Sesión 4: Servicio de almacenamiento

Introducción

Los servicios en la nube deben cubrir una gran variedad de necesidades. En la gran mayoría de los casos de uso, necesitaremos almacenar datos, pero no siempre tendremos los mismos requisitos de espacio, disponibilidad, velocidad, etc. Para intentar ajustarse a nuestras necesidades, AWS ofrece distintos servicios de almacenamiento (también lo hace el resto de proveedores de tecnologías cloud).

Pongamos como ejemplo algunos casos de uso típicos:

- Unidades de almacenamiento para máquinas virtuales EC2.
- Almacenes de datos para análisis Big Data.
- Bases de datos.
- Copias de seguridad.
- Registros de dispositivos IoT (Internet of Things).
- Almacenamiento para albergar una página web estática.

En algunos de estos casos de uso, tener una gran velocidad de acceso será prioritario. Para otros, el coste de almacenamiento a largo plazo puede ser el factor clave que decida la viabilidad del proyecto. En cualquier caso, es necesario conocer los distintos tipos de servicio de almacenamiento ofrecidos por nuestro proveedor para poder seleccionar aquel que mejor se adapte a nuestras necesidades.

Al final de la segunda sesión ya realizamos una introducción a los distintos tipos de servicios de almacenamiento ofrecidos por AWS. Para preparar la actividad práctica de esta cuarta sesión, vamos a profundizar un poquito más en dichos servicios, sus diferencias y casos de uso típicos.

Antes de pasar a ver cada uno de estos tipos, es necesario realizar una distinción entre almacenamientos en bloque y almacenamientos de objetos. La diferencia entre ambos es que los almacenamientos en bloque permiten acceder y modificar cada byte almacenado mientras que los almacenamientos de objetos NO lo permiten. En almacenamiento de objetos, se guarda el objeto entero, en una transacción. Para modificar un objeto es necesario volver a subir el objeto entero (aunque sólo se haya modificado un byte).

Amazon Elastic Block Store (EBS)

Este es el almacenamiento que hemos utilizado en las prácticas anteriores. Es la opción por defecto para los volúmenes de las máquinas virtuales EC2. Permiten que los datos no sean eliminados cuando la máquina se apaga. Como su nombre indica, se trata de un almacenamiento en bloques (permite acceder a cada byte de cada fichero para su lectura o modificación).

Este servicio de almacenamiento cuenta con las siguientes características:

- **Replicación automática** dentro de la zona de disponibilidad.
- **Alta disponibilidad** y durabilidad.
- Rendimiento uniforme de **baja latencia**.
- Permite c**opias de seguridad automáticas** en S3 (llamadas instantáneas).
- Se puede añadir una **capa de cifrado de datos** sin coste adicional.
- Es posible cambiar de tipo de unidad o aumentar la capacidad contratada de forma sencilla (son **elásticos**).

Y sus usos más comunes son:

- Unidades de arranque y almacenamiento de instancias EC2.
- Hosts de bases de datos.
- Almacenamiento genérico de datos con estructura de ficheros y directorios.
- Almacenes de datos para procesado por lotes en IA.

Cuando utilizamos este servicio, podemos seleccionar el tipo de unidad de almacenamiento que AWS aprovisionará por nosotros. Existen para ello dos grandes categorías: volúmenes en unidades SSD y volúmenes en unidades HDD. En líneas generales, las unidades SSD ofrecen un mejor rendimiento, es decir:

1. Mayor velocidad de transferencia: hasta 4000 MiB/s en SSD y hasta 500 MiB/s en HDD.
2. Mayor número de operaciones de entrada salida por segundo (IOPS **I**nput/**O**utput operations **P**er **S**econd): hasta 256000 en SSD y hasta 500 en HDD.

Dentro de estos dos tipos existen también distintas modalidades para afinar todavía mejor la relación coste/rendimiento a las necesidades de nuestro proyecto. Podéis consultar las diferencias entre todos los tipos disponibles, así como el coste asociado a cada uno de ellos en el siguiente enlace:

Tipos de volúmenes EBS

Amazon Simple Storage Service (S3)

Amazon creó este tipo de almacenamiento con el objetivo de que fuese lo más sencillo posible. Al mismo tiempo, los datos almacenados debían ser accesibles desde cualquier ubicación: otros servicios de AWS, sitios web, aplicaciones móviles, etc.

A partir de esta premisa, el equipo de AWS entendió que la forma más sencilla de conseguir sus objetivos era plantear el servicio como un almacén de objetos, que serían escritos o leídos en su totalidad, sin posibilidad de realizar modificaciones parciales.

Esta decisión también tiene implicaciones en la jerarquía de los datos. En lugar de una estructura de directorios tradicional, los objetos de S3 se almacenan en **buckets** (cubos, aunque nadie utiliza el término en español). Un bucket puede contener un **número** virtualmente **ilimitado de objetos**, y cada objeto puede tener un **tamaño de hasta 5 TB**. También es posible almacenar los datos cifrados en S3, de forma totalmente transparente para nosotros (al recuperar los datos serán descifrados en el servidor antes de llegar al usuario).

S3 es un servicio totalmente gestionado (fully managed), lo que quiere decir que, como usuarios, no tenemos que preocuparnos de nada que no sean los datos en sí. Amazon se ocupa:

- De la fiabilidad de los datos (fiabilidad > 99,99999 %, o fiabilidad de cinco nueves). Amazon se encarga de realizar comprobaciones rutinarias, de realizar copias de seguridad y de todo lo necesario para garantizar esta fiabilidad.
- De todo el hardware necesario. No necesitamos saber si hay discos SSDs, HDDs, cintas magnéticas, ni en qué máquinas están montados. Tampoco tenemos que crear ninguna arquitectura adicional (como VPCs, subredes o máquinas EC2). Creamos un bucket, metemos un objeto, configuramos su política de seguridad y ya podremos acceder al objeto desde donde queramos y cuando queramos. La única decisión que debemos tomar es la región donde se creará el bucket.
- De registrar eventos y accesos a los objetos. Lo único que tenemos que hacer es establecer una política de seguridad, para decidir quien tiene permisos para ver o modificar nuestros objetos (subir nuevas versiones completas). Amazon se encargará de garantizar que se cumple la política de seguridad y de dejar un registro de cada uno de los accesos realizados a nuestros datos. También permite configurar eventos para que AWS nos avise cuando un bucket sufre un cambio.
- De la escalabilidad de la solución. No tenemos que preocuparnos por cuántos usuarios accederán a nuestros datos. S3 monitoriza los accesos y escala el hardware subyacente de forma automática para acomodar la arquitectura a la demanda en cada momento.

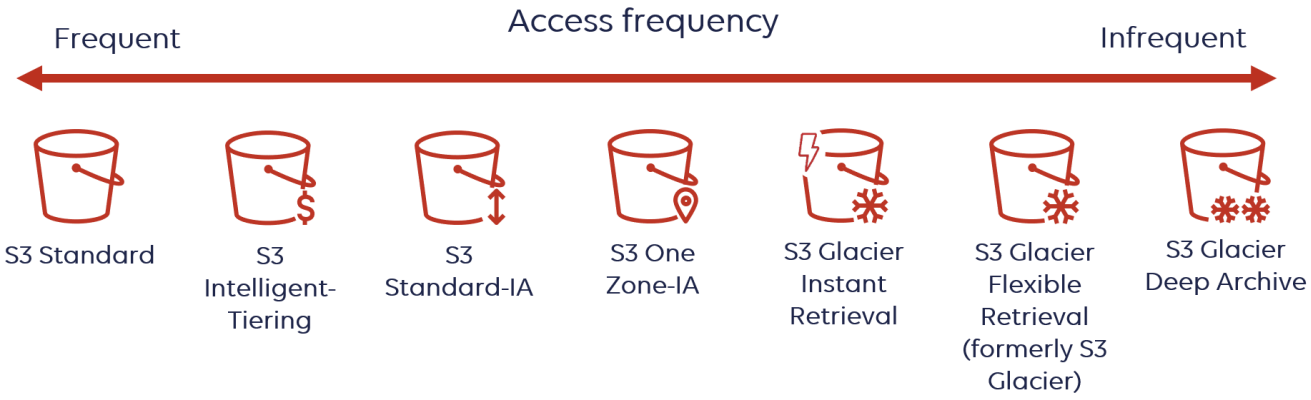
Es importante tener en cuenta que los **nombres de los buckets son universales** y, por tanto, deben ser **únicos a nivel global**. Esto es debido a que el nombre del bucket forma parte de la ruta de acceso a sus datos. También hay que tener en cuenta la región en la que creamos cada bucket, ya que debemos alojarlos cerca de los consumidores de nuestros datos, para minimizar la latencia de acceso. Esta decisión hay que pensarla bien desde el principio, porque **NO es posible cambiar la región de un bucket**. Si tuviésemos esta necesidad, tendríamos que copiar el contenido a un nuevo bucket en la nueva región.

Para adaptarse a los distintos casos de uso, S3 ofrece distintas **clases de almacenamiento**, cada una con unas características únicas y un modelo de costes distinto. El factor clave para decidirnos por una u otra clase es la frecuencia de acceso a los datos.

- **Amazon S3 Standard**. Para datos a los que se accede con mucha frecuencia. Garantiza alta disponibilidad, alta durabilidad y óptimo rendimiento. Casos de uso típicos: sitios webs dinámicos, aplicaciones en la nube, distribución de contenido, videojuegos, procesamiento en streaming para Big Data, etc.
- **Amazon S3 Standard IA** (S3 Infrequent Access). Esta clase ofrece un gran rendimiento y alta disponibilidad, pero con un modelo de coste distinto que favorece los casos de uso en el que el acceso a los datos es menos frecuente. Es ideal para almacenamiento de respaldo que requieran una restauración inmediata como archivos de recuperación ante desastres.
- **Amazon S3 One Zone**. Esta clase ofrece un precio aún más económico que la anterior a costa de bajar los estándares de fiabilidad, puesto que las copias de seguridad automáticas sólo se realizan dentro de una zona de disponibilidad. Se suele utilizar para copias de seguridad secundarias, o para datos que se puedan volver a generar fácilmente en caso de pérdida.
- **Amazon S3 Glacier y Amazon S3 Glacier Deep Archive**. En estas dos clases se juega con el tiempo de recuperación de los datos para obtener un precio de almacenamiento más económico. La lectura de los datos almacenados en un S3 de estas clases se realiza en dos fases: se realiza una solicitud de recuperación y, pasado un tiempo, los datos ya están disponibles para ser leídos. Esto permite a Amazon utilizar tecnologías de respaldo como cintas magnéticas, mucho más económicas que las unidades de disco duro o unidades SSD, pero que requieren un proceso mecánico de carga de la cinta (o las cintas) en robots de lectura. Dentro de estas clases se ofrecen distintas opciones que van desde la recuperación de datos en pocos minutos a varias horas, teniendo un impacto directo en el coste asociado a cada operación. Son las clases más adecuadas para copias de seguridad a largo plazo, o para copias legales (la ley a veces exige guardar datos un determinado periodo de tiempo, aunque rara vez acabamos accediendo a ellos).
- **Amazon S3 Intelligent-Tiering**. Esta clase, a cambio de un coste asociado por la gestión, permite cambiar automáticamente los datos entre el resto de clases vistas anteriormente. S3 monitoriza el acceso a los datos y los mueve a capas más económicas si la frecuencia de acceso es baja o a capas más caras si la frecuencia es alta.

Nota aclaratoria: En la próxima sesión estudiaremos más en detalle el control de coste de los servicios de AWS. En S3, cada clase tiene dos elementos que se cobran: el acceso a los datos y la cantidad de datos almacenada. En las clases diseñadas para un acceso más frecuente el acceso a los datos es más económico pero el almacenamiento en sí es más caro. En el lado opuesto, las clases diseñadas para un acceso más infrecuente tienen un coste de almacenamiento muy barato pero un acceso a los datos más caro.

En la siguiente imagen se puede ver un resumen de las distintas clases, ordenadas según la frecuencia de acceso esperada:



Para más información sobre las diferencias entre las distintas clases de S3, podéis acudir a la documentación oficial de AWS:

Clases de S3

Amazon Elastic File System (EFS)

El servicio EFS (sistema de ficheros elástico) está pensado para cumplir con dos objetivos principales: poder **escalar en capacidad automáticamente** y que sea sencillo **compartir** el volumen entre distintas instancias EC2. Se trata de un servicio totalmente administrado (fully managed), por lo que AWS se encargará de gestionar toda la infraestructura necesaria.

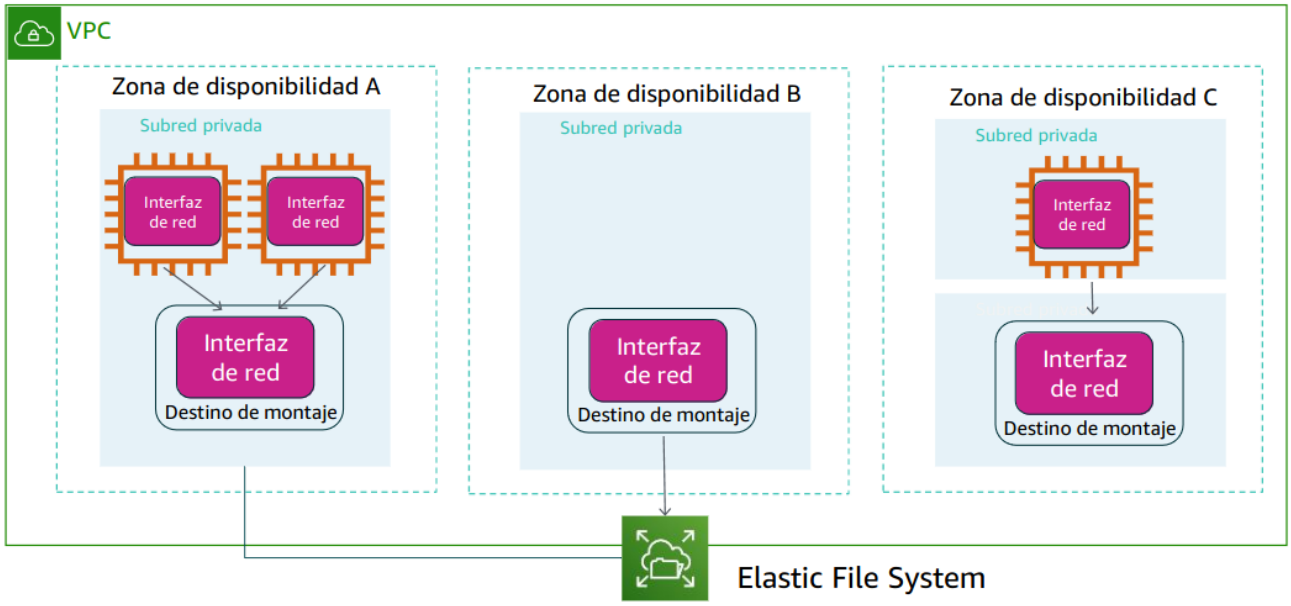
EFS es un sistema de almacenamiento por bloques con sistema de ficheros en jerarquía de directorios, al igual que el servicio EBS que vimos anteriormente. Los volúmenes EBS tienen un tamaño determinado que podemos cambiar en cualquier momento **de forma manual**, por lo que debemos estar atentos al espacio libre restante y actuar en caso de que nos quedemos sin espacio. En cambio con EFS, no seleccionamos una capacidad al generar un volumen, tienen un **espacio virtualmente infinito** (actualmente el límite está en la escala de los Petabytes) y, al mismo tiempo, sólo hay que pagar por el espacio realmente utilizado. Internamente Amazon realizará los cambios hardware oportunos cuando sea necesario, sin interrumpir el servicio en ningún momento.

EFS se implementa como un sistema de ficheros compartidos mediante el **protocolo NFS** (Network File System, versiones 4.0 y 4.1). La infraestructura de AWS garantiza una baja latencia en el acceso a los datos. Un mismo volumen EFS puede estar conectado a miles de instancias y seguir ofreciendo un rendimiento óptimo.

Este servicio **sólo está disponible para las AMIs de EC2 basadas en Linux**.

Las instancias que queramos conectar a un volumen EFS deben estar en la misma VPC. Además, es necesario crear un **destino de montaje** (access point) para cada zona de disponibilidad donde

queramos conectar instancias. Aunque tengamos varias subredes en una zona de disponibilidad sólo será necesario crear un destino para toda la zona.



Para reforzar todos los contenidos aprendidos hasta ahora en el curso, durante esta sesión y la siguiente desarrollaremos un caso práctico en el que se verán involucrados varios servicios de AWS. Se trata de un servidor NextCloud, que alojaremos en la nube. NextCloud es un programa de código abierto para sincronizar y compartir archivos en la nube. Cuando lo tengamos terminado deberíamos tener un servidor con una infraestructura de protección adecuada, un sistema de ficheros automáticamente escalable y una base de datos robusta y flexible. Hemos separado esta práctica en varios partes, para hacerla más accesible. En esta sesión os planteamos las dos primeras, una de creación de la infraestructura y otra para gestionar el almacenamiento de la solución. En Aules están disponibles las guías de cada parte.

Actividad práctica 4.1.

Preparar un servidor NextCloud con una infraestructura de acceso protegido.

Actividad práctica 4.2

Migrar los datos del servidor NextCloud a un sistema de ficheros EFS.