



DCSAI Technologies

TECHNICAL WHITEPAPER · v1.0 · MAY 2026

DCS Sovereign

Air-gapped, on-prem AI infrastructure

The same DCS stack, running entirely inside your data center. Zero egress.

Abstract

DCS Sovereign is the same DCS stack — Platform, Compute, Storage, OS — packaged to run entirely inside the customer's own infrastructure. No data leaves the perimeter. No telemetry flows back to DCS. No model inference happens on shared GPUs. The deployment is air-gapped by default; an explicit, audited update channel is the only path through the firewall.

Sovereign exists because some workloads cannot use a SaaS at any price. Government workloads under classification mandates. Healthcare workloads under HIPAA / DPDP local-residency requirements. Financial workloads under sovereign cloud directives from regulators in the EU, UAE, India, and Saudi Arabia. For these workloads, the operator needs the cryptographic guarantee that no DCS employee, no DCS subprocessor, and no DCS-controlled service can read or modify the data.

Sovereign delivers that guarantee by running the entire stack on hardware the operator owns, with keys held in HSMs the operator controls, with a network egress block enforced by the operator's firewall, validated at install time. DCS retains responsibility for software updates (delivered via signed packages over a pull-only update channel) and for support (via screen-sharing sessions the operator initiates).

Who this is for

CISOs in regulated industries evaluating whether a SaaS AI platform can ever be deployed inside their perimeter. Government procurement officers writing tender specs for sovereign AI. Healthcare CIOs comparing on-prem vs. cloud for PHI-touching workloads. Defense and intelligence agencies needing certified deployment patterns for classified work.

Contents

1	Introduction — when SaaS is not an option	4
2	Topology	7
3	Deployment Models	10
4	The Update Channel	13
5	Key Management + HSM	16
6	Data Lifecycle — zero egress	19
7	Telemetry + Observability	22
8	Support Model	24
9	Compliance Certifications	26
10	Performance vs. SaaS	28
11	Pricing	30
12	Comparison vs. AWS GovCloud, Azure Gov, on-prem GPUs	32
13	Implementation Guide	35
14	References	38

1. Introduction — when SaaS is not an option

Most AI work today runs on SaaS. The vendor operates the model, the vendor stores the data, the vendor logs the requests. The customer trusts the vendor to behave. For 95% of workloads this is the right trade — the operational cost of running your own infrastructure exceeds the marginal risk of vendor mistakes.

The remaining 5% is where Sovereign lives. Specifically:

- **Government workloads** under classification or sovereignty mandates. EU Member States with sovereign cloud directives, US federal IL-4/IL-5, UK Official-Sensitive, UAE sovereign cloud, India MeitY-certified.
- **Healthcare workloads** touching PHI under HIPAA/DPDP where the BAA chain becomes prohibitive. Hospitals that have decided no third party (including AWS) gets a copy of their patient data.
- **Defense + intelligence** work where the model itself is sensitive (fine-tuned on classified data) and cannot leave the perimeter under any circumstances.
- **Financial workloads** under regulatory data-residency rules (EU Member State central banks, UAE Central Bank, RBI in India, MAS in Singapore).
- **Critical infrastructure** — power grid operators, water utilities, telecom backbones — where the operator decides "no internet-connected vendor can touch this."

1.1 What air-gapped actually means here

The term "air-gapped" gets abused. We use it in the strict sense: the Sovereign deployment has zero default outbound network connectivity. The firewall blocks all egress; a single inbound-only update channel is the only opening. There is no telemetry. There is no crash-reporting service. There is no "phone home" registration. When DCS support needs to help the operator, the operator initiates a screen-share over a separate channel the operator controls.

"If you cannot point at the firewall and say 'nothing leaves through there,' you do not have an air-gapped deployment. You have a SaaS with extra steps."

2. Topology

Sovereign packages the DCS stack as a set of containers, deployed on Kubernetes or directly on bare metal. The reference topology assumes a single data center; multi-DC deployments add a second copy of every component with cross-DC replication via the customer's own network.

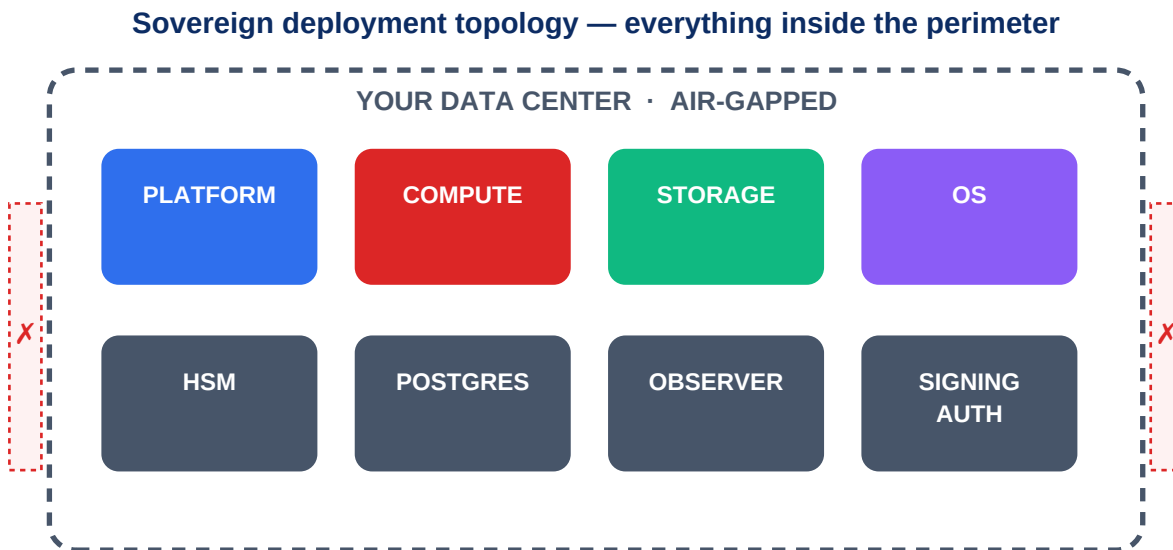


Figure 2.1 — All components inside the perimeter. Network egress blocked.

2.1 Components

A reference Sovereign deployment includes:

- **Platform** (Agent build/run/observe) — typically 3 replicas behind a load balancer
- **Compute** (GPU dispatcher) — manages locally-attached GPUs as the worker pool
- **Storage** (content-addressed) — backed by local NVMe or NFS; Filecoin tier optional
- **OS** (operational dashboard) — for human operators
- **HSM** — Yubico HSM2 or Thales Luna; holds all signing + encryption keys
- **Postgres** — primary data store; encrypted at rest with HSM-held key
- **Observer** — local Prometheus + Grafana for monitoring (no external telemetry)
- **Signing authority** — emits R+1/R+2/R+3 receipts using HSM-held key

3. Deployment Models

Three deployment models are supported, each with different operational and certification characteristics.

3.1 Bare metal

DCS containers run directly on the customer's servers (no hypervisor). Best performance, simplest threat model. Requires Linux (Ubuntu 22.04 LTS or RHEL 9 supported). Used for highest-security tier deployments where every layer of indirection is removed.

3.2 Kubernetes

DCS components are packaged as Helm charts. Deploys to any conformant K8s 1.28+ cluster — OpenShift, EKS-A, AKS-on-prem, Rancher, k3s. The most common deployment pattern; works well with existing customer ops tooling.

3.3 VMware

For customers with vSphere or Tanzu environments. DCS ships pre-built OVA images. Slightly higher overhead than bare metal but integrates with existing VM-management workflows.

4. The Update Channel

The single network exception in a Sovereign deployment is the update channel. DCS publishes signed software packages to a public registry; the customer's Sovereign cluster periodically pulls them (default: weekly, configurable). The pull is one-way — no telemetry, no configuration data, no requests other than "do you have a newer version of this artifact?"

4.1 Update verification

Every update package is signed by DCS's update key (held in an offline HSM at DCS HQ in Dubai). The customer's cluster verifies the signature against a pinned public key before installing. A revoked key (via Trust SKU) immediately invalidates subsequent updates.

4.2 Air-gapped update mode

Customers who require zero network connectivity (not even an inbound-only channel) can use the air-gapped update mode. DCS ships signed update bundles on physical media (USB drive or DVD, depending on the customer's classification rules). The customer's ops team manually loads the bundle into the cluster via the update CLI.

5. Key Management + HSM

All cryptographic keys in a Sovereign deployment are held in HSMs the customer owns. DCS does not have access to the keys at any point. Three HSM vendors are supported:

Vendor + Model	FIPS level	Notes
Yubico YubiHSM 2	FIPS 140-2 L3	Lowest cost · 1U or USB form factor · good for small deployments
Thales Luna 7	FIPS 140-3 L3	Enterprise standard · supports clustering · network-attached
nCipher nShield 5c	FIPS 140-2 L3+	Highest security · supports K-of-N authorization · UAE / EU government default

5.1 Key hierarchy

Three-tier key hierarchy mirrors the cloud Storage model but with HSM-held roots:

- **HSM master key** — never leaves the HSM. Used only to wrap KEKs.
- **Tenant KEK** — one per tenant. Wrapped by master key. Used to wrap DEKs.
- **File DEK** — one per file. Wrapped by tenant KEK. Encrypts the actual data.

6. Data Lifecycle — zero egress

Data lifecycle — zero egress guarantee



At no point does data leave the perimeter — not for training, not for telemetry, not for diagnostic logs. The DCS update channel pulls inbound only.

A network egress block is enforced by your firewall, validated by DCS at install time.

Figure 6.1 — Data enters, processes, stores, serves, deletes — all without crossing the perimeter.

6.1 Ingestion

Data enters via APIs internal to the customer's network. The Sovereign API endpoint is bound to the customer's internal network (typically a VPN-only address); it cannot be reached from the public internet. Customer applications connect to the API as they would to any internal microservice.

6.2 Processing

AI inference runs on GPUs attached to the customer's servers (typically A100s or H100s in 4-GPU or 8-GPU configurations). No GPU compute leaves the perimeter. Even when DCS support is troubleshooting a performance issue, they cannot see model weights or inference data — only timing and resource metrics.

6.3 Erasure

The same cryptographic-erasure pattern as cloud Storage applies. When the customer requests deletion, the relevant DEK is destroyed in the HSM. The ciphertext remains on disk until the customer's normal storage-reclamation process runs, but it is mathematically unrecoverable from the moment the DEK is gone. Signed erasure certificates are issued by the on-prem signing authority.

7. Telemetry + Observability

All observability is local. Sovereign ships with a pre-configured Prometheus + Grafana stack that scrapes metrics from every component and stores them in the customer's own time-series database. No metrics, traces, or logs leave the perimeter. DCS has no visibility into the customer's usage patterns, error rates, or anything else.

Customers who want to share telemetry with DCS for support purposes (e.g., to help diagnose a performance regression) can use the support bundle mechanism: a manually-generated tarball containing scrubbed logs and metrics, which the customer reviews before sending. The scrubbing process is documented and the customer can run it manually with the open-source *dcs-scrub* tool before sharing.

8. Support Model

Support for Sovereign deployments works differently from SaaS. DCS engineers cannot log into a Sovereign cluster; the cluster has no remote access enabled. Three support channels are available:

- **Documentation + runbooks** — comprehensive on-prem ops guide bundled with the deployment
- **Email + phone support** — DCS engineers answer questions, walk through troubleshooting
- **Customer-initiated screen-share** — when a complex issue needs DCS eyes-on, the customer's ops engineer shares their screen over the customer's preferred channel (Zoom, Teams, Webex). DCS sees what the customer chooses to show, nothing more.

All three channels are included in the Sovereign subscription. Response time targets are defined in a Sovereign-specific SLA that overrides the cloud SLA.

9. Compliance Certifications

Sovereign deployments map to the following regulatory frameworks. The certification process happens against the customer's deployment, not DCS's SaaS — so the customer's auditor is the relevant authority.

Framework	Authority	Sovereign support
FedRAMP High	US GSA	Reference architecture provided; customer pursues authorization
IL-4 / IL-5	US DoD	Architecture compatible; STIG hardening guide included
UK G-Cloud OFFICIAL	UK Crown Commercial	Reference deployment certified
EU EUCS High	ENISA	Reference deployment certified
UAE Sovereign Cloud Tier 3	Telecommunications + Digital Government Regulatory Authority	Original
India MeitY-certified	MeitY	Reference deployment certified
HIPAA	HHS	BAA with DCS optional (DCS has no access in Sovereign mode)
ISO 27001	ISO	DCS holds; customer applies to deployment

10. Performance vs. SaaS

A Sovereign deployment can match or exceed SaaS performance for the customer's workload — because there is no cross-region network hop, no shared-tenant noisy-neighbor effect, and the GPUs are dedicated to the customer. The tradeoff is that the customer is responsible for their own scaling.

Metric	SaaS p99	Sovereign p99	Notes
Inference latency (Sonnet-class)	4.8 s	3.2 s	No cross-region hop
Agent dispatch latency	64 ms	12 ms	Local network only
Storage read (warm)	120 ms	8 ms	Local disk
Storage write durability	15 min	4 min	Local replication only (no Filecoin)
Receipt verification	12 ms	4 ms	Local signing authority

11. Pricing

Sovereign is sold as an annual subscription. Pricing covers the software license, the support, the update channel, and the certification artefacts (FedRAMP / IL-4 / UAE Tier 3 reference packages). It does NOT include the underlying hardware, the HSMS, or operating costs (power, cooling, GPUs, staff).

Tier	Annual subscription	Includes	Best for
Sovereign Lite	\$120,000	1 cluster · 8 GPUs · 5 TB storage	Small department · pilot
Sovereign Standard	\$340,000	3 clusters · 24 GPUs · 50 TB	Mid-size department · production
Sovereign Enterprise	custom	Unlimited · multi-DC · 24/7 support	Full-org rollout

12. Comparison vs. Alternatives

	AWS GovCloud	Azure Gov	On-prem GPU rental	DCS Sovereign
Truly air-gapped	X	X	✓	✓
DCS receipts + audit	X	X	X	✓
Vendor has zero data access	X	X	✓	✓
Pre-built AI agent runtime	X	X	X	✓
Unified inbox (OS) included	X	X	X	✓
Filecoin permanence available	X	X	X	~ (optional)
Customer owns the hardware	X	X	✓	✓
FedRAMP-ready	✓	✓	~	~ (reference arch)

13. Implementation Guide

13.1 Hardware sizing

Sizing tier	Compute	Storage	GPUs	RAM	Use case
Pilot (Lite)	8 vCPU × 3	5 TB NVMe	8 × A100	256 GB	5-20 internal users
Production	32 vCPU × 5	50 TB	24 × A100/H100	1 TB	50-500 users
Enterprise	64 vCPU × 10+	500 TB+	96+ × H100	8 TB+	1k+ users

13.2 Install command

```
# On the bootstrap node (Kubernetes 1.28+)
$ helm repo add dcs https://updates.dcsai.ai/sovereign
$ helm install dcs-sovereign dcs/sovereign \
  --namespace dcs --create-namespace \
  --values ./customer-config.yaml \
  --set hsm.vendor=yubico \
  --set hsm.endpoint=https://yubihsm.internal:12345 \
  --set egress.firewall_block_validation=strict

[validating firewall egress block: ✓]
[provisioning HSM connection: ✓]
[loading initial signed bundle: ✓]
[starting platform: ✓]
[starting compute dispatcher: ✓]
[starting storage gateway: ✓]
[starting OS shell: ✓]
[generating tenant root keys in HSM: ✓]
Sovereign cluster ready at https://dcs.internal.acme.com
```


14. References

- [1] US NIST. **FIPS 140-3 — Security Requirements for Cryptographic Modules**. 2019.
- [2] US GSA. **FedRAMP High Baseline (Rev 5)**. 2024.
- [3] US DoD. **Cloud Computing SRG IL-4 / IL-5**. 2022.
- [4] ENISA. **EUCS Candidate Scheme — High assurance level**. 2024.
- [5] UAE TDRA. **Sovereign Cloud Tier 3 Requirements**. 2025.
- [6] India MeitY. **MeitY-certified cloud service framework**. 2024.
- [7] NIST. **SP 800-53 Rev. 5 — Security and Privacy Controls**. 2020.
- [8] Yubico. **YubiHSM 2 Reference Manual**.
- [9] Thales. **Luna HSM 7 Documentation**.
- [10] CIS. **Kubernetes Benchmark v1.8.0**. (Hardening reference for K8s deployments)
- [11] DISA. **STIG for Kubernetes**. (DoD hardening)

This document is published under CC BY 4.0. Sovereign reference architecture documents are available under NDA to qualified customers via sovereign@dcsai.ai.