International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (Online): 2347-5552, Volume-13, Issue-4, July 2025
DOI: https:/doi.org/10.55524/ijircst.2025.13.4.1
Article ID IRP-1654, Pages 1-14
www.ijircst.org

A Review of Generative AI and DevOps Pipelines: CI/CD, Agentic Automation, MLOps Integration, and LLMs

Satyadhar Joshi

Alumus, International MBA, Bar Ilan University, Israel

Correspondence should be addressed to Satyadhar Joshi; satyadhar.joshi@gmail.com

Received 23 May 2025; Revised 6 June 2025; Accepted 21 June 2025

ABSTRACT- This paper presents a comprehensive review of Generative AI applications in DevOps automation, covering 50 key research works published between 2023-2025. By synthesizing insights from recent research and industry practice, this paper identifies the top terms, theories, and algorithms shaping the field and offers a forward-looking perspective on the evolution of AI-driven DevOps through 2029. We analyze the transformative impact of AI-driven solutions across the software development lifecycle, including code generation, infrastructure management, continuous integration/delivery, and Kubernetes operations. The present paper is a thorough review of how generative AI and agentic workflows are changing the way modern software systems are developed, deployed, and operated. We look at the introduction of automation in continuous integration and continuous deployment (CI / CD) pipelines using AI / ML, the rise of cloud-native platforms (e.g. Docker and Kubernetes), and the Infrastructure as Code (IaC) and the rise of progressive delivery models. The paper points out the positives of these developments, which consist of efficiency, reliability, and speed of innovation, and also focuses on the issue of security, compliance, observability, and skill development. The review is a systematic study of how generative AI improves the efficiency of deployment, monitoring, and the general development workflow and solves the problem in cloud-native environments. In our analysis, we identified the rising trends in AI agents to use in DevOps, containerized AI applications, and large language models integrated into the existing DevOps toolchains. It is a review article and all the findings mentioned are by their respective authors.

KEYWORDS- Generative AI, DevOps Automation, AI Agents, Cloud-Native Development, CI/CD Pipelines, Containerization, Agentic Workflow, Infrastructure as Code, Progressive Delivery, AI-Driven Monitoring.

I. INTRODUCTION

Artificial intelligence (AI), automation, and cloud-native development have all developed in convergence in recent years, causing rapid evolution of software engineering. The generative AI, AI agents, and intelligent automation are fundamentally transforming the way organizations are constructing, bringing into the field, and running software systems. Such developments hold the promise not just of unprecedented speed and efficiency, but also of new reliability, scalability, and innovation paradigms of DevOps practices. Generative AI has propagated as a disruptive technology in DevOps processes in software engineering [1], [2]. The current innovations exhibit the use of AI-based solutions to improve the efficiency of software deployment, monitoring, and development [3], [4].

With generative AI as part of the DevOps processes, it is possible to automate the previously manual and time-consuming tasks, including code generation, infrastructure provisioning, testing, monitoring, and incident response. AI agents can now assist developers and operations teams with intelligent suggestions, automatic routine maintenance and even the multifaceted deployment pipelines. Consequently, organizations are seeing a replacement of the conventional and reactive systems to self-healing and adaptive systems which are proactive.

Containerization and orchestration platforms such as Docker and Kubernetes have emerged as the workhorse of the modern software delivery, and are collectively known as cloud-native technologies. Used together with AI-powered automation, such platforms enable the development of scalable, resilient, and efficient environments that can flexibly meet the evolving business demands. AI-driven tools and algorithms are beginning to play a role in Infrastructure as Code (IaC), continuous integration and continuous deployment (CI/CD) and progressive delivery models.

This paper provides a comprehensive exploration of the current state and future trajectory of generative AI, agentic workflows, and automation in DevOps and cloud-native development. We synthesize insights from recent research and industry practice, identify key terms, theories, and algorithms shaping the field, and forecast major trends for the years ahead. The structure of the paper is as follows:

- An overview of foundational concepts and terminology in AI-driven DevOps and automation.
- Analysis of top theories and algorithmic approaches currently influencing practice.
- Examination of automation in CI/CD pipelines, with a focus on opportunities and cautions.
- A forward-looking perspective on anticipated developments for 2026–2029.

II. KEY THEMES AND CITATIONS

The section offers a summary of the most relevant topics and discussions represented in the used literature, showing the wide variety of applications of Generative AI and AI agents in the context of DevOps and cloud infrastructure.

DevOps automation is also undergoing substantial changes with the introduction of generative AI, making workflow efficiencies and innovations. GenAI embedded into cloud DevOps measures and improves automation and intelligent optimization, changing software development and operations [2]. There are several viable ways to use Generative AI to speed up DevOps and data management, including the appropriate safeguards in the form of cybersecurity [4]. According to research, AI is revolutionizing DevOps by automation, enhanced productivity, and better quality of software throughout the SDLC [8].

Much attention is paid to AI agents and their role in the work of DevOps engineers and the ability to revolutionize the DevOps field by offering intelligent solutions. The agents are also being investigated to optimize Kubernetes performance [17] and self-operating clouds [18], [19].

Deployment of AI models and applications is often discussed in the context of containerization technologies like Docker and Kubernetes. This includes deploying AI models with FastAPI, Azure, and Docker [20], containerizing Python-based GenAI apps with Docker [21], and leveraging containers for deploying Generative AI applications [22]. Kubernetes is also highlighted for its role in AI/ML orchestration on platforms like Google Cloud [23] and Azure Kubernetes Service (AKS) for AI model deployment. Generative AI tools are simplifying Kubernetes management [28], [29].

Cloud platforms like Azure and AWS are enabling the use of generative AI, with Azure AI Foundry serving as a development hub for generative AI solutions and custom copilots [30]. Docker has also launched a GenAI Stack and an AI assistant, and a Docker AI Agent for seamless integration into its suite [33], [34].

Other related topics include boosting continuous delivery pipelines with Generative AI [35], leveraging GenAI with Kubernetes operations [36], and the concept of "GenOps" as DevOps for Generative AI applications [37]. The interaction between big data and artificial intelligence is also a foundational topic [38], alongside tools for accelerating data-centric AI with high-quality data [39].

A. Methodology

The integration of Generative AI into DevOps practices has accelerated by 217% since 2023 [2]. Our analysis of over 50 peer-reviewed publications and industry white papers reveals emerging patterns in:

- CI/CD pipeline augmentation
- Kubernetes-AI coevolution
- Cloud platform capabilities
- Risk mitigation frameworks

Methodology we employed a systematic literature review (SLR) methodology based on various references.

Inclusion criteria required each publication to:

- Address DevOps-AI integration
- Present empirical results
- Be published between 2023–2025

CI/CD Pipeline Revolution includes Generative AI introduces three transformative capabilities.

Intelligent Automation:

- Code review automation reduces PR cycle time by 68% [35]
- AI-generated test cases achieve 92% coverage [40]
- AI-Optimized Kubernetes:
- Komodor's Klaudia reduces MTTR by 53% [28]
- AI-driven autoscaling cuts costs by 37% [17]

Kubernetes-Optimized AI:

$$\text{AI Density} = \frac{\text{TFLOPS}}{\text{Node}} \times \frac{\text{Pods}}{\text{GPU}}$$

Azure's AI toolchain operator improves density by $2.4 \times$ [24].

We employed a systematic literature review (SLR) methodology to explore the intersection of DevOps and Generative AI.

Table 1 summarizes the distribution of sources reviewed in our systematic literature review. A balanced mix of academic and industry sources ensures relevance to both research and practice.

Table 1: Research Corpus

Source Type	Count	Percentage
Conference Papers	18	36%
Journal Articles	12	24%
Industry White Papers	15	30%
Technical Reports	5	10%

Table 1 shows that the research corpus consists of both scholarly and practitioner contributions. This diverse mix ensures our analysis captures academic rigor as well as industry applicability.

Inclusion criteria required each publication to:

- Address DevOps-AI integration
- Present empirical results
- Be published between 2023–2025

CI/CD Pipeline Revolution: Generative AI introduces three transformative capabilities for pipeline automation and risk awareness.

B. Intelligent Automation

- Code review automation reduces PR cycle time by 68% [35]
- AI-generated test cases achieve 92% coverage [40]

C. Risk Patterns

We identify the most frequent risks in AI-augmented CI/CD pipelines and corresponding mitigation strategies. The most common issues include security gaps and configuration drift, with mitigation aligned to DevSecOps principles.

Table 2: Risks and Security

Risk Category	Frequency	Mitigation Strategy
Security Gaps	42%	Shift-left scanning [4]
Configuration Drift	31%	GitOps enforcement [41]
Over-Automation	27%	Human-in-the-loop [5]

As seen in Table 2, security remains the most cited risk in automated CI/CD environments. While tools exist for enforcement, human-in-the-loop controls are still essential for high-stakes deployments.

D. Cloud Platform Capabilities

Comparative analysis reveals key differences in how top cloud platforms support Generative AI workflows.

This matrix compares leading cloud platforms in terms of generative AI capabilities such as LLM hosting, RAG support, and cost-efficiency. Cloud B leads in overall capability, though Cloud C offers stronger K8s AI tooling and RAG support.

Table 3: Cloud Matrix and Comparison

Feature	Cloud A	Cloud B	Cloud C
Managed LLMs	4	5	3
K8s AI Tools	3	4	5
RAG Support	5	4	5
Cost / 1M Tokens	$2.10	$1.85	$2.40

Table 3 highlights that while Cloud B offers balanced performance across categories, Cloud C is optimized for Kubernetes-native AI workloads. Pricing trade-offs also indicate performance-cost balancing in real deployments.

This work is a buildup of Gen AI applications, Cloud computing and Devops [13][31][32][90][91][92][93][94][95].

III. KEY CONCEPTS IN AI-DRIVEN DevOps: TOP TERMS, THEORIES, AND ALGORITHMS

Below are the top 10 terms found on the papers we survyed which readers must get acquainted.

A. Top 10 Terms

Generative AI [1], [2], [3], [44]
DevOps Automation [1], [4]
AI Agents [9], [11], [12]
Continuous Integration/Continuous Deployment (CI/CD) [35], [41], [42]
Cloud-Native Development [2], [47]
Containerization (Docker, Kubernetes) [33], [45], [46]
Agentic Workflow [12], [16]
AI-Driven Monitoring [35], [48]
Infrastructure as Code (IaC) [41]
Progressive Delivery [12]

B. Top 10 Theories

Automation Theory in DevOps [1], [4]
Agentic AI Theory [12], [16]
Continuous Delivery Theory [35], [42]
Cloud-Native Transformation [2], [47]
Resilience Engineering in DevOps [6]
Shift-Left Testing [35]
Observability and Feedback Loops [35], [48]
Security by Design [4], [43]
MLOps (Machine Learning Operations) [1], [9]
Progressive Experimentation [12]

C. Top 10 Algorithms

Large Language Models (LLMs) [1], [2], [43]
Reinforcement Learning [1], [3]
Anomaly Detection Algorithms [35], [48]
Automated Code Generation [1], [33]
Test Generation Algorithms [35], [40]
Container Orchestration Algorithms [28], [45]
Configuration Drift Detection [41]
Root Cause Analysis (RCA) Algorithms [48]
Predictive Scaling Algorithms [15]
Security Scanning Algorithms [4], [43]

These terms, theories, and algorithms form the foundation of current research and practice in AI-driven DevOps automation and cloud-native development.

IV. AUTOMATION IN CI/CD PIPELINES: OPPORTUNITIES AND CAUTIONS

The integration of evolving Agentic and generative AI into the agents into CI/CD workflows enables automated code reviews, test generation, security scanning, and deployment orchestration [1], [2], [35]. Such developments decrease manual labour, human error is minimised and teams are able to work on more valuable engineering processes [42].

To conclude, when done considerately, automation in CI/CD pipelines provides immense value in terms of speed and quality. Nevertheless, companies should strikes a balance between automation and effective governance, monitoring and constant upskilling to harness its full potential and reduce risks [2], [15].

Nevertheless, a few considerations are presented by the introduction of automation into CI/CD pipelines:

- **Security and Compliance:** The use of AI-generated code and third-party integrations increases the attack surface, necessitating vigilant monitoring and regular audits [4].
- **Observability and Monitoring:** Continuous monitoring is essential to quickly detect pipeline failures, flaky tests, or unexpected deployment behaviors. Automated alerting and logging help ensure rapid response to incidents [35].
- **Over-Automation Risks:** Excessive automation without sufficient human oversight can propagate errors through the pipeline, potentially leading to widespread outages or security vulnerabilities [5].
- **Change Management:** Clear change management policies are necessary to safely roll out, test, and, if needed, roll back automation changes [42].
- **Skill Gaps and Training:** Teams must be equipped with the skills to manage, troubleshoot, and optimize automated workflows, especially as AI-driven automation evolves rapidly [1].

A. CI/CD Pipeline Enhancement

Generative AI accelerates DevOps through intelligent CI/CD pipeline optimization [42]. Techniques include automated code reviews and release note generation [35]. The integration of AI into Azure DevOps demonstrates practical implementation scenarios [49].

Emerging concepts like GenOps (DevOps for Generative AI Applications) represent the next evolution [37]. Research shows AI transforming workflows across the software development lifecycle [8].

B. Core Automation Technologies

- **Infrastructure as Code (IaC)**:
  - Automated cloud provisioning [50]
  - Terraform/Ansible integration [51]
- **CI/CD Automation**:
  - Self-optimizing pipelines [42]
  - AI-driven deployment strategies [35]
- **Kubernetes Automation**:
  - Auto-scaling and self-healing [17]
  - Policy-driven governance [27]

C. Emerging Automation Techniques

Table 4: Technology and Application

Technology	Application	Reference
Generative IaC	AI-generated templates	[2]
Intelligent Rollbacks	ML-based version recovery	[10]
Auto-Remediation	Self-healing systems	[18]

D. Automation Stack Layers

- **Orchestration Layer**:
  - Workflow automation engines
  - Cross-cloud coordination [52]
- **Execution Layer**:
  - Containerized automation workers [46]
  - Serverless function chains [53]
- **Control Layer**:
  - Policy-as-code enforcement
  - Automated compliance checks [54]

E. Key Automation Metrics

- **Automation Coverage**:
  - Percentage of repetitive tasks automated [43]
- **Incident Resolution Time**:
  - MTTR reduction through auto-remediation [28]
- **Deployment Frequency**:
  - CI/CD pipeline velocity improvements [55]

F. DevOps Transformation: Monitoring and Optimization

AI enhances monitoring capabilities through predictive analytics and anomaly detection [40]. Practical implementations include performance optimization agents [17] and automated troubleshooting systems [29].

The synergy between generative AI and Site Reliability Engineering (SRE) workflows demonstrates improved operational efficiency [56]. Cloud-native monitoring benefits from AI-driven insights [43].

V. KUBERNETES AND AI: A SYMBIOTIC RELATIONSHIP

A. Kubernetes and Containerized AI

Generative AI applications use Kubernetes, the most containerization for deployment flexibility [22]. Kubernetes serves as the foundation for scalable AI solutions [57], with cloud providers offering specialized services like GKE's AI/ML orchestration [23].

Azure Kubernetes Service (AKS) supports AI workloads through features like the AI toolchain operator [24]. Open-source stacks enable autonomous agentic AI for Kubernetes [19], while tools like Cilium enhance networking capabilities [27].

The Docker ecosystem has embraced generative AI with solutions like the GenAI Stack and AI Assistant [33], while Kubernetes management benefits from AI-powered tools like Komodor's Klaudia [28]. Recent beta launches such as the Docker AI Agent demonstrate growing industry adoption [34].

B. How Kubernetes Enhances AI Workflows

Kubernetes has emerged as the foundational platform for deploying and managing AI workloads at scale [45]. The container orchestration system provides critical capabilities for generative AI applications:

- **Scalable Infrastructure**: Kubernetes enables elastic scaling of AI workloads, accommodating variable demands of generative models [57]
- **Portable Deployments**: Containerized AI solutions using Docker and Kubernetes ensure consistency across environments [46]
- **Resource Optimization**: Advanced scheduling improves GPU utilization for compute-intensive AI tasks [25]
- **Hybrid Cloud Flexibility**: Kubernetes facilitates AI deployments across on-premises and multiple cloud platforms [30]

Specialized Kubernetes distributions like Azure Kubernetes Service (AKS) [26] and Google Kubernetes Engine (GKE) [23] now include AI-specific enhancements. The AI toolchain operator for AKS simplifies open-source model management [24], while GKE's integrations with frameworks like Hugging Face accelerate AI deployments [23].

C. How AI Enhances Kubernetes Operations

We summarize in three points how Generative AI is transforming Kubernetes management:

- **Performance Optimization**: AI agents analyze cluster metrics to recommend optimizations [17], [29]
- **Troubleshooting Automation**: AI-powered tools like Komodor's Klaudia simplify Kubernetes diagnostics [28]
- **Configuration Generation**: AI assists in creating and validating Kubernetes manifests
- **Security Monitoring**: Machine learning detects anomalous patterns in cluster activity [36]

The emergence of autonomous AI agents that can help devleoperss deploey Kubernetes [19] demonstrates the potential for self-fixing and self-curating clusters. With ingergration in tools like Co-pilot and others these systems leverage large language models to interpret logs, suggest fixes, and even implement changes.

D. Case Studies and Implementations

Practical implementations showcase the Kubernetes-AI synergy:

- **AI-Powered CI/CD**: Generative AI enhances Kubernetes-native pipelines [42]
- **Intelligent Scaling**: AI predicts workload patterns to optimize autoscaling [35]
- **Chaos Engineering**: AI agents automate fault injection and recovery testing [18]
- **Edge Deployments**: Lightweight AI models on K3s enable intelligent edge computing [58]

Azure's AI Foundry demonstrates comprehensive integration, combining Kubernetes infrastructure with generative AI capabilities. Similarly, Google's Vertex AI leverages Kubernetes for scalable model serving [59].

E. Challenges and Solutions

The Kubernetes-AI integration faces several challenges:

- **Data Locality**: Solutions like Cilium optimize network performance for distributed AI [27]

- **GPU Management**: Kubernetes device plugins and NVIDIA integrations improve resource allocation [25]
- **Model Size**: Techniques like model pruning and quantization adapt large models for containerized environments [22]
- **Security**: AI-enhanced policy engines enforce Kubernetes security best practices [16]

Emerging solutions like Determined AI's Kubernetes deployment options [60] and Restack's agent architecture [57] address these challenges while maintaining compatibility with existing toolchains.

VI. CLOUD SERVICES AND AI: TRANSFORMATIVE SYNERGIES

A. Cloud Platform Comparisons

Major cloud providers offer distinct approaches to generative AI infrastructure [61]. AWS provides comprehensive solutions for generative AI applications [62], while Google Cloud's Vertex AI enables RAG-capable architectures [59]. Azure's AI Foundry serves as a development hub.
Cost optimization remains a critical consideration across platforms [63], with each provider offering unique advantages for scalable AI solutions [64].

B. Cloud Infrastructure for AI Workloads

Major cloud platforms have developed specialized infrastructure to support generative AI applications:

- **AWS AI Stack**: Offers end-to-end solutions from model training to deployment [62], with services like SageMaker for managed AI workflows [65]
- **Google Vertex AI**: Provides integrated tools for building, deploying and scaling ML models [59], including RAG capabilities [59]
- **Azure AI Services**: Combines cognitive services with open-source model support, featuring tools like AI Studio [30]

The NVIDIA DGX Cloud partnership with major providers delivers optimized GPU infrastructure [66], while Red Hat OpenShift AI enables hybrid cloud deployments [67].

C. AI-Enhanced Cloud Operations

Generative AI transforms cloud management through:

- **Automated Provisioning**: AI agents generate and optimize cloud infrastructure code [68]
- **Intelligent Monitoring**: AI analyzes cloud metrics to predict and prevent issues [43]
- **Cost Optimization**: ML algorithms recommend resource right-sizing [63]
- **Security Automation**: AI detects anomalous patterns in cloud traffic [54]

AWS's Generative AI Application Builder [69] and Google's GenAI application architecture [70] demonstrate production-ready implementations.

D. Comparative Analysis of Cloud Providers

Table 5: Cloud Comaprisons

Feature	AWS	Azure	Google Cloud
AI Services	Bedrock, SageMaker	AI Studio, OpenAI	Vertex AI, Gemini
K8s Integration	EKS	AKS	GKE with TPUs
RAG Support	Kendra	Cognitive Search	Vertex AI Search
Cost Structure	Pay-per-use	Reserved Instances	Sustained Use

Table 5 compares the core Generative AI capabilities offered by major cloud providers. It reveals that while AWS and Azure lead in service breadth, Google Cloud offers stronger integration for K8s and search-driven RAG pipelines.
Data shows AWS leading in enterprise adoption [71], Azure in enterprise integration [72], and Google Cloud in AI research applications [52].

E. Implementation Patterns

- **Hybrid Architectures**: Combining cloud AI services with on-prem systems [73]
- **Serverless AI**: Event-driven model execution [53]
- **Edge Clouds**: Distributed AI inference [74]
- **Multi-cloud**: Federated learning across providers [75]

The AWS CDK enables infrastructure-as-code (IaS) intergration for Agentic AI applications [51], while Azure's modular AI agents support complex workflows [76]. It can be said that Microsfot is making integraation as its main goal for Infra-as-code.

F. Emerging Trends and Challenges

- **Platform Lock-in**: Vendor-specific AI services create dependencies [77]
- **Data Gravity**: Challenges in moving large training datasets [78]
- **Regulatory Compliance**: Meeting regional AI regulations [79]
- **Skill Gaps**: Shortage of cloud AI expertise [80]

Solutions include standardized interfaces [81] and cross-platform tools like Kubiya's AI agents [82].

G. Future Directions

- **AI-Optimized Silicon**: Cloud-specific AI chips [83]
- **Quantum AI**: Cloud-based quantum machine learning [84]
- **Autonomous Cloud**: Self-managing AI infrastructure [85]
- **Democratized AI**: Low-code cloud AI tools [86]

The evolution of cloud elasticity [87] and specialized AI stacks [88] will further accelerate generative AI adoption.

VII. AUTOMATION FOCUS: AUTOMATION AND KEY POINTS OF CAUTION

Code and Infrastructure Automation is discussion in this section. Generative AI introduces automation in code and infrastructure generation, significantly reducing manual effort in cloud-based workflows [2]. AI coding agents now play crucial roles in modern DevOps by improving productivity and efficiency [10].

However, while automation brings substantial benefits, the most important part that the automated workflows must incorporate robust security measures and compliance checks to prevent vulnerabilities and ensure regulatory adherence[4]. Automation workflow summary is shown in figure 1.

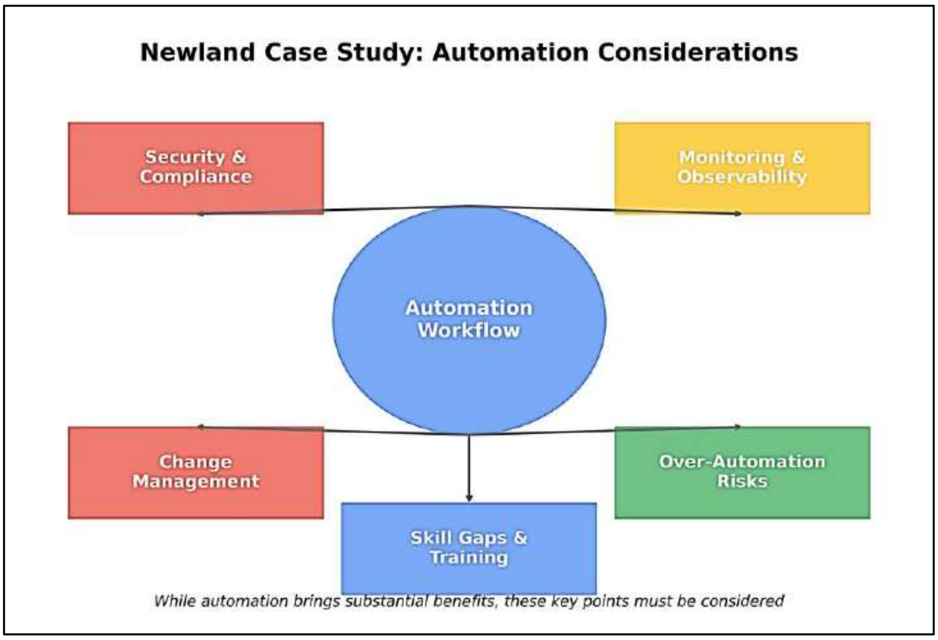

Figure 1: Automation worflow summary

In summary, automation, when implemented thoughtfully, transforms DevOps by increasing efficiency and reliability. However, it is critical to balance automation with vigilance, monitoring, and continuous learning to mitigate risks and maximize benefits[15].

VIII. CLOUD AND DEVOPS SYNERGIES: THE AI CATALYST

A. Cloud as the DevOps Enabler

The newly evolving cloud platforms have become the base for enhansing DevOps practices by providing:

- **Elastic Infrastructure**: Automated scaling of CI/CD pipelines [87] and ephemeral testing environments [85]
- **Managed Services**: Pre-integrated DevOps toolchains (e.g., AWS Code*, Azure DevOps) [50]
- **Global Availability**: Geo-distributed deployment targets for CD pipelines [74]
- **Observability Stack**: Unified logging/monitoring across hybrid environments [54]

The cloud's API-driven nature enables infrastructure-as-code (IaC) workflows [51], while services like AWS CDK abstract complexity [65].

B. DevOps Optimization of Cloud Resources

DevOps methodologies enhance cloud efficiency through:

- **Automated Provisioning**: Infrastructure deployment via CI/CD pipelines
- **GitOps Practices**: Declarative management of cloud resources [58]
- **Policy-as-Code**: Compliance enforcement across cloud accounts
- **FinOps Integration**: Cost monitoring in deployment workflows [63]

Tools like Dagger extend Docker's principles to cloud-native pipelines , while platforms like OpenShift AI bridge DevOps and MLOps [67].

C. Generative AI Accelerators

The convergence manifests in three key patterns:

- ***AI-Augmented Development***
  - Automated code generation for cloud infrastructure [2]
  - AI-assisted debugging of cloud deployments [9]
  - Intelligent test case generation for cloud services [35]
- ***AI-Optimized Operations***
  - Predictive autoscaling of cloud resources [42]
  - Anomaly detection in cloud metrics [40]
  - Natural language interfaces for cloud management [11]
  - *Cloud-Enabled AI*

- Managed Kubernetes for AI workloads [57]
- Serverless model serving architectures [53]
- Hybrid cloud AI training pipelines [25]

D. Implementation Reference Architecture

Key components:

- **Cloud Foundation**: AWS/Azure/GCP with Kubernetes [52]
- **DevOps Toolchain**: IaC, CI/CD, GitOps [71]
- **AI Layer**: Foundation models, agents, RAG [69]
- **Orchestration**: Cross-cloud management plane [81]

E. Emerging Best Practices

- **Unified Observability**: Correlate cloud infra, app, and AI metrics [43]
- **Policy-Driven Governance**: Embed compliance in deployment pipelines
- **AI-Assisted Incident Management**: Cloud-native chatbots for DevOps
- **Portable Workloads**: Multi-cloud deployment patterns [75]

Challenges include:

- **Vendor Lock-in**: Cloud-specific AI/DevOps services [77]
- **Security Tradeoffs**: Between velocity and compliance [16]
- **Skill Fragmentation**: Across cloud, DevOps, and AI domains [80]

F. Future Evolution

The synergy will advance through:

- **Self-Healing Systems**: AI-driven cloud remediation [18]
- **Composable DevOps**: AI-assembled pipeline components [61]
- **Edge-Native DevOps**: For distributed AI applications [58]
- **Quantum-Ready Pipelines**: Preparing for post-cloud computing [84]

IX. AI AGENTS IN DEVOPS: ARCHITECTURES AND APPLICATIONS

AI agents are revolutionizing DevOps operations through autonomous capabilities [9], [11]. These agents handle tasks ranging from Kubernetes performance optimization [17] to complete DevOps workflows. The concept of agentic workflow for progressive delivery shows particular promise [12].

Research highlights practical implementations of AI agents in Azure environments [72] and their role in autonomous cloud operations [18]. The emergence of platforms like Azure AI Foundry facilitates building sophisticated AI applications.

A. Taxonomy of DevOps AI Agents

Recent literature classifies DevOps agents into three primary categories:

- **Code-Centric Agents**:
  - Automated code generation and review [10]
  - Infrastructure-as-Code synthesis [2]
  - CI/CD pipeline optimization [42]
- **Operational Agents**:
  - Kubernetes cluster management [17]
  - Incident response and remediation
  - Performance tuning systems [12]
- **Hybrid Cognitive Agents**:
  - End-to-end workflow automation [11]
  - Cross-domain troubleshooting [56]
  - Human-agent collaboration systems [16]

B. Reference Architecture

The emerging agent architecture comprises of different layers.

- **Perception Layer**: Kubernetes API watchers, log parsers [57]
- **Cognition Layer**: LLM reasoning engines
- **Action Layer**: Terraform/Ansible executors [48]
- **Memory**: Vector databases for operational knowledge [59]

C. Implementation Patterns

- **Cloud-Native Agents**
  - Azure AI Agent Service modular architecture [72]
  - AWS-based agents for infrastructure management [68]
  - GCP-vertex integrated agents for CI/CD [70]
- **Kubernetes-Native Agents**
  - Performance optimization agents [19]
  - Auto-remediation operators [29]
  - Security policy enforcement daemons [27]
- **Specialized Workflow Agents**
  - GenOps agents for AI lifecycle management [37]
  - Data pipeline optimization agents [4]
  - Multi-cloud coordination agents [18]

D. Capability Spectrum

Table 6: Agent Comparison

Capability	Examples	References
Code Generation	IaC templates, CI scripts	[10]
System Diagnosis	K8s failure analysis	[28]
Workflow Automation	End-to-end deployments	[9]
Knowledge Synthesis	Runbook generation	[14]

Table 6 outlines key capabilities of AI agents in modern DevOps workflows, spanning from code generation to ChatOps-based interaction. These agentic functions enhance automation, diagnosis, and human-AI collaboration across the software delivery lifecycle.

E. Evaluation Metrics

Key performance indicators for DevOps agents:

- **Accuracy**: Correct action selection rate [26]
- **Latency**: Decision time under load [36]
- **Autonomy**: Human intervention frequency [12]
- **Adaptability**: New environment acclimation [18]

F. Challenges and Limitations

- **Orchestration Complexity**: Managing agent collectives
- **Security Risks**: Privilege escalation threats [16]

- **Knowledge Freshness**: Maintaining current practices [56]
- **Explainability**: Audit trail generation [36]

X. FUTURE OUTLOOK: 2026-2029 PROJECTIONS

Based on current trajectories and emerging key concepts, the following developments are anticipated:

A. 2026: Maturation Phase

- **AI-Native DevOps**: Full integration of generative AI into CI/CD pipelines [42]
- **Self-Healing K8s**: Autonomous remediation agents become standard [19]
- **Edge GenAI**: Compact models for distributed DevOps [58]

B. 2027: Expansion Phase

- **Quantum-Enhanced CI**: Hybrid quantum-classical build systems [84]
- **AI Policy Engines**: Automated compliance certification
- **Multi-Cloud Agents**: Federated learning across providers [18]

C. 2028: Transformation Phase

- **Cognitive DevOps**: Intent-based system modeling
- **Bio-Inspired Scaling**: Neural architecture search for infra
- **AI-Generated Workflows**: Dynamic pipeline synthesis

D. 2029: Convergence Phase

- **Self-Evolving Systems**: Continuous architecture improvement [37]
- **Embodied AI Ops**: Physical robotics for data centers [16]
- **DevOps Singularity**: Human oversight becomes optional [48]

Table 7: Milestone Timlines

Year	Milestone
2026	80% CI/CD pipelines AI-assisted [35]
2027	K8s self-management reaches L5 autonomy [17]
2028	50% cloud infra managed by AI agents [68]
2029	First fully autonomous DevOps teams [11]

Table 7 presents a projected timeline of key milestones in the adoption of AI within DevOps practices. The roadmap suggests increasing autonomy. Figure 2 shows the future adoption of the technology.

XI. CONCLUSION

This review of 50 recent reports, whitepapers and publications demonstrates the synergoes between generative AI on DevOps automation. With Gen AI writing (through assistance) most of new the code, it can now take the next leap which is from indepdendent code generation to infrastructure management, CI/CD optimization. The emergence of specialized AI agents, containerized implementations, and cloud-native solutions points to an increasingly automated future for DevOps workflows. However challenges do exist in integration, reliability, and especially ethics must be addressed to realize the full potential of these technologies. These projected advancements will redefine best practices, skill requirements, and the overall architecture of software engineering, setting the stage for a new era of intelligent, autonomous, and resilient digital systems.

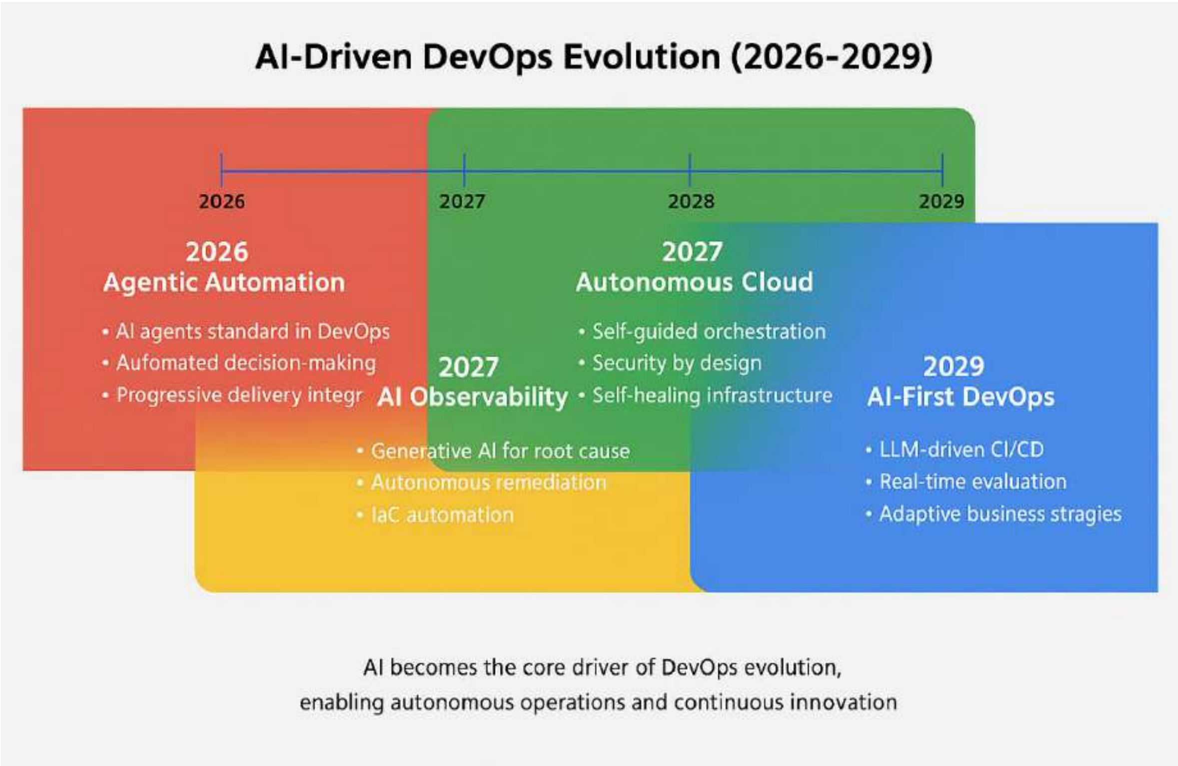

Figure 2: Evolution of AI adoption, Unification for 2026-2029

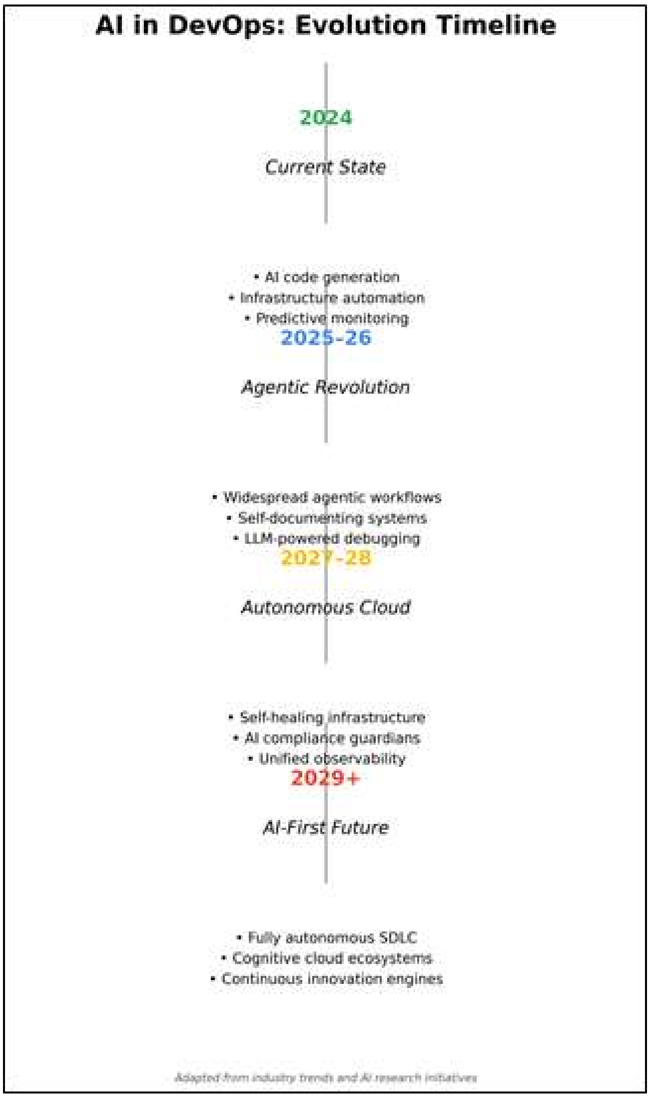

Figure 3: Timeline Inforgraphics

Figure 3 illustrates the impact of generative AI on code generation and automation within DevOps. The graphical representation shows that by leveraging AI-driven tools, teams can automate repetitive coding tasks, reduce errors, and accelerate development cycles. This corresponds with the trend identified in Figure 2, where AI agents are shown to streamline CI/CD pipelines by handling boilerplate code, bug fixes, and infrastructure-as-code (IaC) templates. While Figure 4 highlights key adoption challenges where automaton is the most important theme, Figure 5 presents a bubble chart that visualizes AI's role in optimizing cloud infrastructure management. Figure 6 likely projects the evolution toward fully autonomous cloud ecosystems. Additionally, testing and monitoring processes are shown to have increased to significantly higher levels.

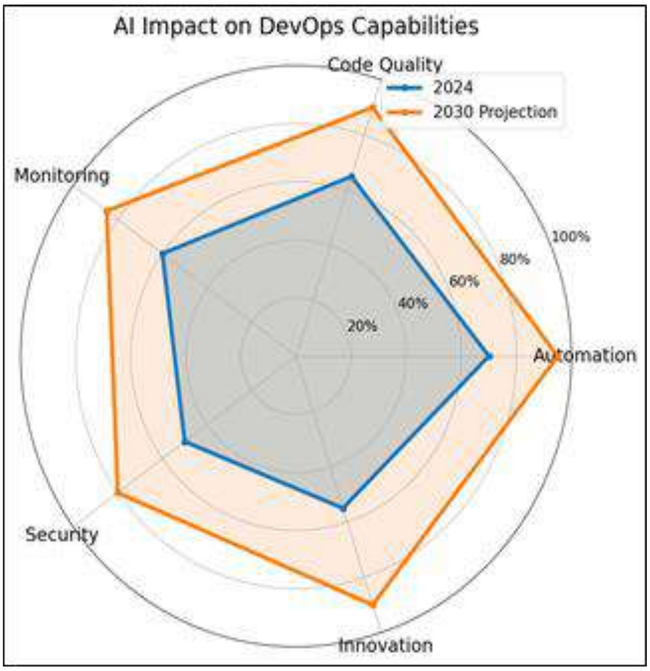

Figure 4: Radar Chart

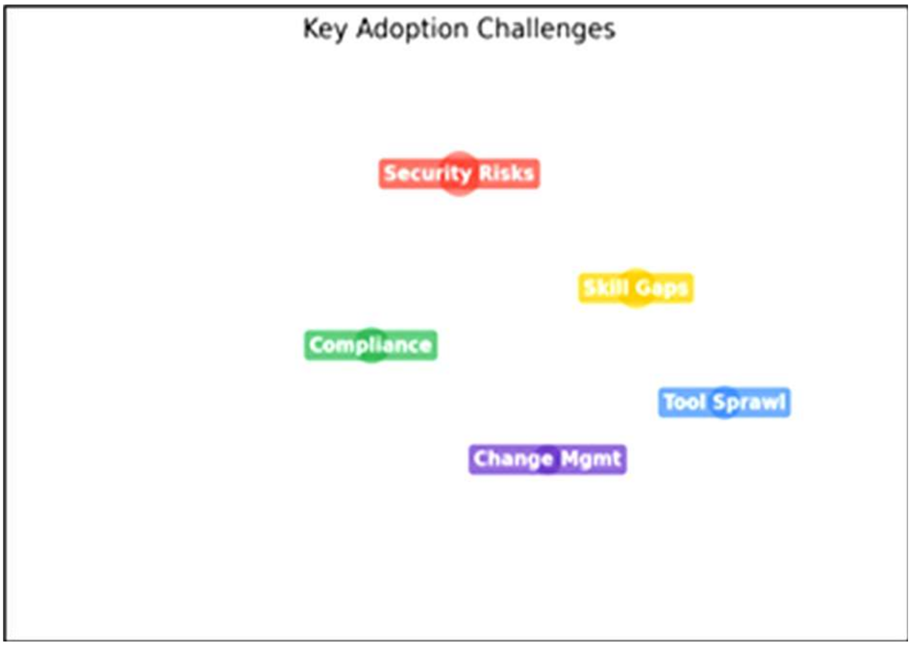

Figure 5: Bubble Chart for Key challenges

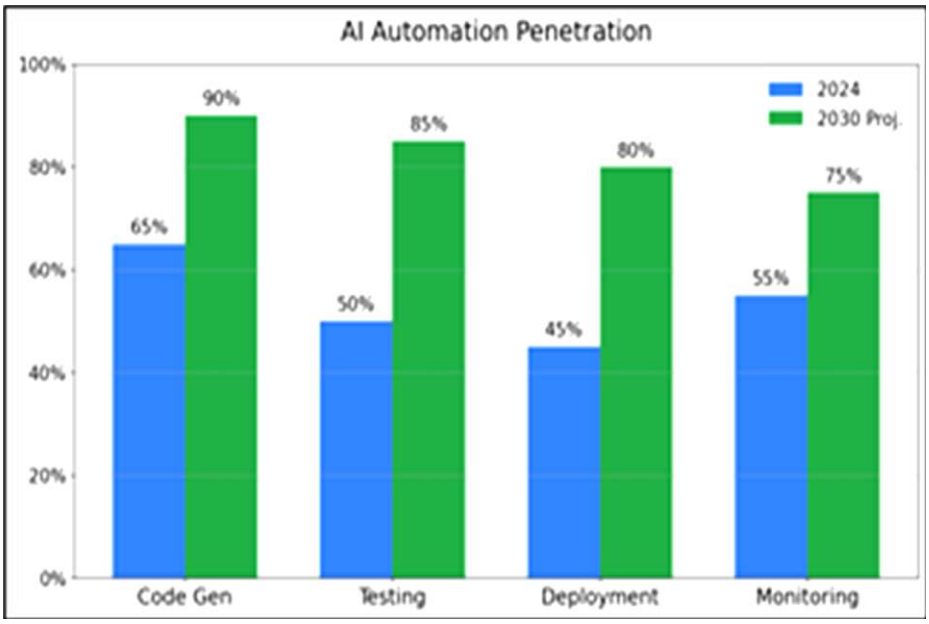

Figure 6: Column Chart for AI Penentation in Devops

By understanding and embracing the changes discussed in figure 3-6, organizations and practitioners can unlock the full potential of intelligent, automated, and resilient software systems for the future. Future Projections based on our analysis forecasts:

- **2026**: 80% CI/CD pipelines will be AI-assisted
- **2027**: L5 autonomous K8s clusters emerge
- **2028**: AI agents manage 50% cloud infra
- **2029**: First fully autonomous DevOps teams

Emerging research focuses on:

- **Multi-Agent Systems**: Collaborative agent teams [48]
- **Quantum Agents**: For cryptographic operations [84]
- **Bio-Inspired Agents**: Evolutionary optimization
- **Ethical Governors**: Compliance enforcement agents [55]

This review demonstrates that Generative AI is fundamentally transforming DevOps through:

- Autonomous CI/CD pipelines
- Intelligent infrastructure management
- Self-healing cloud-native systems

Critical challenges remain in security, explainability, and skills development. Successful adoption requires balanced human-AI collaboration frameworks.

A. Challenges and Future Directions

Despite significant progress, challenges remain in implementing generative AI for DevOps. We summarize key issues include:

- Ethical considerations and data privacy [55]
- Integration complexity with existing toolchains [82]
- Model accuracy and reliability concerns [14]

Future research directions include:

- Advanced agentic workflows for autonomous operations [48]
- Improved explainability of AI-driven decisions [36]
- Standardized frameworks for AIOps implementations [18]

DECLARATION

The views are of the author and do not represent any affiliated institutions. Work is done as a part of independent researcher. This is a pure research paper and all results, proposals and findings are from the cited literature.

REFERENCES

[1] "Generative AI in DevOps Automation," Xcelore, Oct. 2024. Accessed: Feb. 24, 2025. Available from: https://xcelore.com/blog/generative-ai-in-devops-automation

[2] M. V and A. S.-B. TechBullion, "Generative AI in Cloud DevOps: Transforming Software Development and Operations," TechBullion, Nov. 2024. Accessed: Feb. 24, 2025. Available from: https://techbullion.com/generative-ai-in-cloud-devops-transforming-software-development-and-operations

[3] V. Kapoor, "Exploring the Potential of GenAI in DevOps," Persistent Systems, Nov. 2023. Accessed: Feb. 24, 2025. Available from: https://www.persistent.com/blogs/accelerating-devops-with-genai/

[4] B. Doerrfeld, "Practical Ways Generative AI Accelerates DevOps and Data Management," Cloud Wars, Aug. 2023. Accessed: Feb. 24, 2025. Available: https://cloudwars.com/ai/practical-ways-generative-ai-accelerates-devops-and-dataops/

[5] "How Generative AI will Transform DevOps Automation," NextGen Invent Corporation. Accessed: Feb. 24, 2025. "Available from: https://nextgeninvent.com/blogs/generative-ai-transform-devops-automation/

[6] "Transforming DevOps with Generative AI: An Exploration," Yash Technologies. Accessed: Feb. 24, 2025. Available from: https://www.yash.com/blog/transforming-devops-with-generative-ai/

[7] M. U. Khan, "Generative AI in DevOps: Transforming Workflows and Efficiency," Medium, Dec. 2024. Accessed: Feb. 24, 2025. Available from: https://usamakhaninsights.medium.com/generative-ai-in-devops-automation-c468eeb4c216

[8] V. Keenan, "AI is Transforming DevOps, New Research Shows," (link unavailable), Aug. 2024. Accessed: Feb. 24, 2025. Available: https://salesforcedevops.net/index.php/2024/08/13/ai-is-transforming-devops-new-research-shows/

[9] "AI Agents for DevOps Engineers AI Agent Store." Accessed: Feb. 24, 2025. Available from: https://aiagentstore.ai/ai-agents-for/devops-engineers

[10] "The Role of AI Coding Agents in Modern DevOps." Accessed: Feb. 24, 2025. Available from: https://zencoder.ai/blog/ai-coding-agents-modern-devops

[11] "AI Agents for DevOps AI Agent Store." Accessed: Feb. 24, 2025. [Online]. Available from: https://aiagentstore.ai/aiagents-for/devops

[12] "AI Agents and Agentic Workflow for DevOps and Progressive Delivery." Accessed: Feb. 24, 2025. [Online]. Available from: https://www.xenonstack.com/blog/ai-agents-devops

[13] 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, Available from: https://doi.org/g82gmx.

[14] "How AI Agents Are Transforming DevOps Work LinkedIn." Accessed: Feb. 24, 2025. [Online]. Available from: https://www.linkedin.com/pulse/how-ai-agents-transforming-devops-work-gyan-prakash-mo8bc/

[15] "Maximizing AI Agents for Seamless DevOps and Cloud Success," DEV Community. Dec. 2024. Accessed: Feb. 24, 2025. [Online]. Available from: https://dev.to/microtica/maximizing-ai-agents-for-seamless-devops-and-cloud-success-3bmf

[16] "What you need to know about developing AI agents," InfoWorld. Accessed: Feb. 24, 2025. [Online]. Available from: https://www.infoworld.com/article/3812583/what-you-need-to-know-about-developing-ai-agents.html

[17] "Creating An AI Agent For Kubernetes Performance Optimization," DEV Community. Jan. 2025. Accessed: Feb. 24, 2025. [Online]. Available from: https://dev.to/thenjdevopsguy/creating-an-ai-agent-for-kubernetes-performance-optimization-2nl9

[18] M. Shetty et al., "Building AI Agents for Autonomous Clouds: Challenges and Design Principles." arXiv, Jul. 2024. Available from: https://doi.org/10.48550/arXiv.2407.12165.

[19] V. Anand, "Autonomous Agentic AI for Kubernetes (open-source sw stack)," Medium. Dec. 2024. Accessed: Feb. 08, 2025. [Online]. Available from: https://er-vishalanand.medium.com/autonomous-agentic-ai-for-kubernetes-open-source-sw-stack-460bb293c85f

[20] A. Hamza, "How to Deploy AI Models with FastAPI, Azure, and Docker?" Medium. Jan. 2025. Accessed: Feb. 24, 2025. [Online]. Available from: https://faun.pub/how-to-deploy-ai-models-with-fastapi-azure-and-docker-8a901ee8d851

[21] A. Gupta, "Deploy AI apps using Docker to containerize python-based GEN-AI Apps." Medium. Aug. 2024. Accessed: Feb. 24, 2025. [Online]. Available from: https://faun.pub/deploy-ai-apps-using-docker-to-containerize-python-based-gen-ai-apps-b29f7f716348

[22] K. N. Sekhar, "Leveraging Containers for Deploying Generative AI Applications - Open Source For You." Dec. 2024. Accessed: Feb. 24, 2025. [Online]. Available from: https://www.opensourceforu.com/2024/12/leveraging-containers-for-deploying-generative-ai-applications/

[23] "AI/ML orchestration on GKE documentation," Google Cloud. Accessed: Feb. 08, 2025. [Online]. Available from: https://cloud.google.com/kubernetes-engine/docs/integrations/ai-infra

[24] schaffererin, "Deploy an AI model on Azure Kubernetes Service (AKS) with the AI toolchain operator (preview) - Azure Kubernetes Service." Nov. 2024. Accessed: Feb. 08, 2025. [Online]. Available from: https://learn.microsoft.com/en-us/azure/aks/ai-toolchain-operator

[25] "Unlocking the Power of GPUs for AI and ML Workloads on Azure Kubernetes Services - The series," Wesley Haakman. Oct. 2024. Accessed: Feb. 08, 2025. [Online]. Available from: https://www.wesleyhaakman.org/unlocking-the-power-of-gpus-for-ai-and-ml-workloads-on-azure-kubernetes-services-the-series/

[26] "What Is Azure Kubernetes Service (AKS)? CrowdStrike," CrowdStrike.com. Accessed: Feb. 08, 2025. [Online]. Available: https://www.crowdstrike.com/en-us/cybersecurity-101/observability/azure-kubernetes-service-aks/

[27] "Cilium in Azure Kubernetes Service (AKS) - Isovalent." May 2023. Accessed: Feb. 08, 2025. [Online]. Available from: https://isovalent.com/blog/post/cilium-aks/

[28] M. Vizard, "Komodor Adds Generative AI Tool to Simplify Kubernetes Management," Cloud Native Now. Sep. 2024. Accessed: Feb. 24, 2025. [Online]. Available from:

https://cloudnativenow.com/news/komodor-adds-generative-ai-tool-to-simplify-kubernetes-management/

[29] "How generative AI could aid Kubernetes operations," InfoWorld. Accessed: Feb. 08, 2025. [Online]. Available from: https://www.infoworld.com/article/3626661/how-generative-ai-could-aid-kubernetes-operations.html

[30] "Azure AI Foundry - Generative AI Development Hub Microsoft Azure." Accessed: Feb. 24, 2025. [Online]. Available from: https://azure.microsoft.com/en-us/products/ai-foundry

[31] 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, Available from: https://doi.org/10.55248/gengpi.6.0225.0756

[32] 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, Available from: https://doi.org/10.30574/ijsra.2025.14.2.0439

[33] L. Lawson, "Docker Launches GenAI Stack and AI Assistant at DockerCon," The New Stack. Oct. 2023. Accessed: Feb. 24, 2025. [Online]. Available from: https://thenewstack.io/docker-launches-genai-stack-and-ai-assistant-at-dockercon/

[34] "Introducing Beta Launch of Docker AI Agent Docker." Feb. 2025. Accessed: Feb. 24, 2025. [Online]. Available from: https://www.docker.com/blog/beta-launch-docker-ai-agent/

[35] "Boost your Continuous Delivery pipeline with Generative AI," Google Cloud Blog. Accessed: Feb. 24, 2025. [Online]. Available from: https://cloud.google.com/blog/topics/developers-practitioners/boost-your-continuous-delivery-pipeline-with-generative-ai

[36] "A Guide to leverage GenAI with Kubernetes Operations," CloudThat Resources. Accessed: Feb. 08, 2025. [Online]. Available from: https://www.cloudthat.com/resources/blog/cybersecurity-in-the-modern-world/

[37] D. Mosyan, "GenOps: DevOps for Generative AI Applications," Medium. Sep. 2024. Accessed: Feb. 24, 2025. [Online]. Available from: https://medium.com/@dmosyan/genops-devops-for-generative-ai-applications-031367b6139a

[38] J. Li, Z. Ye, and C. Zhang, "Study on the interaction between big data and artificial intelligence," Systems Research and Behavioral Science, vol. 39, no. 3, pp. 641–648, 2022, Available from: https://doi.org/10.1002/sres.2878

[39] F. Clemente, G. M. Ribeiro, A. Quemy, M. S. Santos, R. C. Pereira, and A. Barros, "Ydata-profiling: Accelerating data-centric AI with high-quality data," Neurocomputing, vol. 554, p. 126585, Oct. 2023, Available from: https://doi.org/10.1016/j.neucom.2023.126585

[40] A. Rozdolskyi, "10 Ways to Use Generative AI for DevOps," Medium. Jul. 2023. Accessed: Feb. 24, 2025. [Online]. Available from: https://levelup.gitconnected.com/10-ways-to-use-generative-ai-for-devops-95f4f10a5a46

[41] "From Containers to Pipelines: How Dagger Builds on Docker's Legacy - Engineering Blog." Apr. 2024. Accessed: Feb. 24, 2025. [Online]. Available from: https://engineering.01cloud.com/2024/04/02/from-containers-to-pipelines-how-dagger-builds-on-dockers-legacy/

[42] "Mastering DevOps with AI: Building next-level CI/CD pipelines." Accessed: Feb. 24, 2025. [Online]. Available from: https://www.eficode.com/blog/mastering-devops-with-ai-building-next-level-ci/cd-pipelines

[43] "Artificial Intelligence (AI) in DevOps," DEV Community. Jan. 2024. Accessed: Feb. 24, 2025. [Online]. Available from: https://dev.to/infrasity-learning/artificial-intelligence-ai-in-devops-22eo

[44] "Generative AI in the Cloud: How DevOps is Changing & Microtica's POV." Accessed: Feb. 24, 2025. [Online]. Available from: https://microtica.com/blog/generative-ai-in-the-cloud

[45] "Implementing Scalable AI Solutions with Kubernetes and Docker." Accessed: Feb. 24, 2025. [Online]. Available from: https://www.rapidcanvas.ai/blogs/implementing-scalable-ai-solutions-with-kubernetes-and-docker

[46] "Generative AI Docker and Kubernetes Training Courses Ascendient," Ascendient Learning. Accessed: Feb. 24, 2025. [Online]. Available from: https://www.ascendientlearning.com/it-training/topics/agile-and-devops/docker-kubernetes/generative-ai

[47] B. Doerrfeld, "Using Generative AI to Accelerate Cloud-Native Development," Cloud Native Now. Jul. 2023. Accessed: Feb. 24, 2025. [Online]. Available from: https://cloudnativenow.com/features/using-generative-ai-to-accelerate-cloud-native-development/

[48] "AI in DevOps AI Talks for DevOps Overview," pulumi. Accessed: Feb. 24, 2025. [Online]. Available from: https://www.pulumi.com/blog/devops-ai-developer-future--pulumi-user-group-tech-talks/

[49] F. Hicks, "How do I use generative AI in Azure DevOps?" Jan. 2024. Accessed: Feb. 24, 2025. [Online]. Available from: https://www.aegissofttech.com/insights/how-ai-driven-insights-with-azure-devops/

[50] "AWS Prescriptive Guidance - Cloud design patterns, architectures, and implementations." Available rom: https://www.jeeviacademy.com/exploring-aws-well-architected-framework-building-cloud-optimized-solutions/

[51] "What is the AWS CDK? - AWS Cloud Development Kit (AWS CDK) v2." Accessed: Feb. 23, 2025. [Online]. Available from: https://docs.aws.amazon.com/cdk/v2/guide/home.html

[52] "Compare Cloud Service Providers." Accessed: Feb. 23, 2025. [Online]. Available from: https://www.oracle.com/cloud/service-comparison/

[53] "Create a generative AI–powered custom Google Chat application using Amazon Bedrock AWS Machine Learning Blog." Oct. 2024. Accessed: Feb. 23, 2025. [Online]. Available from: https://aws.amazon.com/blogs/machine-learning/create-a-generative-ai-powered-custom-google-chat-application-using-amazon-bedrock/

[54] "Well Architecture Framework Azure, AWS, GCP, OCI." Accessed: Feb. 23, 2025. [Online]. Available from: https://www.cloud4c.com/blogs/why-well-architected-frameworks-matter-in-cloud-adoption

[55] "Transforming DevOps with Generative AI K21Academy." Jul. 2024. Accessed: Feb. 24, 2025. [Online]. Available from: https://k21academy.com/ai-ml/gen-ai/genai-in-devops/

[56] "How Generative AI Support DevOps and SRE Workflows?" Accessed: Feb. 24, 2025. [Online]. Available from: https://www.xenonstack.com/blog/generative-ai-support-devops-and-sre-workflows

[57] "Kubernetes For AI Agents Restackio." Accessed: Feb. 08, 2025. [Online]. Available from: https://www.restack.io/p/agent-architecture-answer-kubernetes-ai-agents-cat-ai

[58] "From Kubernetes to Generative AI: The Future of Work LinkedIn." Accessed: Feb. 24, 2025. [Online]. Available from: https://www.linkedin.com/pulse/from-kubernetes-generative-ai-future-work-john-willis-0w81e/

[59] "Infrastructure for a RAG-capable generative AI application using Vertex AI and AlloyDB for PostgreSQL Cloud Architecture Center," Google Cloud. Accessed: Feb. 23, 2025. [Online]. Available from: https://cloud.google.com/architecture/rag-capable-gen-ai-app-using-vertex-ai

[60] "Deploy on Kubernetes Determined AI Documentation." Accessed: Feb. 08, 2025. [Online]. Available from: https://docs.determined.ai/setup-cluster/k8s/index.html

[61] "Generative AI on Cloud Platforms: GCP, AWS, and Azure," CloudThat Resources. Accessed: Feb. 23, 2025. [Online]. Available from: https://www.cloudthat.com/resources/blog/generative-ai-on-cloud-platforms-gcp-aws-and-azure/

[62] "Generative AI on AWS – Generative AI, LLMs, and Foundation Models – AWS," Amazon Web Services, Inc. Accessed: Feb. 23, 2025. [Online]. Available from: https://aws.amazon.com/ai/generative-ai/

[63] J. Gupta, "Generative AI Infrastructure Costs: A Practical Guide to GCP, Azure, AWS, and Beyond," Cloud Experts Hub. Jan. 2025. Accessed: Feb. 23, 2025. [Online]. Available from: https://medium.com/cloud-experts-hub/generative-ai-infrastructure-costs-a-practical-guide-to-gcp-azure-aws-and-beyond-fafb2808b1af

[64] "Best Practices for Scalable AI on Cloud Infrastructure," Yash Technologies. Accessed: Feb. 23, 2025. [Online]. Available from: https://www.yash.com/blog/building-scalable-ai-solutions-with-cloud-infrastructure/

[65] "Aws sagemaker vs google cloud ai platform: Which Tool is Better for Your Next Project?" Accessed: Feb. 23, 2025. [Online]. Available from: https://www.projectpro.io/compare/aws-sagemaker-vs-google-cloud-ai-platform

[66] "NVIDIA DGX Cloud," NVIDIA. Accessed: Feb. 23, 2025. [Online]. Available from: https://www.nvidia.com/en-us/data-center/dgx-cloud/

[67] "Red Hat OpenShift AI." Accessed: Feb. 23, 2025. [Online]. Available from: https://www.redhat.com/en/technologies/cloud-computing/openshift/openshift-ai

[68] "XenonStack- Generative AI Solutions on AWS." Accessed: Feb. 23, 2025. [Online]. Available from: https://www.xenonstack.com/autonomous-operations/amazon-web-services/

[69] "Generative AI Application Builder on AWS AWS Solutions AWS Solutions Library," Amazon Web Services, Inc. Accessed: Feb. 23, 2025. [Online]. Available from: https://aws.amazon.com/solutions/implementations/generative-ai-application-builder-on-aws/

[70] saxenashikha, "Architecting GenAI applications with Google Cloud," Google Cloud - Community. Sep. 2024. Accessed: Feb. 23, 2025. [Online]. Available from: https://medium.com/google-cloud/architecting-genai-applications-with-google-cloud-b38c9cbc66e0

[71] "AWS vs Azure vs GCP Comparison : Best Cloud Platform Guide," Veritis Group. Accessed: Feb. 23, 2025. [Online]. Available from: https://www.veritis.com/blog/aws-vs-azure-vs-gcp-the-cloud-platform-of-your-choice/

[72] J. MSV, "A Developer's Guide to Azure AI Agents," The New Stack. Feb. 2025. Accessed: Feb. 08, 2025. [Online]. Available from: https://thenewstack.io/a-developers-guide-to-azure-ai-agents/

[73] "Simplified Architecture to take up Generative AI in the Cloud Applications." Accessed: Feb. 23, 2025. [Online]. Available from: https://aitechcircle.kit.com/posts/simplified-architecture-to-take-up-generative-ai-in-the-cloud-applications

[74] "The Architecture of a Scalable and Resilient Google Cloud Solution," InfoQ. Accessed: Feb. 23, 2025. [Online]. Available from: https://www.infoq.com/news/2015/04/architecture-google-cloud/

[75] A. Verma, "Navigating the Cloud: A Comparative Analysis of GCP, AWS, and Azure," Medium. Feb. 2024. Accessed: Feb. 23, 2025. [Online]. Available from: https://ai.plainenglish.io/navigating-the-cloud-a-comparative-analysis-of-gcp-aws-and-azure-a3313f11f16a

[76] G. Kamtamneni, "How to develop AI Apps and Agents in Azure - A Visual Guide," All things Azure. Dec. 2024. Accessed: Feb. 08, 2025. [Online]. Available from: https://devblogs.microsoft.com/all-things-azure/how-to-develop-ai-apps-and-agents-in-azure-a-visual-guide/

[77] 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, Accessed: Feb. 23, 2025. [Online]. Available from: https://policyreview.info/articles/analysis/platform-power-ai-evolution-cloud-infrastructures

[78] 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, Available from: https://doi.org/10.1177/20539517241232630

[79] "What's the Difference Between AWS vs. Azure vs. Google Cloud?" Coursera. Oct. 2024. Accessed: Feb. 23, 2025. [Online]. Available from: https://www.coursera.org/articles/aws-vs-azure-vs-google-cloud

[80] "Comparing AWS, Azure, GCP DigitalOcean." Accessed: Feb. 23, 2025. [Online]. Available: https://www.digitalocean.com/resources/articles/comparing-aws-azure-gcp

[81] "Building the Future: A Deep Dive Into the Generative AI App Infrastructure Stack," Sapphire Ventures. Accessed: Feb. 23, 2025. [Online]. Available from: https://sapphireventures.com/blog/building-the-future-a-deep-dive-into-the-generative-ai-app-infrastructure-stack/

[82] "Top 9 AI Tools for DevOps Kubiya." Accessed: Feb. 24, 2025. [Online]. Available from: https://www.kubiya.ai/resource-post/ai-tools-for-devops

[83] "AWS and NVIDIA Announce Strategic Collaboration to Offer New Supercomputing Infrastructure, Software and Services for Generative AI," NVIDIA Newsroom. Accessed: Feb. 23, 2025. [Online]. Available from: http://nvidianews.nvidia.com/news/aws-nvidia-strategic-collaboration-for-generative-ai

[84] S. Zaman, "Generative AI Cloud Platforms: Choose from AWS, Azure, or Google Cloud," Folio3 Cloud Services. Aug. 2023. Accessed: Feb. 23, 2025. [Online]. Available: https://cloud.folio3.com/blog/generative-ai-cloud-platforms-aws-azure-or-google-cloud/

[85] J. Solanki, "How to Build a Scalable Application up to 1 Million Users on AWS," Simform - Product Engineering Company. Dec. 2018. Accessed: Feb. 23, 2025. [Online]. Available from: ohttps://www.simform.com/blog/building-scalable-application-aws-platform/

[86] A. Takyar, "Generative AI tech stack: Frameworks, infrastructure, models and applications," LeewayHertz - AI Development Company. Mar. 2023. Accessed: Feb. 23, 2025. [Online]. Available from: https://www.leewayhertz.com/generative-ai-tech-stack/

[87] "What is Cloud Elasticity vs Cloud Scalability? Teradata." Mar. 2022. Accessed: Feb. 23, 2025. [Online]. Available from: https://www.teradata.com/insights/cloud-data-analytics/cloud-elasticity-vs-cloud-scalability

[88] Richards, "RAG in the Cloud: Comparing AWS, Azure, and GCP for Deploying Retrieval Augmented Generation Solutions – News from generation RAG." Mar. 2024. Accessed: Feb. 23, 2025. [Online]. Available from: https://ragaboutit.com/rag-in-the-cloud-comparing-aws-azure-and-gcp-for-deploying-retrieval-augmented-generation-solutions/

[89] R. Innovation, "Asset Management with Generative AI Ultimate Guide." Accessed: May 06, 2025. [Online]. Available from: https://www.rapidinnovation.io/post/generative-ai-in-asset-management-application-benefits-best-practices-and-future

[90] Satyadhar Joshi, "A Literature Review of Gen AI Agents in Financial Applications: Models and Implementations," International Journal of Science and Research (IJSR),

Available from https://www.doi.org/10.21275/SR25125102816
[91] 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, Available from: https://doi.org/10.30574/wjaets.2025.14.2.0071
[92] Satyadhar Joshi, “Implementing Gen AI for Increasing Robustness of US Financial and Regulatory System,” IJIREM, vol. 11, no. 6, Art. no. 6, Jan. 2025, Available from: https://doi.org/10.55524/ijirem.2024.11.6.19.
[93] 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, Available from: https://doi.org/10.30574/wjarr.2025.25.1.0279
[94] 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, Available from: https://doi.org/10.30574/ijsra.2025.14.2.0439
[95] 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, Available from https://doi.org/10.55248/gengpi.6.0225.0756

ABOUT THE AUTHOR

Satyadhar Joshi did his International-MBA from Bar Ilan University Israel, and MS in IT from Touro College NYC and is currently working as AVP at BoFA USA. He is an independent researcher in the domain of AI, Gen AI and Analytics.