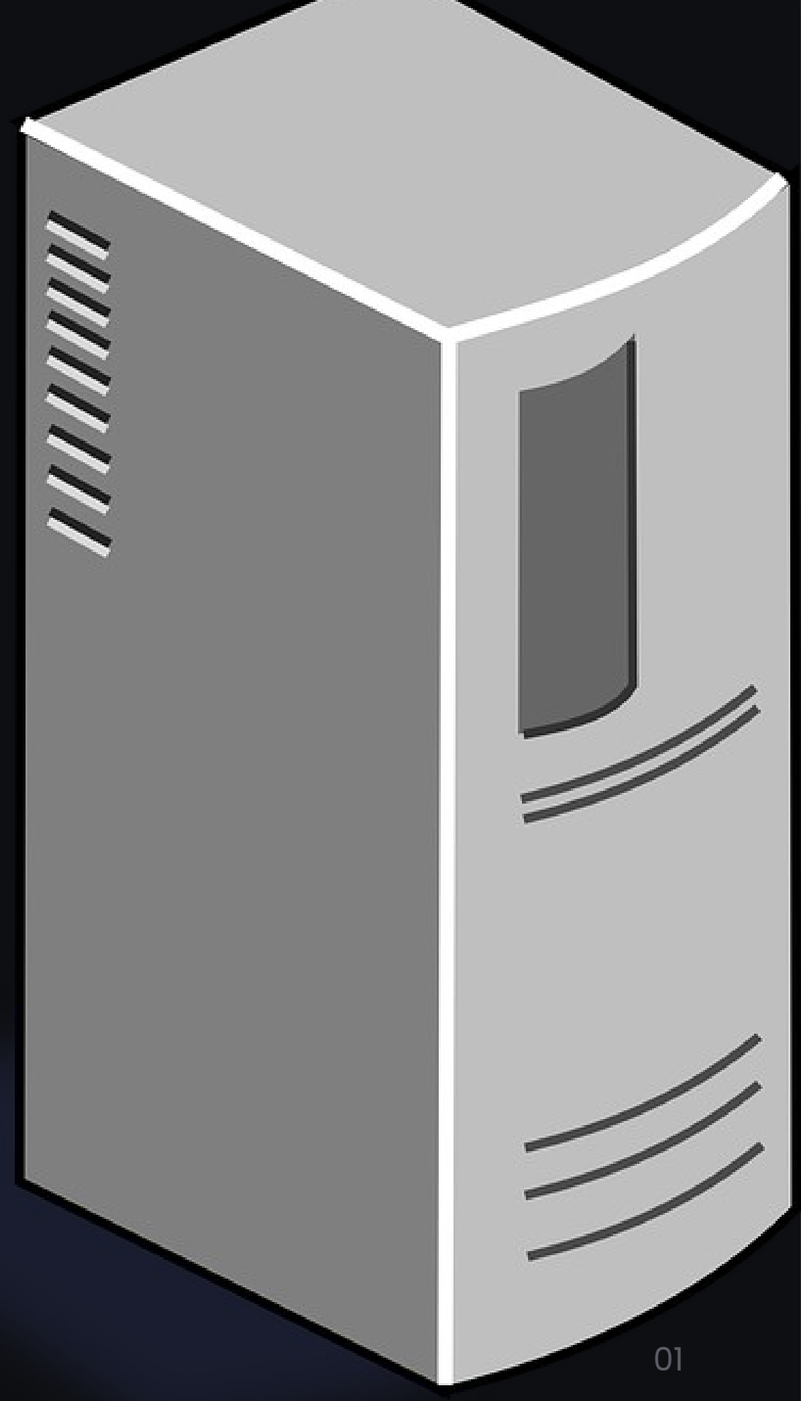
SERVICE · INFRASTRUCTURE

Cloud GPU and AI infrastructure

AI compute capacity with your data hosted on home soil.

Our own datacentre in A Coruña's Ciudad de las TIC, plus elastic deployment across multiple clouds.



cinfo

THE CHALLENGE

Training and serving models without depending on one single cloud

AI needs GPUs, and GPUs force decisions: where the data lives, what a demand peak costs, and what happens when the provider changes prices or terms.

01

Unpredictable cost

Training and inference peaks send the public cloud bill through the roof.

02

Data outside the perimeter

Regulated sectors that cannot process their information in just any jurisdiction.

03

Vendor dependence

Architectures tied to one cloud that are expensive to move.

WHAT IT IS

Our own datacentre plus elastic multi-cloud deployment

cinfo has virtualised its own AI datacentre in A Coruña's Ciudad de las TIC and runs its own elastic computing system to deploy and manage workloads on local servers or on any cloud provider.

- 3,328 high-grade AMD EPYC CPUs and 8 TB of RAM.
- 33 NVIDIA L4 GPUs with 24 GB and 6 NVIDIA A40 GPUs with 48 GB of video memory.
- 114 TB of disk plus 500 TB more on low-latency NAS.
- Automated hybrid deployment on AWS, Azure, Google Cloud, Oracle, Akamai or on-premise.



HOW IT WORKS

BERT: find resources and deploy where it makes sense

cinfo built BERT, its cloud resource finder and elastic deployment manager, which locates capacity and places the workload where it works out best.

01

The workload is sized

Analysis of compute, memory, storage and latency requirements.

02

Resources are located

BERT finds available capacity in our own datacentre and in the contracted clouds.

03

It is deployed in containers

Containerised deployment on Docker Swarm or Kubernetes clusters.

04

It is operated and observed

High availability, lifecycle management and observability dashboards.

CAPABILITIES

Infrastructure built for AI and video workloads

- **Our own datacentre**

  Facilities in A Coruña's Ciudad de las TIC, with the team's offices in the same building.

- **GPUs for training**

  39 NVIDIA GPUs dedicated to model development, training and inference.

- **Low-latency storage**

  114 TB on disk and 500 TB on low-latency NAS for large datasets.

- **Multi-cloud deployment**

  Workloads placed indifferently on public cloud, private cloud or physical servers.

- **Containers and orchestration**

  Docker Swarm and Kubernetes with high availability and lifecycle management.

- **Automated CI/CD**

  From git through integration, test, staging and production.

- **Observability**

  Monitoring with analytics dashboards and alert triggers.

- **Compliance**

  Cybersecurity compliance as part of the development and operations cycle.

BENEFITS

Capacity, cost control and data sovereignty

01

Data on home soil

Processing can stay entirely inside the A Coruña datacentre.

02

Cost under control

Our own capacity for the baseline, public cloud only for the peaks.

03

No single-vendor dependence

The deployment system is independent of the cloud provider.

04

Real GPU experience

Infrastructure used daily for cinfo's own video and AI products.

SCOPE

Ways of contracting the infrastructure

The service ranges from architecture consulting to fully operating the client's platform.

Architecture consulting

Sizing, cloud selection and deployment plan design.

Compute capacity

Use of cinfo's GPU infrastructure for training and inference.

Managed deployment

Platform go-live in containers, with high availability.

Continuous operations

Monitoring, alerting, updates and lifecycle management.

Hosted models

Open-source LLMs served from the cinfo cloud for your applications.

Migration

Moving existing workloads to a hybrid, multi-cloud scheme.

METHODOLOGY

From sizing to operations

01

Assessment

Analysis of current workloads, costs and regulatory requirements.

02

Architecture

Design of the hybrid scheme and the workload split between our cloud and public clouds.

03

Deployment

Containerisation, orchestration and continuous integration pipelines.

04

Operations

Observability, high availability and evolution of the system in production.

cinfo

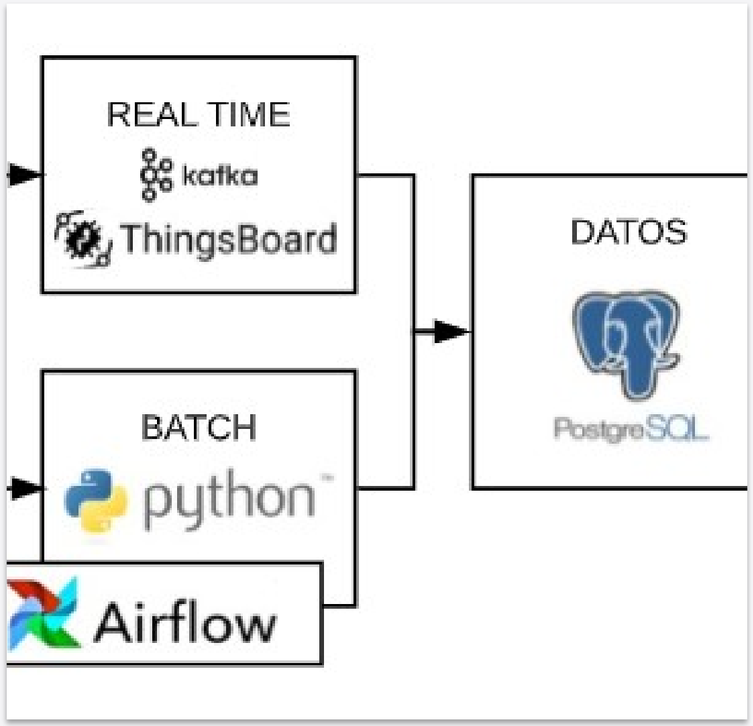
The infrastructure described is cinfo's own datacentre in A Coruña's Ciudad de las TIC. Sizing for each project is defined in the architecture phase.

GALLERY

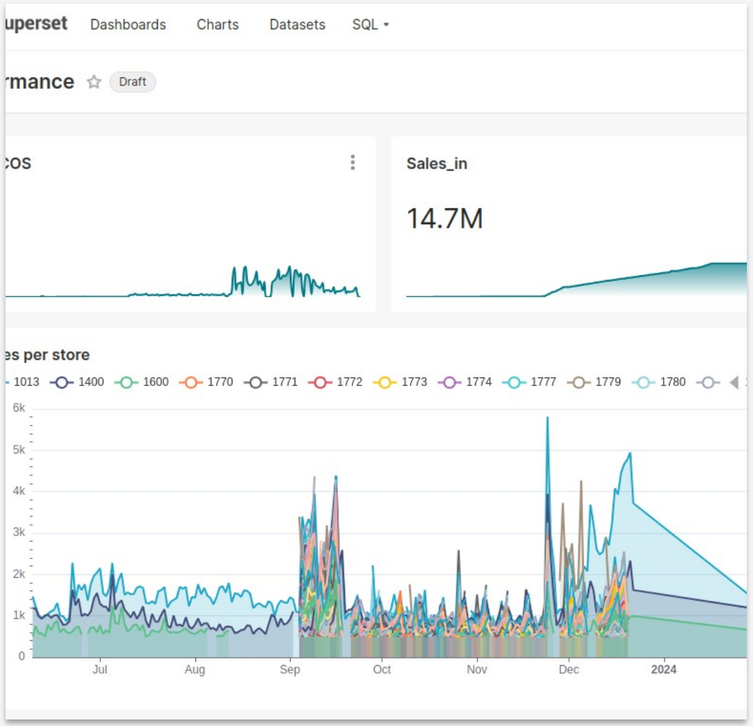
Infrastructure and deployments



The service architecture of a platform in production.



The data architecture of a multi-cloud project.



Operations and observability dashboards.

CASE STUDIES

Infrastructure behind real services

RTVE PLAY

Over 7 million videos served a day

The AI preview service runs on cinfo infrastructure with global CDN delivery and peaks of millions of daily requests.

https://www.rtve.es/play/

TIIVII

750 events produced every week

Automatic production, the mixer and the OTT platforms run on cinfo's elastic infrastructure, with peaks concentrated at weekends.

https://cinfo.es

TIIVII AI

LLMs served from Galicia

cinfo's AI platform serves open-source models from its own datacentre, guaranteeing the client's data never leaves the agreed environment.

https://cinfo.es

CINFO IN NUMBERS

The datacentre in numbers

3.328

high-grade AMD EPYC CPUs

39

NVIDIA L4 and A40 GPUs

8 TB

of available RAM

614 TB

of disk and NAS storage

FREQUENTLY ASKED QUESTIONS

Questions on infrastructure and data sovereignty

Where is the infrastructure physically?

In cinfo's own datacentre in A Coruña's Ciudad de las TIC, where the technical team's offices are as well.

What about demand peaks?

The usual scheme covers the baseline with our own capacity and absorbs peaks in public cloud, which keeps cost under control.

Can I combine your cloud with mine?

Yes, that is the usual scenario: the elastic deployment system places workloads in our datacentre, in public cloud or on the client's servers alike.

Who operates the platform?

Either cinfo as a managed service, or the client's team following knowledge transfer.

Is it for training or only inference?

For both. The GPUs are used daily for development, training and inference of cinfo's own video and AI models.

Does it meet cybersecurity requirements?

Cybersecurity compliance is part of cinfo's development and operations cycle, and is documented for each project.

WHO IT IS FOR

Who needs this capacity

01

Companies with AI projects

Training and inference without depending on one public cloud.

02

Regulated sectors

Cases where the data's location is a project condition.

03

Public administrations

Digital services with sovereignty and traceability requirements.

04

Video platforms

Workloads with concentrated peaks and heavy transcoding needs.

LET'S TALK

Let's size your workload

We analyse what you need to train or serve and propose the split between our cloud and public clouds.

Request a sizing

Explore the datacentre

CONTACT

Email

contact@cinfo.es

Phone / WhatsApp

+34 881 912 538

Web

https://contact.cinfo.es