- ***Data fabric.*** *Data fabric* se centra en proporcionar una capa unificada y coherente sobre los datos dispersos en diferentes sistemas y ubicaciones, independientemente de quién sea responsable de su gestión. Proporciona una infraestructura tecnológica para integrar, gestionar y analizar datos de manera centralizada, lo que facilita el acceso y el análisis de datos en toda la organización.

Implementación

- ***Data mesh.*** La implementación de *data mesh* requiere cambios significativos en la estructura organizativa y en la cultura empresarial, así como en las prácticas de gestión de datos. Se centra en la descentralización de la gestión de datos en cada dominio de negocio al promover la autonomía y la responsabilidad de los equipos en la gestión de sus propios datos.

- ***Data fabric.*** La implementación de *data fabric* requiere integrar múltiples tecnologías y sistemas en una única plataforma unificada. Se centra en proporcionar una capa unificada y coherente sobre los datos dispersos en diferentes sistemas y ubicaciones, lo que facilita su acceso, gestión y análisis de manera eficiente y efectiva.

En resumen, *data mesh* y *data fabric* son enfoques arquitectónicos destinados a abordar los desafíos de la gestión de datos en entornos empresariales cada vez más complejos y distribuidos. Aunque comparten algunos conceptos y objetivos comunes, también presentan **diferencias significativas** en términos de enfoque, implementación y alcance. *Data mesh* se centra en descentralizar la gestión de datos en cada dominio de negocio, mientras que *data fabric* se centra en proporcionar una capa unificada y coherente sobre los datos dispersos en diferentes sistemas y ubicaciones.

3.3. Sistemas de almacenamiento: AWS

Sistema de almacenamiento AWS S3

Amazon S3 es un servicio ofrecido por AWS que proporciona almacenamiento de objetos a través de una interfaz de servicio web. Es uno de los servicios de almacenamiento en la nube más populares y ampliamente utilizados ofrecidos por AWS. Es un servicio de almacenamiento de objetos altamente escalable, seguro y duradero, diseñado para almacenar y recuperar grandes volúmenes de datos de manera eficiente y confiable. S3 es una solución versátil que se adapta a una amplia gama de casos de uso, desde almacenamiento de datos empresariales hasta alojamiento de sitios web estáticos y entrega de contenido multimedia a escala mundial, que ha creado el usuario a dicho archivo y adjunta un identificador personalizado.

Respecto a AWS S3, sus creadores lo definen como un **servicio de almacenamiento multipropósito basado en la nube,** que permite guardar y recuperar datos, cualquiera que sea su volumen, en el que se paga de acuerdo con la capacidad de almacenamiento utilizada y al volumen de trasferencias realizadas. En ese sentido, puede ser utilizado para almacenar sitios web estáticos, realizar cualquier *backup* o copia de seguridad de archivos, alojar imágenes, entre otros.

Características clave de Amazon S3

- **Escalabilidad.** Amazon S3 permite almacenar una cantidad ilimitada de datos y escalar horizontalmente según sea necesario para satisfacer las demandas de almacenamiento de cualquier tamaño.
- **Durabilidad y disponibilidad.** Ofrece una alta durabilidad de datos mediante la replicación y distribución de objetos en múltiples ubicaciones dentro de una región de AWS, lo que garantiza la disponibilidad de datos incluso en caso de fallas de *hardware* o interrupciones del servicio.

- **Seguridad.** Amazon S3 ofrece una amplia gama de características de seguridad, incluyendo control de acceso basado en políticas, cifrado de datos en reposo y en tránsito, y monitoreo de acceso a los datos para proteger la confidencialidad, integridad y disponibilidad de la información almacenada.
- **Facilidad de uso.** S3 proporciona una interfaz de usuario intuitiva y una API robusta que facilita la creación, gestión y automatización de operaciones de almacenamiento de datos.
- **Flexibilidad.** Soporta una variedad de tipos de datos, incluyendo archivos, imágenes, vídeos, documentos y datos estructurados, lo que lo hace adecuado para una amplia gama de aplicaciones y casos de uso.
- **Integración.** S3 se integra estrechamente con otros servicios de AWS, como Amazon EC2, AWS Lambda y Amazon CloudFront, lo que permite construir aplicaciones escalables y altamente disponibles en la nube de AWS.

En resumen, **Amazon S3** es un servicio de almacenamiento en la nube altamente escalable, seguro y duradero, diseñado para almacenar y recuperar grandes volúmenes de datos de manera eficiente y confiable. Con su amplia gama de características y su integración con otros servicios de AWS, S3 es una solución versátil que se adapta a una amplia variedad de casos de uso en entornos empresariales y de desarrollo de aplicaciones.

Un esquema de AWS S3 y sus características lo podemos ver en la Figura 12.

Figura 12. Amazon S3. Fuente: Amazon Web Services, s. f.-a.

Funcionamiento y componentes de AWS S3

Amazon S3 es un servicio de almacenamiento en la nube altamente escalable, seguro y duradero ofrecido por AWS. Su **funcionamiento** se basa en una arquitectura distribuida que permite almacenar y recuperar grandes volúmenes de datos de manera eficiente y confiable. A través de una interfaz simple y robusta, Amazon S3 ofrece una solución versátil para almacenar una amplia variedad de datos, desde archivos estáticos hasta datos dinámicos generados por aplicaciones en tiempo real.

El funcionamiento de Amazon S3 se fundamenta en varios **componentes** clave:

- ***Buckets.*** Los *buckets* son contenedores de nivel superior para almacenar datos en Amazon S3. Cada *bucket* tiene un nombre único a nivel global y puede contener un número ilimitado de objetos. Los *buckets* se utilizan para organizar y gestionar los datos almacenados en S3, y se pueden configurar con políticas de almacenamiento para controlar el acceso y la protección de los datos.

- **Objetos.** Los objetos son los elementos individuales almacenados en un *bucket* de Amazon S3. Cada objeto consiste en datos (generalmente un archivo) y metadatos asociados, como su nombre, tamaño, tipo de contenido y fecha de modificación. Los objetos se identifican mediante una clave única dentro de un *bucket* y se accede a ellos a través de una URL única generada por S3.

- **Claves.** Las claves son los identificadores únicos de los objetos dentro de un *bucket*. Una clave es una cadena de caracteres que especifica la ubicación y el nombre de un objeto dentro de un *bucket*. Por ejemplo, la clave «documentos/factura.pdf» especifica un objeto llamado «factura.pdf» que está almacenado en una carpeta llamada «documentos» dentro de un *bucket*.

- **Consistencia de lectura y escritura.** Amazon S3 ofrece una fuerte consistencia de escritura y una consistencia eventual de lectura. Esto significa que las operaciones de escritura (por ejemplo, subir un nuevo objeto) son consistentes en todas las zonas de disponibilidad inmediatamente, mientras que las operaciones de lectura (por ejemplo, obtener un objeto existente) pueden tomar algún tiempo para propagarse a todas las zonas de disponibilidad.

- **Interfaz de acceso.** Amazon S3 proporciona una interfaz de acceso basada en HTTP/HTTPS que permite a los usuarios y aplicaciones acceder a los datos almacenados en S3 desde cualquier lugar del mundo. Esta interfaz de acceso incluye una API RESTful, así como herramientas de línea de comandos y SDK para varios lenguajes de programación.

- **Regiones y zonas de disponibilidad.** Amazon S3 está disponible en varias regiones geográficas en todo el mundo. Cada región de AWS consta de varias zonas de disponibilidad, que son ubicaciones físicas separadas y aisladas dentro de la región. Los datos almacenados en S3 se replican automáticamente en múltiples zonas de disponibilidad dentro de una región para garantizar la durabilidad y la disponibilidad de los datos.

- **Políticas de almacenamiento.** Amazon S3 permite configurar políticas de almacenamiento para controlar el acceso, la protección y el ciclo de vida de los datos almacenados en S3. Estas políticas pueden aplicarse a nivel de *bucket* u objeto y permiten gestionar de manera flexible el acceso y la protección de los datos almacenados en S3.

En resumen, **Amazon S3** es un servicio de almacenamiento en la nube altamente escalable, seguro y duradero, que permite almacenar y recuperar grandes volúmenes de datos de manera eficiente y confiable. Sus principales componentes incluyen *buckets*, objetos, regiones y zonas de disponibilidad, interfaz de acceso, políticas de almacenamiento y versiones. Estos componentes trabajan juntos para proporcionar una plataforma robusta y flexible para el almacenamiento y la gestión de datos en la nube.

Clases de almacenamiento en S3

Amazon S3 ofrece una variedad de **clases de almacenamiento** entre las cuales puede elegir en función de los requisitos de rendimiento, acceso a los datos, resiliencia y costos de sus cargas de trabajo. Las clases de almacenamiento de S3 se crearon específicamente para brindar el menor costo posible de almacenamiento para los diferentes patrones de acceso. Amazon S3 ofrece diferentes clases de almacenamiento diseñadas para adaptarse a diversos casos de uso y requisitos de rendimiento y costes.

Amazon S3 proporciona el **almacenamiento más duradero** de la nube. Gracias a su arquitectura única, S3 está diseñado para superar el 99,999999999 % (once nueves) de durabilidad de los datos. Además, S3 almacena los datos de forma redundante en un mínimo de tres zonas de disponibilidad de forma predeterminada, lo que proporciona resiliencia integrada contra desastres generalizados. Los clientes pueden almacenar los datos en una única zona de disponibilidad para minimizar los costos de almacenamiento o la latencia, en varias zonas de disponibilidad para evitar la pérdida permanente de todo un centro de datos o en varias regiones de AWS para cumplir con los requisitos de resiliencia geográfica.

Amazon S3 también ofrece capacidades que sirven para administrar sus datos durante todo su ciclo de vida. Una vez configurada una **política de ciclo de vida** de S3, sus datos se transferirán.

A continuación, repasamos las diferentes clases de almacenamiento de S3. En primero lugar, tenemos **Amazon S3 estándar** (S3 estándar). Es la clase de almacenamiento estándar de Amazon S3. Ofrece alta durabilidad, disponibilidad y rendimiento, lo que lo hace ideal para una amplia variedad de casos de uso, incluyendo almacenamiento de datos empresariales, alojamiento de sitios web, distribución de contenido multimedia y copias de seguridad. Sus **características** clave son:

- **Alta durabilidad.** S3 Standard proporciona una durabilidad excepcional para los datos almacenados. Amazon S3 está diseñado para ofrecer una durabilidad del 99,999999999 % (once nueves), lo que significa que es extremadamente poco probable que se pierdan datos.

- **Alta disponibilidad.** Los datos almacenados en S3 Standard están disponibles para su acceso casi instantáneo en todo momento. Amazon S3 está diseñado para ofrecer una disponibilidad del 99,99 %, lo que significa que los datos son accesibles con muy poco tiempo de inactividad.

- **Rendimiento consistente.** S3 Standard ofrece un rendimiento consistente y predecible para el acceso a los datos, independientemente del tamaño del objeto o el volumen de tráfico. Esto lo hace adecuado para una amplia variedad de aplicaciones, desde sitios web hasta aplicaciones empresariales.

- **Escalabilidad ilimitada.** S3 Standard está diseñado para escalar de manera transparente según las necesidades de almacenamiento de tu aplicación. Puedes almacenar cualquier cantidad de datos y S3 se encargará de la escalabilidad y la gestión de la infraestructura subyacente.

- **Acceso a datos en tiempo real.** Los datos almacenados en S3 Standard están disponibles para su acceso en tiempo real a través de una interfaz simple basada en HTTP/HTTPS. Puedes acceder a los datos desde cualquier lugar del mundo utilizando una variedad de herramientas y servicios de AWS.

- **Compatibilidad con versiones y replicación.** S3 Standard admite la versión de objetos y la replicación de datos entre regiones, lo que te permite mantener un historial completo de cambios y revisiones para tus datos y garantizar su disponibilidad en caso de desastre.

- **Tarifas competitivas.** A pesar de ofrecer un rendimiento y una durabilidad excepcionales, S3 Standard tiene tarifas competitivas que lo hacen asequible para una amplia gama de aplicaciones y cargas de trabajo.

Por otro lado, tenemos **Amazon S3 Intelligent-Tiering** (S3 Intelligent-Tiering). La clase de almacenamiento Amazon S3 Intelligent-Tiering es una opción avanzada dentro de Amazon S3 que utiliza aprendizaje automático para analizar el patrón de acceso a los datos y mover automáticamente los objetos entre las clases de almacenamiento *standard* y *standard-IA* (*infrequently accessed*). Esta opción está diseñada para optimizar los costos de almacenamiento al garantizar que los datos se almacenen en la clase de almacenamiento más adecuada en función de su frecuencia de acceso. Entre sus **características** clave tenemos:

- **Ahorros de costos automáticos** para datos con patrones de acceso desconocidos o que cambian constantemente.

- Los **niveles de acceso** frecuente, poco frecuente e instantáneo al archivo tienen la misma baja latencia y alto nivel de rendimiento de S3 Standard.

  - El nivel de acceso **poco frecuente** ahorra hasta un 40 % en costos de almacenamiento.

  - El nivel de acceso **instantáneo** al archivo ahorra hasta un 68 % en costos de

almacenamiento.

- Incluye **capacidades opcionales** de archivo asíncrono para objetos a los que se accede de forma inusual.

Continuamos con **Amazon S3 Express One Zone,** que es una clase de almacenamiento de zona de disponibilidad única y alto rendimiento diseñada específicamente para ofrecer un acceso constante en milisegundos a los datos de un solo dígito para los datos a los que se accede con más frecuencia y las aplicaciones sensibles a la latencia.

Además, puede mejorar diez veces la velocidad de acceso a los datos y reducir los costes de solicitud en un 50 % en comparación con S3 Standard. Los datos se almacenan en un tipo de *bucket* diferente (un *bucket* de directorio de Amazon S3), que admite cientos de miles de solicitudes por segundo. Además, se puede utilizar con servicios, como el entrenamiento de modelos Amazon SageMaker, Amazon Athena, Amazon EMR y el catálogo de datos de AWS Glue, para acelerar sus cargas de trabajo de ML y análisis. Con S3 Express One Zone, el almacenamiento se incrementa o reduce automáticamente en función de su consumo y necesidad, y ya no necesita administrar varios sistemas de almacenamiento para cargas de trabajo de baja latencia. Estas son sus **características** clave:

- **Almacenamiento** de alto rendimiento para los datos a los que se accede con más frecuencia.
- **Latencia** constante de solicitudes de milisegundos de un solo dígito.
- Mejora diez veces las **velocidades** de acceso y reduce los **costos de solicitud** en un 50 % en comparación con S3 Standard.
- Optimizado para grandes **conjuntos de datos** con muchos objetos pequeños.

- Permite utilizar las API existentes de Amazon S3 con un tipo de *bucket* diferente: ***buckets* de directorio.**

- Diseñado para ofrecer una disponibilidad del 99,95 % con un SLA (*service level agreement)* de disponibilidad del 99,9 %

Asimismo, tenemos **Amazon S3 Infrequent Access** (S3 IA o S3 estándar de acceso poco frecuente). La clase de almacenamiento Amazon S3 IA, también conocida como S3 Standard-IA, es una opción diseñada para datos que se acceden con menos frecuencia que los datos almacenados en la clase de almacenamiento S3 Standard, pero que aún requieren un acceso rápido cuando sea necesario. Es una opción intermedia entre S3 Standard y Amazon S3 Glacier, ofreciendo un equilibrio entre rendimiento y costo para datos que no se acceden con tanta frecuencia como los datos en S3 Standard, pero que aún necesitan estar disponibles de manera inmediata cuando se solicitan.

A continuación, tienes algunas **características** clave de la clase de almacenamiento Amazon S3 IA:

- **Acceso rápido.** Aunque los datos se almacenan en una clase de acceso menos frecuente que S3 Standard, aún están disponibles para su acceso inmediato cuando sea necesario. Esto significa que los datos pueden ser recuperados rápidamente sin necesidad de esperar tiempos de recuperación prolongados como en el caso de Amazon S3 Glacier.

- **Bajo coste.** Amazon S3 IA ofrece tarifas más bajas que S3 Standard, lo que la hace una opción económica para datos que no se acceden con tanta frecuencia. Esto permite a las organizaciones reducir costos al optimizar el almacenamiento de datos según su frecuencia de acceso.

- **Durabilidad y disponibilidad.** Los datos almacenados en la clase de almacenamiento Amazon S3 IA disfrutan de la misma durabilidad y disponibilidad que los datos almacenados en S3 Standard. Los objetos se replican

automáticamente en múltiples zonas de disponibilidad dentro de una región para garantizar su integridad y disponibilidad.

- **Política de precios basada en el uso.** Amazon S3 IA se factura en función del volumen de almacenamiento utilizado y de las solicitudes de acceso a los datos. Esto significa que solo pagas por el almacenamiento y el acceso a los datos que realmente utilizas, lo que puede ayudar a reducir los costos operativos.

- **Políticas de ciclo de vida.** Amazon S3 IA es compatible con las políticas de ciclo de vida de Amazon S3, lo que te permite automatizar el movimiento de datos entre clases de almacenamiento en función de criterios predefinidos, como la antigüedad de los datos. Esto te permite optimizar el almacenamiento de datos y reducir costos automáticamente a lo largo del tiempo.

En resumen, **Amazon S3 IA** es una opción de almacenamiento diseñada para datos que se acceden con menos frecuencia que los datos en la clase de almacenamiento S3 Standard, pero que aún requieren un acceso rápido cuando sea necesario. Ofrece un equilibrio entre rendimiento y costo, con tarifas más bajas que S3 Standard y una disponibilidad inmediata cuando se solicita el acceso a los datos. Esto la convierte en una opción económica y práctica para almacenar datos que no se acceden con tanta frecuencia pero que aún necesitan estar disponibles de manera rápida y eficiente.

Continuamos con **Amazon S3 One Zone-Infrequent Access** (S3 One Zone-IA). La clase de almacenamiento S3 One Zone-IA es una opción de almacenamiento en frío ofrecida por AWS que está diseñada para datos que se acceden con menos frecuencia y que no requieren la redundancia de zona de disponibilidad. A diferencia de otras clases de almacenamiento, los datos en S3 One Zone-IA se almacenan en una sola zona de disponibilidad dentro de una región de AWS, lo que la hace más económica pero menos resistente a las fallas.

Estas son algunas **características** clave de S3 One Zone-IA:

- **Coste reducido.** S3 One Zone-IA ofrece tarifas más bajas en comparación con otras opciones de almacenamiento en frío de Amazon S3, como S3 Standard-IA. Al almacenar los datos en una sola zona de disponibilidad, AWS puede reducir los costos operativos y ofrecer precios más competitivos.

- **Menor resistencia a fallas.** A diferencia de S3 Standard-IA, que almacena datos en múltiples zonas de disponibilidad dentro de una región, S3 One Zone-IA almacena datos en una sola zona de disponibilidad. Esto significa que es menos resistente a las fallas y puede estar sujeto a interrupciones si ocurre un evento que afecta a esa zona específica.

- **Disponibilidad inmediata.** A pesar de su menor resistencia a fallas, los datos almacenados en S3 One Zone-IA aún están disponibles para su acceso inmediato cuando sea necesario. Los tiempos de acceso y recuperación son similares a los de otras opciones de almacenamiento en frío de Amazon S3.

- **Política de precios basada en el uso.** S3 One Zone-IA se factura en función del volumen de almacenamiento utilizado y de las solicitudes de acceso a los datos, al igual que otras opciones de almacenamiento en frío de Amazon S3. Esto significa que solo pagas por el almacenamiento y el acceso a los datos que realmente utilizas.

S3 One Zone-IA es adecuado para ciertos **casos de uso** donde la redundancia de zona de disponibilidad no es crítica y los costos son un factor importante. Por ejemplo, puede ser útil para almacenar copias de respaldo de datos que ya se replican en otra ubicación o para datos no críticos que se pueden recuperar fácilmente en caso de una interrupción.

En resumen, **S3 One Zone-IA** es una opción de almacenamiento en frío de bajo costo diseñada para datos que se acceden con menos frecuencia y que no requieren la redundancia de zona de disponibilidad. Ofrece tarifas más bajas en comparación con otras opciones de almacenamiento en frío de Amazon S3, pero es menos

resistente a las fallas al almacenar los datos en una sola zona de disponibilidad. Es adecuado para ciertos casos de uso donde la redundancia de zona de disponibilidad no es crítica y los costos son un factor importante.

Seguimos con **Amazon S3 Glacier,** cuya clase de almacenamiento es una opción de almacenamiento en frío de bajo costo ofrecida por AWS. Está diseñada para archivar datos a largo plazo que rara vez se acceden pero que deben mantenerse disponibles para cumplir con requisitos de cumplimiento, regulaciones o necesidades de negocio. Glacier ofrece una durabilidad excepcional y tarifas extremadamente bajas, lo que lo hace ideal para almacenar grandes volúmenes de datos que no se necesitan con frecuencia pero que deben mantenerse seguros y disponibles.

Aquí tienes algunas **características** clave de Amazon S3 Glacier:

- **Bajo costo.** Glacier ofrece tarifas extremadamente bajas para el almacenamiento de datos en comparación con otras opciones de almacenamiento en Amazon S3. Esto lo hace ideal para archivar grandes volúmenes de datos que no se acceden con frecuencia pero que deben mantenerse disponibles para cumplir con requisitos de cumplimiento o regulaciones.

- **Durabilidad extrema.** Al igual que otras clases de almacenamiento en Amazon S3, Glacier ofrece una durabilidad excepcional para los datos almacenados. Los datos se replican automáticamente en múltiples dispositivos de almacenamiento y ubicaciones geográficas para garantizar su integridad y disponibilidad a lo largo del tiempo.

- **Alta disponibilidad.** Aunque Glacier ofrece tiempos de acceso más largos que otras opciones de almacenamiento en Amazon S3, los datos almacenados en Glacier están disponibles para su acceso en cuestión de minutos u horas, dependiendo de la opción de recuperación seleccionada.

- **Opciones de recuperación.** Glacier ofrece varias opciones de recuperación para adaptarse a diferentes necesidades y requisitos de tiempo. Estas opciones incluyen recuperación estándar, *expedited* y *bulk*, que varían en términos de tiempo de recuperación y tarifas asociadas.

- **Facilidad de gestión.** Glacier se integra estrechamente con otros servicios de AWS, lo que facilita la gestión y el acceso a los datos almacenados en Glacier. Puedes utilizar la consola de administración de AWS, la API de Glacier o herramientas de terceros para gestionar tus datos de forma eficiente.

- **Cumplimiento y seguridad.** Glacier ofrece características avanzadas de cumplimiento y seguridad para proteger los datos almacenados. Esto incluye cifrado de datos en reposo y en tránsito, control de acceso basado en políticas y registros de auditoría para rastrear el acceso y las modificaciones a los datos.

En resumen, **Amazon S3 Glacier** es una opción de almacenamiento en frío de bajo costo y alta durabilidad diseñada para archivar datos a largo plazo que rara vez se acceden pero que deben mantenerse disponibles para cumplir con requisitos de cumplimiento o regulaciones. Ofrece tarifas extremadamente bajas, durabilidad excepcional y varias opciones de recuperación para adaptarse a diferentes necesidades y requisitos de tiempo.

Amazon S3 Glacier Instant Retrieval es una opción de recuperación de datos dentro de Amazon S3 Glacier que permite recuperar datos de forma rápida y en tiempo real. A diferencia de las opciones de recuperación estándar, que pueden llevar horas o incluso días, *expedited retrieval* proporciona acceso inmediato a los datos almacenados en Glacier.

A continuación, se muestran algunas **características** clave de Amazon S3 Glacier Instant Retrieval:

- **Acceso rápido a los datos.** Instant Retrieval ofrece acceso a los datos almacenados en Glacier en cuestión de minutos, lo que permite recuperar información de manera rápida y eficiente cuando sea necesario.
- **Adecuado para cargas de trabajo críticas.** Esta opción de recuperación es ideal para cargas de trabajo que requieren acceso inmediato a los datos almacenados en Glacier, como recuperaciones de emergencia o situaciones en las que el tiempo es esencial.
- **Coste adicional.** Si bien *expedited retrieval* ofrece acceso rápido a los datos, hay un costo adicional asociado con esta opción. El precio puede ser más alto que las opciones estándar de recuperación, por lo que es importante evaluar el equilibrio entre velocidad y costo según los requisitos de tu carga de trabajo.
- **Facilidad de uso.** *Expedited retrieval* se puede activar fácilmente al solicitar la recuperación de datos desde Glacier. Simplemente selecciona la opción de recuperación acelerada y los datos estarán disponibles en minutos.
- **Seguridad.** Al igual que con todas las opciones de almacenamiento en Amazon S3, *expedited retrieval* ofrece seguridad integrada para proteger los datos durante la transferencia y el acceso.

En resumen, **Amazon S3 Glacier Instant Retrieval** es una opción que ofrece acceso rápido a los datos almacenados en Glacier en minutos. Es adecuado para cargas de trabajo críticas que requieren recuperación inmediata de datos, aunque hay un costo adicional asociado. Esta característica proporciona un equilibrio entre velocidad y costo para satisfacer las necesidades de diferentes aplicaciones y cargas de trabajo.

Amazon S3 Glacier Flexible Retrieval ofrece almacenamiento de bajo costo, hasta un 10 % menor que S3 Glacier Instant Retrieval, para los datos de archivo a los que se accede una o dos veces al año y se recuperan de manera asíncrona. Para los datos de archivo que no requieren acceso inmediato, pero necesitan la flexibilidad de recuperar grandes conjuntos de datos sin costo alguno, como los casos de uso de copias de seguridad o recuperación de desastres, S3 Glacier Flexible Retrieval (antes S3 Glacier), es la clase de almacenamiento ideal.

Además, ofrece las opciones de recuperación más flexibles que equilibran el costo con los tiempos de acceso que varían de minutos a horas, con recuperaciones masivas gratuitas. Esto es una solución ideal para las necesidades de copia de seguridad, recuperación de desastres, almacenamiento de datos fuera del sitio y para cuando algunos datos deban recuperarse ocasionalmente en minutos y no desee preocuparse por los costos. Entre sus **características** clave, tenemos:

- Crear **copias de seguridad y archivado de datos** a los que se accede con poca frecuencia y a bajo costo.
- Diseñado para ofrecer una **disponibilidad del 99,99 %** con un SLA de disponibilidad del 99,9 %.
- **Admite SSL** para los datos en tránsito y cifrado de datos en reposo.
- Es ideal para casos de uso de **copia de seguridad y recuperación de desastres** cuando, ocasionalmente, se deban recuperar grandes conjuntos de datos en minutos sin preocuparse por los costos.
- Los **tiempos de recuperación** se pueden configurar de minutos a horas con recuperaciones masivas gratuitas.
- **Tiene la API PUT de S3 para cargas directas a S3 Glacier Flexible Retrieval** y la administración del ciclo de vida de S3 para la migración automática de objetos

Por último, tenemos **Amazon S3 Glacier Deep Archive,** cuya clase de almacenamiento es la opción de almacenamiento en frío de más bajo costo ofrecida por AWS. Está diseñada para archivar datos a largo plazo que rara vez se acceden y que pueden mantenerse en almacenamiento por períodos prolongados, a menudo durante varios años. Glacier Deep Archive ofrece una durabilidad excepcional y tarifas extremadamente bajas, lo que la hace ideal para almacenar grandes volúmenes de datos que no se necesitan con frecuencia pero que deben mantenerse seguros y disponibles para cumplir con requisitos de cumplimiento, regulaciones o necesidades de negocio. Estas son algunas de sus **características** clave:

- **Tarifas extremadamente bajas.** Glacier Deep Archive ofrece las tarifas más bajas entre todas las opciones de almacenamiento en frío de Amazon S3. Estas tarifas son significativamente inferiores a las de otras clases de almacenamiento en frío, lo que la convierte en una opción económica para archivar grandes volúmenes de datos a largo plazo.

- **Durabilidad excepcional.** Al igual que otras clases de almacenamiento en Amazon S3, Glacier Deep Archive ofrece una durabilidad excepcional para los datos almacenados. Los datos se replican automáticamente en múltiples dispositivos de almacenamiento y ubicaciones geográficas para garantizar su integridad y disponibilidad a lo largo del tiempo.

- **Disponibilidad bajo demanda.** Aunque Glacier Deep Archive ofrece tiempos de acceso más largos que otras opciones de almacenamiento en Amazon S3, los datos almacenados en Glacier Deep Archive están disponibles para su acceso en cuestión de horas. Esto significa que los datos pueden ser recuperados de manera rápida y eficiente cuando sea necesario.

- **Política de precios basada en el uso.** Glacier Deep Archive se factura en función del volumen de datos almacenados y de las solicitudes de acceso a los datos. Esto significa que solo pagas por el almacenamiento y el acceso a los datos que realmente utilizas, lo que puede ayudar a reducir los costos operativos.

- **Cumplimiento y seguridad.** Glacier Deep Archive ofrece características avanzadas de cumplimiento y seguridad para proteger los datos almacenados. Esto incluye cifrado de datos en reposo y en tránsito, control de acceso basado en políticas y registros de auditoría para rastrear el acceso y las modificaciones a los datos.

En resumen, **Amazon S3 Glacier Deep Archive** es una opción de almacenamiento en frío de bajo costo diseñada para archivar datos a largo plazo que rara vez se acceden. Ofrece tarifas extremadamente bajas, durabilidad excepcional y disponibilidad bajo demanda, lo que la convierte en una opción económica y práctica para almacenar grandes volúmenes de datos que no se acceden con frecuencia, pero que deben mantenerse disponibles a lo largo del tiempo.

	S3 Standard	S3 Intelligent-Tiering*	S3 Standard-IA	S3 One Zone-IA†	S3 Glacier	S3 Glacier Deep Archive
Designed for durability	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)
Designed for availability	99.99%	99.9%	99.9%	99.5%	99.99%	99.99%
Availability SLA	99.9%	99%	99%	99%	99.9%	99.9%
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum capacity charge per object	N/A	N/A	128KB	128KB	40KB	40KB
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days
Retrieval fee	N/A	N/A	per GB retrieved	per GB retrieved	per GB retrieved	per GB retrieved
First byte latency	milliseconds	milliseconds	milliseconds	milliseconds	select minutes or hours	select hours
Storage type	Object	Object	Object	Object	Object	Object
Lifecycle transitions	Yes	Yes	Yes	Yes	Yes	Yes

Figura 13. Sistemas de almacenamiento AWS S3. Fuente: Jineshkumar, 2021.

AWS S3 Batch Operations

Las **operaciones por lotes** (*batch operations*) en Amazon S3 son un conjunto de herramientas que te permiten realizar cambios masivos en tus objetos de almacenamiento de una manera eficiente y económica. Estas operaciones pueden aplicarse a múltiples objetos a la vez, lo que puede ser útil para tareas como la gestión de versiones, la aplicación de políticas de retención o la modificación de metadatos en grandes conjuntos de datos. Una descripción de los **componentes básicos** de las operaciones por lotes en Amazon S3 es la siguiente:

- ***Job.*** Un trabajo (*job*) en Amazon S3 Batch Operations es una solicitud para realizar una tarea específica en un conjunto de objetos de almacenamiento. Esto puede incluir tareas como cambiar los permisos de acceso, aplicar políticas de retención o eliminar objetos. Los trabajos se pueden configurar a través de la consola de Amazon S3, la API de S3 o la interfaz de línea de comandos (CLI).

- ***Operation.*** La operación es el tipo de acción de API, como copiar objetos, que desea que ejecute el trabajo de las operaciones por lotes. Cada trabajo realiza un único tipo de operación en todos los objetos especificados en el manifiesto.

- **Tarea.** Una tarea es la unidad de ejecución de un trabajo. Una tarea representa una llamada única a una operación de API de Amazon S3 o AWS Lambda para realizar la operación del trabajo en un único objeto. En el transcurso de la vida útil de un trabajo, las operaciones por lotes de S3 crea una tarea para cada objeto especificado en el manifiesto.

- **Manifesto.** Un manifiesto (*manifest*) es un archivo JSON que contiene la lista de objetos a los que se aplicará la operación por lotes. Puedes crear el manifiesto manualmente o generar automáticamente la lista de objetos mediante un prefijo de clave, una expresión de filtro o una lista de identificadores de objeto.

- **Presupuesto de operación.** Antes de ejecutar un trabajo, puedes establecer un presupuesto de operación (*operation budget*) que limita la cantidad máxima de

recursos que se utilizarán para completar la tarea. Esto te permite controlar los costos y evitar gastos inesperados al realizar cambios en grandes conjuntos de datos.

- **Notificaciones de estado.** Una vez que se completa un trabajo, Amazon S3 puede enviar notificaciones de estado (*status notifications*) para informarte sobre el progreso y el resultado de la operación. Puedes configurar notificaciones por correo electrónico, mensajes de texto o integraciones con otros servicios de AWS.

- **Control de acceso y seguridad.** Amazon S3 Batch Operations utiliza los mismos controles de acceso y seguridad que otros servicios de AWS. Esto te permite controlar quién puede crear, modificar o eliminar trabajos, así como acceder a los datos procesados durante las operaciones por lotes.

En resumen, las **operaciones por lotes en Amazon S3** son una herramienta poderosa para realizar cambios masivos en tus objetos de almacenamiento de una manera eficiente y económica. Los componentes básicos incluyen el trabajo en sí, el manifiesto que especifica los objetos afectados, el presupuesto de operación para controlar los costos, las notificaciones de estado para mantenerse informado sobre el progreso y el resultado del trabajo, y los controles de acceso y seguridad para proteger tus datos durante el proceso.

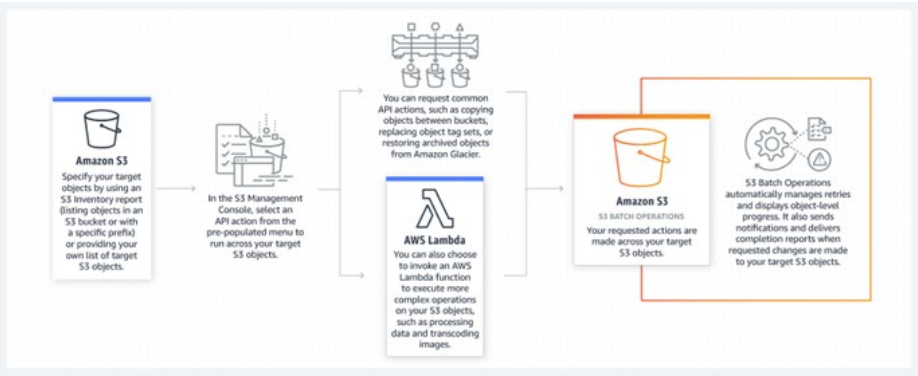

Figura 14. Operaciones por lotes de Amazon S3. Fuente: Amazon Web Services, s. f.-b.

Operaciones admitidas en S3 Batch Operations

Las operaciones admitidas por Amazon S3 Batch Operations varían según la acción que desees realizar en tus objetos de almacenamiento. Aquí tienes una lista de algunas de las operaciones comunes admitidas por S3 Batch Operations:

- **Cambio de permisos (ACL).** Puedes modificar los permisos de acceso de los objetos, como agregar o eliminar usuarios o grupos con permisos de lectura, escritura o eliminación.
- **Cambio de metadatos.** Puedes actualizar los metadatos de los objetos, como el tipo de contenido, el tipo de almacenamiento o cualquier otro atributo personalizado.
- **Aplicación de políticas de retención.** Puedes aplicar o actualizar las políticas de retención de los objetos para garantizar que no sean eliminados o modificados durante un período de tiempo específico.
- **Eliminación de objetos.** Puedes eliminar objetos de almacenamiento según criterios específicos, como su antigüedad, tamaño o ubicación.
- **Copiar objetos.** Puedes realizar copias de seguridad o duplicar objetos de almacenamiento en diferentes ubicaciones o con diferentes atributos.
- **Recuperación de objetos archivados.** Puedes iniciar la recuperación de objetos archivados en Amazon S3 Glacier o S3 Glacier Deep Archive para hacerlos accesibles de nuevo en Amazon S3.

Estas son solo algunas de las **operaciones admitidas** por Amazon S3 Batch Operations. La plataforma ofrece una amplia gama de opciones para realizar cambios masivos en tus objetos de almacenamiento de una manera eficiente y controlada.

Administración de ciclo de vida en AWS S3

AWS S3 ofrece la posibilidad de configurar el ciclo de vida de los objetos para administrarlos de manera que se almacenen de manera económica durante todo su ciclo de vida en **Amazon S3 Lifecycle.** La configuración de S3 Lifecycle es un conjunto de reglas que definen acciones que Amazon S3 aplica a un grupo de objetos. La configuración del ciclo de vida de Amazon S3 es un conjunto simple de reglas que se pueden usar para administrar el ciclo de vida de dichos datos en S3, de manera automatizada. A continuación, se describirá el concepto de control de versiones de AWS S3 y las configuraciones del ciclo de vida de Amazon S3 como concepto. Luego, se mostrará cómo aprovechar la configuración del ciclo de vida para caducar y eliminar objetos a fin de ahorrar costos de almacenamiento en la nube.

¿Qué es el **control de versiones** de objetos de Amazon S3? Para comprender completamente todas las diferentes aplicaciones de las configuraciones del ciclo de vida de Amazon S3, es importante comprender primero el concepto de **control de versiones** del depósito de S3. La función de control de versiones de Amazon S3 permite a los usuarios mantener varias versiones del mismo objeto en un depósito de S3 con fines de reversión o recuperación. Cuando un depósito de Amazon S3 está habilitado para el control de versiones, cada objeto del depósito recibe un identificador de versión que cambia cada vez que el objeto cambia o se sobrescribe. Cuando se sobrescribe un objeto, este nuevo objeto será la versión actual, mientras que el archivo anterior tendrá un ID de versión anterior.

Los usuarios pueden aprovechar las **versiones anteriores** para restaurar los objetos que se eliminan o sobrescriben según sea necesario. Cuando el usuario final ola aplicación elimina un objeto de S3, Amazon S3 insertará un marcador de eliminación. Este marcador de eliminación se convierte en la versión actual del objeto. Si bien el control de versiones de S3 es útil para varios propósitos de recuperación, significa

que una vez habilitados, los objetos más antiguos que se eliminaron o sobrescribieron continúan consumiendo la capacidad de almacenamiento subyacente, lo que genera costos de consumo.

La configuración del **ciclo de vida de AWS S3** es una colección de reglas que definen varias acciones del ciclo de vida que se pueden aplicar automáticamente a un grupo de objetos de Amazon S3. Estas acciones pueden ser **acciones de transición** (que hacen que la versión actual de los objetos de S3 cambie entre varias clases de almacenamiento de S3) o pueden ser **acciones de caducidad** (que definen cuándo caduca un objeto de S3).

Ejemplo de configuración de ciclo de vida en S3 para eliminar objetos

La **configuración del ciclo de vida** de vencimiento de la objeción de S3 se puede crear con varias herramientas diferentes: la herramienta AWS CLI, el SDK de AWS, la consola de Amazon S3 o las llamadas a la API RESTful.

Recomendamos se consulte la **guía del usuario del ciclo de vida** de Amazon S3 para obtener información detallada paso a paso.

La Figura 15 muestra cómo se puede aprovechar la interfaz de usuario de la consola de Amazon S3 (a la que se accede a través de la pestaña «Administración» dentro del depósito de S3) para configurar una regla de ciclo de vida de S3 para caducar la versión actual de los objetos S3.

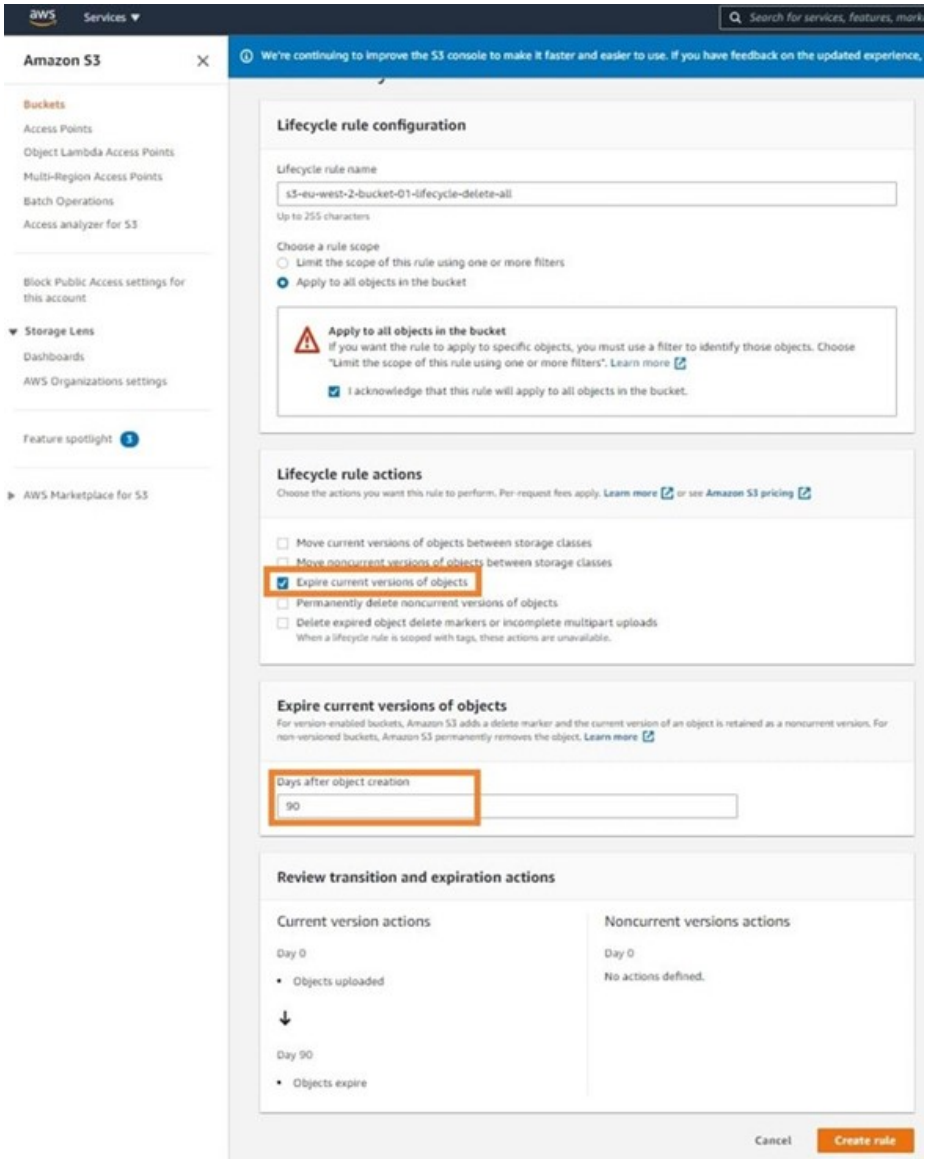

Figura 15. Regla de ciclo de vida de S3. Fuente: elaboración propia.

A los efectos de este ejemplo, se ha configurado la **caducidad** para que se active después de noventa días de la creación del objeto para todo el *bucket* para eliminar la versión actual.

Para eliminar **versiones anteriores** de objetos S3, la misma consola de AWS se puede aprovechar fácilmente para crear la regla del ciclo de vida, como se ilustra en la Figura 16.

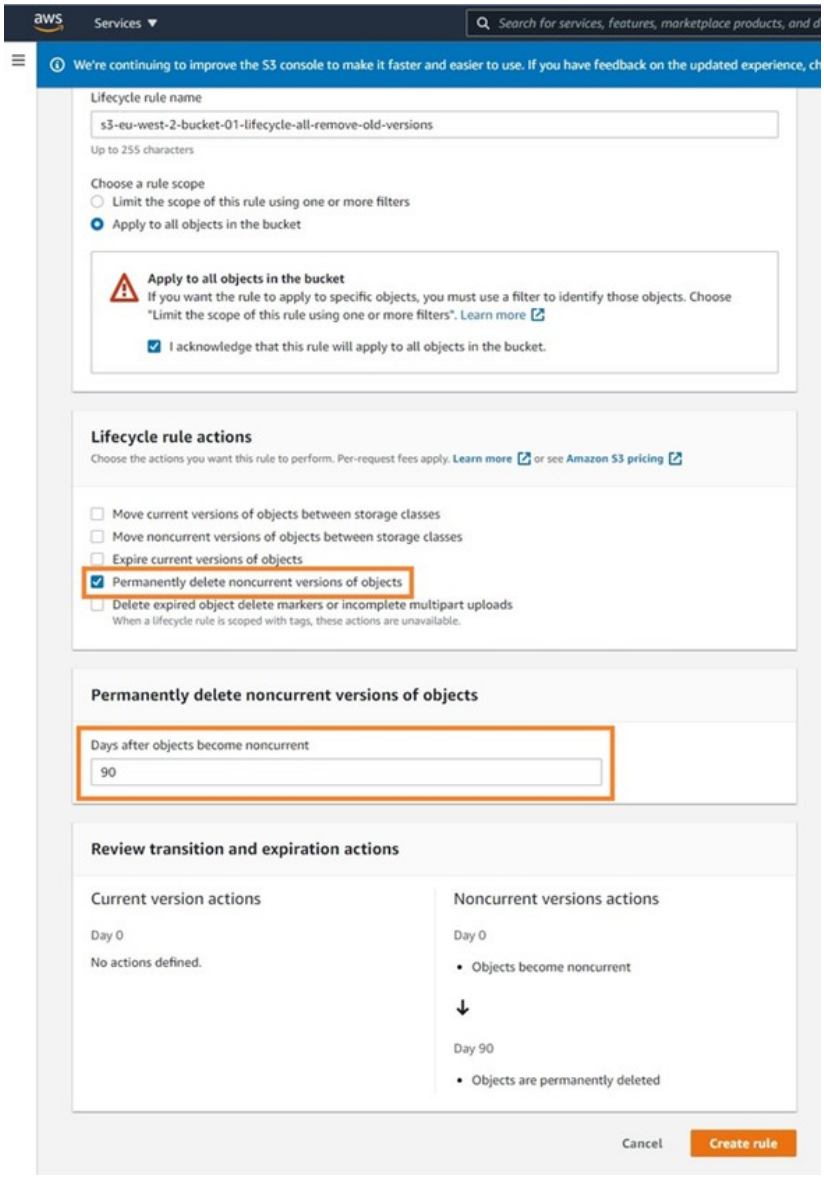

Figura 16. Regla de ciclo de vida de S3. Fuente: elaboración propia.

Replicación en AWS S3

La **replicación** de Amazon S3 es un elemento eficiente y gestionado encargado de replicar objetos almacenados en los *buckets*. Su funcionamiento permite mayor flexibilidad y optimización para el almacenamiento en la nube, lo que a su vez proporciona los controles pertinentes para el cumplimientos de las tareas empresariales.

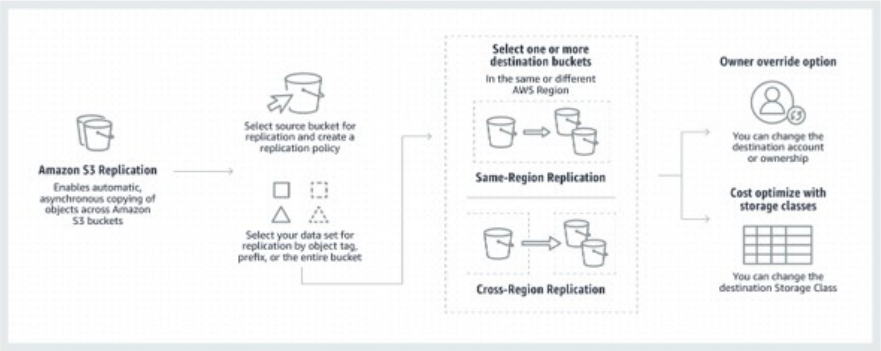

Figura 17. Cómo funciona la replicación de S3. Fuente: Amazon Web Services, s. f.-c.

En cuanto a las **características** de la replicación de objetos en Amazon S3, ofrecer la configuración de Amazon S3 para la copia automática de los objetos del sistema en distintas regiones de AWS a través de la función de réplica entre regiones (CRR) de S3 o dentro de *buckets* en una misma región de AWS por medio de la replicación en una sola región (SRR) de Amazon S3.

La replicación de Amazon S3 provee, igualmente, **notificaciones y métricas** precisas para gestionar el funcionamiento de la réplica de objetos entre *buckets*. Esto permite observar el proceso de la replicación mediante el rastreo de los bytes pendientes y supervisar las tareas pendientes, y la latencia de replicación entre los *buckets* de origen y de destino con la consola de administración de Amazon S3 o AWS CloudWatch.

De la misma manera, esta característica de Amazon S3 proporciona la **configuración de notificaciones de eventos de S3** para obtener avisos sobre replicaciones erróneas y, así, emplear un análisis y reparar los errores de administración de inmediato.

Algunos **casos de uso** de la replicación de objetos son:

- **Replicar objetos conservando los metadatos.** Puede utilizar la replicación para realizar copias de los objetos en las que se conserven todos los metadatos, como la hora de creación del objeto original y los ID de versión. Esta capacidad es importante si necesita asegurarse de que la réplica sea idéntica al objeto de origen.

- **Replicar objetos en diferentes clases de almacenamiento.** Puede utilizar la replicación para colocar objetos directamente en S3 Glacier, S3 Glacier Deep Archive u otra clase de almacenamiento en los *buckets* de destino. También puede replicar los datos en la misma clase de almacenamiento y utilizar las políticas de ciclo de vida en los *buckets* de destino para mover objetos a una clase de almacenamiento con menos actividad conforme adquiere antigüedad.

- **Mantener copias de objetos con distintos propietarios.** Independientemente de quién sea el propietario del objeto de origen, puede indicar a Amazon S3 que cambie la propiedad de la réplica a la cuenta de AWS que posee el *bucket* de destino. Esto se conoce como la opción de invalidación del propietario. Puede usar esta opción para restringir el acceso a las réplicas de objetos.

- **Mantener los objetos almacenados en varias regiones de AWS.** Puede establecer varios *buckets* de destino en diferentes regiones de AWS para garantizar las diferencias geográficas en el lugar donde se guardan los datos. Esto podría ser útil para cumplir ciertos requisitos de conformidad.

- **Reproducir objetos en quince minutos.** Puede utilizar el control de tiempo de reproducción de S3 (S3 RTC) para reproducir sus datos en la misma región de AWS o en distintas regiones dentro de un período predecible. S3 RTC replica el 99,99 %

de los objetos nuevos almacenados en Amazon S3 dentro de un plazo de quince minutos (con el respaldo de un acuerdo de nivel de servicio).

Amazon Web Services. (s. f.-d). *Cumplimiento de los requisitos de conformidad mediante el control de tiempo de replicación de S3 (S3 RTC)*. https://docs.aws.amazon.com/es_es/AmazonS3/latest/userguide/replication-time-control.html

La **reproducción entre las regiones** (CRR) de S3 se utiliza para copiar objetos en *buckets* de Amazon S3 en diferentes regiones de AWS. CRR puede ayudarlo a hacer lo siguiente:

- **Cumplir los requisitos de conformidad.** Aunque Amazon S3 almacena sus datos en diversas zonas de disponibilidad alejadas geográficamente, de forma predeterminada los requisitos de conformidad pueden exigir que almacene los datos en ubicaciones aún más alejadas. La reproducción entre las regiones permite reproducir los datos entre las regiones de AWS alejadas para cumplir con estos requisitos de conformidad.

- **Minimizar la latencia.** Si sus clientes están en dos ubicaciones geográficas, puede minimizar la latencia en el acceso a los objetos mediante el mantenimiento de copias de los objetos en las regiones de AWS que estén geográficamente más cerca de sus usuarios.

- **Aumentar la eficiencia operativa.** Si tiene clústeres de cómputo en dos regiones de AWS diferentes que analizan el mismo conjunto de objetos, puede optar por mantener copias de objetos en dichas regiones.

L a **reproducción en la misma región** (SRR) se utiliza para copiar objetos en *buckets* de Amazon S3 en la misma región de AWS. SRR puede ayudarlo a hacer lo siguiente:

- **Agregar registros en un solo *bucket.*** Si almacena registros en varios *buckets* o en varias cuentas, puede fácilmente replicar registros en un solo *bucket* en la región. Esto permite un procesamiento más simple de los registros en una sola ubicación.
- **Configurar la replicación en vivo entre las cuentas de producción y prueba.** Si usted o sus clientes tienen cuentas de producción y de prueba que utilizan los mismos datos, puede replicar objetos entre esas cuentas múltiples, mientras mantiene metadatos de objeto.
- **Cumplir las leyes de soberanía de datos.** Es posible que tenga que almacenar varias copias de sus datos en las cuentas de AWS separadas dentro de una misma región. La replicación en la misma región puede ayudarle a replicar automáticamente los datos críticos cuando las normativas de conformidad no permiten que los datos salgan de su país.

Apache HDFS

Apache HDFS es un sistema de archivos distribuido diseñado para almacenar grandes conjuntos de datos de manera confiable y escalable en clústeres de servidores estándar. Forma parte del ecosistema Apache Hadoop y es una de las tecnologías clave que permiten el procesamiento distribuido de datos en grandes cantidades. Una descripción de sus **características** sería la siguiente:

- **Distribuido y tolerante a fallos.** HDFS distribuye los datos a través de múltiples nodos de un clúster, lo que permite almacenar grandes volúmenes de datos de manera distribuida. Además, está diseñado para ser tolerante a fallos, lo que significa que puede manejar la pérdida de nodos individuales sin perder datos.

- **Bloques de datos.** HDFS divide los datos en bloques de tamaño fijo (generalmente 128 MB o 256 MB por defecto), que luego se distribuyen y almacenan en diferentes nodos del clúster. Esta división en bloques permite el procesamiento paralelo de datos y facilita la recuperación de fallos.

- **Arquitectura maestro-esclavo.** HDFS sigue una arquitectura maestro-esclavo, donde un conjunto de nodos maestros (NameNode y Secondary NameNode) gestionan el sistema de archivos y un conjunto de nodos esclavos (DataNodes) almacenan los datos reales.

  - **NameNode.** El NameNode es el nodo maestro responsable de mantener el sistema de archivos y el árbol de metadatos. Contiene información sobre la ubicación y el estado de todos los bloques de datos, así como los metadatos asociados con cada archivo.

  - **DataNode.** Los DataNodes son los nodos esclavos que almacenan los bloques de datos reales. Cada DataNode es responsable de gestionar el almacenamiento de los bloques asignados a él y de comunicarse con el NameNode para informar sobre su estado y replicar los bloques según sea necesario.