Certified Tech Developer
The Ultimate Degree

Infraestructura I

Escalabilidad

La escalabilidad se refiere a la capacidad de los recursos para aumentar o disminuir en tamaño o cantidad. Hay mucha infraestructura involucrada para hacer que algo así suceda, por lo que no es una tarea fácil. Justamente, muchos de los servicios en AWS son escalables de manera predeterminada. Esta es una de las razones por las que AWS es tan exitoso. La escalabilidad es bastante simple de definir, por lo que a menudo se le atribuyen algunos de los aspectos de la elasticidad.

Elasticidad

La elasticidad en cloud computing se refiere a la capacidad de incrementar la infraestructura y recursos de los que se dispone en la nube según las necesidades de la empresa —como así también reducirlas cuando ya no se requieran—. Es decir, que es la capacidad de los recursos para escalar en respuesta a los criterios establecidos. Esto es lo que sucede cuando un equilibrador de carga agrega instancias cada vez que una aplicación web recibe mucho tráfico. No todos los servicios de AWS admiten elasticidad, e incluso aquellos que sí a menudo necesitan configurarse de cierta manera.

¡Escalabilidad es necesaria para la elasticidad, pero no al revés!

Escalabilidad	Elasticidad
Capacidad de un sistema para **aumentar** la carga de trabajo en sus recursos de hardware actuales (escalar).	Capacidad de un sistema para aumentar la carga de trabajo en sus recursos de hardware actuales y adicionales —**agregados dinámicamente bajo demanda**— (escalar).
Aumento de la capacidad para cumplir con la carga de trabajo **creciente**.	Aumento o reducción de la capacidad para cumplir con la carga de trabajo.
En un entorno de escala, los recursos disponibles **pueden exceder** para satisfacer las **demandas futuras**.	En el entorno elástico, los recursos disponibles coinciden con las **demandas actuales** lo más cerca posible.
La escalabilidad se adapta solo al **aumento de la carga de trabajo** al **aprovisionar** los recursos de manera **incremental**.	La elasticidad se adapta tanto al **aumento de la carga de trabajo** como a la **disminución**, al aprovisionar y desaprovisionar recursos de manera autónoma.
El **aumento de la carga** de trabajo sirve para aumentar el poder de un solo recurso informático o para aumentar el poder de un grupo de recursos informáticos.	La carga de **trabajo variable** se sirve con variaciones dinámicas en el uso de los recursos de la computadora.
La escalabilidad permite a una empresa satisfacer las demandas esperadas de servicios con **necesidades**	Elasticidad permite a una empresa satisfacer cambios inesperados en la demanda de servicios con

estratégicas a largo plazo.	**necesidades tácticas a corto plazo**.
Está **aumentando** la capacidad de servir a un entorno donde la carga de trabajo está aumentando.	Es la capacidad de **aumentar o disminuir** la capacidad de servir a voluntad.

Elasticidad es la capacidad de un sistema para aumentar (o disminuir) su capacidad de cómputo, almacenamiento, trabajo neto, etc. Por ejemplo, se puede implementar un sistema de fondo que inicialmente tiene un servidor en su clúster, pero configurarlo para agregar una instancia adicional al clúster si la utilización promedio de CPU por minuto de todos los servidores en el clúster excede un umbral determinado (por ejemplo 70%). Del mismo modo, se puede configurar un sistema para eliminar servidores del clúster de fondo si la carga en el sistema disminuye y la utilización promedio de CPU por minuto cae por debajo de un umbral definido (por ejemplo 30%).

Veamos otro ejemplo. Se puede configurar un sistema para aumentar el espacio total en disco de su clúster back end en un orden de 2 si se utiliza más del 80% del almacenamiento total disponible actualmente. Si por alguna razón, en un momento posterior, los datos se eliminan del almacenamiento y, por ejemplo, el almacenamiento total utilizado desciende por debajo del 20%, se puede disminuir el espacio total disponible en el disco a su valor original.

Sin embargo, algunos sistemas (por ejemplo software heredado) no están distribuidos y tal vez solo pueden usar un único núcleo de CPU. Entonces, aunque necesitemos aumentar la capacidad de cómputo disponible a pedido, el sistema no puede usar esta capacidad adicional de ninguna forma. Dichos sistemas son no escalables.

Conclusión

Un sistema escalable no depende de la elasticidad. Tradicionalmente, los departamentos de TI podían reemplazar sus servidores existentes con servidores más nuevos que tenían más CPU, RAM y almacenamiento, y transferir el sistema al nuevo hardware para emplear la capacidad de cómputo adicional disponible. Los entornos en la nube (AWS, Azure, Google Cloud, etc.) ofrecen elasticidad y algunos de sus servicios principales también son escalables desde el primer momento. Además, si creamos un software escalable, podemos implementarlo en estos entornos de nube y beneficiarnos de la infraestructura elástica que proporcionan para aumentar/disminuir automáticamente los recursos de cómputo disponibles.