Auto Scaling is nothing but your going to scale-up your instance and scale down you instance. Depends upon the recruitment.

For eg: you have a couple of instance now

Jenkins

Docker

These instance to be autoscale may be Jenkins need to be autoscale or docker needs to be autoscale depends up on the recruitment. So, initially I have one instance which is created by default and If I am going to enhance you now sorry if I am going to add a scaling group. It should be scaled up and scaled down depends upon your matrix which are going to set usually they scale up and scale down depends upon your memory or cpu usually in real time they will be having you know auto scaling group depends upon your cpu utililization or memory allocation why because you have more cpu utilization or you have more memory utilization your server may be crashed there is high chance of server crash. To avoid that using auto scaling focus on the auto scaling group and come back to the load balancer. So pre request of auto scaling group of your launch configuration or launch template. So, use can create anyone of this. You can click here and create launch configuration. Let me show the both here see I go here create launch configuration give any name may be Test one. Launch configuration is nothing but whatever your going to d it in a creation of eC2 instance manually. So when you trying to launch instances these are the steps you have. Seven steps with which need to be completed. So

1.choose your ami

2.choose instance type

3. configure ec2 instance

4.add storage

5. add tags

6. configure security group

7. review that and going to launch the instance. So, I hope every one will have how to launch a instances. So, if your enabling autoscaling group it going to create automatically a new instance. Creating of new instance you need to do automation to skip few steps.

So these are the steps you have need to skip these step and have it as a template. That it what we call it as launch configuration of launch template we have both of option whatever familiar with or comfortable will you can use that both are same..

Let me show you over her

Let’s go back to launch configuration and create launch configuration give a name Test

2.choose Ami:

If you choose you have plenty of option. This is custom Ami I will show you create a custom AMI in your EBS volume. Leave it as it is

Type AMI follow by AMI ID

!st step completed and next step choose instance type it is going to be t2 micro or t3.micro

Choose instance type

Choose for t2.micro may be your free tier one so choose this.

Additional configuration is not recommended this is spot instances to bet some instances value cost how many. So, this is spot instances nothing but betting the instance amount. If you enable it will give you spotting option you can type the amount over here. You have to give max and min price over here. Then depends on the no which your going over here it is going to create no of instances.

We have IAM instance profile.

We have IAM role.

Spot instance is nothing but betting the instances value. It is going to pay for every hours. Basically going to cost per hour some amount. Spot instances going to offer some value. It will show current value her.

Us-east-1a is subnet

For each and every available zones if I need my instances to be up and running only if it is below 200 or 2000. If you choose this option you can bet over here.

“When price is increase it will instances goes down.”

Test environment

Sandbox environment you don’t want any Impact for your application it will good to have multiple instances.

IAM role is explained about role IAM role is to establish a connection between AWS services.

S3 to EC2 services from

S3 to lambda

So s3 to other services to establish a connection we use IAM to establish a connection we use IAM role. To create one IAM role and attach IAM role over here.

Monitoring & EBS-optimized instance you can leave over here which is going to create one default volume which is xvda. Without default volume we cannot create a instances. So, we need at least one root volume but we can create multiple volume depends upon the requirement. We can add you know how much volume you want you can keep on volume adding it

Security group.

Choose existing security group or you can create a security group as well I choose existing one or you can create a new security group name it over these and add new rules over here. So I am going to use existing one. So I am also going to choose key pair so I have already one. Choose existing key pair which my batch AWS->Batchaws1m and I am acknowledging it and I am creating my launch configuration.

“today you choose add vpc in that so even when you have one tab In your configuration page. You have the option to chose your vpc. For now you going to use the default vpc. But in later session how to going to customize your vpc and how to create a vpc. Something which Is made for configuration.

VPC is one of the major concept in AWS.

Amazon imaging write you can do custom imaging. For eg. If I have configuration in my Jenkins so this is my amazon linux machine. I creating my configure in used Jenkins and set some parameter it may be like. If this Jenkins will have some pre request and we install that it will be store in your EBS volume.

So you can take a snapshot of your volume and create a new image and your creating a new image using this existing Jenkins volume it’s going to create custom image.

Usually image is created to replicate infrastructure .} EBS volume

I have created one launch configuration similar way we have something called template. Launch template which we can create from here. I will show you how to create a launch instances as well if you choose small drop down over here after the launch instances you can see launch templates. Launch instances from template. You can create as a template same thing creating ami steps it going to create as a template. Using this template you can launch a instances withing some minutes which going to create a instances whatever pre conf you have mention is going to fetch that.” So, I have created launch configuration. Now going to launch two instances. Let me show you how to launch templates, choose launch template source

I have created two launch template similar to the instances launch configuration this is also similar launch template is almost similar to launch configuration and If you see here both are almost similar and you can choose to create launch template with instances as well.

I am going to create 2 instances as well out of this dev as the template launch instances it is going to create instances without any step to access it will directly launch the template instances. So it is the easiest way to do it. Let me write it as server1 and I am going to choose as server2. So, I have created two instances and name it ash server 1 and server2.

EC2 dashboard->instances running.

I have created two instances now mean time I have also created launch configuration. Now I am going to set a load balancer sorry I am going to create a auto scaling group for my server1 may be server2. For now take one example only server 1 take eg let’s create a launch auto scaling group. I am going to create auto scaling group.

Auto scaling group is a separate services it is a part of EC2 instances but it is going to be a separate service give a name as Test.

You can see here you can choose from launch template or you can choose from your launch configuration you can switched it over here. From here we have created launch configuration you can see that test over here launch configuration which is going to be created t2.micro with your following ami Id default vpc.

Available zones for now we can leave it as such or we can choose minimal of tow so that it will split out the traffic only in this two not choosing anything you now going to have it a any available 6 zone it can be anywhere. Project requirement only to be available in a particular available zones we can choose any one of the available zone over here. So I hope you will aware inside we have available zones minimum of 2 available zones create one region. Norther virgina 6 available zones over here. Skip these minimal 2 over here it ask in to I select all the subnets over here which is your available zones and for now I have created any load balancer.

We going to create a load balancer and attach to auto scaling group why because depends upon the auto scaling group it is going to create a instances. For eg: it may be depends upon a matrix which we are going to use I show you how to configure the matrix in the next configuration page.

So depends on the matrix it may be 50% above on it may be 30% above you may have created but your traffic will be pointed over a 3 instances which will you have already to split the traffic between all the 5 instances. We are going to use load balancer that is the reason auto scaling load balancer works parallely. If you already created load balancer and attach to it. So now we have created anything now we have to create new one load balancer

Health check

EC2 default checking

ELB you can enable

Only 3 second

Next

Next parameter is to set your number of instances.

Desire capacity

2

Minimum capacity

2

Maximum capacity

5.

Based on project requirement you can increase or decrease

You need to scaling policies

Target tracking scaling policy

You can see here cpu utilization. So what it is going to do. It is going to check first 2 instance desire and minimal if it is going above 50% it is going to create another one instances

It is automatically scale up and

Automatically scale down

To security purpose you are going to have it usually load balancer and auto scaling work together you are going to take snapshot from load balancer also there won’t be any lose but for security purpose the instance will be stopped instead of terminating. Auto scaling group will automatically scale up and scaling down

Why auto scaling deliver your product efficiently to have any impact

CPU utilization getting 50%

Down time in your application is vital thing.

Crash it is nothing but crash in your linux part nothing but it going to stop the instances.

Huge no of use trying to access and entire server will be down. So what exactly happen there is no such concept of auto scaling since it doesn’t have any auto scaling load balancing capacity it is going to affect your application. So to deliver product or to deliver your server efficiently service . you need your traffic to split your work will be splitted in different instances and there won’t be any slowness there won’t be downtime to your application. It will be automatically working so may be at the end of the month lot of EB payments. So may be if they have auto scaling load balancer enabled infrastructure there won’t be any slowness and all the things will be efficiently handle.

If you don’t have anything obviously it going to be downtime. So after 30th day or may be 1st week you don’t traffic splitter for 5 instances. You can have only 3 instances. CPU utilization

50%

You must figure it out where my application is going down. Then depends on that particular parameter you have to create or you have to architect your infrastructure

50% below

Why 5 instances up and running

Instead 5 instances up and running

You can have 2 instances

Next

Next

To check try to terminate any one of instances let me try to terminate one of the instances and see whether it is working

Auto scaling group Is reflected or not it is going to terminate it whatever anything is getting u healthy with 3 instances and if you go back to your auto scaling group see here it is going to be updated

I thing it is already updated. Now you can see something call pending if you see there is 3 instances which is nothing but you can see new instances initializing this how you create auto scaling group.

Virgin

Other purely based on your region

3 types of load balancer available in the market

1. Application load balance(intelligent os 7 layer application)

Automatically deliver your traffic depends upon your

http to https

automatically split the traffic

depends upon your requirement

http🡪point to http application

https🡪 point to https application

Load balancer

Create a load balancer

ALB NLB

GWLB

ALB works on layer 7 in os concept

Have intelligent to deliver your user hit.

ELB- elastic load balancer (entire load balancer)

NLB- huge no of traffic for your application. You are going to use NLB.

-not intelligent is workout on network layer instead of application layer

Traffic to end point

NLB capable of handling higher load no of traffic attach a load balancer to a existing network load balancer. They will create a NLB high utilize application it will create a NLB under the NLB they will attach the ALB. So that your application will be higher no of hits.(ultra high performance) under that attach load balancer as a target.

E beans stack

NLB is capable of efficiently

ALB is user requirement

GWLB majorly used in real time.

No of high hubs

2LB ALB

ALB ALB

Hits over here point to ALB

ALB point ot any one of your instances deliver the traffic efficiently. If your having middleware server again request will be coming from NLB and transfer to your NLB and another ALB will be coming to the hit and deliver middleware application posted in your another ALB. Multiple load balancer in your load

Balancer in that not using single load balancer what Is the use of load balancer going to give traffic split between two server it is ALB to have minimal of 2 instances. So it is going to split the traffic

Server1

Server2

1st goes for server1 and deliver the respond and 2nd hit 2nd input will be deliver to the server2 and that is how delivering the traffic just managing the application efficiently.

No cost efficient

NLB cost higher than ALB

Product efficient

GWLB going to integrate any 3rd party application with your application recently.

Previous version

Classic load balancer

How to create ALB and how do you manage your application.

This is config page

Name Test any name can give over here

Scheme Internet communicate with multiple

Intranet vpc and you can attach data center as well internal to your vpc

Eg you have backend server hosted in your data center and front end hosted in your aws} that’s why internet and you need to give your endpoint her our scenario=internet -facing

Ipaddress type

Ipv4

Ipv6 alpha and numeric

10.0.0.1

10.a

10.b

Default vpc asking whether and make sure you have any balancer load outside of vpc and make sure inside vpc

And choose any one of mapping

Minimal 2 available zones load the traffic in available regions Or

Choose all the mapping previously you have created in your auto scaling group.

Choose security group

Or

Create security group it will be exposed do the particular port

Listener port

http:80

protocol port

http port 80

https 443 not needed

in realtime you have to provide ssl certificate

request new acm certificate manager request some certificate and call certificate over her it is simplest way

without certificate use cannot establish https

AWS global accelerator is nothing but your cloud front acceleration it will deliver your product efficiently end user nothing but if of I have

S3 transfer accelerator to transfer a data fast

Region subnet edge location

Not accessible to see your edge location but eg edge location

Are small data center if I am near to a particular data center or a edge location and may 5 member of access the same website. Same load balancer using over here

Edge location to available zone

Zones data center

Retrieve the information

Virgina I enable global accelerator

5 member access the same thing

Transfer from Mumbai centralized data center platform

Mumbai

Virgina from this going to get my information

This is how my configuration if 2ns user when use to trying to hit 5 members 2nd user access the same load balancer. It is not going to get the information northern virgina, going to store my as such cache in Mumbai. Showing use latency in my application

It going to cost you guys

You should know what is global accelerator

Create load balancer

Create target group

http 80

instances

ip address

lamda function

ALB

We already create two instances which is up and running

Instances

I am going to hit the particular instances and trying to load the traffic

Health check path

Which is up and running instances or not application hosted in a particular path you can choose the particular path.

/ check whether you having route or not unhealthy it will not transfer the date for the particular instances

Healthy 5 mins

Unhealthy 2

Timed out 3 sec

Interval 5 min

For every 5 min going to check the parameter

Success code 200-299

Target group name

Load balancer

Next

Next step choose the no of instances. I can choose the no of instances

Create group

Create target group

Choose Load balancer

Create load balancer

DNS- endpoint

Target group

Connect to instances

Sudo su

yum install httpd -y

cd /var/www/html

vi index.html

:Wq!

Systemctl start httpd.service

Connect to 2nd instances

Sudo su

yum install httpd -y

cd /var/www/html

vi index.html

:Wq!

Systemctl start httpd.service

To check whether the load balancer is working or not so that I have configured it

Load balancing

Load balancers

Target groups

Total target 2

Healthy 2

Test

Copy dns

Instances

Check it running instance click running instances

Monitoring

Everything will shown over here

Integrate load balancer with auto scaling group

Auto scaling groups

Check test click

Check edit load balancer

Load balancing -optional

Load balancer

Check application, network or GWLB

Target groups

Choose target group

Choose load balancer/HTTP

Update

Target group will be updated with auto scaling group

Conf target group updated in your autoscaling group which load balancer will be automatically updated by the target it is going to deliver the traffic

Load balancer

And

Auto scaling works parllely