EDM Webinar

Navigating Unstructured Data: A CDO's Roadmap to GenAI Success

In conversation with…

Jack Berkowitz
Chief Data Officer (CDO)
Securiti AI

Ankur Gupta
Director of Product Marketing
Securiti AI

securiti

EDMCouncil

Today's panel

Moderator

Jim Halcomb
Global Head of
Research & Development
EDM Council

Jack Berkowitz
Chief Data Officer
(CDO)
Securiti AI

Ankur Gupta
Director of
Product Marketing
Securiti AI

EDMCouncil

securiti

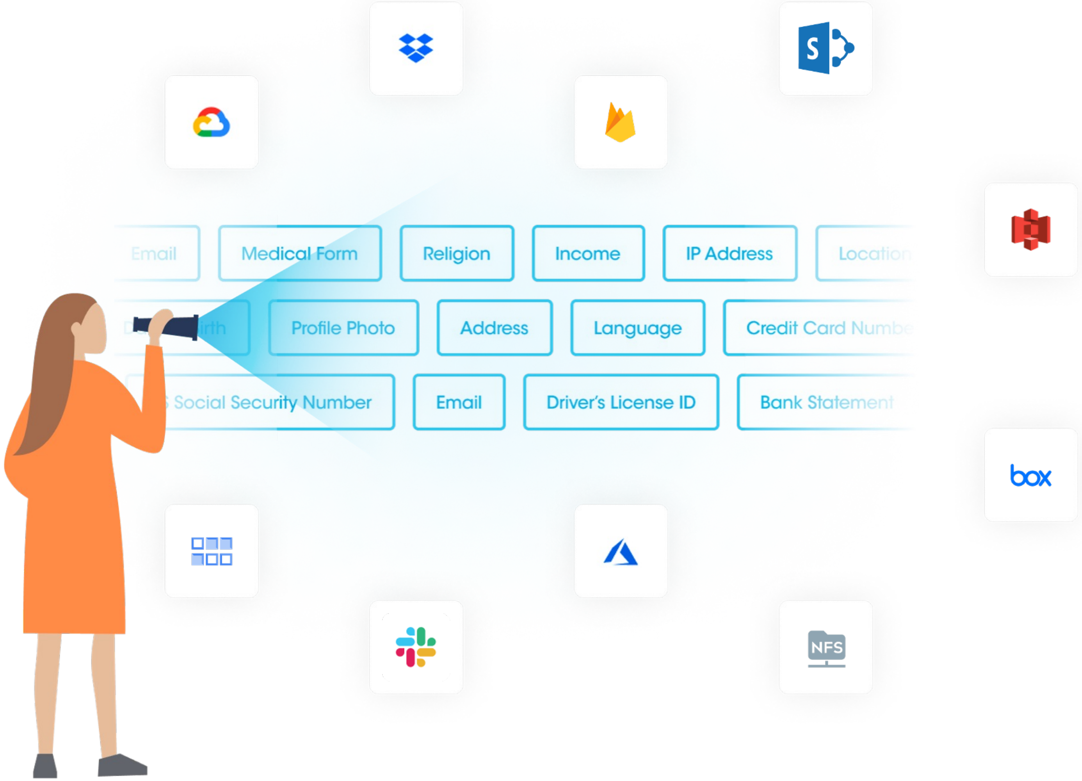
securiti

Poll Question

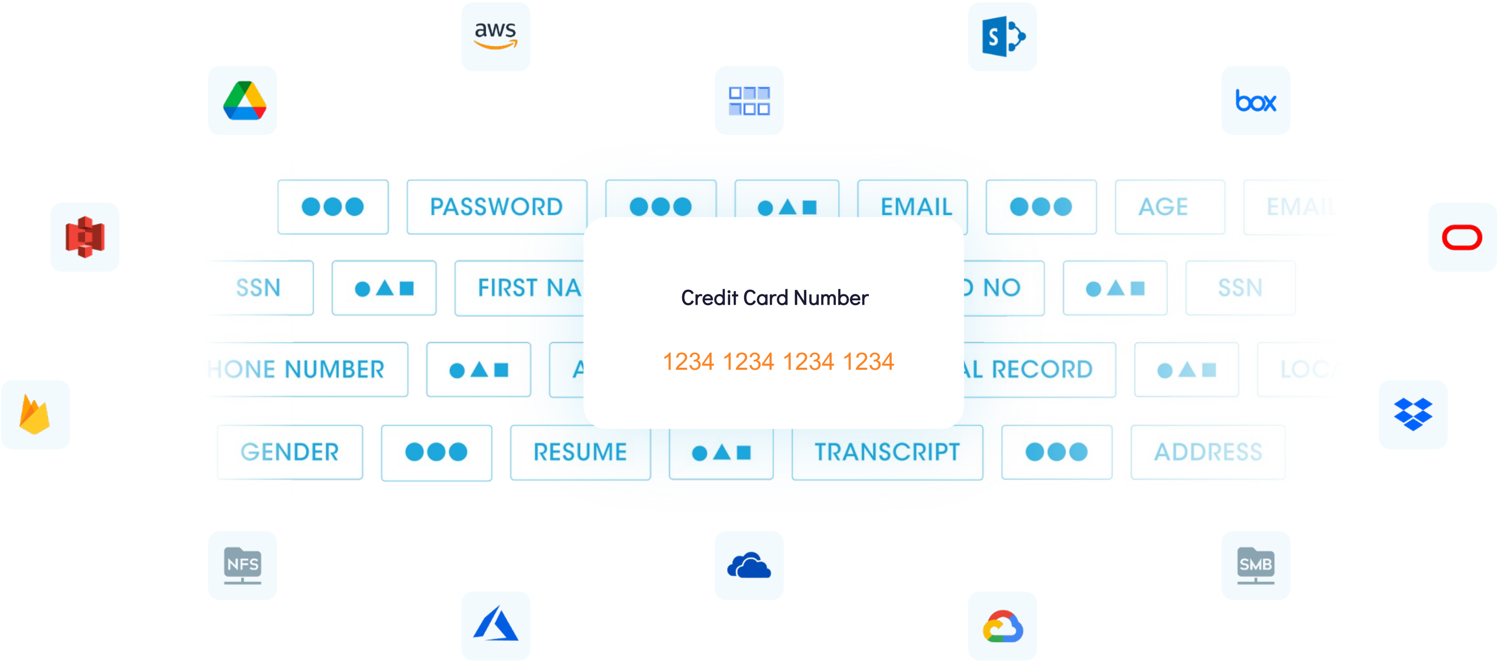
Is your company's unstructured data ready for generative AI?

1. **Fully Ready:** We have a robust data infrastructure and governance in place for unstructured data.
2. **Mostly Ready:** Our unstructured data is well-organized, but we need to implement more governance, privacy, and security controls.
3. **Partially Ready:** We can govern structured data well but need to improve our governance of unstructured data.
4. **Starting Out:** We have just begun preparing our unstructured data for generative AI.
5. **Not Ready**: We need to significantly overhaul our unstructured data systems and data governance strategy.

1. I can't effectively discover, classify, and label unstructured data.



2. I struggle with preventing the exposure of sensitive unstructured data.

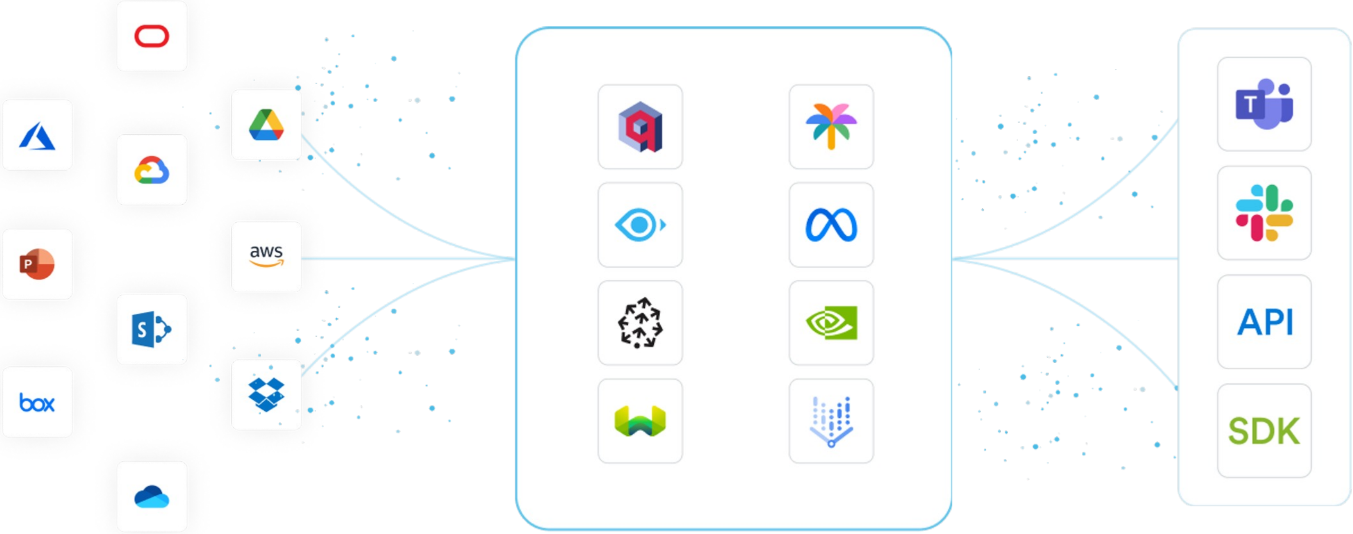


3. I can't easily determine who has access to sensitive data within unstructured data and GenAI systems.

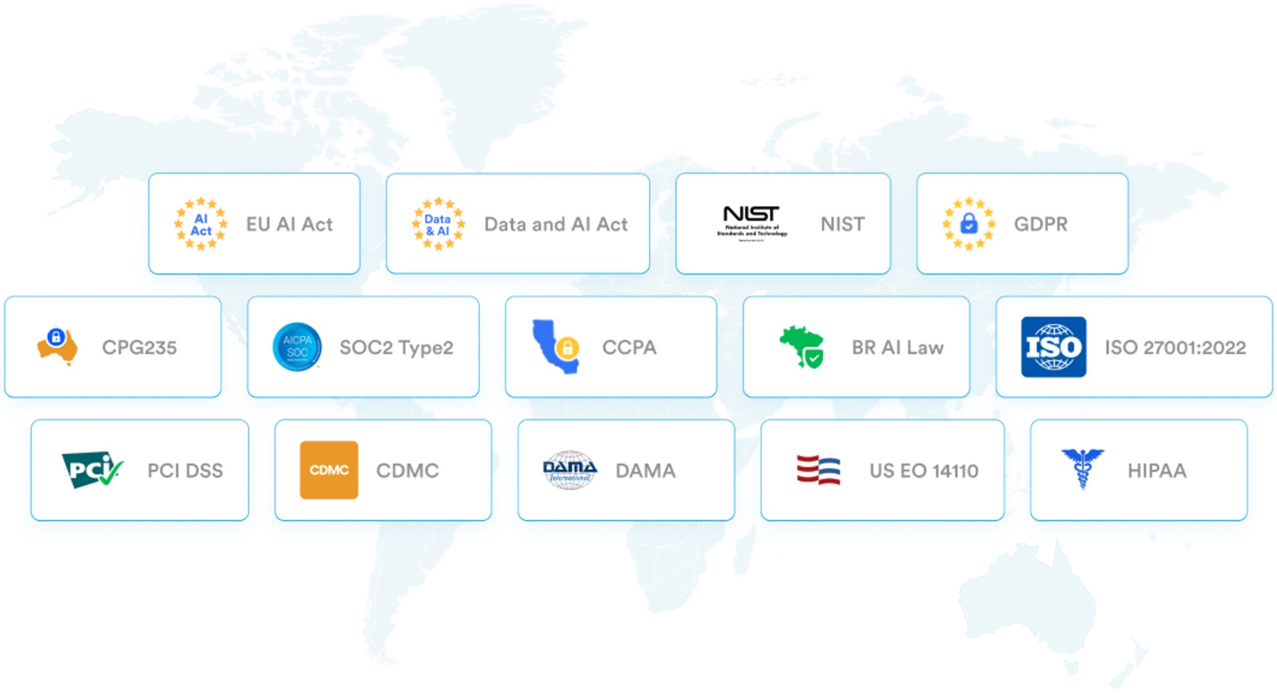
4. I struggle with ensuring my data is relevant, fresh, and free of duplicates.

aws
box
NFS
SMB

5. I can't effectively trace the lineage from files to GenAI models & endpoints.



6. I struggle with understanding what policies and regulations apply to unstructured data.



- EU AI Act
- Data and AI Act
- NIST
- GDPR
- CPG235
- SOC2 Type2
- CCPA
- BR AI Law
- ISO 27001:2022
- PCI DSS
- CDMC
- DAMA
- US EO 14110
- HIPAA

90% of organizational data is now

unstructured, and must

be safely governed when used by GenAI models.

New capabilities are required to govern ‘unstructured data’

1. Discovery of Unstructured Data

Discover unstructured data assets across diverse storages.

2. Cataloging of Unstructured Data

Catalog all files and objects that can be used for genAI projects.

3. Classification of Unstructured Data

Classify all files and documents based on their sensitivity and other attributes.

4. Entitlements of Unstructured Data

Preserve enterprise entitlements at source systems for genAI projects.

5. Lineage of Unstructured Data

Assess and document the origins and uses of data in genAI projects.

6. Curation of Unstructured Data

Curate and auto-label files and objects for use in genAI projects.

7. Extraction of Unstructured Data

Extract data from hundreds of unstructured formats to enhance data utilization.

8. Sanitization of Unstructured Data

Sanitize / redact / mask / sensitive data from files for use in genAI projects.

9. Quality of Unstructured Data

Ensure files are fresh, unique, and relevant before feeding to GenAI models.

10. Security of Unstructured Data

Secure unstructured prompts, retrievals, and responses with distributed, context-aware LLM firewalls.

Data Command Graph: The knowledge graph of enterprise data and its controls

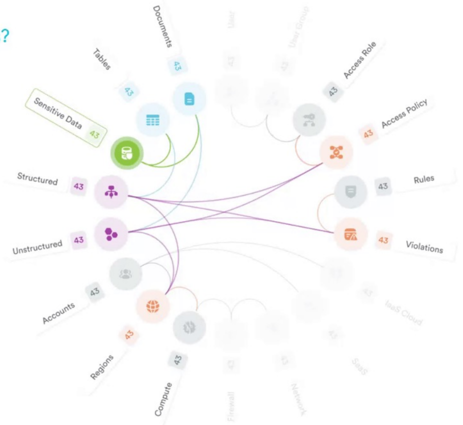
Unstructured Data Intelligence

Data Controls Intelligence

Regulatory Intelligence

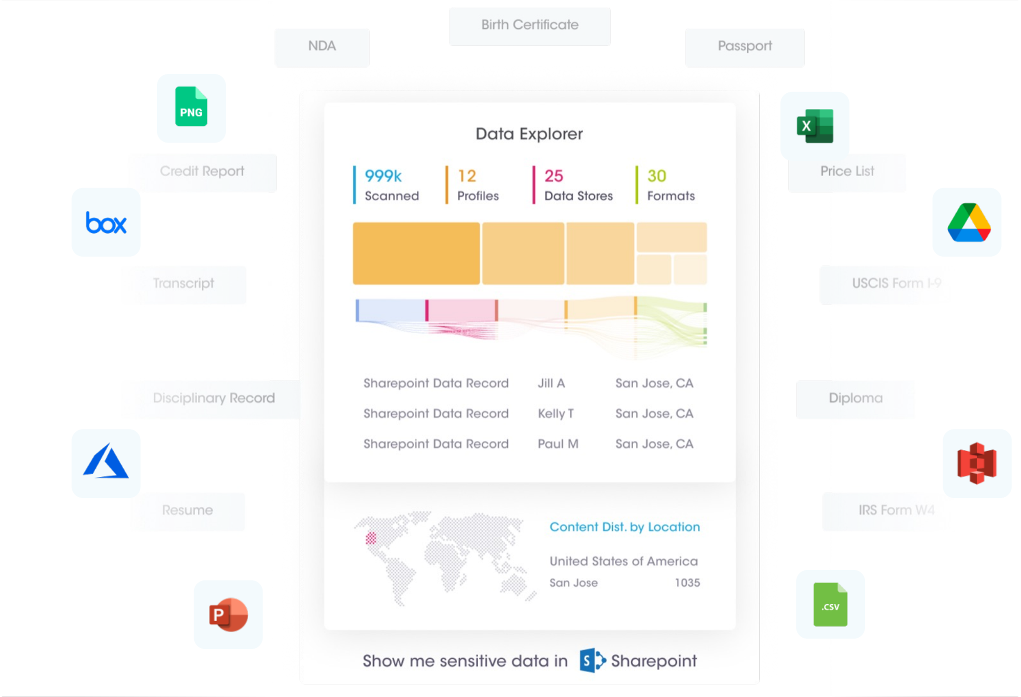
Find answers to your key questions

What sensitive data exists?



Discover files and objects across diverse storages

- Automatically discover files, documents, and objects across various repositories, including file shares, cloud storage, data lakes, and enterprise applications.
- Explore detailed metadata such as vendor info, version info, encryption status, port info, location, owner, size etc.
- By leveraging this metadata, organizations can gain comprehensive visibility into their unstructured data assets.



Data Explorer

999k	12	25	30
Scanned	Profiles	Data Stores	Formats

Sharepoint Data Record	Jill A	San Jose, CA
Sharepoint Data Record	Kelly T	San Jose, CA
Sharepoint Data Record	Paul M	San Jose, CA

Content Dist. by Location

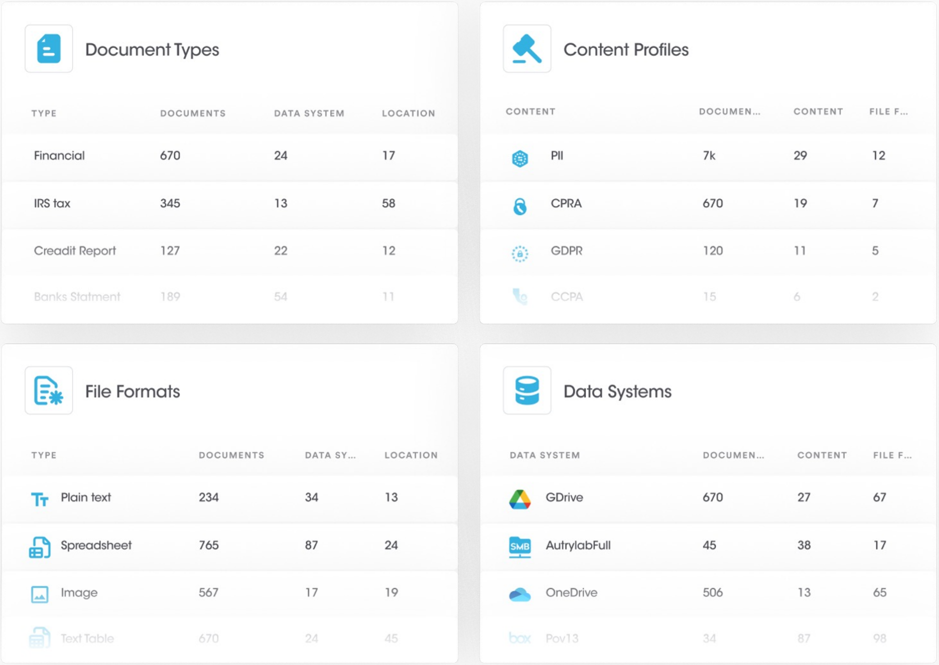
United States of America

San Jose 1035

Show me sensitive data in Sharepoint

Catalog all files and objects for GenAI projects

- Catalog all the files, documents, and objects, enhancing searchability and accessibility across the organization.
- Add tags and metadata to files and objects based on content and context, enabling users to find relevant content quickly.
- Group content by department / function and file formats.



Document Types

TYPE	DOCUMENTS	DATA SYSTEM	LOCATION
Financial	670	24	17
IRS tax	345	13	58
Creadit Report	127	22	12
Banks Statment	189	54	11

Content Profiles

CONTENT	DOCUMEN...	CONTENT	FILE F...
PII	7k	29	12
CPRA	670	19	7
GDPR	120	11	5
CCPA	15	6	2

File Formats

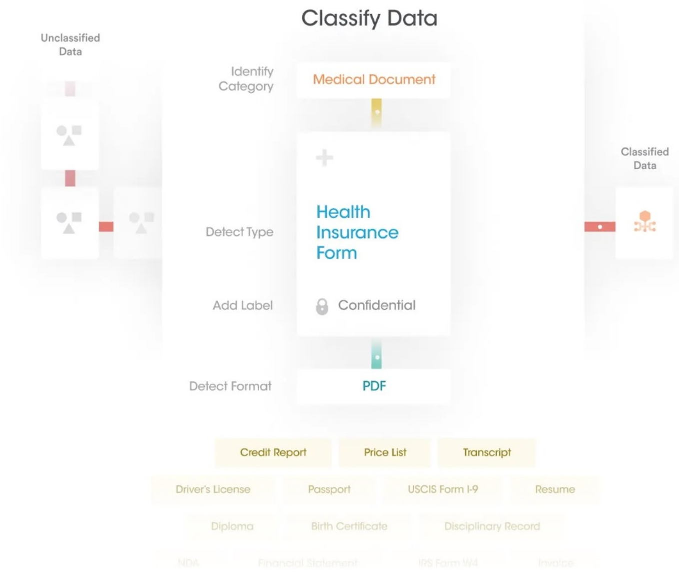
TYPE	DOCUMENTS	DATA SY...	LOCATION
Plain text	234	34	13
Spreadsheet	765	87	24
Image	567	17	19
Text Table	670	24	45

Data Systems

DATA SYSTEM	DOCUMEN...	CONTENT	FILE F...
GDrive	670	27	67
AutrylabFull	45	38	17
OneDrive	506	13	65
Pov13	34	87	98

Automatically classify all files and objects based on their sensitivity and other attributes

- Advanced, out-of-the-box Data Classifiers to auto-classify sensitive data. Advanced NLP and EDM techniques for unstructured data.
- Detect hundreds of sensitive and personal data elements specific to security compliance and global privacy laws.
- Utilize various AI/ML techniques and algorithms that go beyond just pattern and keyword matching.
- Detect sensitive documents such as medical consent forms, insurance forms, tax forms, etc. using out-of-the box ML-based profiles.



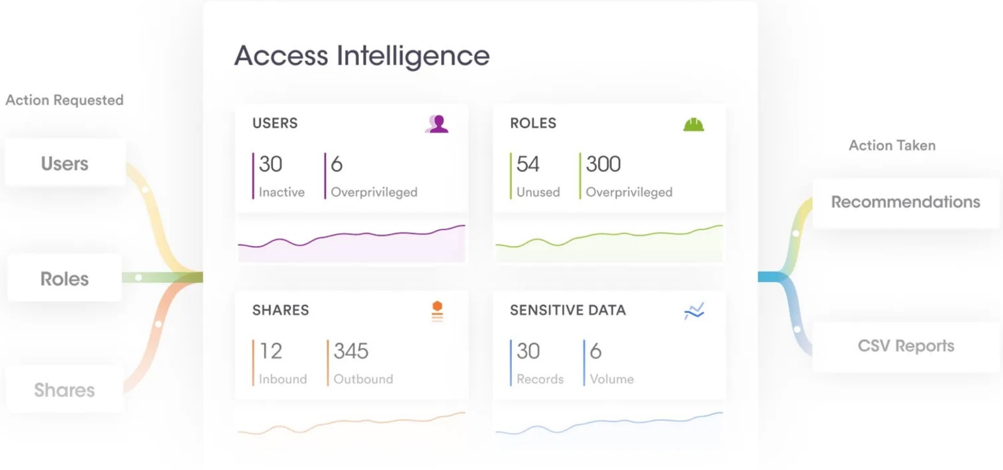
Poll Question

Are you using Copilot, OR building your own custom agents & GenAI applications?

1. Using Copilot exclusively
2. Building custom agents and applications exclusively
3. Using both Copilot and custom solutions
4. Exploring both options, but not actively using them
5. Not using either option currently

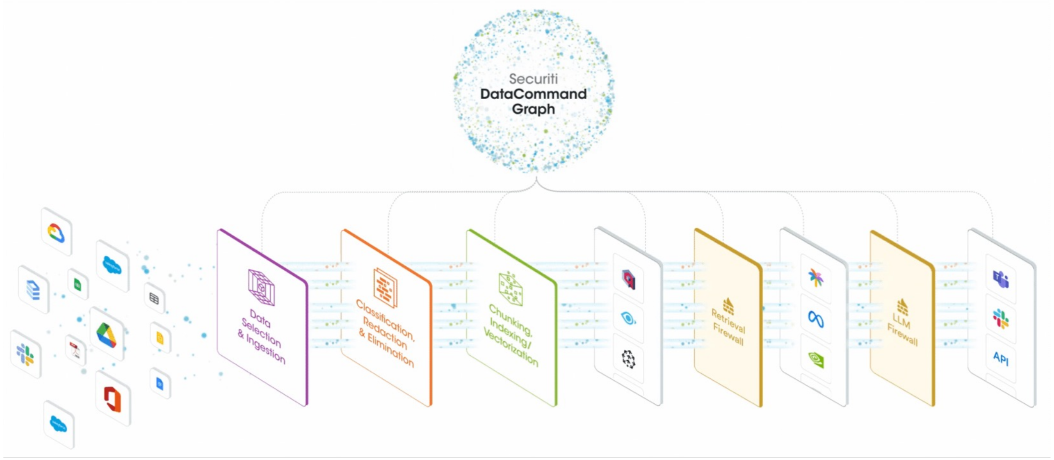
Preserve enterprise entitlements at source systems

- **Know the entitlements** for buckets, folders, files, and documents in data lakes and file shares.
- **Preserve entitlements** when extracting data from source systems for feeding to GenAI models.
- Enforce entitlement information within GenAI pipelines at the prompt level.
- Ensure that only authorized users have access to relevant data by honoring entitlements in AI pipelines and assistants.



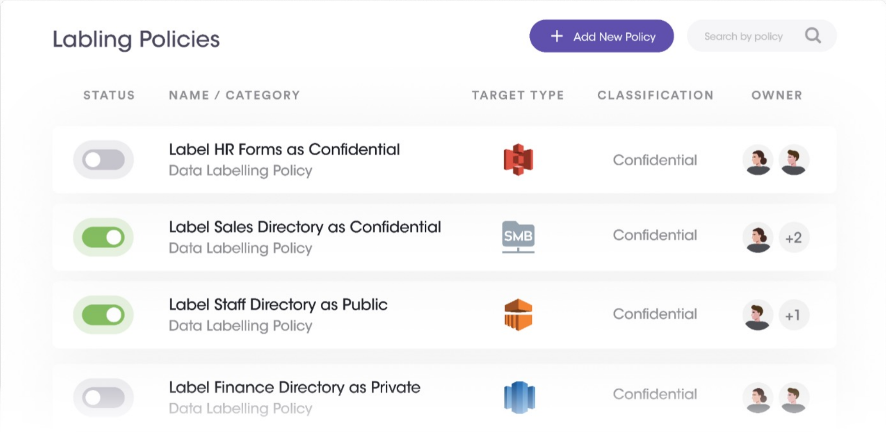
Monitor Unstructured Data Flow and Usage

- Provide **a clear, visual map** of where **unstructured data** originated, and how it was processed and used.
- Verify the source and integrity of each response or output from the AI model or system.
- Ensure GenAI trust by maintaining transparency and compliance.


Securiti
DataCommand Graph
Data Selection & Ingestion
Classification, Redaction & Elimination
Chunking, Indexing/ Vectorization
Retrieval Firewall
LLM Firewall
API

Curate and auto-label files and objects

- Curate data by analyzing content and **automatically adding Data Labels** to files based on content.
- Use an extensible policy framework to automatically apply sensitivity and use case labels within files and documents.
- **Preserve labels and tags** when moving files from source systems for feeding to GenAI models.

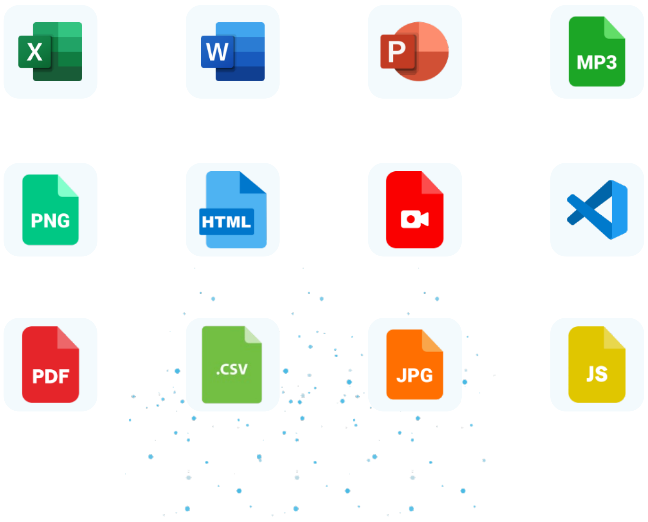


Labling Policies

STATUS	NAME / CATEGORY	TARGET TYPE	CLASSIFICATION	OWNER
○	Label HR Forms as Confidential Data Labelling Policy		Confidential	
●	Label Sales Directory as Confidential Data Labelling Policy	SMB	Confidential	+2
●	Label Staff Directory as Public Data Labelling Policy		Confidential	+1
○	Label Finance Directory as Private Data Labelling Policy		Confidential	

Extract data from diverse formats to enhance utilization

- **Extract Unstructured Data** from hundreds of formats, including
  - Word, PowerPoint, Excel, HTML, PDF
  - Multimedia: Images, Audio, and Video
- **High Fidelity Parsing** that captures a document's visual layout
  - Improves chunking for vectorization
  - Improves LLM comprehension of the data
- **Extract Images from Documents** for OCR - Utilize built-in Optical Character Recognition (OCR) to convert pdf and image documents into text
- **Highly Performant & Scalable** architecture Designed for petabyte-scale scanning

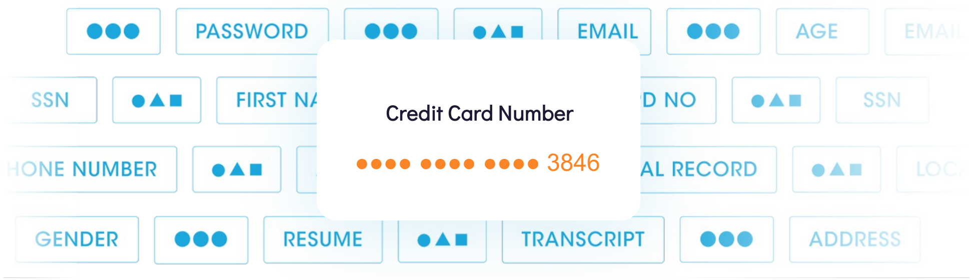


Word, PPT, HTML, PDF, Code

Image, Audio, Video

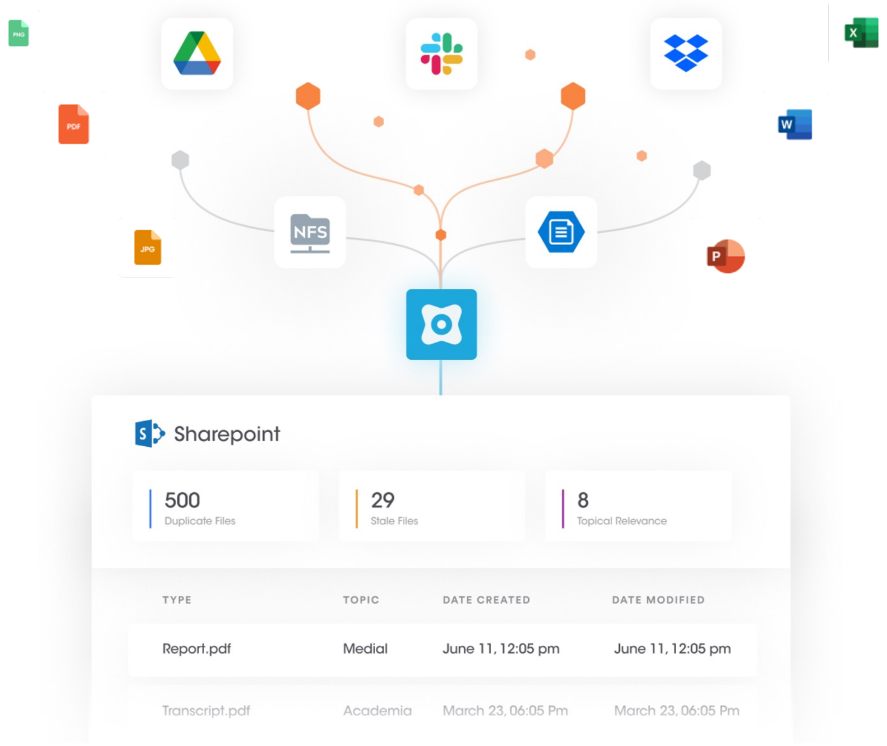
Mask, anonymize, redact, and certify data before use in GenAI pipelines

- Before unstructured data is sent to GenAI models, it must be sanitized. If GenAI models learn from any sensitive information, it remains with them forever, compromising data privacy and security.
- Automatically **mask, anonymize, redact or tokenize** data in-flight in a GenAI pipeline.
- Ensure **compliance** with internal controls and the ever-evolving global data regulations before transferring data for use with LLMs for training, tuning, or inference purposes



Ensure files are fresh, complete, unique, and relevant before feeding to GenAI models

- Infer and analyze metadata on files, such as their recency and topic, to measure data quality
- **Evaluate files inline** to ensure:
  - Freshness
  - Uniqueness
  - Relevance to the topic
  - Reliability of sources
- Develop new data quality measures, such as robustness and non-hallucination of model responses in a non-deterministic world.



Sharepoint

500 Duplicate Files

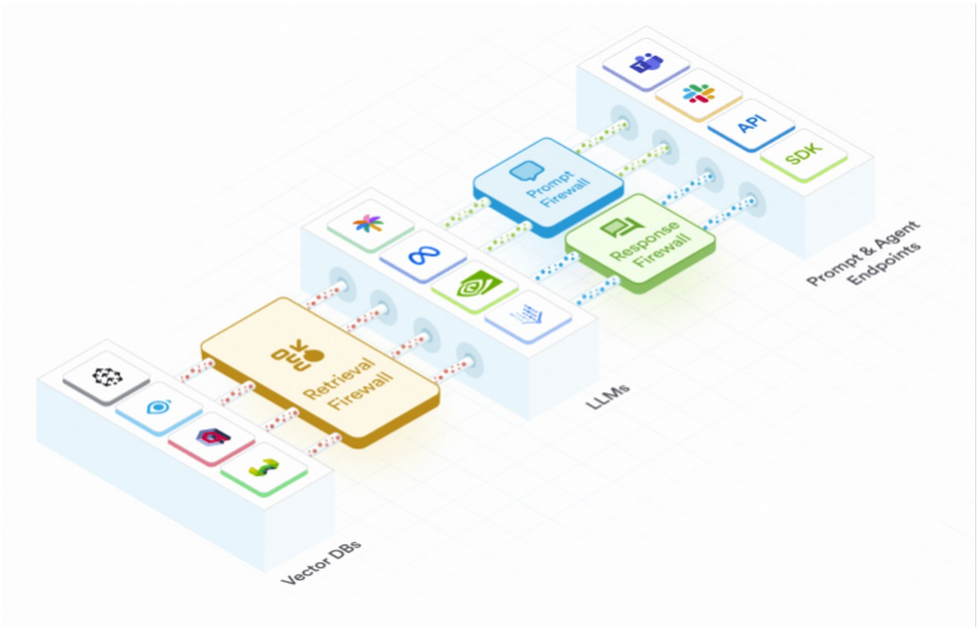
29 Stale Files

8 Topical Relevance

TYPE	TOPIC	DATE CREATED	DATE MODIFIED
Report.pdf	Medial	June 11, 12:05 pm	June 11, 12:05 pm
Transcript.pdf	Academia	March 23, 06:05 Pm	March 23, 06:05 Pm

Secure LLM interactions with inline privacy and security guardrails

- Ensure both models and underlying data systems are properly configured and permissioned to avoid data exposures.
- Provide **continuous visibility into LLM interactions** to protect against attacks, malicious use and mistakes.
- Determine data access at the pipeline level, not just the user level.
- Implement **built-in policies** covering sensitive data, tone, topic, phishing, and attacks.

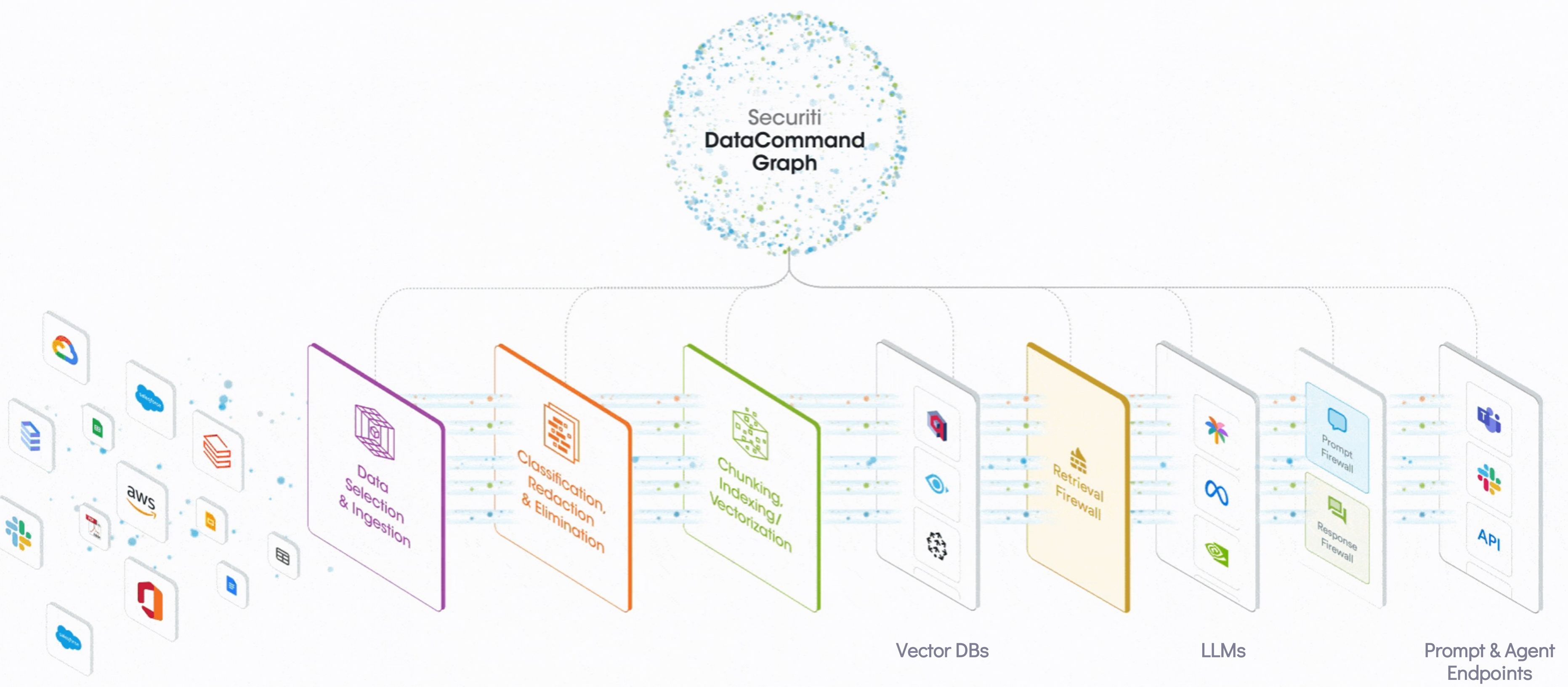


Questions?

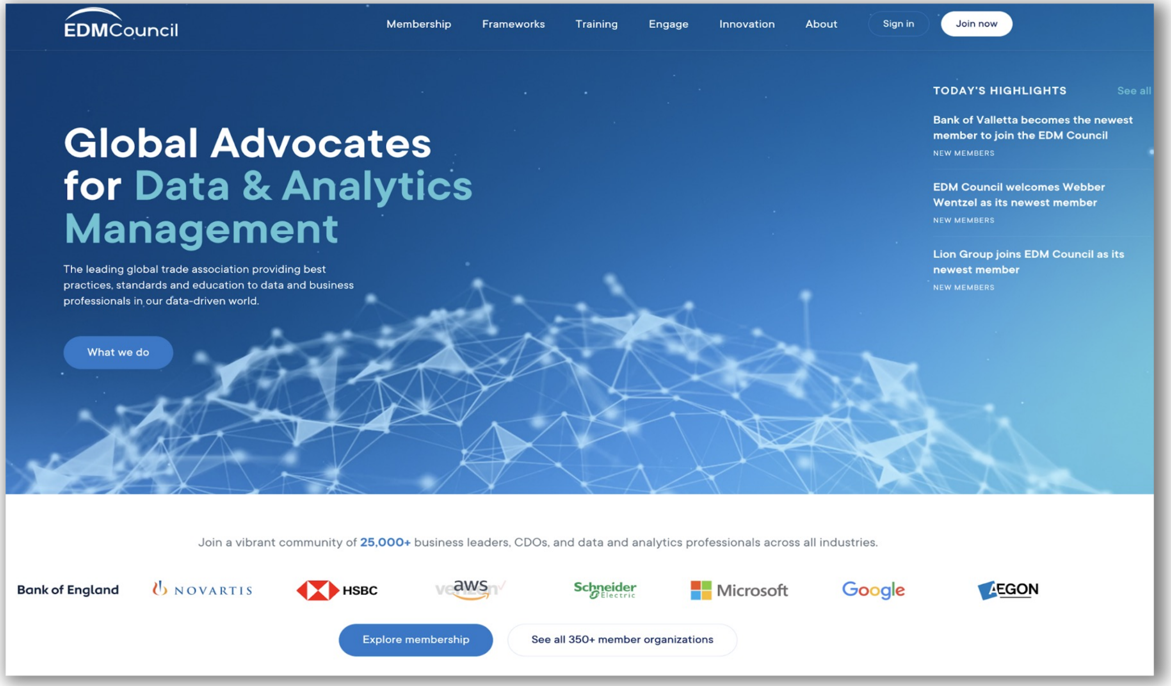
securiti

EDM Webinar

Safe use of ‘unstructured data’ is at the epicenter of GenAI



Join EDM Council and our membership community of companies…



TODAY'S HIGHLIGHTS See all

Bank of Valletta becomes the newest member to join the EDM Council
NEW MEMBERS

EDM Council welcomes Webber Wentzel as its newest member
NEW MEMBERS

Lion Group joins EDM Council as its newest member
NEW MEMBERS

Global Advocates for Data & Analytics Management

The leading global trade association providing best practices, standards and education to data and business professionals in our data-driven world.

What we do

Join a vibrant community of **25,000+** business leaders, CDOs, and data and analytics professionals across all industries.

Explore membership

See all 350+ member organizations

EDMCouncil

350+ Member Firms
Cross-industry, including Regulators

25,000+
Professionals

Worldwide
Americas, Europe, Africa, Asia, Australia

edmcouncil.org

DCAM

CDMC EDMCouncil

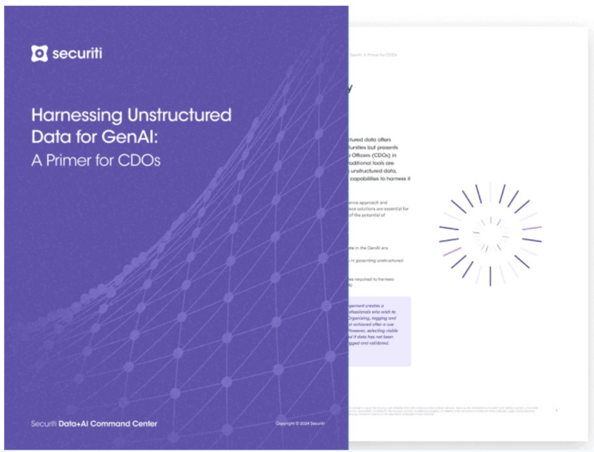
Data ROI EDMCouncil

ESG EDMCouncil

Open Knowledge Graph INNOVATION LAB

elc eLearningCurve

For more information, please check out these resources.

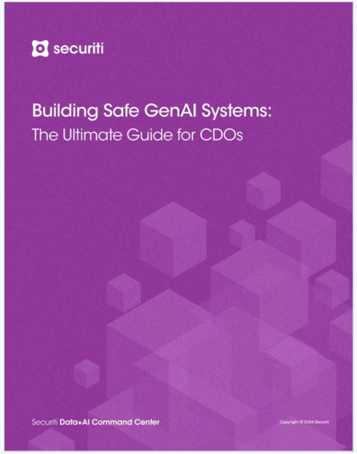


securiti

Harnessing Unstructured Data for GenAI:
A Primer for CDOs

Securiti Data+AI Command Center

Copyright © 2024 Securiti



securiti

Building Safe GenAI Systems:
The Ultimate Guide for CDOs

Securiti Data+AI Command Center

Copyright © 2024 Securiti

Scan Now

Content Bundle

securiti | education

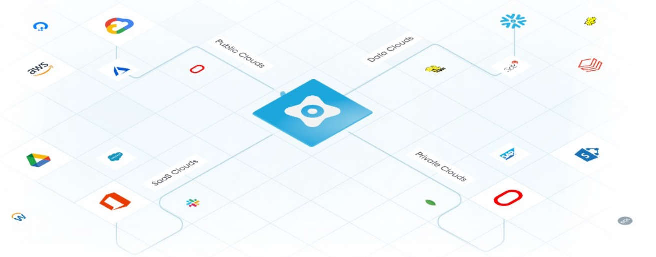
AI Governance Certification

Enroll Now

Your Data Command Center™

Enabling Safe Use of Data & AI
Unified Intelligence & Controls Across Hybrid Multicloud

Security | Privacy | Governance | Compliance



EDM Webinar

Thank you!

FOR MORE INFORMATION, CONTACT:
info@securiti.ai or www.securiti.ai

securiti

EDMCouncil