



Cleanse AI Data to Mitigate Leakage and Strengthen Security Posture

AI is only as secure and trustworthy as the data it's built on. Yet most organizations still rely on raw, unfiltered datasets that contain sensitive, personal, or regulated information — increasing the risk of data leakage, prompt injection, or unintended model exposure.



Cleanse AI Data to Mitigate Leakage and Strengthen Security Posture

Cleanse AI Data to Mitigate Leakage and Strengthen Security Posture

BigID's Data Cleansing for AI helps organizations proactively reduce this risk by automatically redacting or tokenizing sensitive data before it enters generative AI pipelines or large language models (LLMs). Security teams can safeguard against downstream AI misuse while preserving utility for model training — enabling secure innovation without delay.

- ◆ Automatically remove or tokenize sensitive and high-risk data before AI use
- ◆ Reduce exposure to leakage, injection, and unauthorized model access
- ◆ Preserve data structure and context for downstream model performance
- ◆ Cleanse both structured and unstructured data across SaaS, cloud, and files
- ◆ Strengthen data security posture across your entire AI pipeline

Key Outcomes

Sensitive Data Redaction & Tokenization

Mitigate regulated, personal, or sensitive data from entering LLMs and GenAI tools to avoid compliance violations and security gaps.

Strengthen AI Security Posture

Lower the risk of prompt injection, model poisoning, and data misuse with automated upstream cleansing.

Accelerate Trusted AI Adoption

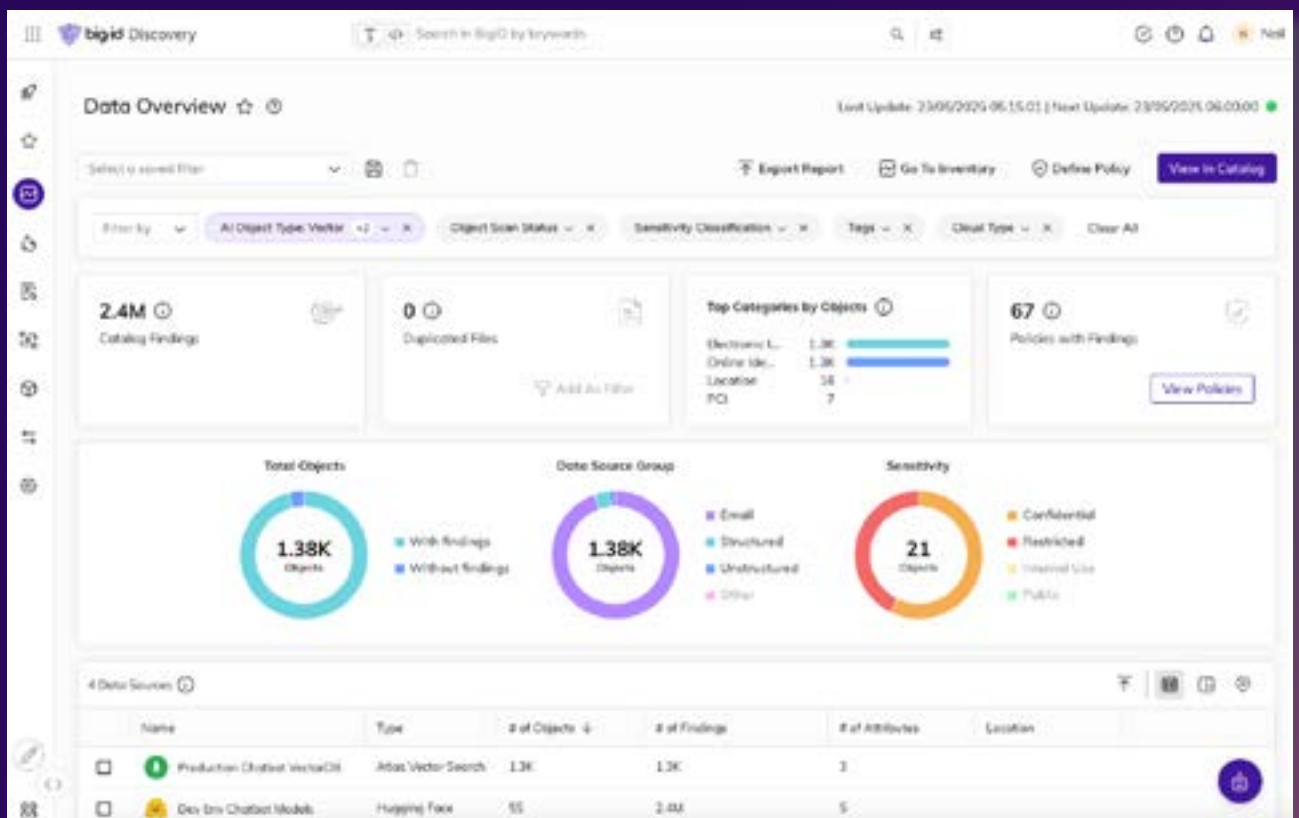
Enable AI and security teams to move faster with confidence — backed by clean, policy-aligned datasets.

Preserve Data Utility for AI

Retain structure, format, and context so cleansed data remains useful for model training and inference.

Operationalize Responsible AI Governance

Enforce data controls that align with internal policies and external obligations across your AI workflows.



Key Capabilities

Sensitive Data Redaction & Tokenization

Automatically identify and cleanse personal, regulated, or sensitive data across both structured and unstructured sources. Choose between redaction or tokenization to neutralize high-risk content while maintaining compliance — ensuring sensitive data never reaches your AI systems.

Unstructured Data Support

Go beyond databases: scan and cleanse emails, PDFs, documents, and cloud file shares where high-risk content often hides. BigID delivers enterprise-wide protection across all data formats, fueling AI models.

Policy-Based Cleansing Controls

Apply cleansing rules based on identity-aware classification, data type, usage context, or AI-specific risk. Policies are enforced automatically to drive consistency and compliance without manual effort.

Context-Preserving Output

Maintain original data formatting and semantics after cleansing to ensure models can still interpret and learn from the data. Ideal for GenAI pipelines that rely on contextual accuracy.

Platform-Native Integration

Built as part of BigID's Secure Data Pipeline, Data Cleansing for AI integrates with GenAI Catalog, Search, and Data Labeling for AI. Together, these capabilities enable end-to-end visibility, classification, control, and remediation across your AI data lifecycle.

About BigID

BigID helps organizations connect the dots across data & AI: for security, privacy, compliance, and AI data management. BigID enables customers to find, understand, manage, protect, and take action on high risk & high value data, wherever it lives.

Customers use BigID to reduce their AI & data risk, automate security and privacy controls, achieve compliance, and understand their data throughout their entire data landscape: from the cloud, on-prem, and everywhere in between.

Connect the Dots Across Data & AI

Security • Compliance • Privacy • AI Data Management

Reduce risk, accelerate time to insight, and get data visibility and control across all your data - everywhere.



Tools like BigID are the future.

Organizations should be leveraging these tools to remove the manual processes from data discovery, provide better visibility, and help with prioritization of controls.



Ryan O'Leary

Future of Trust: Battling Data Discovery Confusion
