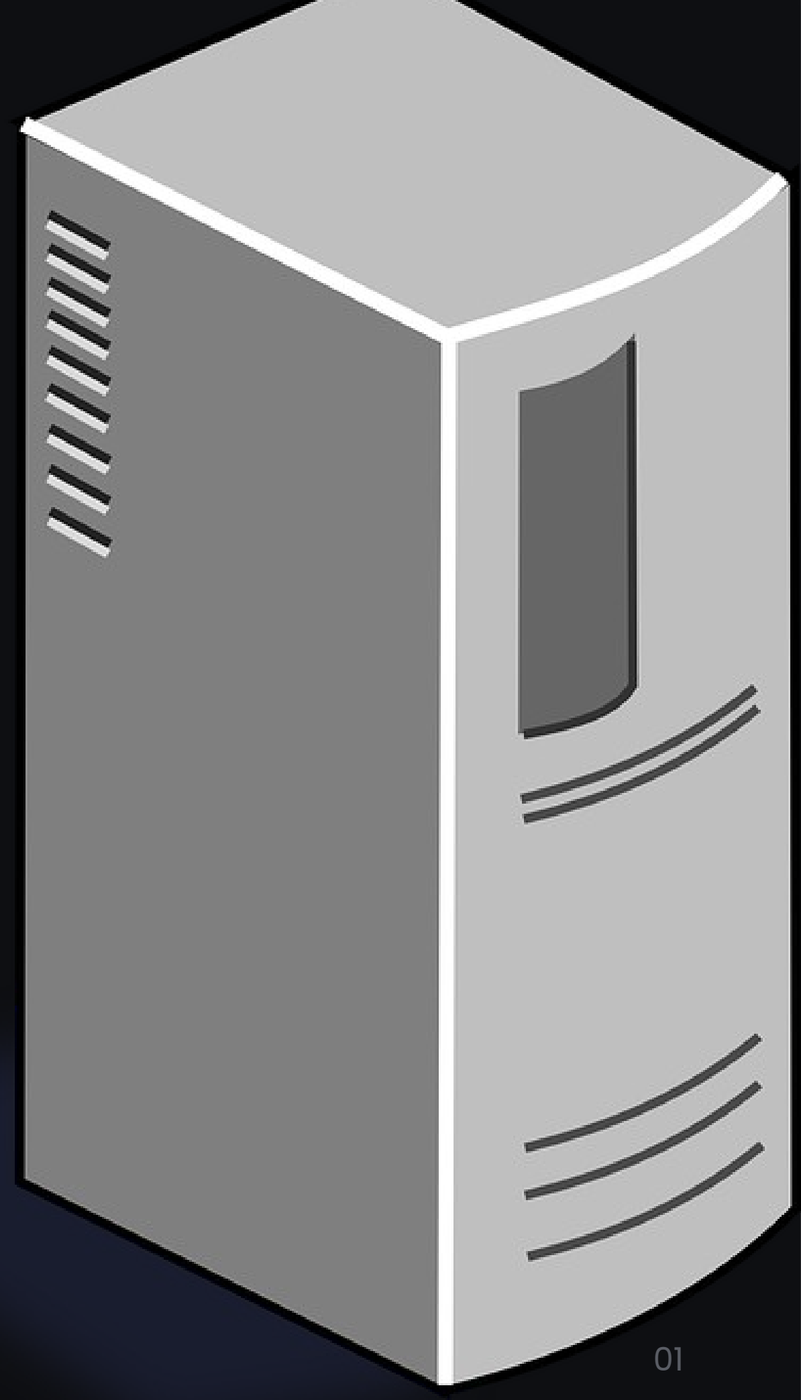
SERVICIO · INFRAESTRUCTURA

Cloud GPU e infraestructura IA

Capacidad de cómputo para IA con los datos alojados en territorio propio.

Datacenter propio en la Ciudad de las TIC de A Coruña y despliegue elástico en múltiples nubes.



cinfo

EL RETO

Entrenar y servir modelos sin depender de una única nube

La IA exige GPU, y la GPU exige decisiones: dónde vive el dato, cuánto cuesta el pico de demanda y qué pasa cuando el proveedor cambia precios o condiciones.

01

Coste imprevisible

Los picos de entrenamiento e inferencia disparan la factura de la nube pública.

02

Datos fuera del perímetro

Sectores regulados que no pueden procesar su información en cualquier jurisdicción.

03

Dependencia de proveedor

Arquitecturas atadas a una única nube que son costosas de mover.

QUÉ ES

Datacenter propio más despliegue elástico multinube

cinfo ha virtualizado su propio datacenter para IA en la Ciudad de las TIC de A Coruña y dispone de un sistema de computación elástica propio para desplegar y gestionar cargas de trabajo en servidores locales o en cualquier proveedor de nube.

- 3.328 CPU de alta gama AMD EPYC y 8 TB de memoria RAM.
- 33 GPU NVIDIA L4 de 24 GB y 6 GPU NVIDIA A40 de 48 GB de memoria de vídeo.
- 114 TB de disco y 500 TB adicionales en NAS de baja latencia.
- Despliegue híbrido automatizado en AWS, Azure, Google Cloud, Oracle, Akamai u on-premise.



CÓMO FUNCIONA

BERT: buscar recursos y desplegar donde convenga

cinfo ha desarrollado BERT, su buscador de recursos en la nube y gestor de despliegue elástico, que localiza capacidad y coloca la carga donde sale mejor.

01

Se dimensiona la carga

Análisis de los requisitos de cómputo, memoria, almacenamiento y latencia.

02

Se localizan recursos

BERT busca capacidad disponible en el datacenter propio y en las nubes contratadas.

03

Se despliega en contenedores

Encapsulación con despliegue en clústeres Docker Swarm o Kubernetes.

04

Se opera y se observa

Alta disponibilidad, gestión del ciclo de vida y cuadros de mando de observabilidad.

CAPACIDADES

Infraestructura pensada para cargas de IA y vídeo

- **Datacenter propio**

  Instalaciones en la Ciudad de las TIC de A Coruña, con oficinas del equipo en el mismo edificio.

- **GPU para entrenamiento**

  39 GPU NVIDIA dedicadas a desarrollo, entrenamiento e inferencia de modelos.

- **Almacenamiento de baja latencia**

  114 TB en disco y 500 TB en NAS de baja latencia para conjuntos de datos grandes.

- **Despliegue multinube**

  Cargas colocadas indistintamente en nube pública, privada o servidores físicos.

- **Contenedores y orquestación**

  Docker Swarm y Kubernetes con alta disponibilidad y gestión del ciclo de vida.

- **CI/CD automatizado**

  Desde git hasta integración, pruebas, preproducción y producción.

- **Observabilidad**

  Monitorización con cuadros de mando de análisis y disparadores de alerta.

- **Cumplimiento**

  Cumplimiento en ciberseguridad como parte del ciclo de desarrollo y explotación.

BENEFICIOS

Capacidad, control de costes y soberanía del dato

01

Datos en territorio propio

El procesamiento puede quedarse íntegramente en el datacenter de A Coruña.

02

Coste bajo control

Capacidad propia para la carga base y nube pública solo para los picos.

03

Sin dependencia de un proveedor

El sistema de despliegue es independiente del proveedor de nube.

04

Experiencia real en GPU

Infraestructura usada a diario para los propios productos de vídeo e IA de cinfo.

ALCANCE

Formas de contratar la infraestructura

El servicio va desde la consultoría de arquitectura hasta la operación completa de la plataforma del cliente.

Consultoría de arquitectura

Dimensionamiento, elección de nube y diseño del plan de despliegue.

Capacidad de cómputo

Uso de la infraestructura GPU de cinfo para entrenamiento e inferencia.

Despliegue gestionado

Puesta en marcha de la plataforma en contenedores, con alta disponibilidad.

Operación continua

Monitorización, alertas, actualizaciones y gestión del ciclo de vida.

Modelos alojados

LLM open source servidos desde la nube de cinfo para tus aplicaciones.

Migración

Traslado de cargas existentes a un esquema híbrido y multinube.

METODOLOGÍA

Del dimensionamiento a la operación

01

Diagnóstico

Análisis de las cargas actuales, de los costes y de los requisitos normativos.

02

Arquitectura

Diseño del esquema híbrido y del reparto de cargas entre nube propia y pública.

03

Despliegue

Contenerización, orquestación y canalización de integración continua.

04

Operación

Observabilidad, alta disponibilidad y evolución del sistema en producción.

cinfo

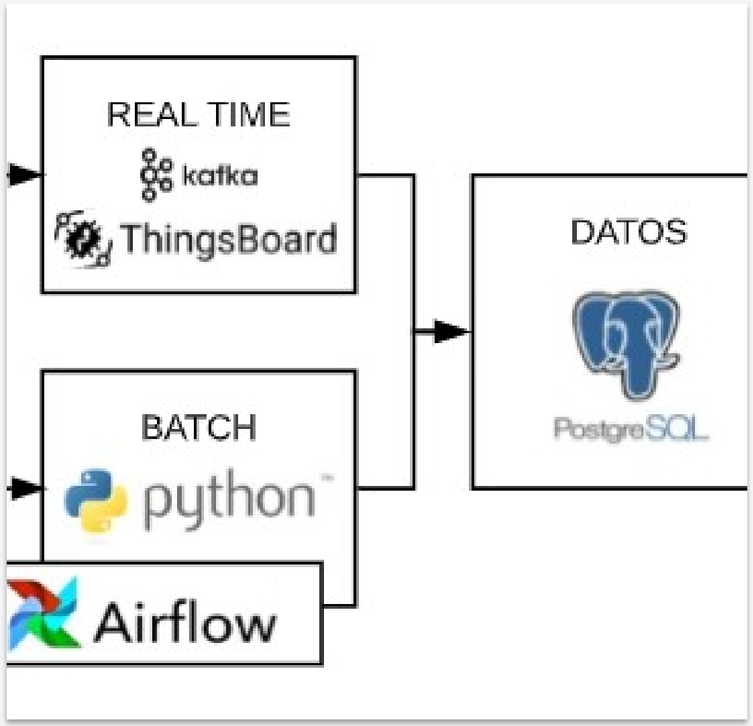
La infraestructura descrita corresponde al datacenter propio de cinfo en la Ciudad de las TIC de A Coruña. El dimensionamiento concreto de cada proyecto se define en la fase de arquitectura.

GALERÍA

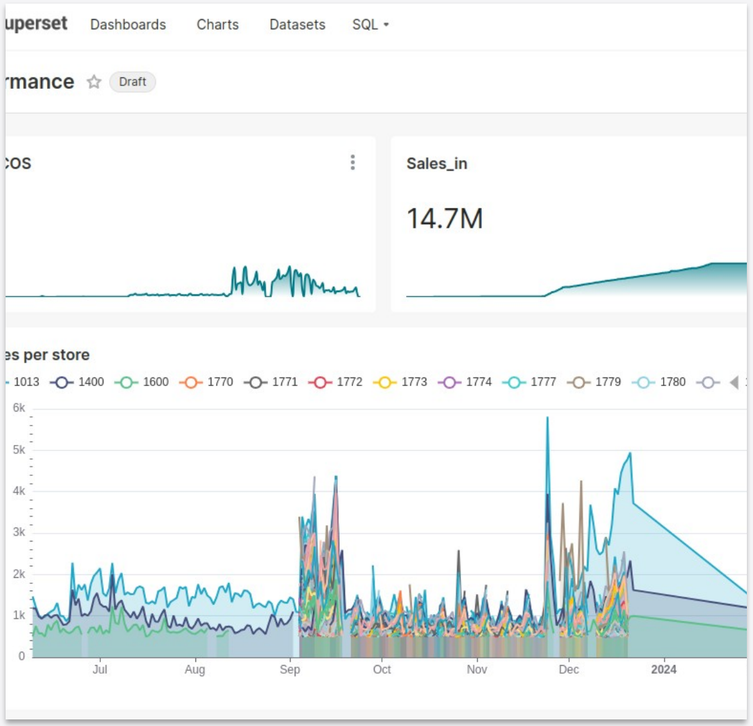
Infraestructura y despliegues



Arquitectura de servicios de una plataforma en producción.



Arquitectura de datos de un proyecto multinube.



Cuadros de mando de explotación y observabilidad.

CASOS DE ÉXITO

Infraestructura que sostiene servicios reales

RTVE PLAY

Más de 7 millones de vídeos servidos al día

El servicio de previsualizaciones por IA se presta desde la infraestructura de cinfo con distribución por CDN global y picos de millones de peticiones diarias.

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

TIIVII

750 eventos producidos cada semana

La producción automática, el mixer y las plataformas OTT se ejecutan sobre la infraestructura elástica de cinfo, con picos concentrados en fin de semana.

https://cinfo.es

TIIVII AI

Modelos LLM servidos desde Galicia

La plataforma de IA de cinfo sirve modelos open source desde el datacenter propio, garantizando que los datos del cliente no salen del entorno acordado.

https://cinfo.es

CINFO EN CIFRAS

El datacenter en números

3.328

CPU AMD EPYC de alta gama

39

GPU NVIDIA L4 y A40

8 TB

de memoria RAM disponible

614 TB

de almacenamiento en disco y NAS

PREGUNTAS FRECUENTES

Dudas sobre infraestructura y soberanía del dato

¿Dónde está físicamente la infraestructura?

En el datacenter propio de cinfo, en la Ciudad de las TIC de A Coruña, donde también están las oficinas del equipo técnico.

¿Puedo combinar vuestra nube con la mía?

Sí, es el escenario habitual: el sistema de despliegue elástico coloca cargas indistintamente en el datacenter propio, en nube pública o en servidores del cliente.

¿Sirve para entrenar modelos o solo para inferencia?

Para las dos cosas. Las GPU se usan a diario para desarrollo, entrenamiento e inferencia de los propios modelos de vídeo e IA de cinfo.

¿Qué pasa con los picos de demanda?

El esquema habitual es cubrir la carga base con capacidad propia y absorber los picos en nube pública, lo que mantiene el coste bajo control.

¿Quién opera la plataforma?

Puede operarla cinfo como servicio gestionado, o el equipo del cliente con la transferencia de conocimiento correspondiente.

¿Cumple los requisitos de ciberseguridad?

El cumplimiento en ciberseguridad forma parte del ciclo de desarrollo y explotación de cinfo, y se documenta para cada proyecto.

PARA QUIÉN

Quién necesita esta capacidad

01

Empresas con proyectos de IA

Entrenamiento e inferencia sin depender de una única nube pública.

02

Sectores regulados

Casos donde la ubicación del dato es una condición del proyecto.

03

Administraciones públicas

Servicios digitales con requisitos de soberanía y trazabilidad.

04

Plataformas de vídeo

Cargas con picos concentrados y necesidades de transcodificación intensiva.

HABLEMOS

Dimensionemos tu carga de trabajo

Analizamos qué necesitas entrenar o servir y te proponemos el reparto entre nube propia y pública.

Solicitar un dimensionamiento

Conocer el datacenter

CONTACTO

Email

contact@cinfo.es

Teléfono / WhatsApp

+34 881 912 538

Web

https://contact.cinfo.es