



Cloud Computing para Inteligencia Artificial

Sesión 1: La Sinergia Fundamental

Departamento de Tecnología Informática y Computación

Grado en Ingeniería en Inteligencia Artificial, 2025-2026

Agenda de la Sesión de Hoy



01

Fundamentos del Cloud

¿Qué es realmente y de dónde viene? Los 5
Los 5 principios NIST.

02

Modelos de Servicio

El espectro de control: IaaS, PaaS y SaaS.

03

Modelos de Despliegue

¿Dónde y cómo se despliega la nube?
(Público, Privado, Híbrido, etc.).

04

Sinergia con la IA

El núcleo de la asignatura. ¿Por qué se necesitan mutuamente?
mutuamente?

05

El Ecosistema

Conociendo a los principales actores y sus consolas.

Cloud Computing: Definición y Evolución

"Un modelo para habilitar el acceso de red ubicuo, conveniente y bajo demanda a un conjunto compartido de recursos informáticos configurables que se pueden aprovisionar y liberar rápidamente con un mínimo esfuerzo de gestión o interacción con el proveedor de servicios".

— National Institute of Standards and Technology (NIST)

Evolución Histórica

J.C.R. Licklider (ARPANET, 1960s)

Impulsor del concepto de *Intergalactic Computer Network*, precursor de Internet. Sentó las bases de la idea de compartir recursos de cómputo a escala global.

Años 90: Virtualización

VMware permite abstraer hardware del software, ejecutando múltiples "máquinas" en un solo servidor físico.

2006: Nacimiento de AWS

Amazon crea AWS para solucionar sus problemas de escalabilidad, lanzando S3 y EC2 y creando el mercado moderno de IaaS.

Microsoft Azure (Ray Ozzie, 2008)

0

Padre del Cloud

John McCarthy (MIT, 1961)
Fue pionero en IA y ya en 1961 predijo que *"la computación algún día podría organizarse como un servicio público"*, de forma similar a la electricidad o el agua.

1

2

Años 50-70: Mainframes IBM

Computación centralizada y compartida. El germen de la idea de idea de "recursos compartidos".

3

4

Salesforce (Marc Benioff, 1999)

Considerado el primer caso comercial de *cloud SaaS*, introdujo el modelo de *"software as a service"* con CRM accesible desde un navegador

5

6

Google (Eric Smidt y Urs Hölzle, 2006-2008)

Google App Engine (2008) introdujo un modelo **PaaS** que permitía desplegar aplicaciones sin preocuparse de la infraestructura.

7



Los 5 Principios Fundamentales (NIST)

Estos cinco puntos no son opcionales. Si un servicio no cumple con los cinco, estrictamente hablando, no es "Cloud Computing".

1

Autoservicio Bajo Demanda

El usuario aprovisiona recursos sin intervención humana.

2

Amplio Acceso a la Red

Los servicios están disponibles sobre la red a red a través de mecanismos estándar (APIs, (APIs, web).

3

Agrupación de Recursos

El proveedor agrupa sus recursos para servir a servir a múltiples clientes (modelo *multi-multi-tenant*).

4

Rápida Elasticidad

Capacidad de escalar recursos hacia arriba y y hacia abajo, de forma rápida y, a menudo, menudo, automática.

5

Servicio Medido

El uso de recursos se mide, controla y reporta. Pagas por lo que usas.

6

Resiliencia

Garantizar la disponibilidad de recursos y la tolerancia a fallos 24x7 a través de la redundancia. redundancia.

Principios 1 & 2: Acceso y Autoservicio

Autoservicio Bajo Demanda

Antes:

Necesitabas un servidor. Abrías un ticket, esperabas a que el equipo de IT lo comprara, lo instalara en el rack, lo cableara y le instalara un SO. Proceso de semanas o meses.

Ahora:

Entras a la consola de AWS, haces unos clics, y en menos de 2 minutos tienes una Máquina Virtual funcionando. Esto cambia radicalmente la velocidad de innovación.

Amplio Acceso a la Red

Los recursos se gestionan a través de protocolos estándar (HTTP/S).

Puedes hacerlo desde:

- La consola web
- Interfaz de Línea de Comandos (CLI) en tu terminal
- Programáticamente usando un Kit de Desarrollo de Software Software (SDK) en Python, Java, etc.

Esto permite la **automatización** (Infraestructura como Código).

Principios 3 & 4: Elasticidad y Recursos Compartidos

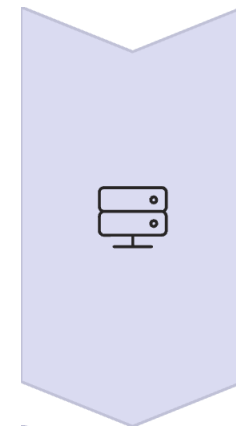
Agrupación de Recursos (Multi-tenancy)

Tus datos y VMs comparten la misma infraestructura física (servidores, (servidores, redes) que los de otros clientes, pero están lógicamente lógicamente aislados y seguros.

Es como un edificio de apartamentos: compartes la estructura, pero tienes tu propio piso con tu propia llave.

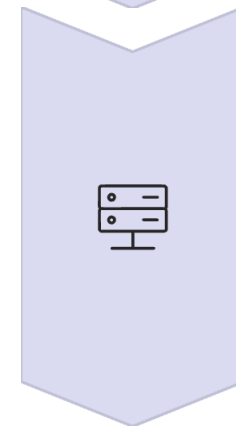
Esto permite economías de escala masivas para el proveedor, lo que se traduce en precios más bajos para ti.

Rápida Elasticidad



Escalado Vertical (Scale Up)

Aumentar la potencia de una instancia (más CPU, más RAM). Ej: Cambiar una instancia t2.micro a una t2.large en AWS.



Escalado Horizontal (Scale Out)

Añadir más instancias. Ej: Pasar de 2 servidores web a 10 a 10 durante un pico de tráfico. El cloud brilla aquí, especialmente con el auto-scaling.

Principio 5: El Modelo de "Utility Computing"

Servicio Medido (Pay-as-you-go)

Es el cambio de paradigma más importante a nivel financiero.

Como la electricidad. No construyes tu propia central eléctrica. Te conectas a la red, usas la electricidad que necesitas y pagas solo por los kilovatios-hora consumidos.

Ejemplos en Cloud:

- **AWS EC2 (Cómputo):** Pagas por segundo/hora que la instancia está encendida.
- **AWS S3 (Almacenamiento):** Pagas por Gigabyte almacenado al mes.
- **AWS Lambda (Serverless):** Pagas por milisegundo de ejecución y número de peticiones.



Principio 6: Confiabilidad

Resiliencia

La capacidad de un sistema Cloud para resistir fallos, adaptarse a interrupciones y recuperarse rápidamente, manteniendo la disponibilidad y continuidad del servicio.

Un sistema resiliente no es aquel que nunca falla, sino aquel que sigue funcionando o se recupera sin impacto crítico para el usuario.

Redundancia de IT en la misma nube o diferentes nubes pero en diferentes localizaciones físicas.

Tolerancia a Fallos

Redundancia

Recuperación Automática

Disponibilidad Continua

De CAPEX a OPEX: Un Cambio Fundamental

CAPEX (Capital Expenditure) Expenditure)

Gasto de capital. La compra de activos físicos (servidores, routers, edificios).

Es una gran inversión inicial, con riesgo de:

- Sobreaprovisionar (comprar de más "por si acaso")
- Infraaprovisionar (quedarse corto)

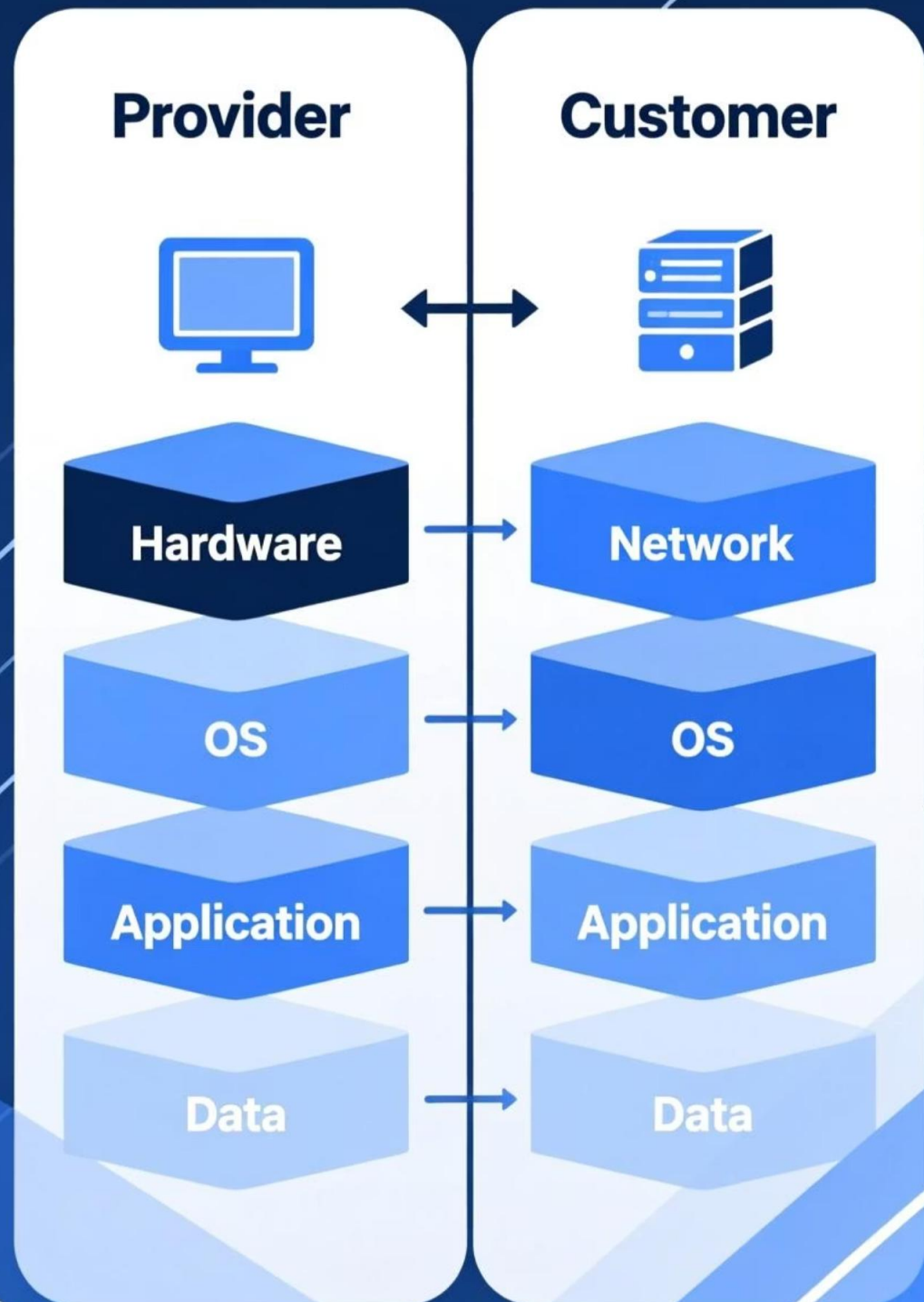
OPEX (Operational Expenditure) Expenditure)

Gasto operativo. El gasto del día a día (electricidad, salarios, y... ¡servicios en la nube!).

El Impacto: El cloud convierte el enorme enorme gasto de CAPEX de montar un centro de datos en un gasto de OPEX predecible y variable.

Esto democratiza el acceso a la infraestructura de nivel mundial. Una startup puede usar la misma tecnología que Netflix desde el primer primer día.





El Modelo de Responsabilidad Compartida

La seguridad y la gestión en la nube son una responsabilidad compartida entre el proveedor y el cliente. La pregunta clave es siempre: "**¿Quién es responsable de qué?**"

El Proveedor (AWS, Azure, GCP)

Es responsable de la seguridad **DE** la nube.

- Infraestructura física (seguridad de los los datacenters)
- Red troncal
- Hardware
- Hipervisor de virtualización

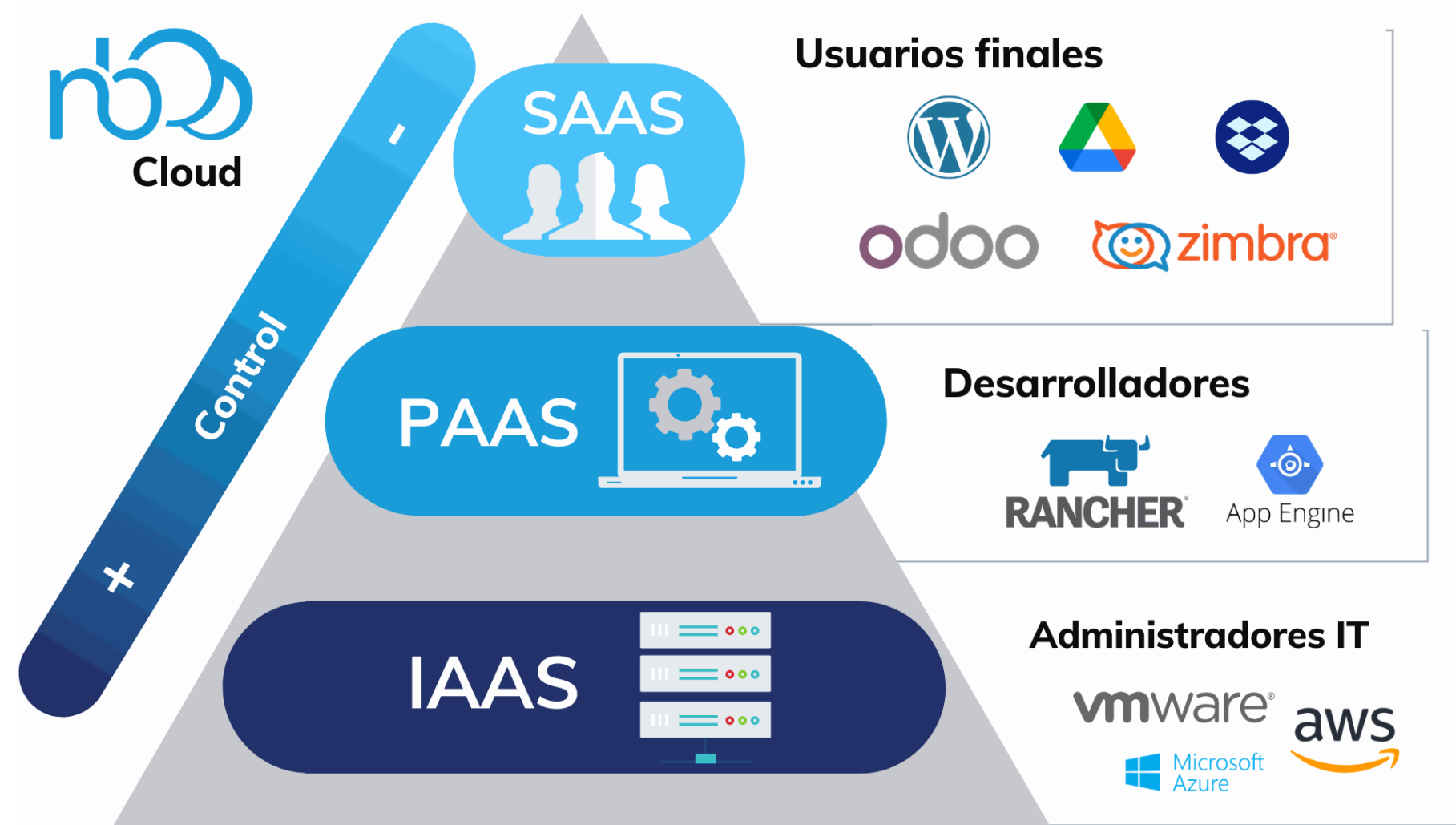
El Cliente (Tú)

Es responsable de la seguridad **EN** la nube.

- Tus datos
- Configuración del sistema operativo
- Gestión de usuarios
- Configuración de firewalls (Grupos de Seguridad)

La línea de responsabilidad cambia según el modelo de servicio (IaaS, PaaS, SaaS).

Modelos de Aprovisionamiento de Servicios Cloud



Software As a Service

Platform As a Service

Infrastructure As a Service

Fuente: Nanobytes

IaaS: Infrastructure as a Service

Definición

Los bloques de construcción fundamentales. Te da el máximo control y flexibilidad, similar a la infraestructura on-premise. **Alquilas el hardware.**

Componentes



Cómputo

Máquinas Virtuales (VMs) donde ejecutas tus aplicaciones y servicios.



Almacenamiento

Almacenamiento en bloque (discos virtuales) y de objetos para tus datos.



Redes

Redes privadas virtuales, balanceadores de carga, firewalls.

Tu Responsabilidad

Eres responsable de instalar y gestionar:

- Sistema Operativo
- Middleware
- Runtime (ej. Python, Java)
- Librerías
- Tu código
- Tus datos

Ideal para:

Migraciones "lift-and-shift", aplicaciones que requieren control granular, cargas de trabajo con software muy específico o licencias complejas.



IaaS en la Práctica: Los Nombres a Conocer

Los conceptos son los mismos en todos los proveedores, solo cambian los nombres comerciales.

AWS

Cómputo: Amazon EC2 (Elastic Compute Cloud)

Almacenamiento en Bloque: EBS (Elastic Block Store)

Redes: VPC (Virtual Private Cloud)

Azure

Cómputo: Azure Virtual Machines

Almacenamiento en Bloque: Azure Disk Storage

Redes: Azure Virtual Network

Google Cloud (GCP)

Cómputo: Compute Engine

Almacenamiento en Bloque: Persistent Disk

Redes: VPC Network



PaaS: Platform as a Service

Definición

Abstrae la infraestructura. El proveedor gestiona el hardware y el sistema operativo. Tú solo te preocupas de tu código y tus datos.

"No te preocupes por los servidores, solo sube tu código."

Tu Responsabilidad

Gestionas tu aplicación y tus datos. El proveedor se encarga de los parches del SO, el escalado de la infraestructura, etc.

Tipos de PaaS

- **aPaaS (Application PaaS):** Para desplegar aplicaciones completas
- **dbPaaS (Database PaaS):** Bases de datos gestionadas (ej. Amazon RDS)
- **FaaS (Functions as a Service / Serverless):** Funciones individuales (ej. AWS Lambda)

Ideal para

Desarrolladores que quieren ser productivos rápidamente, desarrollo de APIs, aplicaciones web estándar, microservicios.

PaaS en la Práctica: Despliegue Simplificado



AWS

aPaaS: AWS Elastic Beanstalk. Subes tu código (Java, Node.js, Python, etc.) y él se encarga de provisionar el balanceador, las instancias EC2, el auto-scaling, etc.

dbPaaS: Amazon RDS (Relational Database Service). Una base de datos MySQL, PostgreSQL, etc., totalmente gestionada.



Azure

aPaaS: Azure App Service.

dbPaaS: Azure SQL Database.



GCP

aPaaS: Google App Engine.

dbPaaS: Cloud SQL.



Elastic Beanslisch

upload

deploy

SaaS: Software as a Service

Definición

Un producto de software completo, listo para usar, al que accedes normalmente a través de un navegador web o una API. El modelo de consumo por excelencia.

"No construyas, consume."

Tu Responsabilidad

Prácticamente ninguna sobre la infraestructura. Solo usas y configuras el software.

Ejemplos

Productividad

- Office 365
- Google Workspace

CRM

- Salesforce

Colaboración

- Slack
- Trello

APIs de IA


- Amazon Rekognition (análisis de imagen)
- OpenAI API

¡Gamificación! Eres el Arquitecto Cloud

Para los siguientes escenarios, elige el modelo de servicio más adecuado: **IaaS**, **PaaS** o **SaaS**.


Escenario 1

"Un equipo de desarrollo móvil necesita una base de datos PostgreSQL. Quieren alta disponibilidad y backups automáticos, pero no tienen un administrador de bases de datos (DBA) en el equipo."

 **Respuesta correcta:** PaaS (dbPaaS como Amazon RDS o Azure Database for PostgreSQL)


Escenario 2

"Una empresa de efectos especiales necesita renderizar escenas 3D usando un software propietario muy específico que solo funciona en una distribución de Linux concreta con drivers de GPU personalizados."

 **Respuesta correcta:** IaaS (Necesitan control total sobre la instancia EC2/VM y el SO)

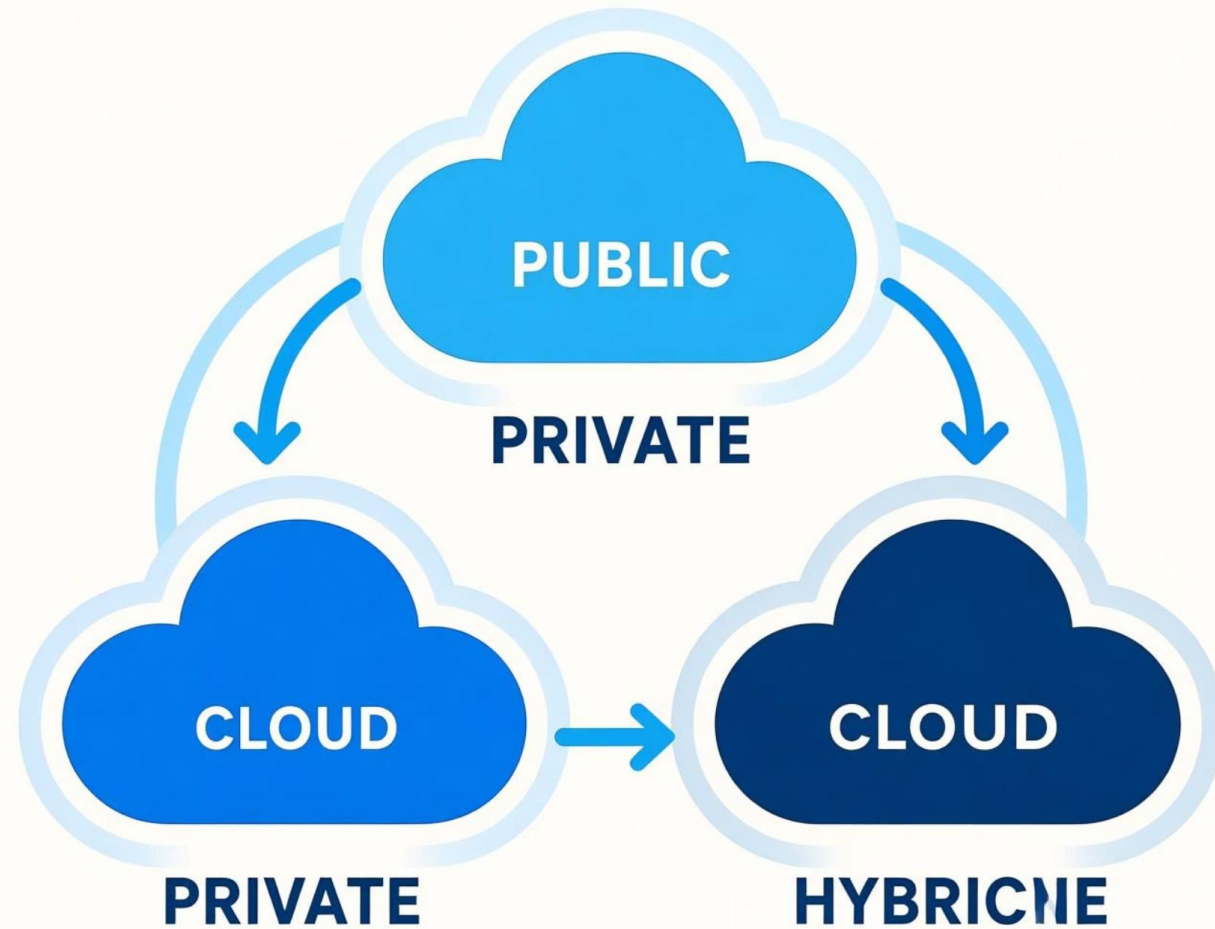
Escenario 3

"Un departamento de marketing quiere añadir a su web la funcionalidad de traducir comentarios de usuarios a varios idiomas en tiempo real, sin contratar a un equipo de NLP."

 **Respuesta correcta:** SaaS (Usar una API como Amazon Translate o Google Translate API)

Modelos de Despliegue: ¿Dónde Vive la Nube?

Esta decisión define quién es el propietario de la infraestructura y dónde reside físicamente.



Cloud Público

La infraestructura pertenece y es operada por un proveedor (AWS, Azure, GCP) y se comparte entre muchos clientes. Es el modelo más común.

Cloud Privado

La infraestructura es de uso exclusivo para una sola organización. Puede estar en su propio centro de datos (*on-premise*) o alojada por un tercero (*hosted private cloud*).

Cloud Híbrido

Un entorno que combina una nube pública y una nube privada, permitiendo que datos y aplicaciones se muevan entre ellas.



Híbrido vs. Multicloud: Una Distinción Crucial

Son términos que a menudo se confunden, pero significan cosas diferentes.

Cloud Híbrido

Definición:

Integra recursos *on-premise* (privados) con recursos de **un** proveedor de nube pública.

Objetivo:

- Extender la infraestructura existente
- Mantener datos sensibles en local (soberanía de datos)
- Recuperación ante desastres

Multicloud

Definición:

Utiliza los servicios de **múltiples** proveedores de nube *pública* (ej. usar AWS para cómputo y GCP para Machine Learning).

Objetivo:

- Evitar la dependencia de un solo proveedor proveedor (*vendor lock-in*)
- Optimizar costes eligiendo el mejor servicio/precio para cada tarea
- Mejorar la resiliencia



Importante: Una arquitectura puede ser ambas cosas a la vez (híbrida y multicloud).

Casos de Uso para Híbrido y Multicloud

Casos de Uso Híbrido

Cloud Bursting (Ráfaga de Nube)

Usar la nube pública para obtener capacidad de cómputo adicional adicional cuando la demanda supera la capacidad del centro de datos de datos privado.

Soberanía de Datos

Procesar datos en la nube pública, pero almacenar los datos brutos y sensibles en la nube privada para cumplir con regulaciones (como GDPR).

Recuperación ante Desastres (DR)

Replicar sistemas críticos del centro de datos privado en la nube pública para una rápida recuperación en caso de fallo.

Casos de Uso Multicloud

Mejor Servicio para la Tarea

Usar BigQuery de GCP para análisis de datos, los servicios de IA de IA de Azure, y la infraestructura global de EC2 de AWS, todo en la en la misma aplicación.

Negociación y Costes

Tener presencia en múltiples nubes da poder de negociación con los con los proveedores y permite optimizar costes.

Resiliencia Global

Distribución de cargas de trabajo críticas entre proveedores para para minimizar el impacto de fallos regionales.

Cloud Comunitario: El Modelo de Nicho

Definición

Un modelo menos común donde la infraestructura de la nube es compartida por varias organizaciones y apoya a una comunidad específica que tiene preocupaciones compartidas (por preocupaciones compartidas (por ejemplo, misión, requisitos de seguridad, política y consideraciones de cumplimiento)).

Ejemplos



Gobierno

Una nube compartida por diferentes agencias gubernamentales. AWS GovCloud o Azure Government son ejemplos comerciales de este concepto.



Investigación

Una nube para un consorcio de universidades para compartir datasets y potencia de cómputo.



Finanzas

Una nube para que varios bancos procesen transacciones cumpliendo con normativas financieras específicas.



Recap del Ciclo de Vida de IA/ML

Como expertos en IA, conocéis este ciclo. Ahora, vamos a pensar en qué parte del ciclo se beneficia (o es directamente imposible sin) el cloud.



Sinergia 1: La Sed Insaciable de Cómputo

El Problema

Entrenar un modelo de Deep Learning moderno (como un Transformer) puede requerir cientos de ExaFLOPs de computación. Hacer esto en local es inviable para casi todos.

La Solución Cloud

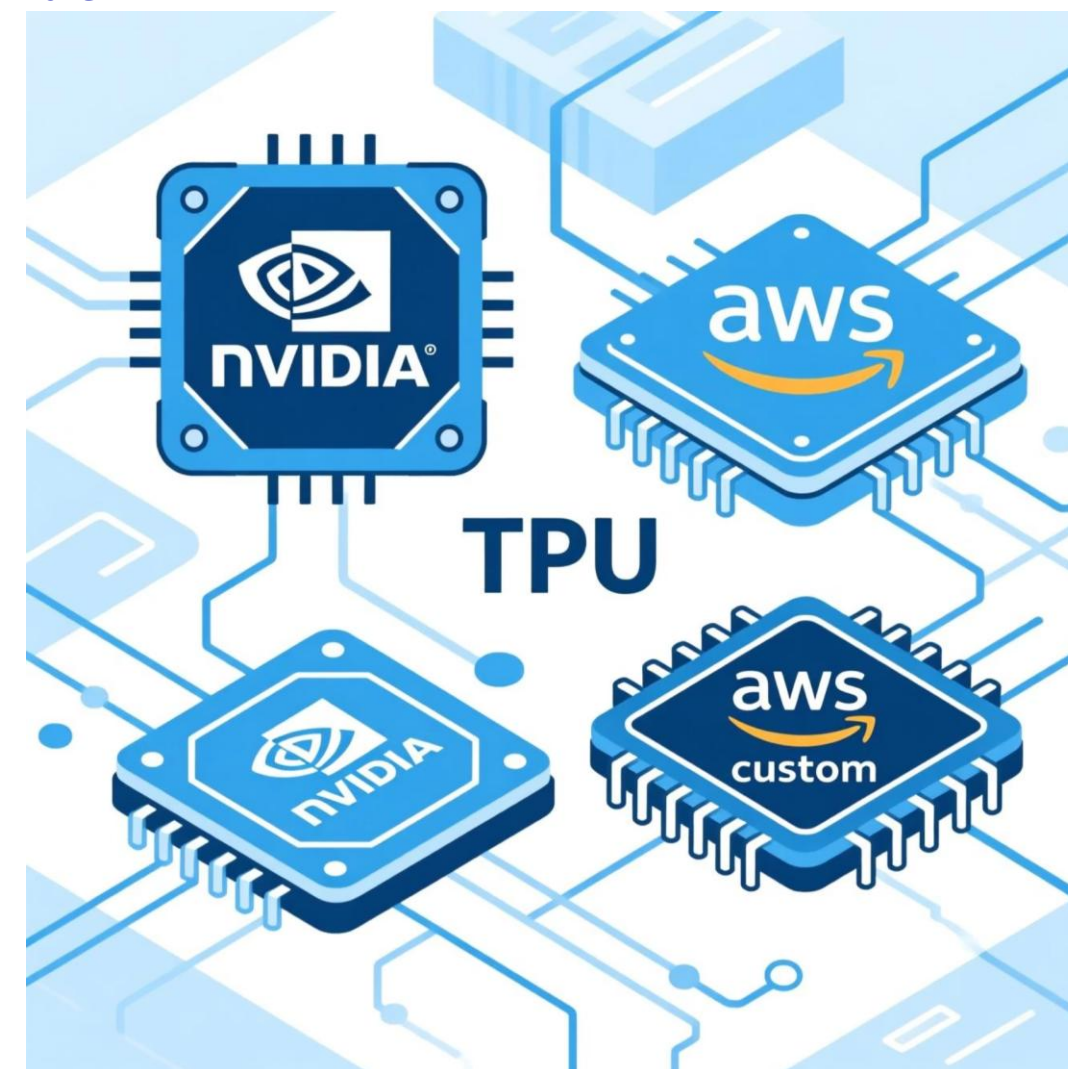
Acceso a Hardware Especializado

- **GPUs:** NVIDIA A100/H100. Disponibles en instancias como las P4d o P5 de [AWS EC2](#), o las ND-series de [Azure](#).
- **TPUs (Tensor Processing Units):** Hardware ASIC de Google, optimizado para TensorFlow/JAX. Disponible solo en [GCP](#).
- **Chips Personalizados:** [AWS](#) diseña sus propios chips para IA: [Trainium](#) (para entrenamiento) e [Inferentia](#) (para inferencia).

Modelo de Pago por Uso

Necesitas entrenar un modelo durante 72 horas en un clúster de 8 GPUs. Lo alquilas, lo usas alquilas, lo usas y lo apagas. El coste es una fracción minúscula de comprar ese hardware.

Ventaja Clave: El cloud democratiza el acceso a hardware especializado que de otra forma estaría fuera del alcance de la mayoría de las organizaciones.



Sinergia 2: El Océano de Datos (Data Lakes)

El Problema

Los modelos de IA se alimentan de datos. A menudo, terabytes o petabytes de datos no estructurados (imágenes, texto, logs). Las bases de datos tradicionales no sirven.

La Solución Cloud: Almacenamiento de Objetos

Concepto: Un sistema de almacenamiento para guardar cantidades virtualmente ilimitadas de datos como "objetos" en "buckets" (contenedores). Es increíblemente barato, duradero y escalable.

Servicios Clave

- **AWS: Simple Storage Service (S3).** El estándar de facto de la industria.
- **Azure: Blob Storage.**
- **GCP: Cloud Storage.**

El Data Lake

Un Data Lake es un repositorio centralizado que te permite almacenar todos tus datos estructurados y no estructurados a cualquier escala. Normalmente, se implementa sobre almacenamiento de objetos. S3 es el corazón de la mayoría de los Data Lakes en AWS.

Sinergia 3: Plataformas de ML Gestionadas (PaaS para IA)

El Problema

Gestionar todo el ciclo de vida de ML es complejo (preparar datos, entrenar, desplegar, escalar, monitorizar). Hacerlo manualmente es propenso a errores.

La Solución Cloud

Plataformas integradas que te dan herramientas para cada paso.



Amazon SageMaker

La plataforma más completa de AWS. Ofrece desde notebooks Jupyter gestionados (SageMaker Studio) hasta entrenamiento distribuido, ajuste de hiperparámetros automático, y despliegue de modelos con un solo clic.



Azure Machine Learning

La propuesta equivalente de Azure, con un fuerte enfoque en un diseñador visual (drag-and-drop) y MLOps.



Google Vertex AI

La plataforma unificada de GCP que combina los servicios de IA de Google, con potentes capacidades de AutoML.

Sinergia 4: APIs de IA (SaaS para IA)

El Problema

No todas las empresas tienen el talento o los datos para entrenar sus propios modelos de visión por computador, NLP o reconocimiento de voz.

La Solución Cloud

Los proveedores han entrenado modelos masivos de altísima calidad y los exponen como simples APIs de pago por uso. Esto democratiza el acceso a la IA.

Ejemplos Comparativos

Visión por Computador

- **AWS:** Amazon Rekognition
- **Azure:** Cognitive Services - Vision
- **GCP:** Vision AI

Traducción de Texto

- **AWS:** Amazon Translate
- **Azure:** Cognitive Services - Translator
- **GCP:** Translation AI

```
import boto3# Crear clienterekognition = boto3.client('rekognition')# Analizar imagenresponse = rekognition.detect_labels( Image={ 'S3Object': { 'Bucket': 'mi-bucket', 'Name': 'imagen.jpg' } })# Ver resultadosfor label in response['Labels']:print(f"{label['Name']}: {label['Confidence']:.2f}%")
```

MLOps: La Pieza Final del Puzzle

Definición

MLOps es a Machine Learning lo que DevOps es al desarrollo de software. Es un conjunto de prácticas que busca unificar el desarrollo de modelos de ML (ML) y su despliegue y operación (Ops).

¿Cómo Ayuda el Cloud?

El cloud proporciona las herramientas para construir los pipelines de MLOps:

CI/CD para ML

Servicios como **AWS CodePipeline**, **Azure DevOps** o **Google Cloud Build** pueden automatizar el re-entrenamiento y despliegue de modelos cuando llega nuevo código o nuevos datos.

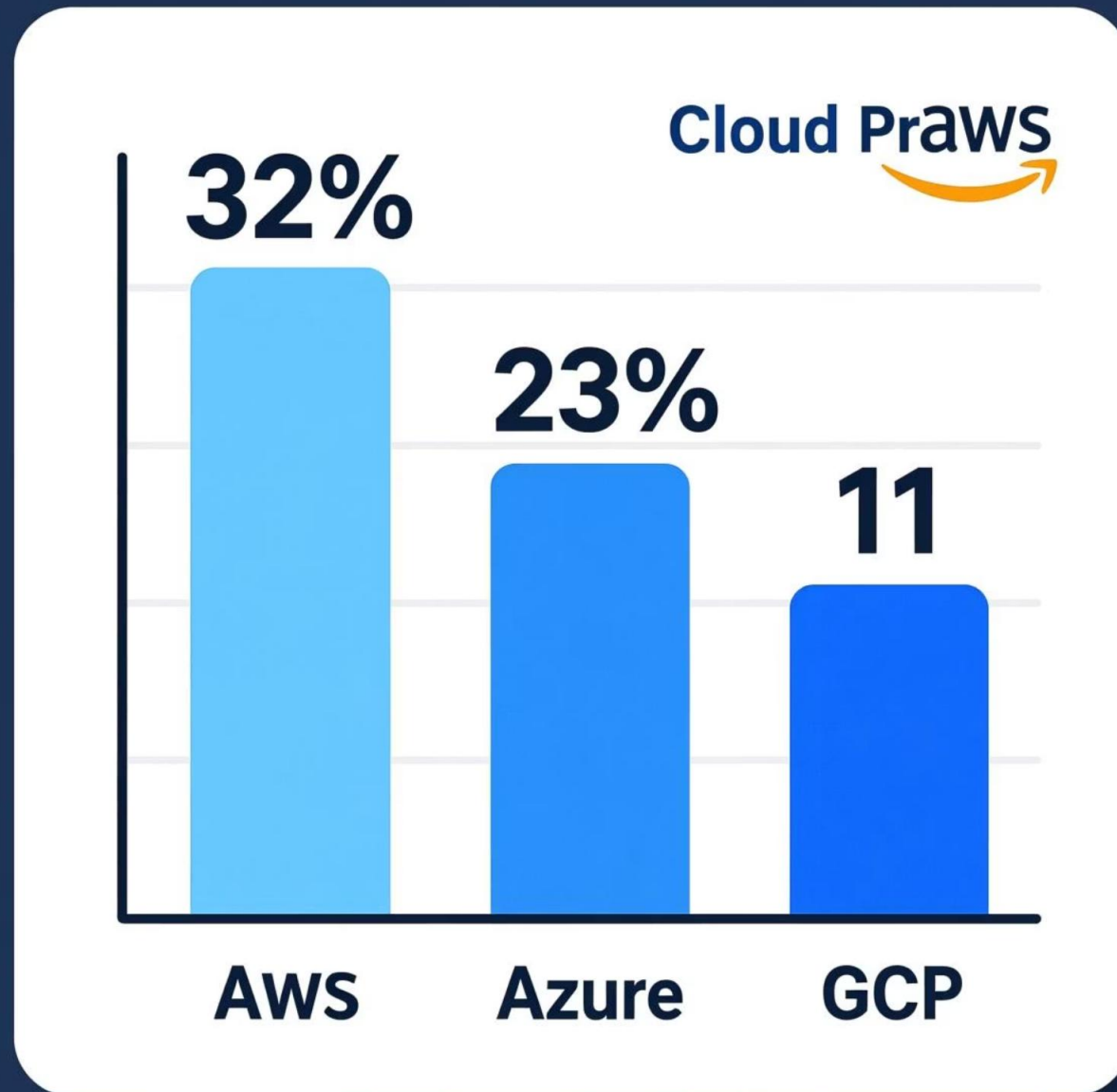
Registros de Modelos

Amazon SageMaker Model Registry, por ejemplo, permite versionar y catalogar los modelos entrenados antes de desplegarlos.

Monitorización

Herramientas como **Amazon SageMaker Model Monitor** detectan automáticamente el "drift" (cuando el rendimiento del modelo en producción se degrada).

Los Tres Grandes: AWS vs. Azure vs. GCP



AWS (Amazon Web Services)

Posición

El líder indiscutible en cuota de mercado (~32%). El pionero.

Fortalezas

La gama más amplia y madura de servicios, enorme comunidad, marketplace muy extenso. Dominante en startups y empresas "nacidas en la nube".

Microsoft Azure

Posición

El segundo (~23%). Crecimiento muy rápido.

Fortalezas

Integración perfecta con el ecosistema empresarial de Microsoft (Windows Server, Office 365, Active Directory). Muy fuerte en el sector corporativo y en soluciones híbridas.

GCP (Google Cloud Platform)

Posición

El tercero (~11%). Visto como el más innovador tecnológicamente.

Fortalezas

Liderazgo en Kubernetes (GKE), analítica de datos (BigQuery), redes y, por supuesto, IA/ML (TensorFlow, TPUs, Vertex AI).

Navegando la Consola: Conceptos Universales

Aunque cada consola es diferente, todas comparten una estructura conceptual similar que debéis entender.

Regiones

Una ubicación geográfica física en el mundo donde agrupan centros de datos (ej: us-east-1 (ej: us-east-1 en N. Virginia, eu-west-1 en Irlanda). La elección de la región afecta a la latencia, el coste y la soberanía de los datos.

Zonas de Disponibilidad (AZs)

Cada Región consta de múltiples AZs. Una AZ es uno o más centros de datos discretos con energía, refrigeración y redes redundantes. Están lo suficientemente cerca para tener baja latencia, pero lo suficientemente lejos para que un desastre en una no afecte a las otras.

| Diseñar para alta disponibilidad implica desplegar en múltiples AZs.

Servicios

La lista de todos los productos (EC2, S3, RDS, etc.) que puedes usar. Cada servicio tiene su propia sección en la consola con opciones de configuración específicas.



Resumen y Conclusiones Clave



El Cloud Computing es un modelo operativo definido por 5 principios clave

No solo "ordenadores de otra persona". Los 5 principios NIST definen claramente qué es y qué no es cloud computing.



El paso de CAPEX a OPEX democratiza la innovación

Convierte grandes inversiones iniciales en gastos operativos predecibles, permitiendo a startups competir con grandes empresas.



La elección entre IaaS, PaaS y SaaS depende del control vs. velocidad

A más control (IaaS), más responsabilidad. A menos control (SaaS), mayor velocidad de desarrollo y menor responsabilidad operativa.



La IA moderna depende del cloud

Para obtener cómputo especializado (GPUs, TPUs), almacenamiento masivo (Data Lakes) y servicios gestionados (MLOps) a escala.



AWS, Azure y GCP dominan el mercado

Cada uno con sus propias fortalezas estratégicas. Comprender los conceptos de Región y AZ es fundamental para cualquier diseño.

Preguntas y Próximos Pasos

Turno de Preguntas Abierto (Q&A)

Aprovechad para resolver cualquier duda sobre los conceptos vistos hoy.



Próxima Sesión (Sesión 2)

Cómputo y Almacenamiento - Los Pilares de IaaS

En la próxima clase, nos arremangaremos. Haremos un *deep dive* en el servicio más fundamental: Amazon EC2. Lanzaremos nuestras primeras instancias, exploraremos los diferentes tipos de familias, y entenderemos el almacenamiento en bloque con EBS.

También sentaremos las bases del almacenamiento de objetos con S3, el corazón de cualquier estrategia cualquier estrategia de datos para IA.



Videos recomendados:

- "AWS In 10 Minutes" (Por Fireship)
- "What is the Shared Responsibility Model?" (Canal Oficial de AWS)