***ELB(ELASTIC LOAD BALANCER) :***

>>ELB is nothing but it will balance the LOAD , it will distribute the load equally to the servers.

>>Share the load into different different machines we are using LOAD BALANCER.

**EXAMPLE :**

Suppose i created two servers like server1 and server2 , in that both servers i installed httpd package and i will install html page in server1 and server2 after that i will configure ELB ,

NOW i will up two servers ,

>> now click on anyone server once you click on that server here it will display one option like SECURITY GROUP or (if you go to networking under ec2 there also will be there security group option )

>>click on security group , by default ssh default it will be enable , now if you want to enable any port number go to actions click on security EDIT INBOUND RULES , by default ssh will be there , now enable port number , select ADD RULE choose CUSTOM TCP and enter port range 8080 or 8081 and save it

after that select server2 and enable port number 8080 , i will connect to that two servers

>>if you click on connect under the example section it will show you one URL select that one go to github paste that url and run or take port number of that instance and go to putty and open , now we logged into that server

--> now i will install httpd ( yum install httpd)

now i will create one index file inside the html directory

$ vi /var/www/html/index.html ---------> $cd /var/www/html by default it will be

vi index.html

SERVER1

:wq!

$service httpd.start

>>now i have to start services , the command is **, $service httpd start**

**$systemctl enable httpd** ---------> if we stop or start servers or reboot servers also automatically servers will be enable

now hit that server of ip address in the browser

>>same this things i will do in server2 also

now go to browser and hit both server ip address ,

>>Now i will do ELB configuration to both servers

>>client servers will request the load to server1 , by using ELB i will share that load into number of servers equally.

request from client side ----> LOAD BALANCER ----> share into different different servers

***HOW TO CONFIGURE ELB INTO SERVERS :***

login to AWS account click on services ----> go to EC2 -----> left side you see LOAD BALANCING option --> under that option we have a option like LOAD BALANCERS click on that here load balancers will display if you already created or click on create load balancer and select CLASSIC LOAD BALANCER

now specify the deatails ,

LOAD BALANCER NAME : myelb1

Create LB inside : -----> by default it will select

CREATE AN INTERNAL LOAD BALANCER :

ENABLE ADVANCED VPC CONFIGURATION : enable this one

down side we have a option like **LOAD BALANCER PROTOCOL :**

**here if you want to enable any port number you can enable here , By default port number is 80**

**while we are creating the ELB it will perform seven stages that is ,**

**Define Load Balancer , Assign Security Group , Configure Security Settings, Add EC2 Instances , Add Tags , Review**

**ADD EC2 INSTANCE :**

in this phase we will add instances into ELB , whatever the AVAILABILITY ZONE you selected in that how many instances are there all instances will be displayed here ,

now select instances and go to next phase ADD TAGS , give tag name and go to review phase and click on create.

edit and all we can do in ELB dash board only

>> Suppose you created load balancer and you have many instances now how to add load balancer into new instances for that go to LOAD BALANCING select load balancer and click on instances( this option will be down side) and click on edit instances and add new instances and save.

---------------------------------------------------------------------------------------------------------------------------

***AUTO SCALLING :***

>> At a time how many servers need to up

EXAMPLE:

server1 , server2 --------> server10

here minimum is server1 and maximum is server10

>>Now server1 is running when server1 CPU utilization crossed 90% then server2 will up with the same configuration similarly till server10 it will up , when we get cpu utilization below 15% automatically we have to scale down it.

>>If server1 cpu utilization crossed 90% automatically how many servers need to scale up either server1 or server2 how many servers need to scale up **designed capacity** maximum till 10 servers.

>>According to design capacity servers will up suppose if i mention design capacityis 2 then when server1 crossed 90% at a time two servers will up with same configuration.

Two groups are there ,

1. Scale up ( scale up means automatically increase the server )

2.Scale down ( scale down means automatically terminate the server )

**PREREQUIREMENTS :**

PREREQUIREMENTS is compulsary we should create AMI ( AMAZON MACHINE IMAGE )

**HOW TO CREATE AMI :**

>>If you want to take backup of any one instance select instance and go to actions click on create image and specify the details ,

>>We can launch another instance using AMI based on the same server

services---->ec2-->launch instance-->left side you can see MY AMI'S

AMI instance will be there in snapshot (EBS) or it will display in AMI

go to services-->EC2-->IMAGES -->AMI

now specify the AMI NAME ,

comming to AUTOSCALLING GROUP :

**HOW TO CREATING AUTOSCALING GROUP :**

go to services ---> ec2---> left side we have a option like AUTO SCALING -->

under the AUTO SCALING we have two options like

**1)LAUNCH CONFIGURATION AND**

**2)AUTO SCALING GROUP**

First click on LAUNCH CONFIGURATION you will get one option like create AUTOSCALING GROUP click on that and click on launch configuration here you can see many instances and MY AMI's so click My AMI'S and select AMI (AMAZON MACHINE IMAGE)

we have some stages to create autoscaling ,

select instance type and under the create launch configuration it will ask name enter name like mylc-1 ----> click on next --->add storage---->security group---->review

finally create launch configuration and select existing key pair and launch

Once you click launch it will take you to AUTO SCALING GROUP here we can mention at a time how many servers need to scaleup and scale down

**2.AUTO SCALING GROUP :**

GROUP NAME : agp1

GROUP SIZE : start with ( give number 1 or 2 ) instances

NETWORK : ------> network mention default

SUBNET : ---------> by default i will give two subnets like a and b

**ADVANCED DETAILS :**

------------------------------------------------

**3.CONFIGURE SCALING POLICIES :**

>>here select the use scaling policies to adjust the capacity of this group

>>scale between 1 and 1 instances , this will be the minimum and maximum size of your group

so here we have to give minimum and maximum number of servers

for example i will give minimum 1 and maximum 5

cpu utilization and alarms and giving the maximum and minimum serevers and increase and decreasing groups everything we will mention under the CONFIGURE SCALING POLICY option.

4.configure notification---->5.configure tag( specify the name) ---> 6.review

finally click on create auto scaling if anything configuration wrong is there it will show here only

>>If you want to stop instance first go to AUTO SCALING group delete auto scaling group and stop the instance otherwise if you delete directly instance , based on the auto scaling group one more instance will be up

-------------------------------------------------------------------------------------------------------------------