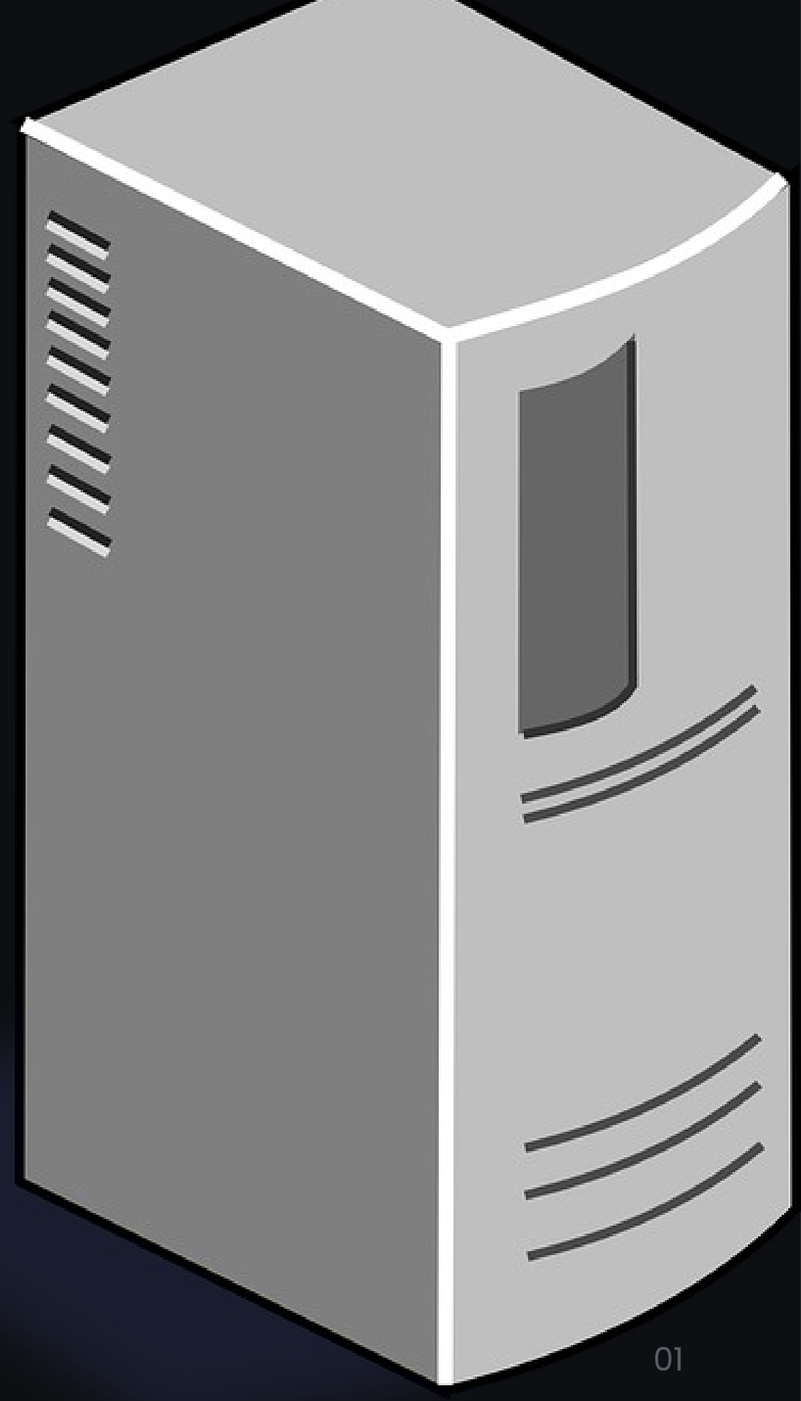
SERVEI · INFRAESTRUCTURA

Cloud GPU i infraestructura IA

Capacitat de còmput per a IA amb les dades allotjades en territori propi.

Datacenter propi a la Ciudad de las TIC de la Corunya i desplegament elàstic en múltiples núvols.



cinfo

EL REPTE

Entrenar i servir models sense dependre d'un únic núvol

La IA exigeix GPU, i la GPU exigeix decisions: on viu la dada, quant costa el pic de demanda i què passa quan el proveïdor canvia preus o condicions.

01

Cost imprevisible

Els pics d'entrenament i inferència disparen la factura del núvol públic.

02

Dades fora del perímetre

Sectors regulats que no poden processar la seva informació en qualsevol jurisdicció.

03

Dependència de proveïdor

Arquitectures lligades a un únic núvol que són costoses de moure.

QUÈ ÉS

Datacenter propi més desplegament elàstic multinúvol

cinfo ha virtualitzat el seu propi datacenter per a IA a la Ciudad de las TIC de la Corunya i disposa d'un sistema de computació elàstica propi per desplegar i gestionar càrregues de treball en servidors locals o en qualsevol proveïdor de núvol.

- 3.328 CPU d'alta gamma AMD EPYC i 8 TB de memòria RAM.
- 33 GPU NVIDIA L4 de 24 GB i 6 GPU NVIDIA A40 de 48 GB de memòria de vídeo.
- 114 TB de disc i 500 TB addicionals en NAS de baixa latència.
- Desplegament híbrid automatitzat a AWS, Azure, Google Cloud, Oracle, Akamai o on-premise.



COM FUNCIONA

BERT: buscar recursos i desplegar on convingui

cinfo ha desenvolupat BERT, el seu cercador de recursos al núvol i gestor de desplegament elàstic, que localitza capacitat i col·loca la càrrega on surt més a compte.

01

Es dimensiona la càrrega

Anàlisi dels requisits de còmput, memòria, emmagatzematge i latència.

02

Es localitzen recursos

BERT busca capacitat disponible al datacenter propi i als núvols contractats.

03

Es desplega en contenidors

Encapsulació amb desplegament en clústers Docker Swarm o Kubernetes.

04

S'opera i s'observa

Alta disponibilitat, gestió del cicle de vida i quadres de comandament d'observabilitat.

CAPACITATS

Infraestructura pensada per a càrregues d'IA i vídeo

- **Datacenter propi**

  Instal·lacions a la Ciudad de las TIC de la Corunya, amb oficines de l'equip al mateix edifici.

- **GPU per a entrenament**

  39 GPU NVIDIA dedicades a desenvolupament, entrenament i inferència de models.

- **Emmagatzematge de baixa latència**

  114 TB en disc i 500 TB en NAS de baixa latència per a conjunts de dades grans.

- **Desplegament multinúvol**

  Càrregues col·locades indistintament en núvol públic, privat o servidors físics.

- **Contenidors i orquestració**

  Docker Swarm i Kubernetes amb alta disponibilitat i gestió del cicle de vida.

- **CI/CD automatitzat**

  Des de git fins a integració, proves, preproducció i producció.

- **Observabilitat**

  Monitoratge amb quadres de comandament d'anàlisi i disparadors d'alerta.

- **Compliment**

  Compliment en ciberseguretat com a part del cicle de desenvolupament i explotació.

BENEFICIS

Capacitat, control de costos i sobirania de la dada

01

Dades en territori propi

El processament es pot quedar íntegrament al datacenter de la Corunya.

02

Cost sota control

Capacitat pròpia per a la càrrega base i núvol públic només per als pics.

03

Sense dependència d'un proveïdor

El sistema de desplegament és independent del proveïdor de núvol.

04

Experiència real en GPU

Infraestructura usada diàriament per als propis productes de vídeo i IA de cinfo.

ABAST

Formes de contractar la infraestructura

El servei va des de la consultoria d'arquitectura fins a l'operació completa de la plataforma del client.

Consultoria d'arquitectura

Dimensionament, elecció de núvol i disseny del pla de desplegament.

Capacitat de còmput

Ús de la infraestructura GPU de cinfo per a entrenament i inferència.

Desplegament gestionat

Posada en marxa de la plataforma en contenidors, amb alta disponibilitat.

Operació contínua

Monitoratge, alertes, actualitzacions i gestió del cicle de vida.

Models allotjats

LLM open source servits des del núvol de cinfo per a les teves aplicacions.

Migració

Trasllat de càrregues existents a un esquema híbrid i multinúvol.

METODOLOGIA

Del dimensionament a l'operació

01

Diagnòstic

Anàlisi de les càrregues actuals, dels costos i dels requisits normatius.

02

Arquitectura

Disseny de l'esquema híbrid i del repartiment de càrregues entre núvol propi i públic.

03

Desplegament

Contenerització, orquestració i canalització d'integració contínua.

04

Operació

Observabilitat, alta disponibilitat i evolució del sistema en producció.

cinfo

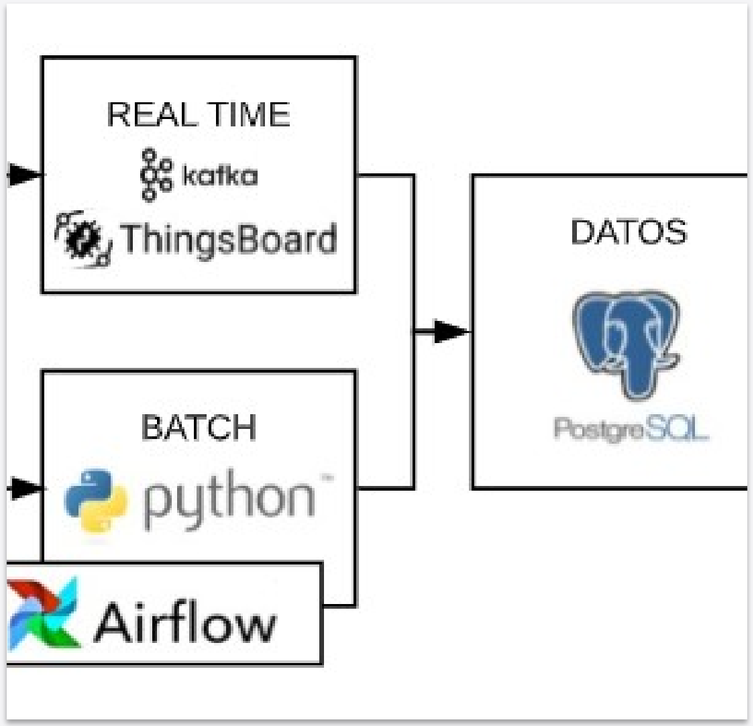
La infraestructura descrita correspon al datacenter propi de cinfo a la Ciudad de las TIC de la Corunya. El dimensionament concret de cada projecte es defineix a la fase d'arquitectura.

GALERIA

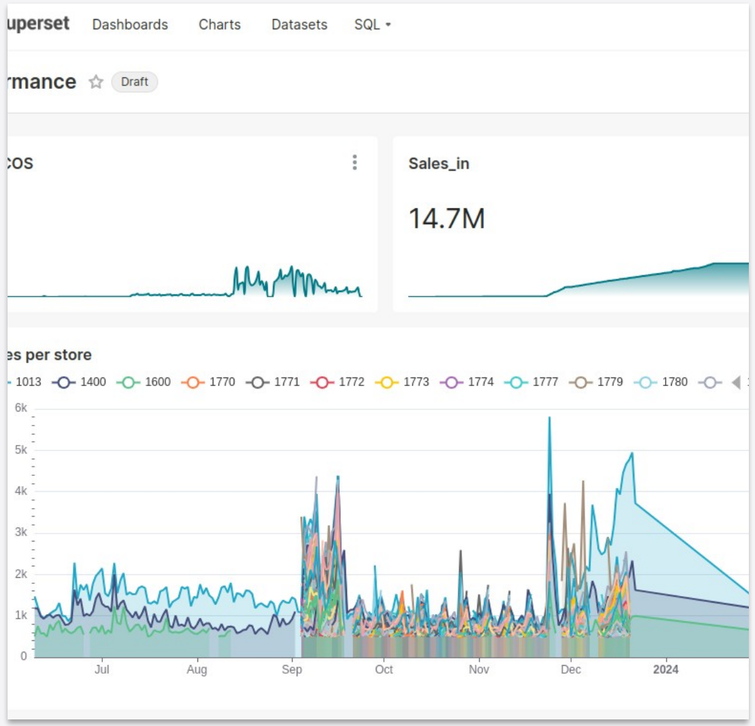
Infraestructura i desplegaments



Arquitectura de serveis d'una plataforma en producció.



Arquitectura de dades d'un projecte multinúvol.



Quadres de comandament d'explotació i observabilitat.

CASOS D'ÈXIT

Infraestructura que sosté serveis reals

RTVE PLAY

Més de 7 milions de vídeos servits al dia

El servei de previsualitzacions per IA es presta des de la infraestructura de cinfo amb distribució per CDN global i pics de milions de peticions diàries.

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

TIIVII

750 esdeveniments produïts cada setmana

La producció automàtica, el mixer i les plataformes OTT s'executen sobre la infraestructura elàstica de cinfo, amb pics concentrats en cap de setmana.

https://cinfo.es

TIIVII AI

Models LLM servits des de Galícia

La plataforma d'IA de cinfo serveix models open source des del datacenter propi, garantint que les dades del client no surten de l'entorn acordat.

https://cinfo.es

CINFO EN XIFRES

El datacenter en números

3.328

CPU AMD EPYC d'alta gamma

39

GPU NVIDIA L4 i A40

8 TB

de memòria RAM disponible

614 TB

d'emmagatzematge en disc i NAS

PREGUNTES FREQÜENTS

Dubtes sobre infraestructura i sobirania de la dada

On és físicament la infraestructura?

Al datacenter propi de cinfo, a la Ciudad de las TIC de la Corunya, on també hi ha les oficines de l'equip tècnic.

Què passa amb els pics de demanda?

L'esquema habitual és cobrir la càrrega base amb capacitat pròpia i absorbir els pics en núvol públic, cosa que manté el cost sota control.

Puc combinar el vostre núvol amb el meu?

Sí, és l'escenari habitual: el sistema de desplegament elàstic col·loca càrregues indistintament al datacenter propi, en núvol públic o en servidors del client.

Qui opera la plataforma?

Pot operar-la cinfo com a servei gestionat, o l'equip del client amb la transferència de coneixement corresponent.

Serveix per entrenar models o només per a inferència?

Per a totes dues coses. Les GPU es fan servir diàriament per al desenvolupament, l'entrenament i la inferència dels propis models de vídeo i IA de cinfo.

Compleix els requisits de ciberseguretat?

El compliment en ciberseguretat forma part del cicle de desenvolupament i explotació de cinfo, i es documenta per a cada projecte.

PER A QUI

Qui necessita aquesta capacitat

01

Empreses amb projectes d'IA

Entrenament i inferència sense dependre d'un únic núvol públic.

02

Sectors regulats

Casos on la ubicació de la dada és una condició del projecte.

03

Administracions públiques

Serveis digitals amb requisits de sobirania i traçabilitat.

04

Plataformes de vídeo

Càrregues amb pics concentrats i necessitats de transcodificació intensiva.

PARLEM-NE

Dimensionem la teva càrrega de treball

Analitzem què necessites entrenar o servir i et proposem el repartiment entre núvol propi i públic.

Sol·licitar un dimensionament

Conèixer el datacenter

CONTACTE

Correu

contact@cinfo.es

Telèfon / WhatsApp

+34 881 912 538

Web

https://contact.cinfo.es