CS 519 Cloud Computing Overview

**VL10: Scale & Load Balance your Architecture**

School of Technology and Computing

## **Instructions**

## For this activity you will be using AWS Academy Labs - **Scale & Load Balance your Architecture.**

Please complete the provided lab.

The difference between Elastic Load Balancing and Auto Scaling is that ELB is the analogy of networking where the system distributes customer requests to instances, whereas auto scaling is about spinning up or shutting down instances based on traffic.

If we take a look at network topology for our cloud server at the beginning and at the end, we see that at the beginning we have a single public subnet compute server with its own private subnet RDS. At the end, we also launch a compute isntance for the Primary RDS in availabilit zone 1. We should also note that at the end, even the computer instances are in private subnets.

We see that we have an EC2 compute instance running. As we have learned in prevous weeks while discussing EC2, we can create custom AMI based on our current instance configuration. We go ahead and do this by creating an image.

Now we want to create a load balancer. As we have talked about in the previous paragraphs, load balancer is an analog for networking. Therefore when we create a load balancer, we have different choices depending on the internet network layer. Network Layer (IP), Transport Layer (TCP/UDP), Application layer (HTTP, HTTPS). We select an application load balancer that routes application packets.

We select LAB VPC and we select multiple public availability zones in order to allow this load balancer to work across multiple availability zones.

We select where do send application requests.

Having an image for an instance is very important in dynamic servers that automatically shut down and spin up servers as opposed to a monolith server that is always on. We don’t want a human to manually change everything and take hours every time a new instance is launched. We want to computerize/codify/automate as many things as possible. This is why we set up the launch configuration.

We select the ami and the instance type and set up other variables. Now we have our launch configuration. We create an auto scaling group using this launch configuration. We select private subnet 1 and 2 to be used for this launch group. Remember that those were in two different availability zones. We choose the load balancer we created previously. We set minimum, maximum, and desired capacity. We tell the auto scaling to keep cpu utilization at around 60%.

Now we go back to instances and we see that 2 instances are launching. We can go to load balancer target groups to see all the groups we have which is one at the moment. If we click this group we see two instances. When a load balancer is used, users don’t access the instances directly, they access the load balancer using its dns name, load balancer assigns an instance to the user. We can see that 2 more instances launched after we loaded the server and stress tested it.

