# ¿Qué es Amazon EC2?

Amazon Elastic Compute Cloud (Amazon EC2) proporciona capacidad de computación escalable en la nube de Amazon Web Services (AWS). El uso de Amazon EC2 elimina la necesidad de invertir inicialmente en hardware, de manera que puede desarrollar e implementar aplicaciones en menos tiempo. Puede usar Amazon EC2 para lanzar tantos servidores virtuales como necesite, configurar la seguridad y las redes, y administrar el almacenamiento. Amazon EC2 le permite escalar hacia arriba o hacia abajo para controlar los cambios en los requisitos o los picos de popularidad, con lo que se reduce la necesidad de prever el tráfico.

*Por defecto, el límite máximo de instancias permitido que pueden ejecutarse al mismo tiempo es 20. Este límite puede ser ampliado por AWS si se lo solicita*

## Características

* Entornos informáticos virtuales (instancias)
* Plantillas preconfiguradas para las instancias (AMI – Amazon Machine Images). Empaquetan el software necesario para el servidor (sistema operativo y software adicional)
* Varias configuraciones de CPU, memoria, almacenamiento y capacidad de red de las instancias (tipos de instancias)
* Información de inicio de sesión segura para las instancias con pares de claves
* Volúmenes de almacenamiento para datos temporales que se eliminan cuando una instancia se detiene, se termina o se pone en hibernación (volúmenes de almacen de instancias – Instance Store Volumes)
* Volúmenes de almacenamiento persistente para los datos usando Amazon Elastic Block Store (Amazon EBS)
* Varias ubicaciones físicas para los recursos, como las instancias, los volúmenes de EBS, conocidas como regiones y zonas de disponibilidad
* Un firewall que permite especificar los protocolos, los puertos y los rangos de direcciones IP que pueden alcanzar las instancias mediante el uso de grupos de seguridad.
* Direcciones IPv4 estáticas, conocidas como Elastic IPs
* Metadatos – etiquetas que se pueden crear y asignar a los recursos de EC2
* Redes virtuales que se pueden crear y que estén aisladas lógicamente del resto de la nube de AWS y que, opcionalmente, se puede conectar a mi propia red, conocidas como nubes virtuales privadas (VPC)

## Acceso a Amazon EC2

* Interfaz de Linea de Comandos (CLI)
* Herramientas para Windows PowerShell
* EC2 permite la creación de recursos utilizando AWS Cloud Formation. Mediante plantillas JSON o YAML que describan los recursos AWS, Cloud Formation aprovisiona y configura esos recursos por nosotros. Se pueden utilizar estas plantillas para aprovisionar estos recursos varias veces, en la misma región y cuenta o en varias regiones y cuentas.
* EC2 ofrece una API de consulta. Estas solicitudes HTTP o HTTPS utilizan los verbos GET y POST de HTTP y un parámetro de consulta denominado Action (*se recomienda utilizar un SDK ya que automatiza tareas como la firma criptográfica de solicitudes, tratamiento de las respuestas de error y reintentos automáticos)*

## Acceso por SSH

Primero, restringir permisos del archivo pem (llave privada)

>> chmod 0400 path/to/archive.pem

En la terminal:

>> ssh -i path/to/archive.pem [ec2-user@138.235.123.111](mailto:ec2-user@138.235.123.111)

## Grupos de Seguridad

Los grupos de seguridad actúan como Firewalls para las instancias asociadas al controlar el tráfico entrante y saliente en el ámbito de la instancia. (Mínimo se debe agregar una regla al firewall que me permita conectarme a mis instancias mediante SSH).

1. Pueden ser adjuntados a múltiples instancias
2. Se encuentran bloqueados a una combinación de Región/VPC específica (por lo que si se cambia de región o de VPC se debe crear otro grupo de seguridad)
3. Es buena práctica mantener un grupo de seguridad solo para el acceso por SSH
4. Se alojan fuera de EC2 (como si fuera el portforwarding del router), por lo que, si cierto tráfico está bloqueado, la instancia ni siquiera lo ve
5. Los errores de Time Out en la conexión SSH, son debido a una mala configuración del grupo de seguridad.
6. Si el error es de Connection Refused, es debido a un error de aplicación dentro de la instancia o no está lanzada

***Si se pretende lanzar instancias en varias regiones, se deberá crear un grupo de seguridad por cada región***

## Referenciando a otros Grupos de Seguridad

## EC2 User Data

* Es posible automatizar algunos comandos en cada instancia que iniciemos, agregando un script en *User Data*
* Este script solo corre una vez, cuando la instancia inicia
* Sirve para automatizar tareas de inicio como por ej
* Instalar actualizaciones
* Instalar software
* Descargar archivos en común
* Cualquier otra cosa
* El script de User data, corre con permisos de usuario root

## Tipos de Lanzamiento EC2

### On demand

Facturación predecible, cargas de trabajo cortas (menos de 1 año).

* Se factura por segundo después del primer minuto
* Tiene el costo mas alto, pero no tiene pago por adelantado
* No existe un compromiso a largo plazo. Por lo que se puede detener y terminar la instancia cuando se desee sin cargo extra.
* **Es aconsejable, cuando el comportamiento de mi aplicación es impredecible y por ende se realiza auto-escalado**

### Instancias Reservadas

Cargas de trabajo prolongadas (mayor o igual a 1 año), por ej: correr una base de datos durante 1 año o 3 años

* 75% de descuento en comparación con On demand
* El pago es por adelantado por lo que existe el compromiso a largo plazo (1 a 3 años)
* Se reserva un tipo específico de instancia
* Se debe estar muy seguro de las necesidades o requerimientos

### Instancias Reservadas Convertibles

Se puede convertir la instancia a cualquier otro tipo de instancia (en tamaño – características) y cuando la carga de trabajo es de larga duración.

* Es una variación del tipo anterior, ya que permite cambiar el tipo de instancia, pero sigue siendo un compromiso a largo plazo. El descuento es del 54%

1. Instancia de Reserva Programada

Se usa cuando se sabe que algo sucederá cada semana, pero solo durante un día o durante una hora especifica. Se reserva algo de capacidad dentro de una ventana de tiempo.

### Instancias Spot

Para cargas de trabajo muy cortas, son muy baratas, pero se tiene el riesgo de perder instancias

* Descuento de hasta el 90% comparado con On-demand
* Se oferta un precio a pagar por la instancia y si alguien (otro cliente) oferta un precio mayor, la instancia se pierde
* Hay una advertencia de 2 minutos antes de perder la instancia
* Se recomienda usar cuando se necesita procesar datos por lotes, análisis de big data o cargas que pueden fallar
* Nunca utilizar para bases de datos

### Instancia Dedicada

Ningún otro elemento de AWS puede compartir el hardware subyacente en el Datacenters con nosotros. Por defecto, al iniciar una instancia EC2, se comparten recursos de hardware con otros clientes

* Solo el hardware es dedicado pero no se tiene control sobre el mismo
* Instancias lanzadas desde la misma cuenta pueden terminar en el mismo hardware
* Al no tener control de la ubicación de la instancia, al apagarla y volver a encenderla, la instancia se puede alojar en otro hardware.

### Host Dedicado

Se reserva un servidor físico completo. La diferencia con el anterior, es que se puede controlar la ubicación de la instancia. (ubicación en un hardware específico)

* Se tiene total control del hardware del servidor
* Visibilidad de los zócalos de hardware y los núcleos físicos (acceso directo a la CPU)
* Periodo de reserva de 3 años
* Son mucho más costosos
* Se usa cuando algunos proveedores usan el modelo BYOL (bring your own license) o para una compañía muy estricta en la seguridad de los datos

## EC2 Pricing

El precio de EC2 se da por horas. Pero este precio cambia según ciertos parámetros como por ej:

* La región
* El tipo de instancia
* El tipo de lanzamiento (on-demand, spot, dedicated, reserved)
* El sistema operativo utilizado
* Se factura por segundo después del primer minuto
* Almacenamiento
* Transferencia de datos
* Tipo de IP utilizada
* Equilibrio de carga (Load balancing)
* Otros

## Prácticas recomendadas de EC2

### Seguridad

* + Administrar el acceso a los recursos y las API de AWS mediante, identidad federada, usuarios y roles de IAM, definir políticas y procedimientos de administración de credenciales para crear, distribuir, rotar y revocar credenciales de acceso de AWS.
  + Implementar las reglas menos permisivas en el grupo de seguridad
  + Corregir, actualizar y proteger con regularidad el sistema operativo y las aplicaciones de la instancia.

### Almacenamiento

* Entender implicaciones del tipo de dispositivo raíz para la persistencia, backup y recuperación de datos.
* Utilizar volúmenes de EBS diferentes para los datos y el sistema operativo. Comprobar la persistencia de los datos en el volumen EBS luego de terminar la instancia
* Utilizar el almacén de instancias disponible para almacenar datos temporales. Estos se eliminan luego de finalizar o poner en hibernación la instancia.
* Cifrar volúmenes y snapshots de EBS.

### Administración de recursos

* Utilizar metadatos de las instancias y las etiquetas de los recursos personalizadas para realizar el seguimiento de mis recursos en AWS e identificarlos
* Visualizar límites actuales para EC2. Planificar la solicitud del aumento de los límites, antes del momento en que los necesite.

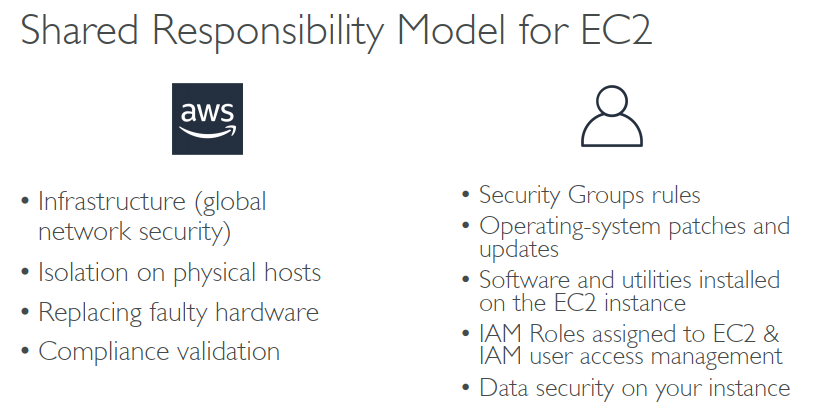
### Copia de seguridad y recuperación

* Realizar con regularidad una copia de seguridad de los volúmenes de EBS utilizando Snapshots de Amazon EBS y creando una imagen de máquina de Amazon a partir de la instancia para guardar la configuración como una plantilla para lanzar instancias en el futuro.
* Implementar componentes fundamentales de mi aplicación en varias zonas de disponibilidad y replicar mis datos de forma adecuada.
* Diseñar aplicaciones para que gestionen las direcciones IP dinámicas cuando mi instancia se reinicie.
* Comprobar que se está preparado para gestionar una conmutación por error.

1. Solución básica: Conectar manualmente una interfaz de red o una dirección IP elástica a una instancia de sustitución.
2. Solución automática: utilizar Amazon EC2 Auto Scaling

### Networking

* Establecer el valor de tiempo de vida (TTL) para las aplicaciones en 255, para IPv4 e IPv6. Si se utiliza un valor menor, existe el riesgo de que el TTL caduque mientras el tráfico de la aplicación está en tránsito, lo que provoca problemas de accesibilidad para las instancias.



Imágenes de máquina de Amazon (AMI)

Una imagen de máquina de Amazon (AMI) es una imagen compatible y mantenida que ofrece AWS y que brinda la información necesaria para lanzar una instancia. Debe especificar una AMI al lanzar una instancia. Cuando necesite varias instancias con la misma configuración, puede lanzarlas desde una misma AMI. Cuando necesite instancias con distintas configuraciones, puede utilizar distintas AMI para lanzarlas.

Una AMI incluye lo siguiente:

* **Una o más instantáneas de Amazon Elastic Block Store (Amazon EBS)** o, para las AMI con respaldo en el almacenamiento de la instancia.
* **Una plantilla para el volumen raíz de la instancia** (por ejemplo, un sistema operativo, un servidor de aplicaciones y aplicaciones).
* **Permisos de lanzamiento** que controlan qué cuentas de AWS pueden utilizar la AMI para lanzar instancias.
  + - **Público:** El propietario concede permiso de lanzamiento a todas las cuentas de AWS
    - **Explícito:** Permisos de lanzamiento solo a cuentas específicas de AWS
    - **Implícito:** El propietario posee permisos de lanzamiento implícitos de una AMI
* **Un mapeo de dispositivos de bloques** que especifica los volúmenes que se van a adjuntar a la instancia cuando se lance

## 

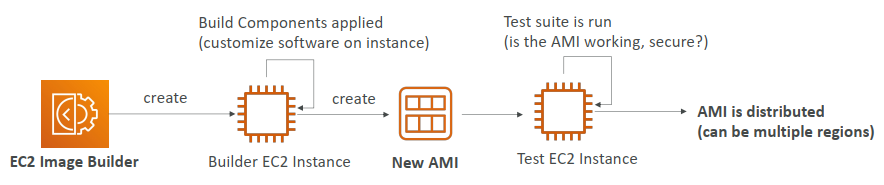
## AMIs personalizadas

* + - Cuando se necesitan paquetes preinstalados
    - Tiempo de arranque rápido (evitar el uso de scripts de datos de usuario muy largos)
    - Maquinas ya configuradas con software empresarial o de monitoreo
    - Control de la máquina en la red desde el momento de arranque
    - Control del mantenimiento y actualización de la AMI a través del tiempo
    - Active Directory Integration
    - Instalar la aplicación propia de antemano (para tiempos de despliegue rápidos cuando se está auto escalando)
    - Utilizar la AMI personalizada de alguien mas que la haya optimizado para algo específico

*Cuando se construye una AMI personalizada, se hace para una región específica*

*(no disponible globalmente)*

## Ec2 Image Builder

* Utilizado para automatizar la creación de Máquinas Virtuales o contenedores de imágenes
* Se puede automatizar la creación, mantenimiento, validación y prueba de AMIs para instancias EC2
* Se puede crear un cronograma para ejecutar este servicio, por ejemplo, para actualizar los paquetes, etc.

¿Cómo se cobra?

Para las AMI con respaldo en el almacén de instancias (Instance store), se cobra por el almacenamiento de las AMI en Amazon S3 y el uso de instancias. Para las AMI con respaldo en Amazon EBS, se cobra por el uso de instancias, por el almacenamiento y el uso del volumen de EBS y por almacenar la AMI como una instantánea de EBS.

Para las AMI con respaldo en el almacén de instancias Amazon EC2, cada vez que se personaliza una AMI y se crea una nueva, todas las partes se almacenan en Amazon S3 por cada AMI. Por lo tanto, el espacio de almacenamiento de cada AMI personalizada es el tamaño completo de la AMI. Para AMI con respaldo en Amazon EBS, cada vez que se personaliza una AMI y se crea una nueva, solo se almacenan los cambios. Por tanto, la huella de almacenamiento de las AMI posteriores que personalice después de la primera es mucho menor, por lo que los cargos de almacenamiento de AMI son menores.

Cuando se detiene una instancia con un volumen de EBS como dispositivo de raíz, no se cobra por el uso de la instancia, sino que se cobra por el almacenamiento del volumen. En cuanto inicie la instancia, cobramos un cargo mínimo de un minuto por uso. Después del primer minuto, solo cobramos por los segundos utilizados. Por ejemplo, si ejecuta una instancia durante 20 segundos y luego la detiene, pasaremos un cargo por un minuto completo. Si ejecuta una instancia durante 3 minutos y 40 segundos, libramos un cargo equivalente exactamente a 3 minutos y 40 segundos de uso. Cobramos por cada segundo, con un mínimo de un minuto, que mantenga la instancia en ejecución, aun cuando permanezca inactiva y no se conecte a ella.

## Redes en Amazon EC2

Amazon VPC permite lanzar recursos de AWS, como las instancias de EC2, en una red virtual dedicada a la cuenta de AWS, conocida como Virtual Private Cloud. Al iniciar una instancia, se puede seleccionar una subred de la VPC. La instancia está configurada con una interfaz de red principal, que es una tarjeta de red virtual lógica. La instancia recibe una IP privada principal de la dirección IPv4 de la subred y se asigna a la interfaz de red principal.

Se puede asociar una IP pública del grupo de direcciones IP públicas de Amazon. Esta dirección se encuentra asociada a la instancia mientras no sea finalizada o detenida. En el caso de necesitar una IP pública persistente, se puede asociar una Elastic IP a la cuenta de AWS y luego asociarla a una instancia o a una interfaz de red. Las Elastic IP están asociadas a la cuenta de AWS hasta que sean liberadas y mientras se puede mover una Elastic IP entre instancias según sea necesario. Se puede reservar un intervalo de direcciones IP.

### Regiones

Cada región de Amazon EC2 se ha diseñado para que esté totalmente aislada de las demás regiones de EC2. Con ello se consigue la mejor tolerancia a errores y estabilidad posibles.

Amazon no replica la información de una región a las demás automáticamente, por lo tanto, cuando se consulta sobre algún recurso solo se puede hacer para los recursos vinculados a una región especifica

Cuando se lanza una instancia, se debe seleccionar una AMI que esté disponible en la misma región. Se puede copiar una AMI de una región a otra, pero se debe tener en cuenta que existen cargos por transferencia de información entre regiones.

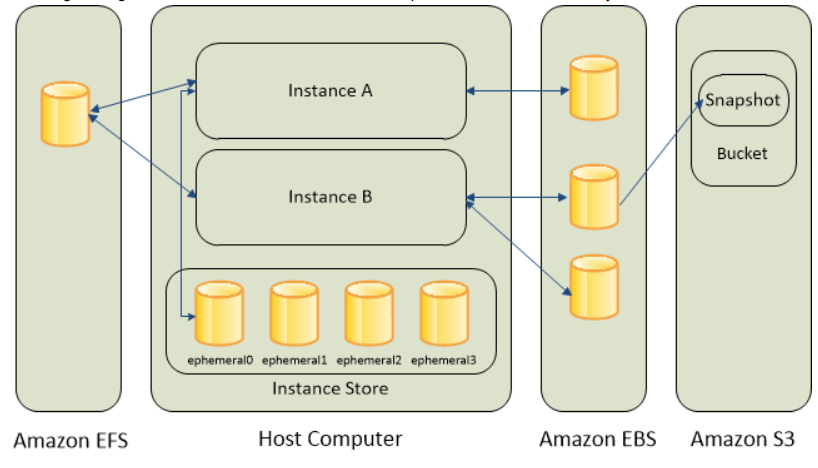
### Zonas de Disponibilidad

Cada región tiene ubicaciones aisladas conocidas como zonas de disponibilidad. Cuando se lanza una instancia se puede elegir en que zona o dejar que lo decida AWS. Si se distribuyen los lanzamientos de instancias en diferentes AZ y una de las instancias genera un error, se puede diseñar la aplicación de forma que otra instancia en otra AZ procese las solicitudes.

Se pueden usar direcciones IP elásticas para enmascarar los errores de una instancia en una AZ volviendo a mapear rápidamente la dirección hacia una instancia en otra AZ

ARN de EC2

# Storage (EBS)

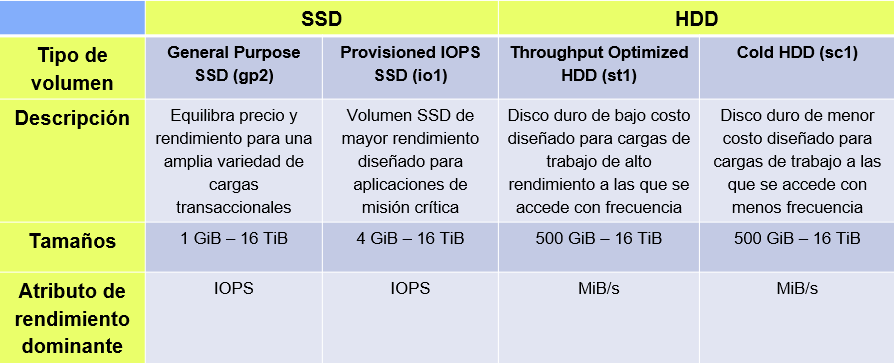


## Amazon EBS Overview

1. EBS ofrece volúmenes de almacenamiento duradero de nivel de bloque que se pueden adjuntar a una instancia en ejecución. EBS es la opción recomendada para almacenar una base de datos en una instancia.
2. Un volumen de EBS se comporta como un dispositivo de bloque externo, sin formatear y sin procesar que se puede adjuntar a una única instancia (en el nivel CCP – Certified Cloud Practitioner). El volumen persiste independientemente de la vida de ejecución de la instancia asociada.
3. Están limitados a una zona de disponibilidad específica
   * Por lo tanto, si se crea un volumen en us-east-1a, no puede ser asociado a una instancia en us-east-1b.
   * Para mover un volumen a otra zona de disponibilidad primero se debe realizar un snapshot del mismo
4. Para realizar un backup o snapshot no es necesario desconectar el volumen, pero es recomendable.
   * Estos snapshots se pueden replicar a través de otras AZ o regiones para ser utilizados en ellas.
5. Es un disco de red (network drive), utiliza la red para comunicarse con la instancia por lo que puede existir algo de latencia.
6. Para crear un volumen se debe especificar la capacidad en GBs y IOPS
7. Un volumen EBS solo se puede incrementar en tamaño en GBs (cualquier tipo de EBS) y en IOPS (solo para IO1). Luego de cambiar el tamaño se debe reparticionar el disco en la instancia, para aprovechar la capacidad agregada.
8. Por defecto, los volúmenes root EBS se eliminan cuando se termina una instancia, pero cualquier otro volumen EBS asociado no es eliminado. Este comportamiento puede ser cambiado por el atributo “Delete on Termination”
9. AWS nos cobra por la capacidad completa provisionada

### Ejemplo de asociaciones:

## Tipos de Volúmenes



GP2

* Recomendado para la mayoría de las cargas de trabajo
* Es el menos costoso para un SSD
* Se puede usar como volumen de arranque del sistema
* Se puede usar para escritorios virtuales
* Aplicaciones interactivas de baja latencia
* Entornos de desarrollo y prueba
* Entre 1GiB y 16TiB
* Volúmenes GP2 pequeños, pueden incrementar las IOPS hasta las 3000
* Las máximas IOPS son 16.000
* Se tienen 3IOPS por c/GB por lo que a los 5.334GB estamos en el máximo de los IOPS. Cualquier tamaño mayor a esté, no incrementará los IOPS

IO1

* Recomendable para aplicaciones comerciales críticas que requieren un rendimiento sostenido de IOPS o mas de 16.000 IOPS por volumen
* Grandes cargas de trabajo de bases de datos
* Tamaño de 4GiB a 16TiB
* IOPS son provisionadas (PIOPS) MIN 100 MAX 64.000
* Por cada GB se pueden provisionar 50IOPS

ST1

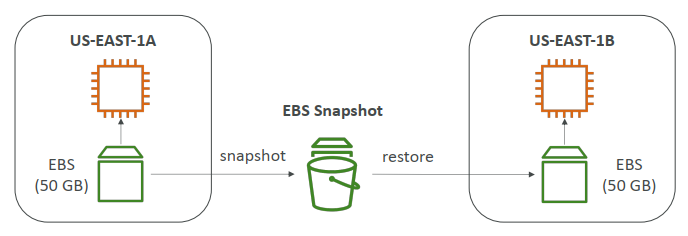
* Para cargas de trabajo de transmisión que necesitan un rendimiento rápido a un precio bajo
* Big Data
* Data Warehouses
* Log processing
* Apache Kafka
* No puede ser un volumen de booteo
* Tamaño desde 500GiB hasta 16TiB
* Max IOPS es 500
* Máximo rendimiento de 500MiB/s - pueden “explotar” (burst)

SC1

* Para grandes volúmenes de datos poco accedidos pero que requieren gran rendimiento
* Costo de almacenamiento bajo
* Tamaño 500GiB – 16TiB
* Max IOPS es 250
* Máximo rendimiento de 250MiB/s – pueden “explotar” (burst)

## EBS Snapshot

* Es una backup del volumen
* El tamaño o peso en GB de estas snapshots solo será de aquel ocupado en el EBS (no se tiene en cuenta el espacio libre)
* Se puede programar la creación de snapshots
* Generalmente se utilizan como backup cuando se quieren realizar tareas como: cambiar el tamaño de volumen EBS hacia abajo, cambiar el tipo de volumen EBS o encriptar un volumen.
* No es necesario desacoplar los discos de las instancias para realizar un snapshot pero si es recomendable.
* Se pueden copiar snapshots a través de Regiones o AZ



## Encriptación de EBS

* Cuando se crea un volumen encriptado:
  + - Se encriptan los datos en el volumen
    - Se encriptan los datos en vuelo, es decir, cuando están en camino desde el EBS o hacia el EBS (SSL encryption)
    - Los snapshots también están encriptados
    - Todos los volúmenes creados a partir de snapshots encriptados, se encriptan
* Todo el proceso de encriptación es transparente para el usuario de AWS
* Impacto mínimo en la latencia
* Se utilizan llaves KMS (AES-256)

EC2 Instance Store

El almacén de instancias ofrece un almacenamiento de nivel de bloques temporal para la instancia. **Este almacenamiento se encuentra en discos que están conectados físicamente al equipo host**. El almacén de instancias es perfecto para el almacenamiento temporal de información que cambia con frecuencia, como los búferes, las cachés, los datos de pruebas y otro contenido temporal, o de datos que se replican en una flota de instancias, como un grupo de servidores web con balance de carga.

Un almacén de instancias consta de uno o varios volúmenes de almacenes de instancias que se exponen como dispositivos de bloques. El tamaño de un almacén de instancias, al igual que el número de dispositivos disponibles, varía por tipo de instancia.

* Son discos de Alto rendimiento, pero limitadas para Inputs/Outputs
* Pierden su almacenamiento si ocurre algún fallo relacionado a la instancia o si es terminada
* Buenos para buffer/cache/contenido-temporario/scratch-data
* Backups y replicación son responsabilidad del usuario

Duración del almacén de instancias

Solo puede especificar volúmenes de almacenes de instancias al lanzar la instancia. No puede separar un volumen de almacén de instancias de una instancia y adjuntarlo a otra instancia.

Los datos en cualquier almacén de instancias se conservan solo durante el ciclo de vida de la instancia asociada. Si una instancia se reinicia (de manera intencionada o no intencionada), los datos del almacén de instancias se conservan. Sin embargo, los datos del almacén de instancias se pierden en cualquiera de las siguientes circunstancias:

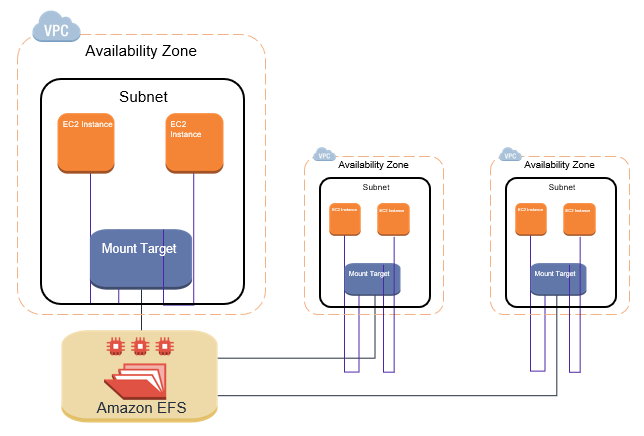
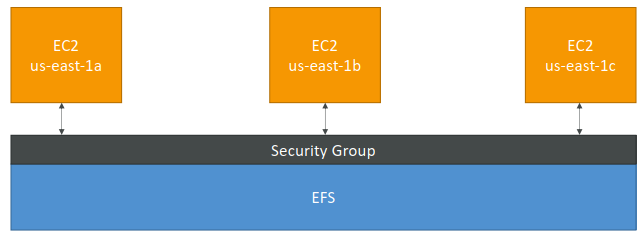
* Falla la unidad de disco subyacente.
* La instancia se para.
* La instancia está hibernando.
* La instancia termina.

Por lo tanto, no confíe en el almacén de instancias para almacenar datos valiosos a largo plazo. En su lugar, utilice un almacenamiento de datos más duradero, como Amazon S3, Amazon EBS o Amazon EFS.

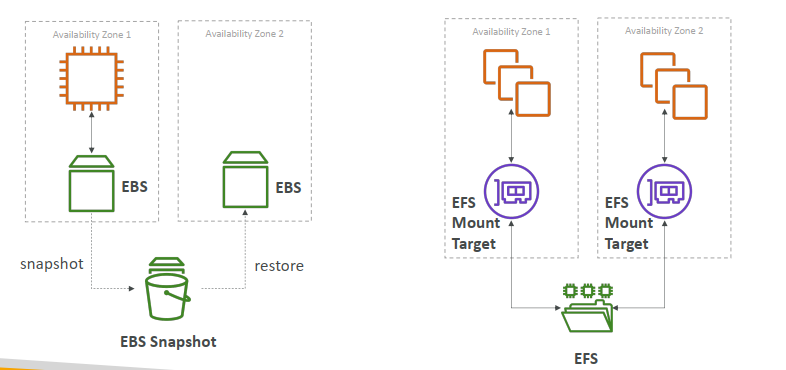
EFS – Elastic File System

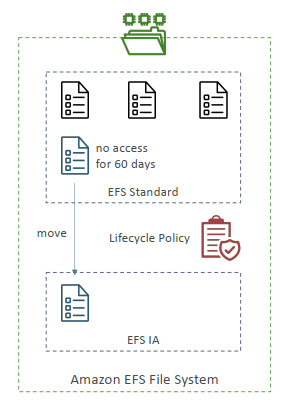
Amazon EFS proporciona almacenamiento de archivos escalable para su uso con Amazon EC2. **Se puede usar un sistema de archivos de EFS como origen de datos común para aplicaciones y cargas de trabajo que se ejecutan en varias instancias (no se admiten instancias de Windows)**

* Sistema de Archivo de Red administrado (Network File System) que **puede ser montado en cientos de instancias EC2**
* EFS funciona con instancias Linux de EC2 **en Múltiples Zonas de Disponibilidad**
* **Altamente disponible, escalable y costoso** (3x gp2), se paga solo por el uso sin planificación de capacidad
* EFS es un servicio que se puede compartir entre Zonas de disponibilidad, Regiones, VPC´s y cuentas

****

## EBS vs EFS





## EFS Acceso Infrecuente (EFS-IA)

* Clase de almacenamiento optimizado en costo para archivos que no son accedidos todos los días.
* Hasta 92% de ahorro comparado con EFS standard
* EFS moverá automáticamente los archivos a EFS-IA en base a la fecha de último acceso
* Se debe habilitar EFS-IA con una Lifecycle-policy

Por ejemplo: “Mover archivos a EFS-IA que no fueron accedidos hace 60 días”

* Es un proceso transparente para las aplicaciones accediendo a EFS

## Modelo de responsabilidad compartida para EC2 Storage

