Assignment-5- Auto Scaling

Problem Statement: You work for XYZ Corporation which uses on-premise solutions and a limited number of systems. With the increase in requests in their application, the load also increases. So, to handle the load the corporation has to buy more systems almost on a regular basis. Realizing the need to cut down on the expenses of systems, they decided to move their infrastructure to AWS.

Tasks To Be Performed:

1. Create a web server AMI with Apache 2 server running in it.

2. Create a launch configuration with this AMI.

3. Use this launch configuration to create an Auto Scaling group with 1 minimum and 3 maximum instances.

Solution

Autoscaling is a process in which we can basically scale up or down the no's of instances depending on the load upon our particular instance. When we said load upon our instances, the load can be anything, it can be CPU Utilization, it can be network in or network out, or depending upon different matrices, we can basically specify whether it needs to increase the number of instances present in our architecture or not.

So ASG (Autoscaling group) enables us to particularly automate our website how exactly does it give us that particular feature, so enabling ASG gives minimum intervention.

AWS autoscaling monitors our applications and adjusts capacity automatically to ensure consistent predictable performance at the lowest possible cost without any manual intervention. Scaling is the process of adding/removing capacity/resources, scale-out means adding the resources and scale-in means removing the resources.

There are two types of scaling 1. Vertical Scaling & 2. Horizontal Scaling. In vertical scaling, we just increase the number of CPUs, RAM, storage, etc on a single instance. The main advantage of vertical scaling is that it will reduce the software cost as we will be using only a single instance and enhancing its capability. Since we are using only one instance so it has more chances of getting downtime.

In horizontal scaling, we are adding more no of machines of the same type to the existing pool of resources. The advantage of horizontal scaling is that it has fault tolerance, which means it is highly available and the latency will be very less. But the disadvantage of Horizontal Scaling is that since we are using more instances so the cost will be high and also the implementation of that architecture is not that much easy.

Here we launch an instance (UBUNTU) and installed an Apache2 web server on it. After that, we will create a launch configuration with this AMI.

Steps for creating a AMI

Select the instance→Action→Image and Templates→ Create image (put an image name)→Create image

Now, go to the Image in the dashboard then AMI, and you will get the AMI that you create just now.

Launch Configuration

Whenever we are going to launch an Auto Scaling Group, the major function is to add and remove instances, but how will I know what kind of instances have to add, what will be the webserver running on it, and what will be the details of that instances? So for that reason, we need to attach this AMI, as we know that this AMI contains all the software details, so if we give this AMI to ASG then it will only know the software specification, that's why Launch Configuration comes in place. Launch configuration gets all the hardware as well as software details.

Steps to Create Launch Configuration

Go to Launch Configuration→Create Launch Configuration→Select the AMI→Select the instance type→Security group (choose the Security group which we used during instance creation)→Select the key pair→ Select the I acknowledge→ Create Launch configuration.

Steps to Create Autoscaling Group

Before going to create the Auto Scaling group, make sure that the AMI must be in the Available state.

Now go to the Auto Scaling group→Create ASG→ Name the ASG→ Launch configuration (switch to LC)→ Select all the availability zone and subnets (this will make the entire infrastructure very highly available)→ Load balancer option (if you select the option "attach to an existing LB" then you must the Load Balancer, then what happens all my targets will be launch inside the target group of my load balancer due to which even though like if my first instance goes down, I will be directly able to access the data of all my instances from the DNS name of the load balancer itself)→ Next→ Group Size (desired=2, minimum-1, maximum=3)→Next→Next →Create ASG

Results