rubrik Pinecone

WHITE PAPER

Powering Secure and Scalable Generative AI for the Enterprise with Rubrik Annapurna and Pinecone

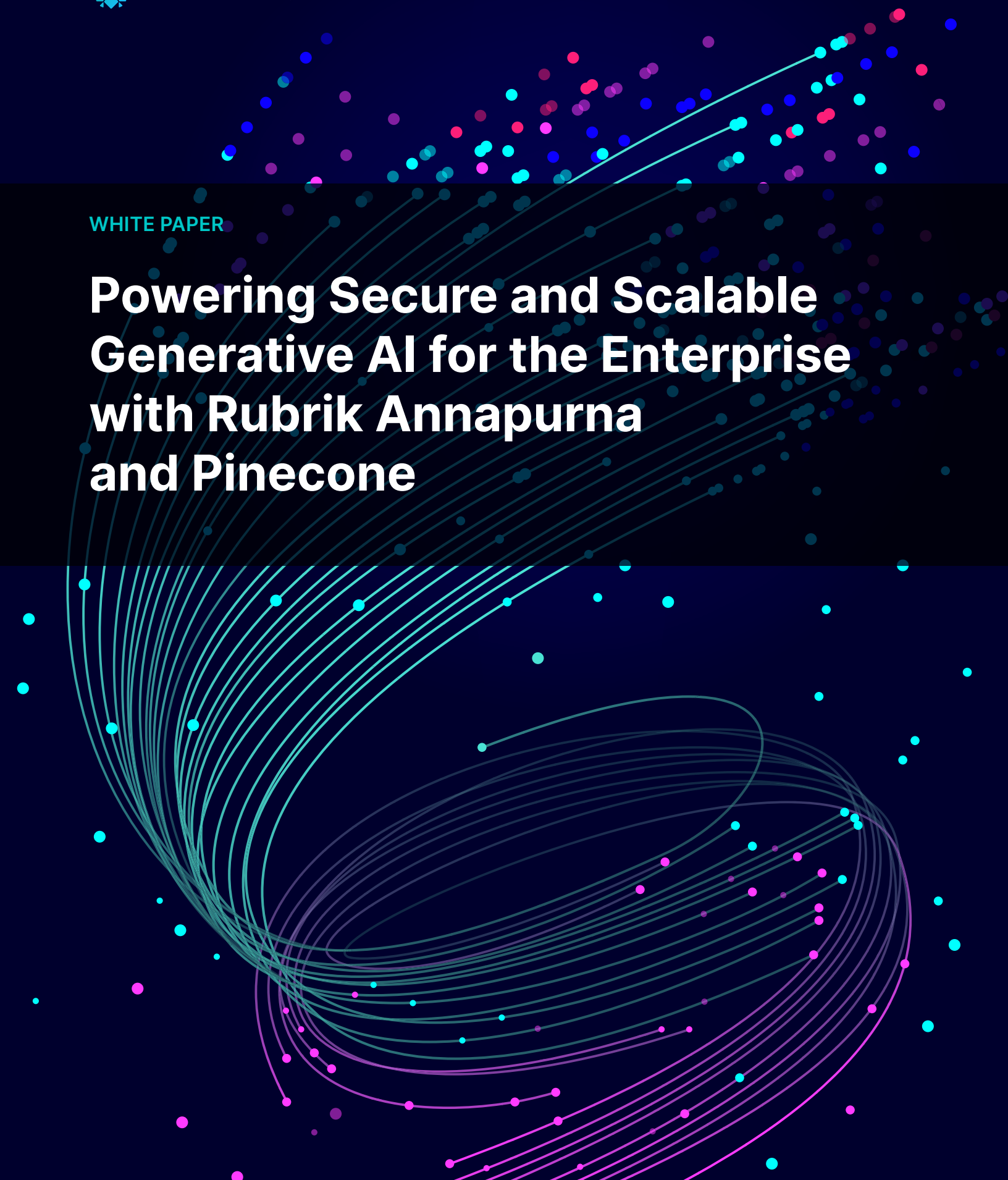


Table of Contents

EXECUTIVE OVERVIEW

INTRODUCTION

Welcome to *Powering Secure and Scalable Generative AI for the Enterprise with Rubrik Annapurna and Pinecone*. In this document, we will explore the growing need and challenges associated with deploying **secure**, scalable AI applications in the modern enterprise. We will also discuss how the combination of Rubrik Annapurna and Pinecone Database empowers organizations to securely leverage enterprise data that resides on-prem, in the cloud, and in SaaS applications. Built-in sensitive data suppression, application aware pre-embeddings and secure access capabilities to accelerate enterprise-scale generative AI adoption will also be discussed.

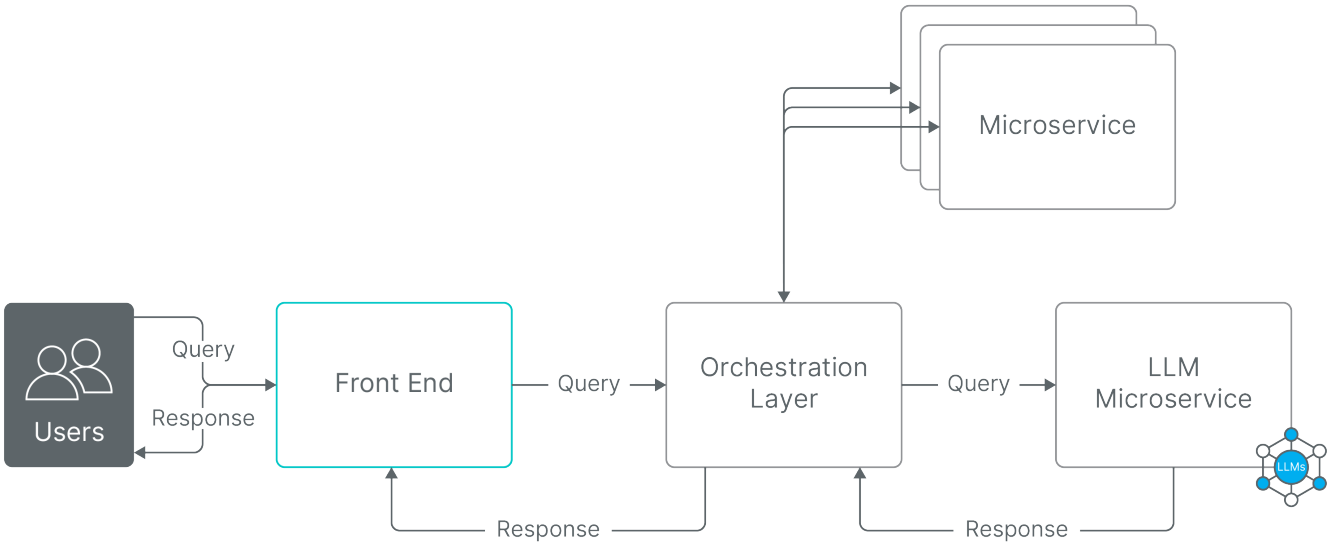
AUDIENCE

This whitepaper is designed for individuals and teams looking to build secure, scalable, and performant enterprise applications powered by Large Language Models (LLMs), particularly those interested in integrating their organization's unique data assets. This includes, but is not limited to, application architects, AI engineers, security professionals, business leaders, and systems administrators working on developing AI solutions. If you are involved in a data-driven AI project, and have a vested interest in security, compliance, or governance, you have come to the right place.

AI APPLICATIONS IN THE ENTERPRISE: OPPORTUNITIES AND CHALLENGES

INTRODUCTION TO AI APPLICATIONS AND DATA PATTERNS

The figure below depicts a simple AI application utilizing a Large Language Model (LLM) service to understand the end user's query and generate an appropriate response. The user issues the query via the application front end—in this case, likely some sort of chatbot. The orchestration layer directs the query to the LLM microservice, and the LLM microservice leverages one or more LLMs to generate a response using natural language. Easy!



Unfortunately, from a data perspective, building AI applications solely with an LLM is fairly rigid and less than ideal. While foundation models and their commercial implementations have been trained on Internet-scale data, this dataset does not include your enterprise-specific or proprietary information. Moreover, relying exclusively on internet-sourced data raises concerns about accuracy, reliability, and timeliness since not all information available online is accurate or up to date. These limitations can lead to several undesirable outcomes, including outdated responses, inaccuracies, and an inability to effectively leverage your organization's unique knowledge. This limitation leads to several undesirable outcomes:

- **Limited knowledge scope:** The application's responses will be constrained to information available up to the LLM's training cutoff date, potentially leading to outdated or incomplete answers. Additionally, the application is unable to incorporate specialized or proprietary information crucial for many business use cases into its responses. This significantly limits its practical utility.
- **Increased risk of hallucinations:** The LLM may generate plausible-sounding but inaccurate information when faced with queries outside its training data, compromising the reliability of the application's responses.
- **Lack of real-time adaptability:** The application will struggle to respond to current events, trends, or rapidly evolving topics, as it cannot integrate new information dynamically.

To address these challenges and add more organizational relevance and accuracy, several approaches have emerged:

- **Building custom LLMs:** Creating a fully custom LLM by training your own model(s) and including your organization's data as part of the training data corpus. However, this approach is extremely resource-intensive, requiring significant computational power, expertise, and time, making it impractical for most organizations. Additionally, incorporating new data or keeping the model current with the latest changes is challenging, as updating the model typically involves retraining or extensive fine-tuning—processes that are both difficult and costly to perform regularly.
- **Fine-tuning:** Taking a pre-trained LLM and further training it on a more specific dataset or task to improve its performance for a specific use case. While less resource-intensive than building a custom LLM, fine-tuning still requires substantial data preparation, can be computationally expensive, and may lead to overfitting or degradation of the model's general capabilities. Fine tuning also suffers from similar shortcomings as training custom LLMs when it comes to incorporating new data or incorporating changes.

Thus, most enterprises have converged on the **Retrieval Augmented Generation** (RAG) approach. This solution was first outlined in a 2020 paper by a team of AI researchers from Meta, who coined the term. RAG addresses the limitations of standalone LLMs by incorporating external knowledge sources into the generation process, allowing applications to access and utilize up-to-date, organization-specific information in their responses.

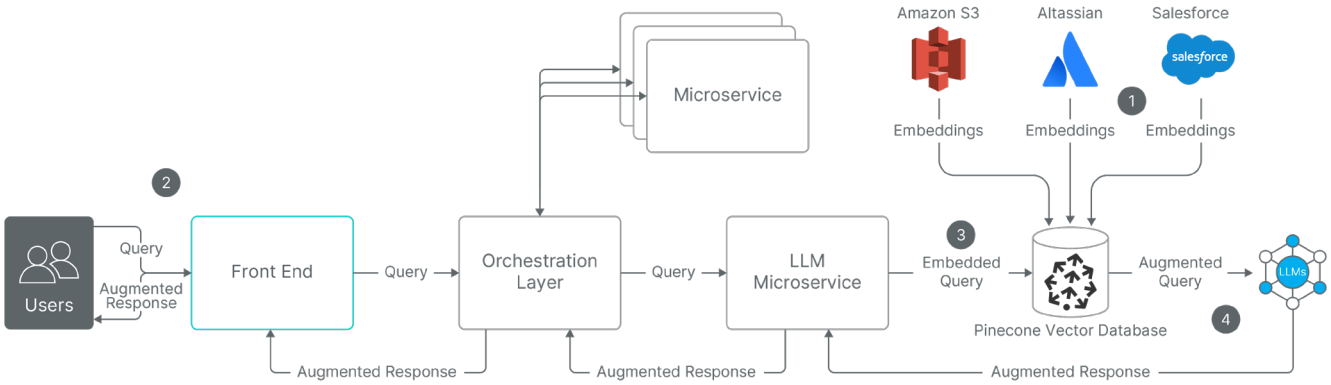
RETRIEVAL-AUGMENTED GENERATION (RAG) IN THE MODERN ENTERPRISE

RAG is an innovative AI framework that enhances the capabilities of LLMs by incorporating external knowledge sources into the generation process. RAG operates by first performing a similarity search against a vector database, such as Pinecone Vector Database, to efficiently retrieve relevant information from a curated knowledge base in response to a user query. This retrieved information is then seamlessly integrated with the original prompt and fed into the LLM, allowing it to generate more accurate, up-to-date, and contextually

appropriate responses. RAG has become a cornerstone technology for enterprises seeking to leverage the power of generative AI while maintaining control over the information sources and helping ensure the reliability and groundedness of AI-generated outputs. RAG technology brings several key benefits to an organization's generative AI efforts including:

- **Cost-effective Implementation:** RAG is a much more cost-effective approach to introducing new data to LLMs than purely leveraging foundation models or fine-tuning. It makes generative AI technology more accessible and usable by reducing computational and financial costs associated with customizing models for organization-specific information.
- **Recent and Proprietary Information:** RAG allows enterprises to provide the latest research, statistics, or news to generative models by connecting LLMs directly to frequently updated information sources that they control. This ensures that the AI can provide up-to-date and domain specific information to end users, maintaining relevancy in rapidly changing environments.
- **Enhanced Trust:** By enabling LLMs to present accurate information with source attribution, RAG increases trust and confidence in generative AI solutions. Users can verify information through provided citations or references, enhancing transparency and reliability.
- **Developer Control:** RAG gives developers greater control over chat applications, allowing them to adapt information sources, restrict sensitive information retrieval, and troubleshoot more efficiently. This flexibility enables organizations to implement generative AI technology more confidently across a broader range of applications.

Let's take a look at this process in greater detail. The diagram below depicts our previous application enhanced with a RAG architecture to incorporate data from external sources like Salesforce, Atlassian, and Amazon S3.



In this example, our application is now augmented to leverage a RAG workflow to incorporate our proprietary business data into its responses. This expands our application's scope of knowledge, reduces the likelihood of hallucinations, and gives us access to information beyond the training horizon of the LLM generating the response. Let's take a look at the steps in this workflow.

1. **Indexing –** Documents stored across various sources are converted into numerical vector representations known as embeddings, which capture their semantic meaning. These embeddings are then stored in a vector database, such as Pinecone, enabling efficient similarity searches. More information on Pinecone Database is presented in the following sections.

2. **Query –** The end user issues a query to the application front end, that query ultimately reaches the LLM microservice responsible for taking the query and returning the response.

3. **Retrieval –** The LLM microservice processes the user query by generating its embedding representation. A similarity search is then performed within the vector database (e.g., Pinecone) to identify and retrieve the documents most closely matching the query embedding.

4. **Prompt Augmentation –** The system enhances the original user query with the retrieved information. This augmented prompt combines both the original query and additional relevant context from the documents returned by the similarity search.

5. **Response Generation –** The augmented prompt is now processed by the LLM(s) and a response is generated. This response incorporates both information from the LLMs training data and the additional context retrieved from the similar documents.

The adoption of RAG represents a significant leap forward in enterprise AI applications. By seamlessly integrating external knowledge sources with the power of LLMs, RAG addresses many of the limitations inherent in traditional LLM implementations. This approach not only expands the scope of knowledge available to AI systems but also enhances their accuracy, relevance, and trustworthiness. The ability to incorporate new, domain-specific information from various enterprise data sources ensures that AI-generated responses are grounded in the organization's current context and proprietary knowledge. However, despite the advancements introduced by RAG, significant challenges persist in its implementation.

CHALLENGES WITH TRADITIONAL RAG

As organizations increasingly adopt AI applications powered by LLMs and RAG, they face a new set of challenges in implementing and scaling these technologies effectively. While RAG offers significant advantages in terms of accuracy, relevance, and up-to-date information, it also introduces complexities in data management, security, performance, and workflow integration.

This section explores four key challenges that organizations commonly encounter when implementing RAG in enterprise environments: data complexity, security gaps, performance at scale, and the integration of agentic workflows. Understanding and addressing these challenges is crucial for organizations seeking to fully leverage the potential of RAG while maintaining robust, secure, and efficient AI applications.

Data Access

RAG systems need to be provided access to enterprise data in order to be useful. Building connectors to each data source and securing those connectors can be problematic and time-consuming for developers. Oftentimes, data sources don't have the performance needed to efficiently supply data to the RAG system. Instead, a data lake needs to be created where the source data is copied to. Building a data lake and storing data in it can be costly, complex and time-consuming. Data lakes also can induce security risks because the original security of the data is removed by the copying process.

Data Refresh

In order for a RAG system to provide timely and accurate results, it must operate on the current or near current copy of data. When the source data can't be directly accessed due to security or performance issues, Extract Transform & Load (ETL) pipelines must be built to copy that data to a place where the RAG solution can access it. Typically, this is done by using a data lake. The building, maintenance and operations of these ETL pipelines can be complex and costly to corporations.

Data Complexity

The implementation of RAG in enterprise environments is often hindered by the complexity of existing data ecosystems. Organizations typically have data scattered across multiple platforms, including legacy on-premises systems, various cloud services, and SaaS applications. This fragmentation creates significant challenges in data integration, as each system may have its own data format, access protocols, and update frequencies. Harmonizing these diverse data sources into a comprehensive knowledge base for RAG requires sophisticated data engineering efforts. Moreover, ensuring data consistency and managing real-time updates across these disparate systems adds another layer of complexity, potentially impacting the accuracy and timeliness of RAG-generated responses.

Data Security and Governance

Implementing RAG while maintaining robust security measures presents a significant challenge for organizations. As RAG systems need to access a wide range of data sources, including potentially sensitive information, ensuring data privacy becomes paramount. Organizations must implement stringent Role-Based Access Control (RBAC) to help ensure that AI systems only access data that users are authorized to see and to help ensure that users can only access data they are authorized to consume via the LLM. This requires a complex mapping of user permissions across various data sources and the AI system itself. Additionally, organizations need to develop sophisticated mechanisms for sensitive data suppression, helping ensure that confidential information is not inadvertently exposed in AI-generated responses. This challenge is further complicated by the need to balance security with the system's ability to provide useful and contextually relevant information, often requiring fine-grained control over data access and usage.

Performance at Scale

As organizations scale their RAG implementations, maintaining high performance becomes increasingly challenging. RAG systems rely on vector databases to store and retrieve relevant information quickly. However, as the volume of data grows, the accuracy and speed of vector search can degrade, potentially leading to slower response times or less relevant results. This challenge is exacerbated in enterprise environments where large amounts of data are constantly being added or updated. Organizations need to invest in advanced vector search technologies and optimized indexing strategies to maintain high-speed retrieval at scale. Without a state of the art solution, they must balance the trade-offs between index update frequency, search accuracy, and query response time. Achieving this balance often requires significant computational resources and sophisticated algorithms, making it a complex and ongoing challenge for many organizations.

Agentic Workflows

The growing adoption of AI agents in enterprise environments introduces new challenges for RAG implementations. These agentic workflows require real-time, secure access to enterprise data to perform complex, multi-step tasks autonomously. Unlike traditional query-response models, AI agents may need to access and process data from multiple sources dynamically, making data retrieval and security more complex. Organizations must develop systems that can provide these agents with secure, context-aware access to

relevant data sources while maintaining strict control over data usage and privacy. This requires not only advanced authentication and authorization mechanisms but also the ability to track and audit data access across complex, multi-step processes. Furthermore, organizations need to ensure that these agentic systems can handle the latency and potential inconsistencies inherent in accessing distributed data sources in real-time, all while maintaining the integrity and coherence of their workflows.

As organizations navigate these challenges in implementing RAG, it becomes clear that traditional approaches may fall short in addressing the complex needs of modern enterprises. The issues of data complexity, security gaps, performance at scale, and support for agentic workflows require innovative solutions that can seamlessly integrate with existing infrastructure while providing robust security and high performance. In the next section, we will explore how cutting-edge technologies like Rubrik Security Cloud and Pinecone Database are revolutionizing the RAG landscape. These advanced solutions offer powerful capabilities to tackle the aforementioned challenges, enabling organizations to harness the full potential of RAG while maintaining data security, ensuring scalability, and supporting the evolving needs of AI-driven applications.

THE SECURE AI EASY-BUTTON: RUBRIK ANNAPURNA AND PINECONE

INTRODUCTION TO RUBRIK SECURITY CLOUD (RSC)

Rubrik offers a comprehensive cybersecurity platform designed to secure and manage data across enterprise, cloud, and SaaS environments. Rubrik Security Cloud offers data protection and cyber resilience in a single solution, built on key principles of policy-driven automation, Zero Trust design, simplicity at scale, and broad ecosystem support.

Rubrik Annapurna offers several key advantages for organizations looking to implement RAG workflows. First, Annapurna consolidates an enterprise's data in a central location leveraging the same technology required for high performance backup operations. For existing customers of Rubrik this eliminates the need to create a separate data copy specifically for RAG. For other organizations, they benefit from the easy to use, no code, mature and highly performant data extraction pipelines that Rubrik has spent a decade building.This saves considerable time, resources, and money that would otherwise be spent on building and maintaining a dedicated data lake.

Second, Rubrik natively scans data in line with extraction for sensitive information, enabling it to identify potential risks associated with inadvertent data exposure through an LLM, helping to maintain data security and compliance.

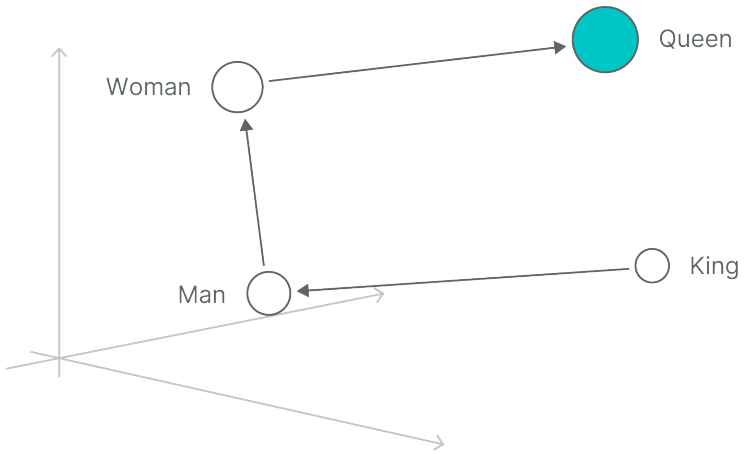
Finally, Rubrik preserves user identity and permissions from production, helping prevent the loss of permissioning context that can occur when data is transferred to a purpose-built data lake. This helps ensure that only authorized users can access specific data segments within generative AI applications, aligning with Zero Trust principles and minimizing the risk of unauthorized data exposure.

By leveraging Rubrik, organizations can streamline data access and compliance management, as well as accelerate the time to production for generative AI projects. The platform's pre-embedded security controls, such as role-based access control and sensitive data filtering, further enhance the security of AI applications. This allows enterprises to confidently deploy scalable and reliable AI applications in real-world settings.

THE ROLE OF VECTOR DATABASES IN RAG

Vector databases store and provide access to data alongside their vector embeddings. Vector embeddings are the data's numerical representation (as a long list of numbers that captures the original data's semantic meaning). Generating these embeddings is typically done via LLMs as well (i.e. you input a string of text and the embedding LLM provides the embeddings.) The data is thus represented as either closer together or further apart in vector space so similarity can be calculated based on the distance between the data object, i.e. the vector search.

As an example, the vector distance between "Queen" and "Princess" would be relatively small due to their shared semantic context of royalty, femininity, and leadership. This close proximity in the vector space would indicate a high similarity between the terms. In contrast, comparing "Queen" to a dissimilar term like "Bicycle" would yield a much larger vector distance, as these words share little semantic context. Similarly, the vector distance between the words "King" and "man" are nearly identical to the distance between "Queen" and "woman". Thus, by using vector math an LLM is easily able to respond to the following prompt "King is to man as Queen is to what?" These simple examples showcase how vector search can identify related concepts and distinguish unrelated ones, even when the exact words differ, providing a more nuanced understanding than traditional keyword-based searches.



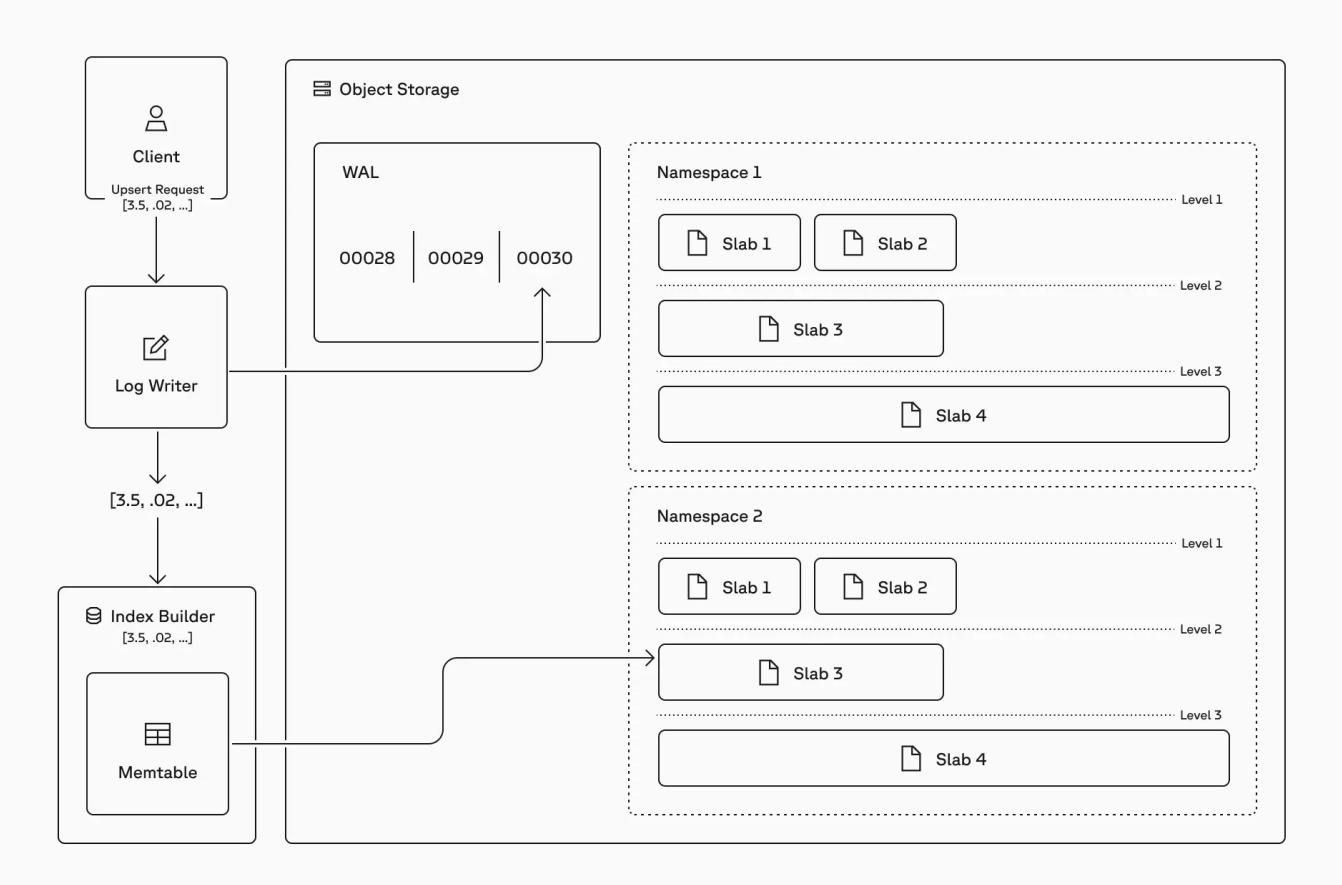
King - Man + Woman = Queen

While traditional databases support storing vector embeddings, vector-databases like Pinecone are AI-native and optimized to perform lightning-fast vector searches at scale, by pre-calculating the distance between data objects and using a vector index.

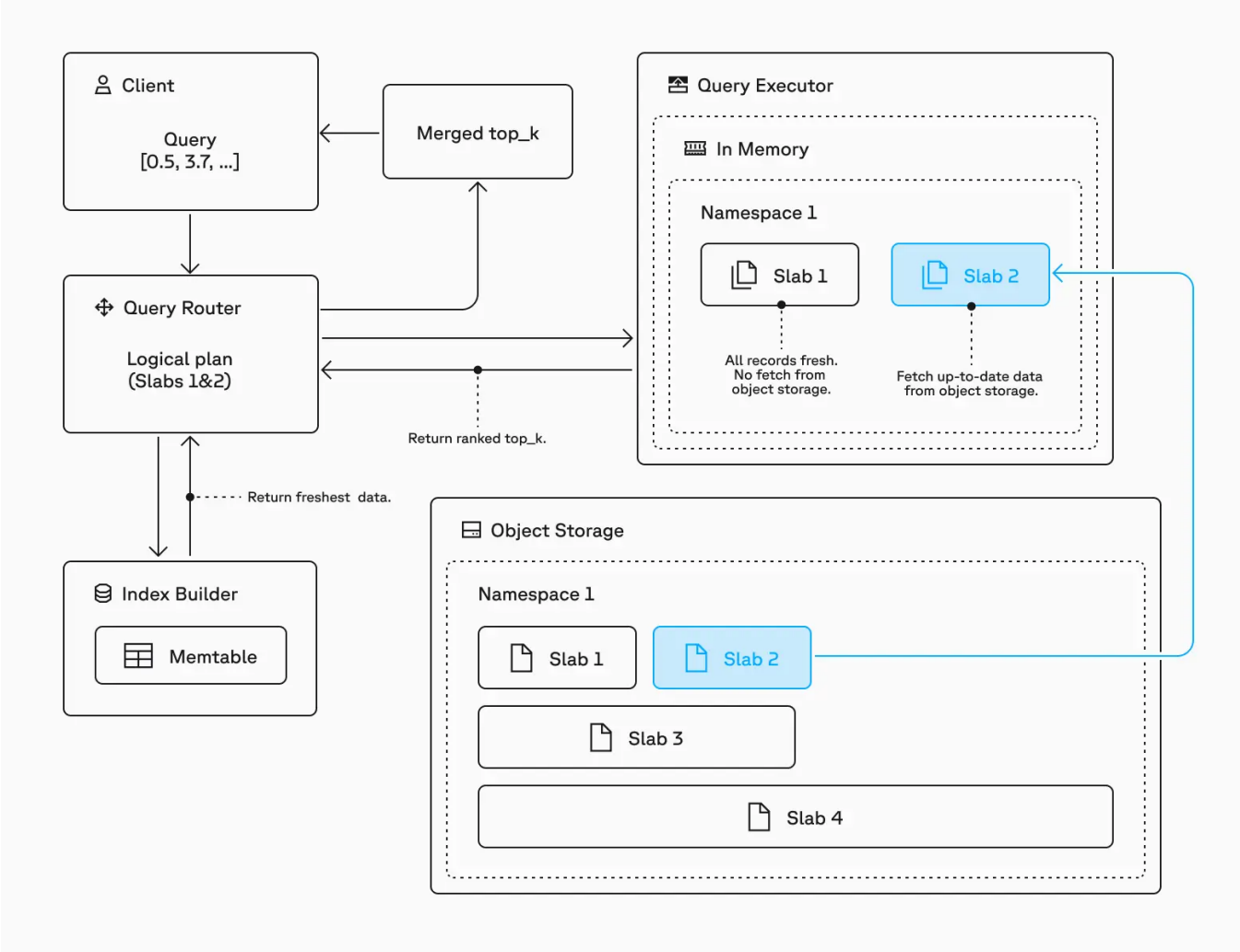
INTRODUCTION TO PINECONE DATABASE

Pinecone is a high-performance vector database that is purpose-built to manage and query vector embeddings at scale. Unlike traditional databases, Pinecone addresses the unique challenges vector data poses, such as high dimensionality, complexity, and the need for approximate nearest neighbor (ANN) searches. It combines cutting-edge indexing techniques, metadata filtering, hybrid search, cascading retrieval, and real-time updates with modern database capabilities, such as integrated reference, making it ideal for building knowledgeable production-grade AI applications.

At the core of Pinecone's architecture is its **serverless slab-based design**, which decouples storage from compute and organizes data into **immutable, log-structured "slabs."** This enables dynamic indexing and resource allocation based on workload demands and data patterns—helping reduce costs by eliminating over-provisioning while ensuring low-latency performance, even under variable query loads. The architecture also supports **predictable caching and efficient metadata filtering** to maintain performance at scale. By abstracting infrastructure management, Pinecone empowers developers to focus on building impactful AI applications without the complexity of provisioning, scaling, or maintaining infrastructure.



Pinecone's **adaptive, slab-aware indexing algorithms** deliver low latency, high recall, and always-fresh data at any scale. The system helps minimize resource usage while ensuring high performance by **applying lightweight algorithms like scalar quantization to small slabs**, and reindexing larger slabs with advanced methods like **clustering, partitioning, or graph-based search**. Each query scans only the most relevant slabs, optimizing compute and memory efficiency. This **adaptive approach dynamically reflects changes in data distribution**, maintaining relevance and freshness for real-time searches, without requiring manual reindexing.



Enterprise-grade security is another cornerstone of Pinecone's platform. Built-in safeguards ensure data is encrypted at rest and in transit, isolated from other workloads, and never used beyond servicing API calls. Pinecone meets stringent compliance standards, including SOC 2 Type II, GDPR, and HIPAA (with a BAA upon request). Features like Single Sign-On (SSO), role-based permissions, and private endpoints provide robust access control, while routine backups protect against accidental deletions or system failures. Continuous monitoring and dedicated support ensure reliability and rapid issue resolution for mission-critical applications.

Finally, Pinecone's support for metadata-driven filtering enhances search relevance by incorporating contextual details like timestamps, categories, or user attributes. This capability is invaluable for personalized recommendations and hybrid search use cases. The platform also supports multitenancy through namespaces, securely partitioning data across workloads while maximizing resource efficiency.

Pinecone allows organizations like Rubrik to win by combining innovative algorithms, serverless scalability, robust security, and operational simplicity to address the demands of modern AI applications. By abstracting the complexities of managing vector data, Pinecone empowers organizations to deliver high-impact solutions like semantic search, generative AI, recommendations, and real-time personalization. With its adaptive clustering, enterprise-grade security, and metadata-rich capabilities, Pinecone is not just a database—it's a strategic enabler for AI-driven innovation.

RUBRIK ANNAPURNA AND PINECONE- SECURE, EFFICIENT AI DEPLOYMENT

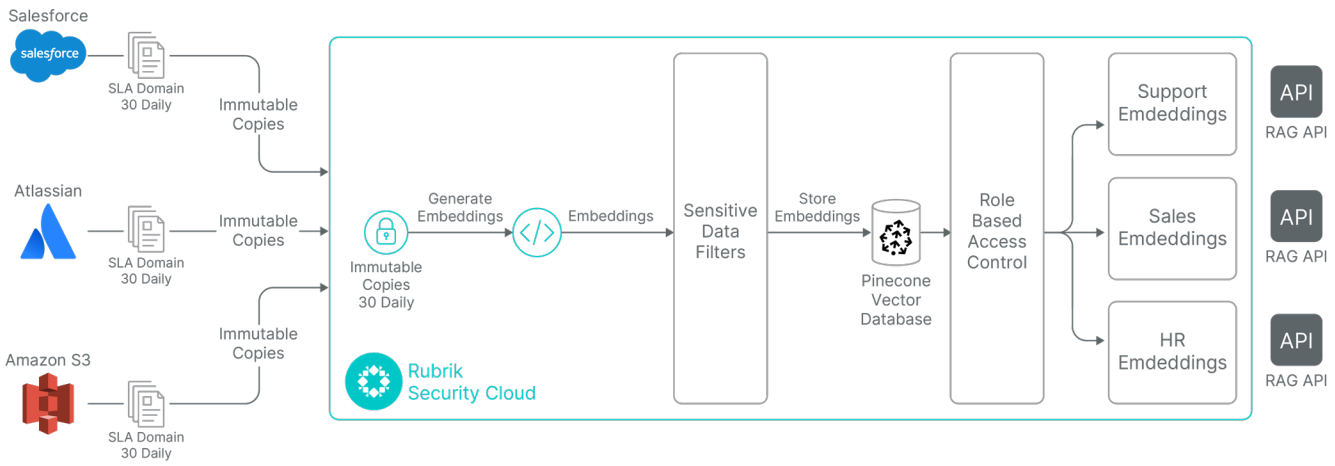
Rubrik Annapurna is a groundbreaking service designed to revolutionize the development of generative AI applications by addressing the key challenges associated with traditional RAG implementations. Annapurna offers a solution to the data complexity, security gaps, performance issues, and agentic workflow challenges that organizations face when implementing RAG. This innovative service provides secure and efficient access to enterprise data across on-premises, cloud, and SaaS environments through a single, unified API.

Annapurna's key features include API-first secure data access with dynamic updates, an application-aware embedding engine, and seamless data mobilization across various platforms without complex ETL processes. By eliminating shadow datastores and simplifying AI development, Annapurna enables organizations to leverage their proprietary knowledge for AI applications while maintaining robust security and compliance. This service represents a significant leap forward in addressing the limitations of traditional RAG approaches, offering a more streamlined, secure, and scalable solution for enterprises looking to harness the full potential of their data in AI-driven applications. In the following section we will explore Rubrik Annapurna's architecture, use cases, and key benefits.

HOW IT WORKS

ARCHITECTURE AND COMPONENTS

At its core, Rubrik Security Cloud is a data security and management platform. Customers connect RSC to all of their data containing assets on premises, in the cloud, and in SaaS platforms. They then configure and assign Service Level Agreement (SLA) Domain policies to match their business requirements around recoverability and assign these policies to the appropriate datasets. Rubrik's data lifecycle engine interprets these policies and creates periodic copies of each dataset, then stores them in an immutable format, air gapped away from the source data repository. These copies are then indexed, and analyzed for sensitive data, threats, and indicators of compromise. This process is depicted in the figure below.



Rubrik Annapurna taps into this vast repository of data already protected by Rubrik Security Cloud. This includes data from on-premises systems, cloud environments, and SaaS applications, providing a rich and diverse dataset for RAG applications. RSC's data lifecycle engine ensures the most up to date data is captured on a regular basis. Annapurna leverages RSC's indexing and security analytics faculties to discover sensitive data, generate secure embeddings, and help make those embeddings available only to the appropriate applications and end users.

With Annapurna, your entire enterprise data estate is now at your fingertips preprocessed for secure use, at scale. Annapurna integrates seamlessly with your LLM or AI application framework of choice by providing a secure embedding API endpoint for any use case. Customers leveraging Rubrik Annapurna realize a number of benefits over traditional RAG including:

Application Aware Embeddings: Utilizing its application-aware embedding engine, Annapurna creates secure embeddings of the protected data. These embeddings are generated with an understanding of common application schemas, ensuring that the context and structure of the data are preserved.

Dynamic Updates: As Rubrik Security Cloud continuously protects and updates data, Annapurna dynamically refreshes the embeddings. This ensures that RAG applications always have access to the most current information without manual intervention.

Permissions and Access Control: Annapurna inherits the robust security model of Rubrik Security Cloud. It respects existing data access permissions, helping ensure that RAG workflows only utilize data that users are authorized to access.

Sensitive Data Handling: Leveraging Rubrik's sensitive data discovery capabilities, Annapurna can identify and handle sensitive information appropriately in RAG workflows, reducing the risk of exposing confidential data.

Metadata Utilization: Annapurna not only uses the protected data itself but also leverages the rich metadata collected by Rubrik Security Cloud. This metadata can provide additional context and insights for RAG applications.

Efficient Data Retrieval: By utilizing Rubrik's efficient data indexing and retrieval mechanisms, Annapurna can quickly access relevant data for RAG workflows, improving the performance of AI applications.

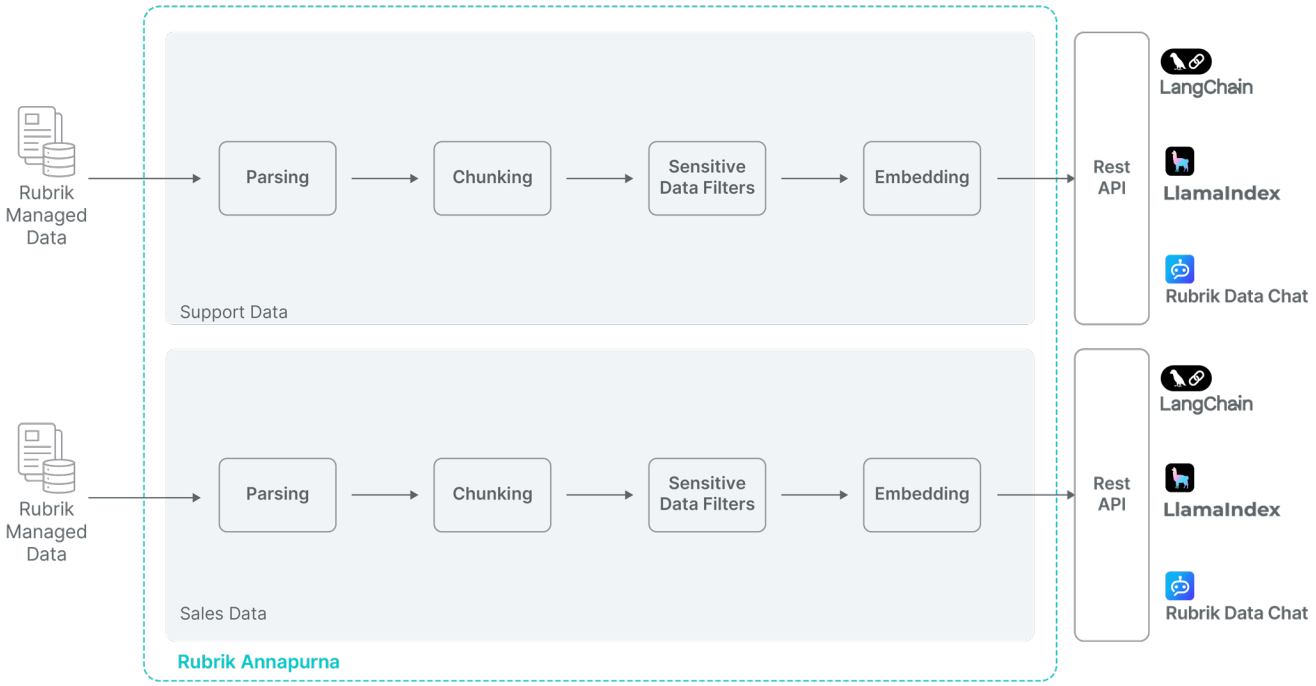
Cross-Platform Integration: Annapurna's ability to seamlessly access data across various protected platforms allows for the creation of comprehensive RAG workflows that can draw insights from multiple data sources.

SECURE DATA PIPELINES

Rubrik Annapurna uses embeddings to make business data readily accessible for LLM applications and agents. Within Rubrik Security Cloud, users can define, on a per-use-case basis:

- **Which** source data is in scope
- **What** sensitive data types are permitted or disallowed
- **Who** can access the generated embeddings

This customizable approach enables organizations to tailor their data access for AI applications while maintaining strict control over data security and compliance. Embeddings act as a bridge between protected enterprise data and AI models, helping ensure that only appropriate, authorized data is used in generative AI workflows by the appropriate entities. The following diagram depicts two example embedding workflows—one built to service a support use case, and one to support a sales use case.



Let's explore the process of creating embeddings in Rubrik Annapurna and consuming their outputs in more detail.

1. An authorized user or service account connects to RSC, receives authorization, and triggers the workflow to create new embeddings. This can be done using a wizard in the RSC UI or by leveraging the RSC API. This workflow allows customers to define the following parameters:
   - Which specific enterprise, cloud and SaaS data sources should be included in the embeddings?
   - What categories and classifications of sensitive data are permitted in the generated embeddings? What, if any, specific data types are disallowed?
   - Which roles within RSC are authorized to access embeddings?

Additionally, Annapurna maintains the permissions context of the original dataset helping ensure that access to datasets remains appropriately authorized.

2. Annapurna ingests the source data and extracts text and metadata from the raw data.
3. The parsed data is broken up into smaller chunks in preparation for embedding.
4. These chunks are filtered based upon the filtering policies chosen by the user.
5. Embeddings are generated and stored in a secure format, accessible only to authorized users.

At this point the embeddings are ready for retrieval and consumption. Rubrik has built retrievers for popular LLM frameworks like LangChain and LlamaIndex to make integration simple. Additionally, if customers want to test the outputs interactively they can leverage RSC's data chat feature to experiment with the outputs via an embedded chat interface and LLM. As of this writing Microsoft Azure OpenAI Service is integrated into Annapurna by default with additional LLMs via a pluggable architecture to follow.

USE CASES

Having explored the core functionality of Rubrik Annapurna, we will now examine several architectural references and example use cases. These scenarios serve as valuable starting points for organizations looking to harness the combined power of Rubrik Annapurna and Pinecone Database. By leveraging these technologies, businesses can build secure, scalable, accurate, and high-performance LLM applications enhanced by RAG.

The following examples illustrate how Annapurna can be applied to solve real-world challenges and drive innovation across various industries and departments.

Basic RAG

A simple RAG workflow will serve well as an example of how many organizations get started with Rubrik Annapurna. This example illustrates how to use Annapurna to build a simple AI assistant for a project or event. In this example, an organization wants to enable its employees to get answers quickly about the annual performance review process. In this scenario, cross functional information about the project, like company policies and guidelines, can be stored in OneDrive and incorporated into the application via RAG. Annapurna's sensitive data policies and document level RBAC capabilities are used to ensure only the appropriate data is accessible via application.

The result is a simple chatbot stakeholders can use to query this knowledge repository via natural language.

Agentic Applications

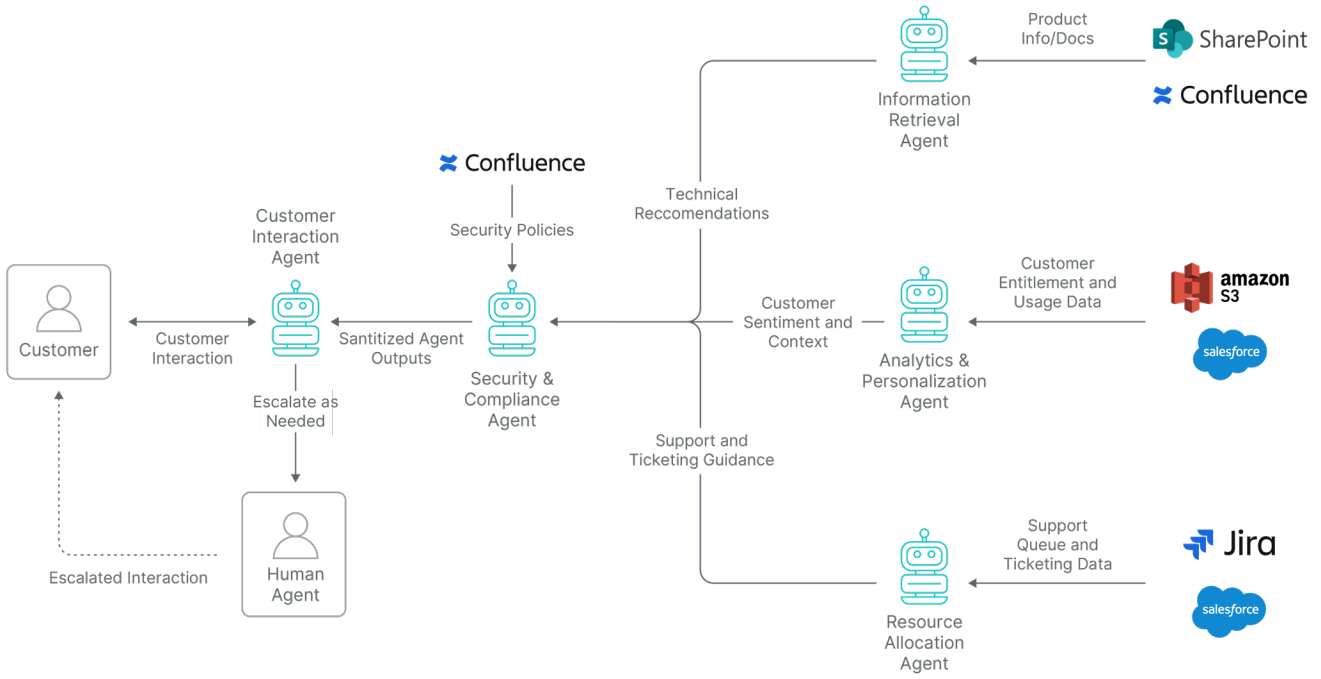
Agentic applications are AI systems designed with enhanced autonomy, decision-making capabilities, and adaptability. Unlike traditional AI applications that rely on a single LLM for specific tasks, agentic applications involve multiple AI agents working together to pursue complex goals and execute multi-step workflows with limited human supervision. These applications can understand context, set appropriate goals, and adapt their actions based on changing conditions. The increased complexity of agentic applications poses unique challenges for organizations, particularly in terms of data management, embeddings, and security. Handling diverse data sources and maintaining up-to-date, relevant information for multiple agents becomes more complex. Embedding generation and management must be more sophisticated to support the varied needs of different agents within the systems. Security concerns are amplified due to the increased attack surfaces, with risks including unauthorized data access, manipulation of decision-making processes, and potential exploitation of tool integration points.

Example Agentic Use Case

INTELLIGENT CUSTOMER SUPPORT AUTOMATION

A simple yet powerful agentic application for a modern enterprise could be an AI-driven customer support system. This system would leverage multiple AI agents to handle complex customer inquiries efficiently and autonomously. There are numerous ways an organization could design and implement such a system, let's explore a simple agentic implementation that would serve this use case.

Intelligent Customer Support Automation represents a cutting-edge approach to customer service, leveraging an ensemble of specialized AI agents to deliver personalized, efficient, and secure support. This system employs the following Agents in concert to deliver rapid world class support to customers.



Customer Interaction Agent: This agent serves as the primary interface between the customer and the Intelligent Customer Support Automation system. It interacts with customers through various channels such as chat, voice, and email, using natural language processing to interpret and understand customer queries. This agent's primary responsibilities include initiating conversations, gathering initial information, and presenting final responses to customers in a clear, conversational manner. It orchestrates the support process by delegating specific tasks to other specialized agents, such as requesting relevant information from the Information Retrieval Agent or personalized insights from the Analytics and Personalization Agent. The Customer Interaction Agent then synthesizes the inputs from these agents, ensuring the information is coherent and tailored to the customer's needs. It also works closely with the Security and Compliance Agent to ensure all interactions adhere to data protection policies. In complex scenarios, it can escalate issues to human agents, providing them with a comprehensive summary of the interaction and relevant context.

Information Retrieval Agent: The Information Retrieval Agent is responsible for quickly finding and extracting relevant information from various enterprise data sources. It processes queries from the Customer Interaction Agent, searching through knowledge bases, product documentation, and FAQs stored in platforms like Confluence and SharePoint. This agent uses advanced search algorithms and natural language processing to understand the context of queries and retrieve the most pertinent information. It then summarizes and structures this information for easy consumption by other agents, particularly the Customer Interaction Agent. The Information Retrieval Agent continuously updates its knowledge base to ensure it provides the most current and accurate information.

Analytics and Personalization Agent: This agent focuses on analyzing customer data to provide personalized insights and recommendations. It accesses customer purchase history, browsing behavior, and demographic data from sources like Salesforce and Amazon S3. By applying machine learning algorithms, it identifies patterns and preferences in customer behavior. The agent generates personalized product or service recommendations and provides valuable insights on customer behavior to other agents, particularly the Customer Interaction Agent. This personalization helps in tailoring responses and proactively addressing customer needs, significantly enhancing the overall customer experience.

Resource Allocation Agent: The Resource Allocation Agent optimizes the use of company resources to best serve customers. It monitors current support queue status in Jira, agent availability, skills and service level agreements in Salesforce. This agent determines the priority of customer issues based on various factors such as urgency, customer value, and complexity. It decides when to allocate human resources for complex issues and works to optimize response times and overall customer satisfaction. The Resource Allocation Agent communicates with the Customer Interaction Agent to ensure seamless handoffs when escalation to human agents is necessary.

Security and Compliance Agent: Acting as a vigilant overseer, the Security and Compliance Agent ensures all interactions and data handling comply with security policies and regulations. It monitors data access and usage across all agents, accessing security policies and compliance requirements from Confluence. It works closely with all other agents, particularly the Customer Interaction Agent, to maintain a balance between providing helpful information and adhering to compliance requirements. The Security and Compliance Agent also maintains detailed audit logs of all system activities for regulatory compliance and internal security reviews.

Streamlining Agentic Data Access and Security with Annapurna

Rubrik Annapurna significantly enhances the efficiency and security of this agentic workflow by providing a unified, secure access point for all necessary data across diverse sources. Instead of implementing individual RAG workflows for each data source (Salesforce, SharePoint, Confluence, Jira, and Amazon S3), Annapurna allows for the creation of pre-configured, use-case specific data access points. With Annapurna, organizations can better handle the complexities of data retrieval and embedding generation, helping ensure that each agent can quickly access relevant, up-to-date information without compromising security or compliance. This centralized approach not only streamlines the development process but also helps ensures consistent data handling and security policies across all agents.

The integration of Pinecone with Annapurna plays a critical role in making this data accessible in a highly performant and scalable manner. As Annapurna guides the generation of vector embeddings from the processed data, Pinecone's vector database efficiently stores and indexes these embeddings. This enables rapid similarity searches across vast amounts of data, allowing agents to quickly retrieve relevant information regardless of the original data source. The combination of Annapurna's secure data processing and Pinecone's high-performance vector search capabilities ensures that the agentic workflow can operate at scale, handling complex queries and large volumes of data while maintaining low latency. This integrated approach significantly outperforms traditional methods, enabling more sophisticated, real-time AI-driven customer support.

CONCLUSION

Rubrik Annapurna powered by Pinecone Database offers a transformative solution for enterprises seeking to deploy secure, scalable, and high-performance generative AI applications. By addressing the challenges of data complexity, security, performance at scale, and agentic workflows, this combination empowers organizations to harness the full potential of their data while maintaining robust security and compliance.

KEY BENEFITS

- **Enhanced Security:** Rubrik Annapurna helps ensure that sensitive data is handled securely, with features like sensitive data suppression and role-based access control, while Pinecone provides enterprise-grade security with encryption and access controls.
- **Scalability and Performance:** Pinecone's vector database enables lightning-fast vector searches at scale, making it ideal for large-scale AI applications. Annapurna's dynamic updates and efficient data retrieval further enhance performance.
- **Simplified Data Access:** Annapurna's embedding workflows simplify data access for AI applications, allowing organizations to leverage their proprietary knowledge securely across multiple platforms.
- **Agentic Workflow Support:** The integration supports complex agentic workflows by providing secure, real-time access to diverse data sources, ensuring that AI agents can operate efficiently and securely.

As AI continues to evolve, the importance of secure, scalable, and performant solutions like Rubrik Annapurna and Pinecone Database will only grow. These technologies are poised to play a critical role in enabling enterprises to unlock the full potential of generative AI while maintaining the highest standards of data security and governance.

LEARN MORE

For organizations interested in exploring how Rubrik Annapurna and Pinecone Database can enhance their AI applications, further resources and case studies are available to provide deeper insights into the capabilities and benefits of these innovative solutions.

SAFE HARBOR STATEMENT

Any unreleased services or features referenced in this paper are not generally available and may not be made generally available on time or at all, as may be determined in our sole discretion.

rubrik

Global HQ
3495 Deer Creek Road
Palo Alto, CA 94304
United States

1-844-4RUBRIK
inquiries@rubrik.com
www.rubrik.com

wp-rubrik-annapurna-and-pinecone / 20250423