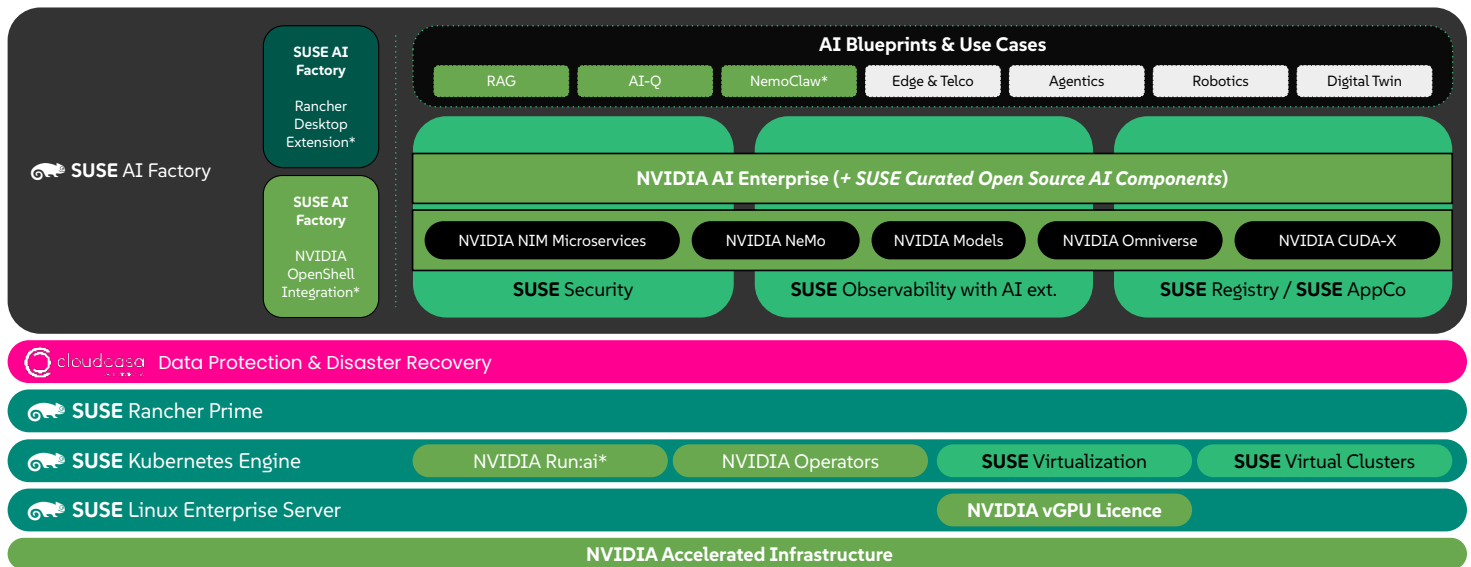


SUSE AI Factory with NVIDIA and CloudCasa

Backup, recovery and workload mobility for AI workloads on SUSE Rancher Prime

SUSE AI Factory gives enterprises a sovereign platform to run NVIDIA AI Enterprise, NIM microservices and RAG blueprints on Rancher Prime and SUSE Kubernetes Engine. CloudCasa protects what those clusters produce and depend on: model weights, fine-tuned checkpoints, vector databases and the Kubernetes configuration that ties them together. Teams recover a failed or compromised AI cluster in hours, not days, keep idle GPU time to a minimum, and keep every backup inside their own infrastructure. CloudCasa runs self-hosted or as SaaS at about one-third the price of competing Kubernetes backup products.



Protect AI data and configuration

Application-consistent backups of persistent volumes holding model weights, checkpoints and vector stores (Milvus, OpenSearch, pgvector) used in RAG blueprints. CloudCasa also captures the Kubernetes state of each deployment: NIM microservices, Helm releases, GPU operator settings, secrets and CRDs. Immutable copies on S3-compatible storage with object lock keep restore points safe from ransomware.

Sovereign and air-gapped operation

CloudCasa self-hosted installs on Rancher-managed clusters and runs fully disconnected, matching air-gapped SUSE AI Factory deployments at the edge, in defense and in regulated industries. Backup data and metadata stay in the customer's data center or sovereign cloud, supporting NIS2, DORA and data-residency requirements. Platform teams manage it from the Rancher Prime UI through the CloudCasa extension.

Workload mobility and fast recovery

Cross-cluster migration (Kubernetes to Kubernetes) moves AI workloads between edge, core data center and cloud clusters, including across storage classes. Full-cluster restore to new GPU infrastructure after an outage cuts the time expensive GPUs sit unused. VMs on SUSE Virtualization are protected alongside containers under one policy engine.

Example Implementations with CloudCasa

SUSE AI Factory runs the AI stack on Kubernetes, and CloudCasa protects it as Kubernetes: backup, restore, migration and disaster recovery for the vector store, object data and application state behind AI services, installed as a Rancher extension. The examples below show three common ways the two fit together.

1. Sovereign AI in regulated environments

Setup. A regulated organisation, such as in financial services, the public sector or healthcare, runs an internal RAG assistant on SUSE AI Factory in its own data centre. Regulations such as DORA or NIS2 apply, and no component may depend on an external SaaS control plane.

With CloudCasa. CloudCasa runs self-hosted, including fully air-gapped, alongside Rancher Prime. Scheduled backups of the AI namespace go to S3-compatible storage the organisation owns, with an immutable copy at a second site. Regular restore tests to an isolated cluster produce timestamped job records.

Result. The control plane and every recovery copy stay on the organisation's own infrastructure and in its jurisdiction, and recovery testing is documented for audit.

2. AI-as-a-Service on a shared platform

Setup. A cloud or managed service provider offers GPU inference and RAG services on SUSE AI to multiple tenants, next to existing Rancher-based services such as managed Kubernetes or databases. Each tenant has its own namespace and vector store.

With CloudCasa. AI tenants use the same backup policies, immutable offsite copies and per-tenant restore as the provider's other stateful services. Restoring one tenant's vector store brings that tenant's assistant back without re-embedding its documents or affecting other tenants. Backup and DR can be offered as a tenant service option.

Result. One protection tool and one runbook across all platform services, at fleet scale.

3. Enterprise AI from pilot to production

Setup. An enterprise moves a RAG application on SUSE AI Factory from pilot to production. A typical stack includes model serving (vLLM), a vector database (Milvus), object storage (MinIO), PostgreSQL, Keycloak and a chat front end. A second site and hardware refreshes are planned.

With CloudCasa. CloudCasa backs up the whole application namespace, Kubernetes resources and persistent volumes, with application hooks for consistent database backups. The stack can be restored in place, or to a different cluster on different storage for hardware refresh, upgrades or DR.

Result. Restoring a vector index is a data copy that takes minutes. Rebuilding it means re-ingesting and re-embedding the corpus on GPU capacity, which grows with the size of the data.

Where it applies

Any AI deployment that keeps its own data in the cluster: RAG document corpora, vector databases, fine-tuned models, chat history. Stateless inference endpoints that can be redeployed from Git need less protection.

What CloudCasa protects in the stack

Applications	Chat front ends such as Open WebUI, RAG pipelines, Keycloak
Models	vLLM and NVIDIA NIM serving configuration, model weights, fine-tuned checkpoints
Data	Milvus vector indexes, MinIO object storage, PostgreSQL, Redis, chat history
Kubernetes	Namespaces, deployments, CRDs, secrets, persistent volumes
CloudCasa: backup, restore, migration and DR for every layer, to any cluster or storage	

SUSE AI Factory with NVIDIA and CloudCasa

**SUSE AI Factory with NVIDIA****Pre-Validated, Use-Case Aligned Blueprints**

- Tightly packaged open-source AI components and 3rd-party ISV software into blueprints tailored for specific workloads (GenAI Inference, RAG, Computer Vision, MLOps)
- Initially aligned with proven SUSE domains (Edge, Telco, SAP) providing turnkey solutions as commercial beachheads

Scalable Deployments Anywhere

- Seamlessly provision AI workloads across any supported environment: tactical edge, core data center, public cloud
- Manage both decentralized, air-gapped footprints and large enterprise fleets through single pane of glass

Unified Governance & Observability

- Single common portal surfacing all critical telemetry
- Fine-grained, real-time visibility into GPU utilization, LLM token throughput, cost modeling, traffic patterns

**Data Protection & Disaster Recovery****Recover AI clusters fast**

Full-cluster and namespace-level restore with PV data to the same or new clusters, so GPU nodes return to production quickly

Disaster recovery

Restore a complete AI cluster, including PVs and NIM deployments, onto standby or new infrastructure in another site or cloud region

Workload mobility

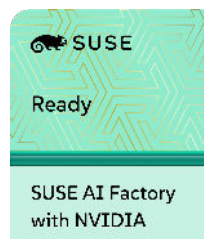
Kubernetes-to-Kubernetes migration across clusters, distributions and storage classes, from edge to core to cloud

Data sovereignty

Self-hosted and air-gapped deployment, with backups on customer-owned S3-compatible storage

Lower cost

About one-third the cost of competitors

[Learn more](#)cloudcasa.io

SUSE is the leading sovereign open infrastructure company where hybrid AI runs. Backed by 30 years of community collaboration, we transform open source innovation into secure, enterprise-grade solutions for mission-critical environments. Built on choice and neutrality, SUSE is the only platform that manages any Linux and Kubernetes distribution. Through our open infrastructure platform and partner ecosystem, IT leaders gain control over their data, budgets, and sovereign AI strategies. Our solutions deploy across hybrid environments, from developer workstations and on-premises data centers to multi-cloud and the farthest edge, and are optimized and certified for SAP. With SUSE, choice happens.

CloudCasa by Catalogic Software provides backup, disaster recovery and workload mobility for Kubernetes and KubeVirt VMs across on-premises, edge and cloud clusters, including AI workloads such as model serving, vector databases and RAG pipelines. Storage-integrated DR works with array-level replication to recover clusters and their data at a second site. Delivered as self-hosted or as SaaS, CloudCasa protects Rancher Prime, SUSE Kubernetes Engine, SUSE Virtualization and SUSE AI Factory environments, integrates with the Rancher UI, and is priced at about one-third of competing solutions.

SUSE S.A.
11-13 Boulevard de la Foire
L-1528 Luxembourg
Grand Duchy of Luxembourg

benelux@suse.com
+31 23-8080010

[SUSE.COM](https://suse.com)

Copyright © 2026 SUSE LLC and/or its affiliates. All rights reserved. SUSE, the SUSE logo, and all SUSE marks are trademarks or registered trademarks of SUSE LLC or its subsidiaries in the United States and other countries. All third-party trademarks are the property of their respective owners.