🕉️

AWS Section 8 - ELB

Scalability means handling load depending on adaptability.

There are two types of scalability:

Vertical scalability: This means increasing the size of the instance. That is if we are using t2.micro instance then we will increase the size of the same instance and convert it into t2.large instance when This scalability is very common for non distributed systems such as database.

Horizontal scalability: This is also called as elasticity. Here the size of the instance is not increased when there is load but here number of instances is increased when there is load. And these instances which are increased are of same type. This is used in distributed systems.

High Availability:

This means the system is divided across two availability zones. So that if one goes down other will still providing the service. This is used to avoid failure of the application. This is like running instances of same applications over different availability zones. Here autoscaling group over multi AZ and load balancer over multi AZ is used. We will see this going further.

Scale in:

Decreasing the number of instances when the load on the instances gets decreased.

Scale out:

Increasing the number of instances when the load on the instances gets increased.

Elastic Load Balancer:

This is a service which balance the incoming load or network traffic across the multiple servers downstream. Consider that there are three instances and there is a load balancer which is connected across these instances.

These servers which are there at the downstream of the load balancer are called as downstream servers or target group. Now These servers are carrying application which user will try to access.

Load balancer will provide a single IP address or a single host name. With this single hostname or single IP address whenever the user makes the request it the load balancer routes its request to any of the server located in the target group. Here user is not knowing to which server it is hitting.

So when the first user makes the request the load balancer routes the request to the first instance.

Now when the second user makes the request the load balancer routes the request to the second instance.

And then when the third user makes the request the load balancer routes the request to the third instance.

This thing goes on. More the user are making the request the more is the load balancer routing the traffic to the instances in downstream. And the request is routed to the instance having minimum load of traffic out of all the instances.

The load balancer provides the single point access to the user. It also checks the health of all the servers across the target group. And then according to the server health check it routes its traffic to it.

Health Check:

Load balancer does the health check. If the response is not 200(OK) then instance is not healthy.

Types of Load Balancer:

Application load balancer - HTTP and HTTPS, WebSocket.

Network load balancer - TCP, TLS, UDP.

Gateway load balancer

Security Group:

So here the traffic to the instances will be coming from the the load balancer and not from some outside user.

So the source of the traffic for the instance in the security group will be security group of the load balancer.