ENTREGA 5 - DESPLIEGUE EN PAAS MIGRACIÓN DE UNA APLICACIÓN WEB A UN PLATAFORMA COMO SERVICIO EN LA NUBE PÚBLICA

Danny Pineda Echeverri, Felipe Alejandro Paez Guerrero, Vihlai Maldonado Cuevas
ISIS4426 - Desarrollo de soluciones cloud
Universidad de los Andes, Bogotá, Colombia
{d.pinedae, f.paezg, v.maldonado1}@uniandes.edu.co
Fecha de presentación: mayo 28 de 2023

1. Arquitectura de la aplicación

A continuación, se muestra el diagrama de arquitectura:

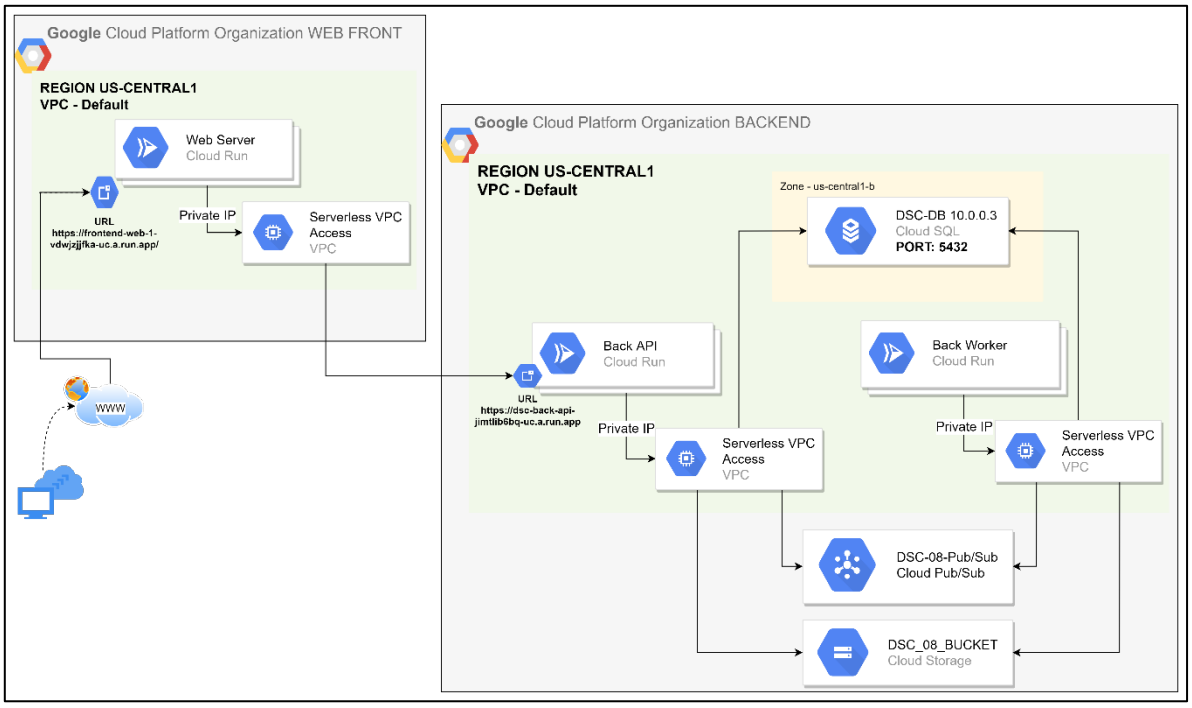

Figura 1. Arquitectura Conectividad GCP-Escalabilidad Web.

Se realizo la configuración del servicio de Nube de GCP Cloud Run para habilitar el despliegue serveless del antes servidor web y del worker como un PaaS. La aplicación web se desarrolló de manera separada entre Backend y el FrontEnd, aprovechando esto y debido a una restricción de presupuesto de la cuenta donde veníamos trabajando, se desplegó el Front en una subscripción diferente a la del Back. Para habilitar la comunicación de los Cloud Run con otros servicios de la VPC, se habilito un conector desde la funcionalidad Serveless VPC Access,

2. Pruebas de estrés

2.1. Escenario 1

El primer escenario considerado busca identificar la máxima cantidad de requests HTTP por minuto que soporta la aplicación web. Para ello se considera:

Pocos usuarios enviando archivos pequeños:

De 1 a 100 usuarios concurrentes enviando archivos pequeños (alrededor de 15 MB).

Tipo de gráfica: Gráfico de líneas, con cantidad de usuarios concurrentes en el eje x y las métricas en el eje y.

Se ejecutarán las pruebas desde una máquina de AWS Academy, con el fin de simular el acceso de los usuarios desde diferentes ubicaciones.

Restricciones:

Tiempo de respuesta máximo: 30 segundos.

Máxima tasa de error: 1%

Cantidad máxima de usuarios concurrentes: 100

Asimismo, cabe resaltar que durante las pruebas se registraron las métricas y tiempos de procesamiento de los archivos que se alcanzaron a procesar previo a la interrupción del servicio.

A continuación, se muestran los resultados de las pruebas de estrés:

Conversión ZIP

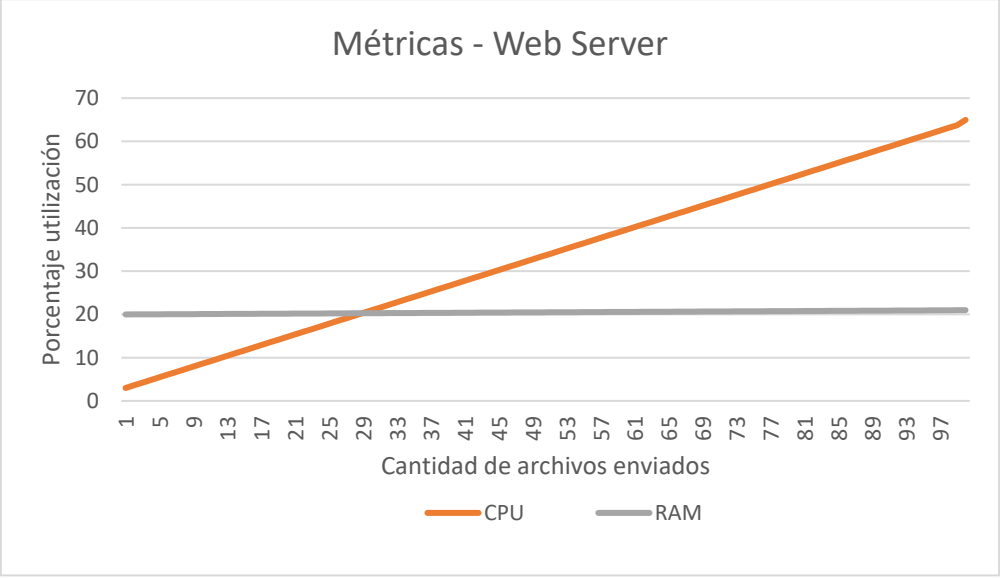

Figura 2. Métricas CPU, RAM Web Server conversión ZIP.

La Figura 2 muestra el consumo de CPU del contenedor dedicado para el Web Server. Se observa que pudo manejar los 100 usuarios concurrentes y el consumo de CPU aumentó en más del 60%, sin embargo, el consumo de memoria se mantuvo constante.

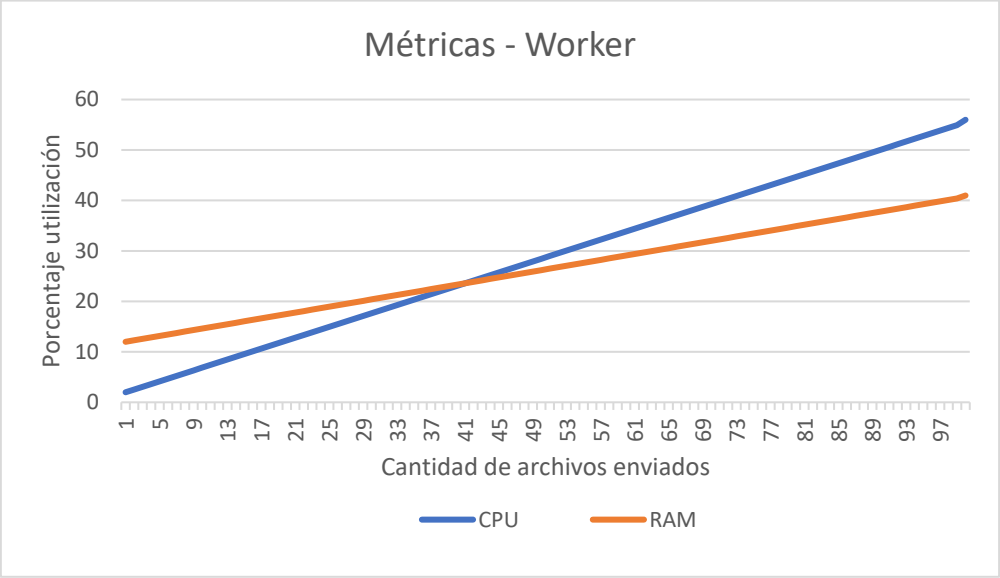

Figura 3. Métricas CPU, RAM Worker conversión ZIP.

La Figura 3 muestra el consumo de CPU, memoria del contenedor dedicado para el Worker. Se observa que hubo un incremento en la utilización de la CPU, en más del 50% y de memoria en más del 40%.

En resumen, en la prueba de compresión tipo .zip se pudieron procesar los 100 archivos. Asimismo, no hubo interrupción en el servicio.

Conversión Tar.bz2

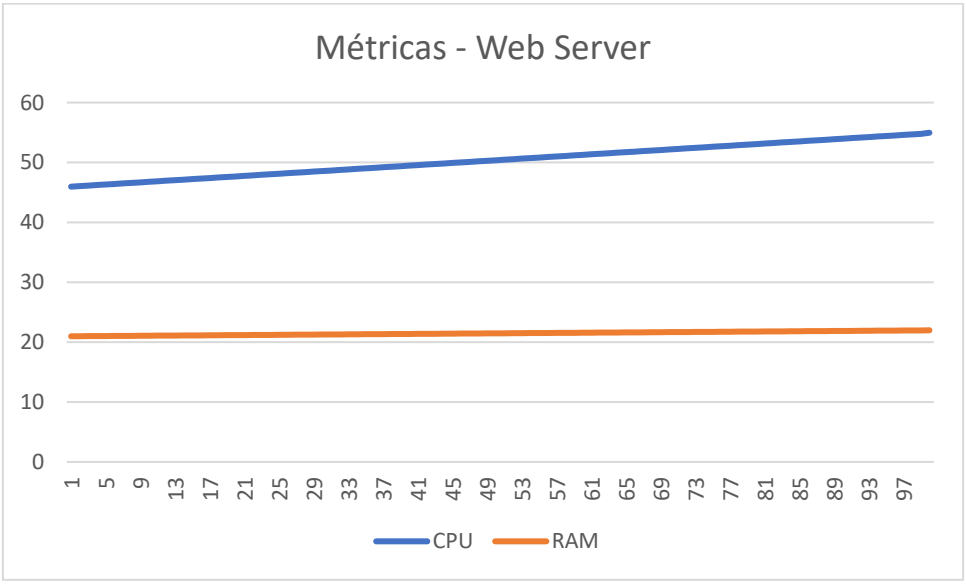

Figura 4. Métricas CPU, RAM y Disco Web Server conversión TAR Bz2.

La Figura 4 muestra el consumo de CPU, memoria del Web Server. Se observa que el consumo de CPU se incrementó hasta llegar cercano a 55%. El consumo de memoria se mantuvo constante.

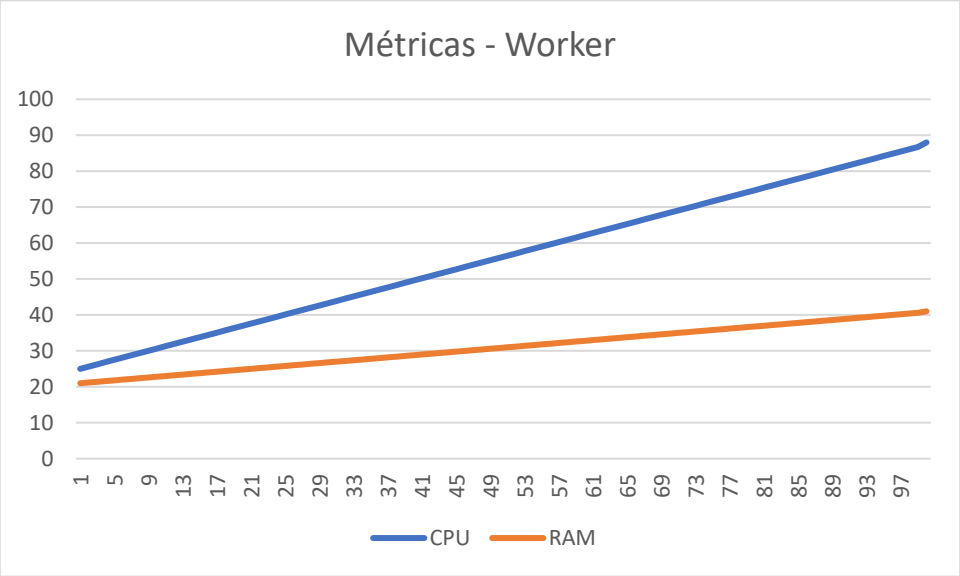

Figura 5. Métricas CPU, RAM Worker conversión TAR Bz2.

La Figura 5 muestra el consumo de CPU, memoria del contenedor del Worker. Se observa que el consumo de CPU aumentó cercano al 90%, mientras el consumo de memoria se mantuvo cercano a 40%.

En resumen, para la prueba con compresión tipo tar.gz2 se pudieron procesar los 100 archivos y no hubo interrupción del servicio.

1.2 **Escenario 2**

El segundo escenario busca identificar la máxima cantidad de archivos que pueden ser procesados por minuto en la aplicación local.

Restricciones:

Tamaño mínimo de archivos: 10MB.

Espera máxima: 30 segundos.

Capacidad de procesamiento mínima: carga de 90 archivos por minuto.

Mínima tasa de transferencia de datos: 75MB por segundo de subida.

Resultados:

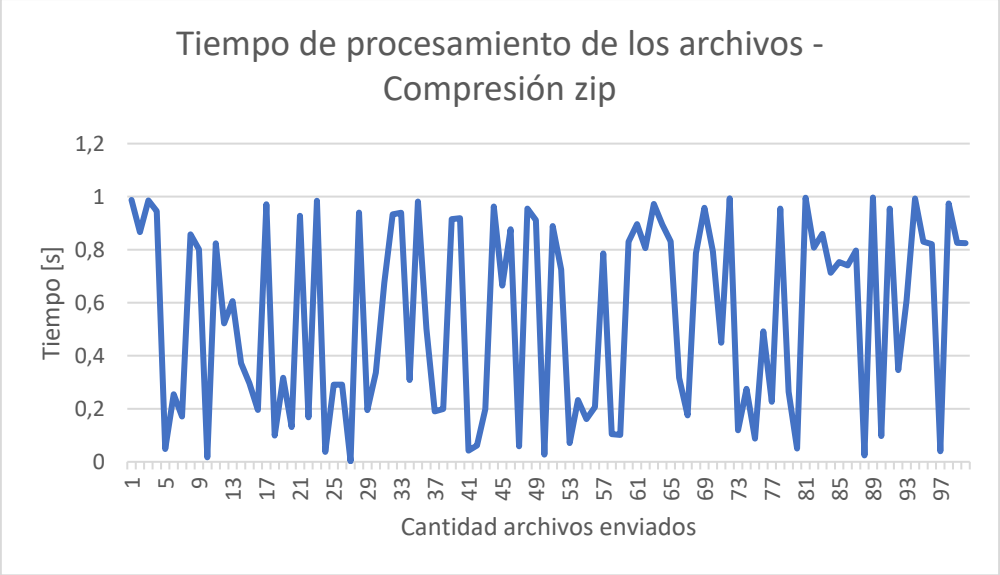

Figura 6. Tiempo Procesamiento ZIP.

En la Figura 6 se grafica el tiempo de procesamiento de los archivos con compresión zip. El tiempo medio es de 0.552 segundos. El tiempo máximo de procesamiento es de 0.997 segundos.

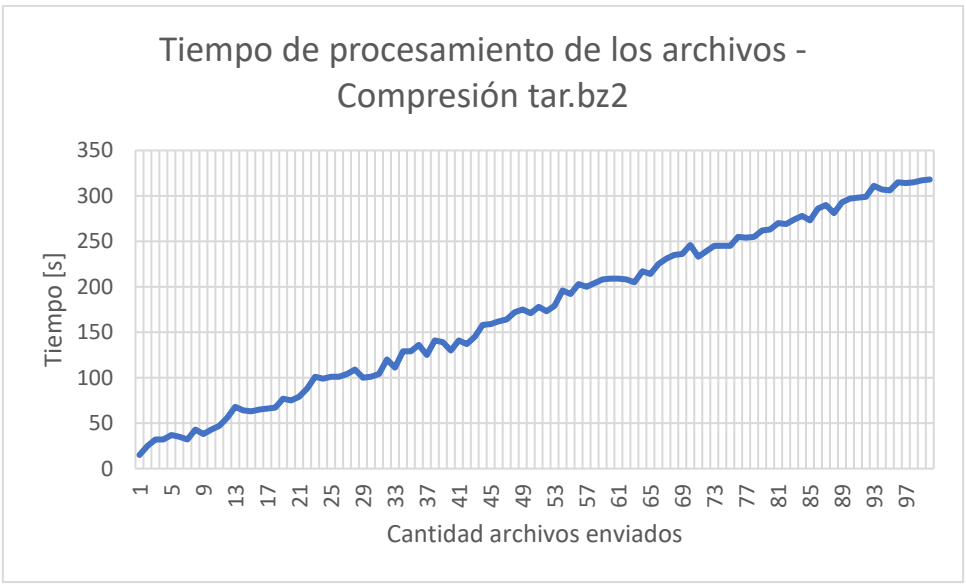

Figura 7.Tiempo Procesamiento TAR Bz2.

En la Figura 7 se observa el tiempo de procesamiento de los 100 archivos con compresión tar.bz2. El tiempo medio es de 173,86 segundos, 2.89 minutos. El tiempo máximo de procesamiento es de 318 segundos (5.3 minutos).

En general se observa que el tiempo de procesamiento para el tipo de compresión .tar.bz2 es más alto que el .zip, igualmente se alcanzaron a procesar los 100 archivos y se observa un comportamiento creciente con correlación positiva entre la cantidad de archivos procesados y el tiempo de procesamiento.

2. CONCLUSIONES

- Se implementó la aplicación en Cloud Run, realizando el despliegue de tres implementaciones serverless, para el servidor backend, frontend, y el worker, lo que permite una administración simplificada, ya que cloud Run se encarga de la infraestructura, aprovisionamiento, la configuración y la escalabilidad.
- La integración de Pub/Sub en la aplicación proporciona una capacidad de mensajería robusta y confiable. Permite la comunicación asíncrona y distribuida entre los componentes de la aplicación, lo que facilita la construcción de sistemas altamente escalables y desacoplados. Los mensajes se envían y reciben a través de los temas y suscripciones de Pub/Sub, lo que garantiza una entrega confiable.