Whitepaper

Exploring the Future of GenAI with Snowflake

by ***Sujith Gopalakrishnan*** *I* ***Su Dogra***

LTIMindtree

In the realm of AI, generative AI's capabilities have started to impact industries significantly. Our research report[i] points out that 49 % of leaders would use Gen AI to 'enhance data and analytics.' It also found that 62 percent of organizations in the UK and continental Europe are allocating 5 to 10 percent of their IT resources to generative AI projects[ii].

Snowflake has rolled out many GenAI functionalities to empower enterprises to utilize GenAI capabilities and also has enabled a platform to build futuristic custom GenAI apps. Snowflake platform is enabling and reshaping the development of generative AI applications for the future, redefining boundaries in data monetization, and facilitating the resolution of diverse business challenges, thus adding significant value to the business.

In this Whitepaper, we explore the harmonious synergy between GenAI and Snowflake and will focus on

01 ***What services does Snowflake offer to enable generative AI?***

02 ***How does Snowflake facilitate the essential aspects required for a platform to build generative AI applications?***

03 ***How can Snowflake and generative AI add value to the business?***

04 ***How can Snowflake Data Cloud and generative AI unlock unlimited possibilities?***

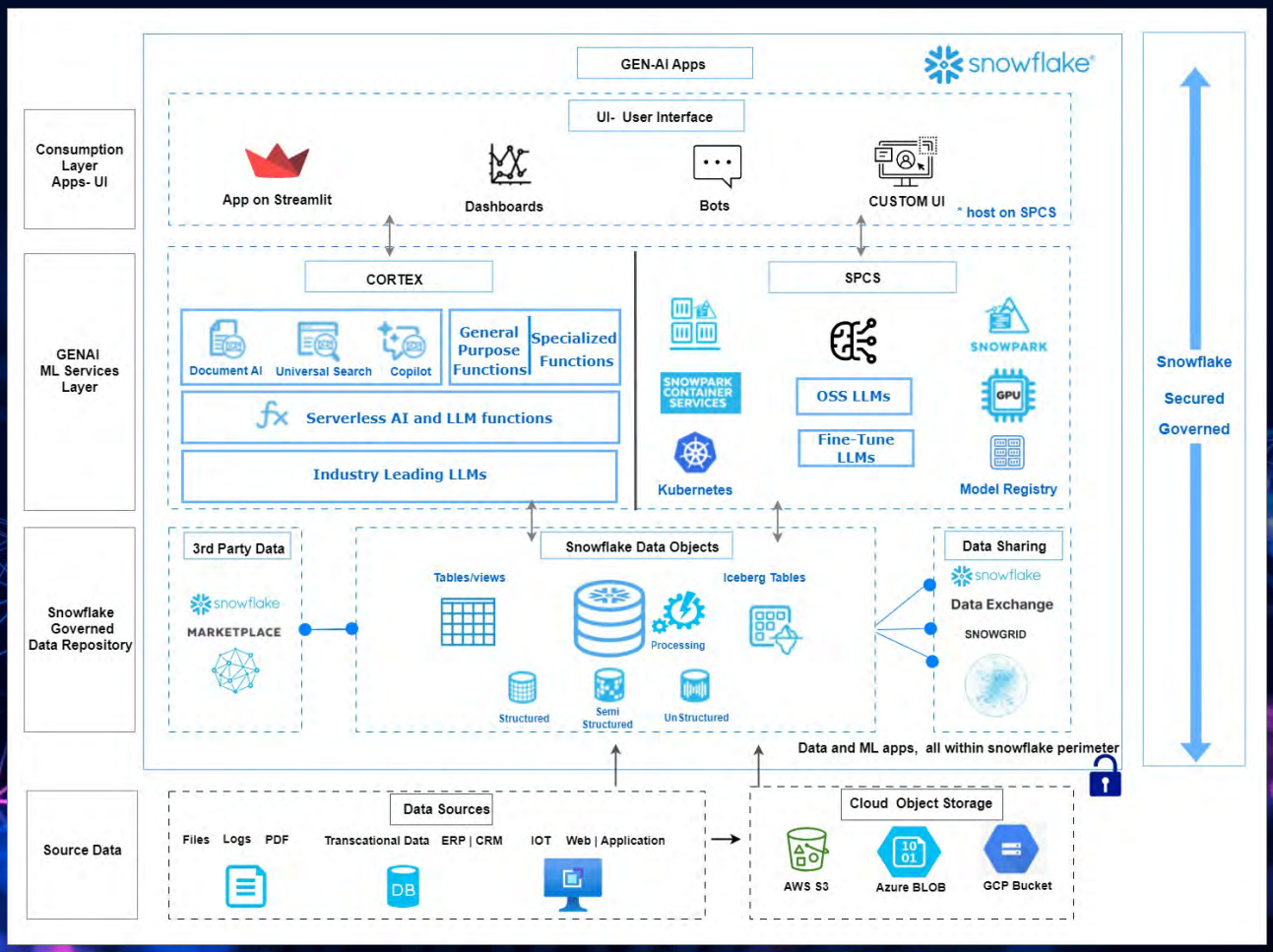
What services does Snowflake offer to enable generative AI?

Snowflake Data Cloud provides you with the flexibility to analyze data and build generative AI apps through:

- A fully managed service, CORTEX, which includes access to industry-leading large language models with out-of-the-box capabilities for data analysis and the rapid development of generative AI apps.
- A platform utilizing Snowpark Container Service (SPCS) to create customized generative AI apps with self-hosted Large Language Models (LLMs) within Snowflake. The platform allows you to host Open Source Software (OSS) LLMs, fine-tune, deploy, and build fully customized generative AI apps.

CORTEX and SPCS functionalities are Public Preview (PuPr) features at the time of publishing this article.

Both the Snowflake services provide the advantage of harnessing generative AI capabilities and building generative AI apps entirely within the Snowflake environment, eliminating the need to send your enterprise data outside Snowflake. This ensures the security and governance of your data and apps.



"Figure 1: An end-to-end reference architecture integrating Snowflakes Data, AI and Apps capability to build GenAI apps

The following matrix serves as a reference guide to determine the suitable service for building your generative AI use case within Snowflake.

	Snowflake Cortex	Platform using SPCS
Persona	Enables data engineers, analysts, and business users with limited AI skills to quickly analyze data using out-of-the-box generative AI capabilities or build AI apps.	Enables platform for ML engineers and data scientists to build fully custom generative AI apps quickly.
Capability	Fully managed service with in-house LLM provides the capability to analyze data or build apps, including: **Serverless LLM Functions:** • Sentiment analysis, extracting answers, summarizing text, translation to a supported language **Services for:** • Implementing search functionality on your Snowflake data objects to democratize data, apps, and marketplace data. • Code generation using natural language. • Extracting insights from unstructured documents. **Cortex ML-Based Functions:** ML-based functions give you automated predictions and insights into your data using machine learning, a few of them being: • Forecasting. • Anomaly Detection. • Contribution Explorer	Platform to bring your preferred LLMs and build 3rd party apps, including the ability to: • Host foundational LLMs • Fine-tune LLMs with your data • Deploy LLMs • Build apps using Streamlit or custom UI, which developers could build and package using code in any programming language (e.g., C/C++, Node.js, Python, R, React, etc.)

	Snowflake Cortex	Platform using SPCS
Operational focus	Managed service with serverless functions that eliminate the need for infrastructure management and Graphics Processing Unit (GPU) planning. Snowflake manages the LLMs.	Unified experience in the end-to-end lifecycle of containerized applications and AI/ML models. Reduces dependency on multiple platforms to deploy, manage, and scale containerized workloads with configurable hardware options, such as GPUs.
Use case examples	Building applications on Snowflake tables that can perform various Snowflake cortex features like summarizing, translating, extracting answers, etc., from given data.	Generative AI apps for industry-specific use cases ranging from interactive chatbots generating insights, enabling data search and discovery, and addressing varied use cases defined in the last section of this blog.
LLMs	Industry-leading LLMs like Llama2 are accessible through Snowflake Cortex. They are self-contained models ready for immediate use.	You can host OSS LLMs, custom-built LLMs, or fine-tune LLMs with your data or partner LLMs within Snowflake.
Security, governance, and cost	Secured as your data doesn't leave the Snowflake perimeter in both cases. Governed, cost-optimized, and applicable for all services across Snowflake	

How does Snowflake facilitate generative AI applications?

When you are looking at building third-party customized generative AI apps using your preferred OSS LLMs, customized with data from Snowflake, Snowflake addresses vital considerations necessary for developing generative AI-related apps using SPCS.

- **Unified platform:** One platform that handles data lake and analytics while hosting generative AI apps.
- **Cost-effective generative AI platform:** Host the LLM model within Snowflake to leverage its cost benefits, including the pay-as-you-go model and compute-optimized warehouse. By extending with functionalities of external tables and iceberg tables, access data from cloud storage, avoiding data replication and reducing cost.
- **Secured platform:** Securely store your data and LLM within the Snowflake perimeter, eliminating the necessity to expose your data outside your data lake for accessing 3rd party LLMs hosted on other platforms.
- **GPU-based processing capability:** Snowflake provides GPU-based clusters for executing models, fine-tuning LLMs, and inference.
- **Ability to fine-tune LLM models:** The Snowflake platform, with SPCS, empowers users to customize and fine-tune open-source LLMs with their enterprise's proprietary data. This capability enables the creation of task-specific LLMs, which are often more cost-effective, smaller, and faster for inferencing.
- **Model training, registration, and deployment:** Snowflake facilitates the fine-tuning of LLMs and addresses the hosting of model registry, vector DB, and LLMs—all within Snowpark container services. This enables the customization and management of any source LLMs within the Data Cloud.
- **User experience Layer/Application:** Utilize Streamlit alongside native Snowflake services or build a custom UI/app that uses any programming language to run with SPCS within your Snowflake account, creating user-friendly, custom LLM-powered apps.

When seeking a platform capable of empowering Generative AI capabilities with instant access to industry-leading large language models (LLMs), Snowflake addresses the key considerations essential in building Generative AI applications through Cortex.

- **Out-of-box capability in the platform:** Addressed by Cortex, a fully managed service that provides access to industry-leading AI models.
- **Rapid LLM app development:** Snowflake Cortex offers a managed service and serverless functions that can be used for inference on leading LLMs using functions like Llama2 to build apps quickly.

- **Reduced dependency on niche skills:** Cortex eliminates infrastructure management overhead and provides access to industry-leading AI models, offering ready-to-use generative AI capabilities for data analysis. This simplifies app-building for data engineers and individuals with limited AI expertise.

- **Elimination of infrastructure overhead:** A managed service with self-contained LLM, eliminating the necessity for hosting LLM, vector DB, or managing GPU infrastructure.

- **In-built services for utilizing generative AI capabilities with your data:** Cortex provides capabilities for answering questions, searching data objects, and generating insights and provides ML-based functions that use machine learning to detect patterns in your data so that you don't have to be a machine learning expert. These functions can help train the ML models on your time series data and provide predictions for ML use cases related to forecasting, anomaly detection, and contribution exploration.

*Cortex has a focused roadmap, and you would find new capabilities and functionalities, being rolled out soon.

How can Snowflake and generative AI add business value?

Enable data monetization

- Build apps that provide insights by answering questions, generating content, and enabling engaging conversations with your stakeholders or clients while supporting data-driven recommendations based on your data.
    - Data apps: Leveraging generative AI with enterprise data can enable building apps on 'data as an asset,' allowing you to establish a subscription-based model for generating recurring revenue.
    - Data insights as an asset: By combining Snowflake's data with marketplace data and the capabilities of LLMs, you can generate content from previously untapped datasets, creating new apps and so new opportunities for generating revenue with your clients, vendors, and partners.
    - Improve customer satisfaction: Gain the capability to answer questions based on your enterprise dataset, thus enhancing conversational insights and enabling a deeper understanding of your customers to address their needs.

Unlock new insights

- Enable users to analyze structured and unstructured data, helping organizations attain new insights and recommendations to make data-driven decisions.
- Finetune OSS LLMs with your enterprise dataset and Snowflake marketplace data to unlock new insights and generate recommendations for running new campaigns and marketing initiatives.

Automate tasks with bots to minimize manual tasks

- Extraction of data/insights from documents such as invoices or receipts
- Chatbots to provide information on data and answer queries from stakeholders
- Automate daily data analysis tasks and monitor data for anomalies
- Summarization of insights or content creation from the organization dataset

Fuel innovation

- Enable deep conversational insights and recommendations to stakeholders, expediting decision-making.
- Build custom generative AI apps to drive new business ventures.

Improve customer satisfaction

- Personalization: Generative AI can personalize user experiences, such as recommending products, content, or services.
- Generate content based on user behavior, preferences, and trends to enable personalization.

Data democratization

- Utilize Cortex search functionality to find database objects, tables, data products, and Snowflake Native Apps from the Snowflake Marketplace.
- Empower your analysts and stakeholders with less AI skills to analyze data using Generative-AI or build apps quickly to generate insights that were inaccessible earlier.

Enable secured, compliance applications

Hosting LLMs outside your Snowflake or data perimeter poses the risk of exposing proprietary data, as it would have to leave Snowflake. To address enterprise security and compliance requirements, Snowflake brings LLMs to your data and hosts Gen AI apps within the Snowflake perimeter.

How can Snowflake Data Cloud and generative AI unlock unlimited possibilities?

Snowflake Data Cloud provides various capabilities to persist, process, and enrich your enterprise data effectively. Combining Snowflake Data Cloud's functionalities with Generative-AI would enable new data ventures in Data. Below are a few possibilities for integrating native Snowflake capability with LLMs to enable varied generative-ai use cases:

1. **Organizations Data in Snowflake Data Cloud (Native Tables, Iceberg Tables on DataLake, Snowflake Data Marketplace) + Chatbot with Cortex + Streamlit for UI all in Snowflake:**
    A. Enables you with AI bots to answer stakeholder questions.
2. **Organizations SQL Code + OSS LLM hosted in Snowflake perimeter finetuned with your code:**
    A. Can assist in generating code snippets and aiding Data Engineers
    B. Could enforce enterprise standards and style guides, thereby implementing best practices and simplifying coding and data engineering.
3. **Organizations data + Snowgrid + LLM:**
    A. Enable true Global insight generation by deriving recommendations from data products across domains and geographies.
    B. Integrate multiple data products whose data persists in various regional languages through language translation from one language to another, generating insights that persist across data products, enabling true enterprise insights, and facilitating cross-cultural communication.
4. **Organizations data in Snowflake Data Cloud + LLM hosted in Snowflake + Snowsight:**
    A. Generate Insights through advanced visualization and recommendations stitching your data with 3rd party external data
5. **Snowflakes data privacy capabilities + LLM:**
    A. Implement generative AI techniques, such as federated learning or differential privacy, to extract insights from distributed data sources without moving the data itself.
    B. Data Clean Room (DCR): Empower insights generated by data brought from multiple companies or divisions of a company in a secure and controlled environment through DCR. Apply LLMs on insights derived to weave new recommendations for advertising and marketing.

Conclusion

When embarking on the journey to enable generative AI in your organization, Snowflake has been empowered to enable the platform that's required to cater to a broad range of personas, from analysts to data scientists. It offers wide offerings, enabling the utilization of out-of-the-box Gen AI features to facilitate the creation of custom apps. A key differentiator is that Snowflake allows you to address diverse workloads, encompassing data lake, analytical, and Gen AI needs, all within a unified platform. This eliminates the risk of moving data outside your enterprise for AI requirements. With Snowflake's native capabilities and generative AI services, you can build multiple possibilities, enabling monetization and adding value to your business.

About the author

Sujith Gopalakrishnan

Managing Principal Architecture
Field CTO, Snowflake Business Unit,
LTIMindtree

Sujith is a Chief Architect with nearly two decades of experience in architecting and building data and analytical applications. He is a TOGAF Certified Enterprise Architect with deep expertise in the areas of data governance, consulting, and implementing data platforms on AWS and Azure. He is also a Snowflake SnowPro Advanced Architect.

Su Dogra

Senior Partner Sales Engineer,
Snowflake

Su Dogra, Senior Partner Sales Engineer at Snowflake focused on supporting Snowflake's partnership with System Integrators.

She has been helping System Integrators with their Data Migration, Data Engineering, and AI/ML efforts. Before joining Snowflake, she worked in the Data Engineering space as Data Architect.

[i] The State of Generative AI Adoption: The Current Landscape and Lessons from Early Adopters (Global Report), LTIMindtree, November 2023: https://www.ltimindtree.info/gen-ai/

[ii] Gen AI Revolution Roadmap – UK and Continental Europe's Journey Unveiled, The State of Generative AI Adoption: The Current Landscape and Lessons from Early Adopters (UK & Continental Europe), LTIMindtree, November 2023: https://www.ltimindtree.info/gen-ai/