EPIC EDGE | SCHEDA PRODOTTO

CephSentinel AI

PRIMO PILOT DISPONIBILE

Monitoring continuo e automazione operativa per flotte Ceph

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

Il problema

Ceph è affidabile ma non è semplice. Un cluster in produzione parla per soglie e codici di stato: dice che qualcosa non va, non perché sta succedendo né cosa fare. Chi gestisce Ceph passa il tempo a tradurre warning in decisioni — e quella traduzione vive nella testa di due o tre persone.

Con più di un cluster il problema cambia natura. Cluster di clienti diversi, in datacenter diversi, alcuni dietro firewall che non espongono nulla. Nessuna vista d'insieme: per sapere dove guardare oggi bisogna aprire un cluster alla volta. E gli errori più costosi non sono i guasti, sono le dimenticanze: il flag noout lasciato acceso per settimane, la finestra di manutenzione che nessuno ha chiuso, il warning critico ignorato per ore perché nessuno stava guardando la dashboard.

La soluzione

CephSentinel AI è il modulo dell'ecosistema SentinelForge AI dedicato alle flotte Ceph. Non è un sistema di alerting in più: è un demone che osserva in autonomia, correla, deduplica, classifica e prioritizza i segnali invece di aspettare che qualcuno faccia una domanda, e gestisce più cluster contemporaneamente — anche geograficamente distinti e non co-locati con il sistema stesso.

Ogni cluster gestito è un oggetto di primo livello, con credenziali dedicate, referenti propri e una policy di automazione configurabile: un cliente può autorizzare più autonomia di un altro sulla stessa operazione. Metriche, eventi, previsioni e audit log sono partizionati per cluster fin dallo schema dati.

Il cuore non è la chat né il modello: è la vista di flotta. Un'unica schermata che risponde a "dove devo guardare oggi", con i cluster ordinati per urgenza. È ciò che un operatore con decine di cluster oggi non ha — e che nessuno strumento tradizionale gli offre.

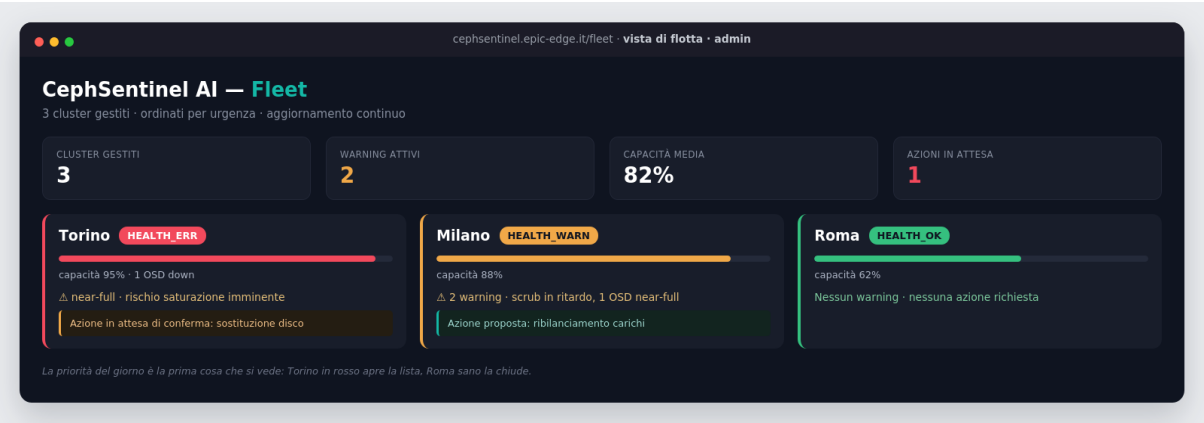

Figura 1 — Vista di flotta: i cluster ordinati per urgenza. La priorità del giorno è la prima cosa che si vede.

Il modello propone. Il motore di automazione esegue.

Come funziona

1 · OSSERVA

Collector per cluster: stato strutturato, topologia, utilizzo per pool, stato dei placement group, metriche SMART. Ciclo continuo più trigger immediato sui warning critici.

2 · INTERPRETA

Livello a regole per le soglie note. Livello LLM solo dove serve davvero contesto: spiegazione in linguaggio naturale calata su quel cluster.

3 · CORRELA E PREVEDE

Collega segnali che le soglie singole non collegherebbero mai — il problema di rete su un rack e il flapping degli OSD su quel rack sono un incidente solo. Proiezioni di capacità sul trend storico.

4 · PROPONE E AGISCE

Piano d'azione leggibile, classificato per rischio secondo la policy di quel cluster. Esecuzione tracciata, mai il modello direttamente sul cluster.

Figura 2 — Anatomia di un incidente: segnali diversi nel tempo correlati in un'unica diagnosi.

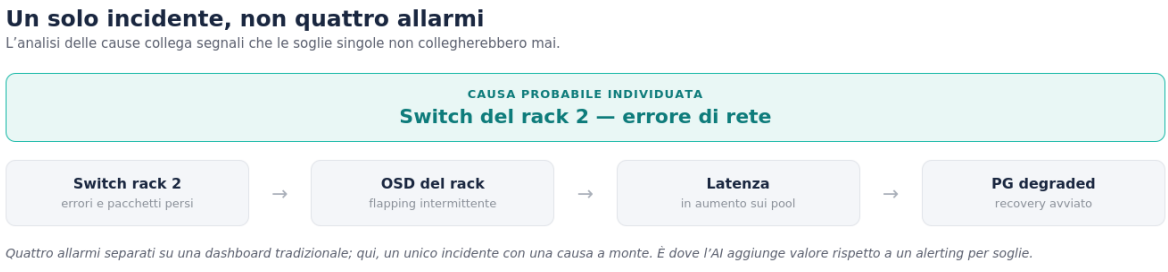

Figura 3 — Analisi delle cause: quattro allarmi separati, una sola causa a monte.

Autonomia proporzionale al rischio

> **AUTONOMO** **Raccolta, diagnostica, segnalazione**
> Health check estesi, verifica delle versioni dei daemon, scrub in ritardo, quorum, scrape delle metriche disco, accensione del LED del disco da sostituire, dry-run e anteprime. Nessun impatto sui dati.

> **SUPERVISIONATO** **Operatività quotidiana, con conferma**
> Rimozione OSD controllata con stima d'impatto, sostituzione disco end-to-end, modalità manutenzione host, orchestrazione dell'espansione del cluster (host e OSD), upgrade rolling per fasi, throttling del recovery tra orario di punta e finestra di manutenzione, flag temporanei con timeout automatico, gestione pool e credenziali.

> **SOLO CONSULENZA** **Il sistema spiega, l'operatore decide**
> Modifiche alla CRUSH map, soglie di riempimento, rimozione forzata di host, purge. Il sistema mostra la simulazione dell'impatto e i passi guidati; il comando finale resta sempre in mano all'operatore.

Figura 4 — I tre livelli di autonomia, configurabili per cluster e per cliente.

Caratteristiche chiave

- **Vista di flotta, non di cluster** — la home risponde a "dove devo guardare oggi": cluster ordinati per urgenza, warning attivi, azioni in attesa di conferma, e un cluster con connettività persa da ore è segnalato in modo diverso da uno sano.
- **Salute reale, non solo il codice Ceph** — un cluster tecnicamente HEALTH_OK ma con una previsione di esaurimento capacità imminente non appare "tutto verde".
- **Due modelli di connettività** — polling diretto verso l'endpoint del cliente, oppure agente leggero installato presso il cliente che si connette in uscita: nessuna porta in ingresso da aprire, requisito spesso non negoziabile in ambienti storage-critical.
- **Nessun vincolo sul metodo di deploy** — opera su qualunque cluster Ceph via CLI, o via API dove disponibili: non solo cluster gestiti da cephadm.
- **Motore a due livelli** — le regole deterministiche coprono i casi noti con latenza minima e zero rischio di interpretazione errata; l'LLM interviene solo dove serve contesto. Non si invoca un modello per un controllo di soglia.

- **Correlazione multi-segnale** — il pezzo dove l'AI aggiunge valore reale rispetto a un alerting tradizionale: due warning distinti che sono in realtà un unico incidente con una causa radice.
- **Il modello non esegue mai comandi sul cluster** — genera un piano strutturato; l'esecuzione passa da un motore separato (AWX + Ansible). È questo disaccoppiamento a rendere possibili audit, dry-run e revoca prima dell'esecuzione.
- **Tracciabilità completa** — ogni playbook eseguito è versionato su Git e legato esplicitamente al cluster di destinazione e al segnale che lo ha generato. Sempre ispezionabile cosa è stato fatto e perché.
- **Isolamento per cliente** — credenziali dedicate per cluster, mai condivise: un cluster compromesso non dà accesso agli altri.
- **Escalation automatica** — se un'azione supervisionata ad alta severità resta senza conferma oltre la soglia, il sistema alza il livello di notifica. Un warning critico ignorato per ore è di per sé un fallimento, va prevenuto strutturalmente.
- **LLM on-premise** — nessun dato di cluster, nessuna topologia, nessuna credenziale esce dal perimetro.

Benefici per il cliente

Una flotta, un operatore
Gestire dieci cluster smette di significare dieci sessioni aperte. La priorità del giorno è la prima cosa che si vede.

Meno dimenticanze costose
Flag con timeout, finestre di manutenzione sorvegliate, escalation sulle conferme mancate: gli errori operativi più comuni diventano strutturalmente difficili.

Competenza Ceph distribuita
La spiegazione contestualizzata rende operativo chi non ha vent'anni di Ceph alle spalle, senza togliere il controllo a chi ce li ha.

Ogni azione difendibile in audit
Cosa è stato eseguito, su quale cluster, per quale segnale, con quale approvazione. Requisito, non accessorio, in contesti regolamentati.

Autonomia a misura di cliente
Ogni cliente autorizza esplicitamente ciò che vuole automatizzare. Si parte prudenti e si allarga con la fiducia.

Sovranità del dato preservata
Motore on-premise: monitorare con l'AI non impone di mandare la telemetria dei propri cluster a un servizio esterno.

Stack e tecnologie

Telemetria	Prometheus + exporter nativo di ceph-mgr — lo standard Ceph, non reinventato
Collector	Servizio Python dedicato per cluster — stato strutturato completo, non solo serie numeriche
Motore LLM	Modelli open-weight eseguiti on-premise via Ollama

Esecuzione	AWX + Ansible, sopra l'orchestratore nativo Ceph dove presente; via CLI o API sui cluster che non lo espongono
Versionamento e audit	GitLab — ogni playbook legato a cluster ed evento di origine
Middleware	FastAPI — REST per stato e conferme, WebSocket/SSE per il push in tempo reale
Dashboard	React — vista di flotta e drill-down per cluster
Notifiche	Email, Slack, Teams, SMS — configurabili per cluster

A chi si rivolge

- **Service provider e MSP** — chi gestisce Ceph per conto di clienti terzi, con referenti, SLA e livelli di autonomia diversi per ciascuno.
- **Telco & ISP** — più cluster in più datacenter, dove la vista d'insieme oggi non esiste.
- **Enterprise multi-datacenter** — storage distribuito su siti diversi, con un team centrale che non può presidiare tutto a mano.
- **Pubblica Amministrazione** — requisiti di sovranità del dato e tracciabilità degli interventi sull'infrastruttura.
- **Finance & Banking** — storage ad alta disponibilità dove ogni modifica deve lasciare una traccia di audit.

Modello di deployment

Stato	Primo pilot disponibile — osservazione multi-cluster e azioni supervisionate
Modalità	On-premise, all'interno del perimetro del cliente
Connettività	Polling diretto verso l'endpoint del cluster, oppure agente leggero in uscita per cluster dietro firewall restrittivi
Compatibilità	Cluster Ceph in produzione, indipendentemente dal metodo di deploy — accesso via CLI o API
Avvio	Registrazione del cluster tramite wizard guidato con test di raggiungibilità immediato; policy iniziale prudente, raffinabile dopo
Autonomia	Configurabile per cluster sui tre livelli di rischio; il perimetro delle automazioni attivate si concorda in fase di attivazione — si parte con tutto supervisionato

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 CephSentinel 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

Ceph in produzione, non in teoria

Cluster reali, incidenti reali, SLA reali. La competenza viene dal campo.

Pattern già collaudati

Motore LLM, esecuzione via AWX+Git, streaming real-time: già validati in produzione con EdgeForge AI.

Open-source e sovrano

Nessun lock-in, nessuna telemetria verso terzi, tutto dentro il perimetro del cliente.

Quanti cluster Ceph state gestendo a mano?

Il primo pilot è disponibile. Registriamo un vostro cluster e vi mostriamo cosa vede — senza toccare nulla, finché non lo decidete voi.

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

EdgeForge AI	CephSentinel AI	CMP	CloudSentinel AI
LIVE	PILOT	IN IMPLEMENTAZIONE	ROADMAP

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