sentra

A Security Practitioners Guide to AI Data Readiness

How to safely adopt Microsoft 365 Copilot and other AI agents

For most enterprises, the first real exposure to generative AI is **Microsoft 365 Copilot** and adjacent copilots tightly wired into productivity suites, SaaS apps, and internal services, not standalone LLMs in a lab. AI agents and copilots inherit years of permission sprawl and oversharing. Suddenly, long-forgotten mailboxes, sites, and shared drives become searchable and synthesizable by tools that were never part of your original access design.

At the same time, most organizations struggle with shadow data, fragmented visibility, and inconsistent labeling across cloud, SaaS, and legacy environments. DLP, IAM, CSPM, and logs each see part of the picture but lack a unified, AI-aware view, especially one that centers on how Copilot and similar assistants actually use data in context windows and workflows.

When AI is deployed on top of that reality, three questions become critical:

- **Which AI systems (copilots, agents, LLM apps) actually exist?**
- **What data can each of them touch, across cloud, SaaS, and on-prem?**
- **How will you prove to regulators and the board that this access is governed before and after you turn on Copilot at scale?**

Legacy controls can't answer these on their own. DLP reasons about files and channels, not context windows; IAM models people and static roles, not agents acting across tenants; logs capture events, not the sensitivity of data flowing through AI prompts and outputs.

An **AI Data Readiness** program starts from a different foundation: an AI-powered data intelligence layer that knows what the data is, where it lives, who or what can access it, and how AI, especially Copilot, actually uses it, then feeds that context into the tools you already own.

AI Data Readiness Challenges

AI copilots are outpacing security; organizations are unaware of hidden risks

- **AI RISK: AI agents (ex. Copilot) inherit broken permissions and oversharing:** years of accumulated data becomes instantly discoverable, dramatically increasing blast radius
- **AI RISK: Unknown sensitive data in M365 content** and inconsistent or missing sensitivity labels generate compliance issues
- **AI RISK: No clear view of what data AI agents will touch** sites, mailboxes, and drives contain regulated data that Copilot can access

The hard truths about AI data readiness

A lot of "AI-ready" offerings amount to a couple of new rules and dashboards on top of legacy architectures. They don't build an AI-aware inventory of copilots and agents, can't reliably show you which datasets those agents can reach, and don't map those datasets to regulated or business-critical content. If the entire AI story is "we integrate with your DLP/E5," you're being handed back the hard work of accurately classifying all data and identifying gaps yourself.

Some "AI visibility" platforms only work if you ship sensitive prompts, logs, or embeddings into their cloud. You gain a dashboard, but you also create a new perimeter, risk breaching residency and compliance requirements, and still miss blind spots for data and AI tools that never get copied out.

An AI-ready approach keeps sensitive data in your environment, uses in-place scanning, and builds its data-context graph from a comprehensive insight. You also can't simply "turn on" AI data readiness in DLP and IAM: these controls weren't designed for training sets and RAG indexes that blend structured and unstructured data, for agents acting autonomously across multiple tenants, or for output behavior where regurgitation of memorized sensitive data is a real risk.

Real readiness starts with AI-aware data understanding and identity modeling, then uses DLP, SSE/CASB, IAM, and AI gateways as enforcement control points informed by that context.

What security teams are actually struggling with

Across financial services, tech/SaaS, retail, healthcare, and media, security leaders describe similar problems as they prepare for Copilot and broader AI adoption.

Hidden AI and unknown exposure

Teams can't answer basic inventory questions: which AI apps and agents exist; which are genuinely AI-specific; which datasets underpin them; and which M365 sites, drives, and mailboxes holding regulated data will suddenly become available to Copilot or other internal agents. The risk isn't only "outside-in" attack. also inside-out and inside-inside exposure as assistants surface data across teams, tenants, and domains that were never meant to be connected.

Shadow AI and governance drift

Business units spin up AI tools faster than security can review them. Security has to decide which tools are acceptable, spot AI-related services hidden in generic SaaS usage, and segment access (for example, by client or business line) so AI doesn't mix data domains that must remain separate.

Training and RAG pipelines with unknown content

Many organizations only discover after the fact that fine-tuning or RAG datasets contain financial data, PII, or confidential records. That creates regulatory exposure and the risk that models will memorize and leak sensitive content back to users.

Unstructured data and broken labels

Few teams have fully tagged their unstructured stores. Many rely on regex and generic LLMs that miss business context, generate high false positives, and require manual review. As a result, labels in M365 and other SaaS platforms can't be trusted to mean what policies think they mean, especially once Copilot and other assistants start consuming that data.

4 steps to AI data readiness

1. Build an AI-aware inventory

Start by making AI visible. Automatically discover and maintain a live inventory of copilots, assistants, and agents, including Microsoft 365 Copilot, custom LLM apps, bots, and third-party services. Then map each one to the data sources it can reach. Identify which of those sources contain regulated or business-critical data.

For Copilot specifically, that means knowing which SharePoint sites, Teams channels, mailboxes, OneDrive folders, and third-party connectors contain sensitive data before you enable the experience for entire business units. This turns AI from an opaque risk into a concrete set of systems and data paths the security team can reason about and constrain.

2. Classify data and AI flows with context

You need accurate, consistent classification across traditional data stores and AI workflows. That means understanding sensitivity and business impact for:

- Cloud and SaaS repositories
- Legacy file shares and on-prem stores
- AI-specific flows like training sets, feature stores, RAG indexes, prompt logs, and outputs

AI-powered classifiers tuned for your business context can dramatically reduce false positives and the manual effort usually tied to unstructured data. For Copilot, this is the difference between “turn it on and hope M365 labels are right” versus “turn it on with confidence that the underlying labels and guardrails actually reflect business and regulatory requirements.”

3. Treat AI as a first-class identity

Model AI agents like any other high-risk identity. Treat copilots and agents as first-class identities, tie their actions back to underlying users and service accounts, and evaluate effective permissions against least-privilege policies for labeled data.

Otherwise it’s easy to end up with an internal assistant that can, for example, pull payroll records and customer data on request because it inherited broader access than any single human was ever meant to have. In the Copilot world, that might look like an executive assistant suddenly being able to summarize salary data, customer complaints, and board materials in a single prompt, even though those sources were never intentionally grouped.

4. Feed data intelligence into enforcement and proof

The same data-context layer should drive both controls and evidence. Context about what the data is, where it lives, who or what can access it, and how it's used should flow into DLP, SSE/CASB, IAM, SIEM/SOAR, and AI gateways, so those systems enforce policies based on real sensitivity and exposure, not just pattern matching. Integrations are an important enabler.

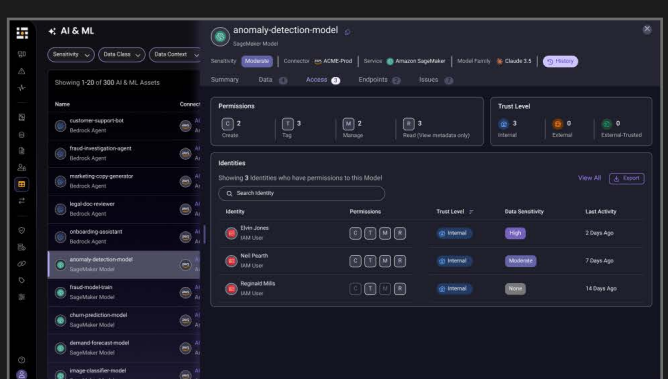
At the same time, the platform should leave you with a clearer map of sensitive data and AI access paths, remediated shadow or obsolete data, and audit-ready evidence you can show regulators and internal stakeholders, particularly as they scrutinize Copilot and other productivity AI deployments.

Drive AI Data Readiness

Secure AI and data governance

Eliminate risk from AI agent and ML model adoption

- **Continuously discover and classify sensitive data exposures** in copilot knowledge bases
- **Inventory AI assistants and agents** and map AI-accessible datasets to minimize exposures
- **Auto-apply labels to sensitive** documents, emails, records holding regulated data
- **Govern copilot access to sensitive data** and institute least pivilege



Sentra ties together **identities, groups, AI service roles, and datasets** into one AI-specific risk view.

"Sentra gave us **AI visibility, eliminated audit surprises,** and **cut cloud storage costs by ~20%** through shadow-data cleanup."

VP, Security Architecture; SoFi

SoFi

AI data readiness checklist

Use this checklist to validate whether you're truly ready for Copilot and broader AI adoption:

Visibility

We have an up-to-date inventory of AI copilots, agents, and apps, mapped to the data sources they can access, including where those sources contain regulated data.

Data understanding

We can answer what our sensitive data is, where it lives (including unstructured and legacy stores), and which AI systems touch it, with consistent labeling across environments.

Identity and access

AI agents are treated as first-class identities, with least-privilege policies applied to labeled data and clear attribution of actions back to users and service accounts.

Controls and proof

DLP, SSE/CASB, IAM, SIEM/SOAR, and AI gateways are all driven by a single data-context layer, and we can produce audit-ready evidence of where regulated data lives and who (including AI assistants like Copilot) can access it—on demand, not just at audit time.

Outcomes

We can show measurable reduction in shadow and obsolete data, along with associated risk and cost (for example, cloud storage savings similar to SoFi's ~20% reduction).

Treating Copilot and other AI systems as core parts of your identity and data fabric—not bolt-on features—lets security practitioners move from AI anxiety to confident, governed adoption grounded in AI Data Readiness.

Gartner Peer Insights Customers' Choice 2025

Gartner Peer Insights™

4.9 ★★★★★

By Sentra in Data Security Posture Management (DSPM)

Setting a New Standard in Data Security

>95% Accuracy

AI-powered classification

10x more efficient

In scanning compared to industry

In less than 1 week

Discover and assess data risks @ PB — scale

sentra

Visit www.sentra.io | Watch a demo

For more information, please contact us at info@sentra.io