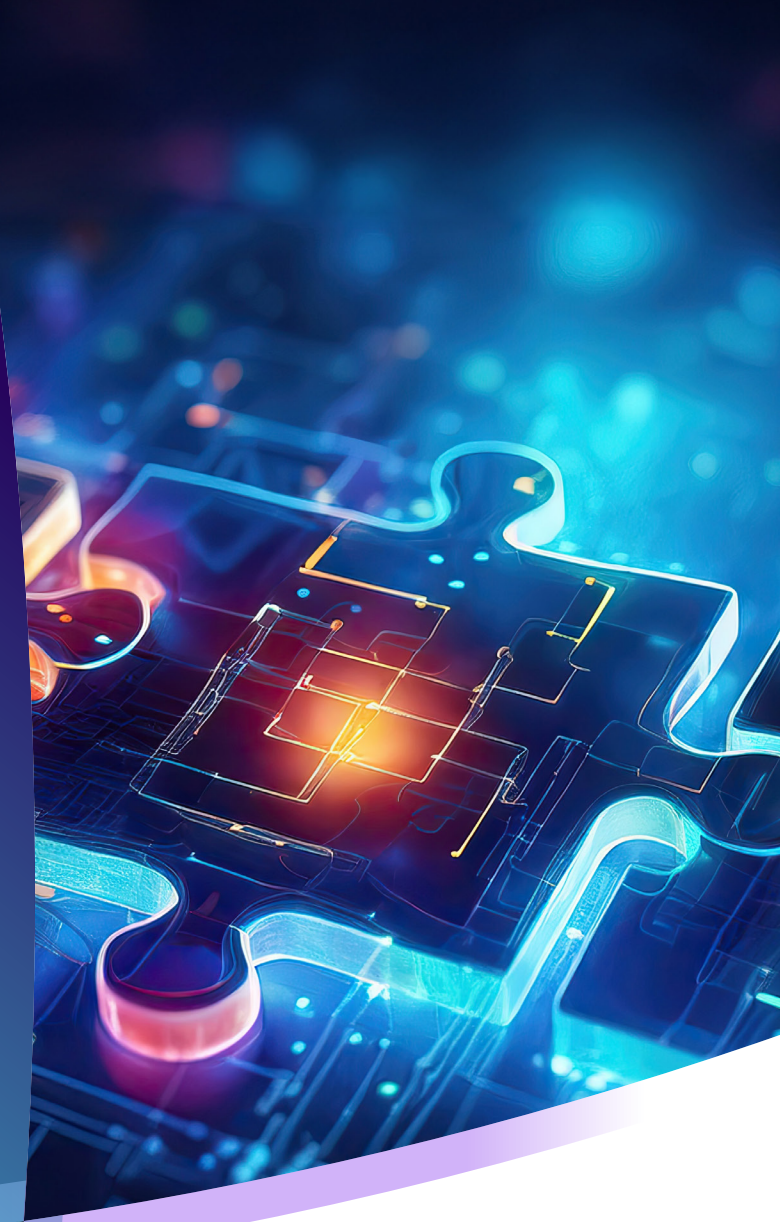
WHITEPAPER

Unlocking the GenAI data dilemma

Powering GenAI solutions with secure, compliant, and up-to-date data sharing



Vendia

The AI revolution promises unprecedented opportunities, but its realization hinges on a critical challenge: harnessing the power of dispersed data. From revolutionizing customer experiences with GenAI-powered chatbots to automating routine tasks for business professionals, AI is transforming industries. However, these advancements are predicated on access to vast amounts of data.

Today, data is scattered across public clouds, SaaS platforms, and third-party systems, posing significant challenges for AI initiatives. Integrating data from diverse sources, ensuring data quality, and maintaining security and compliance are all paramount concerns.

To unlock the full potential of AI, organizations must prioritize modern data automation solutions. By effectively accessing, integrating, reconciling, and transforming data, companies can build robust AI models that deliver exceptional value. This whitepaper delves into the best practices for powering GenAI-based solutions, exploring how effective data-sharing platforms are essential for AI success.

Enhancing GenAI's potential:
From LLMs to RAG

Large language models (LLMs) are the foundational models powering GenAI applications, enabling natural language interactions and basic question answering. However, many of the data sources underpinning these baseline LLMs present significant limitations when applied to enterprise environments.

For starters, LLMs are trained on **publicly available** and **static data** like books, dictionaries, and encyclopedias. Business data is different: it's both **dynamic** and **confidential**, constantly evolving with new customers, financial transactions, and operational changes. This makes it difficult to use a basic LLM for targeted business needs—not only does it lack the necessary proprietary information, but it also can't adapt to the fast-moving nature of business operations.

LLMs are also designed for **general-purpose use**, lacking the depth of knowledge required for specific industries, companies, or even individual customers. In other words, they lack critical business details about nearly everything a company would be interested in, such as inventory levels, delivery statuses, financial transactions, and so forth.

While this genericity can be addressed by augmenting LLMs with company-specific data, the other two problems are structural in nature. LLMs struggle to tailor responses based on who's asking the question AND often rely on outdated information.

These discrepancies become significant hurdles in dynamic business applications such as banking, where data privacy and real-time accuracy are paramount. Advancements in LLM technology are ongoing, yet achieving a comprehensive understanding of every conceivable business domain remains a distant goal.

To bridge the gap between the capabilities of LLMs and the demands of enterprise applications, companies are turning to a best practice known as retrieval augmented generation (RAG). RAG addresses the challenges of data privacy, security, and timeliness by combining the strengths of LLMs with access to up-to-date, relevant information.

> **RAG is the process of optimizing the output of a large language model, so it references an authoritative knowledge base outside of its training data sources before generating a response…all without the need to retrain the model. It is a cost-effective approach to improving LLM output so it remains relevant, accurate, and useful…"**

aws

The following image depicts the basic structure of the RAG pattern:

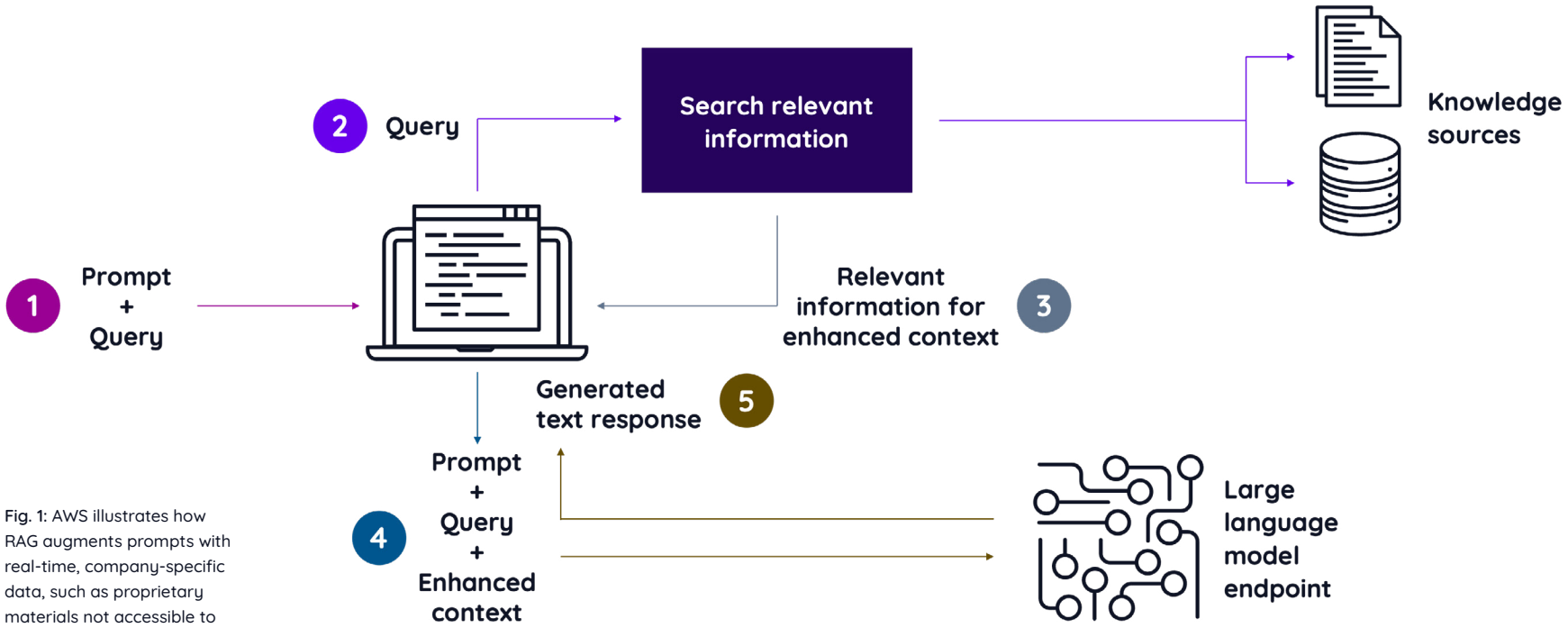


Fig. 1: AWS illustrates how RAG augments prompts with real-time, company-specific data, such as proprietary materials not accessible to the LLM (Source: AWS)

Understanding the data landscape for GenAI

Before delving into the specifics of how RAG uses data, let's look at a simplified view of the "data spectrum" inside an organization as it relates to GenAI. Traditionally, IT and data professionals categorize data into operational and analytical buckets, with additional layers of security classification. However, GenAI introduces a dynamic interplay between these dimensions. For instance, operational data, crucial for predictive models, often carries significant sensitivity.

Figure 2 illustrates this complex landscape. The lower left quadrant represents publicly accessible information that changes infrequently, such as dictionaries or classic literature. Moving rightward, we encounter publicly available data with usage restrictions, such as copyrighted information or trademarked materials.

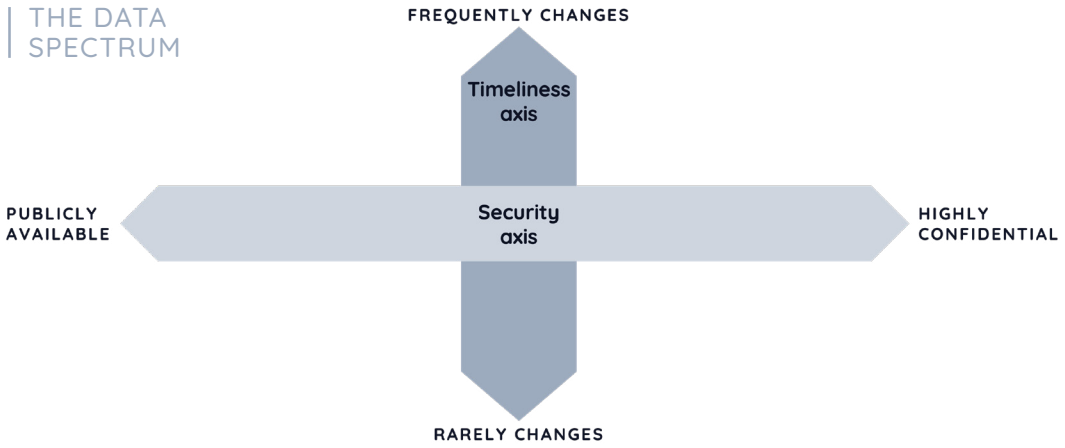


Fig. 2: The two dimensional data spectrum for powering GenAI. On the vertical axis is timeliness (i.e., how frequently the data changes). On the horizontal axis is confidentiality (i.e., how "sensitive" the data is with respect to which audiences are permitted to know it).

In contrast, the upper right quadrant is the "danger zone," containing highly sensitive and rapidly changing data. Examples of data in this quadrant include real-time financial information, personally identifiable information (PII), and protected health information (PHI). Exposing even a fragment of this data could lead to severe consequences, including financial loss, reputational damage, and customer churn.

Components of an effective RAG approach

While public chatbots like ChatGPT can limit themselves to publicly available data—i.e., data in the lower-left quadrant of the spectrum—corporate GenAI solutions require access to data across the entire spectrum. This is where a more comprehensive approach like RAG can help.

Most of the corporate data a company considers highly valuable lives entirely in the upper right quadrant. But in reality, useful data is spread across both axes. To gain access to all this corporate information, an effective RAG approach requires four distinct approaches to data:

1

Leverage a prebuilt LLM base layer

Leverage a prebuilt LLM base layer: Utilize a foundation model trained on a massive corpus of public data to handle general knowledge and language understanding.

2

Extend the base layer with business-specific data

Incorporate publicly available, but potentially copyrighted, data relevant to the company's sector to enrich the model's knowledge. This may include information such as instruction manuals for products, public company documentation, and industry or sector-specific glossaries, and so forth.

3

Utilize vector databases for frequently changing data

Store sensitive, yet essential, data in a secure vector database to optimize query performance and protect information. This category handles the "everyday" information in the enterprise, such as customer profiles, orders, inventory levels, and so forth. These information sources are populated from SaaS services, operational systems, and third-party data and transformed into vector format to improve GenAI query responses and lookup times.

4

Access real-time data on demand

Retrieve the most up-to-date and confidential information directly from source systems to ensure accuracy and security. Because this data can change in seconds—if not milliseconds—and needs to be accurate and/or remain tightly controlled and protected, the RAG approach needs to fetch it only on demand, and then only when authorized to do so. This may limit the ability of the RAG approach to apply the most powerful indexing or comprehension, but for this category of data, security and timeliness trump perfect vector optimization.

THE DATA SPECTRUM

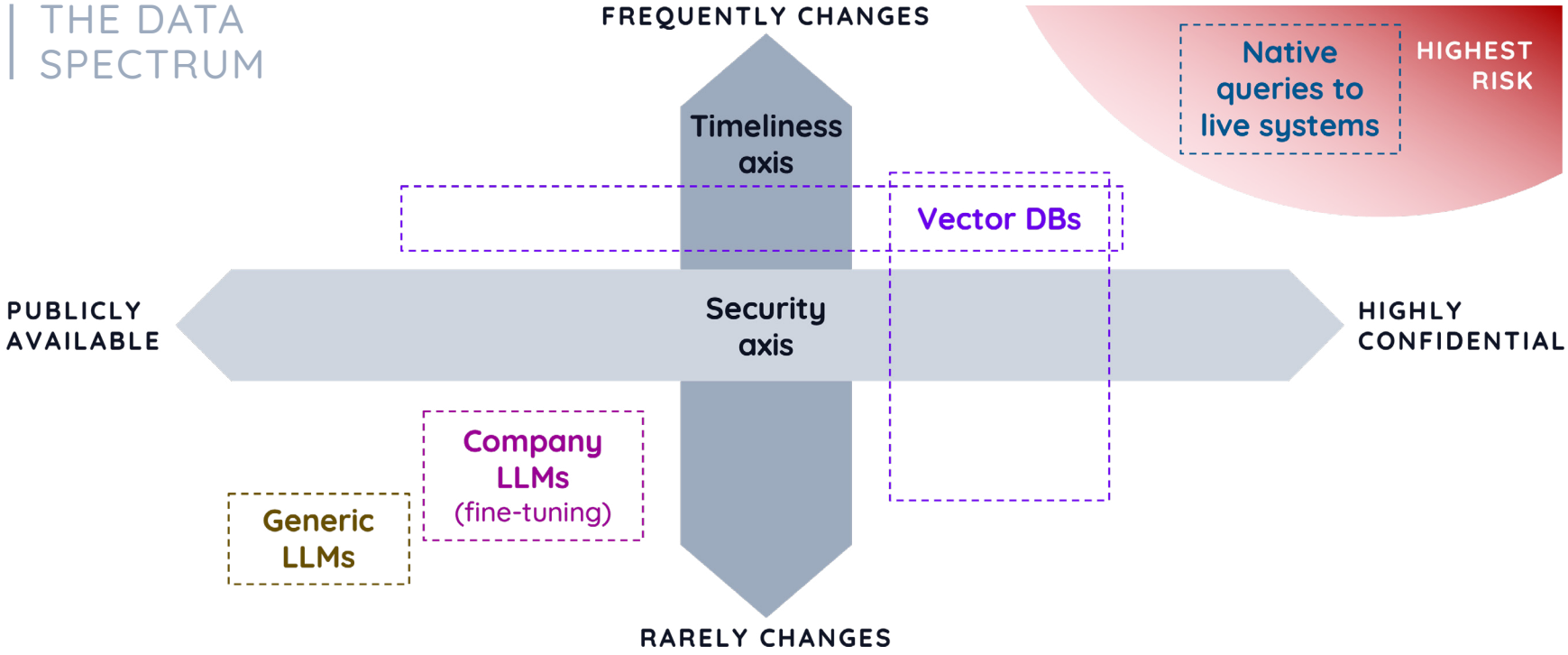


Fig. 3: How the RAG pattern deals with data in different arenas across the spectrum.

Public, static information: Resides in LLMs, including foundational models or a fine-tuned extension for the sector, industry, and company-specific knowledge base.

Operational data: Confidential and/or fast-changing in nature, this information is cached in a vector database for improved recall and indexing.

Sensitive, dynamic data: Accessed directly by RAG, living solely in its native representation and systems to ensure accuracy and security.

Effective data sharing for the RAG lifecycle

As an approach, RAG addresses several key limitations of traditional LLM-based GenAI by providing access to the full spectrum of corporate data while preserving security, privacy, and accuracy. Combining LLMs, vector databases, and direct data access, RAG empowers both employees and customers to harness the potential of natural language interactions for enhanced productivity and insights. Implementing RAG in practice, however, presents significant hurdles:

Continuously updating LLMs to align with the latest advancements is essential.

While foundational LLMs are trained on a largely static and unchanging corpus of material, they still represent an area of rapid innovation and improvement, meaning that they change frequently enough to require updating.

Maintaining and expanding the company-specific knowledge base requires ongoing effort.

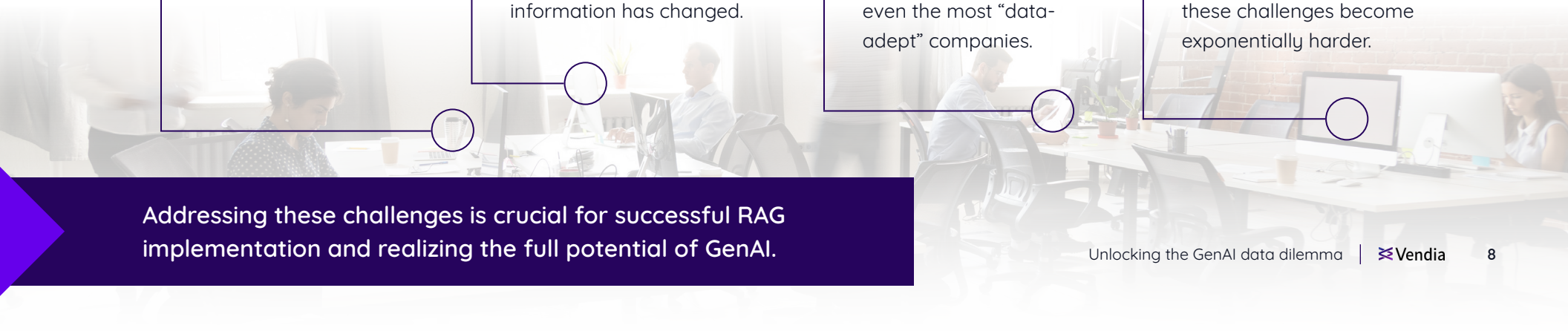
Company-specific extensions to LLMs may be based on public and largely "read-only" information, such as instruction manuals, documentation, and training materials. As a company's products and services evolve, this repository grows larger and additions need to be captured regardless of whether existing information has changed.

Integrating data from diverse sources into vector format is a complex undertaking.

Vector databases, including vector-style support in existing databases, already exist. However, populating these databases from the vast set of heterogeneous systems **internally**, and large array of business partners, data formats, and channels) **externally**, is a huge challenge for even the most "data-adept" companies.

Achieving seamless integration with sensitive, fast-changing data sources is challenging.

Natively accessing internal and external data sources for highly sensitive and/or fast-changing data requires a high degree of data integration, model reconciliation, security and governance, and real-time connectivity. In cases where this information must be retrieved from one or more third parties on the fly, these challenges become exponentially harder.



Addressing these challenges is crucial for successful RAG implementation and realizing the full potential of GenAI.

Overcoming RAG hurdles with modern data technology

Despite advances in existing GenAI infrastructure, including vector databases, these technologies lack the ability to directly access and integrate data from original source systems. Even the largest and technologically savvy enterprises struggle to develop an in-house data-sharing platform capable of handling multiple parties, clouds, regions, and SaaS systems.

Fortunately, modern data sharing platforms offer a solution to these complex data management challenges. GenAI may exacerbate the need for cross-organizational data collaboration, but the core issues—security, data localization, multi-party access control, governance, connectivity, and reconciliation—aren't unique to AI and remain common to a wide array of enterprise needs.

By effectively connecting disparate operational systems and securely integrating third-party data, data sharing platforms such as Vendia provide a competitive edge for RAG implementations.

These platforms also streamline regulatory compliance, reduce audit costs, and optimize resource utilization through data standardization and reduced dataset redundancy.

Data sharing technology: A critical GenAI enabler

How can a data platform be used to implement RAG faster and more efficiently? Building a robust RAG system is complex, requiring the integration of diverse data sources, efficient data management, and real-time access to information. Data platforms like Vendia emerge as a critical enabler, streamlining these processes and accelerating time-to-value. Here's how.

1. **Integrating diverse data sources.** Fine-tuning LLMs to achieve optimal performance requires access to a diverse and comprehensive dataset. This involves integrating a wide array of first- and third-party data sources, including content management systems, industry-specific materials, and proprietary information.

   For instance, a medical LLM may require specialized training on drug interactions, patient records, and clinical trial data, which are typically absent from generic foundational models.

   To address this, data sharing platforms prove invaluable in aggregating, standardizing, and enriching data from various sources, including SaaS applications, databases, and third-party content providers.

   I. **Data quality:** Ensuring data accuracy, consistency, and relevance is paramount for LLM effectiveness.

   II. **Data privacy and security:** Implementing robust measures to protect sensitive information is essential.

   III. **Data diversity:** A broad range of data sources helps prevent biases and enhances model generalizability.

   IV. **Data pre-processing:** Transforming raw data into a suitable format for LLM training is crucial.

2. **Vector database caching.** Vector database caching is the linchpin of RAG architectures: Storing a vast array of first- and third-party data in vectorized format dramatically enhances retrieval speed and accuracy. Data-sharing platforms connect to essential SaaS tools like HubSpot and Salesforce, ingesting critical business data such as customer profiles, orders, and sales metrics.

   Incorporating third-party data from suppliers, logistics providers, and financial partners is also crucial for comprehensive intelligence. This external data enriches the AI-based solution, enabling more accurate and informative outcomes.

   A robust data-sharing platform is essential for aggregating all of this data from various sources, including SaaS applications, cloud storage, third-party APIs, and SFTP depots. This consolidated data is then transformed into vectors and indexed in the vector database, delivering benefits such as:

   I. **Accelerated retrieval:** Vectorized data enables lightning-fast searches, significantly improving response times.

   II. **Cost efficiency:** Optimized data access and processing reduce computational overhead.

   III. **Enhanced accuracy:** By leveraging a rich dataset, the AI model can generate more relevant and informative outputs.

   IV. **Scalability:** The system can handle increasing data volumes and query loads.

3. **Real-time data access.** For highly sensitive, rapidly changing, or geographically dispersed data sources, live requests need to be performed on the fly. This real-time data access is indispensable for applications such as fraud detection, customer support, and financial trading. However, this is especially challenging for RAG applications, which may struggle with the diversity of data formats, channels, and protocols encountered in enterprise environments. Data-sharing platforms can help address these challenges, offering:

   I. **Unified connectivity:** Providing standardized connectors to a wide range of data sources, including cloud services, SaaS applications, databases, and APIs.

   II. **Data transformation:** Converting raw data into a format compatible with RAG systems, ensuring seamless integration.

   III. **Query optimization:** Translating complex queries into efficient data requests, minimizing latency and resource consumption.

By abstracting away the underlying complexities, data-sharing platforms enable organizations to access and leverage real-time data more effectively, improving the accuracy and relevance of RAG-generated outputs.

Safeguarding sensitive corporate data via RAG

While RAG-based systems offer exciting possibilities for leveraging AI, safeguarding sensitive corporate data remains a critical concern. How can organizations ensure that valuable information is protected while still extracting maximum value from GenAI?

Unlike LLMs, which lack the ability to implement fine-grained access controls and allow anyone to ask anything, RAG reinstates familiar role-based security measures that companies have employed for decades to control access to information and systems (Figure 4). By accessing data on demand and enforcing strict access controls at the data source level (i.e., governance "choke points"), RAG effectively mitigates the risk of unauthorized data exposure.

Figure 4 illustrates the three key flows for data integration that typically require combining both internal (i.e., first-party) and external (i.e., third-party) data in order to enable GenAI applications effectively. Each of the blue components illustrates a data integration point where data is required to flow from some system of record, such as a SaaS service or third party API, into RAG-related infrastructure.

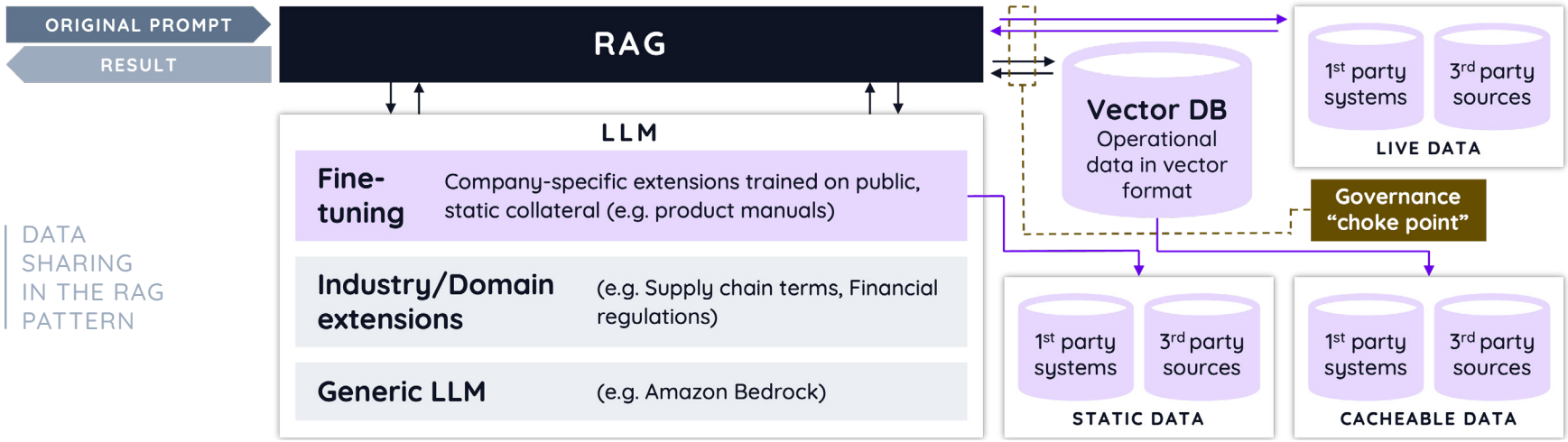


Fig. 4: RAG systems implement access control mechanisms similar to traditional applications, using roles or permissions to determine authorized query actions. Users, whether employees, customers, or systems, must have appropriate privileges to access the vector database or live operational systems. Without these permissions, data retrieval is blocked, and query results are restricted.

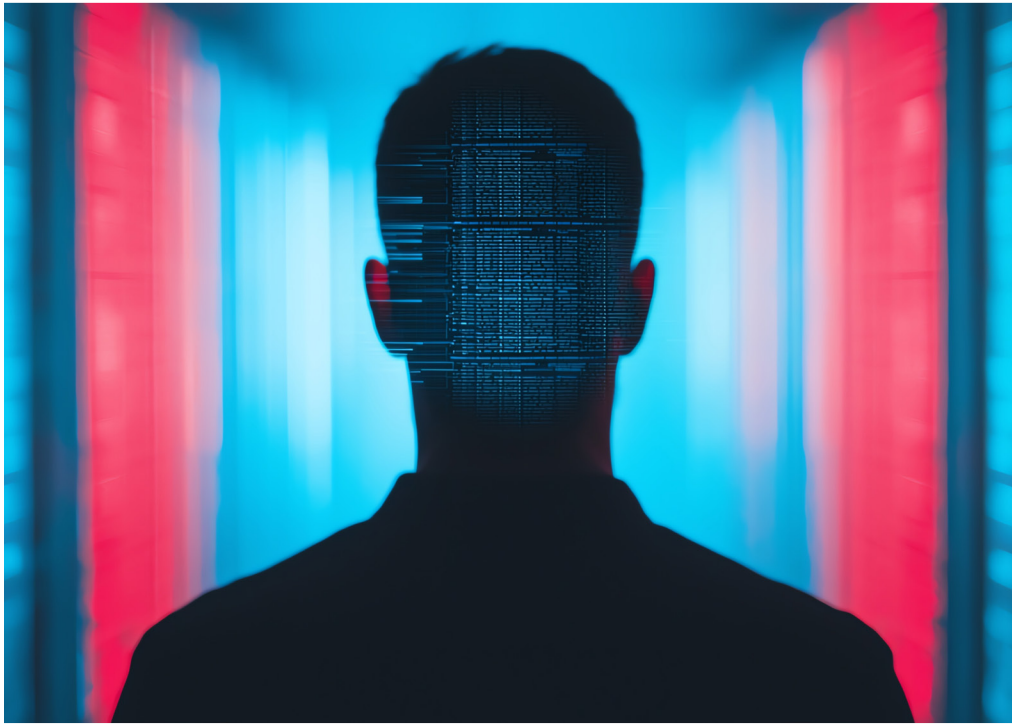
Managing growing data complexity with compliance sidecars

Beyond the foundational mechanisms of data integration and access, employing third-party data effectively within RAG applications requires robust metadata management. Compliance sidecars refer to key business details (i.e., active metadata) that travel with data across external systems, carrying information about its provenance, compliance, timeliness, and other important characteristics.

Modern data sharing platforms excel at capturing, exchanging, and managing both data and metadata, streamlining compliance efforts such as GDPR erasure. Vendors of sidecar-enabling platforms, such as Vendia, also let organizations add further metadata to sidecars to customize and enhance multi-party data sharing, automation, and auditing use cases.

In the context of RAG, sidecars empower business and analytics teams with the necessary information to interrogate results comprehensively, addressing questions such as:

1. Does this result contain PII or PHI?
2. Is this result subject to GDPR laws?
3. Where did this result come from?
4. How trustworthy is this answer?
5. Which other companies contributed to the results I'm seeing?



While building an in-house solution is feasible, the complexity and dynamic nature of metadata often necessitate a specialized platform capable of creating immutable records, such as those offered by distributed ledger technology. This ensures the highest level of data integrity and accountability, fostering trust and confidence in RAG outputs.

Any point within the RAG architecture that utilizes confidential third-party data necessitates meticulous metadata management. Datasets must be not only protected but also comprehensively documented, including origin, lineage, access permissions, and data structure. This metadata is essential for both internal compliance and seamless data sharing with external partners. Figure 6 depicts this “compliance sidecar” process flow.

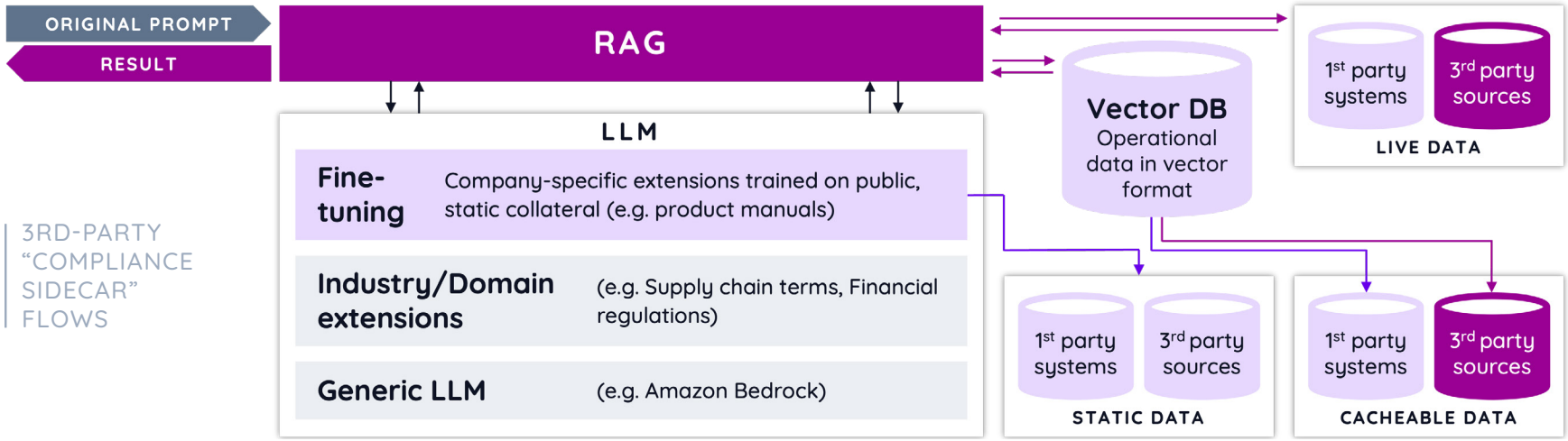


Fig. 5: Metadata-rich sidecars can be seamlessly integrated into the RAG pattern, empowering users to understand the origin, legal implications, and reliability of their results.

Fueling AI with an effective data-sharing foundation

AI, and GenAI in particular, are poised to transform businesses by solving complex challenges and unlocking new opportunities, helping organizations accelerate innovation, enhance customer experiences, and mitigate risks. However, realizing the full potential of AI requires a nuanced understanding of both its capabilities and limitations.

LLMs offer remarkable language generation abilities but can struggle with real-time accuracy and data privacy. RAG-style solutions, fueled by advanced data sharing technology, excel at protecting sensitive information while providing expanded access to relevant business data. A critical factor in successful AI implementation is understanding the two-dimensional nature of enterprise data: its sensitivity and rate of change.

Data sharing platforms provide a robust foundation for managing data complexity by facilitating fine-tuning, vector database population, and integration with live systems. By leveraging these platforms, organizations can accelerate time-to-market, reduce costs, and improve overall AI outcomes.

About Vendia

Vendia is the future of collective data intelligence, combining smart APIs, databases, and distributed ledger technology inside a single platform. Vendia's data automation cloud makes it easy to share data inside and outside of the organization in real time and with full visibility, governance, and control. Companies such as BMW, Delta Airlines, Resolution Life Insurance, and Fannie Mae use Vendia to automate contextual and compliant data flows between any-to-any systems for a harmonized, accurate view of data that unlocks speed, innovation, and cost savings. Learn more about us at Vendia.com and #UnchainYourData with Vendia.

Vendia