Lab 2 Autoscaling

**Group 14**

**Name:**

**Philip Jasionowski    100751888**

# Video Report

Part 2: [youtube video](https://youtu.be/dY0BBUe2HFI)

Part 3: [youtube video](https://youtu.be/KyjbAbbovKs)

# Learning Objectives:

In this lab we learned how to create an autoscaling and a load balancing cluster in google cloud services. We used .YAML files to configure different architecture types

# Discussion:

### Summarize the problem, the solution, and the requirements for the pattern given in part 1. Which of these requirements can be achieved by the procedures shown in parts 2 and 3?

The problem in question is the fact that stakeholders want a way to route multiple services and instances using a single endpoint. For example a stakeholder wants to route multiple different traffic versions to the same endpoint (a Minecraft server that connects to a different version of Minecraft based on the client yet works through 1 ip address would be an example)

The solution is to place a gateway that routes the request to the right version and the appropriate instance. With this pattern the client only needs to know about and communicate with one endpoint. The requirements would be a gateway for the client to connect to and a balancer to automatically decide which service to route the client too.

The gateway and the load balancer are the parts we investigated and implemented in parts 2 and 3.

# Design:

I configured a GKE autoscaler using the following .yaml file:

Graphical user interface, application

Description automatically generated

Afterwards I deployed it with a load balancer and then pushed many requests to it to overwhelm the service. It then automatically scaled upwards as needed.

Below is a screenshot of the GKE load balancer working

Text

Description automatically generated

The difference between a load balancer and an autoscaler is that a load balancer balances requests to the existing resources. Whereas an autoscaler can create and delete resources as necessary.