2025

Fundamentos de la Computación en la Nube

MÓDULO 3

OLIVER BITICA – 1°DAM

Contenido

Leyenda de Colores

- Amarillo: Para conceptos fundamentales, definiciones y puntos clave.
- Verde: Para ventajas, objetivos, propósitos y características positivas.
- Azul: Para tipos, clasificaciones, componentes, estructuras y ejemplos.
- Rojo/Salmón: Para problemas, inconvenientes, limitaciones o advertencias.
- Gris: Para tecnologías específicas, nombres propios, estándares o secciones explícitamente marcadas como "Contenido Prioritario".

Principios y Arquitectura de la Infraestructura Cloud de Alta Disponibilidad: Un Análisis del Modelo de AWS

1.0 Fundamentos de la Alta Disponibilidad en Centros de Datos

1.1 Introducción a la Alta Disponibilidad (HA)

En el ecosistema digital actual, la continuidad operativa es un pilar estratégico. La alta disponibilidad (HA) se refiere a las soluciones diseñadas para que los sistemas de información de una organización estén operativos las 24 horas del día, los 7 días de la semana. Este enfoque garantiza que las aplicaciones y los datos críticos permanezcan accesibles para los usuarios y los procesos de negocio, minimizando las interrupciones que pueden generar pérdidas económicas y de reputación.

El principio de alta disponibilidad se define como la probabilidad de que un sistema funcione normalmente durante un periodo de tiempo establecido. En la práctica, el objetivo principal de una solución de alta disponibilidad es minimizar o mitigar el impacto del tiempo de inactividad. La diferencia entre un sistema con un 99% de disponibilidad y uno con un 99,999% puede parecer marginal, pero su impacto en la operación anual es drástico, como se ilustra a continuación.

Índice de probabilidad	Duración del tiempo de inactividad anual
99%	3 días, 15 horas
99,9%	8 horas, 45 minutos
99,99%	53 minutos
99,999%	5 minutos
99,9999%	32 segundos

Para diseñar sistemas que cumplan con estos exigentes niveles de servicio, es fundamental cuantificar los objetivos de resiliencia mediante dos indicadores clave que miden la recuperación de un sistema tras un incidente.

1.2 Métricas Clave de Recuperación: RTO y RPO

Para alinear las capacidades técnicas de una infraestructura con las necesidades del negocio, es crucial cuantificar los objetivos de recuperación ante desastres. Esto se logra a través de dos métricas fundamentales: RTO y RPO.

El Objetivo de Tiempo de Recuperación (RTO) representa la duración máxima aceptable de una interrupción del servicio. Mide el tiempo transcurrido desde que se produce un fallo hasta que el sistema vuelve a estar operativo. En una estrategia de recuperación por fases, el objetivo inicial puede ser una recuperación parcial (p. ej., reanudar en modo de solo lectura) para mitigar el impacto inmediato, pero el RTO final se mide por la restauración del servicio completo capaz de procesar nuevas transacciones.

El Objetivo de Punto de Recuperación (RPO) mide la pérdida de datos máxima aceptable. Esta métrica cuantifica el intervalo de tiempo entre la última transacción de datos confirmada antes del fallo y los datos más recientes recuperados después del mismo. Un RPO bajo (cercano a cero) implica una pérdida de datos mínima, pero suele requerir una arquitectura más compleja y costosa.

Para poder cumplir con estos exigentes objetivos de RTO y RPO, la industria ha desarrollado estándares que normalizan el diseño físico y la infraestructura de soporte de los centros de datos.

1.3 Estandarización del Diseño: La Norma TIA-942

La norma TIA-942, publicada por la Telecommunications Industry Association (TIA), surge como una respuesta a la necesidad de estandarizar el diseño de los centros de datos, proporcionando un marco de referencia para su construcción y operación.

El propósito principal de esta norma es establecer directrices claras para la infraestructura física, clasificando los centros de datos en diferentes niveles de resiliencia, denominados TIER (I, II, III y IV), donde cada nivel superior ofrece mayor disponibilidad y tolerancia a fallos. Según la norma, la infraestructura de soporte de un centro de datos se compone de cuatro subsistemas críticos:

- **Telecomunicaciones**: Cableado, conectividad y topología de red.
- **Arquitectura**: Diseño estructural, control de acceso y seguridad física.
- **Sistema eléctrico**: Suministro de energía, sistemas de alimentación ininterrumpida (UPS) y generadores.
- **Sistema mecánico**: Climatización, ventilación y sistemas de extinción de incendios.

Estos principios teóricos y estándares de diseño físico son la base sobre la cual se construyen las infraestructuras de nube a gran escala, como la de Amazon Web Services (AWS), que los implementa para ofrecer servicios resilientes a nivel global.

2.0 La Infraestructura Global de AWS como Modelo de Resiliencia

2.1 Visión General y Componentes Fundamentales

Tras establecer los imperativos teóricos de la alta disponibilidad—cuantificados por métricas como RTO y RPO y estandarizados por normativas como TIA-942—analizaremos cómo Amazon Web Services (AWS) traduce estos requisitos en una arquitectura física y lógica a escala planetaria. Su arquitectura se fundamenta en una jerarquía de componentes diseñados para ofrecer resiliencia, seguridad y rendimiento a clientes de todo el mundo.

Esta infraestructura se puede dividir en tres elementos principales que trabajan en conjunto para proporcionar una plataforma de nube robusta:

1. Regiones
2. Zonas de Disponibilidad
3. Puntos de Presencia (que incluyen ubicaciones perimetrales)

A continuación, analizaremos en detalle el primer y más amplio nivel de esta jerarquía: las Regiones de AWS.

2.2 Regiones: La Base Geográfica para la Soberanía y el Rendimiento

Una Región de AWS es una ubicación geográfica física y aislada en el mundo. Este aislamiento es la primera línea de defensa para garantizar la tolerancia a fallos a gran escala y la estabilidad de los servicios. Si un desastre natural o un problema técnico afecta a una Región, las demás permanecen operativas e independientes.

Características Clave de las Regiones:

- Cada Región contiene una o varias Zonas de Disponibilidad, que a su vez alojan los centros de datos.
- Los recursos no se replican automáticamente entre Regiones. Esta responsabilidad del cliente es fundamental para las estrategias de recuperación ante desastres (DR) a gran escala. Mientras que el uso de múltiples AZs protege contra fallos locales, la replicación entre Regiones es la salvaguarda contra eventos que podrían comprometer un área geográfica completa, permitiendo alcanzar los objetivos de RPO y RTO más exigentes a nivel de negocio.

- Las Regiones introducidas después del 20 de marzo de 2019 deben habilitarse manualmente en la cuenta del cliente para poder ser utilizadas.
- Existen regiones con acceso restringido, como AWS (China), que requiere una cuenta específica para las regiones de Pekín y Ningxia, y AWS GovCloud (EE. UU.), diseñada para que los organismos gubernamentales y clientes de EE. UU. puedan alojar cargas de trabajo confidenciales cumpliendo con sus requisitos normativos.

La selección de una Región es un ejercicio de optimización que debe ponderar cuatro factores críticos:

1. **Gobernanza de datos y requisitos legales**: Las leyes locales pueden exigir que ciertos datos permanezcan dentro de límites geográficos específicos. Un ejemplo claro es la Directiva de protección de datos de la Unión Europea (UE), que obliga a muchas organizaciones a procesar y almacenar datos personales dentro de los estados miembros.
2. **Proximidad y Latencia**: Para optimizar la experiencia del usuario, se recomienda ejecutar las aplicaciones en la Región más cercana posible a los usuarios finales. Esto reduce la latencia de red, es decir, el tiempo que tardan los datos en viajar desde el usuario hasta el servidor y viceversa.
3. **Disponibilidad de servicios**: No todos los servicios de AWS están disponibles en todas las Regiones. Es fundamental verificar que los servicios necesarios para una aplicación estén presentes en la Región seleccionada antes de iniciar el despliegue.
4. **Costo**: Finalmente, se debe realizar un análisis de costo-beneficio, ya que los precios de los servicios de AWS varían significativamente entre Regiones. Esta variación puede impactar el TCO (Costo Total de Propiedad) de la solución. Por ejemplo, en el momento en que se redactó la documentación de referencia, ejecutar una instancia t3.medium de Linux bajo demanda costaba 0,0416 USD/hora en la región EE. UU. Este (Ohio), mientras que la misma instancia en Asia-Pacífico (Tokio) costaba 0,0544 USD/hora.

Una vez seleccionada la Región, el siguiente nivel de la arquitectura, las Zonas de Disponibilidad, proporciona los mecanismos para construir aplicaciones de alta disponibilidad dentro de esa área geográfica.

2.3 Zonas de Disponibilidad (AZs): El Pilar de la Alta Disponibilidad

Las Zonas de Disponibilidad (AZs) son ubicaciones aisladas dentro de una Región de AWS. Desempeñan un papel crucial al permitir a los clientes operar aplicaciones y bases de datos con una mayor disponibilidad, tolerancia a fallos y escalabilidad de lo que sería posible en un único centro de datos.

Arquitectura y Características de las AZs:

- Cada AZ puede constar de uno o varios centros de datos, que a escala completa pueden albergar cientos de miles de servidores.
- Están físicamente separadas por muchos kilómetros de distancia para evitar que un mismo desastre afecte a varias AZs simultáneamente, pero se encuentran a menos de 100 km entre sí para garantizar una comunicación de baja latencia.
- Poseen su propia infraestructura de energía, refrigeración y seguridad física, y están conectadas mediante un tejido de red (network fabric) de fibra óptica dedicado, completamente redundante, que garantiza un alto ancho de banda y una latencia ultra-baja.
- Esta separación aísla y protege eficazmente las aplicaciones de desastres locales como incendios, tornados, terremotos o fallos de suministro eléctrico. Esta separación física deliberada es la implementación práctica que permite a las organizaciones diseñar arquitecturas con un Objetivo de Tiempo de Recuperación (RTO) bajo, ya que un fallo catastrófico en una AZ no detiene la operación en las otras.

Es responsabilidad del cliente diseñar arquitecturas que distribuyan los recursos entre múltiples AZs. AWS recomienda explícitamente esta práctica para lograr resiliencia y diseñar sistemas capaces de resistir fallos temporales o prolongados en una única Zona de Disponibilidad.

Dentro de cada AZ reside el núcleo físico de la infraestructura: los centros de datos.

2.4 Centros de Datos: El Núcleo Físico Seguro

Aunque los clientes no eligen centros de datos específicos para desplegar sus recursos —el nivel más granular de control es la Zona de Disponibilidad—, estos son la ubicación física donde residen los datos y la base de toda la infraestructura. Aunque es poco frecuente, pueden ocurrir errores que afecten la disponibilidad de las instancias que están en la misma ubicación. Si aloja todas las instancias en una misma ubicación y se produce un error en ella, ninguna de las instancias estará disponible. Por esta razón, Amazon opera centros de datos de vanguardia diseñados para la máxima seguridad y disponibilidad, agrupados en el constructo lógico de las AZs.

Características de Diseño y Seguridad:

- **Ubicación segura**: Las ubicaciones físicas de los centros de datos no se divulgan públicamente y el acceso está estrictamente restringido para mitigar riesgos tanto físicos como lógicos. Cada ubicación se evalúa cuidadosamente para minimizar el riesgo ambiental.
- **Diseño redundante**: Los componentes críticos del sistema, como la energía y la red, están respaldados en múltiples AZs para tolerar fallos mientras se mantienen los niveles de servicio acordados.
- **Capacidad gestionada**: AWS supervisa continuamente el uso de los servicios para implementar nueva infraestructura de manera proactiva,

garantizando que siempre haya capacidad disponible para satisfacer la demanda de los clientes.
- **Respuesta automatizada**: En caso de un fallo en un componente o sistema, existen procesos automatizados que desvían el tráfico de las zonas afectadas sin intervención humana, garantizando la continuidad del servicio.

Mientras que las Regiones y AZs conforman la infraestructura central, AWS también dispone de una red global de componentes diseñada para acelerar la entrega de contenido a usuarios de todo el mundo.

3.0 Optimización del Rendimiento: La Red Perimetral de AWS

3.1 Puntos de Presencia y la Red Perimetral

Mientras que las Regiones y Zonas de Disponibilidad resuelven el desafío de la resiliencia de la infraestructura central, la red perimetral de AWS aborda un factor igualmente crítico para el negocio: la latencia y la experiencia del usuario final. Esta red optimiza el rendimiento al acercar el contenido y los servicios a la ubicación geográfica de los usuarios mediante una red global de centros de datos más pequeños y distribuidos.

Esta red se compone de Puntos de Presencia (PoPs), que son infraestructuras distribuidas globalmente y que a su vez incluyen componentes como las ubicaciones perimetrales (para el caché de CloudFront) y las cachés de borde regionales. Al medir continuamente la conectividad a Internet y el rendimiento, pueden enrutar las solicitudes de los usuarios de la manera más eficiente posible. Varios servicios de AWS utilizan esta red para mejorar el rendimiento:

- **Amazon CloudFront**: Es la Red de Entrega de Contenido (CDN) de AWS, que almacena en caché copias de contenido (imágenes, vídeos, archivos estáticos) en ubicaciones perimetrales cercanas a los usuarios.
- **Amazon Route 53**: Es el servicio de Sistema de Nombres de Dominio (DNS) de AWS, que traduce los nombres de dominio en direcciones IP, enrutando las solicitudes a la ubicación perimetral más cercana para una resolución más rápida.
- **AWS Shield**: Un servicio de protección contra ataques de denegación de servicio distribuido (DDoS) que opera en el borde de la red.
- **AWS Web Application Firewall (AWS WAF)**: Un firewall de aplicaciones web que protege contra exploits comunes y opera también en las ubicaciones perimetrales.

Adicionalmente, las cachés de borde regionales se utilizan con Amazon CloudFront. Su función es almacenar contenido que no se accede con la frecuencia suficiente como para permanecer en una ubicación perimetral, pero que aun así es más

eficiente recuperarlo de esta caché regional que del servidor de origen, logrando un equilibrio entre costo y rendimiento.

4.0 Síntesis y Beneficios Clave del Modelo de Infraestructura de AWS

La arquitectura multicapa de AWS, que abarca Regiones, Zonas de Disponibilidad y una red perimetral global, se traduce en beneficios tangibles y estratégicos para las empresas que la utilizan, permitiéndoles construir aplicaciones seguras, escalables y altamente disponibles.

Los tres beneficios principales que se derivan de esta infraestructura global son:

1. **Elasticidad y Escalabilidad**: La infraestructura permite que los recursos se adapten dinámicamente a los cambios en la demanda, escalando hacia arriba o hacia abajo según sea necesario. Además, puede adaptarse rápidamente para admitir el crecimiento del negocio sin grandes inversiones iniciales en hardware.
2. **Tolerancia a Fallos**: Gracias a la redundancia integrada en todos sus componentes, desde los centros de datos hasta las Regiones, la infraestructura está diseñada para seguir funcionando incluso si una parte de ella falla.
3. **Alta Disponibilidad**: El diseño global permite construir sistemas que operan con una intervención humana mínima y un tiempo de inactividad mínimo, cumpliendo con los exigentes objetivos de disponibilidad que demandan las aplicaciones modernas.

Para consolidar el conocimiento, a continuación se presentan las conclusiones clave sobre la arquitectura de AWS:

- La infraestructura global de AWS se compone fundamentalmente de Regiones y Zonas de Disponibilidad.
- La elección de una Región se basa principalmente en el cumplimiento normativo (soberanía de datos) o en la necesidad de reducir la latencia para los usuarios finales.
- Cada Zona de Disponibilidad es una entidad físicamente separada con alimentación, redes y conectividad redundantes para garantizar el aislamiento de fallos.
- Las ubicaciones perimetrales mejoran el rendimiento global al almacenar contenido en caché cerca de los usuarios, reduciendo drásticamente la latencia.

En conclusión, el diseño deliberado y jerárquico de la infraestructura global de AWS es la respuesta de ingeniería a los requerimientos teóricos de la alta disponibilidad. Su arquitectura de Regiones aisladas y Zonas de Disponibilidad redundantes implementa a escala planetaria los principios de resiliencia cuantificados por

métricas como RTO y RPO y estandarizados por normativas como TIA-942. Este modelo no es accidental, sino una solución diseñada para consolidar a AWS como un pilar de la computación en la nube resiliente y de alto rendimiento.