Managing Cloud Resources

If you are considering hosting some services in the Cloud, you'll need to learn what the different terms used to configure the services mean.

When deploying a service to the cloud, you will typically create a number of virtual machines that will be the servers in charge of hosting your service. In the usual case, you would start by creating a single machine that will run the service, creating the configuration associated with the machine, verifying that it works, and then turning this into a template that can be used for the creation of many machines as needed.

In order to do this, you'll make use of both **Autoscaling** and **Load Balancing**. Autoscaling means being able to automatically create new instances when the load increases and automatically turn them down when the load decreases. In order for this to be possible, you need to ensure that your instances can be completely configured automatically, and that there's no data being kept in the instances themselves (data can be stored in a database, or in separate drives).

Load Balancing means distributing the load among many servers. There's different approaches to doing load balancing, but the main concept is that there's a load balancing service that will route traffic to the servers in a way that they each get to serve a portion of users, without the users realizing that they are connecting to different machines. In other words, the users will access a single address (e.g. http://www.example.com), which can be served by different servers, in different parts of the world, without the users having to care about that.

Once you have your service set up to scale automatically and balance the load, you'll want to also setup **Monitoring** and **Alerting** for it. Monitoring means checking that the service is healthy, that it's responding to queries as expected and not generating unusual errors. Alerting means sending alerts when things don't happen as expected.

For a simple service, you might go with the monitoring that is already built in by the cloud provider, which will allow you to check that your instance is healthy, but is likely not going to go into much detail as to whether the content is being served correctly. If your service is more complex, you might want to invest more time into making it possible to monitor additional parameters of your service.

Depending on the specific service you are deploying, there might be more concepts that you need to understand before you can actually do it. We recommend reading the documentation offered by the cloud provider you have chosen to figure out what you need to do.

Here are some links with more information from some of the biggest cloud providers:

- https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/
- https://aws.amazon.com/getting-started/
- https://cloud.google.com/docs/overview/

- https://cloud.google.com/docs/overview/