Servicio Orders - Testing de escalabilidad

Propósito:

El objetivo de este testing es probar la capacidad de escalado horizontal del servicio Orders a la hora de recibir una cantidad de peticiones, para asi saber si el mismo puede manejar cargas aumentadas de información.

Contenido:

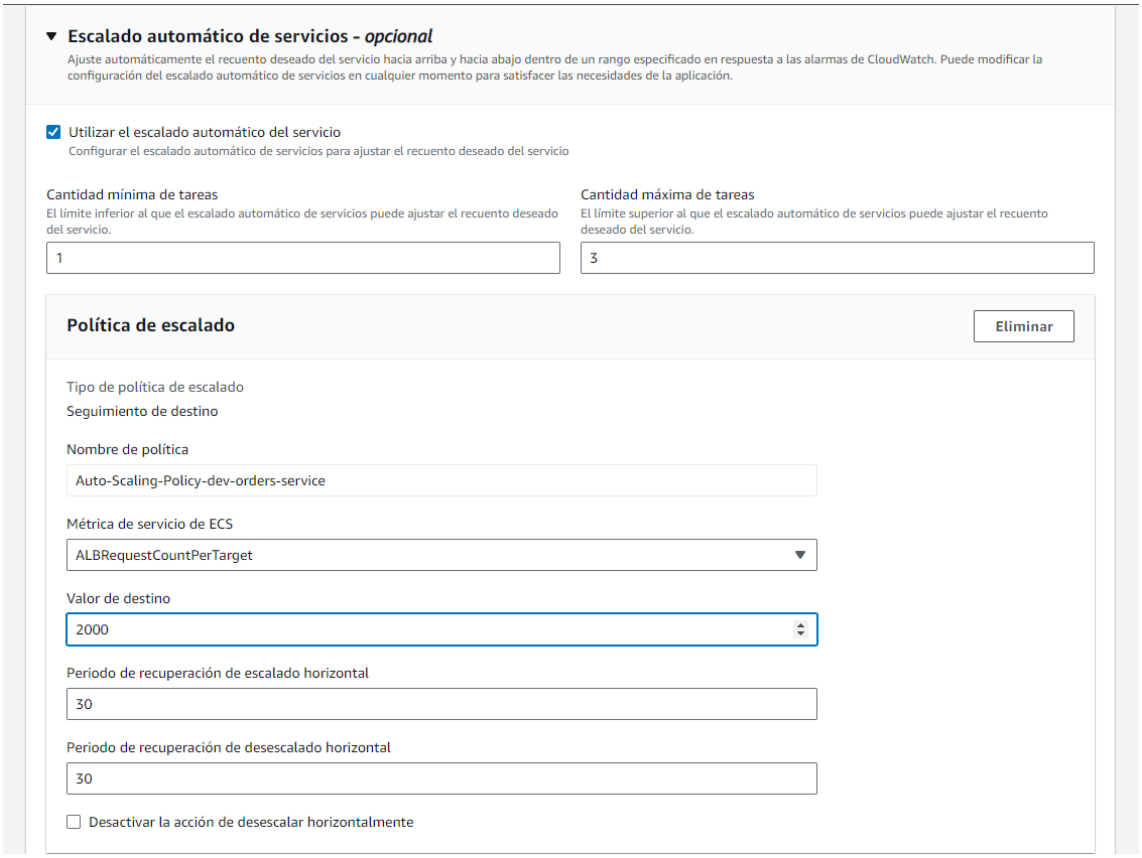
- Objetivos de la prueba: El propósito de la prueba de escalabilidad es validar que el servicio sea capaz de escalar horizontalmente al recibir una mayor cantidad de peticiones de lo normal.
- Resultado Clave: El servicio es capaz de escalar horizontalmente acorde a la demanda de peticiones por los usuarios según la configuración aplicada, sin verse afectado el servicio. Posteriormente a la caída de la demanda, el servicio des escala correctamente.

Entorno de Pruebas

Se realiza las pruebas en instancias con CPU de 256 unidades y 512 MegaBytes de memoria RAM.

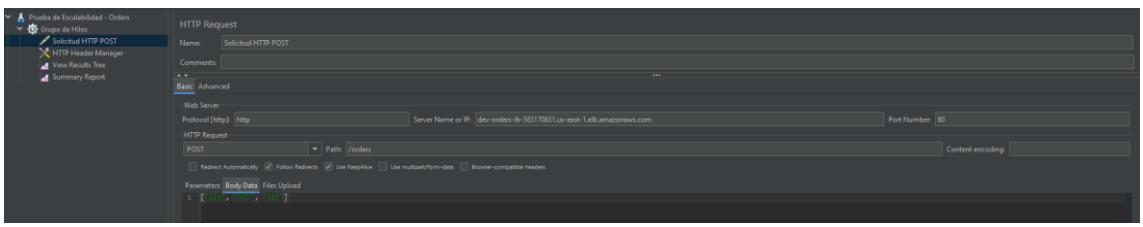
Contenido:

- Descripción del Entorno: Los cluster donde trabajan las instancias son de tipo “FRAGATE”, las configuraciones de red están completamente abiertas tanto la entrada como salida. Cuenta con un balanceador de carga.
- La configuración del auto escalado sea realiza teniendo 1 tarea como mínimo y un máximo de 3 tareas escaladas. La política de escalado es de tipo seguimiento de destino, con una métrica de peticiones con un valor de 2000 peticiones y perdidos de recuperación de escalado y desescalada horizontal de 30 segundos.
- Se utiliza la herramienta JMeter (V 5.6.3) para realizar las pruebas de carga.



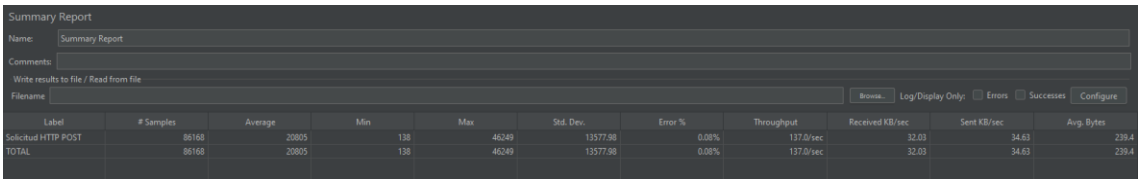
Plan de Pruebas

- Escenario de prueba: 3000 usuarios simultáneos en un tiempo de Ramp up de 30 segundos durante 10 minutos.

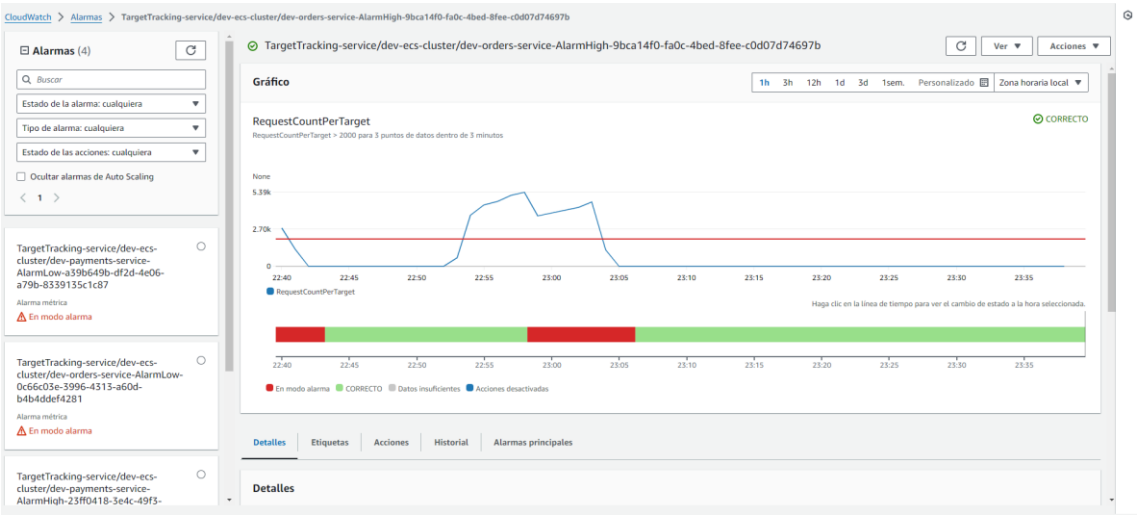


Resultado de la prueba

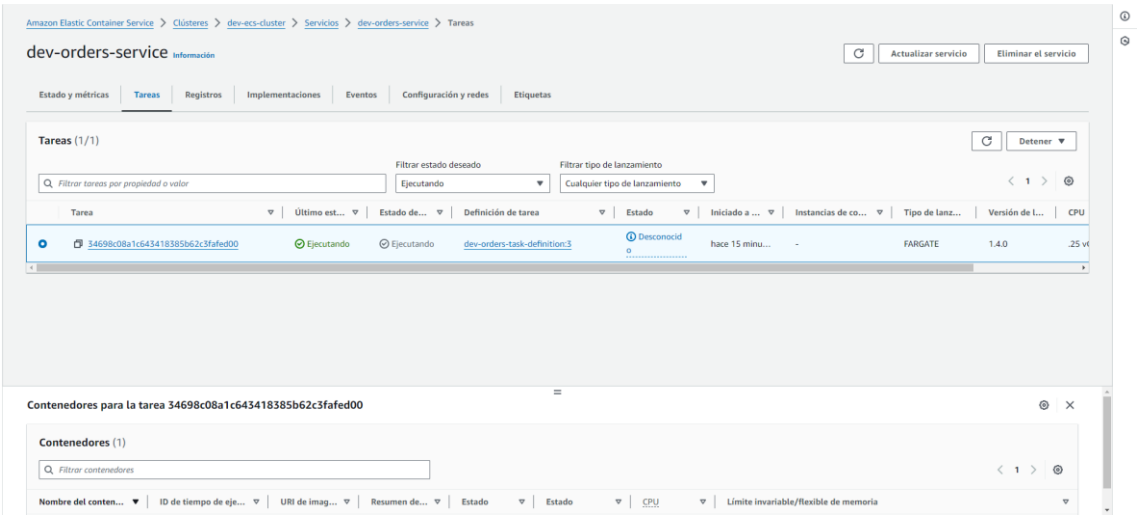
- Se observa en el reporte de Jmeter el tiempo transcurrido de 10 minutos y un envio total de 86 mil peticiones aproximadamente con un índice de éxito del 100%.

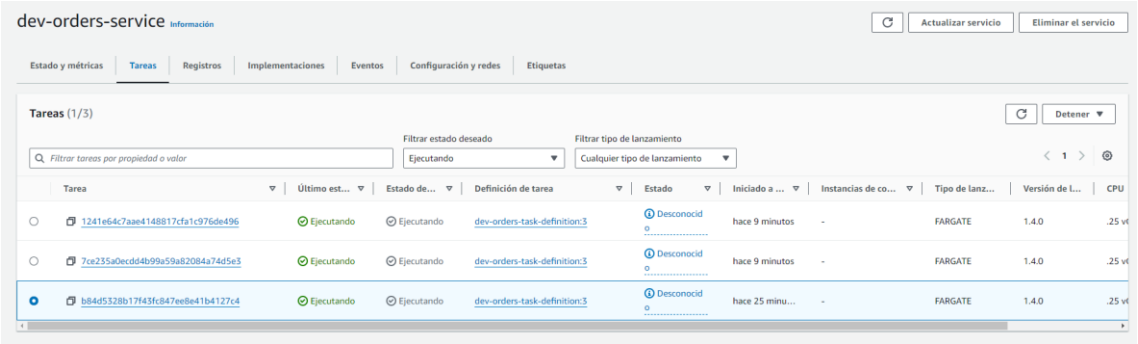


- En cloud watch se observan picos de 5.39 mil peticiones anteriores al escalado.

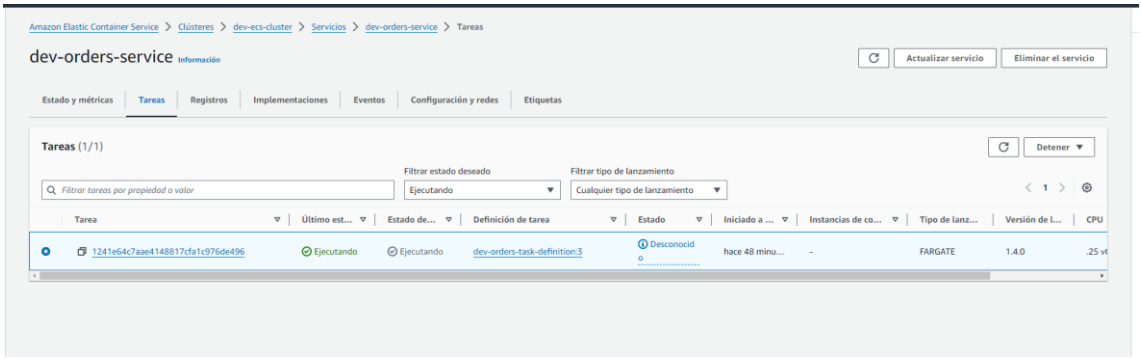


- Se puede observar como de una instancia se pasa a tres instancias en ejecución paralela, indicando el correcto escalamiento según la demanda que se tenía y la configuración.





Desescalado posterior a las pruebas.



Análisis y conclusión de los resultados.

Según se muestra en el análisis proporcionado el servicio se comporto acorde a lo esperado y realizo el debido escalamiento de este, logrando asi picos de 5.39 mil peticiones a la vez, posterior al escalado de los servicios y aun asi mantener un 100% de éxito en estas y sin degradación en el servicio proporcionado.