EDM Webinar

Behind Closed Doors: Silicon Valley Bank's Sensitive Data Management Strategies

Live Date: February 14, 2023

Featuring:

Jennifer Mezzio, Managing Director, HR Divisional Data Officer, Silicon Valley Bank

Peggy Tsai, Chief Data Officer, BigID

Mike Meriton, Moderator, Co-Founder & COO, EDM Council

Recording: View webinar

Presentation: View slide deck

EDM Council Homepage: edmcouncil.org **BigID Homepage: bigid.com**

ADDITIONAL LINKS:

EDM Council's CDMC (Cloud Data Management Capabilities)

WEBINAR Q&A:

For Sensitive Data - its combinations of data that make it sensitive. Should the focus be on identifying underlying individual components in systems or the combinations?

It depends on the organization's classification levels for data. Best practices indicate that individual data elements as well as combinations of data elements should be given risk level ratings that can be defined differently. Normally there are higher levels of risk with certain combinations of data that can be attributed to a specific person which makes it more confidential or critical.

At what level of granularity (i.e. what and how many categories) do you classify your sensitive data?

This is specific to your organization's legal and risk compliance discretion. There may be classification scales for different purposes including security risk, internal usage, information risk etc. These categories vary from company to company depending on the industry and the sensitivity of the data that they create and collect. Examples of categories can include Public, Private, Confidential, Internal Use Only etc.

I assume you have to honor certain data sovereignty and local hosting requirements. How have you navigated these requirements? How many countries are you using to host your data as a result of these restrictions?

Organizations typically utilize local privacy officers to interpret the local sovereignty laws and they have to bear that in mind when developing an enterprise architecture that involves storing local data and moving the data into different data centers.

Based on these outcomes, I approached this aspect through ROI based data driven outcomes so I could get a buy-in for iterative cleanup of data. Any thoughts?

There are multiple benefits from cleaning and adhering to data retention from HR data records. It can help reduce the security risk of the entire organization. The ROI should not be limited to specifically this project. Instead, it should be viewed as the efforts to support privacy and security as well as data management. This type of discovery and data retention effort should be extended to all departments to achieve a higher level of compliance.

Could you please talk more about automated discovery and tagging. How do you assign metadata to records/data?

Metadata for records/data can be assigned based on operational usage and business tagging. Typically metadata is automatically created in database tables or files from the network drives. There are many types of metadata that can be derived from scanning. However the purpose of automated discovery is to leverage ML algorithms to automatically discover the right data and then apply tags based on predefined policies or business rules. For example, you can write a rule that tags all data that is more than 3 years old. This tag is then attached as metadata to that data. Eventually you

want to create a catalog that allows you to search and filter by different metadata including tags to find the information you want.

“Data dictionary” is an old, old term. What does it mean for Jennifer?

To Jennifer, data dictionary refers to a catalog.

Data governance has become fundamental to business and most firms are playing catch up. Do you agree?

Absolutely agree - data governance programs need to demonstrate and prove value to an organization. Organizations have different KPIs for their data governance programs. ROI is not the only measurement of value. Depending on the data maturity of the organization, the KPIs can vary from cost savings in headcount or technology storage to data monetization.

What obfuscation/purging techniques are you considering based on various data types?

Jennifer did not mention specific different obfuscation techniques but typically depending on the retention policy, data is deleted permanently from the organization's systems. There are other policies that define that that are put on legal hold and are held separately from production environments.

How do you approach retention from a transactional perspective versus retention from insight and analytics perspective?

There will be different retention policies for data in transactions versus insights/analytics. These requirements will dictate how it will be treated by an organization. Also different industries and global regions will have overarching retention policies for organizations to follow as well. Any conflicts will be resolved internally by general counsel.

Are you using an Ontology and the Semantic Web to manage your Sensitive data?

Typically organizations do not need an ontology or semantic web to manage sensitive data. There may be a definition of what constitutes sensitive data but it is not mandatory to have an ontology. AI-led discovery can identify and classify sensitive data independently from an ontology.

What is your definition of sensitive data?

For this webinar, sensitive data is defined as personal information (PI) or personally identifiable information (PII). It can also be extended to any critical data for that organization. For example, healthcare companies may store PHI data as well.

What Taxonomy are you applying and path towards automation?

While taxonomies can exist to help identify these data elements, it will vary by industry and policy regulation. Automation is achieved when technology solutions can automatically and accurately find sensitive data embedded in text or inside a cell value. Unstructured data is the most difficult to discern sensitive data.

Can you tell me more about BigID and how it can help identify and manage sensitive data?

Thank you for your interest in BigID. Please reach out to info@bigid.com for more information.

Any insight about leveraging AI to manage subjectivity in the classification and tagging process?

AI can help tag the subjectivity in the classification and tagging process based on predefined rules. The benefit of AI is that it can be done quickly and at large scale so that there are no errors. Manual tagging is subject to errors and can differ by each person reviewing the data. Classification and tagging is the easiest and most efficient application of AI in data management. Also, please note that there may be different classification rules for privacy and security. Therefore, AI can help provide multiple tags to the same data based on different views. This is another benefit of leveraging AI.

As a firm we are now expected to understand assumptions made within machine learning models (I believe there have recent court cases on this). How do we know that given BigIDs IP and patents? How much can client firm's actually see/know?

This is a great question! While assessing any technology, it is important to ensure that any black box computation is transparent and explainable. I recommend that as part of your due diligence in selecting a vendor, that this is clearly explained and that you can provide auditable trails to how the ML algorithms are created and changed. If you have specific questions about BigID's IP and patents inside the product, please feel free to reach out to info@bigid.com.

What is the best way to govern unstructured data? What is involved and how do you measure success?

Unstructured data has been traditionally difficult to govern due to the sheer size, volume and inability to understand the contents inside the file without manual review. This is where AI-driven discovery needs to step in to 1) classify the document 2) automatically categorize the document 3) provide recommendations on actions to take. Success is measured by pre-defined KPIs within your organization. This can include proper deletion, compliance to remediation, reduction in data storage or other measurements.

How would you classify data in log files? Any guidelines on how to treat and process sensitive data in log files?

Classification in log files is dependent on what the contents are inside the log files, whether it contains any sensitive data that requires protection, and lastly how it applies to any existing retention policies within your organization. Log files are treated as unstructured data which makes it difficult to classify without manual review. This is where it is important to utilize a vendor that can read inside unstructured data sources and automatically flag for sensitive data. Lastly, some technology tools only support unstructured data source scanning while others only support structured data sources. Optimally, you want a solution that can cover all types of data sources to maintain a consistent classification view of your data.