In this hands-on will check how Auto scaling group works

For this we require-

Launch configuration

Target group &

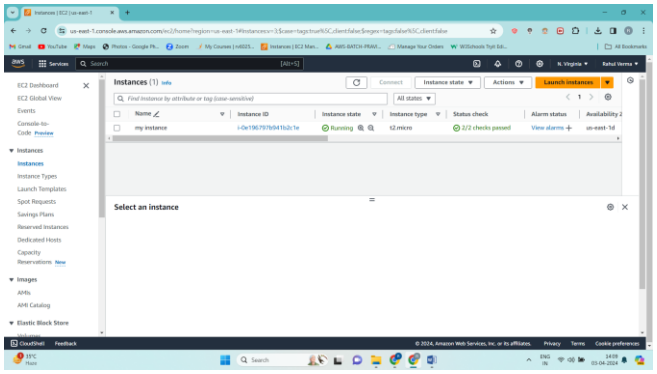
Load balancer & cloud watch alarm

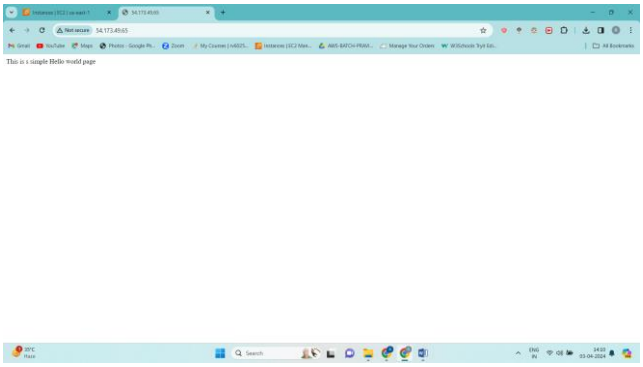
Prerequisite:

1) An ec2 instance having web page

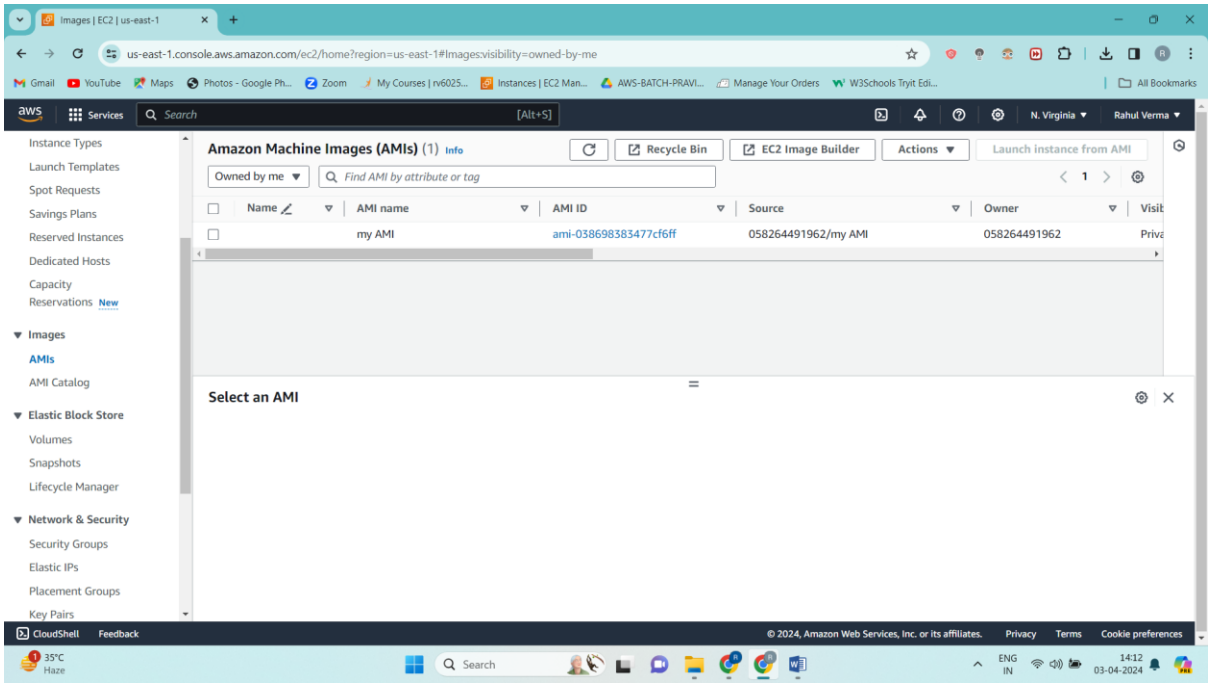
2) Create image of this instance using AMI

1)Instance having web page



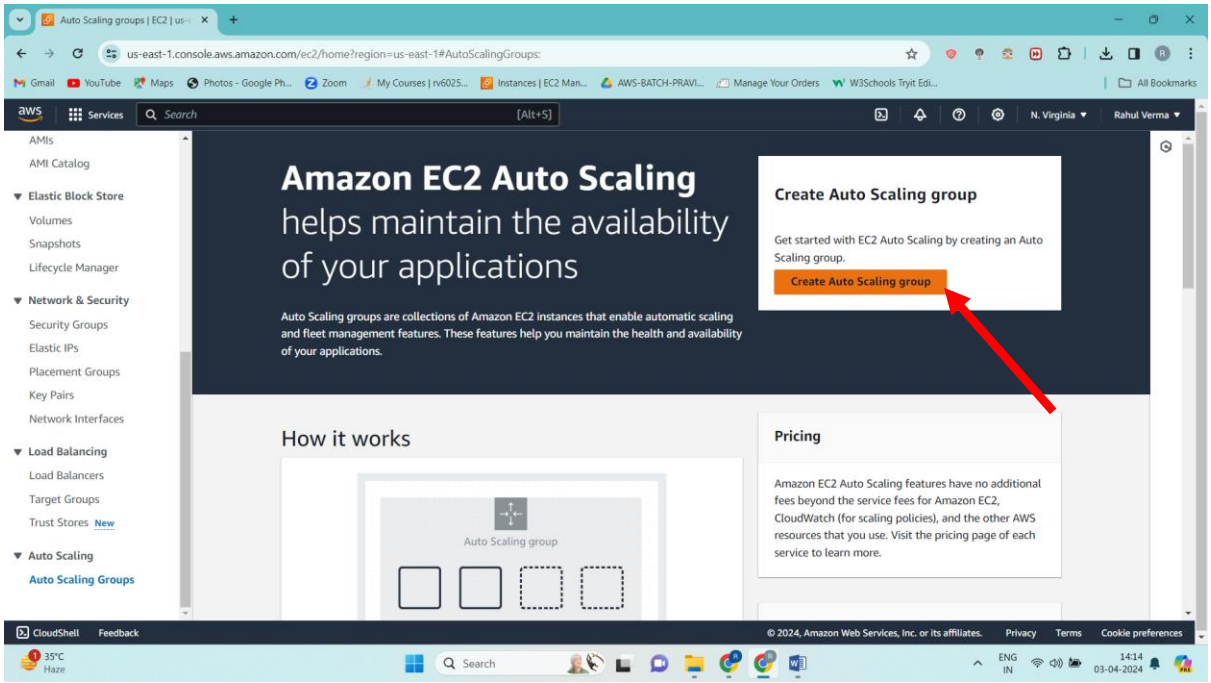


2) AMI of the above Instance

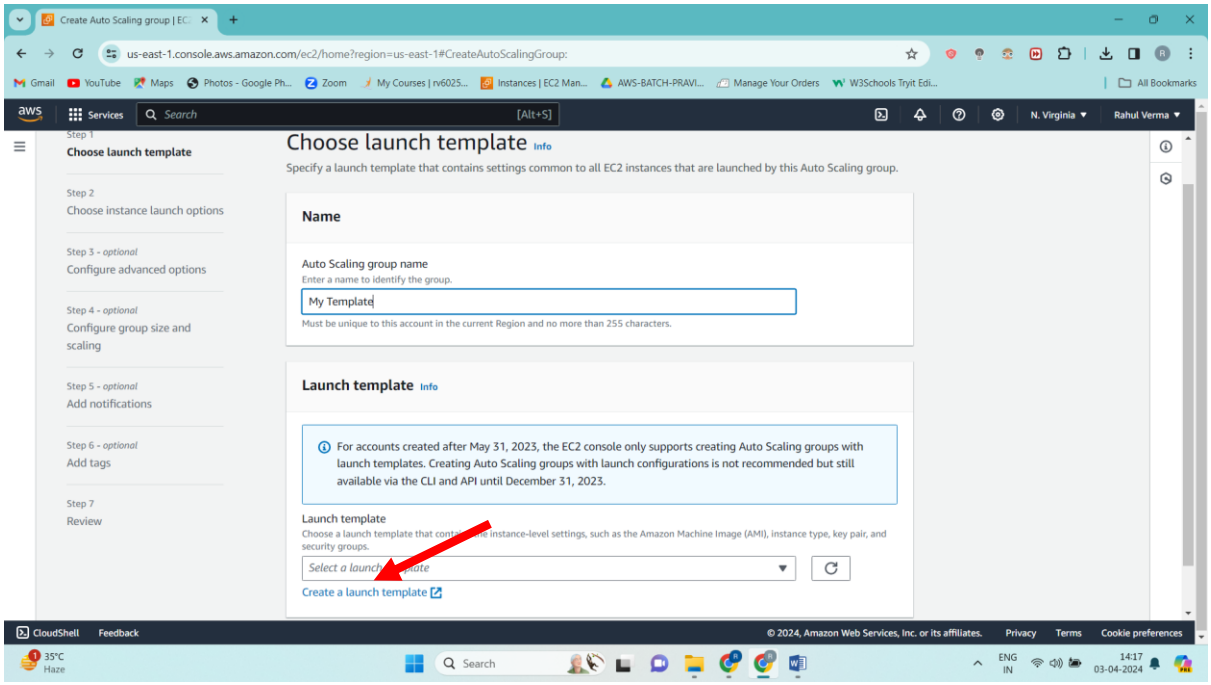


Let’s start with creating launch configuration

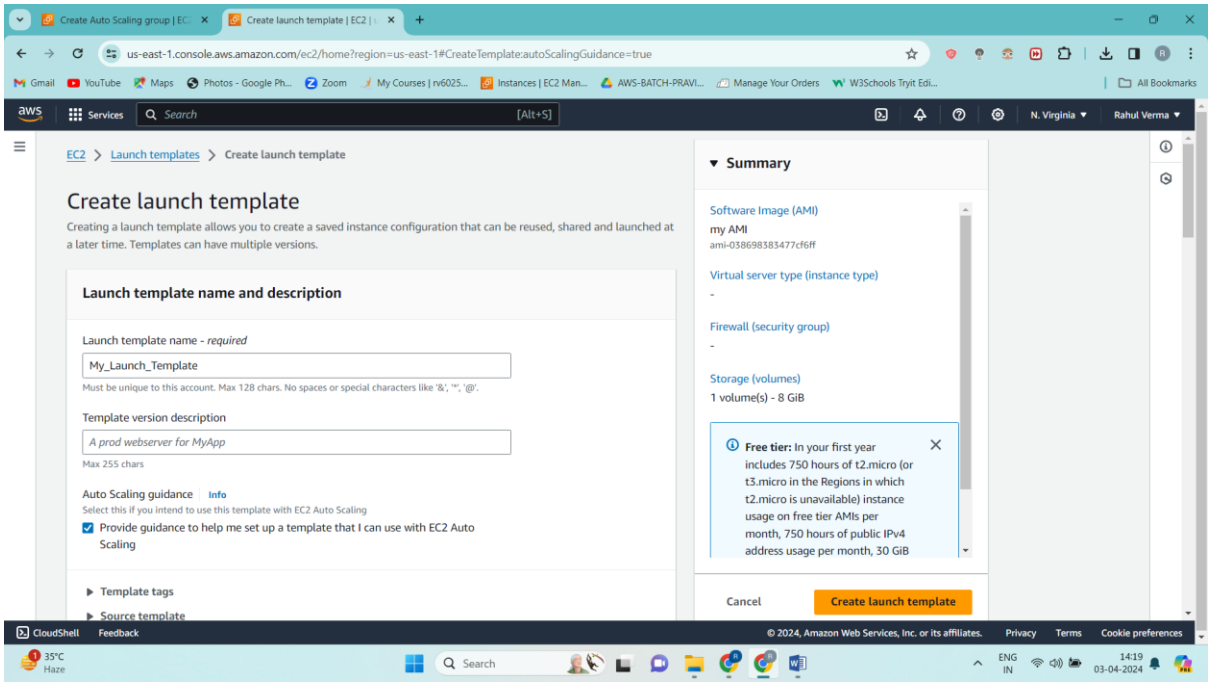
Step 1: scroll down till bottom and click on Auto scaling and click on Auto scaling groups



Step 2: Define it’s name and click on create a launch template

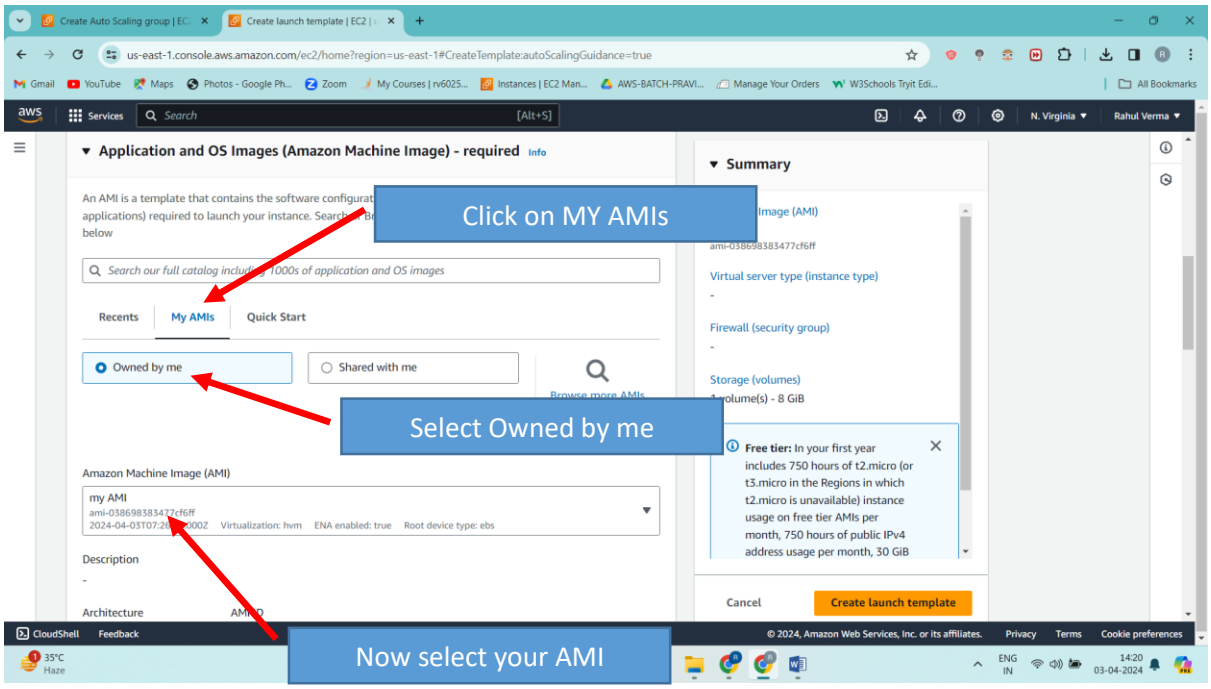


Define template name. I’m giving **My_Launch_Template**

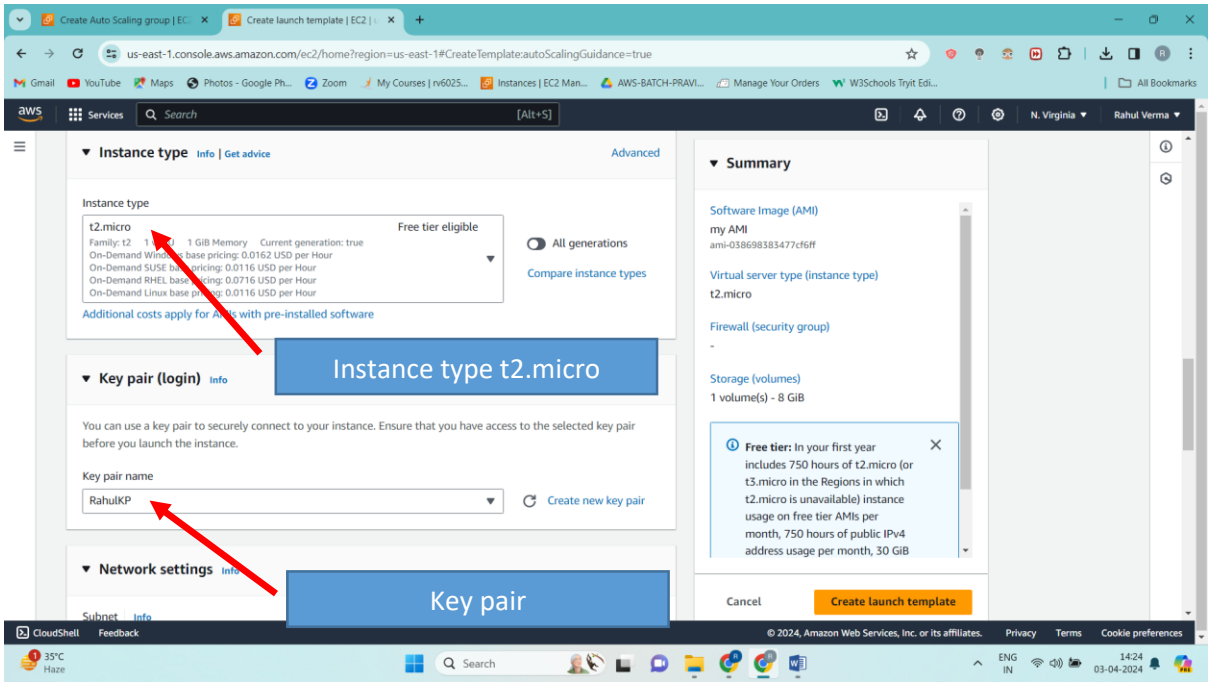


Step 3: Under Application and OS images

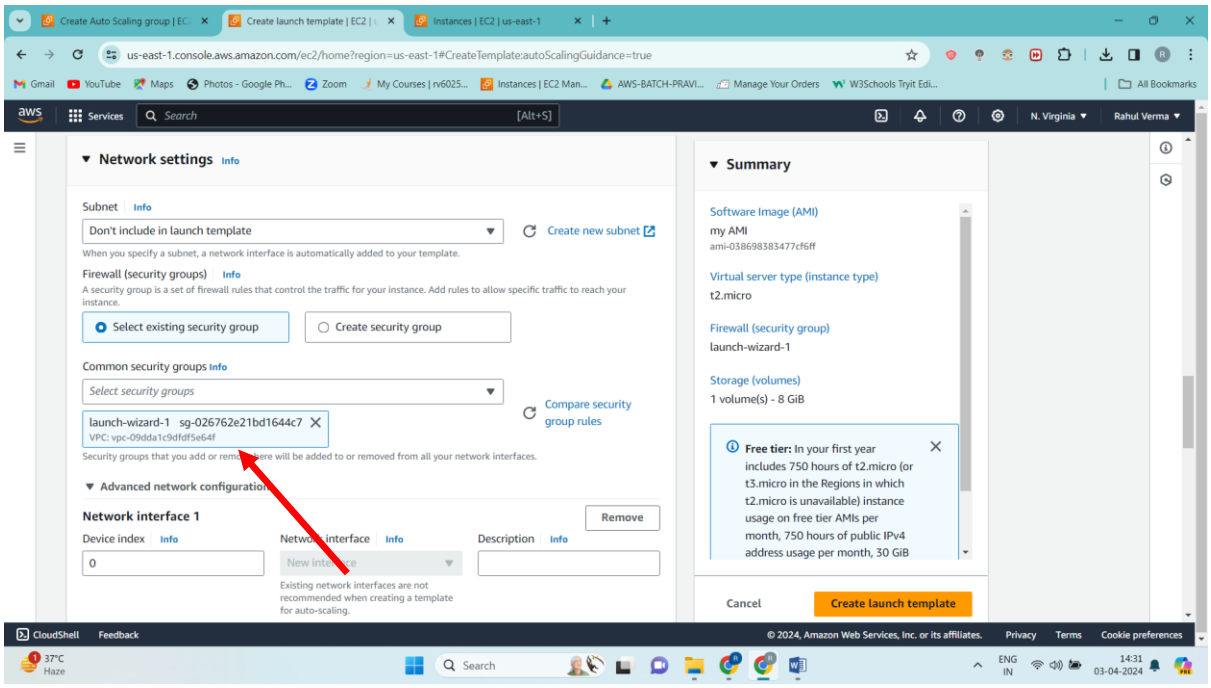
Select your AMI



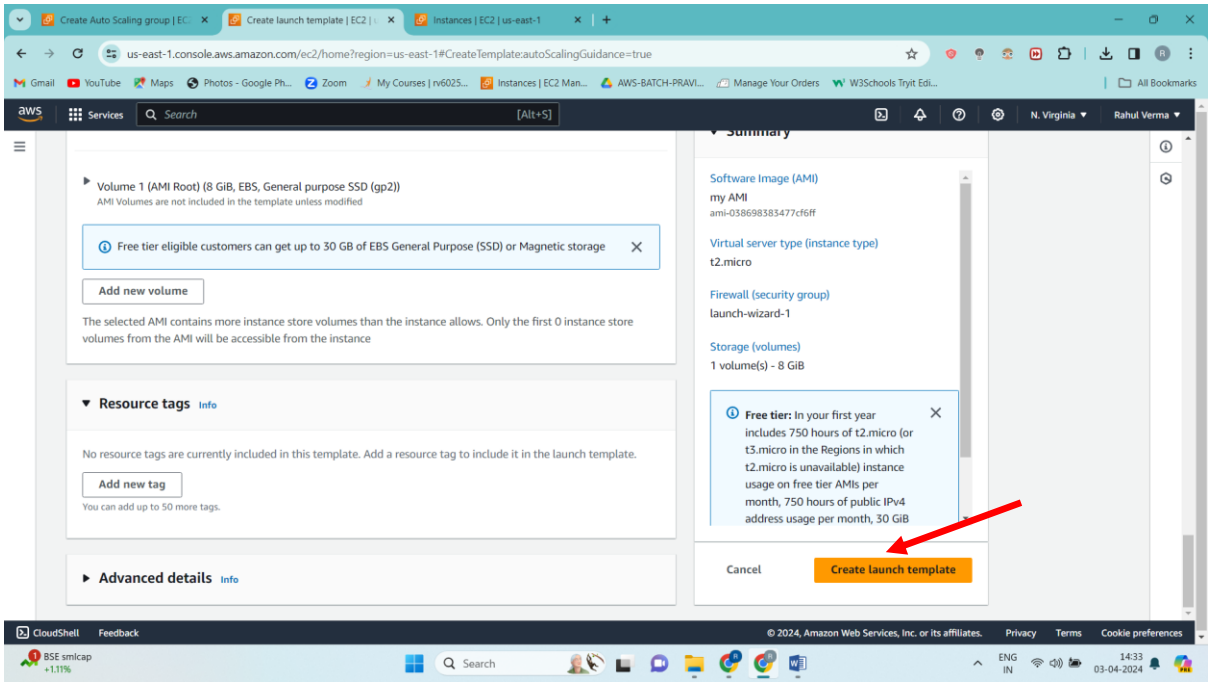
Step 4: Select instance type (for free tier select t2.micro) and select key pair



Step 5: select security group



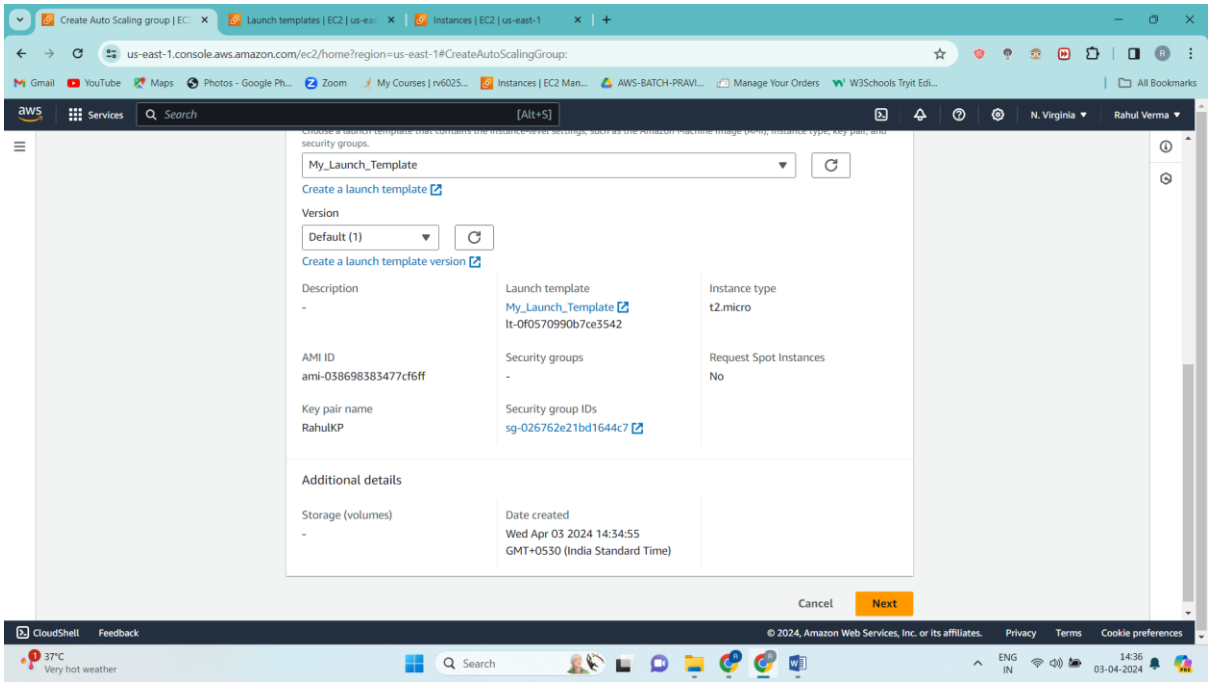
Rest default setting and just click on create launch template



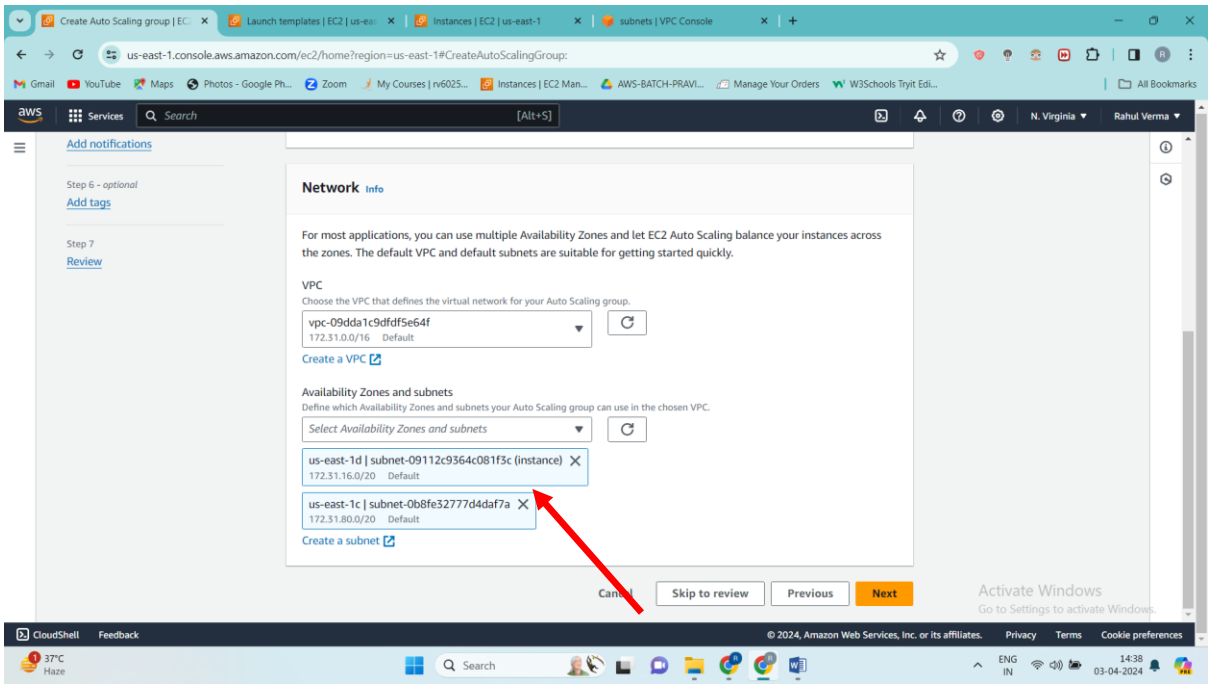
Step 6: now come back to Auto scaling creation page.

Click on this button 

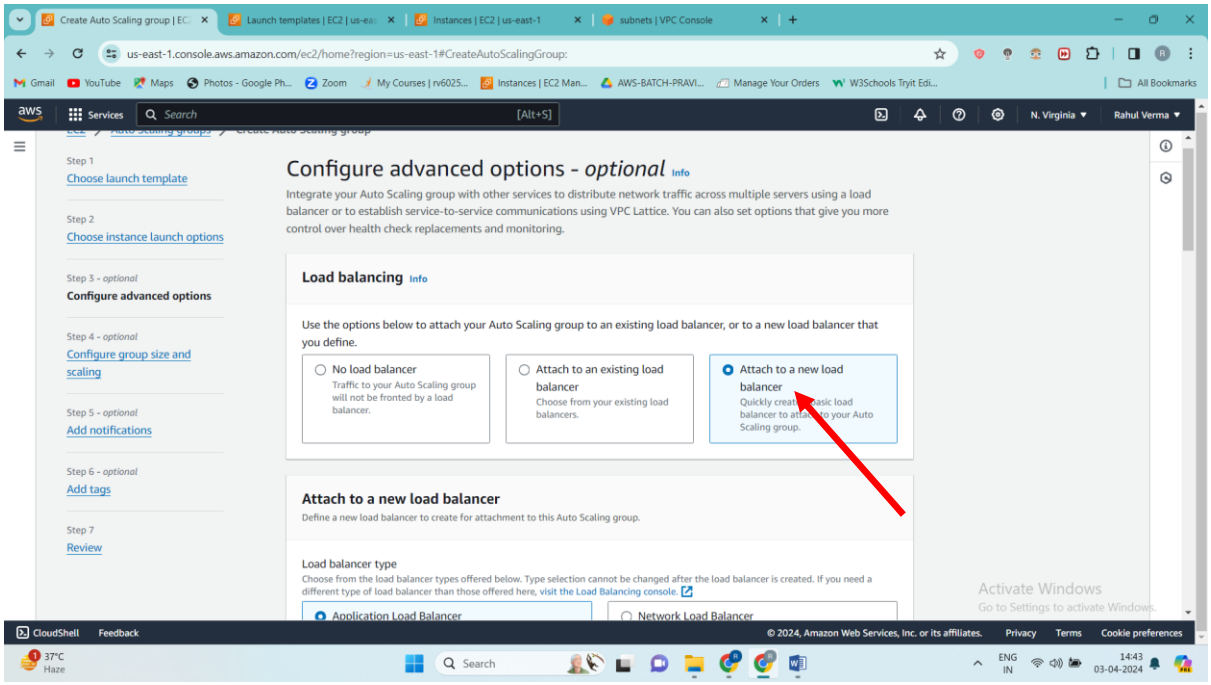
And select your launch template which you have created just now.



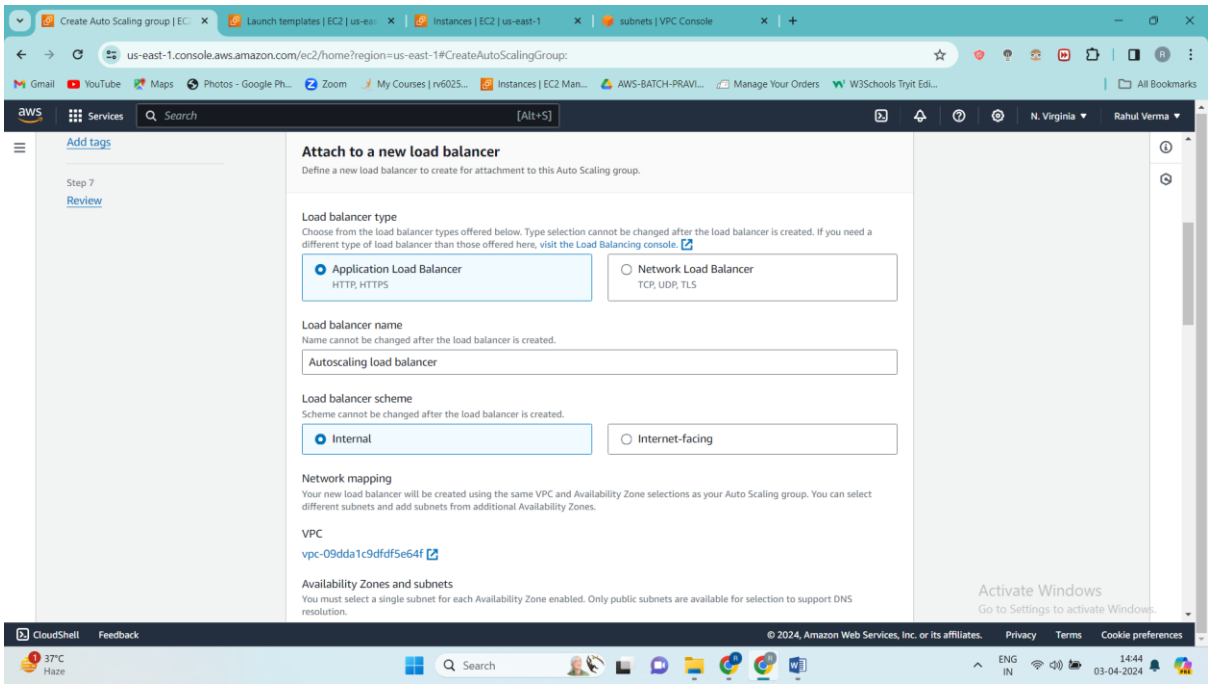
Select availability zone and click on Next



Step 7: now we have to create one Load balancer for that click on attach to a new load balancer

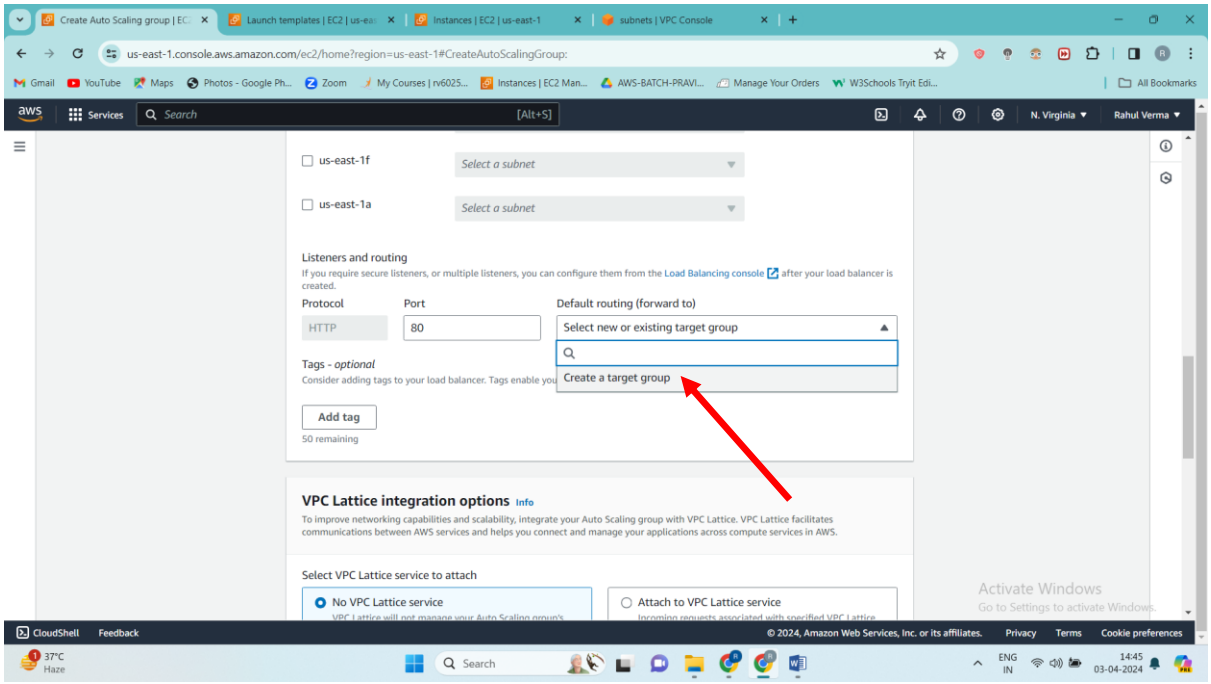


Setup your load balancer

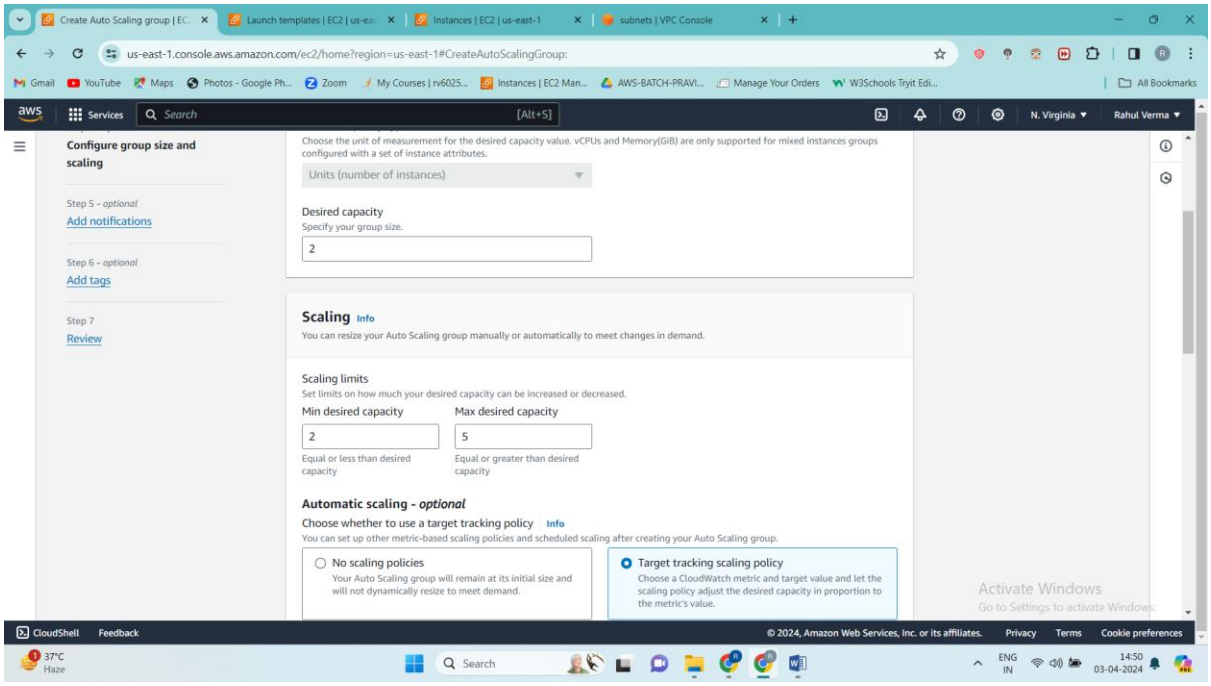


Now create Target group for you load balancer

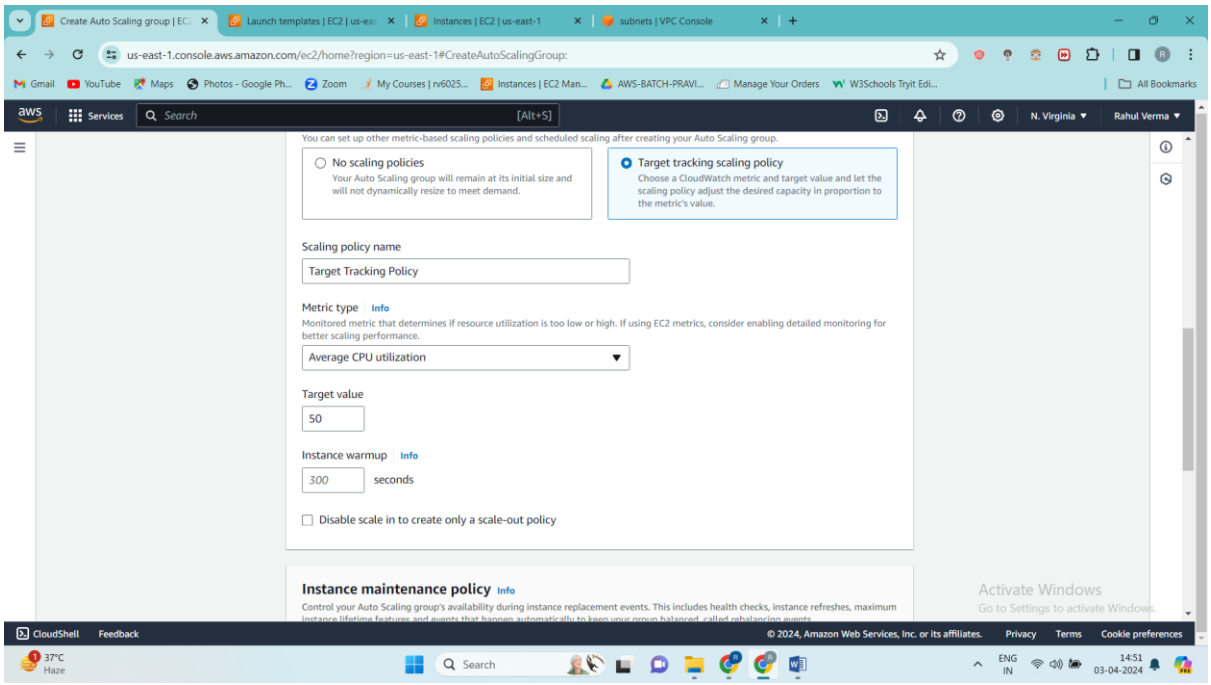
Click on create a target group and click on next



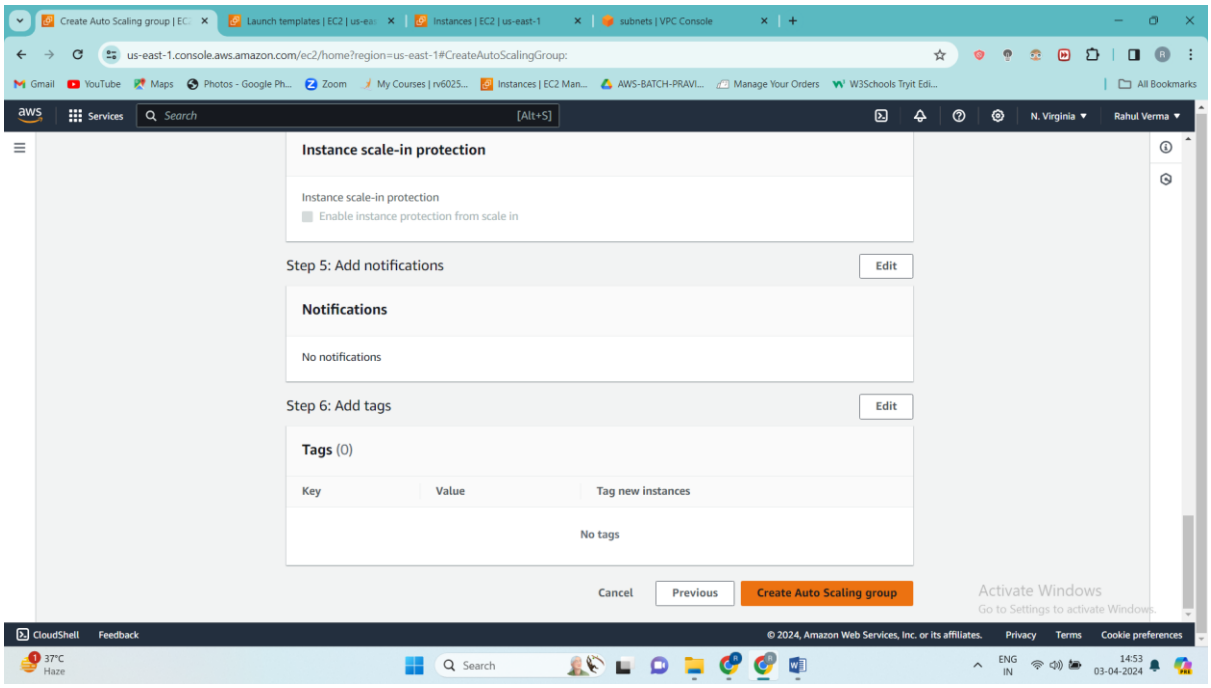
Step 8: Now configure how much desired instances you want for this demo. I'm giving desired capacity 2 and min 2 instance and max 5 instances



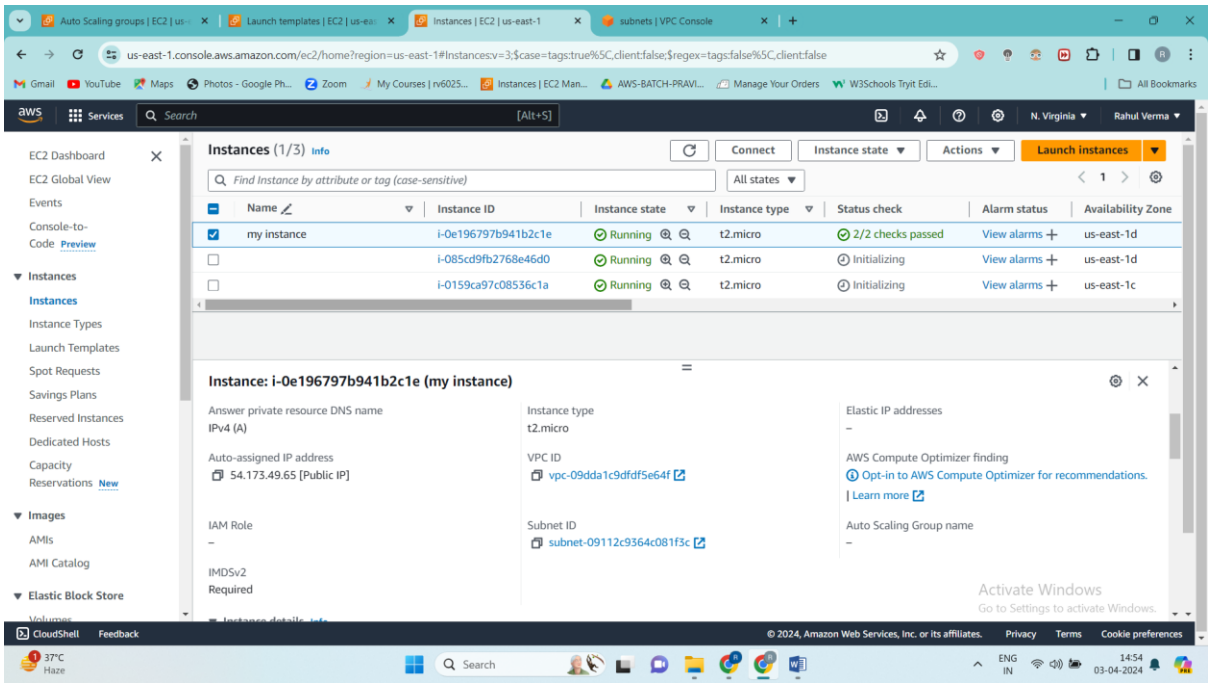
Select Target policy.



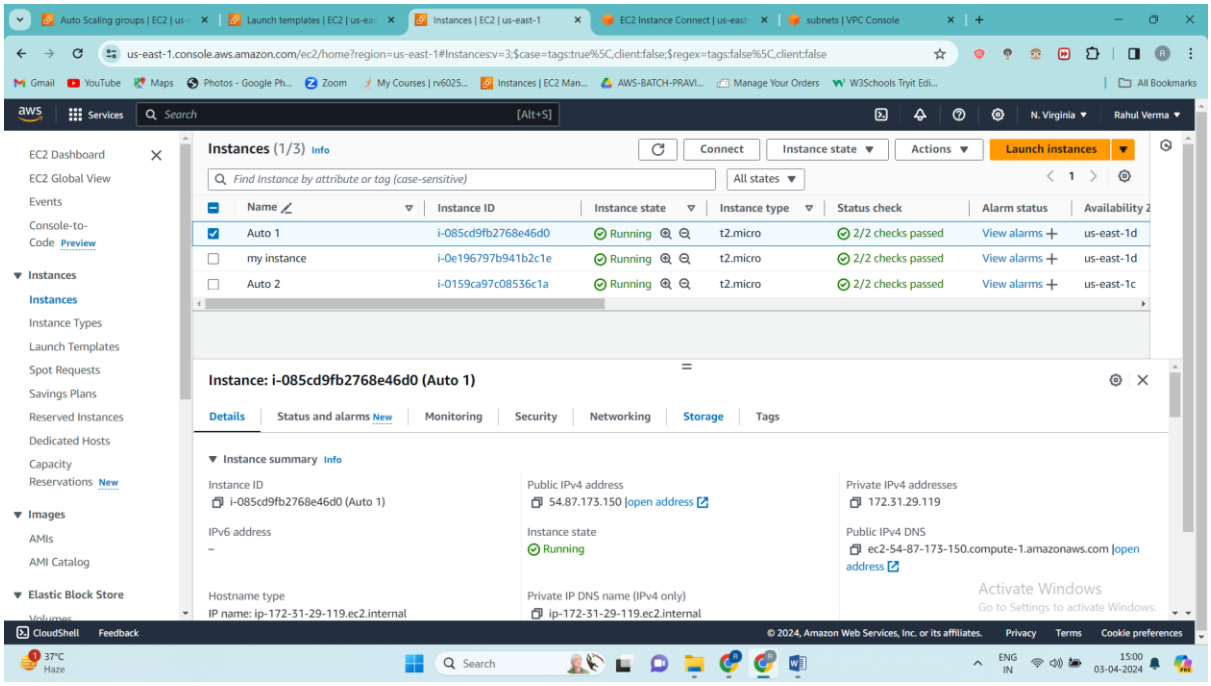
Now click on Next button until it shows create auto scaling group button and click on it.



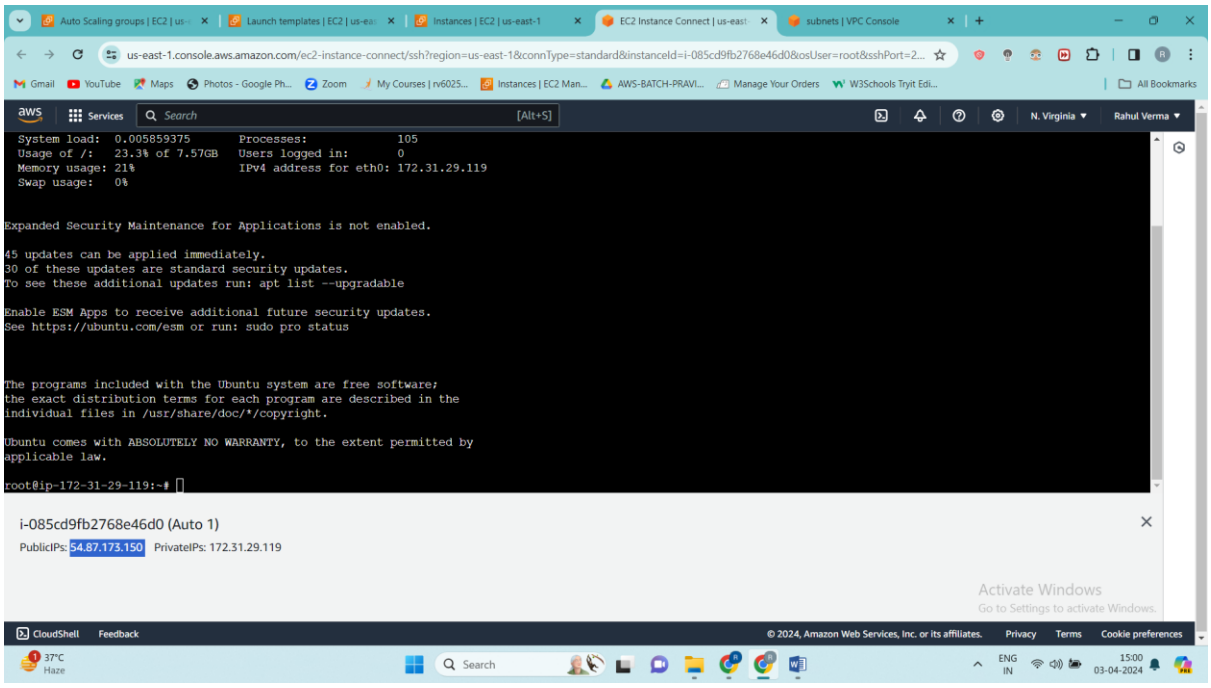
Step 9: now you can see 2 instances are created by auto scaling because we have given desired capacity as 2



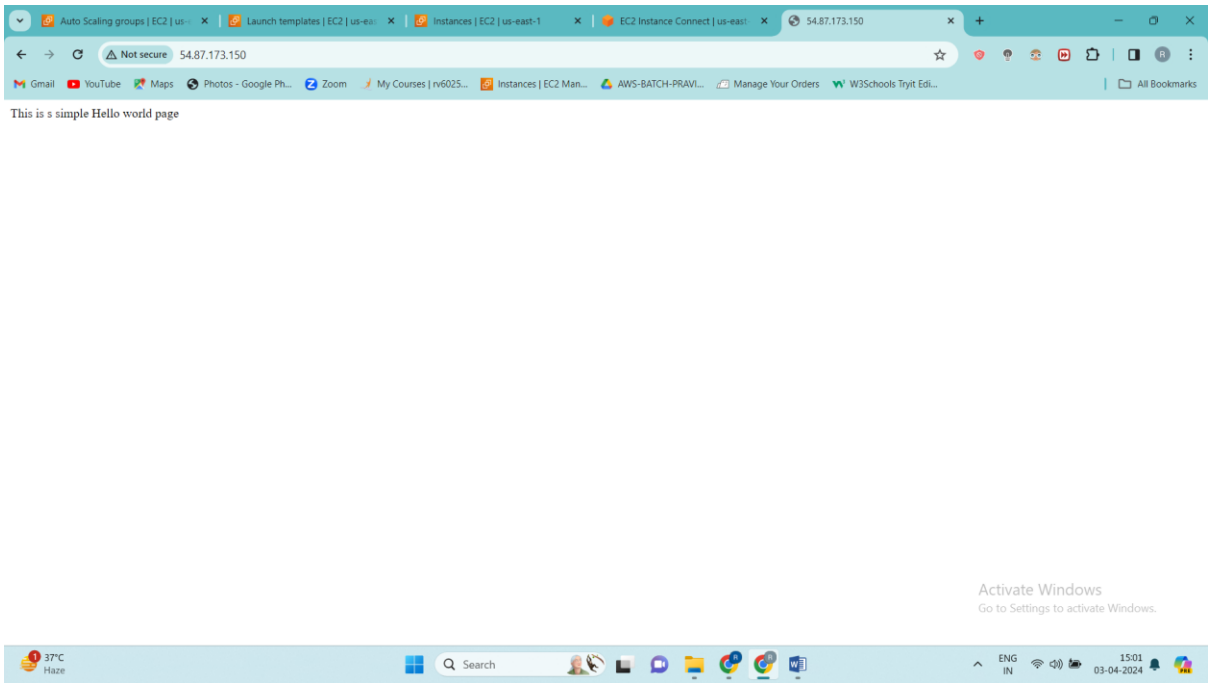
Now let’s connect to one of these instance I have define them as Auto 1 & Auto 2. Let’s connect to Auto 1 and check if our web page is working or not.



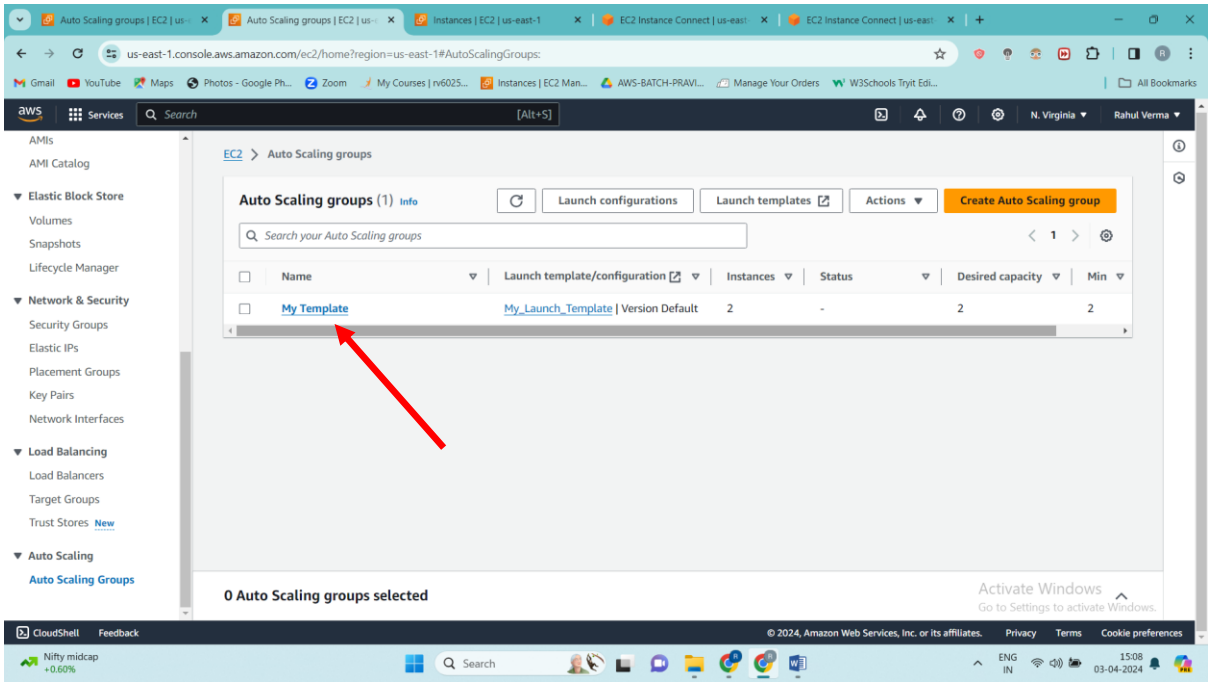
We are connected to Auto 1



And our web is also working good

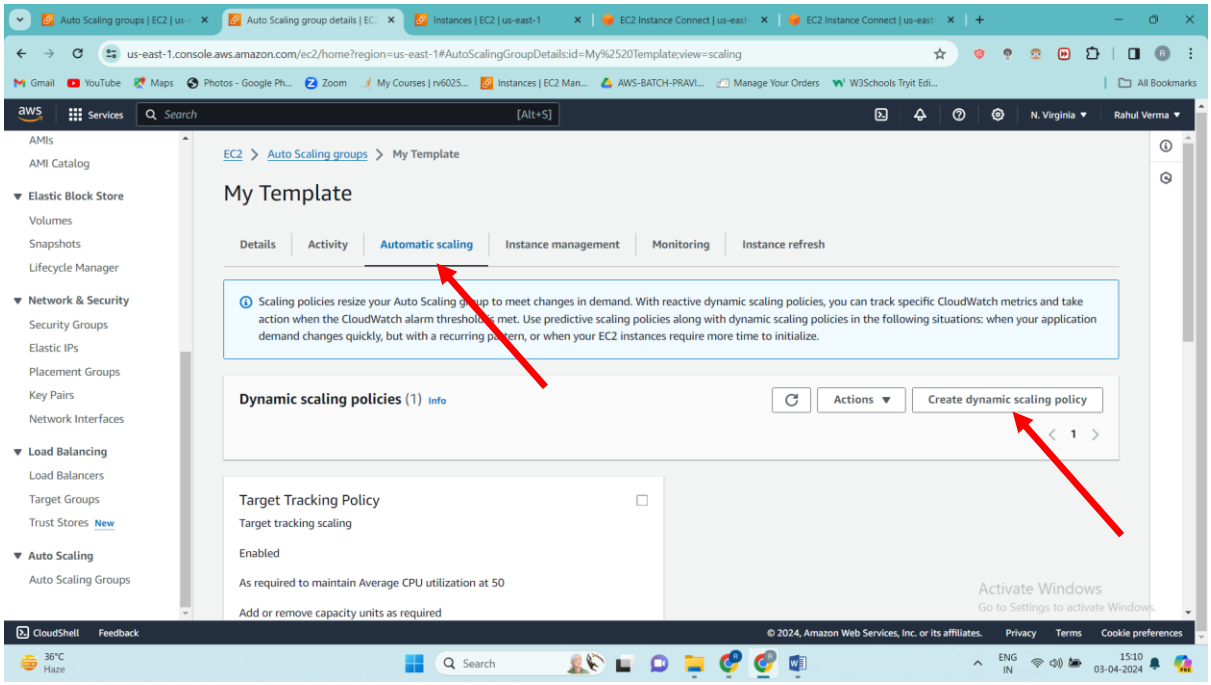


Step 10: now go back to your Auto scaling group click on it

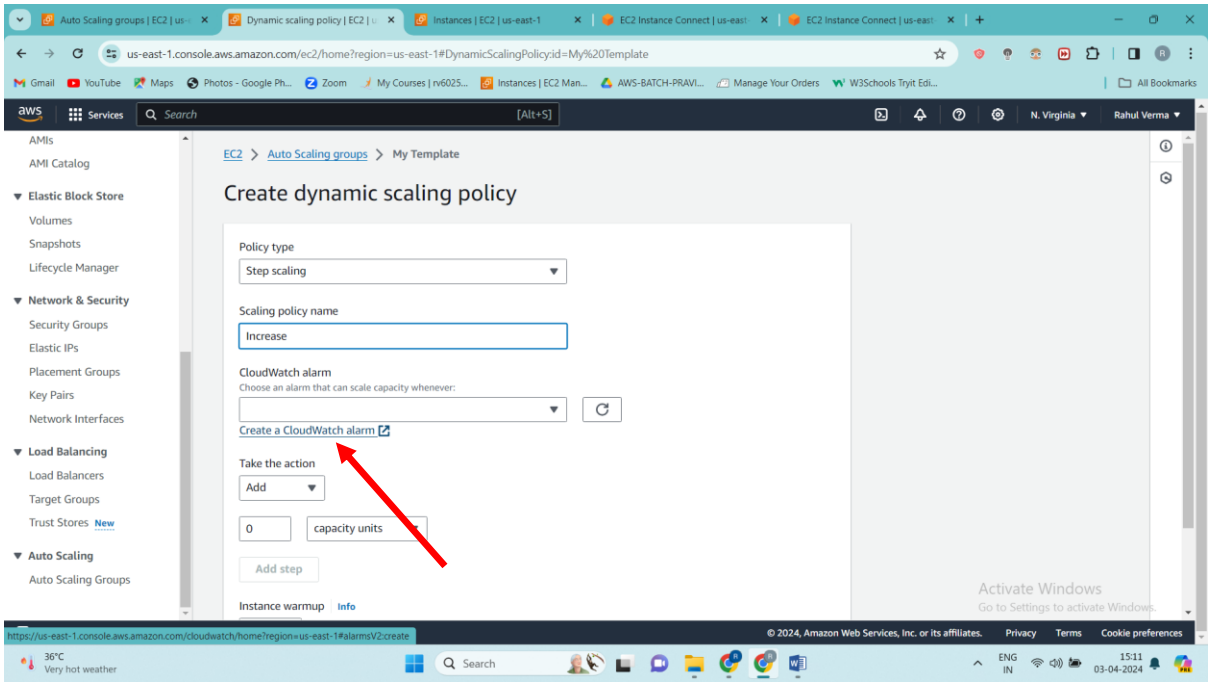


Now go to automatic scaling

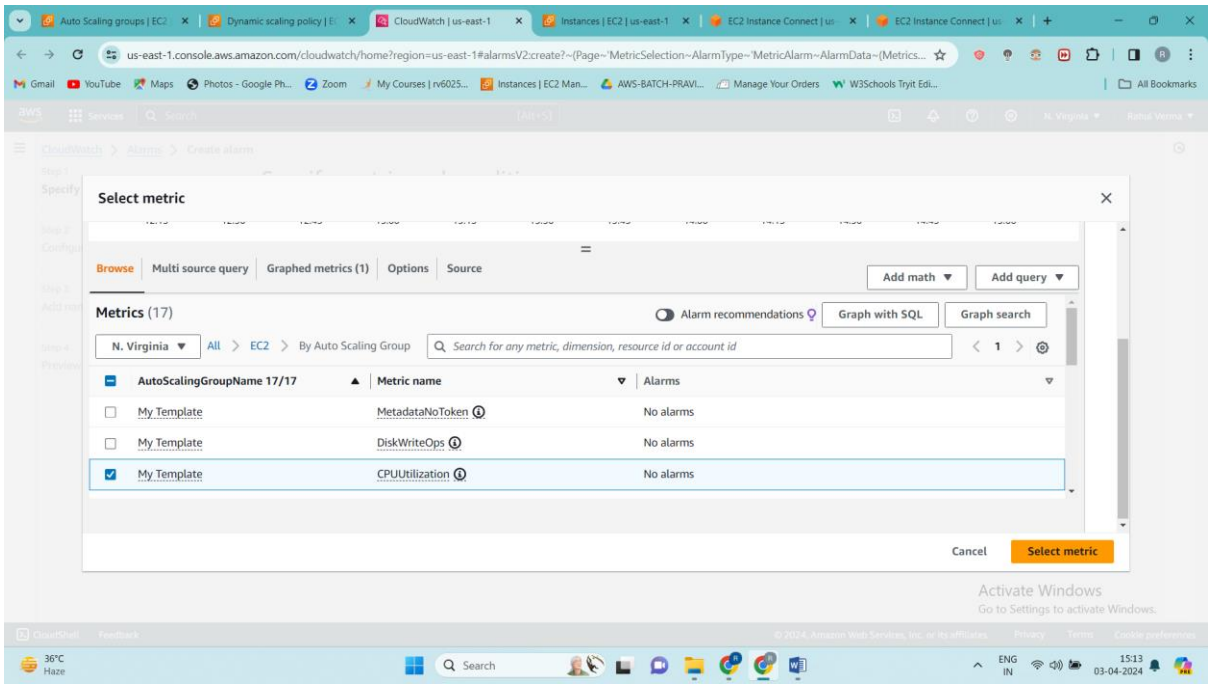
Under that click on create dynamic scaling policy



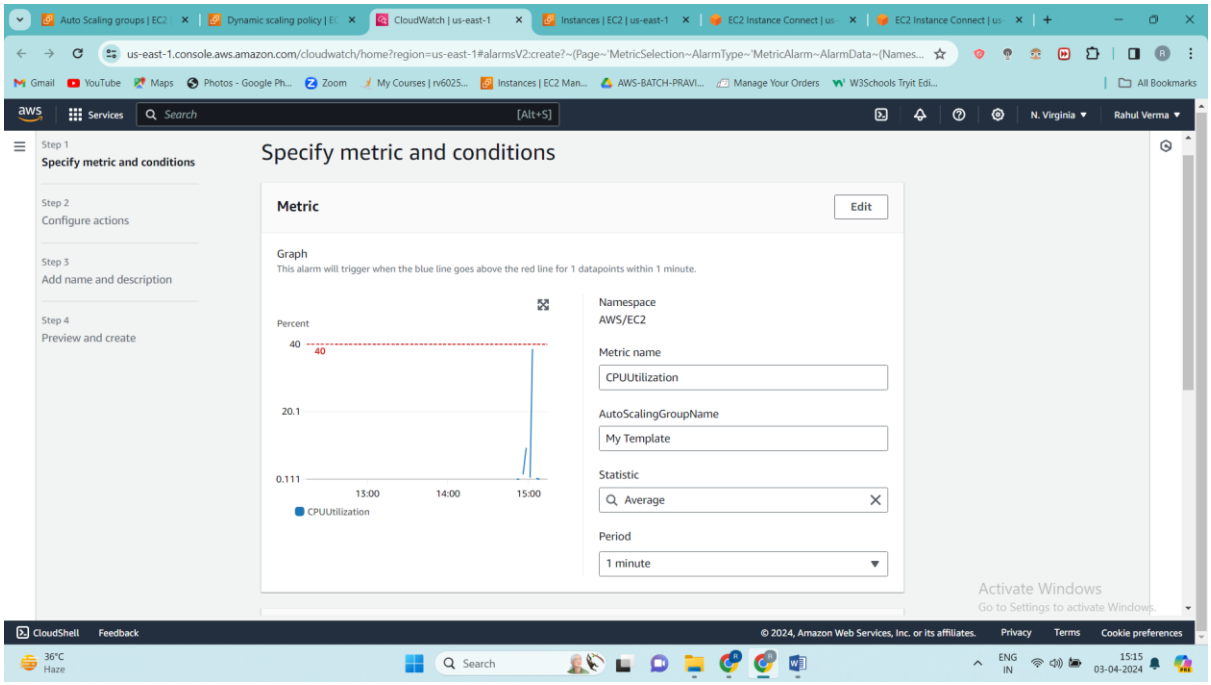
Step 11: define it's name and click on create a cloud watch alarm.



Now browse metrics and select My template (auto scaling group name) CPU utilization

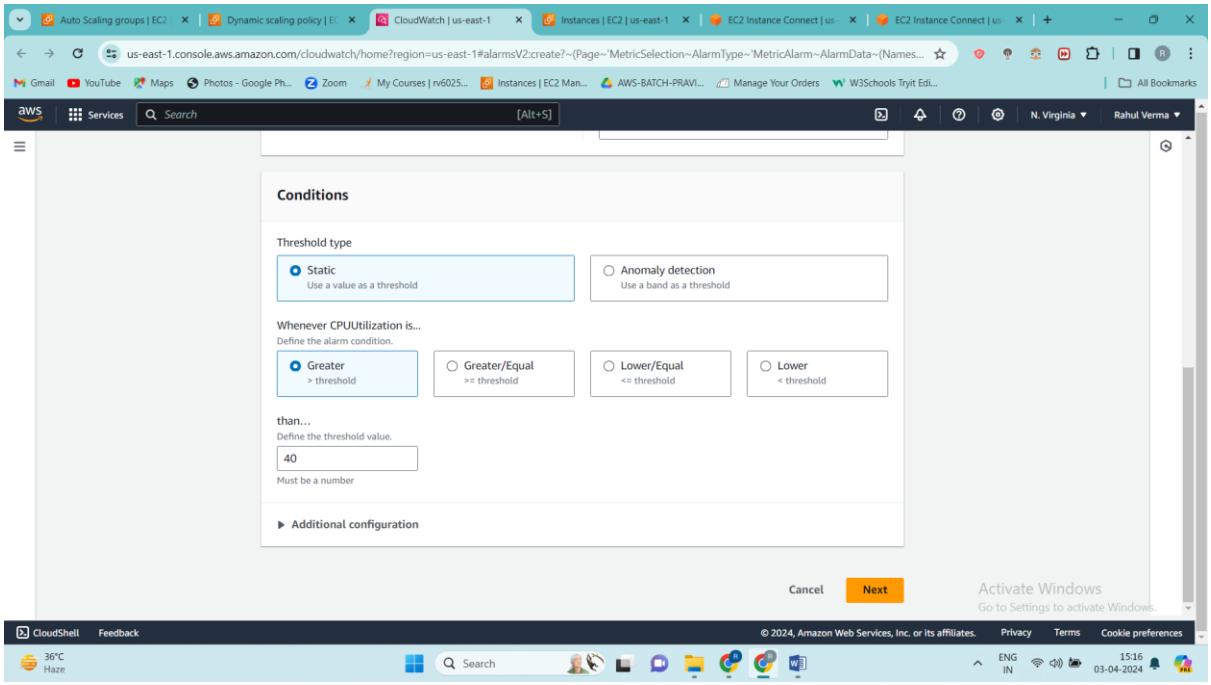


Select the desired settings

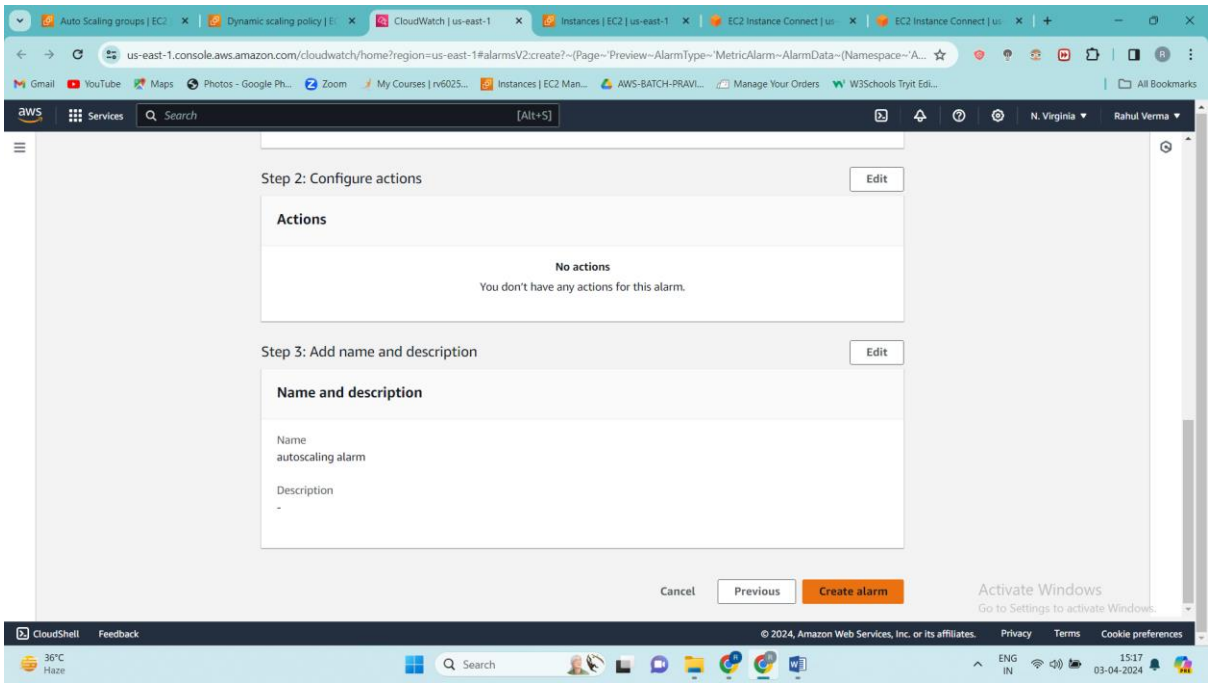


And give some conditions-

If cpu utilization goes more than 40% than alarm will trigger

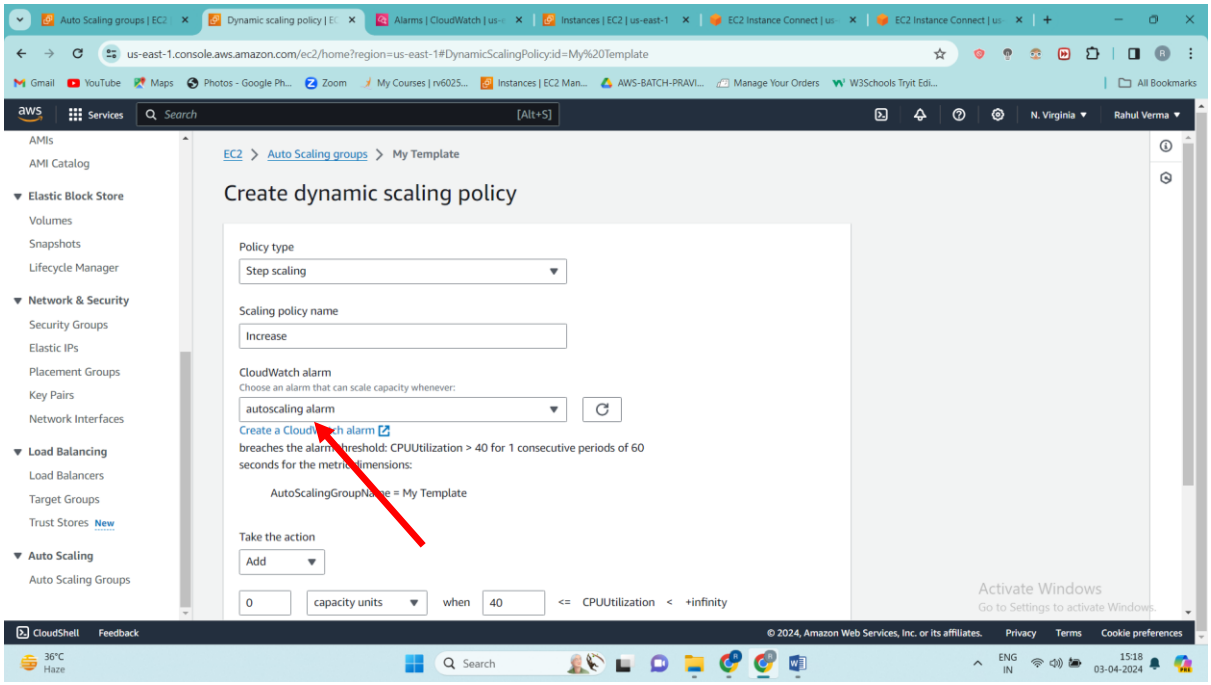


Now preview and click on create alarm

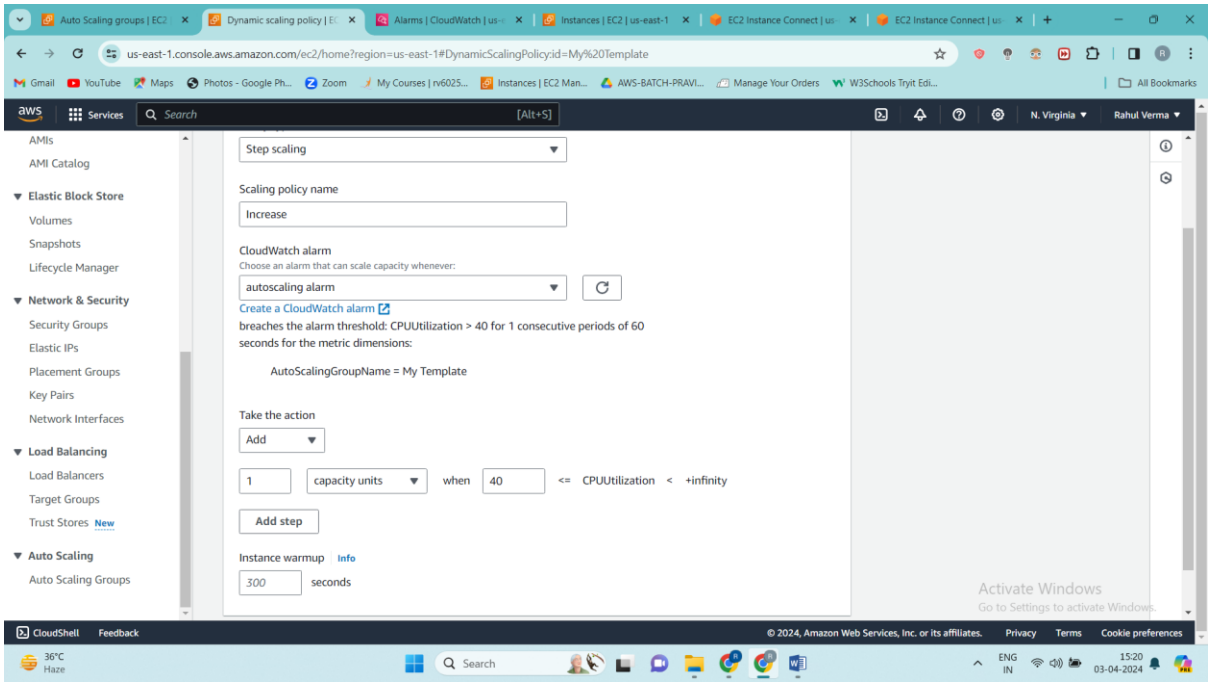


Step 12: now go back to dynamic policy creation page

Select your alarm which you have created just now

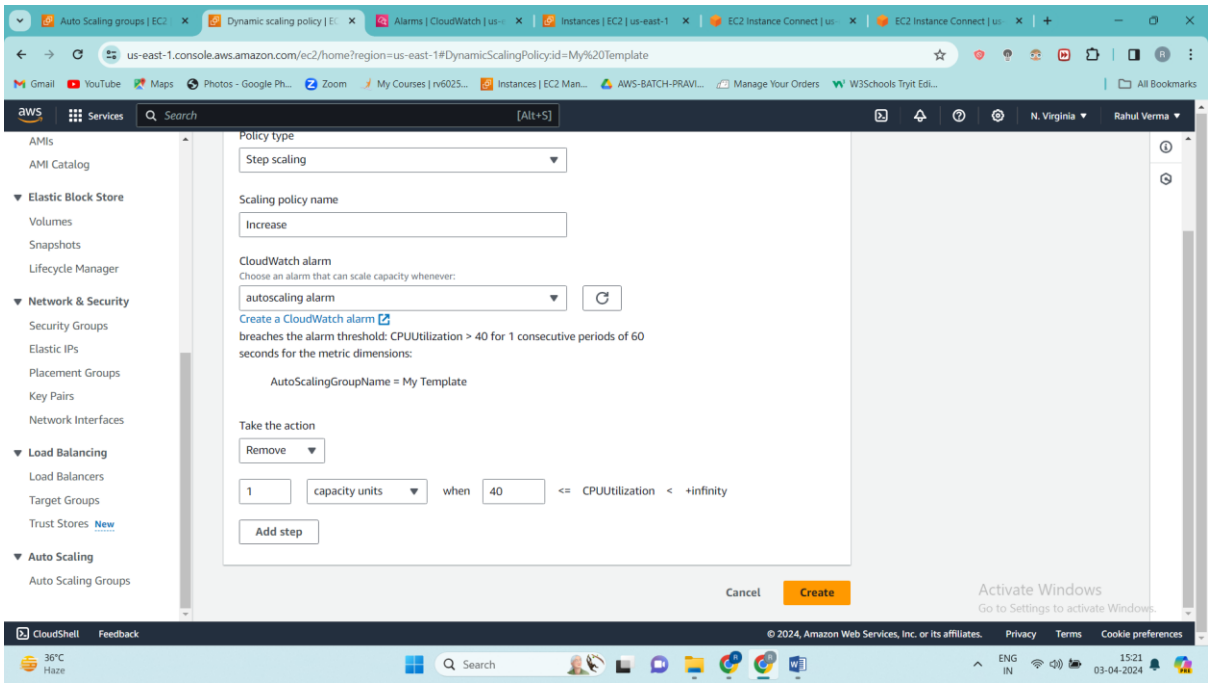


And here we have defined if cpu utilization goes above 40 % it will add one instance.

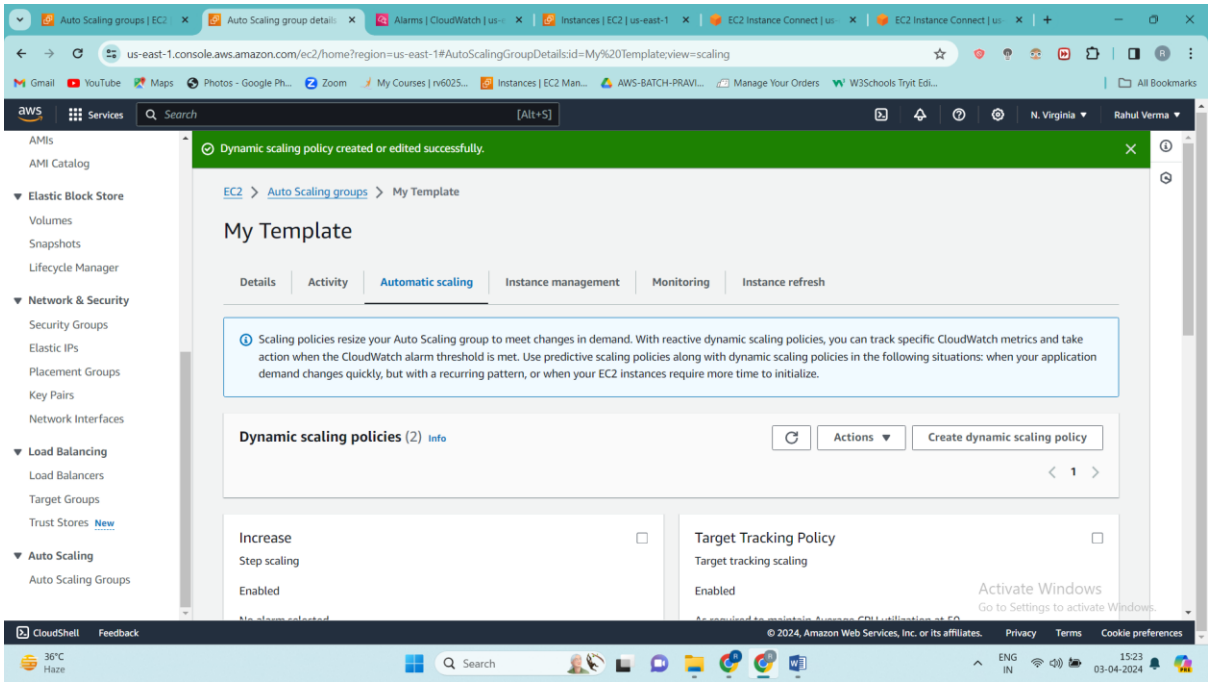


Note: same can be done for remove also for that you have to create another cloud watch alarm

Like If cpu utilization goes below 30% it will delete one instance

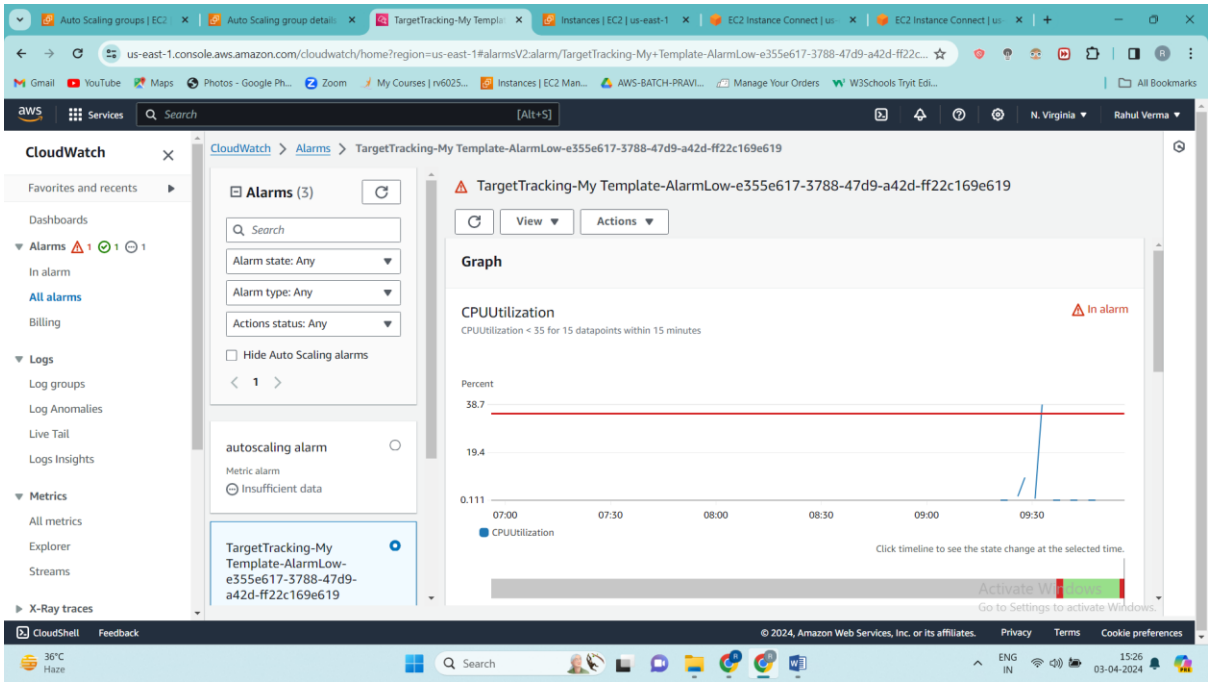


We have created for adding instance only

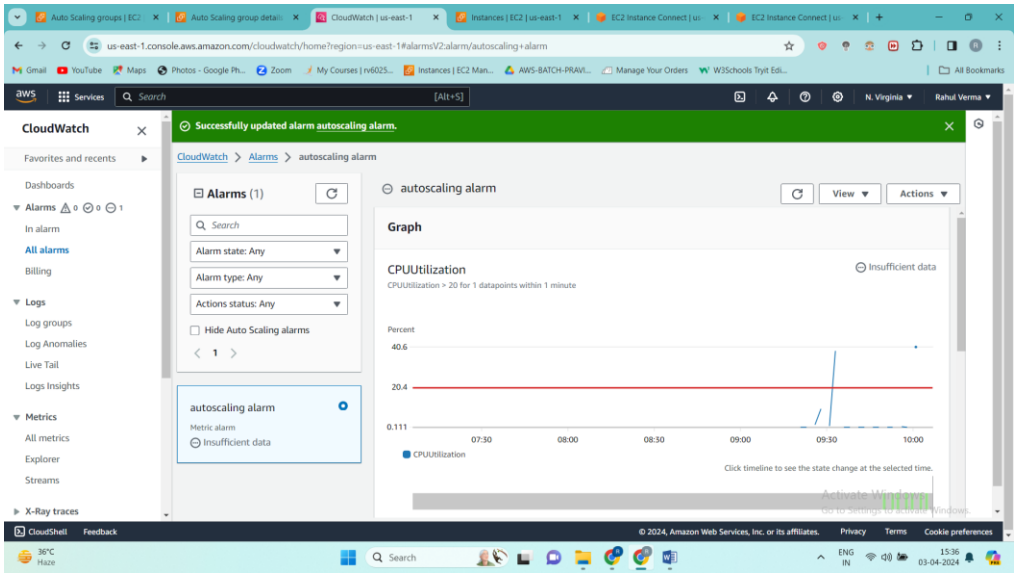


Step 13: let’s increase the load so that it will go above 40% cpu utilization and it should create one instance.

Our alarm is triggered now



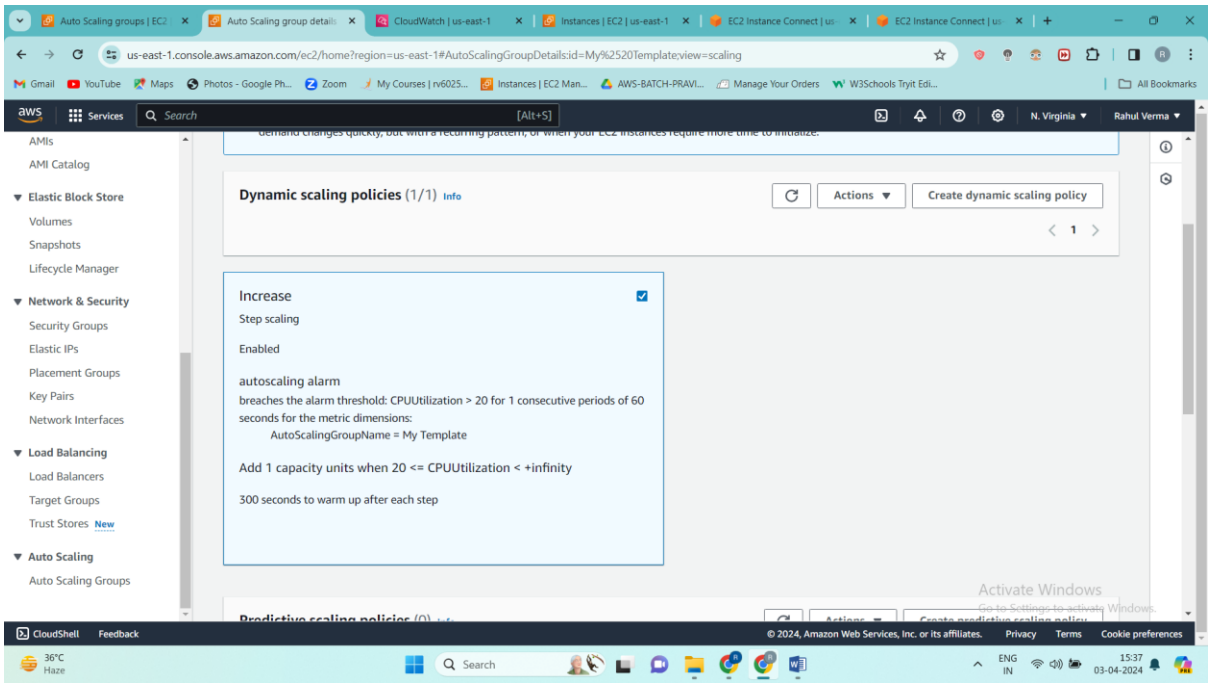
We have updated the alarm to spike the load faster. Instead of 40 we have change it to 20 now.



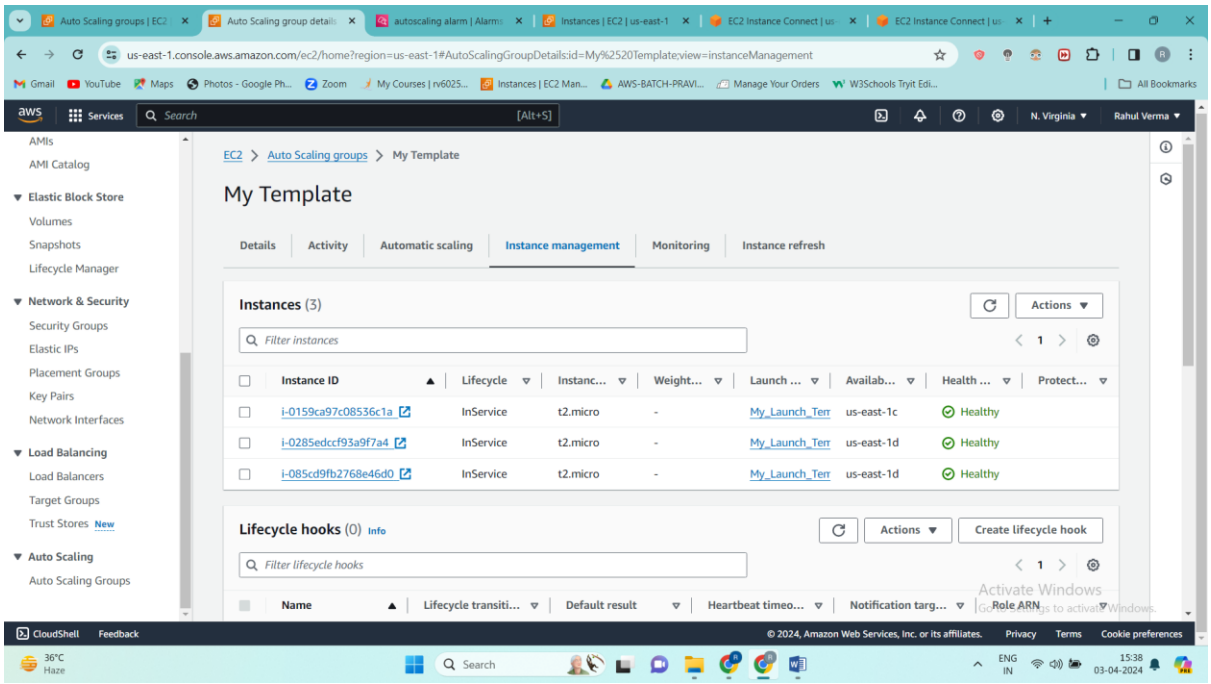
Use following command to spike the load-

yes > /dev/null

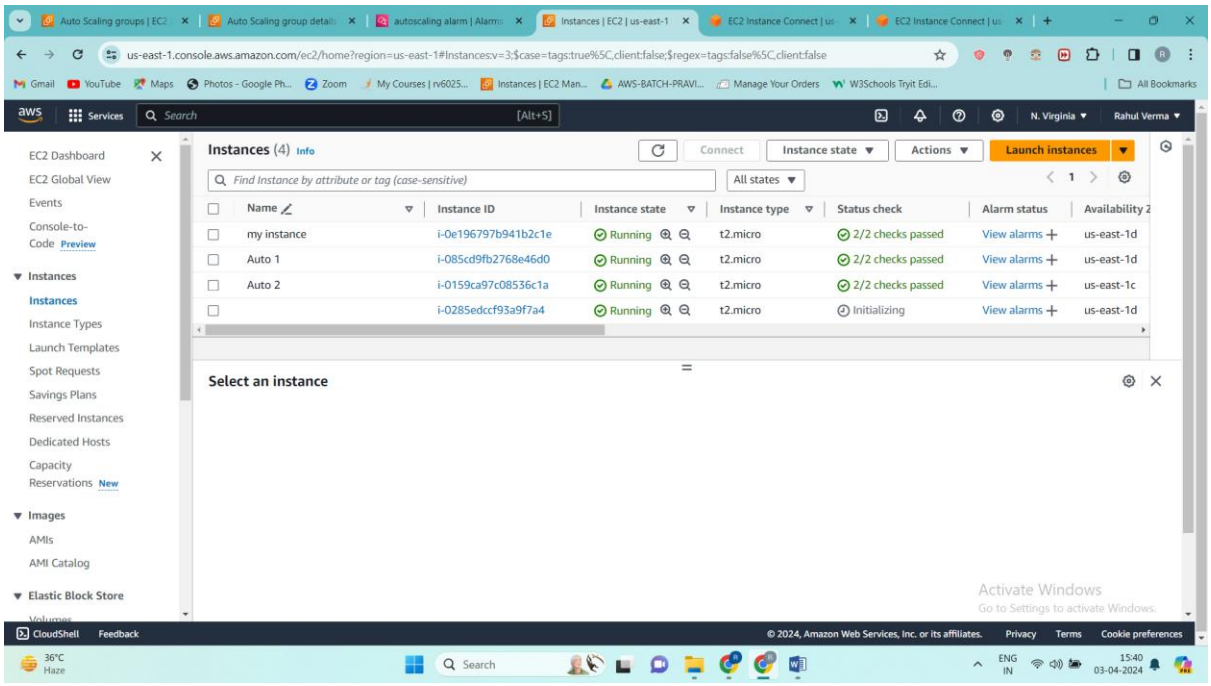
You can see in auto scaling also



And it’s done. It has created third instance



And it’s created one instance automatically



Note: To avoid charges after using the resources do delete them.

Thank You