AWS AUTO SCALING

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

- Enable elasticity by scaling horizontally through adding or terminating EC2 instances.
- Auto scaling ensures that you have the right number of AWS EC2 instances for your needs at all time.
- Auto scaling helps you to save cost by cutting down the number of EC2 instances when not needed and scaling out to add more instances only it is required.

Auto Scaling Components:

1. Launch Configuration : like instance type, AMI, key-pair, security group
2. Auto Scaling Group: group name, group size, VPC, subnet, health check period
3. Scaling Policy : metric type, target value

How to Balance, Attach and Detach EC2 Instances:

Balance:

- If auto scaling finds that the number of EC2 instances launched by ASG into subjects AZs is not balanced (EC2 instances are not evenly distributed), auto scaling do rebalancing activity by itself.
- AS always tries to balance the instances distribution across AZs.
- While rebalancing, ASG launches new EC2 instances where there are less EC2 at present and then terminates the instances from the AZs that had more instances.

What causes imbalance of EC2?

- If we add or remove same subnets/AZ form auto scaling.
- If we manually request for EC2 termination from our ASG.
- An AZ that did not have enough EC2 capacity now has enough capacity and it is one of the auto scaling group.

Attach:

- We can attach a running EC2 instance to an ASG by using AWS console or CLI if the below conditions are meet:
  - Instances must be on running state.
  - AMI used to launch the EC2 still exists.
  - Instances is not the part of another auto scaling group.
  - Instances must be in the same AZ of the same group.
  - If the existing EC2 instances under the ASG, plus the one to be needed, exceeds the maximum capacity of the ASG, the request will fail, EC2 instance would not be added.

Detach:

- You can manually remove EC2 instances from an ASG using AWS console of CLI.
- You can then manage the detached instances independently or attach it to another ASG.
- When you detach an instance you have the option to decrement the ASG desired capacity.
- If you do not, ASG will launch another instance to replace the one detached.
- When you delete an ASG its parameter like maximum, minimum and desired capacity are all set to zero. Hence it terminates it’s all EC2 instances.
- If you want to keep the EC2 instances and manage them independently you can manually detach them first then delete ASG.
- We can attach one more Elastic Load Balancer (ELB) to our ASG.
- The ELB must be in the same region as the ASG.
- Once you do this any EC2 instance existing or added by ASG will be automatically registered with the ASG defined ELB.
- Instances and the ELB must be in the same VPC.
- Auto scaling classifies its EC2 instance health check 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 auto scaling to use both the EC2 status check and SLB health check to determine the instances health check.
- Health check grace period is 300sec by default.
- If we set zero in grace period the instance health is checked once it is in service.
- Until the grace period timer expires any unhealthy status reported by EC22 status check of the ELB attached to the ASG will not be acted upon.
- After grace period expires ASG consider an instance unhealthy in any of the following cases:
  - EC2 status check report to ASG an instance other than running.
  - If ELB health check are configured to be used by the auto scaling then if the ELB report the instance as ‘out of service’.
- Unlike AZ rebalancing, termination of unhealthy instances happen first then auto scaling 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 there to the new instance.

Types of Auto Scaling Policies:

In four situations, ASG sends a SNS email notification:

i. an instance is launched
ii. an instance is terminated
iii. an instance fails to launch
iv. an instance fails to terminate

Merging ASG:

- Can only be done form the CLI not form the AWS console.
- You can merge multiple single AZ ASG into a single, one multi-AZ ASG.
- 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 ASG instances.
- Basic monitoring is every 300sec enabled by default and free of cost.
- You can enabled detailed every 60sec which is chargeable.
- When the launch configuration is done by AWS CLI, detailed monitoring for EC2 instances is enabled by default.

StandBy State:

- You can manually move an EC2 instance form an ASG and put it in standby state.
- Instances in standby state are still managed by auto scaling.
- Instances in standby state are charged as normal in service instances.
- They do not count towards available EC2 instances for workload/app use.
- Auto scaling does not perform health check on instance in standby state.

Scaling Policies:

Generally scaling policy is of two types such as:

1. Manual
2. Dynamic

Again dynamic policy is divided into three categories as follows:

A. Target Tracking
B. Simple Scaling Policy
C. Step Scaling Policy

- Define how much you want to scale based on defined conditions.
- ASG 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 maximum group or below the minimum group.

Predictive/Scheduled/Cycle Scaling: it looks at historic pattern and forecast them into the future to schedule change in the number of EC2 instances. It uses machine learning model to forecast daily and weekly pattern.

Target Tracking Policies: increase or decrease the current capacity of the group based on a target value for specific metric. This is similar to the way that your thermostatic maintain the temperature of your home.

Step Scaling: increase or decrease the current capacity of the group based on a set of scaling adjustment known as step adjustment that vary based on the size of the alarm breach. It does not support/ wait for cool down times. It supports warm-up timer: time taken by newly launched instance to be ready and contribute to the watched metric.

Simple Scaling: single adjustment (up or down) in response to an alarm (cool down timer-300sec by default)

Schedule Scaling: used 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 scheduled action must have a unique data/time. You cannot configure two schedule activities at the same date/time.

Written By:

Nagarjuna Hota

LinkedIn: https://www.linkedin.com/in/nagarjuna-hota-30871017a/

Following Channel for AWS Series: Technical Guftgu

Channel Link: https://www.youtube.com/@TechnicalGuftgu

AWS Playlist: https://www.youtube.com/playlist?list=PLBGx66SQNZ8a_y_CMLHchyHz_R6-6i-i_