# Autoescalado

## KUBERNETES AL COMPLETO

# Kubernetes al completo



- **El auto escalado** significa qué kubernetes puede aumentar o reducir los recursos de forma automática, dependiendo de la carga de trabajo que tengamos en la plataforma
- Dentro de kubernetes se pueden utilizar 3 tipos de auto escalado :
  - HPA- (horizontal POD AutoScaler)
  - VPA-(vertical POD AutoScaler)
  - CA-(Cluster AutoScaler)

# Kubernetes al completo



- HPA- (horizontal pod AutoScaler)
  - Esta opción escala el número de pods que tenemos de un determinado despliegue
  - Es gestionado por el propio Controller manager
  - En cada bucle el Controller compara el uso actual de los recursos con las métricas definidas para cada HP . Estas métricas pueden ser utilizadas por un metric Server o bien se pueden obtener directamente desde los PODS
  - Los pods deberían de tener configurados los requests de recursos

# Kubernetes al completo



- VPA-(vertical post AutoScaler)
  - Esta opción escala de forma vertical , es decir se asigna más recursos a los PODS existentes,  por ejemplo memoria y CPU.
  - También utiliza métricas para determinar si es necesaria este tipo de escalada
  - Solo disponible en determinadas  plataformas
  - Hay un proyecto para implementarlo en un cluster on-premise

https://github.com/kubernetes/autoscaler/tree/master/vertical-pod-autoscaler#installation

  - No deberíamos de utilizarlo tampoco con HPA

# Kubernetes al completo



- CA-(Cluster AutoScaler)
  - Con esta característica podemos aumentar el número de nodos en el clúster sí los nodos existentes no son capaces de afrontar el número de posts que se le solicitan.
  - Esta opción funciona básicamente en entornos Cloud donde se solicita al proveedor nuevas máquinas según nuestras necesidades
  - Podemos utilizar mecanismos manuales para hacer este trabajo pero son realmente complejos en la mayoría de los casos