

Entrega 4 -ESCALABILIDAD DISEÑO E IMPLEMENTACIÓN DE UNA APLICACIÓN WEB ESCALABLE EN NUBE PÚBLICA

Grupo 3

C. Camilo Baquero Gómez, Franklin A. Pinto Carreño, Julian Yamid Torres Torres

Desarrollo de Aplicaciones Cloud

Universidad de los Andes, Bogotá, Colombia

c.baquero@uniandes.edu.co, f.pintoc@uniandes.edu.co, jy.torres@uniandes.edu.co

Fecha de presentación: Mayo 7 de 2023

LINK APLICACIÓN WEB: <http://comprimemelo.com:5000/>

Github: <https://github.com/camilooob/comprimemelo.com>

1. Arquitectura de Aplicación

La aplicación web de compresión de archivos se encuentra implementada bajo el modelo desacoplamiento utilizando una instancia web que se encarga de desplegar el front en un compute engine, y procesar los archivos de compresión en un Worker comunicado mediante una cola de Pubsub, el front y un worker que procesa la compresión de archivos.

Cuando el server web o worker supera un uso de cpu en un 60% se activa la regla de autoscaling y despliega un nodo adicional y lo agrega al Balanceador de Carga. El modelo implementa un conjunto de métodos para crear, modificar, eliminar, consultar, comprimir y descomprimir archivos y tiene acceso directo al motor de persistencia.

La vista está implementada en formato html para los formularios y páginas de presentación en capa web, y para las api rest, se utiliza el formato json, para capturar y responder las peticiones web. El controlador es el intermediario entre el modelo y la vista para interpretar las peticiones y entregar una respuesta a cada petición web realizada por un usuario.

Diagrama de arquitectura

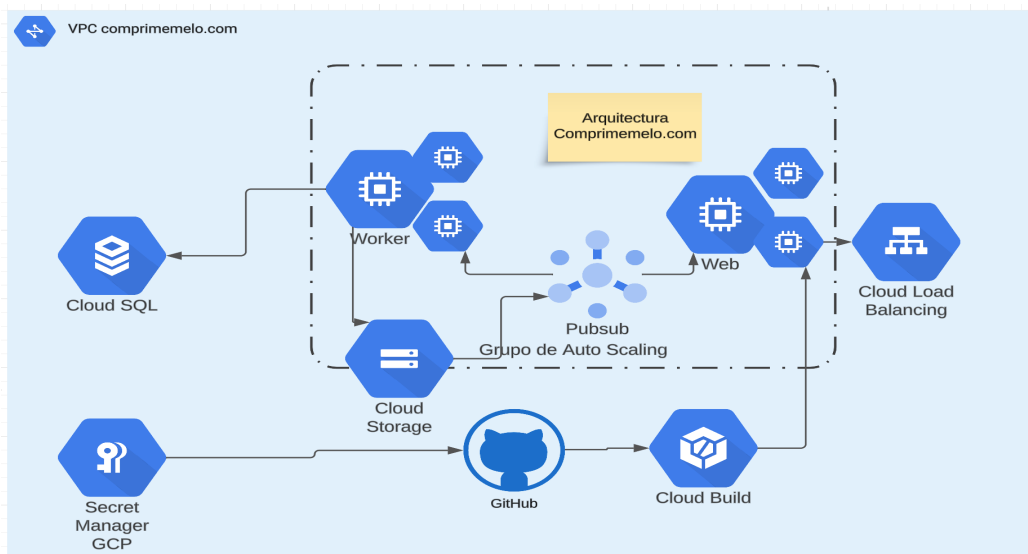


Figura 1. Diagrama de arquitectura aplicación de compresión de archivos

La arquitectura está diseñada para tener un despliegue de la aplicación para aplicaciones escalables sobre GCP, eso incluye el uso de un balanceador de carga hasta de 3 servidores web, políticas de autoscaling, además de una configuración en el servicio de balanceo de carga para poder desplegar en varios servidores web.

Nosotros definimos una estrategia para que los servidores web escalan de manera automática cuando estos alcancen el 60% de cpu , activamos el modo de autoscaling.

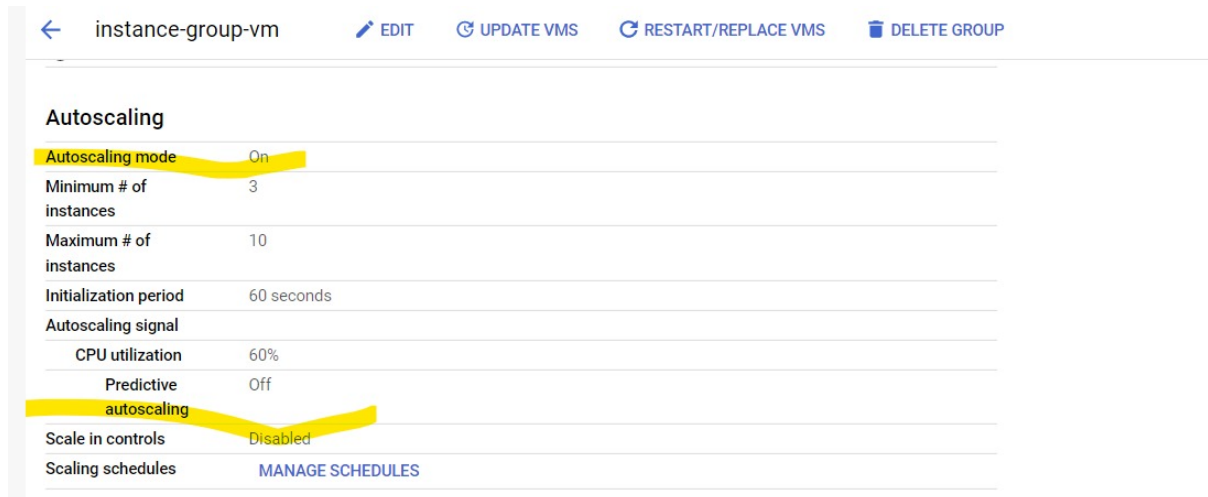
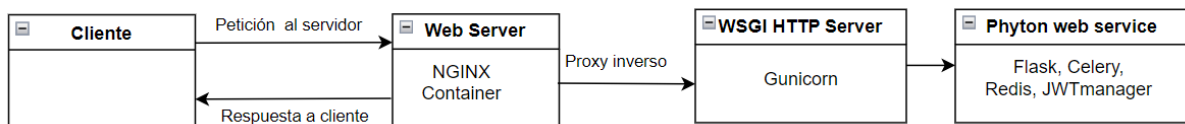


Diagrama de despliegue



La aplicación está corriendo en el dominio www.comprimemelo.com:5000 , desarrollada en Flask la cual está dividida en 3 VM de Compute Engine de la siguiente manera: Worker, Web Server y Cloud Storage. La aplicación Flask (VM Compute Engine Web Server) interactúa con la base de datos Mysql Cloud SQL, el Worker procesa las colas de pubsub de los archivos y Cloud Storage guarda los archivos.



• [Dar clic aquí para autenticarse](#)

Login Registrarme

Login

Username

Password

Login

- **Instancias generadas con el autoscaling:**
- En los buckets se crearon los siguientes ,con una clase de almacenamiento estándar, de esa manera es posible almacenar todos los archivos subidos por los usuarios, tanto los originales como los procesados.

Instancias de VM

Filtro Ingresar el nombre o el valor de la propiedad

<input type="checkbox"/>	Estado	Nombre ↑	Zona	IP interna	IP externa	Red	Conectar
<input type="checkbox"/>	✓	instance-group-front-hsw8	us-central1-c	10.128.0.4 (nic0)	34.31.113.22 (nic0)	comprimemelo-vpc	SSH ▾ ⋮
<input type="checkbox"/>	✓	instance-group-front-jc1p	us-central1-f	10.128.0.2 (nic0)	34.136.199.73 (nic0)	comprimemelo-vpc	SSH ▾ ⋮
<input type="checkbox"/>	✓	instance-group-front-zj2w	us-central1-b	10.128.0.3 (nic0)	35.239.170.215 (nic0)	comprimemelo-vpc	SSH ▾ ⋮
<input type="checkbox"/>	✓	storage-comprimemelo	us-east1-b	storage-comprimemelo-ip (10.142.0.2) (nic0)	35.196.127.54 (nic0)	comprimemelo-vpc	SSH ▾ ⋮
<input type="checkbox"/>	✓	web-comprimemelo	us-west1-b	comprimemelo-ip (10.138.0.2) (nic0)	35.197.22.141 (nic0)	comprimemelo-vpc	SSH ▾ ⋮
<input type="checkbox"/>	✓	worker-comprimemelo	us-west4-a	10.182.0.2 (nic0)	34.125.144.116 (nic0)	comprimemelo-vpc	SSH ▾ ⋮

Acciones relacionadas

[🔍 Buscar por nombre de instancia](#) | [📄 Consultar el informe de](#) | [🔧 Crear una VM](#) | [🔗 Conectar las instancias de VM](#) | [🔧 Crear una](#)

En esta parte nosotros colocamos el siguiente script para iniciar la app , cuando se crea un nuevo nodo con el autoscaling . Es un archivo que contiene comandos que se ejecutan cuando se inicia una instancia de máquina virtual, permitiendo expandir la red, discos, y

administración de los metadatos.

Enlaces de comprime de los

metadatos

Metadatos personalizados

Clave	Valor
startup-script	#!/bin/bash nohup /usr/bin/python3 -m flask --app /root/comprimemelo.com/main --debug run --host=0.0.0.0 --port 5000 & disown

REST EQUIVALENTE

- Pubsub y Secret manager configurado

En el Cloud Pub/Sub utilizamos uno, con el Topic ID pubsubcomprimemelofiles, con el topic name projects/datacompressionprojectfjc/topics/pubsubcomprimemelofiles, de esa forma nosotros diseñamos el sistema de comunicación entre los servidores web y los procesos workers permitiendo que se comunicaran entre ellos. Esto es muy importante porque permite que se añadan o se creen las diferentes solicitudes para procesar nuevos archivos y los workers pueden procesar dicha cola.

Topics [+ CREATE TOPIC](#) [DELETE](#)

LISTMETRICS

Filter Filter topics

Topic ID

↑

Encryption key

Topic name

Retention

pubsubcomprimemelofiles

Google-managed

projects/datacompressionprojectfjc/topics/pubsubcomprimemelofiles

7 days

A continuación podemos ver como se encuentra configurado el servicio de pub/sub

pubsubcomprimemelofiles

EDITAR

+ ACTIVAR CLOUD FUNCTION

IMPORTAR

BORRAR

Las opciones de exportación se trasladaron al menú desplegable Crear suscripción en la pestaña Suscripciones que aparece a continuación.

ENTENDIDO

Nombre del tema

projects/datacompressionprojectfjc/topics/pubsubcomprimemelofiles

Exportar a BigQuery

Exportar datos a una tabla de BigQuery.

EXPORTAR A BIGQUERY

Cómo exportar a Cloud Storage

Crea un trabajo de Dataflow para exportar datos a texto o a un archivo de Avro en Cloud Storage.

EXPORTAR A TEXTOEXPORTAR A AVRO

SUSCRIPCIONES

INSTANTÁNEAS

MÉTRICAS

DETALLES

MENSAJES

Solo se muestran las suscripciones vinculadas a este tema. Una suscripción captura la transmisión de mensajes publicados en un tema específico. También puedes transmitir mensajes a BigQuery o Cloud Storage creando una suscripción desde un trabajo de Cloud Dataflow. [Más información](#)

CREAR SUSCRIPCIÓN

EXPORTAR

Filter Filtrar las suscripciones

ID de la suscripción

Nombre de la suscripción

Proyecto

pubsubcomprimemelofiles-sub

projects/datacompressionprojectfjc/subscriptions/pubsubcomprimemelofiles-sub

datacompressionprojectfjc

pubsubcor

PERMIS

Edita o borra los aparecen a con selecciona "Agi para otorgar aco

Mostra

Filtro

Ingr

Función/Principa

Agente de s

Agente de s

Agente de s

Agente de s

Cuenta de s

Editor (3)

Propietario

datacompressionprojectfjc

secret

Detalles del secreto

EDITAR SECRETO
 BORRAR

Secret: "secrets_comprimemelo"

projects/208785026947/secrets/secrets_comprimemelo

DESCRIPCIÓN GENERAL

VERSIONES

PERMISOS

REGISTROS

Versiones

+ VERSIÓN NUEVA

HABILITAR

INHABILITAR

DESTRUIR

Filtro

Ingresar el nombre o el valor de la propiedad

<input type="checkbox"/>	Versión	Alias	Estado	Encriptación	Fecha de creación ↓	Acciones
<input type="checkbox"/>	2	-	Habilitada	Administrada por Google	7/5/23, 21:44	
<input type="checkbox"/>	1	-	Habilitada	Administrada por Google	7/5/23, 19:44	

No se seleccionaron versiones

- La base de datos Mysql conectada a la VPC interna y recibiendo los registros de la VM.

SQL	Instances	CREATE INSTANCE	MIGRATE DATA	HELP ASSISTANT	SHOW INFO PANEL					
Filter Enter property name or value										
<input type="checkbox"/>	Instance ID	Type	Public IP address	Private IP address	Instance connection name	High availability	Location	Storage used	Labels	Actions
<input type="checkbox"/>	mysql-comprimemelo	MySQL 8.0	34.31.69.88		datacompressionprojectfjc	ADD	us-central1-f	1 GB of 100 GB		

VPC networks

Filter Enter property name or value							
Name	Subnets	MTU	Mode	Internal IP ranges	Gateways	Firewall rules	Global dynamic routing
comprimemelo-vpc	38	1460	Auto			5	On
default	38	1460	Auto			5	Off

Configuración del balanceador de carga de instancia web

El balanceador de carga de red nos permite a nosotros colocar un puerto de origen y de destino, una dirección IP , como también asignar diferentes protocolos para el reenvío de paquetes, de esa forma podemos mejorar el resentimiento de la web

← Detalles del balanceador de cargas

EDITAR

BORRAR

lb-front

Rendimiento web más rápido y mayor protección web con Cloud CDN y Cloud Armor. [Más información](#)

DETALLES

MONITORING

ALMACENAMIENTO EN CACHE

Frontend

Protocolo	IP:Puerto	Certificado	Política de SSL	Nivel de red
HTTP	34.160.219.94:80	-	-	Premium

Normas de enrutamiento

Hosts	Rutas	Backend
Todos los que no coincidan (predeterminado)	Todos los que no coincidan (predeterminado)	back-comp

Backend

Servicios de backend

1. back-comp

Protocolo de extremo	Puerto con nombre	Tiempo de espera	Verificación de estado	Cloud CDN	Registros
HTTP	http	30 segundos segundos	http-web	Habilitada VER DETALLES DE CDN	Inhabilitada

CONFIGURACIÓN AVANZADA

Nombre	Tipo	Alcance	En buen estado	Ajuste de escala automático	Modo de balanceo	Puertos seleccionados	Capacidad
instance-group-front	Grupo de instancias	us-central1	3 de 3	Activado: Objetivo Uso de CPU 60 %	Utilización máxima del backend: 80%	80, 8000, 8080	100%

Plantilla de replicación de instancia web

← instance-template-image-front-2

CREAR VM

CREAR UNA SIMILAR

CREAR GRUPO DE INSTANCIAS

BORRAR

En uso por

[instance-group-front](#)

Reservas

Elegir automáticamente

Etiquetas

Ninguna

Política de posición

No hay políticas

Servicio Confidencial VM

Inhabilitado

Configuración de la máquina

Tipo de máquina	e2-medium
Plataforma de CPU mínima	Ninguna
Arquitectura	—
CPU virtuales para proporción de núcleos	—
Núcleos visibles personalizados	—
Dispositivo de visualización	Inhabilitado
GPU	Ninguna

Redes

Registro PTR del DNS público	Ninguna
Nivel total de ancho de banda de salida	—
Tipo de NIC	—

Firewalls

Tráfico HTTP	Activo
Tráfico HTTPS	Activo

Etiquetas de red

http-server

https-server

Interfaces de red

Nombre	Red	Subred	Dirección IP interna principal	Rangos de alias de IP	IP stack type	Dirección IP externa
nic0	comprimemelo-vpc	—	—	—	IPv4	Efímera

Almacenamiento

Disco de arranque

Nombre	Imagen	Tipo de interfaz	Tamaño (GB)	Nombre del dispositivo	Tipo	Arquitectura	Encriptación
Autogenerated	img-front	—	10	instance-template-image-vm	Disco persistente equilibrado	—	Administrada por Google

La plantilla nos permite obtener una copia de la imagen real de la instancia donde se encuentra configurada la aplicación web, la cual es usada en el grupo de instancias para la replicación en el autoescalamiento.

Adicionalmente con el uso de la plantilla podemos definir el tipo de máquina, imagen del disco de arranque o del contenedor, como también establecer una secuencia de comandos

de inicio. El uso de plantillas de instancias no está vinculado a una zona o región, sin embargo es recomendable especificar cuáles recursos serán utilizados en algunas zonas en específico.

Configuración de grupo de instancias

instance-group-front

EDIT

UPDATE VMs

RESTART/REPLACE VMs

DELETE GROUP

DESCRIPCIÓN GENERAL

DETALLES

SUPERVISIÓN

ERRORES

Instancias por condición

3 instancias

3

Instancia por estado

No configurada

Reparación automática desactivada

Configurar

Ajuste de escala automático

Activo (min 3, max 10)

Based on 1 métrica and 0 programas

Status

Ready

Creation Time

may 7, 2023, 11:15:32 p. m. UTC-05:00

Description

Number of instances

3

Template

instance-template-image-front-2

Location

us-central1 (3/4)

In use by

back-comp

Miembros del grupo de instancias

QUITAR DEL GRUPO

BORRAR INSTANCIA

Filtro

Ingrese el nombre o el valor de la propiedad

Estado	Nombre	Fecha/hora de creación	Plantilla	Zona	Configuración por instancia	IP interna	IP externa	Estado de la verificación de estado	Conectar
<input type="checkbox"/>	instance-group-front-8w6	may 7, 2023, 11:15:47 p. m. UTC-05:00	instance-template-image-front-2	us-central1-c		10.128.0.4	34.91.113.224		SSH
<input type="checkbox"/>	instance-group-front-2jp	may 7, 2023, 11:15:48 p. m. UTC-05:00	instance-template-image-front-2	us-central1-f		10.128.0.2	34.136.199.73		SSH
<input type="checkbox"/>	instance-group-front-q2w	may 7, 2023, 11:15:48 p. m. UTC-05:00	instance-template-image-front-2	us-central1-b		10.128.0.3	35.239.170.215		SSH

El grupo de instancias se basa en la plantilla que previamente se creó y se definen la cantidad mínima de instancias, ciclo de vida de instancias, configuración autoscaling y ubicación de las nuevas máquinas virtuales.

Pruebas de Escalabilidad

Prueba escenario 1. Listar todas las tareas de conversión de un usuario

El servicio entrega el identificador de la tarea, el nombre y la extensión del archivo original, a qué extensión desea convertir y si está disponible o no. El usuario debe proveer el token de autenticación para realizar dicha operación.

Comprimemelo.com / Task List

GET http://comprimemelo.com:5000/api/tasks

Params

Authorization

Headers (8)

Body

Pre-request Script

Tests

Settings

Type

Bearer Token

Heads up! These parameters hold sensitive data. To keep this data secure while working in a collaborative environment, we've masked some values.

The authorization header will be automatically generated when you send the request.

Learn more about authorization

Token

eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJ1b250aW50IjpbImFkbGciXSwiInRhdia...

Body

Cookies

Headers (5)

Test Results

Status: 200 OK

Pretty

Raw

Preview

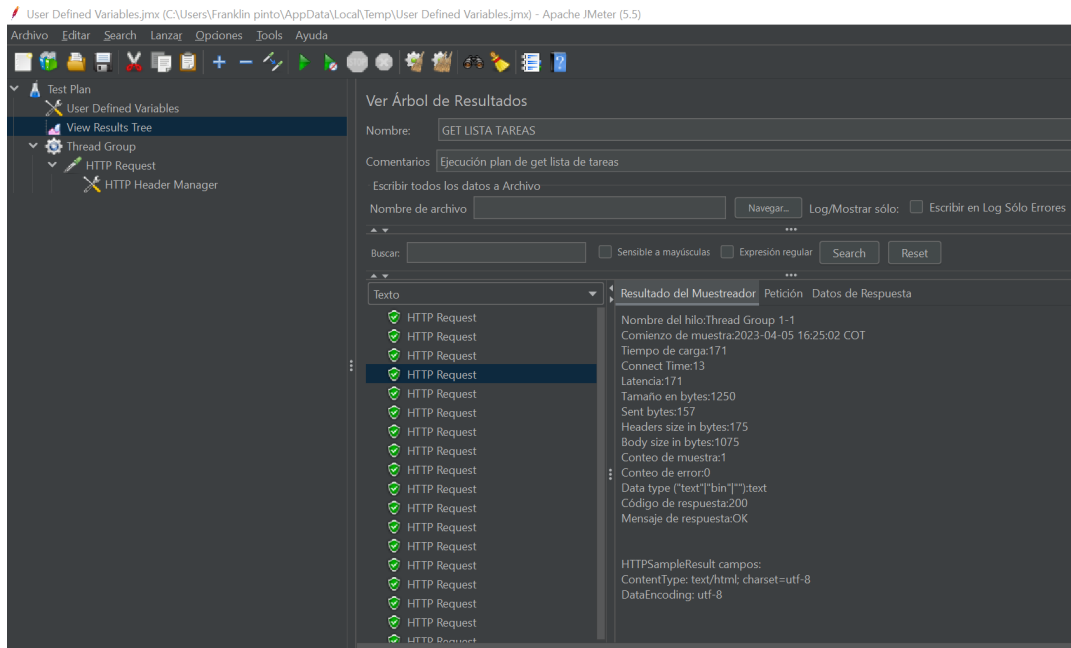
Visualize

JSON

1

{\"tasks\": {\"id\": \"1\" \"original_file\": \"pdf\", \"format\": \"rar\", \"status\": \"processed\"}}

Se ejecuta el plan de pruebas desde JMeter



Con 100,500 y 1000 peticiones en concurrencia no falló ninguna petición mientras que al usar 2000 comenzó a presentar fallos por timeout. Sin embargo, a nivel funcional se afirma que está dentro de los límites, dado que el requerimiento menciona que debe soportar 1000 peticiones concurrentes para este servicio.

Prueba escenario 2. Crear una cuenta de usuario en la aplicación

Para crear una cuenta se deben especificar los campos: usuario, correo electrónico y contraseña. El correo electrónico debe ser único en la plataforma dado que este se usa para la autenticación de los usuarios en la aplicación.

Para crear una cuenta de usuario es posible hacerlo a través de un formulario web ó consumiendo la api rest a través de un cliente http, ejemplo postman..

El siguiente endpoint con método post, permite crear una cuenta de usuario:
<http://comprimemelo.com:5000/api/auth/signup>

POST

http://comprimemelo.com:5000/api/auth/signup

Send

Params

Authorization

Headers (8)

Body

Pre-request Script

Tests

Settings

none

form-data

x-www-form-urlencoded

raw

binary

GraphQL

	KEY	VALUE	DESCRIPTION	...	Bulk Edit
<input checked="" type="checkbox"/>	email	franklin.pintoc@uniandes.edu.co			
<input checked="" type="checkbox"/>	first_name	Franklin			
<input checked="" type="checkbox"/>	last_name	Pinto			

Body

Cookies

Headers (5)

Test Results

201 CREATED

1845 ms

210 B

Save Response

Pretty

Raw

Preview

Visualize

JSON

```

1  {
2    "message": "User created successfully"
3  }

```

← → ↺

127.0.0.1:5000/auth/signup

🔍 ⚙️ ☆ 🌱 📺 📄

Cambiar Tema

Login Registrarme

Registrar Usuario

Nombre

Apellido

Username

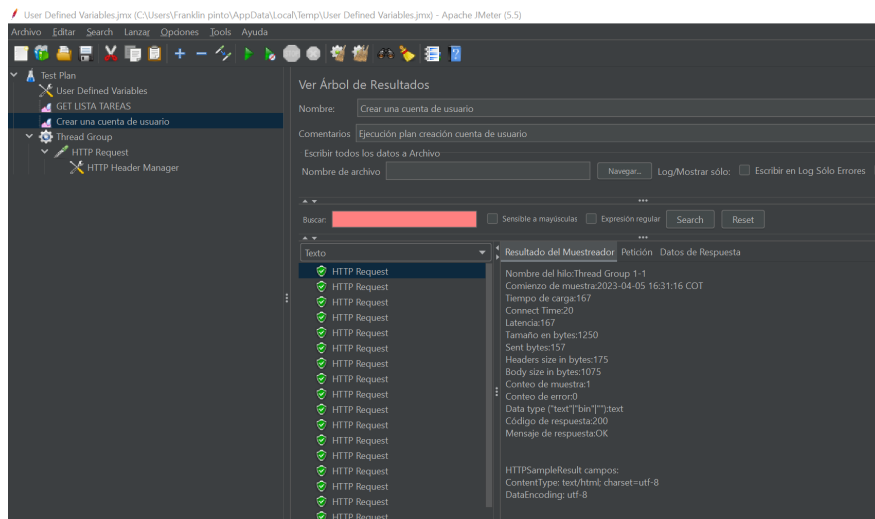
Correo

Celular

Nueva contraseña

Confirmación de contraseña

Registrarme



Para este escenario se identifica que al ejecutar 300 peticiones concurrentes el sistema comienza a responder con timeout, esto posiblemente se debe es por las validaciones de existencia del usuario, creación del mismo y envío de correo simultáneo. *Según los requerimientos del usuario se especifica que con 200 peticiones el sistema responde adecuadamente.*

Prueba escenario 3. Iniciar sesión en la aplicación web

El usuario provee el correo electrónico/usuario y la contraseña con la que creó la cuenta de usuario en la aplicación. La aplicación retorna un token de sesión si el usuario se autenticó de forma correcta, de lo contrario indica un error de autenticación y no permite utilizar los recursos de la aplicación.

Para iniciar sesión en la aplicación se debe autenticarse ingresando las credenciales de usuario a través de la siguiente url <http://comprimemelo.com:5000/api/auth/login>
Para obtener el token de autenticación de usuario a través de la api rest debe usar el siguiente endpoint con metodo POST: <http://comprimemelo.com:5000/api/auth/login>

5. Bibliografía

<https://docs.celeryq.dev/en/stable/>

<https://flask.palletsprojects.com/en/2.2.x/>

<https://flask-jwt-extended.readthedocs.io/en/stable/>