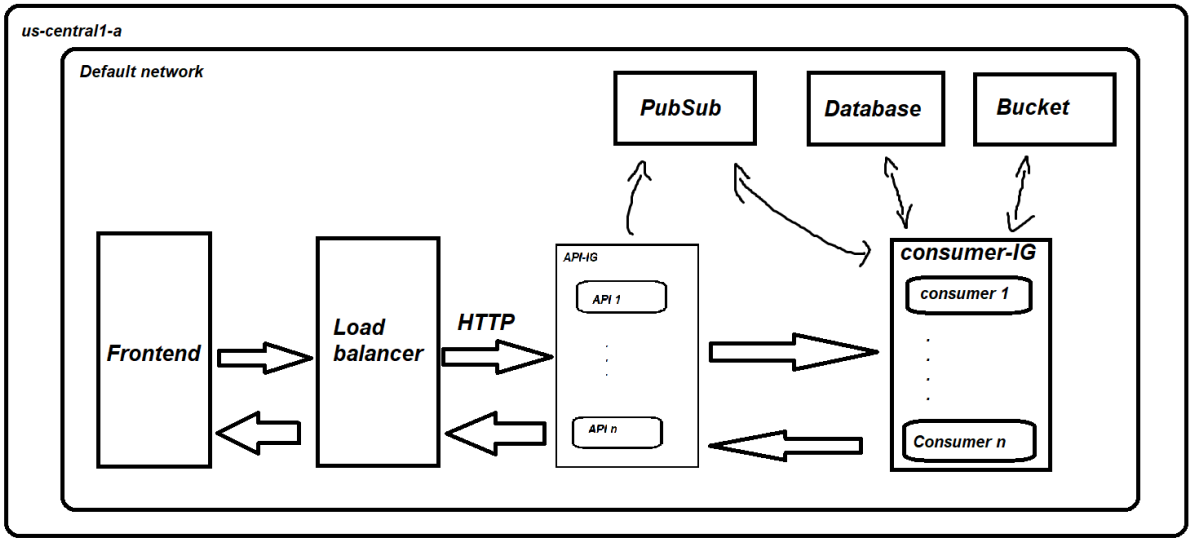
Informe de Pruebas de Estrés y Análisis de Capacidad

Gina Eveling Posada, Martin Daniel Rincón, Juan Camilo Muñoz, Felipe Serrano
MINE semestre 202410
Universidad de los Andes, Bogotá, Colombia
{g.posadas, md.rincon, jc.munozc12, ff.serrano42}@uniandes.edu.co
Fecha de presentación: abril 28 de 2024

Entorno de Prueba

La arquitectura se ilustra en la siguiente imagen. El frontend no tiene mayor cambio respecto a la entrega anterior, solo que ahora apunta a un balanceador de carga que se encarga de distribuir el tráfico http entre las instancias del grupo de API. Tanto el API como el consumer se implementan en contenedores docker que corren en máquinas virtuales; las máquinas de API y consumer se implementan en un "instance group", el cual se encarga de crear o eliminar instancias de acuerdo con las exigencias de recursos computacionales de la aplicación. El manejador de colas que antes se implementaba con rabbit en un contenedor que corría en la misma VM que el consumer, ahora se implementa con el servicio PubSub de GCP; el API escribe en las peticiones que luego son leídas por el consumer, quien se encarga de hacer la conversión y entregar archivos solicitados. La base de datos se implementa con el servicio de postgresql de GCP, su función es almacenar las direcciones en el bucket de los archivos. El bucket almacena los archivos y es accesible desde el consumer para borrar o insertar información.



Características y limitaciones de la infraestructura en producción:

- Proveedor Nube GPC cada estudiante cuenta con 1 cuenta con crédito de 50 usd para consumo en servicios de GCP.
- Redes e infraestructura con limitaciones de tiempo y restricciones que establece el proveedor GCP:
    - Reinicio de ips públicas de manera automática
    - Bloque de puertos que pueda el proveedor de servicios de nube implementar

    - Configuración de puertos de salida y entrada en cada máquina virtual
- Contenedores Docker: La aplicación actualmente está Dockerizado para todo su funcionamiento, es decir que se cuentan un contenedor por cada componente:
    - 1 contenedor Docker para el frontend: Streamlit
    - 1 contenedor Docker para el backend: Fast api
    - 1 contenedor Docker para el Worker consumidor: Python
- Servicios de GCP usados: SQL,PubSub,LoadBalancer,GCP storage.

Preparación para las Pruebas

Criterios de aceptación escenario 1:

Objetivo	Meta	Restricción
Carga de archivos a convertir	• La carga del archivo por parte del usuario debe realizarse en menos de 60 segundos • La disponibilidad de este servicio es superior al 99%	• Archivos con la extensión definida en el enunciado • Archivos de menos de 20 MB de tamaño • Solo se carga el archivo si el usuario ha sido autenticado
Descarga de archivos originales	• La descarga del archivo se realiza en menos de 60 segundos • La disponibilidad de este servicio es superior al 99%	• Velocidad de conexión del usuario • Si se presenta errores, el usuario puede volver a realizar la petición de descarga de archivo • El archivo cargado es menor de 20 MB de tamaño • Solo se carga el archivo si el usuario ha sido autenticado
Consulta de archivos convertidos	• Obtener el listado de los archivos y sus estados se realiza en menos de 2 segundos • La disponibilidad de este servicio es superior al 99% • El 100% de las veces se muestra la información del usuario en especifico	• Solo se carga el listado de documentos si el usuario ha sido autenticado
Conversión de archivos	• La conversión del archivo después de haber sido procesado por la cola de mensajes debe realizarse en menos de 350 segundos • La tasa de conversión de archivos (recibidos vs convertidos) debe ser superior al 97%	• Archivos con la extensión definida en el enunciado • Archivos de menos de 20 MB de tamaño • Solo se carga el archivo si el usuario ha sido autenticado
Login	• La disponibilidad de este servicio es superior al 99% • Al transmitir las credenciales correctas el 100% de las transacciones permiten el ingreso al sistema • La autenticación se realiza en menos de 2 segundos	• El usuario conoce sus credenciales y se transmiten de manera correcta
Creación de usuarios	• La disponibilidad de este servicio es superior al 80% • Si el usuario no está creado en el sistema (nombre+correo) el 100% de las creaciones se hacen de manera correcta	•

Configuración de JMeter:

Para el primer escenario, se configuro cada uno de los thread Groups, se configuro para simular un número diferente de usuarios. Se realizaron grupos de 10, 20, 30 y 40 usuario para simular una cantidad específica de usuarios ejecutando un escenario de prueba (Login, carga y lista), asimismo cada grupo de tuvo 4 escenarios diferentes de carga donde cargaban 10, 20, 30 y 40 archivos según las indicaciones.

Extensión	Cantidad	Min	Máx.
.odt	10	2 KB	6 KB
.docx	10	2 KB	8 KB
.xlsx	10	6 KB	13 KB
.pptx	10	17 KB	20 KB

Es decir, se realizó pruebas donde el grupo de usuario se enfrentaba a cargas en simultaneo de 10 a 40 archivos estas condiciones iban aumentando cada 10 archivos. Cada conjunto de archivos estaba configurado con archivos variados (.odt, .docx, .xlsx, .pptx) y diferentes tamaños en cada uno de los archivos. Las pruebas se organizaron en 4 bloques, el primer bloque 10 usuarios con situaciones de carga diferentes (10 archivos, 20 archivos, 30 archivos y 40 archivos).

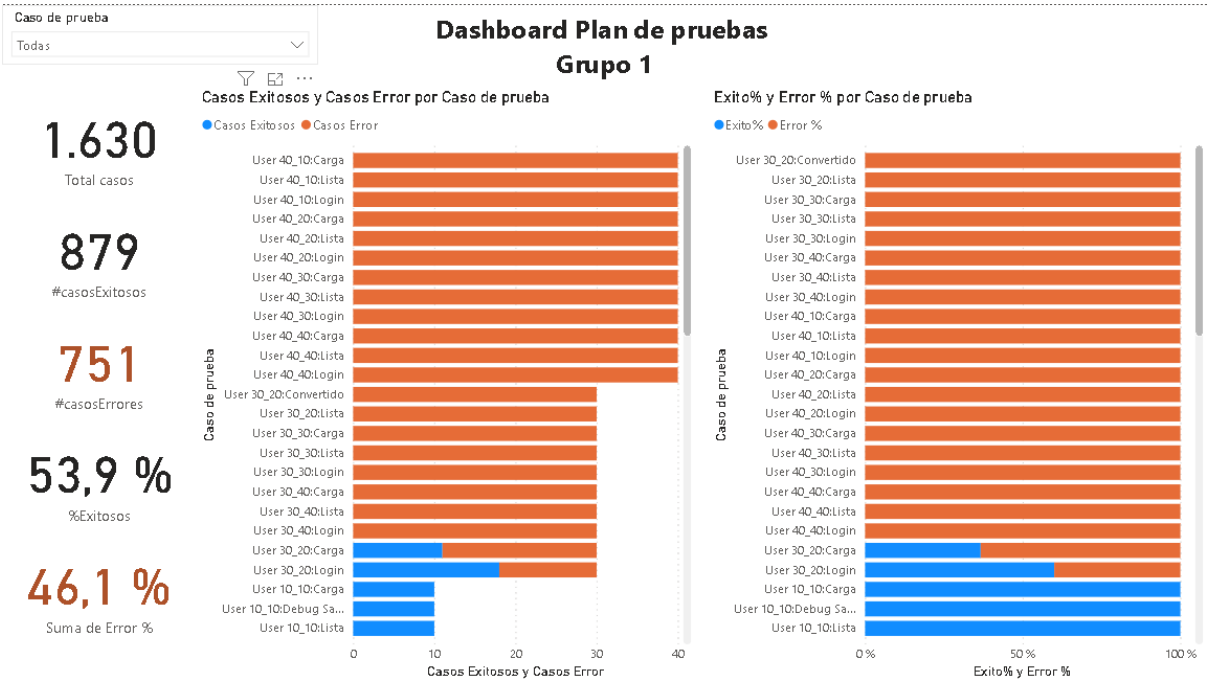
timeStamp	elapsed	label	responseCo	responseMe	threadNam	dataType	success	failureMess	bytes	sentBytes	grpThreads	allThreads	URL	Latency	IdleTime	Connect
1,71393E+1	117	Login	200	OK	User 10_10	text	TRUE		375	281	3	3	http://34.1(	116	0	23
1,71393E+1	117	Login	200	OK	User 10_10	text	TRUE		375	281	3	3	http://34.1(	116	0	23
1,71393E+1	329	Login	200	OK	User 10_10	text	TRUE		375	281	3	3	http://34.1(	328	0	23
1,71393E+1	5	Debug Sam	200	OK	User 10_10	text	TRUE		16897	0	5	5	null	0	0	0
1,71393E+1	5	Debug Sam	200	OK	User 10_10	text	TRUE		17089	0	5	5	null	0	0	0
1,71393E+1	5	Debug Sam	200	OK	User 10_10	text	TRUE		17090	0	5	5	null	0	0	0
1,71393E+1	94	Login	200	OK	User 10_10	text	TRUE		375	281	5	5	http://34.1(	94	0	6

Para este análisis se consideraron los reportes de “View result tree”, que tiene la siguiente estructura de forma detallada se puede ver lo siguientes registros:

No.Usuario y No. de documentos	Acción	Promedio de tiempo transcurrido (ms)	Tasa de éxito (%)	Promedio de latencia (ms)
User 10_10	Carga	476.2	100.0	476.1
User 10_10	Lista	268.1	100.0	262.6
User 10_10	Login	266.5	100.0	266.2
User 10_20	Carga	382.4	100.0	382.3
User 10_20	Lista	205.3	100.0	193.6
User 10_20	Login	184.6	100.0	184.5
User 10_30	Carga	388.0	100.0	387.9
User 10_30	Lista	185.8	100.0	162.2
User 10_30	Login	188.4	100.0	188.3
User 10_40	Carga	362.7	100.0	362.4
User 10_40	Lista	202.6	100.0	177.9
User 10_40	Login	182.2	100.0	182.2
User 20_10	Carga	1087.55	100.0	1087.25
User 20_10	Lista	784.90	100.0	761.55
User 20_10	Login	793.40	100.0	793.25
User 20_20	Carga	1077.60	100.0	1077.40
User 20_20	Lista	849.90	100.0	803.75
User 20_20	Login	779.25	100.0	779.10
User 20_30	Carga	1188.65	100.0	1188.45
User 20_30	Lista	864.45	100.0	816.35
User 20_30	Login	784.45	100.0	784.35
User 20_40	Carga	1201.35	100.0	1201.05
User 20_40	Lista	1154.10	100.0	1000.65
User 20_40	Login	854.20	100.0	854.10
User 30_10	Carga	1778.4	100.0	1778.03
User 30_10	Lista	1510.43	100.0	1510.43
User 30_10	Login	1599.97	100.0	1599.90
User 30_20	Carga	19592.83	36.67	19592.7
User 30_20	Lista	34441.3	0.0	34441.23
User 30_20	Login	12357.0	60.0	12356.97
User 30_30	Carga	30128.03	0.0	30127.97
User 30_30	Lista	37489.13	0.0	37489.13
User 30_30	Login	30110.4	0.0	30110.23
User 30_40	Carga	30136.0	0.0	30135.93
User 30_40	Lista	37584.93	0.0	37584.9
User 30_40	Login	30104.53	0.0	30104.4

User 40_10	Carga	30126.05	0.0	30126.00
User 40_10	Lista	37010.60	0.0	37010.50
User 40_10	Login	30111.60	0.0	30111.50
User 40_20	Carga	30131.13	0.0	30131.03
User 40_20	Lista	37230.30	0.0	37230.30
User 40_20	Login	30111.73	0.0	30111.63
User 40_30	Carga	30129.20	0.0	30129.13
User 40_30	Lista	37824.78	0.0	37824.68
User 40_30	Login	30093.95	0.0	30093.73
User 40_40	Carga	30128.55	0.0	30128.50
User 40_40	Lista	37221.32	0.0	37221.32
User 40_40	Login	30105.38	0.0	30105.30

Tabla 1. Detalle de rendimiento del desarrollo



Como se puede observar durante los grupos de 10 y 20 usuarios el desarrollo colocado a prueba responde satisfactoriamente sin embargo cuando pasamos a 30 usuarios la arquitectura diseñada falló.

No.Usuario	Acción	Promedio de tiempo transcurrido (ms)	Tasa de éxito (%)	Promedio de latencia (ms)
User 10	Carga	402.33	100.0	402.18
User 10	Lista	215.45	100.0	199.08
User 10	Login	205.43	100.0	205.30
User 20	Carga	1138.79	100.0	1138.54
User 20	Lista	913.34	100.0	845.58
User 20	Login	802.83	100.0	802.70
User 30	Carga	20408.82	34.17	20408.66
User 30	Lista	27756.45	25.0	27756.43
User 30	Login	18542.98	40.0	18542.88
User 40	Carga	30128.73	0.0	30128.66
User 40	Lista	37321.75	0.0	37321.70
User 40	Login	30105.66	0.0	30105.54

Tabla 2. Resumen de rendimiento por cantidad de usuarios

De forma consolidada, se puede detallar que las acciones asociadas tienen tiempos de respuesta y latencias relativamente bajos, para los grupos de 10 y 20 usuarios, lo que indica una buena

configuración del entorno de prueba y una gestión eficiente de los recursos, además hay consistencia en las acciones de 'Login', 'Carga', y 'Lista'. Sin embargo, para el caso del grupo de usuarios 30 y 40 se presentan tiempos extremadamente altos tanto en respuesta como en latencia, lo que afectó la tasa de éxito de las acciones solicitadas.

Por otro lado, al usar el “summary report”, los tiempos de respuesta aumentan significativamente en las acciones relacionada a partir del grupo de usuario 30, un aumento del 1000%, es decir, la acción de carga de 10 archivos de 30 usuarios el tiempo promedio de respuesta fie de 1778 ms, mientras para la carga de 20 archivos y la misma cantidad de usuarios el tiempo de respuesta fie de 19592 ms. Este aumento masivo en los tiempos de respuesta implica una significativa degradación del rendimiento a medida que el número total de usuarios simultáneos y archivos cargados aumenta, lo que indica que el sistema llego a su capacidad límite y posiblemente sufriendo de saturación bajo cargas muy altas.

Este aumento de respuesta es consistente con los valores de throughput, notablemente bajos, en su conjunto aspectos como el throughput de recepción (Received KB/sec), throughput de envío (Sent KB/sec), y el tamaño promedio de los bytes (Avg. Bytes). Los valores de Sent KB/sec y Received KB/sec van decreciendo en la medida que aumenta el número de usuarios, especialmente en el grupo de 30 usuarios, esto indica que hay menos eficiencia en la transferencia de datos o que el sistema está bajo una carga más pesada, lo que afecta la tasa de transferencia de datos.

La variabilidad entre las métricas de acciones de carga y lista para los diferentes grupos de usuarios podría deberse a diferentes tamaños de carga, diferentes tiempos de respuesta del servidor, o incluso a la naturaleza de los datos solicitados o enviados.

Frente a los objetivos del escenario se puede considerar que la acción del login, los datos muestran tiempos de respuesta que superan los 2 segundos para ciertos conjuntos de usuarios, lo cual no cumple con la meta establecida en el plan de carga. Por su parte los tiempos de carga, no logran ser medidos de forma directa, sin embargo, con base en los datos de Avg. Bytes y Throughput que son bajos esto es indicativo que tiempos de carga más lentos, entonces este podría ser un área de preocupación

Criterios de aceptación escenario 2:

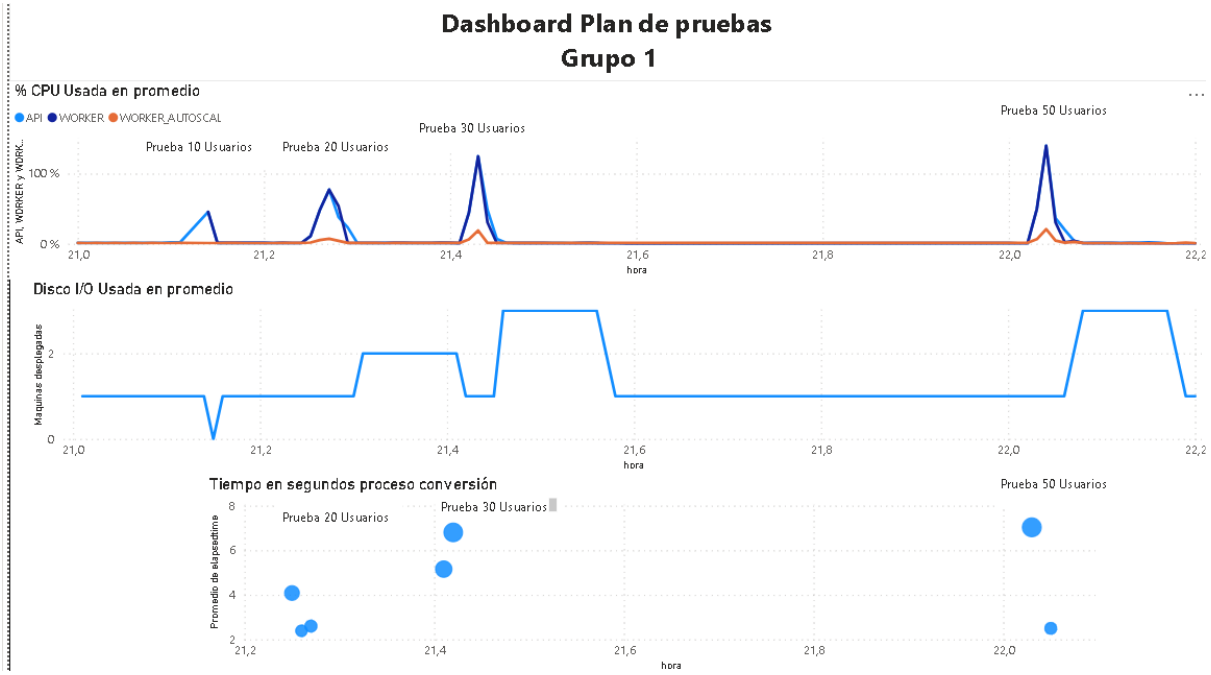
Escenario	Datos de prueba	Métricas que recopilar
Carga de archivos	● Cuentas de usuarios creados en el sistema ● Archivos ficticios de carga por cada usuario	● Tiempo para login de usuario ● Tiempo de carga de cada archivo ● # de login satisfactorios ● # de cargas satisfactorio ● # de peticiones fallidas con error 404 ● Uso de recursos de componente (% utilización)
Descarga de archivo original	● Cuentas de usuarios creados en el sistema ● Lista de archivos cargados en el sistema (ids)	● Tiempo para login de usuario ● Tiempo de descarga de cada archivo ● # de login satisfactorios ● # de descarga satisfactorias ● # de peticiones fallidas con error 404 ● Uso de recursos de componente (% utilización)
Conversión de archivos a PDF.	● Cuentas de usuarios creados en el sistema	● Tiempo para login de usuario

	• Lista de archivos cargados en el sistema (ids)	• Tiempo de conversión de cada archivo • # de login satisfactorios • # de conversión correctas • # de conversiones fallidas • Uso de recursos de componente (% utilización)

Para las estadísticas de rendimiento se realizaron extracciones de las métricas de performance de las máquinas virtuales, donde se puede calcular el tiempo que le lleva al proceso convertir el archivo:

id_document	id_user	source_filename	source_file	source_file_extension	pdf_file	status	upload_datetime	converted_datetime
1096	1	archivo1.odt	https://storage.cloud.google.com/sc-entrega3/archivo1.odt	.odt	https://storage.cloud.google.com/sc-entrega3/archivo1.pdf	Disponible	23/04/2024 23:04	23/04/2024 23:06
1097	1	archivo1.odt	https://storage.cloud.google.com/sc-entrega3/archivo1.odt	.odt	https://storage.cloud.google.com/sc-entrega3/archivo1.pdf	Disponible	23/04/2024 23:07	23/04/2024 23:07
1098	1	archivo10.odt	https://storage.cloud.google.com/sc-entrega3/archivo10.odt	.odt	https://storage.cloud.google.com/sc-entrega3/archivo10.pdf	Disponible	23/04/2024 23:07	23/04/2024 23:07
1099	1	archivo2.odt	https://storage.cloud.google.com/sc-entrega3/archivo2.odt	.odt	https://storage.cloud.google.com/sc-entrega3/archivo2.pdf	Disponible	23/04/2024 23:07	23/04/2024 23:07
1100	1	hoja_calculo8.xlsx	https://storage.cloud.google.com/sc-entrega3/hoja_calculo8.xlsx	.xlsx	https://storage.cloud.google.com/sc-entrega3/hoja_calculo8.pdf	Disponible	23/04/2024 23:07	23/04/2024 23:07
1102	1	presentacion1.pptx	https://storage.cloud.google.com/sc-entrega3/presentacion1.pptx	.pptx	https://storage.cloud.google.com/sc-entrega3/presentacion1.pdf	Disponible	23/04/2024 23:07	23/04/2024 23:07
1101	1	hoja_calculo9.xlsx	https://storage.cloud.google.com/sc-entrega3/hoja_calculo9.xlsx	.xlsx	https://storage.cloud.google.com/sc-entrega3/hoja_calculo9.pdf	Disponible	23/04/2024 23:07	23/04/2024 23:07

También se consideraron datos asociados a disco operaciones de lectura y escritura por contenedor y porcentaje de uso de CPU por contenedor.



En la parte superior del tablero se puede detallar el rendimiento del API_WORKER y WORKER_AUTOSCAL. En consonancia con las pruebas ejecutadas se evidencian los picos de carga de trabajo donde se presenta un uso de la CPU al 50%, 80%, 120% y 160%, que corresponden a las pruebas específicas para 10, 20, 30 y 40 usuarios. Los picos más significativos ocurren durante las pruebas de 30 y 40 usuarios, lo que sugiere un aumento en la carga de trabajo del CPU durante estas pruebas, especialmente en las acciones login y carga de documentos.

En esta prueba el escalado automático (AUTOSCAL) no está incrementando los recursos significativamente, de acuerdo con las exigencias de las tareas, por lo que es necesario revisar las condiciones de este para su configuración.

Por su parte, Disco I/O muestra picos más pronunciados durante las pruebas de 10, 20, 30 y 40 usuarios, lo que sugiere una mayor actividad de escritura/lectura en el disco durante estas pruebas,

para el caso de las pruebas del usuario 10, el umbral del escalado automático definido no se alcanzó durante las pruebas con el usuario 10, por lo que el uso de CPU y recursos no justificó el lanzamiento de más instancias.

Sin embargo, si bien cuando se presenta las acciones del conjunto de usuarios 20, se despliegan 2 máquinas, pero a destiempo de la solicitud, es decir, se despliega después de la solicitud y no da respuesta efectiva a la operación solicitada, esta misma situación ocurre cuando se despliegan 3 máquinas adicionales para los grupos de 30 y 40 usuarios, pero tiempo después del pico de la solicitud. Por lo que es notable que a mayor número de usuarios hay un cuello de botella, que no es intervenido de forma efectiva, lo que impide que el disco sea más utilizado.

Los picos en ambas métricas durante las pruebas indican que las pruebas de carga están generando un aumento en la utilización de recursos, sin embargo, una falta de incremento proporcional en el uso de CPU y disco para las pruebas de mayor cantidad de usuarios podría sugerir un posible cuello de botella o una eficiencia mejorada en el manejo de recursos a medida que aumenta la carga.

Resultados y Análisis

En consideración con lo representado previamente, es necesario considerar la escalabilidad vertical, ya que los recursos de CPU y Disco I/O no aumentan proporcionalmente con el aumento de la carga, lo que lleva a un rendimiento degradado con grupos de usuarios más grandes (30 y 40 usuarios).

Existen cuellos de botella significativos cuando se incrementa el número de usuarios y la carga de trabajo simultáneos, por lo que, los tiempos de respuesta y latencia aumentan dramáticamente con 30 y 40 usuarios, indicando que la infraestructura actual no puede manejar eficientemente cargas altas.

El escalado automático es inefectivo, ya no se activa a tiempo o no se escalan suficientes recursos para manejar los picos de carga, lo cual se muestra en la falta de uso del WORKER_AUTOSCAL para el grupo de 10 usuarios y el despliegue tardío de máquinas adicionales para los grupos de 20, 30 y 40 usuarios.

Es fundamental, revisar la configuración del escalado automático y la capacidad de las máquinas para garantizar que los recursos se escalen de manera proactiva y eficiente en respuesta a los picos de carga. Adicionalmente, los recursos de CPU por contenedor se ven altamente utilizados en cargas mayores, lo cual podría ser un indicador de que la aplicación podría beneficiarse de una mejor distribución de cargas.