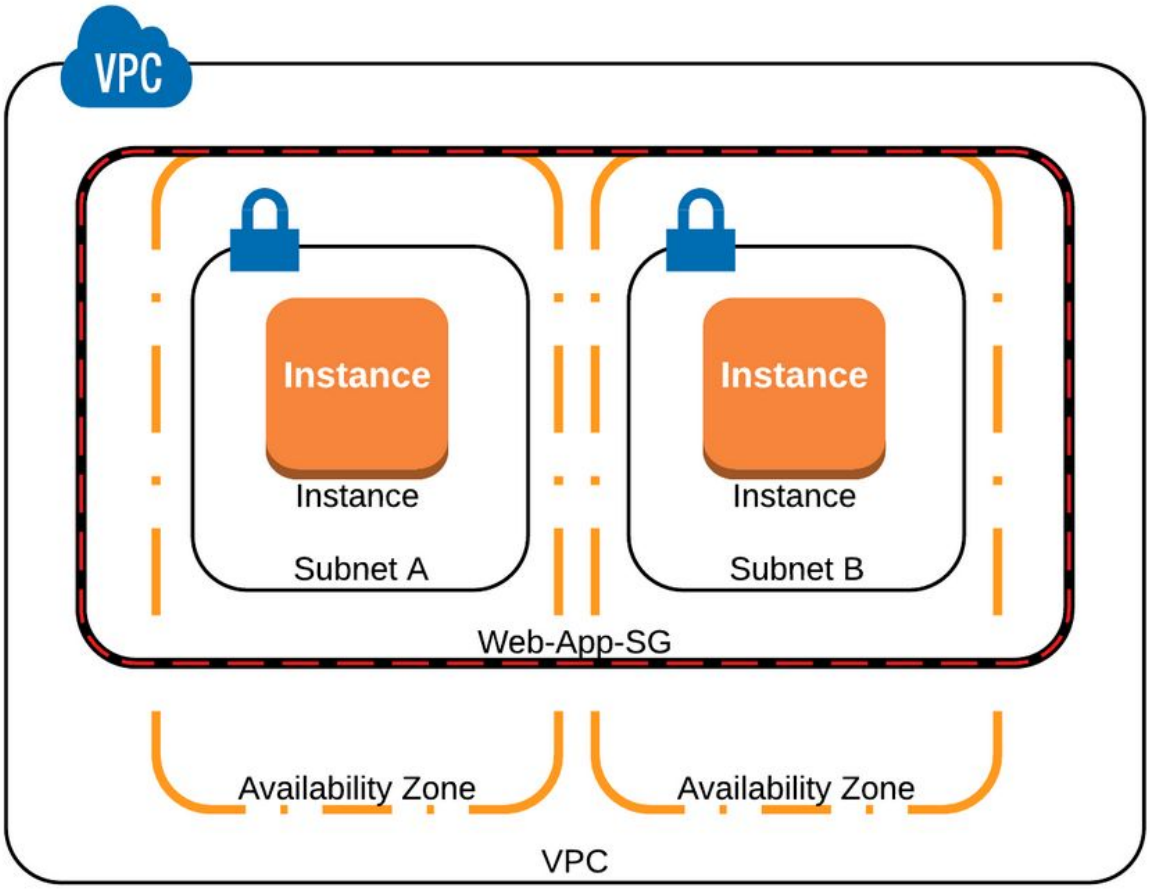
Classic Load Balancer planning

The Elastic Load Balancing Service automatically distributes incoming application traffic across multiple Amazon EC2 instances. It enables you to achieve fault tolerance in your applications, seamlessly providing the required amount of load balancing capacity needed to route application traffic.

Elastic Load Balancing offers two types of load balancers both of which feature high availability, automatic scaling, and robust security. These include the Classic Load Balancer that routes traffic based on either application or network level information, and the Application Load Balancer that routes traffic based on advanced application level information that includes the content of the request. The Classic Load Balancer is ideal for simple load balancing of traffic across multiple EC2 instances, while the Application Load Balancer is ideal for applications needing advanced routing capabilities, microservices, and container-based architectures. Application Load Balancer offers the ability to route traffic to multiple services or load balance across multiple ports on the same EC2 instance. In this lab, we are going to focus on the **Classic Load Balancer**.

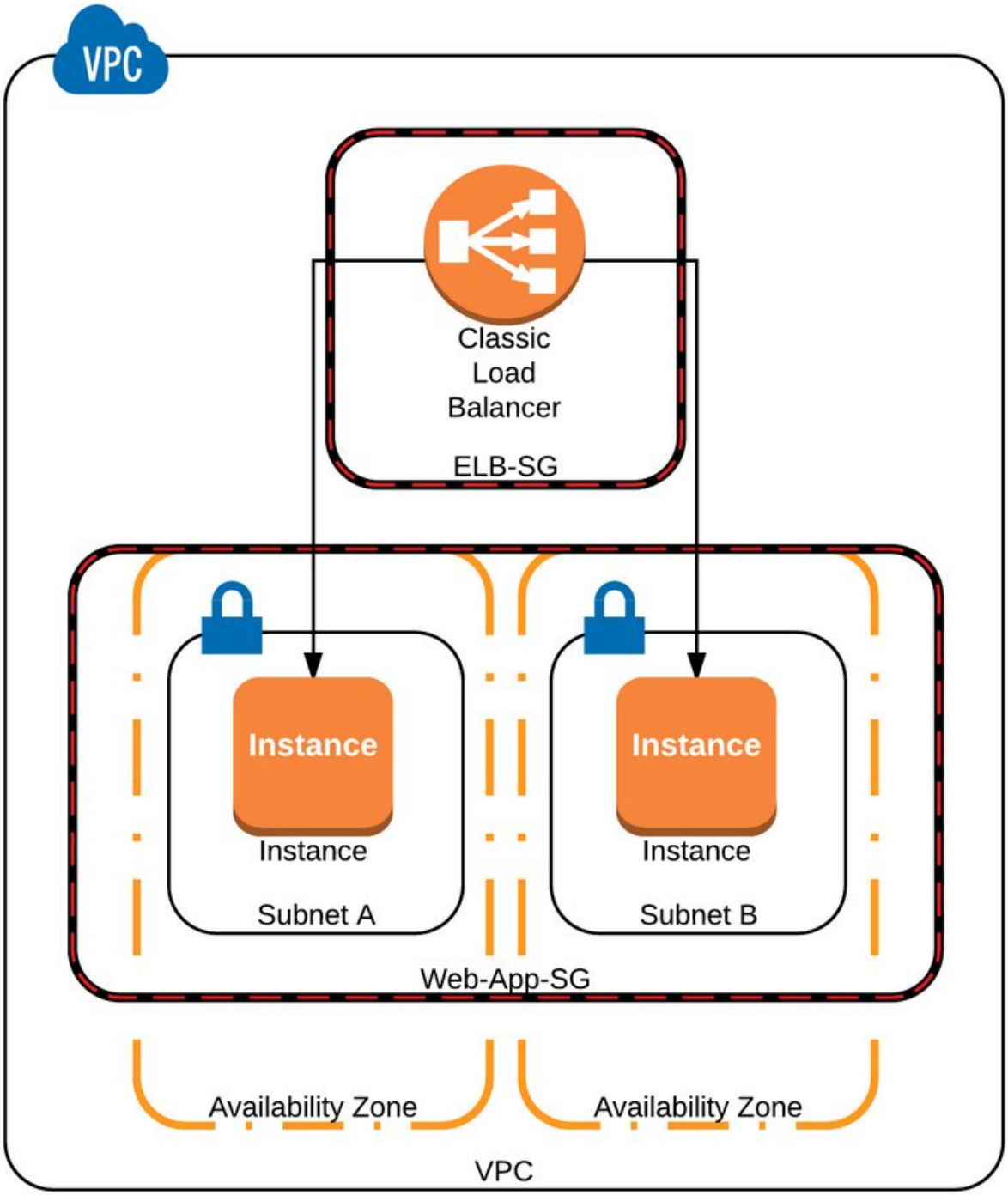
Planning

Before going ahead and creating a Load Balancer (LB), let's take a look at an overview of our current infrastructure. When you connected to the AWS account provided in the former step, you had a few things that were already deployed. This is the current infrastructure that was already deployed for you:



You already have a VPC with some subnets and 2 EC2 instances running inside the VPC in different Availability Zones. Both instances are inside the same Security Group called `Web-App-SG`, which is allowing HTTP access from port 80 to anywhere (0.0.0.0/0). Each EC2 instance is running the same web application. We want to **configure an LB to create a central point of access to our application**, and we also want to configure our architecture in a way that **users can only access the application through the ELB**.

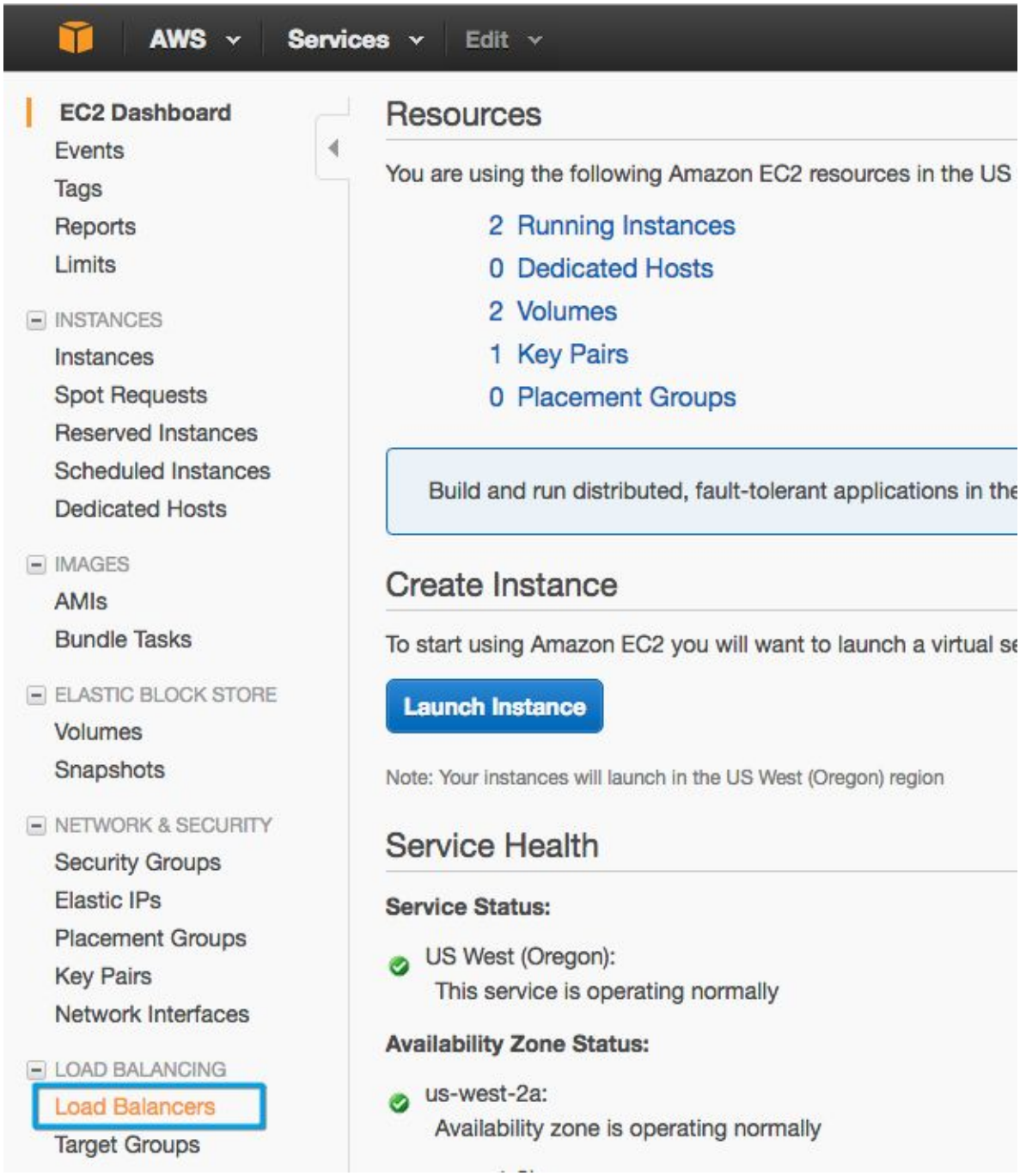
In the end, we should have a solution similar to this one:


VPC
Classic
Load
Balancer
ELB-SG
Instance
Instance
Subnet A
Instance
Instance
Subnet B
Web-App-SG
Availability Zone
Availability Zone
VPC

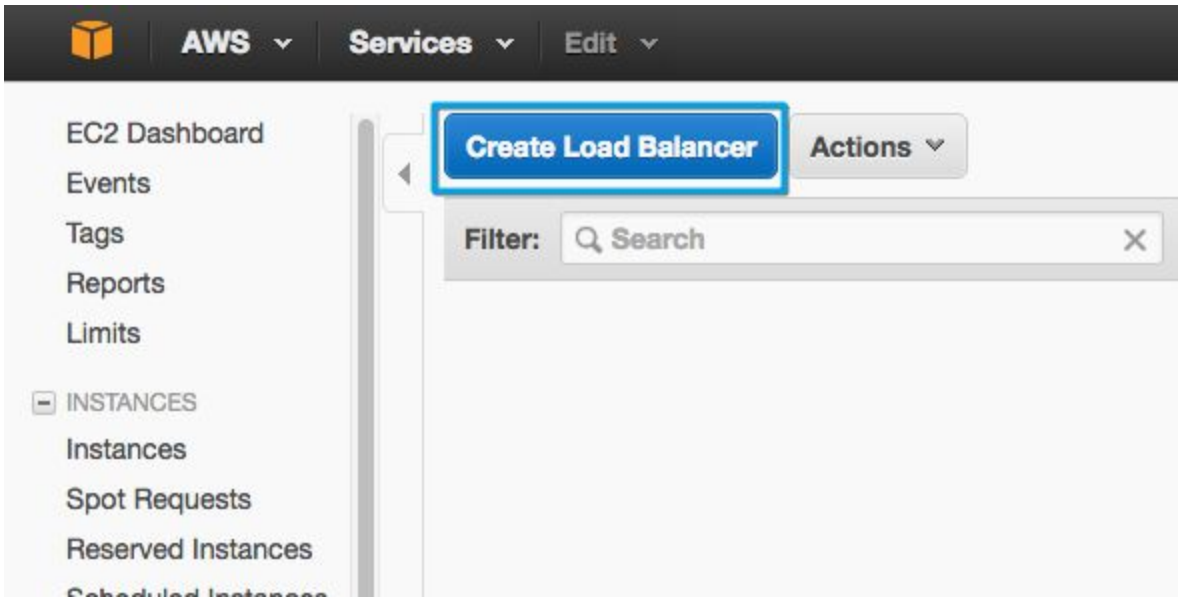
To do that we will have to create and configure a Classic Load Balancer, and properly configure the needed Security Groups to make sure that our application will work as expected.

Create a Classic Load Balancer and register EC2 instances

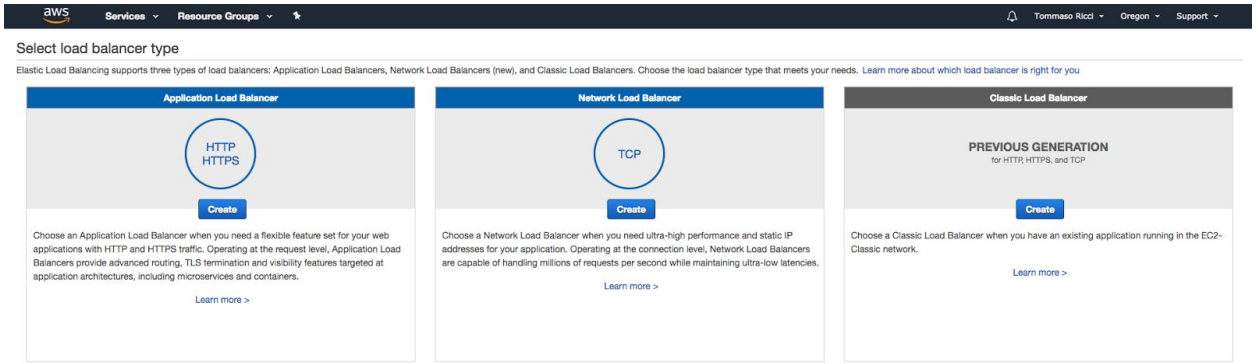
Click on Load Balancers:



Click on **Create Load Balancer**:

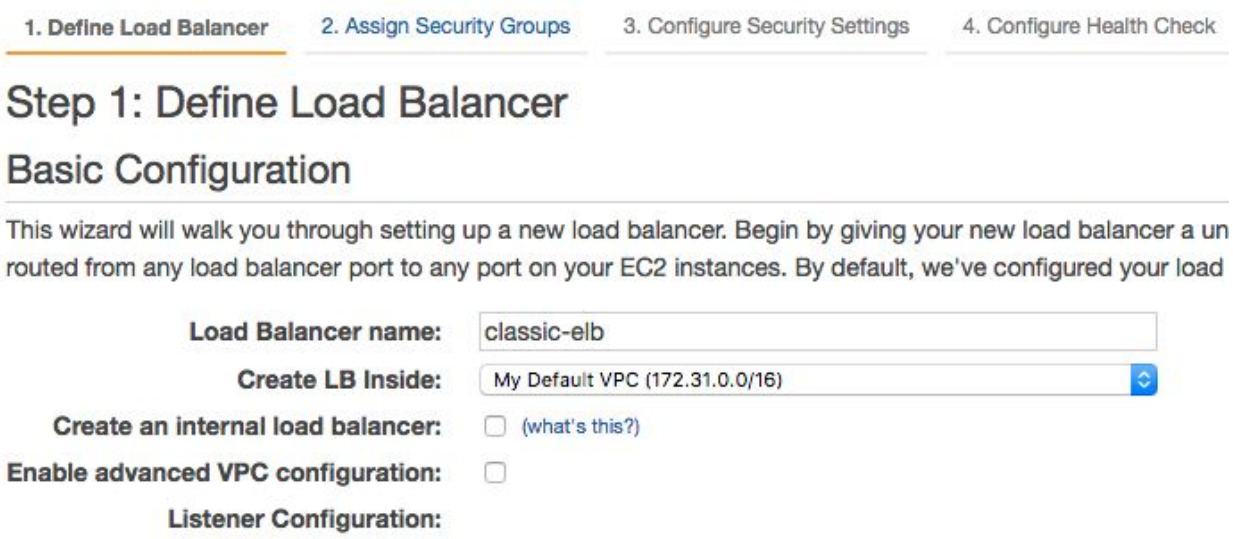


As stated before, you can choose from two flavors of ELBs: Application Load Balancer or Classic Load Balancer. In this lab, we will use the Classic Load Balancer, so simply choose the proper one in this step and click on **Create**.



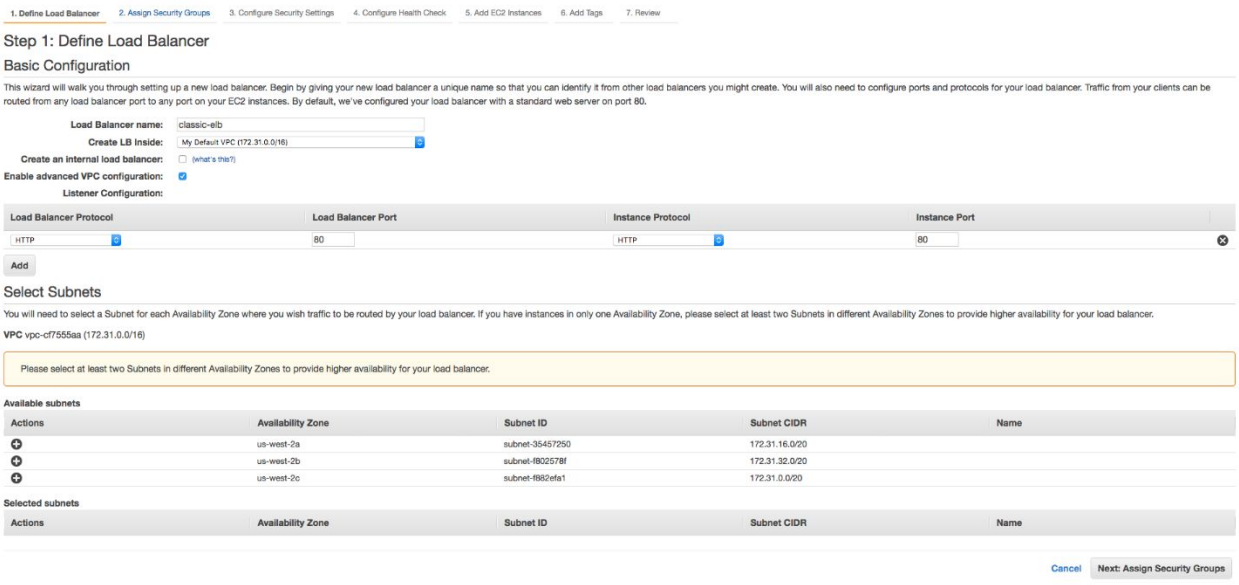
Now to start configuring the specifics of the Load Balancer (LB), you will need to follow a 7-step wizard.

In Step 1, you need to specify a name for the LB; this name can be anything that will make sense for you in the future. But be aware of the limitations (Only a-z, A-Z, 0-9 and hyphens are allowed). You need to select the VPC where the LB will live, this VPC should be the same VPC where the EC2 instances are running, you will probably have only the Default VPC in your account, choose this one. You now should have something like this:

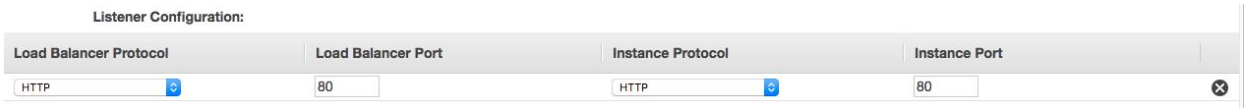


Some info about the next config points. We will create an LB to receive traffic from the internet and forward to our instances, therefore we need a publicly accessible LB. If you select **Create an internal load balancer** you will be creating a load balancer that won't be publicly accessible - in this case, the LB will only be accessible inside the VPC, which is not the goal, so you should leave this box unchecked.

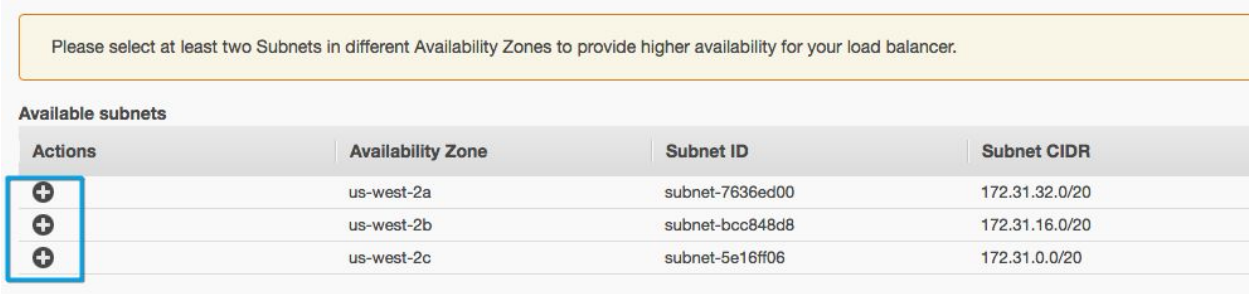
There are instances running in different availability zones, and we need to configure the LB to work in all subnets where we will be launching web instances. To configure this behavior we need to **Enable advanced VPC configuration** in order to select the subnets we want. After that you will be able to see more options, it will look like this:



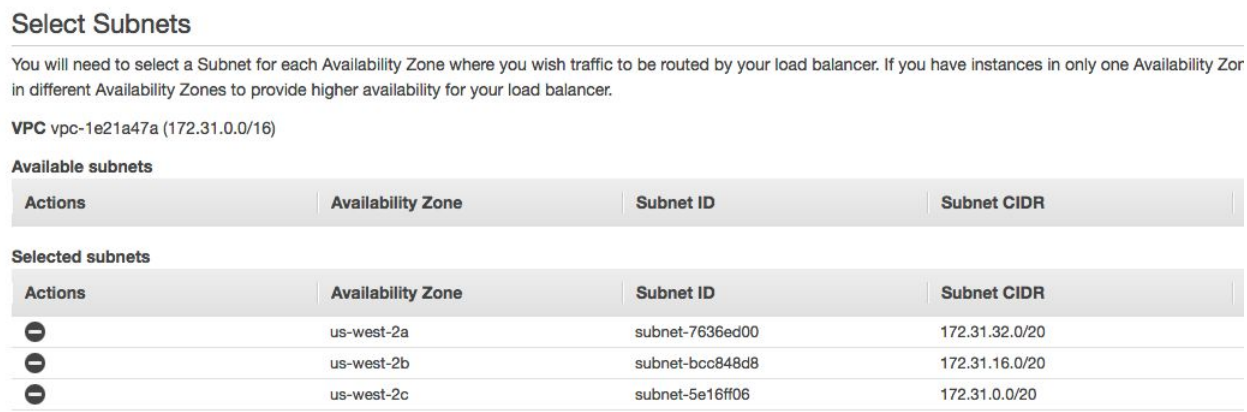
A LB listens to a specific port for requests coming from outside, in this case, the internet, and forwards the request to the instances running behind it on a specific port. In this lab, we will be using the port 80 for HTTP requests for both ELB and the EC2 instances, therefore, the default choice will work for us:



All though there are only 2 instances running in the account, in different Availability Zones, we will want to select ALL the available subnets in this VPC, just in case we want to launch another instance later on in a different Availability Zone. Simply click on the plus button for all the subnets available in this VPC.

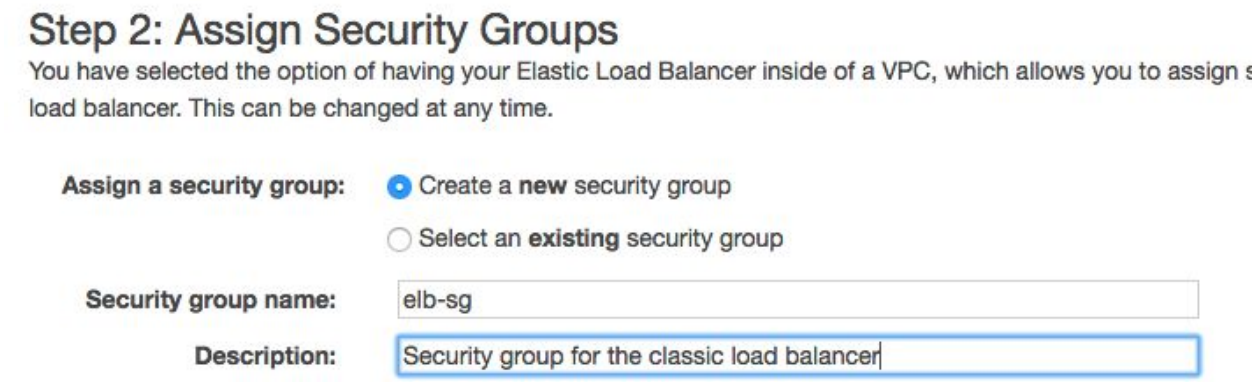


And in the end, you should see something like this:

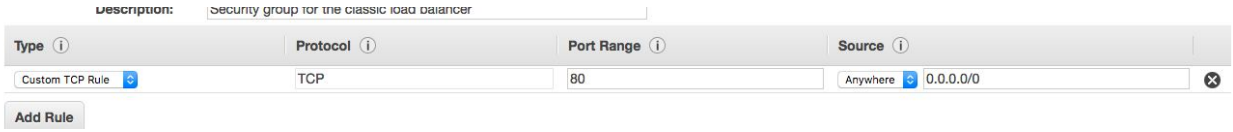


The first step of creating the LB is complete. You can click on **Next: Assign Security Groups**

On step 2, we need to configure a Security Group (SG) for the LB. This SG will be used to manage the security for the LB itself, therefore, since we only defined the port 80 (HTTP) in the listener section of the last step, we will want to create a new SG for the LB that will accept connections coming from anywhere to port 80 of the LB. To do that select **Create a new security group** and provide a name and a quick description for this SG.

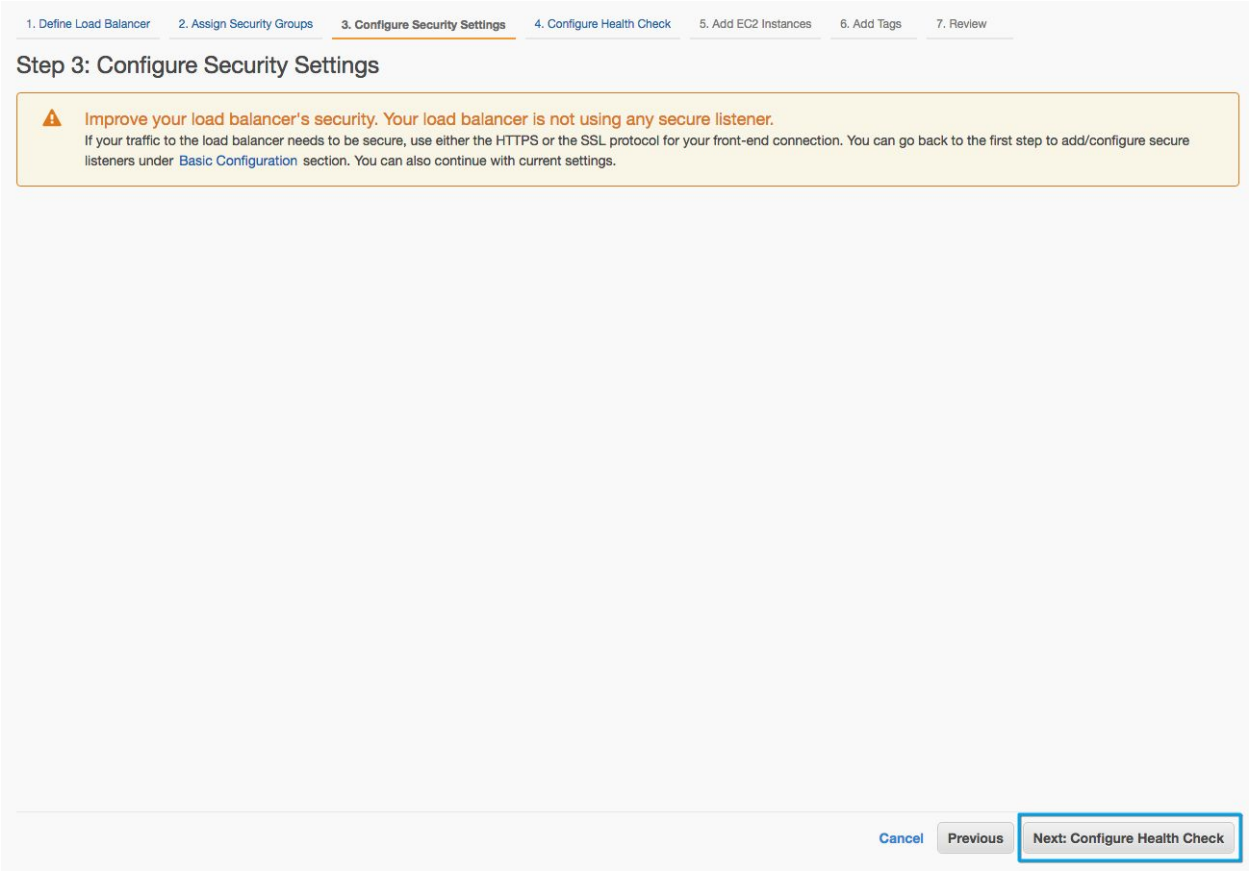


And allow connections coming from Anywhere (0.0.0.0/0) to the port 80 and nothing more

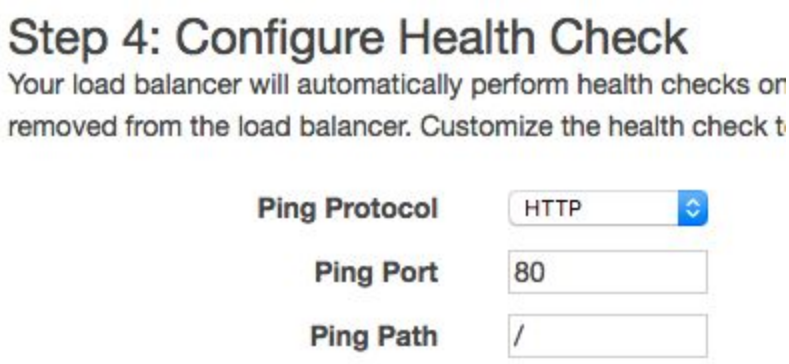


After that, you can click on **Next: Configure Security Settings**

Step 3 consists of configuring the LB to use HTTPS or SSL for security purposes. Although it is highly recommended that you reinforce security in your applications, configuring it is beyond the scope of this lab, therefore, you can simply click on **Next: Configure Health Check**



In step 4, you need to configure a health check. This is how the LB will evaluate the health of an EC2 instance and decide whether to send requests or avoid a particular instance. The first thing to configure in here is the protocol, port, and path that will be used for the health check. The instances running in the account are serving an application in **port 80**, using the **HTTP** protocol, and using the **root path (/). Y**ou should configure this in the first part of this step

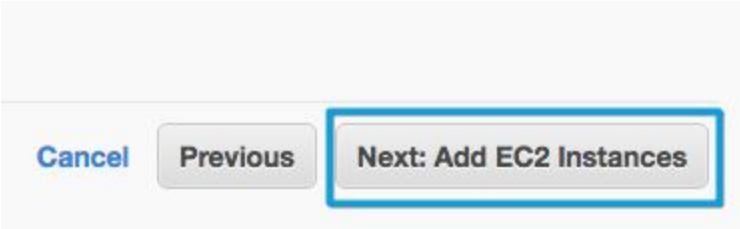


There are some **Advanced details** in this step you can configure, but for the purposes of this lab we will stick with the default settings. However, for your reference, this is what they mean:

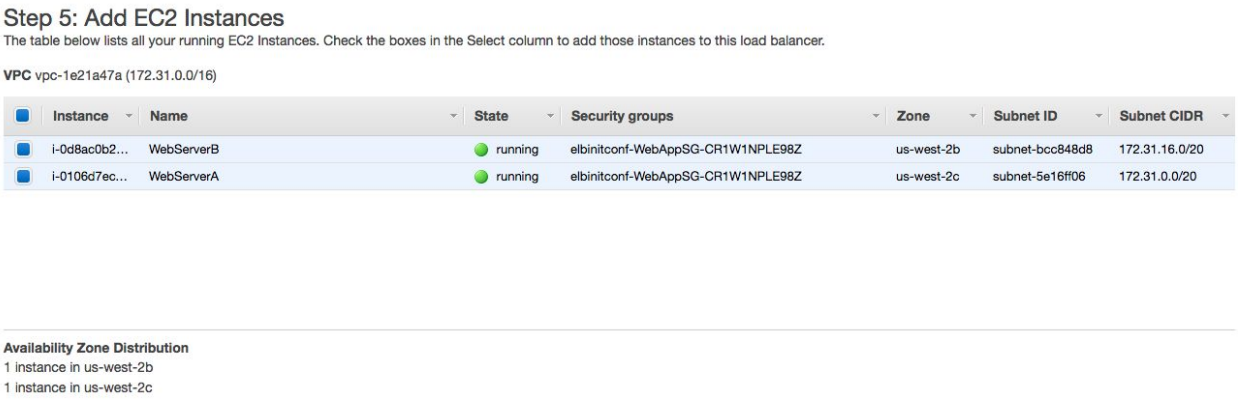
Response Timeout:	The amount of time to wait when receiving a response from the health check, in seconds. Valid values: 2 to 60. Default: 5
HealthCheck Interval:	The amount of time between health checks of an individual instance, in seconds. Valid values: 5 to 300. Default: 30
Unhealthy Threshold:	The number of consecutive failed health checks that must occur before declaring an EC2 instance unhealthy. Valid values: 2 to 10. Default: 2
Healthy Threshold:	The number of consecutive successful health checks that must occur before declaring an EC2 instance healthy. Valid values: 2 to 10. Default: 10

http://docs.aws.amazon.com/elasticloadbalancing/latest/classic/elb-healthchecks.html

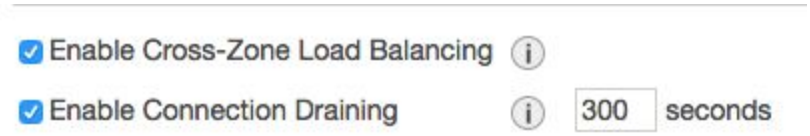
Click on **Next: Add EC2 instances** to move on



In step 5, it is time to add EC2 instances to the LB. The first thing to do is to select the instances called **WebServerA** and **WebServerB**:



There are 2 config points in here as well:

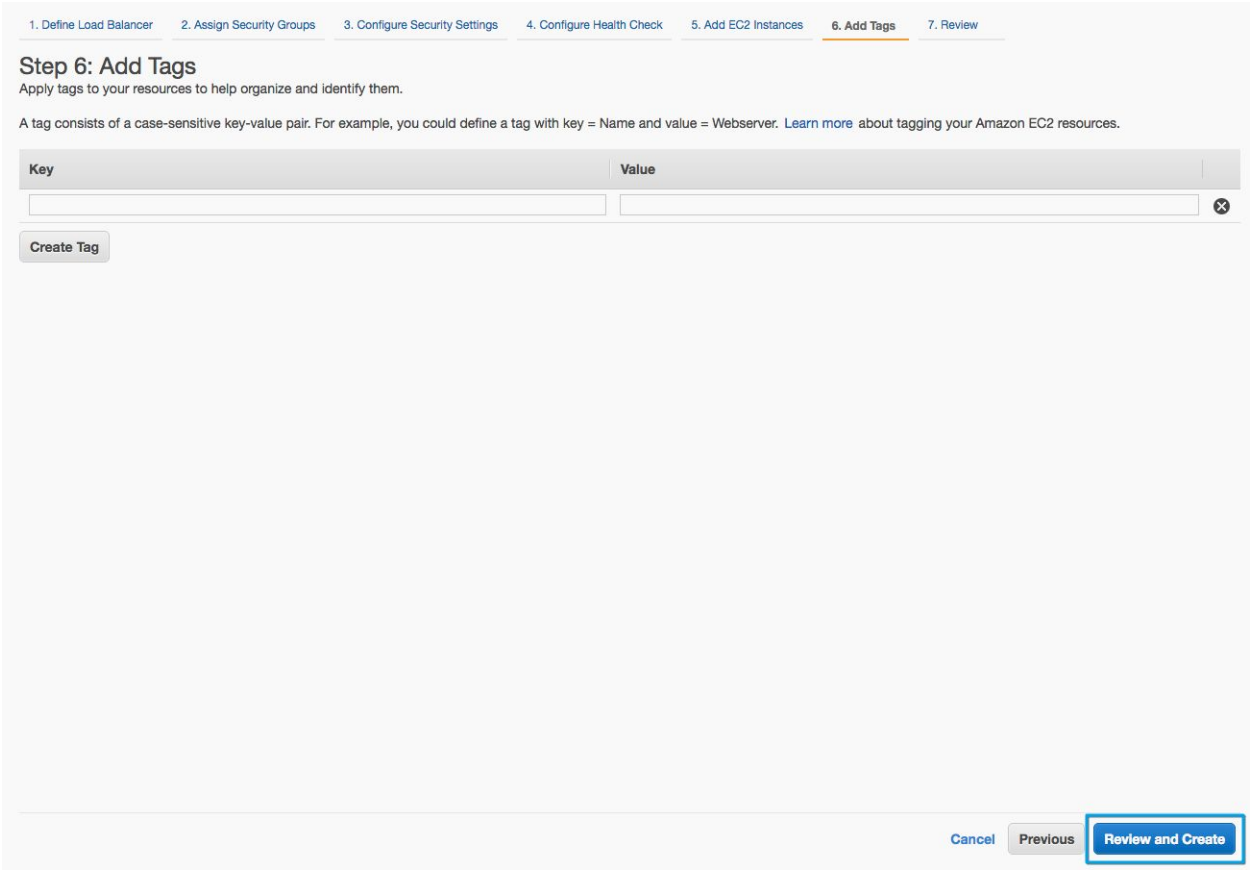


Cross-Zone Load Balancing is used to ensure that your LB distributes incoming requests evenly across all instances in its enabled Availability Zones. That means that the LB will ignore the default of round-robin and will also take into consideration the Availability Zone in which the instance is running. This reduces the need to maintain equivalent numbers of instances in each enabled Availability Zone, and improves your application's ability to handle the loss of one or more instances.

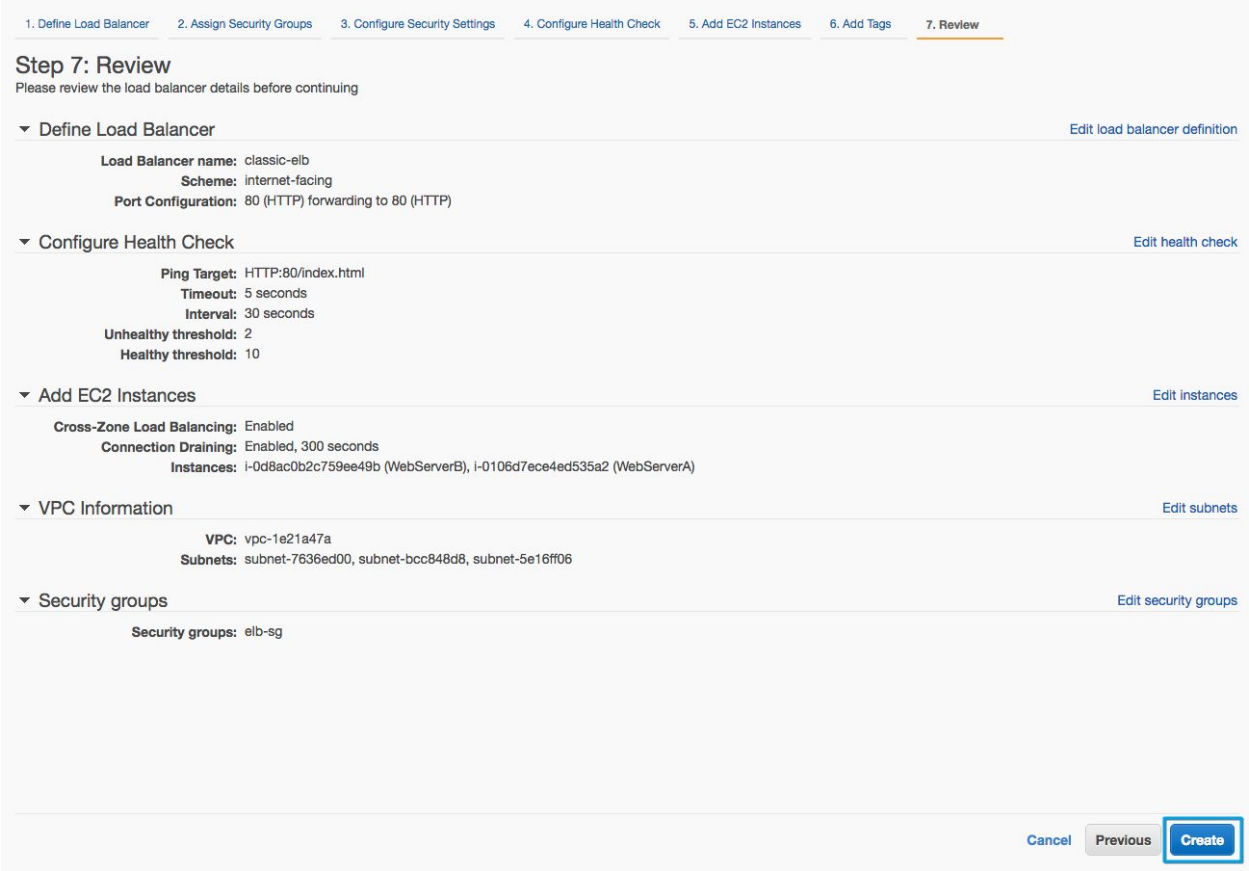
Connection Draining is used to ensure that a Classic Load Balancer stops sending requests to instances that are de-registering or unhealthy while keeping the existing connections open.

For the purposes of this lab, you can use the default settings and click on **Next: Add Tags** to move on.

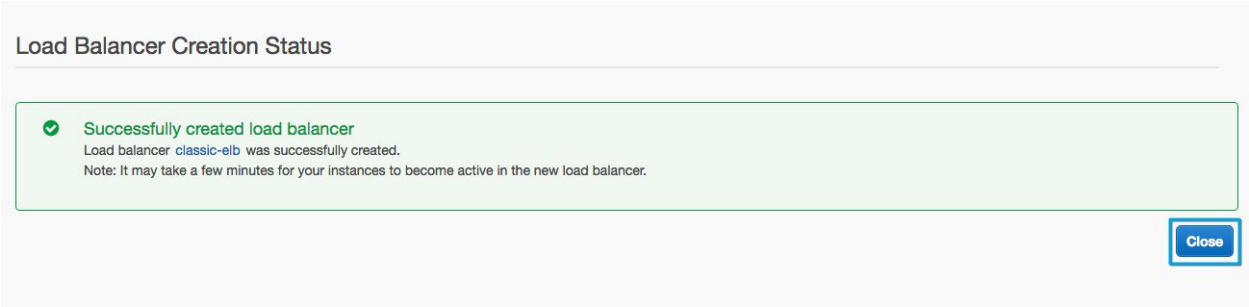
In Step 6, you have the ability to add tags to the LB. You can either leave it in blank or add as many tags as you want and click on **Review and Create** to move on.



In Step 7, it is time to review the LB's settings, double check the config points and click on **Create:**

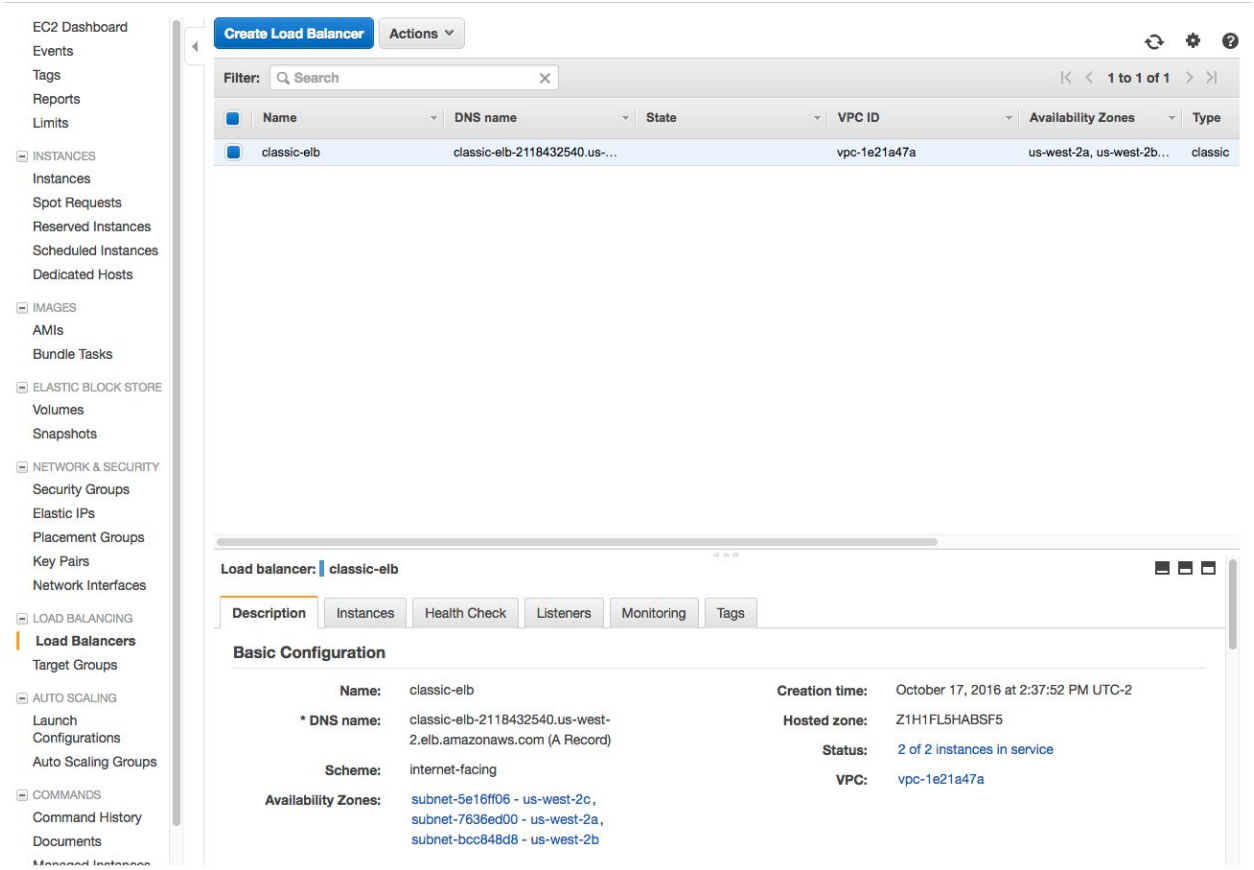


If you see a Success message it means that you LB is created, you can click on **Close** in the AWS console, and move on the the next lab step



Step 4 Configuring security groups for ELB

Now that you completed the creation of you first Classic Load Balancer, you should be seeing this screen:

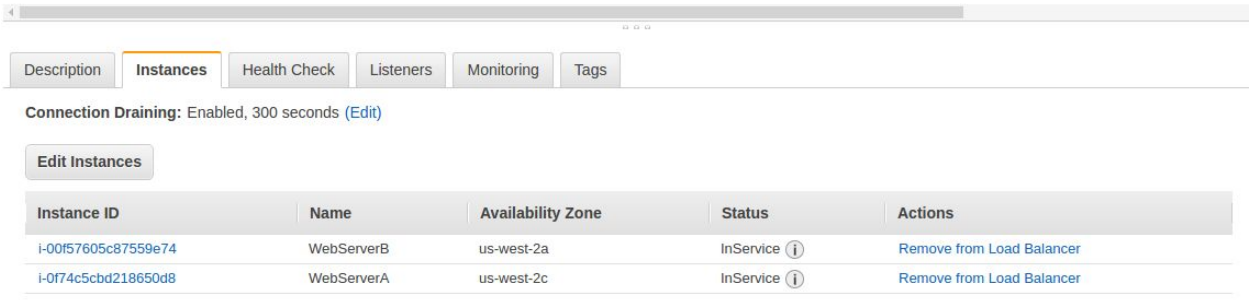


In this step, we will configure the Security Group (SG) associated with the EC2 instances running in the account to accept only connections coming from the LB.

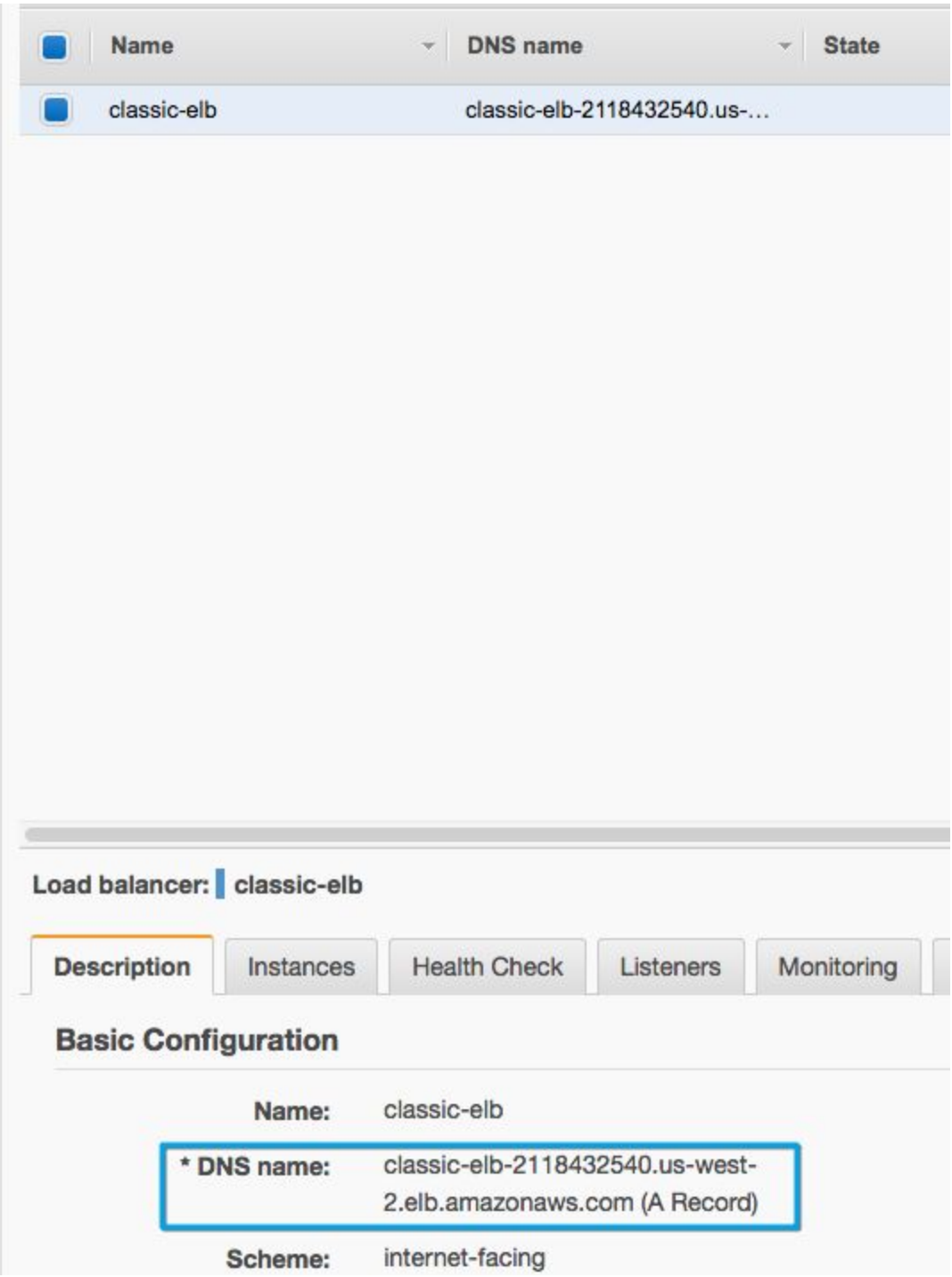
But before that, let's try to access the load balancer through its DNS address.

In order to do it, ensure that both instances are correctly attached to the Load Balancer and their status is *InService*, If not, just wait a few minutes.

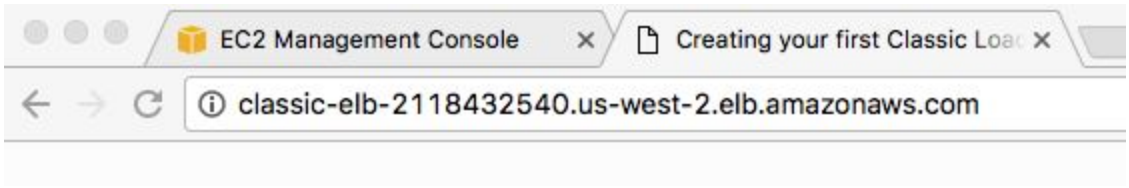
You can check the status of the instances selecting the Load Balancer and going in the *Instances* tab as you can see in the screenshot below.



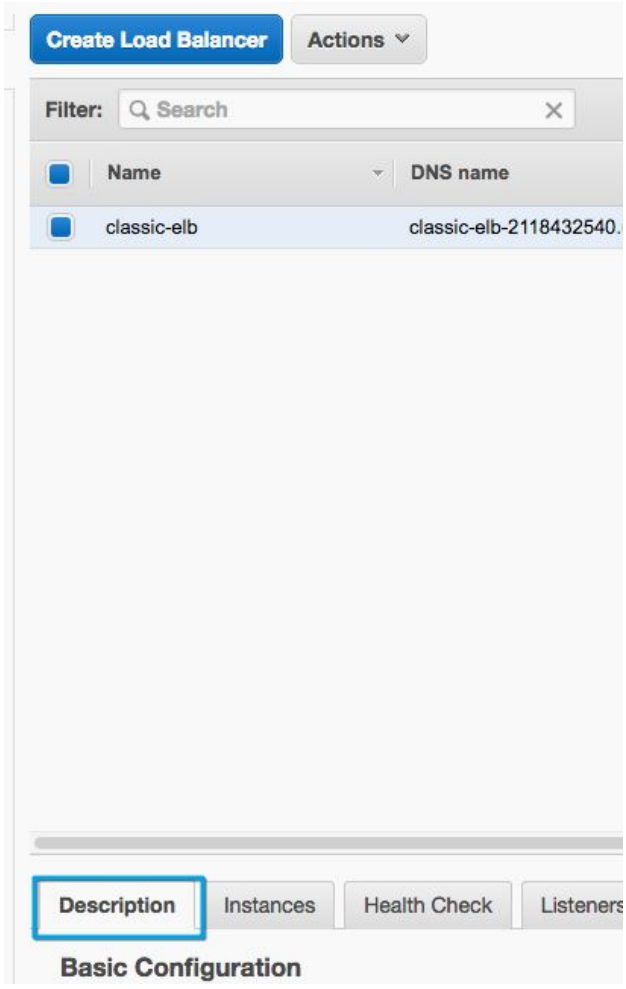
After that both the instances are *InService* copy the Load Balancer **DNS Name.**



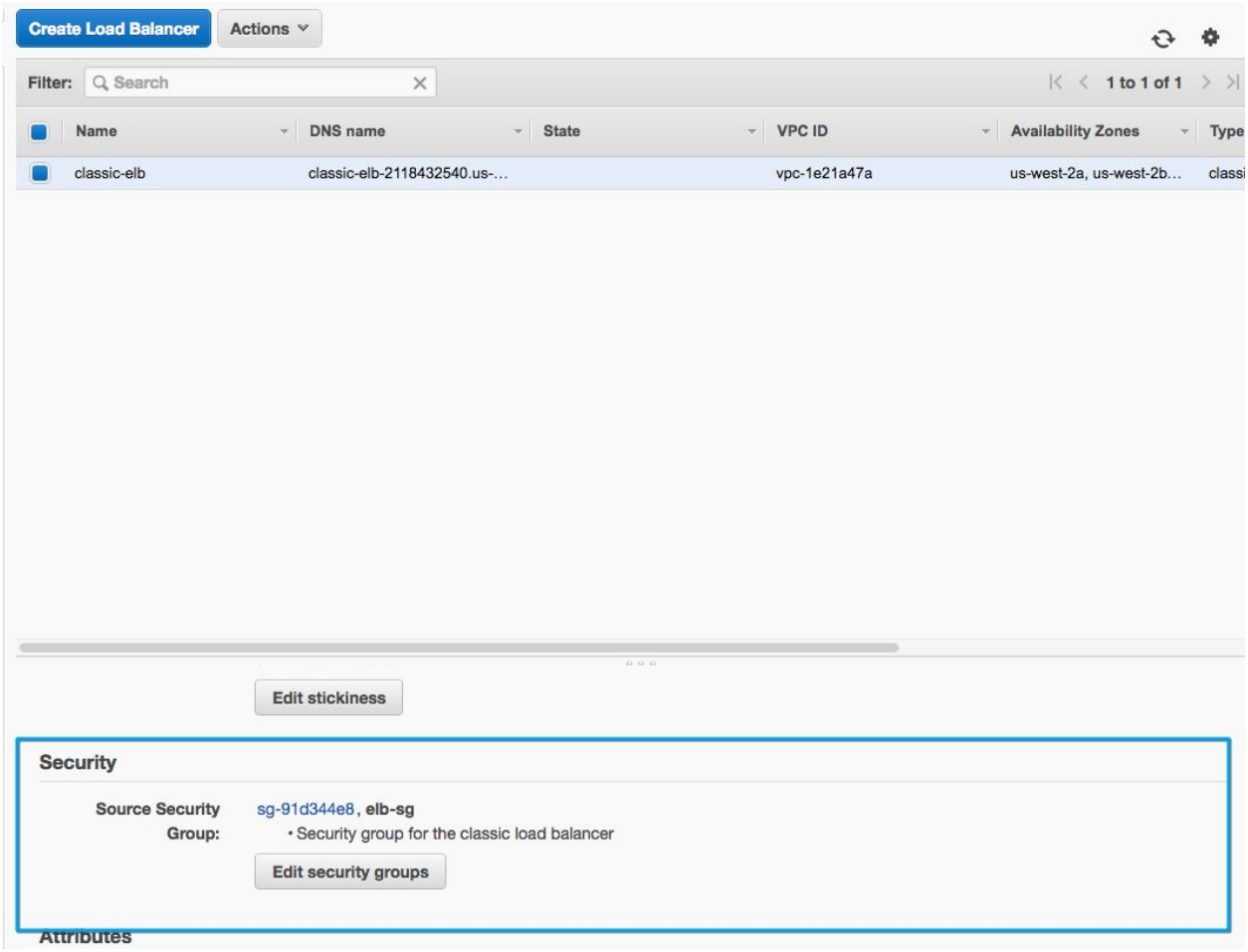
Paste this address into a new tab in your browser (don't include the "A Record" information in parentheses):



Now you are accessing a particular instance directly, and we want to avoid that. To do so we need to change a few things. This is happening because the SG associated with the EC2 instances is allowing access from anywhere to the port 80; we want to change it in a way that the instances will only allow traffic coming from the LB we just created. To configure this go back to the **EC2 management console** on the **Load Balancers** page:



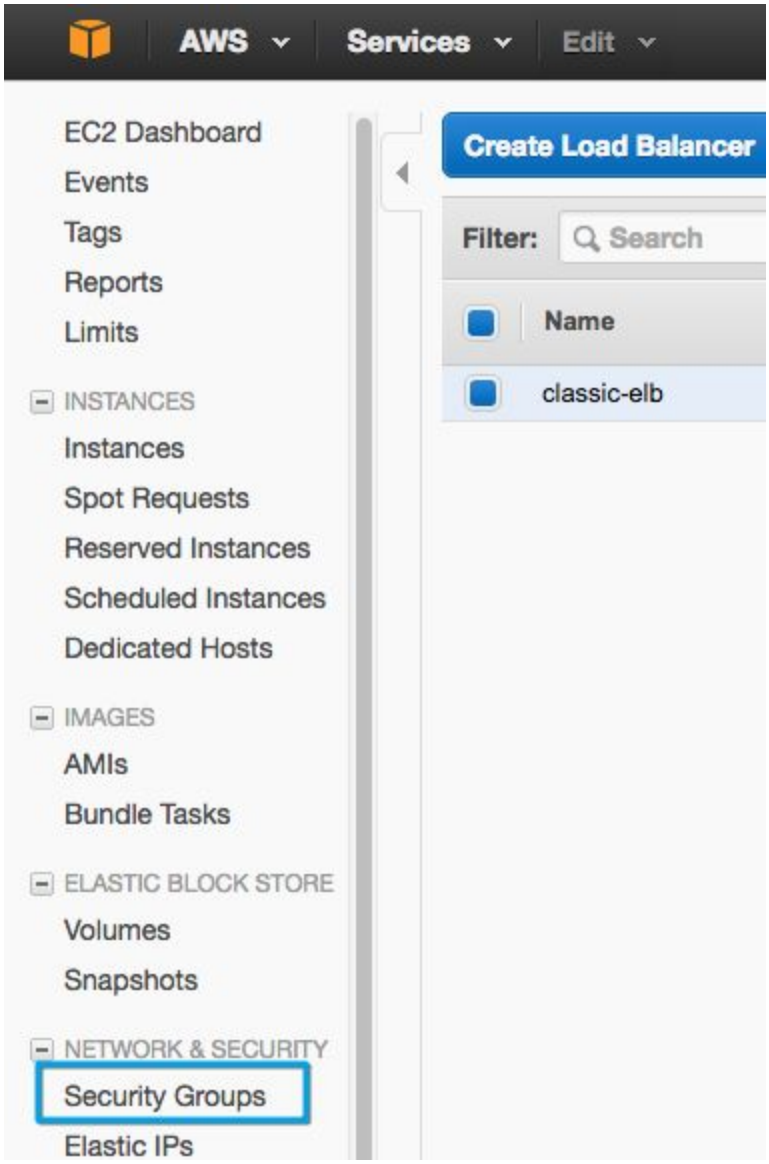
And scroll down until you see **Security**:



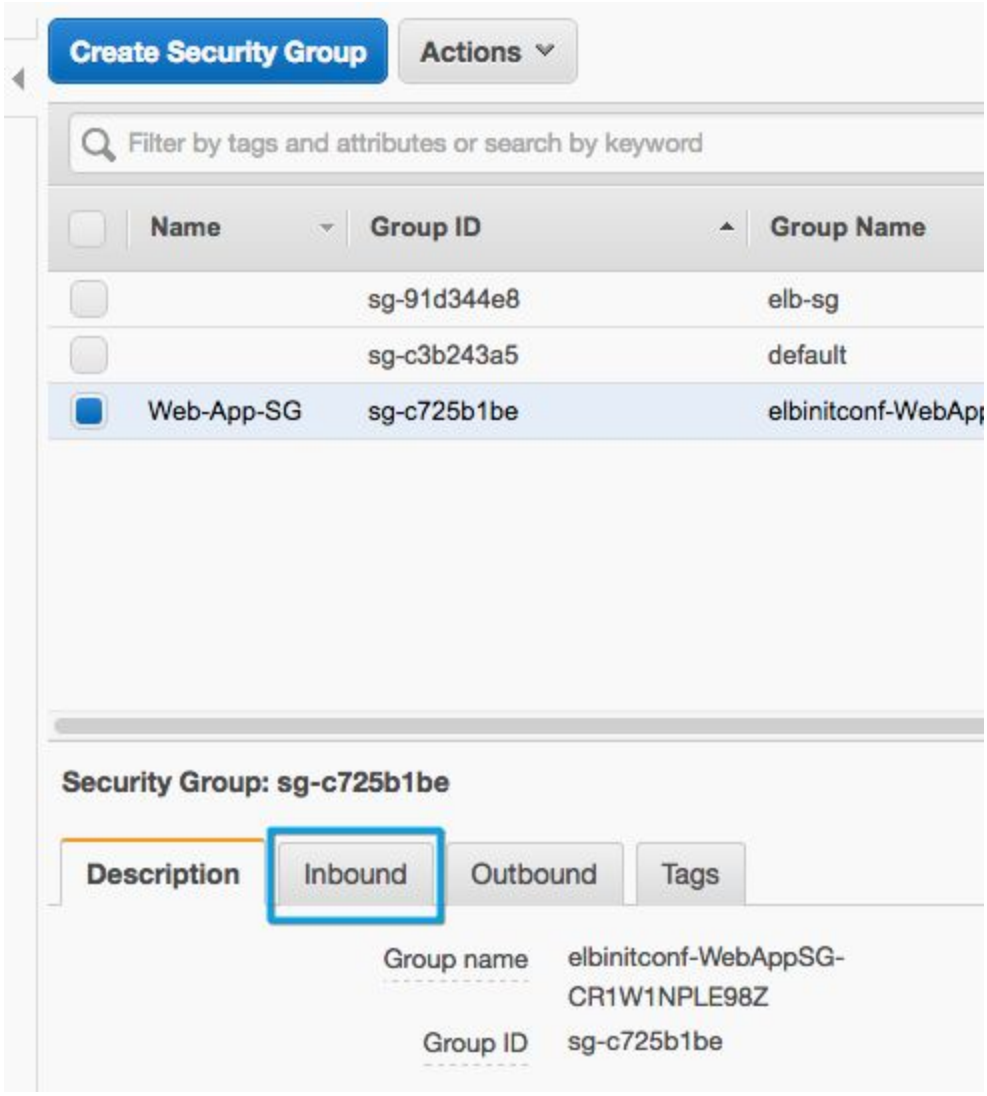
Copy the unique identifier of the LB's Security Group:

We will use this information in just a moment.

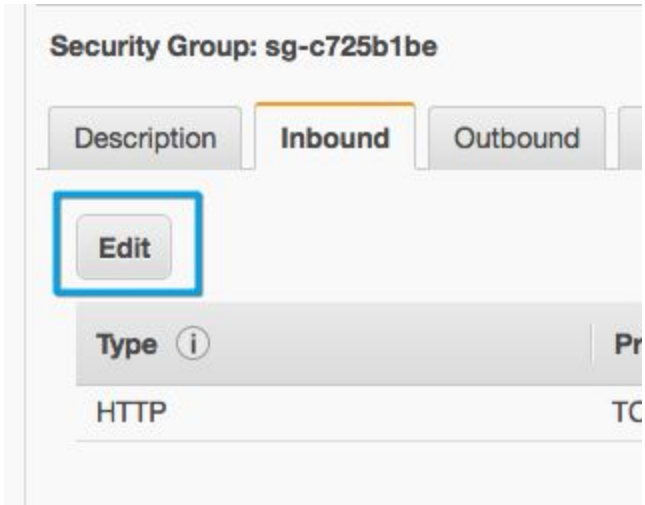
Now click on **Security Groups:**



Select the SG called `Web-App-SG` and click on the **Inbound** tab:

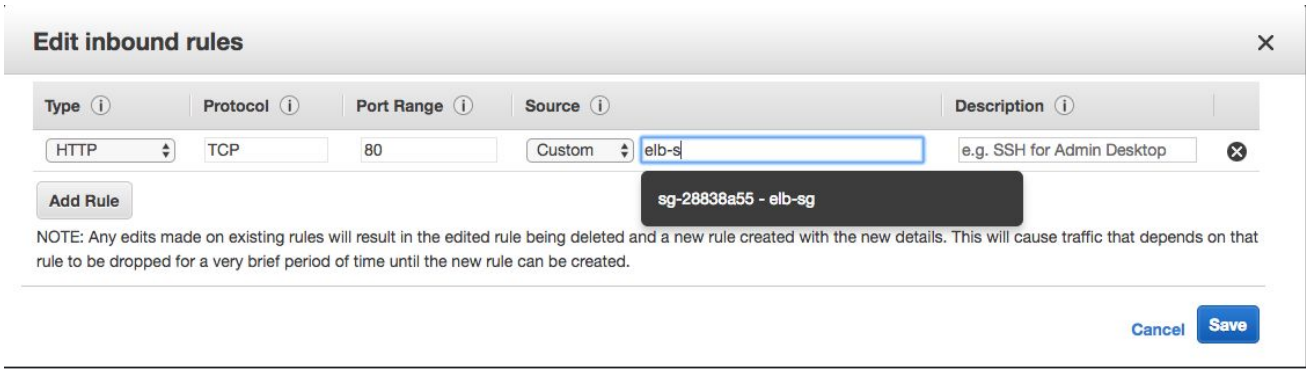


Now, click on **Edit** to change the current rules associated with this SG:

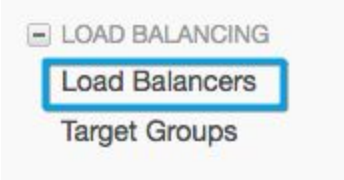


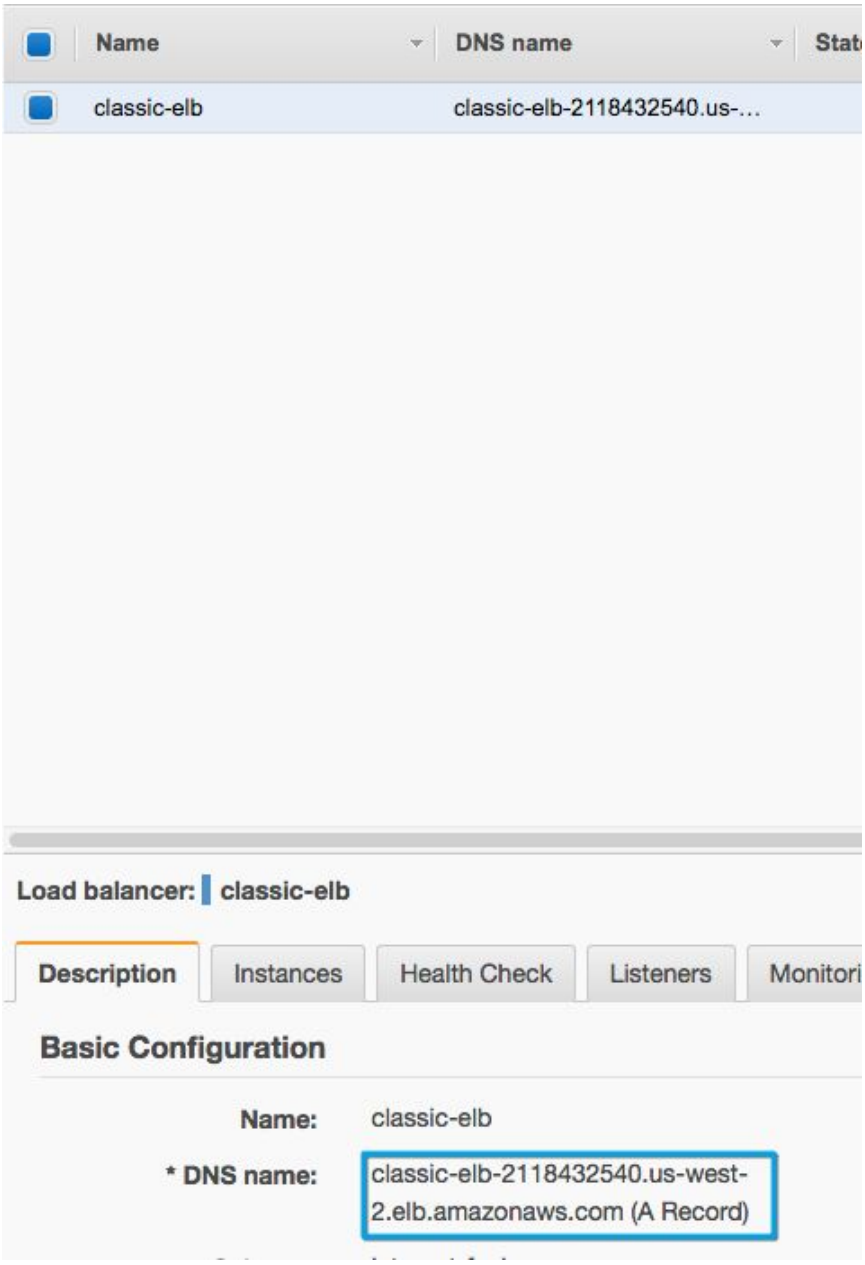
We want to allow only connections coming from the LB to the instances, however, the **LB doesn't have a particular IP address** associated with it so we can't specify an IP address in here. Instead, we will restrict the access by using the SG we just created for the LB. We will change the current rule to deny access to anywhere and allow it only to members of the LB's security group. The process is very straight forward, in the source column, select custom and then simply replace the source in the HTTP rule that is already created with the LB's security group identifier that you just copied:

You can also start typing "sg-" and select the correct SG identifier in the list that will appear. Then click **Save**.

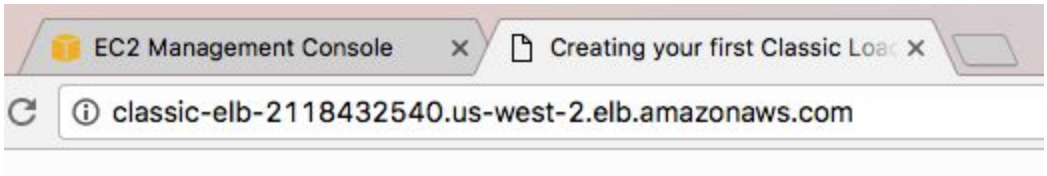


Now let's test the rule. Click on **Load Balancers** to be able to copy the URL for the LB again:

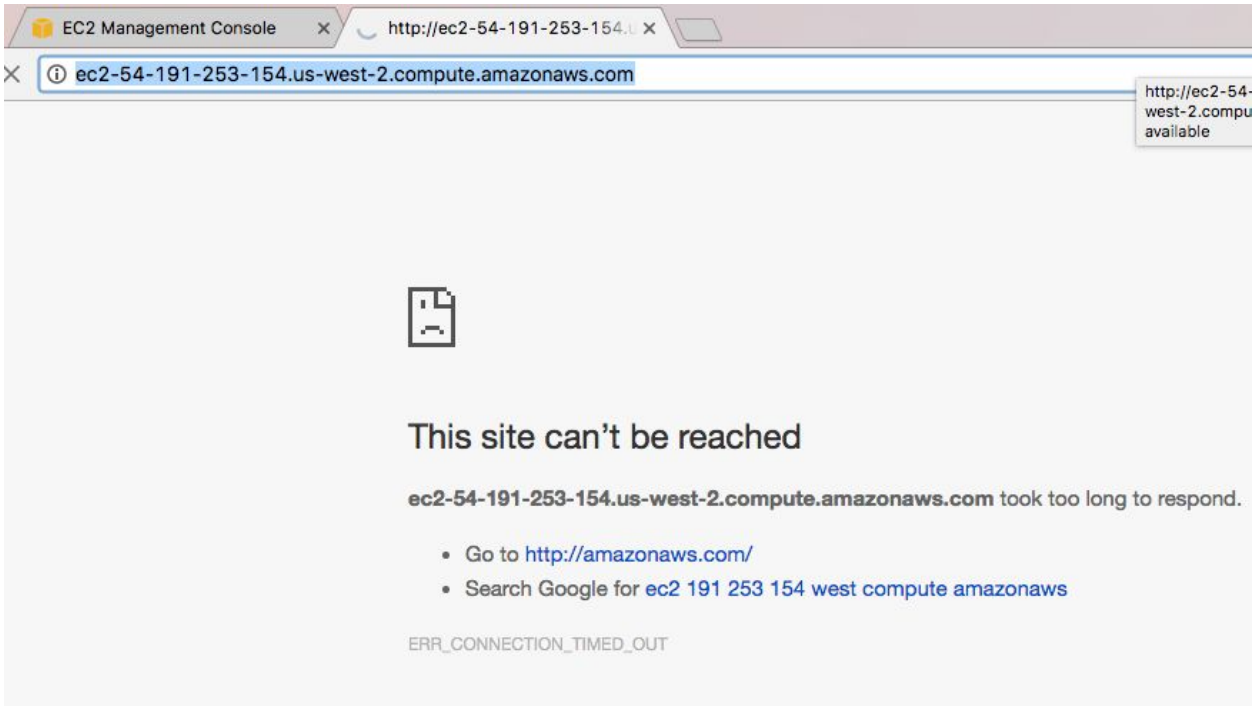




Copy the **DNS Name** address and past it into a new tab in your browser:



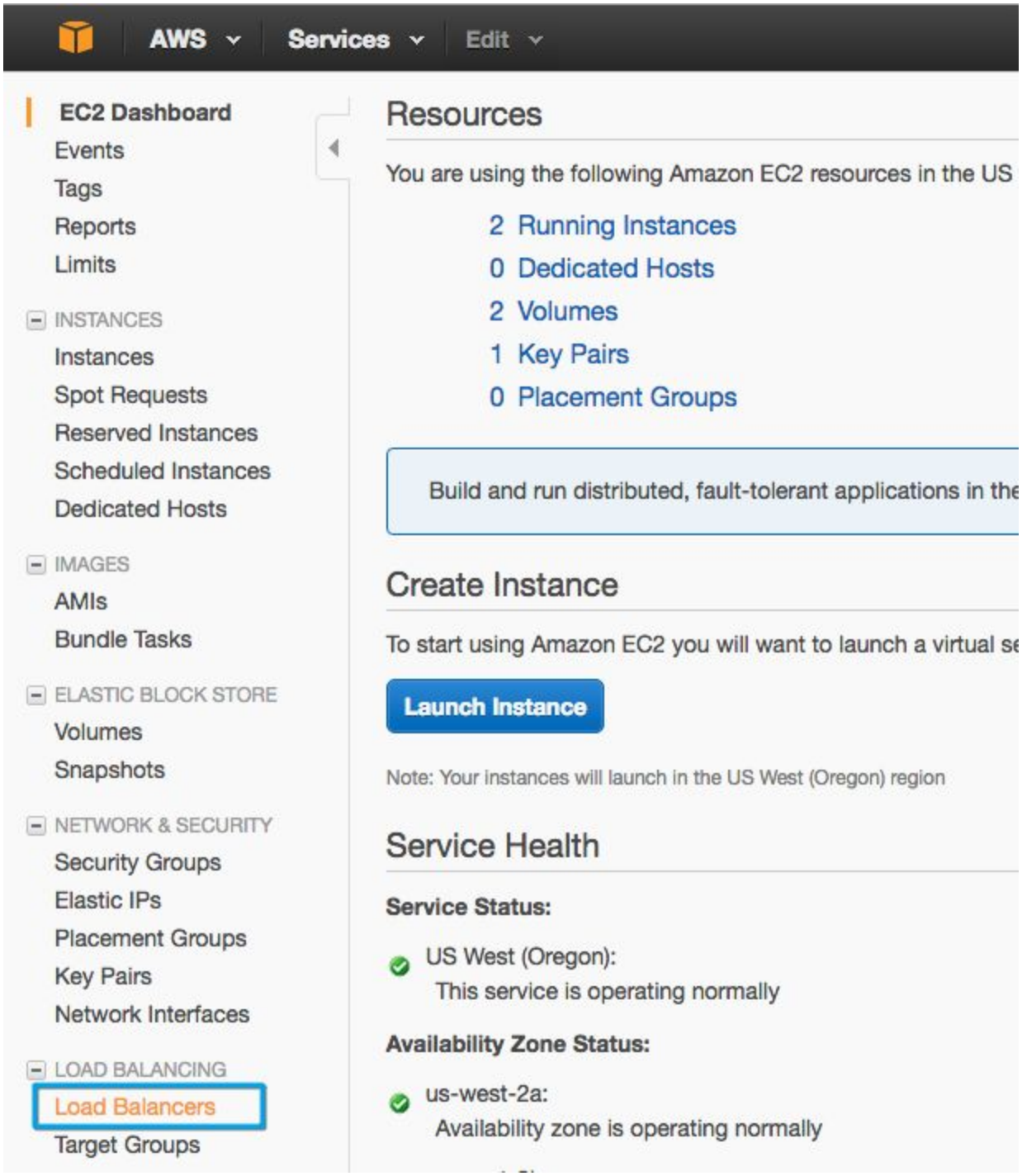
Nothing new here. However, if you try to copy and past the public DNS address of the instance you are accessing, and paste it into your browser, you shouldn't be able to access the app - the connection will timeout.



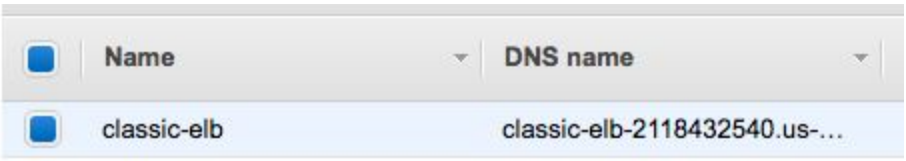
Step 5 Checking a load balancer's behavior during instance failures

In this step, we will make an EC2 instance fail and see how the ELB service responds to that. To do so, go to the EC2 console:

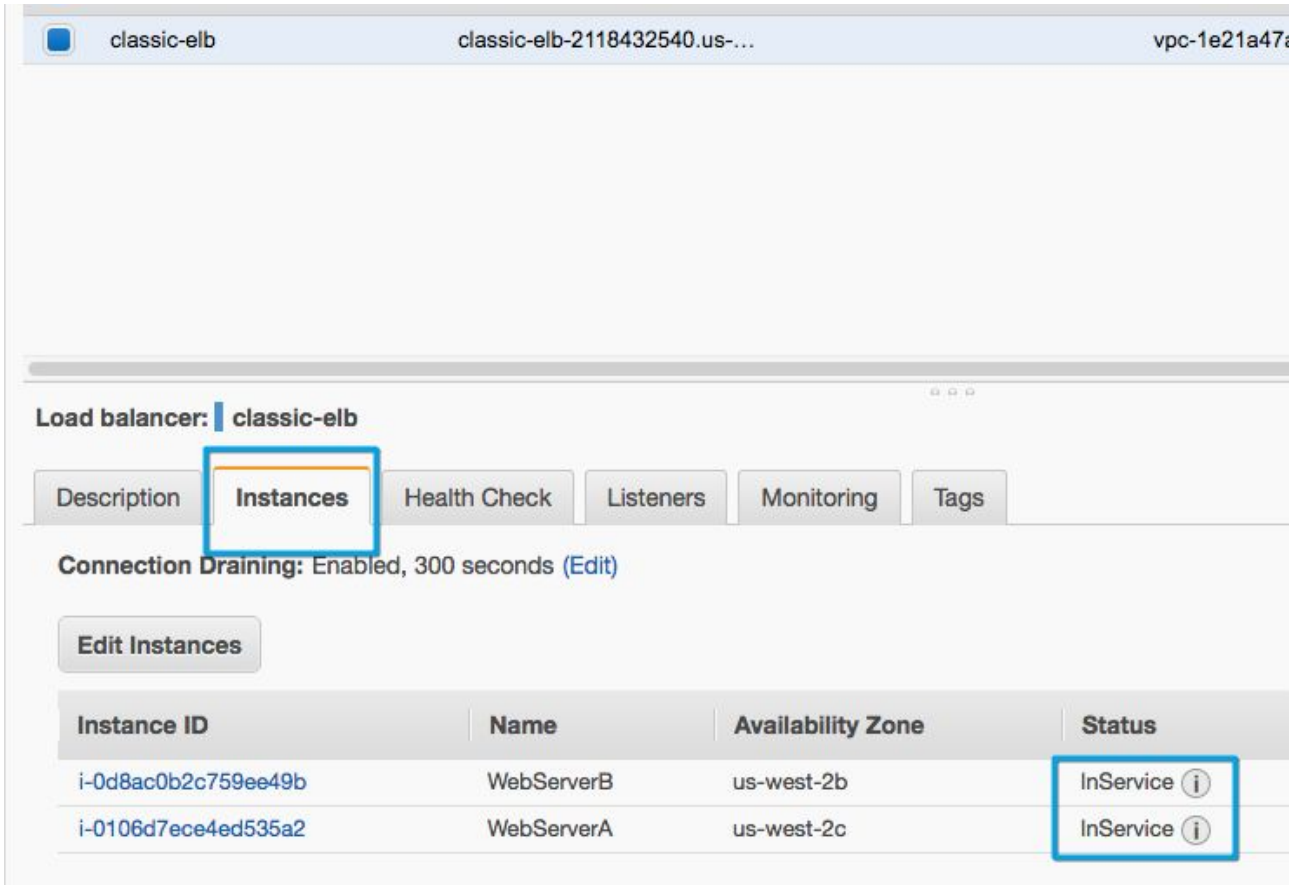
And click on Load Balancers:



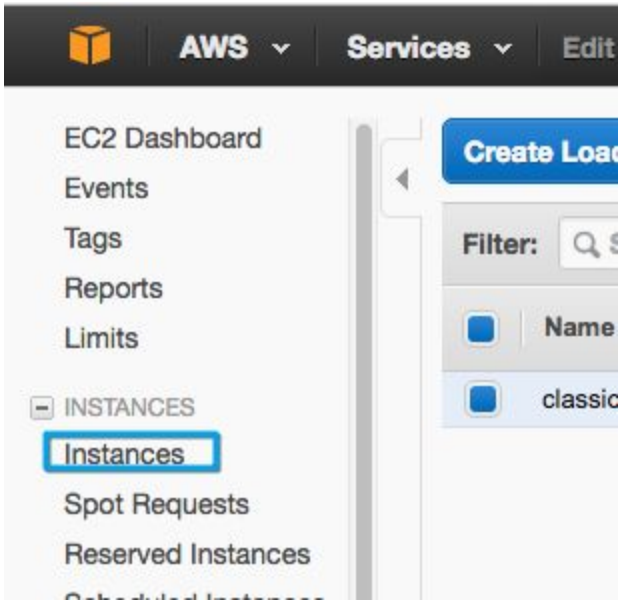
Then select the LB you just created:



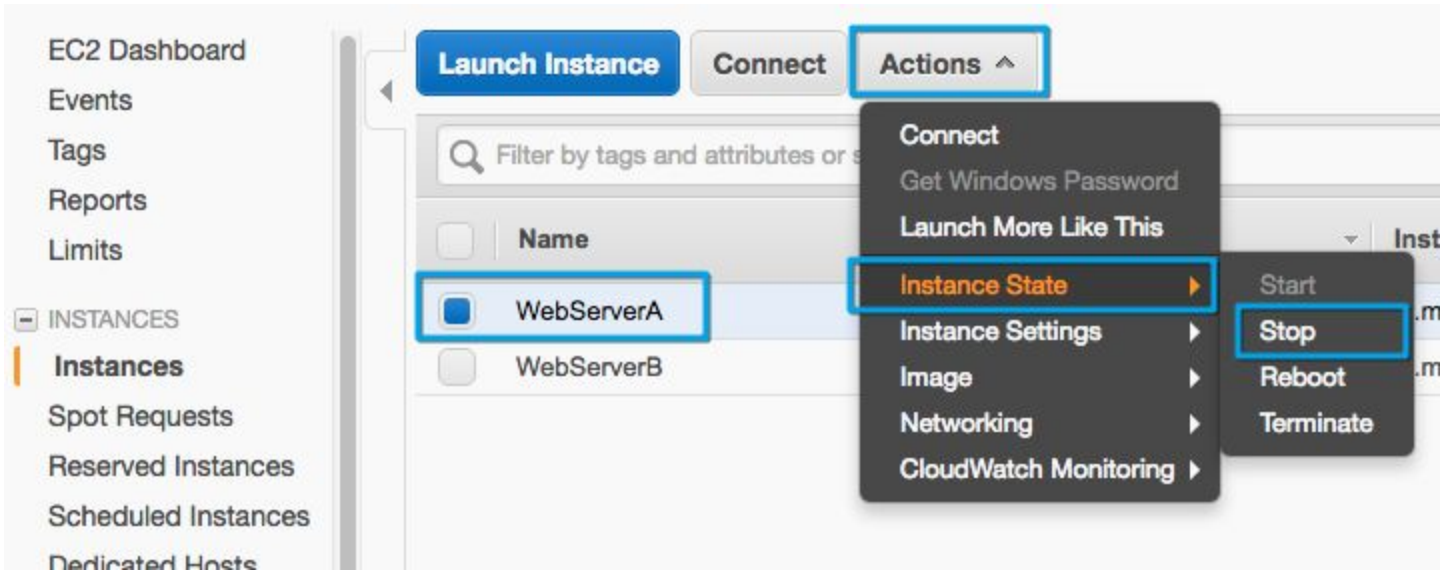
If you click on **Instances** you will be able to see that both instances are currently on service:



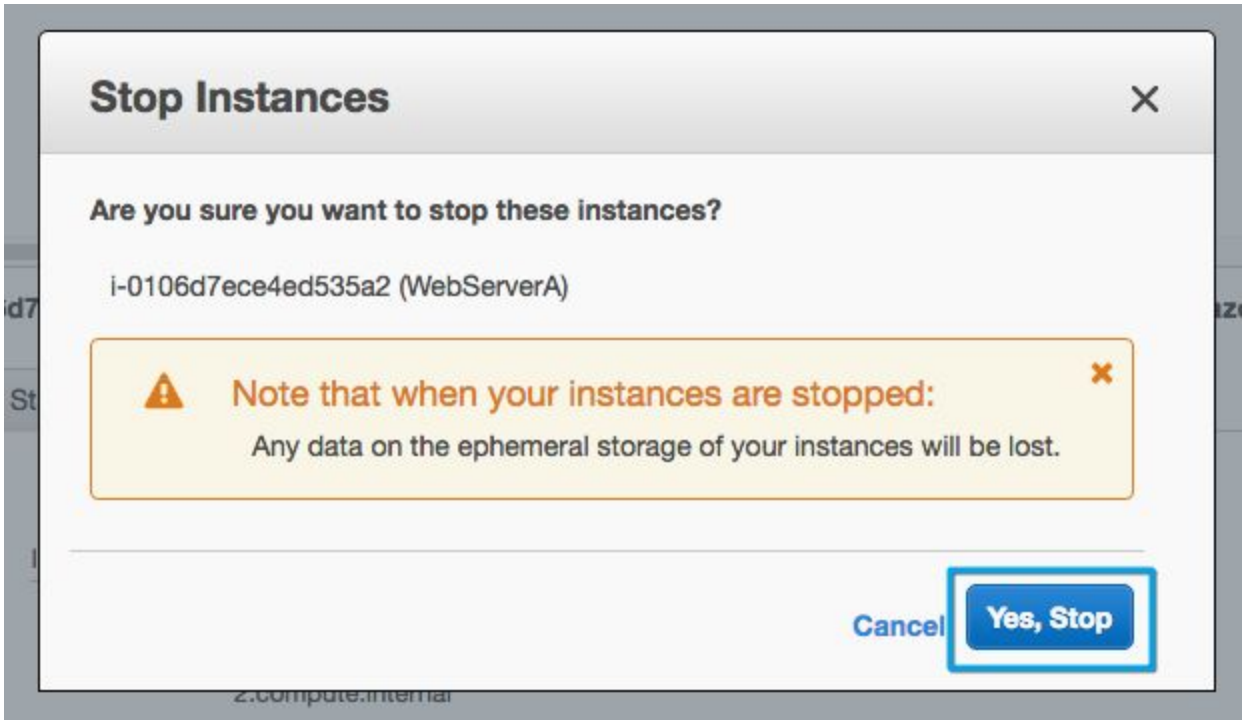
This means that the instances are answering the health checks we configured during the LB's creation, and thus the ELB service has decided to put them **InService.** To see what happens with the LB when there is a problem in a particular instance, let's simulate an instance failure. The easiest way to do this is to stop a running instance, so let's do that. Click on **Instances**:



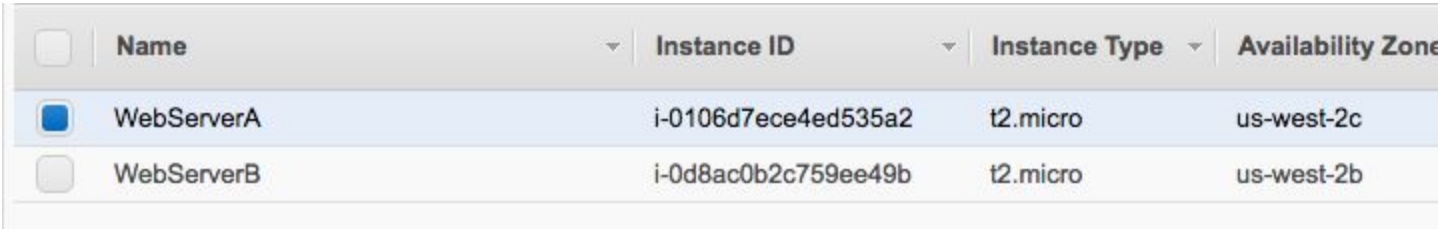
And choose any of the running instances, click on **Actions, Instance State,** then on **Stop**:



Then click on **Yes, Stop** on the dialog box that pops up:



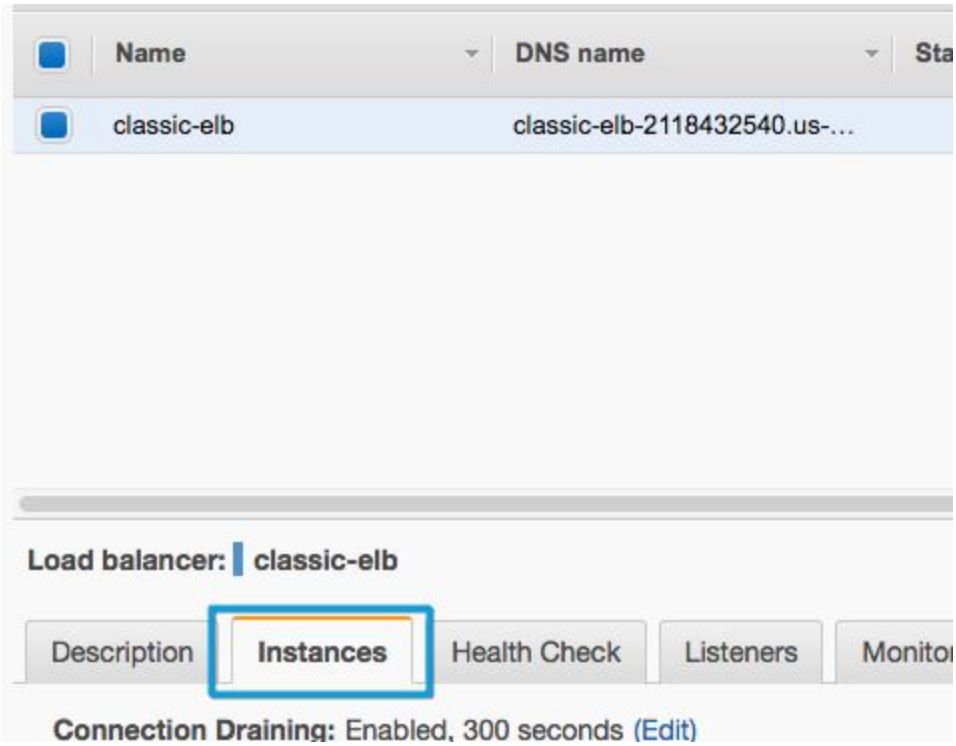
Doing this will stop a particular instance, which will make it fail the ELB's health checks.
Once the **Instance State** changes to **stopped** you can proceed:



Name	Instance ID	Instance Type	Availability Zone
WebServerA	i-0106d7ece4ed535a2	t2.micro	us-west-2c
WebServerB	i-0d8ac0b2c759ee49b	t2.micro	us-west-2b

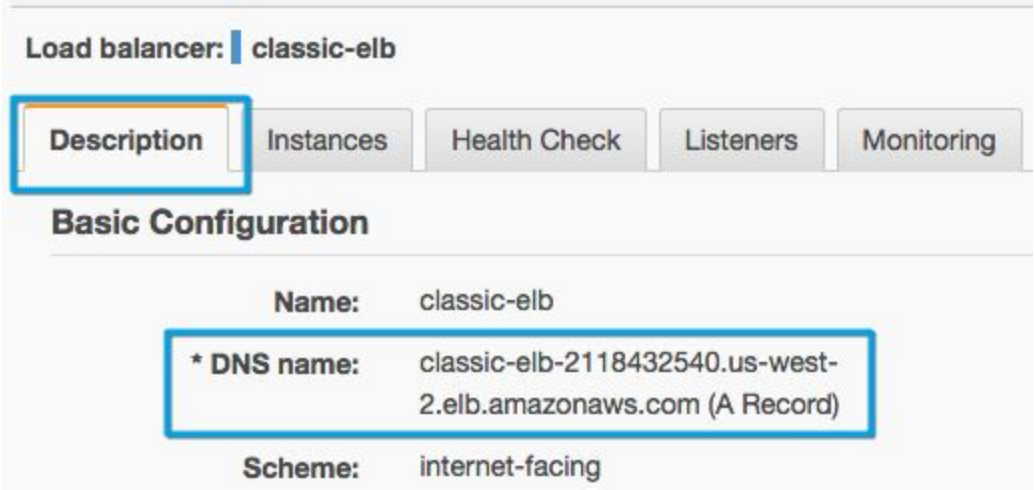
Now click on **Load Balancers**:

Select the LB and click on **Instances**:

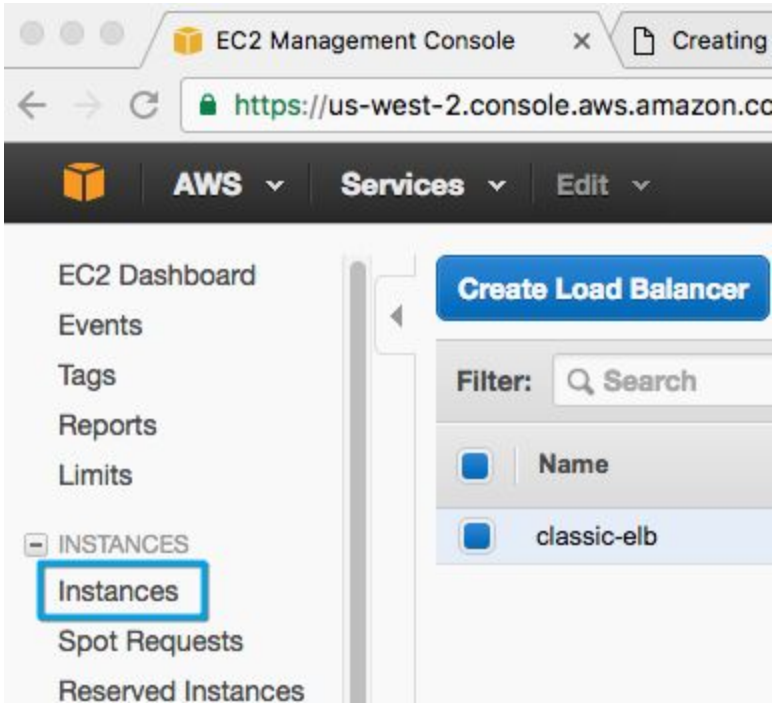


Now you should be able to see an instance with the status **OutOfService**

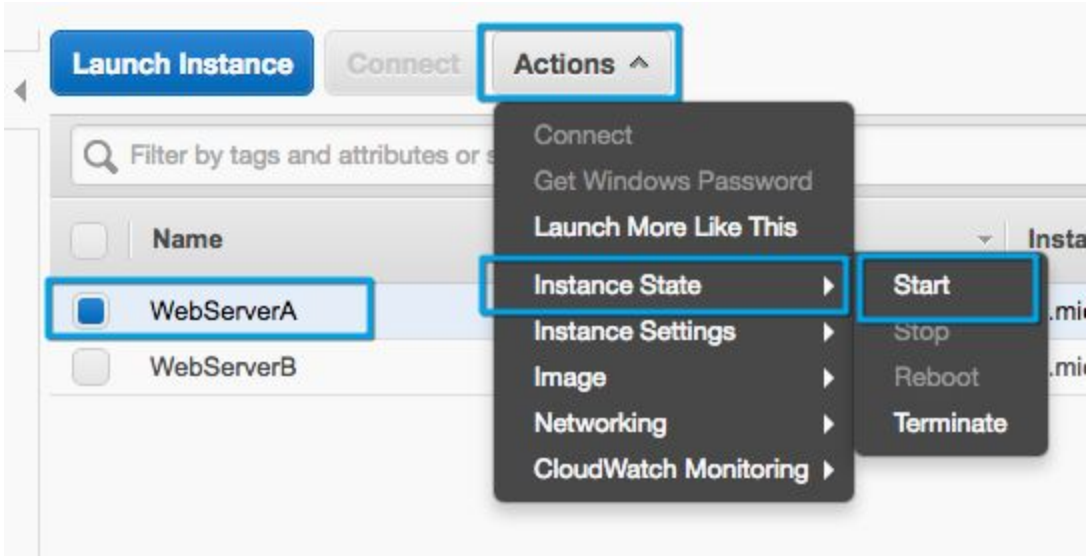
This means that there is only one instance serving the application, and therefore all the requests will be forwarded to the same instance. You can test this behavior by clicking on **Description** and accessing the **DNS name** of the LB:



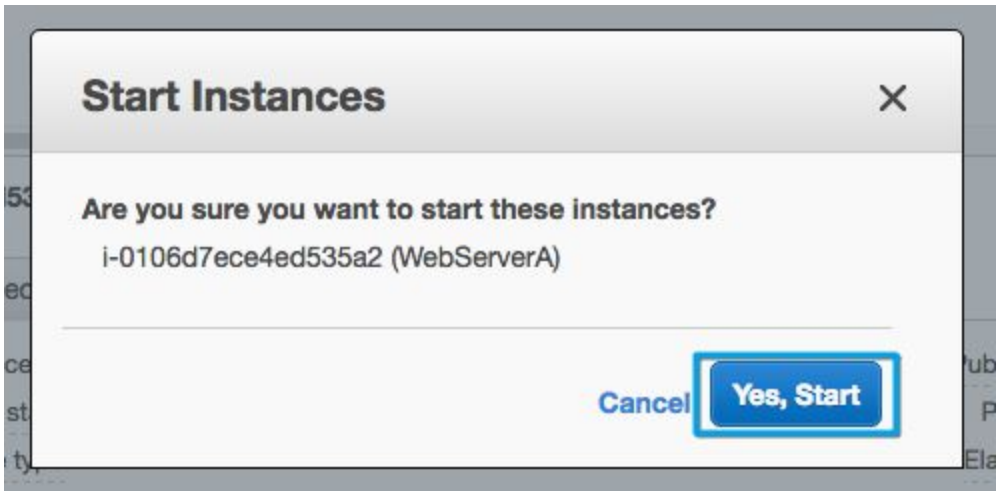
At this stage, no matter how many times you hit the refresh button, you will always be forwarded to the same instance. Hit refresh a few times, then go back to the **EC2 Management Console**, and click on **Instances**:



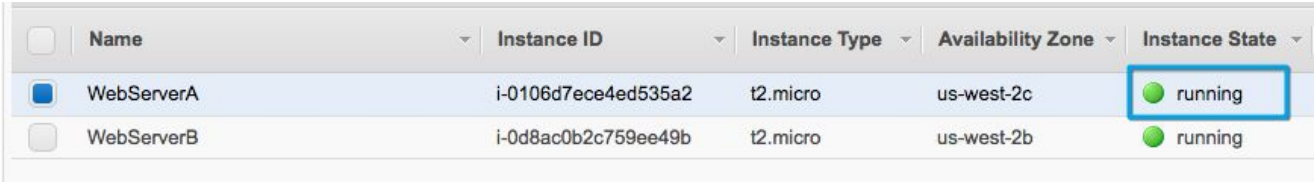
Start the stopped instance again by selecting it, clicking on **Actions, Instance State,** then on **Start**:



Click on **Yes, Start** in the dialogue box that pops up:

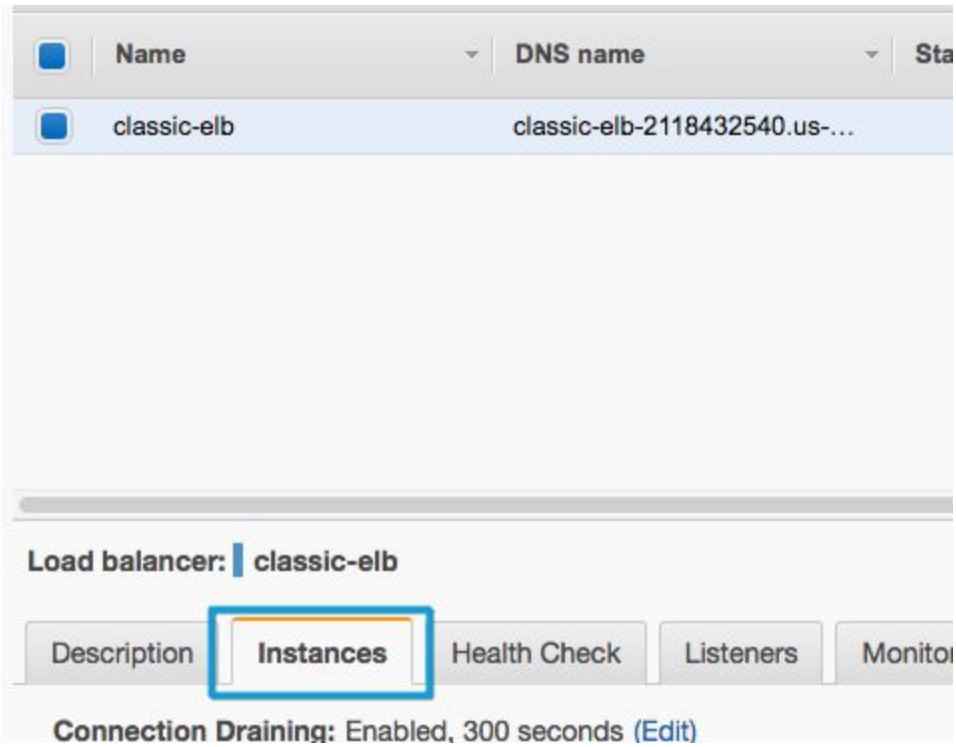


Wait until the **Instance State** changes to **running**:

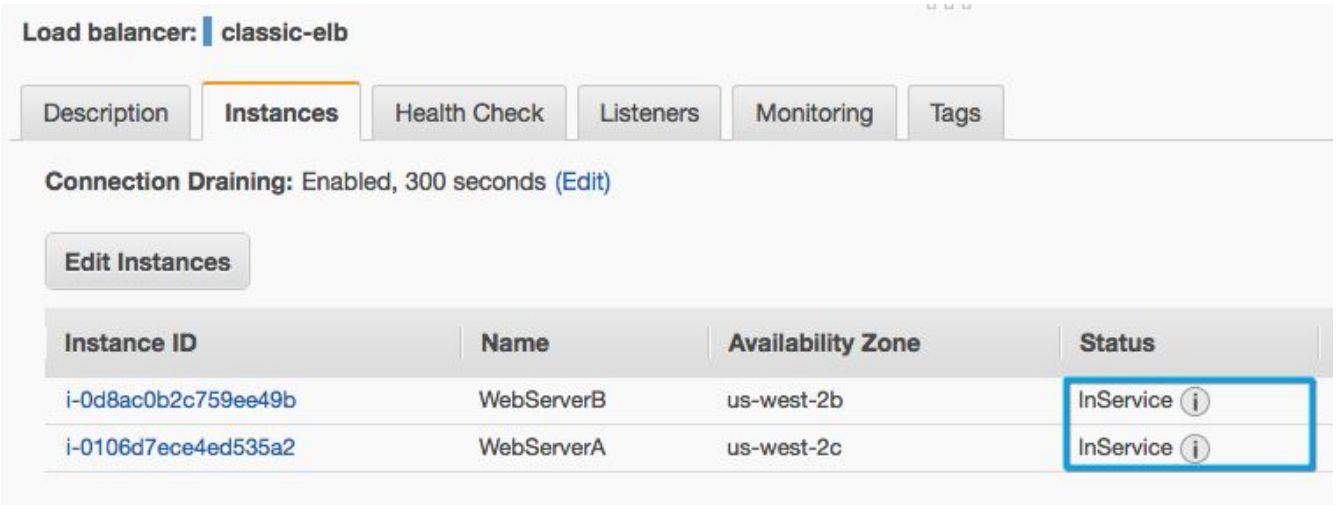


Then click on **Load Balancers**:

Select the LB and click on **Instances**:



And you should be able to see that all the instances have an **InService** status again:



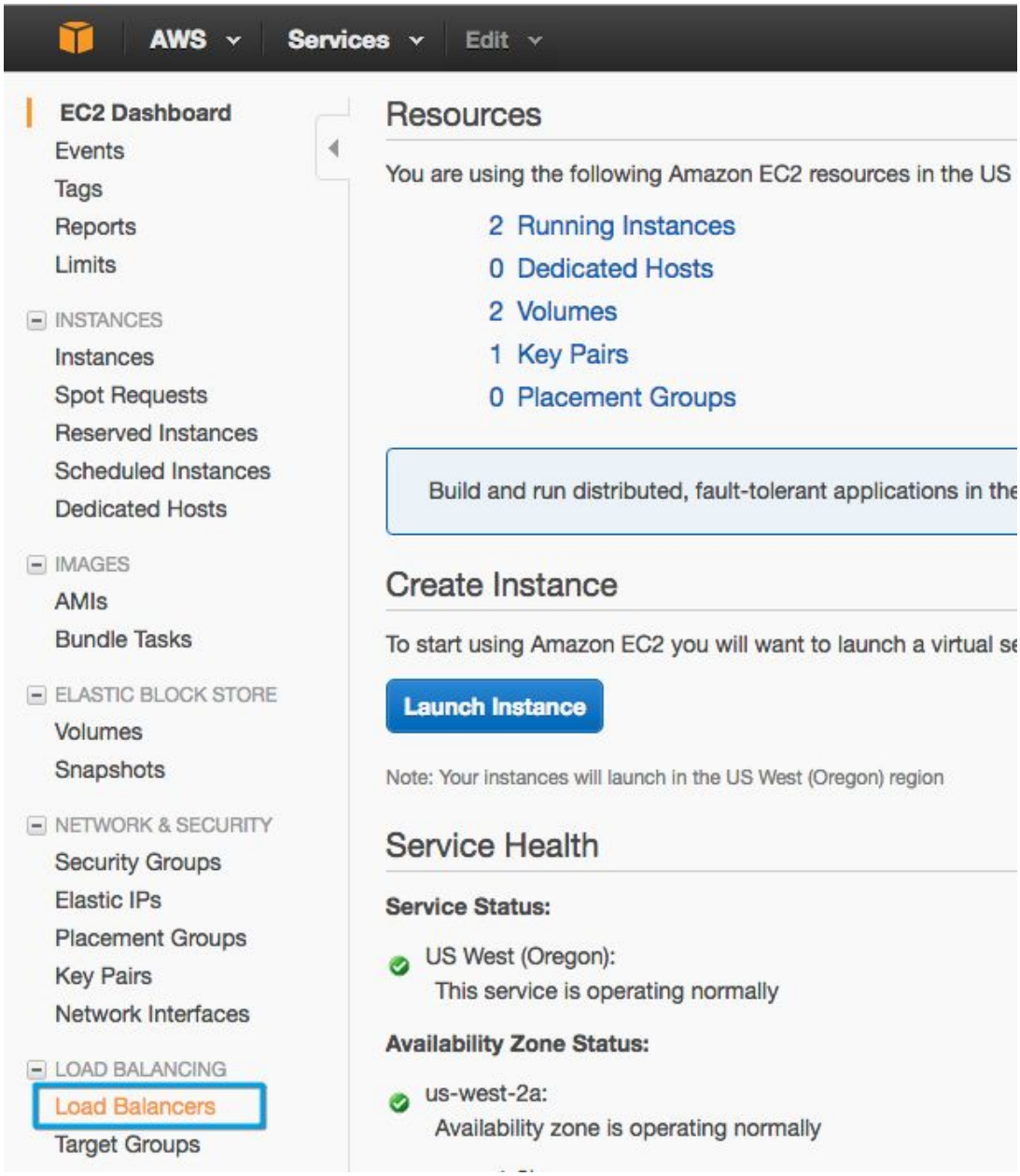
You can test the LB's behavior again by accessing its **DNS Name** in your browser. Once you're done testing move to the next lab step.

Step 6 Monitoring your Classic Load Balancer

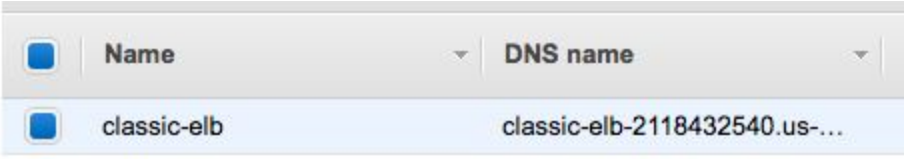
In this step, we will take a quick look at the most common metrics for troubleshooting problems with your Classic Load Balancer and the EC2 instances running behind it.

There are two ways of doing that. One is on the CloudWatch console, which can be a bit frustrating for newcomers because it will hold metrics for ALL LBs that existed within the past 2 weeks in the AWS account you're using, and you might get lost with so many metrics. With that in mind, we will take a different approach, and we will take a look at the metrics related to our LB in the EC2 console. To do so, go to the EC2 console:

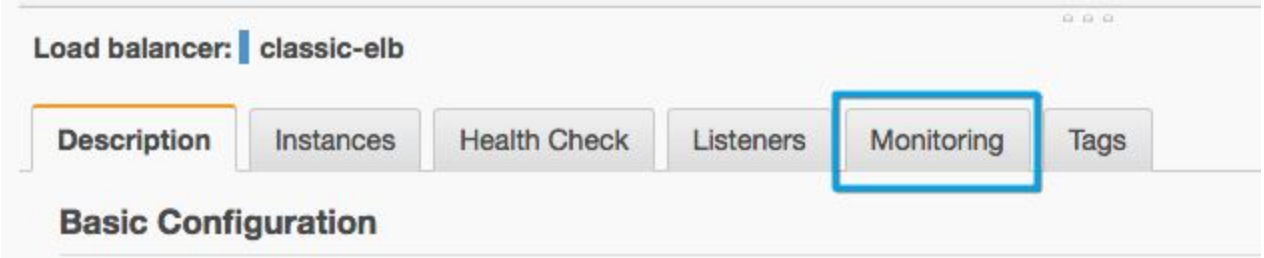
And click on Load Balancers:



Then select the LB you just created:



Click on the **Monitoring** tab to see the metrics of the LB you selected:



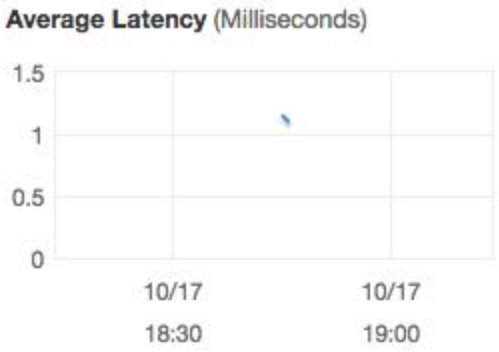
The ELB service reports metrics to CloudWatch only when requests are flowing through the LB. If there are requests flowing through the LB, ELB measures and sends its metrics in 60-second intervals. If there are no requests flowing through the load balancer, or no data for a metric, the metric is not reported. There are a few metrics related to a Classic Load Balancer, and in general they are self-explanatory. However, some of them may be unfamiliar to you, in which case you can take a look at the description for all metrics in here: http://docs.aws.amazon.com/elasticloadbalancing/latest/classic/elb-cloudwatch-metrics.html

The metrics called **HealthyHostCount**, and **UnHealthyHostCount** will count the number of Healthy and Unhealthy instances respectively. These metrics can be useful for you to identify a major problem in your AWS account. For example, you could set up some CloudWatch Alarms to notify you when you have less than 2 instances running your application, though to be clear this is not a general rule: the number of instances that might identify a problem will vary depending on your environment.

Also notice that in these metrics, there is no way of seeing the Availability Zone to which the Healthy/Unhealthy instance belongs. In our lab, we stopped an instance for a few minutes, therefore you should be able to see something like this:

If the **Healthy Hosts** metric reaches 0, that means that people won't see anything when accessing your LB, and it is probable that you have a big problem in your infrastructure.

The **Average Latency** metric might be useful to identify potential issues in your setup. Maybe everything is working in your application, but you notice an increase in this metric. If you haven't changed anything in your application, that can be a potential issue - maybe you haven't provisioned enough EC2 instances, or you even have lots of instances but they don't have enough power to serve your increasing traffic.



The other metrics can be very useful for troubleshooting specific scenarios and will vary depending on your setup.

You have now covered all the Learning Objectives in this lab. You can take some time to play around with the metrics. You can do things such as:

- Stop ALL instances and make requests to the LB
- Make several requests to paths that don't exist, such as <dns name>/app, and <dns name>/users to generate some errors in the responses