

Cloud Computing para Inteligencia Artificial

Sesión 5: Servicios de Almacenamiento I Almacenamiento I - Almacenamiento de Almacenamiento de Objetos y Bloques Bloques

Grado en Ingeniería en Inteligencia Artificial



Agenda de la Sesión

01

El Fundamento de los Datos

Tipos de Almacenamiento en la Nube

03

El Disco Duro de la Nube

Profundización en Almacenamiento de Bloques (AWS EBS)

05

Análisis Comparativo

Vistazo a GCP y Azure

02

Almacenamiento Infinito para Datos No Estructurados

Profundización en Almacenamiento de Objetos (AWS S3)

04

Aplicación Práctica

Escenario de Arquitectura para IA

06

Conclusiones y Preguntas

Resumen y discusión final

La Tríada del Almacenamiento Cloud

Introducción a los tres paradigmas fundamentales de almacenamiento cada uno diseñado para un propósito específico.

Almacenamiento de Objetos

Para escala masiva y datos no estructurados.

Análogo a un servicio de guardería de objetos (valet key): entregas tu "objeto" (fichero) y te dan un ticket (ID único) para recuperarlo.

Almacenamiento de Bloques

El almacenamiento de más alto rendimiento rendimiento para cargas de trabajo estructuradas.

Análogo a un disco duro externo virtual que conectas a tu servidor. El sistema operativo lo controla por completo.

Almacenamiento de Ficheros

Para acceso compartido y jerarquías de directorios.

Análogo a una unidad de red compartida (NAS) en una empresa.



La Tríada del Almacenamiento Cloud

Introducción a los tres paradigmas fundamentales de almacenamiento cada uno diseñado para un propósito específico.

El factor clave es la **latencia** (el tiempo que tarda la información en ir y volver) y se resume así:

Servicio de Almacenamiento	Ubicación Física Relativa	Rendimiento (Latencia)	Diseño
Bloques (EBS)	Almacenamiento dedicado dentro de una única Zona de Disponibilidad (AZ) .	Muy Baja Latencia. Ideal para operaciones de E/S rápidas (IOPS).	Está diseñado para estar virtualmente junto a la instancia EC2 y acoplado por una red ultrarrápida.
Archivos (EFS)	Distribuido y replicado automáticamente a través de múltiples Zonas de Disponibilidad (AZs) .	Latencia ligeramente mayor que EBS, pero aún baja. Varía según el patrón de acceso.	Requiere que el tráfico NFS viaje entre tu EC2 y el <i>cluster</i> de almacenamiento distribuido en la región.
Objetos (S3)	Distribuido a través de múltiples Zonas de Disponibilidad (AZs) y a menudo replicado globalmente (depende de la región).	Mayor Latencia. No diseñado para milisegundos, sino para <i>throughput</i> (transferencia de grandes volúmenes).	El acceso se realiza a través de la API HTTP a través de la red más amplia de AWS.



¿Qué es el Almacenamiento de Objetos?

Definición

Un paradigma de almacenamiento que gestiona datos como "objetos" discretos. No hay jerarquía de carpetas; es una estructura plana (un *flat namespace*).

Componentes de un Objeto

- **Datos:** El contenido del fichero en sí (una imagen, un vídeo, un CSV, un modelo serializado)
- **Identificador Único (Key):** El "nombre" del objeto dentro de su contenedor
- **Metadatos:** Información descriptiva sobre el objeto (tipo de contenido, contenido, fecha de creación, etiquetas personalizadas)

Características Clave

- **Acceso:** Se realiza a través de APIs RESTful (HTTP/S), no montando una unidad
- **Ideal para:** Datos no estructurados, backups, *data lakes*, contenido multimedia
- **Escalabilidad:** Prácticamente ilimitada, desde KB hasta PB



Introducción a Amazon S3

S3: Simple Storage Service

Es uno de los servicios fundacionales de AWS, lanzado en 2006. Diseñado para ofrecer una durabilidad del 99.999999999% (once nueves).

Buckets

Son los contenedores donde se almacenan los objetos.



- **Nombre único globalmente:** Un nombre de bucket mi-dataset-proyecto solo puede existir una vez en todo AWS
- **Recurso Regional:** Aunque el nombre es global, el bucket se crea en una región específica

Objetos

Los ficheros que subimos al bucket.



- Tamaño máximo por objeto: 5TB
- Se identifican por una clave única dentro del bucket
- Pueden tener metadatos personalizados



La "Ilusión" de las Carpetas en S3

S3 tiene un *namespace* plano, pero simula una estructura de carpetas para la comodidad humana y la organización.

Esto se logra mediante el uso de **prefijos** en la clave (nombre) del objeto.

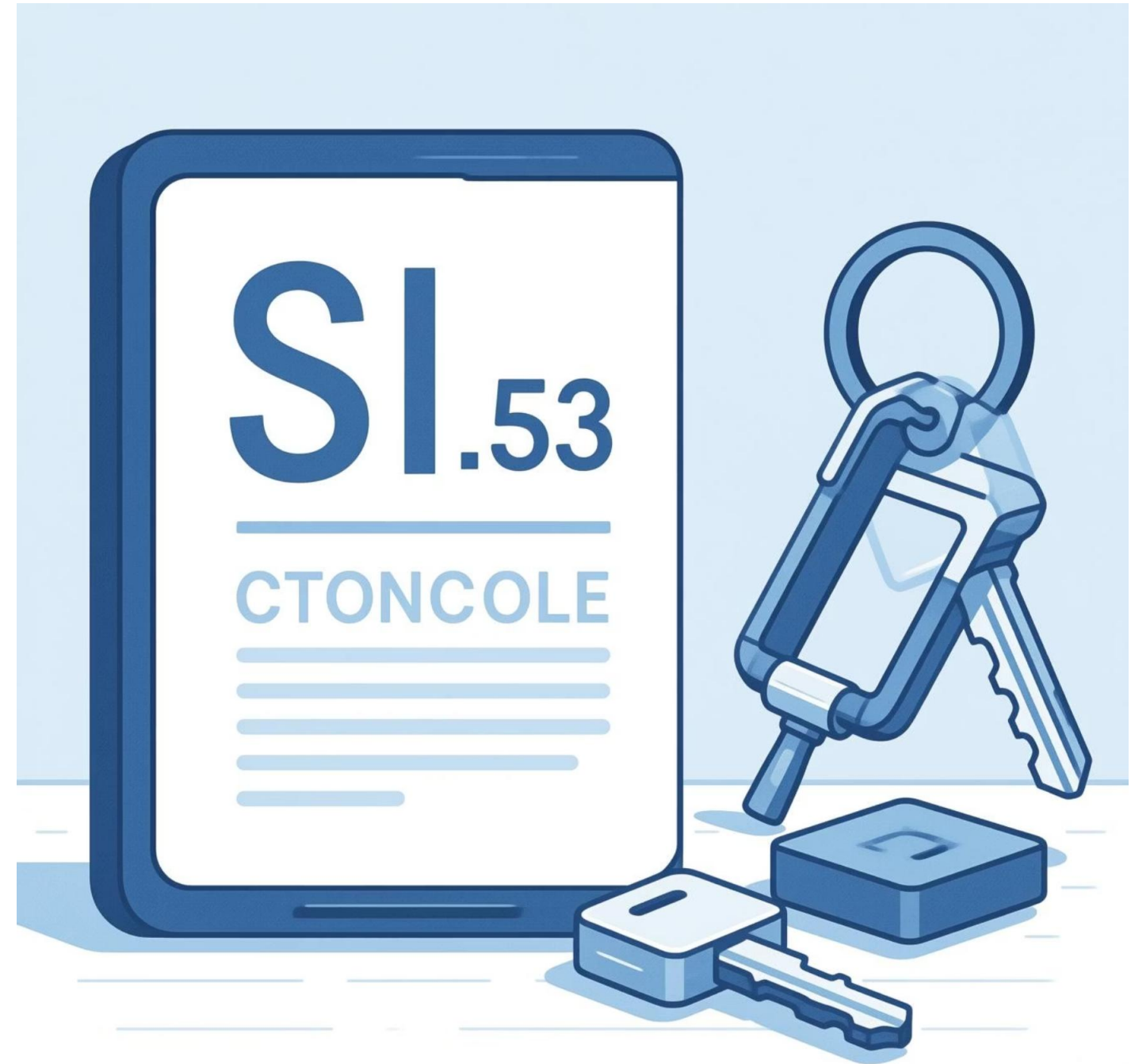
Ejemplo:

- proyecto-ia/datasets/imagenes/gato.jpg
- proyecto-ia/datasets/imagenes/perro.jpg

En este caso, `proyecto-ia/datasets/imagenes/` es el prefijo.

Implicaciones Importantes

- La consola de AWS y las SDKs interpretan el delimitador `/` para mostrar una vista jerárquica
- Para S3, son simplemente dos objetos con nombres largos y complejos
- **No puedes "crear una carpeta vacía"**. Una "carpeta" aparece cuando subes el primer objeto con ese prefijo



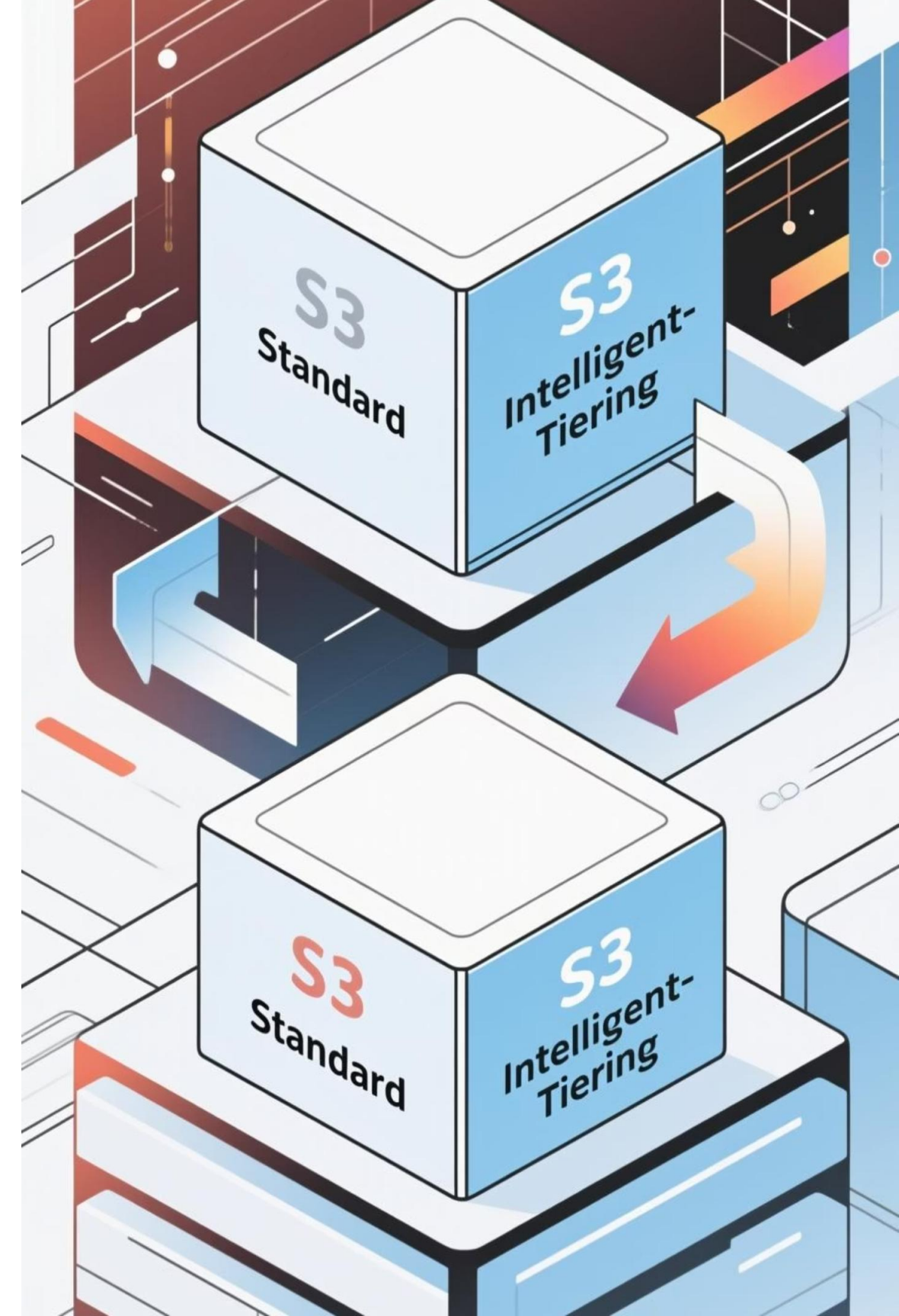
Clases de Almacenamiento (Tiers) - Acceso Frecuente

S3 Standard

- **Uso:** Para datos de acceso muy frecuente que requieren baja latencia y alto rendimiento
- **Durabilidad:** Distribuido en al menos 3 Zonas de Disponibilidad (AZs)
- **Coste:** El más alto en almacenamiento (\$/GB), pero el más bajo en acceso (\$/petición) (\$/petición)
- **Caso IA:** Servir activos de una aplicación web, datasets en uso activo para entrenamiento

S3 Intelligent-Tiering

- **Uso:** Para datos con patrones de acceso desconocidos o cambiantes
- **Funcionamiento:** Monitorea el acceso y mueve automáticamente los objetos entre niveles niveles
- **Coste:** Pequeña tarifa de monitorización, pero potencial ahorro significativo
- **Caso IA:** Datasets de entrenamiento con uso variable a lo largo del tiempo



Clases de Almacenamiento - Acceso Poco Frecuente

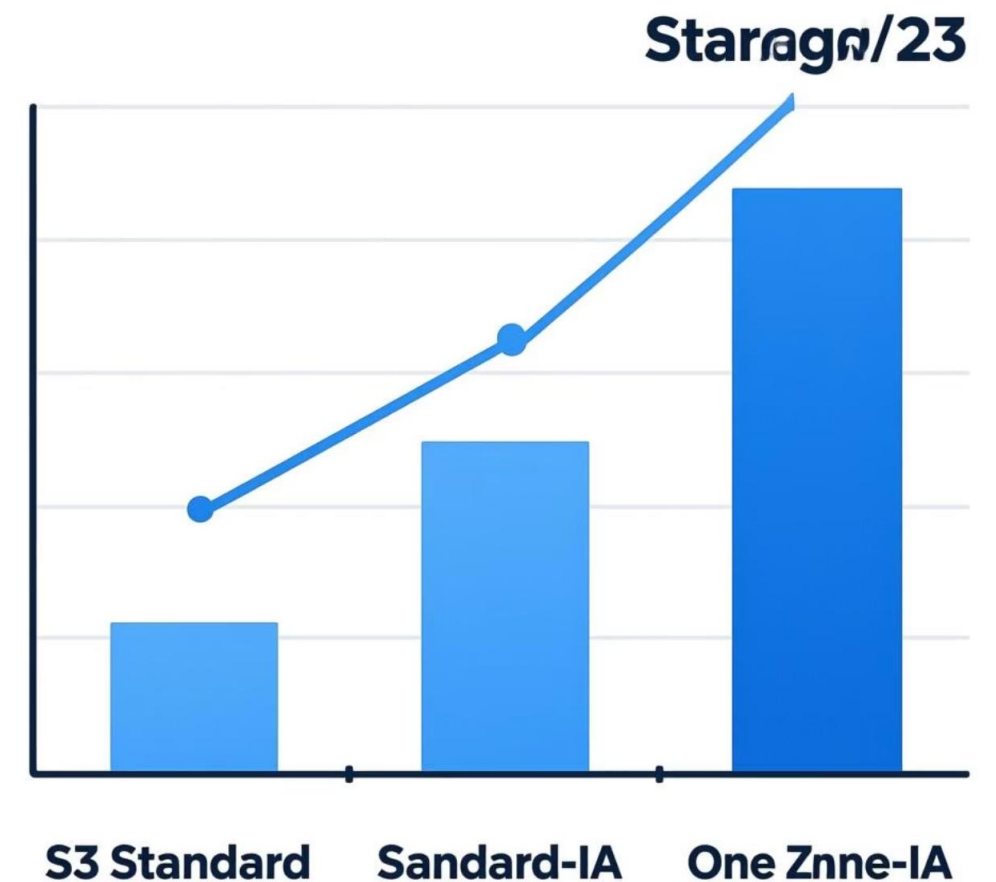
S3 Standard-Infrequent Access (S3 Standard-IA)

- **Uso:** Para datos de acceso menos frecuente, pero que deben estar disponibles rápidamente cuando se necesitan (milisegundos)
- **Durabilidad:** Mínimo 3 AZs
- **Coste:** Almacenamiento más barato que Standard, pero se aplica tarifa de recuperación
- **Caso IA:** Backups, logs antiguos, datasets que se usan esporádicamente

S3 One Zone-Infrequent Access (S3 One Zone-IA)

- **Uso:** Similar a Standard-IA, pero para datos no críticos que son fácilmente reproducibles
- **Durabilidad:** Los datos se almacenan en **una única AZ**
- **Coste:** ~20% más barato que S3 Standard-IA
- **Caso IA:** Resultados intermedios de procesamiento, datos generados que pueden recrearse

AStordee



Clases de Almacenamiento - Archivador Digital (Glacier)

Diseñado para el archivado a largo plazo y el coste más bajo posible.

S3 Glacier Instant Retrieval

Acceso: Milisegundos, como S3 Standard-IA

Uso: Archivos que se necesitan raramente pero de forma inmediata (ej. archivos médicos)

S3 Glacier Flexible Retrieval

Acceso: Flexible, desde minutos (acelerado) hasta horas (bulk)

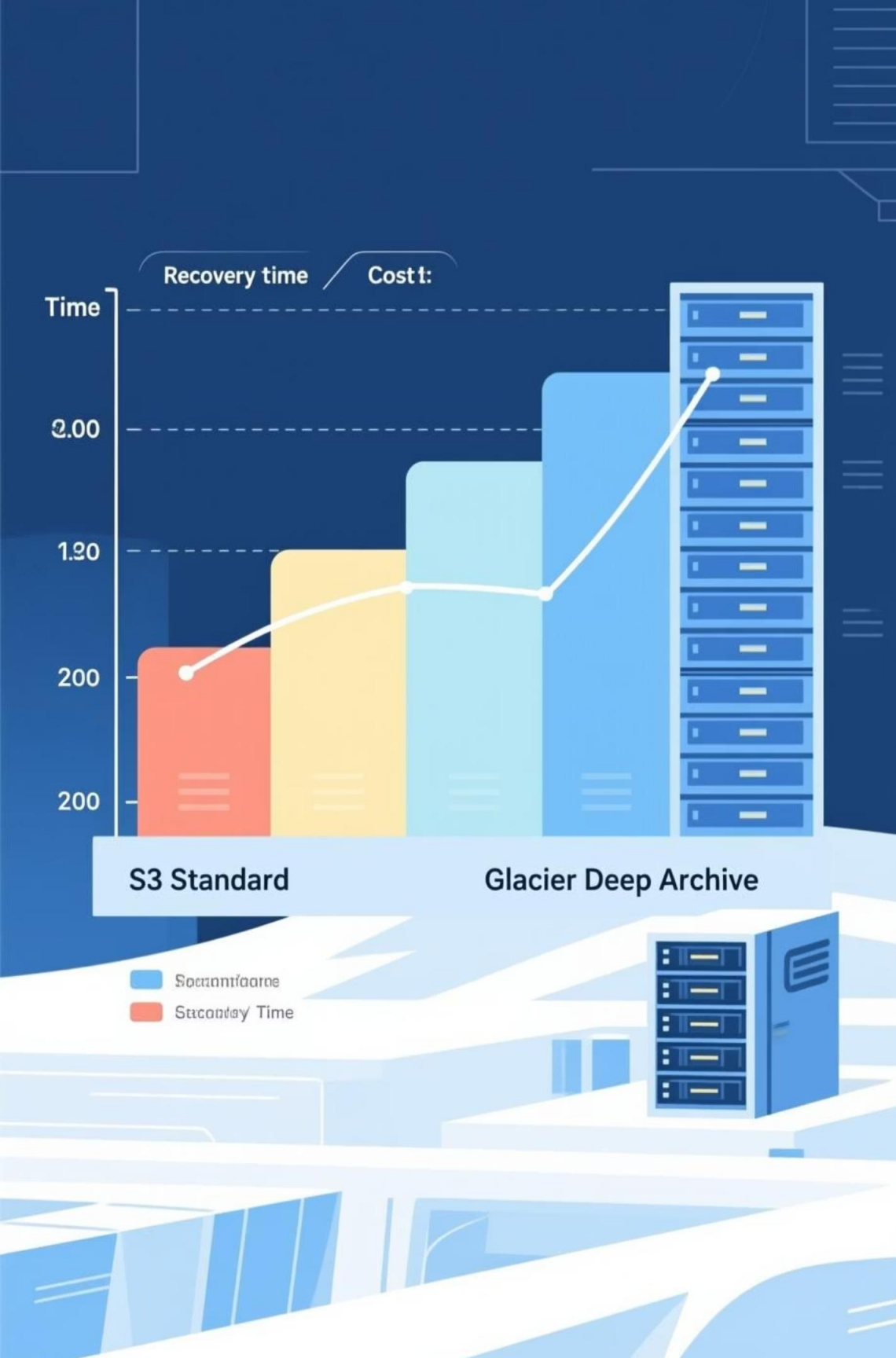
Uso: El archivado estándar. Backups a largo plazo donde un retraso en la recuperación es aceptable

S3 Glacier Deep Archive

Acceso: Lento, típicamente en 12 horas

Uso: El almacenamiento más barato de AWS. Para retención de datos regulatoria o archivado de datos archivado de datos que casi nunca se accederán

Caso IA: Archivar snapshots de modelos muy antiguos o logs de cumplimiento de hace años



Característica Avanzada: Versionado

¿Qué es?

Al activar el versionado en un bucket, S3 guarda todas las versiones de un objeto. No se sobrescribe, se añade una nueva versión.

Protección

- **Contra sobrescritura:** Si subes un fichero con la misma clave, la versión anterior se conserva
- **Contra eliminación accidental:** Si borras un objeto, S3 inserta un "marcador de eliminación" (*delete marker*), pero la versión anterior sigue existiendo

Consideraciones

- **Coste:** Pagas por el almacenamiento de todas las versiones guardadas
- **Implicación para IA:** Fundamental para la reproducibilidad. Puedes recuperar la versión exacta de un dataset o configuración con el que se entrenó un modelo específico



Característica Avanzada: Políticas de Ciclo de Vida

Automatización de costes

Permite definir reglas para gestionar el ciclo de vida de los objetos de forma automática.

Acciones posibles

- **Transición:** Mover objetos entre clases de almacenamiento después de un tiempo
- **Expiración:** Eliminar objetos (o versiones antiguas) permanentemente

Ejemplo de regla para IA

1. "Los logs (logs/) se almacenan en S3 Standard durante 30 días."
2. "Después de 30 días, se mueven a S3 Standard-IA."
3. "Después de 90 días, se archivan en S3 Glacier Flexible Retrieval."
4. "Después de 5 años, se eliminan permanentemente."



Característica Avanzada: Replicación

Copia automática y asíncrona de objetos entre buckets.

Same-Region Replication (SRR)

Replicación a un bucket dentro de la misma región de AWS.

Caso de uso:

- Agregar logs de diferentes cuentas a un bucket central
- Crear una copia de desarrollo/producción en la misma región
- Separar datos procesados de datos originales

Cross-Region Replication (CRR)

Replicación a un bucket en una región de AWS diferente.

Caso de uso:

- **Cumplimiento normativo:** Mantener copias de datos en geografías específicas
- **Minimizar la latencia:** Acercar los datos a usuarios en diferentes partes del mundo
- **Recuperación de desastres (DR):** Tener una copia de seguridad en otra región





Seguridad en S3 - ¡Crítico!

La seguridad en S3 es una responsabilidad compartida y se gestiona en múltiples capas.

Políticas de IAM

Definen permisos para usuarios, grupos y roles.

Ejemplo: "El usuario 'analista' puede leer (GetObject) del bucket 'datasets'".

Políticas de Bucket

Adjuntas al bucket. Definen quién puede acceder al bucket.

Útil para dar acceso entre cuentas o configurar acceso público controlado.

Listas de Control de Acceso (ACLs)

Mecanismo más antiguo y granular (a nivel de objeto).

Generalmente se prefieren las políticas de IAM y Bucket por ser más potentes.
potentes.

S3 Block Public Access

Configuración a nivel de cuenta y de bucket que actúa como un "paraguas de seguridad".

Activado por defecto para prevenir la exposición pública accidental.

Seguridad en S3 - Encriptación

Encriptación en Tránsito

Asegurada mediante SSL/TLS al usar HTTPS en las llamadas a la API.

Encriptación en Reposo (Server-Side Encryption - SSE)

- **SSE-S3:** AWS gestiona las claves de encriptación. Es la opción más sencilla
- **SSE-KMS:** AWS gestiona las claves a través de KMS. Ofrece un rastro de auditoría
- **SSE-C:** *Customer-Provided Keys*. El cliente gestiona sus propias claves

Encriptación en el Cliente

El usuario encripta los datos *antes* de subirlos a S3.

Máximo control, máxima responsabilidad.

Mejor Práctica para IA

Encriptar siempre datos sensibles, especialmente información personal que pueda estar en pueda estar en datasets de entrenamiento.



Automatización con Event Notifications

S3 puede generar notificaciones cuando ocurren eventos específicos en un bucket.

Eventos comunes

- `s3:ObjectCreated:*` (un objeto nuevo es subido)
- `s3:ObjectRemoved:*` (un objeto es eliminado)
- `s3:ObjectTagging` (se modifican las etiquetas)

Destinos de las notificaciones

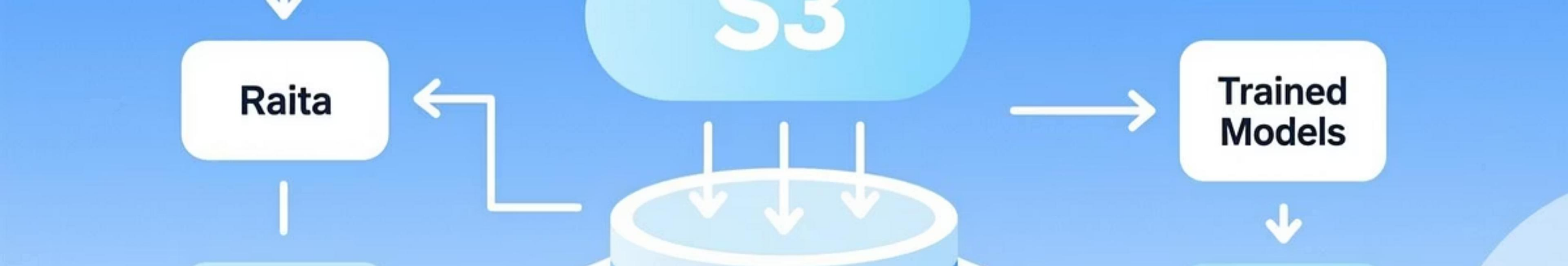
- **AWS Lambda:** El caso de uso más potente
- **Amazon SQS:** Enviar un mensaje a una cola
- **Amazon SNS:** Enviar una notificación a un tópico

Caso de Uso para IA

Iniciar un pipeline de inferencia o pre-procesamiento automáticamente cuando un nuevo dato llega al Data Lake.

Ejemplo: "Cuando una nueva imagen se suba al bucket entradas/, ejecuta una función Lambda para redimensionarla y guardarla en el bucket salidas/".





Casos de Uso de S3 en Inteligencia Artificial



Data Lakes

S3 es el pilar de la mayoría de los Data Lakes. Su escalabilidad y bajo coste permiten almacenar petabytes de datos brutos (logs, JSONs, Parquet, imágenes) de forma centralizada.



Almacenamiento de Modelos

Guardar los artefactos del modelo (ficheros .pkl, .h5, checkpoint) en S3 permite versionarlos, compartirlos y cargarlos en servicios de inferencia como SageMaker.



Logs y Métricas

Centralizar los logs de entrenamiento, inferencia y monitorización de sistemas de IA para su posterior análisis.



Hosting Web Estático

Servir documentación, demos de modelos (con tensorflow.js) o visualizaciones de datos directamente desde S3 a un coste muy bajo.

¿Qué es el Almacenamiento de Bloques?

Definición

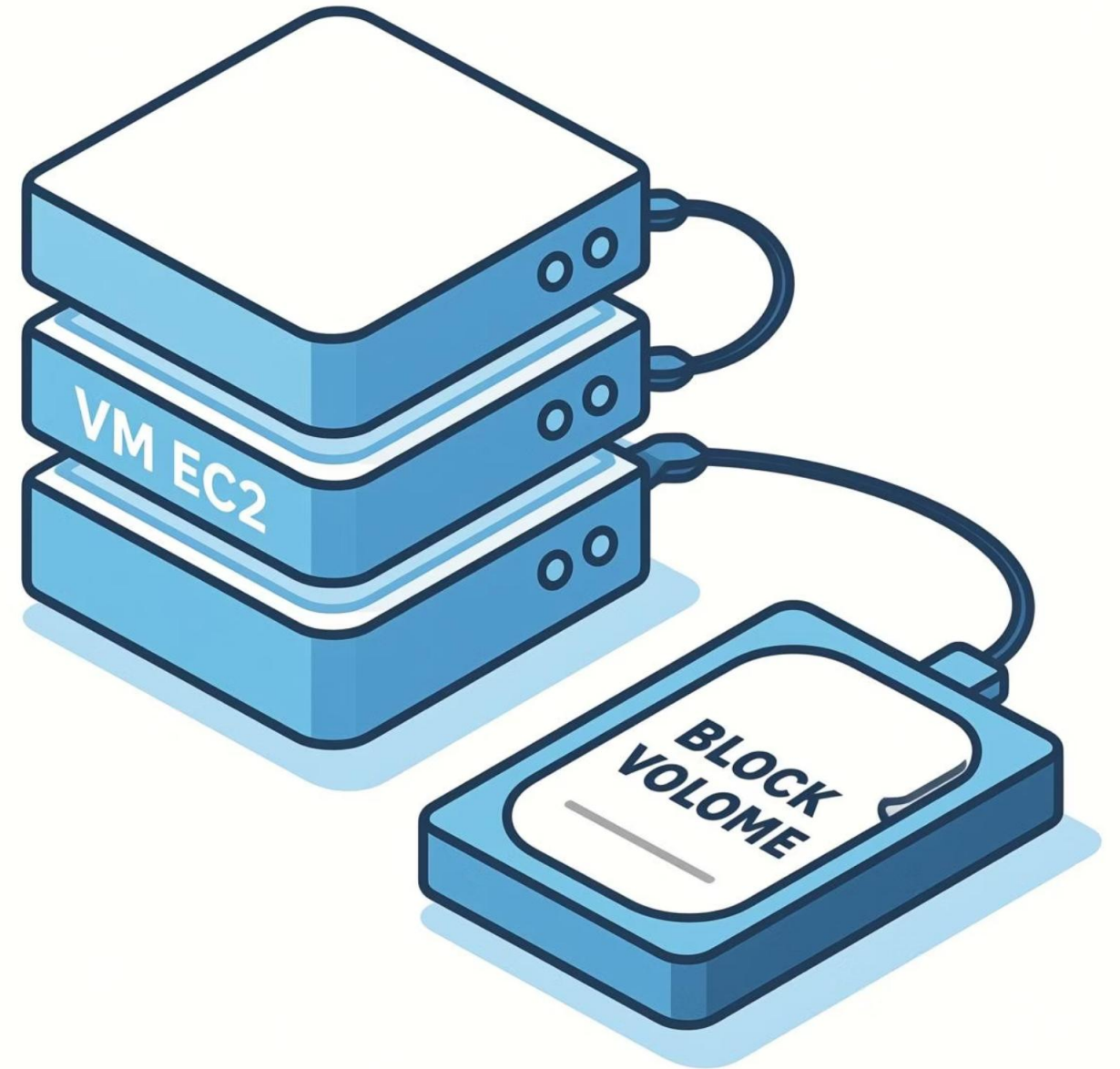
Proporciona volúmenes de almacenamiento (bloques) que se comportan como discos duros físicos.

Funcionamiento

Se "adjuntan" a una máquina virtual (servidor). El sistema operativo de la VM tiene control total sobre el volumen: puede formatearlo con un sistema de ficheros (ext4, ficheros (ext4, NTFS), instalar aplicaciones, bases de datos, etc.

Características

- **Persistente:** Los datos sobreviven al reinicio o terminación de la VM
- **Alto Rendimiento:** Diseñado para baja latencia y altas IOPS
- **Control Total:** El sistema operativo gestiona directamente el espacio
- **Ideal para:** Sistemas de ficheros, bases de datos transaccionales, discos de arranque



Introducción a Amazon EBS

EBS: Elastic Block Store

Es el servicio de almacenamiento de bloques de alto rendimiento para usar con instancias Amazon EC2.



Recurso de Red

Un volumen EBS es un **recurso de red**, no un disco físico conectado directamente al host. Esto permite flexibilidad y durabilidad.



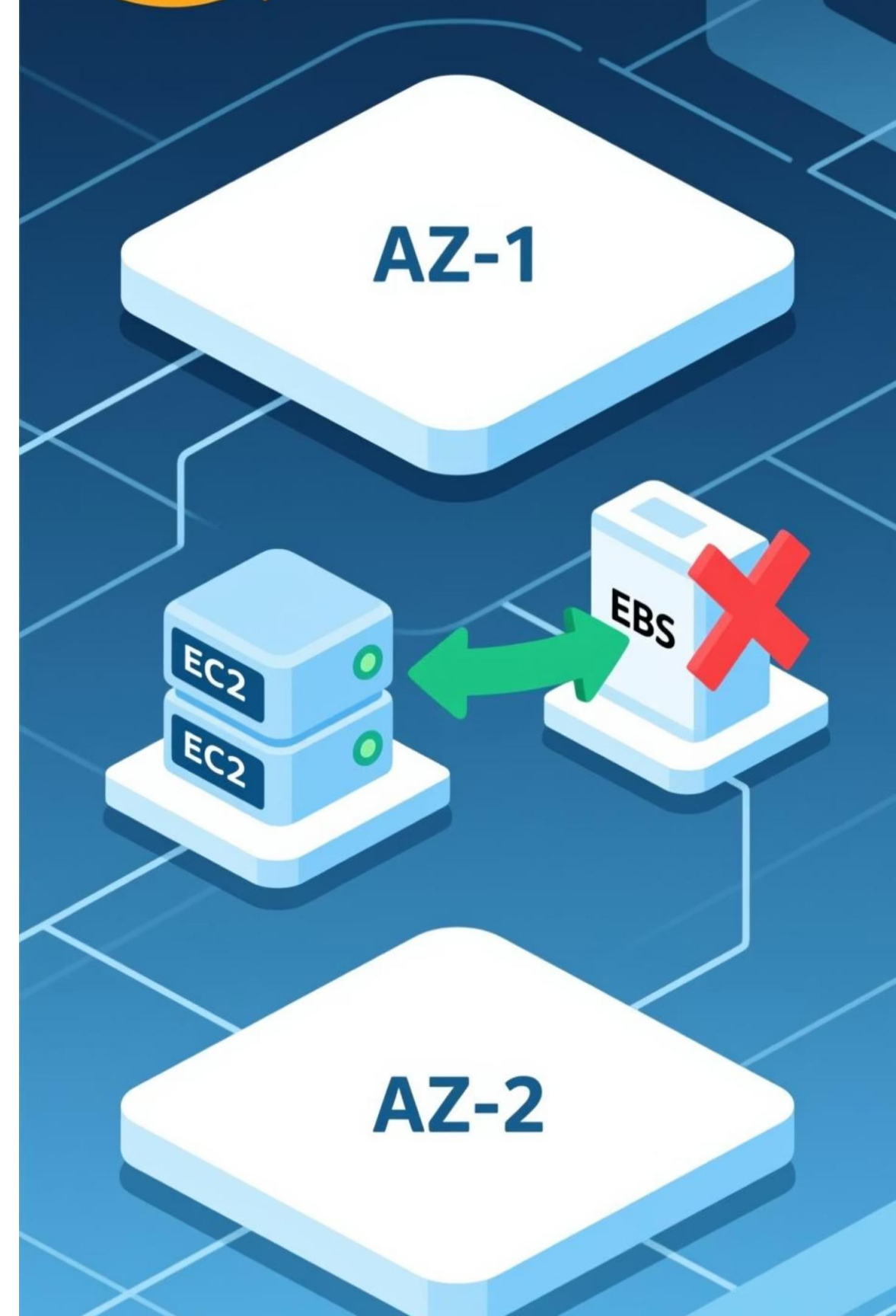
Ligado a una AZ

Un volumen EBS está ligado a una AZ específica. Solo puede ser adjuntado a una instancia EC2 que se encuentre en la **misma AZ**.



Persistencia

Independiente del ciclo de vida de la instancia. Puedes detener una instancia, instancia, desacoplar el volumen, y acoplarlo a otra instancia en la misma AZ. AZ.



Tipos de Volúmenes EBS (Basados en SSD)

SSD de Propósito General (gp2 / gp3)

- **Uso:** El tipo más común. Excelente equilibrio entre precio y rendimiento
- **Aplicaciones:** Discos de arranque, BBDD de tamaño pequeño-medio, entornos de desarrollo
- **Diferencia clave:** **gp3** es la generación más nueva y recomendada. recomendada. Permite provisionar IOPS y *throughput* (rendimiento) (rendimiento) de forma independiente al tamaño

SSD de IOPS Provisionadas (io1 / io2 / io2 Block Express)

- **Uso:** Para cargas de trabajo críticas que requieren el más alto rendimiento y consistencia
- **Aplicaciones:** Grandes bases de datos transaccionales (Oracle, SQL Server), aplicaciones sensibles a la latencia
- **Característica:** Garantiza un nivel específico de IOPS que tú defines y pagas





Tipos de Volúmenes EBS (Basados en HDD)

Optimizados para coste por GB, no para IOPS.

HDD Optimizado para Rendimiento (st1)

- **Uso:** Para cargas de trabajo secuenciales y de alto *throughput* (rendimiento)
- **Aplicaciones:** Big Data, Data Warehousing, procesamiento de logs
- **No es bueno para:** Cargas de trabajo con muchas lecturas/escrituras pequeñas y aleatorias
- **Caso IA:** Almacenamiento temporal para ETL de grandes datasets

HDD en Frío (sc1)

- **Uso:** El almacenamiento de bloques más barato
- **Aplicaciones:** Datos de acceso muy poco frecuente que necesitan ser un sistema de ficheros
- **Coste:** El más bajo \$/GB para almacenamiento de bloques
- **Caso IA:** Almacenamiento a largo plazo de datos intermedios que no requieren acceso rápido

Característica Clave: Snapshots de EBS

¿Qué son?

Son copias de seguridad de un volumen EBS en un punto en el tiempo.

Almacenamiento

Los snapshots se almacenan de forma redundante en [Amazon S3](#), lo que los hace muy duraderos.

Incrementales

El primer snapshot copia todo el volumen. Los snapshots posteriores solo guardan los bloques que han cambiado desde el último snapshot. Esto ahorra tiempo y dinero.

Funcionalidad

- **Restauración:** Puedes crear un nuevo volumen EBS idéntico a partir de un snapshot en cualquier AZ de esa región
- **Migración:** Puedes copiar un snapshot a otra región para crear un volumen allí
- **Caso IA:** Congelar y versionar el estado completo de un experimento de Machine Learning (modelo, código y *datasets*) en un solo punto del tiempo para garantizar la auditoría y la reproducibilidad de la investigación.



Casos de Uso de EBS en Inteligencia Artificial



Discos de Arranque para VMs

Cada instancia EC2 usada para desarrollo, entrenamiento o inferencia necesita un volumen EBS como disco de sistema operativo. Un volumen gp3 es ideal.



Almacenamiento para Bases de Datos

Si ejecutas una base de datos relacional (ej. PostgreSQL para metadatos de experimentos) directamente en una EC2, usarás EBS. Un volumen io2 puede ser necesario para bases de datos de datos de producción con alta carga.



Espacio de Trabajo Temporal

Para tareas de pre-procesamiento de datos en una VM que requieren lecturas/escrituras aleatorias de alta velocidad y un sistema de ficheros.



Software Específico

Algunas aplicaciones científicas o de IA están diseñadas para leer/escribir en un sistema de ficheros POSIX, haciendo que EBS sea la opción necesaria sobre S3.

Gamificación - Elige la Arquitectura Correcta

Eres el arquitecto de una startup de IA que procesa imágenes satelitales para detectar deforestación. Tu sistema tiene tres componentes de datos clave:

1. Ingesta de Datos

Recibes 10TB de imágenes satelitales (.tiff) al mes desde varias fuentes globales. Deben almacenarse de forma duradera y barata para su posterior procesamiento.

2. Base de Datos de Metadatos

Una base de datos PostgreSQL que corre en una instancia EC2. Almacena la geolocalización, fecha y estado de procesamiento de cada imagen. Requiere un acceso muy rápido y consistente.

3. Archivo a Largo Plazo

Las imágenes procesadas y sus resultados deben guardarse durante 7 años por motivos regulatorios, con un acceso muy infrecuente.

Pregunta para Mentimeter:

Para cada componente (1, 2 y 3), ¿qué servicio y tipo/clase de almacenamiento de AWS elegirías y por qué?

Solución al Escenario y Discusión

1. Ingesta de Datos (Imágenes (Imágenes Satelitales))

Servicio: [Amazon S3](#). Ideal para datos no estructurados, escalabilidad masiva y durabilidad.

Clase: [S3 Intelligent-Tiering](#) para empezar, ya que optimizará costes automáticamente. **S3 Transfer**

Acceleration podría usarse para acelerar la subida desde fuentes globales.

2. Base de Datos de Metadatos

Servicio: [Amazon EBS](#) adjunto a la instancia EC2. Las BBDD relacionales necesitan un sistema de ficheros y baja latencia.

Tipo: [EBS gp3](#) si la carga es moderada. **EBS io2** si la consistencia y el alto número de transacciones por segundo es crítico.

3. Archivo a Largo Plazo

Servicio: [Amazon S3](#).

Clase: Una política de ciclo de vida que mueva los datos a [S3 Glacier Deep Archive](#) para el coste de almacenamiento más bajo posible.

Comparativa - Almacenamiento de Objetos

Característica	AWS S3	Google Cloud Storage (GCS)	Azure Blob Storage
Contenedor	Bucket (nombre global)	Bucket (nombre global)	Container (en Storage Account Account)
Clases Frecuentes	Standard	Standard	Hot
Clases Poco Frec.	Standard-IA	Nearline	Cool
Clases Archivo	Glacier, Deep Archive	Coldline, Archive	Archive
Consistencia	Fuerte lectura-tras-escritura	Fuerte global	Fuerte
Característica Única	S3 Intelligent-Tiering	Buckets Multi-regionales	Tipos de Blob (Block, Append, Page)

Comparativa - Almacenamiento de Bloques

Característica	AWS EBS	Google Persistent Disk (PD)	Azure Disk Storage
Unidad	Volumen	Disco	Disco (Managed Disk)
Disponibilidad	Zonal (ligado a una AZ)	Zonal o Regional	Zonal (con opciones de redundancia)
SSD Equilibrado	gp3	Balanced PD	Standard SSD
SSD Rendimiento	io2	SSD (Performance) PD	Premium SSD, Ultra Disk
Característica Única	Elastic Volumes (modificar en caliente)	Regional Persistent Disks (replicación síncrona)	Ultra Disk (rendimiento sub-milisegundo configurable)

Consideraciones de Rendimiento

Latencia

El tiempo que tarda en comenzar la primera operación de datos. Crítico para BBDD y sistemas interactivos.

EBS está diseñado para baja latencia. **S3** tiene una latencia ligeramente mayor.

Más importante para aplicaciones sensibles al tiempo como sistemas interactivos de IA.

Throughput (Rendimiento)

El ancho de banda de datos (MB/s). Importante para procesar grandes ficheros secuencialmente.

Los volúmenes **EBS st1** están optimizados para esto. **S3** ofrece un *throughput* masivamente escalable.

Esencial para la ingesta y procesamiento de grandes datasets de imágenes o vídeo.

IOPS (Operaciones de E/S por segundo)

El número de operaciones de lectura/escritura por segundo. Crítico para cargas de trabajo transaccionales.

Los volúmenes **EBS io2** están optimizados para esto.

Clave para bases de datos que almacenan metadatos de experimentos o características de modelos.

Consideraciones de Coste

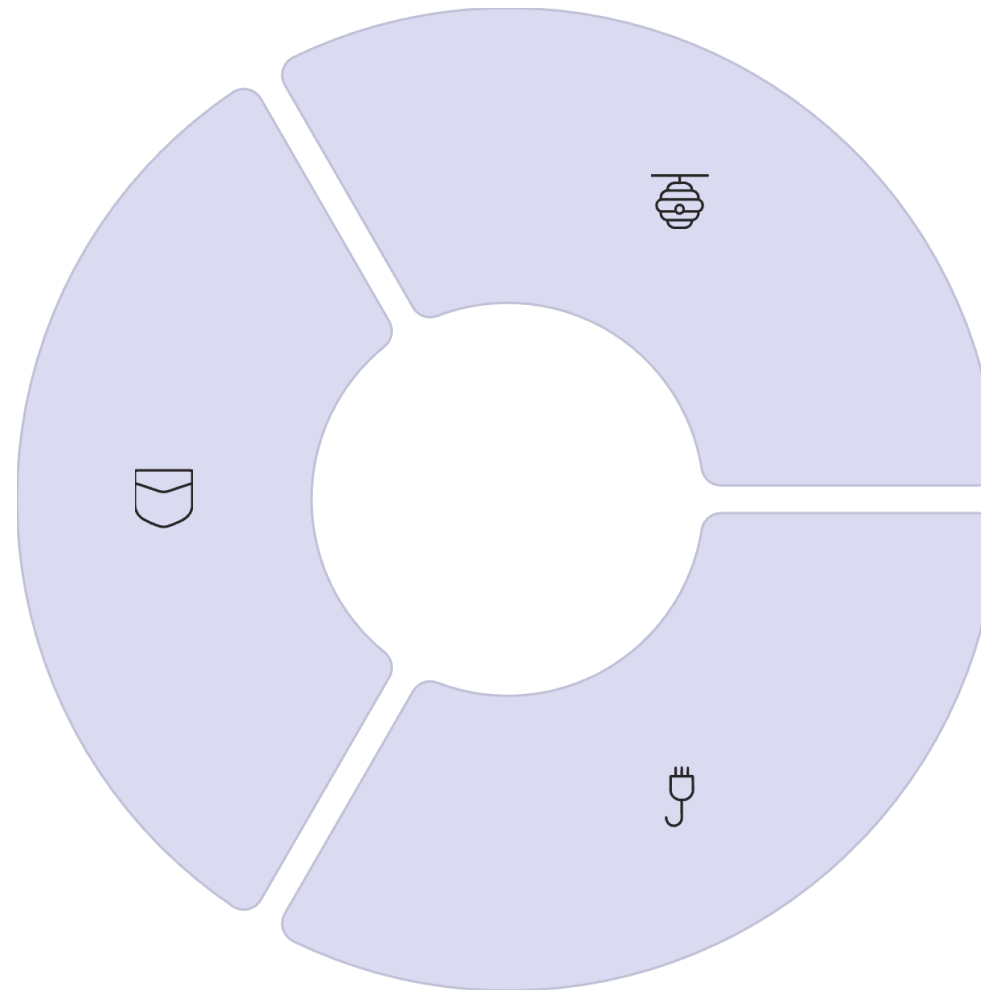
El coste en la nube nunca es una sola cifra. Debes considerar tres ejes:

Coste de Almacenamiento

Lo que pagas por los datos en reposo, medido en
\$/GB/mes.

Varía enormemente entre clases (S3 Standard vs
vs Glacier Deep Archive).

Estrategia: Usar políticas de ciclo de vida para mover
datos automáticamente a clases más baratas según
envejecen.



Coste de Operaciones

Lo que pagas por interactuar con los datos. Medido en
\$/1000 peticiones (GET, PUT, LIST).

En S3, las clases de acceso poco frecuente tienen un
un coste de operación (recuperación) más alto.

Estrategia: Considerar la frecuencia de acceso al elegir
la clase adecuada.

Coste de Transferencia

- Transferencia **entrante** (hacia la nube):
Generalmente **gratis**
- Transferencia **saliente** (desde la nube hacia
Internet): **Nunca es gratis**
- Transferencia entre regiones: Tiene coste
- Dentro de la misma región: Generalmente gratis en
gratis en la misma AZ

Resumen de la Sesión

Almacenamiento de Objetos (S3)

- Para datos **no estructurados** a cualquier escala
- Acceso vía **API**. Duradero, escalable y coste-efectivo
- El **corazón del Data Lake** moderno para IA
- Clases: Standard, IA, Glacier, adaptadas a diferentes patrones de acceso
- Características clave: Versionado, Políticas de Ciclo de Vida, Replicación

Almacenamiento de Bloques (EBS)

- Para datos **estructurados** que necesitan un sistema de ficheros
- Actúa como un **disco duro de red** para tus VMs (EC2)
- El **fundamento para BBDD** y cargas de trabajo de alto rendimiento
- Tipos: gp3 (equilibrado), io2 (rendimiento), st1 (throughput), sc1 (frío)
- Características clave: Snapshots, Elastic Volumes

La Elección Correcta: La clave es entender la naturaleza de tus datos y tu patrón de acceso para elegir el servicio y la clase/tipo correctos, optimizando tanto el rendimiento como el coste.

Próximos Pasos y Preguntas

Próxima Sesión

"Servicios de Almacenamiento II:
Almacenamiento de Ficheros (EFS) y
Bases de Datos Especializadas para IA".

Lecturas Recomendadas

- Whitepaper de AWS: "Storage for AI/ML workloads"
- Documentación de S3 Storage Classes y EBS Volume Types
- AWS re:Invent: Deep Dive on Amazon EBS (última edición)

Recursos adicionales

- Videos didácticos:
 - "AWS S3 Explained in 15 minutes" (TechWorld with Nana)
 - "Deep Dive on Amazon EBS" (AWS Events)
- Laboratorios prácticos en AWS Academy

¡Gracias!

Q&A Abierto

