snowflake®

THE ESSENTIAL GUIDE TO GENERATIVE AI

Key strategies to overcome obstacles and unleash AI success

TABLE OF CONTENTS

INTRODUCTION

Not too long ago, the term "generative artificial intelligence" was known only to data scientists and machine learning specialists. Now it's a topic of cocktail party conversations, comedy monologues and congressional hearings.

This transformative technology became an overnight sensation in December 2022, when OpenAI made available to the public a preview of its text chatbot ChatGPT-3, along with its image-generating cousin DALL-E. Suddenly, anyone with an internet connection could tap into a seemingly sentient AI chatbot to conduct research, compose emails, draft reports, write poetry, solve math problems, generate code or produce original images — all by using conversational text prompts.

ChatGPT captured the public imagination in a way few other emerging technologies have. Within two months, more than 100 million people had signed on to experiment with the tool — making it one of the most rapid adoptions of a digital service in history. By March 2023, competitors like Microsoft, Google and Anthropic had released their own gen AI chatbots to the public, with occasionally embarrassing results. At the same time, we saw a boom in open-source, large language model (LLM) alternatives such as Meta's Llama 2 and Falcon LLM, with performance comparable to commercial, externally hosted models.

These strides of innovation started a race toward greater and greater generative AI (gen AI) advancements, which show no signs of slowing. Companies spanning Amazon to Samsung to X (née Twitter) have all announced plans to unveil their own gen AI products — and a plethora more are on the horizon. It is a frenetic new frontier.

But enterprises are still grappling with how to tap into the enormous potential of this technology while avoiding the worst possible pitfalls. To fully understand how gen AI can help improve productivity and drive business decisions — while skirting the biggest potholes of possible reputational danger — it helps to understand how the technology works and how we got here.

A BRIEF HISTORY OF GEN AI

Though it feels like the gen AI evolution happened quickly, it's the result of decades of research into how to make machines behave more like humans.

The term "artificial intelligence" was coined in the mid-1950s by researchers at MIT and Carnegie Mellon University. Academic interest in the topic has fluctuated over the decades since then; there were multiple "AI winters" where research into the technology was dormant, if not entirely extinct.

Interest in AI began to pick up steam again around 2009, with the emergence of new methods of training machine learning models, as well as increases in computing power available to the neural networks used to create these models. This was the early beginning of the "gen AI era."

Experts began building AI models more quickly when they discovered that the same graphics processing units (GPUs) used to generate 3D images inside video games were ideal for the high-speed, parallel math operations required by neural networks. Suddenly, neural networks became much more powerful and able to analyze larger volumes of data in shorter amounts of time.

1950 - 1960s
THE EARLY YEARS

1950: Turing Test introduced.
1956: Term "artificial intelligence" coined.
1959: Checkers-playing program developed.
1964: ELIZA, the first chatbot, created.

1970 - 1980s
FIRST "AI WINTER," ENTHUSIASM WANES

1990 - 2000s:
RISE OF MACHINE LEARNING

1997: Deep Blue defeats world chess champion Garry Kasparov.
2000s: Neural networks advance.

2000s - 2010
SECOND "AI WINTER"

2010s - 2020
AI RESURGENCE

2011: Apple releases its first iPhone with Siri; Watson beats humans in Jeopardy!
2012: Deep learning with AlexNet.
2016: AlphaGo defeats Go champion Lee Sedol.
2018: Initial GPT LLM released.

2020s
GEN AI, THE NEXT REVOLUTION

2020: GPT-3, the largest LLM at the time, introduced.
2022: ChatGPT released, sparks investment rush and 100M users in 2 months.
2023: AI everywhere

- **61%** of workers already use or plan to use gen AI.
- **49%** of the general population uses gen AI, and **34%** uses it every day.
- Estimated to become a $1.3 trillion market by 2032.

Around 2016, researchers began building new large language models by feeding material scanned from books, websites and other unstructured text into neural networks. In 2017, a new approach to training deep neural networks, called **transformers**, enabled researchers to teach machine learning models the relationships between words based on where they appear within a sentence. Because transformer models work in parallel, aided by fast GPUs, they are both more accurate and require less time than other training approaches.

The transformer approach allowed LLMs to resolve ambiguities in language and understand both literal and metaphorical meanings, as in the classic aphorism, "Time flies like an arrow; fruit flies like a banana." The ability to derive different meanings from the same words, based on the context in which they appear, is what enables users to communicate with these models using conversational English or other languages.

The rise of gen AI has led to an increase in LLMs. Generally speaking, there are two types of LLMs called GPT and BERT. GPT is developed by OpenAI and is based on decoder-only architecture while BERT (Bidirectional Encoder Representations from Transformers) is developed by Google and is an encoder-only pre-trained model. While technically distinct, both types of models are built to perform natural language processing tasks.

WHAT ARE NEURAL NETWORKS AND HOW DO THEY WORK?

Gen AI is made possible by deep neural networks: computer systems modeled on the structure of the human brain, where operations are performed on data by successive layers of "neurons" (mathematical operations). These networks teach themselves to "think" by analyzing data and identifying patterns inside it.

Let's say you wanted to train a neural network to recognize pictures of cats. You would start by feeding it images of felines labeled "cat," along with images of other animals labeled "not cat." The network would then begin to identify what patterns of data the images labeled "cat" have in common: One layer of neurons might identify light and dark pixels in the image, the next layer might identify lines and shapes, the next layer colors, then fine features like whiskers and ears, and so on. Feed the network enough images and it will eventually be able to assess an image it has never seen before and accurately predict whether or not it contains a cat.

Deep neural networks can have anywhere from just a handful of layers to more than 1,000, as well as millions or billions of "parameters" — weights that data scientists apply to operations within each layer to produce more accurate results. But because such networks are self-taught, even the people who built them don't always know exactly how they arrive at their predictions. This is sometimes called "**black box AI**."

As industry watchdogs and regulatory agencies call for more transparency and explainability in AI models and the data used to train them, organizations are starting to deploy simpler, more shallow networks whose output is easier to explain. However, the vast majority of AI models currently in the wild have been created using neural networks whose development remain opaque.

THE BIRTH OF GPTs

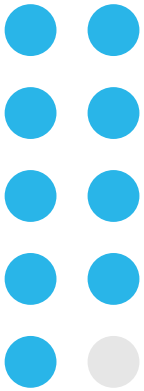
Once these models were pre-trained to understand language, most were then taught how to generate different types of outputs in response to text prompts such as producing code, images, mathematical formulas and so on. (That's why these models are known as generative pre-trained transformers or GPTs.)

The capabilities of a gen AI engine usually correspond to the volume and types of data it has been trained on, as well as **how many parameters (or variables) data scientists** used to fine-tune the model. The larger the model (and the more parameters it uses), the more powerful it becomes. For example, GPT-4 uses a reported 1.75 trillion parameters, or ten times as many as GPT-3; Amazon's recently announced Olympus AI engine may use as many as **2 trillion**.

GPT-4 has demonstrated performance at a level better than humans on the Uniform Bar Examination, the Law School Administration Test (LSAT), Graduate Record Examinations (GRE) test and multiple Advanced Placement (AP) subject exams. **GPT-5**, currently expected to make its debut before the end of 2025, will likely represent another massive leap forward in the capabilities of gen AI.

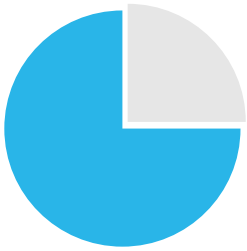
THE 21ST-CENTURY INDUSTRIAL REVOLUTION

Surveys show that **nine out of 10 IT professionals** believe that gen AI will play a prominent role in their companies' near future. Business leaders are equally bullish on the technology's potential. Nearly **three-quarters of Fortune 500 companies** intend to incorporate gen AI into their operations over the next three years to boost employee productivity, improve customer service and automate manual processes.



9 OUT OF 10

BELIEVE THAT GEN AI WILL PLAY A PROMINENT ROLE IN THEIR COMPANIES' NEAR FUTURE.



NEARLY THREE-QUARTERS

OF FORTUNE 500 COMPANIES INTEND TO INCORPORATE GEN AI INTO THEIR OPERATIONS.

There are a vast number of potential use cases for gen AI, spanning virtually every major industry, as seen in the table to the right.

With so much value to gain across industries, virtually every 21st-century enterprise will potentially be looking at deploying gen AI to increase productivity, drive innovations and differentiate their products from those of their competitors.

INDUSTRY	USE CASES
Financial Service Firms	• Analyze insurance claim documents to identify fraud and more accurately gauge maintenance needs • Provide customer-facing wealth advisors with next-best-action insight using models trained on each customer's portfolio data
Retailers	• Analyze customer sentiment to inform data-driven marketing decisions • Build recommendation engines for e-commerce vendors by combining customer purchase histories with data from Snowflake and third parties
Manufacturers	• "Read" quality inspection reports and improve their manufacturing and quality control processes • Aid in performing root-cause analysis of sensor and machine errors
Media and Marketing Companies	• Optimize ad campaigns in real time • Identify which vendors are delivering on the company's core KPIs • Reallocate budgets as needed
Healthcare and Life Sciences Organizations	• Reduce documentation while improving lab and site operations • Accelerate medical research and drug discovery
Telecommunications Companies	• SQL, Python, Java, Scala, SQL APIs, REST API, Dataframes, etc., to access multi-model data (structured, semi-structured, unstructured, various file types, etc.)
Public Sector Agencies	• Uncover fraud and reduce waste • Analyze the effectiveness of government programs

FOUR THINGS TO CONSIDER WHEN NAVIGATING THE RAPID EVOLUTION OF GEN AI

In the rapidly evolving landscape of gen AI, the initial awe and skepticism surrounding its capabilities have given way to a pressing business imperative.

Companies are now racing to adopt gen AI to enhance workforce productivity and profitability. However, the path to implementing effective gen AI solutions is fraught with challenges.

Goutham Belliappa, Managing Director of Strategy and Analytics at Deloitte, highlighted four key considerations for businesses navigating the gen AI terrain in a recent interview with **Data Cloud Now**.

Belliappa emphasizes the critical need for a holistic data strategy, asserting that the fragmented nature of the data market requires organizations to manage data influx on their own terms. A well-defined data strategy becomes the cornerstone for overall business strategies, priorities, and investments, preventing hasty and potentially misguided investments in gen AI capacities.

It's also important for organizations to focus on talent acquisition in the gen AI era because widespread accessibility of the technology can lead to mediocre output. Organizations need to adapt existing talent skill sets to guide gen AI toward producing valuable content.

Organizations also need to understand AI and build trust in its applications. With gen AI's potential to disrupt various business functions, the need for fluency in AI usage becomes paramount. Deloitte is one company that offers fluency courses to address this knowledge gap.

It's also important to build trust in AI, emphasizing the potential biases in training data and the ethical implications of using AI models. Belliappa calls for ethical governance and regulation to safeguard against the misuse of gen AI and underscores the importance of not compromising trust and public safety for the advantages of higher workforce productivity.

To learn more, check out **Belliappa's full interview** on Data Cloud Now.

HOW FINANCIAL SERVICES SHOULD PREPARE FOR GEN AI

In the wake of ChatGPT's unleashed potential to the public in November 2022, the financial services landscape has been swept into a transformative wave of LLMs, reaching into every nook and cranny of modern industry. Vidhya Sekhar, Americas Financial Services AI & Data Leader at Ernst & Young LLP, recently sat down with Data Cloud Now to explore that impact. The following is an excerpt of that conversation.

DEMOCRATIZING INSIGHTS WITH GEN AI/LLMS

Brimming with excitement and curiosity, the financial world has eagerly embraced gen AI and LLMs for their unparalleled abilities in handling vast troves of unstructured data. Unlike their traditional machine learning counterparts, these models boast content synthesis, information extraction, and content generation prowess. In the hands of financial organizations, they become tools not only for automating complex processes but also for rendering technical information comprehensible to both the tech-savvy and the layperson. The democratization of access to insights facilitated by LLMs will offer benefits that extend to content synthesis, information retrieval, and even cross-language content translation.

NAVIGATING RISKS IN THE FINANCIAL TECH SEAS

However, the journey toward democratization is not without its perils. The heavily regulated financial services sector demands a meticulous approach to governance as it grapples with the risks introduced by gen AI, including model output hallucinations and intellectual property concerns. As regulatory bodies worldwide issue guidelines, organizations are urged to reassess their frameworks to align with the evolving legal and regulatory landscape. There will be a need for vigilance and prudence, acknowledging that the short-term experimentation with gen AI will likely evolve into a more deeply ingrained, long-term automation within business processes.

THE ROAD TO INTEGRATION: CHALLENGES AND PROMISES

Transitioning into the world of gen AI is an enticing prospect, but it's far from a straightforward endeavor. Success hinges on the readiness of organizations—both in terms of data infrastructure and governance frameworks. It's a delicate dance, balancing the allure of AI-driven efficiencies with the stark reality of potential pitfalls and regulatory concerns. As gen AI weaves its way into the financial fabric, the transition—while not a sudden disruption—will reshape business processes gradually, necessitating a thoughtful and strategic approach.

In a world where gen AI is not just a buzzword but a force shaping the future of the financial services industry, organizations will need help to navigate the uncharted waters of transformative technology.

To learn more about gen AI in financial services and how Snowflake can help, check out the full interview on DCN's channel.

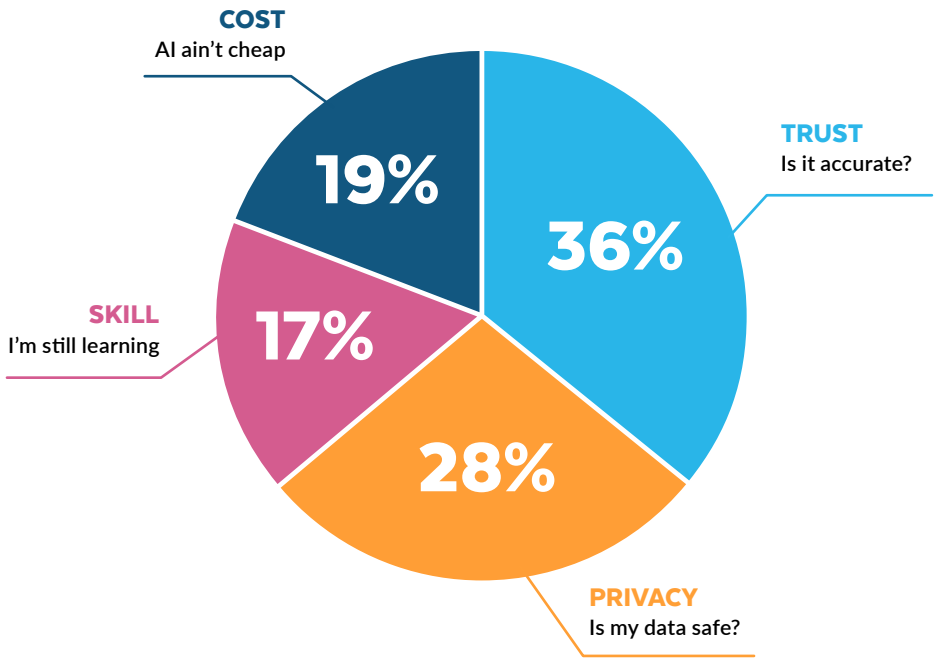
BARRIERS TO THE SUCCESSFUL GEN AI IMPLEMENTATION

This all sounds wonderful, but there is an enormous gap between deciding to implement gen AI capabilities and being prepared to put them into production. While most organizations have more data than they know what to do with, many lack a strategy needed to put that data to effective use.

A survey of more than 300 chief data officers (CDOs) at leading global enterprises found that fewer than half had taken the necessary steps to prepare data for use with gen AI. Data quality is the No. 1 concern, followed closely by the need to establish guardrails around responsible use, ensure data security and privacy, and develop the necessary skills to use gen AI effectively and ethically.

Tom Davenport, distinguished professor of information technology management at Babson College and one of the co-authors of the CDO study, notes that "generative AI is the most dramatic advancement of our age." But he cautions that "if organizations are to succeed with generative AI, they need to increase their focus on data preparation for it, which is a primary prerequisite for success."

WHEN ASKED THEIR BIGGEST CONCERN WHEN BUILDING LLM-POWERED APPS, STREAMLIT USERS REPLIED:



Online surveys were conducted by Streamlit across X/Twitter, YouTube and LinkedIn, August 2023. n = 978

Without a fully developed enterprise data strategy, organizations are likely to face a range of challenges to successfully implementing gen AI. These challenges can be grouped into these four main categories:

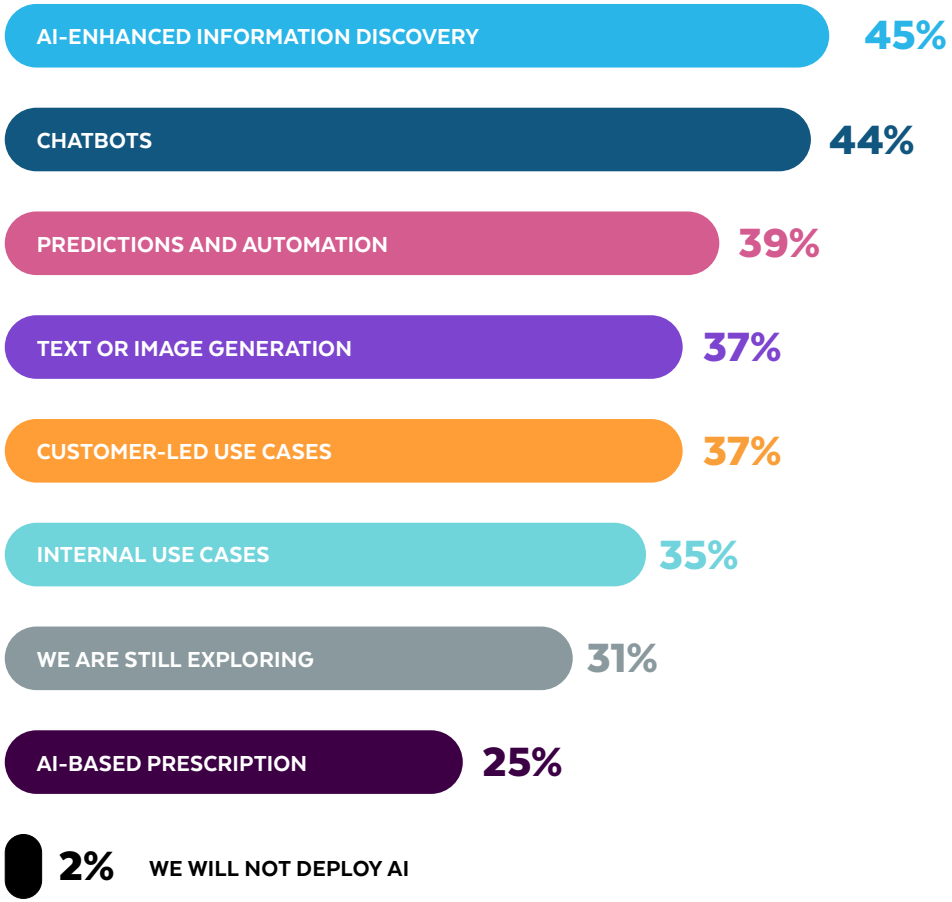
1. Silos across languages, tools and architecture.
Internal teams looking to collaborate on gen AI projects will need to constantly switch contexts, learn new skills, and refactor across different systems and architectural patterns due to a variety of programming language and development tool preferences across teams.

2. Heightened data security threats.
Rushing to capitalize on gen AI can result in data proliferation across many silos and environments. Organizations may lose control over the security of sensitive information, leading to data exfiltration, privacy vulnerabilities and noncompliance with regulatory requirements.

3. Increased operational complexity.
Managing multiple data stores and compute platforms across different cloud environments and regions can be labor-intensive, costly and inefficient.

4. Loss of productivity and business continuity.
Sharing and securing data requires building multiple pipelines, and the copying of data between them, when data and the teams working on that data are siloed. This leads to delays and wasted resources. When inevitable cloud-specific outages occur, that can result in data loss, downtime and gaps in business continuity.

These are just some of the reasons organizations are already using gen AI. Attendees at several Data Cloud World Tour events in 2023 were asked how their businesses were leveraging AI today. They replied:



BEWARE THE PITFALLS OF GEN AI

Like any emerging technology, gen AI has downsides and flaws that users need to be aware of, especially when using publicly available chatbots in enterprise settings.

As Weights & Biases Cofounder and CISO Chris Van Pelt pointed out in a recent Data Cloud Now interview, "These models are probabilistic so they're going to be wrong some percentage of the time by definition."

Here are the most common causes of untrustworthy generative AI results:

Hallucinations. One of the first flaws users discovered with AI chatbots based on LLMs was their tendency to generate material that sounded plausible but wasn't true. Two attorneys in New York discovered this the hard way when they used ChatGPT to research case law and filed briefs citing cases that didn't actually exist. (The attorneys were later fined $5,000 by the court.) The term used by data scientists to describe these fabrications is *hallucinations*. Though more recent iterations of ChatGPT are less prone to hallucination, users should not take anything AI chatbots generate at face value, nor publish any material without fact-checking it first.

Flawed code. By automatically inserting snippets of commonly used code as needed, AI-based programming assistants like GitHub Copilot and Amazon CodeWhisperer offer huge production boosts for developers. But they also increase the risk of introducing outdated, buggy or vulnerable code into enterprise software. A 2022 study by researchers at Stanford University found that developers using AI coding assistants were nearly four times more likely to generate code vulnerable to SQL injection attacks. Developers should always test snippets of AI-generated code to ensure they are safe before putting them into production.

Data leaks. Organizations that upload sensitive data to public-facing AI chatbots may inadvertently share proprietary information or intellectual property. As a result, corporations including Amazon, Apple, Samsung and Verizon have banned the use of ChatGPT among their staff. Organizations that need to apply gen AI to proprietary or sensitive data will likely want to create and train their own LLMs.

Copyright infringement. Some of the largest public LLMs have been trained using data scraped from the internet, much of which may be copyrighted or otherwise owned by third parties. For example, Google's C4 dataset, which has been used to train language models used by Google and Facebook, was found to contain more than 200 million copyright symbols within its text data. Enterprises need to understand the provenance of any data used to train the models they deploy.

Hidden bias. The results produced by AI systems can be skewed due to inadequate or incomplete training data. For example, facial recognition systems are considerably less accurate in identifying people of color, most likely because there was a lack of diversity in their initial training data. This problem is compounded when working with systems where the training data is not transparent and unavailable for review. Enterprises need to ensure the data used to train their LLMs is in their control and as free from bias as possible.

Van Pelt sums up what's needed to avoid these dangers: "The only way to understand how gen AI outputs are wrong is by having a really mature evaluation toolkit. That way the team can actually look at the results and understand what the edge cases are – and then focus their efforts on how to steer or fine-tune the models to minimize the chance that a bad result ends up in users' eyeballs."

HOW SNOWFLAKE'S UNIFIED DATA PLATFORM ADDRESSES THESE ISSUES

Establishing a solid foundation for your enterprise data is essential for taking advantage of the potential benefits gen AI can provide. Snowflake offers several competitive advantages over other providers in this market:

Eliminate silos across architectures, business unit workloads, clouds, languages or tools. Snowflake supports a full spectrum of data formats, architecture patterns and use cases, as well as the ability to work with SQL, Python and other popular languages. It provides a truly flexible platform that allows you to develop using languages and tools already familiar to your organization, as well as the ability to adapt as tools and formats change over time.

Protect your data on a platform with unified and consistent security and governance. Snowflake Horizon is a built-in data governance solution with a unified set of compliance, security, privacy, interoperability and access capabilities. Snowflake's automated continuous compliance monitoring systems can meet even the most stringent and complex security requirements. As raw data is transformed and ingested into the platform — whether it is leveraged for analytics, used to train a large language model or harnessed to build apps — Snowflake secures your data at each step along the way.

Reduce total cost of ownership without any compromises between cost and performance. Snowflake "just works." It can handle a virtually unlimited number of users and jobs, manage data volumes of nearly any size, operate across a wide range of regions and cloud environments, and require near-zero maintenance for teams to focus on development vs. undifferentiated infrastructure tuning and scaling.

Snowflake's consumption-based pricing, when paired with near-instant elasticity that scales up and down as your needs change, provides the best performance at minimal cost. Snowflake also offers regular platform improvements and built-in cost optimizations, providing transparency and cost predictability.

Maximize access and distribution of live, ready-to-query data, while also building cross-cloud resiliency. Snowflake enables organizations to break down data silos and reduce the complexity associated with accessing and distributing data securely across regions and clouds. Snowflake's data sharing also reduces risk to the business; the platform's ability to ensure replication and failover across cloud providers during unplanned business interruptions makes it easy to move data between rival cloud providers such as Amazon Web Services, Microsoft Azure and Google Cloud Platform.

SNOWFLAKE PROVIDES A POWERFUL PLATFORM TO BUILD UPON

Once you've established a solid foundation for your data, you can begin to strategically unleash the power of gen AI for your organization.

The Snowflake cloud platform offers direct access to all raw and curated data, using open formats. You can combine first- and third-party data in a single repository, with sufficient flexibility and control to preserve privacy and ensure effective data governance. Snowflake's secure infrastructure enables enterprises to feed their data into open-source or third-party LLMs and build applications for internal use, putting gen AI and machine learning capabilities into the hands of the entire business.

Here are some of the services available on the Snowflake platform*:

Snowflake Cortex is an intelligent, fully managed service that hosts and serves industry-leading AI models, LLMs and vector functions. It allows you to quickly and securely build AI applications using your enterprise data, generating predictions and insights.

Document AI is a purpose-built multimodal LLM that is natively integrated into the Snowflake platform. It allows customers to extract analytical content from documents and fine-tune the results using natural language and a visual interface.

Snowflake Copilot is an LLM-powered assistant that allows users to generate SQL queries using natural language, then filter down to the insights most relevant to the tasks at hand.

Universal Search allows users to search and discover tables, views, databases, schemas and as well as offerings in the Snowflake Marketplace.

Streamlit is a quick and easy way to build friendly user interfaces for LLM-powered apps. This OSS Python library allows developers to deploy apps at scale, reliably and securely.

*Features and capabilities within each of these categories may be in private or public preview, or may be different than described at time they become generally available.

With AI and data science expertise in increasingly short supply, Snowflake allows organizations to take advantage of advanced analytics without the need for deep AI knowledge or complex integrations. Prebuilt user interfaces and SQL/Python queries do all the heavy lifting, enabling users to get answers to their questions within seconds.

Development teams can build lightweight interactive apps that help put their data to work, creating custom multi-service apps using a common data foundation without risking a loss of intellectual property or introducing new vulnerabilities. Organizations can then share these apps securely with internal teams, business partners and customers across all the major cloud environments.

To learn more about the impact gen AI and other developments dive into our Generative AI & LLM School, where both executives and developers can learn about the power of LLMs and Generative AI on your data.

ABOUT SNOWFLAKE

Snowflake enables every organization to mobilize their data with Snowflake's Data Cloud. Customers use the Data Cloud to unite siloed data, discover and securely share data, and execute diverse artificial intelligence (AI) / machine learning (ML) and analytic workloads. Wherever data or users live, Snowflake delivers a single data experience that spans multiple clouds and geographies. Thousands of customers across many industries, including 647 of the 2023 Forbes Global 2000 (G2K) as of October 31, 2023, use the Snowflake Data Cloud to power their businesses.

Learn more at **snowflake.com**