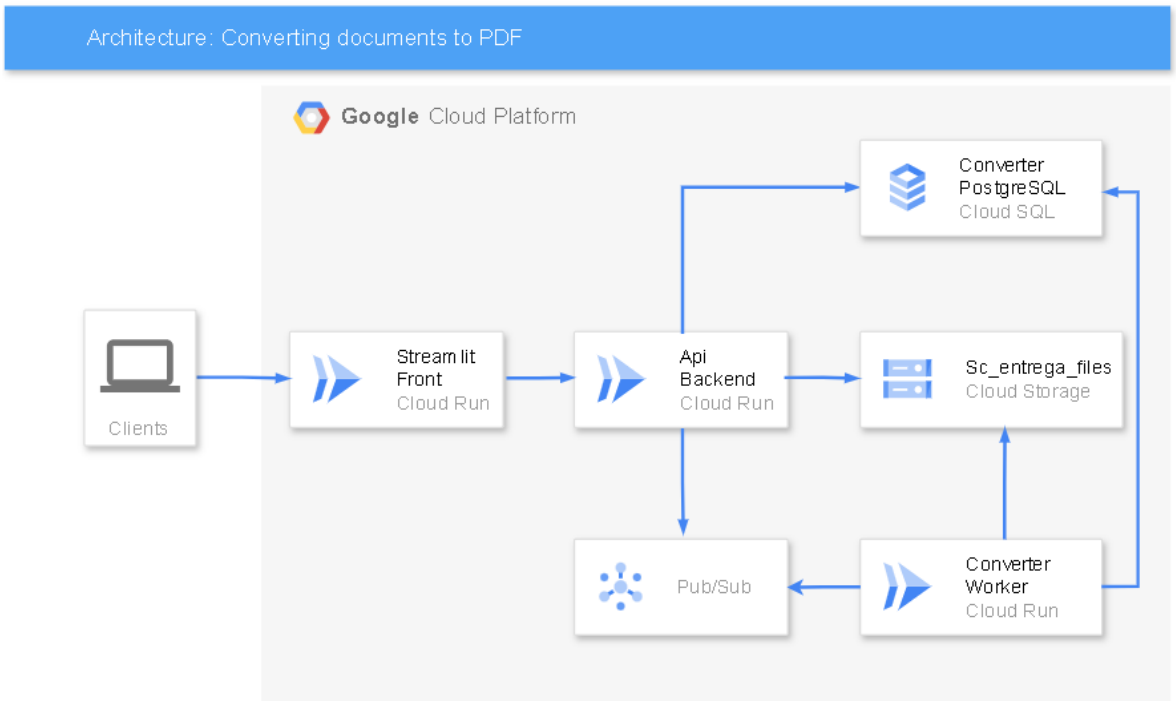
P4 - Informe de Pruebas de Estrés y Análisis de Capacidad

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: mayo 26 de 2024

Entorno de Prueba

La aplicación ha sido desarrollada implementando 3 componentes de Cloud Run, un servicio Cloud SQL, un Cloud Storage para el almacenamiento de archivos y un servicio de Pub/Sub para gestionar el procesamiento desacoplado de la conversión de los archivos.

Por su parte el Frontend, estuvo desarrollado en Streamlit Web API garantizando una interfaz de usuario ágil y eficiente. Permite la interacción fácil y directa con los servicios del backend a través del protocolo HTTP, por su parte la API esta es una imagen de Docker desplegada en Cloud Run y está implementada usando el framework FastAPI, lo que proporciona una estructura robusta para manejar las funcionalidades del negocio. La base de datos es una imagen de Docker desplegada en Cloud Run, lo que proporciona una estructura robusta para manejar las funcionalidades del negocio. Worker (Consumer) se desplego en Cloud Run. Este componente ejecuta un programa en Python en su propio contenedor, que procesa de manera asíncrona los mensajes de la cola PUB/SUB para la conversión de archivos. Finalmente, se bucket de Cloud Storage para alojar tanto los archivos originales como los PDFs convertidos.



- 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 configuró 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.

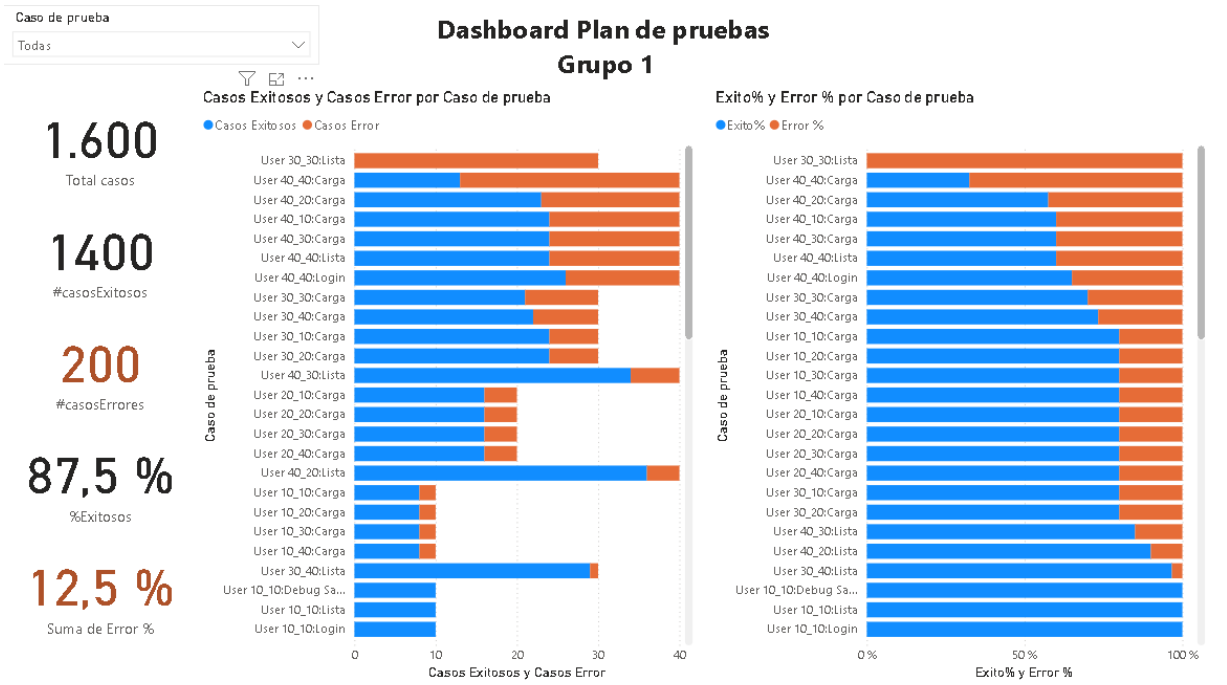
Extensión	Cantidad	Min	Máx.
.odt	10	2 KB	6 KB
.docx	10	2 KB	8 KB
.xlsx	10	6 KB	13 KB
.pptx	10	17 KB	20 KB

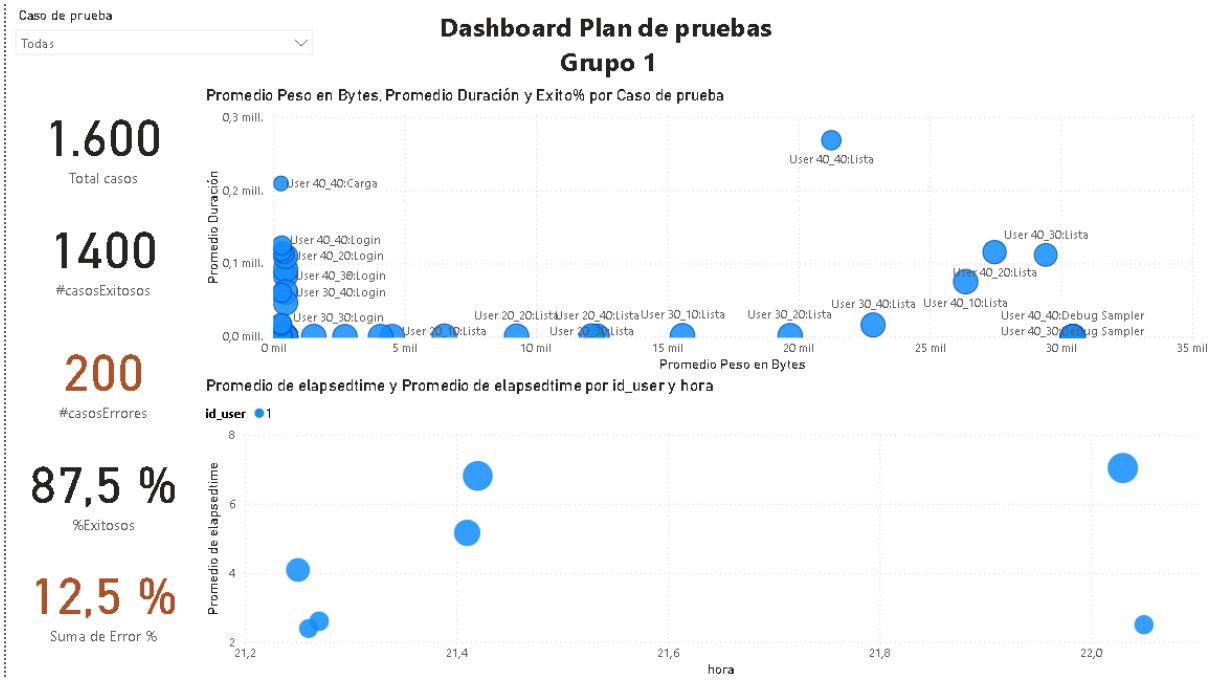
Es decir, se realizó pruebas donde el grupo de usuario se enfrentaba a cargas en simultaneo de 10 a 40 archivos estas condiciones iban aumentando cada 10 archivos. Cada conjunto de archivos estaba configurado con archivos variados (.odt, .docx, .xlsx, .pptx) y diferentes tamaños en cada uno de los archivos. 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).

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:

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





Con un 87.5% de éxito, el plan de pruebas muestra una tasa de éxito general decente, pero el 12.5% de error aún destaca la necesidad de mejoras y correcciones en ciertas áreas. 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. Asimismo, se observan diferencias significativas en el tiempo que tardan las pruebas dependiendo de la hora, lo cual puede indicar variaciones en la carga del sistema, eficiencia de procesamiento o disponibilidad de recursos.

Al momento de revisar la distribución del Elapsed Time se observan diferencias significativas en el tiempo que tardan las pruebas dependiendo del tiempo, lo cual puede indicar variaciones en la carga del sistema, eficiencia de procesamiento o disponibilidad de recursos.

No.Usuario	Acción	Promedio de tiempo transcurrido (ms)	Tasa de éxito (%)	Promedio de latencia (ms)
User 10	Carga	431.50	80.00	431.50
User 10	Lista	270.25	100.00	270.25
User 10	Login	331.00	100.00	331.00
User 20	Carga	972.75	80.00	972.75
User 20	Lista	787.00	100.00	787.00
User 20	Login	739.00	100.00	739.00
User 30	Carga	9325.50	75.83	9325.50
User 30	Lista	25798.00	74.17	25798.00
User 30	Login	27450.25	100.00	27450.25
User 40	Carga	127620.25	52.50	127620.25
User 40	Lista	142665.75	83.75	142665.75
User 40	Login	99769.75	91.25	99769.75

Tabla 1. 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 utilizar el “summary report”, los tiempos de respuesta para ciertas operaciones aumentan significativamente en las acciones a partir del grupo de usuario 30. Específicamente, observamos un aumento en la latencia de más del 1000% para ciertas tareas. Por ejemplo, la acción de carga para los usuarios de User 30 tiene un tiempo promedio de respuesta de 9325.50 ms. Sin embargo, en comparación directa, User 40, bajo un escenario más demandante, registra un tiempo de respuesta aún mayor, llegando a 127620.25 ms para una tarea similar. 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 llegó a su capacidad límite y posiblemente está sufriendo de saturación bajo cargas muy altas.

Este aumento en la respuesta es consistente con los valores de throughput, que son notablemente bajos para operaciones más intensivas. Los valores de Sent KB/sec y Received KB/sec decrecen notablemente en el grupo de usuario 40, lo que 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.

En cuanto a los tiempos de login, los datos muestran que, para el grupo User 40, los tiempos de respuesta superan los 99 segundos para ciertos conjuntos de usuarios, lo cual no cumple con la meta establecida en el plan de carga. Aunque los tiempos de carga no se logran medir de forma directa, con base en los datos de Avg. Bytes y Throughput que son bajos, esto es indicativo de tiempos de carga más lentos, entonces este podría ser un área de preocupación.

Resultados y Análisis

Es vital investigar más a fondo los casos con mayores tasas de error para entender las causas y determinar si son errores sistémicos, problemas de implementación o defectos en el diseño de las pruebas. Según el análisis de la carga de bytes y su impacto en la duración de las pruebas podría ayudar a optimizar el uso de recursos, especialmente durante horas de alta carga o para pruebas que manejan grandes volúmenes de datos. Desde luego se evidencia que existe una correlación entre el peso de bytes y la duración de las pruebas requiere una revisión continua para asegurar que el sistema está escalado adecuadamente y que las pruebas están diseñadas para maximizar la precisión sin comprometer el rendimiento.

Se requiere revisar de forma detallada la distribución de cargas de trabajo revelan diferencias significativas en los pesos de bytes y los tiempos de duración de las pruebas entre diferentes usuarios y tipos de pruebas, lo que implicaría revisar algunas funcionalidades de la arquitectura que pueden estar requiriendo recursos considerables, lo cual podría influir en la duración y posiblemente en la eficiencia de las pruebas.

Finalmente, en algunos tipos de pruebas como "User 30-Lista" y "User 40-Carga" se mostró un número más alto de errores, lo que indica áreas críticas que requieren atención adicional, ya sea en términos de revisión de código, optimización de la configuración de la infraestructura subyacente. Existe un espacio claro para la optimización y mejora continua, tanto en la reducción del porcentaje de errores como en la mejora de la eficiencia del proceso de pruebas, por lo que se requiere implementar estrategias para manejar pruebas que involucran grandes volúmenes de datos, ajustar la planificación de pruebas según la disponibilidad de recursos y la carga del sistema, y abordar las áreas con altas tasas de error son pasos cruciales hacia la mejora del rendimiento del software y la efectividad de las pruebas.