AI & ML FEB 12, 2025 | 10 MIN READ

Your Enterprise Data Needs an Agent

Authors

Harshal Pimpalkhute

Arun Agarwal

data insights by orchestrating across structured and unstructured datasets. Cortex Agents streamlines agentic application data access and orchestration for more reliable AI-driven decisions by building on top of enhancements to our **Cortex AI** retrieval services:

- Cortex Analyst, now generally available with Anthropic Claude as a key LLM powering agentic text-to-SQL for high-quality structured data retrieval
- Cortex Search has achieved state-of-the-art quality unstructured data retrieval accuracy, beating OpenAI embedding models by at least 12% across a diverse set of benchmarks including (NDCG@10)

AI agents, autonomous systems that perform tasks using AI, can enhance business productivity by handling complex, multi-step operations in minutes. Agents need to access an organization's ever-growing structured and unstructured data to be effective and reliable. As data connections expand, managing access controls and efficiently retrieving accurate information—while maintaining strict privacy protocols—becomes increasingly complex.

Agentic outputs are only as good as the quality of the underlying data and the accuracy of the retrieval systems that help ground them. Yet organizations struggle to pave a path to production due to an AI and data mismatch. LLMs excel at unstructured data, but many organizations lack mature preparation practices for this type of data; meanwhile, structured data is better managed, but challenges remain in enabling LLMs to understand rows and columns.

> **“At Luminate, we're revolutionizing how we deliver precise, data-driven insights to our clients through generative AI applications. Snowflake's unified data and AI platform provides our developer team with scalable processing and retrieval for both structured and unstructured data — the critical building blocks for developing, deploying and orchestrating data agents powering our applications. Using Snowflake Cortex AI brings advanced AI within the same security and governance perimeter as our data and saves us countless development hours, allowing us to unlock the full potential of entertainment industry data with agentic AI.”**

Snowflake customers now have a unified platform for processing and retrieval of both structured and unstructured data with high accuracy out-of-the-box. End-to-end unified governance, from ingestion to application, enables teams to deliver a new wave of data agents. Customers can build scalable solutions while enforcing access and privacy controls.

The need for data agents

At Snowflake, we believe that AI agents will soon be essential to the enterprise workforce, enhancing productivity for teams across customer support, field technicians, analytics, engineering and more. They will free up valuable employee time to focus on higher-value challenges facing the business. Data agents, a specialized category of AI agents, combine data and tools to deliver more accurate, grounded insights by effectively selecting the right data sources and tools for retrieval.

For AI agents to work at scale, they need secure connection with enterprise data and unified governance to manage their access, similar to existing controls for your teams. They must follow data policies, access multiple sources efficiently, and retrieve accurate information to deliver reliable, high-value outcomes.

However, we understand that this agentic future has challenges proportional to its potential. While model quality increases and inference costs decrease, we see the same set of challenges among companies trying to deploy trustworthy agentic systems at scale:

- **Accuracy:** In terms of quality, there is a high bar for agentic output in enterprise apps; the margin for error is low, especially in business-critical functions like finance or engineering.
- **Trust and security:** As customers build more data-intensive AI applications, meeting security and governance policies is increasingly challenging.
- **Governed data access:** Agents need access to a wide variety of data sources so they can operate reliably on business context, including both unstructured (e.g., text, audio) and structured (e.g., tables, views) data sources — which are often spread across multiple systems.

The key to scaling agentic workflows that tap into data is the seamless interaction between models and data while maintaining accuracy, trust and compliance. For example, a financial analyst may need to combine revenue data (structured) with financial reports and market research (unstructured). These enterprise use cases need secure access to data and a way to surface the right information to AI with end-to-end governance.

Cortex Agents, now available in public preview, orchestrates across structured and unstructured data sources — whether it be Snowflake tables or PDF files stored in object storage — to deliver insights. They break down complex queries, retrieve relevant data and generate precise answers, using Cortex Search, Cortex Analyst and LLMs. This enables accuracy, efficiency and governance at every step.

What are Cortex Agents?

Cortex Agents plan tasks, use tools to execute them, and reflect on results to improve responses. Available as a convenient REST API, Cortex Agents can seamlessly integrate into any application. Agents use Cortex Analyst (structured SQL) and Cortex Search (unstructured data) as tools, along with LLMs, to analyze and generate answers. The workflow involves four key components:

1. Planning: Applications often switch between processing data from structured and unstructured sources. For example, consider a conversational app designed to answer user queries. A business user may first ask for top distributors by revenue (structured) and then switch to inquiring about a contract (unstructured). Cortex Agents can parse a request to orchestrate a plan and arrive at a response:

- **Explore options:** When the user poses an ambiguous question (e.g., "Tell me about Acme Supplies"), the agent considers different permutations — products, location, or sales personnel — to disambiguate and improve accuracy.
- **Split into subtasks:** Cortex Agents can split a task or request (e.g., “What are the differences between contract terms for Acme Supplies and Acme Stationery?”) into multiple parts for a more precise response.
- **Route across tools:** The agent selects a tool — Cortex Analyst, Cortex Search or SQL generation from natural language — to facilitate governed access and enable compliance with enterprise policies.

2. Tool use: With a plan in place, the agent can retrieve data efficiently. Cortex Search extracts insights from unstructured sources, while Cortex Analyst generates SQL to process structured data. A comprehensive support for tool identification and tool execution enables the delivery of sophisticated applications grounded in enterprise data.

4. Monitor and iterate: After deployment, customers can track metrics, analyze performance and refine behavior for continuous improvements. On the client application developers can use TruLens to monitor the Agent interaction. By continuously monitoring and refining governance controls, enterprises can confidently scale AI agents while maintaining security and compliance.

Combined with other Snowflake offerings, Cortex Agents now provide an end to end solution for retrieving, processing and governing both structured and unstructured data at scale.

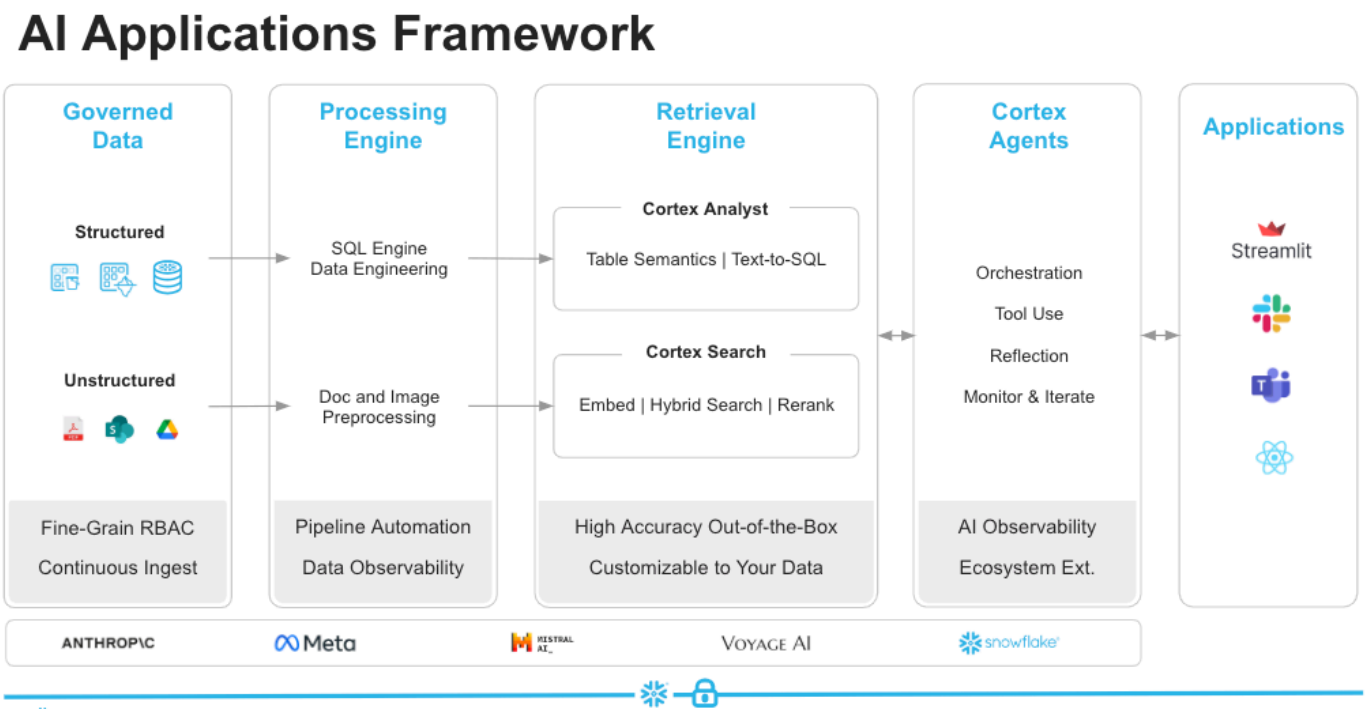

Figure 1: Snowflake enables an end-to-end solution for agentic applications

Now, let's explore how Cortex Analyst enhances structured data analysis and the latest innovations improving its capabilities.

Cortex Analyst: AI-powered SQL generation, with semantic understanding

Cortex Analyst can be used as a tool within Cortex Agents.

Unlike typical text-to-SQL systems that rely only on pattern matching, Cortex Analyst uses a semantic model to map business terms to underlying data. This unique approach improves precision in real-world use cases that involve complex multi-table environments.

advanced JOIN validation mitigates common issues, such as JOIN hallucinations and double counting, which often arise in complex queries. This allows Cortex Analyst to support multi-table queries without compromising precision.

2. Semantic model generation and monitoring

Our public preview of the new Analyst Admin UI in Snowsight simplifies the process of building and refining semantic models. Admins can select tables and columns, and use LLMs (running within Snowflake's secure perimeter) to generate a starting Semantic Model YAML file.

The admin interface also monitors user engagement and feedback. This allows customers to track usage, and make informed improvements to semantic models over time.

3. Customization for business-specific logic

With Custom Instructions now in GA, users can tailor Cortex Analyst to their unique business needs using natural language in the Semantic Model file. Common use cases include specifying fiscal year start dates, explaining internal naming conventions and prioritizing key tables during SQL generation.

4. Proven performance on benchmarks

Based on internal benchmarks, we have achieved 90% accuracy for text-to-SQL use cases. With Anthropic's Claude 3.5 Sonnet, we are able to further enhance the performance for improved experience. Cortex Analyst, running on Claude, outperforms other models on real-world queries by using information stored in the semantic model.

With these updates, Cortex Analyst enhances structured data analysis and simplifies admin setup for agentic applications.

Cortex Search: High-quality context engine for unstructured data

Cortex Agents use Cortex Search to retrieve unstructured data (e.g., text, audio, image, video). Cortex Search is a natively hybrid search, a combination of vector and lexical (keyword) search, with an additional semantic reranking step, to deliver high-quality, low-latency retrieval at scale.

OpenSearch, ElasticSearch), using both hybrid search with OpenAI's Text Embedding 3 Large, as well as keyword-only search.

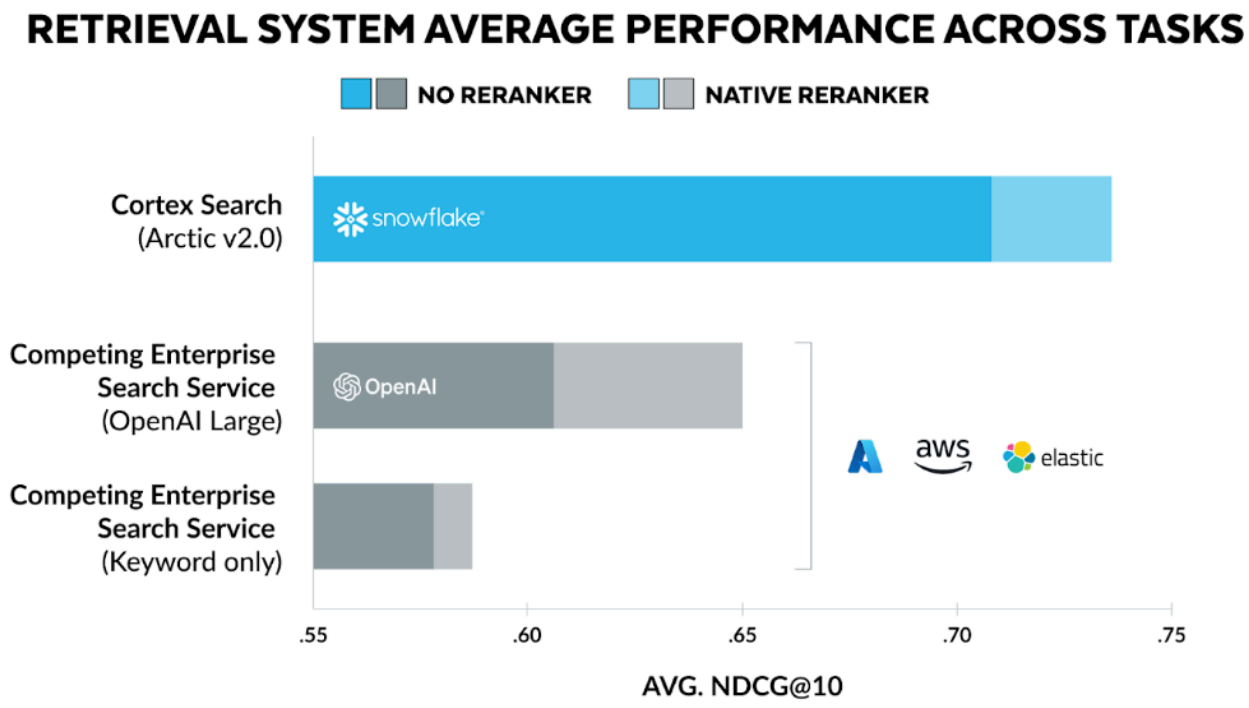

Figure 2: Cortex Search benchmark results

What's new with Cortex Search?

1. Increased scale and affordability

Cortex Search now supports indexing hundreds of millions of rows. Additionally, serving costs for Cortex Search have been reduced by 30% as a result of infrastructure optimizations.

2. Improved customizability

Cortex Search now provides the ability to select the vector embedding model for semantic search. This includes two multilingual models, snowflake-arctic-embed-l-v2.0 and voyage-multilingual-2. Additionally, Cortex Search supports date-range filtering on metadata columns.

3. New preview features

and agentic applications built on Snowflake data.

Anthropic models: The SOTA models powering Cortex Agents

Anthropic's most intelligent LLM, the updated Claude 3.5 Sonnet, runs inside Snowflake to bring advanced reasoning, coding and complex workflow execution to enterprise applications using the same governance framework as the data. This model now powers Cortex Analyst, and can be used in Cortex Agents, resulting in more accurate retrieval, advanced structured data analysis and efficient agentic workflows, all while maintaining governance at scale.

With Claude 3.5 Sonnet, Cortex Agents can plan, orchestrate, reflect and monitor AI-driven tasks with greater precision. Agentic capabilities of Cortex are improved with the support for tool use and structured output. Customers can use the multimodal capability in Claude 3.5 Sonnet to get insights from a broader set of unstructured data, including images. All interactions happen within Snowflake's secure environment, facilitating controlled access and unified governance across use cases.

Customers can use Claude 3.5 in Cortex Agents to deliver accurate, efficient and governed AI at scale and expedite the delivery of generative AI applications.

AI Observability: Evaluation and tracing of AI Agents

AI observability brings reliability, performance and trust to generative AI applications. With proper evaluations and monitoring, businesses can get more accurate results, optimize costs and address their governance needs.

What's new with Cortex AI Observability?

Cortex AI Observability on Snowflake is powered by TruLens and will be available in public preview soon.

1. End-to-end evaluation

AI Observability can evaluate the performance of agents and apps, using techniques such as LLM-as-a-judge. It can report metrics such as relevance, groundedness and harmfulness, giving customers the ability to quickly iterate and refine the agent for improved performance.

3. Comprehensive tracing

Customers can enable logging for every step of agent executions across input prompts, tool use and final response generation. This allows easy debugging and refinement for accuracy, latency and cost.

Effective governance and processing of both structured and unstructured data within Snowflake are crucial for creating AI-ready datasets that retrieval services can utilize. Snowflake's support for unstructured data includes capabilities to store, access, process, manage, govern and share such data. The **Snowflake Connector for SharePoint** checks that existing permissions are respected to secure access controls. Furthermore, Snowflake's **acquisition of Datavolo** enhances the platform's ability to handle multimodal data integration, reinforcing its commitment to robust data governance and processing.

With these capabilities, Cortex AI Observability makes AI applications more efficient and trusted for enterprise use.

The future of AI agents

AI agents are moving beyond basic automation, dynamically handling multi-step actions and reasoning. This is a significant improvement over the mostly reactive software tools available today. As LLMs continue to advance, agents will collaborate, plan, execute, and refine tasks, driving efficiency and reducing costs. Agents have the potential to reduce both software and labor expenses by orders of magnitude.

Cortex Agents, using Cortex Analyst, Cortex Search, Anthropic's Claude models and AI Observability, bring intelligence on top of a unified governance framework and efficient processing engine for both structured and unstructured data. Using these building blocks, developers can build and deploy data agents that can be integrated to their application of choice using the REST API interface. Additionally, organizations can leverage the solutions built by our partners **Sema4.ai** and **Seek AI**.

Learn more

- Try Cortex Agents: **Build your first Cortex Agent.**
- Watch the demo: **See Cortex Agents in action.**

Secrets of Gen AI Success

Discover how leaders like Bayer and Siemens Energy use gen AI to increase revenue, improve productivity and better serve customers.

DOWNLOAD NOW

Subscribe to our blog newsletter

Get the best, coolest and latest delivered to your inbox each week

elliott.botwick@snowflake.com

SUBSCRIBE NOW

LIVE DEMO

Product

PRODUCT CATEGORIES

Platform

Analytics

AI

Data Engineering

Applications & Collaboration

FEATURED CAPABILITIES

Cortex AI

Data Clean Rooms

Horizon

Marketplace

Native Apps

Notebooks

Snowpark

Streamlit

Snowflake ML

Solutions

INDUSTRIES

Advertising, Media & Entertainment

Financial Services

Healthcare & Life Sciences

Manufacturing

Public Sector

Retail & Consumer Goods

Technology

Telecom

Travel & Hospitality

DEPARTMENTS

Marketing

IT

Finance

Cybersecurity

ENABLEMENT SOLUTIONS

Migrate to the AI Data Cloud

Professional Services

PARTNER SOLUTIONS

Snowflake Partner Network

Partner Finder

Event Partnership Opportunities

Why Snowflake

Why Snowflake

CONNECT

Blog

Engineering Blog

Community

Events

Support

Contact

LEARN

Resource Library

Training

Webinars

Certifications

Live Demos

Snowflake University

Hands-on Labs

Guides

Trending

Developers

Developers Overview

Solutions Center

Open Source

Builder Education

Downloads

Company

Careers

Newsroom

ESG

Snowflake Ventures

End Data Disparity

Pricing

Pricing Options

Cost & Performance Optimization

Snowflake Performance Index

Sign Up for Our Newsletter

elliott.botwick@snowflake.com

United States