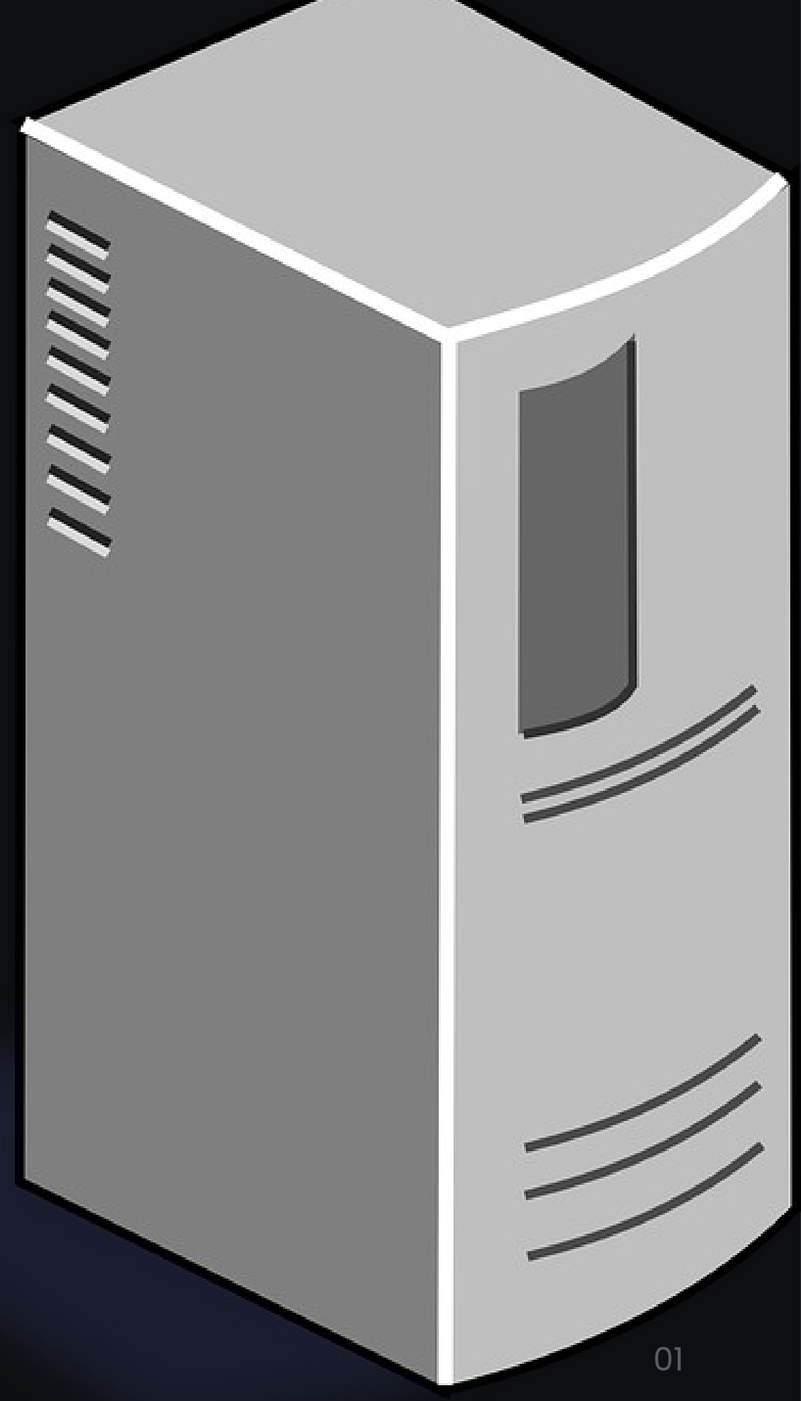
SERVIZO · INFRAESTRUTURA

Cloud GPU e infraestrutura IA

Capacidade de cómputo para IA cos datos aloxados en territorio propio.

Datacenter propio na Cidade das TIC da Coruña e despregamento elástico en múltiples nubes.



cinfo

O RETO

Adestrar e servir modelos sen depender dunha única nube

A IA esixe GPU, e a GPU esixe decisións: onde vive o dato, canto custa o pico de demanda e que pasa cando o provedor cambia prezos ou condicións.

01

Custo imprevisible

Os picos de adestramento e inferencia disparan a factura da nube pública.

02

Datos fóra do perímetro

Sectores regulados que non poden procesar a súa información en calquera xurisdición.

03

Dependencia de provedor

Arquitecturas atadas a unha única nube que son custosas de mover.

QUE É

Datacenter propio máis despregamento elástico multinube

cinfo virtualizou o seu propio datacenter para IA na Cidade das TIC da Coruña e dispón dun sistema de computación elástica propio para despregar e xestionar cargas de traballo en servidores locais ou en calquera provedor de nube.

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



COMO FUNCIONA

BERT: buscar recursos e despregar onde conveña

cinfo desenvolveu BERT, o seu buscador de recursos na nube e xestor de despregamento elástico, que localiza capacidade e coloca a carga onde sae mellor.

01

Dimensiónase a carga

Análise dos requisitos de cómputo, memoria, almacenamento e latencia.

02

Localízanse recursos

BERT busca capacidade dispoñible no datacenter propio e nas nubes contratadas.

03

Desprégase en contedores

Encapsulación con despregamento en clústeres Docker Swarm ou Kubernetes.

04

Opérase e obsérvase

Alta dispoñibilidade, xestión do ciclo de vida e cadros de mando de observabilidade.

CAPACIDADES

Infraestrutura pensada para cargas de IA e vídeo

- **Datacenter propio**

  Instalacións na Cidade das TIC da Coruña, con oficinas do equipo no mesmo edificio.

- **GPU para adestramento**

  39 GPU NVIDIA dedicadas a desenvolvemento, adestramento e inferencia de modelos.

- **Almacenamento de baixa latencia**

  114 TB en disco e 500 TB en NAS de baixa latencia para conxuntos de datos grandes.

- **Despregamento multinube**

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

- **Contedores e orquestración**

  Docker Swarm e Kubernetes con alta dispoñibilidade e xestión do ciclo de vida.

- **CI/CD automatizado**

  Desde git ata integración, probas, preprodución e produción.

- **Observabilidade**

  Monitorización con cadros de mando de análise e disparadores de alerta.

- **Cumprimento**

  Cumprimento en ciberseguridade como parte do ciclo de desenvolvemento e explotación.

BENEFICIOS

Capacidade, control de custos e soberanía do dato

01

Datos en territorio propio

O procesamento pode quedar integramente no datacenter da Coruña.

02

Custo baixo control

Capacidade propia para a carga base e nube pública só para os picos.

03

Sen dependencia dun provedor

O sistema de despregamento é independente do provedor de nube.

04

Experiencia real en GPU

Infraestrutura usada a diario para os propios produtos de vídeo e IA de cinfo.

ALCANCE

Formas de contratar a infraestrutura

O servizo vai desde a consultoría de arquitectura ata a operación completa da plataforma do cliente.

Consultoría de arquitectura

Dimensionamento, elección de nube e deseño do plan de despregamento.

Capacidade de cómputo

Uso da infraestrutura GPU de cinfo para adestramento e inferencia.

Despregamento xestionado

Posta en marcha da plataforma en contedores, con alta dispoñibilidade.

Operación continua

Monitorización, alertas, actualizacións e xestión do ciclo de vida.

Modelos aloxados

LLM open source servidos desde a nube de cinfo para as túas aplicacións.

Migración

Traslado de cargas existentes a un esquema híbrido e multinube.

METODOLOXÍA

Do dimensionamento á operación

01

Diagnóstico

Análise das cargas actuais, dos custos e dos requisitos normativos.

02

Arquitectura

Deseño do esquema híbrido e do reparto de cargas entre nube propia e pública.

03

Despregamento

Contedorización, orquestración e canalización de integración continua.

04

Operación

Observabilidade, alta dispoñibilidade e evolución do sistema en produción.

cinfo

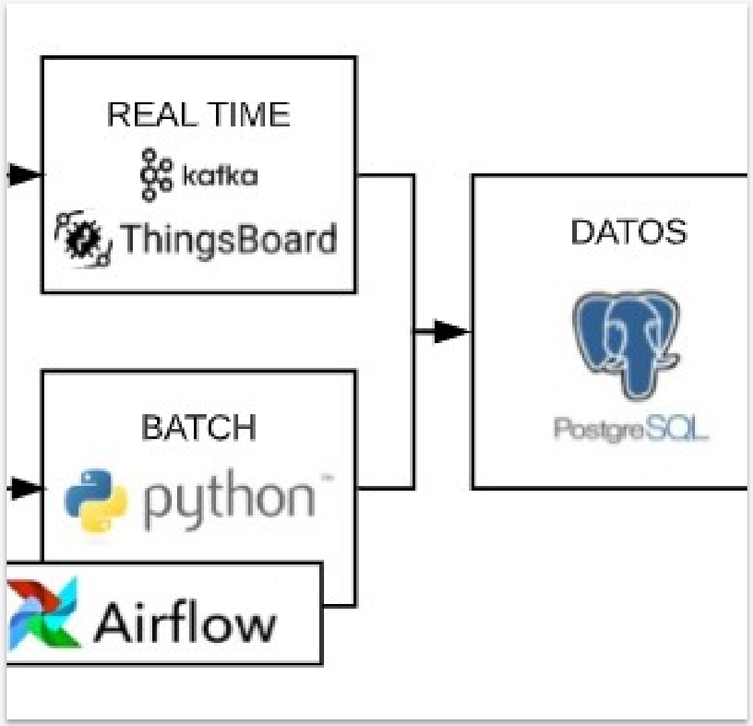
A infraestrutura descrita corresponde ao datacenter propio de cinfo na Cidade das TIC da Coruña. O dimensionamento concreto de cada proxecto defínese na fase de arquitectura.

GALERÍA

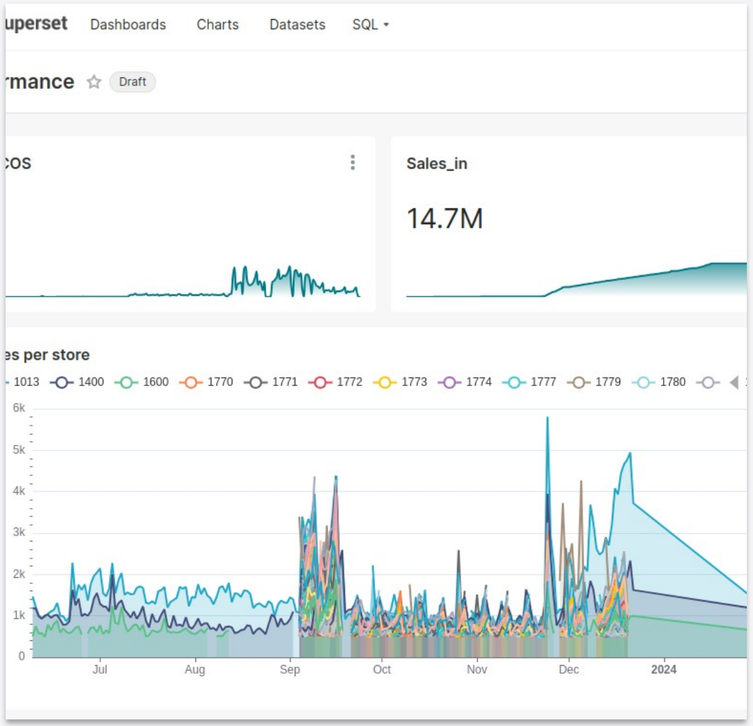
Infraestrutura e despregamentos



Arquitectura de servizos dunha plataforma en produción.



Arquitectura de datos dun proxecto multinube.



Cadros de mando de explotación e observabilidade.

CASOS DE ÉXITO

Infraestrutura que sostén servizos reais

RTVE PLAY

Máis de 7 millóns de vídeos servidos ao día

O servizo de previsualizacións por IA préstase desde a infraestrutura de cinfo con distribución por CDN global e picos de millóns de peticións diarias.

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

TIIVII

750 eventos producidos cada semana

A produción automática, o mixer e as plataformas OTT execútanse sobre a infraestrutura elástica de cinfo, con picos concentrados na fin de semana.

https://cinfo.es

TIIVII AI

Modelos LLM servidos desde Galicia

A plataforma de IA de cinfo serve modelos open source desde o datacenter propio, garantindo que os datos do cliente non saen do contorno acordado.

https://cinfo.es

CINFO EN CIFRAS

O datacenter en números

3.328

CPU AMD EPYC de alta gama

39

GPU NVIDIA L4 e A40

8 TB

de memoria RAM dispoñible

614 TB

de almacenamento en disco e NAS

PREGUNTAS FRECUENTES

Dúbidas sobre infraestrutura e soberanía do dato

Onde está fisicamente a infraestrutura?

No datacenter propio de cinfo, na Cidade das TIC da Coruña, onde tamén están as oficinas do equipo técnico.

Podo combinar a vosa nube coa miña?

Si, é o escenario habitual: o sistema de despregamento elástico coloca cargas indistintamente no datacenter propio, en nube pública ou en servidores do cliente.

Serve para adestrar modelos ou só para inferencia?

Para as dúas cousas. As GPU úsanse a diario para desenvolvemento, adestramento e inferencia dos propios modelos de vídeo e IA de cinfo.

Que pasa cos picos de demanda?

O esquema habitual é cubrir a carga base con capacidade propia e absorber os picos en nube pública, o que mantén o custo baixo control.

Quen opera a plataforma?

Pode operala cinfo como servizo xestionado, ou o equipo do cliente coa transferencia de coñecemento correspondente.

Cumpre os requisitos de ciberseguridade?

O cumprimento en ciberseguridade forma parte do ciclo de desenvolvemento e explotación de cinfo, e documéntase para cada proxecto.

PARA QUEN

Quen necesita esta capacidade

01

Empresas con proxectos de IA

Adestramento e inferencia sen depender dunha única nube pública.

02

Sectores regulados

Casos onde a localización do dato é unha condición do proxecto.

03

Administracións públicas

Servizos dixitais con requisitos de soberanía e trazabilidade.

04

Plataformas de vídeo

Cargas con picos concentrados e necesidades de transcodificación intensiva.

FALEMOS

Dimensionemos a túa carga de traballo

Analizamos que necesitas adestrar ou servir e propoñémosche o reparto entre nube propia e pública.

Solicitar un dimensionamento

Coñecer o datacenter

CONTACTO

Correo

contact@cinfo.es

Teléfono / WhatsApp

+34 881 912 538

Web

https://contact.cinfo.es