snowflake® | aws

DEPLOYING AI AGENTS AT SCALE

3 PATTERNS TO MOVE BEYOND POCS

Learn how to move beyond experimentation and implement AI agents in your enterprise with practical patterns, tooling strategies, and a secure, AI-ready data foundation

TABLE OF CONTENTS

THE MANDATE TO MOVE FROM AI PROOF OF CONCEPT TO PRODUCTION

Artificial intelligence (AI) innovation is happening fast. The pace of change and investment is unprecedented and shows no signs of slowing down.

In 2024, 37% of venture capital (VC) funding went to AI startups, and 57% of business and IT leaders said in a recent survey that they've deployed generative AI solutions and the reason is simple.

"Generative AI is going to reinvent virtually every customer experience we know, and enable altogether new ones about which we've only fantasized," Amazon CEO Andy Jassy wrote in a recent shareholder letter.

Renowned American VC Mary Meeker's annual internet trends report is a must-read for anyone looking for an early preview of what's next in technology. Meeker wastes no time contextualizing the impact this year's biggest trend will have.

"Consider now that AI user and usage trending is ramping materially faster … and the machines can outpace us," Meeker wrote. "The pace and scope of change related to the artificial intelligence technology evolution is indeed unprecedented, as supported by the data."

Let's dive into some of that data, starting with Amazon. The technology giant projects over $100 billion in capital expenditure in 2025 alone to expand data centers for its cloud services business to keep pace with demand for AI compute. The large investments from VCs and the world's biggest companies, coupled with the pace of innovation, makes the mandate from leaders clear: It's time for AI to move from proof of concept to production. And the rewards are clear, with 92% of AI "early adopters" who say they're seeing the value of AI, according to a recent Snowflake survey.

GENERATIVE AI VS. AI AGENTS

Generative AI: Gen AI can create content and communication personalized to a customer's inquiries by analyzing large data collections and adapting the tone and style of its outputs to match brand guidelines or customer needs.

AI agents: Agents act as active problem solvers that can handle interactive and multi-step engagements with customers using a natural language interface.

Agentic AI: This combines the capabilities of traditional and generative AI. It has massive potential to unlock efficiency in enterprises. Why? AI agents are like computer programs that can perform tasks independently using artificial intelligence. They can significantly boost business productivity by handling complex, multi-step operations quickly and taking action on behalf of the requestor.

But early adoption is not simply a set-it-and-forget-it motion. AI is changing use cases across industries from search and coding to how we shop to how we carry out cancer and drug research and so much more. The pace of innovation is moving faster in some areas than others, but organizations that aren't planning ahead will be left behind quickly.

"How soon?" Jassy wrote in his letter. "It won't all happen in a year or two, but it won't take 10 either. It's moving faster than almost anything technology has ever seen."

Don't be one of those organizations. In this ebook, we'll explore three common types of agents that you can implement today, how to set them up and the value they can unlock. We'll look at:

1. Data agents
2. Conversational agents
3. Multi-agent systems

This book will explore how these agents can be used to automate tasks, generate insights and improve customer experiences — and why they're critical to unlocking the full potential of AI in the enterprise. By understanding the various types of agents and how to implement them effectively, organizations can begin to realize tangible value from their AI investments and maintain a competitive edge — while separating AI reality from hype.

"The hype cycle will continue, but for me it's a tale of two cities — one is hype and the other is let's go get some work done," Snowflake CEO Sridhar Ramaswamy recently said. "And we are firmly in the latter."

In that spirit, we hope this work helps your organization get some work done.

But first, some assumptions

In this book, we'll walk you through three paths to implementing agents at your organization, including technical considerations, the tools and services to use and practical considerations designed to help you get agents into production. But we need to start by making a few assumptions:

You have a modern, AI-ready data estate: To work effectively, agents need access to an organization's data, which can be both structured — patient records at hospitals or transaction details at a department store — and unstructured data — video files or social media posts, just to name two examples. All this data needs to be ready in near real time, available to bidirectionally "talk" with your AI agents. But without an AI-ready data foundation, agents won't have the information they need to take action. Think about your agentic use case and the structured, unstructured and semi-structured data assets your agents will need to use. Are they readable, harmonized and unified?

> **53% of business leaders cited data quality or timeliness and 48% cited data silos or data integration as challenges their organization faces when it comes to deploying AI at scale, while 59% of business leaders cited data governance, security or privacy as a challenge.**
>
> Data Strategies for AI Leaders, MIT Technology Review Insights

Security and governance matter to you: These agentic implementation patterns place unified governance and security standards at their foundation. They assume you're not comfortable with your data entering public domains or ever leaving the walls of your organization's infrastructure, even when using features and services from multiple vendors.

You like keeping things flexible and scalable: Businesses are complex, with different lines of business and teams requiring different tools based on skillset, job function, preference and more. The architecture patterns in this book assume you may want to mix and match tools depending on the job and/or use case. And that the impetus is on AI service providers to provide a flexible, integrated data platform with services that work together, eliminating manual, complex configurations for your team, helping you secure and govern your data, and setting up your business for growth and change.

THE THREE COMMON TYPES OF AGENTS EXPLAINED

Now that we're past the assumptions, let's get to the agents. There are two categories of agents at a basic level, depending on their function. There are **personal agents,** which are used for consumer and individual data, and then there are **enterprise agents,** which are used for an organization's data.

As you think about how to architect your agent, start with its function, because agents work best when they have a specific and focused use case. Ask: What is the agent going to do within your business?

In this book, we'll focus on enterprise agents, which have nearly endless possibilities for how they can be utilized across an organization. A few basic examples are:

- Customer service agents, that can build travel itineraries and book travel
- Paralegal agents, that can perform case research and prepare legal documents
- Data entry agents, that extract information from unstructured data and put it into structured formats

Now let's make sure we're all on the same page before we get into the agentic implementation patterns of the three types of enterprise agents we'll explain in this book. It's important to understand the basic level of what they do and their benefits:

1. **Data agents:** This is a category of AI agents focused on efficiently combining data and tools to deliver data-grounded insights with an extreme focus on the accuracy of the results. Data agents must be able to plan and route requests to the correct structured or unstructured data source to retrieve data complete a request. When doing so, data agents must enforce data access and governance policies.
2. **Conversational agents:** They generate informed responses for a human end user in natural language within the workflow or platform the end user is familiar with. Often, the workflow or platform isn't a technical space but instead it could be a messaging tool like Slack or email, a content repository like Box, SharePoint or Google Drive, or a custom application like a customer service window within a website.
3. **Multi-agent systems:** Their focus is to orchestrate multiple agents to complete jobs with multiple steps, requiring different, often specialized agents to gather the information needed to complete each step, then consolidate it into a response.

Building and deploying agents in your organization

In this chapter, we'll show you sample reference architectures and technical implementation patterns to deploy agents at your company using Snowflake and Amazon Web Services (AWS), an integrated cloud service provider. We'll cover them in order of complexity, starting with patterns that are relatively easy to deploy and manage and ending with patterns that require more configuration. We'll end by sharing real-world examples from customers that have deployed agents into production.

By using native integrations between Snowflake and AWS, organizations can use cutting-edge, fully managed agentic services from both providers without managing complex data pipelines or configuring integrations from scratch using API calls. They can leverage features powered by models from industry leaders like Anthropic — without their data ever leaving their secure, governed cloud boundary.

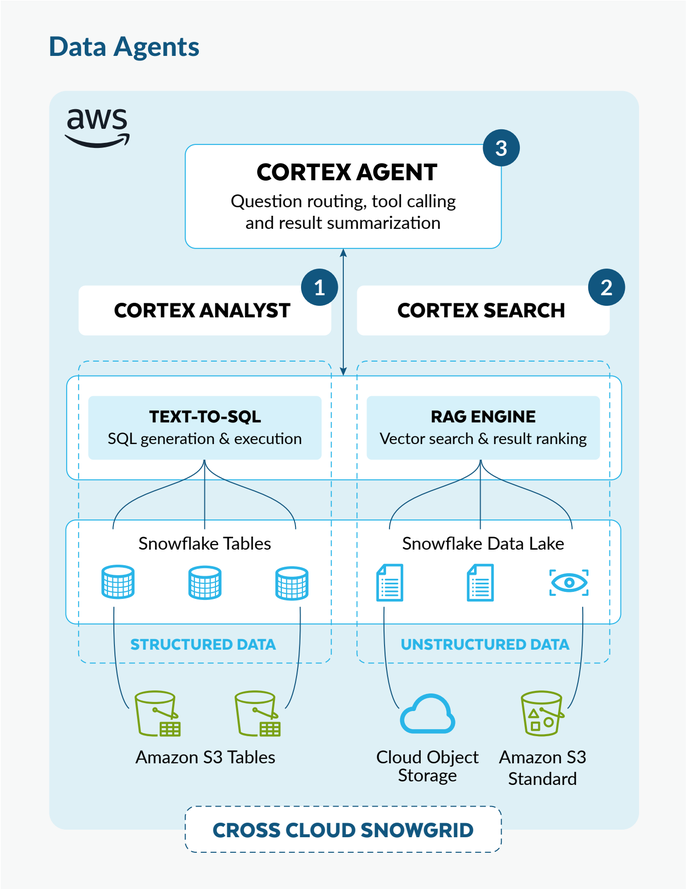
1. DATA AGENTS

Data agents are primarily used by technical personas in organizations, like data analysts and data engineers. They're comfortable using cloud-based data platforms, like Snowflake, where accuracy and grounding in data matters. Data agents provide them an efficient way to generate insights from structured and unstructured data across cloud regions and providers at scale in an easy-to-use and fully managed experience.

In this example, we employ Cortex Agents as our data agent. It works with two specialized AI features designed with the data source in mind.

1. First, Cortex Analyst is designed for structured data retrieval using SQL. It can read structured data in the form of tables that live in a Snowflake warehouse or in any externally hosted, Iceberg REST compatible table, including Amazon S3 Tables or Snowflake Managed Iceberg tables. In either case, Cortex Analyst reads and processes the structured data to generate its response.
2. Next, Cortex Search is designed for unstructured data calling using RAG-based vector search and results ranking. It can read unstructured data in the form of files, images and contracts. These files can live in a Snowflake data lake or in any cloud-based object storage, including Amazon S3. In either case, Cortex Search reads and processes the unstructured data to generate its response.
3. Using the combined skillsets of the purpose-built analyst and search features, the data agent can efficiently generate insights with high accuracy. It returns these insights to the end user.

The data agent in this example analyzes data hosted by multiple cloud providers through a single, governed frontend, uniting the data insights to generate a fully informed response. This is critical because many organizations choose to store their data with multiple cloud providers for a variety of reasons, like maintaining business continuity for critical workflows, preserving choice for different business units, or complying with data residency controls.



Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. Customers of all sizes and industries can use Amazon S3 to store and protect any amount of data for a range of use cases, such as data lakes, websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics.

Amazon S3 Tables is the first cloud object store with built-in Apache Iceberg support that streamlines storing tabular data at scale.

Cortex Agents, powered by leading large language models (LLMs) like Anthropic's Claude and offered by Snowflake, are automated, AI-driven tools that analyze an organization's structured and unstructured data, optimize queries, and generate responses that enable data-grounded insights.

Cortex Analyst provides a conversational interface to interact with structured data in Snowflake. Behind the scenes, it leverages the power of Anthropic's Claude and other LLMs to generate highly accurate text-to-SQL queries so users can answer questions about structured data in natural language.

Cortex Search is a hybrid (vector and keyword) search service in Snowflake that enables low-latency, high-quality "fuzzy" search on unstructured data, without having to worry about embedding, infrastructure maintenance, search quality parameter tuning, or ongoing index refreshes.

Snowflake Managed Iceberg Tables combine the performance and query semantics of typical Snowflake tables with external cloud storage that you manage. They are ideal for existing data lakes that you cannot, or choose not to, store in Snowflake. Iceberg tables use the Apache Iceberg™ open table format specification, which provides an abstraction layer on data files stored in open formats.

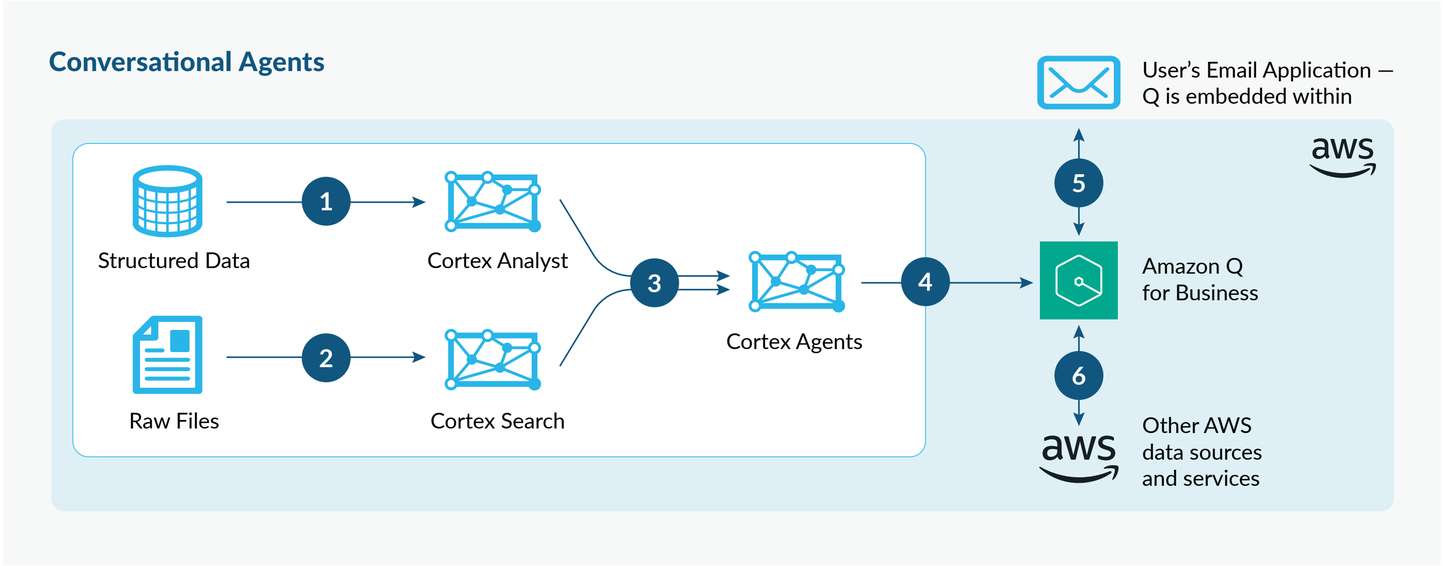
Snowgrid is a Snowflake feature that enables customers to securely share and collaborate on data across different regions, clouds and organizations, providing a single, unified platform for data management, analytics and collaboration.

2. CONVERSATIONAL AGENTS

Conversational agents help nontechnical users by giving them access to a conversational assistant to help them get the answers they need from their data. Your agent needs to live in the places where business users work — tools like email, chats and content repositories. When deploying conversational agents in your enterprise, there are several technical considerations to keep in mind. One key consideration is how to integrate the conversational assistant with your existing data and services, which we'll get into below.

In this example, the architecture embeds Amazon Q, a conversational assistant, directly into users' email platforms, enabling interaction as a personal AI assistant within familiar tools and workflows. Amazon Q then leverages its integration with Cortex Agents to retrieve the necessary business data for generating responses, without moving it.

1. Start by creating a Cortex Analyst service to analyze structured data.
2. Next, create a Cortex Search service to analyze unstructured data.
3. Then, create a Cortex Agent that brokers the decision on when to use each service.
4. Connect Amazon Q to the Cortex Agents service using a plugin that has Open Authorization (OAuth) authentication.
5. Connect Amazon Q to other AWS data and services as needed to inform your assistant's responses.
6. Embed Amazon Q within the frontend application of choice using out-of-the-box connectors or Amazon Q Embedded for trusted websites.



Amazon Q Business is an enterprise-grade generative AI service that enables businesses to create, customize, and manage their own secure AI assistants. These assistants can be seamlessly integrated into web browsers and applications, helping employees find information and complete tasks faster while maintaining enterprise security and compliance standards.

3. MULTI-AGENT SYSTEMS

Multi-agent systems automate and manage complex workflows and tasks, providing streamlined processes, increased efficiency and reduced manual effort, enabling them to focus on higher-value activities.

Keep in mind that when you use multi-agent systems, your workflow will be complex because you need at least three specialized agents to work together seamlessly to complete the job. Ideally, you want to mix and match specialized agentic tools from various providers. Aiding inter-agent connection and communication are open standards, which enable secure and authorized interactions between agents and the tools they use to complete jobs — as well as among the agents themselves — regardless of their underlying frameworks or vendors. We'll explain more below.

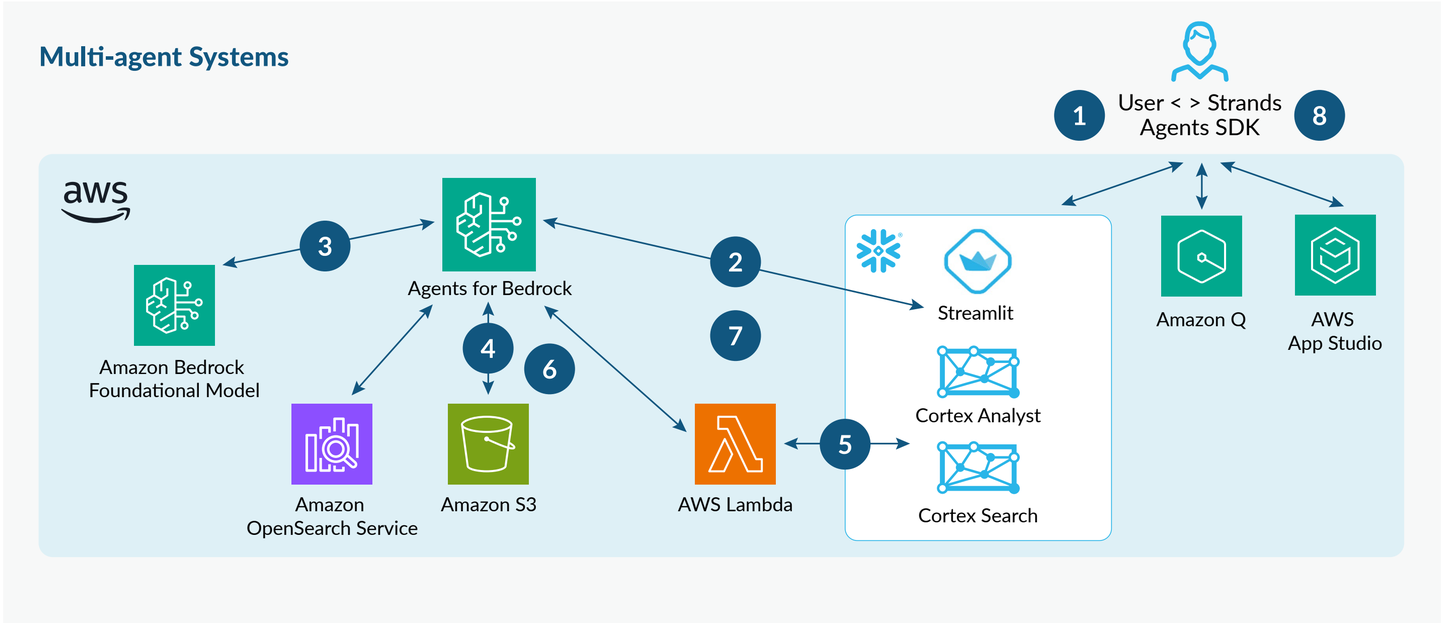
> "Maximizing the potential of agentic AI depends on open protocols like A2A and the shared knowledge and best practices they provide."
>
> **—DWARAK RAJAGOPAL**
> VP of AI Engineering & Research, Snowflake

AGENTIC INTEROPERABILITY

The field of AI agent communication is rapidly evolving with two key open standards gaining traction: the model context protocol (MCP) and agent2agent (A2A). MCP was designed to connect AI agents to tools, providing a structured way for agents to access external data, resources and functionalities. It leverages the OAuth framework to help ensure interactions between agents and their tools are secure and authorized. A2A was specifically created to connect agents to agents, enabling AI agents to discover, communicate and collaborate with each other. While distinct in their primary focus, MCP and A2A can work together. The choice between them (or using them together) depends on the specific needs of the agent system.

In this example, we use Amazon Bedrock Agents to orchestrate complex workflows that employ multiple agents working together to complete a task. The multi-agent system employs a model to act as a decision maker and broker by receiving the prompt and deciding how to take action. One example would be how to break the job down into subtasks and farm those tasks out to the right agent. Multi-agent systems may ask, "What type of data is this?" while also thinking about factors like where the data lives and what specialist agents are on their team to inform the decision and instruct the other agents on how to work together to complete the job.

1. Start by building your multi-agent system by using Strands Agents, a software development kit (SDK) that helps you build out your agentic workflow by writing code that defines three things:
    a. Your model, or the "brains" of multi-agent system
    b. The tools that are deployed and orchestrated to finish the tasks
    c. And the prompt, or the frontend application/natural language interface
2. For your frontend application or conversational assistant, you can use tools like Streamlit in Snowflake, Amazon Q, AWS App Studio or a custom frontend of your choice.
3. Next, the Amazon Bedrock Agent receives the prompt and uses the model to make a decision. In this case, we're using a model from Amazon Bedrock.
4. The Amazon Bedrock Agent then sends the action to the downstream agents and additional required tools. In this case, the agent uses specialist agents (Cortex Analyst, Cortex Search) and retrieves data with Amazon OpenSearch Service and directly from Amazon S3.
5. AWS Lambda receives the action and, using a Python SDK, prompts Cortex Search and Analyst.
6. The agents generate their responses and return them to the Amazon Bedrock Agent, again via AWS Lambda.
7. The Amazon Bedrock Agent again leverages its model to consolidate the responses, generate its reply in natural language, and return it to the application and end user.



Strand Agents are an open source SDK that integrates A2A standards and takes a model-driven approach to building and running AI agents with just a few lines of code. Strands can help users scale from simple to complex agent use cases and from local development to deployment in production quickly.

Amazon Bedrock Agents is a service for building and configuring autonomous agents within applications by connecting with company systems, APIs and data sources. It uses the reasoning of foundation models, APIs and data to break down user requests, gather relevant information and efficiently complete tasks.

AWS App Studio is a generative AI-powered service that uses natural language to build business applications without requiring deep software development skills.

Amazon OpenSearch Service is an AWS-managed service that lets users run and scale OpenSearch clusters without worrying about managing, monitoring and maintaining infrastructure or having OpenSearch clusters operation expertise.

AGENTS DELIVERING BUSINESS VALUE IN PRODUCTION

LUMINATE

Luminate Amps Entertainment Analytics with Rich, AI-Powered Insights

Luminate is the entertainment industry's trusted data partner. Everyone — from major record labels and production companies to studios, networks and talent agencies — relies on Luminate for verified, independent information that reflects industry trends so their own customers can make data-driven decisions.

Data is Luminate's entire business. The company powers both the Billboard music charts and Variety Streaming Originals charts every week via the more than 4TB of data it receives every day, in different formats, from different sources. Luminate sought to revolutionize how its users interact with these massive data sets, enabling them to ask questions in natural language and generate fast insights with the help of agents.

Delivering on this vision required a powerful, secure and unified platform that could handle both structured and unstructured data at scale. To achieve this, Luminate underwent a massive data modernization project, transitioning its data infrastructure from legacy on-premises systems to Snowflake and AWS. Luminate has centralized its massive volumes of data into Amazon S3 and Snowflake while continuously bringing in fresh data through a pipeline built using dbt and Snowpark. These improvements allow Luminate to easily generate timely cross-industry insights that were previously impossible to create.

With an AI-ready data foundation in place, Luminate embedded AI agents directly in its enterprise products using Snowflake and AWS. The company uses Amazon Bedrock Agents to orchestrate sophisticated, multi-step agentic workflows. The system employs a combination of agents, including Cortex Analyst — powered behind the scenes by Anthropic's latest Claude models — and Cortex Search to retrieve more accurate, high-quality insights from data. By relying on a combination of specialized agents, Luminate can tailor the model to the use case, striking the right balance of performance, cost and latency to meet its business needs.

The technical architecture leverages Snowflake Cortex's semantic understanding capabilities to parse complex entertainment industry queries and translate them into executable SQL operations against massive datasets. Cortex Analyst functions as an intelligent SQL generation layer that understands the nuanced relationships between artists, genres, streaming platforms and geographic markets. This agent can automatically join disparate tables, apply appropriate filters and generate aggregations while maintaining data lineage and governance controls. The system's ability to handle time-series analysis is particularly powerful for tracking chart performance, streaming trends and seasonal patterns across different entertainment verticals.

Amazon Bedrock Agents complement Snowflake's analytical capabilities by providing sophisticated reasoning and planning abilities that enable multi-step workflows. These agents can decompose complex user requests into a series of coordinated actions, such as first identifying relevant artists based on similarity metrics, then cross-referencing their performance data with venue capacity and geographic preferences, and finally synthesizing recommendations with supporting rationale. Amazon Bedrock's function calling capabilities allow these agents to interact with external APIs, trigger data pipeline updates, and even initiate automated reporting processes, creating a truly autonomous analytical environment.

The synergy between Snowflake Cortex and Amazon Bedrock creates a powerful feedback loop for continuous model improvement. Amazon Bedrock Agents can analyze user interaction patterns and query success rates to identify knowledge gaps, which then inform Cortex's semantic model refinements. This collaborative intelligence enables the system to learn from each interaction, gradually improving its understanding of entertainment industry terminology, seasonal patterns and user preferences. The architecture also supports A/B testing of different agent configurations, allowing Luminate to optimize response quality and latency based on real-world usage patterns.

Security and governance are paramount in this agentic architecture, with the built-in guardrails of Amazon Bedrock working in tandem with Snowflake's row-level security and dynamic data masking capabilities. Each agent operates within predefined boundaries, helping keep sensitive client data protected while still enabling comprehensive cross-industry analysis. The system maintains detailed audit logs of all agent interactions, providing full transparency into how insights are generated and verifying compliance with industry regulations and client confidentiality requirements.

Luminate's data teams are now more productive since the fully managed service and native integrations provided by Snowflake and AWS eliminate the need for time-consuming infrastructure maintenance. And by building its AI applications where its data already lives, Luminate enables robust levels of data quality, security and control that its customers demand.

With these innovations in place, Luminate provides users with faster, deeper and more personalized insights. Before, it was a manual, time-consuming process to piece together disparate data points to answer nuanced questions like "What are the top three Japanese TV shows similar to Netflix's 'Stranger Things'?" Now, its AI agents can combine various data sources, including text and graph embeddings, to quickly deliver these complex recommendations.

These recommendations heavily influence key decisions for Luminate's users. An artist manager, for example, could ask where a specific artist should tour next to maximize physical album sales. The agent would then tap into Luminate's historical sales data down to the ZIP code level, potentially pinpointing specific venues and markets where the artist has historically performed well or shows high popularity, helping them make smarter tour decisions that directly impact their bottom line.

> **"At Luminate, we're revolutionizing how we deliver data-driven insights through generative AI. Snowflake and AWS provide a unified, scalable platform for processing and retrieval for both structured and unstructured data, which is the foundation for building and orchestrating AI agents using their embedded AI services."**
>
> **—GLENN WALKER**
> Chief Data Officer, Luminate Data

eSENTIRE

eSentire Democratizes Elite Cybersecurity Expertise Through Agentic Intelligence

As a global leader in Managed Detection and Response (MDR) services, eSentire helps more than 2,000 organizations worldwide anticipate, withstand and recover from cyber attacks.

eSentire has accumulated 25 years of deep investigative data and refined expertise from elite SOC analysts who've seen every attack vector and threat evolution. But in cybersecurity, these seasoned SOC analysts are sparse, making it challenging for organizations to access the deep expertise needed for sophisticated threat analysis.

To democratize access to world-class cybersecurity expertise, eSentire developed a significant breakthrough, embedding the collective intelligence of their most elite Tier 3 SOC analysts into AI agents that deliver comprehensive, expert-level threat investigations. These investigations, which include the depth of analysis equivalent to five hours of expert effort, now take place within the critical 15-minute window that security incidents demand — a feat that was previously impossible within operational time constraints.

eSentire's solution, Atlas AI, is an advanced platform that doesn't just automate tasks, but actually thinks like the company's most experienced SOC analysts. eSentire uses Snowflake as a data mesh to centralize and normalize vast amounts of security data across all customers, while their low-code platform enables programmatic access to that data as well as security tools. Using AWS Lambda to orchestrate complex agent workflows, Atlas AI integrates Anthropic's Claude Sonnet models with Amazon Bedrock to create AI agents that investigate threats with the same sophistication as elite human experts.

Since launching Atlas AI, eSentire has seen remarkable results, including:

- **Expert-level investigations within critical timeframes:** Atlas AI enables the comprehensive analysis that would ideally require five hours of expert investigation — but was previously impossible within the critical 15-minute window that security incidents demand.
- **95% expert alignment:** Atlas AI investigations demonstrate 95% alignment with Tier 3 SOC analyst decisions, proving AI can genuinely replicate expert judgment.
- **99.3% first-host containment:** Nearly perfect threat isolation demonstrates the speed and precision of expert-level AI investigations.
- **Upskilled local talent:** Partners can elevate their existing analysts by giving them access to AI agents that deliver elite SOC expertise, enabling junior and mid-level analysts to deliver expert-level investigations.
- **Improved service delivery:** Instead of competing for scarce expert talent, partners can upskill their teams with AI agents that guide them through sophisticated threat analysis with expert-level precision.

Success by the numbers

- Expert-level investigations within critical 15-minute timeframes
- 95% alignment with expert SOC analyst decisions
- 99.3% of threats isolated at first host
- 99% noise reduction across environments
- 96% SOC analyst retention, with an average tenure of 6 years

> "AI on its own is somewhat valuable, but for us the real value comes in the stack of capabilities, and that comes from bringing all the data into one unified platform where we can operate AI on it at scale."
>
> **—DUSTIN HILLARD**
> CTO, eSentire

SNOWFLAKE AND AWS FOR PRODUCTION-READY AI AGENTS AT ENTERPRISE SCALE

Designed for interoperability, Snowflake and AWS provide over 50 integrated features and services for data and AI that reduce manual configuration and maintain security and governance while providing users with both flexibility and choice in how they store, read, process and activate all types and volumes of data across their cloud environments.

With this data foundation in place, production AI for real business ROI is within reach — and organizations are already seeing this value with AI agents. This ebook provided a deeper dive into common agentic implementation patterns seen across industries to help you get started.

Whether deploying specialized data agents to extract structured and unstructured data insights, embedding conversational assistants, or orchestrating multi-agent systems, Snowflake and AWS offer integrated, fully managed services for agentic AI that drive transformational business outcomes with zero data movement or complex pipeline management required. Securely activate enterprise data for agentic AI workloads faster with Snowflake and AWS to unlock efficiencies, accelerate insights and drive transformational outcomes.

Learn more about Snowflake and AWS to get started with AI agents today.

Snowflake is the platform for the AI era, making it easy for enterprises to innovate faster and get more value from data. More than 11,000 companies around the globe, including hundreds of the world's largest, use Snowflake's AI Data Cloud to build, use and share data, applications and AI. With Snowflake, data and AI are transformative for everyone.

Learn more at **snowflake.com**

(NYSE: SNOW)

Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud platform, offering over 240 fully featured services from data centers globally. Millions of customers — including the fastest-growing startups, largest enterprises, and leading government agencies — trust AWS to power their infrastructure, become more agile, and lower costs.

Learn more at **aws.amazon.com**

(NYSE: AMZN)