



GenAI Headaches: **The Cure for CIOs**

Pressure is mounting to create Generative AI solutions—but cautions abound. Here's how to meet your goals, while reducing your costs and risks.



Consumers' meteoric fascination for ChatGPT has led to a rapid demand for enterprise adoption. Whether to surface technical instructions, offer product advice, or provide step-by-step assistance, Generative AI (GenAI) has quickly been prioritized on CIOs' must-do lists.

CIOs are in the difficult position of figuring out the best ways to take advantage of the quickly developing tech while making sure to mitigate the risks. Here we take a look at the solutions to the headaches CIOs need to address when implementing GenAI so they can fully realize the tool's potential for their businesses.



Glossary of Terms

A list of some of the terms that will come up in this guide.



Auditable Prompts and Responses:

Refers to the fact that all prompts and responses generated by the AI model are tracked and recorded for auditing purposes.



Grounding Context:

Refers to the input data, information, or authorized document chunks, which are extracted from the index and used to guide the output of the gen AI model.



Retrieval Augmented Generation (RAG):

A feature of LLM systems that allows companies to use their own data to generate answers, making the LLM more cost-effective and enabling learning in context without costly adjustments.



Zero Retention:

A policy that ensures that data (queries/prompts) is not stored beyond a certain period of time nor used for training the LLM.





Privacy of Public Generative Engines

Risk: Public generative engines save and store chat history to further train their models, potentially exposing proprietary information.

ChatGPT's privacy policy reveals the collection of crucial user data, including IP addresses, browser information, and browsing activities across websites, which might be shared with third-party entities.

Solution

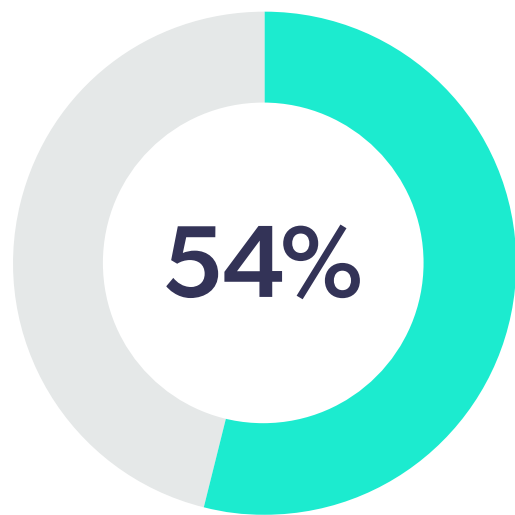


Banning a public tool is the easy option; but getting people to avoid those tools is hard, unless you offer a secure alternative.

A locked down environment will help ensure necessary governance against unauthorized access, use, disclosure, disruption, modification, or destruction of any enterprise data. And a secure environment must comply with regulations like AICPA, SOC 2 Type II, HIPAA, Cloud Security Alliance, and ISO 27001.

A **zero retention** approach is crucial, too. This includes having contractual agreements with LLM vendors to not train or retain any data longer than necessary and helps maintain user information privacy and security.

Within your perimeter, an individual's query is cached only for that person, preventing the questions they ask from becoming part of the model's training. Ensuring customer data is handled in a responsible and secure manner, it should not be kept by your LLM provider nor used for LLM training.



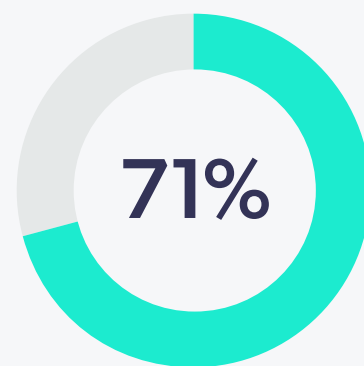
of companies have
not yet rolled out a policy
on the use of AI solutions.

– Coveo survey of 2,000 workers

2 ▶ Proprietary Enterprise Data

Risk: If enterprises want GenAI technology to provide answers about their specific products, services, and organization, that information needs to be indexed. But, building secure connectors to every complex system where your data lives is difficult, time-consuming, and expensive.

If not set properly, public GenAI platforms like ChatGPT store your data, prompts, documents, and everything else you send to them, for future learning. This is a dealbreaker for most enterprises because it leaves their confidential information vulnerable to data leaks.



of Senior IT leaders believe that GenAI could introduce new security risks.

– Coveo survey of 2,000 workers

Solution



To maintain your data's integrity, you need to retain complete ownership over it. Make sure that contractually you remain the sole owner of your information. This includes both your index and usage analytics. You have full visibility of the queries, clicks, and generative responses. Again, your data should not be used to train the model.

HTTPS ensures that communication between client and server is encrypted and secure, preventing potential attacks such as eavesdropping, tampering or data theft.

TLS endpoints use cryptographic protocols to provide authentication, confidentiality, integrity and non-repudiation services, allowing secure web communication between Coveo and Microsoft Azure's servers.

Providing an audit trail can also promote responsible and ethical use. **Auditable prompts and responses** allow organizations to understand how the AI model is being used to generate search results and to ensure that it is being used in a responsible and ethical manner.

3 Security of Generated Content

Risk: Without secure content retrieval built into the content index, you may risk generating answers from content that the user should not have access to.

Solution

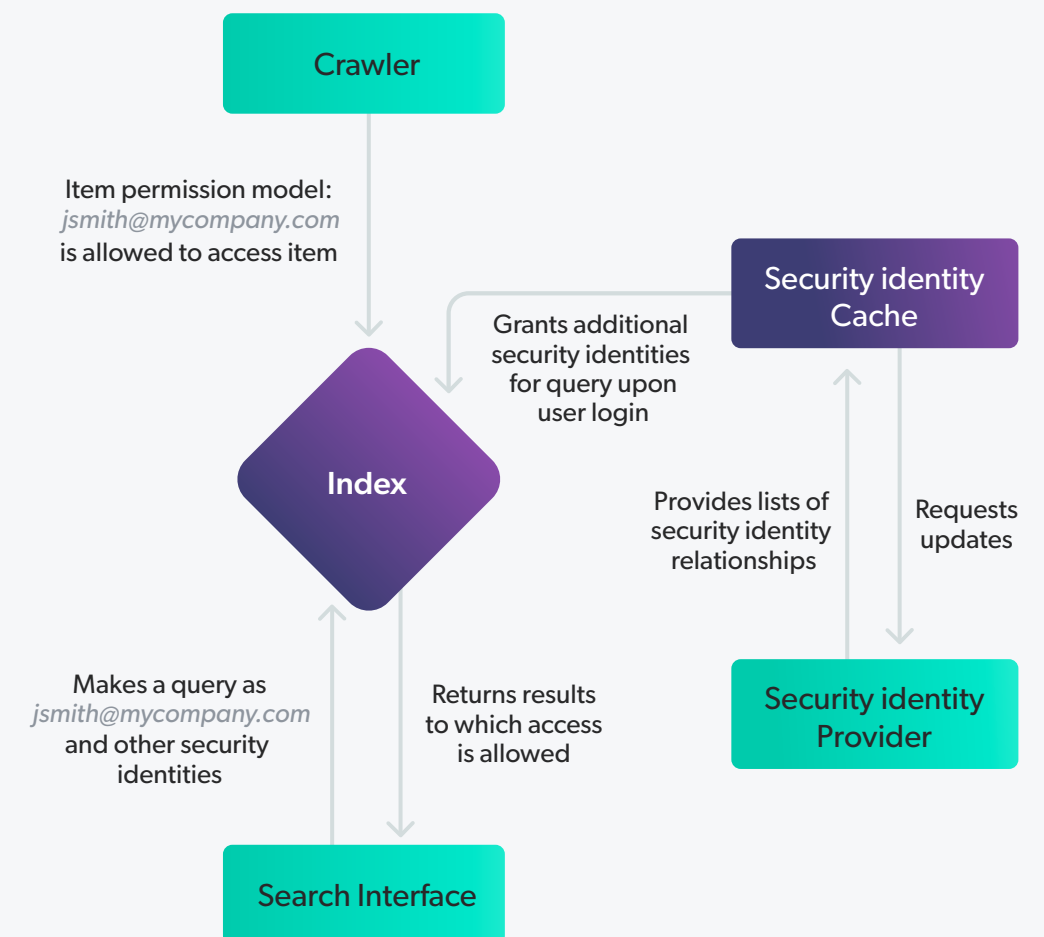


Your generative solution must consider each user's access rights and only generate answers from the content they have access to. Highly regulated industries and companies with PII, public safety, or technical IP all need to prioritize the security of generated content.

This requires secure content retrieval, with the security cache at its core, as it helps to ensure that the AI system's output is relevant, accurate, consistent, and secure.

At **crawling time**, user permission settings should be saved in order to determine which content can be accessed and made searchable, individually or as a group. This ensures that sensitive or confidential information is not inadvertently exposed through search results or generated responses.

At **query time**, the security cache should be used to handle permissions for each user accessing the search index. This ensures that users only see search results or generated responses that are relevant and accessible based on their specific permissions.



4 Multiple Content Sources

Risk: Lack of content reduces the quality and relevance of a generated answer.

Solution



By using a variety of content connectors feeding into a unified index, your organization can expand its knowledge base and improve the quality of Gen AI responses.

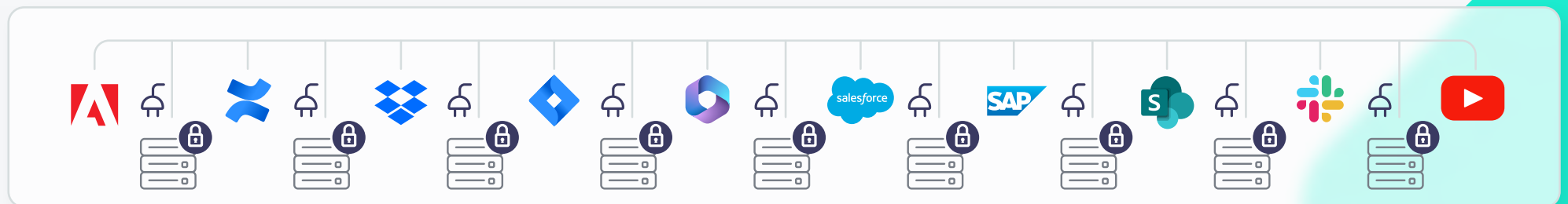
A staple of a unified search engine, a unified index centralizes access to all connected content repositories. This includes the option to crawl or push content stored in the cloud or on-prem via connector library or APIs—and ranks the entire corpus in a single user interface, applying relevance across the board to deliver meaningful results.

This index can be populated by many sources, each configured to regularly retrieve items from a specific content repository, possibly along with permissions and security identities to replicate a secured system.

In particular, this involves maximizing the segmentation of documents as part of the grounding context process before they are routed to the GenAI model. The LLM then provides a relevant response from these divided chunks of information, based specifically on your organization's knowledge.

This contextualization plays an essential role in the security of generative responses. It ensures that the results of the AI system are relevant, accurate, consistent and safe.

Content and Data Layer





Always Up-to-Date Content

Risk: Presenting customers or employees with out-of-date information can have major consequences, especially in highly regulated industries such as financial services and healthcare.

Solution



For a GenAI experience, the results should display the latest changes that were made to the content in your repository. Refresh and rescan operations ensure that the searchable content is up to date. For a great user experience, changes in your repository must be searchable and displayed in your search interface in near real-time after they've been saved.

There are **three update operations** you can use to keep your source content up to date. When you have the **required privileges**, you can schedule **refreshes** and **rescans**.

- ▶ **A refresh crawls a content repository and re-indexes any content that has been modified since the last source update.**
- ▶ **A rescan crawls a content repository, but only re-indexes new or updated content.**
- ▶ **A rebuild crawls and re-indexes an entire content repository.**

6 Accuracy at Scale

Risk: Generative AI often produces grammatically correct but factually incorrect text. This can be a significant risk for organizations and individuals as the output can seem highly coherent and convincing, leading to potential misinformation or incorrect conclusions.

Solution

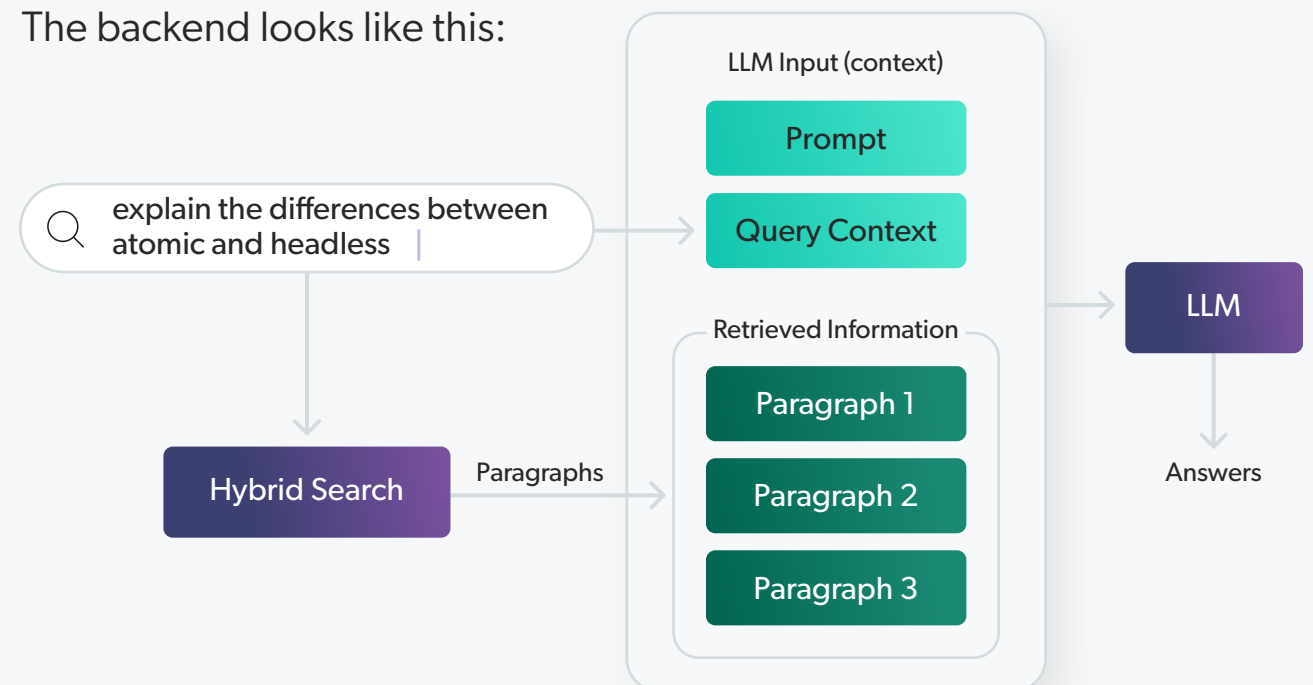


By using grounding context, you can ensure that the AI system's output—based on the quality of the content in your enterprise—is relevant, accurate, consistent and secure—at any scale.

Here's an example:

Let's say a developer is interested in understanding the differences between headless and atomic. They navigate to the Coveo Docs site—but instead of having to dig through multiple documents to get an idea of how these frameworks differ, they can just query: “explain the differences between atomic and headless”.

The backend looks like this:



Coveo parses the query and adds additional grounding context to ensure relevant chunks are returned from the index. This allows for a summarized query that can even be presented in bullet form for easy digestion.

7 Verifiability of Accuracy

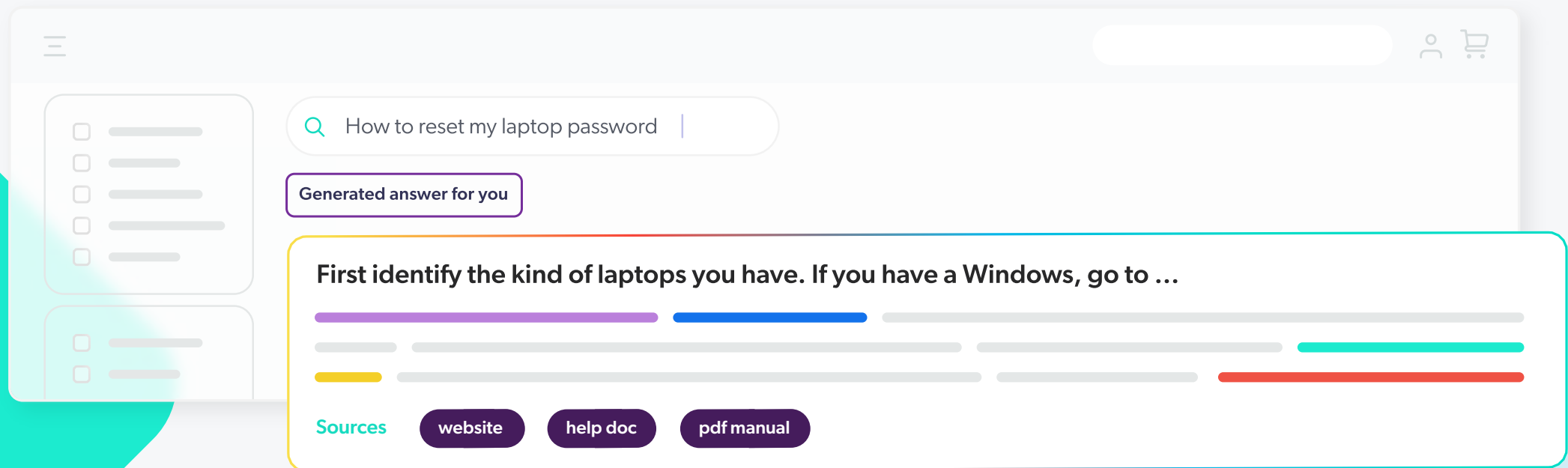
Risk: For businesses, inaccuracies can have significant consequences including customer trust erosion, as well as ethical and legal ramifications.

Solution



Ensuring the accuracy of training data and placing human intervention in key points to validate the model's outputs can help address this problem.

A best practice should be to provide links to the sources (citations) so the provenance of the data is available—and fully transparent. The sources are displayed directly under the generated answer.





Consistency Across Search and Chat Channels

Risk: If an enterprise is providing both chat experiences and search experiences (either through applications or standalone) customers should not be receiving different answers.

Solution



A unified index and headless integration into your front end applications will ensure this consistency of experience.

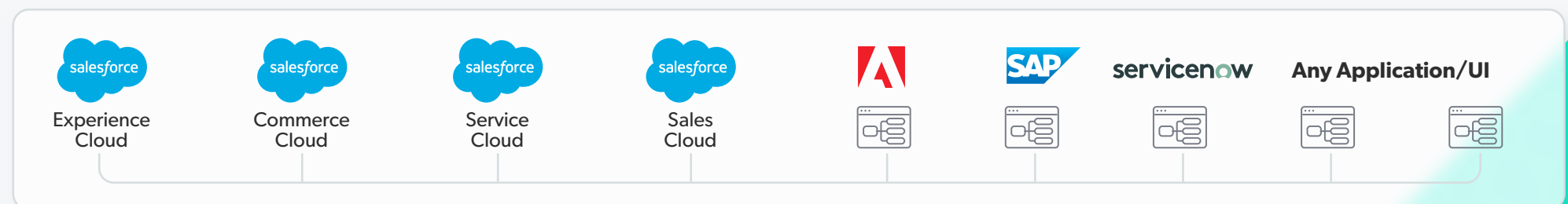
When your search relies on a unified search engine, relevance is applied across the entire corpus of content and can even self-optimize with the application of machine learning.

That means that results will always be ranked by relevance, regardless of the source.

When you deploy a **unified search engine**, a better user experience is guaranteed. Search results are blended and pages can be built from a component library or APIs. Sorting, filtering, faceting and other capabilities of the search UX are built on top of the unified index.

Coveo AI combines lexical, semantic and behavioral signals to understand user intent AND indexed documents to deliver the best experiences.

Engagement Apps Layers

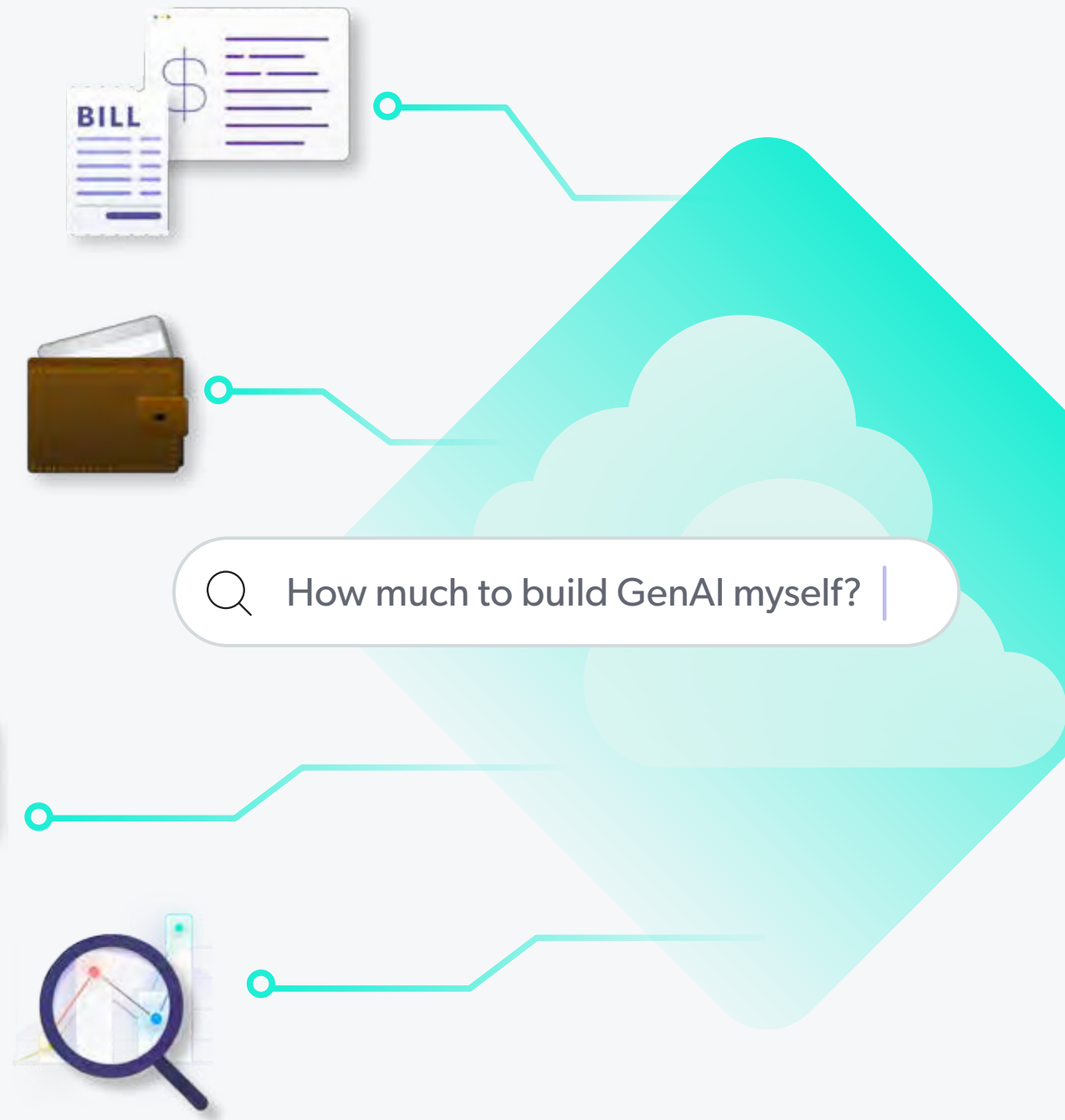


9 ▶ High Costs of GenAI

The promise of Generative AI has CIOs scrambling to create a sound, secure solution. But there are heavy build costs and ongoing development and maintenance required to produce a similar Generative Answering solution.

Aside from high costs, choosing to build in house significantly slows time to value. And, with resources tied up with building, you have less time and money to innovate.

Thankfully, there are tested and proven options on the market, backed by vendors who offer years of experience...





The Coveo Relevance Cloud™ offers

- 1 ▶ Relevant experiences **in any app across a digital journey** (engagement apps layer)
- 2 ▶ **In-session, personalized** search, navigation, content recommendations, chats and conversations (relevance intelligence layer)
- 3 ▶ **Secure, unified search that extends** across internal and external content sources (content and data layer)

Built for Enterprise **Safety**—and Performance



Secure Content Retrieval

Secure access to +100 cloud and on-premise content sources.

Structured and unstructured data with document-level security in the index.



Grounding Context

The LLM prompt context is grounded with passages from secure, personalized search results.

Helps to ensure that the output is relevant, accurate, consistent and secure.



Auditable Prompts and Responses

All prompts and responses generated by the model are tracked and recorded for auditing purposes.

Ensures the model is being used to generate answers in a responsible and ethical manner.



Zero Retention

Data (queries) is not stored by the LLM provider—nor used for training the LLM.

This helps to maintain the privacy and security of an organization's information.

Ensuring: Mitigation of Financial Security Risks + Best-in-Class Practices in Governance

Providing you with:

Cost-Effective at Scale

- ▶ Out-of-the-box relevance—including secured unified search and enrichment layer native connectors with end-to-end encryption, security cache, machine learning models
- ▶ Predictable subscriptions
- ▶ Covered hosting costs
- ▶ Low in-house maintenance cost

A Subscription to Innovation at Scale

- ▶ Without the worry of cost and complexity of building and maintaining in-house
- ▶ That leverages the expertise of 250 full-time data scientists, search experts and AI thought leaders

Accelerated Time to Value

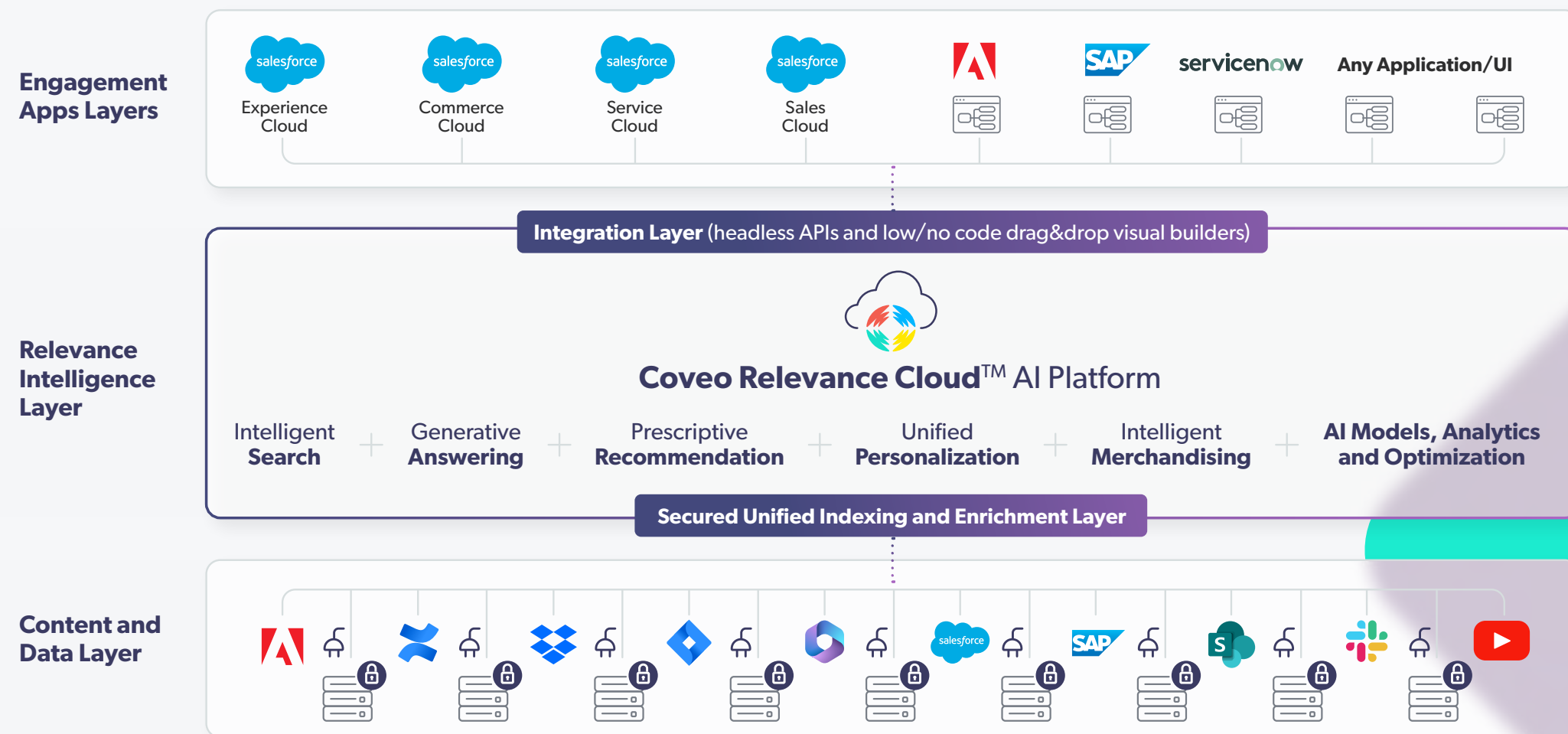
Partnering with Coveo provides businesses with:

- ▶ A reliable, scalable, and secure Generative Answering solution that delivers value from day one, enabling them to stay ahead of the competition and focus on their core operations.
- ▶ The ability for customers to tap into Coveo's upcoming and continuous GenAI releases and new capabilities e.g., website, workplace and commerce offerings.

Summary

GenAI is the next make-or-break moment for technology leaders. GenAI will bring significant leaps in employee proficiency and customer satisfaction.

Companies can't, and don't have to, solve for this alone. The Coveo Relevance Cloud™ is the leading, trusted AI platform that has helped global enterprises become AI proficient over the last 10 years — paving the way for the next evolution of AI — Generative AI.



Coveo can help CIOs harness GenAI for their organizations — while mitigating the risks outlined in this guide.

Coveo has been building enterprise-ready AI for the last 10 years

With over a decade of AI expertise and a proven track record of successful LLM deployments in large enterprises, Coveo continues to push the boundaries of innovation with its new enterprise-ready GenAI solution.

Factually accurate, contextually relevant question answering. Enterprise-grade security. Multiple content sources. All combined to bring you GenAI you can trust.

© 2023. All rights reserved.

All trademarks are the property of their respective owners.



Learn more about Coveo

Coveo, a leading provider of enterprise AI platforms that enable individualized, connected, and trusted digital experiences at scale with semantic search, AI recommendations, and GenAI answering.

Contact us



coveo.com

