EPIC EDGE | PRODUCT SHEET

CloudSentinel AI IN DEVELOPMENT — 2027

Predictive analytics, anomaly detection and energy optimization for cloud infrastructure

Module of the SentinelForge AI ecosystem · on-premise · sovereign · open-source · AI-native

The problem

Cloud infrastructure generates a continuous stream of metrics, logs and events. The problem is not a lack of data: too much data is equivalent to no data at all. Traditional monitoring tools signal everything with the same urgency — and those managing the infrastructure learn to ignore notifications because there are too many, they are redundant, and they rarely point to the real cause.

An imminent compute node failure rarely announces itself with a single alert: it manifests as a distributed pattern across different metrics over time. Anyone reading only point-to-point alerts cannot see the pattern. Those managing Ceph, OpenStack and Kubernetes clusters in parallel have even less chance of correlating signals arriving from systems with separate dashboards.

The cost of an unplanned outage is not just the downtime: it is the time wasted searching for the root cause among hundreds of uncorrelated notifications while the problem worsens.

The solution

CloudSentinel AI is the SentinelForge AI ecosystem module dedicated to predictive observability and cloud infrastructure optimization. It is not another monitoring system: it is an intelligence layer that sits on top of existing tools — Prometheus, Grafana, Zabbix, ELK — collecting signals, correlating events and identifying patterns that a single threshold cannot see.

CloudSentinel AI observes the infrastructure as a whole: compute, storage, networking, containers. When it detects an anomaly, it does not simply flag it — it classifies it by probable impact, correlates it with concurrent events and proposes a readable, risk-graduated action plan. The operator already sees the likely root cause and the suggested action, not just the raw alert.

The energy optimization component works continuously: it analyses load patterns over time, identifies underutilised nodes, proposes workload consolidation and active consumption reduction — with the same graduated autonomy logic that characterises the SentinelForge AI ecosystem.

The model proposes. The automation engine executes.

How it works

1 · COLLECTS

Unified telemetry pipeline: Prometheus metrics, structured logs, events from OpenStack, Ceph and Kubernetes. Continuous cycle with immediate trigger on critical events.

2 · CORRELATES

Two-level engine: deterministic rules for known cases, on-premise LLM where context is needed. Two separate alerts on the same rack are recognised as a single incident.

3 · PREDICTS	4 · PROPOSES AND ACTS
Historical series analysis to anticipate resource exhaustion, hardware degradation and overload patterns. The projection includes a time window and confidence margin.	Readable action plan, classified by risk. Execution via AWX + Ansible: the model never touches the infrastructure directly. Every action is tracked on Git.

Energy optimisation

CloudSentinel AI includes a dedicated module for optimising infrastructure energy consumption. The analysis is continuous: the system observes load patterns over a configurable time horizon, identifies recurrently underutilised compute nodes and proposes active workload consolidation.

Load analysis	Workload consolidation	Automatic restart
Identifies nodes with low average utilisation over configurable time windows. Distinguishes structural underutilisation from episodic underutilisation.	Proposes live migration of workloads to active nodes, with impact estimation before execution. The emptied node is physically powered off.	When load increases again, the system powers nodes back on, verifies software consistency, updates them if necessary and reintroduces them into the cluster.

Graduated autonomy

Like every module in the SentinelForge AI ecosystem, CloudSentinel AI operates on three autonomy levels configurable per environment and operation type.

● AUTONOMOUS — Collection, analysis, alerting

Continuous health checks, metric collection, anomaly detection, capacity projections, alert classification by impact, action plan generation. No impact on infrastructure.

● SUPERVISED — Optimisation and remediation, with confirmation

Workload consolidation, node shutdown and restart, resource throttling, application of optimised configurations, guided rolling upgrades. Every action is presented with impact estimation and requires explicit confirmation.

● ADVISORY ONLY — The system explains, the operator decides

Architectural changes, high-impact configuration modifications, interventions on degraded components. The system provides impact simulation and guided steps; the final command always remains in the operator's hands.

Key features

- **Unified observability** — metrics, logs and events from OpenStack, Ceph and Kubernetes in a single correlated view: no separate dashboard for each stack.
- **Multi-signal correlation** — where AI adds real value over traditional alerting: two distinct warnings that are actually one incident with a single root cause.
- **AI-driven anomaly detection** — the system learns the normal patterns of the infrastructure and signals significant deviations, not just individual thresholds exceeded.

- **Capacity projections** — historical trend analysis to anticipate resource exhaustion with time window and confidence margin.
- **Automated Root Cause Analysis** — for every detected anomaly, the system produces a natural-language explanation tailored to that specific environment.
- **Active energy optimisation** — workload consolidation, shutdown of underutilised nodes and automatic restart with consistency verification.
- **Fully on-premise LLM** — no infrastructure data, no metrics, no credentials leave the perimeter.
- **The model never executes commands** — it generates a structured plan; execution goes through AWX + Ansible. Audit, dry-run and revocation before execution.
- **Data collection already underway** — Prometheus + Zabbix + ELK pipeline running on an internal cluster to train the model.
- **Complete traceability** — every executed playbook is versioned on Git and explicitly linked to the event that triggered it.

Customer benefits

Fewer alerts, more signal Multi-signal correlation reduces noise: the operator sees incidents, not notifications. Less time on triage, more time on the root cause.	**Failures anticipated** Capacity projections and anomaly detection identify problems before they become outages. Planned intervention rather than emergency response.
Reduced energy costs Active workload consolidation reduces the number of nodes running during low-load hours. Less hardware on, less cooling, lower operational costs.	**Data sovereignty preserved** On-premise LLM: AI automation does not require exposing metrics or topology to an external service.
Every action defensible in audit What was executed, on which infrastructure, for which signal, with which approval. A requirement, not an accessory, in regulated contexts.	**Unified stack** One system observing OpenStack, Ceph and Kubernetes in correlation. No separate dashboard to open for each layer.

Stack and technologies

Telemetry	Prometheus + native exporters for OpenStack, Ceph-mgr and Kubernetes — the standard, not reinvented
LLM engine	Open-weight models running on-premise via Ollama — no dependency on external APIs
Data pipeline	ELK Stack + Zabbix for structured logs and system metrics
Execution	AWX + Ansible — the model generates the plan, the engine executes it; every step tracked
Versioning and audit	GitLab — every playbook linked to the source event and target infrastructure
Middleware	FastAPI — REST for state and confirmations, WebSocket/SSE for real-time push

Dashboard	React — unified multi-stack view with drill-down by layer and correlated alerts
Notifications	Email, Slack, Teams, SMS — configurable per environment and severity

Target audience

- **Telco & ISP** — operators with multi-layer infrastructure (OpenStack + Ceph + K8s) where cross-stack correlation is critical for incident diagnosis.
- **Cloud & Service Providers** — those managing infrastructure on behalf of third-party customers who need aggregated visibility with per-customer isolation.
- **Public Administration** — AI automation compatible with data sovereignty and intervention traceability requirements. Nothing leaves the perimeter.
- **Finance & Banking** — contexts where every infrastructure intervention must leave a complete audit trail and every action must be explicitly approved.
- **Sustainability-focused Enterprise IT** — organisations with targets to reduce IT infrastructure energy consumption.

Deployment model

Status	In development — planned release 2027. Data collection underway on internal clusters to train the model.
Mode	On-premise, within the customer's perimeter. No mandatory SaaS component.
Prerequisites	Existing OpenStack/Ceph/Kubernetes infrastructure; resources dedicated to the on-premise inference engine.
Integration	Positioned above existing monitoring tools (Prometheus, Grafana, ELK) — no replacement of the current observability stack.
Customisation	Thresholds, autonomy policies and correlation rules configurable per environment and infrastructure type.
Support	Provided by Epic Edge across the entire stack (OpenStack, Ceph, Kubernetes, AI engine), not just the module.

The Epic Edge ecosystem

Every Epic Edge product is part of SentinelForge AI: distinct modules on the same substrate — on-premise LLM, tracked execution, autonomy policies, audit, multi-provider middleware.

Architectural principles

Building an AI automation platform for infrastructure means holding together things that usually live apart: integrating technologically diverse providers, maintaining independent adapters that evolve without touching the rest, ensuring a complete audit of every action, preserving data sovereignty and — above all — separating the engine that decides from the engine that executes.

This is why CloudSentinel AI is born as a module of the SentinelForge AI ecosystem: not a function added to an existing tool, but an architecture designed by those who have lived with those constraints in production.

Why Epic Edge

25+ years of operations	Battle-tested patterns	Open-source and sovereign
Real data centres, real SLAs, real enterprise customers. Not a startup imagining what the cloud is.	LLM engine, AWX+Git execution, real-time streaming: already validated in production with EdgeForge AI and CephSentinel AI.	No lock-in, no telemetry to third parties, everything within the customer's perimeter.

Want to know when CloudSentinel AI will be available?

We are in active development and data collection. If you manage an OpenStack, Ceph or Kubernetes infrastructure and are interested in joining the early access programme, get in touch.

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

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

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