JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

Cloud-Native DevOps for GenAI-Powered Insurance Risk & Fraud Intelligence

Ethan Williams
Asst Professor

Abstract

Cloud-native DevOps for AI pipelines in insurance risk intelligence and fraud detection promotes replicability and composability across use cases. Support for scalable workloads addresses high costs, disk and memory limits, and specific cost and time-saving demands. Cloud-native principles—composability, elasticity, portability, observability, security, formal compliance, and data governance—are applied to underpin a generative AI-oriented insurance architecture. The DevOps component tailors continuous integration and continuous delivery (CI/CD) practices to be responsible, reproducible, reliable, and safe. Infrastructure depends on containerization and orchestration technologies to fully contemplate reproducibility, observability, isolation, and multi-cloud portability. AI workload provisioning adheres to a service model. For safe operation, specialized AI CI/CD pipelines provide model versioning and rollback, data versioning and validation, evaluation, performance metrics and safety gates for canary releases, and go/no-go decisions based on business risk appetite. Their design highlights domain-specific operational risk controls.

Different generative AI applications in insurance cover a wide spectrum of functionalities, underlining the fundamental basis for the creation of these models and the capability of these applications to adapt to different environments. Nevertheless, models must be handled with care in order to mitigate potential ethical, social, legal, regulatory, and business-related issues. Risk management and operational risk must be part of the implementation which, when properly completed, can provide a safe way to integrate these types of applications. Several risk intelligence use cases point to specific implementations demonstrated during the CapGemini FBDX programme and depict high-priority paths in the integration of such approaches, illustrating data origin and target indicators.

Keywords: Cloud-Native DevOps for AI, Insurance Risk Intelligence Systems, Fraud Detection Analytics, Generative Artificial Intelligence in Insurance, AI CI/CD Pipelines, MLOps for Regulated Industries, Containerization and Orchestration Technologies, Multi-Cloud Portability, Model Versioning and Rollback Controls, Data Versioning and Validation, Observability and Safety Gates, Canary Releases in AI Deployment, Compliance-by-Design Architectures, Operational Risk Controls in AI Systems, Scalable Insurance AI Workloads.

1. Introduction

In insurance, risk intelligence functions use analytics to inform decisions on accepting, renewing, and pricing insurances while fraud detection functions employ analytics to identify fraudulent activities. Generative AI is a defining capability of artificial intelligence that enables machines to generate new content and therefore revolutionises risk intelligence and fraud detection functions. For example, generative AI can be applied to confidential information such as operational risks during underwriting, fraud detection in public data or transient data such as claims. However, to be effectively harnessed, the use of generative AI across both risk intelligence and fraud detection requires specialised governance controls around the ethical and legal ramifications of using confidential data for AI, the assessment of the reliability and accuracy of AI outputs, the operational capability of implementing AI in production, and finally feeding AI-generated insights back into the risk and fraud pipelines, supported by cloud native DevOps practices for generative AI development and deployment.

Cloud-native DevOps provides scalable, automated, reproducible and governed processes and infrastructure for developing and operating software. In insurance risk intelligence and fraud detection function pipelines,

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

these capabilities allow the use of scaling AI to support operationally intensive functions over highly transient public or untrusted data that may include yet unexposed and therefore still credible insights. Four key areas are required to enable efficient and accurate AI-based risk-fraud insight generation: the infrastructure hosting the AI workloads must be designed to support peak workloads, the lifecycle of the AI workloads must incorporate governance controls, the productionisation of the AI workloads must include governing decisions for both the organisation and the capability of operating AI effectively, and risk-fraud pipelines must be updated to store any insights generated from AI workload execution. For all four areas, support for using scaling AI across risk intelligence and fraud detection is discussed.

1.1. Context and Overview

Insurance is one of the most heavily regulated industries in the world, serving important broad societal functions. Therefore, it is imperative that the processing of sensitive personal data, particularly for uses that involve AI-generative models, follows legal and ethical guidelines. The application of cloud computing, DevOps, and AI technologies collectively represents a significant opportunity for the insurance industry, tax management, nation-states, commercial banks, serving ethical and social purposes based upon community cooperations. Continuous adaptations for changing environments and parameters constitute best practices.

Fig 1: Synergistic Governance: A Cloud-Native DevOps Framework for Ethical Generative AI in Insurance Risk and Fraud Analytics

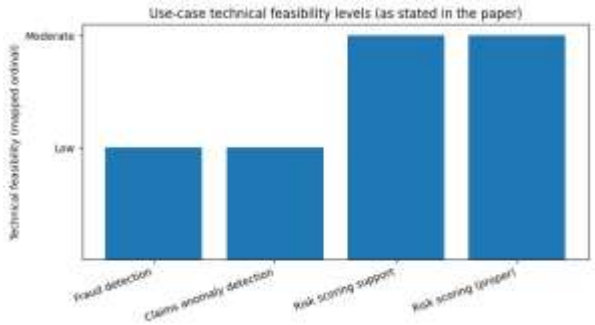
Organisations, such as (re)insurers, need a consolidated cloud-native approach for all generative AI use cases and applications specifically within insurance risk intelligence and fraud analytics as applied in the underwriting, claims and financial areas. DevOps is key to make these AI pipelines cloud-native and continuously adapt to data and user needs. Such a holistic approach focuses on both risk intelligence and fraud analytics, spanning the full range of AI-generative capabilities for insurance processes, associated data supply and constantly changing environments. The scaling of AI pipelines is an important issue, with current demand not only from (re)insurers but also from other financial services and related sectors.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

2. Foundations of Cloud-Native DevOps for Generative AI

Systematic use of cloud-native DevOps practices throughout the AI creation and operationalization lifecycle is essential for success. Cloud-native DevOps integrates rigorous software-development practices—often encapsulated in the agile, lean, and DevOps movement—with the architectural, operational, and service-use principles enshrined in cloud-native design. Cloud-native approaches also play an enabling role in the use of generative AI within risk intelligence and fraud detection pipelines in external annexed report economics, steering both the architecture and operational considerations for AI workload ranges. The discipline steers the systematic use of code-base creation and management for AI pipelines.

AI models have a much shorter useful operational lifetime than conventional business applications and must therefore be Internet and canary-release using real-world traffic to enable reinforcement and generative tuning. While strong governance of pipelines is critically important, a model-centric view must therefore extend beyond that of traditional DevOps. A containerized, container-orchestration-centric view is instead more supportive, supporting operational workload isolation for both generative and reinforcement pipelines by applying traditional segregation-of-duty across agent pipelines and operational AI traffic.



Equation 1) Resource provisioning for containerized AI workloads (Kubernetes-style)

Goal: compute total requested resources.

Let each workload $w \in \{1, \dots, W\}$ run with:

- r_w replicas (pods)
- CPU request per replica c_w (cores)
- Memory request per replica m_w (GiB)

Step 1: Total CPU requested

$$C_{\text{total}} = \sum_{w=1}^{W} r_w \; c_w$$

Step 2: Total Memory requested

$$M_{\text{total}} = \sum_{w=1}^{W} r_w \; m_w$$

Step 3: Capacity constraint (cluster)
If cluster allocatable CPU is $C_{\max}$ and memory is $M_{\max}$, then feasibility requires:

$$C_{\text{total}} \le C_{\max}, \qquad M_{\text{total}} \le M_{\max}$$

2.1. Architectural Principles

Modularization, elasticity, portability, observability, security, compliance, and data governance direct the architecture of cloud-native pipelines that scale the use of AI across product lifecycle phases supported by multiple cloud service providers. Distributing dev/test/staging/production pipelines across clouds with scalability considerations decouples AI workloads from other downstream control/production operations. Containers describe artifacts, their configurations, and the automations needed to build them. Multi-cloud container pipelines leverage patterns for safely hosting local and external pipelines within a shared central CI/CD service while Kubernetes organizes container

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

deployments across service dependencies and workloads in dedicated clusters, thereby speeding up expeditious resource selection. Full stack components can thus be efficiently containerized/contained, infrastructure costs closely monitored, and emerging technologies like data-centric AI reproducibly explored. Deploying within cloud provider restrictions, containers provide CPU/memory thresholds for controlling hosting costs, while embedding machine learning frameworks speeds up isolated 1+1 real-time interaction testing. Further cementing reproducibility, the provisioning model also allows easy switching to a black-box strategy in tightly controlled infrastructures.

Continuous integration and delivery (CI/CD) pipelines for AI empower safety gates that closely monitor predictive behaviour through metrics such as data quality and drift. A model versioning pipeline provides the tests and tooling combined with the governance structures needed to safely experiment with emerging technologies like data-centric AI. Canary release experimentations are simplified to direct comparisons of product quality attributes. Features like Data Build Tooling help to independently detect and correct bad data assumptions during production. Closely related ML observability tooling spans data sources with automated logging of input data, parameters, and model predictions for downstream explorations and incident responses.

	AI workload types (paper)	Infrastructure level	Description (paper)
0	Management model		
1	Data / non-model services		
2	Data pipelines & shared services (SLA non-critical)		
3	Model training		
4	Inference / prediction		

3. Generative AI in Insurance: Applications and Constraints

Generative AI capabilities align with core insurance functions, especially for filling analytic gaps, improving efficiency, and enabling new services. Constraints arise from ethical and legal concerns, the readiness of generative models for production use, and operational support, especially for safety-critical pipelines. Compliance with ethical and legal frameworks is essential in order to mitigate risks associated with incorrect model predictions. Composite or multi-step models composed of several pipelines often pose a higher-level risk that is more difficult to evaluate at individual steps.

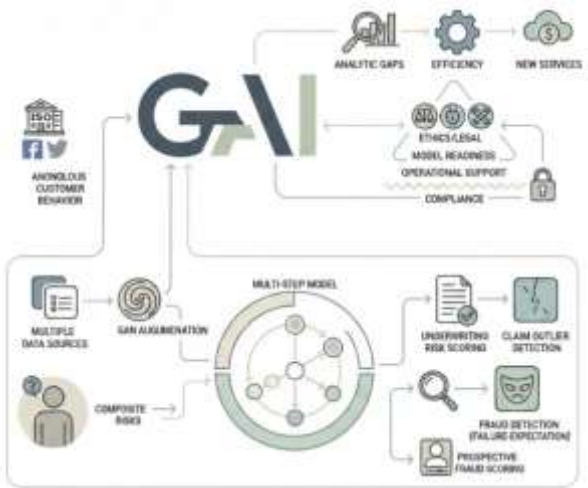


Fig 2: Integrated Risk Intelligence: Architecting Generative AI and GAN Augmentation for Safety-Critical Insurance Pipelines

Concrete risk-intelligence use cases provide guidance on applying generative AI in insurance. Generative models can fill gaps in commercial underwriting risk scoring, claim detection can leverage outlier detection, and fraud detection can rely upon a failure-expectation model. Fraudulent claims can be scored prospectively

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

based on the profile produced by a representation model. Multiple data sources have been identified for the main pipelines, and scoring performance improves relative to a baseline when using Generative Adversarial Network (GAN) augmentation. A concrete example of detecting anomalous customer behaviour in insurance claims complements the main risk-scoring pipelines, and adds Insurance Services Office (ISO) classification and social media-feeding networks as additional external sources.

Equation 2) Horizontal Pod Autoscaling (elasticity for peak workloads)

A common Kubernetes autoscaling rule uses observed metric vs target metric.

Let:

- current replicas r
- observed utilization/metric u (e.g., CPU utilization or requests per second per pod)
- target u^*

Step 1: proportional scaling

$$r_{\text{raw}} = r \cdot \frac{u}{u^*}$$

Step 2: replica count must be integer

$$r_{\text{new}} = \lceil r_{\text{raw}} \rceil$$

Step 3: enforce min/max replicas

$$r_{\text{final}} = \min\left(r_{\max}, \max(r_{\min}, r_{\text{new}})\right)$$

3.1. Risk Intelligence Use Cases

Concrete use cases exemplify the threat landscape and the potential contributions of risk intelligence pipelines. These include risk scoring support for underwriting, fraud detection, anomaly detection for claims, and risk scoring to identify strategic trends. Underlying AI-enabled risk analytics depend on pipelined data and scalable capacity for training and evaluation, with emphasis on rapid data delivery from diverse providers in a mix of on-premise and cloud locations, sufficient model versioning to ensure reliable service, and evaluation metrics for both functional and operational aspects of candidate models.

Prioritizing the investment of cloud resources requires transparent measurements of AI feasibility and future value. Technical feasibility is assessed by the effort and time required to implement a use case; these are considered low for the emerging fraud detection and claims anomaly detection use cases, and moderate for risk scoring support and risk scoring proper. Creating an evaluation framework offers a repeatable means to compare current and new models and quantitatively express support for a go-no-go decision.

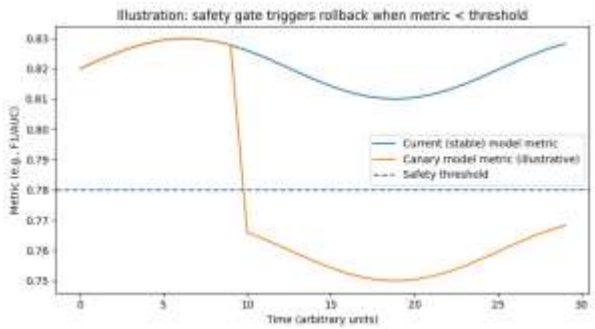
4. Cloud-Native Infrastructure for Scalable AI Workloads

Infrastructure models, service boundaries, and scalability considerations are discussed, supporting the need for cloud-native DevOps as enabler for scalable AI workloads for risk modeling, fraud detection, claim anomaly detection, and an integrated risk score. Scalable AI workloads for risk and fraud analytics demand robust infrastructure, including training and testing pipelines for support of continuous integration and delivery. Such requirements can be addressed by defining workload types, modelling resource provisioning, determining service boundaries, considering SLA, security, compliance, budget, and support for training/retraining pipelines. For risk intelligence and fraud detection, these aspects lead to supporting AI workload types—management model, data/non-model services, data pipelines and other shared services (SLA non-critical), model training, inference prediction, and non-AI services—and a four-level model: (1) production clustering environment supporting training/retraining pipelines of AI

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

workloads, (2) controlled training/retraining Kubernetes cluster managed by the specific AI workload with sensitive data stored unencrypted, (3) controlled training/retraining Kubernetes cluster managed by the specific AI workload, and (4) shared services (i.e. non-AI, data pipelines and other SLA non-critical services) on a PaaS or IaaS infrastructure.

Cloud-native infrastructures allow for scalability. Containerization and orchestration are used to provide reproducible and isolated services. Container images are produced via a pipeline integrating Git and an artifact repository, while deployment is managed by Kubernetes. The multi-cloud strategy ensures robustness with respect to IaaS disruptions (e.g. downtime, budget overrun). Kubernetes reusable patterns (Terraform code) implement HA, backup, workload-security for KMS and GCS, audit, and DLP, while pipeline controllability includes cluster-prod vs. cluster-dev separation strategy, security concerning Prod workload pipelines by Kubernetes Security Context together with SOC-team controls incenting accurate DLP data classification, and Data Protection by Design by pipeline deployers.



Equation 3) Cloud cost model under CPU/memory thresholds

Assume unit prices:

- p_C: cost per CPU-core-hour
- p_M: cost per GiB-hour
- runtime duration T hours

Using totals from Section (1):

Step 1: CPU cost

$$\text{Cost}_{CPU} = C_{\text{total}} \cdot p_C \cdot T$$

Step 2: Memory cost

$$\text{Cost}_{MEM} = M_{\text{total}} \cdot p_M \cdot T$$

Step 3: Total

$$\text{Cost}_{\text{total}} = \text{Cost}_{CPU} + \text{Cost}_{MEM}$$

4.1. Containerization and Orchestration

Container images encapsulate all dependencies required to execute code within a directory, enabling build processes to remain isolated from host environments and their respective libraries. These images are immutable, facilitating reproducibility and auditability. Container images are frequently constructed for every check-in, generally containing the model binary and any supplementary files required for dependencies, evaluators, or asset preparation.

Container orchestration abstracts resource deployment (compute, network, storage) required to execute applications. Kubernetes (K8s) and Amazon Elastic Kubernetes Service (EKS) provide a cloud-agnostic method of deploying AI workloads via a common orchestration standard, even across multi-cloud configurations. K8s employs resources such as secrets for sensitive information, volumes for shared file storage, and jobs for batch processing. K8s Helm charts define the environment and data dependencies, while associated templates lay out performance-related configurations for the model. Deployment patterns for K8s pipelines use either the operator library that integrates AI frameworks (currently supporting TensorFlow and PyTorch) or the KServe library that simplifies inference-serving workloads.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

Before online deployment, continuous integration (CI) tools set up a K8s-in-kind local cluster for compilation in a Docker file, executing unit tests and loading the model repository with the latest image. Once staged for external deployment, a separate test K8s cluster installed into the target environment serves as the last test location before public consumption, accompanied by performance and A/B tests.

Time	Baseline_metric	Canary_metric	Rollback? (Canary<threshold)
0	0.82	0.82	False
1	0.822	0.822	False
2	0.825	0.825	False
3	0.827	0.827	False
4	0.828	0.828	False
5	0.829	0.829	False

5. DevOps Practices for Generative AI Pipelines

Addressing the ethical, social, and governance challenges of deploying generative Ai requires considering these concerns holistically within a coherent governance framework. The core concerns are data, cybersecurity, intellectual property, reputation, compliance, insurance, and reliability. Given that these concerns touch on multiple aspects of an organization, establishing an appropriate governance framework for these concerns is essential. However, such a framework alone is not sufficient, as organizations also need to embed appropriate AI-specific considerations throughout the DevOps process to address the various concerns during Ai service generation.

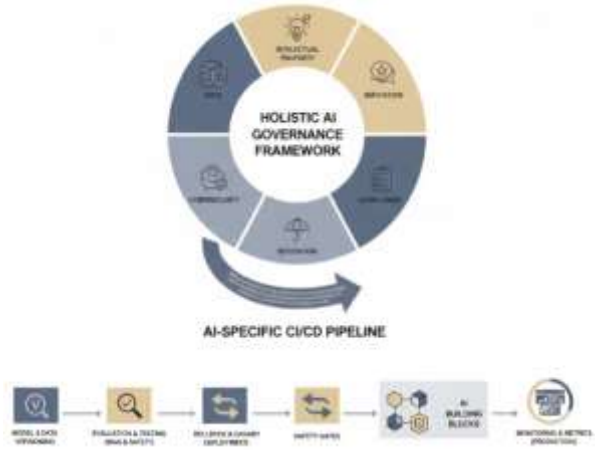


Fig 3: Embedding Ethics in the Pipeline: A Holistic Governance and AI-Specific CI/CD Framework for Generative AI Reliability

Although Ai is often portrayed as different from traditional software engineering, the methodologies that have been developed and deployed in these areas over the past few decades have direct relevance to the Ai product lifecycle. Following best practices from model evaluation and testing can help minimize bias, safety, and security issues. However, organizations also need to consider putting in place appropriate Ai-specific continuous integration and continuous delivery (CI/CD) practices to address these issues when deploying and operating generative AI services. This CI/CD framework addresses the areas of model versioning, data versioning, evaluation, rollback, canary deployment, safety gates, AI building blocks, and applicable metrics for monitoring the service once in production.

5.1. Continuous Integration and Delivery for AI

Infrastructure Pipelines for Scalable Workloads

Tuning AI systems requires reliable assessments of new model versions and associated data sets. For easy evaluation, AI solutions can be instrumented with

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

version gates that monitor performance of newly deployed model and/or data versions. When performance drops below a safety threshold, the previously active model can be automatically restored. Such safety gates also enable controlled canary releases, where a small fraction of the workload uses the newly deployed model version. The trigger for rolling back a new version can vary beyond simple performance losses, as it ultimately depends on the real-time risk tolerance of the regulated business owners.

The go-live decision for AI models and data sets can further benefit from multi-stakeholder control gates that formalise its approval or rejection. Incorporating expert knowledge specific to a given AI task into the continuous delivery ensures that AI solutions do not solely rely on performance data. These control gates also inspire the go-live decision documented within the formal risk framework of the organisation, thus addressing the AI risk triangle. To complement safety gates, these multi-stakeholder control gates should also be integrated to enable automatic rollback and controlled canary releases.

Equation 4) Canary release evaluation + safety gate rollback rule

Let:

- S_{old}: score/metric of current model (AUC/F1/etc.)
- S_{new}: score of candidate model on canary traffic
- τ: minimum acceptable score ("safety threshold")
- optionally, Δ: maximum tolerated drop vs current

Step 1: define acceptance conditions
Common acceptance logic:

- absolute threshold: $S_{\text{new}} \geq \tau$
- relative degradation guard: $S_{\text{new}} \geq S_{\text{old}} - \Delta$

Step 2: combine into one "go/no-go" gate

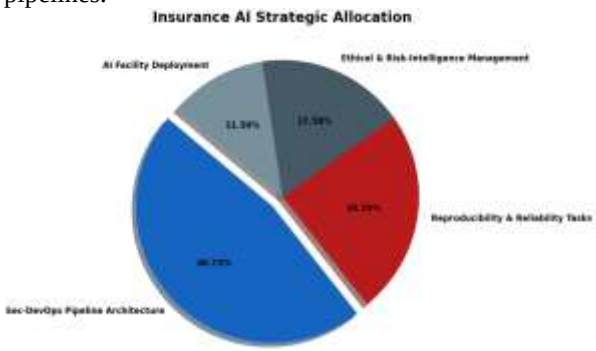
$$\text{GO} \Leftrightarrow (S_{\text{new}} \geq \tau) \wedge (S_{\text{new}} \geq S_{\text{old}} - \Delta)$$

Step 3: rollback condition

$$\text{ROLLBACK} \Leftrightarrow \neg\text{GO}$$

6. Conclusion

Deploying and operationalizing cloud-native DevOps for scalable Generative AI applied to risk intelligence and fraud-detection pipelines in insurance is essential, but neither straightforward nor trivial. Insurance pipelines populated by AI-generated content can introduce hitherto unseen reputational, ethical, safety, and legal risks; a cloud-native architecture together with Sec-DevOps CAN mitigate such risk. The approach neither centres on the generative models nor their fine-tuning but rather remains focused on the broader architectural and operational aspects of the pipelines themselves; hence, the Sec-DevOps practices cover special provisions for these pipelines, not the underlying models. The successful implementation of cloud-native DevOps for the considered pipelines relies on commonly accepted Sec-DevOps best practices and extends such practices with added emphasis on reproducibility and reliability—tasks that typically require extra time investments for machine-learning pipelines.



JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

Fig 4: Insurance AI Strategic Allocation

Despite the tangible benefits of cloud-native Sec-DevOps, several aspects remain challenging. First, large-scale cloud-native AI capabilities have security, compliance, ethical, and risk-intelligence implications that must be managed; the associated Sec-DevOps practices remain a work in progress. Second, special facilities must be deployed when working with such AI capabilities; deploying these amenities within the organization introduces further operational rigor that Cummins has started to develop but is not yet complete. Third, Cummins performed the preceding analysis and planning for the AI-intense pipelines; implementing these plans and realizing the anticipated gains, which should be substantial, is indeed a daunting prospect. Finally, reverberations from the use of generative models can disrupt institutions, markets, and society. Although the potential benefits of using such AI capabilities are vast, careful consideration of the myriad surrounding consequences is therefore essential.

6.1. Summary and Future Directions

This work establishes and exemplifies the foundations of cloud-native DevOps for scalable generative AI in risk-intelligence and fraud-detection pipelines, showing how organization, infrastructure, and practices can be curated to support elastic AI workloads in domains with ethical, legal, and operational constraints. Key takeaways include how DevOps governs the entire AI lifecycle (from development to production), the motives for a cloud-native platform in a regulated industry, and the interests behind applying generative AI in insurance. The analysis positions AI capabilities in direct relation to traditional insurance functions and maps risks into the system being developed, making them an integral part of the design.

The analysis underscores how cloud-native principles such as modularity support infrastructure-as-code approaches for reproducibility and control, and how containerization and orchestration at multiple levels enable the careful control of A-model deployment while mitigating the operational burden associated with multiple algorithms. These aspects are particularly critical in a regulated industry. However, they have broad applicability in highly ethical contexts, where ambiguity is undesirable. Nonetheless, many aspects of the AI pipelines are also be common in less-regulated industries, and DevOps concepts presented here can be integrated into such pipelines even when the range of risks to AI-supported decision-making is more limited. The work concludes with a discussion of emerging practice-orientated questions: Who does the security validation of A-models? What new roles are required to ensure that A-models can be safely used in a production role? Who is responsible for continuous operation and visible governance of A-model use?

7. References

1. Gomes, C., Jin, Z., & Yang, H. (2021). Insurance fraud detection with unsupervised deep learning. Journal of Risk and Insurance, 88(3), 591–624.
2. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2021). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 34.
3. Balamurugan, J., Bhuvaneswari, M., Aitha, A. R., Nagaraju, S., Viswanathan, R., & Dhasarathan, N. (2025, November). Explanatory Transformer-Based Sequential Recommendation: Assessing BERTRec with SASRec for Customer Behaviour Prediction. In 2025 International Conference on Emerging Engineering Technologies and Applications (IC-EETA) (pp. 470-475). IEEE.
4. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

5. Aslam, F., Hunjra, A. I., Ftiti, Z., Louhichi, W., & Shams, T. (2022). Insurance fraud detection: Evidence from artificial intelligence and machine learning. Research in International Business and Finance, 62, 101744.
6. Amistapuram, K., Pandiri, L., Raju, V. R., Paleti, S., Singireddy, S., & Sheelam, G. K. (2025). AI-Based Cloud Infrastructure and MLOps Frameworks for Scalable Data Engineering Across Banking and Insurance. In 2025 IEEE International Conference on Communication Networks and Computing (CNC) (pp. 186–192). IEEE. 2025 IEEE International Conference on Communication Networks and Computing (CNC). https://doi.org/10.1109/cnc68716.2025.11484532
7. Xia, H., Zhou, Y., & Zhang, Z. (2022). Auto insurance fraud identification based on a CNN-LSTM fusion deep learning model. International Journal of Ad Hoc and Ubiquitous Computing, 39(1–2), 37–45.
8. Hewage, N., & Meedeniya, D. (2022). Machine learning operations: A survey on MLOps tool support. arXiv preprint arXiv:2202.10169.
9. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
10. Inala, R. (2025). A Unified Framework for Agentic AI and Data Products: Enhancing Cloud, Big Data, and Machine Learning in Supply Chain, Insurance, Retail, and Manufacturing. EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR, 46(1), 1614-1628.
11. Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36.
12. Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22.
13. Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. IEEE Access, 11, 31866–31879.
14. Goel, A. V., Kumari, P., Vaghela, K., Nagabhyru, K. C., Karichalil, R. A., Salman, S. A., & Brahmane, P. (2025). STRATEGIC CHANGE MANAGEMENT IN THE ERA OF DIGITAL DISRUPTION: AN INTERDISCIPLINARY STUDY ON ORGANISATIONAL ADAPTABILITY AND INNOVATION CULTURE. Scientific Culture, 11(4), 236.
15. Ruf, P., Madan, M., Reich, C., & Olex, A. (2023). Demystifying MLOps: A systematic literature review. IEEE Access, 11, 110111–110132.
16. Settipalli, L., & Gangadharan, G. R. (2023). WMTDBC: An unsupervised multivariate analysis model for fraud detection in health insurance claims. Expert Systems with Applications, 215, 119259.
17. Debener, J., Heinke, V., & Kriebel, J. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. Journal of Risk and Insurance, 90(3), 743–768.
18. Davuluri, P. S. L. (2023). AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems. AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems (December 15, 2023).
19. Maiano, L., Montuschi, A., Caserio, M., Ferri, E., Kieffer, F., Germanò, C., Baiocco, L., Ricciardi Celsi, L., Amerini, I., & Anagnostopoulos, A. (2023). A deep-learning-based antifraud system for car-insurance claims. Expert Systems with Applications, 231, 120644.
20. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations.

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

21. Isukapatla, M., Davuluri, P. N., Satya, G. S., & Prakash, P. R. (2026, April). An Attention-Enhanced YOLO Framework with IMU Confidence for Road Surface Defect Analysis. In 2026 2nd International Conference on Intelligent Systems and Computational Networks (ICISCN) (pp. 1-8). IEEE.
22. Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., & Liu, Y. (2023). Prompt injection attack against LLM-integrated applications. arXiv preprint arXiv:2306.05499.
23. Schreiner, M., & colleagues. (2023). The pipeline for the continuous development of artificial intelligence models—Current state of research and practice. Journal of Systems and Software, 199, 111615.
24. Yandamuri, U. S., Loganathan, R., Davuluri, P. N., Rani, P. S., Kolla, S. H., & Nagubandi, A. R. (2026, June). Adaptive Intelligence Networks for Humancentered Enterprise Automation and Governance. In 2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS) (pp. 678-683). IEEE.
25. Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C., Bensalem, S., Mu, R., Qi, Y., Zhao, X., Cai, K., Zhang, Y., Wu, S., Xu, P., Wu, D., Freitas, A., & Mustafa, M. A. (2024). A survey of safety and trustworthiness of large language models through the lens of verification and validation. Artificial Intelligence Review, 57, Article 175.
26. Wang, Y., Wang, M., Manzoor, M. A., Liu, F., Georgiev, G. N., Das, R. J., & Nakov, P. (2024). Factuality of large language models: A survey. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 19519–19529.
27. Priyanka, R. P., Annapareddy, V. N., Yenugu, B. C., Nagabhyru, K. C., & Kapila, D. (2025, September). Optimization of Battery Management Systems Using Machine Learning. In 2025 International Conference on Computing and Communications (COMPUTINGCON) (pp. 1-6). IEEE.
28. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), Article 39.
29. Kumar, P. (2024). Large language models (LLMs): Survey, technical frameworks, and future challenges. Artificial Intelligence Review, 57, Article 260.
30. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 8048–8057.
31. Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. Journal of Computer Science and Technology, 39, 1–25.
32. Lebcir, I., Mageswari, S. D., Bhosale, Y. H., Nagubandi, A. R., & Mahabooba, M. (2025). Agile Strategic Management in the Age of Disruption: Leveraging AI and Data Analytics for Competitive Advantage. Advances in Consumer Research, 2(6), 2581.
33. Vorobyev, I. (2024). Fraud risk assessment in car insurance using claims graph features in machine learning. Expert Systems with Applications, 251, 124109.
34. Schrijver, G., Sarmah, D. K., & El-hajj, M. (2024). Automobile insurance fraud detection using data mining: A systematic literature review. Intelligent Systems with Applications, 21, 200340.
35. Khalil, A. A., Liu, Z., Fathalla, A., Ali, A., & Salah, A. (2024). Machine learning based method for insurance fraud detection on class imbalance datasets with missing values. IEEE Access, 12, 155451–155468.
36. Ganesh, S. K., Subbareddy, K., Gopi, A., Davuluri, P. N., Mannar, B. R., & Karnawat, A. T. (2026, June). Fraudulent Credit Card Transaction Detection Using a Hybrid Ensemble-Anomaly

JOURNAL OF INNOVATIVE ACADEMIC RESEARCH
ISSN : 3049-2122
VOLUME: 04 ISSUE: 02
RECEIVED: APRIL 17 REVISED: MAY 08 ACCEPTED: MAY 20 PUBLISHED: JUNE 15

Detection Framework. In 2026 6th International Conference on Intelligent Technologies (CONIT) (pp. 1-6). IEEE.

37. Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. Artificial Intelligence Review, 57, Article 243.
38. Kumar, P., Ekin, T., Park, C., Markey, M. K., Barner, J. C., & Rascati, K. (2024). Using a Bayesian belief network to detect healthcare fraud. Expert Systems with Applications, 238, 122241.
39. Kumar, P. A., & Sountharrajan, S. (2025). Insurance claims estimation and fraud detection with optimized deep learning techniques. Scientific Reports, 15, Article 27296.
40. Garapati, R. S. (2025). Real-Time Monitoring and AI-Based Control of Industrial Robots Using Cloud-Hosted Web Applications. Available at SSRN 5612491.
41. Yankol-Schalck, M. (2026). Auto insurance fraud detection: Machine learning and deep learning applications. Journal of Risk and Insurance, 93(2), 534–568.