## Conceptos previos

### Escalabilidad

Escalabilidad es la propiedad de un-Sistema o aplicación para manejar cargas de trabajo cada vez mayores según la demanda.

Existen 2 tipos de Escalabilidad:

* **Horizontal (=elasticidad)**
* **Vertical**

La Escalabilidad está relacionada pero no es lo mismo que la Alta Disponibilidad de un Sistema.

### Escalabilidad Vertical

Aplicado al concepto de instancias en la nube, la Escalabilidad vertical significaría un aumento en las capacidades de la instancia (mayor procesamiento, mayor memoria, mayor almacenamiento, etc)

* La Escalabilidad vertical es muy común para sistemas no distribuidos, por ejemplo, una base de datos.
* Existe un límite para la Escalabilidad vertical relacionado a límites de hardware

### Escalabilidad Horizontal

Aplicado al concepto de instancias en la nube, la Escalabilidad horizontal significaría un aumento en la cantidad de instancias procesando una carga de trabajo, o corriendo una aplicación/Sistema.

* La Escalabilidad horizontal implica sistemas distribuidos
* Este tipo de escalabilidad es más fácil de lograr gracias a la nube.
* Auto Scaling Group y Load Balancer

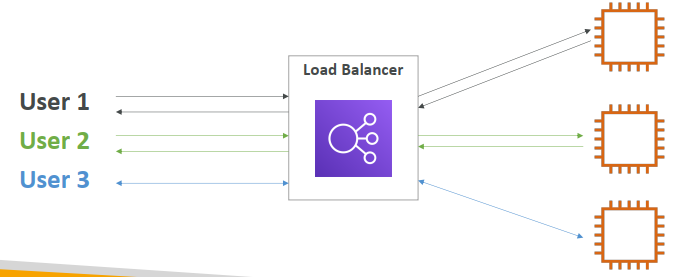
### Alta Disponibilidad

* Usualmente va de la mano con la escalabilidad horizontal
* La Alta disponibilidad usualmente implica alojar mi aplicación en al menos 2 zonas de disponibilidad
* Auto Scaling Group Multi AZ
* Load Balancer Multi AZ

### Escalabilidad vs Elasticidad vs Agilidad

* Escalabilidad: Es la habilidad de procesar una carga grande mediante el aumento de capacidades del hardware (scale up) o mediante el aumento de nodos de procesamiento (scale out).
* Elasticidad: Cuando un sistema es elástico implica que habrá un auto escalado y por lo tanto, el sistema escalará en base a la carga. Esto es “cloud-friendly”: pago por uso, coincide con la demanda, optimiza los costos.
* Agilidad: Habilidad de reducir el tiempo en tener disponibles recursos de IT necesarios para los desarrolladores.

# Elastic Load Balancing (ELB)

Son servidores que redirigen el tráfico de internet entrante, a múltiples servidores (EC2 instances) aguas abajo.

## ¿Por qué utilizar un Load Balancer?

* Distribuye la carga equitativamente a diferentes instancias aguas abajo
* Expone al Internet un solo punto de acceso (DNS) para mi aplicación
* Puede manejar los fallos de las instancias a su cargo
* Realiza chequeos de salud periódicamente a las instancias para comprobar su disponibilidad
* Provee terminación SSL/TLC (HTTPS) para mi sitio web
* Se puede utilizar un mismo Load Balancer a través de múltiples zonas de disponibilidad

### ¿Por qué utilizar un Elastic Load Balancer?

* Un Elastic Load Balancer (ELB) es un Load Balancer administrado por AWS
  + AWS garantiza que estará funcionando
  + AWS se hace cargo de las actualizaciones, mantenimiento y alta disponibilidad
  + AWS solo provee unas pocas configuraciones, por lo que nosotros debemos proveer configuraciones extra para el comportamiento del ELB
* Es mas costoso que armar nuestro propio LB en EC2, lo cual es posible, pero implica un mayor uso de nuestro tiempo en mantenimiento, integraciones, actualizaciones, etc
* AWS ofrece 4 tipos de Load Balancer
  + Application Load Balancer (solo para HTTP y HTTPS) – Layer 7
  + Network Load Balancer (ultra-high performance, para TCP) – Layer 4
  + Classic Load Balancer – Obsoleto
  + Gateway Load Balancer