

Universidad de los Andes

Facultad de Ingeniería

Desarrollo de Soluciones Cloud

Plan de pruebas de carga

Grupo 1

Juan Sebastian Colmenares - 202312351

William Fernando Alarcón - 202423127

Jhon Mario Forero – 20123466

Contenido

Plan de Pruebas de Carga	3
1. Descripción de la Aplicación	3
2. Entorno de Pruebas	3
3. Criterios de Aceptación	3
4. Escenarios de Prueba	3
5. Parámetros de Configuración.....	4
6. Justificación de la Herramienta Seleccionada	4
7. Definición de los Recursos de la Instancia de Pruebas	4
8. Identificación y Explicación de Métricas Clave	5
Resultados pruebas de estrés	5

Plan de Pruebas de Carga

1. Descripción de la Aplicación

La aplicación consta de un backend en FastAPI y un frontend en Angular. El backend gestiona la carga, análisis y resumen de documentos con PostgreSQL y Redis. Se usa el modelo LLM web de Google (Gemini Flash 2.0) para el análisis de documentos. El frontend permite a los usuarios interactuar con la plataforma de análisis.

2. Entorno de Pruebas

Las pruebas de proyecto "LLM Analyze Documents" se ejecutarán en una plataforma implementada sobre Google Cloud Platform, la cual tiene los siguientes recursos:

- **Web Server** (Compute Engine - 4 vCPU, 4 GiB RAM, 20 GiB disco): Encargado de exponer la interfaz web y recibir peticiones REST API de los usuarios.
- **Worker Server** (Compute Engine - 4 vCPU, 4 GiB RAM, 50 GiB disco): Realiza procesamiento intensivo, embedding y chunking de los documentos.
- **File Server (NFS)** (Compute Engine - 2 vCPU, 2 GiB RAM, 20 GiB disco): Proporciona almacenamiento compartido para los documentos subidos.
- **DB Server** (PostgreSQL - 2 vCPU, 2 GiB RAM, 20 GiB disco): Guarda metadatos, logs y resultados del análisis de documentos.

Gemini Flash 2.0 API: LLM externo que recibe los chunks de texto para su análisis

3. Criterios de Aceptación

- Tiempo de respuesta óptimo para cargas estándar.
- Capacidad de procesamiento adecuada para múltiples solicitudes concurrentes.
- Uso eficiente de los recursos de hardware.

4. Escenarios de Prueba

Se han definido escenarios representativos que cubren rutas críticas de usuario tanto en la capa web como en el procesamiento por lotes. La variedad y realismo de los escenarios garantizan una evaluación precisa del rendimiento del sistema.

Capa Web:

- **Autenticación:**
 - Registro de usuario (/api/v1/auth/register)
 - Inicio de sesión (/api/v1/auth/login)
- **Gestión de Documentos:**
 - Carga de documentos (/api/v1/documents/upload)
 - Listado de documentos (/api/v1/documents/listDocuments)

- **Análisis de Documentos:**
 - Resumen de documentos (/api/v1/analysis/summarize/{document_id})
 - Preguntas sobre documentos (/api/v1/analysis/ask/{document_id})
- **Conversaciones:**
 - Listado de conversaciones (/api/v1/conversations/list)
 - Detalle de conversación (/api/v1/conversations/byId/{conversation_id})
- **Monitoreo del servicio:**
 - Verificación del estado (/health)

Procesamiento por Lotes:

- Simulación de carga masiva en el backend para evaluar el impacto del procesamiento en lotes sobre el rendimiento del sistema.
- Análisis del tiempo de ejecución para múltiples peticiones simultáneas de resumen y preguntas sobre documentos.

5. Parámetros de Configuración

Las pruebas utilizarán Apache JMeter con la siguiente configuración:

- **Número de usuarios simulados:** 50-100
- **Duración de las pruebas:** 10 minutos por escenario
- **Monitoreo del uso de CPU y memoria**

6. Justificación de la Herramienta Seleccionada

Se ha elegido **Apache JMeter** por su adecuación al contexto del ejercicio y su facilidad de configuración para pruebas de carga y estrés. Entre sus ventajas destacan:

- Soporte para pruebas de múltiples protocolos (HTTP, WebSockets, JDBC, etc.).
- Facilidad para simular cargas concurrentes y obtener métricas detalladas.
- Integración con herramientas de monitoreo y generación de reportes.
- Flexibilidad para realizar pruebas en entornos locales y en la nube.

7. Definición de los Recursos de la Instancia de Pruebas

Para garantizar resultados representativos, se han definido los siguientes recursos:

- **CPU:** 4 núcleos para manejar múltiples hilos simultáneamente.
- **Memoria RAM:** 4 para soportar grandes volúmenes de peticiones sin degradación.
- **Disco SSD:** 512GB para optimizar tiempos de acceso a datos y almacenamiento temporal.

Estos recursos han sido seleccionados considerando la carga esperada y las pruebas de estrés que se realizarán en JMeter.

8. Identificación y Explicación de Métricas Clave

Para evaluar el rendimiento del sistema, se analizarán las siguientes métricas clave:

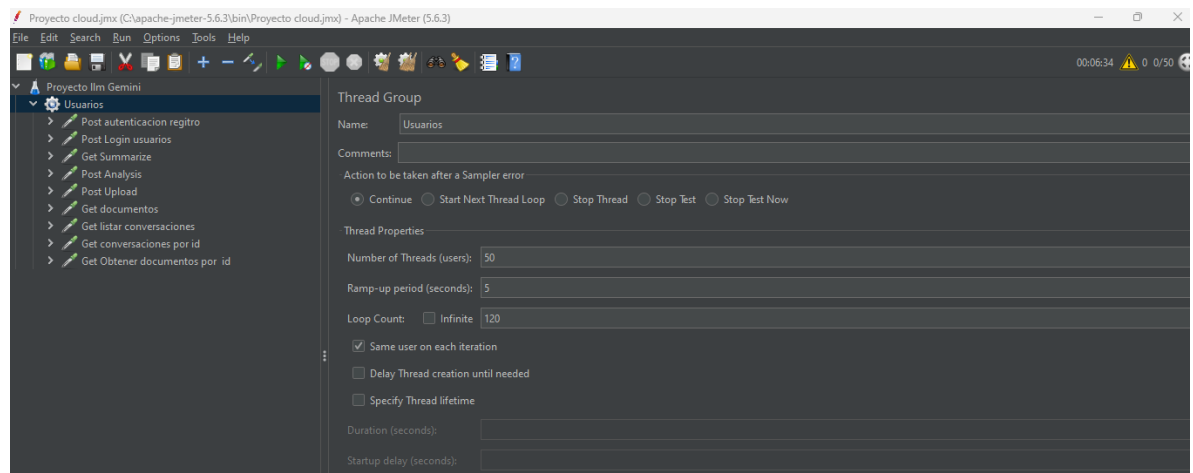
- **Throughput:** Número de peticiones procesadas por segundo. Indicador clave de la capacidad del sistema.
- **Response Time:** Tiempo promedio de respuesta de cada endpoint bajo diferentes niveles de carga.
- **Utilización de Recursos:** Consumo de CPU, memoria y disco durante las pruebas.
- **SLIs relevantes:** Se definirán umbrales aceptables de tiempo de respuesta y tasa de éxito de solicitudes para evaluar el cumplimiento de los acuerdos de nivel de servicio (SLAs).

Con este enfoque, se asegura una evaluación completa del rendimiento del sistema bajo distintos niveles de carga y escenarios de uso.

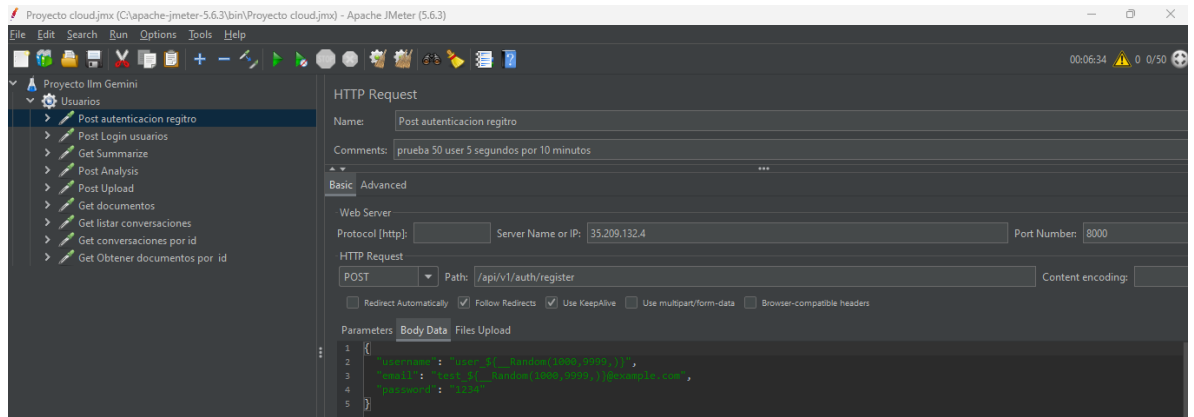
Resultados pruebas de estrés

Criterios de las pruebas de cada uno de los servicios:

50 peticiones cada 5 segundos durante 120 ciclos para un total de 6000 peticiones en 10 minutos.



- Registro de usuario:
<http://35.209.132.4:8000/api/v1/auth/register>
Prueba jmeter registro de usuarios, request



Resultados:

En la siguiente imagen se puede evidenciar que en 10 minutos hubieron 3027 registros exitosos, 2037 request rechazados por el servidor y 936 que no fueron fallas si no que los usuarios ya existían.

Projecto cloud.jmx (C:\apache-jmeter-5.6.3\bin\Projecto cloud.jmx) - Apache JMeter (5.6.3)

File Edit Search Run Options Tools Help

Projecto Ilim Gemini

- Usuarios
 - Post autenticacion registro
 - headers POST
 - Resultados en Arbol
 - Resultados en tabla
 - Post Login usuarios
 - Get Summarize
 - Post Analysis
 - Post Upload
 - Get documentos
 - Get listar conversaciones
 - Get conversaciones por id
 - Get Obtener documentos por id

View Results in Table

Name: Resultados en tabla

Comments:

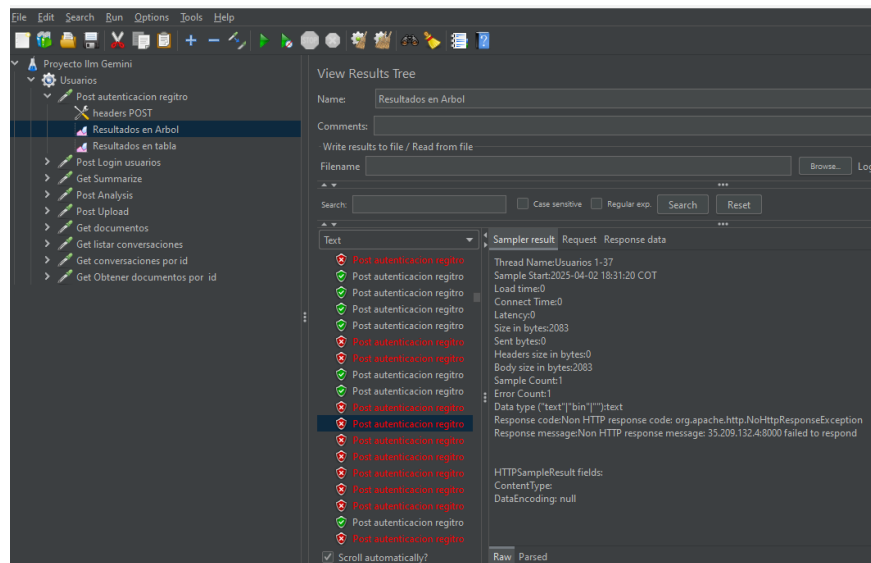
Write results to file / Read from file

Filename: Browse... Log/Display Only: ☐ Errors ☐ Successes ☐ Configure

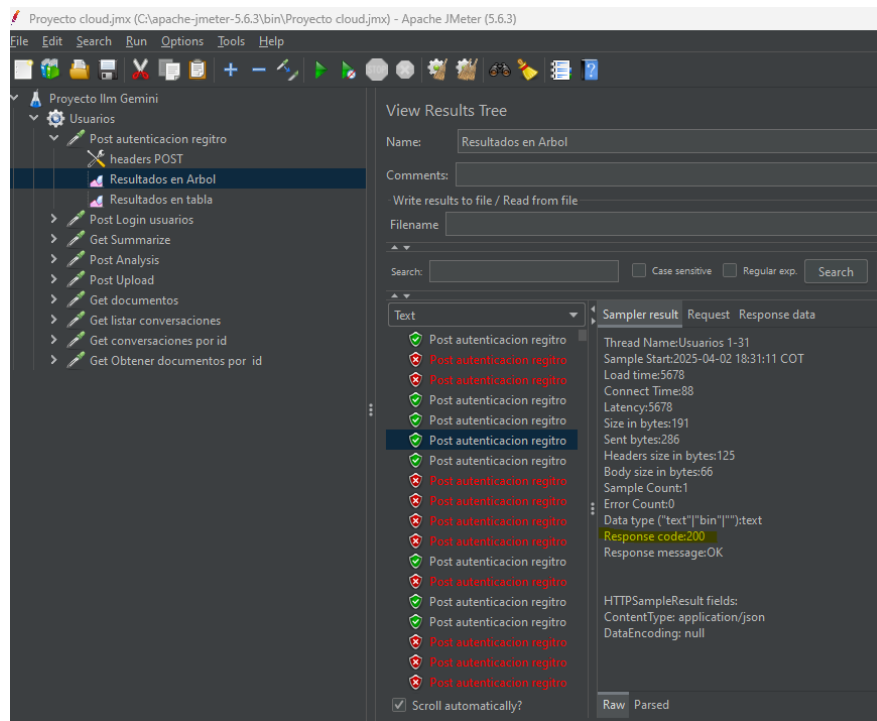
Sample #	Start Time	Thread Name	Label	Sample Tim...	Status	Bytes	Sent Bytes	Latency	Connect T...
5574	18.31.17.238	Usuarios 1-26	Post autenticacion registro	2864	Success	191	286	2864	0
5573	18.31.17.238	Usuarios 1-30	Post autenticacion registro	6563	Success	191	286	6563	0
5572	18.31.17.238	Usuarios 1-47	Post autenticacion registro	2850	Success	191	286	2850	0
5571	18.31.17.238	Usuarios 1-9	Post autenticacion registro	2820	Success	191	286	2820	0
5570	18.31.19.952	Usuarios 1-33	Post autenticacion registro	1	Error	2083	0	0	0
5569	18.31.17.630	Usuarios 1-33	Post autenticacion registro	2322	Error	174	286	2322	79
5568	18.31.18.484	Usuarios 1-6	Post autenticacion registro	1454	Error	166	286	1454	82
5567	18.31.17.237	Usuarios 1-22	Post autenticacion registro	2696	Error	181	286	2696	0
5566	18.31.18.484	Usuarios 1-45	Post autenticacion registro	1439	Error	166	286	1439	0
5565	18.31.18.483	Usuarios 1-23	Post autenticacion registro	1437	Error	166	286	1437	0
5564	18.31.16.116	Usuarios 1-35	Post autenticacion registro	3787	Success	191	286	3787	0
5563	18.31.16.977	Usuarios 1-5	Post autenticacion registro	2925	Success	191	286	2925	99
5562	18.31.16.117	Usuarios 1-3	Post autenticacion registro	3793	Success	191	286	3793	0
5561	18.31.17.238	Usuarios 1-31	Post autenticacion registro	2864	Success	191	286	2864	0
5560	18.31.19.829	Usuarios 1-32	Post autenticacion registro	0	Error	2083	0	0	0
5559	18.31.15.984	Usuarios 1-32	Post autenticacion registro	3781	Error	174	286	3781	79
5558	18.31.17.629	Usuarios 1-39	Post autenticacion registro	2997	Success	191	286	2997	0
5557	18.31.16.975	Usuarios 1-29	Post autenticacion registro	2751	Success	191	286	2751	101
5556	18.31.14.443	Usuarios 1-27	Post autenticacion registro	4312	Success	191	286	4312	80
5555	18.31.15.981	Usuarios 1-2	Post autenticacion registro	2850	Success	191	286	2850	0

☒ Scroll automatically? ☐ Child samples? No of Samples: 600 Latest Sample: 1 Success: 2037 Response: 937

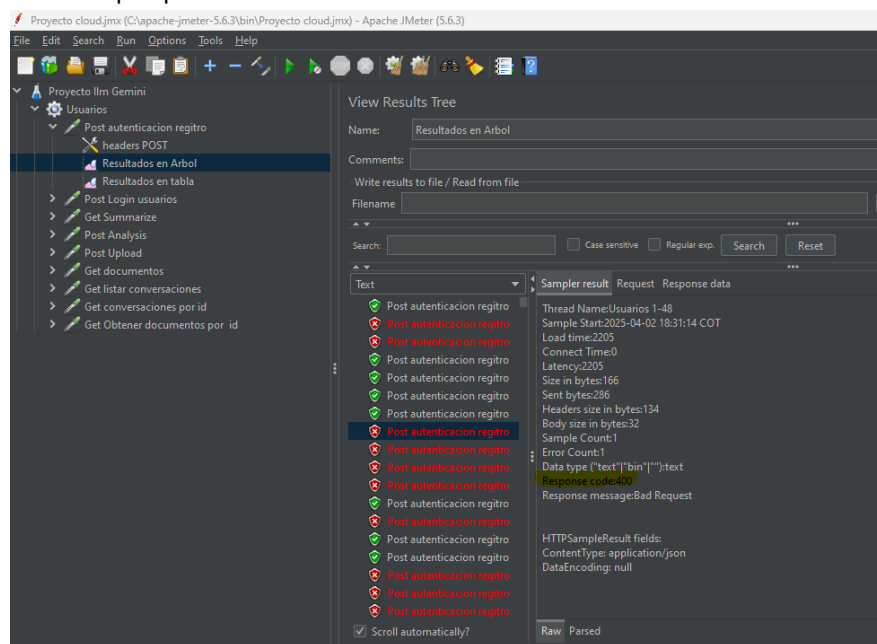
Respuesta servidor El error **org.apache.http.NoHttpResponseException** con el mensaje **"35.209.132.4:8000 failed to respond"** significa que el servidor no respondió a la solicitud de JMeter:



Respuesta exitosa como se muestra:



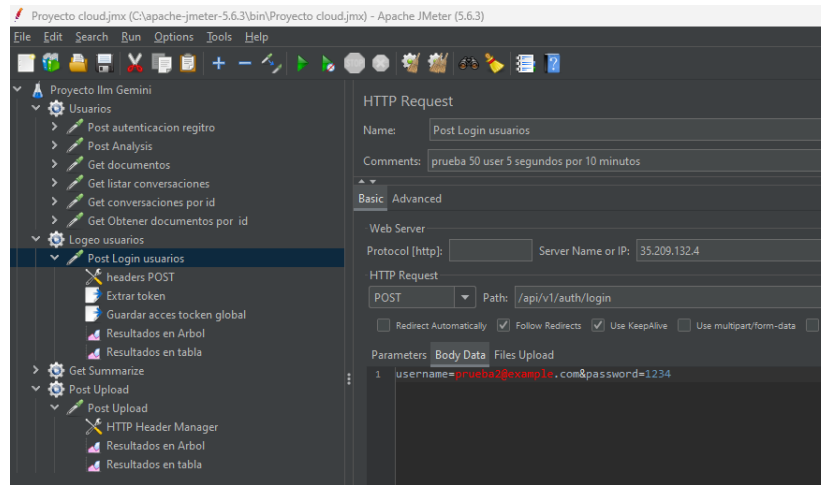
Respuesta de error, el servidor El error **400 - Bad Request** indica que el servidor rechazó la solicitud porque está malformada o tiene datos incorrectos.:



- Login de usuarios:

<http://35.209.132.4:8000/api/v1/auth/login>

Request que se ejecuto:

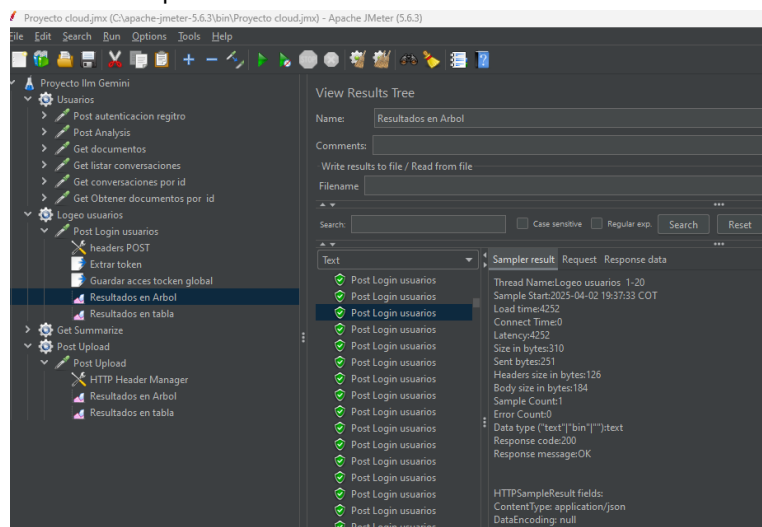


Se puede evidenciar que las 6000 peticiones fueron exitosas donde la petición no genero ningún error genero error , sin embargo hay una desviación de 500 ms lo cual indica que algunas peticiones tardan más en responder.

The screenshot shows the 'View Results in Table' window in Apache JMeter. The table displays the results of the 6000 'Post Login usuarios' requests. The columns are: Sample #, Start Time, Thread Name, Label, Sample Time, Status, Bytes, Sent Bytes, Latency, and Connect Time. The table shows that all 6000 samples were successful (Status: OK) and completed within the expected time frame. The 'Sample Time' column shows values ranging from 597 to 2674 ms. The 'Latency' column shows values ranging from 0 to 2674 ms. The 'Connect Time' column shows values ranging from 0 to 0 ms.

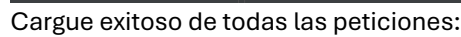
Sample #	Start Time	Thread Name	Label	Sample Time	Status	Bytes	Sent Bytes	Latency	Connect Time
6000	19.38.18.698	Logeo usuarios...	Post Login us...	597	OK	310	251	597	0
5999	19.38.18.661	Logeo usuarios...	Post Login us...	610	OK	310	251	610	0
5998	19.38.18.648	Logeo usuarios...	Post Login us...	623	OK	310	251	623	0
5997	19.38.17.971	Logeo usuarios...	Post Login us...	727	OK	310	251	727	0
5996	19.38.17.911	Logeo usuarios...	Post Login us...	784	OK	310	251	784	0
5995	19.38.17.877	Logeo usuarios...	Post Login us...	664	OK	310	251	664	0
5994	19.38.17.850	Logeo usuarios...	Post Login us...	943	OK	310	251	943	0
5993	19.38.17.911	Logeo usuarios...	Post Login us...	689	OK	310	251	689	0
5992	19.38.16.538	Logeo usuarios...	Post Login us...	1772	OK	310	251	1772	0
5991	19.38.16.097	Logeo usuarios...	Post Login us...	1216	OK	310	251	1216	0
5990	19.38.16.065	Logeo usuarios...	Post Login us...	1473	OK	310	251	1473	0
5989	19.38.16.013	Logeo usuarios...	Post Login us...	1277	OK	310	251	1277	0
5988	19.38.16.136	Logeo usuarios...	Post Login us...	2037	OK	310	251	2037	0
5987	19.38.16.402	Logeo usuarios...	Post Login us...	1683	OK	310	251	1683	0
5986	19.38.15.476	Logeo usuarios...	Post Login us...	2537	OK	310	251	2537	0
5985	19.38.15.163	Logeo usuarios...	Post Login us...	2813	OK	310	251	2813	0
5984	19.38.15.370	Logeo usuarios...	Post Login us...	2401	OK	310	251	2401	0
5983	19.38.15.223	Logeo usuarios...	Post Login us...	2248	OK	310	251	2248	0
5982	19.38.15.480	Logeo usuarios...	Post Login us...	2470	OK	310	251	2470	0
5981	19.38.15.129	Logeo usuarios...	Post Login us...	2674	OK	310	251	2674	0

Todos los response fueron exitosos sin fallas



- Cargar archivo:

Request que se ejecutó en este se está realizando la carga de archivos masivos:



La desviación estándar es alta (2512 ms), lo que significa que los tiempos de respuesta varían mucho entre las solicitudes. Algunas son rápidas, pero otras son mucho más lentas.

Response de la data del archivo:

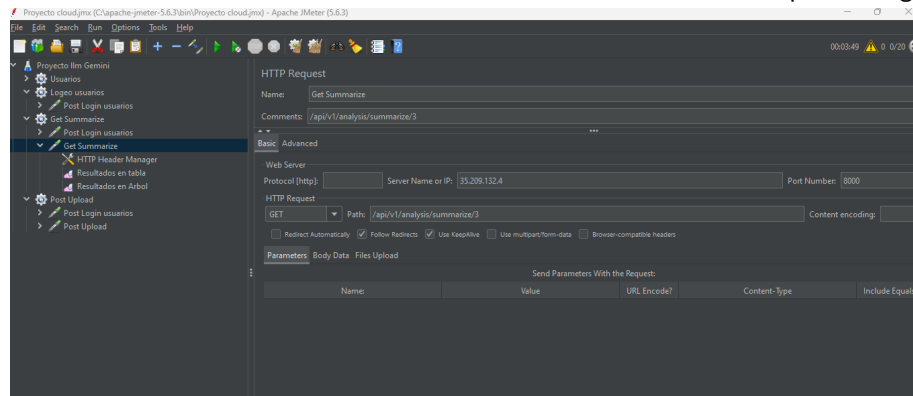
The screenshot shows the Burp Suite interface. On the left, the 'Project Item Gemini' tree is expanded to 'Post Login usuarios', which contains 'headers POST', 'Extra token', 'Guardar acces token global', 'Resultados en Arbol', and 'Resultados en tabla'. The 'View Results Tree' panel on the right shows a list of results for the selected item. The 'Response Body' tab is active, displaying a JSON response:

```
{
  "id": 34,
  "title": "test document doc",
  "content": "Internet m\u00f3vil rompe r\u00e9cords, 45 millones de accesos, mientras el 5G avanza lento y la telefon\u00eda fija desaparece en el crecimiento de los servicios de telecomunicaciones en Colombia ha mostrado un avance significativo en los \u00faltimos a\u00f1os con un aumento sustancial en los accesos fijos y m\u00f3viles a Internet, \u00e1s\u00ed como en la penetraci\u00f3n de la telefon\u00eda m\u00f3vil. El Bolet\u00edn Trimestral del Ministerio TIC para el tercer trimestre de 2024 revela datos que reflejan tanto el progreso del sector como los desaf\u00edos que a\u00fan persisten en la reducci\u00f3n de la brecha digital. Un n\u00famero de accesos fijos a Internet alcanz\u00f3 los 9,10 millones, con un crecimiento interanual de 187.600 nuevas conexiones. Este aumento se traduce en una penetraci\u00f3n de 7,7 accesos por cada 100 habitantes, lo que indica que, si bien hay un avance, el pa\u00eds enfrenta barreras para ampliar la conectividad, especialmente en zonas rurales. En este sentido, la brecha entre regiones sigue siendo un reto, con ciudades como Bogot\u00e1 liderando el acceso con 29 conexiones por cada 100 habitantes, mientras que otras regiones se encuentran rezagadas. Un n\u00edndice de acceso a Internet m\u00f3vil tambi\u00e9n muestra un crecimiento notable, alcanzando los 45,1 millones de accesos, con un 85,4% operando en tecnolog\u00eda 4G y un 5,8% en 5G. La expansi\u00f3n de la red 5G es particularmente relevante, ya que en el \u00faltimo a\u00f1o se sumaron m\u00e1s de 640.000 nuevos accesos, lo que sugiere una transici\u00f3n progresiva hacia tecnolog\u00edas m\u00e1s avanzadas. No obstante, la cobertura de 5G sigue siendo limitada a"
```

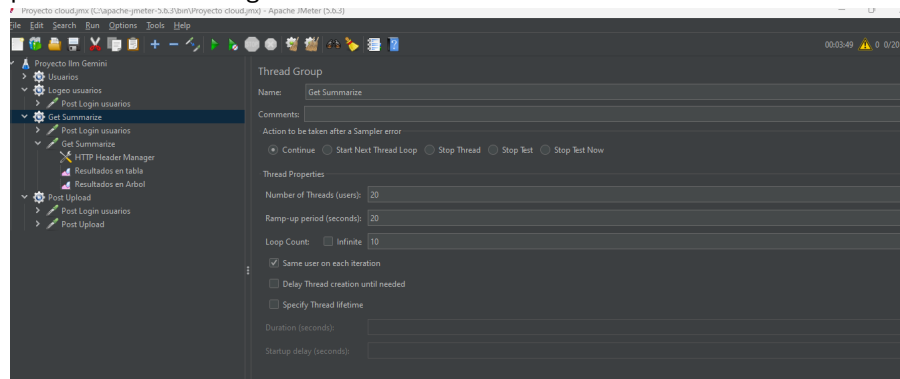
- Resumen del archivo cargado:

<http://35.209.132.4:8000/api/v1/analysis/summarize/3>

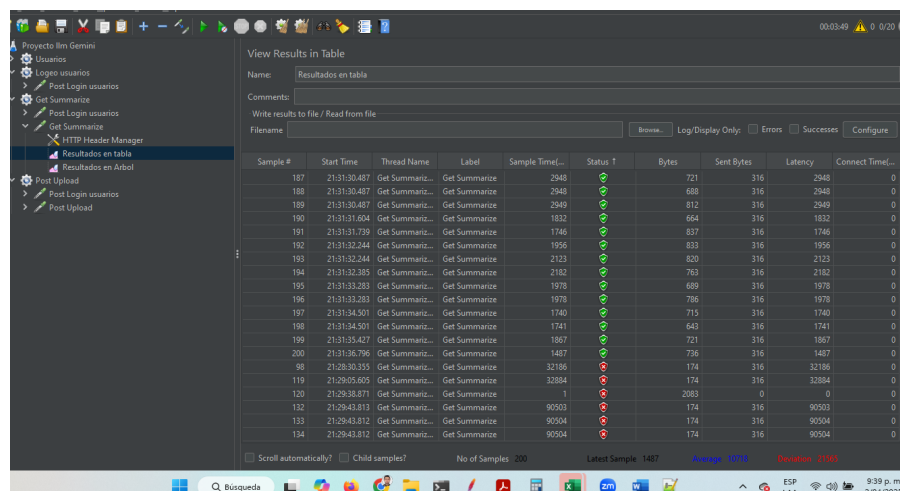
Al ser un método Get trae la información de acuerdo con el archivo que se cargó previamente:



Toco realizar una configuración diferente, en los parámetros de carga debido a que bajaba el servidor, numero de usuarios 20 peticiones por segundo 20 y 10 ciclos para un total de 200 peticiones en 200 segundos.



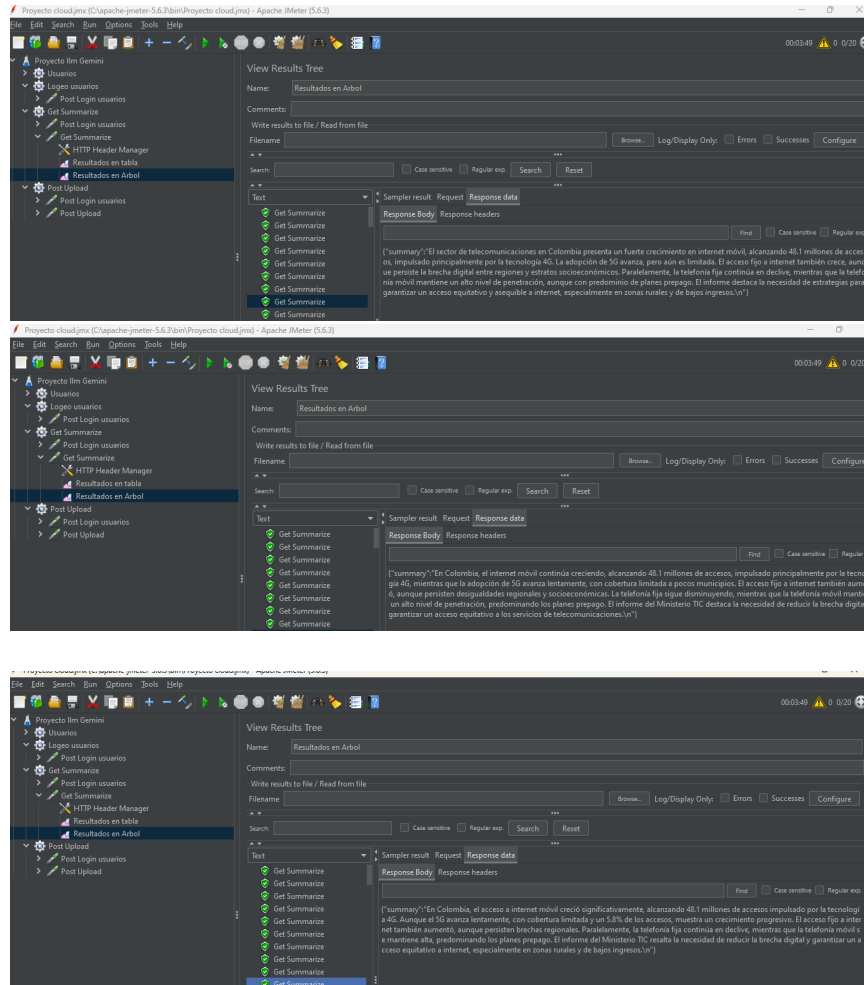
Se evidenciaron 7 peticiones fallidas y 193 correctas con esta configuración de prueba de estrés:



Conclusión:

1. Promedio alto (10.7 segundos): Esto indica que el servidor está tardando mucho en responder.
2. Desviación estándar muy alta (21.5 segundos): Hay una gran inconsistencia en los tiempos de respuesta, lo que sugiere que algunas solicitudes son rápidas y otras extremadamente lentas.
3. Errores en algunas solicitudes (90504 ms, 32.8 s, etc.): Algunas peticiones han fallado o han tomado demasiado tiempo.

Al momento de realizar el response este trae la información diferente de cada petición del mismo texto debido a que el resumen se hace de forma diferente en cada petición:



Posibles problemas y soluciones:

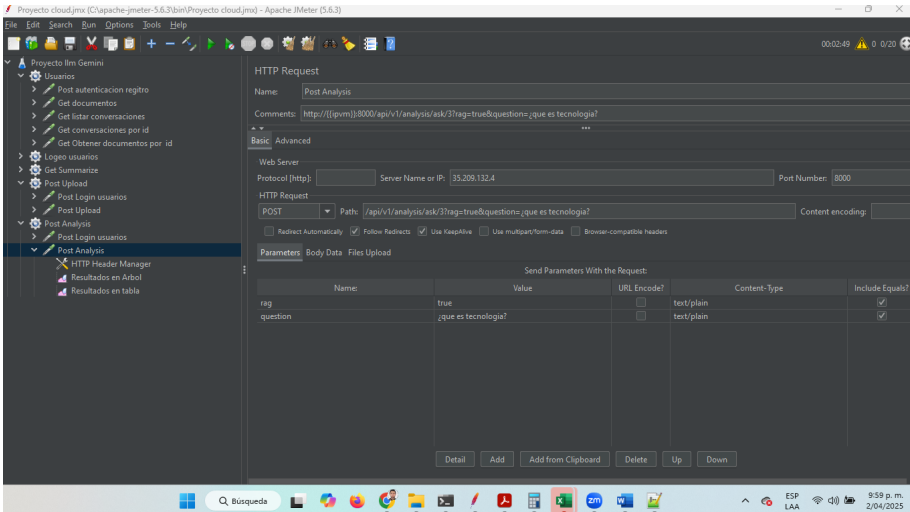
Sobrecarga del servidor: El backend puede estar saturado. Se recomienda monitorear CPU, memoria y latencia en el servidor.

Cuellos de botella en la base de datos: Si este endpoint depende de consultas, revisa los tiempos de ejecución en la base de datos.

Problemas de concurrencia: Si hay muchas solicitudes simultáneas, prueba con menos hilos para ver si mejora.

Optimización del código: Usa caching, mejora la estructura de datos o implementa procesamiento en segundo plano si es posible.

- Pregunta sobre el archivo cargado:
 http://35.209.132.4:8000/api/v1/analysis/ask/3?rag=true&question=¿que es tecnologia?
 Envía una pregunta referente al archivo que se cargó previamente, con 20 peticiones cada 20 segundos en un ciclo de 10:



Genero 187 casos exitosos y 13 fallidos al momento de realizar los diferentes envíos de petición:

View Results in Table

Name: Resultados en tabla

Comments:

Write results to file / Read from file

Filename: Log/Display Only: ☐ Errors ☐ Successes

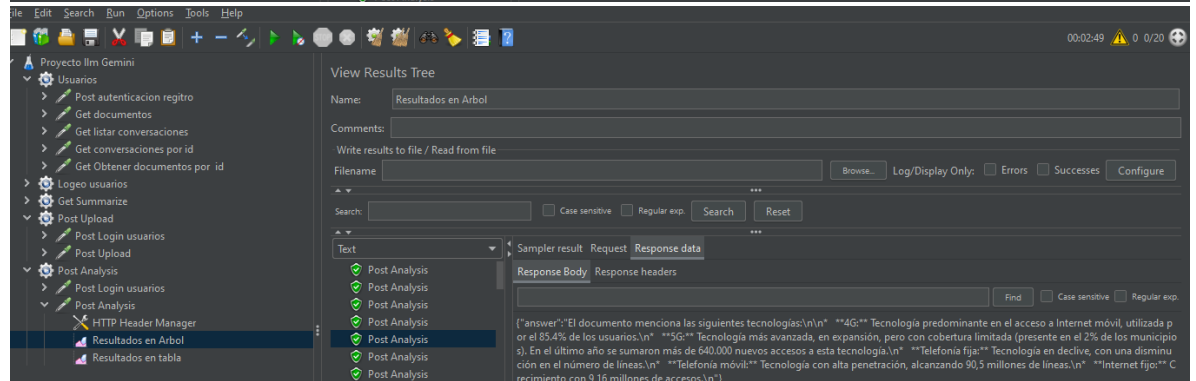
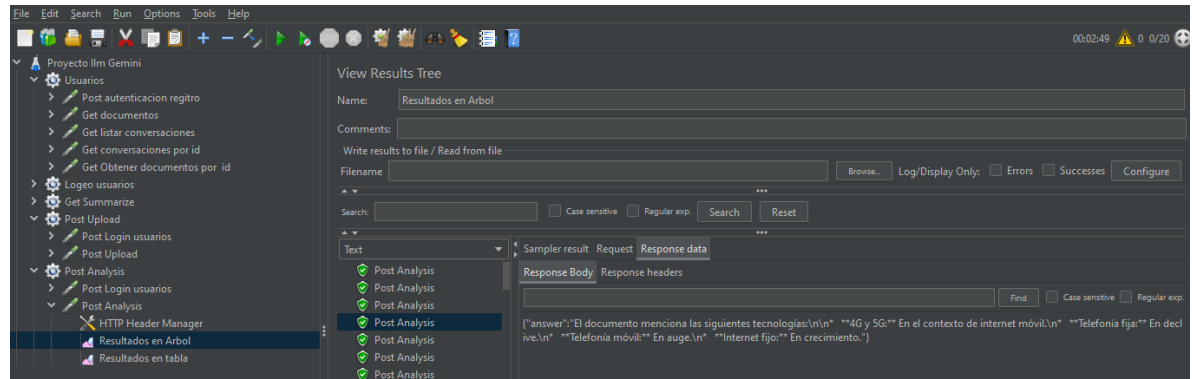
Sample #	Start Time	Thread Name	Label	Sample Time(...)	Status ↓	Bytes	Sent Bytes	Latency	Connect Time(...)
55	21:54:31.424	Post Analysis ...	Post Analysis	60578	✖	174	486	60578	0
56	21:54:31.491	Post Analysis ...	Post Analysis	60511	✖	174	486	60511	0
60	21:55:32.004	Post Analysis ...	Post Analysis	231	✖	174	357	231	117
61	21:55:32.236	Post Analysis ...	Post Analysis	3980	✖	174	357	3980	97
72	21:55:36.939	Post Analysis ...	Post Analysis	873	✖	174	357	873	98
77	21:55:37.813	Post Analysis ...	Post Analysis	1262	✖	174	357	1262	101
81	21:55:39.076	Post Analysis ...	Post Analysis	204	✖	174	357	204	98
83	21:55:39.280	Post Analysis ...	Post Analysis	3374	✖	174	357	3374	98
86	21:55:43.238	Post Analysis ...	Post Analysis	1477	✖	174	357	1477	98
100	21:55:44.715	Post Analysis ...	Post Analysis	689	✖	174	357	689	100
101	21:55:45.526	Post Analysis ...	Post Analysis	30360	✖	174	486	30360	0
105	21:56:16.407	Post Analysis ...	Post Analysis	4063	✖	174	357	4063	81
109	21:56:20.470	Post Analysis ...	Post Analysis	223	✖	174	357	223	79
1	21:54:05.324	Post Analysis ...	Post Analysis	1751	✔	511	486	1751	0
2	21:54:06.338	Post Analysis ...	Post Analysis	2157	✔	785	486	2157	0
3	21:54:07.429	Post Analysis ...	Post Analysis	1344	✔	486	486	1344	0
4	21:54:08.818	Post Analysis ...	Post Analysis	1840	✔	359	486	1840	0
5	21:54:07.665	Post Analysis ...	Post Analysis	2993	✔	769	486	2993	0
6	21:54:09.343	Post Analysis ...	Post Analysis	1750	✔	366	486	1750	0
7	21:54:09.343	Post Analysis ...	Post Analysis	1861	✔	408	486	1861	0

☐ Scroll automatically? ☐ Child samples? No of Samples: 200 Latest Sample: 1457 Average: 6200 Deviation: 10817

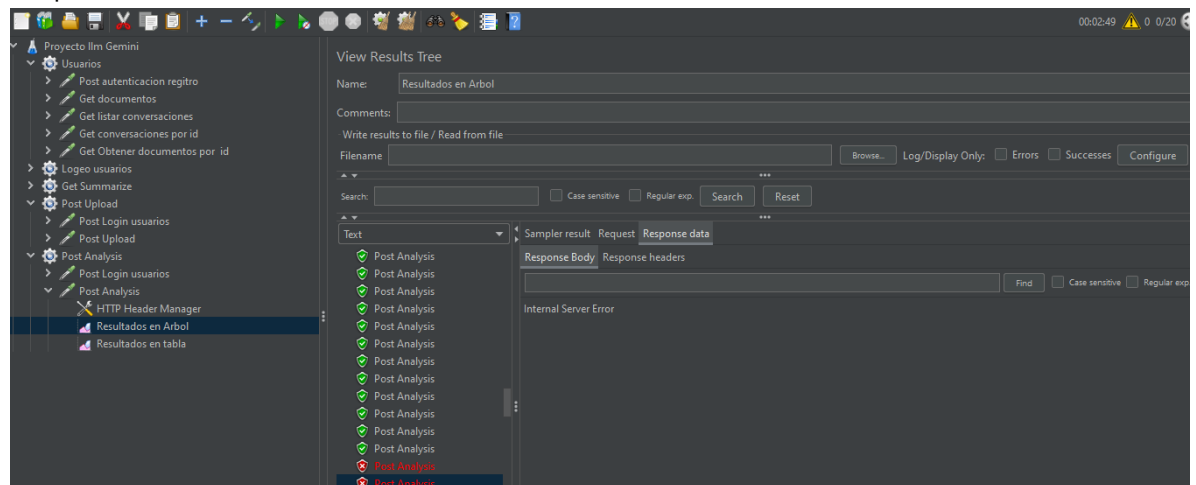
Conclusiones:

1. Hay tiempos de respuesta extremadamente altos, como 60,578 ms (~60 segundos) en algunas solicitudes.
2. Otras respuestas tardan solo unos cientos de milisegundos.
3. Esto sugiere que algunas solicitudes se quedan atascadas antes de fallar.
4. Tiempo promedio: 6200 ms (~6.2 segundos).
5. Desviación estándar: 10,617 ms, lo que indica que los tiempos de respuesta son muy variables.

Response de casos exitosos:



Response caso fallido:



Posibles causas del problema

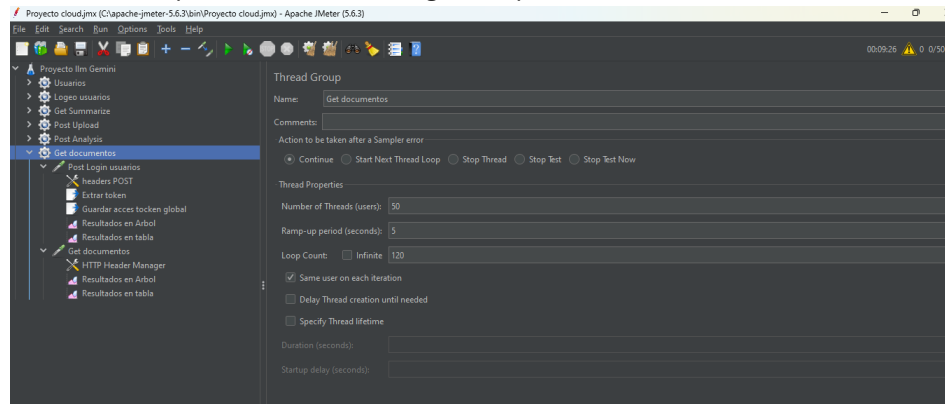
1. Sobrecarga del servidor: Puede estar procesando más solicitudes de las que soporta.

2. Problemas en la base de datos: Consultas lentas o bloqueos pueden estar retrasando las respuestas.
3. Errores en la API: La lógica de negocio o el backend pueden estar devolviendo errores internos.
4. Timeouts: Las peticiones pueden estar tardando demasiado y fallando por configuración de JMeter o el servidor.

- Trae los documentos de la sesión correspondiente:

<http://35.209.132.4:80008000/api/v1/documents/listDocuments>

Se realizan 50 peticiones cada 50 segundos por 120 ciclos funcionando correctamente:

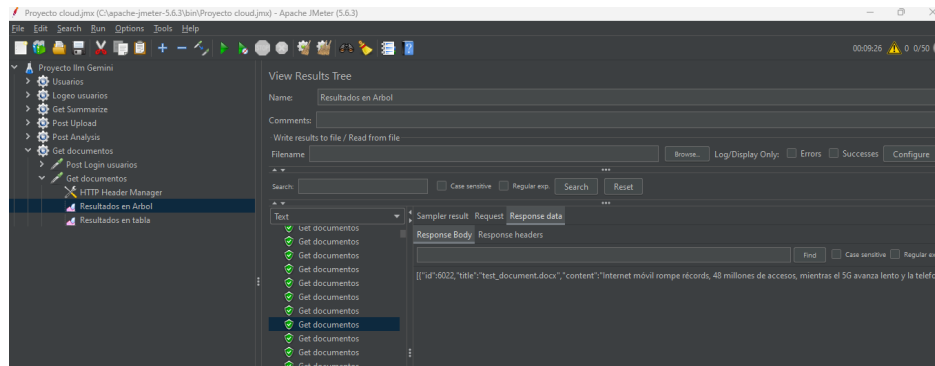


La ejecución fue exitosa en sus 6000 peticiones.

The screenshot shows the 'View Results in Table' window in Apache JMeter. It displays a table of test results for the 'Resultados en tabla' thread group. The table has columns for Sample #, Start Time, Thread Name, Label, Sample Time, Status, Bytes, Sent Bytes, Latency, and Connect Time. The status for all samples is 'Success' (green checkmark). The total number of samples is 6000, with the latest sample being 287. The table shows a list of samples with their respective times and data sizes.

Sample #	Start Time	Thread Name	Label	Sample Time...	Status	Bytes	Sent Bytes	Latency	Connect Time...
1	22:43:47.156	Get document...	Get document...	254	Success	3638	319	121	0
2	22:43:47.224	Get document...	Get document...	249	Success	3638	319	104	0
3	22:43:47.353	Get document...	Get document...	272	Success	3638	319	140	0
4	22:43:47.473	Get document...	Get document...	322	Success	3638	319	195	0
5	22:43:47.539	Get document...	Get document...	287	Success	3638	319	151	0
6	22:43:47.625	Get document...	Get document...	231	Success	3638	319	103	0
7	22:43:47.751	Get document...	Get document...	293	Success	3638	319	159	0
8	22:43:48.067	Get document...	Get document...	348	Success	3638	319	209	0
9	22:43:48.188	Get document...	Get document...	345	Success	3638	319	104	0
10	22:43:48.387	Get document...	Get document...	564	Success	3638	319	432	0
11	22:43:48.653	Get document...	Get document...	702	Success	3638	319	561	0
12	22:43:48.792	Get document...	Get document...	654	Success	3638	319	564	0
13	22:43:48.939	Get document...	Get document...	572	Success	3638	319	438	0
14	22:43:48.968	Get document...	Get document...	589	Success	3638	319	456	0
15	22:43:49.029	Get document...	Get document...	624	Success	3638	319	491	0
16	22:43:49.167	Get document...	Get document...	574	Success	3638	319	446	0
17	22:43:49.328	Get document...	Get document...	499	Success	3638	319	371	0
18	22:43:49.479	Get document...	Get document...	886	Success	3638	319	751	0
19	22:43:49.481	Get document...	Get document...	907	Success	3638	319	768	0
20	22:43:49.495	Get document...	Get document...	900	Success	3638	319	764	0

Podemos ver el titulo del documento cargado y su contenido en formato de texto.



Conclusiones:

1. Tiempo promedio: 1,832 ms (~1.8 segundos).
2. Desviaci\u00f3n est\u00e1ndar: 596 ms, lo que significa que los tiempos de respuesta son relativamente consistentes.
3. La mayor\u00eda de los tiempos de respuesta est\u00e1n entre 1 y 2 segundos, con algunas variaciones.
4. Buen desempe\u00f1o de la API: No hay errores y los tiempos de respuesta son aceptables.
5. Algunas variaciones en latencia: Aunque no son cr\u00edticas, pueden mejorar con optimizaci\u00f3n de red o del backend.
6. Posible optimizaci\u00f3n: Si el objetivo es reducir tiempos de respuesta por debajo de 1 segundo, podr\u00edas revisar mejoras en la infraestructura o cach\u00e9.