Navigating the PII Minefield:

Tools and Tactics for Ultimate Data Security

Personally Identifiable Information (PII) serves as a unique digital fingerprint, identifying individuals via data such as Social Security, Driver's License and Passport details This data originates from various sources, including online forms, financial transactions, and official documents. PII acts as a crucial link, connecting disparate data points to create a detailed profile of an individual. Safeguarding PII is paramount, as its exposure can result in privacy breaches and identity theft.

Identifying PII within an organization starts with finding it in large data sets. As the volume of data continues to increase across many sources, flowing through various locations, and undergoing frequent updates, pinpointing PII data becomes a complex challenge [1]. The difficulty is compounded by the potential harm caused if such information is lost or disclosed without authorization. An organization becomes liable for such incidents, emphasizing the need for organizations to employ suitable tools to find and secure PII data.

The Impact of Data Breaches and the Need for PII Protection

The consequences of not finding and protecting PII data are serious. In 2022, Thales Group published a global survey on cloud security, revealing that, 45% of US organizations had a data breach in 2022, which is greater than 10% compared to the previous year. [2]

Their most recent report from 2023 revealed that 47% said ransomware attacks have increased, and the volume or severity of security risks is rising. 37% of respondents said that their organization had been targeted by ransomware assaults, and more than 32% said that they had experienced a data breach in the previous year.

Data breaches, such as the significant one Microsoft faced in 2020, underscore the severity of the issue. The United States tops the list for the most data breaches, with California leading in breaches and exposed data over a 15-year period [2]. To address the rising threat of data breaches and PII exposure, major cloud companies like Amazon, Google, and Microsoft have taken steps to provide PII scanning tools. These tools are designed to find and secure sensitive information within datasets, helping organizations mitigate the risks associated with data breaches and unauthorized disclosures. [2]

siLab's Analysis of PII Scanning Tools

In response to these challenges, our Innovation Lab *(siLab)* Team conducted an in-depth analysis and evaluation of various PII scanning tools to help you decide which tools align best with your needs and provide effective PII protection. Here are the results of our analysis which will help an organization pick the PII tool for their needs.

Amazon Macie

Amazon Macie, an AWS service, uses machine learning and pattern matching to automatically find and safeguard sensitive data in S3. It provides interactive maps and dashboards for security risk prioritization, sending findings to Amazon Event Bridge and AWS Security Hub. [3]

Pros

- ML-driven automated detection.
- Interactive visualizations for risk prioritization.

Cons

- Limited to data stored in S3 buckets.

Amazon Comprehend

Amazon Comprehend employs natural language processing and machine learning to extract insights from unstructured data. It can scan millions of documents, recognizing entities, key phrases, sentiments, and masking sensitive information. [4]

Pros

- Efficient processing of unstructured data.
- Pre-packaged PII detection functionality.

Cons

- Limited to document analysis.

Azure PII Detection Cognitive Skill

Azure's PII detection tool in Cognitive Service for Language uses NLP to analyze input text. It can find, categorize, and redact sensitive data with options for web-based or API integration. [5]

Pros

- Cloud-based NLP features.
- Options for model customization.

Cons

- Requires integration for application use.

Azure Information Protection (AIP)

Azure Information Protection helps find, categorize, label, and protect data in various environments. It offers encryption, file permissions, and flexible label application for data protection. [6]

Pros

- Comprehensive data protection features.
- Manual and automatic label application.

Cons

- Some users find setup challenging.

Google Automatic Cloud DLP

Google's Automatic Cloud DLP is a fully managed service for discovering, organizing, and protecting sensitive data. It provides audit reports, dashboards, and data discovery capabilities. [7]

Pros

- Fully managed service.
- Integration with cloud services.

Cons

- Limited availability independently as an API.

Python PII Catcher

PII Catcher, a Python tool, detects sensitive data in databases and file systems using regular expressions and NLP libraries. It supports incremental scans and is available as a docker image, command-line app, or API. [8]

Pros

- Programmatic detection.
- Supports multiple databases.

Cons

- Requires some coding knowledge.
- Currently limited to detecting only a few PII elements and the backend pretrained NLP model needs improvement.

Digital Guardian Powered by AWS

Digital Guardian is a data protection platform offering data discovery, classification, and control across endpoints, networks, and the cloud. It addresses various IT areas and supports data protection for different information levels.

Pros

- Comprehensive data protection.
- Integration with AWS.

Cons

- Some users find the initial setup challenging.

IBM Security Guardium

IBM Security Guardium continuously monitors data access to protect sensitive data across various contexts. It supports a zero-trust philosophy, checking data access based on contextual data. [9]

Pros

- Continuous monitoring.
- Complete data protection across various platforms.

Cons

- Some users may find it complex.

Varonis

Varonis is a data security platform that identifies risks, reduces exposure to sensitive data, and ensures compliance. It offers features like data discovery, access governance, and GDPR compliance support. [10]

Pros

- Comprehensive data security features.
- Suitable for both on-premises and cloud environments.

Cons

- Initial setup can be challenging.
- It is costly.
- Requires additional licenses for some features.

Netwrix Auditor

Netwrix Auditor provides a centralized console for analyzing, alerting, and reporting on changes to IT infrastructure. It supports information governance, data security, eDiscovery, and compliance. [11] [12]

Pros

- Single console for analysis.
- Suitable for various IT events monitoring.

Cons

- Some known problems with integrating Active directory, slow in upgrading etc.

Conclusion

While many PII tools exist, this article focuses on evaluating leading solutions. Safeguarding sensitive data is inherently challenging, and tools, though crucial, are just one aspect. The newest, flashiest tool won't suffice without a robust data policy and strict enforcement, but effectively managing sensitive data entails compliance with privacy laws, employee education on data governance, and selecting tools aligned with the company's needs. Assessing your organization's approach to protecting PII involves understanding where such data is stored. Team Synectics stands ready to aid in finding and securing sensitive PII data.

Are you ready to Identity and protect your organization’s PII data? Contact us at smdi.com to fortify your data protection strategy.

References:

[1] https://www.imperva.com/learn/data-security/personally-identifiable-information-pii/

[2] https://cpl.thalesgroup.com/about-us/newsroom/thales-cloud-data-breaches-2022-trends-challenges

Also see: https://www.thalesgroup.com/en/worldwide/security/press_release/2023-thales-data-threat-report-reveals-increase-ransomware-attacks for the latest statistics.

[3] https://aws.amazon.com/macie/

[4] https://aws.amazon.com/comprehend/

[5] https://learn.microsoft.com/en-us/azure/search/cognitive-search-skill-pii-detection

[6] https://learn.microsoft.com/en-us/azure/information-protection/what-is-information-protection

[7] https://cloud.google.com/blog/products/identity-security/automatic-dlp-for-bigquery

[8] https://github.com/tokern/piicatcher

[9] https://www.ibm.com/downloads/cas/VJEXYRZK

[10] https://www.varonis.com/

[11] https://www.netwrix.com/data_classification_software.html

[12] https://helpcenter.netwrix.com/bundle/Auditor_10.0/page/Content/Release_Notes/NA_Release_Notes/Known_Issues.htm