**SCALABILITY AND AVAILABILITY.**

Scalability means that an application can handle a large number of users or data by adapting.

There are two types of scalability:

- Vertical Scalability

- Horizontal Scalability or Elasticity.

Vertical Scalability: This entails increasing the size of an instance. A developer can increase/scale its instance from a t2.micro to t6.large.

Vertical scalability are compatible or common with non-distributed systems such as databases. There is a limit to how much an engineer can scale vertically, that is the hardware limit.

Horizontal Scalability: A developer can scale horizontally by operating/creating a new server/instance.

This is used for load balancers in the same AZ and auto scaling group

This type of scalability is common in web applications.

High availability is use to work hand in hand with horizontal scaling. This entails having more than instance up in different AZ. This concept is used to curtail data loss.

**LOAD BALANCERS**

Load balancers are servers that forward internet traffic to other servers (EC2 Instance) downstream.

Importance of load balancers

- Load balancers help us to distribute a load of internet traffic downstream

- Load balancers exposes a single DNS access.

- Load balancers help us to perform health check on the instances, so as to make sure they are alive and well.

- High availability across AZ.

- They provide SSL for your website.

- It acts as a proxy, in the sense that it separate public and private traffic. It differentiates between outgoing and incoming traffics.

Types of Load Balancers on AWS.

- Classic Load Balancer – HTTP, HTTPS, TCP (V1 old gen)

- Application Load Balancer – HTTP, HTTPS, WEBSOCKET (V2 new gen)

- Network Load Balancer – TCP, TLS (secured TCP), UDP (V2 new gen)

Classic Load Balancer: Supports TCP (Layer 4) while HTTP/HTTPS (Layer 7).

The CLB is accesses only IP and TCP layer.

Application Load Balancer: Supports HTTP (Layer 7). It also supports redirect from HTTP-HTTPS

ALB is useful in micro-services and container based application.

Unlike CLB where we need to have multi load balancers to house servers, in ALB one load balancer can host multiple servers/instances