Data Governance for Generative AI

A conversation with…

Peter Wang,
Staff Product Manager,
Databricks

Jocelyn Houle
Sr. Director, Product Management,
Securiti
Today’s panel

**Moderator**

Eric Bigelsen  
Head of Industry Engagement  
EDM Council

**Jocelyn Houle**  
Sr. Director, Product Management  
Securiti

**Peter Wang**  
Staff Product Manager  
Databricks
One New Variable of GenAI Impacts Almost Everything!

GenAI’s Probabilistic Foundations of Upcoming AI Agents for Knowledge Creation, Automation & Autopilots

- Use of Data & Resources
- Generation of Data & Knowledge
- Interaction with Customers & Service Delivery

It affects many internal Data Controls.
New Level of Visibility and Controls at Every Layer of Data Stack
Data + AI will drive major transformations in human history.

Your Proprietary Data Anywhere

Unstructured Data
- Google Drive
- Google Cloud Storage
- GCP Firebase
- Slack
- Azure Files
- OneDrive
- Amazon S3
- Box
- DropBox
- NFS
- SMB
- Oracle Object Storage

Data Command Graph
Single Source of Truth

Structured Data
- GCP BigQuery
- Snowflake
- MongoDB
- IBM DB2 Cloud
- MariaDB
- Salesforce
- Azure Cosmos DB
- Apache Hive
- Amazon DynamoDB
- Workday
- Databricks
- MySQL

GenAI in Databricks

AI Models + Agents

GenAI Apps & Innovation
- Copilots & Assistants
- Business Process Automations
- New Category of Apps
- Insights & Analytics

Enabling Safe Use of Data and AI
90% of organizational data is now unstructured, and must be safely governed when used by GenAI models.
# Top 10 Considerations for Enabling Safe Use of Unstructured Data with GenAI

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Discover unstructured data assets across diverse storages.</td>
<td>Catalog all files and objects that can be used for genAI projects.</td>
<td>Classify all files and documents based on their sensitivity and other attributes.</td>
<td>Preserve enterprise entitlements at source systems for genAI projects.</td>
<td>Assess and document the origins and uses of data in genAI projects.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Curate and auto-label files and objects for use in genAI projects.</td>
<td>Extract data from hundreds of unstructured formats to enhance data utilization.</td>
<td>Sanitize / redact / mask / sensitive data from files for use in genAI projects.</td>
<td>Ensure files are fresh, unique, and relevant before feeding to GenAI models.</td>
<td>Secure unstructured prompts, retrievals, and responses with distributed, context-aware LLM firewalls</td>
</tr>
</tbody>
</table>
Poll

Is unstructured data like documents, videos and sound files part of your data estate?

- Yes - it is a big part
- No - we are pretty standard on structured
- I don’t know
Build AI Apps with Model Serving

Low TCO

- Low TCO

Accelerated deployments

- Automatically capture requests and responses to monitor and debug models
- MLflow integration to seamlessly evaluate & deploy AI models

Fewer Mistakes

- Automated vector and feature lookups simplifies integration and reduces mistakes

Feature & Vector Store

Retraining or Fine-Tuning

Evaluation & Deployment

Model Serving

Unity Catalog
Govern all data and ML assets

Cloud Data Lake
All structured and unstructured data

⭐
- Automated vector and feature lookups simplifies integration and reduces mistakes

⭐
- Automatically capture requests and responses to monitor and debug models

⭐
- MLflow integration to seamlessly evaluate & deploy AI models

Microsoft Azure, Amazon Web Services, Google Cloud
Data Command Graph provides deep contextual insights for Unstructured + Structured Data. A baseline need for GenAI pipelines & use cases.
Data Command Graph provides deep contextual insights for Unstructured + Structured Data. A baseline need for GenAI pipelines & use cases

- **Data Command Graph** is a Knowledge Graph that automatically captures all key metadata and relationships between them:
  - Data Systems
  - Files/ Folders
  - Tables/ Columns
  - Data Sensitivity
  - Access & Entitlements
  - Internal Policies & Controls
  - Applicable Regulations
  - Related AI Models / Pipelines

- Advanced out-of-the-box Data Classifiers to auto-classify sensitive data. Advanced NLP and EDM techniques for unstructured and structured data.

- Answer any question around Data+AI and their relationships with a multidimensional graph model. This enables the safe use of data to go from POCs to Production.
Enabling Safe Use of Data and AI

Unstructured Data takes the center stage with GenAI!

Contextual Insights for Unstructured Data from all key perspectives, a baseline need for Safe GenAI Pipelines & Use Cases

- Files of all types (docs, audio, video, images etc.). CLOBs.
- File categories (legal, finance, HR etc.) based on content
- Access and user entitlements
- Sensitive objects within a file
- Regulations applicable on file content
- File quality (freshness, relevance etc.)
- Lineage of files & embeddings used in GenAI pipes

Curate data by analyzing content and automatically adding Data Labels to files based on content

Empower the use of Unstructured Data for LLM Training/Tuning, RAG and all related GenAI Apps
Comprehensive Data Intelligence for Structured Data

Contextual Insights for Structured Data from all key perspectives, empowering AI Agents & AI Use Cases

- Tables, Columns, Schemas
- Business Context & Glossary
- Access and user entitlements
- Sensitive Data Columns & Elements
- Regulations applicable to Data
- Quality
- Lineage

Structured Data, combined with Unstructured Data enables variety of GenAI use cases, AI Agents and GenAI Apps.
Poll

Do you need granular data access down to the row, column or element level?

- Yes
- No
- I don’t know
Enabling Safe Use of Data and AI

Data Access & Entitlement Intelligence empowers safe GenAI pipelines

- Identify users and roles with access to
  - **Unstructured data** - buckets, folders, and files or documents in data lakes and file shares
  - **Structured data** - databases, tables, and sensitive data in data warehouses (like Databricks)

- Map the relationships between users, roles, and the permissions they have for accessing documents or datasets.

- Enables Dynamic Masking of Structured Data for Certain Roles and Users, for safe use with AI.

- Entitlement information can be used within GenAI pipelines to enforce same entitlements at the prompt level.

Honor data entitlements to prevent data leakage and ensure safe data utilization
Regulatory context of data enables its compliant use

Gain insights into regulatory risks based on applicable Data+AI laws

- Categorize data based on sensitivity levels (as per GDPR, HIPAA, and other data protection regulations) and determine what data should be anonymized or redacted before feeding into AI models.

- Identify and mitigate risks related to data privacy, security, and other regulatory concerns.

- Enable the compliance of GenAI projects with applicable AI Regulations (EU AI Act, NIST AI RMF) and Data Regulations (GDPR, CPRA, GPD etc).
Safe Enterprise AI with Databricks & Securiti AI

Building Blocks of a GenAI Pipeline

Securiti Data Command Graph

Select and ingest unstructured & structured data from hundreds of sources
Classify and redact sensitive data on the fly
Create and ingest embeddings in the vector DB of your choice
Pick the foundational AI Model of your choice
Safeguard the prompts and interactions with LLMs

Safe Enterprise AI with Databricks & Securiti AI

Building Blocks of a GenAI Pipeline

Securiti Data Command Graph

Select and ingest unstructured & structured data from hundreds of sources
Classify and redact sensitive data on the fly
Create and ingest embeddings in the vector DB of your choice
Pick the foundational AI Model of your choice
Safeguard the prompts and interactions with LLMs

Safe Enterprise AI with Databricks & Securiti AI

Building Blocks of a GenAI Pipeline

Securiti Data Command Graph

Select and ingest unstructured & structured data from hundreds of sources
Classify and redact sensitive data on the fly
Create and ingest embeddings in the vector DB of your choice
Pick the foundational AI Model of your choice
Safeguard the prompts and interactions with LLMs
Questions?
For more information, please check out these resources.
Thank you!

FOR MORE INFORMATION,
CONTACT:
info@securiti.ai

OR VISIT ONLINE:
www.securiti.ai