EPICEDGE | SCHEDA PRODOTTO

CloudSentinel AI

IN SVILUPPO · 2027

Osservabilità predittiva, anomaly detection e ottimizzazione energetica per l'infrastruttura 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 equivalgono a nessun dato. Gli strumenti tradizionali segnalano tutto con la stessa urgenza, e chi gestisce l'infrastruttura impara a ignorare notifiche troppo numerose, ridondanti e che raramente indicano la causa reale.

Un guasto imminente raramente si annuncia con un singolo alert: si manifesta come un pattern distribuito su metriche diverse nel tempo. Chi gestisce cluster Ceph, OpenStack, Proxmox VE e Kubernetes in parallelo ha poche possibilità di correlare segnali che arrivano da sistemi con dashboard separate. Il costo di un'interruzione non è solo il downtime, ma il tempo sprecato a cercarne la causa 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.

Osserva l'infrastruttura nel suo insieme — compute, storage, networking, container — e quando individua un'anomalia non si limita a segnalarla: la classifica per impatto, la correla con eventi concomitanti e propone un piano d'azione leggibile, graduato per rischio. In più, la componente di ottimizzazione energetica analizza i pattern di carico, consolida i workload sui nodi attivi e riduce il consumo, con la stessa logica di autonomia graduata dell'ecosistema.

Il modello propone. Il motore di automazione esegue.

Come funziona

1 · RACCOGLIE

Telemetria unificata: metriche Prometheus, log strutturati, eventi da OpenStack, Proxmox VE, Ceph e Kubernetes. Ciclo continuo con trigger immediato sugli eventi critici.

2 · CORRELA

Motore a due livelli: regole deterministiche per i casi noti, LLM on-premise dove serve contesto. Due alert distinti sullo stesso rack diventano un unico incidente.

3 · PREVEDE

Analisi delle serie storiche per anticipare esaurimento risorse, degrado hardware e sovraccarico, con 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. Ogni azione è tracciata su Git.

«Perché il cluster Ceph-prod ha avuto latenze anomale stanotte, e cosa conviene fare?»

CloudSentinel AI →

Incidente correlato: flapping OSD sul rack B3 + saturazione rete · causa probabile: link degradato · impatto: medio

Piano d’azione proposto, graduato per rischio | esecuzione solo su conferma



Figura 1 — Dal rumore al segnale: CloudSentinel AI correla centinaia di alert da stack diversi in un unico incidente, con causa probabile e piano d’azione.



Figura 2 — Ottimizzazione energetica: consolidamento dei workload sui nodi attivi, spegnimento dei nodi scarichi e riattivazione automatica quando il carico risale.

Caratteristiche chiave

- **Osservabilità unificata** — metriche, log ed eventi di OpenStack, Proxmox VE, Ceph e Kubernetes in un'unica vista correlata: nessuna dashboard separata per ogni stack.
- **Correlazione multi-segnale** — il punto in cui 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 l'esaurimento delle 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 dei workload, spegnimento dei nodi sottoutilizzati e riattivazione automatica con verifica di consistenza.
- **LLM interamente on-premise** — nessun dato di infrastruttura, nessuna metrica, nessuna credenziale lascia il perimetro dell'organizzazione.
- **Autonomia graduata** — tre livelli configurabili per ambiente: sola segnalazione, remediation previa conferma, oppure esecuzione automatica limitata alle azioni a basso rischio.

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

Proiezioni di capacità e anomaly detection individuano i problemi prima che diventino interruzioni: intervento pianificato, non reazione d'emergenza.

Riduzione dei consumi energetici

Il consolidamento attivo riduce i nodi accesi nelle ore di basso carico: meno hardware acceso, meno climatizzazione, meno costi operativi.

Stack unificato

Un solo sistema osserva OpenStack, Proxmox VE, Ceph e Kubernetes in correlazione: nessuna dashboard da aprire separatamente per ogni layer.

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 e con quale approvazione: requisito, non accessorio, nei contesti regolamentati.

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
Dashboard e notifiche	React — vista unificata multi-stack con drill-down per layer; notifiche via Email, Slack, Teams, SMS

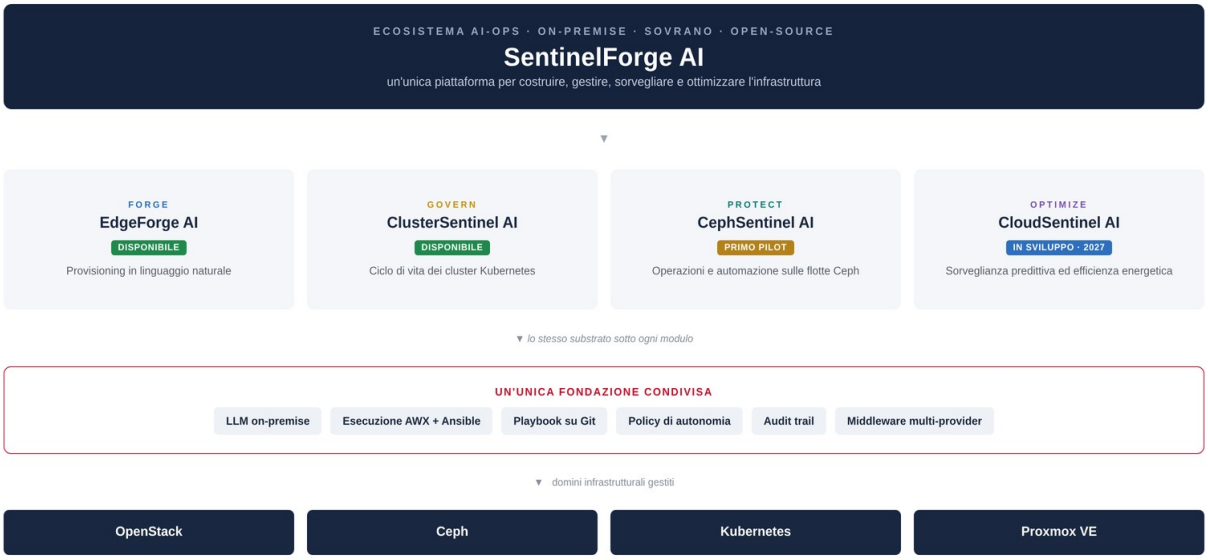
A chi si rivolge

- **Telco & ISP** — operatori con infrastrutture multi-layer (OpenStack + Ceph + K8s) dove la correlazione tra stack è critica per la diagnosi degli incidenti.
- **Cloud & Service Provider** — chi gestisce infrastruttura per conto di clienti terzi e ha bisogno di visibilità aggregata con isolamento per cliente.
- **Pubblica Amministrazione** — automazione AI compatibile con requisiti di sovranità del dato e tracciabilità degli interventi: nulla esce dal perimetro.
- **Finance & Banking** — contesti dove ogni intervento deve lasciare una traccia di audit completa e ogni azione va approvata esplicitamente.
- **Enterprise IT** — organizzazioni con obiettivi di riduzione dei consumi energetici dell'infrastruttura IT e di 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 / Proxmox VE / Ceph / Kubernetes esistente; risorse GPU 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, Proxmox VE, 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, ClusterSentinel 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, Proxmox VE, Ceph o Kubernetes e vuoi partecipare al programma early access, scrivici.

info@epic-edge.it · www.epic-edge.it/epic-edge-platform

