and Redshift Spectrum can query the data in the data lake directly. The information can be pushed to a staging service or accessed directly from the data lake, creating curated data that can live in a data warehouse or be accessed directly by data analytics and visualization applications.

Data processing, analytics solutions, including machine learning, dashboards, reporting, and solutions can be part of the DIDL. Machine learning models are available to use for predictive and prescriptive analytics. Business analysts and others can build interactive visualizations and derive value from DIDL data in BI dashboards and reports. APIs and real-time analytics capabilities enable the extraction of data from the DIDL for log analytics, full-text search, application monitoring, clickstream analytics, and much more.

Monitoring

A key aspect of operating a data lake environment is understanding how all the components that comprise the data lake are operating and performing, and generating notifications when issues occur or operational performance falls below predefined thresholds. Using AWS and partner solutions, DIDL can collect and track metrics, collect and monitor log files, set thresholds, and trigger alarms that enable automatic and proactive reaction to issues. These can include changes in DIDL resources, compute and database instances, and custom metrics generated by other data lake applications and services.

DIDL also has the capability to continuously monitor and retain events related to API calls across the AWS services that comprise a data lake. It provides a history of AWS API calls for an account, including API calls made through the AWS Management Console, AWS SDKs, command line tools, and most Amazon S3-based data lake services. Organizations can identify which users and accounts made requests or took actions against AWS services that support CloudTrail, the source IP address the actions were made from, and when the actions occurred.
The DIDL data governance process

What does the process of data governance with DIDL look like? The DIDL data governance process consists of discovery, classification, and implementation. Discovery determines where PII exists. Classification—raw data, portal data, de-identified data—determines which data fields are needed and where. Implementation enables organizations to extract what they need from every data source safely.

This process can be further broken down into five steps: data discovery and cataloging, data de-identification, data matching, data stewardship, and information quality.

Step 1: Data discovery and cataloging

The root cause of a data breach is the collection and storage of PII. The use and proliferation of PII increases data breach risk. Once PII leaves a collection system and lands in the data lake as a single source of truth, the proliferation of PII has an exponential risk effect. Where once only a small number of people had access to PII, it is now available to potentially hundreds or thousands of people. Therefore, before de-identifying someone’s personal data, discovery of all that person’s data and how and where it is collected is a must. It is not enough to find the data—understanding all the places where it resides is also critical. Is it in internal databases? Files? What about in images, video, healthcare MRIs and EMRs that show the patient’s name? What are the source systems? Where has it been collected?

The DIDL process, therefore, starts with data discovery and cataloging, combing through systems, databases, data lakes, SaaS offerings, and anywhere else PII might be hiding, including high-risk data environments. The architecture supports a dual-test process of using a partner discovery tool and then ensuring no PII has made it into the DIDL using Amazon Macie. The data cataloging process enables companies to record all those places where a person’s data is created, captured, and stored. Cataloging and metadata management ensure appropriate governance of data including lineage, masking, tokenization, data quality and lifecycle management. Companies know what data they have, where it resides, and whether they have permission to use it.

To transmit data to the catalog, the AWS partner ETL service sends manifest files that include a description of the data but not the data itself. The kind of information included will be the source it was pulled from, what was used to translate it and how, where it will land, who owns it, and who can use it. The manifest also includes audit data that helps in mapping lineage.
Step 2: Data de-identification

The actual PII de-identification process in the DIDL relies on data masking. It’s tempting to write data prior to removing sensitive data or PII. However, this approach carries the risk that additional actors may be able to access that PII, capturing the data before the masking is complete.

DIDL uses a hashing process to obfuscate personal information, which creates a unique hash for each field. In order to avoid a dictionary attack, the system employs a process of hash slicing. This method takes a field, multiplies it by a value, sometimes adding a “salt” value, and creates a unique hash. The primary attack of hashing is to rehash every possible string until that string matches the hash. Adding salt to the string reduces the possibility of this attack being successful, and this measure is further strengthened by hash-slicing.

Step 3: Data matching

Thanks to tokenization, hashing, and slicing, the PII is now unavailable. And, if an organization’s only goal was to make sure that no one accessed PII, that might be sufficient. But business does not work that way. Companies exchange PII with one another for all kinds of purposes, including payroll, taxes, compensation, benefits, and more.

The DIDL data matching process enables the exchange of de-identified PII between different parties. A matching service and governance capabilities link identities across different contributors without disclosing data or putting it at risk of being misused. In addition, everything is fully auditable, which helps build confidence and trust in the process.
Step 4: Data stewardship and classification

The data in the DIDL must be cared for before, during, and after de-identification, and this requires human intervention. Therefore, a person often called a data steward is in charge of the data. The data steward authorizes people to do discovery and cataloging and sets permissions on parts of the data (file, object, all rows and columns, cells), assigns the level of authorization and depth, and maintains records of the permitted use of data.

Data stewards have another critical role: classifying data into levels and defining whether the use is permitted legally or ethically. A common classification model has four levels: public, internal, confidential, and restricted.

Step 5: Analytics

Just as the DIDL process of data matching enables the exchange of de-identified data between parties as if it were PII, the DIDL analytics process helps organizations to get value from data while still complying with data privacy regulations such as GDPR, CCPA, and the Australian and Canadian data policies. In fact, organizations can keep using the same analytics tools.

Organizations can update a source, feed a recommendation engine or provide data scientists with information that can inform decisions throughout the enterprise. Nearly 100% of reports and data analysis can be completed as usual without the need for high-risk consumer identities.
DIDL: The art of the possible

So, how does all this work? Let’s consider a scenario. A global software company headquartered in the U.S. houses all its R&D, engineering, and development functions in Europe, and its customer base spans the globe. This puts it squarely in the sights of GDPR and CCPA. In 2016, the company moved from siloed data storage in various data warehouses and databases to a data lake for ease of analytics and reporting.

Fast forward to 2018 and GDPR. Although the company followed all the required steps to comply with GDPR, numerous questions remain. Do they have data that is not in the data lake? Have they anonymized all their data properly? Are people receiving email “spam” when they shouldn’t? Is the PII used in the company’s development and testing environment protected? How sure are they that payroll information is being properly shared with the company they’ve contracted with to handle pay and benefits? What about the company they’ve outsourced their sales compensation management (SCM) to?

Data discovery and cataloging provided by AWS Data and Analytics Competency Partners searches the company’s data lake—as well as its marketing automation system, CRM, email databases, development and staging environments—for PII, anonymizes it through masking, loads it into an AWS DIDL and creates a catalog of all its sources in a partner data catalog. Now they know what data they have, where it is stored, how they can use it, and more. Data matching enables them to share their employee information with the payroll company and the outsourced SCM without transmitting any PII, and U.S. customers still receive nurture emails while European citizens and developers can “opt out.” With the AWS partner solutions in DIDL, the Chief Data Officer classifies the information accurately. The company’s data scientists use data analytics to study past performance, map trends, predict sales, and build dashboards and data visualizations. These are put in a successful presentation to investors while demonstrating that no PII is at risk. Meanwhile, any issues with new data sets or PII are reported automatically with automated mitigation in most cases.
Conclusion

DIDL approaches data privacy challenges by de-identifying and protecting sensitive information before it even enters a data lake. By minimizing the storage and use of PII, it can help your organization:

- **Reduce risk.** DIDL makes it possible to remove PII information before it enters a data lake, replacing the information with functional and consistent characters. The data remains usable, but without putting privacy or compliance at risk.

- **Understand all data.** DIDL can create a catalog of all your data. As a result, you can see what data you have, where it is located, and how you can use it based on your customer's consent.

- **Reduce compliance costs.** DIDL fully automates the discovery, classification, de-identification, and ongoing monitoring of data across an organization.

- **Continue to use analytics as normal.** AWS and partner analytic services run as normal on DIDL, including interactive analysis, big data processing, data warehousing, real-time analytics, dashboards, and visualizations.

- **Increase business agility.** The flexible and scalable DIDL architecture helps your organization to incorporate and secure new sources of data quickly so you can conduct productive experiments and uncover new business opportunities.

The result? You can maintain customer trust, which is essential to business success. When customer data is obtained by unauthorized parties, it damages customers and the organization’s reputation. DIDL protects your customers and your business by implementing a de-identification policy that removes PII from records before entering the data lake, ensuring your data is de-identified.

In addition, you’re able to implement a data governance strategy. By identifying issues, defining standards, and implementing changes, you have ground rules for what can and can’t be done with your company’s data that enables you to adhere to constantly evolving compliance regulations. And because your organization has aligned people, processes, and technology with a centralized data strategy, you can begin to leverage your data to benefit larger business goals.
Ready to get started?

Visit our data governance compliance featured partner solutions page to learn more about how to reduce security and privacy risks and meet compliance needs with a de-identified data lake.