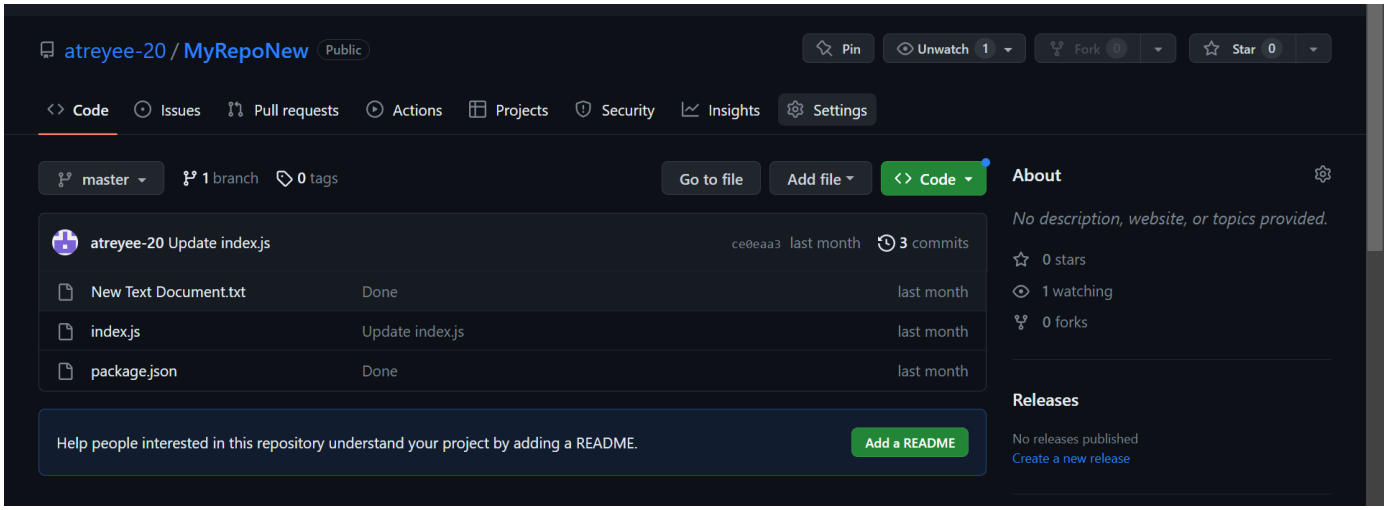
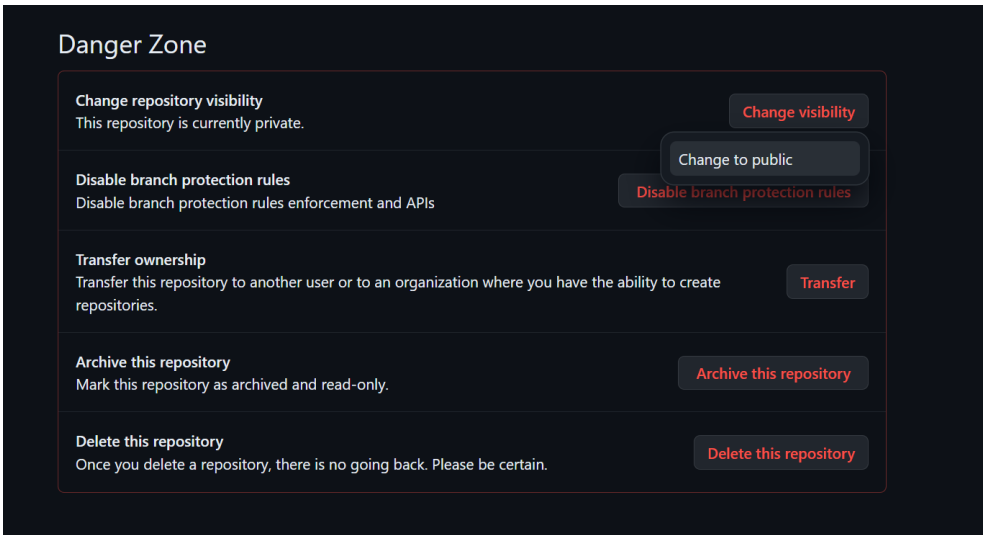
ASSIGNMENT-11

Build Scaling plans in AWS that balance load on different EC2 instances.

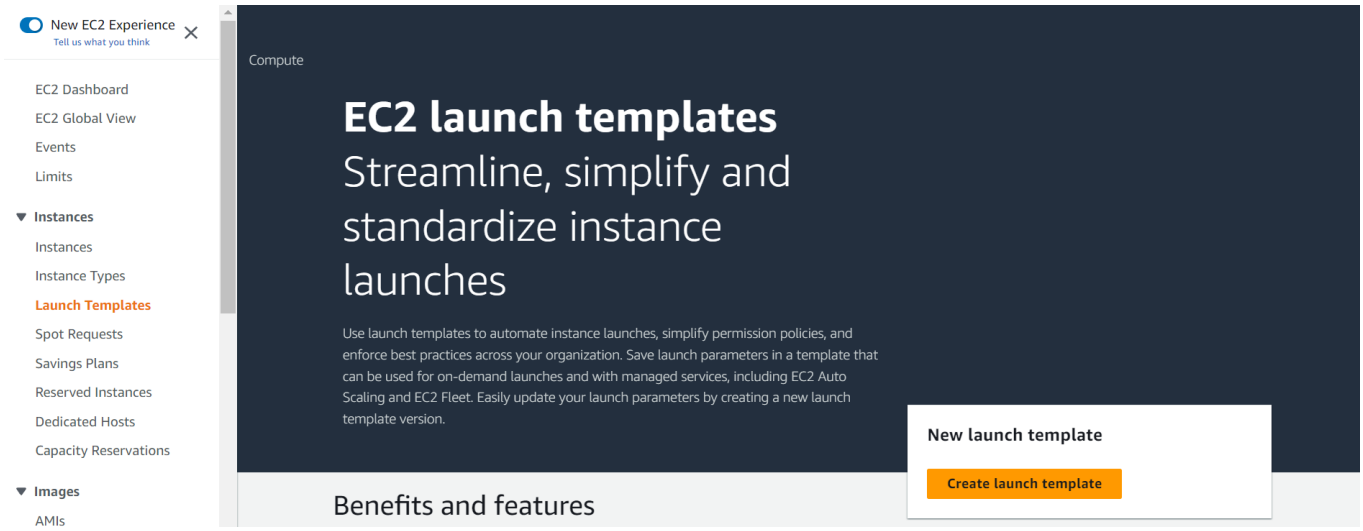
Step 1: Sign in to your GitHub account. Go to your repository. Then go to settings.



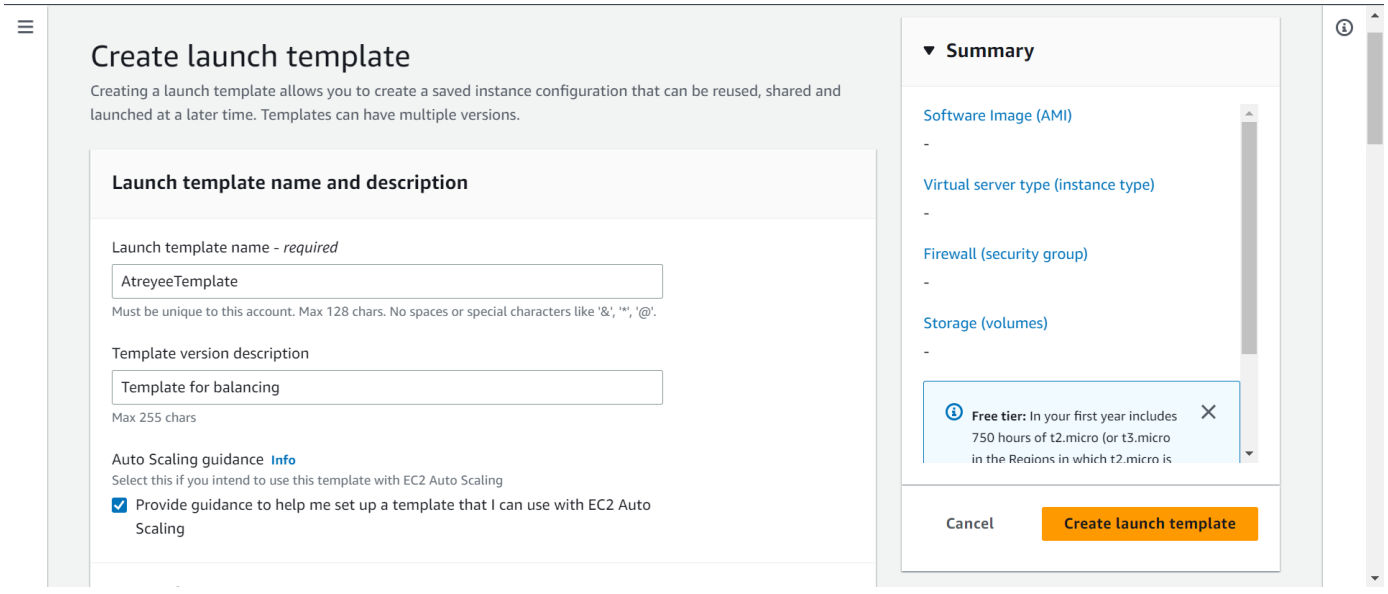
Step 2: Scroll down to Danger Zone. Click on Change visibility. Then Change to public.



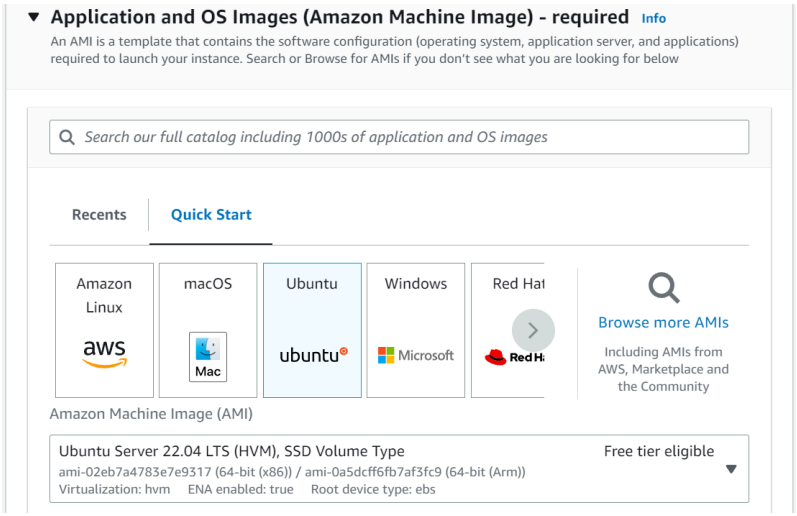
Step 3: Now Sign in to your AWS account as a root user. Then go to EC2 Dashboard. Click on Launch Templates.



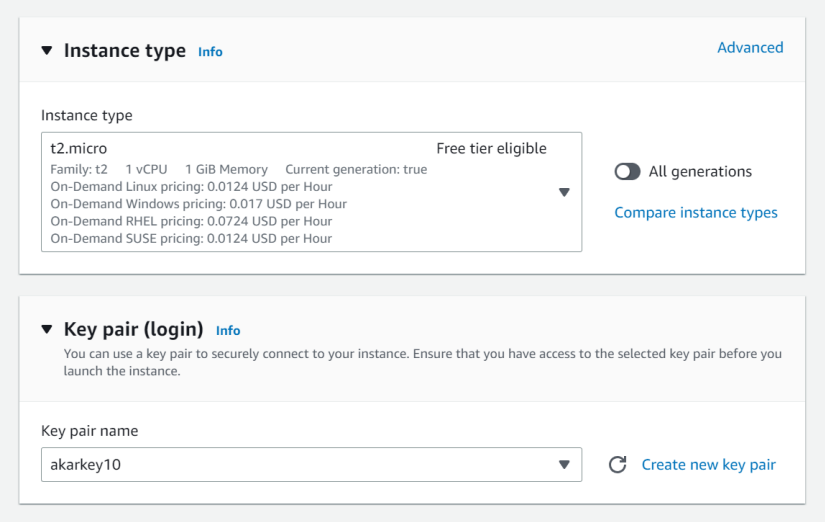
Step 4: Click on Create Launch Template. Give name and description. Then check Auto Scaling guidance box.



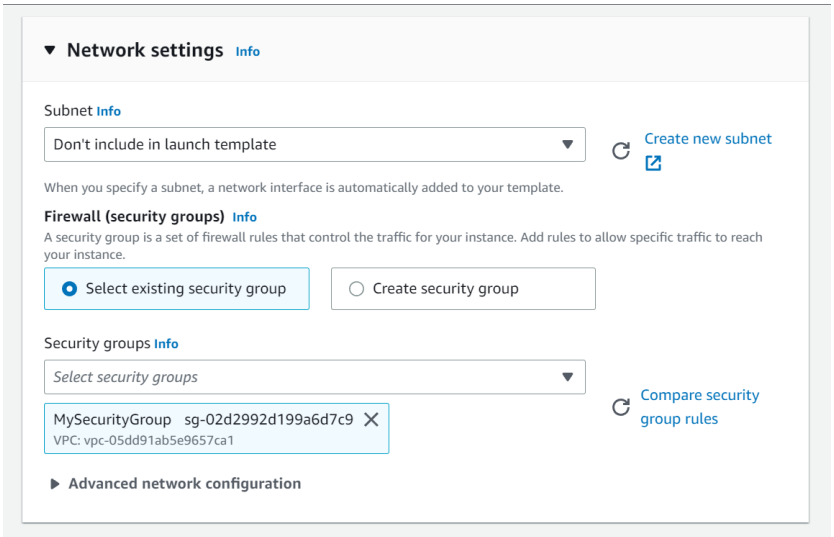
Step 5: Select Ubuntu as OS.



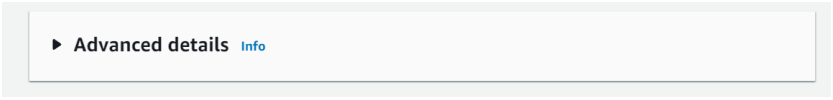
Step 6: Select t2.micro as Instance type and provide a key pair.



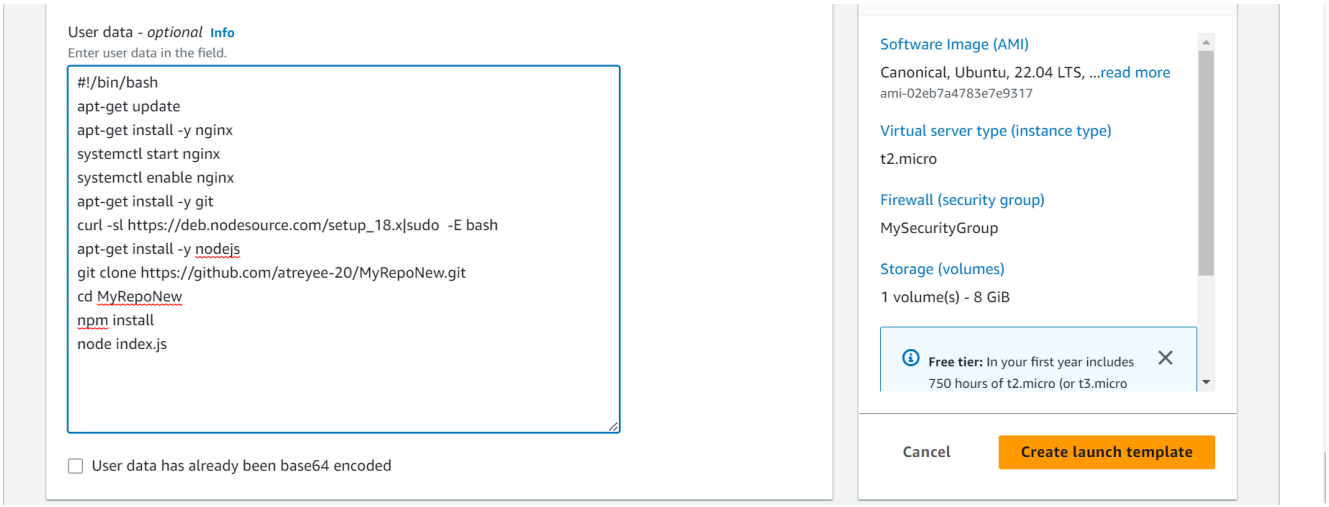
Step 7: Either select existing security group (If there is any) or create new Security Group.



Step 8: Go to Advanced details section and then scroll down to User data.

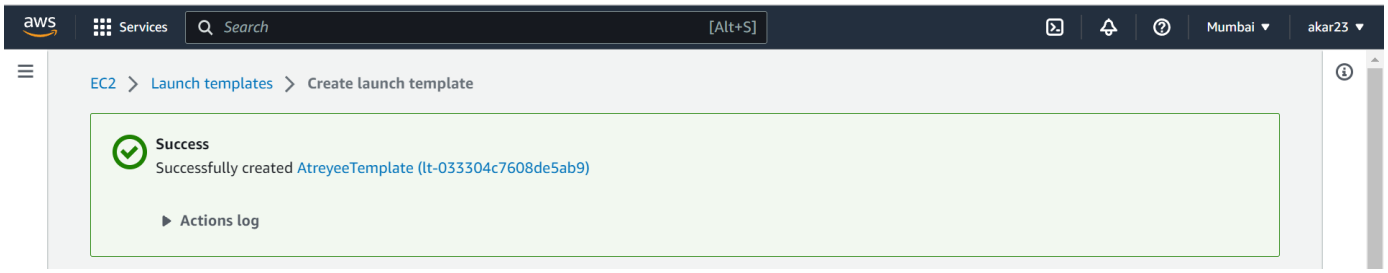


Step 9: Provide the following commands in User data as shown:

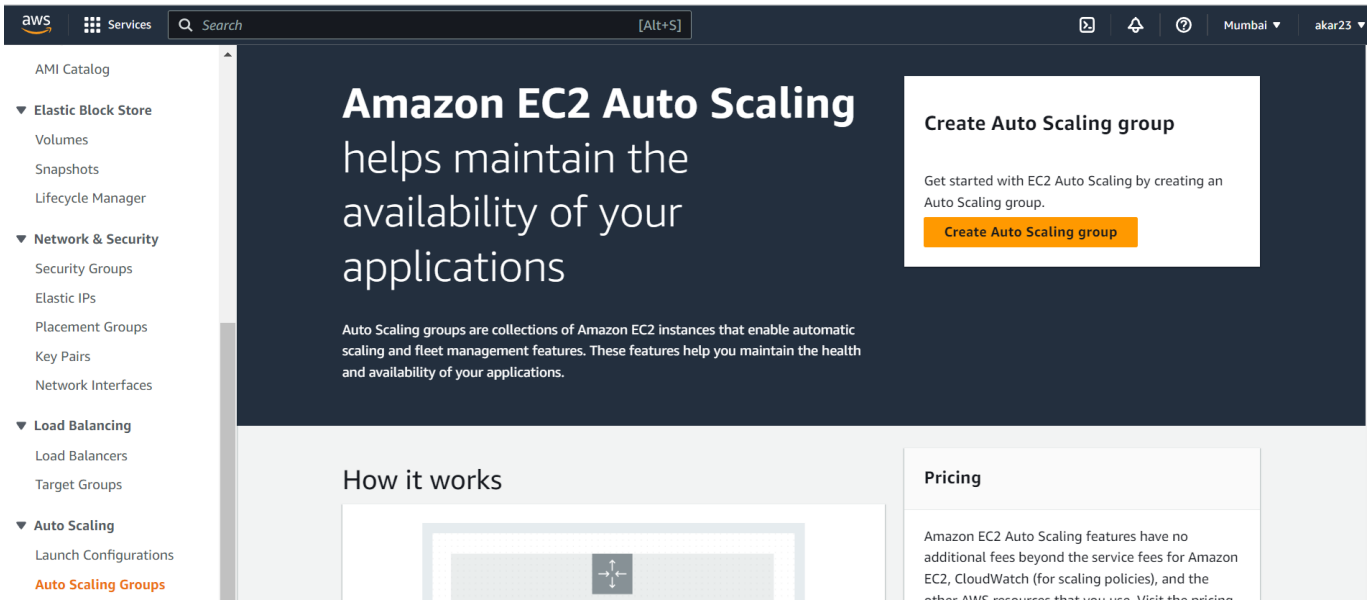


Then Click on Create launch template.

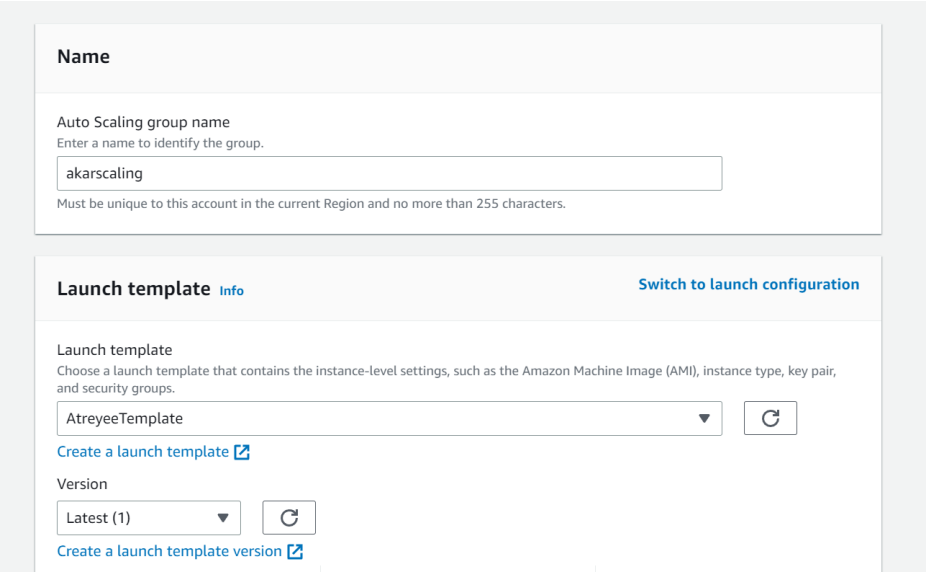
Step 10: New template is created.



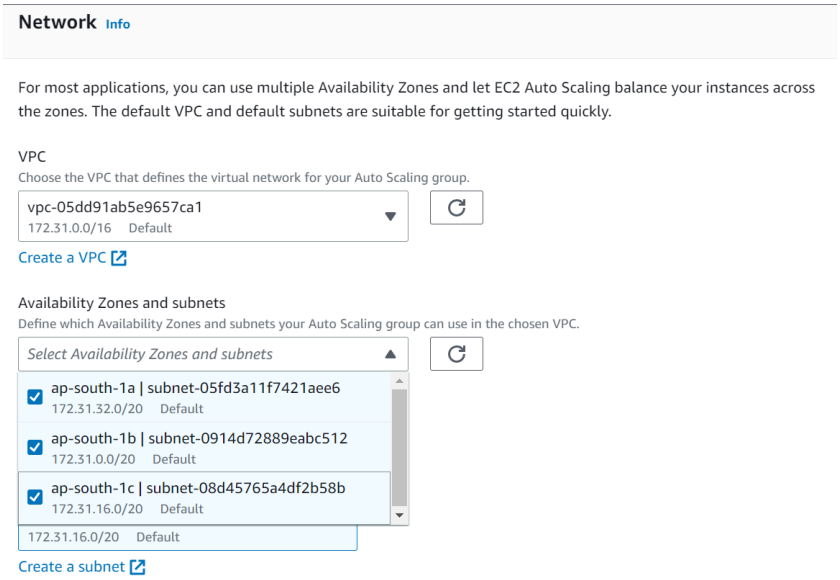
Step 11: Again go to EC2 dashboard and click on Auto Scaling Groups. Then click on Create Auto Scaling group.



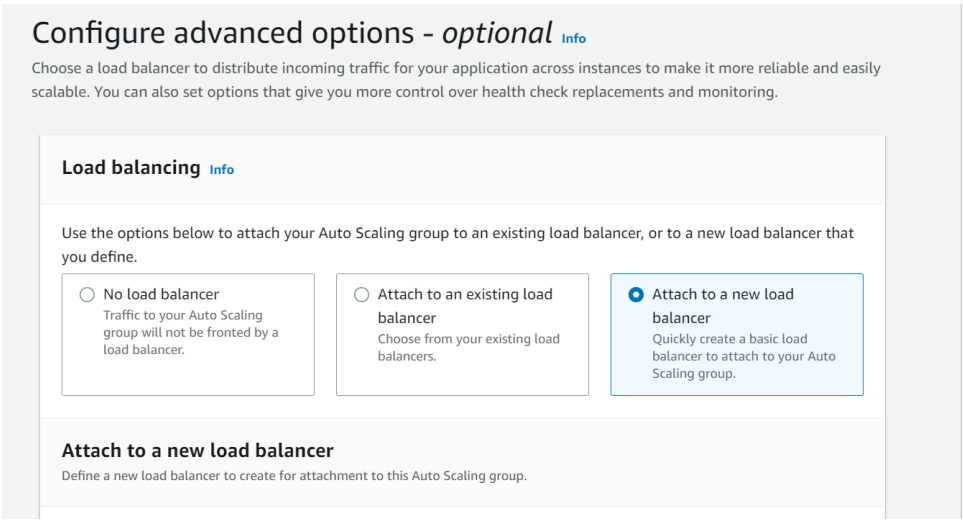
Step 12: Give a name. Select the template you just created and choose its latest version. Then click on Next.



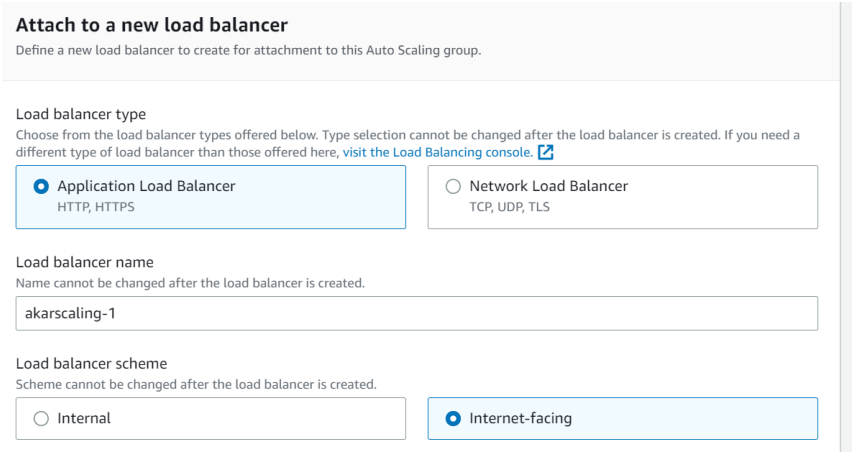
Step 13: Now go to Availability Zones and subnets and select all. Then click on Next.



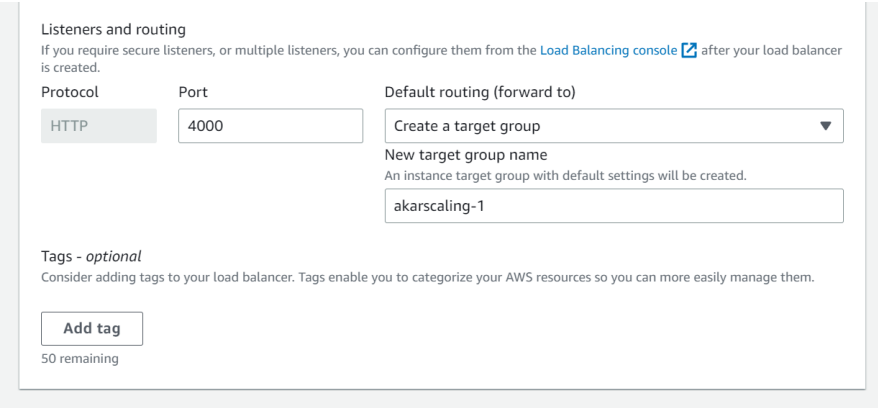
Step 14: Select Attach a new load balancer.



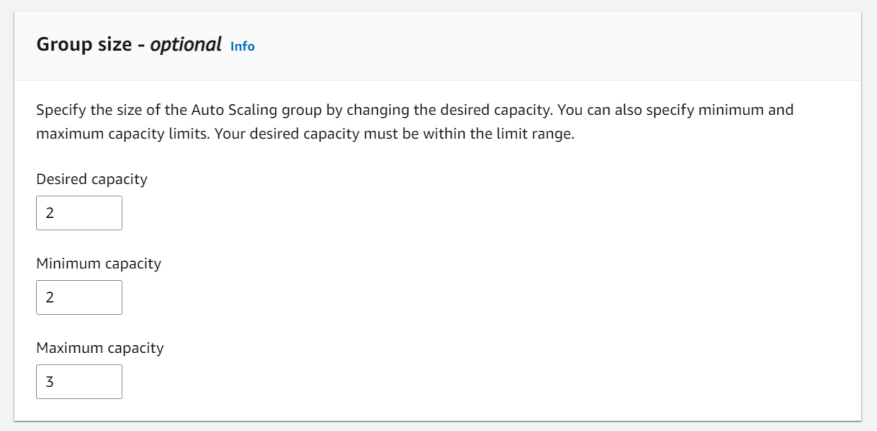
Step 15: In Attach a new load balancer, select balancer type as Application Load Balancer and Load Balancer scheme as Internet-facing.



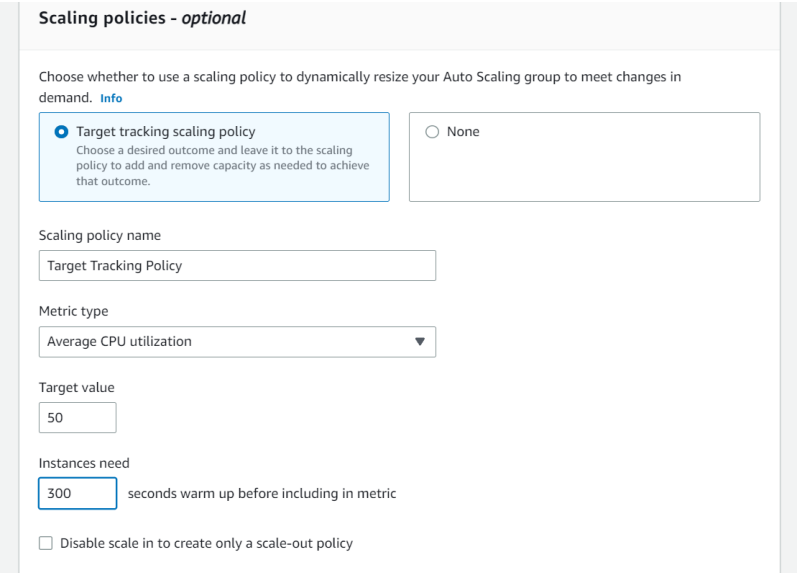
Step 16: In Listeners and routing give the port number (For my project, it is 4000). Set the Default routing as “Create a target group”. Automatically New target group name will appear. Then click on Next.



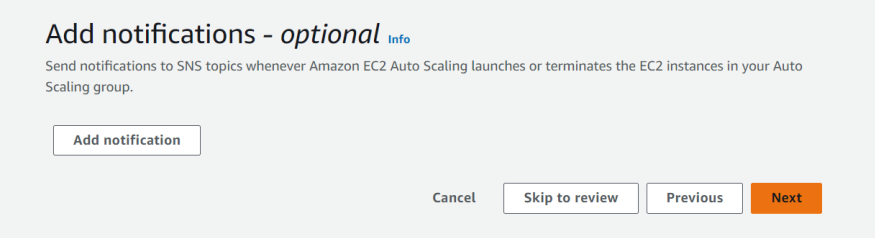
Step 17: Set Group size.



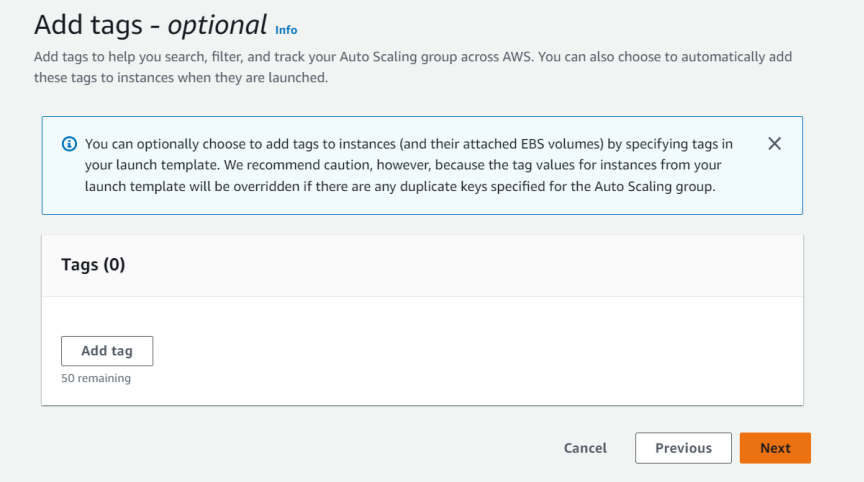
Step 18: Set scaling policy as Target tracking scaling policy and provide the time needed for instances. Then click on Next.



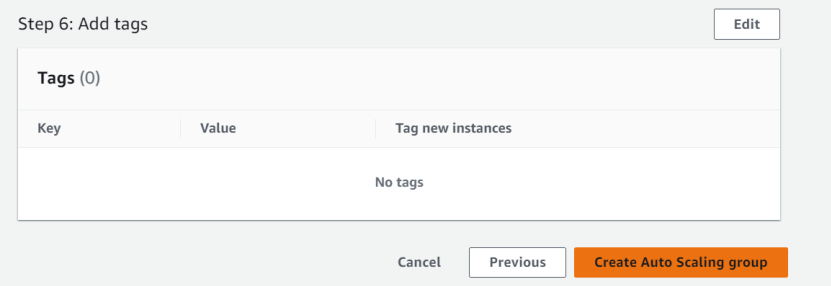
Step 19: Again click on Next.



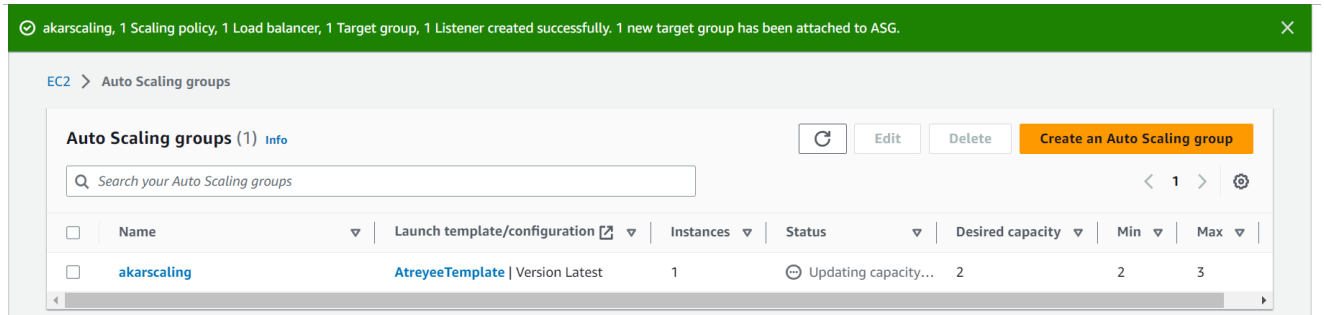
Step 20: Again click on Next.



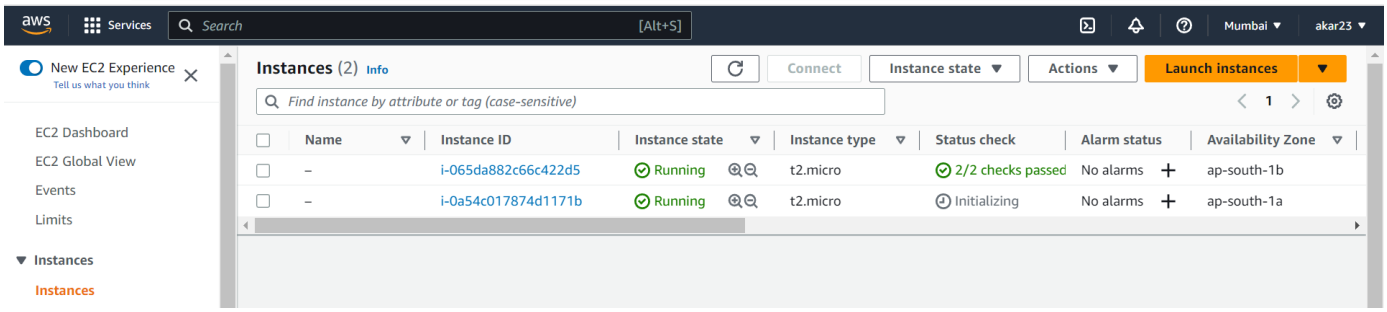
Step 21: Review everything and click on “Create Auto Scaling group”.



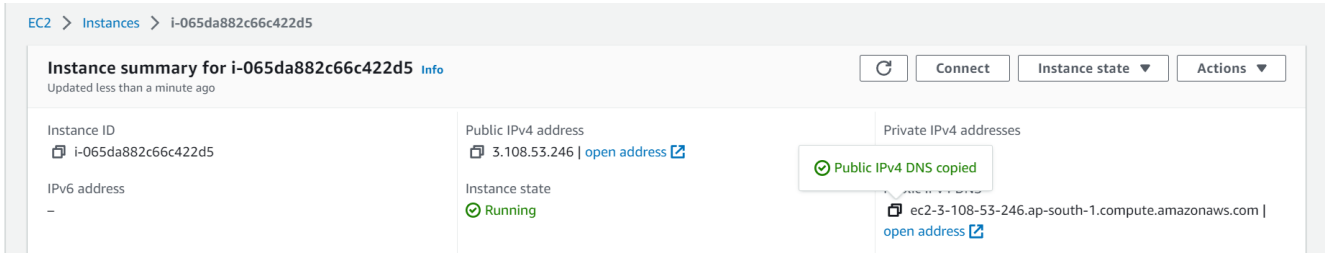
Step 22: Now Auto Scaling group is created.



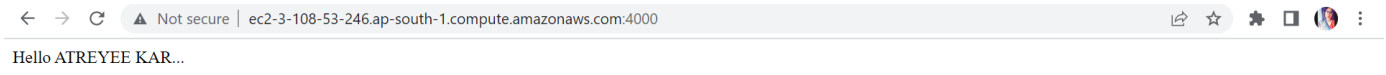
Step 23: Again go to EC2 dashboard. Then go to Instance. You can see that two instance is already created.



Step 24: Open any of the instance and copy the Public IPv4 DNS.



Step 25: Open it in a new browser. Give the port number (:4000) after the address. The content can be seen.

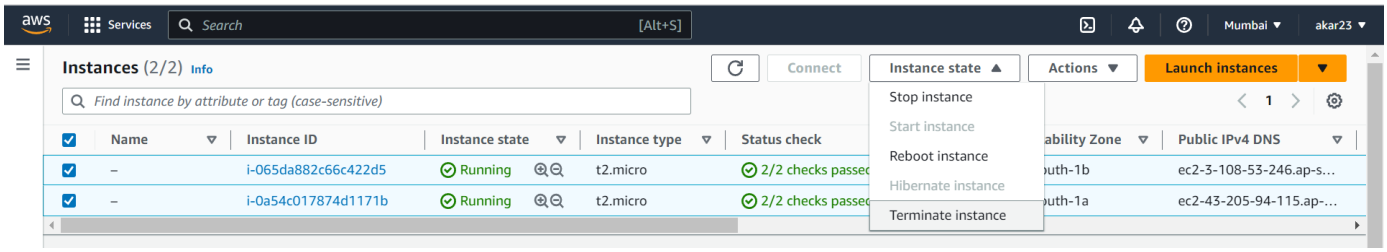


Hello ATREYEE KAR...

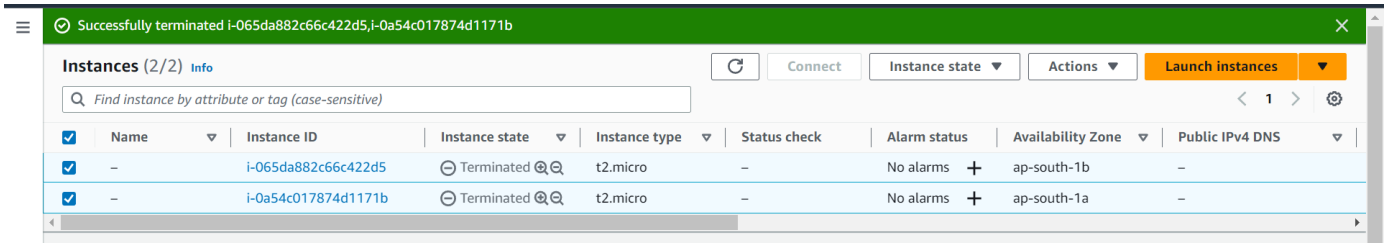
To test whether our Auto-Scaling Group actually works we need to crash or overload the existing instance servers.
Then only our Auto-Scaling Group will provide fresh instance servers automatically in case of crash or it can provide extra servers to handle overloads.

We will now **CRASH THE SERVER INSTANCES** manually by terminating them.

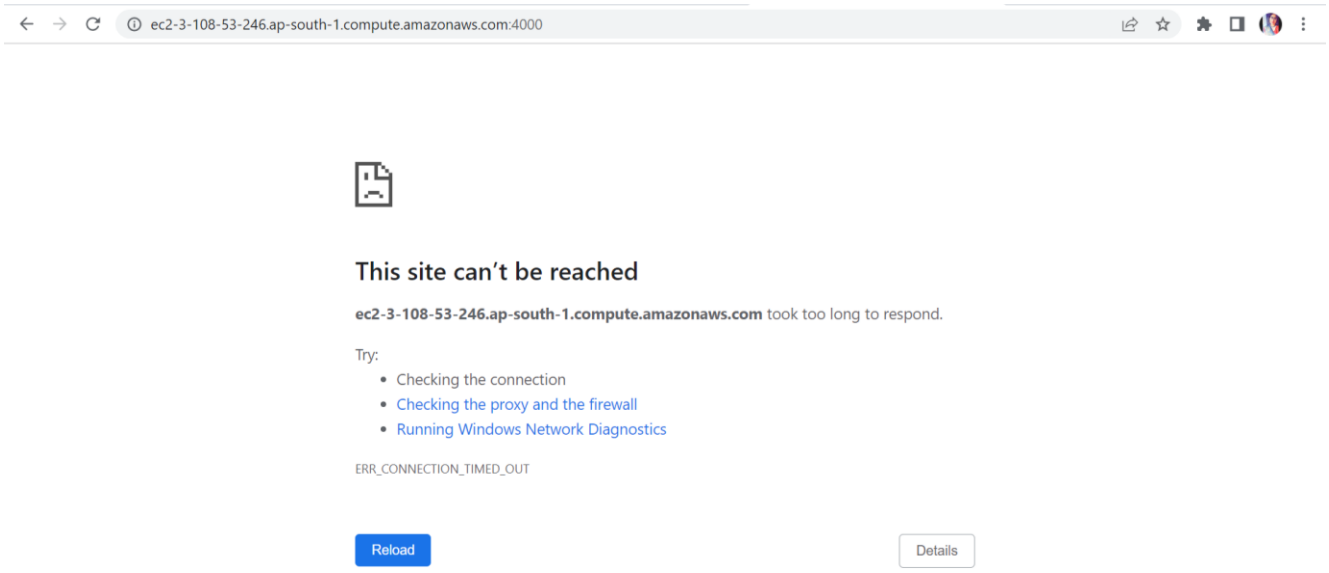
Step 1: Select both the instances. Go to Instance state and select Terminate instance.



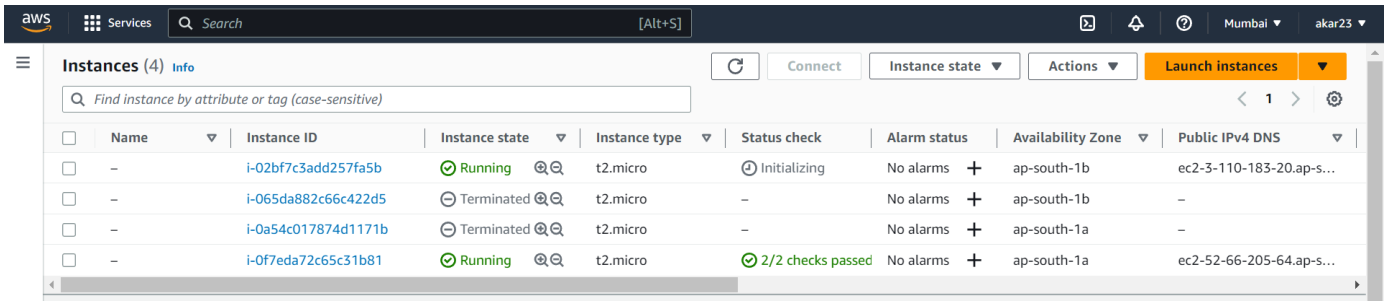
Step 2: Both the instances are terminated.



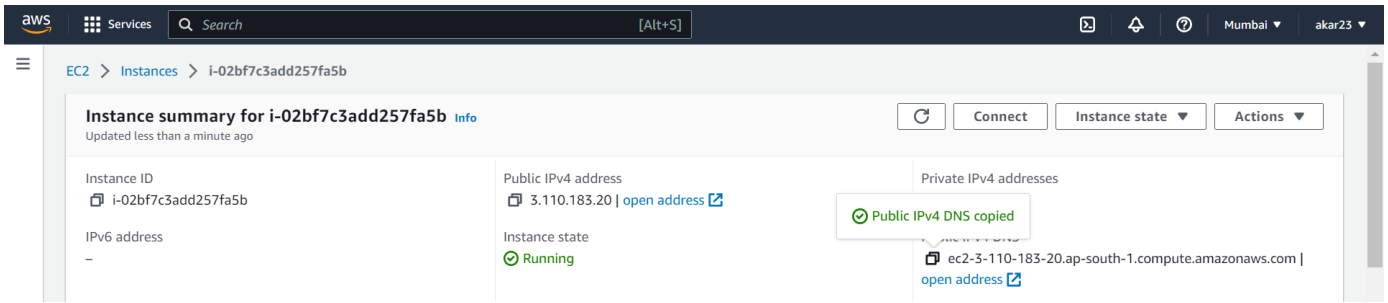
Step 3: Now refresh the browser. Now, we cannot reach the site.



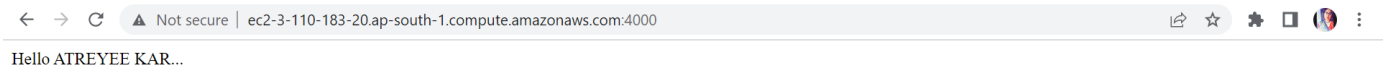
Step 4: Wait for few seconds. Then refresh the Instances page. We can see two new instances are automatically created.



Step 5: Now open any of the running instance. Copy the public IPv4 DNS address.



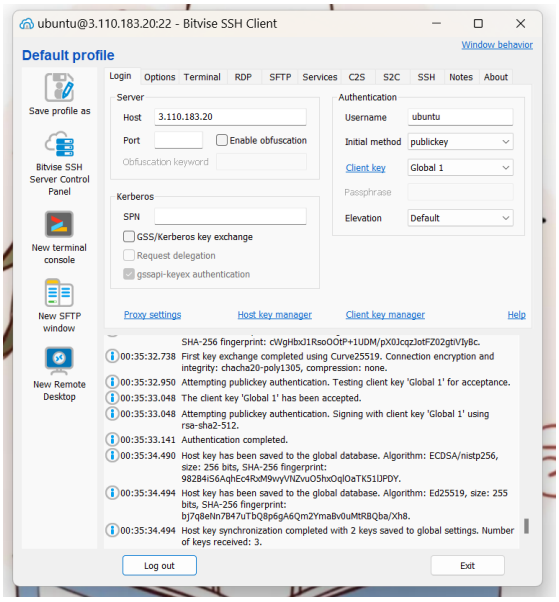
Step 6: Open it in a new browser. Give the port number after the address using ":". Again we can see the content.



So, our Auto-Scaling Group can handle instance crashing by providing new fresh instances.

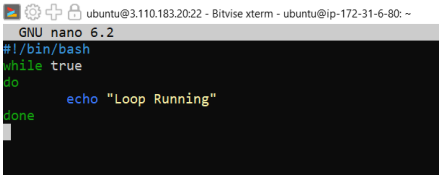
Now, we will **CRASH THE SERVER INSTANCES** by overloading them by running scripts.

Step 1: Now Log in to Bitvise SSH Client.



Step 2: Now go to New Terminal Console and enter the command : **nano infil.sh**

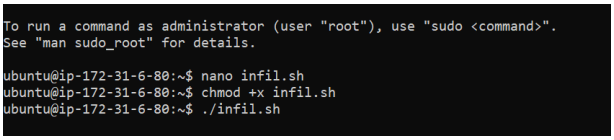
Step 3: A nano Editor will open. Write the following commands in it:



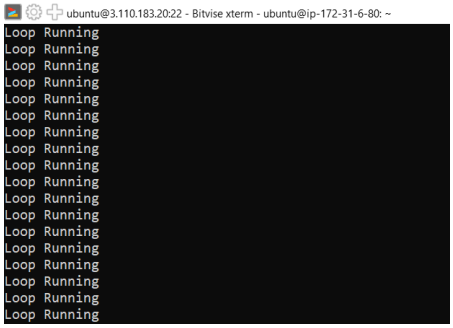
Step 4: Now, to save and close the shell script we need to press the following shortcuts and keys sequentially: **Ctrl X → Y → Enter.** Then you will return to the terminal.

Step 5: Give execute permission to the .sh file. Then execute it.

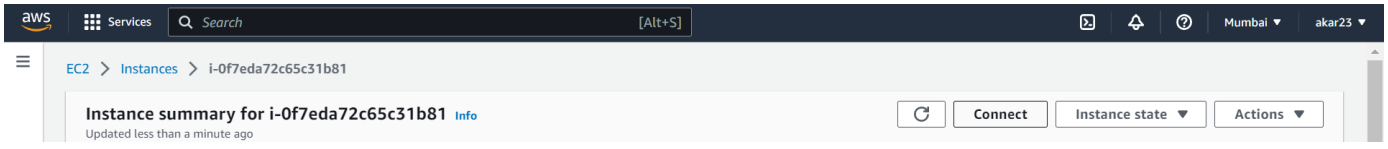
- **chmod +x infil.sh –** To give execute permission to the file.
- **./infil.sh –** To execute the file.

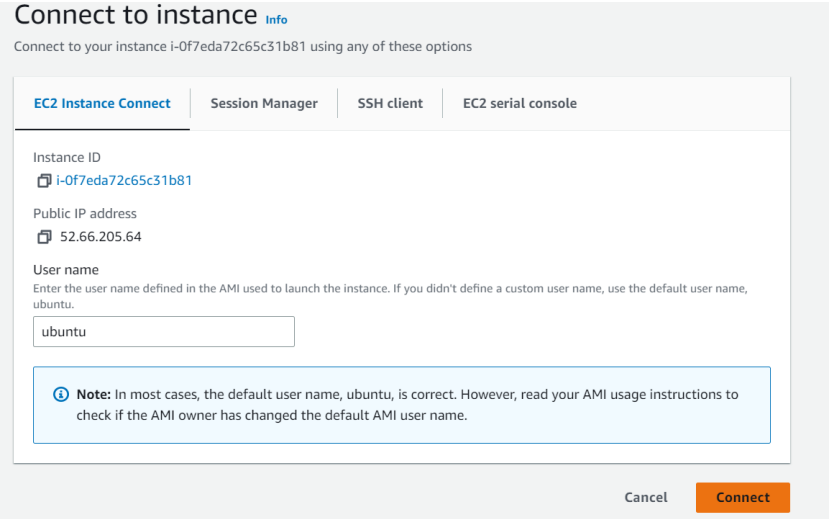


Step 6: The file is executing. Therefore, the first instance is running infinitely. Keep it as it is.

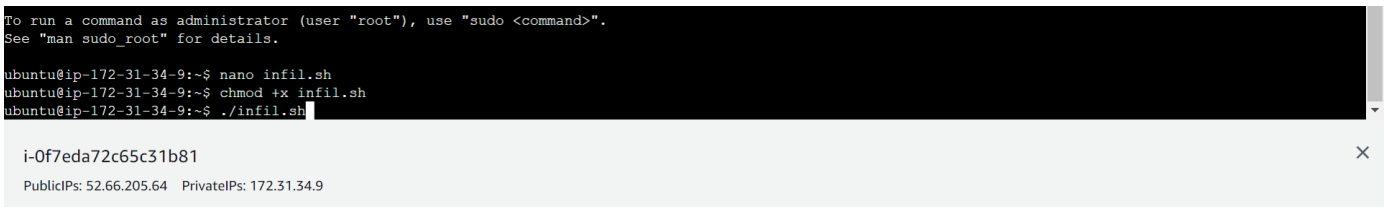


Step 7: Now go to the other running instance. And click on Connect. Then again click on Connect.

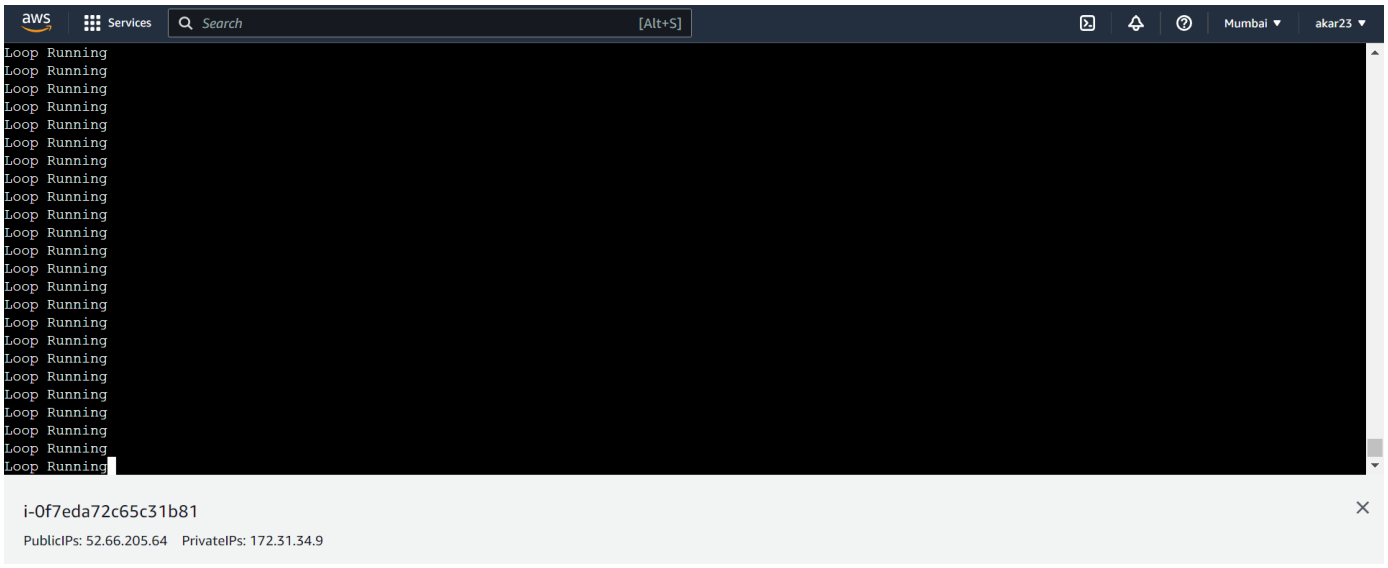




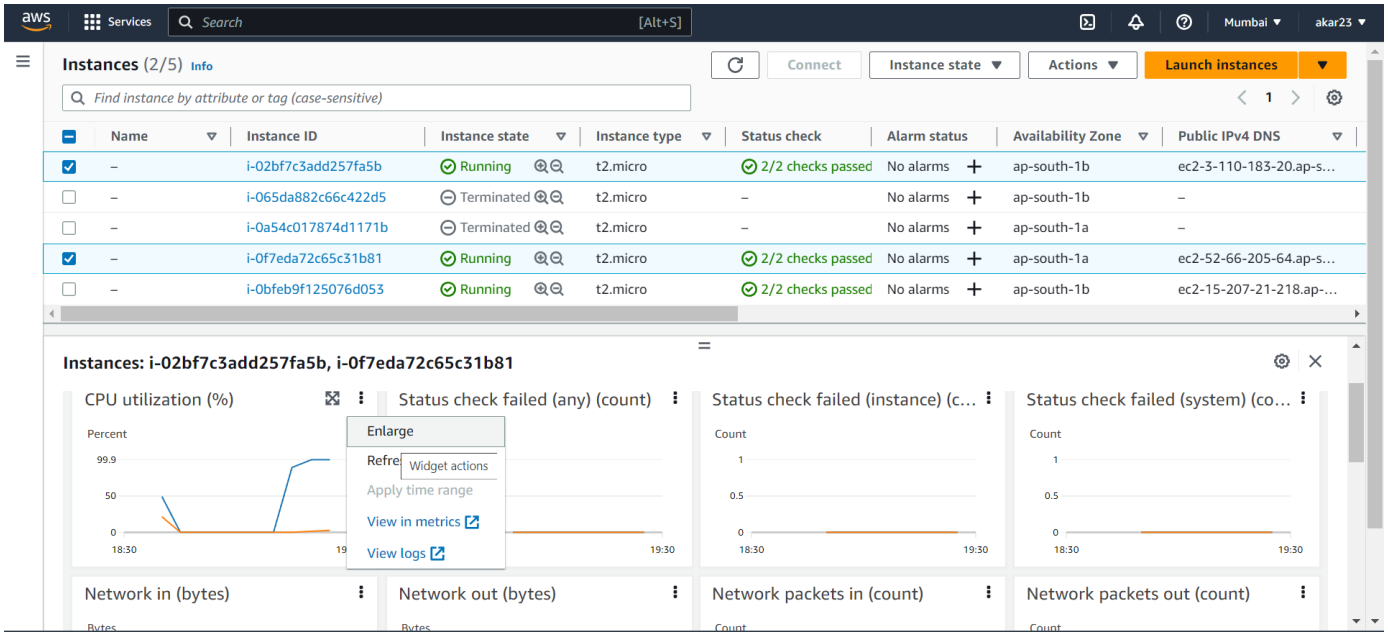
Step 8: A connect terminal will open. Repeat Step-2 to Step-5 here.



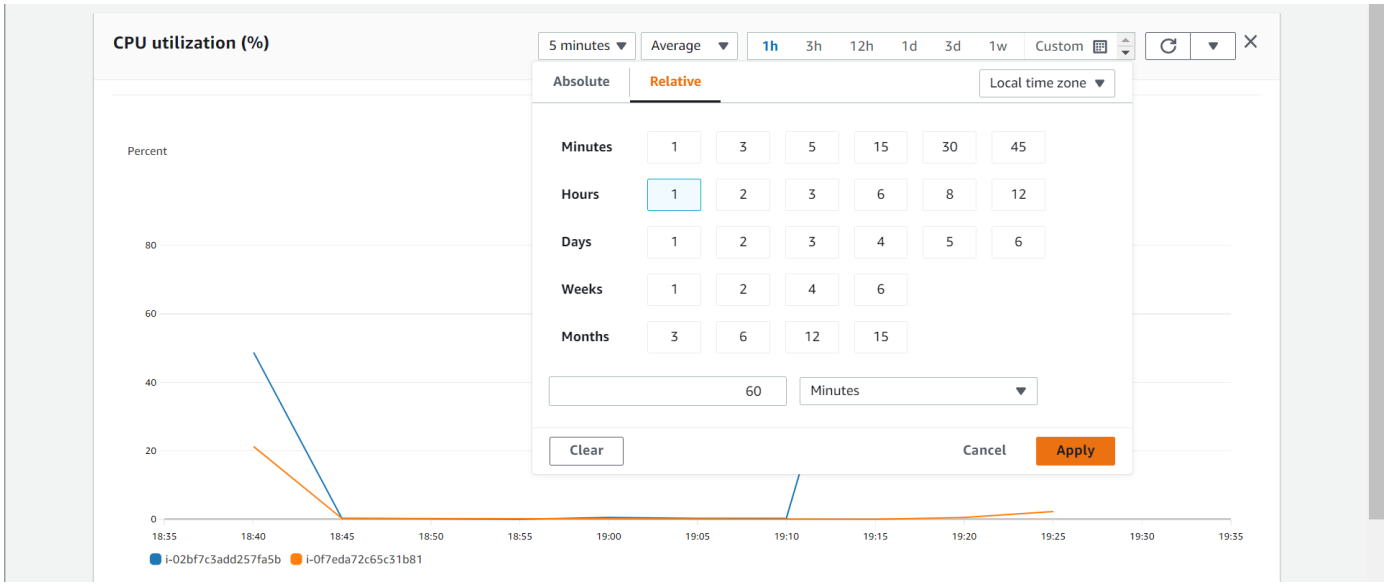
Step 9: Now the second instance is also running infinitely. Keep it as it is.



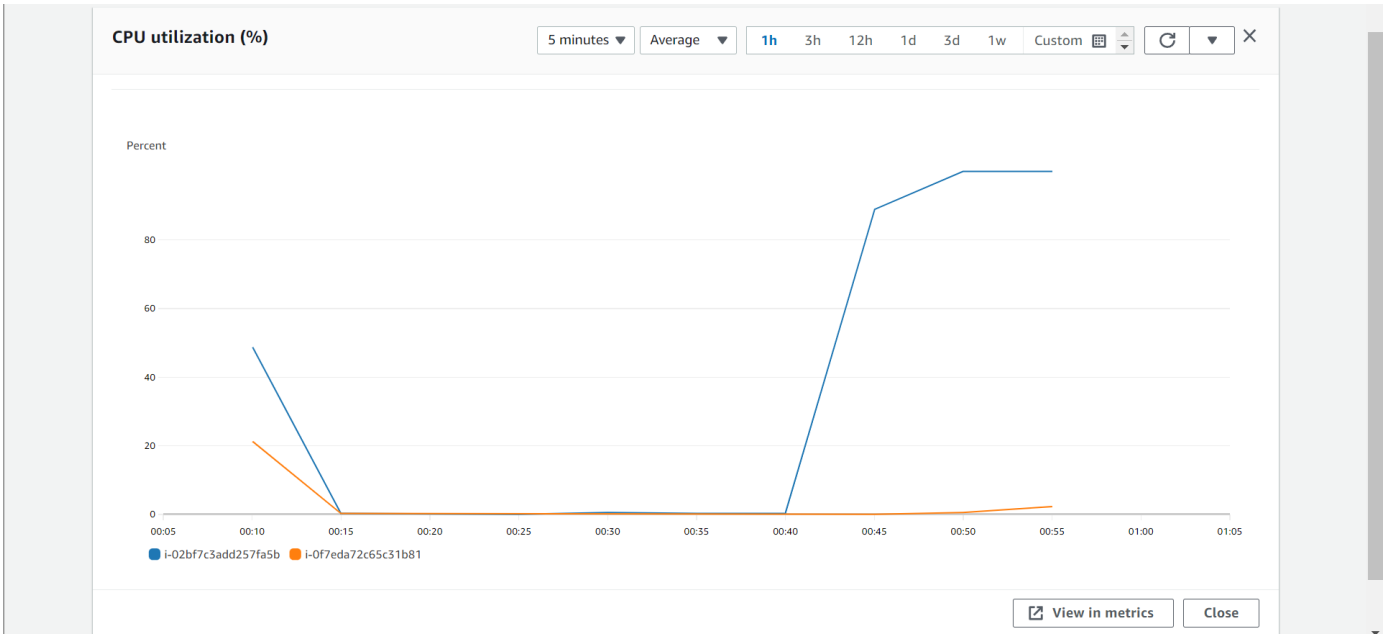
Step 10: Now go to Instance page. Select both the instances. Then scroll down in the "Monitoring" section. Go to CPU Utilization. Click on the three dots. Click on Enlarge.



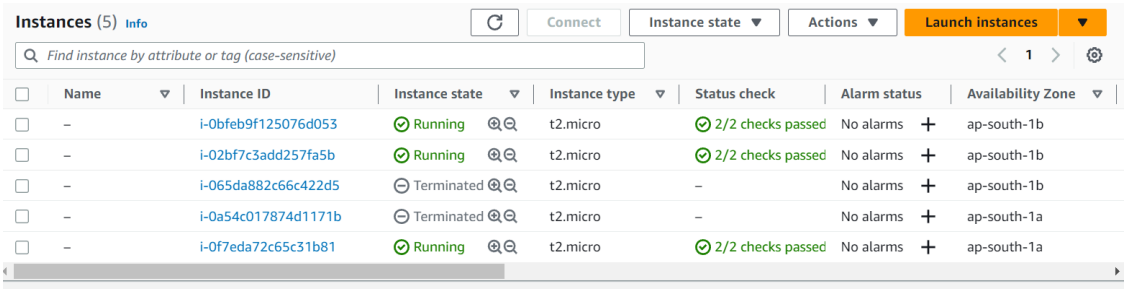
Step 11: Go to Custom. Select Local time zone. Then click on Apply.



Step 12: We can see our first instance has already reached 50% utilization.



So, our Auto-scaling group has created another instance to compensate for the overload.



Our webpage was not at all disconnected in this whole process. You can check it by refresing the webpage.

If we want to delete the instances permanently, we have to follow the sequence given below. Otherwise, the loop will continue and the instances will be created automatically.

Deleting Sequence:

- Delete Auto-Scaling groups
- Delete Load Balancers
- Delete Target groups
- Delete EC2 Instances

Thus, we have built Scaling plans in AWS that balance load on different EC2 instances.