EPIC EDGE | SCHEDA PRODOTTO

CloudSentinel AI IN SVILUPPO — 2027

Predictive analytics, anomaly detection e ottimizzazione energetica per infrastrutture cloud

Modulo dell'ecosistema SentinelForge AI · on-premise · sovrano · open-source · AI-native

Il problema

Le infrastrutture cloud generano un flusso continuo di metriche, log ed eventi. Il problema non è la mancanza di dati: è che troppi dati sono equivalenti a nessun dato. Gli strumenti tradizionali di monitoraggio segnalano tutto con la stessa urgenza --- e chi gestisce l'infrastruttura impara a ignorare le notifiche perché sono troppe, sono ridondanti e raramente indicano la causa reale.

Un guasto imminente di un nodo di compute raramente si annuncia con un singolo alert: si manifesta come un pattern distribuito su metriche diverse nel tempo. Chi legge solo alert punto-a-punto non vede il pattern. Chi gestisce cluster Ceph, OpenStack e Kubernetes in parallelo ha ancora meno possibilità di correlare segnali che arrivano da sistemi con dashboard separate.

Il costo di un'interruzione non pianificata non è solo il downtime: è il tempo sprecato a cercare la causa in mezzo a centinaia di notifiche non correlate, mentre il problema si aggrava.

La soluzione

CloudSentinel AI è il modulo dell'ecosistema SentinelForge AI dedicato all'osservabilità predittiva e all'ottimizzazione dell'infrastruttura cloud. Non è un sistema di monitoraggio in più: è un livello di intelligenza che si posiziona sopra gli strumenti esistenti --- Prometheus, Grafana, Zabbix, ELK --- raccogliendo i segnali, correlando gli eventi e identificando i pattern che una soglia singola non può vedere.

CloudSentinel AI osserva l'infrastruttura nel suo insieme: compute, storage, networking, container. Quando individua un'anomalia, non si limita a segnalarla --- la classifica per impatto probabile, la correla con eventi concomitanti e propone un piano d'azione leggibile, graduato per rischio. L'operatore vede già la causa probabile e l'azione suggerita, non solo l'alert grezzo.

La componente di ottimizzazione energetica lavora in modo continuo: analizza i pattern di carico nel tempo, identifica i nodi sottoutilizzati, propone la consolidazione dei workload e la riduzione attiva del consumo --- con la stessa logica di autonomia graduata che caratterizza l'ecosistema SentinelForge AI.

Il modello propone. Il motore di automazione esegue.

Come funziona

1 · RACCOGLIE	2 · CORRELA
Pipeline di telemetria unificata: metriche	Motore a due livelli: regole deterministiche per

Prometheus, log strutturati, eventi da OpenStack, Ceph e Kubernetes. Ciclo continuo con trigger immediato sugli eventi critici.	i casi noti, LLM on-premise dove serve contesto. Due alert distinti sullo stesso rack vengono riconosciuti come un unico incidente.
3 · PREVEDE Analisi delle serie storiche per anticipare esaurimento risorse, degrado hardware, pattern di sovraccarico. La proiezione include finestra temporale e margine di confidenza.	**4 · PROPONE E AGISCE** Piano d'azione leggibile, classificato per rischio. Esecuzione tramite AWX + Ansible: il modello non tocca mai l'infrastruttura direttamente. Ogni azione è tracciata su Git.

Ottimizzazione energetica

CloudSentinel AI include un modulo dedicato all'ottimizzazione dei consumi energetici dell'infrastruttura. L'analisi è continua: il sistema osserva i pattern di carico su un orizzonte temporale configurabile, identifica i nodi di compute sottoutilizzati in modo ricorrente e propone la consolidazione attiva dei workload.

Analisi del carico	Consolidamento workload	Riattivazione automatica
Identifica i nodi con utilizzo medio basso su finestre temporali configurabili. Distingue il sottoutilizzo strutturale da quello episodico.	Propone la migrazione live dei workload verso nodi attivi, con stima dell'impatto prima dell'esecuzione. Il nodo svuotato viene spento fisicamente.	Quando il carico risale, il sistema riaccende i nodi, verifica la consistenza del software, li aggiorna se necessario e li reintroduce nel cluster.

Autonomia graduata

Come ogni modulo dell'ecosistema SentinelForge AI, CloudSentinel AI opera su tre livelli di autonomia configurabili per ambiente e per tipo di operazione.

● AUTONOMO — Raccolta, analisi, segnalazione

Health check continui, raccolta metriche, anomaly detection, proiezioni di capacità, classificazione degli alert per impatto, produzione del piano d'azione. Nessun impatto sull'infrastruttura.

● SUPERVISIONATO — Ottimizzazione e remediation, con conferma

Consolidamento workload, spegnimento e riaccensione nodi, throttling delle risorse, applicazione di configurazioni ottimizzate, upgrade rolling guidati. Ogni azione viene presentata con stima d'impatto e richiede conferma esplicita.

● SOLO CONSULENZA — Il sistema spiega, l'operatore decide

Modifiche architetturali, variazioni di configurazione ad alto impatto, interventi su componenti in stato degradato. Il sistema fornisce la simulazione dell'impatto e i passi guidati; il comando finale resta sempre in mano all'operatore.

Caratteristiche chiave

- Osservabilità unificata — metriche, log ed eventi di OpenStack, Ceph e Kubernetes in un'unica vista correlata: nessuna dashboard separata per ogni stack.

- Correlazione multi-segnale — il pezzo dove l'AI aggiunge valore reale rispetto all'alerting tradizionale: due warning distinti che sono in realtà un unico incidente con una causa radice.
- Anomaly detection AI-driven — il sistema impara i pattern normali dell'infrastruttura e segnala le deviazioni significative, non le singole soglie superate.
- Proiezioni di capacità — analisi del trend storico per anticipare esaurimento risorse con finestra temporale e margine di confidenza.
- Root Cause Analysis automatizzata — per ogni anomalia rilevata, il sistema produce una spiegazione in linguaggio naturale calata su quell'ambiente specifico.
- Ottimizzazione energetica attiva — consolidamento workload, spegnimento nodi sottoutilizzati e riattivazione automatica con verifica di consistenza.
- LLM interamente on-premise — nessun dato di infrastruttura, nessuna metrica, nessuna credenziale lascia il perimetro.
- Il modello non esegue mai comandi — genera un piano strutturato; l'esecuzione passa da AWX + Ansible. Audit, dry-run e revoca prima dell'esecuzione.
- Raccolta dati già in corso — pipeline Prometheus + Zabbix + ELK attiva su cluster interno per il training del modello.
- Tracciabilità completa — ogni playbook eseguito è versionato su Git e legato esplicitamente all'evento che lo ha generato.

Benefici per il cliente

Meno alert, più segnale La correlazione multi-segnale riduce il rumore: l'operatore vede incidenti, non notifiche. Meno tempo a triage, più tempo sulla causa.	**Guasti anticipati** Le proiezioni di capacità e l'anomaly detection identificano i problemi prima che diventino interruzioni. Intervento pianificato anziché reazione d'emergenza.
Riduzione consumi energetici Il consolidamento attivo dei workload riduce il numero di nodi accesi nelle ore di basso carico. Meno hardware acceso, meno climatizzazione, meno costi operativi.	**Sovranità del dato preservata** LLM on-premise: l'automazione AI non impone di esporre metriche o topologia a un servizio esterno.
Ogni azione difendibile in audit Cosa è stato eseguito, su quale infrastruttura, per quale segnale, con quale approvazione. Requisito, non accessorio, in contesti regolamentati.	**Stack unificato** Un solo sistema che osserva OpenStack, Ceph e Kubernetes in correlazione. Nessuna dashboard da aprire separatamente per ogni layer.

Stack e tecnologie

Telemetria	Prometheus + esporter nativi di OpenStack, Ceph-mgr e Kubernetes — lo standard, non reinventato
Motore LLM	Modelli open-weight eseguiti on-premise via Ollama — nessuna dipendenza da API esterne
Pipeline dati	ELK Stack + Zabbix per log strutturati e metriche di sistema
Esecuzione	AWX + Ansible — il modello genera il piano, il motore lo esegue; ogni step tracciato

Versionamento e audit	GitLab — ogni playbook legato all'evento di origine e all'infrastruttura di destinazione
Middleware	FastAPI — REST per stato e conferme, WebSocket/SSE per il push in tempo reale
Dashboard	React — vista unificata multi-stack con drill-down per layer e alert correlati
Notifiche	Email, Slack, Teams, SMS — configurabili per ambiente e per severità

A chi si rivolge

- **— operatori con infrastrutture multi-layer (OpenStack + Ceph + K8s) dove la correlazione tra stack diversi è critica per la diagnosi degli incidenti.**Telco & ISP
- **— chi gestisce infrastruttura per conto di clienti terzi e ha bisogno di visibilità aggregata con isolamento per cliente.**Cloud & Service Provider
- **— automazione AI compatibile con requisiti di sovranità del dato e tracciabilità degli interventi. Nulla esce dal perimetro.**Pubblica Amministrazione
- **— contesti dove ogni intervento sull'infrastruttura deve lasciare una traccia di audit completa e ogni azione deve essere approvata esplicitamente.**Finance & Banking
- **— organizzazioni con obiettivi di riduzione dei consumi energetici dell'infrastruttura IT.**Enterprise IT con focus sulla sostenibilità

Modello di deployment

Stato	In sviluppo — rilascio previsto 2027. Raccolta dati in corso su cluster interni per il training del modello.
Modalità	On-premise, all'interno del perimetro del cliente. Nessuna componente SaaS obbligatoria.
Prerequisiti	Infrastruttura OpenStack/Ceph/Kubernetes esistente; risorse dedicate al motore di inferenza on-premise.
Integrazione	Si posiziona sopra gli strumenti di monitoraggio esistenti (Prometheus, Grafana, ELK) — nessuna sostituzione dell'observability stack attuale.
Personalizzazione	Soglie, policy di autonomia e regole di correlazione configurabili per ambiente e per tipo di infrastruttura.
Supporto	Erogato da Epic Edge sull'intero stack (OpenStack, Ceph, Kubernetes, motore AI), non solo sul modulo.

L'ecosistema Epic Edge

Ogni prodotto Epic Edge fa parte di SentinelForge AI: moduli distinti sullo stesso substrato — LLM on-premise, esecuzione tracciata, policy di autonomia, audit, middleware multi-provider.

Principi architetturali

Realizzare una piattaforma di automazione AI per l'infrastruttura significa tenere insieme cose che di solito vivono separate: integrare provider tecnologicamente diversi, mantenere adapter indipendenti che evolvono senza toccare il resto, garantire un audit completo di ogni azione, preservare la sovranità del dato e — soprattutto — separare il motore che decide dal motore che esegue.

È per questo che CloudSentinel AI nasce come modulo dell'ecosistema SentinelForge AI: non una funzione aggiunta a uno strumento esistente, ma un'architettura pensata da chi quei vincoli li ha vissuti in produzione.

Perché Epic Edge

25+ anni di operations Data center reali, SLA reali, clienti enterprise reali. Non una startup che immagina il cloud.	**Pattern già collaudati** Motore LLM, esecuzione via AWX+Git, streaming real-time: già validati in produzione con EdgeForge AI e CephSentinel AI.	**Open-source e sovrano** Nessun lock-in, nessuna telemetria verso terzi, tutto dentro il perimetro del cliente.

Vuoi sapere quando CloudSentinel AI sarà disponibile?

Siamo in fase di sviluppo e raccolta dati. Se gestisci un'infrastruttura OpenStack, Ceph o Kubernetes e sei interessato a partecipare al programma early access, scrivici.

andrea.martra@epic-edge.it · www.epic-edge.it/epic-edge-platform

Epic Edge S.r.l. · Via Michele Buniva, 26 — 10064 Pinerolo (TO), Italia · info@epic-edge.it · www.epic-edge.it

OPEN-SOURCE • SOVEREIGN • ENTERPRISE-GRADE • AI-DRIVEN