Auto Scaling

==============

Outline:

1. What is Auto Scaling (Horizontal and Vertical)

2. Scale-out and Scale-in

3. Amazon EC2 Auto Scaling

4. Benefits of Auto Scaling

-----------------------------------------------------------------------

1. What is Auto Scaling (Horizontal and Vertical)

Scaling --> scaling is nothing, but we can increase and decrease as per our need. it is simply known as "scaling".

==> In AWS we can increase and decrease the resources depends upon the requirement. when the demand is very high at the time we can increase, and the demand is low we can decrease the number of resources. it is simply known as "Scaling".

Scaling has two types 1) Horizontal Scaling and 2) Vertical Scaling

Horizontal Scaling --> as per the horizontal scaling it will add more servers when demand increases.

Vertical Scaling --> In vertical Scaling it will increase the size of the server when demand increases (like RAM, CPU).

Horizontal Scaling has two terms, 1) Scale-Out and 2) Scale-In

--> Scale-out: It is Horizontal scaling where more instances are added when the demand increases.

--> Scale-In: It is scaling where the instances are removed when the demand decreases.

----------------------------------------------------------------------------------

What is Auto Scaling?

--> In AWS the resources are increase and decrease automatically depends upon the requirement. when the demand is very high at the time it will increase, and the demand is low it will decrease the number of resources automatically. it is simply known as "Auto Scaling".

--> In Simply Auto Scaling allows you to scale our amazon EC2 instances increase and decrease automatically as per the instructions set by user.

--> Amazon EC2 Auto Scaling helps you ensure that you have the correct number of Amazon EC2 instances available to handle the load for our application.

A picture containing text, screenshot, font

Description automatically generated

🡪 Initially we can set up the 1 instance for our application because we don’t have any demand.

🡪 After some days the demand will go very high, we can scale the number of instances from 1 to 4.

🡪 In case the demand will decrease we need to reduce the number of instances as per our requirements.

🡪 Health check grade periods is most important in our EC2 auto scaling. It will automatically check in some periods of time it will check the health of the EC2 instance, it will good or not. Weather the health of the Ec2 instance is not good it will replace our existing EC2 instance with new instance.

🡪 Auto Scaling group can’t contain instances from the multiple regions. It must be of same region along with multiple availability zones. We couldn’t distribute along multiple regions.

Benefits of Auto Scaling:

🡪 Maintaining Performance Automatically: AWS Auto Scaling frequently monitoring resources underlying our application to make sure that they are operating at your desired performance level. When the demand will increase, AWS Auto Scaling Automatically adds number of instances so that we can maintain high quality of services.

🡪 Take Care of Replacing Unhealthy instances: Auto Scaling detect immediately when an instance is not healthy. The instance can be terminated, and another instance can be replaced on it.

🡪 Reducing Bill: Auto Scaling can reduce bills, because most cloud providers are charge based on total usage rather than maximum capacity for infrastructure hosted in the cloud.

🡪 Auto Scaling can offer greater uptime and more availability in case where production workloads are variable and unpredictable.

🡪 Auto Scaling keep track od the demands of our application and makes sure that the application has correct amount of capacity.

**Practically we need to create a Auto Scaling in AWS Console**

**🡪** Let login into the AWS console and go to EC2 Dashboard.

🡪 Scroll down at the bottom we have Launch Configuration and Auto Scaling.

A screenshot of a computer

Description automatically generated

click on Auto Scaling 🡪 click on create Auto Scaling Group 🡪 here Click on Create a Launch Template 🡪 Name (whatever u want) 🡪 AMI (in case u have any instance create an AMI {or} select anything from here) 🡪 Instance type (select what u want) 🡪 In key pair (choose existing key pair) 🡪 In existing key pair (select your existing key pair) 🡪 Click on Acknowledge box 🡪 click on create launch configuration.

Now click on Auto scaling group 🡪 Give name (Whatever you want) 🡪 Launch Configuration (Select your launch configuration created above) 🡪 click on Next 🡪 Select the VPC 🡪 Select the multiple Availability zones 🡪 Click on Next 🡪 Here if u did not need any LB u may select no LB { if who have LB select Attach to New LB 🡺 here what ever you want select I will select ALB 🡺 Select Default Routing click on create target group} 🡪 click on Next. 🡪 In Group Size (Desired capacity=3, Minimum capacity=2, Maximum capacity=4) 🡪 Scaling Policies {if demand goes high the servers will increase automatically, demand goes down it will decrease automatically depends upon CPU, Network in, Network Out, LB request Count} 🡪 select Target tracking Scaling Policy 🡪 Metric type (CPU Utilization) 🡪 Target value (give what value you want) 🡪 Instance need =100 seconds {if our server reached the specific CPU utilization it will wait for 100 seconds and will create new server automatically} 🡪 Click on Next 🡪 click on Next 🡪 click on Next 🡪 Verify all the steps and click on Auto Scaling group.

It is Successfully launched 🡪 Click on Name 🡪 we see the details of Auto Scaling here. 🡺 we need to check this Auto Scaling is working properly or not.

🡺 Go to EC2 instances 🡪 Here the 2 instances are created successfully.

Note: Some additional process is there for Auto Scaling we can connect with servers and create a file on it and run that file the CPU will utilise and Auto Scaling Will work here.