System and Method to Securely Exchange Data Across Organizational Boundaries for AI Language Query Processing

Abstract

This paper presents a comprehensive system and method for securely exchanging information across organizational boundaries within an AI-enabled Retrieval-Augmented Generation (RAG) pipeline. As enterprises increasingly rely on external Large Language Models (LLMs) to process identity governance and access management queries, protecting sensitive information contained in organizational knowledge graphs becomes a critical requirement. The proposed method introduces a cryptographically enforced data-obfuscation framework that transforms sensitive identity attributes into one-way hashed tokens before any data crosses the organizational trust perimeter. All graph queries are executed strictly within the organization's secure infrastructure, while external AI models are limited to handling only obfuscated representations of the data. This approach ensures full data confidentiality while still enabling sophisticated AI-driven reasoning, query generation, and workflow automation. By separating sensitive data from AI processing, the system maintains compliance with enterprise security policies and global privacy regulations, while supporting seamless integration with third-party AI technologies.

---

1. Introduction

Organizations across the world are rapidly adopting AI-driven systems to interpret and process natural-language queries related to identity governance, entitlement visibility, policy validation, and access risk assessments. These systems provide users with intuitive interactions that previously required navigating complex IAM consoles or understanding specialized query languages. However, the underlying identity data involved in these processes—typically stored in knowledge graphs—contains deeply sensitive information such as user identities, group memberships, roles, policy bindings, separation-of-duty constraints, and audit relationships. Exposing this data directly to external AI services poses an unacceptable privacy and security risk, especially when those AI services operate outside the organization's infrastructure.

Modern AI innovation is overwhelmingly driven by large general-purpose models trained and maintained by third-party providers. While these models offer exceptional natural-language understanding and reasoning capabilities, they cannot be trusted with raw enterprise identity data, nor can organizations guarantee how such information would be logged, stored, or used once transmitted outside their boundary. Even when these models run on secure vendor infrastructure, sensitive enterprise data leaving the organizational perimeter often violates regulatory obligations and internal compliance requirements.

The challenge, therefore, is not simply about using AI to improve identity governance; it is about using AI in a way that respects data sovereignty and ensures that organizations maintain full control over their identity information. This white paper introduces a secure architectural approach that solves this problem through a combination of one-way hashing, strict boundary controls, policy-driven execution, and an identity-aware obfuscation workflow tailored specifically for cross-boundary AI interactions.

---

2. Problem Statement

In a typical AI-assisted identity governance workflow, user queries are sent to a model that interprets the request and constructs an appropriate knowledge graph query (such as a Cypher or Gremlin query). The resulting graph query is then executed against a highly sensitive identity dataset containing user accounts, entitlement mappings, privileged roles, administrative relationships, and audit metadata. This data is not only proprietary—it is often subject to strong compliance requirements such as GDPR, HIPAA, SOC2, and government-specific privacy mandates. Any leakage of this information, even in small fragments, may constitute a violation of privacy regulations, confidentiality agreements, or industry compliance frameworks.

Organizations face several difficult constraints. First, they cannot directly transmit sensitive data—such as personal names, access roles, policy identifiers, or privilege assignments—to external AI models without creating the possibility of unintended exposure. Second, even when organizations attempt to run open-source models within their secure perimeter, the cost of maintaining GPU hardware, continually updating models, applying security patches, and managing the operational lifecycle becomes prohibitively high. Third, most identity governance datasets include deeply interconnected relationship structures that, when transmitted in raw form, may reveal organizational structure, role hierarchies, operational responsibilities, or sensitive administrative pathways. Revealing such structures could expose the organization to security vulnerabilities or targeted attacks.

Traditional data-masking strategies are insufficient because they often rely on reversible transformations that still reveal patterns or maintain semantic correlations. Furthermore, identity governance queries frequently reference specific individuals or roles whose names or identifiers cannot be shared externally. Without a secure abstraction layer, organizations must choose

between forgoing the benefits of AI or accepting significant data leakage risk. This paper proposes an architectural framework that allows organizations to gain the full benefits of external AI reasoning without ever exposing the underlying identity information.

---

3. Proposed Solution: Identity-Aware Obfuscated Data Exchange Framework

The proposed solution is designed around the principle that external AI systems should never be allowed to view or infer actual identity data. Instead, sensitive information is transformed within the organization's boundary using a one-way hashing mechanism that produces irreversible tokens. These tokens preserve structural meaning and positional relevance but reveal nothing about the original values. The external model operates entirely on these hashed placeholders, generating queries and performing reasoning without any knowledge of the true underlying data.

When a user submits a natural-language question related to identity governance, the system first identifies sensitive entities contained in the query. This includes personal names, group names, role titles, resource identifiers, and policy constructs. The Data Obfuscation Layer immediately transforms these sensitive values into unique, salted cryptographic hashes. The hashed query, now stripped of any identifiable information, becomes suitable for transmission to external AI models. Because the transformation is one-way and irreversible, the external model has no method—mathematically or logically—to reconstruct sensitive details.

After the model generates a graph query, this query returns to the organizational infrastructure where it is executed against the actual knowledge graph. Importantly, the external model never interacts directly with the real graph. The internal system retrieves the resulting dataset and again applies the same hashing mechanism to sensitive fields before sending the output back to the external model for final reasoning or summarization. Once the model produces its final answer, the Data Obfuscation Layer uses its internal mapping table to replace the hashed tokens with the original values, ensuring the final output delivered to the user is complete, accurate, and semantically faithful.

This approach ensures that the external AI model becomes a reasoning engine rather than a data repository. It receives only obfuscated values, never raw identity information. All query execution remains inside the secure perimeter, and all sensitive data is only ever accessible to internal systems. This separation of responsibilities preserves data confidentiality while enabling full use of advanced AI capabilities.

---

4. Architecture Overview

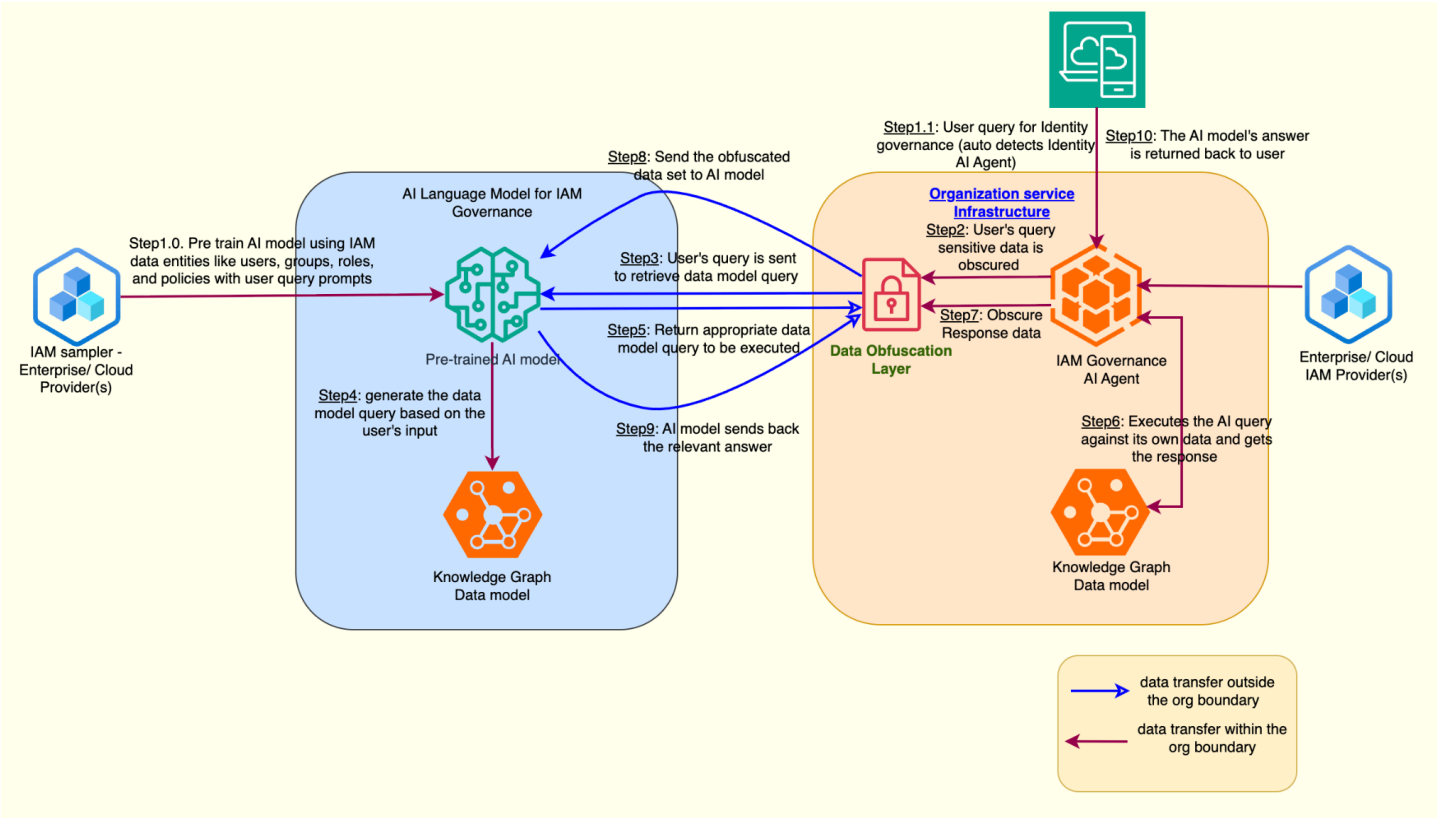
The architecture supporting this model consists of several interconnected components, each responsible for preserving confidentiality while enabling meaningful AI-driven processing. At the top of the flow, the user interacts with the system through a natural-language interface. The system automatically detects the intent behind the request and recognizes whether the query pertains to identity governance. If the system determines the request is identity-related, it routes the query toward the Data Obfuscation Layer.

Before any part of the query is exposed to an external AI model, the Data Obfuscation Layer performs entity recognition to identify sensitive fields and applies strong cryptographic hashing to each element. It also maintains a secure internal mapping that allows the system to reconstitute the original values later. This mapping is never shared externally and never leaves the organization's secure infrastructure.

The obfuscated query is sent to the external AI model, represented as the Pre-Trained AI Model in the architecture diagram. This model leverages its extensive training to generate a well-formed graph query, relying entirely on the structural relationships preserved in the hashed tokens. The model is unaware of any real identity values, organizations, or contextual meaning beyond the shape and structure of the obfuscated inputs.

Once the external model produces the graph query, the query is transmitted back into the organization's secure environment. Here, the Organization Service Infrastructure takes responsibility for evaluating the query under policy constraints. Identity governance policies ensure that the query respects access controls, operational boundaries, and compliance requirements. The graph database executes the query and retrieves the necessary data, which again consists of deeply sensitive information.

Before this data is allowed to leave the secure environment, the Data Obfuscation Layer applies the hashing mechanism once more. This ensures that even the output dataset is safe for external processing. The hashed results are sent back to the AI model for final reasoning or summarization. After the model completes its output, the internal system substitutes the hashed tokens with the original values using the secure mapping table, producing a clean, human-readable response for the user.



---

5. Case Study: Identity Governance AI Query Processing

To validate the architecture, a case study was conducted using an open-source AI model (Ollama) integrated with a custom identity governance knowledge graph. The objective was to simulate real-world identity governance queries and evaluate whether operational workflows could be supported without exposing sensitive identity data.

A representative query, such as “What are the roles assigned to user Anna D’Souza?”, provides a meaningful example. In this scenario, the user's name is immediately transformed into a cryptographic hash before being sent to the external AI model. The external model encounters only a hashed representation of the name and uses this token to generate a Cypher query. When this query is executed internally, the resulting roles—such as “Lab Technician” or “Receptionist”—are themselves considered sensitive. These values are therefore also hashed before they are sent to the AI for summarization. Only at the very end of the processing pipeline, after the AI has produced its response, does the internal system re-insert the true values by reversing the hashed placeholders using the internal secure mapping.

The result is a seamless and secure workflow: the AI model never sees “Anna D’Souza,” never sees “Lab Technician,” and never interacts with the identity graph directly. Yet despite these restrictions, the system produces accurate responses to identity governance questions in a natural-language format, demonstrating the viability of this cross-boundary obfuscation framework.

---

6. Security Considerations

The security model supporting this architecture is rooted in the principle of zero trust. At every boundary crossing, sensitive data is intentionally removed or transformed, ensuring that external systems can never infer or reconstruct the underlying identity information. One-way hashing provides strong mathematical guarantees that the original values cannot be recovered from the hashed output. The system uses salted and optionally peppered hashing to prevent collisions and eliminate the possibility of dictionary or correlation attacks. Because external AI models receive only hashed values with no additional contextual hints, they are unable to derive any meaningful insights about the organization’s internal structure or identity relationships.

Additionally, the system enforces strict data minimization principles consistent with leading industry recommendations, such as those articulated in Salesforce’s guidance on protecting customer data in LLM interactions. Only the minimal amount of data necessary to formulate a proper AI response is ever transmitted outside the organizational boundary. No raw identity fields, no metadata, no logs containing sensitive information, and no policy identifiers are shared externally. Even the communication between internal and external components is tightly controlled through secure APIs, audit logs, and outbound filtering policies that ensure adherence to compliance requirements.

---

7. Conclusion

As enterprises move toward AI-assisted identity governance, the need for secure, privacy-preserving data exchange across organizational boundaries becomes increasingly urgent. Traditional approaches to AI integration either compromise on security by sharing sensitive data externally or compromise on capability by limiting reliance on advanced AI models. The system described in this paper offers a path forward that reconciles both goals. By grounding the architecture in one-way hashing, strict boundary enforcement, and internal execution of all sensitive operations, organizations can adopt next-generation AI features without sacrificing data confidentiality or compliance posture.

This framework provides a scalable, future-proof foundation for AI-driven identity systems, ensuring that the world's most sensitive identity data remains protected while still enabling AI to play a transformative role in governance, automation, and security operations.