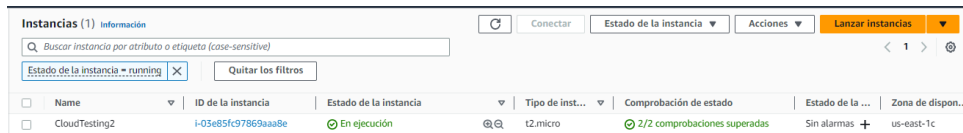


ANÁLISIS DE CAPACIDAD DESPLIEGUE EN PAAS

Las pruebas presentadas en este documento fueron realizadas en una instancia t2.micro de AWS.



Instancias (1) Información

Conectar Estado de la instancia Acciones Lanzar instancias

Buscar instancia por atributo o etiqueta (case-sensitive)

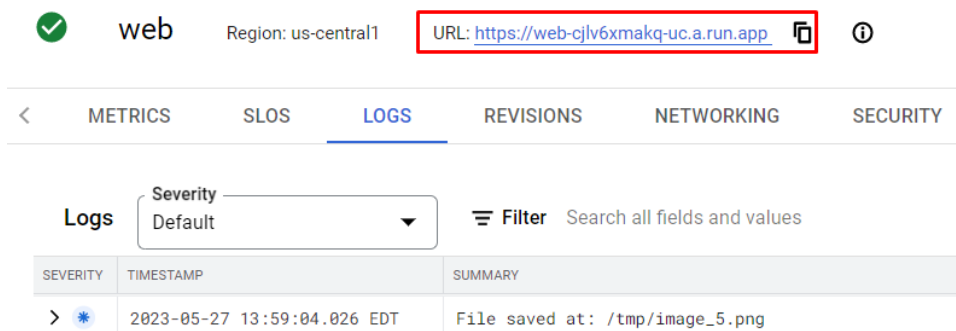
Estado de la instancia: running Quitar los filtros

Name	ID de la instancia	Estado de la instancia	Tipo de inst...	Comprobación de estado	Estado de la ...	Zona de dispon...
CloudTesting2	i-03e85fc97869aaa8e	En ejecución	t2.micro	2/2 comprobaciones superadas	Sin alarmas +	us-east-1c

Las propiedades de este tipo de instancia se muestran en la siguiente imagen:

Instancia	CPU virtual*	Créditos por hora de CPU	Memoria (GiB)	Almacenamiento	Rendimiento de red
t2.nano	1	3	0,5	Solo EBS	Bajo
t2.micro	1	6	1	Solo EBS	De bajo a moderado

Para la realización de dichas pruebas las herramientas de Postman y Jmeter apuntaron a la URL del servicio de Cloud Run donde fue desplegado el servicio web.



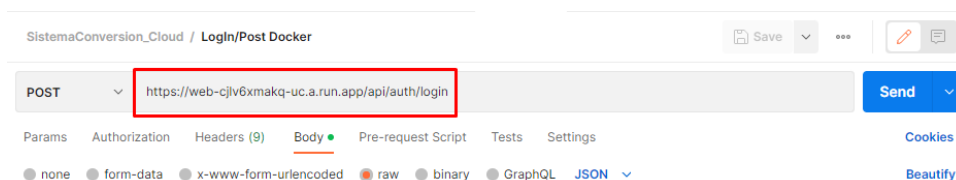
web Region: us-central1 URL: <https://web-cjlv6xmakq-uc.a.run.app>

METRICS SLOS LOGS REVISIONS NETWORKING SECURITY

Logs Severity: Default Filter Search all fields and values

SEVERITY	TIMESTAMP	SUMMARY
> *	2023-05-27 13:59:04.026 EDT	File saved at: /tmp/image_5.png

Url base Utilizada en PostMan



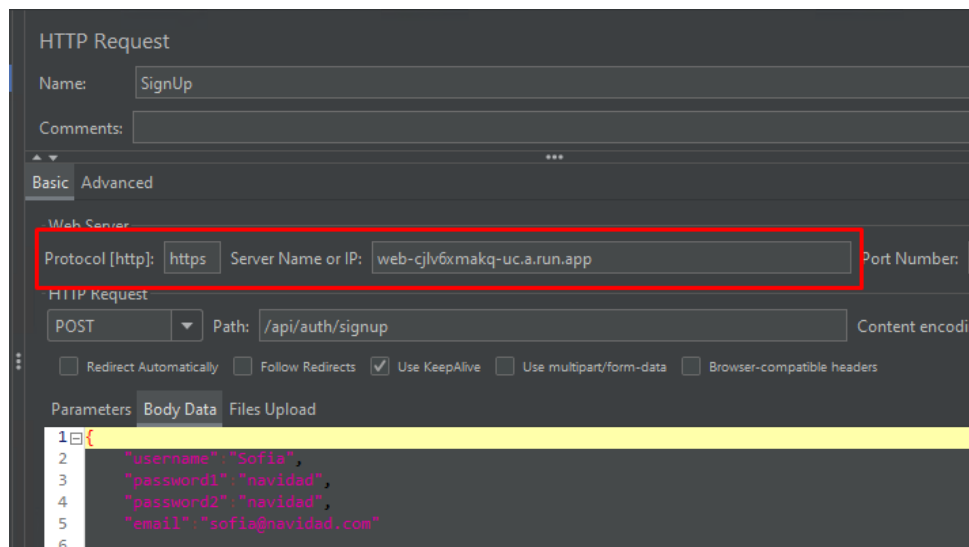
SistemaConversion_Cloud / Login/Post Docker Save ...

POST <https://web-cjlv6xmakq-uc.a.run.app/api/auth/login> Send

Params Authorization Headers (9) Body Pre-request Script Tests Settings Cookies Beautify

none form-data x-www-form-urlencoded raw binary GraphQL JSON

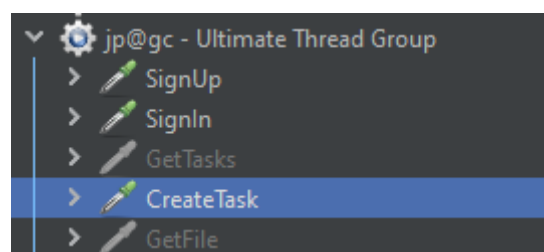
Url base Utilizada en JMeter



Se realizaron pruebas de rendimiento de acuerdo a los escenarios posteriormente descritos:

Escenario 1. Se deberá definir un escenario donde se pueda probar cuál es la máxima cantidad de requests HTTP por minuto que soporta la aplicación web con usuarios. Para hacer pruebas de estrés se debe utilizar la herramienta Apache Bench (ab) o JMeter. Las pruebas de estrés deberán realizarse desde otros equipos diferentes a los utilizados para ejecutar el servidor web y el servidor de base de datos. El escenario y los resultados de las pruebas de estrés deberán ser documentados con gráficas que ilustran cómo se comporta el sistema a medida que el número de clientes accediendo a la aplicación se incrementa hasta llegar al punto de degradar completamente el rendimiento de esta.

Para realizar esta prueba, se configuraron en JMeter las peticiones correspondientes al proceso de SignUp, SignIn, y Create Task.



Las peticiones de SignUp y SignIn eran necesarias en el grupo de peticiones ya que apartir de la información de respuesta de dichas peticiones se obtiene el requerido actualmente por la aplicación para poder realizar la petición de Create Task.

A continuación, se presentan los datos usados en las peticiones en mención:

SignUp:

HTTP Request

Name:

Comments:

Basic Advanced

Web Server

Protocol [http]: Server Name or IP:

HTTP Request

POST Path:

☐ Redirect Automatically ☐ Follow Redirects ☒ Use KeepAlive ☐ Use multipart/form-data ☐ Browser-compatible headers

Parameters Body Data Files Upload

```
1 {
2   "username": "Sofia",
3   "password1": "navidad",
4   "password2": "navidad",
5   "email": "sofia@navidad.com"
}
```

Sign In:

HTTP Request

Name:

Comments:

Basic Advanced

Web Server

Protocol [http]: Server Name or IP:

HTTP Request

POST Path:

☐ Redirect Automatically ☐ Follow Redirects ☒ Use KeepAlive ☐ Use multipart/form-data ☐ Browser-compatible headers

Parameters Body Data Files Upload

```
1 {
2   "username": "Sofia",
3   "password": "navidad"
4 }
5
```

Create Task:

HTTP Request

Name:

Comments:

Basic Advanced

Web Server

Protocol [http]: Server Name or IP:

HTTP Request

POST Path:

☐ Redirect Automatically ☐ Follow Redirects ☒ Use KeepAlive ☐ Use multipart/form-data ☐ Browser-compatible headers

Parameters Body Data Files Upload

Send Parameters With the Request:

Name:	Value	URL Encode?	Content-Type
newFormat	ZIP	<input type="checkbox"/>	text/plain

HTTP Request

Name: CreateTask

Comments:

Basic Advanced

Web Server

Protocol [http]: https Server Name or IP: web-cjlv6xmakq-uc.a.run.app

HTTP Request

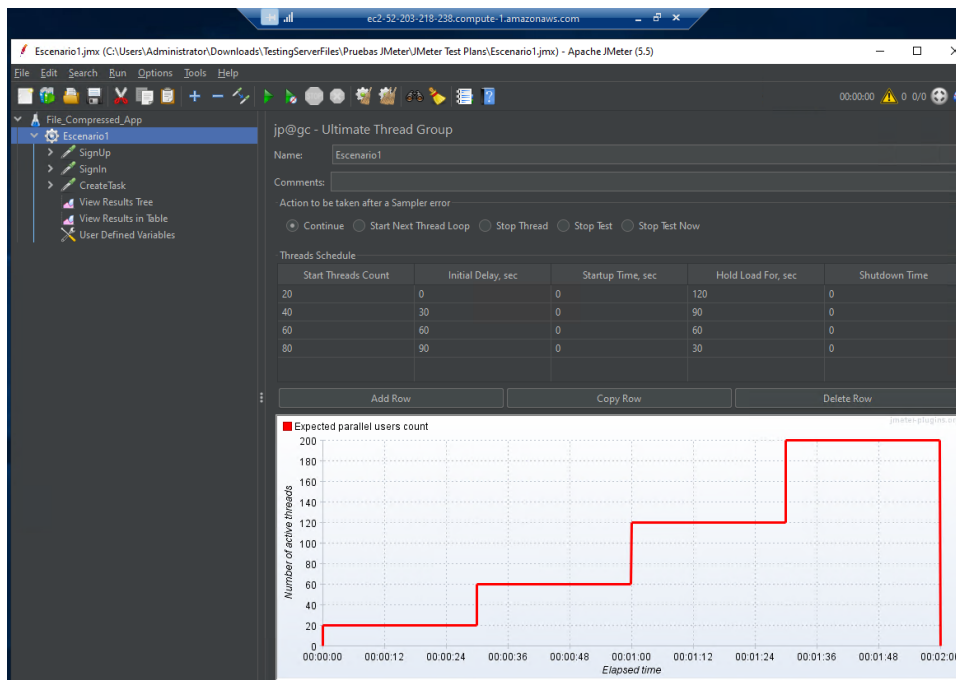
POST Path: /api/tasks

☐ Redirect Automatically ☐ Follow Redirects ☒ Use KeepAlive ☐ Use multipart/form-data ☐ Browser-compatible headers

Parameters Body Data Files Upload

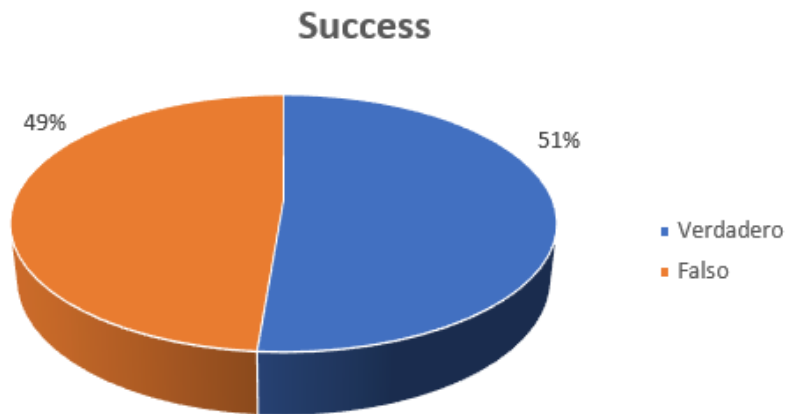
File Path	Parameter Name	Value
C:\Users\Administrator\Downloads\Testing...	fileName	application/pdf

La rampa de carga utilizada en este escenario se presenta en la siguiente imagen:

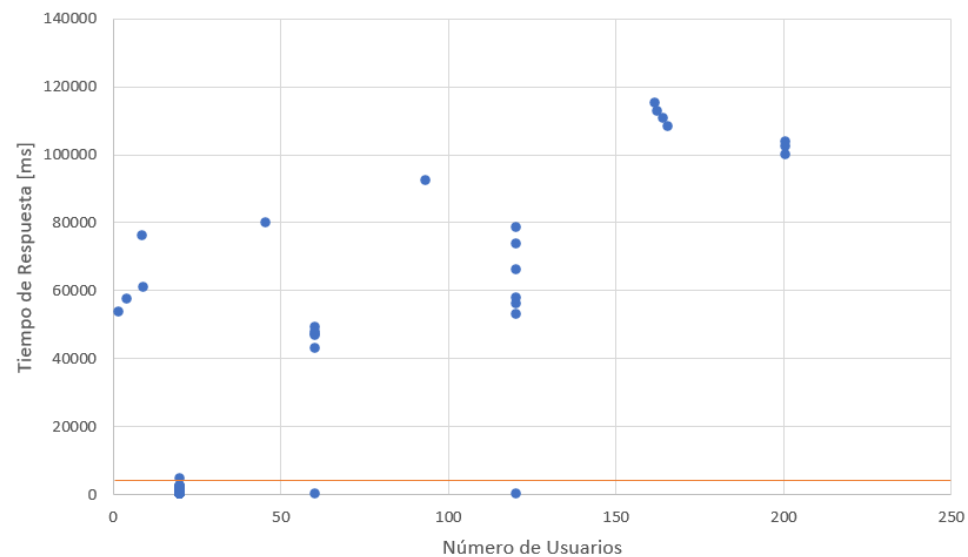


A continuación se presentan los resultados obtenidos para las pruebas realizadas en el Despliegue previo Alta disponibilidad del Web Server + Servicio de Mensajes Cloud Pub/Sub + Worker con Autoscaling y el Despliegue en PAAS (Entrega Actual).

Porcentaje de Peticiones Exitosas – Despliegue Previo

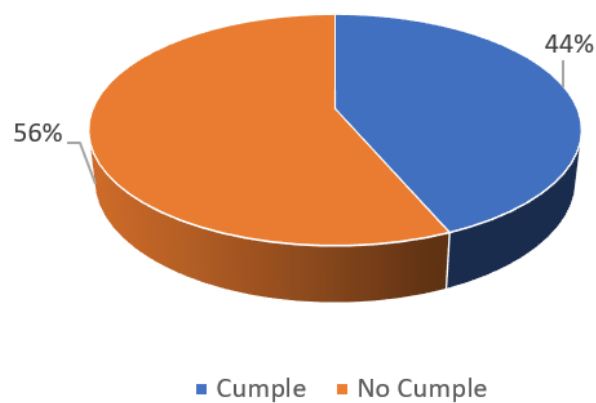


Grafica Tiempo de Respuesta Vs Usuarios – Despliegue Previo

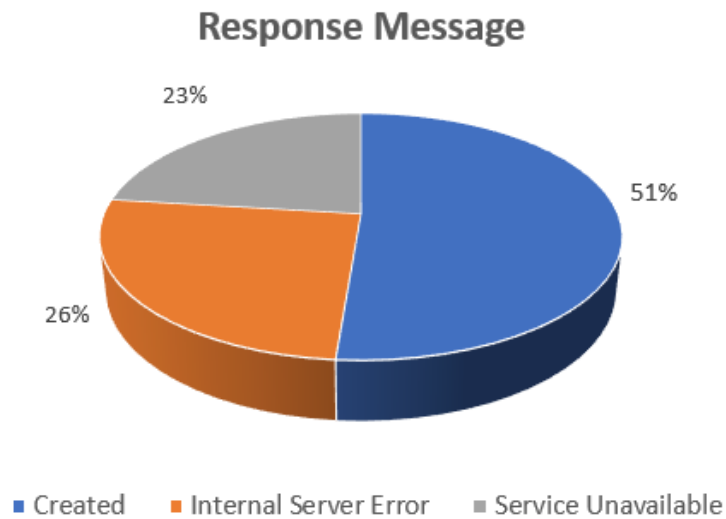


Porcentaje Cumplimiento tiempo de respuesta – Despliegue Previo

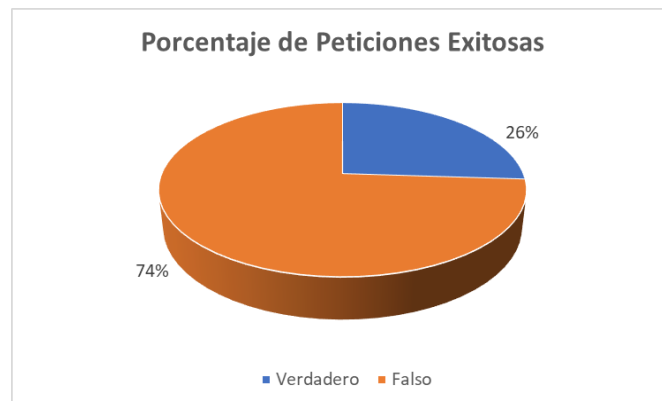
Porcentaje de Cumplimiento Requerimiento Tiempo de Respuesta



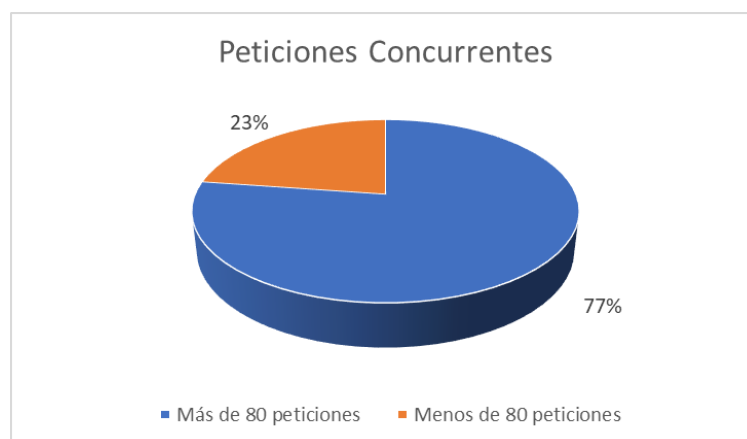
Tipos de mensajes de respuestas obtenidos durante la prueba- Despliegue Previo



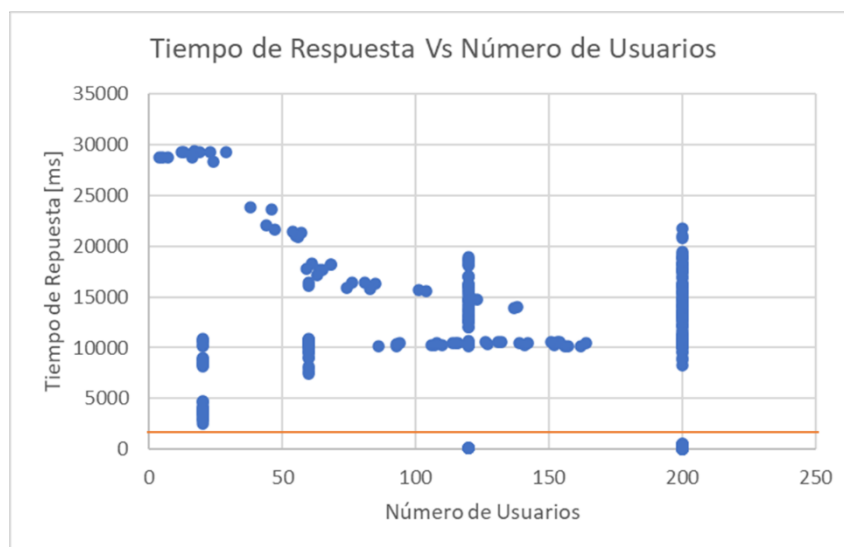
Porcentaje de Peticiones Exitosas – Despliegue Actual



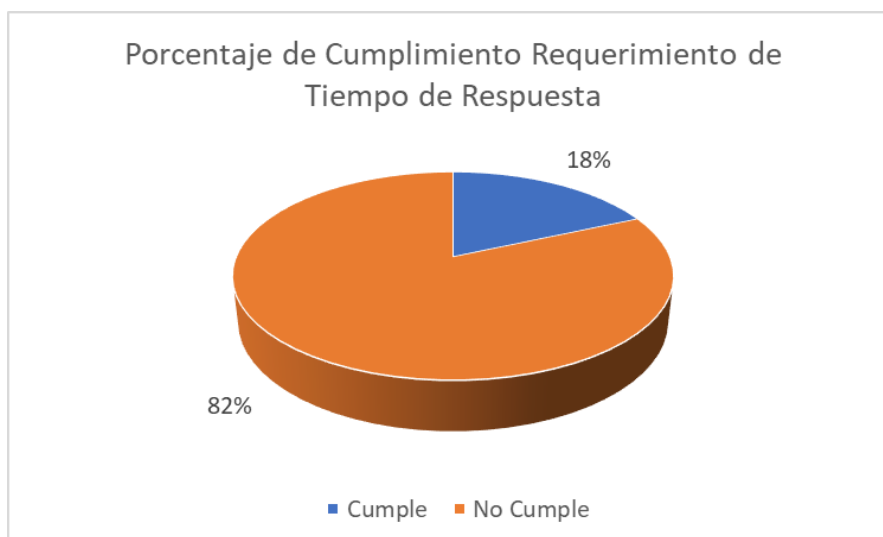
Peticiones Concurrentes – Despliegue Actual



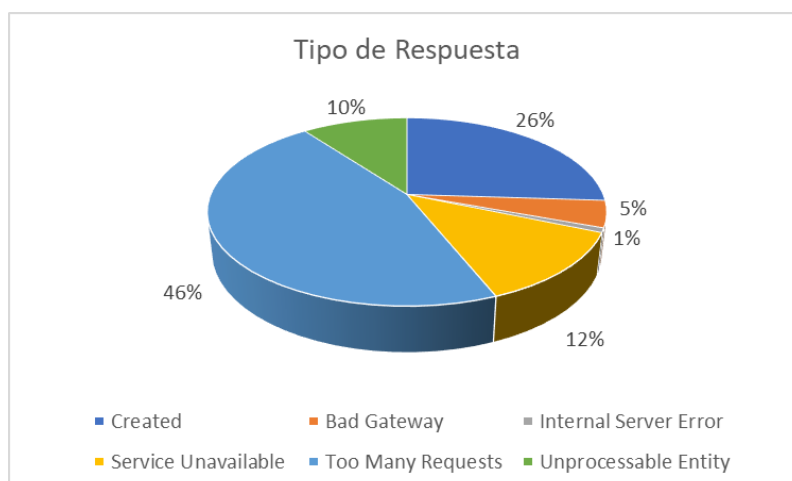
Grafica Tiempo de Respuesta Vs Usuarios – Despliegue Actual



Porcentaje Cumplimiento tiempo de respuesta – Despliegue Actual



Tipos de mensajes de respuestas obtenidos durante la prueba- Despliegue Actual



Análisis de Resultados Escenario 1

Los resultados obtenidos muestran que el 26% de las peticiones realizadas durante la prueba correspondiente al escenario 1 tuvieron una respuesta exitosa. Al analizar los porcentajes del tipo de respuesta obtenidos durante la prueba se identifica que un número significativo de peticiones (46%) obtuvo como respuesta *Too Many Requests* esta respuesta asociada al número de peticiones concurrentes ya que como se indica en la descripción de Google Cloud Run¹ el número máximo de peticiones concurrentes que puede manejar por default es de aproximadamente 88. Lo anterior se comprueba con los resultados mostrados en la gráfica de peticiones concurrentes donde solo para el 23% de las peticiones realizadas el número de peticiones concurrentes era menor a 80.

En relación al porcentaje de peticiones que cumplen con el tiempo de respuesta requerido (18%) este comportamiento se asocia a un known issue² registrado por el equipo de Cloud Run referente a la alta latencia que presenta el servicio.

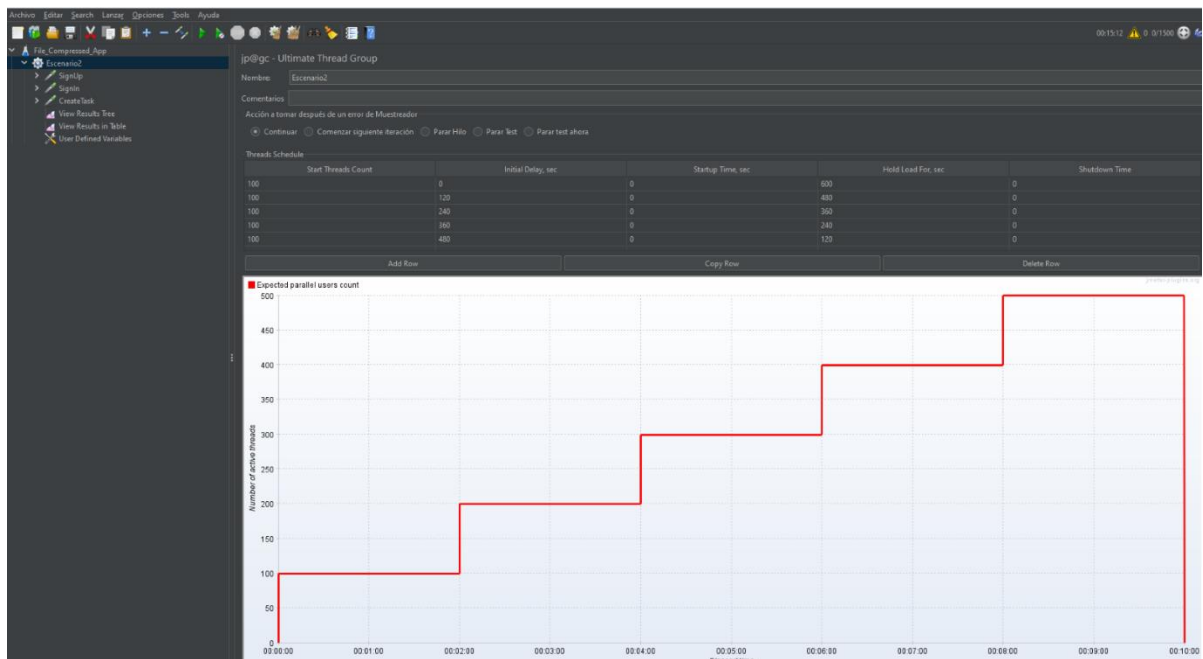
Escenario 2. Se deberá definir un escenario donde se pueda probar cuál es la máxima cantidad de archivos que pueden ser procesados por minuto en la aplicación local. Para hacer pruebas de estrés se recomienda utilizar la herramienta Apache Bench (ab) o JMeter. Las pruebas de estrés deberán realizarse desde otro equipo diferente a los utilizados para ejecutar el servidor web y el servidor de base de datos. El escenario y los resultados de las pruebas de estrés deberán ser documentados con gráficas que ilustran cómo se comporta la aplicación a medida que el número de usuarios procesando archivos se incrementa, hasta llegar al punto en que el tiempo para iniciar el procesamiento de un archivo enviado por un usuario supere los 10 minutos (600 segundos). Restricciones del escenario. El archivo enviado a convertir durante las pruebas debe ser de un tamaño mínimo de 15 MiB y un máximo de 20 MiB.

Para la realización de esta prueba, teniendo en cuenta la restricción que presenta Cloud Run referente al número de peticiones concurrentes, se decidió modificar la rampa de carga utilizada en el escenario ya que al usar la rampa usada en pruebas anteriores el 99,5% de las peticiones registraban error de tipo *Too Many Requests*.

¹ <https://cloud.google.com/run/docs/about-concurrency?hl=es-419#:~:text=By%20default%20each%20Cloud%20Run,can%20lower%20the%20maximum%20concurrency.>

² <https://cloud.google.com/run/docs/issues?hl=es-419>

Rampa usada en pruebas previas.



Errores de tipo *Too Many Requests* registrados al intentar usar la misma carga en Cloud Run

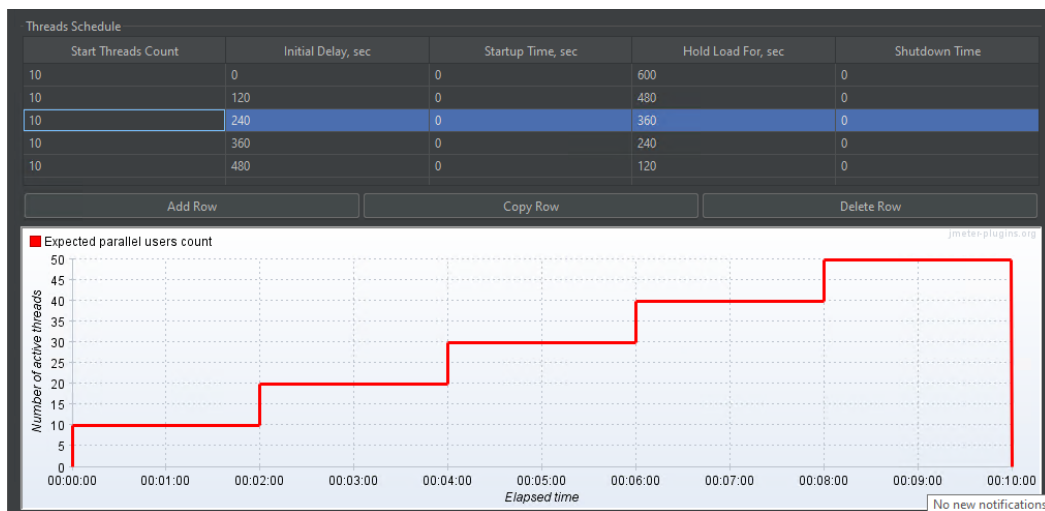
The screenshot shows the Apache JMeter results view. The 'View Results Tree' is expanded, showing a list of test results. The 'Sampler result' for the 'Request' is highlighted, showing the following details:

- Thread Name: Escenario2 1-9
- Sample Start: 2023-05-26 18:00:13 UTC
- Load time: 229
- Connect Time: 0
- Latency: 229
- Size in bytes: 269
- Sent bytes: 260
- Headers size in bytes: 255
- Body size in bytes: 14
- Sample Count: 1
- Error Count: 1
- Data type ("text"|"bin"|""): text
- Response code: 429
- Response message: Too Many Requests

The 'HTTPSampleResult fields' section shows:

- ContentType: text/html
- DataEncoding: null

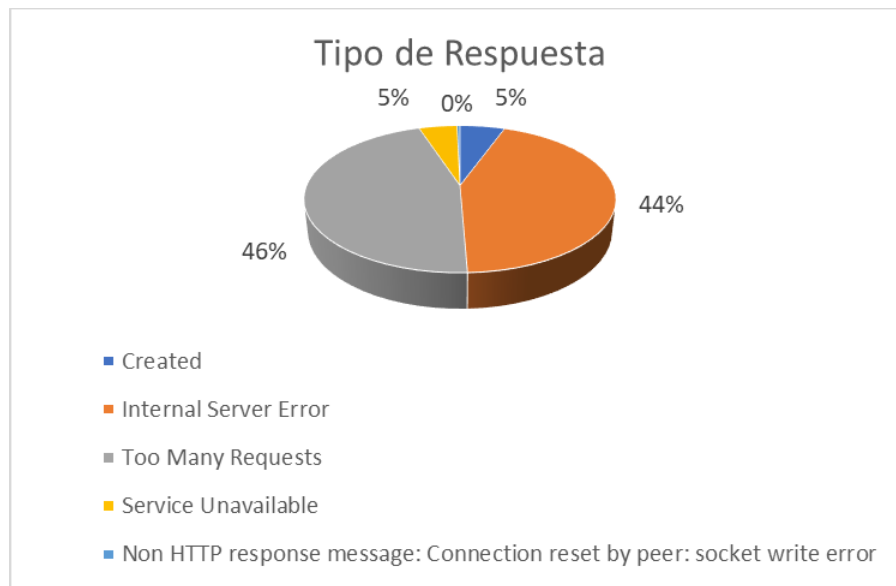
Rampa Modificada para el Escenario 2 (Despliegue PaaS)



A continuación, se presentan los resultados obtenidos para las pruebas realizadas en el Despliegue Previo Alta disponibilidad del Web Server + Servicio de Mensajes Cloud Pub/Sub + Worker con Autoscaling y la Entrega Actual (Despliegue PaaS).

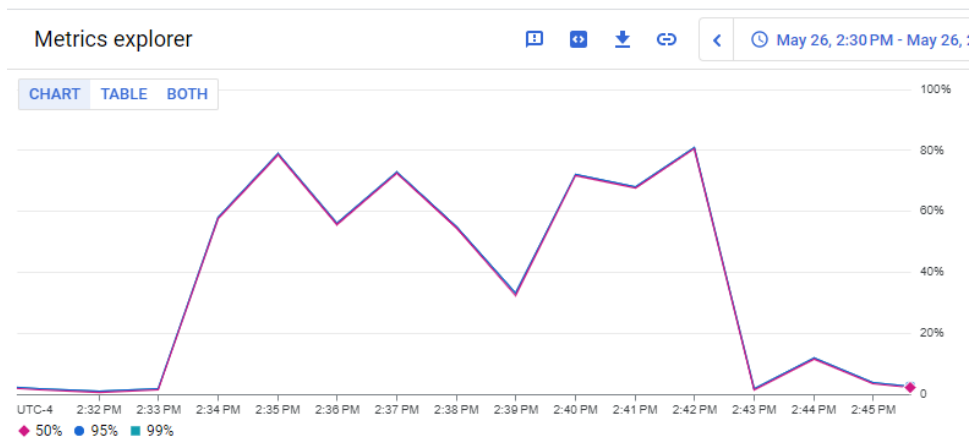
Para las pruebas de este escenario *Despliegue Actual* se realizó la misma metodología usada en la entrega anterior. Una vez se corrieron las pruebas en JMeter se esperó 10 minutos y se hizo una consulta a la base de datos para obtener todas las tareas creadas hasta el momento con su respectivo estado.

Inicialmente se realizó una clasificación de las respuestas obtenidas durante el escenario. Los resultados de esta clasificación se muestran en la siguiente imagen:



Debido a que un porcentaje considerable de las respuestas no fueron satisfactorias (90%) se realizó una revisión de las métricas de la utilización de CPU y Memoria del container con el propósito de identificar posibles causas de dicho comportamiento.

Container CPU Utilización



Container Memory Utilization



Referente a las tareas creadas, se realizó una consulta a la base de datos a través de Postman usando en el endpoint de Tasks, la respuesta de la petición fue exportada en formato Json. Durante la prueba del escenario 2 se crearon 61 tareas de conversión, al realizar la consulta a la base de datos todas las tareas se encontraban en estado *Completed*.

Consultas [1]

response_Final

✕ ✓ f/x = Origen{1564}

fecha_modificacion	2023-05-26T18:43:40.566478
id	1590
usuario	1
extension_final	Record
estado_conversion	Record
usuario_id	1
nombre_archivo	ArchicoTets_Cloud.pptx
estado_tarea	Record
fecha_creacion	2023-05-26T18:43:24.340756

< >

llave: COMPLETED

valor: 3

Análisis de Resultados

Aunque la rampa de carga para el escenario 2 fue modificada para poder ajustar al escenario a las restricciones de Cloud Run referentes al número de peticiones concurrentes los resultados obtenidos muestran que el porcentaje de peticiones exitosas fue bajo (5%).

La revisión de la utilización de CPU del Container permite identificar que durante la ejecución de la prueba del escenario 2 estos valores oscilan entre el 60% y 80% durante la mayor parte de la prueba. Esta podría ser la causa de la falla en las peticiones.

Otra posible causa asociada al comportamiento evidenciado puede ser un known issue³ reportado por el equipo de Cloud Run de acuerdo al cual puede presentarse baja disponibilidad en el servicio cuando se usan 3 instancias o menos.

³ <https://cloud.google.com/run/docs/issues?hl=es-419>

Conclusiones

- Al diseñar despliegues en Cloud Run es importante considerar el número de peticiones concurrentes que una instancia puede manejar para evitar fallas en el servicio que pueden afectar la experiencia del usuario.
- De acuerdo a los known issues registrados por el equipo de Cloud Run sería recomendable considerar tener más de 4 instancias especialmente en escenarios donde la disponibilidad sea considerada un factor crítico para el diseño del despliegue.
- Aunque la cantidad máxima predeterminada de solicitudes simultáneas por instancia es de 80, Cloud Run permite configurar la cantidad máxima. Configurar este valor de simultaneidad permite que Cloud Run restrinja el envío de solicitudes a una instancia en caso de que la CPU de la instancia presente un alto uso.
- Para el caso del escenario 2 se infiere que el alto uso de CPU está asociado al tamaño del archivo enviado (20 MB) ya que al comparar los resultados obtenidos con el escenario 1 se identifica que para dicho escenario los valores de consumo de CPU oscilan entre 10% y 20% (con una rampa que incluye un mayor número de peticiones).