Lectura 4.3: Optimización de soluciones en AWS

¿Qué es la disponibilidad?

La disponibilidad de un sistema generalmente se expresa como un porcentaje de tiempo de actividad en un año determinado o como un número de nueves. A continuación, puede ver una lista de los porcentajes de disponibilidad en función del tiempo de inactividad por año, así como su notación en nueves.

Disponibilidad (%)	Tiempo de inactividad (por año)
90% ("uno nueve")	36,53 días
99% ("dos nueves")	3,65 días
99,9% ("tres nueves")	8.77 horas
99,95% ("tres nueves y medio")	4.38 horas
99,99% ("cuatro nueves")	52,60 minutos
99,995% ("cuatro nueves y medio")	26.30 minutos
99,999% ("cinco nueves")	5,26 minutos

Para aumentar la disponibilidad, necesita redundancia. Esto generalmente significa más infraestructura: más centros de datos, más servidores, más bases de datos y más replicación de datos. Puede imaginar que agregar más de esta infraestructura significa un costo más alto. Los clientes quieren que la aplicación esté siempre disponible, pero debe trazar una línea en la que agregar redundancia ya no sea viable en términos de ingresos.

Mejore la disponibilidad de la aplicación

En la aplicación actual, solo se usa una instancia EC2 para alojar la aplicación, las fotos se sirven desde Amazon Simple Storage Service (S3) y los datos estructurados se almacenan en Amazon DynamoDB. Esa única instancia EC2 es un único punto de falla para la aplicación. Incluso si la base de datos y S3 tienen alta disponibilidad, los clientes no tienen forma de conectarse si la única instancia deja de estar disponible. Una forma de resolver este problema de punto único de falla es agregar un servidor más.

Usar una segunda zona de disponibilidad

La ubicación física de ese servidor es importante. Además de tener problemas de software a nivel del sistema operativo o de la aplicación, puede haber un problema de hardware. Podría estar en el servidor físico, el bastidor, el centro de datos o incluso la zona de disponibilidad que aloja la máquina virtual. Una forma fácil de solucionar el problema de la ubicación física es implementar una segunda instancia EC2 en una zona de disponibilidad diferente. Eso también resolvería problemas con el sistema operativo y la aplicación. Sin embargo, tener más de una instancia trae nuevos desafíos.

Administre la replicación, la redirección y la alta disponibilidad

Crear un proceso para la replicación El primer desafío es que necesita crear un proceso para replicar los archivos de configuración, los parches de software y la aplicación en sí misma en todas las instancias. El mejor método es automatizar donde puedas.

Redirección de direcciones de clientes El segundo desafío es cómo permitir que los clientes, las computadoras que envían solicitudes a su servidor, conozcan los diferentes servidores. Hay diferentes herramientas que se pueden utilizar aquí. El más común es usar un Sistema de Nombres de Dominio (DNS) donde el cliente usa un registro que apunta a la dirección IP de todos los servidores disponibles. Sin embargo, el tiempo que lleva actualizar esa lista de direcciones IP y que los clientes se den cuenta de dicho cambio, a veces llamado propagación, suele ser la razón por la que este método no siempre se usa.

Otra opción es utilizar un equilibrador de carga que se encargue de las comprobaciones de estado y distribuya la carga en cada servidor. Al estar entre el cliente y el servidor, el balanceador de carga evita problemas de tiempo de propagación. Hablaremos de los balanceadores de carga más adelante.

Comprenda los tipos de alta disponibilidad El último desafío que debe abordar cuando tiene más de un servidor es el tipo de disponibilidad que necesita, ya sea un sistema activo-pasivo o activo-activo.

- *Activo-Pasivo:* Con un sistema activo-pasivo, solo una de las dos instancias está disponible a la vez. Una ventaja de este método es que para las aplicaciones con estado donde los datos sobre la sesión del cliente se almacenan en el servidor, no habrá ningún problema ya que los clientes siempre se envían al mismo servidor donde se almacena su sesión.
- *Activo-Activo:* Una desventaja del activo-pasivo y donde brilla un sistema activo-activo es la escalabilidad. Al tener ambos servidores disponibles, el segundo servidor puede soportar algo de carga para la aplicación, lo que permite que todo el sistema soporte más carga. Sin embargo, si la aplicación tiene estado, habría un problema si la sesión del cliente no está disponible en ambos servidores. Las aplicaciones sin estado funcionan mejor para sistemas activo-activo.

Recursos

- Sitio externo: alta disponibilidad y escalabilidad en AWS

- Sitio externo: alta disponibilidad y escalabilidad en AWS
- *Sitio externo:* AWS: Pilar de confiabilidad de AWS: marco de buena arquitectura de AWS

Marcar como completo

Me gusta No me gusta Informar de un problema