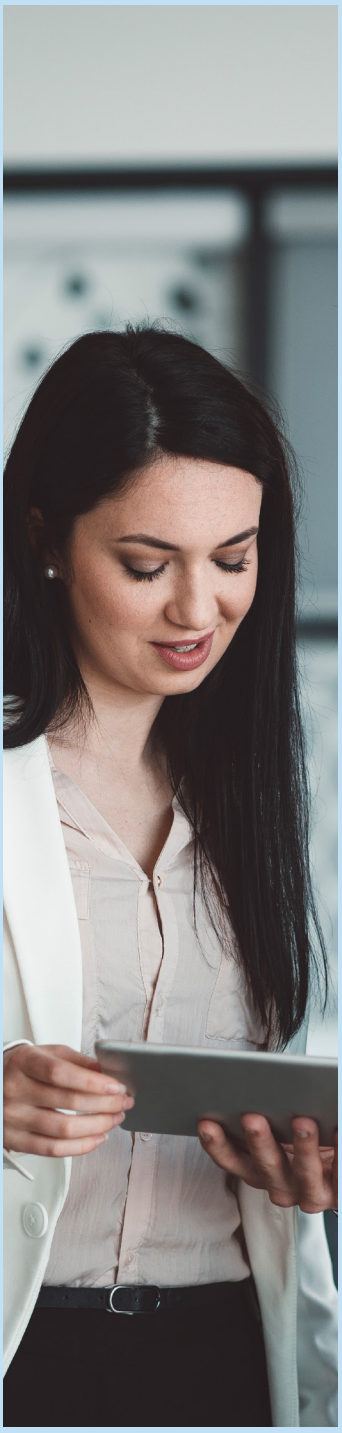
Infosys topaz

VIEW POINT



AI-DRIVEN INNOVATION: EMERGING DEMANDS ON CORTEX SEARCH IN ENTERPRISE AI

Infosys®
Navigate your next

Overview

In today's data-centric era, the fusion of artificial intelligence and advanced data platforms is reshaping industries. AI-powered tools are becoming indispensable for extracting meaningful insights from ever-growing data volumes, driving innovation, and enhancing decision-making processes. As enterprise systems grow more complex, the landscape of data and AI continues to expand rapidly, offering unprecedented opportunities for businesses to leverage their data assets for competitive advantage.

Traditional on-premises systems are now being modernized to cloud-based solutions and proliferation of data sources is fueling exponential growth in data consumption, driven by diverse consumers, platforms, and analytical needs. This evolution is not only fostering innovation in data consumption, sharing, and monetization but also presenting new opportunities and challenges for organizations.

As enterprise data grows, finding the required information has become increasingly challenging. Traditional systems often fail to effectively index and retrieve insights from unstructured data, resulting in inefficiencies. Additionally, achieving low-latency, high-performance search results over large data volumes has been a significant hurdle. These issues highlight the need for advanced search solutions that can handle the complexity and scale of modern enterprise data environments. Snowflake Cortex Search addresses these challenges by providing a robust, high-quality search service that ensures efficient and accurate data retrieval.

Snowflake Cortex Search is a fully managed service that enhances data discovery by providing high-quality, low-latency search capabilities over structured and unstructured data stored in Snowflake. It integrates seamlessly with Snowflake's ecosystem, offering advanced semantic and keyword search functionalities. This service is ideal for various applications, including AI chatbots and enterprise search, ensuring precise and relevant results. By surfacing the right data at the right time, Cortex Search significantly boosts operational efficiency and innovation. It supports diverse data sources, making it a versatile tool for modern enterprises.

Why do we need Cortex search for Enterprise AI

Search and retrieval Semantic search has evolved significantly over the years, aiming to improve search accuracy by understanding the searcher's intent and the contextual meaning of terms. It began in the with keyword search and statistical methods like TF-IDF. Then we saw the introduction of NLP techniques such as stemming and lemmatization. The Semantic Web concept introduced later aimed to make web data more meaningful and machine-readable. Innovations later on significantly advanced semantic search, including the launch of a major knowledge database, algorithm updates and the introduction of machine learning-based components. Today, semantic search continues to integrate advanced AI and machine learning to provide more relevant and accurate results.

The integration of large language models (LLMs) into enterprise tech stacks has highlighted the importance of robust search solutions. LLMs can enhance applications with its knowledge but require reliable search partners to ensure their outputs are relevant and trustworthy.

The adoption of the retrieval-augmented generation (RAG) stack for contextualized LLMs in enterprises exemplifies this need. Traditional keyword-based search systems are being enhanced by representation models, such as embeddings and cross-encoders, which significantly improve retrieval accuracy.

Snowflake Cortex search combines vector and keyword search to deliver precise and relevant results, supporting natural language queries. The service enhances large language models (LLMs) by providing contextualized, up-to-date data, crucial for applications like AI chatbots and enterprise search. Cortex Search integrates seamlessly with Snowflake, simplifying setup and maintenance with a single SQL statement. It ensures quick data retrieval, robust security, and compliance with industry regulations, making it a powerful tool for modern enterprises.

Key benefits of Snowflake Cortex Search include:

Fully Managed Infrastructure:

Frees customers from infrastructure and operations responsibilities, offering seamless, incremental ingestion with low-latency search results.

High-Quality Search:

Provides high-precision, high-recall ranked results out of the box, reducing the need for extensive search quality tuning.

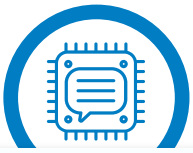
Security and Governance:

Maintains strong security and governance controls, ensuring that RAG applications are secure and governed like the rest of Snowflake data.

In this post, we will explore how Cortex Search is designed to handle RAG workloads and beyond, embodying this vision for enterprise search and retrieval.

Key traits of Cortex Search Offering

Snowflake Cortex Search offers a robust set of features designed to enhance data accessibility and analysis within the Snowflake Data Cloud. Here are some of its key traits:



Natural Language Queries:

Users can ask questions in plain English or their preferred language, eliminating the need for complex SQL syntax. Cortex Search understands the context of queries, providing more accurate and relevant results.

Seamless Integration:

It integrates seamlessly with Snowflake applications, making it a natural extension of existing workflows and reducing the need for switching between tools, thereby improving productivity.

Comprehensive Search:

Cortex Search searches across a wide range of data sources, including tables, views, files, PDFs, external stages and SharePoint. It can search for specific data points or patterns within larger datasets. Leveraging advanced algorithms, Cortex Search understands the context of queries and returns relevant results, reducing the likelihood of irrelevant or inaccurate search results.

Customization:

Users can tailor Cortex Search to their specific needs, defining search indexes and configuring search settings, allowing organizations to optimize search performance and results based on their unique data and use cases.

Governance and security:

It adheres to Snowflake's robust security measures to protect sensitive data and supports compliance with industry regulations and standards.

How Cortex Search works:

Cortex Search empowers everyone in the organization with a robust search engine. It automatically indexes and embeds data incrementally, processing only the changed rows from the underlying data source. This service abstracts the operational complexity of building a search engine into a single SQL statement for service creation. By eliminating the need to manage multiple processes for ingestion, embedding, and serving, it frees up valuable time for developing advanced AI applications.

Once the service is set up, querying it is straightforward via REST or Python APIs. This functionality is available for applications hosted within Snowflake (e.g., Streamlit in Snowflake) or in external environments.

Solution Pattern 1

Cortex Search combines both vector and keyword search capabilities. This allows it to handle a wide range of search queries effectively, from simple keyword searches to more complex semantic searches. One of the primary use cases for Cortex Search is as a retrieval engine for RAG applications. This means it can retrieve relevant data from a knowledge base to enhance the responses generated by large language models (LLMs). This is particularly useful for creating chatbots that provide contextualized and accurate responses based on up-to-date data. Once a Cortex Search Service is created, a REST API endpoint is provisioned and can query this endpoint to get search results, making it easy to integrate with other applications.

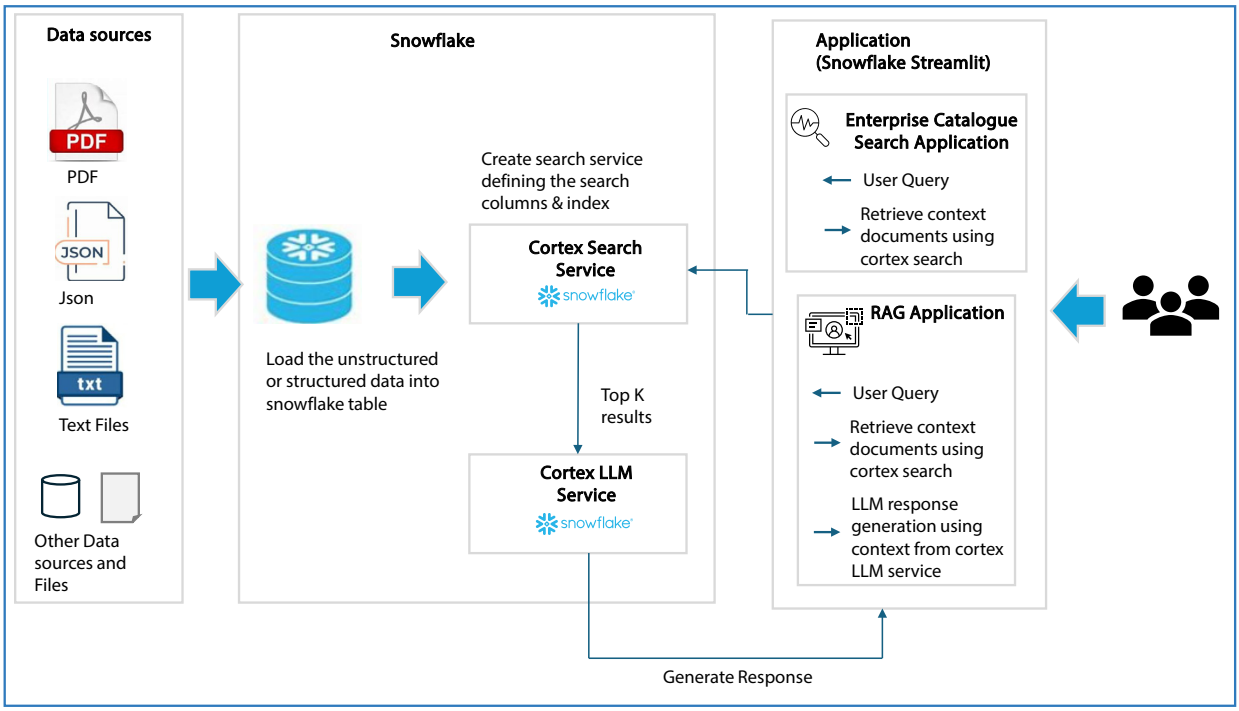


Figure 1: Solution Pattern 1

Solution Pattern 2

The Snowflake Connector for SharePoint connects the Microsoft 365 SharePoint site with Snowflake. It ingests files and user permissions from SharePoint into Snowflake, keeping them up to date. After installing and configuring the connector, it starts ingesting content from SharePoint. It can be specified whether to ingest files from all folders or specific folders within the SharePoint site. Once the content is ingested, Cortex Search service can be used to query these documents. This allows to build chat and search applications that can interact with or query the SharePoint documents. The responses from the Cortex Search service can be filtered to restrict results to documents that a specific user has access to in SharePoint. This ensures that search results adhere to the access controls specified in the SharePoint. The Cortex Search service provides a REST API endpoint that can be queried to get search results. This makes it easy to integrate with other applications and use the ingested SharePoint content in various ways.

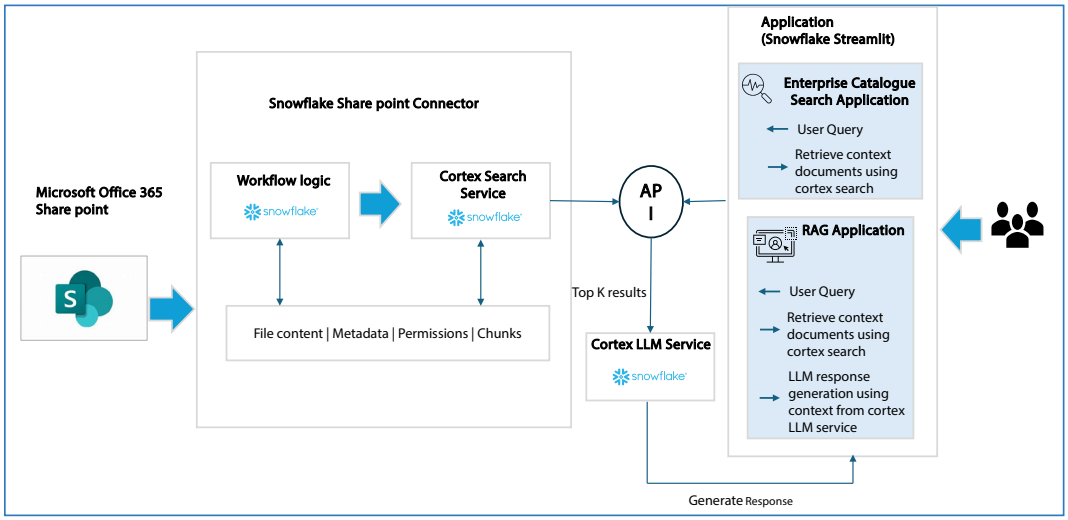


Figure 2: Solution Pattern 2

Solution Pattern 3

Third-party knowledge extensions available in the Snowflake Marketplace can be integrated with existing Snowflake account. These extensions provide access to various external data sources, such as news articles, research publications, and more. Once integrated, the data from these third-party sources is ingested into the Snowflake environment. This allows the user to leverage this external data alongside with existing datasets. With the data ingested, the Cortex Search service can be used to query this information. Cortex Search combines both vector and keyword search capabilities, enables to perform both simple and complex searches over the ingested data. One of the primary use cases is using Cortex Search as a retrieval engine for RAG applications. This means relevant data can be retrieved from the third-party sources to enhance the responses generated by large language models (LLMs). This is particularly useful for creating chatbots that provide contextualized and accurate responses based on up-to-date information. Setting up and using these knowledge extensions is straightforward. Snowflake provides a user-friendly interface and REST API endpoints to facilitate the integration and querying processes.

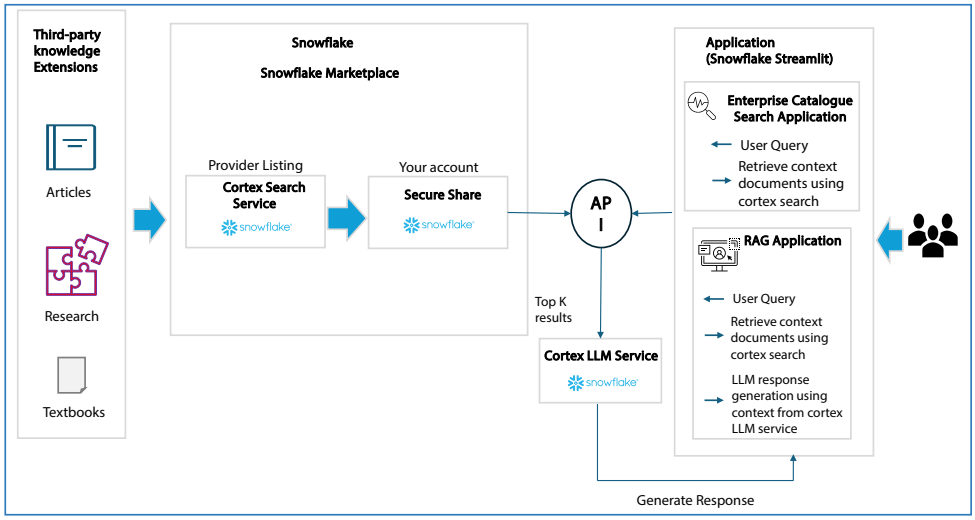
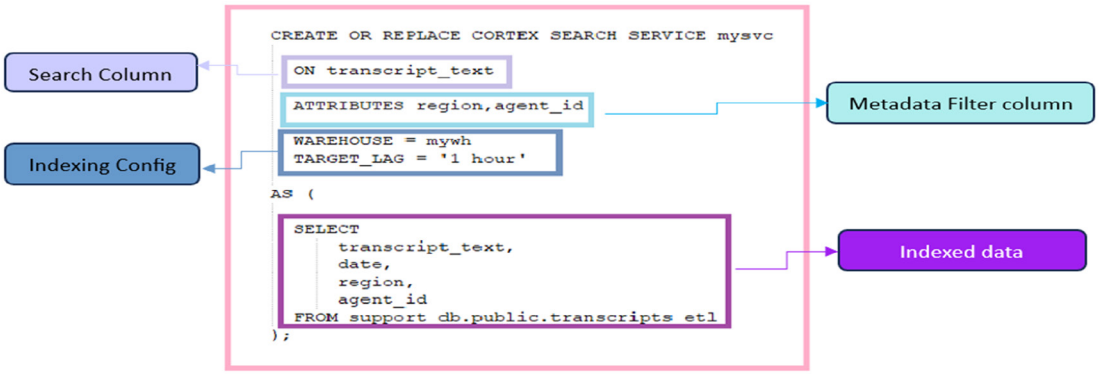


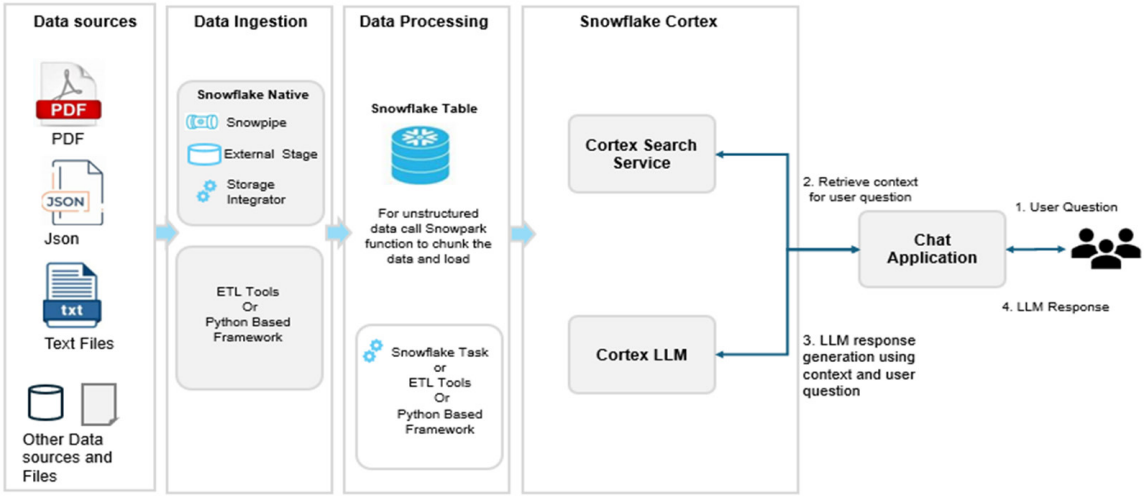
Figure 3: Solution Pattern 3

Cortex search implementation using Intelligence Assistant on enterprise data



Features	Description
Search Column	Specifies the text column in the base table that you wish to search on. This column must be a text value
Metadata Filter Column	Specifies comma-separated list of columns in the base table that you wish to filter on when issuing queries to the service. Attribute columns must be included in the source query, either via explicit enumeration or wildcard, (*).
Index Configuration	Specifies the warehouse to use for running the source query, building the search index, and keeping it refreshed per the TARGET_LAG target.
Indexed Data	Define the indexed data source declaratively with a SQL query.

End to End integration for Cortex Search:



Data sources:

Includes structured and unstructured data

Data ingestion:

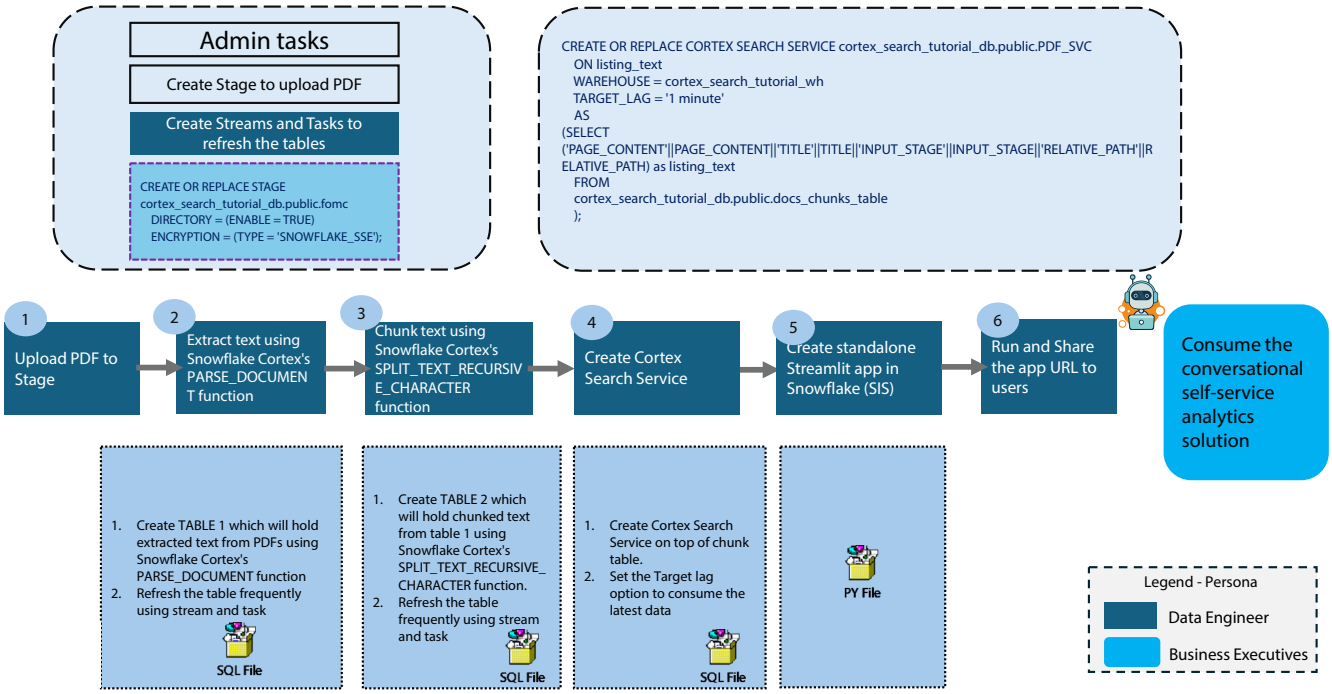
ingest the data into snowflake table using snowflake native solutions or any ETL tool or custom framework

Data processing:

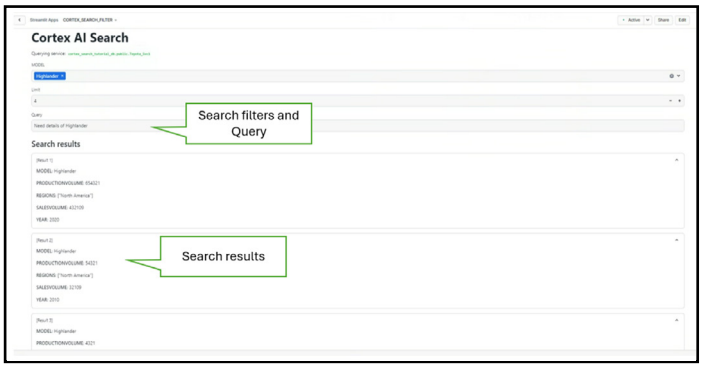
Preprocess the data, in case of unstructured data split it into chunks and load the snowflake table using native snowflake functions. Use snowflake task or custom tools to continuously ingest the data.

Cortex & Consumption:

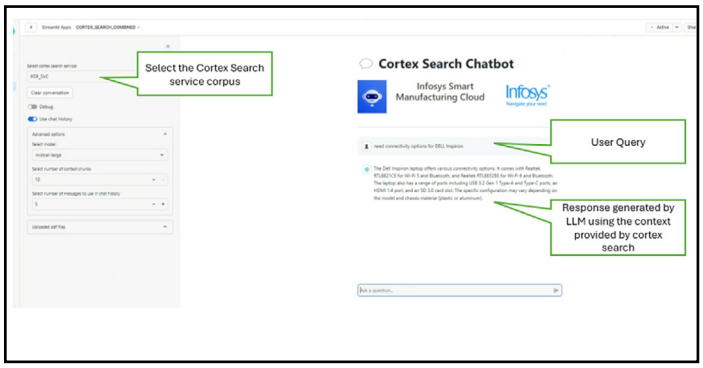
Application queries snowflake search service to get the context data and sent it to LLM for comprehensive response



Cortex Seach query executed using streamlit app



RAG application using cortex search



Our learnings in implementing solutions using Cortex Search Service assistants

Out of the document questions may lead cortex search to pass the whole chunk into LLM without filtering. Hence LLM must response based on the user query.

Cortex search service needs to be recreated whenever base table is re-created.

Why do we need fully managed Agentic solution like Snowflake Cortex Search

Using cloud-based solutions like Snowflake Cortex offer managed services for NLP, machine learning, and infrastructure for the solutions. This provides benefits of scalability and reduce TCO of AI agents.

Customer data resides within the customers Snowflake environment and doesn't get transferred for LLMs

This feature helps to build solutions that improve Human-AI Collaboration and augment human resource potential in organizations

Use cases that would benefit from Cortex Search

Cortex Search, with its natural language query capabilities and integration with Snowflake's Data Cloud, can be highly beneficial across various industries and use cases. Cortex Search can be used to improve efficiency, decision-making, and innovation across various industries. The tool's ability to understand natural language and access a wide range of data sources makes it an asset for organizations seeking to unlock the potential of their data.

Here are some examples:

- Customer Service Chatbot
- Human Resources policy retrieval
- Healthcare Benefits and claim search
- Supply Chain Management for enhanced visibility and Data retrieval
- Marketing and Sales Summary
- Research and Development

Authors

Carlos Carrero

Global Principal Architect - AI/ML - Partners, Snowflake

Dharmendra Shavkani

Principal Data Cloud Architect, Snowflake

Balaji Ramanujam

Distinguished Technologist, Data analytics and AI, Infosys

Dhanus Kodi Paulraj

Principal Technology architect, Data analytics and AI, Infosys

Jagan Kumar Krishna Pillai

Digital solutions specialist, Data analytics and AI, Infosys

For more information, contact askus@infosys.com

Infosys® Navigate your next