Amazon EC2 Auto Scaling

❖ AMAZON EC2-AUTO-SCALING:

- Auto Scaling helps you maintain application availability and allows you to scale your Amazon EC2 capacity up or down automatically according to conditions you define.
- Auto scaling will increase or decrease the number of instances based on chosen CloudWatch metrics.
- Dynamic scaling responds to changing demand and predictive scaling automatically schedules the right number of EC2 instances based on predicted demand.
- Dynamic scaling and predictive scaling can be used together to scale faster.
- Automatic tracking and maintaining instance pool.
- Autoscaling has two components.

FEATURES AND BENEFITS:

- Improve fault tolerance
- Increase application availability
- Automatically scale in and out
- Choose when and how to scale
- Fleet management
- Scheduled scaling
- Dynamic scaling

➢ LAUNCH TEMPLATE / CONFIGURATION:

- A launch configuration is an instance configuration template that an Auto Scaling group uses to launch EC2 instances. When you create a launch configuration, you specify information for the instances.
- Templates can have multiple versions.

➢ AUTO SCALING GROUP:

- An Auto Scaling group contains a collection of Amazon EC2 instances that are treated as a logical grouping for the purposes of automatic scaling and management.

- An Auto Scaling group also enables you to use Amazon EC2 Auto Scaling features such as health check replacements and scaling policies.
- The size of an Auto Scaling group depends on the number of instances that you set as the desired capacity.
- For example, the following Auto Scaling group has a minimum size of one instance, a desired capacity of two instances, and a maximum size of four instances. The scaling policies that you define adjust the number of instances, within your minimum and maximum number of instances, based on the criteria that you specify.

➢ EC2 VERTICAL SCALING & HORIZONTAL SCALING:

VERTICAL SCALING:

- With vertical scaling, the solution automatically adjusts capacity to maintain steady, predictable performance at the lowest possible cost.
- The solution can resize your instances by restarting your existing instance.
- It means Increase instance size (= scale up / down)

From: t2.micro - 1G RAM, 1 vCPU → **To:** t2.small – 2GB RAM, 1 vCPUs

HORIZONTAL SCALING:

- A "horizontally scalable" system is one that can increase capacity by adding more computers to the system.
- Horizontally scalable systems are oftentimes able to outperform vertically scalable systems by enabling parallel execution of workloads and distributing those across many different computers.
- It means Increases number of instances (= scale out / in)
  - Auto Scaling Group
  - Load Balancer

HIGH AVAILABILITY (MULTI AZ'S):

- Run instances for the same application across multiple-AZ's.

Elastic Load Balancing: you can launch several EC2 instances and distribute traffic between them.

Auto Scaling: use auto-scaling to detect when loads increase, and then dynamically add more instances.