



- 1 It scales your infrastructure whenever needed.
- Automatically increase and decrease the size of auto scaling group (mainly EC2 instances)
- Adds new instances when load reaches a certain level and spin down instances when no longer needed.



- It automatically routes incoming application traffic across dynamically changing number of multiple EC2 instances.
  - It acts as a single point of contact for all incoming traffic to the instances in your auto scaling group
- It used to monitor the health of the registered instances so that the load balancer send the requested traffic to healthy instance.