Escalamiento Automático en GCP

Elaborada por: M. en C. Ukranio Coronilla

GCP tiene disponible el servicio de escalabilidad automática (autoscaling) el cual ajusta de manera automática la cantidad de instancias en función de la demanda. Esto significa que el número de instancias puede aumentar cuando hay más carga de trabajo o disminuir cuando hay menos, lo que optimiza el uso de recursos y ayuda a reducir costos.

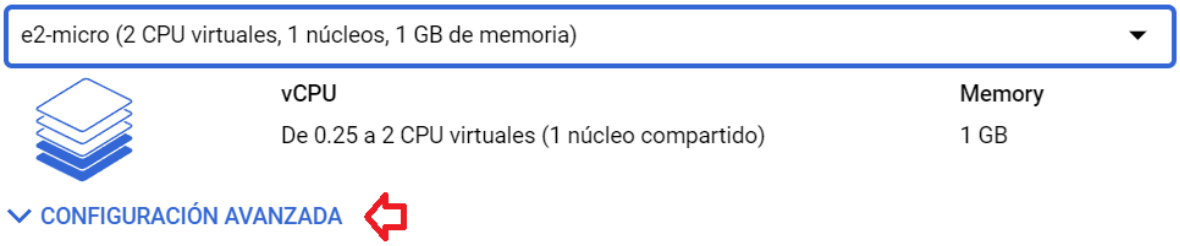
Para el escalamiento se utiliza alguno de los siguientes indicadores:

- Uso de CPU: Si el uso promedio de CPU en las instancias alcanza un cierto umbral, se inician automáticamente más instancias para manejar la carga adicional.
- Uso de memoria: Se agregan más instancias cuando el uso de memoria aumenta por encima de un límite.
- Número de solicitudes: Si las instancias están recibiendo más solicitudes de las que pueden manejar se agregan más instancias para procesar el tráfico.

El autoscaling no tiene un costo como característica, pero si se cobran los recursos adicionales que se generen de manera automática al escalar hacia arriba.

Vamos a probar el autoscaling utilizando como indicador el uso de CPU, para lo cual retomaremos el código del servidor HTTP con el endpoint **/cpu** del cual crearemos el archivo JAR correspondiente y lo subiremos a un bucket de nuestro proyecto. Después probaremos que funciona correctamente en una instancia y puede ser accedido desde nuestra LAP con curl. Mande una captura que incluye la terminal ssh de la instancia y la terminal de su LAP probando el endpoint **/cpu** (**Item 1**).

Posteriormente crearemos una **plantilla de instancias** de nombre **plantilla-escalamiento** que utilice un núcleo compartido de la serie E2 **e2-micro** y le vamos a dar click en CONFIGURACIÓN AVANZADA:

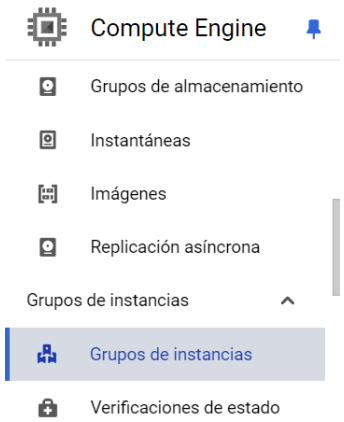


Para seleccionar:

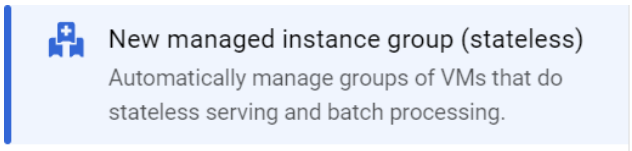
En la sección de Firewall permitimos el tráfico HTTP y en el menú de **Opciones avanzadas** -> **Gerenciamiento** dentro del campo **Automatización** coloque su script para ejecutar automáticamente su servidor y de click en Crear.

Para la creación del grupo de instancias administrado con la característica de **autoscaling** llevaremos a cabo el siguiente procedimiento:

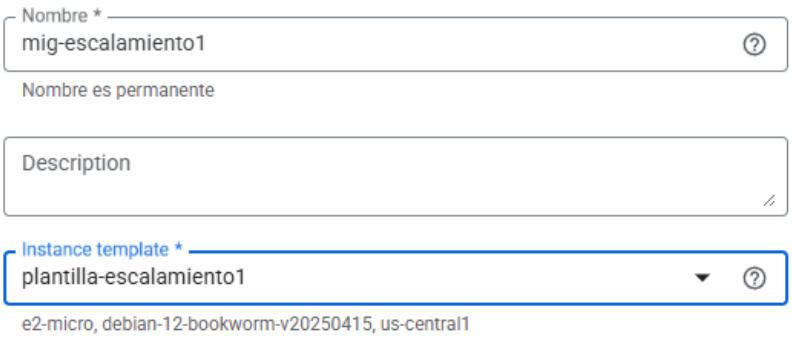
En el recurso Compute Engine de GCP damos click en Grupos de instancias:



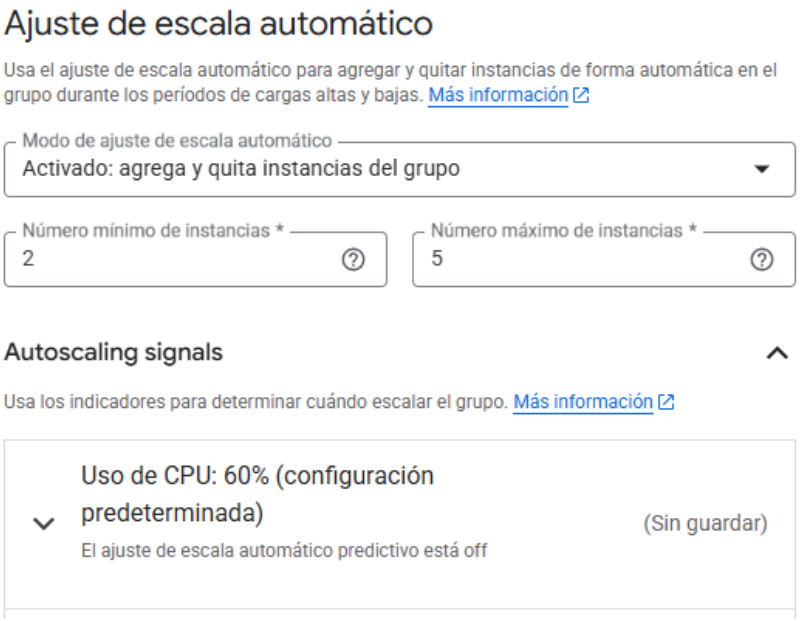
Y damos click en Crear grupo de instancias donde por default queda seleccionada la creación de un nuevo grupo de instancias administrado sin estado:



Le ponemos un nombre y seleccionamos la plantilla que hemos creado con anterioridad:



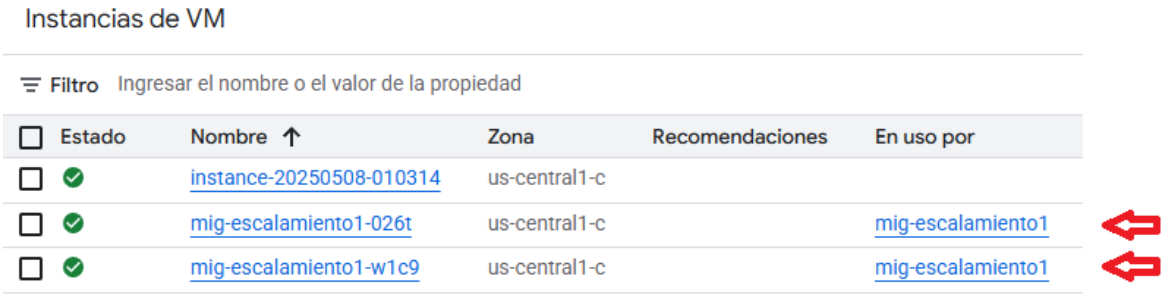
En **Ajuste de escala automático** verificamos que esté en modo **Activado: agrega y quita instancias del grupo**, en el **Número mínimo de instancias** vamos a dejar dos y en máximo vamos a poner cinco como límite. En **Autoscaling signals** dejamos la configuración predeterminada de 60% con lo cual al usarse más de 60% de CPU provocará un escalamiento hacia arriba o al decrecer por debajo de 60% un escalamiento hacia abajo, quedando como sigue:



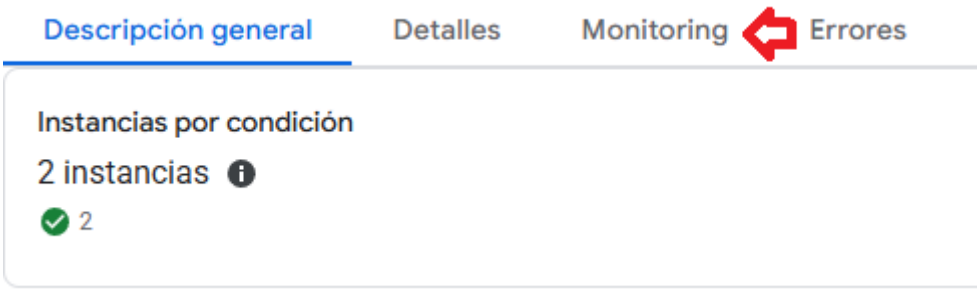
En **Periodo de inicialización** va el número de segundos que el escalador automático debería esperar después de que una máquina virtual ha iniciado, antes de que el escalador comience a obtener información nueva del uso de CPU. En nuestro caso ponemos dos minutos (120 segundos) para evitar hacer un monitoreo en el momento que se levanten automáticamente una o varias instancias, pues podría estarse consumiendo mucho CPU al estarse ejecutando la sucesión de comandos de la plantilla, provocando que se levanten más instancias de manera recursiva.

Al final damos click en  en la parte inferior izquierda de la ventana, con lo que se crea nuestro grupo de instancias.

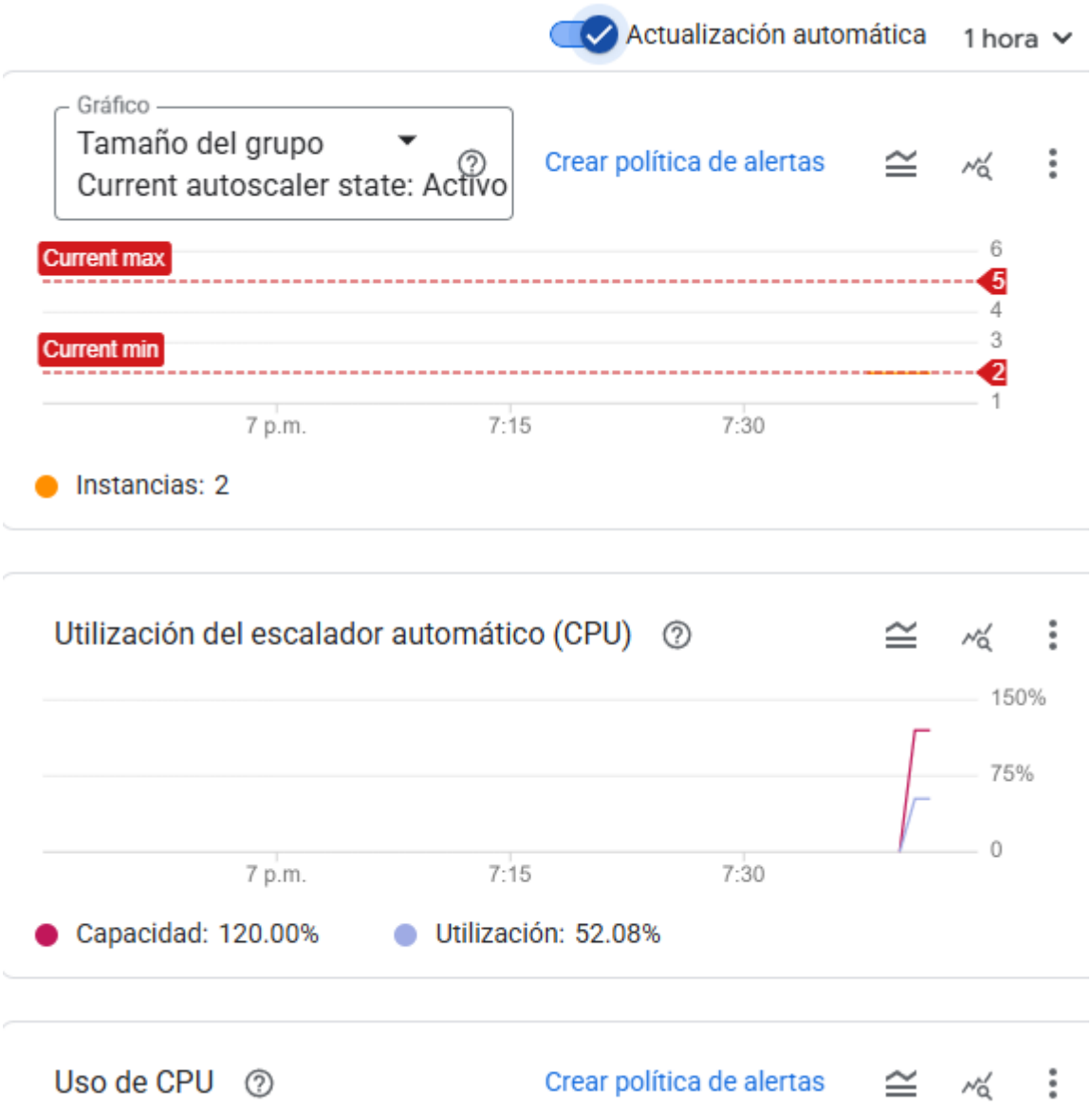
Si esperamos un minuto podremos ver en **Compute Engine->Instancias de VM** las dos instancias creadas, las cuales forman parte del grupo de instancias administrado que acabamos de crear:



Ahora damos click en el nombre del grupo donde señala la flecha en rojo y obtendremos una pantalla con la descripción general del grupo de instancias administrado, en la cual le vamos a dar click a la pestaña **Monitoring** :



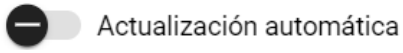
Donde se muestran los gráficos de uso de diversos recursos (se recomienda activar el botón de **Actualización automática** y esperar unos minutos para ver los cambios):



Probaremos como se realiza el monitoreo accediendo al endpoint **/cpu** desde nuestra computadora con la herramienta curl solicitando un uso de cpu del 10% (para no activar el autoscaling pues tenga en cuenta que el sistema operativo también ejecuta otros procesos haciendo uso de CPU también) durante 30 segundos y al mismo tiempo en segundo plano ejecutamos el comando date para mantener un registro de la hora en que iniciamos la prueba. Observe los cambios que se dan en el gráfico **Uso de CPU** el cual debería reflejarse de inmediato, sin embargo, hay tres inconvenientes que podrá constatar:

- El uso de CPU no está desglosado para cada instancia del grupo, sino que grafica la utilización promedio.

- La información presentada en el gráfico tiene un retraso de tiempo de algunos minutos (verifique cuantos son).
- No se refresca la información en el gráfico de manera continua, lo cual podemos “forzar” activando y desactivando el botón



Mande llamar al endpoint **/cpu** desde más terminales para obtener el escalamiento automático a 3, 4 y 5 instancias y también que se observe el escalamiento hacia abajo. Agregue la captura de pantalla del gráfico de instancias correspondiente (**Item 2**).

Como ejemplo el siguiente gráfico muestra sólo el escalamiento hacia arriba a 3 y luego a 5 instancias: