Solution Brief

THALES

CYBERSECURITY

Thales Retrieval Augmented Generative (RAG) AI Data Protection Solutions

cpl.thalesgroup.com

THALES

CYBERSECURITY

The Rise of AI and the Emergence of RAG Applications

Artificial Intelligence (AI) has rapidly evolved from an academic pursuit to one of the most transformative technologies shaping the global economy. In recent years, the advent of large language models (LLMs) such as GPT, LLaMA, and PaLM, has demonstrated the remarkable ability of AI to generate human-like text, analyze complex information, and support decision-making across industries. These models, trained on vast corpora of public data, have enabled breakthroughs in productivity, customer engagement, and knowledge management.

However, despite their potential, LLMs face critical limitations. Because their training data is static and often lacks proprietary or domain-specific knowledge, their outputs can be inaccurate, outdated, or incomplete. For enterprises, this creates a significant barrier: organizations need AI systems that not only demonstrate language fluency but also provide reliable, context-aware, and current insights rooted in their own data.

This challenge has fueled the rise of **Retrieval-Augmented Generation** (RAG), a new class of AI applications designed to extend the power of LLMs by integrating them with external, trusted data sources. RAG introduces a dynamic retrieval layer that accesses enterprise knowledge bases, document repositories, real-time feeds, or regulatory archives. By augmenting a model's reasoning with authoritative and up-to-date information, RAG bridges the gap between general-purpose AI and enterprise-grade intelligence.

The growing interest in RAG reflects a broader shift in the AI landscape. Enterprises increasingly recognize that success with AI is not measured by raw generative capabilities alone, but by the accuracy, transparency, and trustworthiness of results. In industries such as finance, healthcare, government, and critical infrastructure, the risks of "hallucinated" answers or mishandled sensitive data can outweigh the benefits of automation. RAG directly addresses this by grounding AI outputs in evidence, reducing hallucinations, and ensuring that responses are explainable and verifiable.

Furthermore, the scalability of RAG applications allows enterprises to modernize how employees, partners, and customers interact with information. Complex knowledge retrieval that once required manual research or specialized expertise can now be delivered instantly through natural language queries. From accelerating research and development to streamlining compliance reporting and improving customer service, RAG represents a practical pathway to enterprise-wide AI adoption.

As organizations accelerate their digital transformation journeys, the adoption of RAG-based systems is expected to grow rapidly. Analysts forecast that AI systems capable of integrating real-time enterprise knowledge will become a cornerstone of competitive advantage, transforming not only how data is accessed but also how strategic decisions are made. In this context, data protection, compliance, and governance become essential. The same mechanisms that make RAG powerful - its ability to ingest, store, retrieve, and generate from sensitive enterprise information, also introduce new risks. Without strong safeguards, enterprises may expose proprietary data, intellectual property, or personal information to unauthorized access. This is why securing RAG workflows end-to-end is not optional but foundational to responsible enterprise AI.

Understanding the RAG Workflow

RAG consists of two core components: a retrieval mechanism and a generation model. Unlike traditional LLMs, which rely solely on training data, RAG retrieves relevant information from trusted, enterprise-specific knowledge bases, such as internal documents, databases, or live news feeds, In response to a query. By combining the retrieved information with the original query, RAG produces precise and contextually relevant responses.

Vector databases are the foundation of RAG systems, enabling efficient semantic search for retrieval. The typical RAG workflow includes the following stages:

- **Ingestion**: Enterprises input documents (e.g., PDFs, articles, internal manuals) into the RAG system. The text is segmented into smaller, manageable units called chunks.
- **Embedding:** An embedding model transforms each chunk into a vector, a numerical representation that captures semantic meaning. Chunks with similar meanings have vectors positioned close to one another in a high-dimensional space.
- **Storage:** The vector database stores thousands or millions of these vectors alongside their associated text chunks.
- **Retrieval:** When a user submits a query, it is converted into a vector using the same embedding model. The vector database then performs a similarity search, typically using algorithms like k-Nearest Neighbors (KNN), to identify the top k most relevant vectors.
- **Generation:** The system retrieves the text chunks linked to these top vectors and provides them, along with the original query, to the LLM, which generates a precise, contextually relevant response.

Thales RAG Data Protection Solutions

Thales provides a comprehensive suite of RAG data protection solutions that safeguard sensitive data throughout its lifecycle within an enterprise AI application's RAG system, from ingestion and storage to retrieval. Enterprises can select the appropriate solution based on their technology stack, deployment environment, data classification needs, and security requirements.

Use Case 1: Pre-Ingestion Data Discovery and Protection

The CipherTrust Data Discovery and Classification (DDC) solution enables enterprises to identify and classify sensitive data before ingestion. Additionally, cloud-based tools such as Google Data Loss Prevention (DLP), Azure Text Analytics, or AWS Comprehend can detect sensitive information within individual documents. Once identified sensitive data can be secured using the Thales CipherTrust platform through tokenization, encryption, or masking, in line with policies defined by the enterprise's information security team.

Use Case 2: Data Protection in the Vector Database

For cases where sensitive data cannot be protected during ingestion or embedding, Thales offers solutions to secure data stored in the vector database. Two approaches are available: Key protections include:

- **CipherTrust Transparent Encryption (CTE)**
  CTE encrypts the entire storage used by the vector database ensuring that only authorized processes can access the encrypted data. It operates transparently, requiring no direct integration with specific database providers. This approach is best suited for enterprises that deploy their own vector databases rather than consuming them as a Software-as-a-Service (SaaS).
- **CipherTrust Cloud Key Management (CCKM)**
  When enterprises use a vector database as a SaaS offering, CTE may not be feasible. In such cases, CCKM enables independent management of encryption keys, separate from the SaaS provider. While this requires integration with each database provider, most services support cloud-based key management solutions (e.g., AWS Key Management Service External Key Store) that are compatible with CCKM.

Use Case 3: End-to-End Data Activity Monitoring for Compliance and Threat Detection

Thales Imperva Data Activity Monitoring (DAM), part of Data Security Fabric, delivers continuous, real-time monitoring of all interactions with databases and unstructured data across the RAG lifecycle. This includes ingestion, storage, retrieval, and generation-related access. This capability is especially critical for regulated industries such as finance or healthcare, where RAG systems process sensitive data like customer records or intellectual property.

DAM provides:

- Behavioral baselining: Using machine learning to profile normal activity patterns, such as ingestion rates, retrieval query frequency, and access behaviors, for users, applications, and operations.
- Audit and compliance reporting: Detailed logs capture who accessed which databases or files, when, and from where. These records support forensic analysis and compliance with regulations such as GDPR, SOX, or HIPAA, with customizable reports and dashboards.

Key protections

- Real-time anomaly detection to identify threats such as insider misuse (e.g., excessive retrieval of sensitive data) or external probes (e.g., high-volume queries suggesting exfiltration attempts), with automated alerts and blocking.
- Integration with CipherTrust for context-aware monitoring, for example, validating decryption events during authorized retrievals or flagging unauthorized attempts to access encrypted storage.
- Vulnerability assessments and rights management to detect misconfigurations, enforce least-privilege access, and minimize risks from over-privileged AI service accounts.

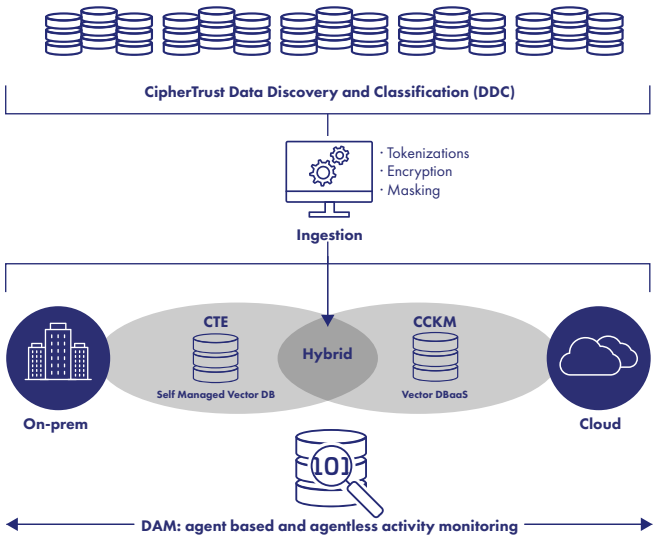
DAM supports both agent-based and agentless deployments across cloud, on-premises, and hybrid environments. It delivers comprehensive visibility into data flows throughout the RAG pipeline while minimizing performance impact on high-throughput workloads.

Securing the Future of Enterprise RAG

As enterprises embrace Retrieval-Augmented Generation to unlock new levels of intelligence, efficiency, and competitiveness, they must also confront the risks associated with processing sensitive data across complex AI workflows. RAG represents the future of enterprise AI, bridging the gap between general-purpose language models and mission-critical business needs, but its success depends on trust.

Thales enables enterprises to build secure, compliant, and resilient RAG systems by providing solutions that protect sensitive data before ingestion, secure storage in vector databases, and monitor activity throughout the entire lifecycle. By embedding security at every stage, enterprises can confidently deploy RAG applications that meet the highest standards of accuracy, compliance, and governance.

The rise of RAG is not just a technological milestone—it is a business transformation. Organizations that implement secure RAG today will be positioned to lead in tomorrow's AI-driven economy, where knowledge is instant, intelligence is contextual, and trust is paramount.



cpl.thalesgroup.com

Contact us – For contact information, please visit cpl.thalesgroup.com/contact-us