AWS Marco Well-Architected

Recuperación de cargas de trabajo ante desastres en AWS: recuperación en la nube

Recuperación de cargas de trabajo ante desastres en AWS: recuperación en la nube: AWS Marco Well-Architected

Table of Contents

Recuperación de cargas de trabajo ante desastres en AWS: recuperación en la nube

Fecha de publicación: 12 de febrero de 2021 () Historial de documentos

La recuperación ante un desastre es el proceso de preparación y recuperación ante un desastre. Se considera desastre un evento que impide que una carga de trabajo o un sistema cumpla sus objetivos empresariales en su ubicación de despliegue principal. Este paper describe las mejores prácticas para planificar y probar la recuperación ante desastres para cualquier carga de trabajo en la que AWS se despliegue y ofrece diferentes enfoques para mitigar los riesgos y cumplir el objetivo de tiempo de recuperación (RTO) y el objetivo de punto de recuperación (RPO) para esa carga de trabajo.

En este documento técnico, se explica cómo implementar la recuperación ante desastres para las cargas de trabajo de una empresa. AWS Consulte la sección Recuperación ante desastres de aplicaciones locales AWS para obtener información sobre su uso AWS como sitio de recuperación ante desastres para cargas de trabajo locales.

Introducción

La carga de trabajo debe realizar la función prevista de forma correcta y coherente. Para lograrlo, debe diseñar una arquitectura basada en la resiliencia. La resiliencia es la capacidad de una carga de trabajo para recuperarse de las interrupciones en la infraestructura, el servicio o las aplicaciones, adquirir recursos informáticos de forma dinámica para satisfacer la demanda y mitigar las interrupciones, como los errores de configuración o los problemas transitorios de la red.

La recuperación ante desastres (DR) es una parte importante de su estrategia de resiliencia y se refiere a la forma en que su carga de trabajo responde ante un desastre (un desastre es un suceso que tiene un impacto negativo grave en su empresa). Esta respuesta debe basarse en los objetivos empresariales de su organización, que especifican la estrategia de su carga de trabajo para evitar la pérdida de datos, lo que se conoce como objetivo de punto de recuperación (RPO), y reducir el tiempo de inactividad cuando la carga de trabajo no está disponible para su uso, lo que se conoce como objetivo de tiempo de recuperación (RTO). Por lo tanto, debe implementar la resiliencia en el diseño de sus cargas de trabajo en la nube para cumplir sus objetivos de recuperación (RPO y RTO) en caso de un desastre único. Este enfoque ayuda a su organización a mantener la continuidad empresarial como parte de la planificación de la continuidad empresarial (BCP).

Este paper se centra en cómo planificar, diseñar e implementar arquitecturas AWS que cumplan los objetivos de recuperación ante desastres de su empresa. La información que se comparte aquí está destinada a quienes desempeñan funciones tecnológicas, como los directores de tecnología (CTOs), los arquitectos, los desarrolladores, los miembros del equipo de operaciones y las personas encargadas de evaluar y mitigar los riesgos.

Disponibilidad y recuperación ante desastres

La recuperación ante desastres se puede comparar con la disponibilidad, que es otro componente importante de su estrategia de resiliencia. Mientras que la recuperación ante desastres mide los objetivos para eventos únicos, los objetivos de disponibilidad miden los valores medios durante un período de tiempo.

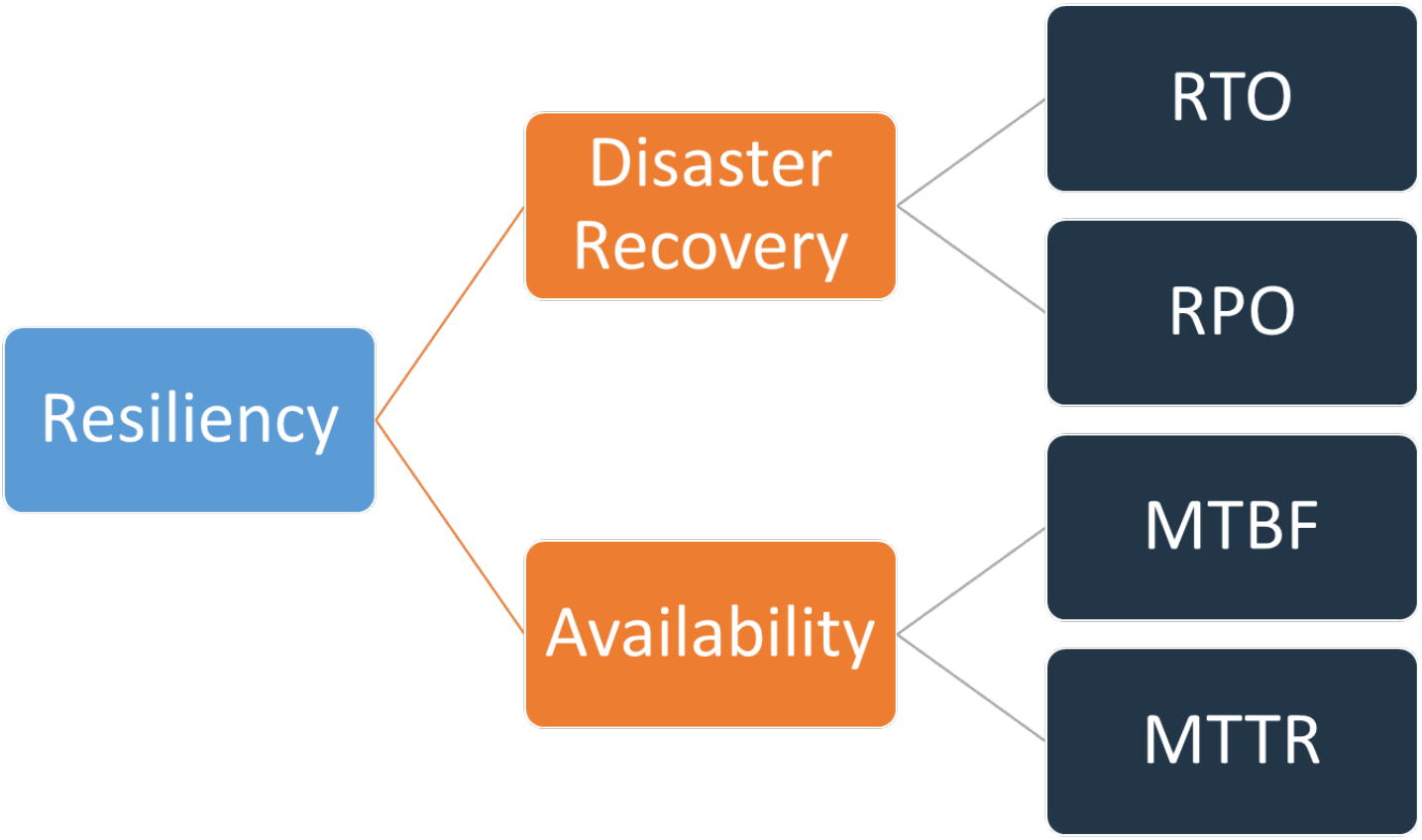

Figura 1: Objetivos de resiliencia

La disponibilidad se calcula utilizando el tiempo medio entre fallos (MTBF) y el tiempo medio de recuperación (MTTR):

$$Availability = \frac{Available\ for\ Use\ Time}{Total\ Time} = \frac{MTBF}{MTBF + MTTR}$$

Este enfoque suele denominarse «nueves», mientras que un objetivo de disponibilidad del 99,9% se denomina «tres nueves».

Para su carga de trabajo, puede ser más fácil contar las solicitudes correctas y las fallidas en lugar de utilizar un enfoque basado en el tiempo. En este caso, se puede utilizar el siguiente cálculo:

$$Availability = \frac{Successful\ Responses}{Valid\ Requests}$$

La recuperación ante desastres se centra en los desastres, mientras que la disponibilidad se centra en las interrupciones más comunes y de menor escala, como los fallos de los componentes, los problemas de red, los errores de software y los picos de carga. El objetivo de la recuperación ante desastres es la continuidad empresarial, mientras que la disponibilidad se refiere a maximizar el tiempo que una carga de trabajo está disponible para realizar la funcionalidad empresarial prevista. Ambos deben formar parte de su estrategia de resiliencia.

¿Dispone de Well-Architected?

El AWS Well-Architected Framework le ayuda a comprender las ventajas y desventajas de las decisiones que toma al crear sistemas en la nube. Los seis pilares del marco le permitirán aprender las prácticas recomendadas de arquitectura para diseñar y utilizar sistemas fiables, seguros, eficientes, rentables y sostenibles. Con la herramienta AWS Well-Architected Tool, disponible de forma gratuita en la Consola de administración de AWS, puede comparar sus cargas de trabajo con estas prácticas recomendadas respondiendo a una serie de preguntas para cada pilar.

Los conceptos que se tratan en este documento técnico amplían las mejores prácticas incluidas en el documento técnico sobre el pilar de la confiabilidad, específicamente la pregunta REL 13, «¿Cómo se planifica la recuperación ante desastres (DR)?». Tras implementar las prácticas de este documento técnico, asegúrese de revisar (o volver a revisar) su carga de trabajo con la herramienta AWS Well-Architected Tool.

Modelo de responsabilidad compartida para la resiliencia

La resiliencia es una responsabilidad compartida entre usted AWS y usted, el cliente. Es importante que comprenda cómo funcionan la recuperación ante desastres y la disponibilidad, como parte de la resiliencia, en este modelo compartido.

Responsabilidad de AWS: «Resiliencia de la nube»

AWS es responsable de la resiliencia de la infraestructura que ejecuta todos los servicios que se ofrecen en la nube de AWS. Esta infraestructura comprende el hardware, el software, las redes y las instalaciones que ejecutan los servicios en la nube de AWS. AWS hace todos los esfuerzos comercialmente razonables para que estos servicios en la nube de AWS estén disponibles, garantizando que la disponibilidad del servicio cumpla o supere los acuerdos de nivel de servicio de AWS (SLAs).

La infraestructura de nube global de AWS está diseñada para permitir a los clientes crear arquitecturas de cargas de trabajo altamente resilientes. Cada región de AWS está completamente aislada y consta de varias zonas de disponibilidad, que son particiones de la infraestructura aisladas físicamente. Las zonas de disponibilidad aíslan errores que podrían afectar la resiliencia de la carga de trabajo y les impide repercutir en otras zonas en la región. Pero, al mismo tiempo, todas las zonas de una región de AWS están interconectadas con redes de gran ancho de banda y baja latencia, a través de fibra metropolitana dedicada y totalmente redundante, lo que proporciona redes de alto rendimiento y baja latencia entre zonas. Todo el tráfico entre las zonas está cifrado. El rendimiento de la red es suficiente para llevar a cabo la replicación sincrónica entre zonas. Cuando se divide una aplicación AZs, las empresas están mejor aisladas y protegidas de problemas como cortes de energía, rayos, tornados, huracanes y más.

Responsabilidad del cliente: «Resiliencia en la nube»

Su responsabilidad estará determinada por los servicios en la nube de AWS que seleccione. Esto determina la cantidad de trabajo de configuración que debe llevar a cabo como parte de sus responsabilidades de resiliencia. Por ejemplo, un servicio como Amazon Elastic Compute Cloud (Amazon EC2) requiere que el cliente realice todas las tareas de configuración y administración de la resiliencia necesarias. Los clientes que implementan EC2 instancias de Amazon son responsables de implementar EC2 las instancias en varias ubicaciones (como las zonas de disponibilidad de AWS), implementar la autorreparación mediante servicios como Amazon EC2 Auto Scaling y utilizar

las mejores prácticas de arquitectura de carga de trabajo resiliente para las aplicaciones instaladas en las instancias. En el caso de los servicios gestionados, como Amazon S3 y Amazon DynamoDB, AWS opera la capa de infraestructura, el sistema operativo y las plataformas, y los clientes acceden a los puntos de enlace para almacenar y recuperar datos. El cliente es responsable de administrar la resiliencia de sus datos, incluidas las estrategias de copia de seguridad, control de versiones y replicación.

La implementación de la carga de trabajo en varias zonas de disponibilidad de una región de AWS forma parte de una estrategia de alta disponibilidad diseñada para proteger las cargas de trabajo al aislar los problemas en una zona de disponibilidad y utiliza la redundancia de las demás zonas de disponibilidad para seguir atendiendo las solicitudes. Una arquitectura de varias zonas de disponibilidad también forma parte de una estrategia de DR diseñada para que las cargas de trabajo estén mejor aisladas y protegidas de problemas como interrupciones de alimentación eléctrica, tormentas eléctricas, tornados, terremotos, etc. Las estrategias de recuperación ante desastres también pueden utilizar varias regiones de AWS. Por ejemplo, en una configuración activa/pasiva, el servicio de la carga de trabajo pasará por error de su región activa a su región de DR si la región activa ya no puede atender las solicitudes.

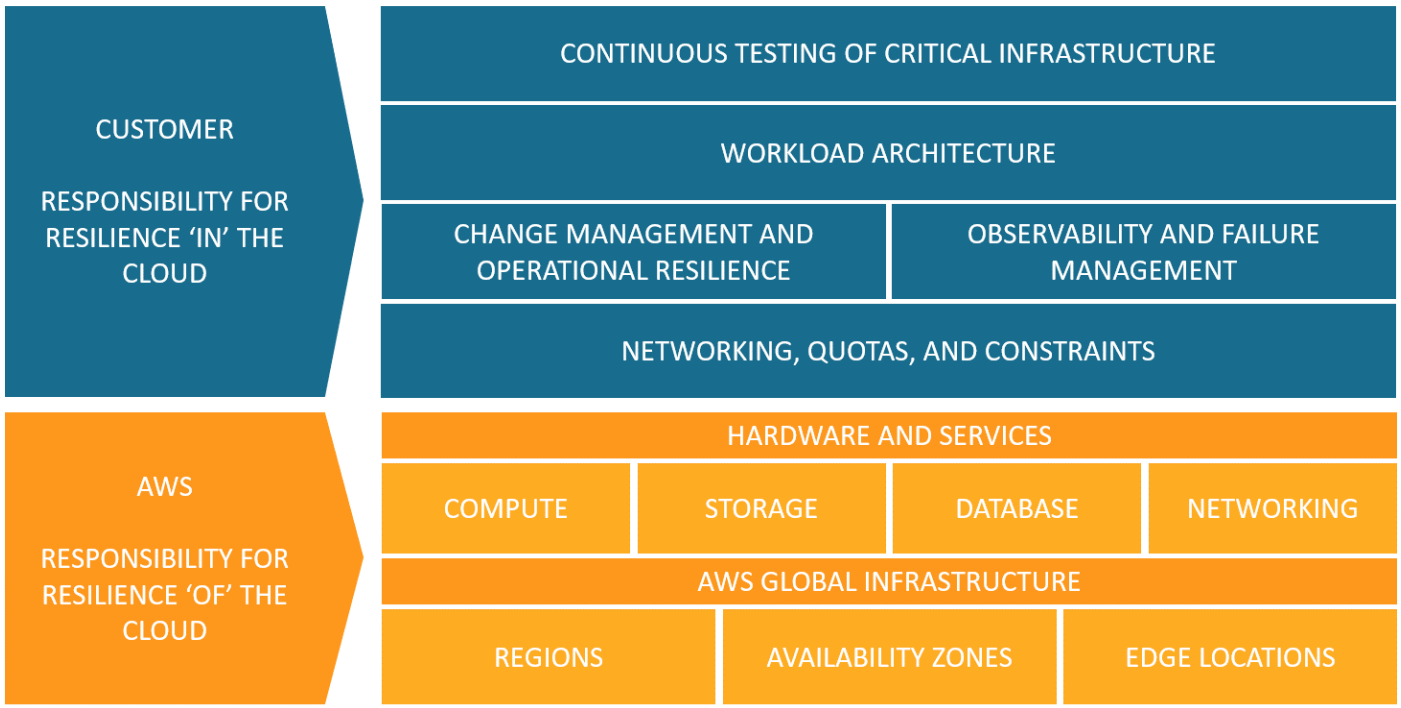

Figura 2: La resiliencia es una responsabilidad compartida entre AWS y el cliente

¿Qué es un desastre?

Cuando planifique la recuperación ante un desastre, evalúe su plan para estas tres categorías principales de desastres:

- Desastres naturales, como terremotos o inundaciones
- Fallos técnicos, como un corte de energía o de conectividad de red
- Acciones humanas, como una configuración incorrecta involuntaria o el acceso o la modificación por parte de unauthorized/outside terceros

Cada uno de estos posibles desastres también tendrá un impacto geográfico que puede ser local, regional, nacional, continental o global. Tanto la naturaleza del desastre como el impacto geográfico son importantes a la hora de considerar su estrategia de recuperación ante un desastre. Por ejemplo, puede mitigar un problema de inundación local que provoque una interrupción en el centro de datos mediante una estrategia de zonas de disponibilidad múltiples, ya que no afectaría a más de una zona de disponibilidad. Sin embargo, un ataque a los datos de producción requeriría que invocara una estrategia de recuperación ante desastres que conmute por error la copia de seguridad de los datos en otra región de AWS.

La alta disponibilidad no es recuperación ante desastres

Tanto la disponibilidad como la recuperación ante desastres se basan en algunas de las mismas prácticas recomendadas, como la supervisión de los fallos, la implementación en varias ubicaciones y la conmutación automática por error. Sin embargo, la disponibilidad se centra en los componentes de la carga de trabajo, mientras que la recuperación ante desastres se centra en las copias independientes de toda la carga de trabajo. La recuperación ante desastres tiene objetivos diferentes a los de la disponibilidad, ya que mide el tiempo de recuperación después de eventos de mayor escala que se consideran desastres. En primer lugar, debe asegurarse de que su carga de trabajo cumpla sus objetivos de disponibilidad, ya que una arquitectura de alta disponibilidad le permitirá satisfacer las necesidades de los clientes en caso de que los eventos afecten a la disponibilidad. Su estrategia de recuperación ante desastres requiere enfoques diferentes a los de la disponibilidad, y se centra en la implementación de sistemas discretos en varias ubicaciones, de modo que pueda conmutar por error toda la carga de trabajo si es necesario.

Debe tener en cuenta la disponibilidad de su carga de trabajo al planificar la recuperación ante desastres, ya que influirá en el enfoque que adopte. Una carga de trabajo que se ejecuta en una sola EC2 instancia de Amazon en una zona de disponibilidad no tiene alta disponibilidad. Si un problema de inundación local afecta a esa zona de disponibilidad, este escenario requiere una conmutación por error a otra zona de disponibilidad para cumplir los objetivos de recuperación ante desastres. Compare este escenario con una carga de trabajo de alta disponibilidad desplegada en varios sitios activa/activa, en la que la carga de trabajo se despliega en varias regiones activas y todas las regiones atienden al tráfico de producción. En este caso, incluso en el improbable caso de que un desastre masivo deje una región inutilizable, la estrategia de recuperación ante desastres consiste en enrutar todo el tráfico a las regiones restantes.

La forma de abordar los datos también es diferente entre la disponibilidad y la recuperación ante desastres. Considere una solución de almacenamiento que se replique continuamente en otro sitio para lograr una alta disponibilidad (por ejemplo, una active/active carga de trabajo de varios sitios). Si uno o varios archivos se eliminan o dañan en el dispositivo de almacenamiento principal, esos cambios destructivos se pueden replicar en el dispositivo de almacenamiento secundario. En este escenario, a pesar de la alta disponibilidad, la capacidad de conmutación por error en caso de eliminación o corrupción de los datos se verá comprometida. En cambio, también se requiere una point-in-time copia de seguridad como parte de una estrategia de recuperación ante desastres.

Plan de continuidad empresarial (BCP)

Su plan de recuperación ante desastres debe ser un subconjunto del plan de continuidad empresarial (BCP) de su organización, no debe ser un documento independiente. No tiene sentido mantener objetivos ambiciosos de recuperación ante desastres para restaurar una carga de trabajo si los objetivos empresariales de esa carga de trabajo no se pueden alcanzar debido al impacto del desastre en elementos de la empresa distintos de la carga de trabajo. Por ejemplo, un terremoto podría impedirle transportar los productos comprados en su aplicación de Ecommerce; incluso si una DR efectiva mantiene su carga de trabajo en funcionamiento, su BCP debe adaptarse a las necesidades de transporte. Su estrategia de recuperación ante desastres debe basarse en los requisitos, las prioridades y el contexto de la empresa.

Análisis del impacto empresarial y evaluación de riesgos

Un análisis del impacto empresarial debe cuantificar el impacto empresarial de una interrupción en sus cargas de trabajo. Debe identificar el impacto que tiene en los clientes internos y externos el hecho de no poder utilizar sus cargas de trabajo y el efecto que eso tiene en su empresa. El análisis debería ayudar a determinar la rapidez con la que se debe disponer de la carga de trabajo y la cantidad de pérdida de datos que se puede tolerar. Sin embargo, es importante tener en cuenta que los objetivos de recuperación no se deben establecer de forma aislada; la probabilidad de interrupción y el costo de la recuperación son factores clave que ayudan a determinar el valor empresarial de proporcionar recuperación ante desastres para una carga de trabajo.

El impacto en el negocio puede depender del tiempo. Es posible que desee considerar la posibilidad de tener esto en cuenta en su planificación de recuperación ante desastres. Por ejemplo, es probable que la interrupción de su sistema de nómina tenga un impacto muy alto en la empresa justo antes de que todos reciban el pago, pero puede tener un impacto bajo justo después de que todos hayan recibido el pago.

Una evaluación del riesgo del tipo de desastre y su impacto geográfico, junto con una visión general de la implementación técnica de la carga de trabajo, determinarán la probabilidad de que se produzcan interrupciones en cada tipo de desastre.

En el caso de cargas de trabajo muy críticas, podría considerar la posibilidad de implementar una infraestructura en varias regiones con replicación de datos y copias de seguridad continuas para minimizar el impacto empresarial. Para las cargas de trabajo menos críticas, una estrategia válida puede ser no implementar ningún tipo de recuperación ante desastres. Y para algunos escenarios

de desastre, también es válido no contar con ninguna estrategia de recuperación ante desastres como una decisión informada basada en la baja probabilidad de que se produzca el desastre. Recuerde que las zonas de disponibilidad de una región de AWS ya están diseñadas con una distancia significativa entre ellas y una planificación cuidadosa de la ubicación, de modo que los desastres más comunes solo afecten a una zona y no a las demás. Por lo tanto, es posible que una arquitectura Multi-AZ dentro de una región de AWS ya satisfaga gran parte de sus necesidades de mitigación de riesgos.

Se debe evaluar el costo de las opciones de recuperación ante desastres para garantizar que la estrategia de recuperación ante desastres proporcione el nivel correcto de valor empresarial, teniendo en cuenta el impacto y el riesgo empresarial.

Con toda esta información, puede documentar la amenaza, el riesgo, el impacto y el costo de los diferentes escenarios de desastre y las opciones de recuperación asociadas. Esta información debe usarse para determinar sus objetivos de recuperación para cada una de sus cargas de trabajo.

Objetivos de recuperación (RTO y RPO)

Al crear una estrategia de recuperación ante desastres (DR), las organizaciones suelen planificar el objetivo de tiempo de recuperación (RTO) y el objetivo de punto de recuperación (RPO).

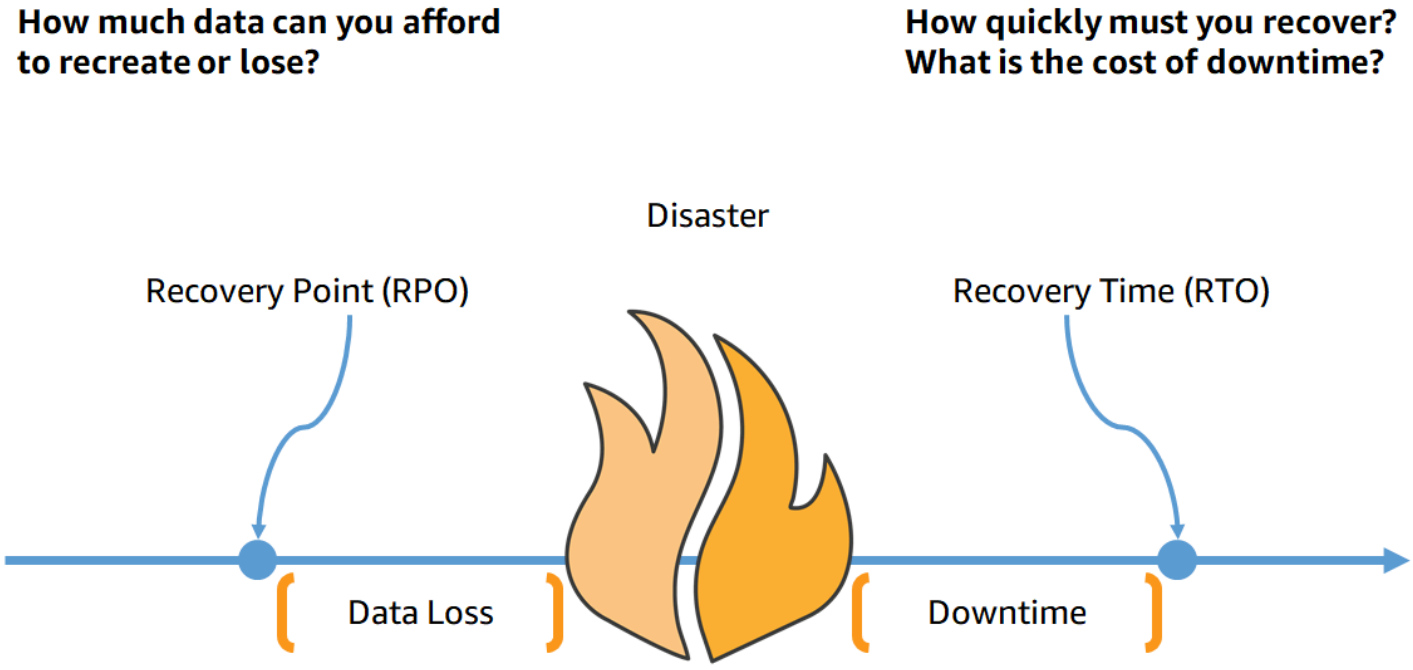

Figura 3: Objetivos de recuperación

El objetivo de tiempo de recuperación (RTO) es el retraso máximo aceptable entre la interrupción del servicio y el restablecimiento del servicio. Este objetivo determina qué período de tiempo se considera aceptable cuando el servicio no está disponible y lo define la organización.

En general, en este documento se analizan cuatro estrategias de recuperación ante desastres: copia de seguridad y restauración, piloto piloto, espera en caliente y multisitio active/active (consulte Opciones de recuperación ante desastres en la nube). En el siguiente diagrama, la empresa ha determinado su RTO máximo permitido, así como el límite de lo que puede gastar en su estrategia de restauración del servicio. Teniendo en cuenta los objetivos de la empresa, las estrategias de recuperación ante desastres Pilot Light o Warm Standby cumplirán tanto el RTO como los criterios de coste.

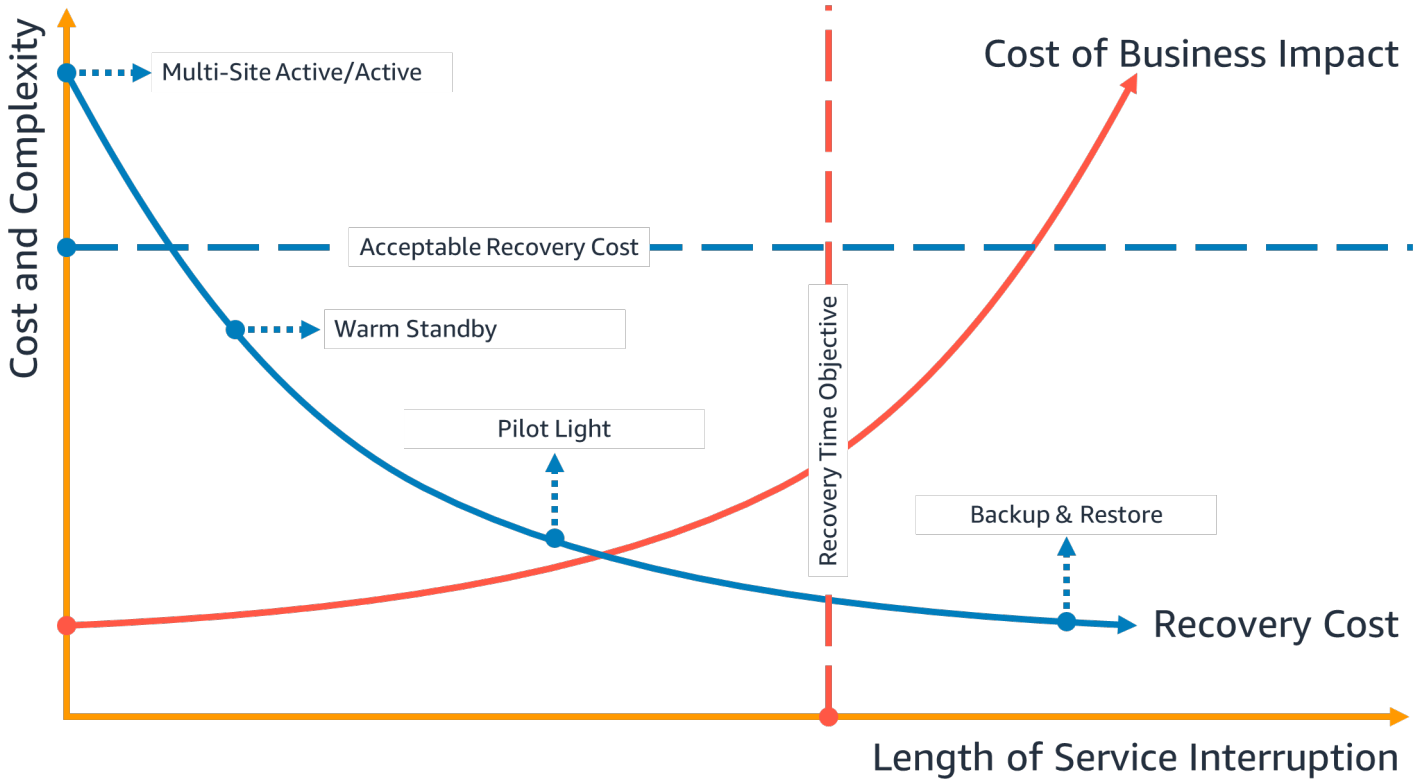

Figura 4: Objetivo de tiempo de recuperación

El objetivo del punto de recuperación (RPO) es el tiempo máximo aceptable desde el último punto de recuperación de datos. Este objetivo determina qué se considera una pérdida de datos aceptable entre el último punto de recuperación y la interrupción del servicio, y lo define la organización.

En el siguiente diagrama, la empresa ha determinado su RPO máximo permitido, así como el límite de lo que puede gastar en su estrategia de recuperación de datos. De las cuatro estrategias de

recuperación ante desastres, tanto la estrategia Pilot Light como la Warm Standby cumplen los dos criterios de RPO y coste.

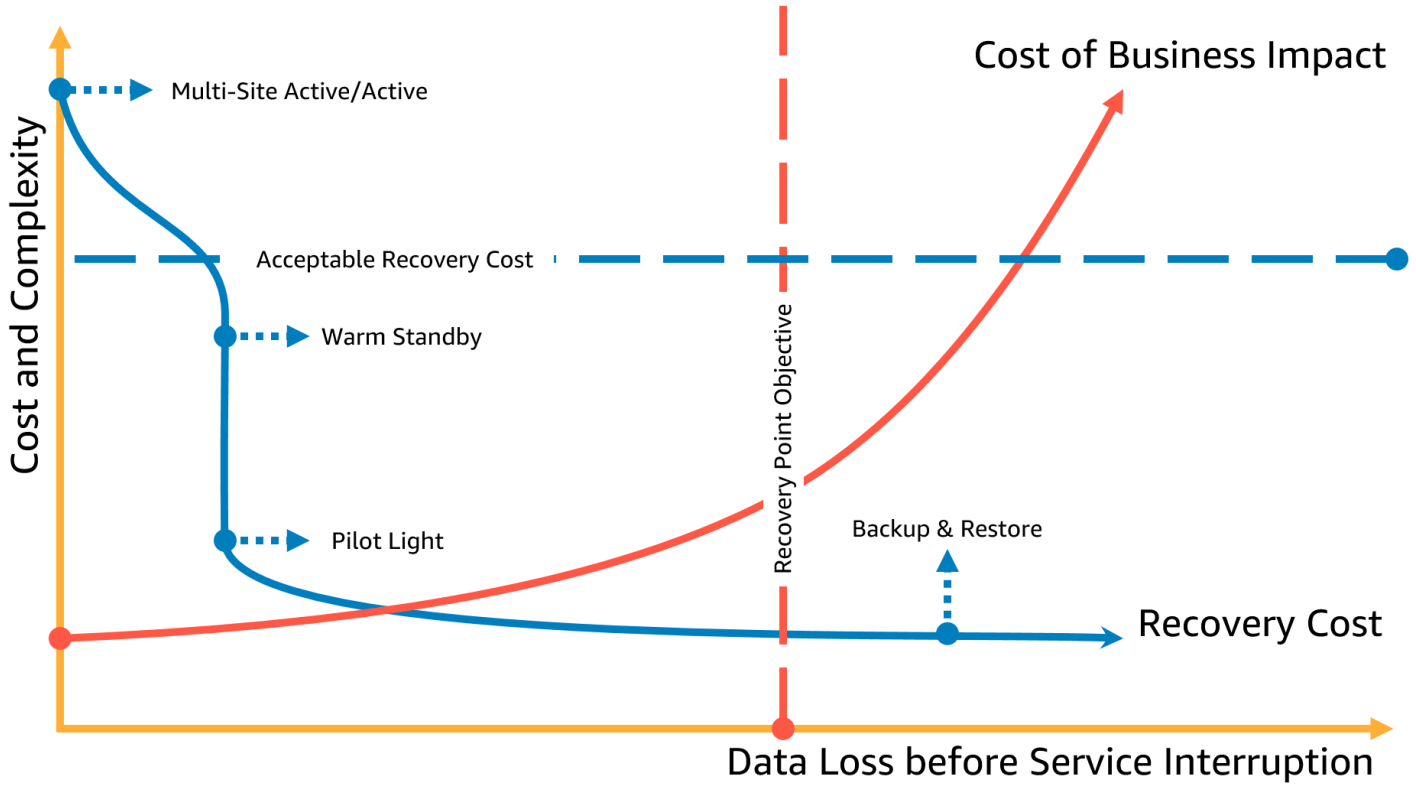

Figura 5: Objetivo del punto de recuperación

> **Note**
>
> Si el coste de la estrategia de recuperación es superior al coste de la avería o la pérdida, no debería adoptarse la opción de recuperación a menos que haya un factor secundario, como los requisitos reglamentarios. Tenga en cuenta las estrategias de recuperación de costes variables al realizar esta evaluación.

La recuperación de desastres es diferente en la nube

Las estrategias de recuperación ante desastres evolucionan con la innovación técnica. Un plan de recuperación ante desastres local puede implicar el transporte físico de las cintas o la replicación de datos a otro sitio. Su organización debe volver a evaluar el impacto empresarial, el riesgo y el costo de sus estrategias anteriores de recuperación ante desastres para cumplir sus objetivos de recuperación ante desastres en AWS. La recuperación ante desastres en la nube de AWS incluye las siguientes ventajas en comparación con los entornos tradicionales:

- Recupérese rápidamente de un desastre con una complejidad reducida
- Las pruebas sencillas y repetibles le permiten realizar pruebas con mayor facilidad y frecuencia
- Una menor sobrecarga de administración reduce la carga operativa
- Las oportunidades de automatización disminuyen las posibilidades de error y mejoran el tiempo de recuperación

AWS le permite cambiar los gastos de capital fijos de un centro de datos de backup físico por los gastos operativos variables de un entorno en la nube del tamaño adecuado, lo que puede reducir considerablemente los costes.

Para muchas organizaciones, la recuperación ante desastres in situ se basaba en el riesgo de interrumpir una o varias cargas de trabajo en un centro de datos y en la recuperación de los datos replicados o respaldados en un centro de datos secundario. Cuando las organizaciones implementan cargas de trabajo en AWS, pueden implementar una carga de trabajo bien diseñada y confiar en el diseño de la infraestructura de nube global de AWS para ayudar a mitigar el efecto de dichas interrupciones. Consulte el documento técnico AWS Well-Architected Framework: pilar de confiabilidad para obtener más información sobre las prácticas recomendadas de arquitectura para diseñar y operar cargas de trabajo confiables, seguras, eficientes y rentables en la nube. Úselo AWS Well-Architected Toolpara revisar sus cargas de trabajo periódicamente y asegurarse de que siguen las mejores prácticas y las directrices del Well-Architected Framework. La herramienta está disponible de forma gratuita en. Consola de administración de AWS

Si sus cargas de trabajo están en AWS, no tiene que preocuparse por la conectividad del centro de datos (con la excepción de la posibilidad de acceder a él), la alimentación, el aire acondicionado, la extinción de incendios y el hardware. Todo esto lo gestiona usted mismo y tiene acceso a varias zonas de disponibilidad aisladas en caso de averías (cada una compuesta por uno o más centros de datos independientes).

Región única de AWS

En el caso de un desastre provocado por la interrupción o la pérdida de un centro de datos físico, la implementación de una carga de trabajo de alta disponibilidad en varias zonas de disponibilidad dentro de una sola región de AWS ayuda a mitigar los desastres naturales y técnicos. El respaldo continuo de los datos dentro de esta única región puede reducir el riesgo de amenazas humanas, como un error o una actividad no autorizada que podría provocar la pérdida de datos. Cada región de AWS se compone de varias zonas de disponibilidad, cada una aislada de los errores de las demás zonas. Cada zona de disponibilidad, a su vez, consta de uno o más centros de datos físicos discretos. Para aislar mejor los problemas más importantes y lograr una alta disponibilidad, puede dividir las cargas de trabajo en varias zonas de la misma región. Las zonas de disponibilidad están diseñadas para ofrecer redundancia física y ofrecer resiliencia, lo que permite un rendimiento ininterrumpido, incluso en caso de cortes de energía, tiempo de inactividad de Internet, inundaciones y otros desastres naturales. Consulte AWS Global Cloud Infrastructure para descubrir cómo lo hace AWS.

Al realizar la implementación en varias zonas de disponibilidad en una sola región de AWS, su carga de trabajo está mejor protegida contra los fallos de un solo centro de datos (o incluso de varios). Para mayor seguridad con su implementación en una sola región, puede hacer copias de seguridad de los datos y la configuración (incluida la definición de la infraestructura) en otra región. Esta estrategia reduce el alcance de su plan de recuperación ante desastres para incluir únicamente la copia de seguridad y la restauración de datos. Aprovechar la resiliencia multirregional realizando copias de seguridad en otra región de AWS es sencillo y económico en comparación con las demás opciones multirregionales que se describen en la siguiente sección. Por ejemplo, hacer copias de seguridad en Amazon Simple Storage Service (Amazon S3) le da acceso a la recuperación inmediata de sus datos. Sin embargo, si su estrategia de DR para partes de sus datos tiene requisitos más relajados en cuanto a los tiempos de recuperación (de minutos a horas), el uso de Amazon Glacier o Amazon Glacier Deep Archive reducirá significativamente los costos de su estrategia de respaldo y recuperación.

Algunas cargas de trabajo pueden tener requisitos reglamentarios de residencia de datos. Si esto se aplica a su carga de trabajo en una localidad que actualmente solo tiene una región de AWS, además de diseñar cargas de trabajo Multi-AZ para una alta disponibilidad, como se ha mencionado anteriormente, también puede utilizar las AZs de esa región como ubicaciones discretas, lo que puede resultar útil para cumplir con los requisitos de residencia de datos aplicables a su carga de trabajo dentro de esa región. Las estrategias de recuperación ante desastres que se describen en

las siguientes secciones utilizan varias regiones de AWS, pero también se pueden implementar mediante zonas de disponibilidad en lugar de regiones.

Varias regiones de AWS

En el caso de un desastre que implique el riesgo de perder varios centros de datos a una distancia significativa entre sí, debería considerar opciones de recuperación ante desastres para mitigar los desastres naturales y técnicos que afectan a toda una región dentro de AWS. Todas las opciones descritas en las siguientes secciones se pueden implementar como arquitecturas multirregionales para protegerse contra este tipo de desastres.

Opciones de recuperación de desastres en la nube

Las estrategias de recuperación ante desastres que tiene a su disposición en AWS se pueden clasificar a grandes rasgos en cuatro enfoques, que van desde el bajo costo y la baja complejidad de realizar copias de seguridad hasta estrategias más complejas que utilizan varias regiones activas. Active/passive las estrategias utilizan un sitio activo (como una región de AWS) para alojar la carga de trabajo y atender el tráfico. El sitio pasivo (por ejemplo, una región de AWS diferente) se utiliza para la recuperación. El sitio pasivo no atiende tráfico de forma activa hasta que se desencadena un evento de conmutación por error.

Es fundamental evaluar y probar periódicamente su estrategia de recuperación ante desastres para tener confianza a la hora de utilizarla en caso de que sea necesario. Utilice AWS Resilience Hub para validar y realizar un seguimiento continuo de la resiliencia de sus AWS cargas de trabajo, incluida la probabilidad de que cumpla sus objetivos de RTO y RPO.

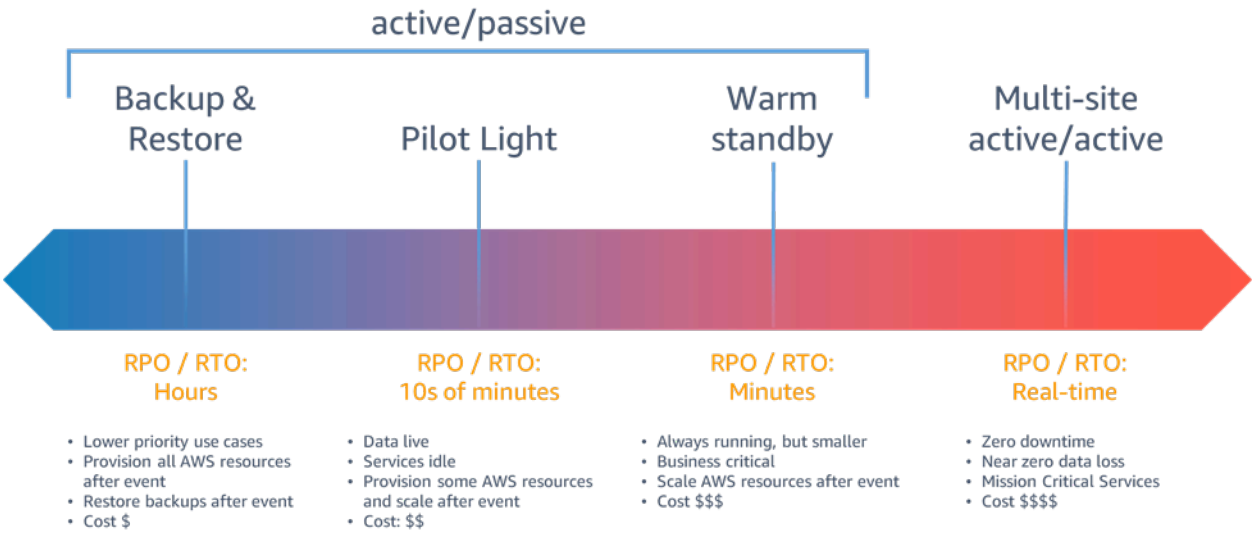

Figura 6: Estrategias de recuperación ante desastres

En el caso de un desastre provocado por la interrupción o la pérdida de un centro de datos físico debido a una carga de trabajo bien diseñada y de alta disponibilidad, es posible que solo necesite un enfoque de respaldo y restauración para la recuperación ante desastres. Si su definición de desastre va más allá de la interrupción o pérdida de un centro de datos físico y se aplica a la de una región o si está sujeto a los requisitos reglamentarios que así lo exigen, entonces debería considerar la opción Pilot Light, Warm Standby o Multi-Site Active/Active.

Al elegir su estrategia y los recursos de AWS para implementarla, tenga en cuenta que, en AWS, solemos dividir los servicios en el plano de datos y el plano de control. El plano de datos se encarga de entregar el servicio en tiempo real mientras que el plano de control se utiliza para configurar el entorno. Para obtener la máxima resiliencia, debe utilizar únicamente las operaciones del plano de datos como parte de la operación de conmutación por error. Esto se debe a que los planos de datos suelen tener objetivos de diseño de disponibilidad más altos que los planos de control.

Copia de seguridad y restauración

El backup y la restauración son un enfoque adecuado para mitigar la pérdida o la corrupción de datos. Este enfoque también se puede utilizar para mitigar un desastre regional mediante la replicación de datos en otras regiones de AWS, o para mitigar la falta de redundancia de las cargas de trabajo implementadas en una única zona de disponibilidad. Además de los datos, debe volver a implementar la infraestructura, la configuración y el código de la aplicación en la región de recuperación. Para permitir que la infraestructura se vuelva a implementar rápidamente y sin errores, siempre debe implementarla utilizando la infraestructura como código (IaC) utilizando servicios como o el. AWS CloudFormationAWS Cloud Development Kit (AWS CDK) Sin el IaC, puede resultar complejo restaurar las cargas de trabajo en la región de recuperación, lo que aumentará los tiempos de recuperación y, posiblemente, superará el RTO. Además de los datos del usuario, asegúrate de hacer copias de seguridad del código y la configuración, incluidas las Amazon Machine Images (AMIs) que utilizas para crear EC2 instancias de Amazon. Se puede utilizar AWS CodePipelinepara automatizar la redistribución del código y la configuración de la aplicación.

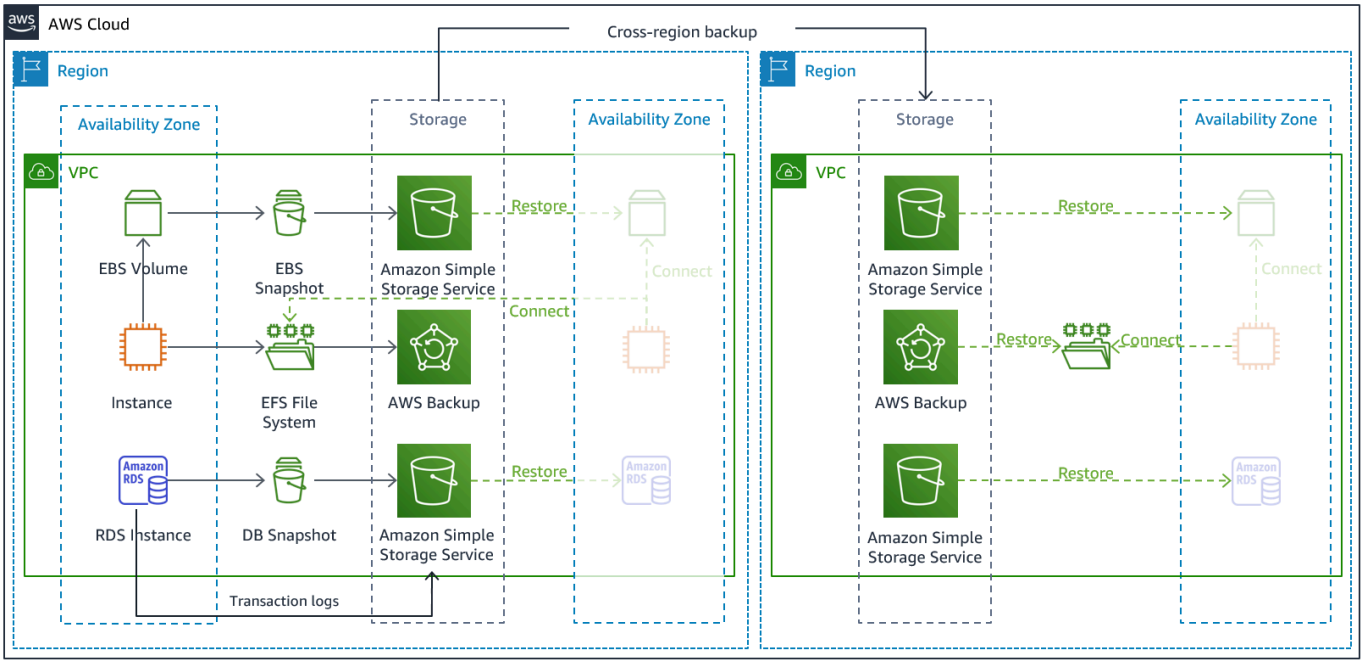

Figura 7: Arquitectura de backup y restauración

Servicios de AWS

Los datos de su carga de trabajo requerirán una estrategia de respaldo que se ejecute periódicamente o sea continua. La frecuencia con la que ejecute el backup determinará el punto de recuperación alcanzable (que debe ajustarse a su RPO). La copia de seguridad también debe ofrecer una forma de restaurarla hasta el momento en que se realizó. El backup con point-in-time recuperación está disponible a través de los siguientes servicios y recursos:

- Instantánea de Amazon Elastic Block Store (Amazon EBS)
- Respaldo de Amazon DynamoDB
- Instantánea de Amazon RDS
- Instantánea de base de datos de Amazon Aurora
- Respaldo de Amazon EFS (cuando se usa AWS Backup)
- Instantánea de Amazon Redshift
- Instantánea de Amazon Neptune
- Amazon DocumentDB
- Amazon FSx para Windows File Server, Amazon FSx para Lustre, Amazon FSx para NetApp ONTAP y Amazon FSx para OpenZFS

Para Amazon Simple Storage Service (Amazon S3), puede utilizar Amazon S3 Cross-Region Replication (CRR) para copiar objetos de forma asíncrona a un bucket de S3 en la región DR de forma continua y, al mismo tiempo, proporcionar el control de versiones de los objetos almacenados para que pueda elegir el punto de restauración. La replicación continua de los datos tiene la ventaja de ser el tiempo más corto (casi nulo) para realizar copias de seguridad de los datos, pero es posible que no proteja contra desastres, como la corrupción de los datos o los ataques malintencionados (como la eliminación no autorizada de datos), ni tampoco contra las point-in-time copias de seguridad. La replicación continua se describe en la sección Servicios de AWS para Pilot Light.

AWS Backupproporciona una ubicación centralizada para configurar, programar y monitorear las capacidades de respaldo de AWS para los siguientes servicios y recursos:

- Volúmenes de Amazon Elastic Block Store (Amazon EBS)
- EC2Instancias de Amazon
- Bases de datos de Amazon Relational Database Service (Amazon RDS) (incluidas las bases de datos de Amazon Aurora)
- Tablas de Amazon DynamoDB
- Sistemas de archivos Amazon Elastic File System (Amazon EFS)
- AWS Storage GatewayVolúmenes de
- Amazon FSx para Windows File Server, Amazon FSx para Lustre, Amazon FSx para NetApp ONTAP y Amazon FSx para OpenZFS

AWS Backup permite copiar copias de seguridad entre regiones, por ejemplo, en una región de recuperación ante desastres.

Como estrategia adicional de recuperación ante desastres para sus datos de Amazon S3, habilite el control de versiones de objetos de S3. El control de versiones de objetos protege sus datos en S3 de las consecuencias de las acciones de eliminación o modificación, ya que conserva la versión original antes de la acción. El control de versiones de objetos puede ser una forma útil de mitigar los desastres provocados por errores humanos. Si utiliza la replicación de S3 para realizar copias de seguridad de los datos en su región de DR, de forma predeterminada, cuando se elimina un objeto del bucket de origen, Amazon S3 añade un marcador de eliminación únicamente en el bucket de origen. Este enfoque protege los datos de la región de DR de las eliminaciones malintencionadas en la región de origen.

Además de los datos, también debe realizar una copia de seguridad de la configuración y la infraestructura necesarias para volver a implementar su carga de trabajo y cumplir su objetivo de tiempo de recuperación (RTO). AWS CloudFormationproporciona infraestructura como código (IaC) y le permite definir todos los recursos de AWS de su carga de trabajo para que pueda implementarlos y volver a implementarlos de manera confiable en varias cuentas de AWS y regiones de AWS. Puede hacer copias de seguridad de EC2 las instancias de Amazon utilizadas por su carga de trabajo como Amazon Machine Images (AMIs). La AMI se crea a partir de instantáneas del volumen raíz de la instancia y de cualquier otro volumen de EBS adjunto a la instancia. Puede usar esta AMI para lanzar una versión restaurada de la EC2 instancia. Una AMI se puede copiar dentro de una región o de una región a otra. O bien, puede utilizar AWS Backuppara copiar copias de seguridad entre cuentas y otras regiones de AWS. La función de copia de seguridad multicuenta le ayuda a protegerse de posibles desastres, como las amenazas internas o la puesta en peligro de la cuenta. AWS Backup también agrega capacidades adicionales para la EC2 copia de seguridad: además de los volúmenes EBS individuales de la instancia, AWS Backup también almacena y rastrea los siguientes metadatos: tipo de instancia, nube privada virtual (VPC) configurada, grupo de seguridad, rol de IAM, configuración de monitoreo y etiquetas. Sin embargo, estos metadatos adicionales solo se utilizan al restaurar la EC2 copia de seguridad en la misma región de AWS.

Todos los datos almacenados en la región de recuperación ante desastres como copias de seguridad deben restaurarse en el momento de la conmutación por error. AWS Backup ofrece la capacidad de restauración, pero actualmente no permite la restauración programada o automática. Puede implementar la restauración automática en la región de DR utilizando el SDK de AWS, si así lo APIs solicita AWS Backup. Puede configurarlo como un trabajo periódico o activar la restauración cada vez que se complete una copia de seguridad. En la siguiente figura se muestra un ejemplo de restauración automática mediante Amazon Simple Notification Service (Amazon SNS) y. AWS Lambda Implementar una restauración de datos periódica y programada es una buena idea, ya que la restauración de datos a partir de una copia de seguridad es una operación del plano de control. Si esta operación no estuviera disponible durante un desastre, seguiría teniendo almacenes de datos operativos creados a partir de una copia de seguridad reciente.

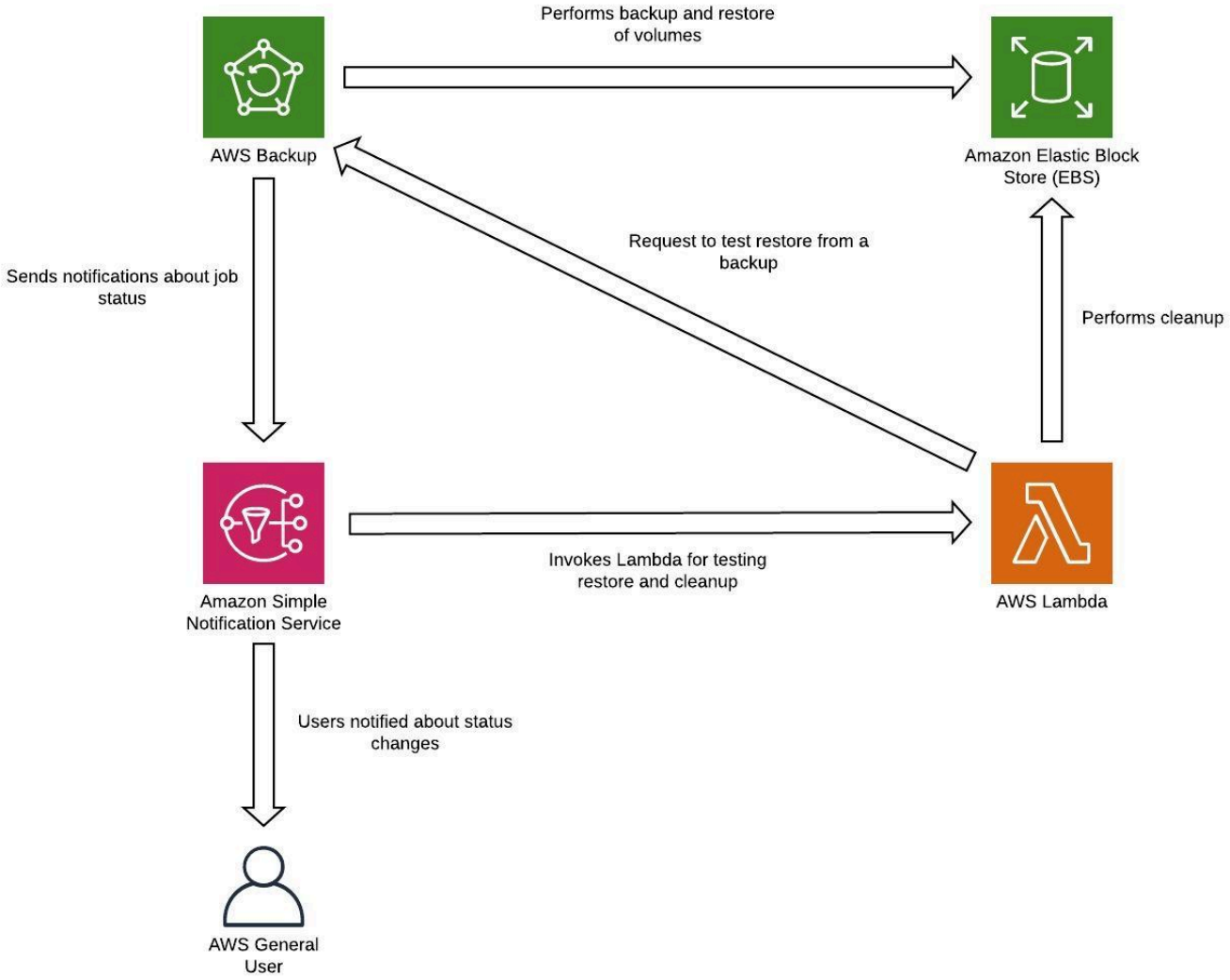

Figura 8: Restauración y prueba de copias de seguridad

> **Note**
>
> La estrategia de copia de seguridad debe incluir la prueba de sus copias de seguridad. Consulte la sección Probar la recuperación ante desastres para obtener más información. Consulte el AWS Well-Architected Lab: Testing Backup and Restore of Data para ver una demostración práctica de la implementación.

Luz piloto

Con el enfoque basado en la modalidad piloto, puede replicar sus datos de una región a otra y aprovisionar una copia de su infraestructura de carga de trabajo principal. Los recursos necesarios para permitir la replicación y copia de seguridad de los datos, como el almacenamiento de bases de datos y objetos, están siempre disponibles. Otros elementos, como los servidores de aplicaciones, se cargan con el código y las configuraciones de la aplicación, pero están «apagados» y solo se utilizan durante las pruebas o cuando se invoca la conmutación por error de recuperación ante desastres. En la nube, tiene la flexibilidad de desaprovisionar recursos cuando no los necesite y aprovisionarlos cuando los necesite. Una buena práctica para «desconectar» es no desplegar el recurso y, a continuación, crear la configuración y las capacidades necesarias para desplegarlo («encenderlo») cuando sea necesario. A diferencia del enfoque de backup y restauración, su infraestructura principal está siempre disponible y siempre tiene la opción de aprovisionar rápidamente un entorno de producción a gran escala activando y ampliando sus servidores de aplicaciones.

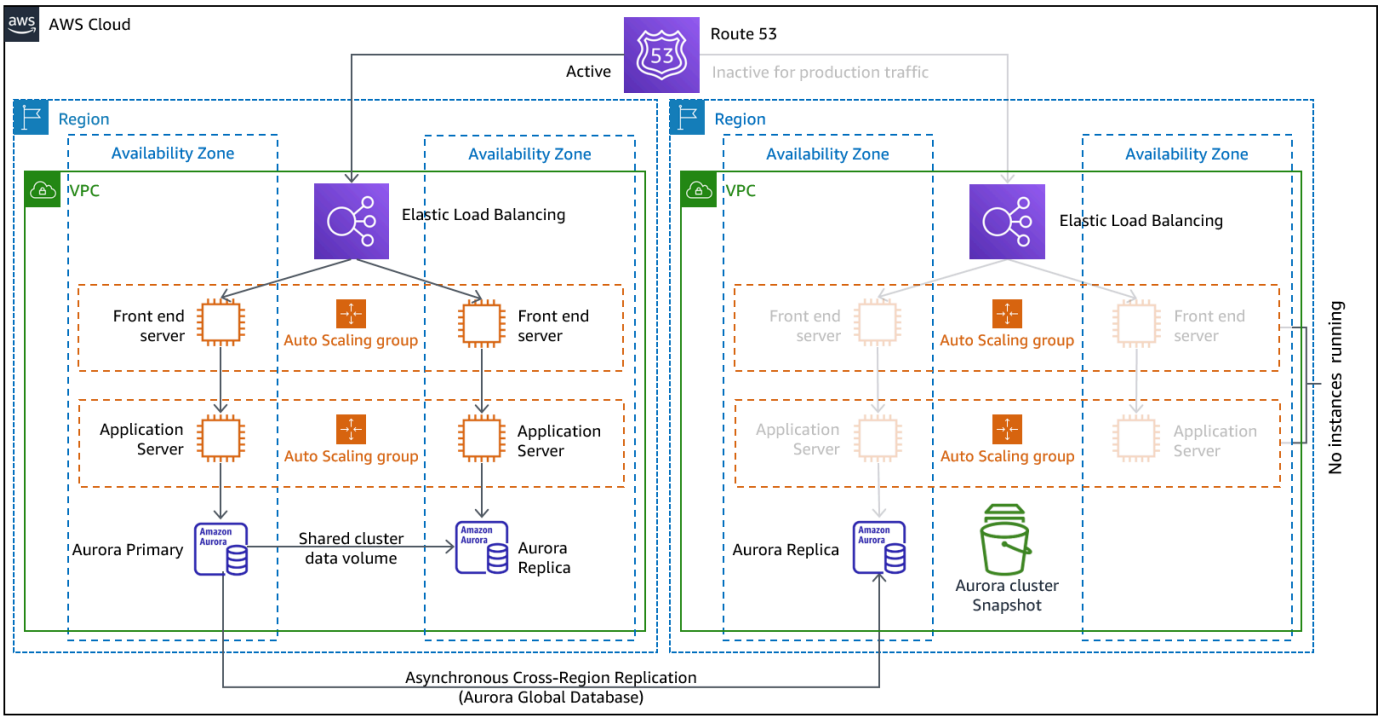

Figura 9: Arquitectura de luz piloto

Un enfoque piloto ligero minimiza el costo continuo de la recuperación ante desastres al minimizar los recursos activos y simplifica la recuperación en el momento de un desastre, ya que todos los requisitos de infraestructura básicos están establecidos. Esta opción de recuperación requiere que cambie su enfoque de implementación. Debe realizar cambios en la infraestructura principal de

cada región e implementar los cambios en la carga de trabajo (configuración, código) de forma simultánea en cada región. Este paso se puede simplificar automatizando las implementaciones y utilizando la infraestructura como código (IaC) para implementar la infraestructura en varias cuentas y regiones (implementación completa de la infraestructura en la región principal e implementación de infraestructura reducida o desconectada en las regiones de DR). Se recomienda utilizar una cuenta diferente por región para proporcionar el máximo nivel de aislamiento de recursos y seguridad (en el caso de que las credenciales comprometidas también formen parte de sus planes de recuperación ante desastres).

Con este enfoque, también debe evitar un desastre de datos. La replicación continua de los datos lo protege contra algunos tipos de desastres, pero es posible que no lo proteja contra la corrupción o la destrucción de los datos, a menos que su estrategia también incluya el control de versiones de los datos almacenados o las opciones de point-in-time recuperación. Puede hacer una copia de seguridad de los datos replicados en la región del desastre para crear point-in-time copias de seguridad en esa misma región.

Servicios de AWS

Además de utilizar los servicios de AWS descritos en la sección Backup and Restore para crear point-in-time copias de seguridad, considere también los siguientes servicios para su estrategia piloto.

A modo de prueba, la replicación continua de datos en bases de datos y almacenes de datos activos de la región de DR es el mejor enfoque para un RPO bajo (si se utiliza además de las point-in-time copias de seguridad descritas anteriormente). AWS proporciona una replicación de datos asincrónica, continua y entre regiones mediante los siguientes servicios y recursos:

- Replicación del Amazon Simple Storage Service (Amazon S3)
- Réplicas de lectura de Amazon RDS
- Bases de datos globales de Amazon Aurora
- Tablas globales de Amazon DynamoDB
- Clústeres globales de Amazon DocumentDB
- Almacén de datos global para Amazon ElastiCache (Redis OSS)

Con la replicación continua, las versiones de sus datos están disponibles casi de inmediato en su región de DR. Los tiempos de replicación reales se pueden monitorear mediante funciones

de servicio como S3 Replication Time Control (S3 RTC) para objetos de S3 y funciones de administración de las bases de datos globales de Amazon Aurora.

Si no puede ejecutar su read/write carga de trabajo desde la región de recuperación ante desastres, debe promover una réplica de lectura de RDS para que se convierta en la instancia principal. En el caso de las instancias de base de datos distintas de Aurora, el proceso tarda unos minutos en completarse y el reinicio forma parte del proceso. Para la replicación entre regiones (CRR) y la conmutación por error con RDS, el uso de la base de datos global Amazon Aurora ofrece varias ventajas. La base de datos global utiliza una infraestructura dedicada que deja sus bases de datos completamente disponibles para servir a su aplicación y puede replicarse en la región secundaria con una latencia típica de menos de un segundo (y dentro de una región de AWS es mucho menor a 100 milisegundos). Con la base de datos global Amazon Aurora, si su región principal sufre una disminución del rendimiento o una interrupción, puede promover una de las regiones secundarias para que asuma responsabilidades de lectura/escritura en menos de un minuto, incluso en el caso de que se produzca una interrupción regional total. También puede configurar Aurora para que supervise el tiempo de retraso del RPO de todos los clústeres secundarios y asegurarse de que al menos un clúster secundario permanezca dentro de la ventana de RPO de destino.

Debe implementar una versión reducida de su infraestructura de carga de trabajo principal con menos o menos recursos en su región de DR. Con él AWS CloudFormation, puede definir su infraestructura e implementarla de manera uniforme en todas las cuentas y regiones de AWS. AWS CloudFormation utiliza pseudoparámetros predefinidos para identificar la cuenta de AWS y la región de AWS en las que se implementa. Por lo tanto, puede implementar la lógica de condiciones en sus CloudFormation plantillas para implementar solo la versión reducida de su infraestructura en la región de DR. Por EC2 ejemplo, en las implementaciones, una Amazon Machine Image (AMI) proporciona información como la configuración del hardware y el software instalado. Puede implementar una canalización de Image Builder que cree las que AMIs necesite y copiarlas tanto en la región principal como en la de respaldo. Esto ayuda a garantizar que estas regiones AMIscuenten con todo lo que necesita para volver a implementar o ampliar su carga de trabajo en una nueva región, en caso de que se produzca un desastre. Las EC2 instancias de Amazon se implementan en una configuración reducida (menos instancias que en su región principal). Para ampliar la infraestructura para soportar el tráfico de producción, consulte Amazon EC2 Auto Scaling en la sección Warm Standby.

En el caso de una active/passive configuración como la de un semáforo piloto, todo el tráfico se dirige inicialmente a la región principal y pasa a la región de recuperación ante desastres si la región principal ya no está disponible. Esta operación de conmutación por error se puede iniciar automática o manualmente. La conmutación por error que se inicia automáticamente y que se basa en controles de estado o alarmas debe utilizarse con precaución. Incluso si se utilizan las prácticas

recomendadas aquí, el tiempo y el punto de recuperación serán superiores a cero, lo que provocará cierta pérdida de disponibilidad y datos. Si realiza una conmutación por error cuando no es necesario (falsa alarma), incurre en esas pérdidas. Por tanto, la conmutación por error iniciada manualmente es la que se suele utilizar. En este caso, debe seguir automatizando los pasos de la conmutación por error, de modo que la iniciación manual sea como pulsar un botón.

Hay varias opciones de administración del tráfico que se deben tener en cuenta al utilizar AWS los servicios.

Una opción es utilizar Amazon Route 53. Con Amazon Route 53, puede asociar varios puntos de enlace IP en una o más regiones de AWS a un nombre de dominio de Route 53. A continuación, puede dirigir el tráfico al punto final correspondiente con ese nombre de dominio. En caso de conmutación por error, debe dirigir el tráfico al punto final de recuperación y alejarlo del punto final principal. Los controles de estado de Amazon Route 53 supervisan estos puntos de conexión. Con estas comprobaciones de estado, puede configurar la conmutación por error de DNS que se inicie automáticamente para garantizar que el tráfico se envíe únicamente a los puntos finales en buen estado, lo que constituye una operación muy fiable que se realiza en el plano de los datos. Para implementarlo mediante una conmutación por error iniciada manualmente, puede utilizar Amazon Application Recovery Controller (ARC). Con ARC, puede crear comprobaciones de estado de Route 53 que, en realidad, no comprueban el estado, sino que actúan como interruptores de encendido/apagado sobre los que usted tiene pleno control. Con la AWS CLI o el AWS SDK, puede programar la conmutación por error mediante esta API de plano de datos de alta disponibilidad. Su script activa estos conmutadores (las comprobaciones de estado de Route 53) y le indica a Route 53 que envíe el tráfico a la región de recuperación en lugar de a la región principal. Otra opción de conmutación por error iniciada manualmente que algunos han utilizado consiste en utilizar una política de enrutamiento ponderado y cambiar los pesos de las regiones principal y de recuperación para que todo el tráfico vaya a la región de recuperación. Sin embargo, tenga en cuenta que se trata de una operación del plano de control y, por lo tanto, no es tan resistente como el enfoque del plano de datos que utiliza Amazon Application Recovery Controller (ARC).

Otra opción es utilizar AWS Global Accelerator. Con la AnyCast IP, puede asociar varios puntos de enlace en una o más regiones de AWS con la misma dirección o direcciones IP públicas estáticas. AWS Global Accelerator a continuación, enruta el tráfico al punto final correspondiente asociado a esa dirección. Los controles de estado de Global Accelerator supervisan los puntos finales. Con estas comprobaciones de estado, AWS Global Accelerator comprueba el estado de las aplicaciones y redirige el tráfico de usuarios automáticamente al punto final de la aplicación en buen estado. En el caso de la conmutación por error iniciada manualmente, puede ajustar el punto final que recibe el tráfico mediante los diales de tráfico, pero tenga en cuenta que se trata de una operación del plano

de control. Global Accelerator ofrece latencias más bajas en el punto final de la aplicación, ya que utiliza la amplia red perimetral de AWS para colocar el tráfico en la red troncal de AWS lo antes posible. Global Accelerator también evita los problemas de almacenamiento en caché que pueden producirse con los sistemas DNS (como Route 53).

Amazon CloudFront ofrece la conmutación por error de origen, mediante la cual, si falla una solicitud determinada al punto de enlace principal CloudFront , la envía al punto de enlace secundario. A diferencia de las operaciones de conmutación por error descritas anteriormente, todas las solicitudes posteriores siguen yendo al punto final principal y la conmutación por error se realiza por cada solicitud.

AWS Elastic Disaster Recovery

AWS Elastic Disaster Recovery (DRS) replica continuamente las aplicaciones alojadas en el servidor y las bases de datos alojadas en el servidor desde cualquier fuente para AWS utilizar la replicación a nivel de bloques del servidor subyacente. Elastic Disaster Recovery te permite usar una región Nube de AWS como destino de recuperación ante desastres para una carga de trabajo alojada localmente o en otro proveedor de nube, así como para su entorno. También se puede usar para la recuperación ante desastres de cargas de trabajo AWS alojadas si están compuestas únicamente por aplicaciones y bases de datos alojadas en ellas EC2 (es decir, no en RDS). Elastic Disaster Recovery utiliza la estrategia Pilot Light, que mantiene una copia de los datos y los recursos «desconectados» en una Amazon Virtual Private Cloud (Amazon VPC) que se utiliza como área de preparación. Cuando se desencadena un evento de conmutación por error, los recursos preparados se utilizan para crear automáticamente un despliegue a plena capacidad en la Amazon VPC de destino utilizada como ubicación de recuperación.

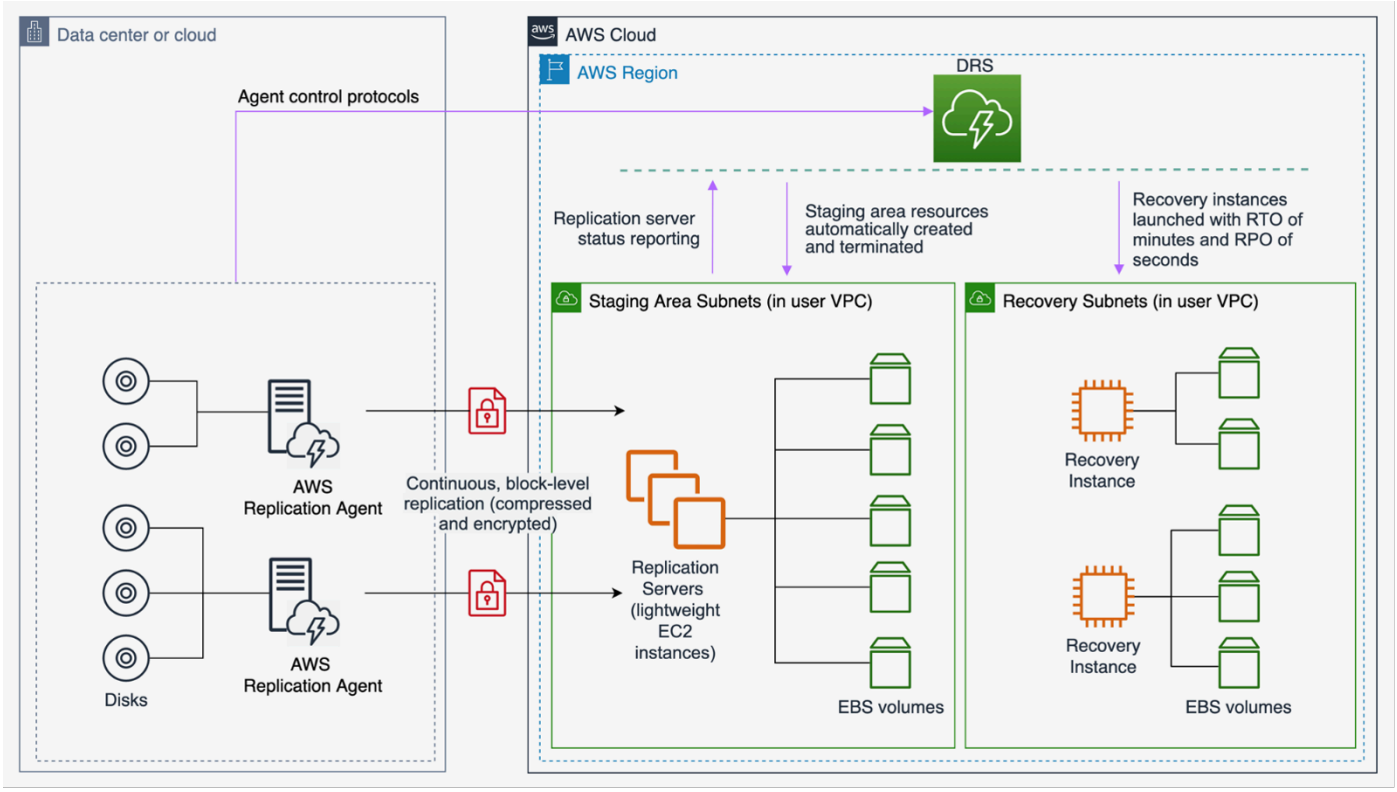

Figura 10: Arquitectura de AWS Elastic Disaster Recovery

Espera semiactiva

El enfoque de espera semiactiva implica garantizar que haya una copia reducida, pero completamente funcional, del entorno de producción en otra región. Este enfoque extiende el concepto de luz piloto y reduce el tiempo de recuperación, ya que su carga de trabajo tiene disponibilidad permanente en otra región. Este enfoque también le permite realizar pruebas con mayor facilidad o implementar pruebas continuas para aumentar la confianza en su capacidad de recuperación tras un desastre.

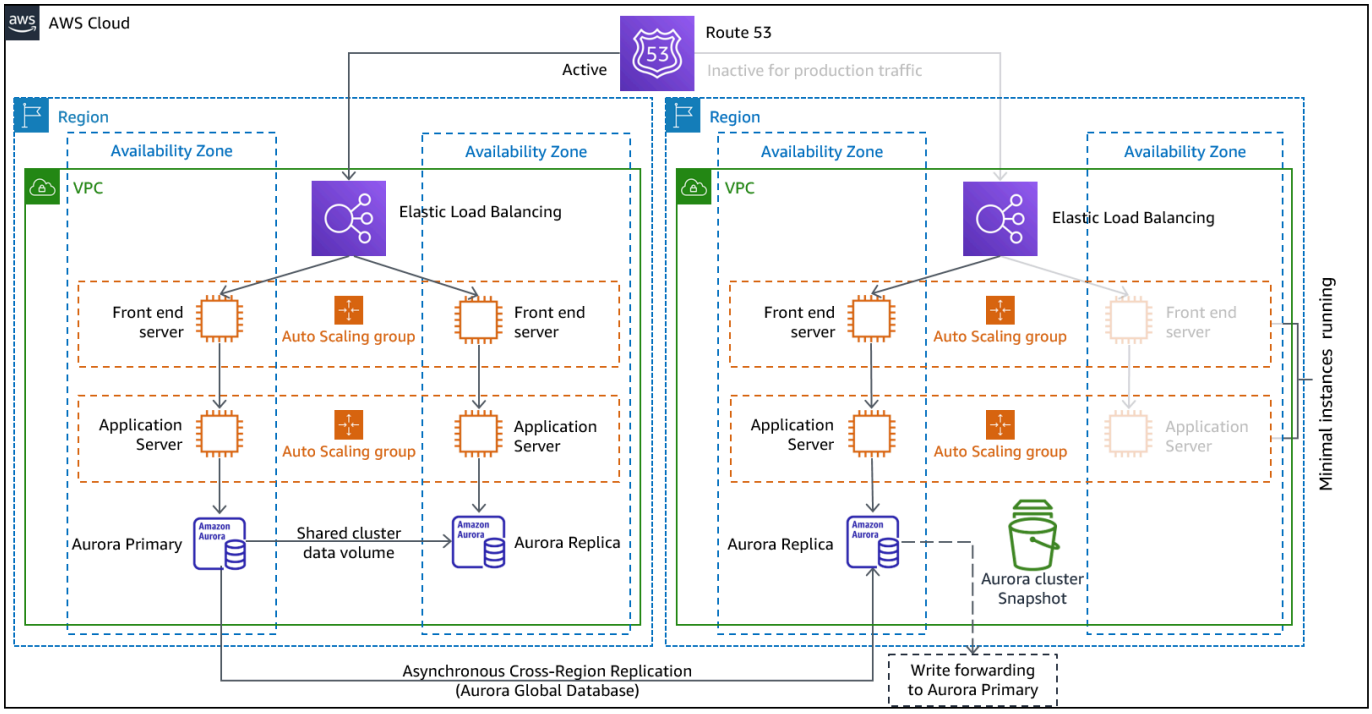

Figura 11: Arquitectura de modo de espera en caliente

Nota: La diferencia entre la luz piloto y la estación de espera cálida a veces puede resultar difícil de entender. Ambos incluyen un entorno en su región de RD con copias de los activos de su región principal. La diferencia es que Pilot Light no puede procesar las solicitudes sin antes tomar medidas adicionales, mientras que en modo de espera en caliente se puede gestionar el tráfico (con niveles de capacidad reducidos) de forma inmediata. El enfoque piloto requiere «encender» los servidores, posiblemente implementar una infraestructura adicional (no esencial) y ampliarlos, mientras que en modo de espera temporal solo es necesario ampliarlos (todo ya está desplegado y funcionando). Utilice sus necesidades de RTO y RPO para ayudarle a elegir entre estos enfoques.

Servicios de AWS

Todos los servicios de AWS incluidos en las secciones de copia de seguridad y restauración y Pilot Light también se utilizan en modo de espera caliente para la copia de seguridad de datos, la replicación de datos, el enrutamiento del active/passive tráfico y el despliegue de la infraestructura, incluidas EC2 las instancias.

Amazon EC2 Auto Scaling se utiliza para escalar recursos, incluidas las EC2 instancias de Amazon, las tareas de Amazon ECS, el rendimiento de Amazon DynamoDB y las réplicas de Amazon Aurora dentro de una región de AWS. Amazon EC2 Auto Scaling escala la implementación de la

EC2 instancia en todas las zonas de disponibilidad de una región de AWS, lo que proporciona resiliencia dentro de esa región. Utilice Auto Scaling para ampliar su región de DR hasta alcanzar la capacidad de producción total, como parte de una estrategia piloto de espera en condiciones de luz o en condiciones cálidas. Por ejemplo EC2, para aumentar la configuración de capacidad deseada en el grupo Auto Scaling. Puede ajustar esta configuración manualmente a través de Consola de administración de AWS, automáticamente a través del SDK de AWS o volviendo a implementar la AWS CloudFormation plantilla con el nuevo valor de capacidad deseado. Puede usar AWS CloudFormation parámetros para facilitar la reimplementación de la plantilla. CloudFormation Asegúrese de que las cuotas de servicio en su región de RD sean lo suficientemente altas como para no limitar su capacidad de ampliación a la capacidad de producción.

Como Auto Scaling es una actividad del plano de control, depender de él reducirá la resiliencia de su estrategia de recuperación general. Se trata de una compensación. Puede optar por aprovisionar una capacidad suficiente para que la región de recuperación pueda gestionar toda la carga de producción tal como se haya desplegado. Esta configuración estable desde el punto de vista estático se denomina modo de espera activa (consulte la siguiente sección). O puede optar por aprovisionar menos recursos, lo que le costará menos, pero dependerá de Auto Scaling. Algunas implementaciones de DR desplegarán recursos suficientes para gestionar el tráfico inicial, lo que garantizará un RTO bajo, y luego dependerán de Auto Scaling para aumentar el tráfico posterior.

Activa-activa multisitio

Puede gestionar su carga de trabajo simultáneamente en varias regiones como parte de una estrategia activa/activa o pasiva multisitio o en modo de espera activa o pasiva. Los sitios múltiples active/active atienden el tráfico de todas las regiones en las que están desplegados, mientras que el modo de espera activa solo atiende el tráfico de una sola región y las demás regiones solo se utilizan para la recuperación ante desastres. Con un active/active enfoque multisitio, los usuarios pueden acceder a su carga de trabajo en cualquiera de las regiones en las que esté desplegada. Este enfoque es el más complejo y costoso de la recuperación ante desastres, pero puede reducir el tiempo de recuperación prácticamente a cero en la mayoría de los casos si se utilizan la tecnología y la implementación correctas (sin embargo, la corrupción de los datos puede tener que depender de las copias de seguridad, lo que normalmente se traduce en un punto de recuperación distinto de cero). El modo de espera activo utiliza una active/passive configuración en la que los usuarios solo se dirigen a una sola región y las regiones de DR no reciben tráfico. La mayoría de los clientes consideran que, si quieren instalar un entorno completo en la segunda región, tiene sentido utilizarlo activo/activo. Como alternativa, si no desea utilizar ambas regiones para gestionar el tráfico de

usuarios, Warm Standby ofrece un enfoque más económico y menos complejo desde el punto de vista operativo.

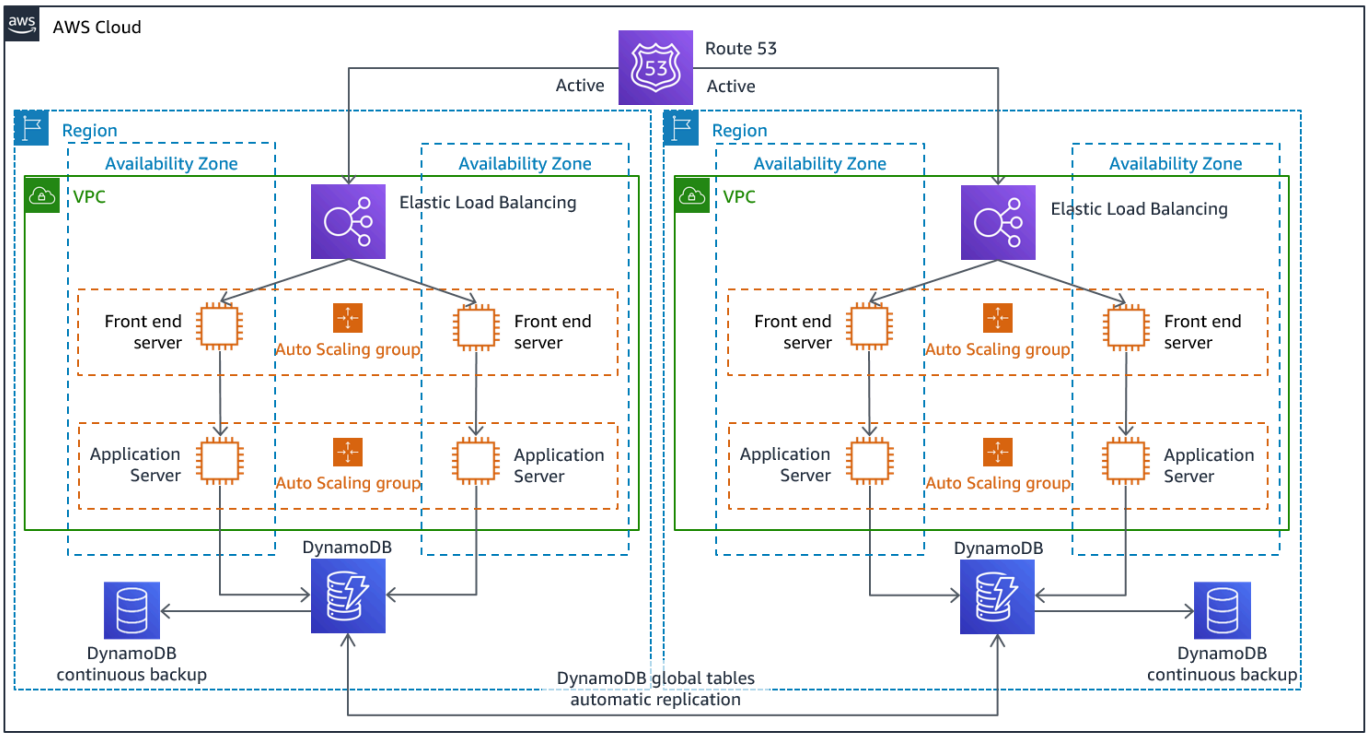

Figura 12: active/active Arquitectura multisitio (cambie una ruta activa a inactiva para el modo de espera activo)

Si se opta por un enfoque multisitio active/active, because the workload is running in more than one Region, there is no such thing as failover in this scenario. Disaster recovery testing in this case would focus on how the workload reacts to loss of a Region: Is traffic routed away from the failed Region? Can the other Region(s) handle all the traffic? Testing for a data disaster is also required. Backup and recovery are still required and should be tested regularly. It should also be noted that recovery times for a data disaster involving data corruption, deletion, or obfuscation will always be greater than zero and the recovery point will always be at some point before the disaster was discovered. If the additional complexity and cost of a multi-site active/active (o en modo de espera activa), es necesario mantener los tiempos de recuperación prácticamente nulos, por lo que se deben realizar esfuerzos adicionales para mantener la seguridad y evitar los errores humanos a fin de mitigar los desastres humanos.

Servicios de AWS

Todos los servicios de AWS incluidos en respaldo y restauración, Pilot Light y Warm standby también se utilizan aquí para el respaldo de point-in-time datos, la replicación de datos, el enrutamiento active/active del tráfico y el despliegue y escalado de la infraestructura, incluidas EC2 las instancias.

Para los active/passive escenarios descritos anteriormente (Pilot Light y Warm Standby), tanto Amazon Route 53 como Amazon AWS Global Accelerator pueden usarse para enrutar el tráfico de la red a la región activa. Para la active/active estrategia en este caso, ambos servicios también permiten definir políticas que determinan qué usuarios van a qué punto final regional activo. Configura un dial de tráfico para controlar el porcentaje de tráfico que se dirige a cada punto final de la aplicación. AWS Global Accelerator Amazon Route 53 admite este enfoque porcentual y también varias otras políticas disponibles, incluidas las basadas en la geoproximidad y la latencia. Global Accelerator aprovecha automáticamente la amplia red de servidores perimetrales de AWS para incorporar el tráfico a la red troncal de AWS lo antes posible, lo que reduce las latencias de las solicitudes.

La replicación asíncrona de datos con esta estrategia permite un RPO prácticamente nulo. Los servicios de AWS, como la base de datos global Amazon Aurora, utilizan una infraestructura dedicada que deja sus bases de datos totalmente disponibles para servir a su aplicación y pueden replicarse en hasta cinco regiones secundarias con una latencia típica de menos de un segundo. active/passive strategies, writes occur only to the primary Region. The difference with active/ activeEstá diseñando cómo se gestiona la coherencia de los datos con las escrituras en cada región activa. Es habitual diseñar las lecturas de los usuarios para que las entreguen desde la región más cercana a ellos, lo que se conoce como lectura local. Con las escrituras, tienes varias opciones:

- Una estrategia global de escritura dirige todas las escrituras a una sola región. En caso de que esa región fracase, se promovería a otra región para que aceptara escrituras. La base de datos global Aurora es una buena opción para la escritura global, ya que admite la sincronización con réplicas de lectura en todas las regiones y puede promover que una de las regiones secundarias read/write asuma responsabilidades en menos de un minuto. Aurora también admite el reenvío de escritura, que permite a los clústeres secundarios de una base de datos global de Aurora reenviar sentencias SQL que realizan operaciones de escritura al clúster principal.
- Una estrategia local de escritura enruta las escrituras a la región más cercana (igual que las lecturas). Las tablas globales de Amazon DynamoDB permiten esta estrategia, ya que permiten leer y escribir en todas las regiones en las que esté desplegada la tabla global. En las tablas globales de Amazon DynamoDB, el último escritor gana la reconciliación entre las actualizaciones simultáneas.

- Una estrategia de escritura particionada asigna las escrituras a una región específica en función de una clave de partición (como el ID de usuario) para evitar conflictos de escritura. La replicación de Amazon S3 configurada de forma bidireccional se puede utilizar en este caso y, actualmente, admite la replicación entre dos regiones. Al implementar este enfoque, asegúrese de habilitar la sincronización de las modificaciones de las réplicas en los depósitos A y B para replicar los cambios en los metadatos de las réplicas, como las listas de control de acceso a los objetos (ACLs), las etiquetas de objetos o los bloqueos de objetos en los objetos replicados. También puede configurar si desea replicar o no los marcadores de eliminación entre los depósitos de sus regiones activas. Además de la replicación, su estrategia también debe incluir point-in-time copias de seguridad para evitar que los datos se corrompan o destruyan.

AWS CloudFormation es una herramienta poderosa para aplicar una infraestructura implementada de manera uniforme entre las cuentas de AWS en varias regiones de AWS. AWS CloudFormation StackSetsamplía esta funcionalidad al permitirle crear, actualizar o eliminar CloudFormation pilas en varias cuentas y regiones con una sola operación. Aunque AWS CloudFormation utiliza YAML o JSON para definir la infraestructura como código, AWS Cloud Development Kit (AWS CDK)permite definir la infraestructura como código utilizando lenguajes de programación conocidos. El código se convierte y CloudFormation , a continuación, se utiliza para implementar recursos en AWS.

Detección

Es importante saber lo antes posible que sus cargas de trabajo no están produciendo los resultados empresariales que deberían ofrecer. De esta forma, puede declarar rápidamente un desastre y recuperarse de un incidente. Para los objetivos de recuperación exigentes, este tiempo de respuesta, junto con la información adecuada, es fundamental para cumplir los objetivos de recuperación. Si su objetivo de tiempo de recuperación es de una hora, debe detectar el incidente, notificar al personal correspondiente, iniciar los procesos de escalamiento, evaluar la información (si la tiene) sobre el tiempo previsto de recuperación (sin ejecutar el plan de recuperación ante desastres), declarar un desastre y recuperarse en una hora.

> **Note**
>
> Si las partes interesadas deciden no recurrir a la DR a pesar de que la RTO estaría en peligro, reevalúe los planes y objetivos de la DR. La decisión de no invocar los planes de DR puede deberse a que los planes son inadecuados o a una falta de confianza en la ejecución.

Es fundamental tener en cuenta la detección, la notificación, la escalación, el descubrimiento y la declaración de incidentes en la planificación y los objetivos para ofrecer objetivos realistas y alcanzables que aporten valor empresarial.

AWS publica la mayor parte de la up-to-the-minute información sobre la disponibilidad de los servicios en el Service Health Dashboard. Compruebe en cualquier momento si desea obtener información sobre el estado actual o suscríbase a una fuente RSS para recibir notificaciones sobre las interrupciones de cada servicio individual. Si tiene un problema operativo en tiempo real con uno de nuestros servicios que no aparece en el Service Health Dashboard, puede crear una solicitud de soporte.

Panel de AWS HealthProporciona información sobre AWS Health los eventos que pueden afectar a su cuenta. La información se presenta de dos formas: en un panel donde se muestran los eventos recientes y próximos organizados por categorías, y en un log de eventos que contiene todos los eventos de los últimos 90 días.

Para cumplir con los requisitos de RTO más estrictos, puede implementar una conmutación por error automatizada basada en controles de estado. Diseñe controles de estado que sean representativos de la experiencia del usuario y se basen en los indicadores clave de rendimiento. Los controles

de estado exhaustivos ejercitan las funciones clave de su carga de trabajo y van más allá de los controles superficiales de los latidos del corazón. Utilice controles de estado exhaustivos basados en múltiples señales. Tenga cuidado con este enfoque para no activar falsas alarmas, ya que una conmutación por error cuando no es necesaria puede por sí sola introducir riesgos de disponibilidad.

Probar la recuperación ante desastres

Pruebe la implementación de la recuperación ante desastres para validar la implementación y pruebe periódicamente la conmutación por error en la región de DR de su carga de trabajo para asegurarse de que se cumplen el RTO y el RPO.

Un patrón que se debe evitar es desarrollar rutas de recuperación que rara vez se ejecuten. Por ejemplo, puede tener un almacén de datos secundario que se utilice para consultas de solo lectura. Cuando escribe en un almacén de datos y se produce un error del almacén principal, es posible que quiera conmutar por error al almacén de datos secundario. Si no se prueba frecuentemente esta conmutación por error, es posible que sus suposiciones sobre las capacidades del almacén de datos secundario sean incorrectas. Es posible que la capacidad de la secundaria, que podría haber sido suficiente en la última prueba, ya no pueda tolerar la carga en este escenario o que las cuotas de servicio en la región secundaria no sean suficientes.

Nuestra experiencia ha demostrado que la única forma de recuperación de errores que funciona es aquella que se prueba constantemente. Esta es la razón por la que es mejor tener un número reducido de rutas de recuperación.

Puede establecer patrones de recuperación y probarlos con frecuencia. Si tiene una ruta de recuperación compleja o crítica, aún debe ejecutar periódicamente esa falla en la producción para validar que la ruta de recuperación funciona.

Gestione los cambios de configuración en la región de DR. Asegúrese de que la infraestructura, los datos y la configuración sean los necesarios en la región de DR. Por ejemplo, compruebe que las cuotas de servicio AMIs y las cuotas de servicio lo sean up-to-date.

Puede utilizarlo AWS Configpara supervisar y registrar continuamente las configuraciones de sus recursos de AWS. AWS Config puede detectar desviaciones y activar AWS Systems Manager Automation para corregir las desviaciones y activar las alarmas. AWS CloudFormationtambién puede detectar la desviación en las pilas que haya desplegado.

Conclusión

Los clientes son responsables de la disponibilidad de sus aplicaciones en la nube. Es importante definir qué es un desastre y contar con un plan de recuperación ante desastres que refleje esta definición y el impacto que puede tener en los resultados empresariales. Cree un objetivo de tiempo de recuperación (RTO) y un objetivo de punto de recuperación (RPO) en función del análisis del impacto y las evaluaciones de riesgos y, a continuación, elija la arquitectura adecuada para mitigar los desastres. Asegúrese de que la detección de los desastres sea posible y oportuna: es vital saber cuándo los objetivos están en riesgo. Asegúrese de tener un plan y valide el plan realizando pruebas. Los planes de recuperación ante desastres que no se hayan validado corren el riesgo de no implementarse debido a la falta de confianza o al incumplimiento de los objetivos de recuperación ante desastres.

Colaboradores

Los colaboradores de este documento son:

- Alex Livingstone, director práctico de operaciones en la nube de AWS Enterprise Support
- Seth Eliot, arquitecto principal de soluciones de confiabilidad, Amazon Web Services

Documentación adicional

Para obtener información adicional, consulte:

- AWS Centro de arquitectura
- El pilar de la confiabilidad, AWS Well-Architected Framework
- Lista de verificación del plan de recuperación ante desastres
- Implementación de controles de salud
- Arquitectura de recuperación ante desastres (DR) en AWS, parte I: Estrategias de recuperación en la nube
- Arquitectura de recuperación ante desastres (DR) en AWS, parte II: Backup y restauración con recuperación rápida
- Arquitectura de recuperación ante desastres (DR) en AWS, parte III: Pilot Light and Warm Standby
- Arquitectura de recuperación ante desastres (DR) en AWS, parte IV: activo/activo en varios sitios
- Creating Disaster Recovery Mechanisms Using Amazon Route 53
- Minimizing Dependencies in a Disaster Recovery Plan
- Laboratorios prácticos de AWS recuperación ante desastres de Well-Architected
- AWS Implementaciones de soluciones: Multi-Region Application Architecture
- AWS re:Invent 2018: patrones de arquitectura para aplicaciones activas-activas en varias regiones (09-R2) ARC2

Historial del documento

Para recibir notificaciones sobre las actualizaciones de este documento técnico, suscríbase a la fuente RSS.

Cambio	Descripción	Fecha
Actualizaciones menores	Correcciones de errores y numerosos cambios menores en todo momento.	1 de abril de 2022
Documento técnico actualizado	Actualizaciones editoriales menores.	21 de marzo de 2022
Documento técnico actualizado	Se agregó información sobre el plano de datos y el plano de control. Se agregaron más detalles sobre cómo implementar la active/passive conmutación por error. Se reemplazó CloudEndure Disaster Recovery por AWS Elastic Disaster Recovery.	17 de febrero de 2022
Actualización menor	AWS Well-Architected Tool información agregada.	11 de febrero de 2022
Publicación inicial	Documento técnico publicado por primera vez.	12 de febrero de 2021

Avisos

Es responsabilidad de los clientes realizar su propia evaluación independiente de la información que contiene este documento. El presente documento: (a) tiene solo fines informativos, (b) representa las ofertas y prácticas actuales de los productos de AWS, que están sujetas a cambios sin previo aviso, y (c) no supone ningún compromiso ni garantía por parte de AWS y sus filiales, proveedores o licenciantes. Los productos o servicios de AWS se proporcionan “tal cual” sin garantías, declaraciones ni condiciones de ningún tipo, ya sean expresas o implícitas. Las responsabilidades y obligaciones de AWS con respecto a sus clientes se controlan mediante los acuerdos de AWS y este documento no forma parte ni modifica ningún acuerdo entre AWS y sus clientes.

AWS Glosario

Para obtener la AWS terminología más reciente, consulte el AWS glosario de la Glosario de AWS Referencia.

Las traducciones son generadas a través de traducción automática. En caso de conflicto entre la traducción y la version original de inglés, prevalecerá la version en inglés.