СИСТЕМА ВИЯВЛЕННЯ НЕБЕЗПЕЧНИХ КОНФІГУРАЦІЙ У KUBERNETES НА БАЗІ AI АГЕНТІВ: ПРАКТИЧНЕ ЗАСТОСУАННЯ

Мазепа А.Д., Фісенко Д. М., Пантєлєєв В.О.

Кафедра інфокомунікаційної інженерії ім. В.В. Поповського,
Харківський національний університет радіоелектроніки,
Україна

E-mail: artem.mazepa@nure.ua
dmytro.fisenko@nure.ua
vadym.pantielieiev@nure.ua

Abstract

This paper introduces an AI-agent-based system for detecting and mitigating dangerous configurations in Kubernetes clusters, such as excessive privileges, unsecured secrets, and vulnerable network policies. Leveraging a locally deployed LLM model (e.g., Llama-3.1-8B) calibrated for security expertise, the system enables real-time scanning of YAML manifests and runtime states via multi-agent architecture integrated with tools like Falco and Kyverno. Comparative analysis shows superior detection accuracy (96%) over alternatives like Kubescape, though with trade-offs in processing speed and resource usage. Future enhancements include resource optimization and expanded agents for broader cluster monitoring, transforming Kubernetes security into a proactive, intelligent framework for enterprise environments.

У наш час Kubernetes є золотим стандартом для проектування IT-інфраструктури рівня Enterprise. Включаючи в себе величезну кількість інструментів, він дозволяє точково налаштувати архітектуру системи для конкретного випадку, що дає змогу проектам будь-якої складності рости без особливих проблем.

Однак така гнучкість є, одночасно, й головним мінусом Kubernetes, оскільки для ефективного управління ним потрібна досить досвідчена команда, щоб уникнути неправильного налаштування параметрів безпеки та, як наслідок, компрометації всієї системи. Та навіть з великою, та досвідченою командою доволі складно помітити незначні помилки конфігурації по причині великого розміру самої системи.

У цій роботі буде розглянуто агентну систему ШІ, яка дозволить вирішити ці проблеми і полегшити оперування системою досвідченим адміністраторам.

Використання ШІ у кібербезпеці

За останні роки ШІ в кібербезпеці тільки набирає обертів і не думає зупинятися найближчим часом через почастішання випадків, коли зловмисники самі починають використовувати ШІ з метою атаки.

Наприклад, у 2023 році кількість зареєстрованих кібератак з використанням ШІ зросла на 47% у глобальному масштабі, а 87% організацій повідомили про пережиті атаки на базі ШІ за останній рік. Крім того, 82,6% фішингових електронних листів зараз використовують ШІ в тій чи іншій формі, глобальні атаки з використанням ШІ перевищать 28 мільйонів інцидентів у 2025 році, а 50% респондентів з організацій критичної інфраструктури вже зіткнулися з атаками на базі ШІ за останній рік. При цьому такі атаки досить складно відбити звичайними засобами, які вже використовуються у індустрії роками.

Отже, автоматичні системи виявлення вразливостей на базі AI є більш ніж виправданими і являють собою наступний ступінь у розвитку кібербезпеки.

Опис системи

Система використовує ШІ-агент, що перевіряє конфігурації кластера Kubernetes на наявність небезпечних або неправильних налаштувань безпеки, таких як надмірні привілеї, незахищені секрети чи вразливі мережеві політики.

Модель, яка використовується у агенті, є варіантом відкритої LLM, Llama-3.1-8B, адаптованою для ефективного локального запуску з мінімальними ресурсами (наприклад, на стандартних серверах без потужних GPU). Це дозволяє агенту автономно аналізувати YAML-маніфести, runtime-логи. та приймати рішення щодо посилення політик (наприклад, RBAC, NetworkPolicies) на основі принципів найменших привілеїв.

Вибір локальної моделі є обов'язковим для запобігання витоку критичної інформації, як може статися під час використання комерційних моделей, таких як ChatGPT або Grok. Крім того, комерційні моделі є витратними в експлуатації та не містять спеціалізованих знань для виконання унікальних завдань, характерних для окремих кластерів. Для розв'язання подібних завдань локальна модель може бути перекалібрована та розгорнута в кластері з інтеграцією необхідної інформації.

Оскільки дана модель виступає центром прийняття рішень, необхідно забезпечити належний захист цього компонента для уникнення компрометації всієї системи. З цією метою модель проходить обов'язкову перевірку на вразливість до типових атак, таких як prompt-injection, перед кожним розгортанням нової версії агента в системі.

Агент може функціонувати в двох режимах. Проактивний режим передбачає безперервний моніторинг усіх конфігураційних файлів з метою виявлення вразливостей та автоматичного їх усунення відповідно до норм безпеки, засвоєних під час розгортання. Спостережний режим, навпаки, фіксує дефектні конфігураційні файли та, замість автоматичного виправлення, надсилає сповіщення адміністратору на попередньо визначений електронний поштовий адрес або інший канал зв'язку. Таким чином, адміністратор кластера отримує можливість оперативно реагувати на виявлені загрози.

Ця система розгортається безпосередньо в кластері за допомогою Helm-чартів, що забезпечує швидкий і стандартизований початок роботи в будь-якому середовищі Kubernetes, мінімізуючи залежність від зовнішніх інструментів і полегшуючи інтеграцію з існуючими пайплайнами CI/CD.

Заміри ефективності

Побудована система була порівняна за іншими подібними системами та альтернативами.

Табл. 1. Порівняння системи з альернативами

Метрика	Запропонована система (AI-агент на базі Llama-3.1-8B)	Kubescape (статичний сканер конфігурацій)	Falco (runtime-моніторинг конфігурацій)	Kyverno (політики конфігурацій)
Точність виявлення (Accuracy)	96%	85%	88%	87%
Швидкість обробки (час на сканування конфігурацій кластера)	20-26 хвилин	10-15 хвилин	9-13 хвилин	8-11 хвилин
Навантаження на CPU під час сканування конфігурацій (%)	20-26%	6-8%	9-12%	7-9%

Як можна побачити з порівняльних результатів, точність виявлення вразливостей у запропонованій системі становить 96%, що робить її найточнішою серед розглянутих альтернатив. Цей показник перевищує відповідні метрики традиційних інструментів, таких як Kubescape (85%) та Falco (88%), забезпечуючи вищий рівень ідентифікації небезпечних конфігурацій, включаючи надмірні привілеї, незахищені секрети та вразливі мережеві політики.

Однак для досягнення такої високої точності було зроблено компроміс щодо швидкості обробки та споживання ресурсів під час повного сканування кластера: середній час аналізу збільшується на 25-30% порівняно з статичними сканерами, а пікове навантаження на CPU може сягати 26% у режимі інтенсивного обчислення.

Перспективи

У перспективі подальшого розвитку системи передбачено оптимізацію моделі для зменшення споживання обчислювальних ресурсів, а також реалізацію додаткових агентів для моніторингу інших аспектів кластера Kubernetes. Зокрема розширення мультиагентної архітектури шляхом впровадження спеціалізованих агентів для моніторингу інших критичних аспектів, таких як оптимізація ресурсів (наприклад, автоматичне масштабування подів на базі reinforcement learning), аналіз логів на аномалії (з використанням graph neural networks для виявлення патернів атак) та управління витратами (інтеграція з інструментами на кшталт Kubecost для AI-допомоги в оптимізації бюджетів).

Висновки

Запропонована система продемонструвала високі результати ефективності у виявленні та усуненні небезпечних конфігурацій у кластері Kubernetes завдяки налаштованому AI-агенту. Система мінімізує ручні втручання та відкриває перспективи для оптимізації ресурсів і розширення на інші аспекти моніторингу кластера. Загалом, інтеграція AI трансформує безпеку Kubernetes у більш інтелектуальну та стійку до еволюціонуючих загроз.

Література

1. Djenna, A., Bouridane, A., Rubab, S. and Marou, I.M., 2023. Artificial intelligence-based malware detection, analysis, and mitigation. Symmetry, 15(3), p.677.

2. Guembe, B., Azeta, A., Misra, S., Osamor, V.C., Fernandez-Sanz, L. and Pospelova, V., 2022. The emerging threat of ai-driven cyber attacks: A review. Applied Artificial Intelligence, 36(1), p.2037254.

3. J. Zhang et al., “When LLMs meet cybersecurity: a systematic literature review”, vol. 8, no. 1, Feb. 2025, doi: https://doi.org/10.1186/s42400-025-00361-w.

4. H. Xu et al., “Large Language Models for Cyber Security: A Systematic Literature Review,” arXiv.org, May 09, 2024. https://arxiv.org/abs/2405.04760.

5. R. K. Jones and A. J. Jones, “Integration of LLMs With Traditional Security Tools,” Advances in Information Security, Privacy, and Ethics, pp. 295–334, Nov. 2024, doi: https://doi.org/10.4018/979-8-3693-9311-6.ch008.