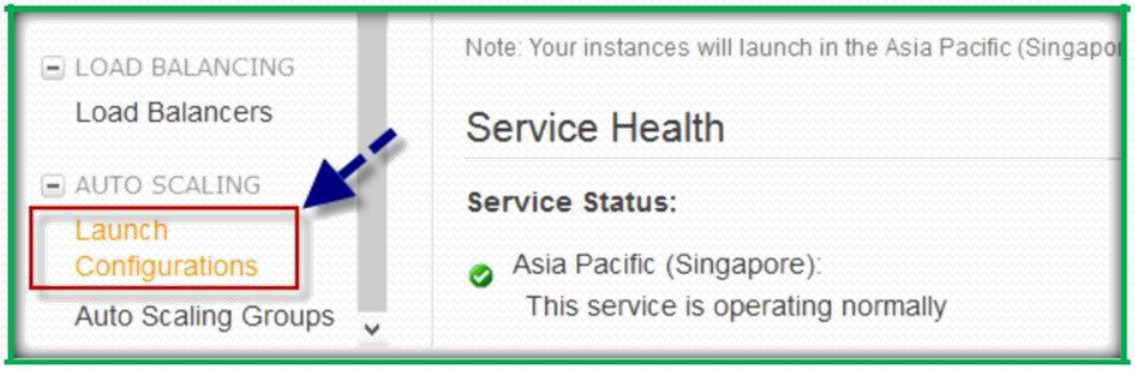
20. HORIZONTAL VS. VERTICAL SCALING

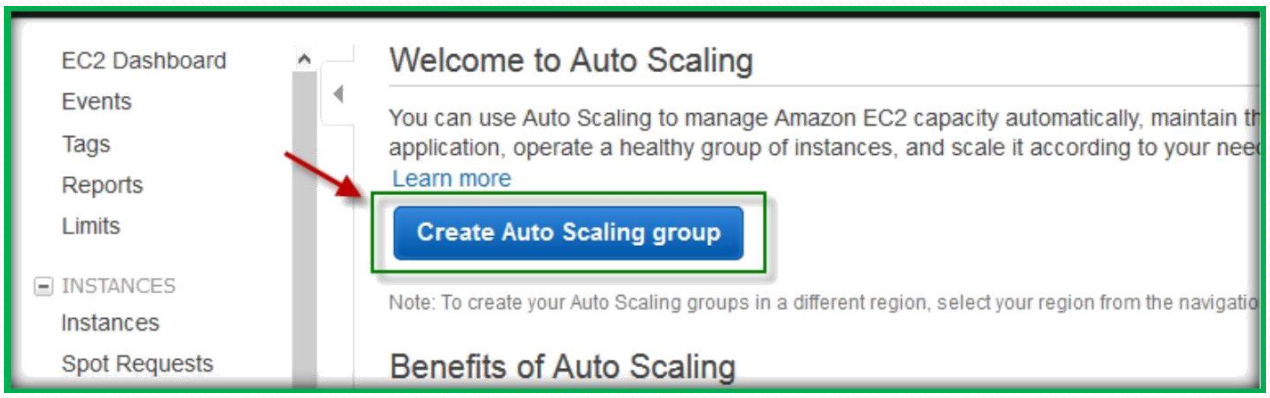
Horizontal scaling means that you scale by adding more machines into your pool of resources.

Vertical scaling means that you scale by adding more power (CPU, RAM) to your existing machine

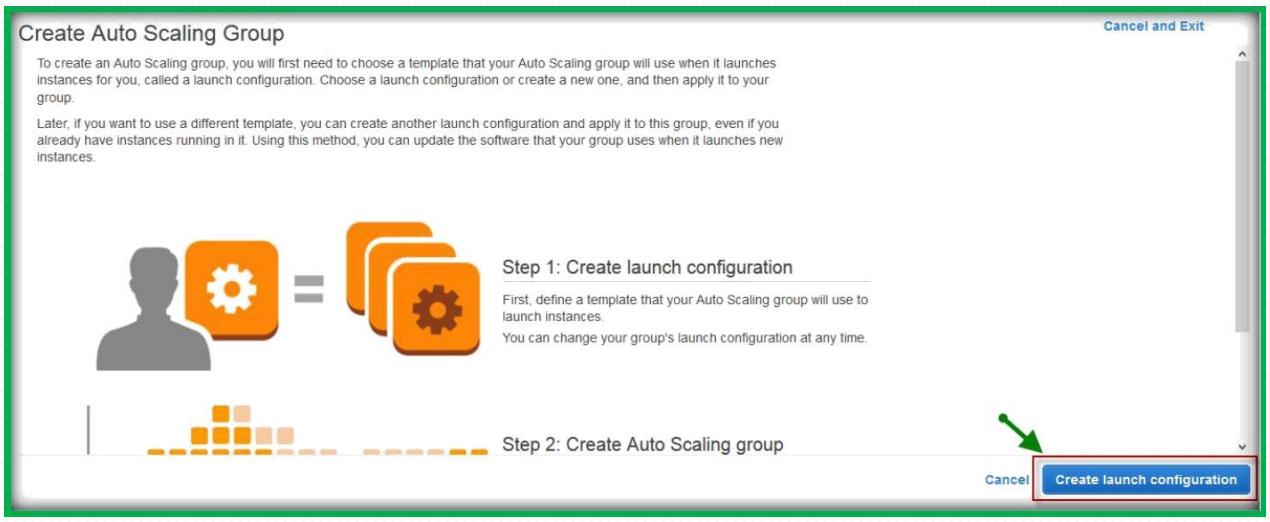
Navigate to the EC2 dashboard from the AWS Console and select Launch Configurations, located in the left bar under Auto Scaling.



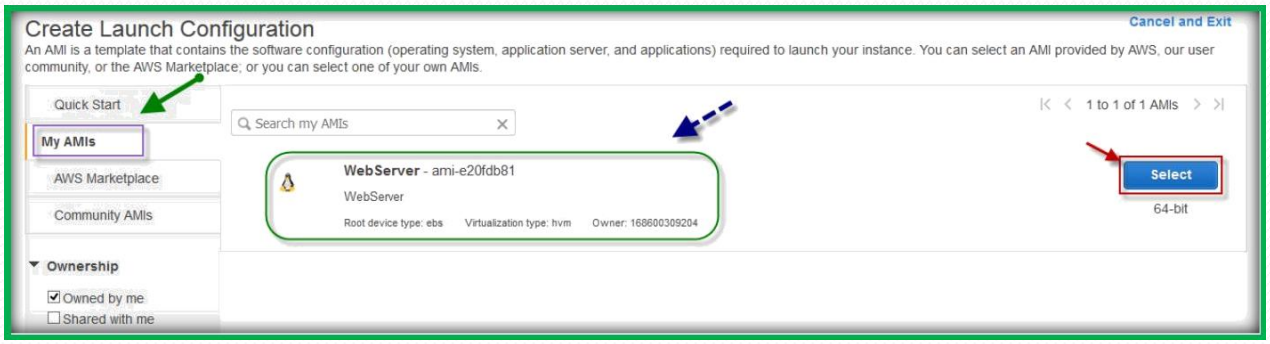
Choose Create Auto Scaling group under Welcome to Auto scaling page.



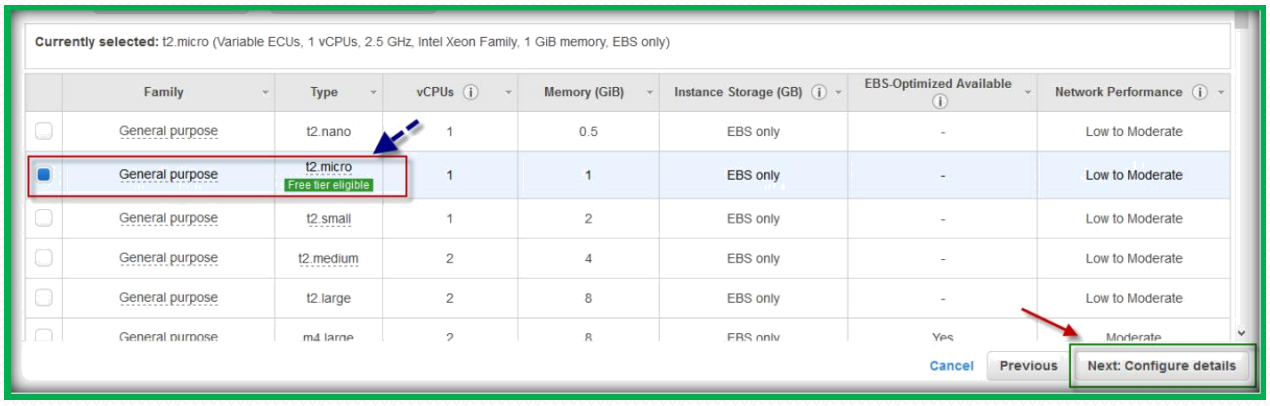
AWS will provide you with a page giving you an overview of Auto Scaling group creation. Click **Create launch configuration**.



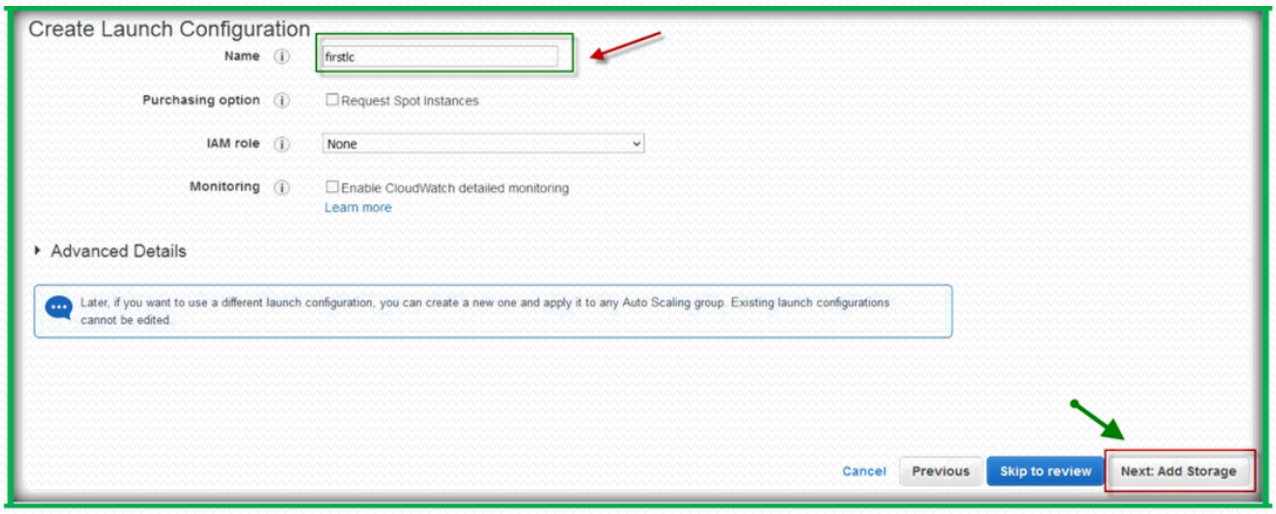
Under Create Launch Configuration page, choose My AMIs from left pane, choose your AMI by clicking on Select.



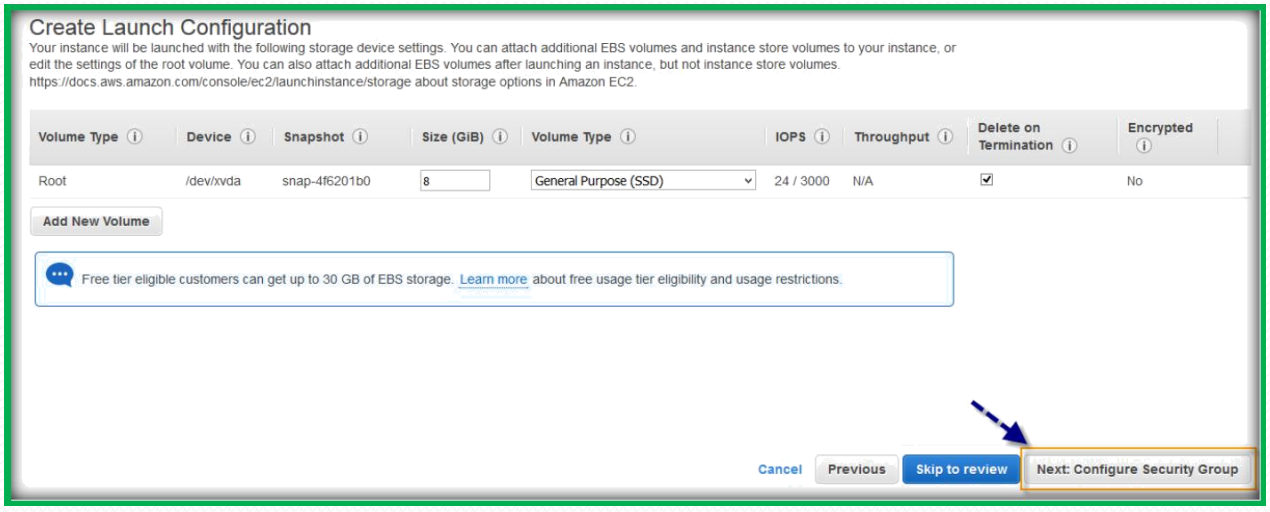
Choose your instance type and click on Next Configure Details.



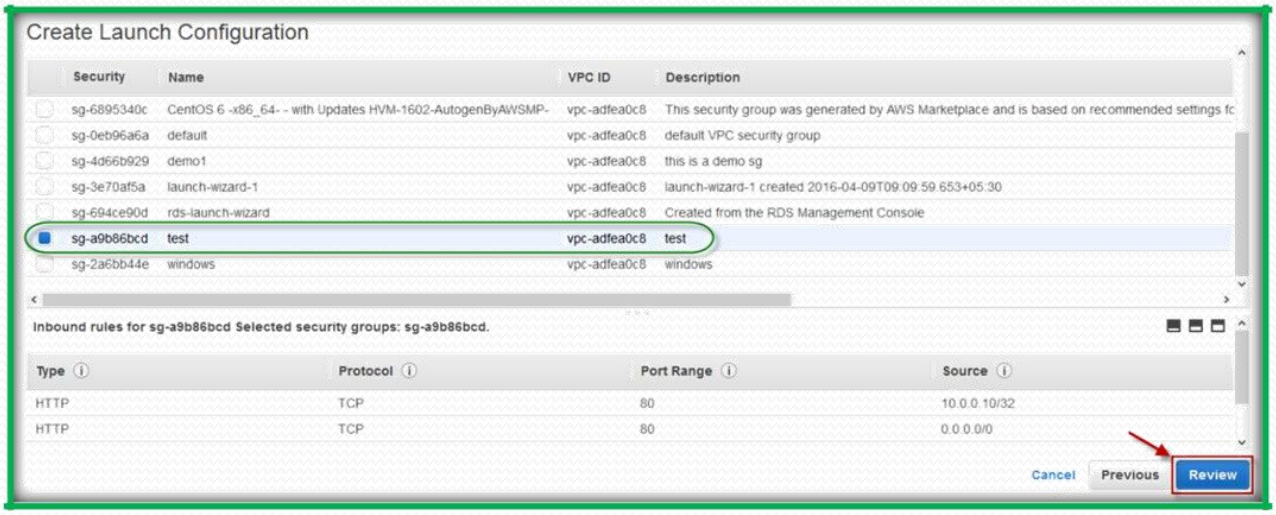
Specify a name for Launch Configuration, do not check **Request Spot Instances**, and leave the **IAM role** set to **none**. Also, leave **Monitoring** unchecked and choose Next Add Storage.



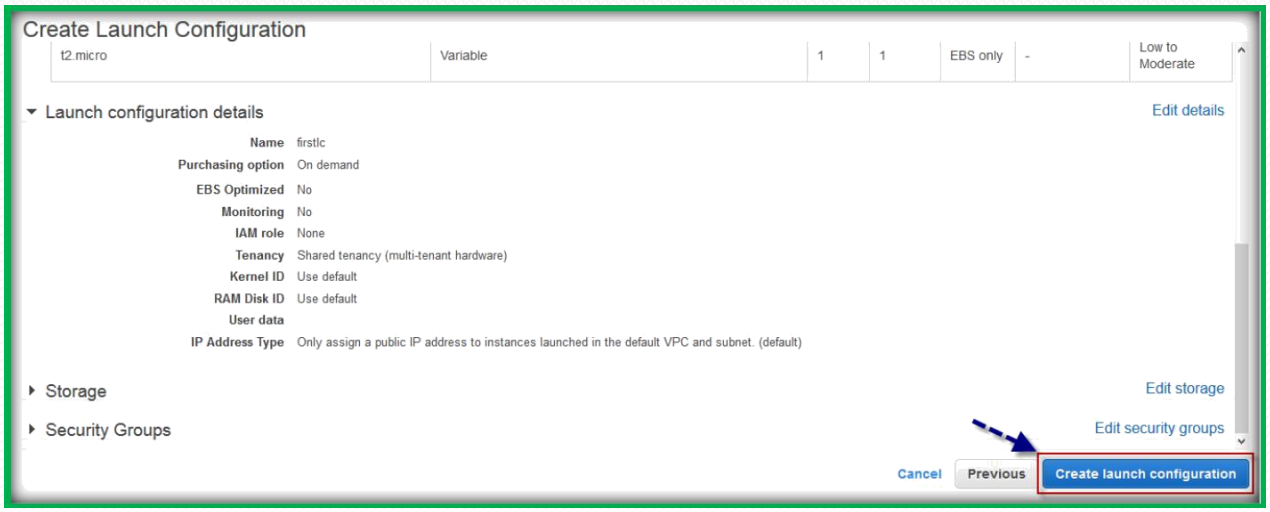
leave everything with the default settings. Go to **Next: Configure Security Group**



Choose existing security group or create a new one by adding required ports and choose **Review.**

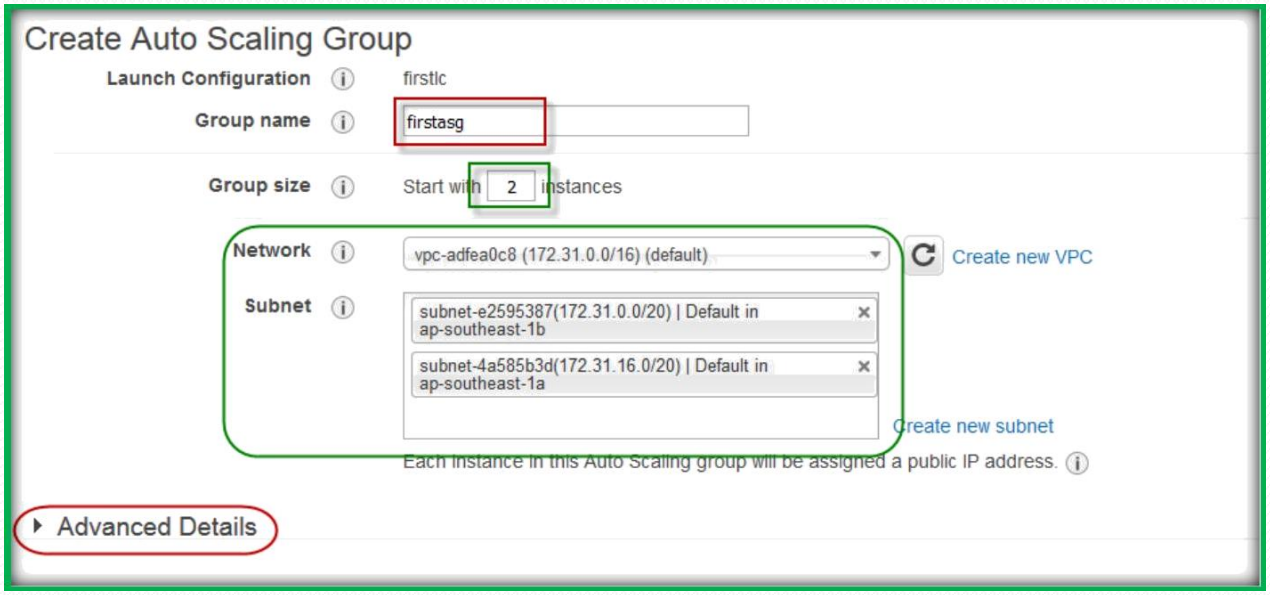


AWS will provide you with a page giving you review of all your settings, click **Create launch configuration.**

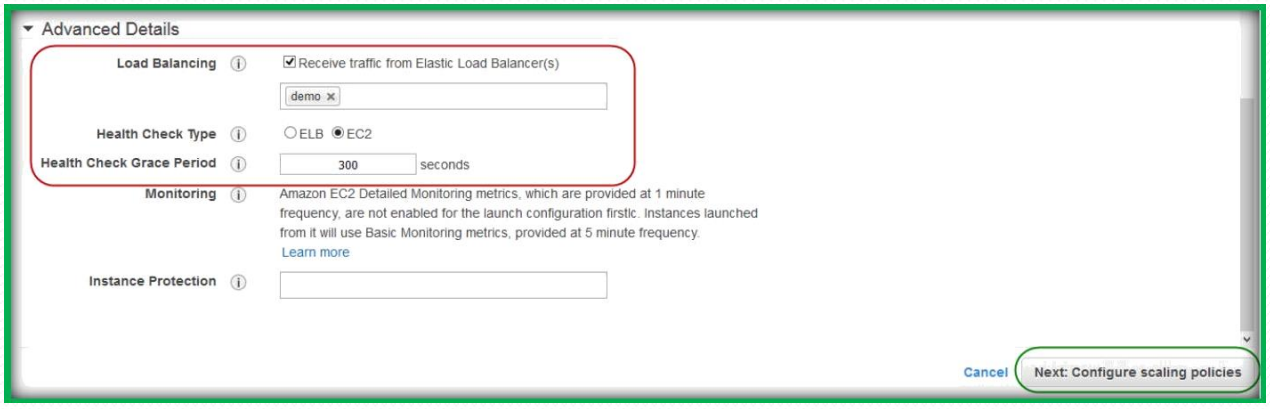


You will be asked for Key pair choose existing or create a new one, then choose **Create launch configuration**.

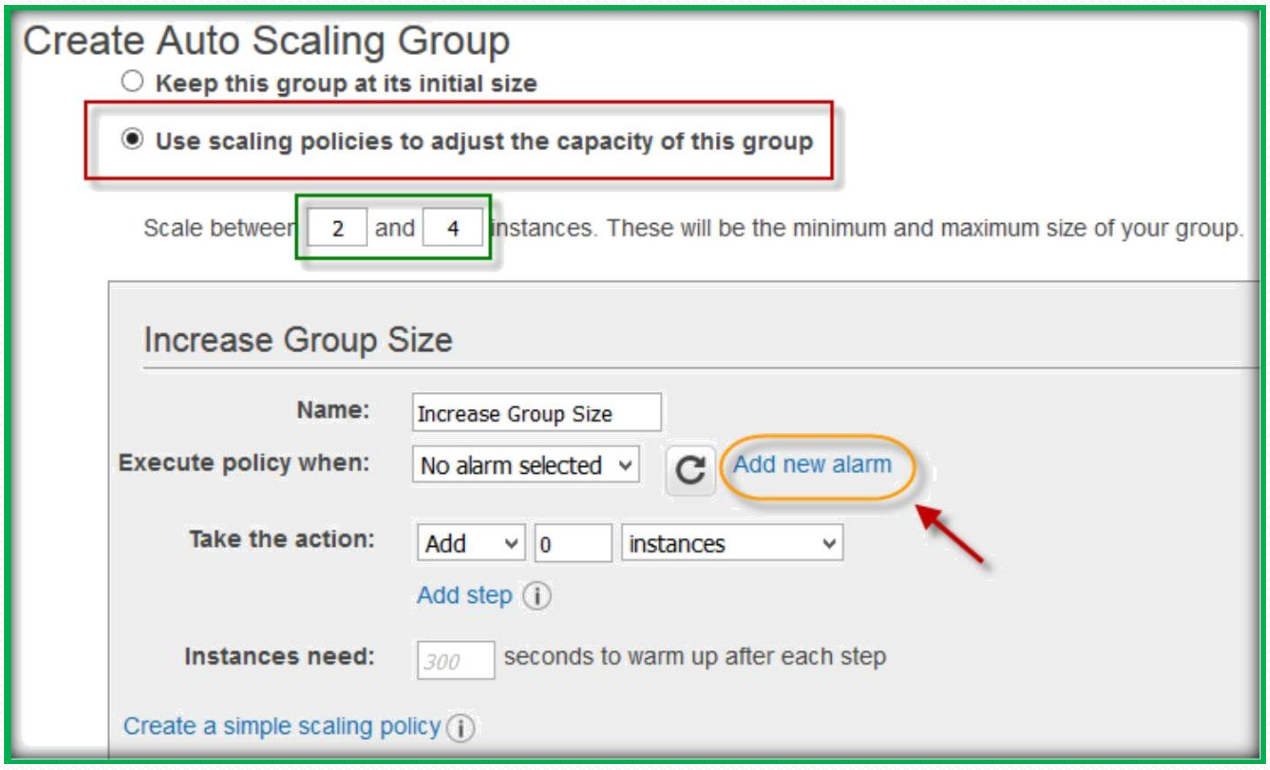
AWS will provide you with a page to Create Auto Scaling group. Specify a group name, we will start with 2 instances for high availability.
Select VPC from the VPC drop down list, add subnets under subnet section.
Then expand Advanced Details.



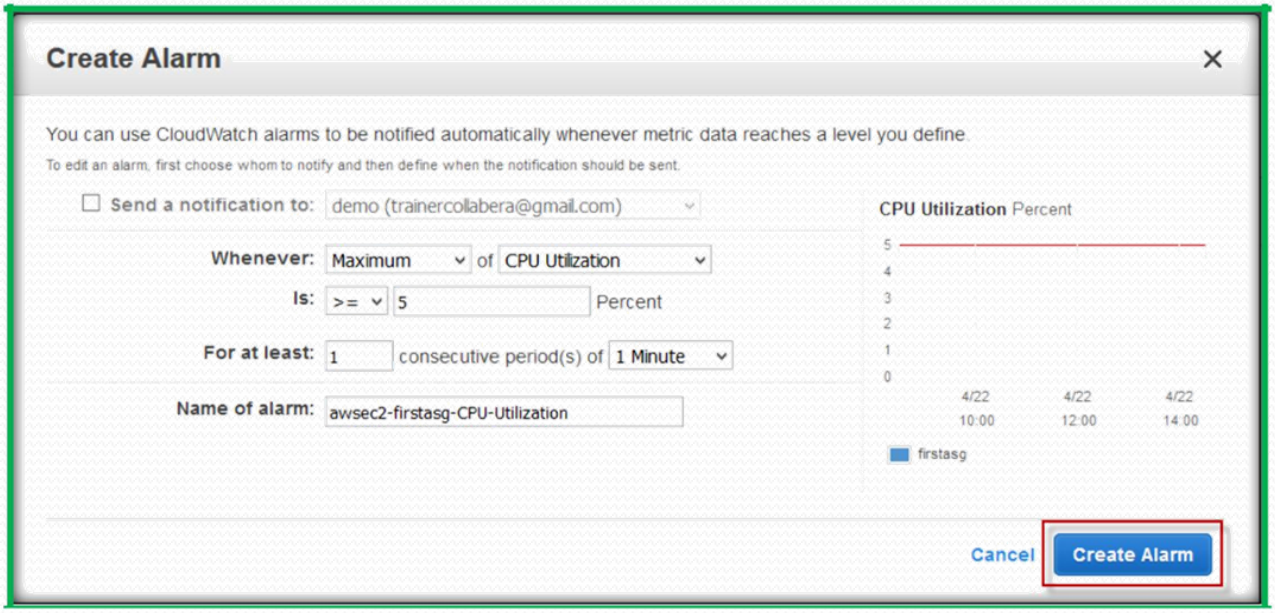
Under Subnet area is an Advanced Details section. Expand this so we can configure load balancing portion of the application. Check Receive traffic from Elastic Load Balancer(s) and in the box below, select the single ELB available. Set the Health Check Type to ELB. You can leave the Health Check Grace Period at the default 300 seconds.
Then choose next configure scaling policies.



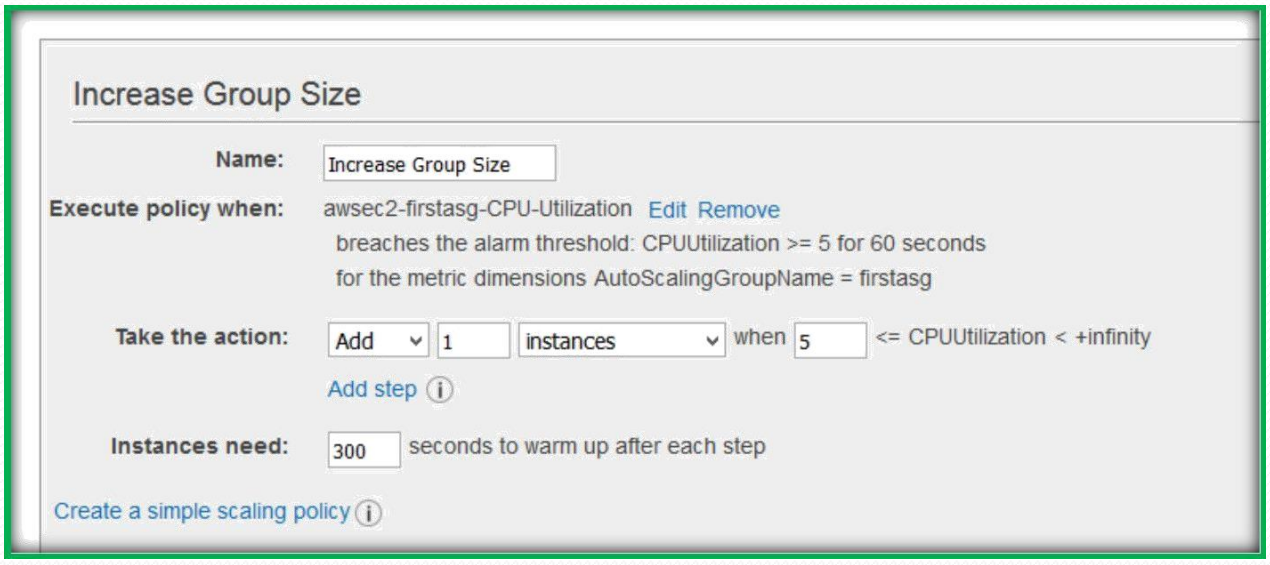
We want to **Use scaling policies to adjust the capacity of this group**. You will be presented with two options for actions and alerts: Increasing and Decreasing the group size.
First, we must define the minimum and maximum amount of instances, however. Set it to **Scale between 2 and 4 instances**.
Within the **Increase Group Size** area, press **Add new alarm**.



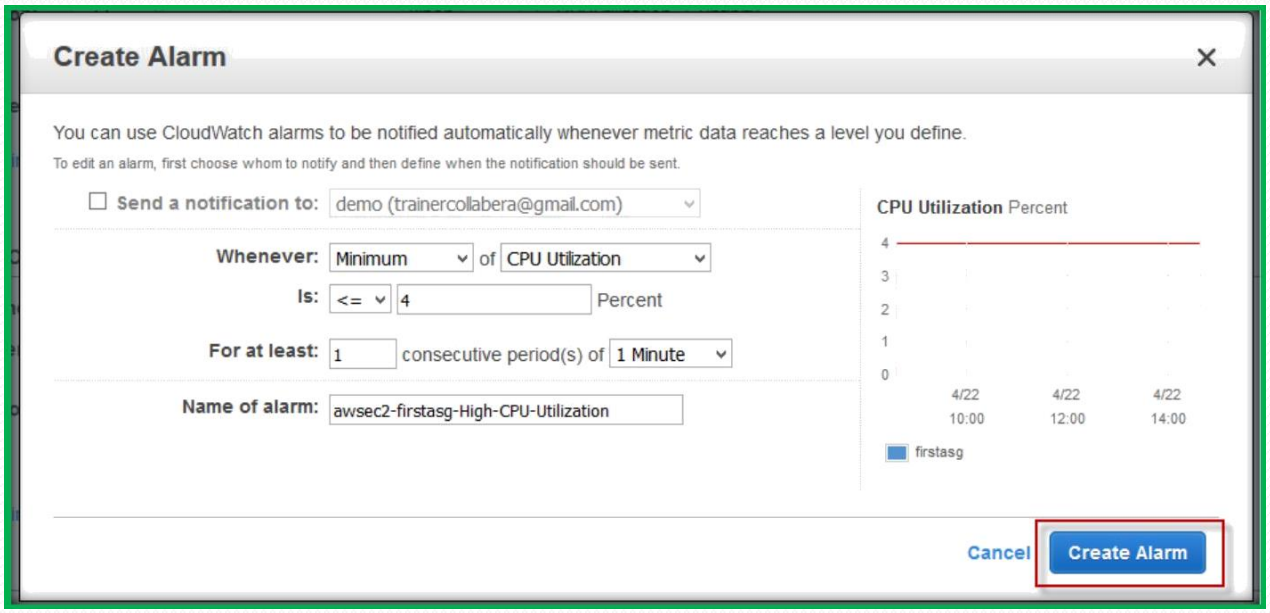
Uncheck the option to send out a notification, and change Whenever to be a Maximum of CPU Utilization [that] is >= 5 Percent. Set for at least to be 1 consecutive period(s) of 1 Minute. Press Create Alarm.



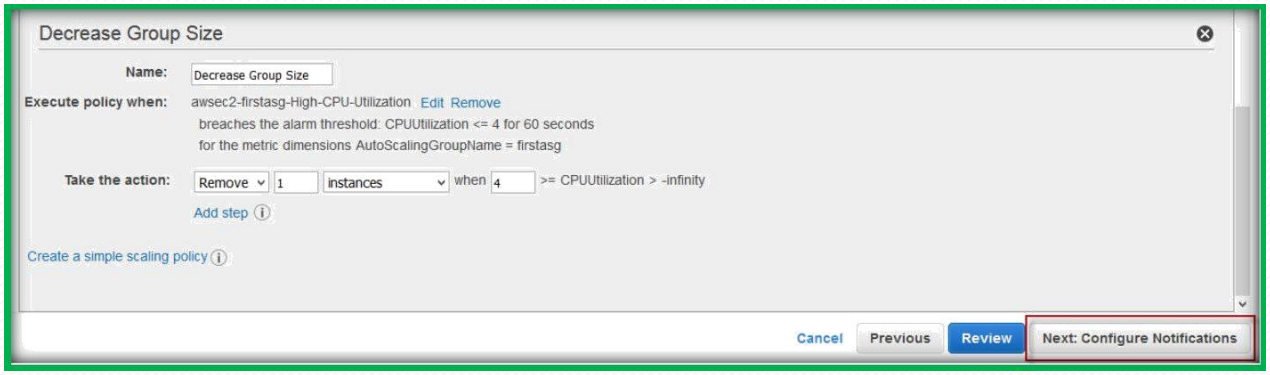
From here, we now need to define the action we want AWS to take when the alarm threshold is hit.
In the Take the action area, we want to Add 1 instance. Set the Instances needed to 300 seconds to warm up after each step.



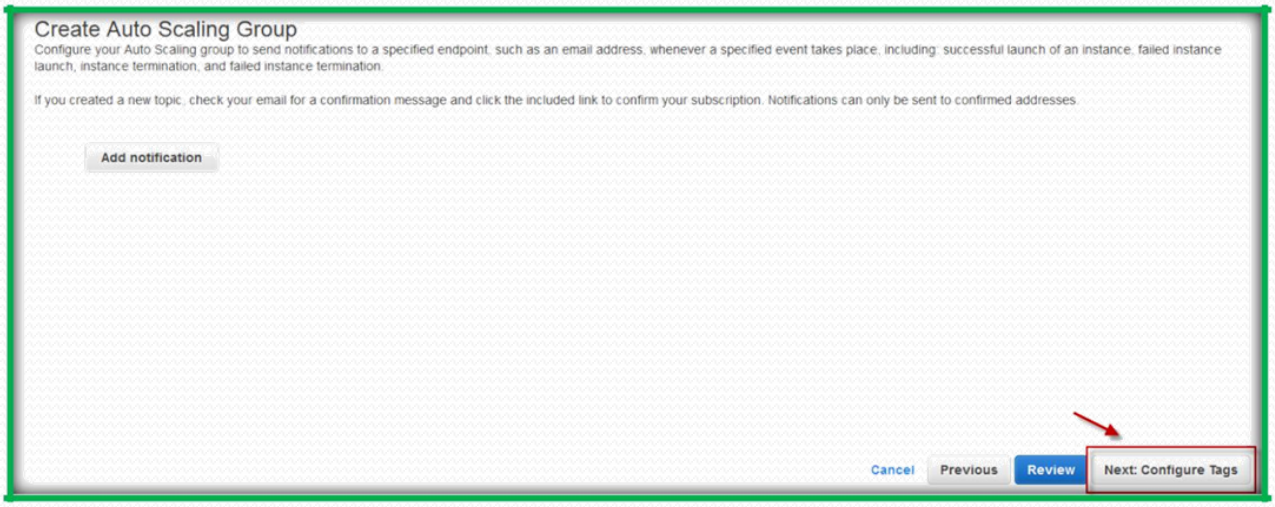
Under Decrease Group Size, also Add new alarm.
Again, deselect the send notification option.
Set Whenever to a Minimum of CPU Utilization [that] is <= 19 Percent for at least 1 consecutive period of 1 Minute. Create Alarm.



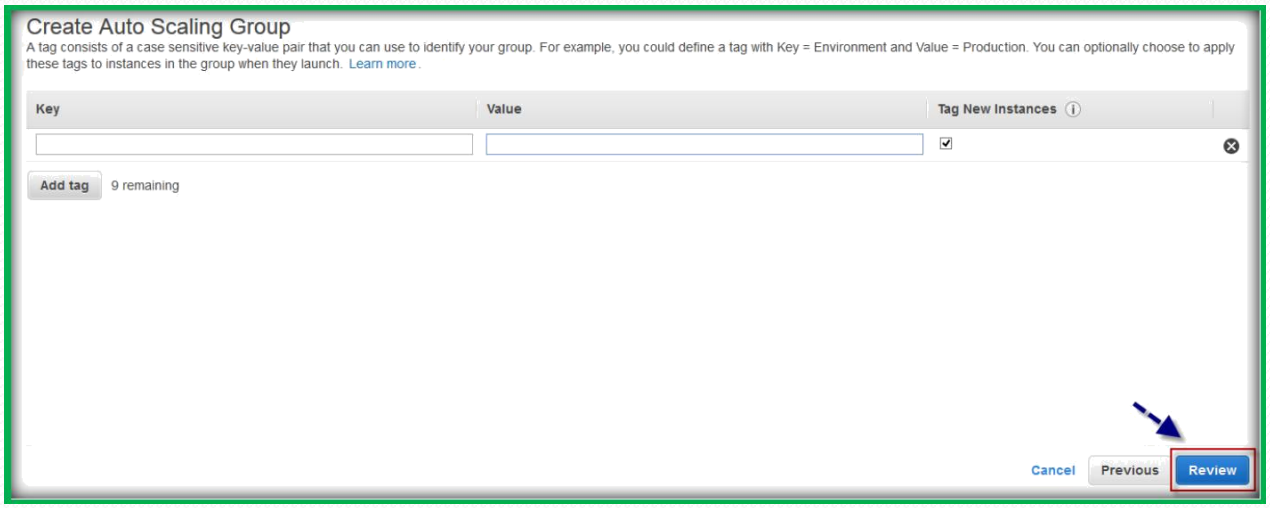
Then set Take the action to Remove 1 instances. Press next configure notifications.



Do nothing on the next page, choose Next: Configure tags.

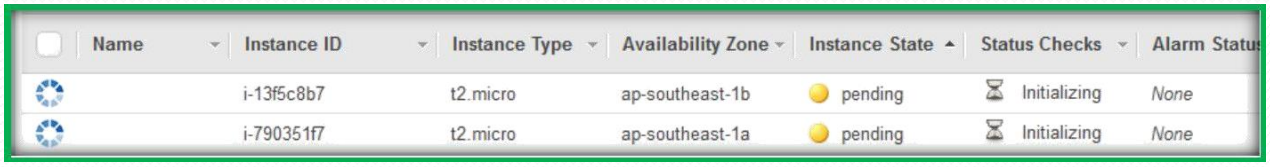


On the next page, choose Review, as we do not want to add any tags to these instances.

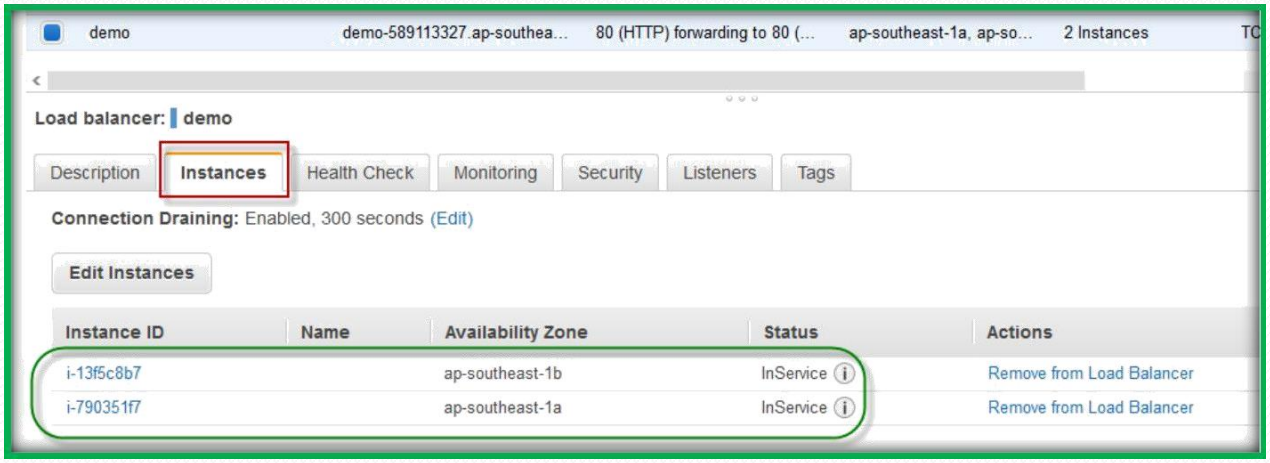


AWS will provide you with a page giving you review of all your settings, click **Create Auto Scaling Group.**

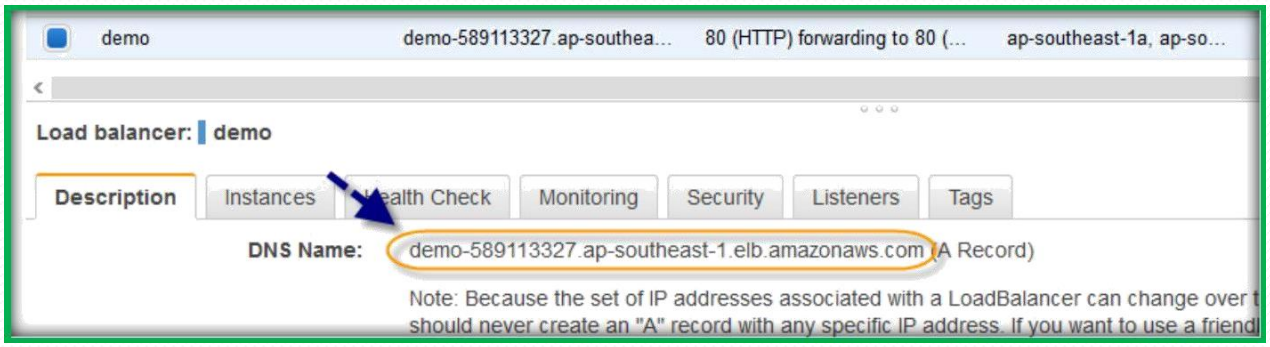
Then go to Instances page to see the instances which were started creating.



Then go to Load Balancers section, select your load balancer, choose instances tab, you can see instances from two AZ's attached to ELB and status is in service.



Then select Description tab, you can see the DNS name for ELB, copy and browse
your application which is auto scaled and high available.



If you want to test self-healing, we can delete one instance which is created by Auto Scaling, Auto Scaling will automatically launch a new instance to meet the minimum requirement which put as 2 instances.