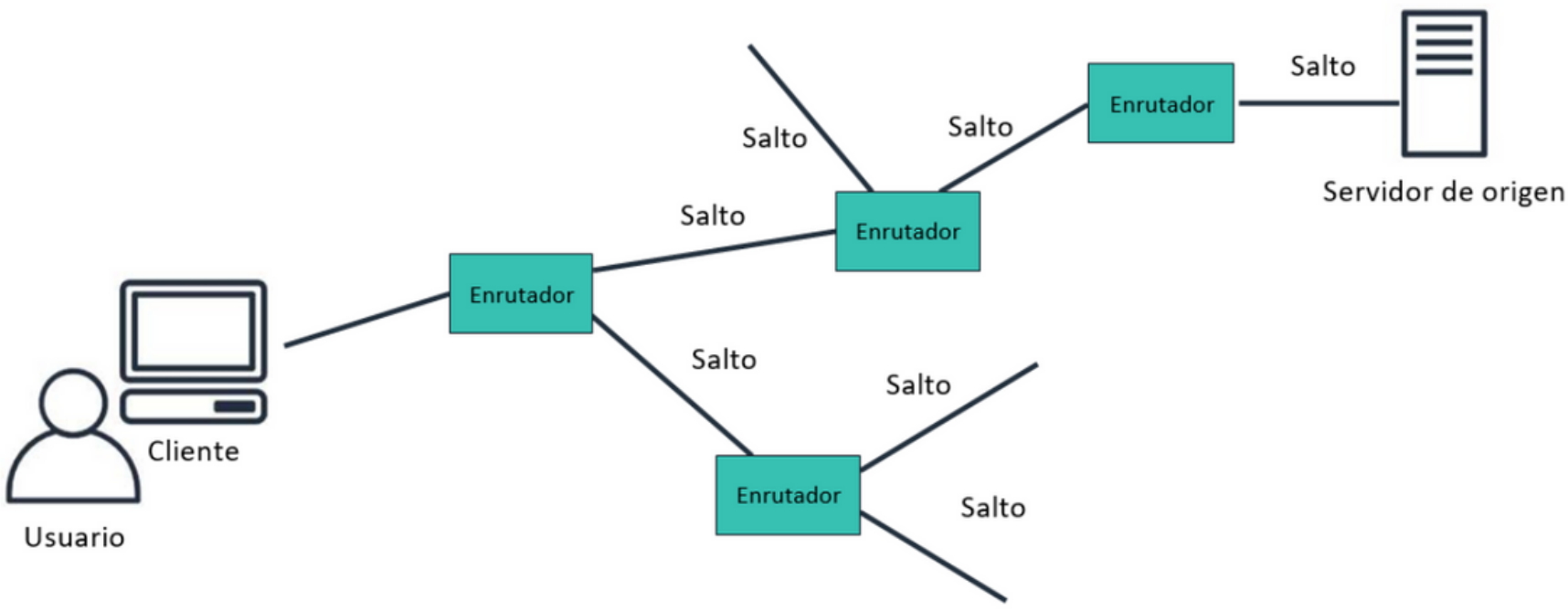
Lección 3
Almacenamiento en caché perimetral

Latencia de red



Cuando alguien busca su sitio web o utiliza su aplicación, su solicitud viaja a través de muchas redes diferentes para llegar a su servidor de origen. El servidor de origen, que también se conoce como origen, almacena las versiones originales y definitivas de sus objetos (por ejemplo, objetos web, imágenes y archivos multimedia). El número de saltos de red y la distancia que debe recorrer la solicitud afectan significativamente el rendimiento y la capacidad de respuesta del sitio web.

Además, la latencia de red puede depender de la ubicación geográfica del servidor de origen. Cuando el tráfico web está geográficamente disperso, no siempre es viable (ni rentable) replicar la infraestructura completa en todo el mundo. En este caso, una red de entrega de contenido (CDN) puede ser útil.

Red de entrega de contenido (CDN)

- Es un sistema distribuido a nivel mundial de servidores de almacenamiento en caché
- Almacena en caché copias de archivos solicitados habitualmente (contenido estático)
- Entrega una copia local del contenido solicitado desde un borde de caché cercano o punto de presencia
- Mejora el rendimiento y el escalado de las aplicaciones

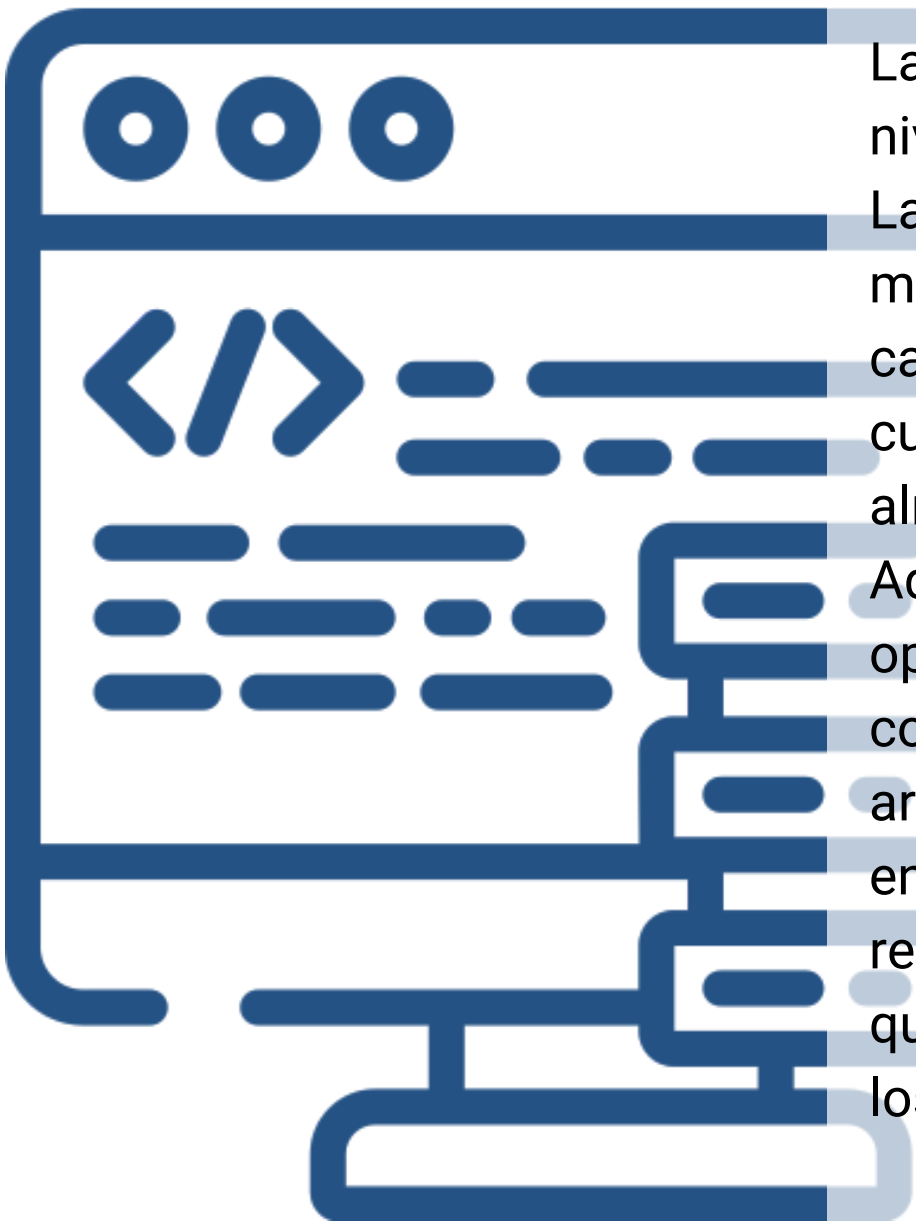
Una red de entrega de contenido (CDN) es un sistema de servidores de almacenamiento en caché distribuido a nivel mundial. Una CDN almacena en caché copias de archivos solicitados comúnmente que están alojados en el servidor de origen de la aplicación. Estos archivos pueden incluir contenido estático, como archivos HTML, CSS, JavaScript, de imagen y de video. La CDN entrega una copia local del contenido solicitado desde un borde de caché o un punto de presencia que ofrece la entrega más rápida al solicitante.

Para obtener más información acerca del almacenamiento en caché con CDN, consulte Content Delivery Network (CDN) Caching.

Amazon CloudFront

- Es la CDN global de Amazon.
- Está optimizado para todos los casos de uso de entrega, con una caché de varios niveles de forma predeterminada y amplia flexibilidad.
- Le brinda una capa adicional de seguridad para sus arquitecturas.
- Admite métodos WebSockets y HTTP o HTTPS.

Amazon CloudFront es un servicio de CDN global que acelera la entrega de contenido a los usuarios. Dicho contenido puede ser contenido estático y dinámico, archivos multimedia que utilizan HTTP o HTTPS, y video en streaming (tanto video bajo demanda como streaming en directo). Al igual que otros servicios de AWS, CloudFront es un producto de autoservicio y de pago por uso que no requiere compromisos a largo plazo ni tarifas mínimas.

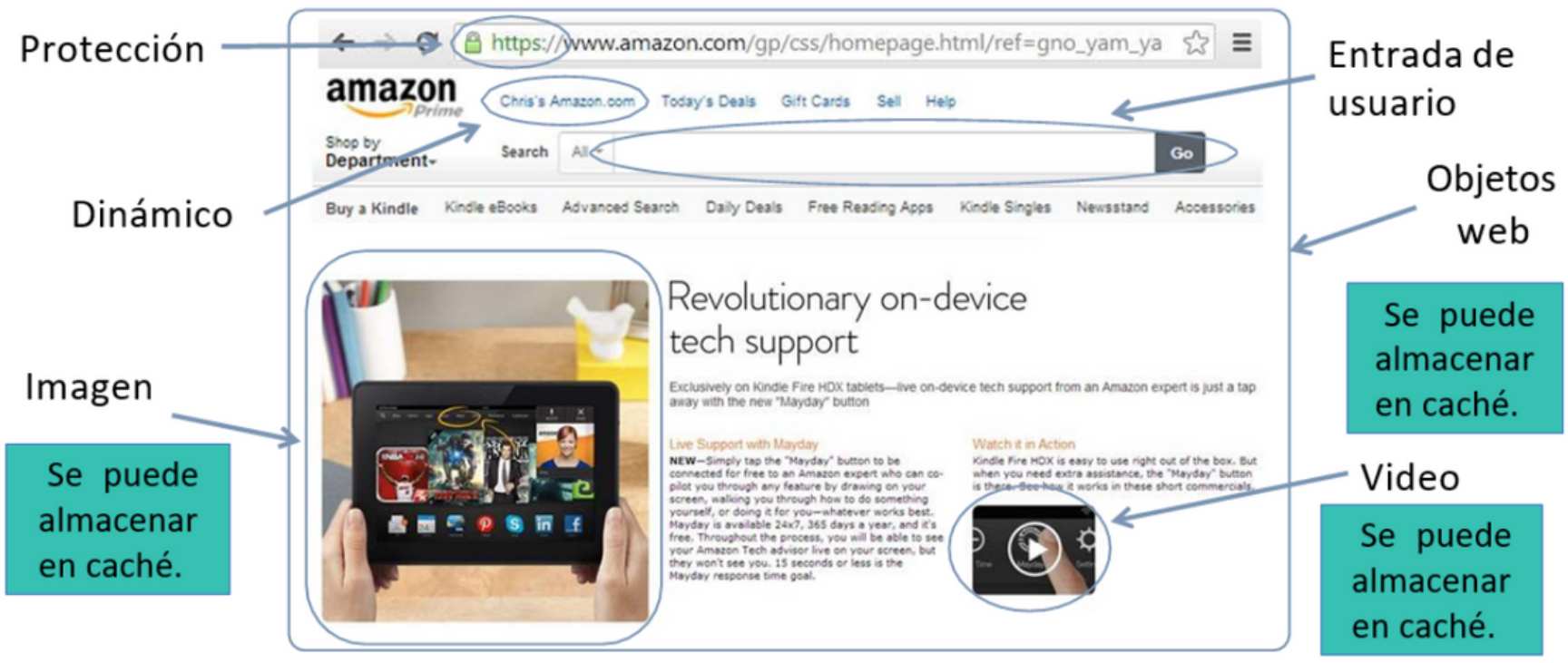


La CDN ofrece una caché de varios niveles de forma predeterminada. Las cachés perimetrales regionales mejoran la latencia y reducen la carga en los servidores de origen cuando el objeto aún no está almacenado en caché en el borde. Además, la CDN ofrece múltiples opciones para transmitir su contenido multimedia, tanto archivos pregrabados como eventos en directo. Los ofrece con el rendimiento elevado y sostenido que se requiere para la entrega 4K a los espectadores de todo el mundo.

CloudFront brinda protección tanto a nivel de red como a nivel de aplicación. Las aplicaciones y el tráfico se ven beneficiados por distintas protecciones integradas, como AWS Shield Standard, sin costo adicional. También puede utilizar características configurables, como AWS Certificate Manager (ACM), para crear y administrar certificados SSL personalizados sin costo adicional. CloudFront admite los protocolos capa de conexión segura y Transport Layer Security (SSL/TLS).

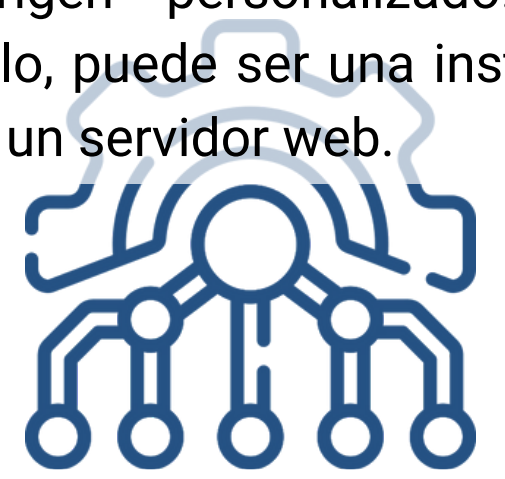
CloudFront admite la comunicación bidireccional en tiempo real a través del protocolo WebSocket. Esta conexión persistente permite a los clientes y los servidores enviar datos en tiempo real entre sí sin incurrir en el costo de establecer conexiones repetidas veces. Esto es especialmente útil para aplicaciones de comunicación, como chat, colaboración, juegos y negociaciones financieras. CloudFront también admite métodos HTTP (DELETE, GET, HEAD, OPTIONS, PATCH POST, PUT), que mejoran el rendimiento de los sitios web dinámicos. Estos sitios web tienen formularios web, casillas de comentarios e inicio de sesión, botones de agregar al carro y otras características que cargan datos de los usuarios. Por lo tanto, puede utilizar un único nombre de dominio para entregar todo el sitio web a través de CloudFront, lo que acelera tanto la descarga como la carga de las partes de su sitio web.

¿Qué tipo de contenido se puede almacenar en caché en una caché perimetral?



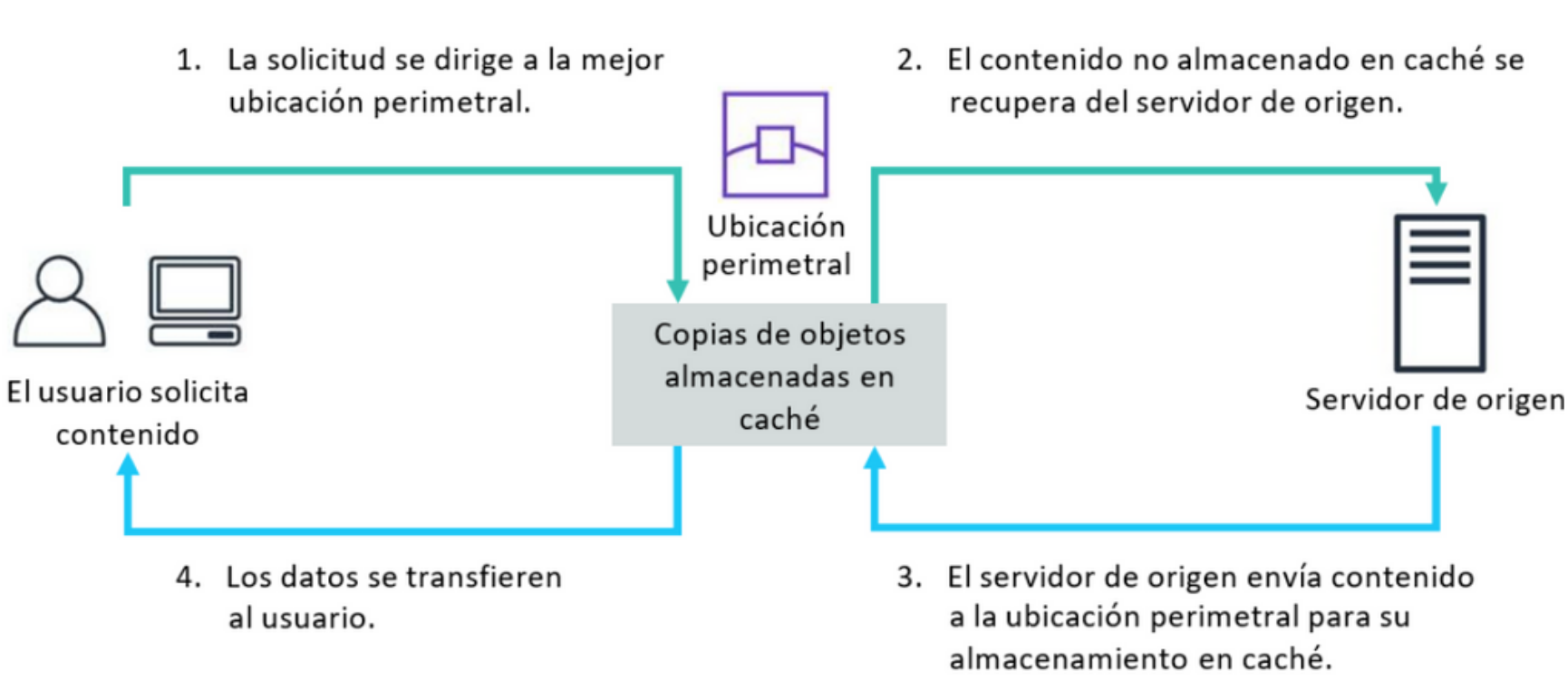
Este ejemplo de una página web Amazon.com muestra cómo el contenido estático y dinámico puede componer una aplicación web dinámica. Esta aplicación web se entrega con el protocolo HTTPS para el cifrado de las solicitudes de página de usuario y las páginas que se devuelven desde un servidor web. Puede usar una CDN o una caché perimetral para almacenar el contenido estático. Dicho contenido puede incluir objetos web (por ejemplo, documentos HTML, hojas de estilo CSS o archivos JavaScript), archivos de imagen y archivos de video.

No se puede almacenar en caché contenido generado de manera dinámica ni datos generados por el usuario. Sin embargo, puede configurar CloudFront para que entregue esta información desde una aplicación que se ejecuta en un origen personalizado. Por ejemplo, puede ser una instancia EC2 o un servidor web.



Además, puede configurar CloudFront para exigir a los espectadores que utilicen HTTPS al solicitar sus objetos, de modo que las conexiones se cifren cuando CloudFront se comunique con los espectadores. También puede configurar CloudFront para usar HTTPS a la hora de obtener objetos del origen, de modo que las conexiones se cifren cuando CloudFront se comunique con su origen.

Funcionamiento del almacenamiento en caché en Amazon CloudFront

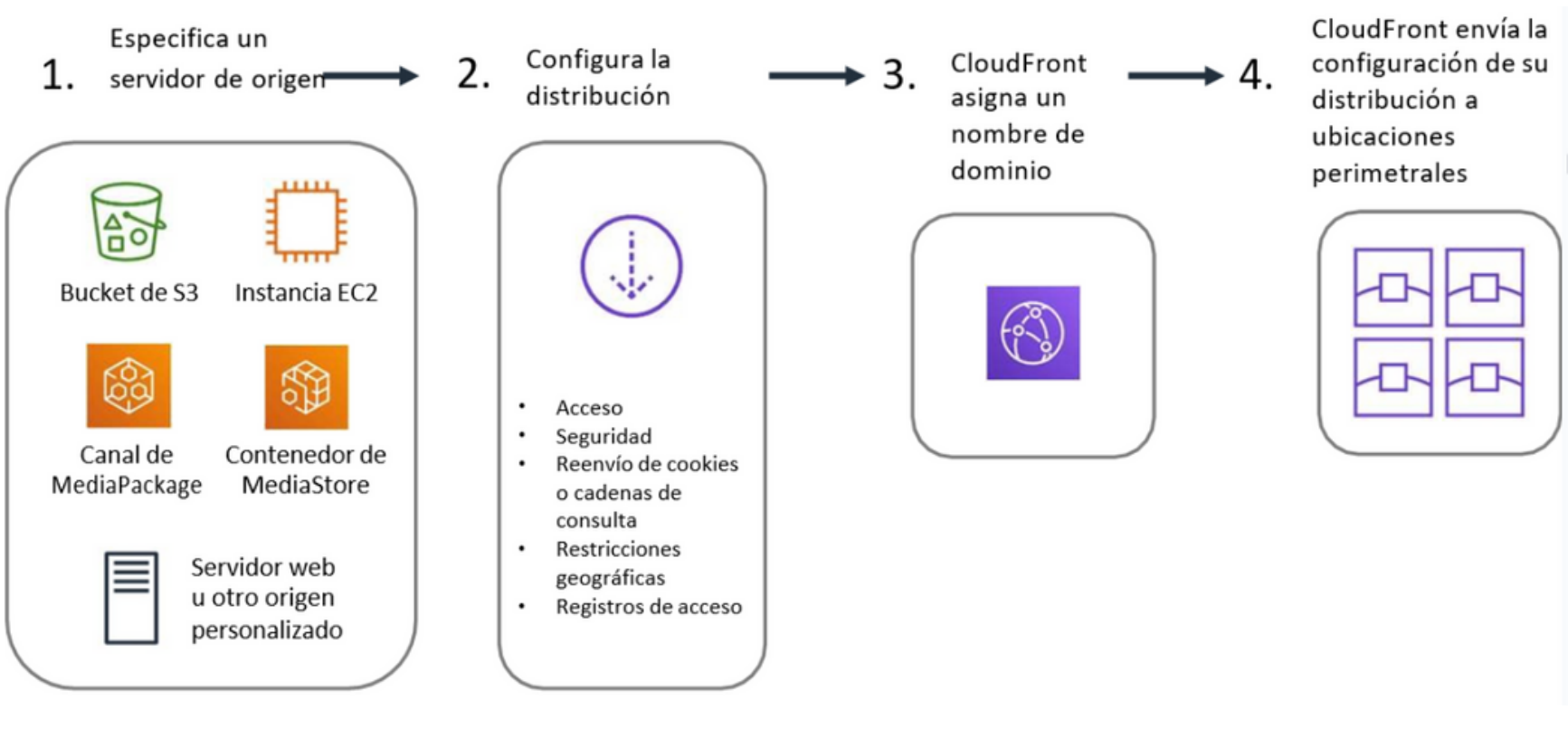


Amazon CloudFront entrega su contenido a usuarios a través de una red mundial de centros de datos que reciben el nombre de ubicaciones perimetrales.

Cuando un usuario solicita contenido que usted atiende con CloudFront, el DNS dirige la solicitud a la ubicación perimetral que mejor atienda la solicitud. Normalmente, es la ubicación perimetral más cercana la que tiene la latencia más reducida. CloudFront revisa la caché para comprobar si dispone del contenido solicitado (paso 1). Si el contenido está en la caché, CloudFront lo entrega inmediatamente al usuario (paso 4). Es posible que el contenido no esté actualmente en la caché. Si no lo está, CloudFront reenvía la solicitud al servidor de origen que usted identificó como origen de la versión definitiva de su contenido (paso 2). El servidor de origen envía el contenido de vuelta a la ubicación perimetral (paso 3) y CloudFront luego reenvía el contenido al usuario (paso 4). También agrega el contenido a la caché en la ubicación perimetral para la próxima vez que alguien lo solicite.

A medida que los objetos pierden popularidad, las ubicaciones perimetrales individuales podrán eliminarlos a fin de liberar espacio para el contenido más solicitado. Para el contenido menos popular, CloudFront dispone de cachés perimetrales regionales. Las cachés perimetrales regionales son ubicaciones de CloudFront que se implementan a nivel mundial y están cerca de sus espectadores. Están ubicadas entre el servidor de origen y las ubicaciones perimetrales globales que entregan contenido directamente a los espectadores. Una caché perimetral regional tiene una memoria caché más grande que una ubicación perimetral individual, por lo que los objetos permanecen en la caché perimetral regional durante más tiempo. Esta disposición ayuda a mantener más contenido cerca de los espectadores. Reduce la necesidad de que CloudFront vuelva a su servidor de origen y mejora el rendimiento general para los espectadores.

Configuración de una distribución de CloudFront



Si desea utilizar CloudFront para distribuir su contenido, debe crear una distribución.

1 Especifica el servidor de origen que aloja los archivos. El servidor de origen puede ser un bucket de S3, un canal de AWS Elemental MediaPackage, un contenedor de AWS Elemental MediaStore o un origen personalizado. Por ejemplo, un origen personalizado puede ser una instancia EC2 o su propio servidor web.

2 A continuación, especifica detalles sobre cómo realizar un seguimiento de la entrega de contenido y cómo administrarla. Por ejemplo, puede especificar si desea que los archivos estén disponibles para todos o solo para determinados usuarios. También puede especificar si desea que CloudFront cumpla las siguientes funciones: crear registros de acceso que muestren la actividad del usuario, reenviar cookies o cadenas de consulta a su origen, o exigir a los usuarios que utilicen HTTPS para acceder a su contenido.

3 CloudFront asigna un nombre de dominio a la nueva distribución.

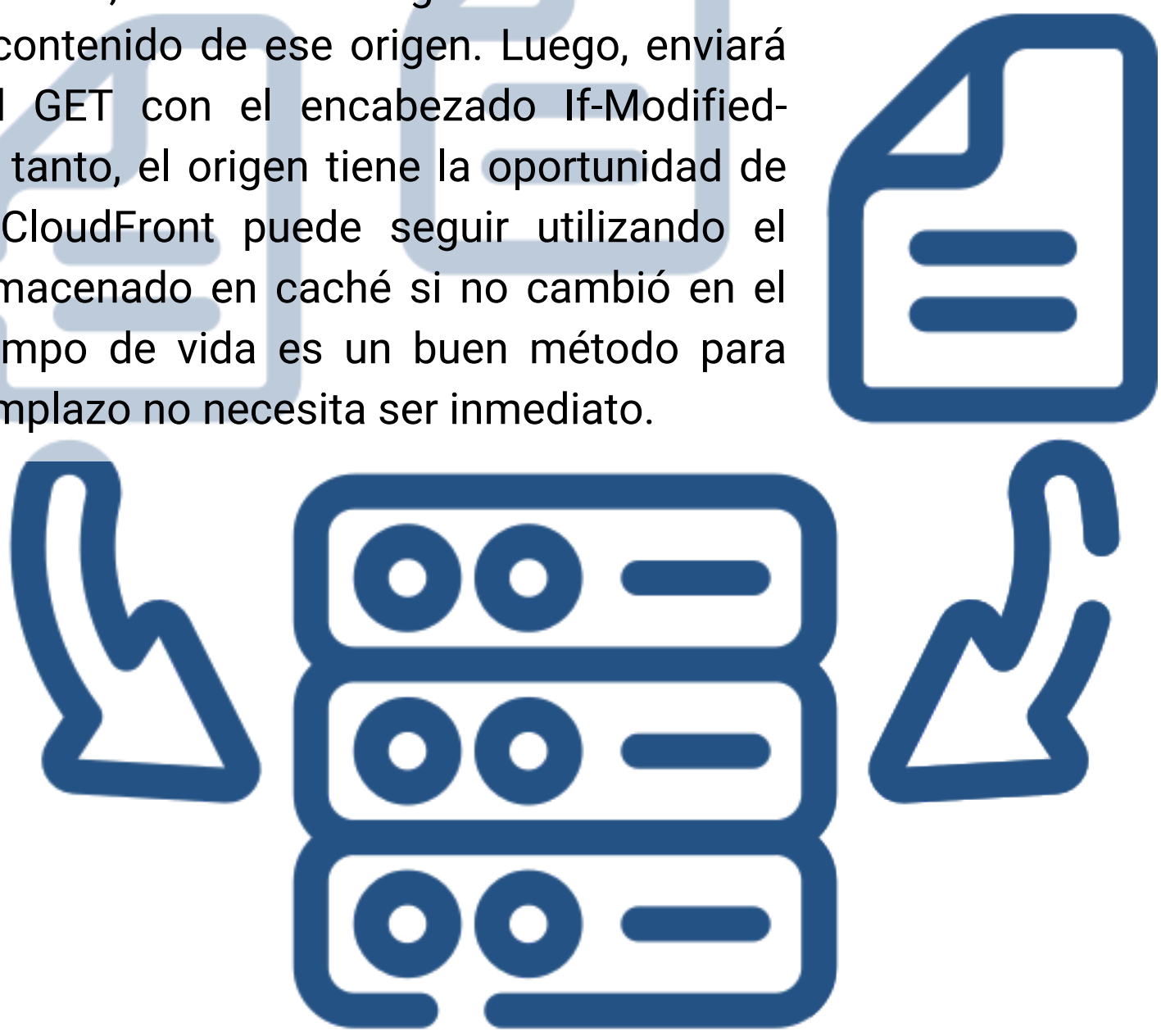
4 CloudFront envía la configuración de su distribución, pero no el contenido, a todas las ubicaciones perimetrales.

Vencimiento del contenido

- Tiempo de vida (TTL)
  - Se trata de un periodo fijo (periodo de vencimiento)
  - Usted lo define
  - La solicitud GET al origen de CloudFront utiliza el encabezado If-Modified-Since
- Cambio de nombre del objeto
  - Header-v1.jpg pasa a llamarse Header-v2.jpg
  - El nuevo nombre fuerza la actualización inmediata
- Invalidación del objeto
  - Último recurso: ineficiente y costoso

Puede hacer que el contenido almacenado en caché caduque de tres maneras:

Tiempo de vida (TTL): con este método, puede controlar durante cuánto tiempo los archivos permanecen en una caché de CloudFront antes de que dicho servicio reenvíe otra solicitud al origen. Acortar la duración le permite ofrecer contenido dinámico. Extender la duración implica que sus usuarios podrán disfrutar de un mejor rendimiento, ya que es más probable que sus archivos se ofrezcan directamente desde la caché perimetral. Una duración extendida también reduce la carga en el origen. Si define el tiempo de vida de un origen específico como 0, CloudFront seguirá almacenando en caché el contenido de ese origen. Luego, enviará una solicitud GET con el encabezado If-Modified-Since. Por lo tanto, el origen tiene la oportunidad de indicar que CloudFront puede seguir utilizando el contenido almacenado en caché si no cambió en el origen. El tiempo de vida es un buen método para usar si el reemplazo no necesita ser inmediato.

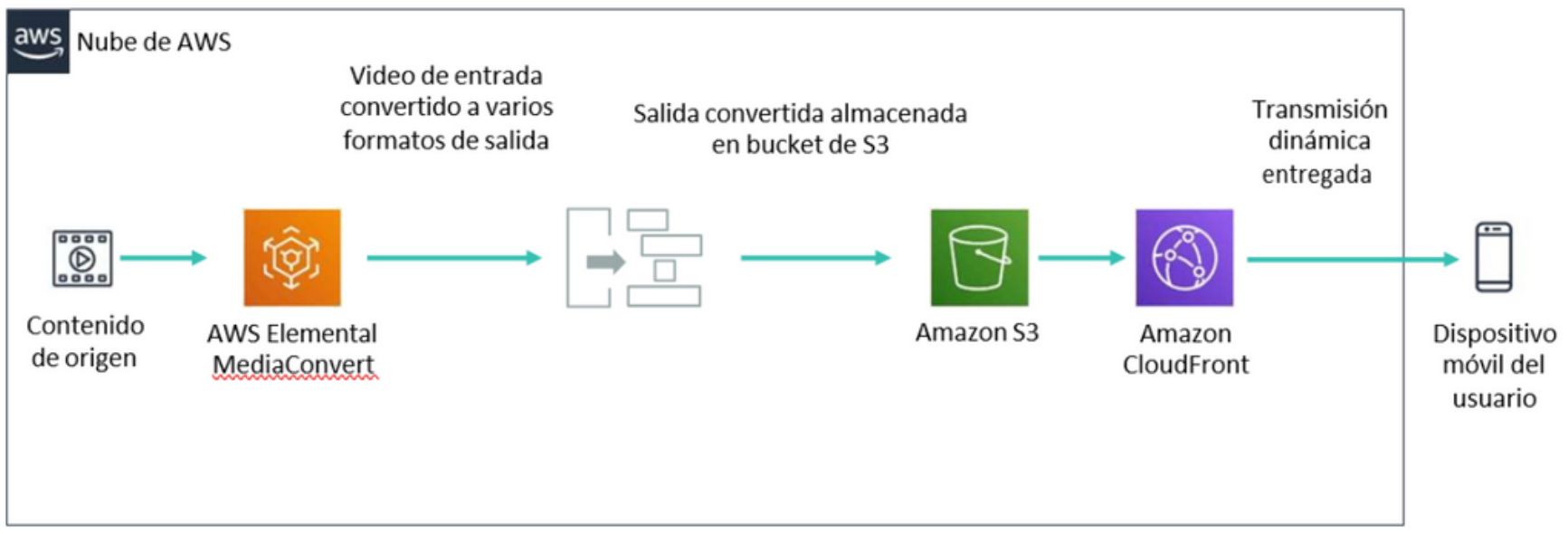


Cambio de nombre del objeto: este método requiere más esfuerzo, pero el reemplazo es inmediato. Aunque puede actualizar los objetos existentes en una distribución de CloudFront y utilizar los mismos nombres de objeto, no se recomienda hacerlo. CloudFront distribuye los objetos a ubicaciones perimetrales solo cuando estos se solicitan y no cuando se agregan objetos nuevos o actualizados al origen. Por ejemplo, podría actualizar un objeto existente en su origen con una versión más reciente que tenga el mismo nombre. En ese caso, una ubicación perimetral no obtendrá esa nueva versión de su origen hasta que se produzcan los dos eventos enumerados.

Invalidación del objeto: este método es una mala solución porque el sistema debe interactuar por la fuerza con todas las ubicaciones perimetrales. Debe usar este método con moderación y sólo para objetos individuales.

Para obtener más información acerca de cómo se vence el contenido de la caché, consulte Administración de cuánto tiempo se mantiene el contenido en una caché perimetral (vencimiento).

Ejemplo: streaming de video bajo demanda



Como aprendió, puede utilizar CloudFront para ofrecer video en streaming, tanto video bajo demanda como streaming en directo.

En el caso del streaming de video bajo demanda, debe utilizar un codificador para formatear y empaquetar el contenido del video antes de que CloudFront pueda distribuirlo. Algunos ejemplos de codificadores son AWS Elemental MediaConvert y Amazon Elastic Transcoder. El proceso de empaquetado crea segmentos, que son archivos estáticos, los cuales incluyen el contenido de audio, video y subtítulos. También genera archivos de manifiesto, los cuales describen qué segmentos reproducir y el orden específico para reproducirlos. Los formatos de paquete incluyen Dynamic Adaptive Streaming over HTTP (DASH o MPEG-DASH), HTTP Live Streaming (HLS) de Apple, Microsoft Smooth Streaming y Common Media Application Format (CMAF).

Después de convertir el video a los formatos de salida, aloja el contenido convertido en un bucket de S3, que es su servidor de origen. A continuación, utiliza CloudFront para entregar los archivos de segmento a usuarios de todo el mundo.

Para obtener más información acerca del streaming de video con CloudFront, consulte Video bajo demanda y streaming de video en directo con CloudFront.

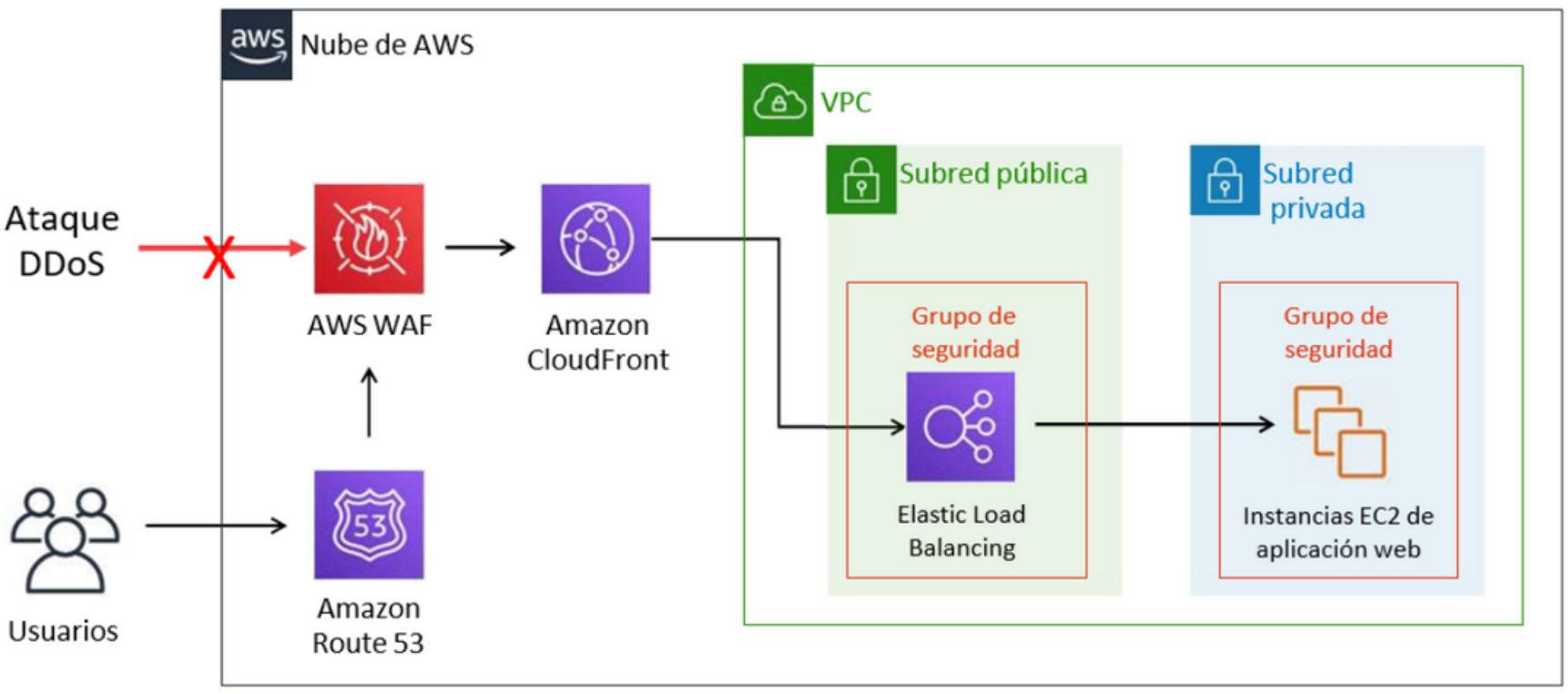
Ejemplo: contenido generado de manera dinámica

Caso de uso: mosaicos de mapas
Problema: se necesita un tiempo de respuesta más rápido en la base de datos

En general, se almacena en caché solo el contenido estático. Sin embargo, puede tener contenido que parece estático (por su URL), pero se crea dinámicamente la primera vez que se necesita. Esta compilación dinámica puede ser útil cuando el contenido es reutilizable, pero su creación puede ser costosa y puede cambiar con poca frecuencia.

Los mosaicos de mapas son el ejemplo clásico. En este caso, almacena en caché lugares que los usuarios ven con frecuencia (como ciudades principales), pero no lugares como áreas remotas. Generar todas las combinaciones posibles de mosaicos sería demasiado costoso y excesivo porque casi nunca se solicitaría la mayoría de los mosaicos. En su lugar, el componente de ruta de la URL de cada mosaico puede incluir los parámetros necesarios para generar el mosaico. Si el mosaico ya está presente en una ubicación perimetral de CloudFront en particular, entonces se ofrece directamente. De lo contrario, se genera, se devuelve a la ubicación perimetral y, luego, se utiliza para satisfacer futuras solicitudes.

Ejemplo: contenido generado de manera dinámica



Puede utilizar CloudFront para potenciar la resiliencia de las aplicaciones que se ejecutan en AWS ante ataques de denegación de servicio distribuidos (DDoS). Un ataque DDoS es un intento deliberado de hacer que su sitio web o aplicación no estén disponibles para los usuarios, por ejemplo, mediante inundación con tráfico de red. Con este objetivo en mente, los atacantes utilizan varios orígenes para organizar un ataque contra un objetivo. Estos orígenes pueden incluir grupos distribuidos de equipos, enrutadores, dispositivos de Internet de las cosas (IoT) y otros puntos de enlace infectados por malware.

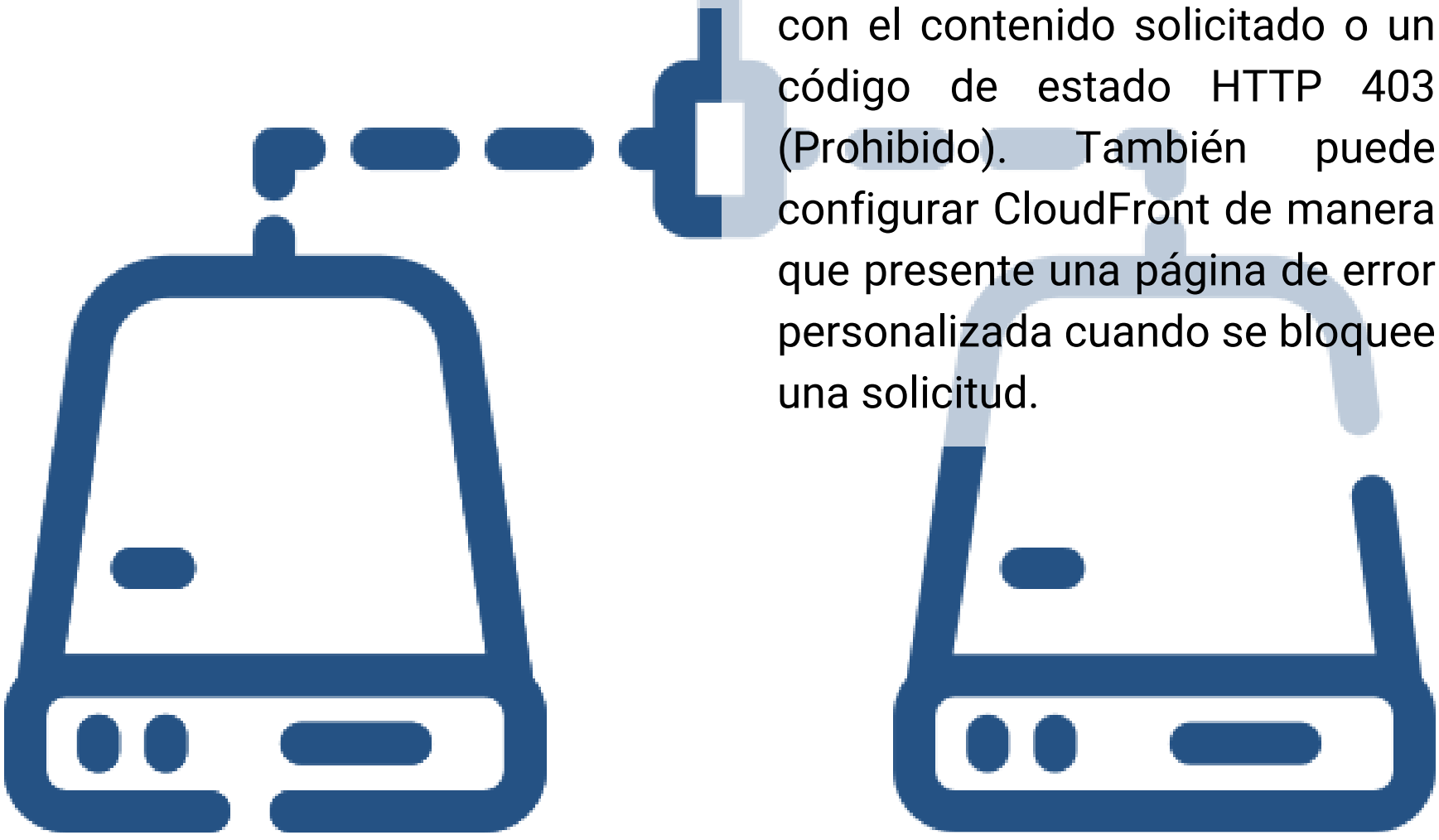
El siguiente ejemplo muestra una arquitectura resistente que puede ayudar a prevenir o mitigar los ataques DDoS.

Un servicio de DNS, como Amazon Route 53, puede conectar de manera eficaz las solicitudes de los usuarios a una distribución de CloudFront. A continuación, la distribución de CloudFront actúa como proxy de las solicitudes de contenido dinámico ante la infraestructura que aloja los puntos de enlace de su aplicación. Tanto las solicitudes de DNS de Route 53 como el tráfico de aplicación subsiguiente que se dirigen a través de CloudFront se inspeccionan allí mismo. El monitoreo siempre activo, la detección de anomalías y la mitigación frente a ataques DDoS de infraestructura comunes están integrados a Route 53 y CloudFront.

Los ataques de infraestructura comunes incluyen inundaciones de sincronización/confirmación (SYN/ACK), inundaciones del protocolo de datagramas de usuario (UDP) y ataques de reflejo. Cuando se supera el límite de ataques de inundación SYN, las cookies de SYN se activan para evitar que se caigan las conexiones con los clientes legítimos. El filtrado de paquetes determinista elimina paquetes TCP con formato incorrecto y solicitudes de DNS no válidas, además de permitir que el tráfico pase únicamente si es válido para el servicio. La detección de anomalías basada en la heurística evalúa atributos, como el tipo, el origen y la composición del tráfico. El tráfico se califica según muchas dimensiones y solo se descarta el más sospechoso.

Este método le permite evitar falsos positivos mientras protege la disponibilidad de las aplicaciones. Route 53 también está diseñado para resistir inundaciones de consultas de DNS. Las inundaciones de consultas de DNS son solicitudes de DNS reales que pueden persistir durante horas e intentar agotar los recursos del servidor de DNS. Route 53 utiliza el particionamiento aleatorio y la fragmentación anycast para distribuir el tráfico DNS entre ubicaciones perimetrales y ayudar a proteger la disponibilidad del servicio.

AWS WAF es un firewall para aplicaciones web. Le permite monitorear las solicitudes HTTP y HTTPS que se reenvían a una API de Amazon API Gateway, CloudFront o un balanceador de carga de aplicaciones. AWS WAF también le permite controlar el acceso a su contenido. Por ejemplo, puede especificar condiciones, como las direcciones IP de las que se originan las solicitudes o los valores de las cadenas de consulta. En función de estas condiciones, API Gateway, CloudFront o un balanceador de carga de aplicaciones responden con el contenido solicitado o un código de estado HTTP 403 (Prohibido). También puede configurar CloudFront de manera que presente una página de error personalizada cuando se bloquee una solicitud.



Para obtener más información acerca de cómo potenciar la resiliencia de las aplicaciones que se ejecutan en AWS ante los ataques DDoS, consulte los siguientes recursos:

- Documento técnico de AWS: Prácticas recomendadas de AWS para la resiliencia ante DDoS
- Publicación del blog de seguridad de AWS: How to Help Protect Dynamic Web Applications Against DDoS Attacks by Using Amazon CloudFront and Amazon Route 53

Estos son algunos de los aprendizajes clave de esta lección:

- Amazon CloudFront es un servicio de CDN global que acelera la entrega de contenido, incluido el estático y el de video, a los usuarios sin compromisos de uso mínimo.
- CloudFront utiliza una red mundial que comprende ubicaciones perimetrales y cachés perimetrales regionales para entregar contenido a los usuarios.
- Para utilizar CloudFront en la entrega de contenido, especifica un servidor de origen y configura una distribución de CloudFront. CloudFront asigna un nombre de dominio y envía la configuración de su distribución a todas sus ubicaciones perimetrales.
- Puede utilizar Amazon CloudFront para potenciar la resiliencia de las aplicaciones que se ejecutan en AWS ante los ataques DDoS.

Ahora, completará el Laboratorio guiado de la unidad: Streaming de contenido dinámico con Amazon CloudFront.

Laboratorio guiado: caso

Amazon Elastic Transcoder

Amazon CloudFront

En este laboratorio, utiliza Amazon Elastic Transcoder para convertir un video de origen a varias velocidades de bits. Utiliza Amazon CloudFront para entregar la transmisión dinámica de varias velocidades de bits a un dispositivo conectado mediante el protocolo HTTP Live Streaming (HLS) de Apple.

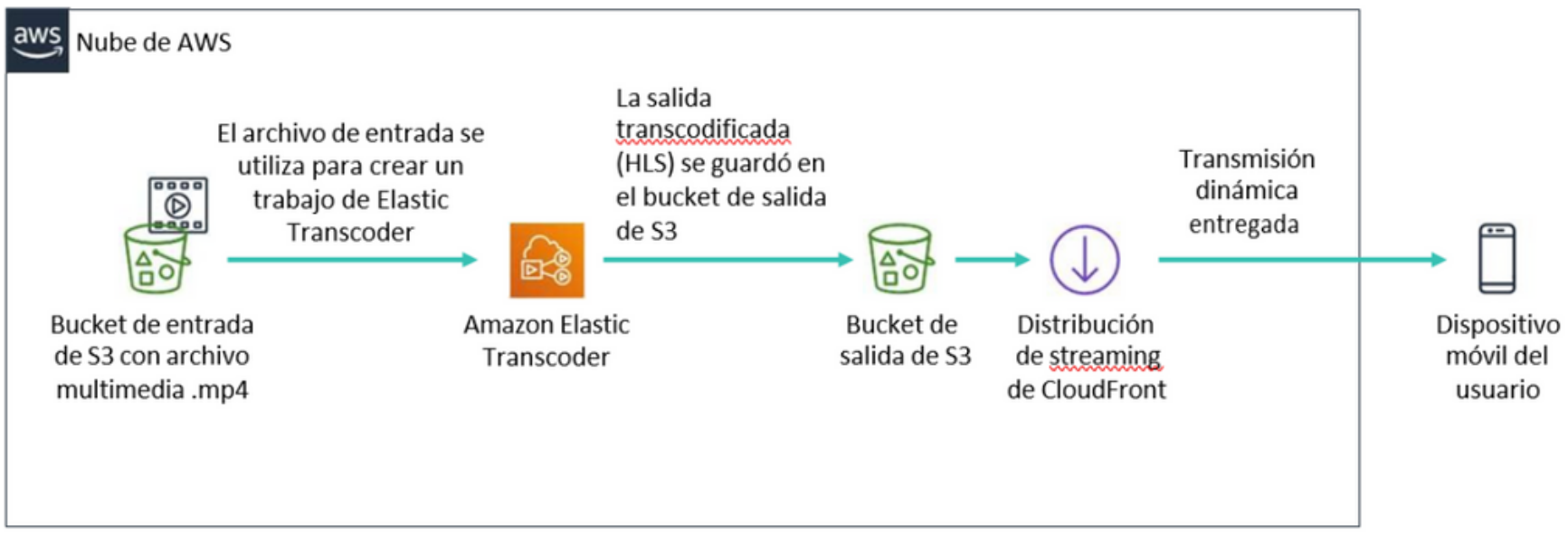
En este laboratorio, utiliza Amazon Elastic Transcoder para convertir un video de origen a varias velocidades de bits. Utiliza Amazon CloudFront para entregar la transmisión dinámica de varias velocidades de bits a un dispositivo conectado mediante el protocolo HTTP Live Streaming (HLS) de Apple. La transmisión se puede reproducir en cualquier navegador que admita el protocolo HLS.

HLS de Apple puede ajustar de forma dinámica la calidad de reproducción de películas de manera que coincida con la velocidad disponible de las redes cableadas o inalámbricas utilizando un servidor web normal. Funciona creando diferentes transmisiones de calidad. Luego, cada transmisión se divide en segmentos que se transmiten de manera secuencial a un dispositivo cliente. Del lado del cliente, puede seleccionar transmisiones de diversas velocidades de bits, las cuales permiten que las sesiones de streaming se adapten a diferentes velocidades de red.

En este laboratorio guiado, realizará las siguientes tareas:

1. crear una distribución de Amazon CloudFront
2. crear una canalización de Amazon Elastic Transcoder
3. probar la reproducción de la transmisión dinámica (varias velocidades de bits)

Laboratorio guiado: producto final



En el diagrama, se resume lo que habrá creado después de terminar el laboratorio.

Llegó la hora de iniciar el laboratorio guiado.

El instructor podría mediar una conversación sobre los aprendizajes clave de este laboratorio guiado, una vez que haya terminado.