ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Scaling Generative AI in Enterprise IT Operations: Challenges and Opportunities

Sriramaraju Sagi

NetApp, USA

ABSTRACT

This research paper delves into industries that can benefit from the implementation of AI, such, as IT operations, healthcare, finance and gaming. The potential of AI lies in its ability to enhance customer service create personalized content assist in research endeavors optimize investment strategies and create virtual environments. However businesses often encounter challenges when moving from AI proof of concept to implementation. These challenges involve scalability, integration with existing systems, data governance considerations aligning objectives and meeting requirements. To tackle these concerns head on this paper suggests the utilization of a converged infrastructure platform that combines computing power with network capabilities and storage resources using NVIDIA GPUs. It recommends evaluating existing AI proof of concepts and providing the infrastructure to transform each concept into an use case. The primary goal of this research study is to explore the process involved in converting AI proof of concepts into use cases while understanding its significance in harnessing AI capabilities, within organizations.

***Corresponding author**

Sriramaraju Sagi, NetApp, USA.

Received: January 05, 2024; **Accepted:** January 13, 2024; **Published:** January 20, 2024

Introduction

Generative AI, which harnesses machine learning algorithms to generate content, like text, images and videos has the power to bring about changes across various industries, including enterprise IT operations. One area of promise within IT operations is its application in natural language processing for the creation of chatbots, virtual assistants and automated customer service systems. These technologies can enhance customer service by delivering accurate responses to inquiries while also lightening the workload of representatives. Moreover generative AI can be leveraged in the healthcare sector to aid in research and drug discovery. Through analyzing amounts of data it can identify patterns. Make predictions that may lead to breakthroughs in treatment options. Additionally generative AI has the potential to revolutionize the field by generating innovative designs, music compositions and artworks that push the boundaries of human creativity. In addition to this AI can find utility in the sector by optimizing investment strategies and detecting activities through analysis of extensive financial data sets. Furthermore generative AI holds promise, for transforming the gaming industry by creating worlds and adaptable characters that respond dynamically to players actions-offering a more engaging and interactive gaming experience.

Current State of GenAI Use Cases in Enterprises

The integration of AI, in businesses is progressing rapidly as it becomes deeply embedded in aspects of operations to streamline processes improve efficiency and drive innovation. One area where AI applications are especially prominent is customer service, where chatbots are used to handle inquiries and provide assistance. Additionally AI plays a role in optimizing inventory levels and improving delivery times within supply chain management. In the healthcare sector it helps analyze data for diagnoses and the development of treatment plans. Furthermore AI is applied in finance to identify activities and strengthen risk management efforts. The widespread use of AI across functions within enterprises continues to expand with expectations for growth.

Enterprises that have made progress in developing GenAI Proof of Concept (POC) projects now face the challenge of transitioning these projects into operational and scalable GenAI use cases that are ready for production. This challenge arises from factors. POCs are typically designed for small scale experimentation. May not possess the scalability required for real world operations. Integrating GenAI models, into existing IT infrastructure and workflows presents obstacles that can lead to bottlenecks. Additionally ensuring high quality data that is well governed for GenAI models remains a challenge. Insufficient monitoring techniques and detecting model drift or maintaining processes further undermine the reliability of projects.

A major challenge, in AI projects is the lack of alignment among data scientists IT teams and business stakeholders regarding objectives and expectations. This complexity is further exacerbated by the need to keep up with evolving requirements in the AI landscape. To address this issue it is crucial to establish a structured framework and approach that facilitates a transition from proof of concept to fully operational use cases for GenAI projects. This involves modifying and enhancing GenAI models and infrastructure for scalability, devising integration plans and tools enforcing data governance practices implementing quality checks to ensure data integrity establishing monitoring and maintenance procedures

as well as fostering effective collaboration among data science, IT and business units to align objectives. Additionally adapting GenAI practices to meet changing standards is essential. Solving this problem successfully brings about benefits such as improved efficiency through refined processes and optimized models. It allows organizations to make decisions through operational GenAI use cases while gaining an advantage by deploying GenAI projects into production. Moreover compliance with regulatory standards mitigates legal risks while maximizing return, on investment by leveraging the potential of GenAI investments.

To make progress, in adopting AI its recommended for enterprises to assemble a team consisting of individuals from the data science, IT and business departments. This team will assess the existing GenAI proof of concepts (POCs) develop a plan for integrating each POC into scenarios and diligently implement the proposed strategies while closely monitoring their progress. Converting GenAI POCs into use cases is vital to harness the potential of AI, in businesses enhance decision making abilities and maintain competitiveness in todays data driven world.

This research study focuses on the process of turning GenAI proof of concepts, into applications highlighting their significance in leveraging AI capabilities within organizations. It highlights the importance of assessing existing GenAI proof of concepts and establishing a defined platform and infrastructure for integrating them into real world scenarios. The study also explores the techniques and essential components of a convergent infrastructure, such as computing power, networking and storage which play a role in transforming GenAI proof of concepts into applications. These elements are crucial, for establishing a platform that supports and enhance decision making abilities and staying competitive in a data driven environment.

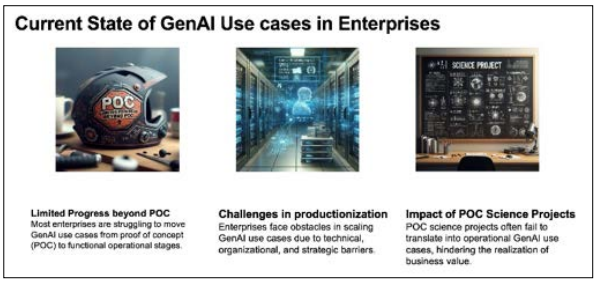

Figure 1: Current State of GenAI Use cases in Enterprises

Literature Review

Generative AI encompasses a range of applications spanning from business to software engineering. However there is growing concern, about its misuse in settings [1]. To address this issue it is crucial to establish AI guidelines that consider the social implications [2]. The papers main findings cover the progress and implementation of AI models the ethical and social implications associated with these models as well as practical strategies for effectively utilizing responsible AI techniques. This study provides a overview of text and image generation models with a focus on the essential requirements for responsible AI linked to these models. It primarily examines applications of AI in various domains such as media generation writing assistance, copywriting, code generation and conversational support. Furthermore it presents approaches and guidelines for implementing AI techniques. Additionally it discusses insights gained from deploying AI methods, in generative AI applications and highlights research challenges that remain unresolved.

The importance of explainability, in AI for coding cannot be understated within the field of software engineering. To ensure the development of AI features human centered techniques serve as a guiding principle [3]. The primary findings of this research paper revolve around exploring the requirements for explainability in AI (GenAI) across three software engineering applications; converting natural language to code code translation and code auto completion. By leveraging centered methodologies this paper showcases how explainable AI can advance technically in unexplored domains. Despite challenges generative AI solutions like Bard, ChatGPT and CoPilot have the potential to significantly improve software production processes [4]. This article thoroughly examines the adoption of AI tools within the software industry and discusses their impact on software productivity as well as their relationship with software development. It also addresses risks involved and provides advice and real life case examples, for practitioners.

Various applications are examined to analyze the functionalities of AI infrastructure. Xu and Du both delve into the use of AI generated content, in networks. Xus research focuses on the structure of cloud edge mobile systems while Dus study explores the interaction between artificial intelligence and networking systems [7,8]. The main findings of their work include the generation of data using AI algorithms deploying AI and graphical computing applications on mobile edge networks as well as analyzing innovative applications driven by artificial intelligence and graphical computing. Furthermore they discuss the challenges related to implementation, security and privacy aspects. Jeong proposes solutions for overcoming knowledge gaps by leveraging large Language Models [9].

The development of a Retrieval Augmented Generation (RAG) model is highlighted as an advancement in improving information storage and retrieval for enhanced content generation. The research demonstrates how this method can be practically applied in organizations that utilize LLMs thereby advancing the field of AI and encouraging utilization of LLM based services, in corporate settings.

Methodology

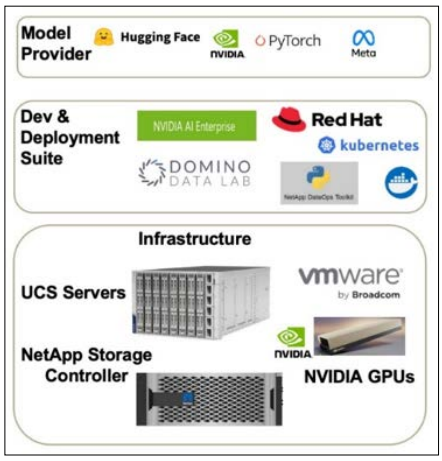

Figure 2: Converged Infrastructure from Cisco and NetApp

In our study report we conducted research using a converged infrastructure platform that combines components, from Cisco and NetApp. Cisco UCS is a computing architecture designed to simplify and speed up the deployment of applications. It covers

areas such as virtualization, cloud computing, scaling workloads analyzing data in memory, edge computing, remote locations, branch offices and IoT data. This platform is designed for expansion. Can support multiple chassis. It brings together all resources under one management domain. On the hand NetApps AFF A Series controllers offer performance and high quality data services for both on premises data centers and the cloud in shared environments. These systems deliver performance, flexibility, top tier data services and seamless integration with the cloud by utilizing NetApp ONTAP data management software. They are pioneers in supporting both NVMe over Fibre Channel (FC NVMe) and NVMe over TCP (NVMe TCP) which enhances performance through network connectivity options. With the industrys latency for an enterprise all flash array these systems are well suited for demanding workloads and applications. Customers can effectively handle workloads, with faster response times by implementing a straightforward software upgrade without any interruption or need to migrate data.

The NetApp DataOps Toolkit is a Python toolkit designed specifically for developers, data scientists and engineers to carry out data management tasks. This functionality allows users to create data volumes, clone data and generate a NetApp Snapshot copy, for the purpose of tracking changes and establishing a baseline. The library can be used as either a command line tool or a set of functions, for Python programs or Jupyter Notebooks.

NVIDIA AI Enterprise is a software platform that boosts the speed of data science and simplifies the creation and implementation of AI models, like generative AI, computer vision and speech AI in real world scenarios. With over 50 frameworks, pre trained models and development tools the platform aims to make AI more accessible to businesses of all kinds. This helps them progress in the field of AI at a pace. The NVIDIA A100 Tensor Core GPU delivers performance in intelligence, data analytics and high performance computing. It is compatible with PCIe Express Gen 4 doubling the bandwidth for high speed network interfaces. The Multi Instance GPU feature divides the GPU into seven instances allowing multiple networks to operate simultaneously. For acceleration in AI applications, data analytics and high performance computing tasks with mathematical precisions consider using the NVIDIA H100 Tensor Core GPU. It also offers enterprise support. If you need a graphics processing unit for operation in enterprise data centers that can handle artificial intelligence applications, like 3D graphics and rendering with improved performance and reliability while maintaining uptime check out the NVIDIA L40S GPU.

Specification	A100	H100	L40S
Architecture	Ampere	Hopper	Ada Lovelace
Release Year	2020	2022	2023
FP32	19.5 TFLOPS	67 TFLOPS	91.6 TFLOPS
TF32 Tensor Core	312 TFLOPS	989 TFLOPS	183 | 366* TFLOPS
GPU Memory	80 GB HBM2e	80 GB	48GB GDDR6 with ECC
GPU Memory Bandwidth	2,039 Gbps	3.35 Tbps	864 Gbps
Form Factor	SXM	SXM	4.4" (H) x 10.5" (L), dual slot
Interconnect	NVLink: 600 GB/s PCIe Gen4: 64 GB/s	NVLink: 900GB/s PCIe Gen5: 128GB/s	PCIe Gen4 x16: 64GB/s bidirectional
CUDA® Cores	6,912	16,896	18,176

Figure 3: NVIDIA GPU Types Comparison

Results

Enterprises can utilize this research and analysis to acquire a pre-trained LLM model and further refine it using their own private data. This can be done by hosting the model on a straightforward, secure, and automated infrastructure configuration. For instance, if an enterprise possesses comprehensive documentation regarding their firm or goods, they can optimize the LLM by incorporating more layers and leveraging their proprietary data. Implementing this will enhance the model's efficacy in processing queries related to your dataset. Organizations have the option to utilize public cloud models to do inferencing on their private data sets or create their own collection of models that represent different business activities.

By developing distinct qualities, businesses can gain a competitive edge and distinguish themselves from rivals. The AI infrastructure presented in this study provides a versatile IT solution that streamlines the administration of AI workloads, enabling the incorporation of GPUs, CPUs, and VMs within VMware vCenter. Additionally, it streamlines the process of allocating storage resources and empowers data scientists and engineering teams to enhance the efficiency of data services. The platform is capable of handling data in any format, allowing for precise adjustments and making inferences on diverse data sets. AI initiatives can be expanded to the cloud using first-party services provided by Amazon, Microsoft, and Google. The AI Control Plane, a comprehensive AI data and experiment management reference design, is improved by simplifying the management of Kubernetes containers.

This AI infrastructure speeds up the implementation and distribution of AI resources making it easier to replicate data across hybrid multicloud systems. Furthermore it improves the efficiency of computing and storage hardware, for AI tasks maximizes parallel processing capabilities and ensures the infrastructure operates continuously with automatic resilience features. The Zero Trust solution provides an approach to safeguarding logical and information security in order to prevent tampering, counterfeiting and unauthorized access. It offers security for SaaS operations by ensuring that users, administrators and processes are verified and authenticated before being granted access.

A user friendly platform for runtime generative AI use cases in enterprise IT operations has advantages. It simplifies integration and scalability by combining storage, computing and networking capabilities to improve efficiency. It optimizes performance and resource utilization by consolidating resources to provide computing power for processing datasets efficiently. It promotes agility by harmonizing IT components to reduce complexities when scaling AI applications. By offering a platform it supports alignment and collaboration while simplifying operations so that enterprises can prioritize their strategic goals such, as implementing AI rather than managing systems. Additionally it enhances resource allocation and management to ensure that GenAI projects have computing power as well as storage and network capabilities. By consolidating IT components, it simplifies things. Speeds up the time, to market process.

Conclusion

The research paper provides an analysis of how Generative AI (GenAI) can be integrated and optimized within enterprise IT operations. It focuses on moving from proof of concept, to application and highlights the challenges involved, such as scalability, data governance and regulatory compliance. The study suggests the need for a converged infrastructure that combines computing, networking, storage resources and NVIDIA GPUs to facilitate this transition effectively. Our findings demonstrate that

a structured approach is crucial for deploying GenAI applications. This includes refining models implementing data governance practices and aligning IT with business units strategically. By addressing these factors organizations can fully leverage the potential of GenAI resulting in improved efficiency enhanced decision making capabilities and a competitive advantage in the market. Looking ahead the research emphasizes the significance of collaboration and continuous monitoring while adapting to evolving AI standards. The convergence of these elements is pivotal not for GenAI deployment but also for sustaining and evolving these capabilities amidst the ever changing landscape of enterprise IT operations. This study establishes a foundation for research and development, in this field while encouraging further exploration of the vast possibilities GenAI offers across various business sectors.

References

1. Charlotte B, Eddie LU, Atoosa K (2020) Business (mis) Use Cases of Generative AI. ArXiv https://arxiv.org/pdf/2307.05543.pdf.
2. Kenthapadi K, Himabindu L, Nazneen FR (2023) Generative AI meets Responsible AI: Practical Challenges and Opportunities. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 5805-5806.
3. Jiao S, Liao QV, Michael M, Mayank A, Stephanie H, et al. (2022) Investigating Explainability of Generative AI for Code through Scenario-based Design. 27th International Conference on Intelligent User Interfaces https://arxiv.org/abs/2202.04903.
4. Ebert C, Panos L (2023) Generative AI for Software Practitioners. IEEE Software 40: 30-38.
5. Minrui X, Hongyang D, Dusit N, Jiawen K, Zehui X, et al. (2023) Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services. ArXiv https://arxiv.org/abs/2303.16129.
6. Hongyang D, Dusit N, Jiawen K, Zehui X, Ping Z, et al. (2023) The Age of Generative AI and AI-Generated Everything. ArXiv https://arxiv.org/abs/2311.00947.
7. Jeong CS (2023) A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. Adv Artif Intell Mach Learn 3: 1588-1618.
8. Kirelli Y (2023) Analysis of Factors Affecting Common Use of Generative Artificial Intelligence-Based Tools by Machine Learning Methods. International Journal of Computational and Experimental Science and Engineering 9: 233-237.
9. Erik B, Danielle L, Lindsey RR (2023) Generative AI at Work. NBER https://www.nber.org/papers/w31161.
10. Wang YC, Jintang X, Chengwei W, Kuo CCJ (2023) An Overview on Generative AI at Scale With Edge–Cloud Computing. IEEE Open Journal of the Communications Society 4: 2952-2971.
11. Borg M, Markus B (2023) Pipeline Infrastructure Required to Meet the Requirements on AI. IEEE Software 40: 18-22.
12. Tommaso F, Mario V, Fabio LP, Andrea L, Matteo B, et al. (2023) AI and data-driven infrastructures for workflow automation and integration in advanced research and industrial applications. Ital-IA https://ceur-ws.org/Vol-3486/121.pdf.