Cloud Computing para Inteligencia Artificial

Sesión 2: Conceptos Clave de Infraestructura Cloud y Primeros Pasos en AWS

Agenda de la Sesión

La Huella Global del Cloud

Regiones, Zonas de Disponibilidad y Edge Locations

Seguridad: El Modelo de Responsabilidad Compartida

Quién es responsable de qué en la nube

Gestión de Identidad y Acceso (IAM)

¿Quién puede hacer qué?

Redes Virtuales (VPC)

Construyendo tu fortaleza en la nube

Primeros Pasos Prácticos en la Consola de AWS

Configuración inicial y seguridad

Vistazo Comparativo

AWS vs. GCP vs. Azure

Objetivos de Aprendizaje

Al finalizar esta sesión, serás capaz de:

Describir

La arquitectura de la infraestructura global de un proveedor cloud y su impacto en el diseño de aplicaciones de IA.

Explicar

El Modelo de Responsabilidad Compartida y tus obligaciones de seguridad en la nube.

Diseñar

Políticas de acceso seguras utilizando los componentes fundamentales de IAM.

Esbozar

Una arquitectura de red básica y segura en la nube usando VPCs y subredes.

Realizar

La configuración de seguridad inicial en una nueva cuenta de AWS.

Infraestructura Global: El Mundo a tu Disposición

Los proveedores de cloud no operan desde un único lugar. Han construido una **red masiva y distribuida de centros de datos** por todo el mundo.

Analogía: Piensa en ello como una red logística global. Si quieres entregar un entregar un paquete (tus datos o aplicación) a un usuario en Japón, no lo no lo envías desde Madrid; usas un centro de distribución local en Asia. Asia.

Esta infraestructura física es la base sobre la que se construyen todos los servicios cloud. Comprender su estructura es clave para diseñar sistemas:

- Resilientes
- De baja latencia
- Que cumplan con la normativa

Regiones Cloud

Definición

Una **Región** es un área geográfica física y separada en el mundo donde se agrupan centros de datos.

Ejemplos de AWS

- eu-west-1 (Irlanda)
- us-east-1 (N. Virginia)
- ap-southeast-2 (Sídney)

Criterios para su establecimiento

Latencia

Estar cerca de los usuarios finales para para reducir el tiempo de respuesta.

Soberanía de datos

Leyes como el GDPR en Europa exigen exigen que los datos de los ciudadanos ciudadanos residan físicamente en la UE. la UE.

Disponibilidad de recursos

Acceso a energía, conectividad de red y estabilidad geopolítica.

Zonas de Disponibilidad (AZs)

Definición

Una **Zona de Disponibilidad (AZ)** es uno o más centros de datos datos discretos con energía, refrigeración y redes redundantes redundantes dentro de una Región.

Aislamiento Físico

Las AZs están lo suficientemente lejos como para no verse afectadas afectadas por un desastre único (incendio, inundación), pero lo pero lo suficientemente cerca como para tener una latencia de red de red muy baja entre ellas (<10ms).

Clave para la Alta Disponibilidad

El pilar del diseño de arquitecturas tolerantes a fallos. Si una AZ falla, AZ falla, tu aplicación puede seguir funcionando en otra AZ de la de la misma región sin interrupción para el usuario.

Ejemplo práctico para IA

Entrenas un modelo de Computer Vision en una instancia en la AZ en la AZ eu-central-1a. Para asegurar que el servicio de inferencia inferencia esté siempre disponible, despliegas instancias idénticas en idénticas en las AZs eu-central-1b y eu-central-1c detrás de un un balanceador de carga.

Edge Locations / Points of Presence (PoPs)

Definición

Son una red de centros de datos mucho más pequeños y numerosos que las regiones, distribuidos distribuidos en las principales ciudades del mundo. mundo.

Propósito Principal

Acercar el contenido al usuario final para minimizar la minimizar la latencia. No se usan para ejecutar tus tus aplicaciones principales (como modelos de IA), IA), sino para acelerar la entrega de contenido.

Servicios Clave que los utilizan

Amazon CloudFront (CDN)

Almacena en caché copias de tu contenido cerca de los usuarios. Un usuario en Madrid recibe el contenido desde un PoP en Madrid, no desde la región de Irlanda.

Amazon Route 53 (DNS)

Resuelve las peticiones de DNS desde el PoP PoP más cercano, acelerando la conexión inicial. inicial.

Alcance de los Servicios: Global vs. Regional vs. Zonal Zonal

No todos los servicios de AWS operan de la misma manera en relación a la infraestructura global. Es crucial entender su alcance:

Servicios Globales

Operan sin necesidad de seleccionar una una región. Son únicos y universales en tu en tu cuenta.

Ejemplos: IAM (usuarios y permisos son los son los mismos en todo el mundo), Route 53 Route 53 (DNS es global), CloudFront.

Servicios Regionales

Se despliegan dentro de una región específica, pero están diseñados para replicarse o utilizar múltiples AZs automáticamente para alta disponibilidad. disponibilidad.

Ejemplos: Amazon S3 (los buckets se crean en una región, pero los datos son altamente duraderos), Amazon DynamoDB.

Servicios Zonales

Se despliegan en una AZ específica. Si esa AZ falla, el recurso falla. Eres responsable de diseñar la redundancia entre AZs.

Ejemplo: Amazon EC2 (una instancia de de máquina virtual se lanza en una AZ concreta de una región).

Modelo de Responsabilidad Compartida

Concepto Central

La seguridad en la nube no es una responsabilidad exclusiva del proveedor; es un proveedor; es un **acuerdo compartido** entre el proveedor (AWS, Google, Google, Microsoft) y tú, el cliente.

Analogía

AWS te alquila una casa increíblemente segura (con muros de hormigón, alarma perimetral y guardias). Ellos se encargan de la seguridad **DEL** edificio. Pero tú eres responsable de cerrar la puerta con llave, no dejar las ventanas abiertas y decidir a quién le das una copia de la llave. Esa es la seguridad **EN** la casa.

AWS es responsable de la seguridad DEL cloud.

Tú eres responsable de la seguridad EN el EN el cloud.

Desglose del Modelo (IaaS, PaaS, SaaS)

La línea que divide la responsabilidad cambia según el tipo de servicio que uses:

IaaS (Infrastructure as a Service - Ej: EC2, VPC)

AWS: Infraestructura física (regiones, AZs), computación, almacenamiento, red y el hipervisor.

Cliente: TODO lo demás. Seguridad del del S.O. (parches), configuración de firewalls (Security Groups), gestión de de identidades (IAM), cifrado de datos y la datos y la seguridad de la aplicación.

PaaS (Platform as a Service - Ej: Ej: AWS Lambda, RDS)

AWS: Gestiona también el S.O., el middleware y el runtime.

Cliente: Se centra en la seguridad de su su código, la gestión de los datos y la configuración del servicio (ej: quién puede puede invocar una función Lambda).

SaaS (Software as a Service - Ej: Gmail, Amazon SageMaker)

AWS: Gestiona casi todo.

Cliente: Responsable de la gestión de de usuarios y sus datos dentro de la aplicación.

Implicaciones para Arquitecturas de IA

Datos de Entrenamiento

Tú eres responsable de cifrar los datasets sensibles en S3, controlar el controlar el acceso mediante políticas de bucket e IAM, y asegurar que asegurar que cumplen con normativas de privacidad. AWS asegura que el asegura que el disco físico donde se guarda no falle.

Modelos Entrenados

El artefacto de tu modelo (ej. model.pth o saved_model.pb) es tu responsabilidad. Debes protegerlo contra el acceso no autorizado, tanto en reposo (S3) como en tránsito.

Entorno de Cómputo (EC2/SageMaker)

Si usas una instancia EC2 para entrenar, eres responsable de parchear el S.O., configurar las reglas del firewall (Security Groups) para que solo acepten tráfico necesario y gestionar las credenciales de acceso.

Endpoints de Inferencia

Tú configuras la autenticación y autorización para la API que expone tu expone tu modelo. AWS se encarga de que el servidor que ejecuta la API ejecuta la API no se caiga.

IAM: La Piedra Angular de la Seguridad

Definición

IAM (Identity and Access Management) es el servicio de AWS que te permite permite gestionar de forma segura el acceso a los servicios y recursos de AWS. recursos de AWS.
IAM responde a la pregunta fundamental: **¿Quién (identidad) puede hacer qué (permisos) en qué recursos?**

Autenticación vs. Autorización

Autenticación (¿Quién eres?)

El proceso de verificar tu identidad. Se hace mediante usuario/contraseña, claves de acceso, o credenciales temporales. temporales.

Autorización (¿Qué puedes hacer?)

El proceso de determinar qué permisos tiene una identidad autenticada. autenticada. Se gestiona mediante políticas.

IAM es un servicio **global**. Un usuario creado en IAM existe en todas las regiones.

Componentes Fundamentales de IAM

Usuarios (Users)

Una entidad que representa a una una persona o una aplicación que que interactúa con AWS.

Credenciales: Contraseña para la la consola, y Claves de Acceso (Access Key ID y Secret Access Key) Key) para el acceso programático programático (CLI, SDKs).

Grupos (Groups)

Una colección de usuarios. Es una una forma de asignar permisos a a múltiples usuarios a la vez en lugar lugar de hacerlo uno por uno.

Ejemplo: Un grupo DataScientists con permisos para acceder a S3 y SageMaker.

Roles

Una identidad con permisos específicos que puede ser "asumida" temporalmente por una una entidad de confianza (un usuario, una aplicación, otro servicio servicio de AWS).

Ventaja clave: No requiere credenciales a largo plazo como las como las claves de acceso, usa credenciales de seguridad temporales.

Profundizando en los Roles de IAM

Los roles son uno de los conceptos más potentes y seguros de IAM.

Caso de Uso 1: Acceso entre servicios de AWS

Escenario

Tienes una instancia EC2 que necesita leer datos de un bucket de S3 para entrenar un modelo

Solución Insegura

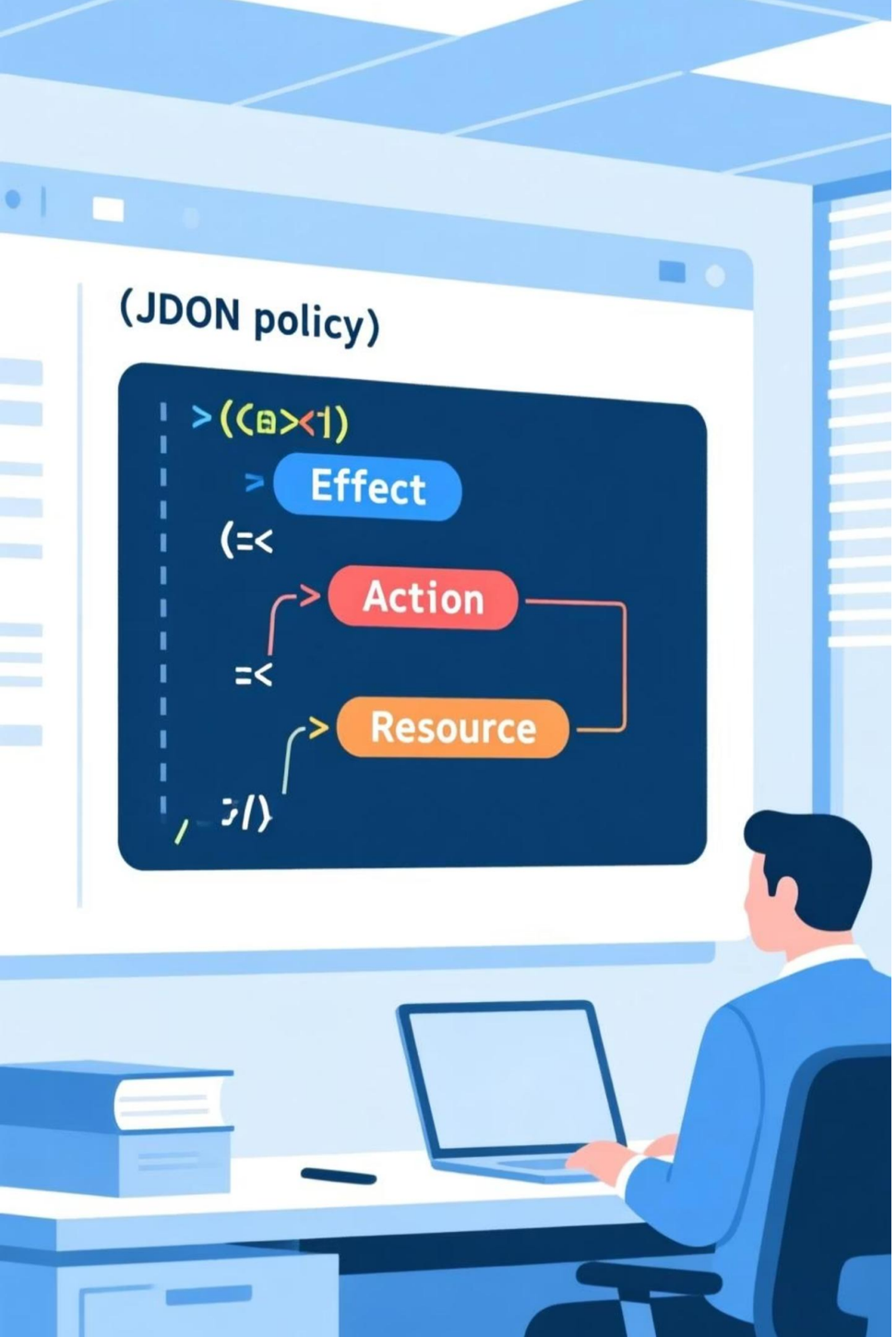
Guardar claves de acceso de un usuario IAM en la instancia EC2. ¡Muy mala práctica!

Solución Segura (con Rol)

Creas un Rol que tiene permiso para leer de S3. Asocias ese Rol a la instancia EC2. instancia EC2. La instancia "asume" el rol y obtiene credenciales temporales para temporales para acceder a S3.

Caso de Uso 2: Acceso para usuarios federados

Permite que usuarios de tu directorio corporativo (ej. Active Directory) accedan a la consola a la consola de AWS sin necesidad de crearles usuarios IAM individuales.



Políticas: El Lenguaje de los Permisos

Definición

Una **Política** es un documento JSON que que define explícitamente los permisos. Todo Todo en IAM se basa en políticas.

Estructura de una Política JSON

- **Effect:** Allow o Deny. Deny siempre prevalece.
- **Action:** La API call del servicio (ej. ec2:StartInstances).
- **Resource:** El ARN (Amazon Resource Resource Name) del recurso.
- **Condition:** (Opcional) Condiciones para que aplique la política.

```
{ "Version": "2012-10-17",
"Statement": [ { "Effect": "Allow",
"Action": [ "s3:GetObject",
"s3:ListBucket" ], "Resource": [
"arn:aws:s3:::my-ai-datasets",
"arn:aws:s3:::my-ai-datasets/*" ],
"Condition": { "IpAddress": {
"aws:SourceIp": "198.51.100.0/24" } }
} ]}
```

Gamificación: ¡Decodifica la Política!

```
{  "Version": "2012-10-17",
"Statement": [    {      "Effect":
"Allow",      "Action":
"sagemaker:CreateTrainingJob",
"Resource": "*"    },    {
"Effect": "Deny",      "Action":
"ec2:TerminateInstances",
"Resource": "arn:aws:ec2:eu-west-1:
123456789012:instance/i-
0abcdef1234567890"    }  ]}
```

Actividad Interactiva

La siguiente política IAM se adjunta a un grupo de JuniorDataScientists. ¿Qué les permite hacer?

1. Pueden crear trabajos de entrenamiento en SageMaker y terminar cualquier instancia EC2.
2. Pueden crear trabajos de entrenamiento en SageMaker, pero no pueden terminar una instancia EC2 específica.
3. Solo pueden terminar una instancia EC2 específica.
4. No pueden hacer nada porque la política es es inválida.

Principios y Mejores Prácticas de IAM

Principio de Mínimo Privilegio

Otorga solo los permisos mínimos necesarios para que una identidad realice su tarea, y nada tarea, y nada más. No des s3:* si solo necesitas s3:GetObject.

Autenticación Multifactor (MFA)

¡Actívala siempre! Añade una capa extra de seguridad al requerir un segundo factor de autenticación además de la contraseña.

Mejores Prácticas Clave

NO uses el usuario root para las tareas diarias. Úsalo solo para crear tu primer usuario administrador y para tareas de facturación.

Crea **usuarios IAM individuales** para cada persona. No compartas credenciales. Usa **grupos** para asignar permisos.

Usa **roles** para aplicaciones y servicios AWS. **Audita** los los permisos regularmente con herramientas como IAM IAM Access Analyzer.

VPC: Tu Centro de Datos Virtual y Privado

Definición

Una **Virtual Private Cloud (VPC)** es una red privada, lógicamente aislada, dentro de la nube de la nube de AWS. Es tu propio trozo de la red de AWS donde puedes lanzar recursos. recursos.

Analogía

Piensa en el cloud de AWS como una gran ciudad. Una VPC es como si compraras un terreno compraras un terreno y levantaras un muro perimetral. Dentro de ese perímetro, tienes tienes control total: puedes construir edificios (subredes), establecer calles (tablas de rutas) (tablas de rutas) y decidir quién entra y sale por la puerta principal (gateways).

Direccionamiento IP

Se define un rango de IPs privadas para tu VPC usando la notación CIDR (Classless Inter-Domain Routing).

Ejemplo: 10.0.0.0/16 te da 65,536 direcciones IP para usar dentro de tu VPC.

Subredes: Segmentación de tu Red

Definición

Una **Subred** es un segmento de rango de IP de tu VPC. Las subredes te permiten agrupar recursos según sus necesidades de seguridad y enrutamiento.

Cada subred debe estar asociada a **una única única Zona de Disponibilidad (AZ)**. Esto es Esto es clave para la alta disponibilidad.

Tipos de Subredes

Subred Pública

Una subred cuyo tráfico puede ser enrutado directamente desde y hacia internet a través de un Internet Gateway. Ideal para servidores web o balanceadores de carga.

Subred Privada

Una subred que no tiene una ruta directa a internet. Ideal para bases de bases de datos, backends o, muy importante, para las instancias de entrenamiento de modelos de IA que que no necesitan ser expuestas al mundo.

Tablas de Rutas y Gateways

Tablas de Rutas

Conjunto de reglas, llamadas rutas, que determinan a dónde se dirige el tráfico de red desde tu subred. Cada subred está asociada a una tabla de rutas.

Gateways (Las puertas de tu VPC)

Internet Gateway (IGW)

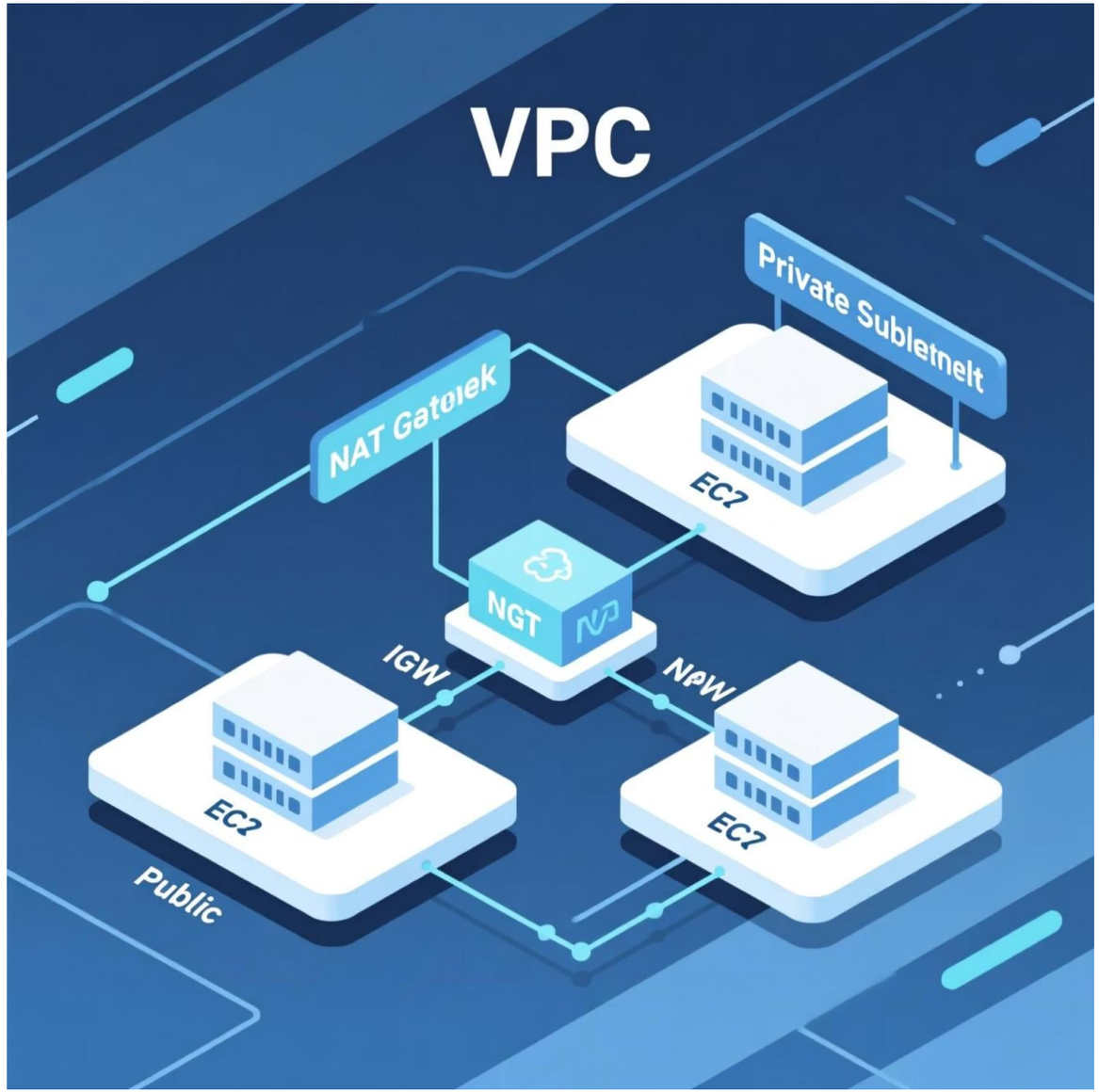
Permite la comunicación bidireccional entre instancias en tus subredes públicas e internet. Es la puerta principal.

NAT Gateway

Se coloca en una subred pública y permite a las instancias en subredes privadas iniciar tráfico saliente a internet, pero impide que internet inicie conexiones entrantes.

Virtual Private Gateway (VPG)

Conecta tu VPC a tu red on-premise a través de una conexión VPN segura.



Arquitectura VPC Típica para IA IA

Un diseño común para una aplicación de Machine Learning:

VPC con 2 AZs para alta disponibilidad. En cada AZ, una subred pública y una pública y una privada.

Subredes Públicas: Contienen un balanceador de carga que recibe las peticiones de inferencia desde internet.

Subredes Privadas: Contienen las instancias EC2/SageMaker que alojan el alojan el modelo. Estas instancias reciben tráfico solo desde el balanceador de balanceador de carga.

S3: Los datos y modelos se almacenan en S3. El acceso desde la VPC se puede hacer de forma segura y privada usando un *VPC Endpoint para S3*.

Gamificación: Elige la Arquitectura Correcta

Escenario de Debate

> "Estás diseñando la infraestructura para un sistema de análisis análisis de sentimientos en tiempo real que procesa tweets. Los tweets. Los datos son públicos, pero el modelo que has desarrollado es propiedad intelectual de tu empresa y debe debe estar protegido. La aplicación debe ser altamente disponible."

Pregunta

¿En qué tipo de subred (pública o privada) desplegarías los siguientes componentes y por qué?

Servidor web que recibe los tweets

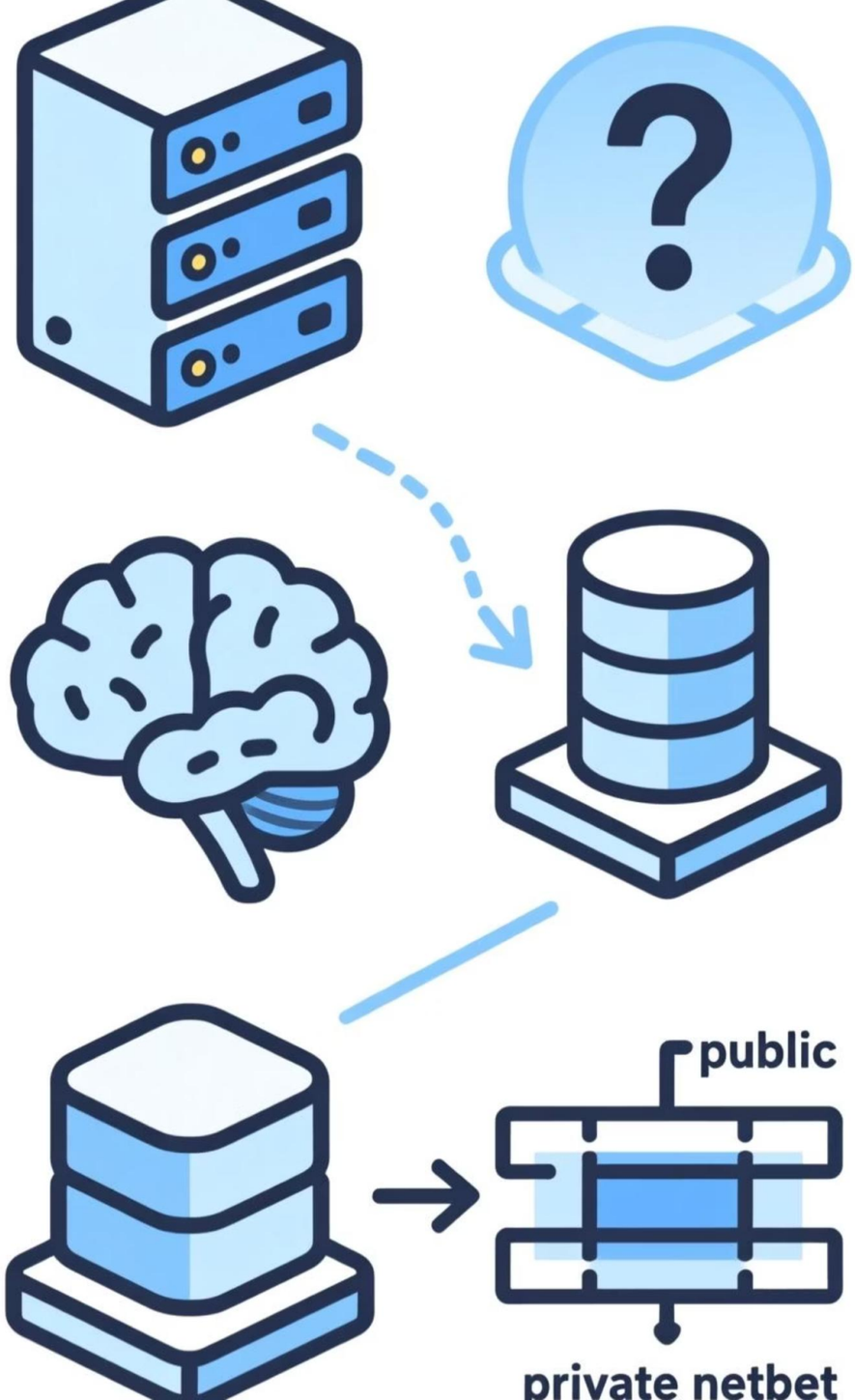
¿Pública o privada?

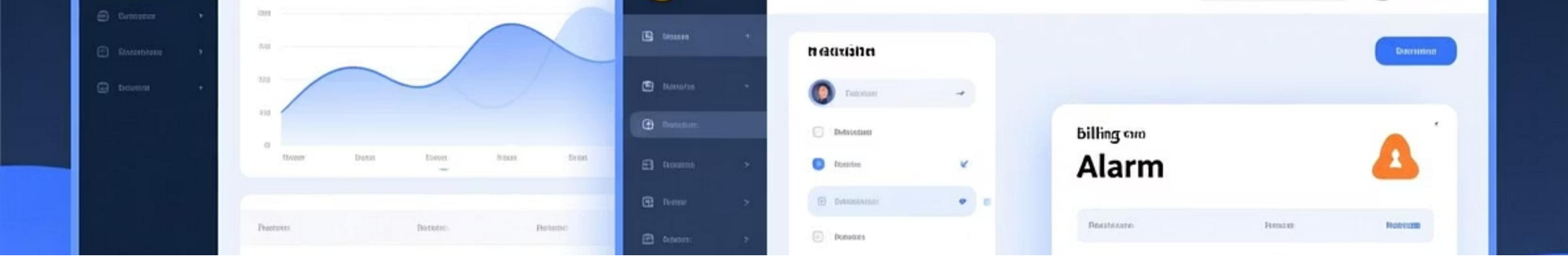
Instancias que ejecutan el modelo de análisis análisis

¿Pública o privada?

Base de datos que almacena los resultados

¿Pública o privada?





Creación de Cuenta y Navegación en la Consola AWS

Paso 1: Crear una cuenta de AWS

Ve a aws.amazon.com. Necesitarás una tarjeta de crédito (no se cobra nada si te nada si te mantienes en la capa gratuita) y un número de teléfono.

Paso 2: ¡Activa las Alertas de Facturación!

Es lo primero que debes hacer. Ve al servicio **CloudWatch -> Billing** y crea una crea una alarma que te notifique por email si tu gasto estimado supera un umbral un umbral (ej. 5$).

Navegación por la Consola

Barra de Búsqueda

La forma más rápida de encontrar cualquier servicio.

Selector de Región

¡Importante! Asegúrate siempre de estar trabajando en la región correcta. correcta. Lo que creas en una región no aparece en otra.

Tarea Práctica: Asegura tu Cuenta

Objetivo: Realizar la configuración de seguridad inicial y fundamental.

1

No uses el usuario root

El usuario root tiene acceso completo a todos los servicios y recursos. Solo debe usarse para tareas específicas.

2

Configura MFA para el usuario root

Inicia sesión como root, ve al servicio **IAM -> Dashboard** y activa el MFA.

3

Crea tu primer usuario Administrador

En IAM, ve a **Users -> Add users**. Dale un nombre, permite el acceso a la consola, crea un grupo Administrators y adjunta la política AdministratorAccess.

4

Cierra sesión de root e inicia con tu usuario admin

A partir de ahora, usa siempre este usuario para las tareas administrativas.

5

Configura MFA también para tu usuario admin

Añade esta capa adicional de seguridad para todas tus identidades administrativas.

Introducción a AWS CLI y SDKs

Aunque la consola es útil para empezar, el verdadero poder del cloud para un ingeniero de IA está en la automatización y el acceso programático.

AWS Command Line Interface (CLI)

Una herramienta para controlar los servicios de servicios de AWS desde tu terminal.

Permite crear scripts para automatizar tareas repetitivas (ej. lanzar 10 instancias para un hyperparameter tuning).

Ejemplo: `aws s3 cp my_model.pth s3://my-models-bucket/`

AWS Software Development Kits (SDKs)

Librerías para interactuar con las APIs de AWS desde tu lenguaje de programación preferido.

Esencial para integrar los servicios de AWS en tus en tus aplicaciones.

Ejemplo para IA (Python - Boto3)

```
import boto3s3 =
boto3.client('s3')s3.download_fi
le('my-ai-datasets',
'dataset.csv',
'local_dataset.csv')
```

Comparativa de Conceptos: AWS vs. GCP vs. Azure Azure

Los conceptos fundamentales existen en todos los proveedores, pero con nombres y estructuras diferentes.

Organización y Aislamiento de Recursos

AWS	GCP	Azure
Cuenta (Account) es el principal contenedor. La organización se puede puede hacer con **AWS Organizations**. **Organizations**.	El contenedor fundamental es el **Proyecto Proyecto (Project)**. Los proyectos se pueden agrupar en **Carpetas (Folders)** y **(Folders)** y **Organizaciones (Organizations)**. Cada proyecto tiene su su propia facturación, APIs y recursos. recursos.	El contenedor principal es la **Suscripción Suscripción (Subscription)**. Los recursos recursos dentro de una suscripción se se organizan en **Grupos de Recursos (Resource Groups)**, que actúan como un como un ciclo de vida para los recursos recursos que contienen.

Comparativa de IAM y Redes

Gestión de Identidades

AWS	GCP	Azure
IAMUsuarios, Grupos, Roles, Políticas JSON	**Cloud IAM**Miembros como Usuarios, Grupos, Cuentas de Servicio. Los permisos se agrupan en **Roles** - Primitivos, Predefinidos, Personalizados	**Azure Active Directory (Azure AD)**Las identidades (Usuarios, Grupos, Service Principals) obtienen acceso a recursos a través de asignaciones de **Roles (RBAC)**

Redes Virtuales

AWS	GCP	Azure
VPC (Virtual Private Cloud) Son de ámbito **Regional**. Las subredes son **Zonales**	**VPC Network**Son de ámbito **Global**. Puedes tener subredes en diferentes regiones dentro de la misma VPC. Las subredes son **Regionales**	**VNet (Virtual Network)**Son de ámbito **Regional**. Las subredes se definen dentro de la VNet

Resumen y Puntos Clave

Infraestructura Global

La base de todo. El diseño basado en Regiones y AZs es crucial para la disponibilidad y el cumplimiento normativo.

Responsabilidad Compartida

La seguridad es un trabajo de equipo. AWS asegura la nube, tú aseguras TUS cosas EN la nube.

IAM es la Clave

El principio de mínimo privilegio no es una sugerencia, es una regla. Usa roles siempre que sea posible y activa el MFA.

VPC es tu Fortaleza

Aísla tus recursos. Usa subredes públicas y privadas para proteger tus proteger tus componentes críticos, como los modelos de IA y las bases de las bases de datos.

La Práctica es Fundamental: La teoría es importante, pero la configuración real de estos servicios es donde ocurre el verdadero aprendizaje. aprendizaje.

Próximos Pasos

Para la próxima sesión

Revisaremos los conceptos de cómputo en la nube, centrándonos en máquinas virtuales (EC2) y contenedores.

Lanzaremos nuestra primera máquina virtual y la configuraremos para un un entorno básico de ciencia de datos.

Tarea/Lectura recomendada

- Finalizar la configuración de seguridad de tu cuenta de AWS (MFA, usuario admin). usuario admin).
- Leer el whitepaper de AWS: "Introducción a la seguridad de AWS".

Enlaces a recursos adicionales:

- AWS re:Invent: Introduction to AWS Identity and Access Management Management (IAM)
- Google Cloud Tech: Virtual Private Cloud (VPC) deep dive

¿Preguntas?

Contacto:

profesor@universidad.es

Horario de tutorías: Lunes y Miércoles de 15:00 a 17:00

Oficina: Edificio de Informática, Despacho 2.15