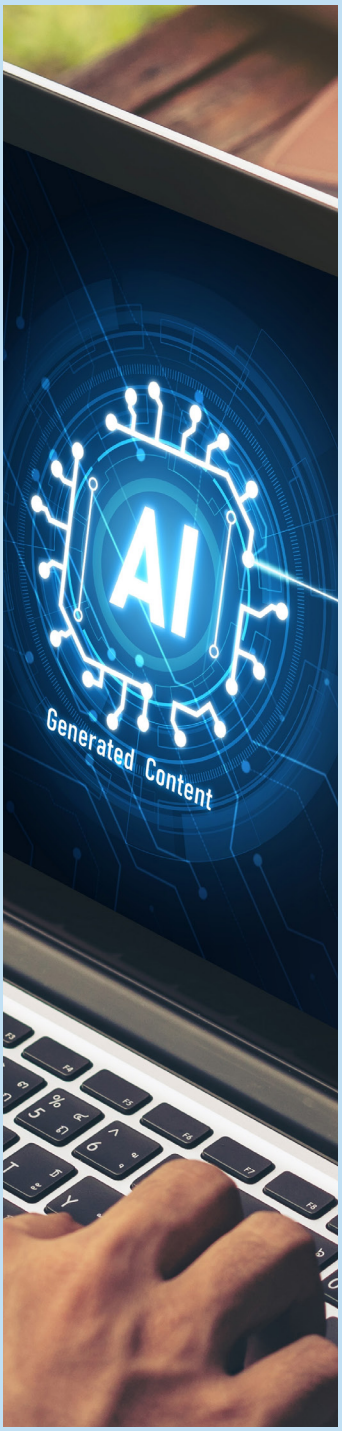
Infosys topaz

VIEW POINT



SCALING GENERATIVE AI
THE PLATFORM WAY

Abstract

Enterprises are in a race to embrace generative AI for competitive advantage. However, adopting this technology without incorporating aspects of efficiency and scale could lead to breakdowns during implementation with losses to the business. In this paper, we discuss how a unified, platform-centric approach can lay the foundation for a robust generative AI strategy that can be easily adopted and scaled for the benefit of the organization and all its stakeholders.

Infosys®
Navigate your next

Table of Contents

Introduction

Even as generative AI (GenAI) makes waves across the world of business, not everyone is on board. While many companies see immense value in adopting it, some remain unconvinced. Questions remain around whether organizations should dive headfirst and embrace this revolutionary technology or wait for proven results. One thing is for certain - Generative AI is here to stay. While early adoption might hold the key to staying ahead, a cautious and strategic approach is crucial.

Scaling GenAI effectively across a large enterprise requires a well-defined strategy. Proponents argue it can boost efficiency and creativity but implementing it organization-wide poses multiple challenges. Opponents – fearing that the hype surrounding GenAI overshadows potential drawbacks – recommend a wait-and-watch approach.

Finding the right balance is key. Organizations must assess their needs and develop a plan to leverage GenAI strategically. Rushing into adoption without a plan can lead to wasted resources and missed opportunities. A centralized approach ensures consistency and avoids duplication of efforts when implementing this technology. With careful planning and a measured approach, organizations can successfully harness the power of GenAI to gain a competitive edge.

The Rise of Generative AI in Enterprises

Over the past couple of years, there has been a surge in enterprise adoption of GenAI for various purposes. From search enhancements to generating insights, organizations have experimented with a wide range of use cases. This decentralized approach has fostered innovation and awareness about AI technology. However, leaders now recognize the need for strategic planning to scale GenAI effectively.

Having collaborated with multiple clients for generative AI implementation, Infosys has gathered vast experience and expertise as well as a range of best practices to help organizations embrace GenAI at scale. By adopting our platform-centric approach, businesses can unlock the full potential of this transformative technology.

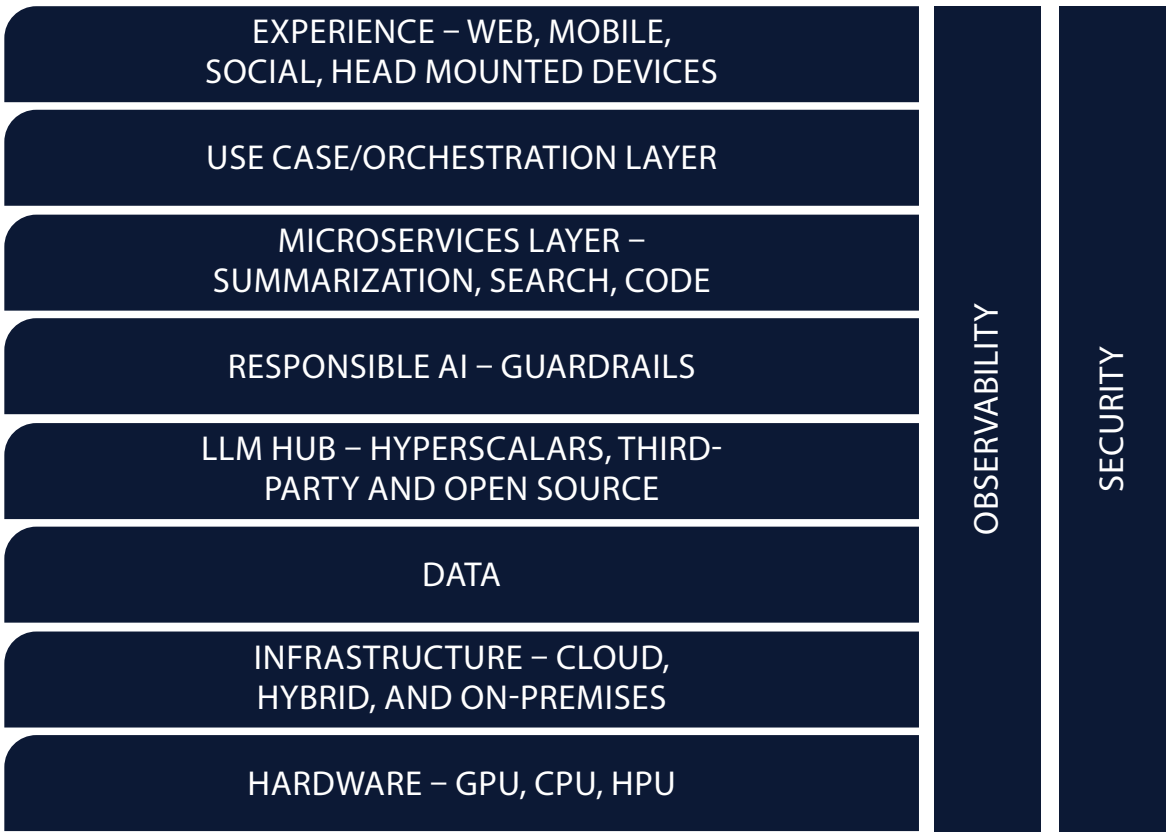


Figure 1: A logical view of an enterprise GenAI platform

An enterprise platform for GenAI does not require building everything from scratch. Instead, the platform brings together different AI tools and resources in a way that is easy to use and control. Advantages of using an enterprise GenAI platform include:

Efficiency:

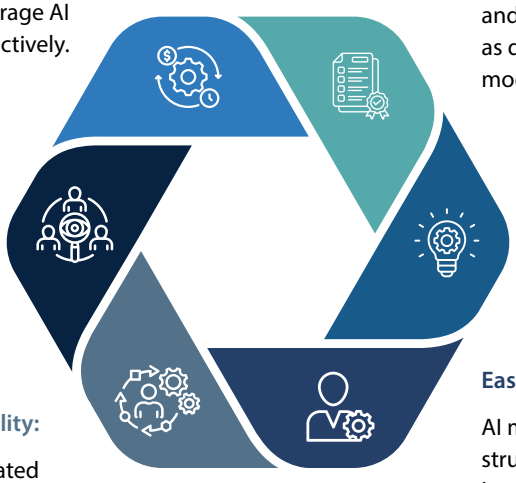
Organizations can leverage AI capabilities quickly and effectively.

Transparency:

Everyone in the company understands how AI is being used.

Adaptability:

The platform can easily be updated with new models and hardware.



Standardization and guardrails:

The business can enforce company-wide rules and best practices for AI development and deployment. This includes aspects such as data security, prevention of hallucination, model bias detection, and explainability.

Futureproofing and innovation:

IT teams can stay up-to-date with latest AI advancements and integrate them seamlessly. This ensures the organization remains competitive and leverages the latest advancements.

Ease of management:

AI microservices is an architectural style that structures an application as a collection of loosely coupled, independently deployable services. Each service is responsible for a specific business capability. Using microservices, accessing and utilizing the platform becomes much simpler. Users can focus on their specific tasks and interact with the appropriate microservice to get their job done. This approach also simplifies managing the AI landscape by allowing each microservice to independently handle a distinct functionality.

Key Considerations For an Enterprise AI Platform

Before building and deploying an AI platform, organizations must conduct necessary due diligence keeping these five key factors in mind.

1.Technology

In the rapidly evolving field of generative AI where there is no single dominant player, maintaining flexibility and avoiding dependence on specific vendors is important. From startups to large hyperscalers to in-house custom models, AI technology providers are constantly innovating and introducing new solutions. To navigate this dynamic landscape, organizations should adopt a poly AI approach when building their AI platforms.

For poly AI adoption, it is essential to create a layer of abstraction that acts as a buffer between the underlying technological complexities and organizational requirements. The abstraction layer enables organizations to:

- Seamlessly switch between different technologies
- Benefit from best-of-breed solutions
- Mitigate vendor lock-in
- Stay adaptable
- Optimize performance
- Maintain control

Infosys' Applied AI Cloud helps democratize AI and scale the adoption of GenAI efficiently. It specifically addresses the platform needs of organizations by bringing together and enabling access to AI hardware, open-source AI software-as-a-service on hybrid cloud infrastructure, and edge AI capabilities.

2.Responsible AI

The generative AI journey is fraught with challenges, including ethical dilemmas, legal complexities, transparency issues, biases, IP violations, and traceability concerns. These obstacles have been further aggravated by cybersecurity threats such as prompt injection, jailbreaks, profanity, and toxicity, keeping security teams in a perpetual state of vigilance.

Considering these challenges, it is crucial for organizations to adopt a proactive stance and establish a robust policy framework around responsible AI. Defining a policy framework will help simplify the complex landscape and enable organizations to navigate these issues with greater ease and clarity. Moreover, a platform-centric approach to responsible AI can be a game-changer, offering a centralized GenAI platform that ensures organization-wide adherence to responsible AI practices.

Infosys champions a forward-thinking responsible AI first (RAI-first) philosophy, positioning itself at the forefront of the RAI domain and actively shaping the future through collaboration with industry bodies, governments, policymakers, and influencers. As a trailblazer in this space, Infosys has received the industry-first ISO 42001:2023 Artificial Intelligence AI Management Certification, underscoring its commitment to excellence and leadership in AI management practices. Further solidifying its dedication, Infosys has established a Responsible AI Office, a testament to our pledge to become an RAI-first enterprise.

The establishment of a clear, proactive policy for responsible AI, coupled with a platform-centric approach, is not just a strategic imperative but a necessity for organizations aiming to harness the full potential of AI responsibly and ethically. It paves the way for innovation with integrity, ensuring that AI serves the greater good while aligning with the organization's strategic objectives and societal norms.

3.Cost

In the realm of generative AI, the total cost of ownership (TCO) is a pivotal factor for organizations considering the adoption of AI at scale. While the allure of generative AI's benefits is undeniable, the financial implications of implementing, maintaining, and sustaining generative AI initiatives cannot be overlooked. A siloed approach to generative AI may incur inefficiencies, rendering the investment less cost-effective. However, a centralized generative AI platform can streamline processes, yielding cost efficiencies and traceability that enhance the overall value proposition.

A centralized approach not only consolidates resources and expertise, leading to better management of costs but also provides a clear trajectory for scaling AI initiatives. It allows for the meticulous tracking of expenses, ensuring that every dollar spent is accounted for and contributes to the organization's strategic goals. In an era where CXOs are under pressure to integrate generative AI, a well-articulated TCO may not be the deciding factor, but it certainly equips organizations with the insights needed to make informed decisions. Ultimately, a centralized GenAI platform can transform the cost narrative from a potential barrier to a strategic advantage, enabling responsible and sustainable adoption of AI technologies.

4.Sustainability

The sustainability of generative AI, particularly in the context of building a centralized platform for an organization to adopt AI at scale, is a multifaceted issue that encompasses energy efficiency and carbon emission goals. The hype around GenAI often overlooks the significant energy requirements not only for training LLMs but also for inference activities. The lifecycle of these activities could potentially impact the environmental, social, and governance (ESG) goals of many organizations.

A well-thought-out strategy to mitigate this risk is essential. The choice of models will play a crucial role. Opting for power-efficient CPUs over GPUs for certain generative AI tasks could lead to substantial energy savings. Centralized management of these resources will help organizations to measure and maintain their ESG targets effectively.

5.Observability

For large enterprises employing generative AI at scale, observability ensures comprehensive monitoring and management of systems. It enables the tracking of performance metrics, system health, and unusual patterns in real-time, which are crucial for maintaining the integrity and efficiency of AI models.

Implementing observability in an enterprise platform helps:

- Monitor costs effectively to ensure that resources are utilized efficiently, and expenses are kept under control.
- Enhance security by providing detailed oversight of system activities to help detect and mitigate potential threats promptly.
- Gain insights through analytics to drive informed decision-making and strategic planning.
- Automate alerts to respond quickly to changes or issues, minimizing downtime and maintaining system integrity.
- Ensure data security by tracking access and modifications to safeguard sensitive information against unauthorized use.

By integrating these five elements, a central platform can support generative AI adoption at scale, providing a robust and transparent environment that fosters innovation while maintaining operational excellence.

Benefits of the Platform Approach

The unification of diverse components within a single platform has the potential to deliver substantial advantages. Organizations can garner significant rewards from the amalgamation of capabilities and resources, yielding a spectrum of both quantifiable and qualitative benefits as under:

Scalability and flexibility in a central platform for GenAI adoption at scale offer significant benefits. This includes unified infrastructure to avoid building separate systems for each use case, leading to improved resource efficiencies and lower costs.

Consistency and integration are achieved through standardization of data handling, model training, and deployment, along with seamless integration of AI services across use cases.

Cost efficiency is realized through economies of scale, making it more cost-effective than separate solutions. In addition, centralized management reduces maintenance overhead and saves operational and capital expenses.

Accelerated innovation is facilitated by rapid deployment capabilities and a sandbox environment for experimentation and testing.

Enhanced security and compliance are ensured with unified security protocols and simplified auditing, improving adherence to regulations.

Improved data management comes from consolidated data repositories and advanced analytics for better decision-making.

Enhanced user experience is provided by a unified user interface and a consistent user journey across applications, improving usability and adoption rates.

Conclusion

The adoption of a centralized platform-based approach to integrate generative AI within an enterprise is critical for ensuring efficiency, transparency, adaptability, and standardization. This future-proof strategy establishes the guardrails necessary to foster innovation while consolidating benefits such as scalability and cost efficiency. By centralizing AI efforts, enterprises can leverage responsible AI practices, ensuring ethical and accountable use of technology. The centralized platform serves as a foundation for sustainable growth, enabling organizations to respond swiftly to evolving market demands and technological advancements. Ultimately, this approach not only streamlines operations but also empowers enterprises to harness the full potential of GenAI, driving significant competitive advantage in the digital era.

At Infosys, we have collaborated with global clients across industries to enhance the integration of generative AI across their enterprises with a platform-centric strategy. For a prominent telecommunications company, the establishment of a unified AI platform has led to significant cost savings and accelerated the adoption of GenAI. For another major client in the United States, the creation of an enterprise AI platform has facilitated the widespread use of generative AI in their software development life cycle (SDLC), chatbots, and contract management processes.

Infosys Topaz plays a pivotal role in actualizing the concept of a platform for the integration of generative AI. It offers essential best practices, technological resources, reference architectures, and other vital tools that can expedite the development and implementation of such a platform.

About the Author

Guruprasad NV

AVP, Senior Principal Technology Architect

Guruprasad N V (Guru) is an Associate Vice President at Infosys. he has been with Infosys for over 18 years and currently leads the Hyperscalers partnership and solutions team within Infosys Topaz, the company's flagship offering focused on AI-first services, solutions, and platforms built using generative AI technologies.

He strategically engages with Hyperscalers to secure early access to their advanced technologies, facilitating the development of innovative solutions through collaborative partnerships. he coordinates joint go-to-market strategies to effectively launch these solutions and leverages deep insights into the latest AI technologies.

Infosys Topaz is an AI-first set of services, solutions and platforms using generative AI technologies. It amplifies the potential of humans, enterprises and communities to create value. With a vast repository of AI assets, pre-trained AI models, 10+ AI platforms steered by AI-first specialists and data strategists, and a 'responsible by design' approach, Infosys Topaz helps enterprises accelerate growth, unlock efficiencies at scale and build connected ecosystems.

For more information, contact askus@infosys.com

Infosys®
Navigate your next