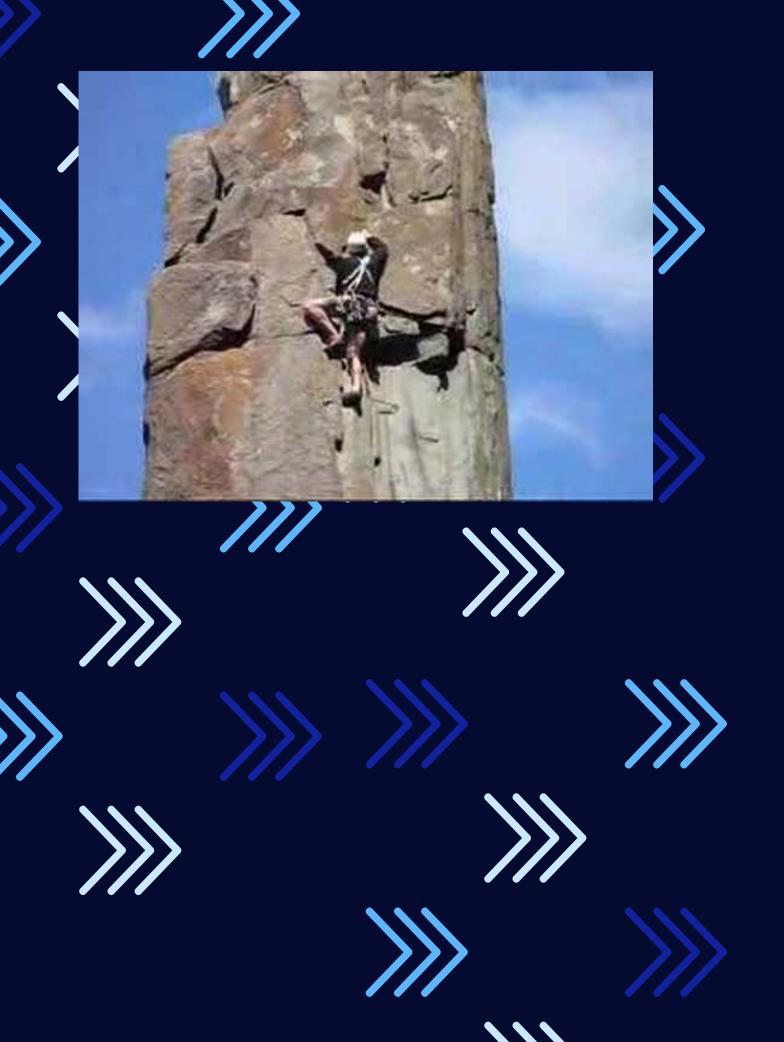


ESCALADO EN, MONOLÍTICO

Eddie Girón Julio Ruíz



¿CÓMO ESCALAMOS UN MONOLITO?

Puntos a mejorar para escalar una aplicación monolítica:

- Mejora de capacidad de carga
- Optimización general de código
- Recursos adicionales
- Implementación de técnicas de escalabilidad
 - Refactorización
- Migración a otra arquitectura

Escalar una arquitectura monolítica puede ser más complicado y caro que otras arquitecturas.

¿PUEDE ESCALAR VERTICALMENTE?

El escalado vertical es una forma directa de manejar un aumento en la carga; sin embargo, tiene límites físicos y financieros. Eventualmente, agregar más recursos se vuelve costoso o técnicamente inviable.

- Aumento de recursos de hardware
- Utilización de Caché

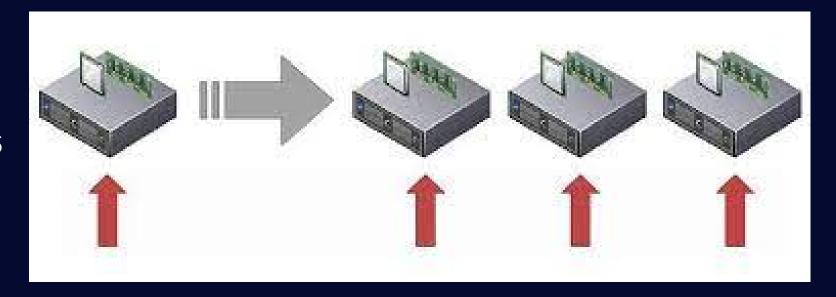


¿PUEDE ESCALAR HORIZONTALMENTE?

Aumento en la capacidad de una aplicación distribuyendo la carga de trabajo entre múltiples instancias de la misma.

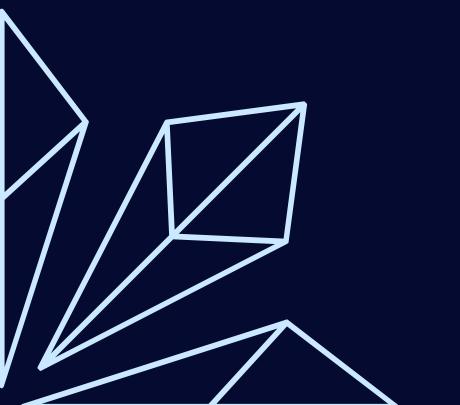
Escalamiento en aplicación de log-in:

- Persistencia de datos
- Automatización de despliegues

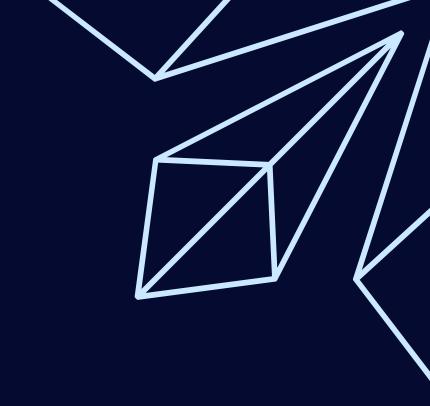


¿ARQUITECTURA MONOLÍTICA CON ORLEANS?

Puede hacerse, sin embargo no se aprovecharía el potencial completo de Orleans, aunque si se hace con fines de escalabilidad futura puede ser conveniente.



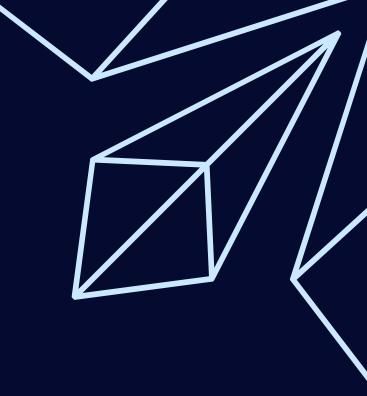
Git: https://github.com/Edd1enator/SimpleLog_In.git



HIPOTESIS

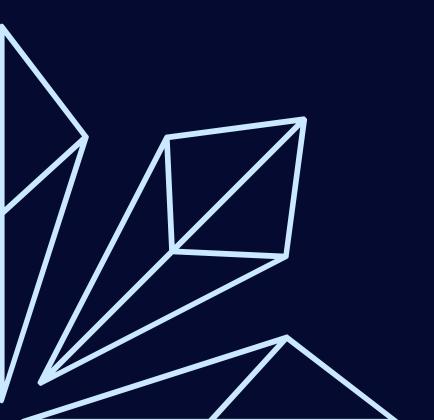
Hipótesis de Rendimiento: La aplicación maneja un número específico de usuarios concurrentes (por ejemplo, 1400 usuarios simultáneos) sin degradar significativamente el rendimiento (por ejemplo, el tiempo de respuesta promedio permanecerá por debajo de 2 segundos).

RECURSOS

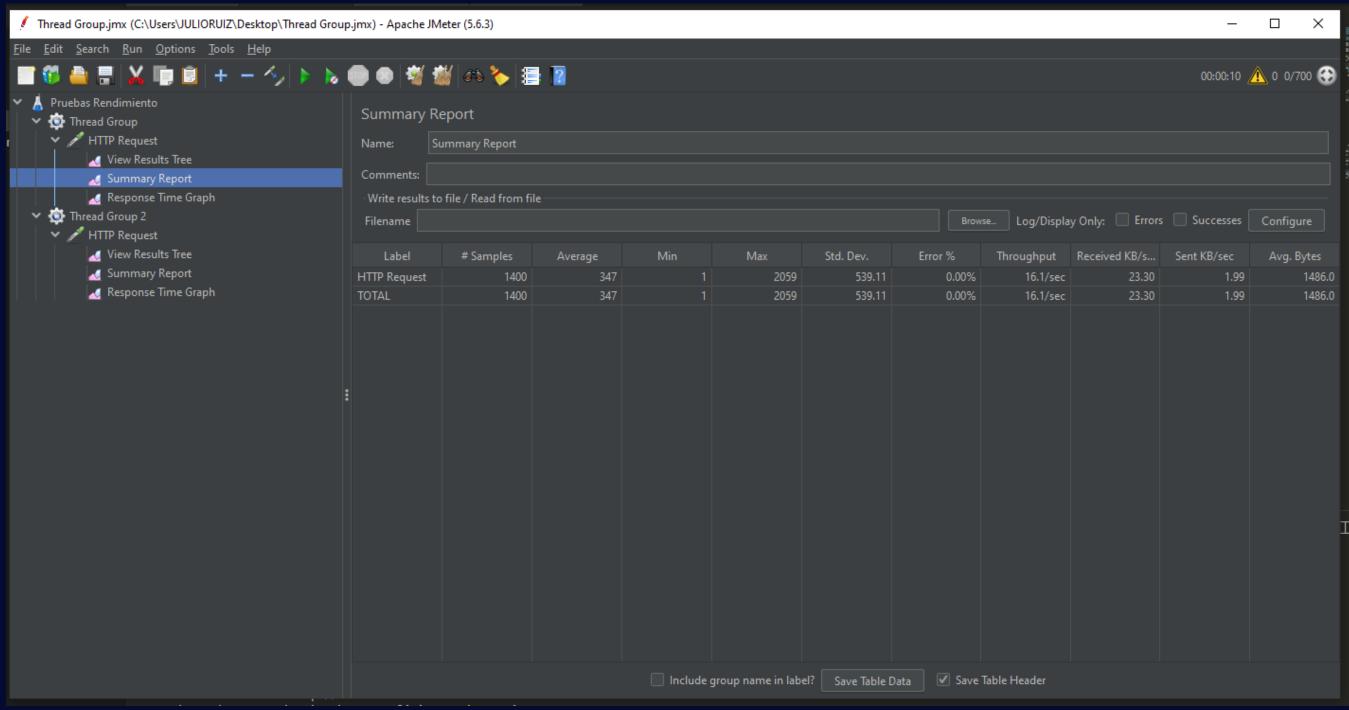


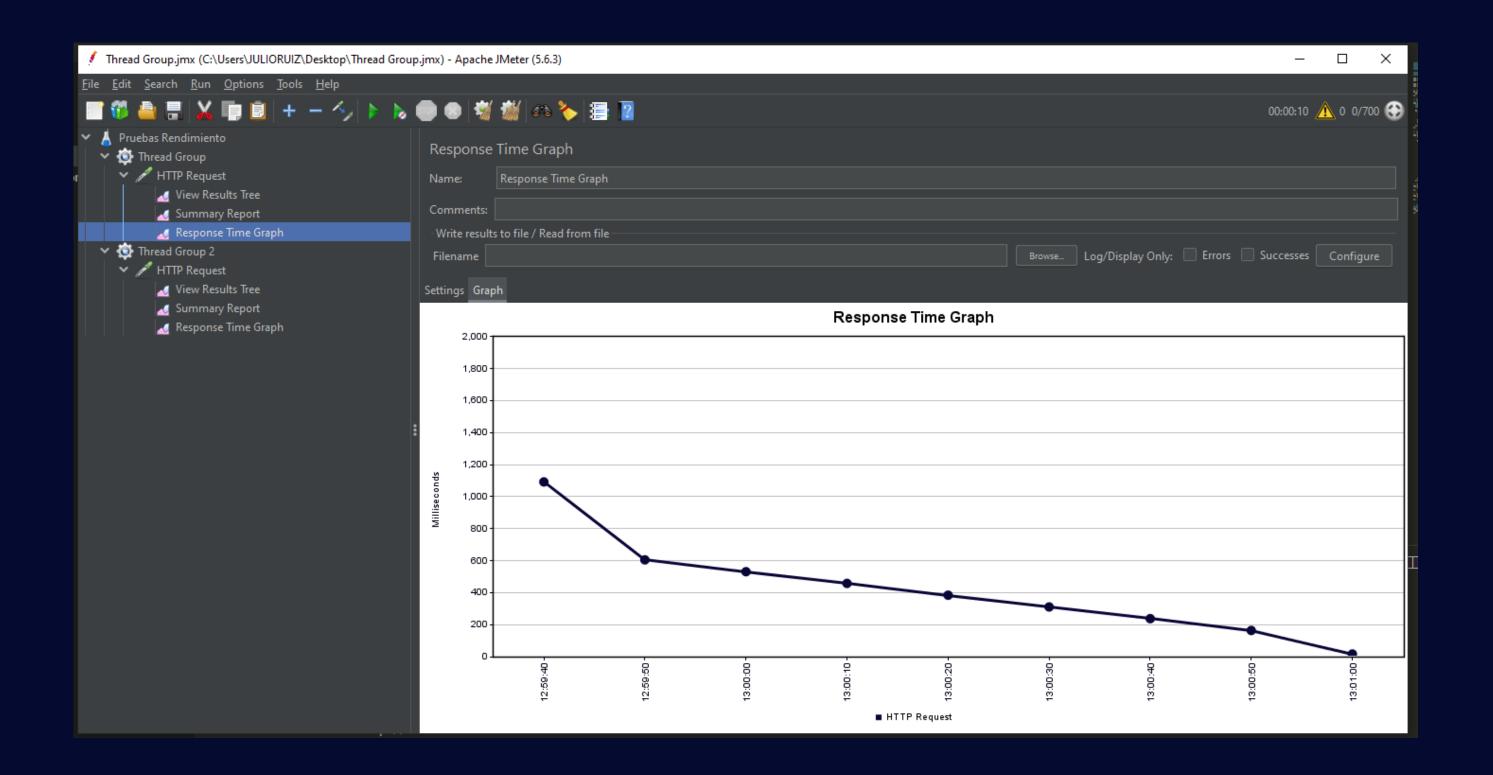
2 CPU

4 RAM

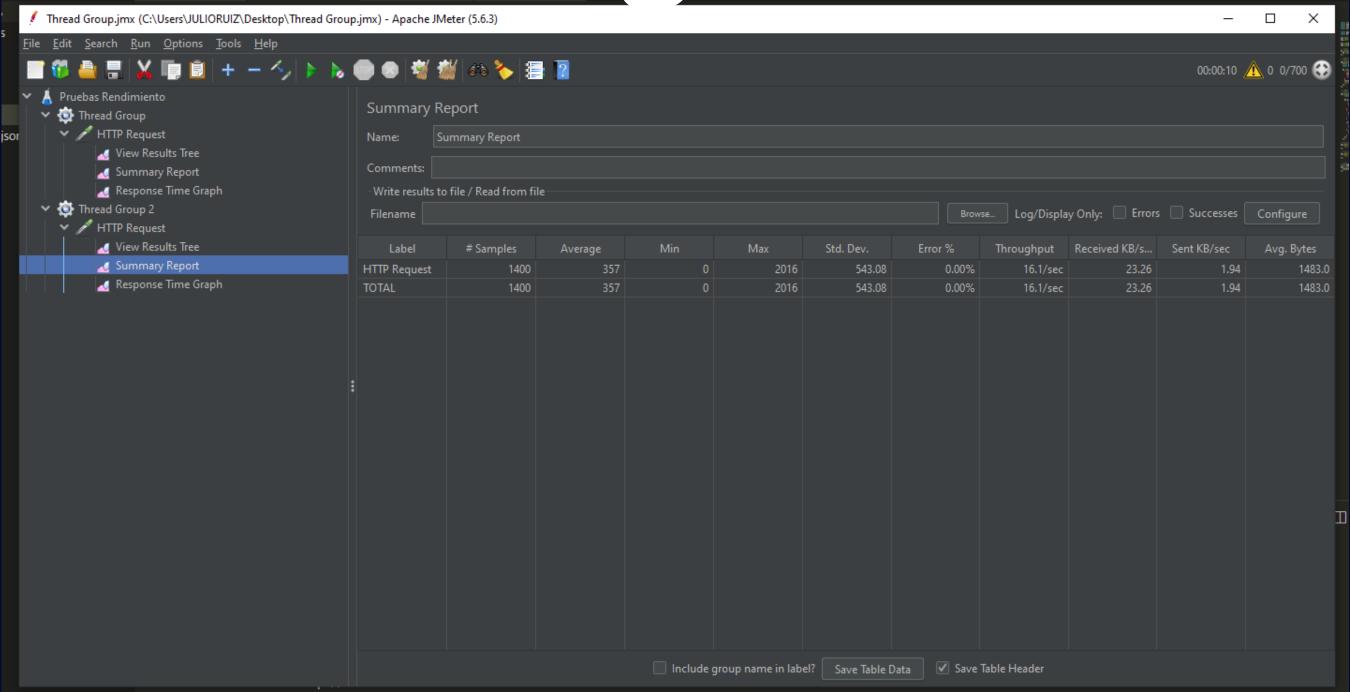


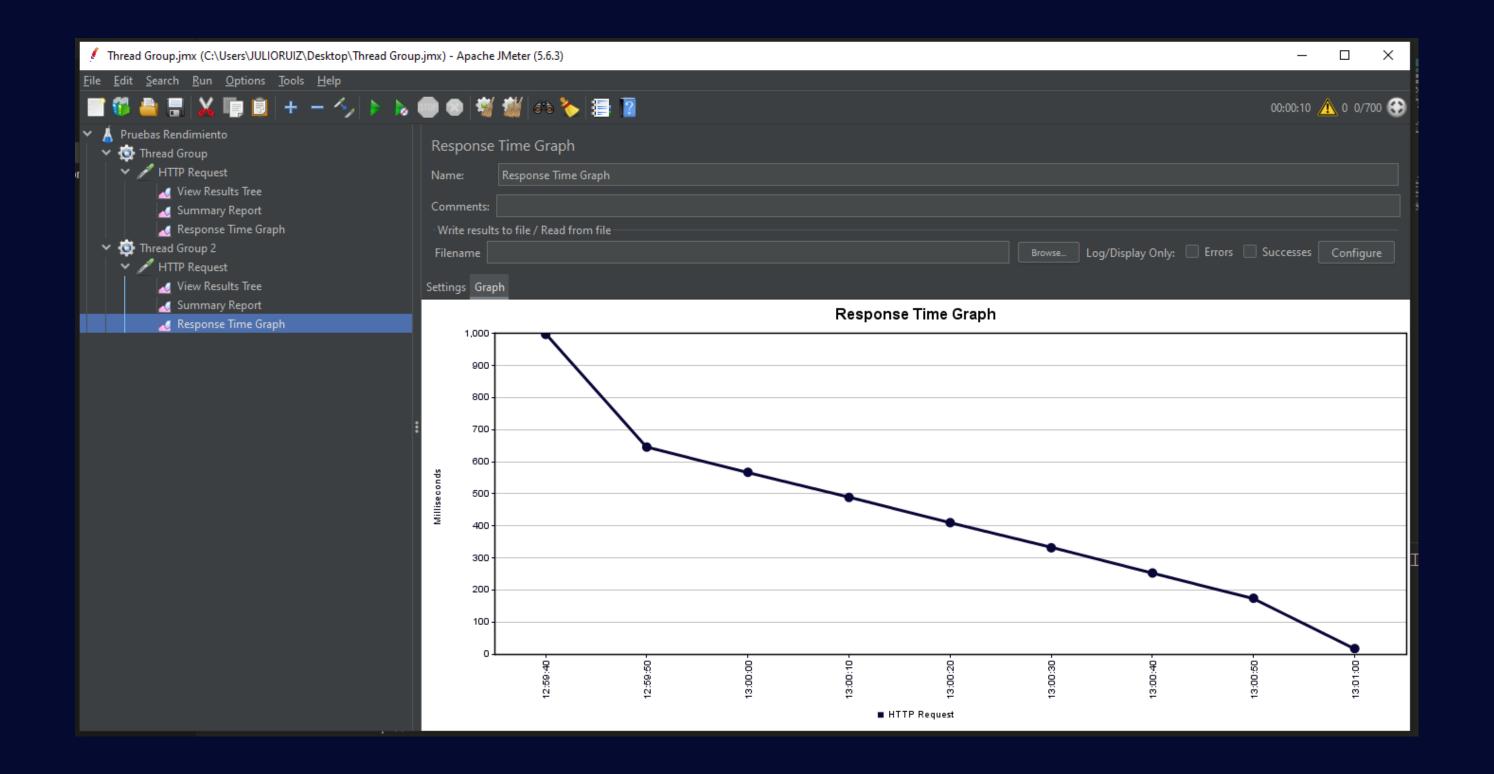
/register



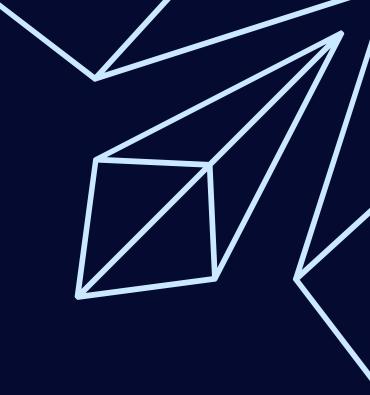


/login





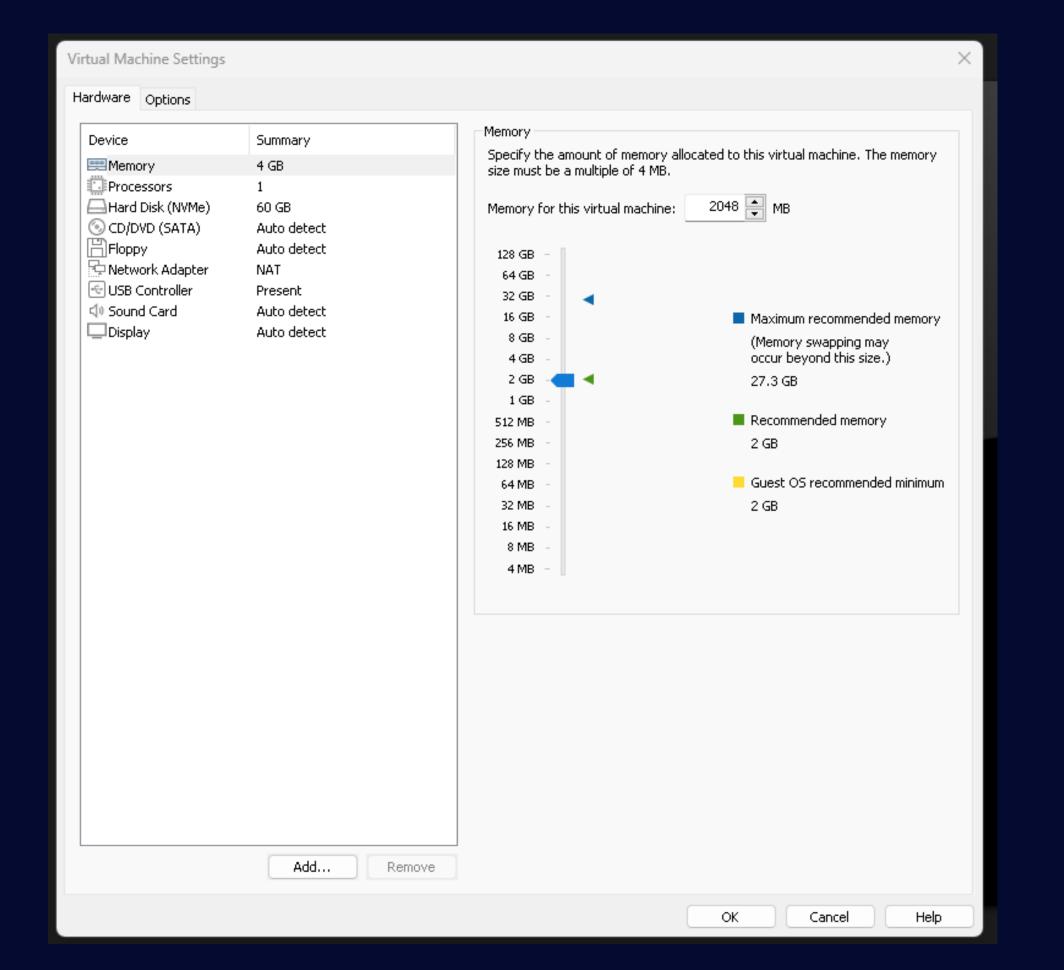
RECURSOS



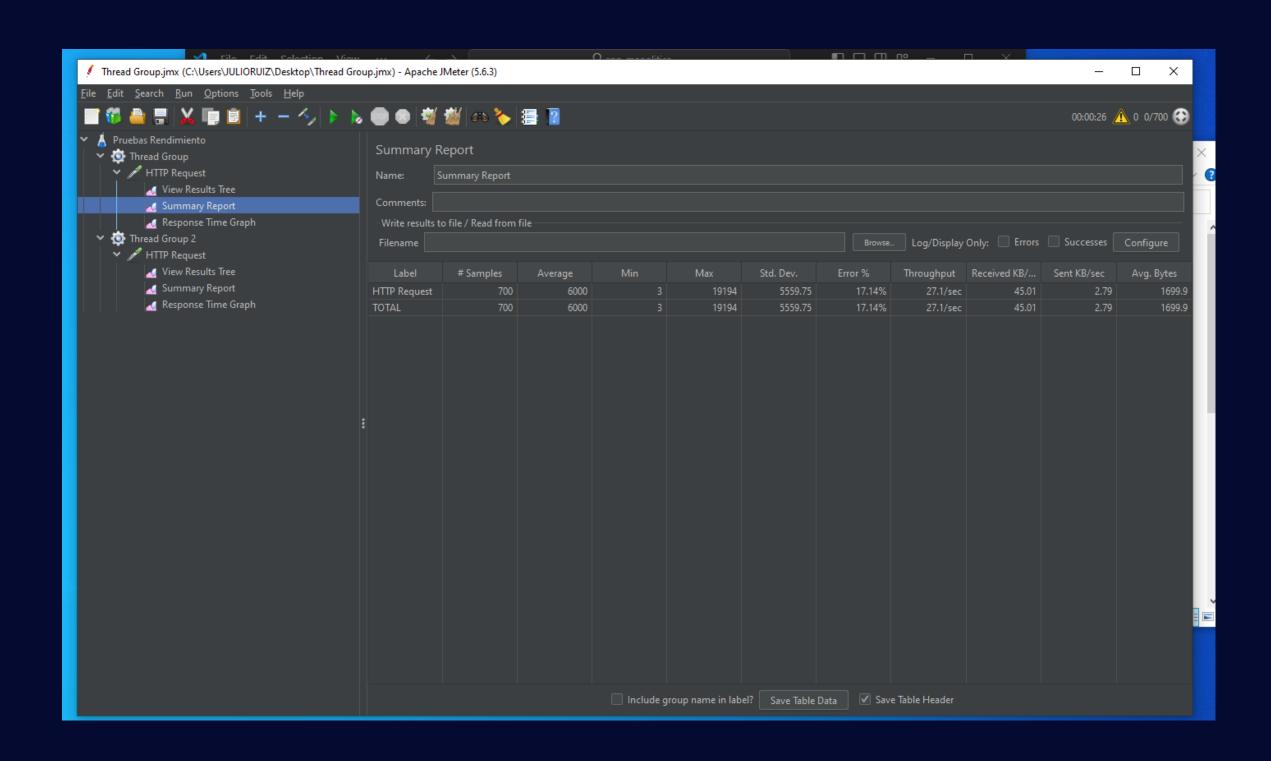
1 CPU

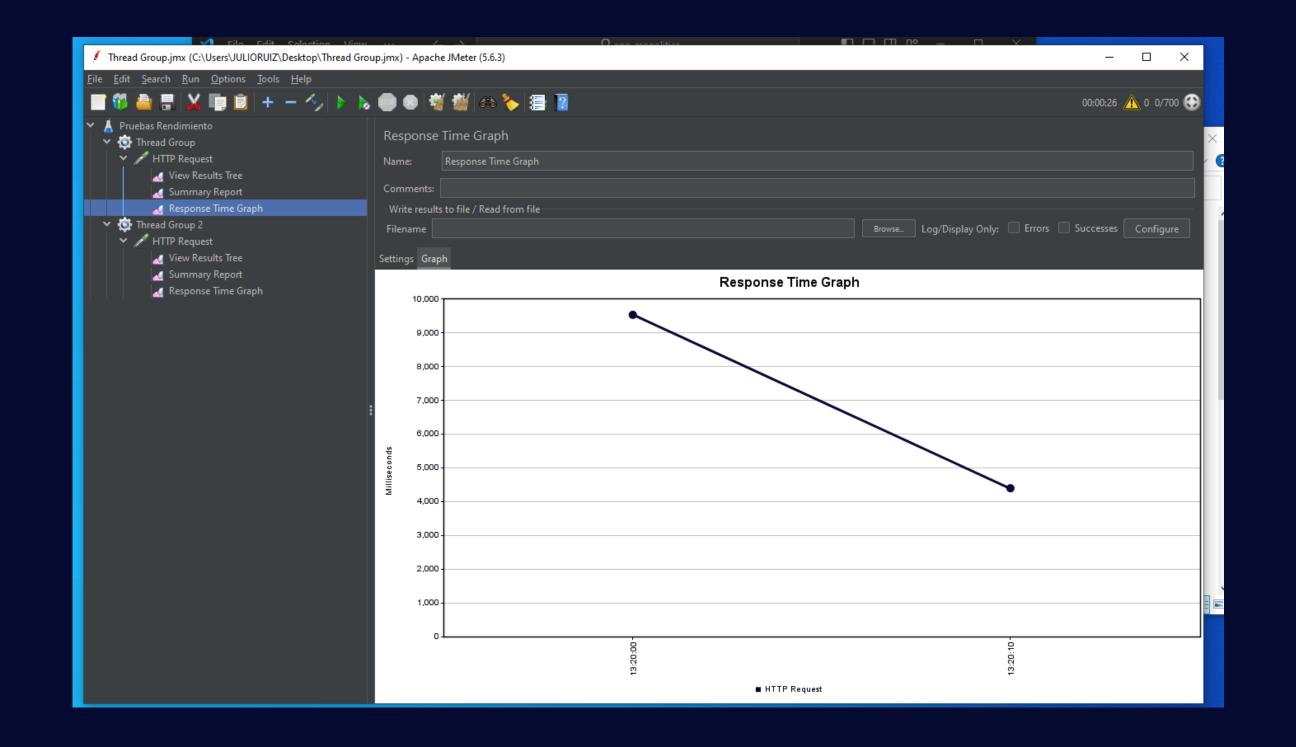
2 RAM



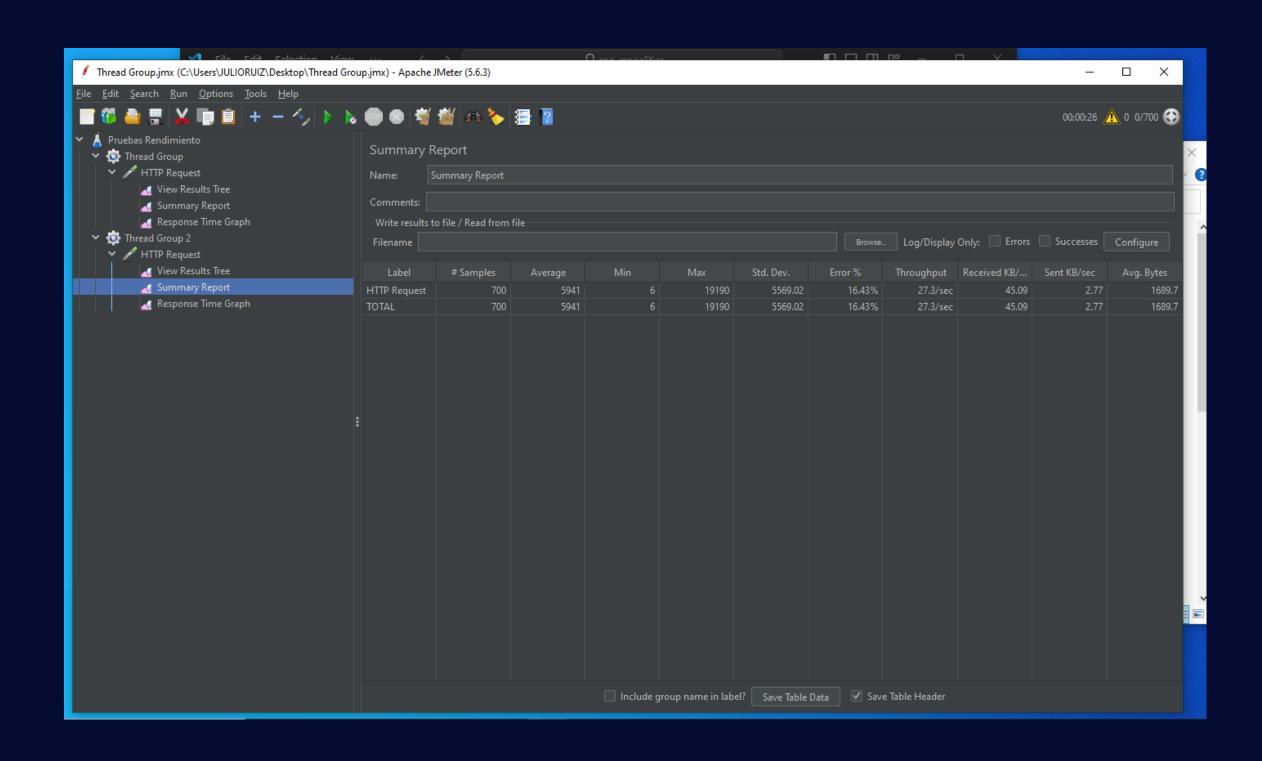


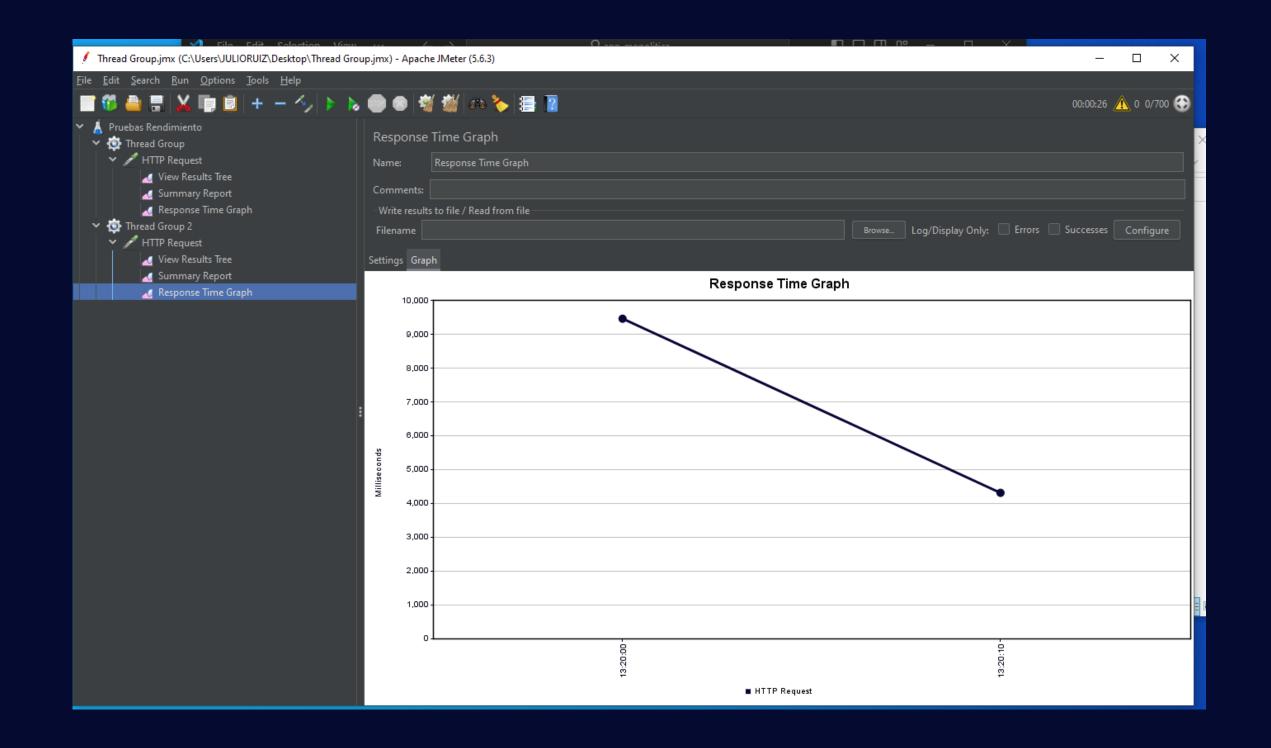
/register primera prueba



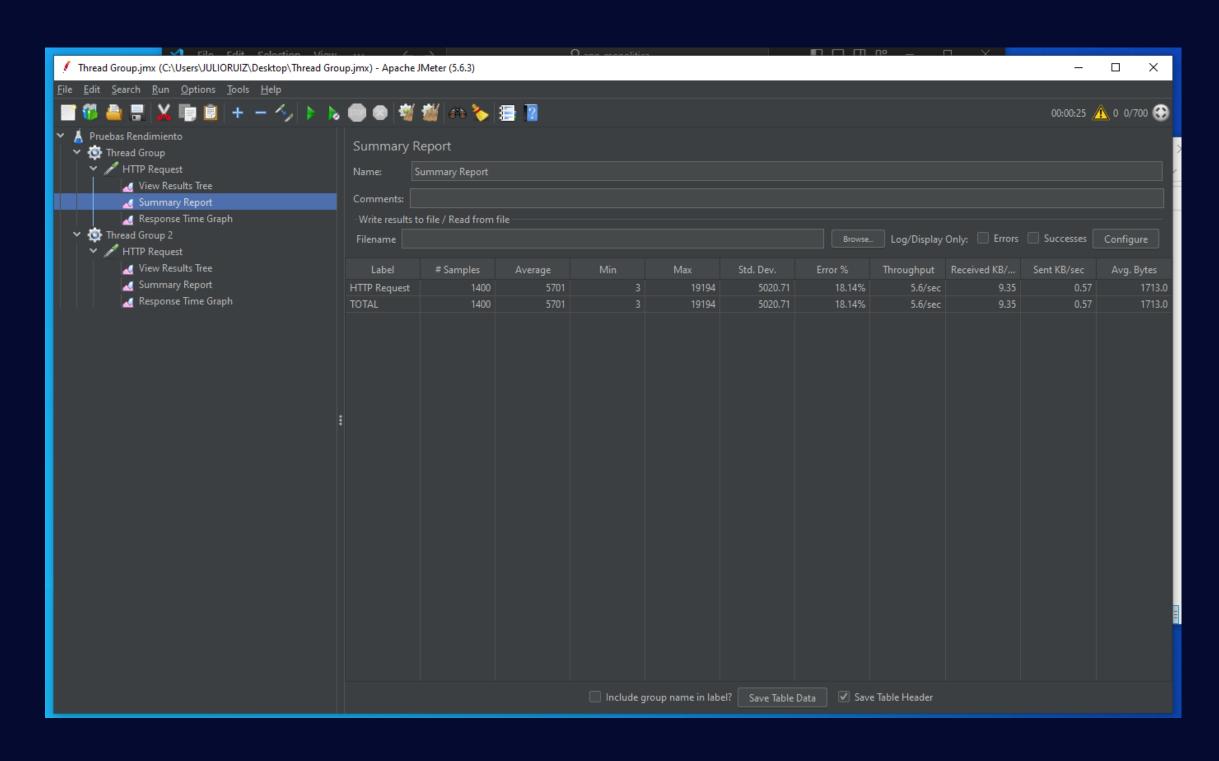


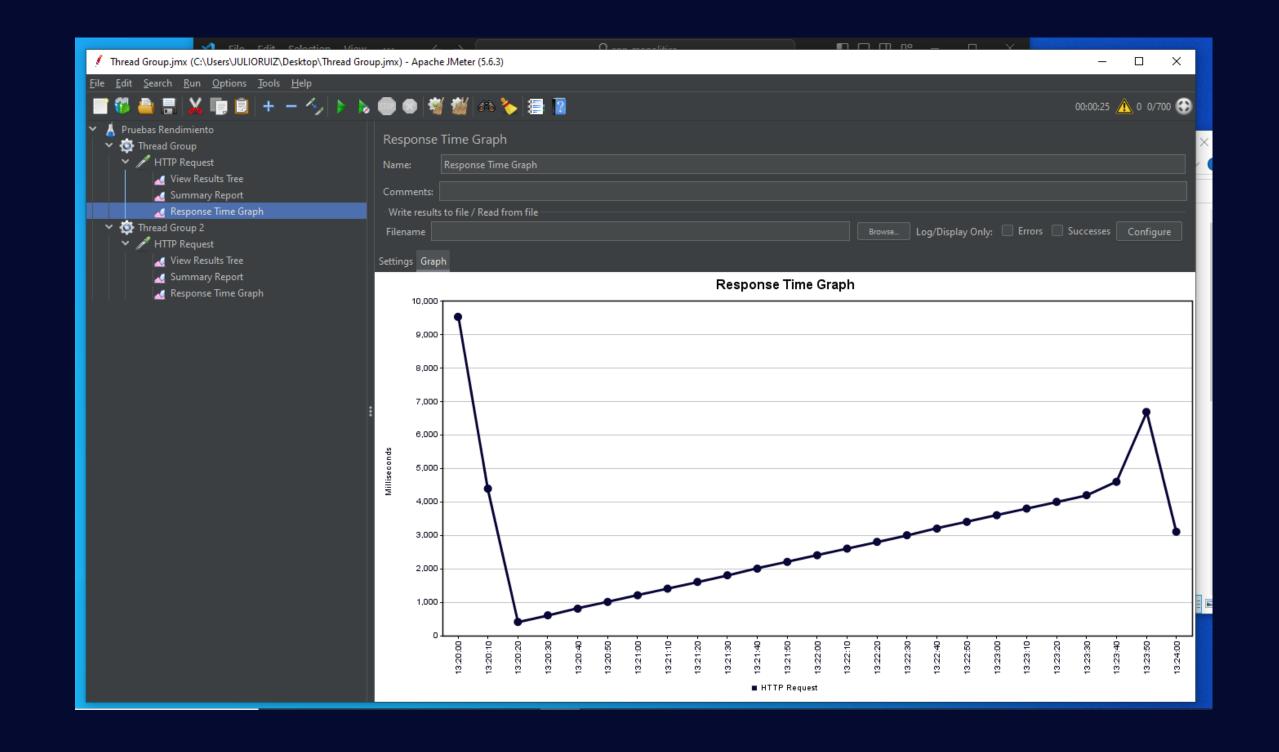
/login primera prueba



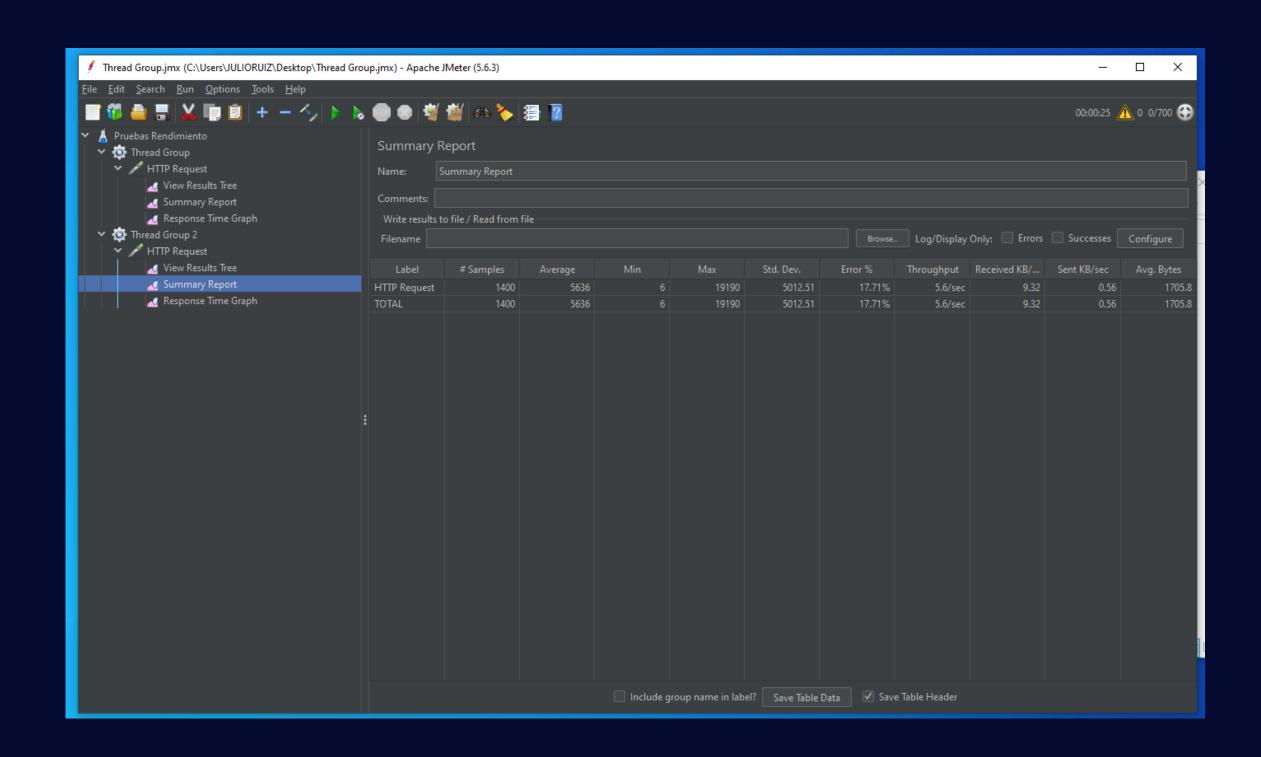


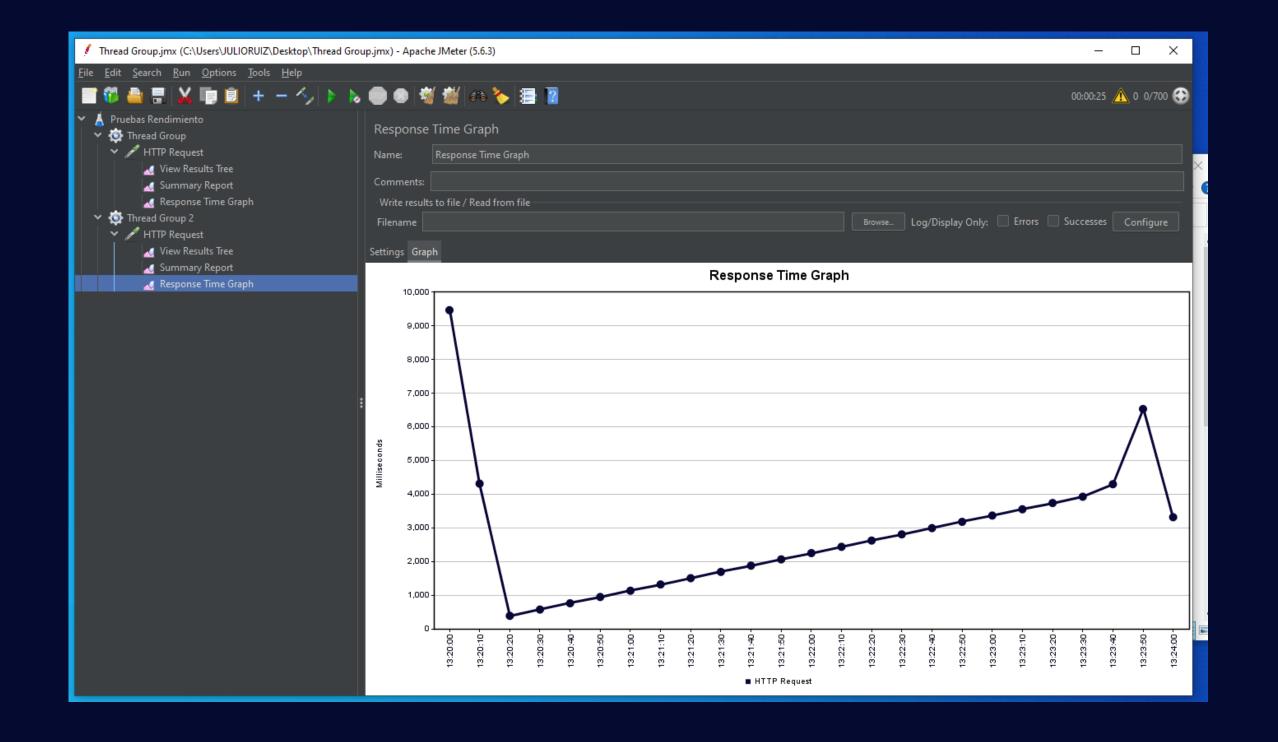
/register segunda prueba





/login segunda prueaba





CONCLUSIONES

Con un servidor más robusto (2 CPUs, 4 GB de RAM), la aplicación tomó un promedio de 16.1 segundos para completar operaciones tanto de registro como de login con 1400 usuarios concurrentes. Este tiempo de respuesta es bastante alto, lo que sugiere que la experiencia del usuario podría verse afectada negativamente bajo esta carga. En un entorno de producción, querrías apuntar a tiempos de respuesta inferiores a varios segundos.

Al reducir a la mitad los recursos del servidor (1 CPU, 2 GB de RAM), y también a la mitad la carga de usuarios (700 usuarios), el tiempo de respuesta aumentó significativamente a más de 27 segundos para ambas operaciones. Esto es un deterioro sustancial y señala que la aplicación es posiblemente más sensible a la disponibilidad de CPU que a la memoria, aunque ambos son factores críticos.

- La mejora en el tiempo de respuesta durante la "segunda vuelta" podría sugerir que hay un efecto de calentamiento en juego. Este efecto puede deberse a la caché de la base de datos que se llena, la JIT (Just-In-Time) compilation en el servidor de aplicaciones, o simplemente que ciertas operaciones iniciales pesadas ya se han realizado y no necesitan repetirse.
- Es interesante que, incluso con recursos reducidos, la segunda vuelta tuvo tiempos de respuesta de 5.6 segundos, lo cual es mejor que la primera vuelta con más recursos. Esto podría indicar que la aplicación tiene una etapa inicial pesada o que se beneficia significativamente de la caché o también debido al menor rendimiento del sistema se puede apreciar que existe una mayor cantidad de operaciones fallidas, las cuales provocarían un menor tiempo de ejecución.