Enterprise Strategy Group by TechTarget

SHOWCASE

SEPTEMBER 2024

Developing and Deploying Cloud-native GenAI Solutions at Scale

Mike Leone, Practice Director and Principal Analyst

Abstract: Developing and deploying generative AI (GenAI) applications offers numerous challenges, many of which center on having access to high-performance, cost-efficient hardware and software resources. In order to develop, accelerate, and scale cloud-native AI applications, organizations need a purpose-built cloud environment to speed time to value for GenAI use cases.

Introduction: Accelerating GenAI Adoption for a Widening Range of Use Cases

AI in general, and GenAI specifically, has quickly crossed the threshold from promising technology to a transformative set of solutions delivering tangible value. According to TechTarget's Enterprise Strategy Group, more than a third (34%) of organizations already believe AI is fully embedded in their culture, and another 27% said they are expanding AI more broadly across the business.[1] But as organizations ramp up their GenAI development efforts and put new solutions to work, pressure is mounting for organizations to derive tangible value from their AI projects and to unearth insights that deliver sustainable business benefits.

Perhaps the most impressive sign of GenAI's accelerated adoption rates is the growing number of diverse use cases where the technology is used to solve a wide range of problems and address new business opportunities. Enterprise Strategy Group research pointed out that data insights, chatbots, workforce productivity, content creation, and business process improvement are the top priorities for GenAI projects and use cases.[2]

Market Insights

The top five priorities of GenAI projects:

1. Data insights.
2. Chatbots.
3. Employee productivity and tasks.
4. Content creation.
5. Business process improvement.

Developing GenAI Apps Can Be Complex, Time-consuming, and Expensive

Even with the promising evolution of GenAI in real-world settings, organizations are pushing hard for ways to accelerate time to value from their investments. The problem is with the number of challenges that organizations must face, seeing value quickly can often feel like a pipe dream. Between skills gaps, data quality issues, and security concerns, there is a lot to think about when pursuing GenAI initiatives. One of the biggest areas of concern when it comes to GenAI focuses on the underlying infrastructure stack and computational resource requirements to

[1] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.
[2] Source: Enterprise Strategy Group Research Report, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

support it. In fact, nearly 40% of organizations cited computational resource requirements for GenAI techniques and integration complexity with existing infrastructure and tools as their top concerns when building GenAI applications.[3]

Market Insight

Nearly 40% of organizations cited computational resource requirements and integration complexity with existing infrastructure and tools as their top concerns when building GenAI applications.

How can developers expect to build efficient, high-performant GenAI applications powered by large language models if they don't have access to the right resources to develop, test, train, and tune them? Simply put, they can't. As demands on the model increase or as more models are put into production, applications must scale to meet the demand. The underlying infrastructure must accommodate distributed and diverse computing requirements, data storage, and real-time inferencing, all of which must be monitored to maintain consistent performance and a positive end-user experience. In this case, the scalability challenges are quite clear.

The complexity of the infrastructure are heightened by the need for specialized hardware, which not only incurs high costs but also demands expertise for optimization and maintenance. When developing GenAI applications, accessibility and cost considerations present significant challenges. The initial investment can be substantial, as the necessary infrastructure often requires expensive hardware, software licenses, and cloud resources. Additionally, the financial burden doesn't stop after deployment; ongoing expenses for cloud services, energy consumption, and the upkeep of specialized infrastructure can accumulate quickly. This makes effective cost management a constant concern for developers, potentially limiting access to these powerful technologies, especially for smaller organizations or startups.

Then comes the struggle with ensuring environment consistency and managing system integrations through the development and deployment of GenAI solutions. Development to production is often hindered by differences in infrastructure configurations, which can result in performance discrepancies when models, as well as the applications that rely on the models, are transitioned to production. Dependency management across various environments can lead to conflicts and necessitate rigorous testing and validation. The integration of GenAI with legacy systems further complicates matters, as it often requires custom interfaces or substantial modifications to existing infrastructure.

And much of this occurs before addressing the security and ethical concerns associated with GenAI, an area in which many developers have limited experience. Bias mitigation, responsible use of generated content, and monitoring model manipulation all pose challenges, especially as organizations try to navigate evolving regulatory guidelines. Addressing these challenges requires not only a robust understanding of AI methodologies and software development practices but also ongoing training and skill development to deliver viable solutions that both the business and their customers can trust.

The Essentials for Reliably Developing GenAI Apps

To overcome these and other challenges, many organizations are evaluating and embracing flexible environments designed specifically for GenAI application development and deployment. This approach enables both internal and third-party developers to leverage best-in-class hardware and software resources that are essential to rapid, secure, and cost-efficient GenAI development and deployment.

So what should organizations be considering as they look for the ideal GenAI development environment?

[3] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

- **Robust infrastructure.** Access to scalable compute resources, including CPUs, GPUs, and accelerators to manage diverse and demanding GenAI workloads, whether training and/or inferencing. Utilizing a cloud services architecture enables flexible, on-demand resource availability, while tools that offer containerization and orchestration help manage microservices and maintain environment consistency throughout the development, testing, and deployment of GenAI applications.
- **Model accessibility and support.** Model accessibility and support are vital to support the custom needs of businesses across industry spectrums. Developers must be enabled to efficiently integrate, modify, and deploy GenAI models across applications and environments. Ensuring that models are easily accessible, whether third-party developed, proprietary, or open source, can enable development teams to focus on building applications rather than managing model logistics. And while ensuring support for the training of models is important to those looking to build custom models, using pre-trained models can greatly save time and reduce costs. Developers can then be empowered to fine-tune these models on specific data or incorporate enterprise data using retrieval-augmented generation, enabling customization without starting from scratch.
- **Observability and automation tools.** Implementing CI/CD pipelines for GenAI development can streamline the deployment process by automating re-training, testing, and monitoring. This helps to facilitate the rapid iteration and deployment of models at scale. Access to techniques like model versioning tools can also enable improved management of different model versions and experiments across environments, simplifying the tracking of performance, accuracy, and reliability over time.
- **Data management.** Efficient data labeling and preprocessing tools are essential for quickly preparing datasets for GenAI use. Techniques like data cleaning, labeling, and augmentation are step one, but areas like automated data pipeline creation and management can streamline workflows by facilitating data ingestion, processing, and storage.
- **Security, compliance, and responsibility.** Table stakes to any development environment, key security features like encryption, authentication, and data anonymization to comply with evolving data and AI regulations are must-haves for any development team building GenAI applications. This extends from development to deployment, regardless of where these applications reside, whether on premises, in the cloud, or hybrid. Protecting sensitive data through the GenAI development cycle is critical, especially as organizations prioritize the responsible and ethical development of AI more than ever before. In fact, 74% of organizations believe responsible AI is one of, if not the most important priority as they pursue GenAI initiatives.[4]

Market Insight

74% of organizations believe responsible AI is one of, if not the most important priority as they pursue GenAI initiatives.

- **Cost management.** Implementing budget-tracking tools and cost estimation frameworks enables organizations to monitor spending in real time, identify inefficiencies, and optimize resource allocation. This is especially important as organizations just get started with generative AI application development. Furthermore, leveraging scalable, cost-effective, and right-sized infrastructure can dramatically reduce costs while maintaining performance.

One aspect of GenAI development that should not be overlooked is the people—the developers themselves that must stay on top of a fast-moving industry. Between new models, new architectures, new use cases, and new regulations, organizations must ensure developers are collaborating and sharing knowledge. It starts by extending the basics of version control systems that can enable teams to track changes over time. And of course, providing comprehensive documentation and internal training resources not only facilitates faster onboarding but also streamlines the development processes. But close collaboration with subject matter experts will remain critical to ensure that model outputs align with business objectives and fulfill specific domain needs.

[4] Source: Enterprise Strategy Group Research Report, *Evaluating the Pillars of Responsible AI*, August 2024.

Empowering Development Teams With Intel Tiber Developer Cloud

Intel Tiber Development Cloud provides an enterprise-class solution for developers, data scientists, and AI engineers looking for cost-efficient, reliable, and scalable development and deployment of GenAI and other applications. This environment is available as a managed service and helps organizations develop, accelerate, and scale AI using Intel's high-performance AI infrastructure and optimized open source software: foundational models, frameworks, tools, and more.

Prediction Guard Uses the Intel Tiber Developer Cloud to Help Customers Mitigate Risks of LLM Application Development

Challenge: Large language model (LLM) output can be unpredictable, resulting in reliability risks and potentially higher costs. Hallucinations might occur, and security vulnerabilities such as prompt injections can result in data breaches or other threats.

Solution: Intel's cloud-enabled Prediction Guard to offer customers access to multiple LLMs while leveraging tools that address harmful inputs and outputs in a secure, private, cloud environment. The development teams could learn, test, and run applications on Intel-based hardware clusters, taking full advantage of Intel's high-performance AI infrastructure and the latest software development tools without having to make major Capex commitments.

Outcome: Developers fine-tuned their LLM development faster and with greater security, scalability, and reliability. They also were able to reduce costs using Intel Gaudi accelerators in Intel's cloud, as much as doubling processing throughput for some LLMs. This resulted in the faster, more cost-efficient development and deployment of new GenAI applications, improving developers' cost visibility and profitability.

For more information, check out the complete case study and video.

Developers working in Intel's cloud benefit in several ways, including the ability to build and deploy AI applications at scale, maximize AI compute infrastructure with high performance and cost efficiency, and leverage an open software/open platform model for choice and flexibility.

A key advantage for Intel is their ability to deliver customers a complete AI stack. In fact, they are the only compute provider that offers a full, vertically integrated stack that includes design, operations, hardware, and services, presenting a significant savings opportunity for customers. Further, with no charge for egress, organizations gain peace of mind knowing they have complete flexibility.

Intel Tiber Developer Cloud provides a specialized cloud environment for developers to build, experiment, and deploy GenAI solutions in a scalable, cost-efficient manner. It also enables organizations to develop and deploy anywhere with a standards-based programming framework and a vendor-agnostic hardware model that enables developers to select the right infrastructure from their AI workload. The service also promotes flexibility for developers and other users because cloud-optimized solutions can run in Intel's cloud, run in other cloud environments, or be migrated back on premises.

One of the key aspects of Intel's cloud is access to Intel's many market-leading AI hardware infrastructure solutions, including its Intel Gaudi AI accelerators. The new Intel Gaudi 3 accelerator offering is particularly important, as it delivers higher performance and more enhanced scalability than previous Gaudi versions.

Gaudi 3 is optimized for performance-intensive and energy-efficient GenAI workloads—both essential for compute-centric AI applications. It also supports a wide range of industry standards, such as open source software frameworks and popular foundation models, with employed Ethernet networking to improve scalability, reliability, and cost efficiency. Gaudi's AI-dedicated compute engine includes a heterogeneous architecture supporting 64 AI-custom and programmable Tensor Processor Cores and eight Matrix Multiplication Engines.

Conclusion

Even as GenAI becomes a more strategic and integral part of organizations' overarching technology framework, developers will continue to need to confront and overcome significant challenges to bring their solutions to a production state quicker and more cost-efficiently.

This means organizations will need purpose-built environments to help them develop, deploy, and scale GenAI solutions in order to keep costs under control, meet the intense performance requirements associated with GenAI development, scale their solutions rapidly, and ensure safe and secure treatment of sensitive and private data for GenAI models.

Intel Tiber Developer Cloud is an innovative approach for organizations to develop and deploy GenAI applications in a cloud-native setting in order to ease complexity, reduce costs, and bring use cases online faster and more efficiently.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com