Autoscaling In AWS

Creating group of EC2 instances that can scale up or down depending on conditions you set.

→ Enable elasticity by scaling Horizontally through adding or terminating EC2 instance.

→ Autoscaling ensures that you have the right number of AWS EC2 instances for your needs at all time.

→ Autoscaling helps you save cost by cutting down the number of EC2 instances when not needed, and scaling out to add more instance only when it is required.

Autoscaling Components

① Launch Configuration :→ like instance type, AMI, keypair Security group.

② Autoscaling group :→

Group Name, Group size, VPC, Subnet, Health check period.

③ Scaling policy.

Metric type (cpu utilization), Target value.

→ If Autoscaling finds that the number of EC2 instance launched by ASG into subjects AZs is not balanced (EC2 instances are not evenly distributed) Autoscaling do Rebalancing Activity by itself.

→ Autoscaling always tries to balance the Instance distribution across AZs.

→ while Rebalancing, ASG launches new EC2 instances, where there are less EC2 at present and then terminates the instance from the AZ, that had more instances.

What causes Imbalance of EC2

→ If we add or Remove some subnets/AZ from Autoscaling group.

→ If we manually Request for EC2 termination from our ASG.

→ An AZ that didnot have enough EC2 Capacity new has enough capacity & it is one of our Autoscaling group AZ.

→ we can attach a Running EC2 instances to an Autoscaling group by using AWS console or CLI, if the below conditions are met —

→ Instances must be in Running state.
(Not terminated or stopped)

→ AMI used to launch the Ec2 still Exist.

→ Instance is not part of another Autoscaling group.

→ Instance is in the same AZ of the same group.

→ If the existing Ec2 instances under the Autoscaling group, plus the one to be needed, exceed the max. Capacity of the Autoscaling group,. the Request will fail, Ec2 instance would not be added.

→ you can manually remove EC2 instances from an autoscaling group using AWS console or CLI.

→ you can then manage the detached instances independently or attach it to another Autoscaling group.

→ when you detach an Instance, you have the option to decrement the Autoscaling group desired capacity.

→ If you do not, the autoscaling group will launch another instance to Replace the one detached.

when you delete an autoscaling group, its parameters Like maximum, minimum and desired Capacity are all set to zero. Hence, it terminates all its EC2 instances.

→ If you want to keep the EC2 instances and manage them independently, you can manually detach them first, then delete the ASG.

→ we can attach one or more elastic load Balancer to our ~~Aus~~ ASG.

→ The elastic load Balancer must be in the same Region as ASG.

→ Once you do this, any EC2 instance existing or added by the Autoscaling group will be automatically registered with the ASG defined ELB.

→ you do not need to Register those instances manually on the Autoscaling group defined ELB.

→ Instance and ~~the~~ the ELB must be in the same vpc.

→ Autoscaling classifies its EC2 instance health status as either ~~e~~ healthy or unhealthy.

→ By default, AS uses EC2 status checks only to determine the health status of an Instance.

→ when you have one or more ELB defined with the ASG, you can configure Autoscaling ~~grou~~ to use 'both' the EC2 health check. & the ELB health check to determine the instance health check.

→ Health check grace period in 300sec by default

→ If we set 'zero' in Grace period, the instance health in checked once it is in Service.

→ until the grace period timer Expires any unhealthy status Reported by EC2 status checks, or the ELB attached to the autoscaling group, will not be acted upon.

→ After grace period expires, Autoscaling group would consider an Instance unhealthy in any of the following Cases ~~e~~ :-

→ EC2 status check report to autoscaling group an instance status other than Running

→ If ELB health check are configured to be used by the Autoscaling, then if the ELB Report the Instance as 'Out of service'.

→ Unlike AZ Rebalancing, termination of unhealthy instance happen first, then Autoscaling attempt to launch new instance to Replace the ones terminated.

→ Elastic IP and EBS volumes gets detached from the terminated instances. you need to manually attach then to the new instances.

In four situation, ASG sends a SNS email notifigation:-

① An Instance is launched.

② An Instance is terminated

③ An instance fails to launch

④ An Instance ~~fails~~ fails to terminate.

Merging Autoscaling group —

→ Can only be done from the CLI (not AWS console)

→ you can merge multiple, single, on multi-AZ autoscaling group.

→ Scale-out means launching more EC2 instances.

→ scale-in means terminating one or more EC2 instances by scaling Policy.

→ It is always Recommended to create a scale-in event for each scale-out event you create.

→ AWS EC2 services sends EC2 metrics to cloudwatch about the ASG instances.

→ Basic monitoring (every 300 sec) enabled by default & free of cost)

→ you can enabled detailed (every 60 sec - chargeable)

→ when the launch configuration is done by AWS CLI, detailed monitoring for EC2 instances in enabled by default.

Standby state

→ you can manually move an instance from an ASG and put it in standby state.

→ Instances in standby state are still managed by Autoscaling.

→ Instances in standby state are still managed by Autoscaling.

→ Instances in standby state are charged as Normal, in service instance

→ They do not count towards available EC2 instances for workload /App use.

→ Autoscaling does not perform health check on instances in standby state.

Scaling Policies

→ Define how much you want to scale based on defined conditions.

→ Autoscaling group uses alarms and policies to determine scaling.

→ For simple or step scaling, a scaling adjustment can't change the capacity of the group above the max. group size or Below the min group size.

Predictive Scaling → It looks at historic pattern & forecast them into the future to schedule change in the no. of EC2 instances. It uses Machine learning model to forecast daily & weekly pattern.

Target Tracking policies :→ Increase or decrease the current capacity of the group based on a target value for a Specific Metric. This is similar to the way that your thermostat mantain the temp. of your home.

Step Scaling :→ Increase or decrease the current capacity of the group based on set of scaling adjustment known as step adjustment, that vary based on the size of the alarm Breach.

→ Does not support/wait for a cool-down times.

→ Support a warm-up timer Time taken by newly launch instances to be ready & contribute to the watched Metric.

Simple Scaling :→

→ single adjustment (up or down) in response to an alarm (cooldown timer 300 sec default)

Schedule Scaling :→

→ use for predictable load change.

→ you need to Configure a schedule action for a scale out at a specific date/time and to a Required Capacity.

→ A schedule action must have a unique date/time you cannot config two schedule activities at the same time/data.