IJARIIT
INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 1 - V11I1-1395)
Available online at: https://www.ijariit.com

Review, Tutorials and Introduction to Cloud Platforms for Agentic GenAI: A Comparative Studies

Satyadhar Joshi
satyadhar.joshi@gmail.com
Bank of America, USA

ABSTRACT

This paper presents a comparative analysis of leading cloud platforms for Generative AI applications. We evaluate performance, scalability, cost, and ecosystem support for AI workloads. The rapid evolution of generative artificial intelligence (AI) has significantly increased the demand for scalable and robust cloud infrastructure. This paper presents a comparative analysis of major cloud platforms, including Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP), focusing on their capabilities to support generative AI applications. We examine key aspects such as infrastructure scalability, cost efficiency, and the availability of specialized AI services. Furthermore, we discuss the importance of well-architected frameworks and best practices for deploying scalable AI solutions. The paper also explores the strategic collaborations and advancements in supercomputing infrastructure that are driving the future of generative AI. Generative AI (GenAI) is rapidly transforming various industries, demanding scalable and cost-effective infrastructure. Cloud platforms like Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and Oracle Cloud Infrastructure (OCI) are vying to provide the necessary tools and services. This literature review examines recent publications and articles discussing the capabilities, architectures, and cost considerations of these platforms in the context of GenAI application development and deployment. We categorize these resources based on their focus: (1) comparative analyses of cloud platforms, (2) GenAI infrastructure and application development, (3) Retrieval-Augmented Generation (RAG) solutions, and (4) scalability and cost optimization strategies. This review aims to provide a comprehensive overview of the current state of GenAI in the cloud, highlighting the strengths and weaknesses of each platform and identifying key trends and challenges.

Keywords: *Generative AI, Cloud Computing, AWS, Azure, GCP, Oracle Cloud, RAG, Scalability, Cost Optimization*

INTRODUCTION

Generative AI (GenAI) is rapidly transforming various industries, demanding scalable and cost-effective infrastructure. The increasing sophistication of GenAI models, coupled with their growing adoption across diverse sectors, necessitates a robust and adaptable cloud infrastructure. Cloud platforms like Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and Oracle Cloud Infrastructure (OCI) are actively developing and promoting their services to capitalize on this demand. This competition fosters innovation and provides businesses with a range of options tailored to their specific needs. This literature review examines recent publications and articles discussing the capabilities, architectures, and cost considerations of these platforms in the context of GenAI application development and deployment. We categorize these resources based on their focus: (1) comparative analyses of cloud platforms, (2) GenAI infrastructure and application development, (3) Retrieval-Augmented Generation (RAG) solutions, and (4) scalability and cost optimization strategies. By synthesizing the findings from these sources, this review aims to provide a valuable resource for researchers, developers, and decision-makers seeking to navigate the complex landscape of GenAI in the cloud. Furthermore, it identifies areas for future research and exploration within this rapidly evolving domain. Generative AI (GenAI) has transformed multiple industries, from finance to healthcare. Deploying these models efficiently requires robust cloud infrastructure. This paper evaluates leading cloud providers, analyzing their suitability for GenAI applications.

The landscape of artificial intelligence is being revolutionized by generative AI, which enables the creation of new content and data. This advancement necessitates robust and scalable cloud infrastructure to support the computational demands of large language models (LLMs) and other generative AI applications. Major cloud providers are competing to offer the most advanced solutions, driving innovation and collaboration in this rapidly evolving field [1], [2], [3].

This paper aims to provide a comprehensive comparison of leading cloud platforms, focusing on their suitability for deploying scalable generative AI applications. We will explore the architectural guidelines, infrastructure capabilities, and specialized services offered by AWS, Azure, and GCP, and discuss the importance of cost optimization and efficient resource utilization [4]. We compare AWS, Azure, and GCP across key parameters: compute performance, cost, ease of integration, and AI ecosystem support. Each cloud platform demonstrates unique strengths. AWS offers flexibility but is cost-intensive. Azure has seamless enterprise integration. GCP's TPU-based acceleration provides superior deep-learning performance. This work is build on our previous work [37-51].

2. COMPARATIVE ANALYSES OF CLOUD PLATFORMS

This section provides a comparative analysis of the major cloud providers – Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure, and Oracle Cloud Infrastructure (OCI) – with a focus on their offerings for Generative AI. The comparison will consider infrastructure, AI services, and specific tools for building and deploying GenAI applications.

Several sources offer comparative analyses of the leading cloud providers, providing valuable insights into their respective strengths and weaknesses. [5], [6], [7], [8] provide broad comparisons of AWS, Azure, and GCP, evaluating their services, pricing models, and overall performance. These comparisons often consider factors such as compute power, storage capabilities, AI/ML services, and geographic availability. [8] stands out by including Oracle Cloud Infrastructure in its comparison, highlighting Oracle's specific advantages in areas like database management and enterprise solutions. [7] takes a practical approach, focusing on guiding businesses in selecting the most suitable cloud platform based on their unique requirements, considering factors like budget, security needs, and compliance requirements. [5] offers a more granular comparison, examining networking infrastructure, specific cloud solutions, and storage options across the major providers. It identifies use cases where each platform excels, enabling businesses to make informed decisions. [6] takes a broader perspective by comparing AWS, Azure and GCP, whilst considering other cloud providers such as DigitalOcean and IBM cloud. This expanded comparison emphasizes that choosing the right platform depends on the scale of the business, industry requirements and pricing. The paper also describes edge computing, serverless computing and containerization offerings from each of the cloud platform offerings.

A comparative analysis of these cloud platforms reveals key differences in their service offerings, pricing models, and infrastructure capabilities. The choice of platform depends on specific project requirements, cost considerations, and the level of integration with existing systems [5], [6], [7], [8], [9], [10].

2.1 Infrastructure

All major cloud providers offer a range of compute instances suitable for GenAI workloads, including CPU-based and GPU-based options. AWS offers EC2 instances with NVIDIA GPUs, optimized for machine learning tasks, and has announced further collaboration with NVIDIA to offer supercomputing infrastructure [11], [12]. GCP provides Compute Engine instances with NVIDIA GPUs and TPUs (Tensor Processing Units), which are custom-designed ASICs for machine learning [13]. Azure offers virtual machines with NVIDIA GPUs and also supports specialized hardware accelerators. Oracle Cloud Infrastructure also provides GPU-based instances, competing with other major cloud providers [8], [10].

2.2 AI Services and Platforms

Each cloud provider offers a suite of AI services and platforms designed to simplify the development and deployment of GenAI applications. AWS provides SageMaker, a comprehensive machine learning platform that supports model training, deployment, and monitoring [14], [15]. It also offers pre-trained AI services for tasks such as natural language processing and computer vision. Google Cloud offers Vertex AI, a unified platform for building, deploying, and managing machine learning models [14], [16], [17]. Azure provides Azure Machine Learning, a cloud-based service for building, training, and deploying machine learning models. Red Hat OpenShift AI [18] offers AI tools on their platform for quickly building and deploying AI models. The platforms all aim to assist in building and deploying AI tools in a streamlined manner.

2.3 Retrieval-Augmented Generation (RAG) Support

Retrieval-Augmented Generation (RAG) is a critical technique for improving the accuracy and relevance of GenAI models. AWS, Azure, and GCP all offer services and tools to support RAG implementations [19]. Google Cloud's Vertex AI integrates with AlloyDB for PostgreSQL to provide a vector store for efficient knowledge retrieval [20]. AWS offers integrations with services like Kendra and OpenSearch for knowledge retrieval and vector search [21], [22]. The overall aim is to provide tools to improve AI models and augment them with external knowledge.

2.4 Cost and Scalability

Cost and scalability are essential considerations when choosing a cloud platform for GenAI. Gupta [4] provides a practical guide to managing GenAI infrastructure costs across major cloud platforms, offering strategies for optimizing resource allocation and selecting cost-effective instance types. AWS, Azure, and GCP all offer various pricing models and autoscaling capabilities to help manage costs and ensure scalability [13], [23], [24]. Cloud providers often offer tools for cost analysis, optimization, and automation, helping users manage and reduce GenAI costs.

2.5 Ecosystem and Community

The ecosystem and community surrounding a cloud platform can significantly impact the ease of development and adoption of GenAI technologies. AWS, Azure, and GCP have large and active communities, providing extensive documentation, tutorials, and support resources [25]. These communities contribute to the development of open-source tools and libraries, further accelerating GenAI innovation. Smaller cloud provider communities exist but do not have the same level of support as the main three offerings.

Cloud Infrastructure for Generative AI

The core of generative AI applications lies in the underlying cloud infrastructure. The ability to scale resources dynamically is crucial for handling the fluctuating demands of AI workloads. Cloud elasticity and scalability are key factors in ensuring optimal performance and cost efficiency [26].

2.6 Amazon Web Services (AWS)

AWS offers a wide range of services tailored for generative AI, including Amazon Bedrock, SageMaker, and specialized infrastructure for LLMs [15], [21], [22]. The strategic collaboration between AWS and NVIDIA to provide supercomputing infrastructure highlights the commitment to advancing generative AI capabilities [11], [12], [27]. AWS also provides prescriptive guidance for designing robust cloud architectures [28].

2.7 Microsoft Azure

Azure provides a comprehensive suite of AI services, including Azure AI and specialized infrastructure for deep learning. The platform's focus on enterprise-grade solutions and integration with other Microsoft services makes it a strong contender in the cloud AI space [5], [6], [29].

2.8 Google Cloud Platform (GCP)

GCP offers Vertex AI, a unified platform for building and deploying machine learning models, and specialized infrastructure for generative AI workloads [16], [17], [20]. Google's emphasis on scalability and resilience is evident in its architectural guidelines for cloud solutions [13].

2.9 Summary of Comparative Analysis

Each cloud provider offers a comprehensive set of services and infrastructure for building and deploying GenAI applications. The best choice depends on the specific requirements of the application, the expertise of the development team, and the organization's existing cloud investments. AWS, Azure, and GCP are the leading platforms, while Oracle Cloud Infrastructure provides competitive offerings, especially for organizations with Oracle-centric environments [8]. The cloud provider ecosystems all seek to simplify the development and building of AI services.

3 SCALABILITY AND BEST PRACTICES

Building scalable AI solutions requires adherence to best practices and the adoption of well-architected frameworks. This includes designing for elasticity, optimizing resource utilization, and implementing robust monitoring and management systems [23], [24].

3.1 Well-Architected Frameworks

Well-architected frameworks, such as those provided by AWS, Azure, and GCP, offer guidelines for designing secure, reliable, and efficient cloud solutions [30]. These frameworks emphasize the importance of operational excellence, security, reliability, performance efficiency, and cost optimization.

3.2 Scalable AI Application Infrastructure

The infrastructure stack for generative AI applications includes models, frameworks, and deployment strategies. The ability to scale these components is crucial for handling the demands of large-scale AI workloads [31], [32], [33].

3.3 Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) is an advanced technique that enhances the capabilities of generative AI models by integrating them with information retrieval systems. Cloud providers offer specialized services and infrastructure for deploying RAG solutions [19], [20].

3.4 The Future of Generative AI in the Cloud

The future of generative AI in the cloud is marked by continuous innovation and strategic collaborations. The increasing demand for AI solutions is driving the development of specialized hardware and software, as well as the expansion of cloud infrastructure [18], [34].

3.5 Platform Power and Industrialization of AI

The major cloud providers are playing a pivotal role in the industrialization of AI, shaping the future of AI development and implementation. Their ecosystems for cloud AI are crucial in operationalizing infrastructural power [25], [35].

4 QUANTITATIVE ANALYSIS OF CLOUD PLATFORM COSTS FOR GENERATIVE AI

To provide a quantitative perspective on cloud platform selection for generative AI, we focus on cost analysis, a critical factor for organizations deploying such applications. Gupta (2025) [4] emphasizes the importance of strategic cost management across major cloud platforms, particularly for optimizing generative AI workloads.

4.1 Cost Considerations Across Platforms

The cost of generative AI infrastructure varies significantly across AWS, Azure, and GCP. Key factors influencing these costs include compute resources (GPUs, CPUs), storage, data transfer, and specialized AI services.

Compute Resources: The demand for high-performance computing, particularly GPUs, is substantial in generative AI. Cloud providers offer various GPU-enabled instances with different pricing models. A quantitative comparison would involve analyzing the cost per GPU hour for comparable instance types across platforms.

Storage and Data Transfer: Large datasets required for training and inference contribute to significant storage and data transfer costs. A quantitative assessment would involve comparing the cost per GB of storage and data transfer rates across platforms.

Specialized AI Services: Services like Amazon SageMaker, Azure AI, and Google Vertex AI offer managed environments for AI development and deployment. A quantitative analysis would compare the pricing structures of these services, considering factors such as model training, inference, and deployment costs.

4.2 Illustrative Cost Modeling

To illustrate the quantitative aspects, consider a hypothetical scenario: training a large language model (LLM) requiring 1000 GPU hours.

AWS: Assuming an AWS instance with a cost of $3 per GPU hour, the total compute cost would be $3000.

Azure: If Azure offers a comparable instance at $2.8 per GPU hour, the total compute cost would be $2800.

GCP: With GCP's offering at $2.9 per GPU hour, the total compute cost would be $2900.

This simplified model demonstrates the potential cost variations. However, a comprehensive quantitative analysis would require detailed workload characterization, including data storage, transfer, and specialized service utilization. Gupta (2025) [4] provides a practical guide to optimizing generative AI workloads by cost management across major cloud platforms, which should be consulted for a more in-depth quantitative cost analysis.

Add the content in the font size 10 and justified font. All paragraphs further should have same styling. Some content related to your research work in running paragraphs. Some content related to your research work in running paragraphs. Some content related to your research work in running paragraphs.

5 TUTORIALS AND PRACTICAL GUIDES FOR GENERATIVE AI DEPLOYMENT

Several resources within the cited bibliography offer practical tutorials, step-by-step guides, and instructional content aimed at facilitating the deployment and utilization of generative AI on various cloud platforms. These resources provide valuable hands-on experience and knowledge transfer, bridging the gap between theoretical concepts and practical implementation. Access to comprehensive tutorials and training resources is crucial for fostering wider adoption and effective utilization of Generative AI (GenAI) technologies within cloud environments. While many of the listed resources focus on infrastructure and platform comparisons, some provide insights into available learning opportunities.

5.1 AWS-Focused Tutorials

AWS provides a wealth of tutorials and guides for deploying generative AI applications. The "Create a Generative AI–Powered Custom Google Chat Application Using Amazon Bedrock" tutorial [21] offers a detailed, step-by-step approach to building a custom application, specifically using Amazon Bedrock. This blog post is categorized as "Advanced (300)," indicating its suitability for users with some prior AWS experience. Similarly, the "Generative AI Application Builder on AWS" [22] provides a solution implementation guide, enabling users to deploy production-scale generative AI applications. Additionally, the "What Is the AWS CDK?" documentation [36] serves as a manual to understand and use the AWS Cloud Development Kit, a tool to define cloud infrastructure as code. The guide "How to Build a Scalable Application up to 1 Million Users on AWS" [23] provides a step by step guide for building a scalable application on AWS.

5.2 Google Cloud Platform (GCP) Guides

GCP offers practical guidance through its architectural center and community resources. The "Infrastructure for a RAG-capable Generative AI Application Using Vertex AI and AlloyDB for PostgreSQL" guide [20] provides detailed steps for designing infrastructure to run a generative AI application with retrieval-augmented generation. Saxena (2024) [17] provides guidance on architecting GenAI applications with Google Cloud, providing practical insights and methodologies. Hutcherson (2023) [16] provides a guide on building LLM applications with Redis on Google's Vertex AI.

5.3 Conceptual and Comparative Guides

Beyond platform-specific tutorials, several resources offer conceptual and comparative guides. The "Generative AI Tech Stack: Frameworks, Infrastructure, Models and Applications" [32] provides an overview of the generative AI tech stack, including models, frameworks, tools, and deployment strategies. These guides offer a broader understanding of the generative AI landscape and facilitate informed decision-making when selecting cloud platforms and tools.

5.4 Platform-Specific Training

Each major cloud provider offers extensive documentation, tutorials, and training programs to help developers learn how to use their AI services and infrastructure. AWS provides a wealth of resources through its AWS Training and Certification program, covering topics such as machine learning, deep learning, and GenAI [12], [15]. These resources often include step-by-step tutorials, code samples, and hands-on labs. Google Cloud offers similar training programs through Google Cloud Skills Boost and its AI Platform Training courses [17]. Microsoft Azure provides learning paths and certifications for AI and machine learning through Microsoft Learn.

5.5 Community-Driven Tutorials and Examples

In addition to official training programs, a vibrant community of developers and researchers contributes to a wealth of tutorials, blog posts, and open-source projects that can help users learn how to build GenAI applications in the cloud. Resources like the AWS Machine Learning Blog [21] and community forums provide practical examples and guidance on using cloud services for specific GenAI tasks. Platforms like Medium and Towards Data Science host numerous tutorials and articles on GenAI topics, often focusing on specific cloud platforms and tools.

5.6 Specialized Training Platforms

Platforms like NVIDIA DGX Cloud [34] and Red Hat OpenShift AI [18] offer specialized training and development environments for AI and machine learning. These platforms often provide access to pre-configured hardware and software, as well as expert support, making it easier for developers to get started with GenAI.

5.6 Considerations for Choosing Training Resources

When selecting training resources for GenAI in the cloud, it is important to consider the following factors:

* **Level of Expertise:** Choose resources that are appropriate for your current skill level, whether you are a beginner or an experienced AI practitioner. * **Specific Use Case:** Look for tutorials and examples that are relevant to your specific GenAI application. * **Cloud Platform:** Select resources that focus on the cloud platform you are using, whether it is AWS, Azure, GCP, or another provider. * **Cost:** Consider the cost of training programs and certifications, and whether they are worth the investment.

5.7 Gap in the Literature

While the reviewed resources provide a general overview of available training opportunities, there is a gap in the literature regarding comparative evaluations of different training programs and their effectiveness. Future research could focus on assessing the impact of various training resources on developer skills and GenAI application outcomes. Furthermore, there is a need for more resources that address the ethical considerations and societal impacts of GenAI, ensuring that developers are equipped to build responsible and beneficial AI systems.

6 FUTURE TRENDS IN GENERATIVE AI CLOUD INFRASTRUCTURE (2025-2030)

The trajectory of generative AI and its cloud infrastructure is poised for significant advancements in the coming years. Based on the cited literature, we can project several key trends through 2030.

The field of cloud-based Generative AI (GenAI) is rapidly evolving, with significant potential for future research and development. Based on current trends and projections, this section outlines potential directions for future work in this area, spanning the period from 2025 to 2030.

6.1 2025: Enhanced Specialization and Collaboration Focus on Cost Optimization and Scalability

In 2025, we anticipate further specialization in cloud offerings for generative AI. Strategic collaborations, such as the AWS and NVIDIA partnership [11], will become more prevalent, driving the development of optimized hardware and software stacks. Cloud providers will continue to refine their managed AI services, like Amazon Bedrock [15], Azure AI, and Google Vertex AI, focusing on ease of use and cost efficiency. Gupta (2025) [4] highlights that 2025 will see an increased emphasis on optimizing costs, as organizations seek to scale generative AI workloads efficiently.

In 2025, a key focus will likely be on optimizing the cost and scalability of GenAI applications. Gupta [4] emphasizes the importance of managing GenAI infrastructure costs across major cloud platforms, suggesting that organizations will increasingly seek strategies for optimizing resource allocation and selecting cost-effective instance types. Furthermore, research will likely focus on developing more efficient algorithms and models that can reduce the computational requirements of GenAI applications. Scalability will remain a critical concern, with researchers exploring architectural patterns and techniques for building GenAI systems that can handle increasing workloads and data volumes [13], [23], [24].

6.2 2026-2027: Emphasis on RAG and Knowledge Integration, Edge AI and Hybrid Deployments

By 2026-2027, Retrieval-Augmented Generation (RAG) will likely become an even more prominent technique for enhancing the accuracy and relevance of GenAI models. Richards [19] compares AWS, Azure, and GCP for deploying RAG solutions, highlighting the importance of efficient knowledge retrieval and integration with LLMs. Future work in this area will likely focus on developing more sophisticated RAG techniques that can handle complex knowledge sources and improve the quality of generated content. Researchers may explore novel approaches for building and managing vector databases, as well as developing more effective methods for extracting and integrating knowledge from unstructured data [20].

By 2026, edge AI will gain traction, enabling real-time generative AI applications on devices and at the network edge. This trend will be facilitated by advancements in hardware acceleration and model compression. Hybrid cloud deployments will become more common, allowing organizations to leverage on-premises infrastructure for sensitive data and cloud resources for scalable compute. The "Well Architecture Framework" [30] will be even more critical in designing these hybrid deployments.

6.3 2027: AI Model Marketplaces and Ecosystem Expansion

2027 will witness the rise of AI model marketplaces, where pre-trained models and AI components can be easily accessed and integrated into applications. This will foster a vibrant ecosystem of AI developers and businesses. Cloud providers will expand their offerings to include comprehensive AI development platforms, supporting the entire lifecycle of AI applications. Luitse (2024) [35] and van der Vlist (2024) [25] highlight the increasing "platform power" of major cloud providers, and by 2027 this will be a fully realized ecosystem.

6.4 2028: Autonomous AI and Advanced RAG

In 2028, autonomous AI systems will become more prevalent, capable of self-learning and adapting to changing environments. Retrieval Augmented Generation (RAG) will evolve significantly, with advanced techniques for integrating external knowledge sources and improving the accuracy and relevance of generative AI outputs. Richards (2024) [19] discusses the importance of RAG, and this will be an area of active innovation.

2028-2030: Focus on Ethical Considerations and Societal Impact

Looking further ahead to 2028-2030, ethical considerations and societal impact will likely become increasingly important topics in GenAI research and development. As GenAI technologies become more powerful and pervasive, it is crucial to address potential risks and biases, ensuring that these technologies are used responsibly and ethically. Luitse [35] highlights the platform power in AI, suggesting that future research should explore the political economy of AI and the potential for cloud platforms to shape the development and deployment of these technologies. Future work may also focus on developing methods for detecting and mitigating biases in GenAI models, as well as exploring the social and economic implications of these technologies [25]. There is a call for future research and development of standardized benchmarks for GenAI performance and exploration of novel techniques for optimizing resources and reducing environmental impacts.

6.5 2030: Quantum-Enhanced AI and Ethical AI Frameworks

By 2030, quantum computing may begin to impact generative AI, enabling the development of more powerful and efficient AI models. Ethical AI frameworks will become essential, addressing concerns related to bias, privacy, and security. Cloud providers will integrate these frameworks into their AI platforms, promoting responsible AI development and deployment. The "Simplified

Architecture to Take up Generative AI in the Cloud Applications" [33] will need to evolve to consider the impact of quantum computing, and ethical frameworks.

In summary, the years leading to 2030 will witness a continuous evolution of generative AI cloud infrastructure, driven by technological advancements, strategic collaborations, and the growing demand for AI-powered solutions.

6.6 Overall Trends for the Next Five Years

Several overall trends are expected to shape the future of cloud-based GenAI:

* **Increased Specialization:** Cloud providers will likely offer more specialized AI services and infrastructure tailored to specific GenAI use cases [11]. * **Greater Automation:** Automation will play an increasingly important role in managing GenAI infrastructure, reducing costs, and improving efficiency. * **Expanded Ecosystems:** The ecosystems surrounding cloud-based GenAI will continue to grow, with more open-source tools, pre-trained models, and community resources becoming available [31], [32].

By addressing these challenges and pursuing these opportunities, researchers and developers can help to ensure that cloud-based GenAI technologies are used to create a more innovative, equitable, and sustainable future.

7 ARCHITECTURAL CONSIDERATIONS FOR GENERATIVE AI APPLICATIONS IN THE CLOUD

The design and implementation of cloud architectures are pivotal for deploying scalable and resilient generative AI applications. Several references provide insights into architectural considerations, highlighting the importance of well-defined patterns and frameworks. Developing and deploying Generative AI (GenAI) applications in the cloud requires careful consideration of architectural patterns and design choices to ensure scalability, resilience, and cost-effectiveness. This section explores the architectural considerations discussed in the provided literature, focusing on the recommendations and best practices for building GenAI solutions on different cloud platforms.

7.1 GenAI Infrastructure and Application Development

A significant number of resources delve into the infrastructure and tools offered by each cloud platform for developing and deploying GenAI applications, showcasing the platforms' commitment to fostering GenAI innovation. [12], [15] highlight AWS's comprehensive offerings, emphasizing its purpose-built AI services, such as SageMaker, and its infrastructure optimized for computationally intensive GenAI workloads, including GPU-powered instances and high-performance networking. These resources showcase AWS's commitment to providing developers with the tools and resources needed to build and deploy GenAI applications at scale. [1], [2] offer a comparative analysis of the GenAI capabilities of AWS, Azure, and GCP, examining their respective AI platforms, model marketplaces, and developer tools. These comparisons help developers understand the nuances of each platform and choose the one that best aligns with their specific needs and expertise. [17] provides a focused guide on architecting GenAI applications specifically on Google Cloud, offering practical advice on leveraging Google's AI services, such as Vertex AI, and its infrastructure to build robust and scalable GenAI solutions. This paper's detailed architectural guidance is invaluable for developers seeking to maximize their utilization of Google Cloud for GenAI applications. [22] describes a solution for deploying production-scale GenAI applications on AWS, providing a blueprint for building and deploying GenAI applications that can handle real-world workloads. This detailed solution guide can dramatically accelerate the development and deployment process for AWS users. [33] describes a simplified architecture for incorporating GenAI into cloud applications, making it accessible for applications to leverage GenAI capabilities with minimal modification. [31], [32] provide broader overviews of the GenAI technology stack, encompassing frameworks, models, and deployment strategies, although not exclusively tied to a single cloud provider. These overviews provide a valuable context for understanding the broader GenAI landscape and the role of cloud platforms within it. [34] showcases NVIDIA's DGX Cloud, an AI platform specifically targeted towards enterprise developers, offering pre-configured hardware and software for accelerated AI development. [18] presents Red Hat OpenShift AI, a platform designed for developing, training, and serving AI models, promoting open-source principles and providing a flexible environment for AI innovation.

7.2 Retrieval-Augmented Generation (RAG) Solutions

Retrieval-Augmented Generation (RAG) is a key technique for enhancing the accuracy, contextuality, and relevance of GenAI models, enabling them to access and incorporate external knowledge sources. [19] compares AWS, Azure, and GCP for deploying RAG solutions, evaluating their offerings in terms of vector databases, knowledge retrieval tools, and integration capabilities with LLMs. This comparison is essential for developers seeking to implement RAG-based GenAI applications in the cloud. [20] details the infrastructure for a RAG-capable GenAI application using Google Cloud's Vertex AI and AlloyDB, highlighting the use of AlloyDB as a vector store for efficient knowledge retrieval. This detailed infrastructure blueprint provides developers with a practical guide for building RAG solutions on Google Cloud. [21] shows how to create a GenAI powered application using Amazon Bedrock, demonstrating practical ways to leverage AWS services to build intelligent applications.

7.3 Scalable and Resilient Architectures

Several resources emphasize the importance of building scalable and resilient architectures for GenAI applications. The InfoQ article on the architecture of a scalable and resilient Google Cloud solution [13] provides architectural guidelines that can be adapted for deploying web applications on various cloud platforms. Solanki [23] offers a guide to building scalable applications on AWS for up to a million users, providing practical advice on designing systems that can handle high traffic and data volumes. Best practices for scalable AI on cloud infrastructure are discussed in [24], emphasizing the importance of optimizing resource utilization and minimizing costs.

7.4 Reference Architectures for GenAI Applications

Some resources provide reference architectures for building specific types of GenAI applications in the cloud.

The AWS Solutions Library offers a Generative AI Application Builder solution [22] that helps developers deploy production-scale GenAI applications on AWS. Google Cloud's Architecture Center provides a blueprint for building RAG-capable GenAI applications using Vertex AI and AlloyDB for PostgreSQL [20]. These reference architectures can serve as a starting point for developers seeking to build similar applications on their respective cloud platforms. The simplified Architecture Take [33] provides a simple option for adopting Generative AI into cloud applications.

7.5 GenAI Tech Stack

Takyar [32] provides a broader overview of the GenAI technology stack, encompassing frameworks, infrastructure, models, and applications. This overview highlights the different components that are typically involved in a GenAI application and how they interact with each other. The Sapphire Ventures article [31] dives deep into the Generative AI App Infrastructure Stack, providing a detailed market map and analysis of the emerging GenAI landscape.

7.6 Cloud Platform Specific Architectures

Saxenashikha [17] offers a focused guide on architecting GenAI applications specifically on Google Cloud, providing practical advice on leveraging Google's AI services and infrastructure to build robust and scalable GenAI solutions. Luitse [35] explores how AWS, Microsoft Azure, and Google Cloud strategically attempt to operationalize infrastructural power in AI development and implementation through their ecosystems for cloud AI.

7.7 Considerations for Choosing an Architecture

When choosing an architecture for a GenAI application in the cloud, it is important to consider the following factors:
* **Scalability Requirements:** How many users and data volumes will the application need to support? * **Resilience Requirements:** How important is it that the application remains available in the event of a failure? * **Cost Constraints:** What is the budget for building and running the application? * **Cloud Platform:** Which cloud platform will the application be deployed on?
By carefully considering these factors, developers can choose an architecture that meets the specific needs of their GenAI application and ensures its success in the cloud.

7.8 Scalability and Resilience in Cloud Architectures

The architecture of a scalable and resilient cloud solution is paramount, especially for generative AI workloads that demand high availability and performance. Google's architectural guidelines, as discussed in [13], emphasize building robust systems capable of handling fluctuating demands. These guidelines, while specific to Google Cloud, offer valuable insights applicable across different cloud platforms. Solanki (2018) [23] provides a step by step guide for building a scalable application on AWS.

7.9 Well-Architected Frameworks and Prescriptive Guidance

Cloud providers offer well-architected frameworks that serve as blueprints for designing secure, reliable, and efficient cloud architectures. These frameworks, discussed in [30], cover operational excellence, security, reliability, performance efficiency, and cost optimization. AWS, for example, provides prescriptive guidance on cloud design patterns and architectures [28]. These guides help developers implement best practices and avoid common pitfalls.

Architectural Patterns for Generative AI

The deployment of generative AI applications introduces unique architectural challenges. For instance, the "Infrastructure for a RAG-capable Generative AI Application Using Vertex AI and AlloyDB for PostgreSQL" [20] showcases an architecture tailored for retrieval-augmented generation (RAG), emphasizing the integration of vector databases with AI services. Saxena (2024) [17] discusses general architectural considerations for GenAI applications on GCP.

7.10 Comparative Architectural Considerations

Comparing architectures across different cloud providers reveals variations in service offerings and implementation strategies. For instance, AWS's emphasis on modular services and serverless computing contrasts with GCP's focus on containerization and Kubernetes. Azure, on the other hand, integrates its AI services with its broader ecosystem of enterprise solutions.

AWS: Architectures often leverage services like Lambda, S3, and SageMaker, promoting a decoupled and scalable design. The AWS CDK [36] allows for infrastructure as code, which adds a layer of repeatability.

Azure: Architectures typically integrate Azure AI with other Azure services, such as Azure Kubernetes Service (AKS) and Azure Storage.

GCP: Architectures frequently utilize Google Kubernetes Engine (GKE), Vertex AI, and Cloud Storage, emphasizing containerized deployments and managed AI services.

7.11 Simplified Architectural Approaches

Simplified architectural approaches are also discussed, aiming to make generative AI more accessible. The "Simplified Architecture to Take up Generative AI in the Cloud Applications" [33] highlights the need for streamlined designs that reduce complexity and accelerate deployment.

7.12 Architectural Considerations for RAG

The architecture for RAG systems is of particular importance. As discussed in [19], the integration of vector databases, LLMs and data retrieval services are core to RAG architecture. The database and the LLM must be optimized for low latency interactions.

In conclusion, architectural considerations are crucial for building robust and scalable generative AI applications. Cloud providers offer a variety of services and frameworks to support these architectures, and developers must carefully evaluate their options to select the most suitable approach for their specific needs.

Cloud Platform Tools and Sub-tools for Data Engineering Steps
Data engineering is a crucial aspect of building and deploying Generative AI (GenAI) applications in the cloud. It involves various steps, including data ingestion, storage, processing, transformation, and serving. This section outlines the tools and sub-tools offered by each major cloud platform for different data engineering tasks, based on the provided resources.

7.13 Amazon Web Services (AWS)

AWS provides a comprehensive suite of services for data engineering, catering to various needs and skill levels.
* **Data Ingestion:** * **AWS Data Pipeline:** A service for moving data between different AWS compute and storage services, as well as on-premises data sources. * **AWS Kinesis:** A platform for streaming data, enabling real-time data ingestion from various sources. * **AWS Glue:** A fully managed ETL (extract, transform, load) service that simplifies data preparation and transformation. * **Data Storage:** * **Amazon S3 (Simple Storage Service):** A highly scalable and durable object storage service for storing large volumes of data. * **Amazon RDS (Relational Database Service):** A managed relational database service supporting various database engines, such as MySQL, PostgreSQL, and Oracle. * **Amazon DynamoDB:** A NoSQL database service for high-performance applications. * **Data Processing and Transformation:** * **AWS Glue:** For ETL tasks, data cleaning, and data transformation. * **Amazon EMR (Elastic MapReduce):** A managed Hadoop service for big data processing and analysis. * **AWS Lambda:** A serverless compute service that can be used for data transformation and processing tasks. * **Data Serving and Analytics:** * **Amazon Redshift:** A fast, fully managed data warehouse service for large-scale data analysis. * **Amazon Athena:** An interactive query service that enables querying data stored in S3 using SQL. * **Amazon SageMaker:** While primarily an AI/ML platform, SageMaker can be used for data exploration, feature engineering, and model deployment [14]. * **AI/ML:** * **Amazon Sagemaker:** Sagemaker has different tools for building and deploying and model building [14]. * **Amazon Bedrock**: Bedrock can be used for GenAI-powered custom applications [21]. * **AWS AI Services:** A collection of pre-trained AI services for tasks such as natural language processing, computer vision, and speech recognition [15].

7.14 Google Cloud Platform (GCP)

GCP offers a range of tools and services for data engineering, designed for scalability and performance.
* **Data Ingestion:** * **Cloud Dataflow:** A fully managed stream and batch data processing service. * **Cloud Pub/Sub:** A messaging service for real-time data ingestion and distribution. * **Data Storage:** * **Cloud Storage:** A scalable object storage service for storing unstructured data. * **Cloud SQL:** A managed relational database service supporting various database engines, such as MySQL, PostgreSQL, and SQL Server. * **Cloud Spanner:** A globally distributed, scalable, and strongly consistent database service. * **AlloyDB for PostgreSQL:** AlloyDB can be used with Vertex AI for RAG [20]. * **Data Processing and Transformation:** * **Cloud Dataflow:** For ETL tasks, data transformation, and data enrichment. * **Cloud Dataproc:** A managed Hadoop and Spark service for big data processing and analysis. * **Cloud Functions:** A serverless compute service that can be used for data transformation and processing tasks. * **Data Serving and Analytics:** * **BigQuery:** A fast, fully managed data warehouse service for large-scale data analysis. * **Looker:** A business intelligence and data visualization platform that integrates with BigQuery. * **AI/ML** * **Vertex AI:** A unified platform for building, deploying, and managing machine learning models [16], [17].

7.15 Microsoft Azure

Azure provides a comprehensive set of tools and services for data engineering, with a focus on integration with other Microsoft products and services.
* **Data Ingestion:** * **Azure Data Factory:** A cloud-based ETL service for data integration and transformation. * **Azure Event Hubs:** A scalable event ingestion service for real-time data streaming. * **Data Storage:** * **Azure Blob Storage:** A scalable object storage service for storing unstructured data. * **Azure SQL Database:** A managed relational database service based on SQL Server. * **Azure Cosmos DB:** A NoSQL database service for high-performance applications. * **Data Processing and Transformation:** * **Azure Data Factory:** For ETL tasks, data cleaning, and data transformation. * **Azure Databricks:** An Apache Spark-based analytics service for big data processing and machine learning. * **Azure Functions:** A serverless compute service that can be used for data transformation and processing tasks. * **Data Serving and Analytics:** * **Azure Synapse Analytics:** A fully managed data warehouse service for large-scale data analysis. * **Power BI:** A business intelligence and data visualization platform that integrates with Azure Synapse Analytics.

7.16 Oracle Cloud Infrastructure (OCI)

While the provided resources offer less detail on OCI's specific data engineering tools, Oracle Cloud provides the following tools:
* **Data Integration**: Oracle Cloud Infrastructure Data Integration is a fully managed, serverless, native cloud data integration service that extracts, loads, transforms, cleanses, and reshapes data from a variety of data sources into OCI data stores * **Autonomous Data Warehouse:** A self-driving, self-securing, and self-repairing data warehouse service in the cloud

7.17 Cross-Platform Tools

In addition to platform-specific tools, several cross-platform data engineering tools can be used in conjunction with cloud services:
* **Apache Spark:** A unified analytics engine for big data processing. * **Apache Kafka:** A distributed streaming platform for building real-time data pipelines. * **Airflow:** Airflow can be used to programmatically author, schedule and monitor workflows. This section provides a comprehensive overview of the tools and sub-tools offered by each major cloud platform for different data engineering tasks. By leveraging these tools effectively, developers can build robust and scalable data pipelines for GenAI applications in the cloud.

8 DOMAIN-SPECIFIC APPLICATIONS OF GENERATIVE AI AND CLOUD PLATFORMS

While many of the reviewed resources provide a general overview of Generative AI (GenAI) and cloud platforms, some touch upon domain-specific applications or offer insights that can be extrapolated to various industries. This section explores the potential of GenAI in specific domains, highlighting relevant resources and identifying areas for future research.

8.1 Healthcare

The reviewed resources do not explicitly discuss GenAI applications in healthcare. However, we can draw parallels from other domains and suggest potential applications such as:
* **Drug Discovery:** GenAI models could be used to accelerate drug discovery by generating novel molecules and predicting their efficacy. * **Personalized Medicine:** GenAI could be used to analyze patient data and develop personalized treatment plans. * **Medical Image Analysis:** GenAI could be used to automate the analysis of medical images, such as X-rays and MRIs, improving the accuracy and efficiency of diagnoses. * **Medical Summarization**: GenAI can be used to summarize lengthy medical reports and research documents, saving time for healthcare professionals.
Future research should focus on developing and evaluating GenAI applications in healthcare, addressing the unique ethical and regulatory considerations of this domain, such as patient privacy and data security.

8.2 Other Potential Domains

Beyond finance and healthcare, GenAI has the potential to transform many other industries:
* **Education:** GenAI could be used to create personalized learning experiences and automate administrative tasks. * **Manufacturing:** GenAI could be used to optimize manufacturing processes and improve product quality. * **Retail:** GenAI could be used to personalize customer experiences and improve inventory management. * **Media and Entertainment:** GenAI could be used to create new forms of content, such as music, art, and videos.
Future research should explore the potential of GenAI in these and other domains, addressing the specific challenges and opportunities of each industry.

9 GENERATIVE AI APPLICATIONS IN FINANCE: OPPORTUNITIES AND CONSIDERATIONS

As discussed previously, while none of the sources focus *specifically* on finance, we can infer potential applications based on GenAI's capabilities. Gupta [4] highlights the importance of cost optimization, which is a key concern for financial institutions deploying AI. Furthermore, the scalable architectures discussed in [13], [23] are relevant to the high-volume transaction processing requirements of the finance industry. Future research should explore specific applications such as fraud detection, risk management, and algorithmic trading, as well as the unique regulatory and security requirements of this domain.
While the provided resources offer a broad overview of GenAI and cloud platforms, specific references to applications within the finance industry are limited. However, we can infer potential opportunities and considerations for leveraging GenAI in finance based on the broader themes discussed in the literature. This section outlines these potential applications and highlights the need for further research in this area.

9.1 Potential Applications of GenAI in Finance

Based on the general capabilities of GenAI and the trends discussed in the literature, potential applications in finance include:
* **Fraud Detection:** GenAI models could be trained to identify fraudulent transactions and patterns by analyzing vast amounts of financial data. This could help to improve the accuracy and efficiency of fraud detection systems. Tools for data engineering such as AWS and GCP discussed earlier may be used [21], [24]. * **Risk Management:** GenAI could be used to assess and manage financial risks by simulating different market scenarios and predicting potential losses. This could help financial institutions to make more informed decisions and mitigate risks more effectively. * **Algorithmic Trading:** GenAI models could be used to develop more sophisticated trading algorithms that can adapt to changing market conditions and generate higher returns. * **Customer Service:** GenAI-powered chatbots could be used to provide personalized customer service and support, answering customer queries and resolving issues more efficiently. * **Content Creation for Financial Reports:** GenAI can be used to automate the generation of financial reports and summaries, freeing up analysts to focus on more strategic tasks. * **Personalized Financial Advice:** GenAI can be used to provide personalized financial advice to customers based on their individual financial situations and goals.

9.2 Considerations for Implementing GenAI in Finance

Implementing GenAI in finance requires careful consideration of the following factors:
* **Data Quality:** The accuracy and reliability of GenAI models depend on the quality of the data they are trained on. Financial institutions must ensure that their data is clean, accurate, and up-to-date. * **Regulatory Compliance:** The financial industry is heavily regulated, and any GenAI applications must comply with all applicable regulations. * **Ethical Considerations:** The use of GenAI in finance raises ethical considerations, such as the potential for bias and discrimination. Financial institutions must ensure that their GenAI applications are fair and unbiased. * **Security:** Financial data is highly sensitive, and GenAI applications must be secured against unauthorized access and cyberattacks. * **Transparency and Explainability:** The decision-making processes of GenAI models should be transparent and explainable, allowing regulators and customers to understand how decisions are made.

9.3 Gap in the Literature

The lack of specific references to finance in the provided resources highlights a gap in the literature. Future research should focus on exploring the specific opportunities and challenges of using GenAI in the finance industry, addressing the unique requirements and regulatory considerations of this sector. More research is needed to examine data regulations, model training, and data governance in AI for finance.

While the provided resources do not explicitly delve into finance applications, the landscape of GenAI in finance is rapidly evolving and presents a compelling area for future research and practical implementation.

10 IDENTIFIED GAPS AND FUTURE WORK

A review of the provided references reveals several gaps in the current landscape of generative AI cloud infrastructure and suggests avenues for future research and development.

The rapidly evolving field of Generative AI (GenAI) in cloud computing presents numerous opportunities for future research. This section identifies key gaps in the current literature and proposes directions for future work, based on the insights and suggestions found in the reviewed resources.

10.1 Scalability and Cost Optimization

Scalability and cost are crucial considerations for deploying GenAI applications, as these models often require significant computing resources and can incur substantial costs. [13] provides architectural guidelines for creating scalable and resilient solutions on Google Cloud, offering advice on designing systems that can handle fluctuating workloads and maintain high availability. [24] discusses best practices for scalable AI on cloud infrastructure in general, providing platform-agnostic guidance on optimizing resource utilization and minimizing costs. [23] offers a guide to building scalable applications on AWS for up to a million users, providing concrete examples and recommendations for achieving scalability on the AWS platform. [4] provides a practical guide to managing GenAI infrastructure costs across major cloud platforms, offering actionable strategies for optimizing resource allocation, selecting cost-effective instance types, and leveraging cloud-native cost management tools. The paper also discusses methods to optimize the infrastructure costs, engineering costs and other costs like the cost of safety.

10.2 Cost Optimization and Granular Resource Management

Gupta (2025) [4] emphasizes the importance of cost optimization, but a significant gap remains in providing granular resource management tools. Future work should focus on developing dynamic resource allocation strategies that adapt to the fluctuating demands of generative AI workloads. This includes fine-tuning GPU utilization, optimizing data transfer costs, and providing detailed cost breakdown analysis.

10.3 Standardization and Interoperability

The cloud AI landscape is fragmented, with each provider offering proprietary services and APIs. This lack of standardization hinders interoperability and portability of AI models and applications. Future work should explore the development of open standards and frameworks that facilitate cross-platform deployment and integration. This will enable organizations to leverage the best of each cloud platform and avoid vendor lock-in.

10.4 Ethical AI and Responsible Deployment

While ethical AI is increasingly recognized as a crucial consideration, there is a gap in translating ethical principles into practical tools and guidelines. Future work should focus on developing robust AI governance frameworks, bias detection and mitigation techniques, and transparent model explainability tools. These tools should be seamlessly integrated into cloud AI platforms to promote responsible AI deployment.

10.5 Edge AI and Real-time Inference

The potential of edge AI for real-time generative AI applications is significant, but there is a gap in developing efficient model deployment and inference strategies for resource-constrained edge devices. Future work should explore lightweight model architectures, hardware acceleration techniques, and distributed inference frameworks that enable real-time generative AI at the edge.

10.6 Quantitative Performance Benchmarking

While some references provide qualitative comparisons of cloud platforms, there is a lack of comprehensive quantitative performance benchmarks for generative AI workloads. Future work should focus on developing standardized benchmarks that measure the performance of different cloud platforms across various generative AI tasks, including model training, inference, and data processing.

10.7 RAG and Knowledge Integration

Richards (2024) [19] and related articles point to the power of RAG, but there is still a gap in fully automatic and dynamic integration of knowledge. Future work should explore methods to more effectively and dynamically integrate knowledge into LLMs and generative AI models, improving accuracy and relevance of outputs.

10.8 Long-Term Impact of Platform Power

Luitse (2024) [35] and van der Vlist (2024) [25] discuss the growing "platform power" of major cloud providers, but a gap exists in fully understanding the long-term implications of this trend. Future work should investigate the potential for market dominance, the impact on innovation, and the need for regulatory frameworks to ensure fair competition and prevent anti-competitive practices.

10.9 Standardized Benchmarks for GenAI Performance

One significant gap in the current research is the lack of standardized benchmarks for evaluating GenAI performance across different cloud platforms. Van der Vlist et al. [25] highlight the need for more comprehensive studies on the industrialization of AI, which includes developing standardized metrics for assessing GenAI applications. Future research should focus on:

* Developing standardized performance metrics for GenAI applications * Creating benchmark datasets that represent real-world scenarios * Conducting comparative studies of GenAI performance across different cloud platforms

Cost Optimization Strategies

While Gupta [4] provides insights into managing GenAI infrastructure costs, there is a need for more in-depth research on cost optimization strategies. Future work in this area could include:

* Developing predictive models for estimating GenAI infrastructure costs * Exploring novel techniques for optimizing resource allocation in GenAI workloads * Investigating the long-term cost implications of different GenAI architectures

Ethical Considerations and Societal Impact

Luitse [35] touches on the political economy of AI, but there is a need for more comprehensive research on the ethical considerations and societal impact of cloud-based GenAI. Future research should address:

* Developing frameworks for ethical GenAI development and deployment * Investigating the potential biases in GenAI models and proposing mitigation strategies * Studying the long-term societal impacts of widespread GenAI adoption

10.10 Advanced RAG Techniques

While Richards [19] compares RAG solutions across cloud platforms, there is room for further research on advanced RAG techniques. Future work could focus on:

* Developing more sophisticated RAG algorithms that can handle complex, multi-modal knowledge sources * Investigating the integration of RAG with other AI techniques, such as transfer learning and meta-learning * Exploring the use of RAG in specific domains, such as finance or healthcare

Scalability and Resilience

Although some resources discuss scalability [13], [23], there is a need for more research on building highly scalable and resilient GenAI systems. Future work could include:

* Developing novel architectural patterns for scaling GenAI applications to millions of users * Investigating techniques for ensuring the resilience of GenAI systems in the face of failures or attacks * Exploring the use of edge computing and federated learning in GenAI applications

10.11 Integration with Emerging Technologies

There is a gap in research on integrating GenAI with other emerging technologies. Future work could explore:

* The intersection of GenAI with blockchain technology for secure and transparent AI systems * The use of quantum computing to enhance GenAI capabilities * The integration of GenAI with Internet of Things (IoT) devices for edge AI applications

10.12 Domain-Specific GenAI Applications

While some resources discuss general GenAI applications [17], [22], there is a need for more research on domain-specific applications. Future work could focus on:

* Developing and evaluating GenAI applications in specific industries such as healthcare, finance, or education * Investigating the unique challenges and opportunities of GenAI in different domains * Creating domain-specific benchmarks and evaluation criteria for GenAI applications

Environmental Impact

There is a notable gap in research on the environmental impact of large-scale GenAI deployments. Future work should address:

* Assessing the carbon footprint of GenAI training and inference * Developing techniques for reducing the energy consumption of GenAI systems * Exploring the use of renewable energy sources for powering GenAI infrastructure

By addressing these gaps and pursuing these research directions, the field of cloud-based GenAI can continue to advance, leading to more powerful, efficient, and responsible AI systems.

CONCLUSION

Selecting a cloud provider for GenAI depends on workload characteristics, cost constraints, and infrastructure preferences. Future research should explore hybrid-cloud solutions. This paper has provided a comparative analysis of major cloud platforms for scalable generative AI applications. The capabilities of AWS, Azure, and GCP have been examined, highlighting their strengths and weaknesses. The importance of scalability, cost efficiency, and best practices has been emphasized. The continuous advancements and strategic collaborations in the cloud AI space are driving innovation and shaping the future of generative AI.

The reviewed literature highlights the diverse approaches and offerings of major cloud platforms in the rapidly evolving field of Generative AI. AWS, Azure, and GCP are at the forefront, offering comprehensive AI services, robust infrastructure, and a wide array of developer tools. Oracle Cloud is also emerging as a viable alternative, particularly for organizations with existing Oracle investments. RAG and cost optimization are identified as critical areas of focus, reflecting the growing demand for accurate, context-aware, and cost-effective GenAI solutions. The cloud infrastructure is constantly adapting to keep pace with AI advancements. This leads to organizations needing to think strategically about the AI investment to make. Further research is needed to rigorously evaluate the long-term performance, cost-effectiveness, security, and ethical implications of GenAI applications deployed on these platforms. Future research should focus on developing standardized benchmarks for evaluating GenAI performance across different cloud platforms, as well as exploring novel techniques for optimizing resource utilization and minimizing the environmental impact of large-scale GenAI deployments. The political economy of AI is also rapidly evolving leading to the dominance of tech giants.

REFERENCES

[1] “Generative AI on Cloud Platforms: GCP, AWS, and Azure,” CloudThat Resources.

[2] J. MSV, “Generative AI Cloud Platforms: AWS, Azure, or Google?” The New Stack. Jun. 2023.

[3] S. Zaman, “Generative AI Cloud Platforms: Choose from AWS, Azure, or Google Cloud,” Folio3 Cloud Services. Aug. 2023.

[4] J. Gupta, “Generative AI Infrastructure Costs: A Practical Guide to GCP, Azure, AWS, and Beyond,” Cloud Experts Hub. Jan. 2025.
[5] “AWS vs Azure vs GCP Comparison : Best Cloud Platform Guide,” Veritis Group.
[6] “Comparing AWS, Azure, GCP DigitalOcean.” https://www.digitalocean.com/resources/articles/comparing-aws-azure-gcp.
[7] Kanerika, “AWS Vs Azure Vs Google Cloud: How to Choose the Best Cloud Platform?” Kanerika. Sep. 2024.
[8] “Compare Cloud Service Providers.” https://www.oracle.com/cloud/service-comparison/.
[9] A. Verma, “Navigating the Cloud: A Comparative Analysis of GCP, AWS, and Azure,” Medium. https://ai.plainenglish.io/navigating-the-cloud-a-comparative-analysis-of-gcp-aws-and-azure-a3313f11f16a, Feb. 2024.
[10] “Explore Oracle Cloud Infrastructure.” https://www.oracle.com/in/cloud/.
[11] “AWS and NVIDIA Announce Strategic Collaboration to Offer New Supercomputing Infrastructure, Software and Services for Generative AI,” NVIDIA Newsroom. http://nvidianews.nvidia.com/news/aws-nvidia-strategic-collaboration-for-generative-ai.
[12] “Generative AI Infrastructure at AWS AWS Compute Blog.” https://aws.amazon.com/blogs/compute/generative-ai-infrastructure-at-aws/, Jan. 2024.
[13] “The Architecture of a Scalable and Resilient Google Cloud Solution,” InfoQ. https://www.infoq.com/news/2015/04/architecture-google-cloud/.
[14] “Aws sagemaker vs google cloud ai platform: Which Tool is Better for Your Next Project?” https://www.projectpro.io/compare/aws-sagemaker-vs-google-cloud-ai-platform.
[15] “Generative AI on AWS – Generative AI, LLMs, and Foundation Models – AWS,” Amazon Web Services, Inc. https://aws.amazon.com/ai/generative-ai/.
[16] T. Hutcherson, “Building LLM Apps with Redis on Google’s Vertex AI,” Redis. https://redis.io/blog/building-llm-applications-with-redis-on-googles-vertex-ai-platform/, Aug. 2023.
[17] saxenashikha, “Architecting GenAI applications with Google Cloud,” Google Cloud - Community. Sep. 2024.
[18] “Red Hat OpenShift AI.” https://www.redhat.com/en/technologies/cloud-computing/openshift/openshift-ai.
[19] D. Richards, “RAG in the Cloud: Comparing AWS, Azure, and GCP for Deploying Retrieval Augmented Generation Solutions – News from generation RAG.” Mar. 2024.
[20] “Infrastructure for a RAG-capable generative AI application using Vertex AI and AlloyDB for PostgreSQL Cloud Architecture Center,” Google Cloud. https://cloud.google.com/architecture/rag-capable-gen-ai-app-using-vertex-ai.
[21] “Create a generative AI–powered custom Google Chat application using Amazon Bedrock AWS Machine Learning Blog.” https://aws.amazon.com/blogs/machine-learning/create-a-generative-ai-powered-custom-google-chat-application-using-amazon-bedrock/, Oct. 2024.
[22] “Generative AI Application Builder on AWS AWS Solutions AWS Solutions Library,” Amazon Web Services, Inc. https://aws.amazon.com/solutions/implementations/generative-ai-application-builder-on-aws/.
[23] J. Solanki, “How to Build a Scalable Application up to 1 Million Users on AWS,” Simform - Product Engineering Company. Dec. 2018.
[24] “Best Practices for Scalable AI on Cloud Infrastructure,” Yash Technologies. https://www.yash.com/blog/building-scalable-ai-solutions-with-cloud-infrastructure/.
[25] F. van der Vlist, A. Helmond, and F. Ferrari, “Big AI: Cloud infrastructure dependence and the industrialisation of artificial intelligence,” Big Data & Society, vol. 11, no. 1, p. 20539517241232630, Mar. 2024, doi: 10.1177/20539517241232630.
[26] “What is Cloud Elasticity vs Cloud Scalability? Teradata.” https://www.teradata.com/insights/cloud-data-analytics/cloud-elasticity-vs-cloud-scalability, Mar. 2022.
[27] “XenonStack- Generative AI Solutions on AWS.” https://www.xenonstack.com/autonomous-operations/amazon-web-services/.
[28] “AWS Prescriptive Guidance - Cloud design patterns, architectures, and implementations.”
[29] “What’s the Difference Between AWS vs. Azure vs. Google Cloud?” Coursera. https://www.coursera.org/articles/aws-vs-azure-vs-google-cloud, Oct. 2024.
[30] “Well Architecture Framework Azure, AWS, GCP, OCI.” https://www.cloud4c.com/blogs/why-well-architected-frameworks-matter-in-cloud-adoption.
[31] “Building the Future: A Deep Dive Into the Generative AI App Infrastructure Stack,” Sapphire Ventures. https://sapphireventures.com/blog/building-the-future-a-deep-dive-into-the-generative-ai-app-infrastructure-stack/.
[32] A. Takyar, “Generative AI tech stack: Frameworks, infrastructure, models and applications,” LeewayHertz - AI Development Company. https://www.leewayhertz.com/generative-ai-tech-stack/, Mar. 2023.
[33] “Simplified Architecture to take up Generative AI in the Cloud Applications.” https://aitechcircle.kit.com/posts/simplified-architecture-to-take-up-generative-ai-in-the-cloud-applications.
[34] “NVIDIA DGX Cloud,” NVIDIA. https://www.nvidia.com/en-us/data-center/dgx-cloud/.
[35] D. Luitse, “Platform power in AI: The evolution of cloud infrastructures in the political economy of artificial intelligence,” Internet Policy Review, vol. 13, no. 2, Jun. 2024.
[36]“What is the AWS CDK? - AWS Cloud Development Kit (AWS CDK) V2.” https://docs.aws.amazon.com/cdk/v2/guide/home.html.
[37] Satyadhar Joshi, “A Literature Review of Gen AI Agents in Financial Applications: Models and Implementations,” International Journal of Science and Research (IJSR), doi: https://www.doi.org/10.21275/SR25125102816.
[38] Satyadhar Joshi, “Advancing innovation in financial stability: A comprehensive review of ai agent frameworks, challenges and applications,” World Journal of Advanced Engineering Technology and Sciences, vol. 14, no. 2, pp. 117–126, 2025, doi: 10.30574/wjaets.2025.14.2.0071.
[39] Satyadhar Joshi, “Agentic Generative AI and the Future U.S. Workforce: Advancing Innovation and National Competitiveness,” Feb. 03, 2025, Social Science Research Network, Rochester, NY: 5126922. doi: 10.2139/ssrn.5126922.

[40] Satyadhar Joshi, "Generative AI: Mitigating Workforce and Economic Disruptions While Strategizing Policy Responses for Governments and Companies," Feb. 12, 2025, Social Science Research Network, Rochester, NY: 5135229. doi: 10.2139/ssrn.5135229.
[41] Satyadhar Joshi, "Implementing Gen AI for Increasing Robustness of US Financial and Regulatory System," IJIREM, vol. 11, no. 6, Art. no. 6, Jan. 2025, doi: 10.55524/ijirem.2024.11.6.19.
[42] Satyadhar Joshi, "Leveraging prompt engineering to enhance financial market integrity and risk management," World J. Adv. Res. Rev., vol. 25, no. 1, pp. 1775–1785, Jan. 2025, doi: 10.30574/wjarr.2025.25.1.0279.
[43] Satyadhar Joshi, "Retraining US Workforce in the Age of Agentic Gen AI: Role of Prompt Engineering and Up- Skilling Initiatives," Communication and Technology, vol. 5, no. 1, 2025.
[44] Satyadhar Joshi, "Review of autonomous systems and collaborative AI agent frameworks," International Journal of Science and Research Archive, vol. 14, no. 2, pp. 961–972, 2025, doi: 10.30574/ijsra.2025.14.2.0439.
[45] Satyadhar Joshi, "Review of Data Engineering and Data Lakes for Implementing GenAI in Financial Risk A Comprehensive Review of Current Developments in GenAI Implementations," Jan. 01, 2025, Social Science Research Network, Rochester, NY: 5123081. doi: 10.2139/ssrn.5123081.
[46] Satyadhar Joshi, "Review of Data Engineering Frameworks (Trino and Kubernetes) for Implementing Generative AI in Financial Risk," Int. J. Res. Publ. Rev., vol. 6, no. 2, pp. 1461–1470, Feb. 2025, doi: 10.55248/gengpi.6.0225.0756.
[47] Satyadhar Joshi, "Review of Data Pipelines and Streaming for Generative AI Integration: Challenges, Solutions, and Future Directions".
[48] Satyadhar Joshi, "The Synergy of Generative AI and Big Data for Financial Risk: Review of Recent Developments," IJFMR - International Journal For Multidisciplinary Research, vol. 7, no. 1, doi: https://doi.org/g82gmx.
[49] Satyadhar Joshi, "The Transformative Role of Agentic GenAI in Shaping Workforce Development and Education in the US," Feb. 01, 2025, Social Science Research Network, Rochester, NY: 5133376. Accessed: Feb. 17, 2025. [Online]. Available: https://papers.ssrn.com/abstract=5133376
[50] Satyadhar Joshi, "Review of Data Engineering Frameworks (Trino and Kubernetes) for Implementing Generative AI in Financial Risk," Int. J. Res. Publ. Rev., vol. 6, no. 2, pp. 1461–1470, Feb. 2025, doi: 10.55248/gengpi.6.0225.0756.
[51] Satyadhar Joshi, "Review of autonomous systems and collaborative AI agent frameworks," International Journal of Science and Research Archive, vol. 14, no. 2, pp. 961–972, 2025, doi: 10.30574/ijsra.2025.14.2.0439.