Tasks To Be Performed:

1. Manage the scaling requirements of the company by: a. Deploying multiple compute resources on the cloud as soon as the load increases and the CPU utilization exceeds 80% b. Removing the resources when the CPU utilization goes under 60%

2. Create a load balancer to distribute the load between compute resources.

3. Route the traffic to the company’ domain

So the soln. here is 🡪

We will create a ASG(Auto scaling group) here with policy regarding the CPU utilization .

Also we will attach a Load Balancer Here.

In Last we will attach our Load Balancer to the Route . using a private domain.

[www.intellipatassignment.com](http://www.intellipatassignment.com).

So step 1 🡪 Lets create ASG.

Demo🡪 go into ec2 🡪left side🡪 Auto Scaling Group(after load balancer)



Click on create auto scaling group.

First step is to choose or create template. As we don’t have any lets create.

A screenshot of a computer

Description automatically generated

On clicking on create launch template a popup will open.

A screenshot of a computer

Description automatically generated

Here template version description is like version.

If we are updating our template later we can follow versioning , if we want to old one we can also do it later.

Now if you scroll down we can see the things similar to when we create our ec2 instance.

Like select Ami , memory , network security group etc.

We don’t need to describe these steps , as we are pro in it now.

A screenshot of a computer

Description automatically generated

See same they are , lets finish this creation step and come back and attach to it our ASG.

Now in step 2 we have to choose instance launch options.

Here we have to select network on which we have to deploy.

So select the vpc and all the availability zones with the respective subnets.

A screenshot of a computer

Description automatically generated

We have select all zones as we don’t want to keep our instances in single zone , so that if one corrupt other should be available.

Now our 3 step 🡪select or create Load Balancer , as we also want to distribute load on targets.

Lets create new as you already know steps and create one ALB.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

This step lets skip as it is we will learn it later when we learn VPC.

Now the most important one point to do health check.

A screenshot of a computer

Description automatically generated

So we have enabled here ELB health checks.

Now what is use of these health checks?

We have concept of desired, min, max in ASG.

These are the number of instances.

Desired 🡪 the number of instances when we start our application. By default these many instances create.

Min 🡪 No of instances when scale in happens.

Max 🡪 No of instances when scale out happens.

Desired automatic comes equal to min when scale in happens.

Lets see below diagram desired is always in between min and max .

A blackboard with yellow and red writing

Description automatically generated

So now when we perform health checks if any ec2 instance is not working ASG automatic create a new one and detach old one.

A screenshot of a computer

Description automatically generated

Now in this step we have enable group metrics for our ASG .

So whatever instances we are having at that time it will group their logs on basis of average time, network ,memory , cpu utilization etc.

Warmup time is when we are having new instance how much time we should wait to send it data in grouping to cloudwatch console.

Now next step is to configure group size and scaling. The place where we define desire , min ,max.

A screenshot of a computer

Description automatically generated

Lets put desire 2 , min 1 and max=5.

For Now lets put other things default.

A screenshot of a computer

Description automatically generated

We will create policy later and attach to it.

Skip all other things for now and create it.

Now if you check automatic 2 instance are created , one load balancer with target groups.

Now if u hit the dns of your load balancer we can see URLs are working perfectly fine.

A yellow mark on a white background

Description automatically generated

Now lets create autoscaling policy🡪

Go to your ASG🡪autoscaling 🡪create dynamic scaling policy.

A yellow and blue text on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

A popup will open

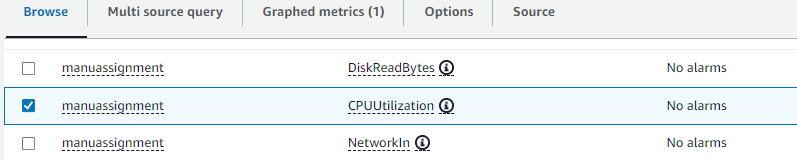
Before doing that first create two alarms on our ASG with respect to cpu utilization >80 and <60.

Go in cloud watch🡪metrics🡪scaling group.

Search our scaling group.

Select **Metric >> EC2 >> By Auto scaling group >> Select the Autoscaling group name “manuassignment” and the metric name “CPUUtilization” >> Select metric**

Create two alarms one for cpu utilization more than 80 and one less than 60.



For scale out

A screenshot of a computer

Description automatically generated

If greater than 80 percent we will scale out so created an alarm.

Same for scale in 60 percent.

A screenshot of a computer

Description automatically generated

Now back to create dynamic scaling policy.

A screenshot of a computer

Description automatically generated

Same way create scale in also.

A screenshot of a computer

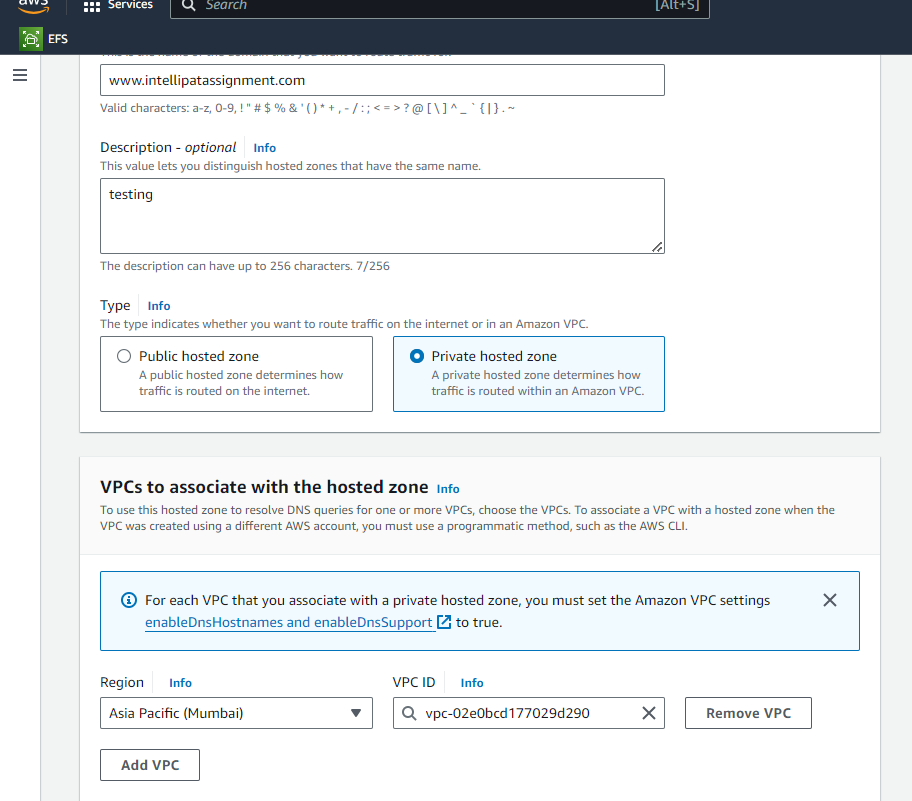
Description automatically generated

We have attached now our both policy to our ASG.

Now the main point is to route to our domain.

[www.intellipatassignment.com](http://www.intellipatassignment.com)

go to route53 dashboard

1. First we have to create our private hosted zone for this domain.
2. 
3. Now in this hosted zone we will create the record of A type to redirect it to our LOAD balancer .

First copy the DNS of our load balancer.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Now create the record.

Now connect to our instance and try to curl on our domain.