Preparación de cara a eventos de infraestructura

Directrices y prácticas recomendadas de AWS

Julio de 2017

Avisos

Este documento se suministra únicamente con fines informativos. Representa la oferta actual de productos y prácticas de AWS a partir de la fecha de publicación de este documento. Dichas prácticas y productos pueden modificarse sin previo aviso. Los clientes son responsables de realizar sus propias evaluaciones independientes de la información contenida en este documento y de cualquier uso de los productos o servicios de AWS, cada uno de los cuales se ofrece “tal cual”, sin garantía de ningún tipo, ya sea explícita o implícita. Este documento no genera ninguna garantía, declaración, compromiso contractual, condición ni certeza por parte de AWS, sus filiales, proveedores o licenciantes. 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.

Contenido

Resumen

En este documento técnico se describen las directrices y las prácticas recomendadas para los clientes con cargas de trabajo de producción implementadas en Amazon Web Services (AWS) que desean diseñar y aprovisionar sus aplicaciones basadas en la nube para gestionar los eventos de escalado planificados, como, por ejemplo, lanzamientos de productos o picos de tráfico estacionales con fluidez y de forma dinámica. Abordamos los principios generales de diseño y proporcionamos prácticas recomendadas específicas y ayuda en diversas áreas conceptuales de la planificación de eventos de infraestructura. A continuación, describimos las consideraciones y prácticas de disponibilidad operativa y las actividades después del evento.

Introducción

La preparación de cara a eventos de infraestructura consiste en el diseño y la preparación de eventos previstos y significativos que afectan a su negocio. Durante estos eventos es fundamental que el servicio web de la empresa sea fiable, con capacidad de respuesta y con una gran tolerancia a errores, bajo cualquier condición y durante los cambios en los patrones de tráfico. Estos eventos pueden incluir la expansión a nuevos territorios, lanzamientos de nuevos productos o funciones, eventos estacionales o anuncios comerciales o eventos de marketing importantes.

Un evento de infraestructura que se haya planeado correctamente puede afectar de forma negativa a la reputación, a la continuidad o a las finanzas de la empresa. Los eventos de infraestructura pueden verse afectados por fallos como, por ejemplo, fallos de servicio no previstos, pérdida de rendimiento relacionada con la carga, latencia de red, limitaciones en la capacidad de almacenamiento, como, por ejemplo, límites para las llamadas al API, una cantidad determinada de direcciones IP disponibles, una comprensión insuficiente del comportamiento de los componentes de una pila de aplicaciones debida a una supervisión insuficiente, dependencias imprevistas de un servicio o componente de terceros que no está configurado para el escalado, o cualquier otro tipo de error imprevisto.

Para minimizar el riesgo de errores no previstos durante un evento importante, las empresas deben invertir tiempo y recursos en la planificación, preparación y capacitación de los empleados, así como para diseñar y documentar los procesos relevantes. La cantidad de inversión en la planificación de eventos de infraestructura para una aplicación o conjunto de aplicaciones en la nube en concreto puede variar en función de la complejidad y del alcance global del sistema. Independientemente del alcance o de la complejidad de la presencia en la nube de una empresa, los principios de diseño y las directrices de las prácticas recomendadas proporcionados en este documento técnico son los mismos.

Con Amazon Web Services (AWS), la empresa puede escalar su infraestructura para prepararse de cara a un evento de escalado previsto de forma dinámica, adaptable y utilizando el método de pago por uso. La amplia gama de productos y servicios elásticos y programables de Amazon ofrece a su empresa acceso a la misma infraestructura segura, fiable y rápida que utiliza Amazon para el

funcionamiento de su propia red internacional y permite a su empresa adaptarse con agilidad para responder a los cambios de sus requisitos empresariales.

En este documento técnico se describen las prácticas recomendadas y los principios de diseño para guiarle en la planificación y ejecución de eventos de infraestructura, y se muestra cómo puede utilizar los servicios de AWS para garantizar que sus aplicaciones estarán preparadas para escalar en función de sus necesidades empresariales.

Planificación de la preparación de cara a eventos de infraestructura

En esta sección se describe en qué consiste un evento de infraestructura planificado y el tipo de actividades que normalmente se producen durante este evento.

¿Qué es un evento de infraestructura planificado?

Un *evento de infraestructura planificado* es un evento previsto y programado que responde a una necesidad de la empresa, y durante el cual mantener un servicio web con una gran capacidad de respuesta, altamente escalable y tolerante a fallos resulta de vital importancia. Este requisito puede estar motivado por campañas de marketing, eventos de noticias relacionadas con el tipo de actividad comercial de la empresa, lanzamientos de productos, una expansión territorial o cualquier actividad similar que provoque un aumento en el tráfico en las aplicaciones web de la empresa y en la infraestructura subyacente.

¿Qué ocurre durante un evento de infraestructura planificado?

La principal preocupación en la mayoría de los eventos de infraestructura planificados es la posibilidad de añadir capacidad a su infraestructura web para satisfacer una mayor demanda de tráfico. En un entorno tradicional en las instalaciones del cliente que cuenta con sus propios recursos de computación, almacenamiento y recursos de red, el departamento de TI de la empresa tendría que aprovisionar capacidad adicional en función de sus estimaciones de un pico

máximo teórico. Esto comporta el riesgo de que la provisión de capacidad sea insuficiente y la empresa sufra una pérdida de negocio por culpa de la sobrecarga de los servidores web, de tiempos de respuesta lentos y de otros errores de tiempo de ejecución.

Dentro de la nube de AWS, la infraestructura es programable y elástica. Esto significa que se puede aprovisionar con rapidez para responder a la demanda en tiempo real. También significa que se puede configurar para que responda a métricas del sistema de manera automatizada, inteligente y dinámica, de modo que los recursos como clústeres de servidor web, rendimiento aprovisionado, capacidad de almacenamiento, núcleos de computación disponibles, número de particiones de streaming, etc., aumenten o disminuyan según sea necesario.

Además, muchos de los servicios de AWS están completamente administrados. Estos servicios incluyen almacenamiento, bases de datos, análisis, aplicaciones y servicios de implementación. Esto significa que los clientes de AWS no tienen que preocuparse de las complejidades de la configuración de estos servicios para eventos de tráfico elevado. Los servicios completamente administrados de AWS han sido diseñados para ofrecer escalabilidad y una elevada disponibilidad.

Normalmente, para preparar un evento de infraestructura planificado, los clientes de AWS realizan una revisión del sistema para evaluar su arquitectura de aplicaciones y la disponibilidad operativa, teniendo en cuenta tanto la escalabilidad como la tolerancia a fallos. Se tiene en cuenta la estimación de tráfico y se compara con el rendimiento de la actividad normal y se determinan las métricas de capacidad y las estimaciones de capacidad adicional necesaria. Se identifica y se soluciona cualquier posible cuello de botella, así como las dependencias de subida y bajada de terceros. También se tiene en cuenta la ubicación geográfica si el evento planificado incluye una expansión de territorio o la introducción de nuevas audiencias. La expansión a otras zonas de disponibilidad o regiones de AWS se realiza antes del evento planificado. También se realiza una revisión de la configuración del sistema dinámico de AWS del cliente, como Auto Scaling, el equilibrio de carga, el enrutamiento geográfico, la alta disponibilidad y medidas de conmutación por error para garantizar que están configurados para gestionar correctamente los aumentos previstos de volumen y de tasa de transacciones. Los ajustes estáticos como, por ejemplo, los límites de recursos de AWS y la ubicación de los servidores de origen red de entrega de contenido (CDN) también se tienen en cuenta y modifican según sea necesario.

También se revisan los mecanismos de supervisión y de notificación y se mejoran, según sea necesario, para proporcionar transparencia en tiempo real de los eventos a medida que se producen y para el análisis posterior una vez el evento planificado se ha completado.

Durante el evento planificado, los clientes de AWS también pueden abrir casos de soporte técnico con AWS si algún asunto requiere asistencia en tiempo real, como, por ejemplo, un servidor que deja de funcionar. Los clientes suscritos al plan de AWS Enterprise Support cuentan con la flexibilidad adicional que supone poder hablar con ingenieros de soporte inmediatamente para plantear casos graves que requieran una respuesta rápida.

Tras el evento, los recursos de AWS han sido diseñados para reducirse automáticamente hasta el nivel adecuado para adaptarse el nivel de tráfico o bien seguir escalándose en función de los eventos.

Principios de diseño

La preparación para eventos programados comienza con un buen diseño al inicio de cualquier implementación de una pila de aplicaciones o carga de trabajo basadas en la nube.

Cargas de trabajo diferenciadas

Un buen diseño es esencial para una gestión eficaz de las cargas de trabajo del evento planificado tanto con el nivel de tráfico normal como cuando aumenta. Asegúrese de diseñar desde el principio agrupaciones funcionales de recursos distintas e independientes centradas en aplicaciones o productos comerciales específicos. En esta sección se describen los distintos aspectos que se deben tener en cuenta para lograr este diseño.

Etiquetado

Las etiquetas se utilizan para marcar y organizar los recursos. Son un componente esencial de la gestión de recursos de infraestructura durante un evento de infraestructura programado. En AWS, las etiquetas las gestiona el cliente, el cual aplica etiquetas basadas en valores clave a un recurso administrado individual, como, por ejemplo, un balanceador de carga o una instancia de Amazon Elastic Compute Cloud (EC2). Al mencionar etiquetas bien definidas que se hayan adjuntado a los recursos de AWS, podrá identificar

fácilmente qué recursos dentro de su infraestructura general componen la carga de trabajo del evento planificado. A continuación, utilizando esta información, puede ejecutar el análisis de cara a la preparación del evento. Las etiquetas también se pueden utilizar para asignar costos.

Las etiquetas se pueden utilizar para organizar instancias EC2, imágenes realizadas con la imagen de máquina de Amazon (AMI), balanceadores de carga, grupos de seguridad, recursos de Amazon Relational Database Service (RDS), recursos de Amazon Virtual Private Cloud (VPC), comprobaciones de estado de Amazon Route 53 y buckets de Amazon Simple Storage Service (S3), por ejemplo.

Para obtener más información sobre estrategias de etiquetado eficaces, consulte AWS Tagging Strategies.[1]

Para ver ejemplos de cómo crear y gestionar etiquetas e incluirlas en grupos de recursos, consulte Resource Groups and Tagging for AWS.[2]

Bajo acoplamiento

Cuando diseña para la nube, debe proyectar cada componente de la pila de aplicaciones para que opere de la forma más independiente posible. De este modo, se proporciona a las cargas de trabajo basadas en la nube las ventajas de la resiliencia y la escalabilidad.

Puede reducir las dependencias entre los componentes de una pila de aplicaciones basada en la nube al diseñar cada componente como una caja negra con interfaces bien definidas para entradas y salidas (por ejemplo, API RESTful). Si los componentes no son aplicaciones, pero son servicios que juntos componen una aplicación, reciben el nombre de *arquitectura de microservicios*. Para la comunicación y la coordinación entre componentes de una aplicación, puede utilizar los mecanismos de notificación basados en eventos como, por ejemplo, las colas de mensajes de AWS para transmitir mensajes entre los componentes, tal y como se muestra en la figura 1.

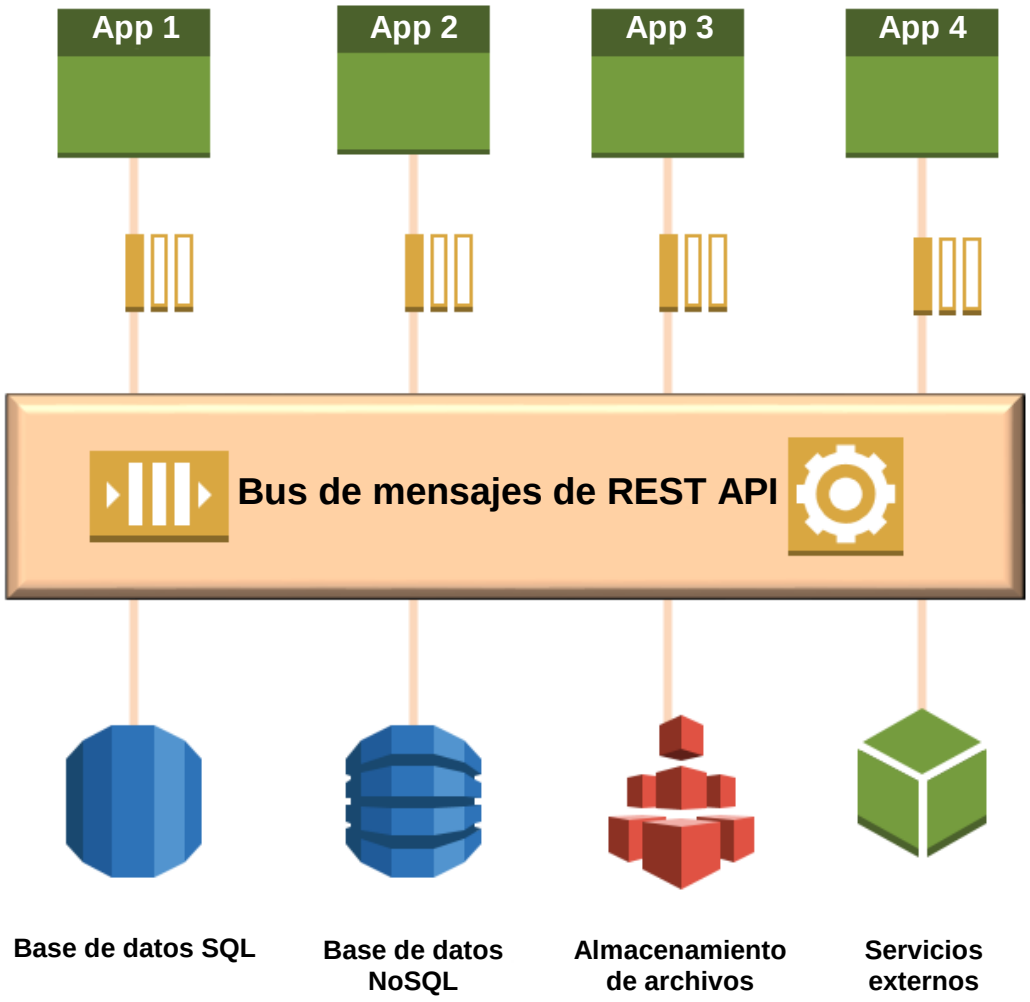

Figura 1. Bajo acoplamiento con interfaces RESTful y colas de mensajes

Cuando se utilizan mecanismos como estos, un cambio o un error en un componente tiene muchas menos posibilidades de afectar en cascada a otros componentes. Por ejemplo, si un servidor en una pila de aplicaciones multinivel deja de responder, las aplicaciones con un bajo acoplamiento pueden diseñarse para que omitan el nivel que no responde o para que pasen a usar transacciones alternativas de modo degradado.

Los componentes de la aplicación con un bajo acoplamiento que utilicen colas de mensajes intermedias también se pueden diseñar más fácilmente para la integración asíncrona. Como los componentes de una aplicación no utilizan una comunicación directa de punto a punto, sino que utilizan una capa de mensajería persistente intermedia (por ejemplo, una cola de Amazon Simple Queue Service (SQS) o a un mecanismo de streaming de datos como Amazon

Kinesis Streams), pueden resistir aumentos repentinos de la actividad en un componente mientras que los componentes de bajada procesan la cola entrante. O, si se produce un error en un componente, los mensajes persisten en las colas o flujos hasta que el componente se pueda recuperar.

Para obtener más información sobre colas de mensajes y los servicios de notificación que ofrece AWS, consulte Amazon Simple Queue Service.[3]

Servicios, no servidores

Los servicios administrados y los puntos de enlace de servicio le liberan de la necesidad de preocuparse por la seguridad o el acceso, las copias de seguridad o las restauraciones, la gestión de parches o el cambio de control, la configuración de la supervisión o la elaboración de informes o de la necesidad de administrar muchos de los detalles de gestión de los sistemas tradicionales. Estos recursos en la nube se pueden aprovisionar con antelación para ofrecer un alto nivel de disponibilidad y flexibilidad, mediante el uso de diversas configuraciones de zona de disponibilidad (o, en algunos casos, múltiples regiones). Se pueden aumentar o reducir, a menudo sin tiempo de inactividad, y puede configurarlos sobre la marcha a través de AWS Management Console o de llamadas a API/CLI.

Los servicios administrados y los puntos de enlace de servicio se pueden utilizar para dotar a las pilas de aplicaciones del cliente de funciones como, por ejemplo, sistemas de bases de datos NoSQL y relacionales, almacenamiento de datos, notificación de eventos, almacenamiento de archivos y objetos, streaming en tiempo real, análisis de big data, aprendizaje automático, búsqueda, transcodificación y muchas otras. Un punto de enlace es una URL que constituye el punto de entrada de un servicio de AWS. Por ejemplo, https://dynamodb.us-west-2.amazonaws.com es un punto de entrada para el servicio de Amazon DynamoDB.

Cuando utiliza servicios administrados y sus puntos de enlace de servicio, puede beneficiarse de la potencia de los recursos listos para la producción como parte de su solución de diseño para hacer frente a un volumen, un alcance y unas tasas de transacciones mayores durante el evento. No necesita aprovisionar ni administrar sus propios servidores cuando estos realicen las mismas funciones que los servicios administrados.

Para obtener más información acerca de los puntos de enlace de servicio de AWS, consulte Regiones y puntos de enlace de AWS.[4] Consulte también Amazon EMR,[5] Amazon RDS[6] y Amazon ECS[7] para obtener ejemplos de servicios administrados que tienen puntos de enlace.

Arquitecturas sin servidor

El uso de AWS Lambda es otra estrategia que puede hacer frente de forma eficaz a la necesidad de dar respuesta a las cargas de procesamiento cambiantes de forma dinámica durante un evento de infraestructura planificado. Lambda es una plataforma informática sin servidor impulsada por eventos. Se trata de un servicio invocado de forma dinámica que ejecuta código Python, Node.js o Java en respuesta a eventos (a través de notificaciones) y que gestiona automáticamente los recursos de computación especificados por el código. Lambda no requiere el aprovisionamiento previo de recursos de computación de Amazon EC2. Amazon Simple Notification Service (Amazon SNS) se puede configurar para que active funciones de Lambda. Para obtener más información sobre Amazon SNS, consulte Amazon Push Notification Service.[8]

Las funciones sin servidor Lambda pueden ejecutar código que invoca o accede a otros servicios de AWS como, por ejemplo, las operaciones de la base de datos, transformaciones de datos, recuperación de objetos o archivos o, incluso, escalar operaciones en respuesta a eventos externos o a métricas de carga del sistema interno. AWS Lambda también puede generar notificaciones o eventos nuevos por sí mismo e incluso lanzar otras funciones de Lambda.

AWS Lambda ofrece la posibilidad de ejercer un control preciso sobre las operaciones de escalado durante un evento de infraestructura planificado. Por ejemplo, Lambda puede utilizarse para ampliar la funcionalidad de las operaciones de Auto Scaling con el fin de realizar acciones como, por ejemplo, la notificación de sistemas de terceros que también necesitan escalarse o para incluir interfaces de red adicionales a instancias nuevas a medida que se suministran. Consulte Using AWS Lambda with Auto Scaling Lifecycle Hooks[9] para obtener ejemplos de cómo utilizar Lambda para personalizar las operaciones de escalado.

Para obtener más información sobre AWS Lambda, consulte What is AWS Lambda?[10]

Automatización

Auto Scaling

Un componente crucial de la planificación de eventos de infraestructura es la funcionalidad Auto Scaling. La posibilidad de escalar automáticamente la capacidad de una aplicación para aumentarla o reducirla, en función de unas condiciones predefinidas, ayuda a mantener la disponibilidad de la aplicación durante las fluctuaciones en los patrones de tráfico y de volumen que se producen durante un evento de infraestructura planificado.

AWS proporciona capacidad de Auto Scaling en muchos de sus recursos, incluidas las instancias EC2, la capacidad de base de datos, los contenedores, etc.

Se puede utilizar Auto Scaling para escalar agrupaciones de instancias, como una flota de servidores que componen una aplicación basada en la nube, para que se puedan escalar automáticamente en función de determinados criterios. También se puede utilizar Auto Scaling para mantener un número fijo de instancias incluso cuando una instancia presente problemas. Este escalado automático, junto a la capacidad de mantener la cantidad de instancias, constituyen la funcionalidad principal del servicio de Auto Scaling.

Auto Scaling mantiene el número de instancias que haya especificado mediante comprobaciones de estado periódicas en las instancias del grupo. Si una instancia presenta problemas, el grupo termina dicha instancia y lanza otra para sustituirla.

Las políticas de Auto Scaling pueden utilizarse para aumentar o reducir automáticamente el número de instancias EC2 en ejecución en un grupo de servidores para satisfacer condiciones cambiantes. Cuando la política de escalado entra en vigor, el grupo de Auto Scaling ajusta la capacidad deseada del grupo y lanza o termina las instancias según sea necesario, ya sea de forma dinámica o alternativa siguiendo un calendario, si hay un tráfico que cambia de volumen constantemente.

Reinicio y recuperación

Un elemento de diseño importante en cualquier evento de infraestructura planificado es contar con los procedimientos y la automatización para gestionar las instancias o servidores comprometidos y poder recuperarlos o reiniciarlos sobre la marcha.

Las instancias EC2 se pueden configurar para recuperarse automáticamente cuando falla una comprobación de estado del sistema del hardware subyacente. La instancia se reiniciará (en hardware nuevo si es necesario), pero conservará su ID de instancia, la dirección IP, las direcciones IP elásticas, los adjuntos de volumen de Amazon Elastic Block Store (EBS) y otros detalles de configuración. Para obtener más información sobre la recuperación automática de instancias EC2, consulte Auto Recovery of Amazon EC2.[11]

Gestión de la configuración / Orquestación

Un elemento fundamental en una estrategia de planificación de eventos de infraestructura que resulte sólida, fiable y con capacidad de respuesta es la incorporación de herramientas de administración y orquestación de la configuración para la administración del estado de recursos individuales y la implementación de pilas de aplicaciones.

Las herramientas de administración de la configuración normalmente gestionan el aprovisionamiento y la configuración de instancias de servidor, los balanceadores de carga, Auto Scaling, la implementación de aplicaciones individuales y la supervisión del estado de la aplicación. También proporcionan la capacidad para integrar servicios adicionales como, por ejemplo, bases de datos, volúmenes de almacenamiento y capas de almacenamiento en caché.

Las herramientas de orquestación, que se sitúan una capa de abstracción por encima de la gestión de la configuración, proporcionan los medios para especificar las relaciones de estos recursos, lo cual permite a los clientes aprovisionar y gestionar varios recursos como una infraestructura de aplicación en la nube unificada, sin tener que preocuparse de las dependencias de recursos.

Dado que estas herramientas definen y describen los recursos individuales, así como sus relaciones como código, este puede controlarse mediante versiones, lo que facilita la posibilidad de restaurar versiones anteriores o probar nuevas ramas de código con el fin de experimentar nuevas soluciones de desarrollo. También es posible definir orquestaciones y configuraciones optimizadas para un evento de infraestructura y, a continuación, volver a la configuración estándar cuando se produzca dicho evento.

Amazon Web Services recomienda las siguientes herramientas para lograr implementaciones y orquestaciones de hardware como código:

- **AWS Config with Config Rules** o un socio de AWS Config para proporcionar un inventario de los recursos de AWS detallado, visual y que admita búsquedas, así como el historial de configuración y el cumplimiento de la configuración de recursos.
- **AWS CloudFormation** o herramientas de orquestación de recursos de AWS de terceros para gestionar, actualizar y terminar el aprovisionamiento de recursos de AWS.
- **AWS OpsWorks, Elastic Beanstalk** o herramientas de administración de la configuración del servidor de terceros para gestionar los cambios en el sistema operativo (SO) y en la configuración de la aplicación.

Consulte Infrastructure Configuration Management para obtener más información sobre cómo gestionar el hardware como código.[12]

Diversidad / Resiliencia

Eliminación de puntos de error únicos y cuellos de botella

Cuando planifique un evento de infraestructura debe analizar las pilas de aplicaciones en busca de puntos de error únicos (SPOF) o cuellos de botella que afecten al rendimiento. ¿Hay alguna instancia única de algún servidor, volumen de datos, base de datos, portal NAT o balanceador de carga, por ejemplo, que pueda provocar que toda la aplicación o partes considerables de la misma dejen de funcionar debido a un error?

En segundo lugar, ya que la aplicación basada en la nube aumenta de escala en cuanto a tráfico o volumen de transacciones, ¿existe alguna parte de la infraestructura que pueda encontrarse con un límite físico o restricción como, por ejemplo, ancho de banda de red o ciclos de procesamiento de la CPU a medida que el volumen de datos crezca junto con la ruta de flujo de datos?

Estos riesgos, una vez identificados, se pueden reducir de diversas formas.

Diseño preparado para los errores

Como se ha mencionado anteriormente, el uso de un bajo acoplamiento y colas de mensajes con interfaces RESTful es una buena estrategia para lograr una elevada resiliencia frente a errores de recursos individuales o fluctuaciones en el tráfico o el volumen de las transacciones. Otra aspecto que se debe tener en

cuenta en un diseño resiliente consiste en configurar los componentes de la aplicación para que permanezcan sin estado en la medida de lo posible.

Las aplicaciones sin estado no necesitan tener en consideración transacciones anteriores y cuentan con una baja dependencia de otros componentes de la aplicación. No almacenan ninguna información de sesión. Una aplicación sin estado puede escalar horizontalmente, como miembro de un grupo o clúster, ya que cualquier solicitud puede gestionarla cualquier instancia dentro del grupo o clúster. Simplemente añada más recursos según sea necesario a través de Auto Scaling y de los criterios de comprobación de estado para gestionar de forma programática la capacidad de computación, de almacenamiento y los requisitos de rendimiento fluctuantes. Cuando una aplicación se ha diseñado con un protocolo sin estado, puede redistribuirse si se desea en una arquitectura sin servidor, mediante el uso de funciones de Lambda en lugar de instancias EC2. Las funciones de Lambda también cuentan con capacidad integrada de escalado dinámico.

En la situación en la que un recurso de la aplicación como, por ejemplo, un servidor web no pueda evitar obtener datos sobre el estado de las transacciones, debe considerarse la opción de diseñar las aplicaciones de manera que las partes de la aplicación con estado estén desvinculadas de los servidores en sí. Por ejemplo, los datos de estado de cookies HTTP o equivalentes pueden almacenarse en una base de datos, como, por ejemplo, DynamoDB, o en un bucket de S3 o volumen de EBS.

Si tiene un flujo de trabajo complejo con múltiples pasos donde sea necesario realizar un seguimiento del estado actual de cada uno de ellos, puede utilizar Amazon Simple Workflow Service (SWF) para almacenar de forma centralizada el historial de ejecución y convertir estas cargas de trabajo a cargas sin estado.

Otra medida de resiliencia consiste en utilizar el procesamiento distribuido. Para casos de uso que requieren el procesamiento de grandes cantidades de datos de manera puntual en un único recurso de computación que no puede satisfacer la demanda, puede diseñar sus cargas de trabajo a fin de que las tareas y los datos se dividan en fragmentos más pequeños y se ejecuten en paralelo en un clúster de instancias de recursos de computación. El procesamiento distribuido no tiene estado, ya que los nodos independientes en los que se están procesando los datos particionados y las tareas pueden fallar. En este caso, el motor de programación de procesamiento distribuido se

encarga de reiniciar de forma automática las tareas fallidas en otro nodo del clúster de procesamiento distribuido.

AWS ofrece diversos motores de procesamiento de datos distribuidos como Amazon EMR, Amazon Athena y Amazon Machine Learning, cada uno de los cuales es un servicio administrado que proporciona puntos de enlace de servicio y blindaje frente a cualquier proceso complejo que conlleve la aplicación de parches, tareas de mantenimiento, escalado, conmutación por error, etc.

Para el procesamiento en tiempo real de datos de streaming, Amazon Kinesis Streams puede dividir los datos en varios fragmentos que se pueden procesar mediante diferentes consumidores de datos, como las funciones de Lambda o instancias EC2.

Para obtener más información sobre estos tipos de cargas de trabajo, consulte Big Data Analytics Options on AWS.[13]

Zonas y regiones múltiples

Los servicios de AWS están alojados en varias ubicaciones por todo el mundo. Estas ubicaciones están compuestas por regiones y zonas de disponibilidad. Una región es una zona geográfica independiente. Cada región cuenta con varias ubicaciones aisladas, denominadas zonas de disponibilidad. AWS ofrece a sus clientes la posibilidad de colocar recursos como, por ejemplo, instancias y datos, en múltiples ubicaciones.

Debe diseñar sus aplicaciones para que se distribuyan por múltiples zonas de disponibilidad y regiones. En combinación con la distribución y la replicación de recursos en las regiones y zonas de disponibilidad, debe diseñar las aplicaciones utilizando mecanismos de balanceo de carga y de conmutación por error, para que las pilas de aplicaciones redirijan automáticamente los flujos de datos y el tráfico a estas ubicaciones alternativas en caso de que se produzca algún error.

Balanceo de carga

Con el servicio Elastic Load Balancing (ELB), se puede adjuntar una flota de servidores de aplicaciones a un balanceador de carga y aún puede distribuirse a través de varias zonas de disponibilidad. Cuando fallan las comprobaciones de estado de instancias EC2 de una zona de disponibilidad concreta que se encuentran detrás de un balanceador de carga, el balanceador de carga deja de

enviar el tráfico a estos nodos. En combinación con Auto Scaling, el número de nodos en buen estado se reequilibra automáticamente con otras zonas de disponibilidad y no se requiere intervención manual.

También es posible que el balanceado de carga se realice entre regiones mediante el uso de Amazon Route 53 y algoritmos de direccionamiento DNS basados en la latencia. Consulte Latency Based Routing para obtener más información.[14]

Estrategias de diseminación de carga

El concepto de *diseminación de carga* en infraestructuras basadas en la nube consiste en redirigir o enviar mediante un servidor proxy el tráfico a otra ubicación para reducir la presión en los sistemas principales. En algunos casos, la estrategia de diseminación de carga puede obligarnos a establecer prioridades, haciéndonos rechazar determinadas transmisiones de tráfico o reducir la funcionalidad de las aplicaciones para aligerar la carga de procesamiento y poder atender a, por lo menos, una parte de las solicitudes entrantes.

Existen numerosas técnicas que se pueden utilizar para llevar a cabo la diseminación de carga. Uno de estos métodos es el direccionamiento DNS basado en la latencia. Otro método consiste en utilizar el almacenamiento en caché. El almacenamiento en caché puede realizarse cerca de la aplicación, utilizando una capa de caché dentro de la memoria como, por ejemplo, Amazon ElastiCache. También puede utilizar una capa de caché que esté más cerca del extremo del usuario, mediante una red de distribución de contenido global como, por ejemplo, Amazon CloudFront.

Para obtener más información sobre ElastiCache y CloudFront, consulte la sección Introducción y ElastiCache[15] y Amazon CloudFront CDN.[16]

Optimización de costos

Instancias reservadas frente a instancias de subasta y bajo demanda

La capacidad de controlar los costos de aprovisionamiento de recursos en la nube está estrechamente ligada a la capacidad para aprovisionar recursos en la nube de forma dinámica en función de las métricas de rendimiento y otros criterios de comprobación del estado y del rendimiento. Con Auto Scaling, la

utilización de recursos puede ajustarse de forma precisa a las necesidades de almacenamiento y procesamiento reales para minimizar los gastos innecesarios y los recursos infrautilizados.

Otro aspecto del control de costos en la nube es la posibilidad de elegir entre instancias bajo demanda, instancias reservadas o instancias de subasta. También hay una función de reserva de capacidad para DynamoDB.

Con las instancias bajo demanda solo paga por las instancias EC2 que utilice. Las instancias bajo demanda le permiten pagar por la capacidad de computación por hora, sin necesidad de fijar compromisos a largo plazo.

Las instancias reservadas Amazon EC2 ofrecen un descuento importante (hasta del 75 %) en comparación con los precios de las instancias bajo demanda, y proporcionan una reserva de capacidad cuando se utilizan en una zona de disponibilidad específica. Sin embargo, aparte de la reserva de disponibilidad y del descuento en la facturación, no hay diferencias funcionales entre las instancias reservadas y las instancias bajo demanda.

Las instancias de subasta le permiten pujar por capacidad de computación libre de Amazon EC2. Las instancias de subasta a menudo se ofrecen con descuentos respecto a los precios bajo demanda, lo cual reduce significativamente el costo de ejecución de las aplicaciones basadas en la nube.

Cuando se diseña para la nube, algunos casos de uso son más adecuados que otros para las instancias de subasta. Por ejemplo, como las instancias de subasta se pueden retirar en cualquier momento en cuanto una puja supere su oferta, debería considerar la ejecución de las instancias de subasta únicamente con pilas de aplicaciones de escalado horizontal y sin estado, en la medida de lo posible. Para las aplicaciones con estado o para cargas de procesamiento caras, las instancias reservadas o las instancias bajo demanda pueden ser una opción mejor. Para aplicaciones críticas en las que no puede arriesgarse a tener limitaciones de capacidad, las instancias reservadas son la opción ideal.

Consulte Reserved Instances[17] y Spot Instances[18] para obtener más información.

Proceso de gestión de eventos

La planificación de un evento de infraestructura es una actividad de grupo que involucra a los desarrolladores de aplicaciones, a los administradores y a los grupos interesados de la empresa. Semanas antes de llevar a cabo un evento de infraestructura, deberá haber establecido una cadencia de reuniones recurrentes que involucren al personal técnico clave que posee y opera cada uno de los principales componentes de la infraestructura del servicio web.

Programación de un evento de infraestructura

La planificación de un evento de infraestructura debería empezar varias semanas antes de la fecha del evento. En la figura 2 se muestra un calendario típico del ciclo de vida de un evento planificado.

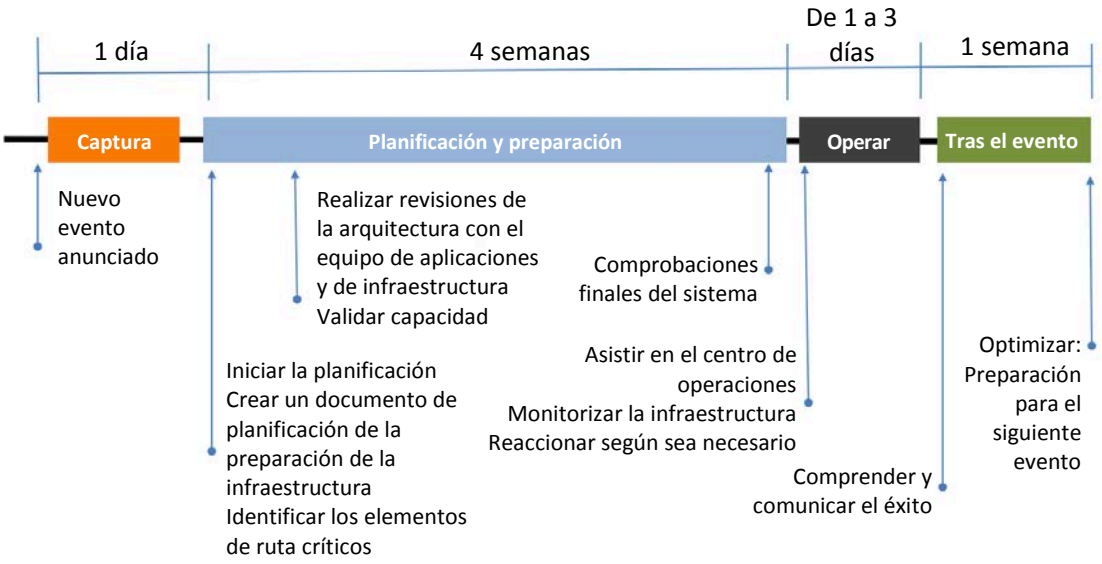

Figura 2. Calendario típico de un evento de infraestructura

Planificación y preparación

Programación

Le recomendamos el siguiente programa de actividades en las semanas anteriores al evento de infraestructura:

Semana 1:

- Designe un equipo para impulsar la planificación y el diseño del evento de infraestructura.
- Lleve a cabo reuniones entre las partes interesadas para comprender los parámetros del evento (escala, duración, tiempo, alcance geográfico, cargas de trabajo afectadas) y los criterios para valorar el éxito de la operación.
- Implique a cualquier socio o proveedor de cualquier nivel.

Semanas 2-3:

- Revise la arquitectura y realice los ajustes necesarios.
- Lleve a cabo una revisión operativa y ajuste según sea necesario.
- Siga las prácticas recomendadas que se describen en este documento y en las referencias a pie de página.
- Identifique los riesgos y desarrolle planes de mitigación.
- Elabore una guía para el evento planificado.

Semana 4:

- Revise todos los servicios basados en la nube de su proveedor que deban escalarse en función de la carga esperada.
- Compruebe los límites del servicio, y auméntelos según sea necesario.
- Configure un panel de monitorización y alertas para los umbrales definidos.

Revisión de la arquitectura

Una parte esencial de la preparación de un evento de infraestructura consiste en revisar la arquitectura de la pila de aplicaciones que experimentará el aumento de tráfico. El objetivo de la revisión es verificar e identificar posibles áreas de riesgo que afecten a la escalabilidad o la fiabilidad de la aplicación, así como identificar posibles formas de optimizar el conjunto antes del evento.

AWS proporciona a sus clientes de Enterprise Support un marco para la revisión de las pilas de aplicaciones del cliente que se centra en cinco pilares de diseño. Estos pilares son: seguridad, fiabilidad, eficiencia del rendimiento,

optimización de costos y excelencia operativa, tal y como se describe a continuación.

Tabla 1: Pilares de las aplicaciones bien diseñadas

Nombre del pilar	Definición del pilar	Área de interés
Seguridad	La capacidad para proteger la información, los sistemas y los activos, a la vez que se aporta valor empresarial a través de evaluaciones de riesgos y estrategias de mitigación.	Gestión de identidad, cifrado, monitorización, registros, gestión de claves, instancias dedicadas, cumplimiento, gobernanza
Fiabilidad	La capacidad de un sistema para recuperarse de un error en la infraestructura o en el servicio, adquirir de forma dinámica recursos de computación para satisfacer la demanda y mitigar interrupciones como, por ejemplo, configuraciones erróneas o problemas de red transitorios.	Límites de servicio, diversas zonas y regiones de disponibilidad, escalabilidad, comprobación de estado/monitorización, copia de seguridad/recuperación de desastres, redes, automatización de la reparación automática
Eficiencia del rendimiento	La capacidad de utilizar los recursos de computación de forma eficiente para satisfacer los requisitos del sistema y para mantener la eficiencia a medida que cambia la demanda y las tecnologías evolucionan.	Servicios de AWS adecuados, utilización de recursos, arquitectura del almacenamiento, almacenamiento en caché, requisitos de latencia
Optimización de costos	La capacidad de evitar o eliminar costos innecesarios o recursos no aprovechados al máximo.	Instancias reservadas y de subasta, ajuste del entorno, selección de servicios, ajuste del volumen, gestión de cuentas, facturación consolidada, retirada de recursos
Excelencia operativa	La capacidad de ejecutar y supervisar sistemas para ofrecer un valor empresarial y para mejorar los procesos y los procedimientos de soporte continuamente.	Guías, guías prácticas, CI/CD, días de juego, infraestructura como código, RCA

Encontrará una lista detallada de los elementos de la arquitectura para su revisión, que puede utilizar para revisar una pila de aplicaciones basada en AWS, disponible en el Anexo de este documento técnico.

Revisión operativa

Además de una revisión de la arquitectura, que se centra principalmente en los componentes de diseño, debe revisar las operaciones en la nube y las prácticas de gestión para evaluar la eficacia con la que aborda la gestión de las cargas de trabajo en la nube. El objetivo de la revisión consiste en identificar las

deficiencias y problemas operativos y emprender acciones antes del evento para poder reducirlos al mínimo.

AWS ofrece una revisión de operaciones en la nube a sus clientes de Enterprise Support. Esta revisión puede constituir una herramienta importante de cara a la preparación de un evento de infraestructura. La revisión se centra en la evaluación de las siguientes áreas:

- Preparación: debe disponer de la combinación correcta de estructura organizativa, procesos y tecnología. Debe tener claras las funciones y responsabilidades definidas para el personal que gestiona la pila de aplicaciones. Deben definirse los procesos por adelantado para que se alineen con el evento. Deben automatizarse los procesos siempre que sea posible.
- Supervisión: la supervisión mide de forma eficaz si una aplicación cumple su cometido de forma adecuada. La monitorización es crucial para detectar anomalías antes de que se conviertan en problemas, y ofrece oportunidades para minimizar el efecto que puedan causar los contratiempos.
- Operaciones: las actividades operativas deben realizarse de forma oportuna y fiable aprovechando la automatización siempre que sea posible, además de abordar los eventos operativos inesperados que necesiten escalarse.
- Optimización: realice un análisis post mortem utilizando las métricas recopiladas, las tendencias operativas y las lecciones aprendidas para detectar y notificar las oportunidades de mejora para eventos futuros. La optimización, junto a una adecuada preparación, crea un sistema de información que se autoalimenta para abordar los problemas operativos y evitar que vuelvan a suceder.

Conocimiento de los límites de los servicios de AWS

Durante un evento de infraestructura planificado, es de vital importancia no superar cualquier límite de servicio que pueda imponer un proveedor de nube a la hora de escalar una aplicación o una carga de trabajo.

Los proveedores de servicios en la nube suelen tener límites para los distintos recursos que pueda utilizar. Estos suelen ser impuestos en función de la cuenta o de la región. Los recursos afectados incluyen, entre otros, instancias,

volúmenes, transmisiones, invocaciones sin servidor, snapshots, número de VPC, reglas de seguridad. Se utilizan como medida de seguridad frente a código desbordado o personas malintencionadas intenten abusar de los recursos; sirve asimismo como mecanismo de control para ayudar a minimizar riesgos de facturación excesiva.

Algunos límites del servicio surgen automáticamente con el tiempo a medida que amplía su presencia en la nube, aunque la mayoría de estos servicios requieren que solicite un aumento del límite mediante la apertura de un caso de soporte. Aunque algunos límites del servicio puede aumentarse a través de casos de soporte, otros servicios tienen límites que no pueden modificarse.

AWS ofrece a los clientes de Enterprise Support y Business Support un Trusted Advisor, que proporciona un panel de comprobación de límites para permitir a los clientes gestionar de forma proactiva todos los límites del servicio.

Para obtener más información acerca de los límites para los diferentes servicios de AWS y cómo comprobarlos, consulte Límites de los servicios de AWS[19] y Trusted Advisor.[20]

Comprensión de patrones

Puntos de referencia

Debe documentar los valores para las métricas clave que indican que el sistema funciona correctamente antes del inicio de un evento de infraestructura. Esto le ayudará a determinar cuándo una aplicación o servicio regresa a los valores normales de forma segura tras completar un evento. Por ejemplo, identificar que la tasa de transacciones normales a través de un balanceador de carga es de 2500 solicitudes por segundo le ayudará a determinar cuándo es seguro empezar a desmantelar los procedimientos después del evento.

Flujos de datos y dependencias

Entender cómo fluyen los datos a través de los distintos componentes de una aplicación le ayuda a identificar posibles cuellos de botella y dependencias. ¿Los niveles de aplicaciones o los componentes que consumen datos se encuentran en un flujo de datos de tamaño adecuado, y se han configurado correctamente para escalarse automáticamente si los niveles o componentes en una pila de aplicaciones que producen datos aumentan? En el caso de que se produzca un error en un componente, ¿los datos pueden esperar en cola hasta que se

recupere el componente? ¿Alguno de los proveedores o consumidores de datos en sentido ascendente o descendente puede escalarse en respuesta a su evento?

Proporcionalidad

Otra aspecto que se debe considerar durante la preparación de un evento de infraestructura es la proporcionalidad de escalado necesaria para los distintos componentes de una pila de aplicaciones. Esta proporcionalidad no es siempre de uno a uno. Por ejemplo, un aumento de las transacciones por segundo diez veces superior a lo normal a través de un balanceador de carga puede necesitar una capacidad hasta veinte veces superior de almacenamiento, o bien del número de fragmentos de streaming de datos o del número de operaciones de lectura y escritura de la base de datos, debido al procesamiento que realice la aplicación en primer plano.

Plan de comunicación

Antes del evento, debe desarrollar un plan de comunicación. Elabore una lista de las partes interesadas y grupos de soporte de la empresa, e identifique a quien se debe contactar en las distintas fases del evento en diversas situaciones, como, por ejemplo, al principio del evento, durante el evento, al final del evento, durante el análisis después del evento, así como los contactos de emergencia, los contactos durante las situaciones de resolución de problemas, etc.

Las personas y grupos con quienes contactar pueden ser, entre otros:

- Partes interesadas
- Directores de operaciones
- Desarrolladores
- Equipos de soporte
- Equipos de proveedores de servicios en la nube
- Equipo del centro de operaciones de red (NOC)

Cuando reúna una lista de contactos internos, también debe desarrollar una lista de contactos de las partes interesadas externas que juegan un papel en el desempeño en tiempo real de la aplicación. Entre estas partes interesadas se encuentran los socios y proveedores que ofrecen soporte a los componentes clave de la pila, así como los proveedores de toda la cadena que proporcionan los servicios externos, fuentes de datos, servicios de autenticación y demás.

Esta lista de contactos externos también debe incluir las siguientes figuras:

- Proveedores de alojamiento de la infraestructura
- Proveedores de telecomunicaciones
- Socios de streaming de datos en directo
- Contactos de marketing de relaciones públicas
- Socios de publicidad
- Consultores técnicos implicados en la ingeniería de servicios

Solicite la información siguiente a cada proveedor:

- Puntos de contacto directo durante la hora del evento
- Contacto de soporte crítico y proceso de escalado
- Nombre, número de teléfono y dirección de correo electrónico
- Verificación de disponibilidad en directo de los contactos técnicos

Los clientes de AWS suscritos a Enterprise Support también cuentan con gestores de cuentas técnicas (TAM) asignados a su cuenta, quienes pueden coordinar y verificar que el personal de AWS Support dedicado está al corriente del evento y se encuentra preparado para ofrecer soporte durante el mismo. Los TAM también se encuentran disponibles para recibir llamadas durante el evento, estar presentes en el centro de operaciones y dirigir escaladas de soporte si fuera necesario.

Preparación del centro de operaciones de red

Antes del evento, debe instruir a su equipo de operaciones o de desarrollo para que cree un panel de métricas que monitorice en tiempo real cada componente crítico del servicio web en producción mientras se lleve a cabo el evento. Idealmente, el panel debe presentar automáticamente métricas actualizadas cada minuto o, con cualquier intervalo que resulte adecuado y eficaz durante el evento.

Considere llevar a cabo la monitorización de los siguientes componentes:

- Uso de recursos de cada servidor (uso de la CPU, del disco y de la memoria)

- Tiempo de respuesta del servicio web
- Métricas del tráfico web (usuarios, visualizaciones de páginas, sesiones)
- Tráfico web por región del usuario (segmentos de clientes globales)
- Uso del servidor de la base de datos
- Embudos de conversión de flujo de marketing, como por ejemplo la tasa de conversión y el porcentaje de bajas
- Registros de errores de las aplicaciones
- Monitorización precoz

Amazon CloudWatch ofrece un mecanismo para recopilar la mayoría de estas métricas de recursos de AWS en una única pantalla mediante paneles de CloudWatch personalizados. Además, CloudWatch ofrece la capacidad para importar métricas personalizadas a CloudWatch en el caso de que AWS no esté proporcionando dicha métrica automáticamente. Consulte la sección de monitorización de este documento para obtener más información acerca de las herramientas y capacidades de monitorización de AWS.

Preparación de una guía de ejecución

Debe elaborar una guía de ejecución como preparación para el evento de infraestructura. Una *guía de ejecución* es un manual operativo que contiene una recopilación de los procedimientos y operaciones que los operadores realizarán durante el evento. Las guías de ejecución de eventos pueden ser una ampliación de una guía de ejecución existente que se use para operaciones rutinarias y en la gestión de excepciones. Normalmente, una guía de ejecución contiene los procedimientos para iniciar, detener, supervisar y depurar un sistema. También debe describir los procedimientos para la gestión de contingencias y eventos inesperados.

Una guía de ejecución debe incluir las siguientes secciones:

- **Información del evento**: Describe brevemente el evento, los criterios de éxito, la cobertura de medios, las fechas del evento y datos de contacto de las principales partes interesadas por parte del cliente y de AWS.

- **Lista de servicios de AWS**: Enumera todos los servicios de AWS que se utilizarán durante el evento. Además, incluye la carga prevista para estos servicios, las regiones afectadas y los ID de cuenta.
- **Revisión de la arquitectura y de la aplicación**: Documenta los resultados de las pruebas de carga, cualquier punto de sobrecarga en la infraestructura y en el diseño de la aplicación, puntos de fallo individuales y los posibles cuellos de botella.
- **Revisión operativa**: Aspectos destacados de la configuración de la monitorización, criterios de estado, mecanismos de notificación y procedimientos de restauración del servicio.
- **Lista de comprobaciones del nivel de preparación**: Incluye consideraciones como las comprobaciones de los límites del servicio, el precalentamiento de los componentes de la pila de aplicaciones como, por ejemplo, los balanceadores de carga, el aprovisionamiento de recursos como fragmentos de streaming de datos, particiones DynamoDB y particiones S3 entre otras. Para obtener más información, consulte la lista de comprobaciones detallada para la revisión de la arquitectura en el Anexo de este documento técnico.

Monitorización

Plan de monitorización

La monitorización de la base de datos, de la aplicación y del sistema operativo es crucial para garantizar el éxito del evento. Se deben configurar unos sistemas de monitorización exhaustiva de modo que se puedan detectar de forma eficaz incidentes graves y responder inmediatamente durante el evento de infraestructura. A grandes rasgos, una estrategia de monitorización eficaz garantiza que las herramientas de monitorización se hayan instalado en el nivel adecuado para una aplicación en función de su importancia para el negocio. Una estrategia de gestión de incidencias eficaz incorporará tanto los datos de monitorización de AWS como del cliente con sus herramientas y procesos de gestión de eventos e incidencias. La implementación de un plan de monitorización que permita recopilar datos de monitorización de todos los segmentos de la solución de AWS de forma colectiva le será de enorme ayuda durante la depuración de un error complejo en caso de que se produzca.

El plan de monitorización debería abordar las siguientes preguntas:

- ¿Qué herramientas de monitorización y paneles se deben configurar para el evento?
- ¿Cuáles son los objetivos de monitorización y los umbrales permitidos? ¿Qué eventos dispararán acciones?
- ¿Qué recursos y qué métricas de estos recursos se monitorizarán y con qué frecuencia se deberán sondear?
- ¿Quién se encargará de realizar las tareas de monitorización? ¿De qué alertas de monitorización dispone? ¿A quién se avisará?
- ¿Qué planes de reparación se han establecido para errores comunes y esperados? ¿Cómo se gestionarán los eventos inesperados?
- ¿Cuál es el proceso de escalado en caso de cualquier error?

Las siguientes herramientas de monitorización de AWS se pueden utilizar como parte de esta estrategia:

- **Amazon CloudWatch**: Una solución lista para utilizar para las métricas del panel de AWS, la monitorización, las alertas y el aprovisionamiento automatizado.
- **Métricas personalizadas de Amazon CloudWatch**: Se utilizan para la recopilación de métricas de sistemas operativos, aplicaciones y empresariales. La API de Amazon CloudWatch permite la recopilación de prácticamente cualquier tipo de métrica personalizada.
- **Estado de las instancias Amazon EC2**: Se utiliza para visualizar las comprobaciones de estado y para programar eventos para sus instancias en función de su estado, como, por ejemplo, reiniciar automáticamente o de forma manual una instancia.
- **Amazon SNS**: Se utiliza para configurar, operar y enviar notificaciones basadas en eventos.
- **AWS X-Ray**: Ayuda en la depuración y el análisis de aplicaciones distribuidas y de arquitectura de microservicios mediante el análisis de los flujos de datos entre los componentes del sistema.
- **Amazon Elasticsearch Service**: Se utiliza para la recopilación de un registro centralizado y el análisis de registros en tiempo real. Para una detección de problemas heurística y rápida.

- **Herramientas de terceros:** Se utilizan para el análisis en tiempo real y la monitorización y visibilidad de toda la pila.
- **Herramientas estándar de monitorización del sistema operativo**: Se utilizan para la monitorización a nivel del sistema operativo.

Para obtener más información sobre las herramientas de monitorización de AWS, consulte Automated and Manual Monitoring.[21] Consulte también Uso de los paneles de Amazon CloudWatch[22] y Publicación de métricas personalizadas.[23]

Notificaciones

Un elemento operativo crucial para el diseño de los eventos de infraestructura es la configuración de alarmas y notificaciones que se integren con sus soluciones de monitorización. Estas alarmas y las notificaciones se pueden utilizar con servicios como AWS Lambda para iniciar acciones en función de la alerta. Las respuestas automatizadas a eventos operativos son un elemento clave para permitir la mitigación, restauración y recuperación con la máxima capacidad de respuesta.

También deben existir herramientas para monitorizar de forma centralizada las cargas de trabajo y crear alertas y notificaciones adecuadas en función de los registros y las métricas disponibles relacionadas con los indicadores operativos clave. Esto incluye las alarmas y notificaciones de anomalías fuera de los límites, así como errores de servicio o del componente. Idealmente, cuando se superan los umbrales de bajo rendimiento o se producen errores, el sistema se habrá diseñado para repararse automáticamente o escalar en respuesta a dichas notificaciones y alertas.

Como hemos indicado anteriormente, AWS ofrece servicios (Amazon SQS y Amazon SNS) para garantizar un sistema de alertas y notificaciones adecuado en respuesta a eventos operativos imprevistos, así como para permitir las respuestas automáticas.

Preparación operativa (día del evento)

Ejecución del plan

El día del evento, el equipo principal relacionado con el evento de infraestructura deberá realizar una conferencia en directo mientras monitoriza

los paneles en tiempo real. Las guías de ejecución deberán estar totalmente acabadas y a su disposición. Asegúrese de que el plan de comunicación está bien definido y que todas las partes interesadas y el personal de soporte lo conocen; compruebe asimismo que se dispone de un plan de emergencia.

Centro de operaciones

Durante el evento, habilite un sistema de comunicación en directo con los siguientes participantes:

- El responsable principal de los equipos de aplicación y de operaciones
- El líder del equipo de operaciones
- Operadores técnicos de socios externos directamente relacionados con los aspectos técnicos
- Las partes interesadas de la empresa

Durante la mayoría del evento, la conversación mediante esta conferencia debería ser la mínima indispensable. Si se produce una anomalía operativa, las personas clave que pueden responder al evento estarán ya conectadas y listas para actuar y ofrecer asesoramiento.

Informes para los responsables

Durante el evento, envíe un correo electrónico cada hora a los responsables principales de las partes interesadas. Esta actualización debe incluir lo siguiente:

- Resumen de estado: Verde (como previsto), amarillo (se han encontrado problemas), rojo (problema importante)
- Actualización de las métricas principales
- Problemas detectados, estado del plan de resolución, tiempo estimado para la resolución
- Número de teléfono para contactar con el centro de operaciones (en caso de que alguien desee unirse)

Al finalizar el evento, se debe enviar un correo electrónico de resumen final con un formato similar.

Plan de emergencia

Cada paso del proceso de preparación para el evento debe tener un plan de restauración correspondiente que haya sido verificado en un entorno de pruebas.

Considere las siguientes preguntas a medida que elabore el plan de restauración:

- ¿Cuáles son las peores situaciones que pueden producirse durante el evento?
- ¿Qué tipo de eventos pueden afectar negativamente a las relaciones públicas?
- ¿Qué componentes y servicios de terceros pueden fallar durante el evento?
- ¿Qué métricas deben monitorizarse que puedan indicar que se está produciendo una situación problemática?
- ¿Cuál es el plan de restauración para cada situación posible?
- ¿Cuánto tiempo requiere cada proceso de restauración? ¿Cuál es el objetivo de punto de recuperación (RPO) y el objetivo del tiempo de recuperación (RTO) que se considera aceptable? (Consulte Using AWS for Disaster Recovery[24] para obtener más información sobre estos conceptos).

Considere los siguientes tipos de restauración:

- **Implementación azul/verde**: Si está desplegando un nuevo entorno o aplicación de producción, conserve la versión de producción en línea y disponible para poder volver a utilizarla rápidamente.
- **Piloto preparado**: Lance un entorno mínimo en una región secundaria que pueda escalarse rápidamente si fuera necesario. En caso de que se produzca un error en la región principal, puede escalar rápidamente en la región de copia de seguridad y cambiar el tráfico hacia la región secundaria.
- **Páginas de error del modo de mantenimiento**: Consulte las instalaciones y disparadores de la página de error en cada capa de su

servicio web. Prepárese para añadir un mensaje más específico en estas páginas de error según sea necesario.

Pruebe y documente cada plan de restauración para cada posible situación de error.

Actividades posteriores al evento

Análisis post mortem

El análisis post mortem se omite con mucha frecuencia, ya que los clientes suelen estar ansiosos por volver a las operaciones normales. Sin embargo, le recomendamos que solicite un análisis post mortem como parte del ciclo de vida de gestión de cualquier evento de infraestructura. Los análisis post mortem le permiten colaborar con cada equipo implicado e identificar áreas que puedan necesitar optimización como, por ejemplo, procedimientos operativos, detalles de la implementación, procedimientos de conmutación por error y recuperación, etc. Esto es especialmente importante si una pila de aplicaciones ha sufrido interrupciones durante el evento. Un análisis post mortem del evento también ayuda a proporcionar documentación en caso de que sea necesario desarrollar documentos de análisis de la causa principal (RCA).

Proceso de cese gradual

Inmediatamente después de la finalización del evento de infraestructura, debería iniciarse el proceso de cese gradual. Durante este período, es aconsejable seguir monitorizando las aplicaciones y servicios relevantes para garantizar que el tráfico ha vuelto a los niveles de producción normales. Utilice los paneles de estado creados durante la fase de preparación para verificar la normalización del tráfico y de las tasas de transacción. Los períodos de cese gradual de algunos eventos pueden ser lineales y sencillos, mientras que otros pueden experimentar una reducción de volumen irregular o más gradual. Algunos patrones de tráfico pueden mantenerse. Por ejemplo, la recuperación de un aumento repentino de tráfico, por lo general, requiere procedimientos de cese gradual sencillos, mientras que la implementación de una aplicación o la expansión a una nueva región geográfica pueden tener efectos duraderos que requieran una monitorización cuidadosa de los nuevos patrones de tráfico, así como aplicar un sistema de monitorización adicional como parte fija de la pila de aplicaciones.

En algún momento, tras la finalización del evento, debe determinar cuándo es seguro finalizar las operaciones de gestión de eventos. Consulte los valores "normales" de las métricas clave documentados previamente para ayudarle a determinar cuándo puede declarar que un evento se ha completado o ha finalizado. Recomendamos dividir las actividades de cese gradual en dos vertientes, que pueden tener calendarios distintos. Centre la primera vertiente en la gestión operativa del evento, como, por ejemplo, el envío de las comunicaciones a las partes interesadas y socios, tanto internos como externos, y el restablecimiento de los límites del servicio. Centre la segunda vertiente en los aspectos técnicos del cese gradual, como, por ejemplo, reducir los procedimientos, validar el estado del entorno de TI y los criterios para determinar si los cambios de arquitectura deben ser recuperados o confirmados.

El calendario asociado a cada una de las vertientes puede variar en función de la naturaleza del evento, las métricas clave y la comodidad de los clientes. Hemos indicado algunas tareas comunes asociadas con cada vertiente en la siguiente tabla para ayudarle a determinar el momento adecuado para concluir la gestión de un evento.

Tabla 2: Tareas de cese gradual operativo

Tarea	Descripción
Comunicaciones	Notificación a las partes interesadas internas y externas de que el evento ha finalizado. La comunicación sobre el momento de finalización debe ir en consonancia con la definición de la realización del evento. Utilice las métricas que indican que el sistema funciona correctamente para determinar cuándo es conveniente finalizar la comunicación. De forma alternativa, puede finalizar la comunicación por niveles. Por ejemplo, puede concluir la comunicación con el centro de operaciones, pero dejar los procedimientos de escalado del evento activos por si se producen errores después del evento.
Límites de servicio / Contención de costos	Aunque puede ser tentador conservar un límite de servicio elevado después de un evento, tenga en cuenta que los límites de servicio también se utilizan como una red de seguridad. Los límites de servicio le protegen a usted y a sus costos al evitar un uso excesivo del servicio, que puede producirse por una cuenta comprometida o una automatización configurada erróneamente.
Informes y análisis	La recopilación de datos y de métricas de eventos, acompañada de narrativas analíticas que muestran patrones, tendencias, áreas problemáticas, procedimientos llevados a cabo con éxito, procedimientos ad-hoc, el calendario del evento y si se han cumplido los criterios de éxito o no son procedimientos que deben desarrollarse y distribuirse a todas las partes internas identificadas en el plan de comunicaciones. También debe desarrollarse un análisis detallado del costo para mostrar el gasto operativo de ofrecer soporte al evento.

Tarea	Descripción
Tareas de optimización	Las organizaciones empresariales evolucionan con el tiempo a medida que siguen mejorando sus operaciones. La optimización operativa requiere la recopilación constante de métricas, tendencias operativas y lecciones aprendidas en eventos para descubrir las oportunidades que permitan mejorar. La optimización está ligada a la preparación para formar un bucle de retroalimentación para abordar los problemas operativos y evitar que vuelvan a suceder.

Tabla 3: Tareas técnicas de cese gradual

Tarea	Descripción
Límites de servicio / Contención de costos	Aunque puede ser tentador conservar unos límites de servicio elevados después de un evento, tenga en cuenta que los límites de servicio también funcionan como una red de seguridad. Los límites de servicio protegen sus operaciones y los costos operativos al impedir un uso excesivo del servicio, ya sea a través de una actividad maliciosa derivada de una cuenta comprometida o a través de una automatización configurada erróneamente.
Procedimientos de reducción de escala	Revertir los recursos que se habían escalado durante la fase de preparación. Estos elementos son exclusivos de su arquitectura, pero los siguientes ejemplos son comunes: Tamaño de la instancia EC2/RDS Configuración de Auto Scaling Capacidad reservada IOPS provisionadas
Entorno de validación de estado	Compare las métricas de referencia y revise el estado de la producción para comprobar que una vez que el evento y la reducción de la escala de los procedimientos se han completado, los sistemas afectados muestran un comportamiento normal.
Disposición de los cambios de arquitectura	Puede valer la pena conservar algunos de los cambios realizados para preparar el evento, en función de la naturaleza del evento y del cumplimiento de las métricas operativas. Por ejemplo, la expansión a una nueva región geográfica puede requerir un aumento permanente de los recursos en dicha región o elevar determinados límites de servicio o parámetros de configuración, como el número de particiones en una base de datos o de fragmentos en una transmisión de PIOPS en un volumen, pueden ser una medida de ajuste del rendimiento que debe conservarse.

Optimización

Quizá el componente más importante de la gestión de un evento de infraestructura es el análisis tras el evento y la identificación de los retos de funcionamiento y arquitectura que se han observado, así como las oportunidades para mejorar. Los eventos de infraestructura no suelen suceder una sola vez. Pueden ser estacionales o coincidir con nuevas versiones de una aplicación, o bien pueden formar parte del crecimiento de la empresa a medida que se expande hacia nuevos mercados y territorios. Por lo tanto, todos los

eventos de infraestructura son una oportunidad para observar, mejorar y preparar con mayor eficacia el siguiente.

Conclusión

AWS proporciona bloques de construcción en forma de productos y servicios programables y elásticos que su empresa puede unir para dar soporte a prácticamente cualquier escala de carga de trabajo. Con las directrices y prácticas recomendadas para los eventos de infraestructura de AWS, junto con nuestro conjunto completo de servicios de alta disponibilidad, su empresa puede diseñar y prepararse para eventos importantes y asegurarse de que se cumplen las demandas de escalado sin problemas y de forma dinámica, de manera que se garantice una rápida respuesta y un alcance global.

Colaboradores

Las siguientes personas y organizaciones han participado en la redacción de este documento:

- Presley Acuna, AWS Enterprise Support Manager
- Kurt Gray, AWS Global Solutions Architect
- Michael Bozek, AWS Sr. Technical Account Manager
- Rovan Omar, AWS Technical Account Manager
- Will Badr, AWS Technical Account Manager
- Eric Blankenship, AWS Sr. Technical Account Manager
- Greg Bur, AWS Technical Account Manager
- Bill Hesse, AWS Sr. Technical Account Manager
- Hasan Khan, AWS Sr. Technical Account Manager
- Varun Bakshi, AWS Sr. Technical Account Manager

Documentación adicional

Para leer más sobre las prácticas recomendadas para el funcionamiento y la arquitectura, consulte Operational Checklists for AWS.[25] Le recomendamos a los lectores que consulten AWS Well Architected Framework[26] para ver un

documento bien estructurado para evaluar sus pilas de entrega de aplicaciones basadas en la nube. AWS ofrece Infrastructure Event Management (IEM) como una oferta de soporte premium para los clientes que deseen una implicación más directa de los ingenieros de soporte y del AWS Technical Account Manager en las operaciones de diseño, planificación y ejecución del evento. Para obtener más información sobre la oferta IEM de AWS Premium Support, consulte Infrastructure Event Management.[27]

Anexo

Lista detallada de comprobaciones de revisión de la arquitectura

Sí-No-N/A	Seguridad
☐-☐-☐	Rotamos nuestras claves de acceso de AWS Identity and Access Management (IAM) y la contraseña y de las credenciales del usuario en los recursos implicados en nuestra aplicación como máximo cada 3 meses de acuerdo con las prácticas recomendadas de AWS en materia de seguridad. Aplicamos la política de contraseñas en cada cuenta y usamos dispositivos de autenticación multifactorial virtual (MFA) o de hardware.
☐-☐-☐	Disponemos de procesos y controles de seguridad internos para controlar el acceso único, basado en funciones y de privilegios mínimos a las API de AWS aprovechando IAM.
☐-☐-☐	Hemos eliminado cualquier información de carácter confidencial o privado como instancias públicas/privadas de pares de claves integradas y hemos revisado todos los archivos de claves SSH autorizados de cualquier imagen de máquina de Amazon (AMI) personalizada.
☐-☐-☐	Utilizamos las funciones de IAM para instancias EC2 según sea necesario en lugar de incrustar las credenciales dentro de las AMI.
☐-☐-☐	Segregamos los privilegios administrativos de IAM a partir de los privilegios de usuario normales mediante la creación de la función administrativa IAM y la restricción de las acciones IAM de otros roles funcionales.
☐-☐-☐	Aplicamos los parches de seguridad más recientes en nuestras instancias EC2 para instancias de Windows o Linux. Utilizamos controles de acceso al sistema operativo como las normas del grupo de seguridad de Amazon EC2, las listas de control de acceso a la red VPC, el endurecimiento del sistema operativo, la protección del firewall basada en host, la prevención y detección de intrusiones, la monitorización, la configuración del software y el inventario de host.
☐-☐-☐	Garantizamos que la conectividad de red hacia y desde AWS de la organización y de los entornos corporativos utiliza un cifrado de protocolos de transporte.

Sí-No-N/A	Seguridad
☐-☐-☐	Aplicamos una solución de gestión de registro centralizado y auditoría para identificar y analizar cualquier patrón de acceso inusual o de ataques maliciosos en el entorno.
☐-☐-☐	Disponemos de gestión, correlación y procesos de generación de informes sobre eventos e incidentes de seguridad en vigor.
☐-☐-☐	Nos aseguramos de que no hay un acceso ilimitado a los recursos de AWS en cualquiera de nuestros grupos de seguridad.
☐-☐-☐	Utilizamos un protocolo de seguridad (HTTPS o SSL), políticas de seguridad actualizadas y protocolos de cifrado para las conexiones front-end (del cliente al balanceador de carga). Las solicitudes se cifran entre los clientes y el balanceador de carga, que es más seguro.
☐-☐-☐	Configuramos nuestro registro de recursos Amazon Route 53 MX para disponga de un registro de recursos TXT que contenga el valor de marco de directivas de remitente (SPF) correspondiente para especificar los servidores que están autorizados para enviar correo electrónico para nuestro dominio.

Sí-No-N/A	Fiabilidad
☐-☐-☐	Desplegamos nuestra aplicación en una flota de instancias EC2 que se implementan en un grupo de Auto Scaling para garantizar un escalado horizontal automático en función de los planes de escalado predefinidos. Más información.
☐-☐-☐	Utilizamos una comprobación de estado de Elastic Load Balancing en la configuración del grupo de Auto Scaling para garantizar que el grupo de Auto Scaling actúa en el estado de las instancias EC2 subyacentes. (Se aplica únicamente si utiliza los balanceadores de carga en los grupos de Auto Scaling).
☐-☐-☐	Desplegamos los componentes críticos de nuestras aplicaciones en varias zonas de disponibilidad, que replican debidamente los datos entre zonas. Comprobamos cómo afectan los errores dentro de estos componentes sobre la disponibilidad de la aplicación mediante Elastic Load Balancing, Amazon Route 53, o cualquier herramienta de terceros adecuada.
☐-☐-☐	En la capa de la base de datos implementamos nuestras instancias de Amazon RDS en varias zonas de disponibilidad para mejorar la disponibilidad de la base de datos mediante la replicación sincrónica a una instancia en reposo en una zona de disponibilidad distinta.
☐-☐-☐	Hemos definido los procesos para la conmutación por error automática o manual en caso de cualquier interrupción o disminución del rendimiento.
☐-☐-☐	Utilizamos registros de CNAME para asignar los nombres DNS a nuestros servicios. NO utilizamos registros A.
☐-☐-☐	Hemos configurado un valor inferior de tiempo de vida (TTL) para nuestro conjunto de registros de Amazon Route 53. Esto evita retrasos a la hora de que los sistemas de resolución DNS soliciten los registros DNS actualizados al volver a enrutar el tráfico. (Por ejemplo, esto puede producirse cuando la conmutación por error de DNS detecta y responde a un error de uno de los puntos de enlace).

Sí-No-N/A	Fiabilidad
☐-☐-☐	Disponemos de al menos dos túneles de VPN configurados para proporcionar redundancia en caso de interrupciones o mantenimiento previsto de los dispositivos en el punto de enlace de AWS.
☐-☐-☐	Utilizamos AWS Direct Connect, y tenemos dos conexiones de Direct Connect configuradas en todo momento para ofrecer redundancia en caso de que un dispositivo no esté disponible. Las conexiones se aprovisionan en diferentes ubicaciones de Direct Connect para proporcionar redundancia en caso de que una ubicación no esté disponible. También hemos configurado la conectividad a la gateway privada virtual para que tenga múltiples interfaces virtuales configuradas en varias conexiones y ubicaciones de Direct Connect.
☐-☐-☐	Utilizamos instancias de Windows y nos aseguramos de que utilizamos los controladores de PV más recientes. El controlador de PV ayuda a optimizar el rendimiento del mismo y a minimizar los problemas de tiempo de ejecución y de seguridad. También nos hemos asegurado de que ejecuta la última versión del agente EC2Config en nuestra instancia de Windows.
☐-☐-☐	Tomamos snapshots de nuestros volúmenes de Amazon Elastic Block Store (EBS) para garantizar una recuperación a un momento dado en caso de que se produzca algún error.
☐-☐-☐	Utilizamos volúmenes de Amazon EBS independientes para los datos del sistema operativo y de la aplicación/base de datos cuando procede.
☐-☐-☐	Aplicamos los parches de kernel, software y de controladores más recientes en cualquier instancia Linux.

Sí-No-N/A	Eficiencia del rendimiento
☐-☐-☐	Comprobamos completamente nuestros componentes de aplicaciones alojadas en AWS, incluidas las pruebas de rendimiento, antes de su lanzamiento. También realizamos pruebas de carga para garantizar que hemos utilizado el tamaño de instancia EC2 correcto, el número de IOPS, el tamaño de instancia de base de datos de RDS, etc.
☐-☐-☐	Ejecutamos un informe de comprobación de uso en comparación con nuestros límites de servicio y nos aseguramos de que el uso actual de los servicios de AWS es igual o inferior al 80 % de los límites del servicio. Más información
☐-☐-☐	Utilizamos una red de distribución/entrega de contenido (CDN) para utilizar el almacenamiento en caché de la aplicación (Amazon CloudFront) y como forma de optimizar la entrega del contenido y la distribución automática del contenido a la ubicación de borde más cercana al usuario.
☐-☐-☐	Somos conscientes de que algunos encabezados de solicitudes HTTP dinámicas que recibe Amazon CloudFront (usuario-agente, fecha, etc.) pueden influir en el rendimiento al reducir la tasa de impacto e incrementar la carga en el origen. Más información
☐-☐-☐	Garantizamos que el rendimiento máximo de una instancia EC2 es mayor que la suma del rendimiento máximo de los volúmenes de EBS adjuntos. También utilizamos instancias optimizadas para EBS con volúmenes de EBS PIOPS para obtener el rendimiento esperado fuera de los volúmenes.

Sí-No-N/A	Eficiencia del rendimiento
☐-☐-☐	Garantizamos que el diseño de la solución no tiene ningún cuello de botella en la infraestructura ni puntos de esfuerzo en la base de datos o en la aplicación.
☐-☐-☐	Implementamos la monitorización de los recursos de la aplicación y configuramos alarmas basadas en cualquier interrupción del rendimiento mediante Amazon CloudWatch o herramientas de socios de terceros.
☐-☐-☐	Nuestro diseño evita el uso de un gran número de normas en cualquier grupo de seguridad conectado a nuestras instancias de aplicación. Un gran número de normas en un grupo de seguridad puede dañar el rendimiento.

Sí-No-N/A	Optimización de costos
☐-☐-☐	Comprendemos que el evento de infraestructura puede implicar un cierto exceso en el aprovisionamiento de capacidad que deberá depurarse tras el evento para evitar costos innecesarios.
☐-☐-☐	Utilizamos el tamaño adecuado para todos los componentes de la infraestructura, incluido el tamaño de la instancia EC2, el tamaño de la instancia de base de datos de RDS, el tamaño y los números de los nodos de clúster de caché, el tamaño y los números de los nodos de clúster Redshift y el tamaño de volumen de EBS.
☐-☐-☐	Utilizamos instancias de subasta cuando resulta conveniente. Las instancias de subasta son ideales para cargas de trabajo que tienen un horario de inicio y finalización flexibles. Entre los casos de uso típicos para las instancias de subasta se encuentran: El procesamiento por lotes, la generación de informes y las cargas de trabajo de computación de alto rendimiento.
☐-☐-☐	Contamos con requisitos mínimos de capacidad de aplicación predecibles y aprovechamos las instancias reservadas. . Las instancias reservadas permiten reservar capacidad informática de Amazon EC2 a cambio de un descuento significativo en la tarifa por hora en comparación con los precios de las instancias bajo demanda.

Notes

[1] https://aws.amazon.com/answers/account-management/aws-tagging-strategies/

[2] https://aws.amazon.com/blogs/aws/resource-groups-and-tagging/

[3] https://aws.amazon.com/sqs/

[4] http://docs.aws.amazon.com/general/latest/gr/rande.html

[5] https://aws.amazon.com/emr/

[6] https://aws.amazon.com/rds/

[7] https://aws.amazon.com/ecs/

[8] https://aws.amazon.com/sns/

[9] https://aws.amazon.com/blogs/compute/using-aws-lambda-with-auto-scaling-lifecycle-hooks/

[10] http://docs.aws.amazon.com/lambda/latest/dg/welcome.html

[11] https://aws.amazon.com/blogs/aws/new-auto-recovery-for-amazon-ec2/

[12] https://aws.amazon.com/answers/configuration-management/aws-infrastructure-configuration-management/

[13] https://d0.awsstatic.com/whitepapers/Big_Data_Analytics_Options_on_AWS%20.pdf

[14] http://docs.aws.amazon.com/Route53/latest/DeveloperGuide/routing-policy.html#routing-policy-latency

[15] https://aws.amazon.com/elasticache/

[16] https://aws.amazon.com/cloudfront/

[17] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts-on-demand-reserved-instances.html

[18] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-spot-instances.html

[19] https://docs.aws.amazon.com/general/latest/gr/aws_service_limits.html

[20] https://aws.amazon.com/about-aws/whats-new/2014/07/31/aws-trusted-advisor-security-and-service-limits-checks-now-free/

[21] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/monitoring_automated_manual.html

[22] http://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/CloudWatch_Dashboards.html

[23] http://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/publishingMetrics.html

[24] https://aws.amazon.com/blogs/aws/new-whitepaper-use-aws-for-disaster-recovery/

[25] http://media.amazonwebservices.com/AWS_Operational_Checklists.pdf

[26] http://d0.awsstatic.com/whitepapers/architecture/AWS_Well-Architected_Framework.pdf

[27] https://aws.amazon.com/premiumsupport/iem/