Manual Despliegue de Sistemas Canalización de Datos para análisis de Información en AWS

Contenido

[1. Introducción 3](#_Toc169287934)

[2. Instalación de Agente de CloudWatch y crear el archivo de configuración 4](#_Toc169287935)

[2.1 Instalación del agente en el Servidor Web 4](#_Toc169287936)

[2.2 Creación del archivo de configuración 4](#_Toc169287937)

[2.3 Configurar el archivo httpd.conf para que los log se registren en formato JSON 5](#_Toc169287938)

[2.4 Usar el archivo de configuración actualizado con el CloudWatch Agent 7](#_Toc169287939)

[3. Comprobación del funcionamiento del agente de CloudWatch 8](#_Toc169287940)

[3.1 Uso de logs de ejemplo generados manualmente 10](#_Toc169287941)

[4. Usar la función de CloudWatch Log Insight para el Análisis 12](#_Toc169287942)

[4.1 Determinar el número de visitantes con log Insight al menu 12](#_Toc169287943)

[4.2 Determinar el número de visitantes que han realizado una compra con log Insight 14](#_Toc169287944)

[5. Ajustar las consultas y paneles de CloudWatch para obtener información más detallada 15](#_Toc169287945)

[5.1 Crear las tablas para observar la información. 16](#_Toc169287946)

[6. Guardar los registros en servicio S3 para posteriores ampliaciones y otorgar flexibilidad con otros servicios. 21](#_Toc169287947)

[6.1 Comprobar que podemos realizar consultas con el servicio de S3 Select 22](#_Toc169287948)

# 1. Introducción

En este manual, se explicara de manera detallada como llevar a cabo los pasos necesarios para montar un servicio de canalización de datos en una instancia EC2 de AWS, utilizando los servicios de CloudWatch y almacenamiento en S3.

Es necesario dejar algunas especificaciones claras del escenario sobre el que se va a realizar el manual antes de proceder con la implementación:

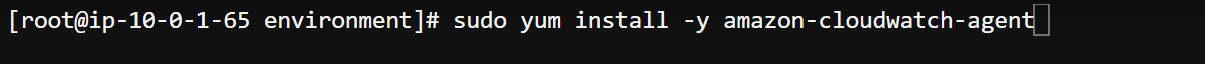
* El sistema operativo de la instancia en el que se realizara la implementación será Amazon Linux 2, que se trata de un SO parecido a CentOS, de manera que los comandos serán específicos de esa distribución de Linux, los cuales pueden variar si se utiliza otro SO.
* Se cuenta con una página web previamente implementada con apache2 que será la que se utilice como método de registro de logs
* En la implementación se usara Cloud9 para la facilitación de la manipulación de archivos en la instancia EC2

# 2. Instalación de Agente de CloudWatch y crear el archivo de configuración

## 2.1 Instalación del agente en el Servidor Web

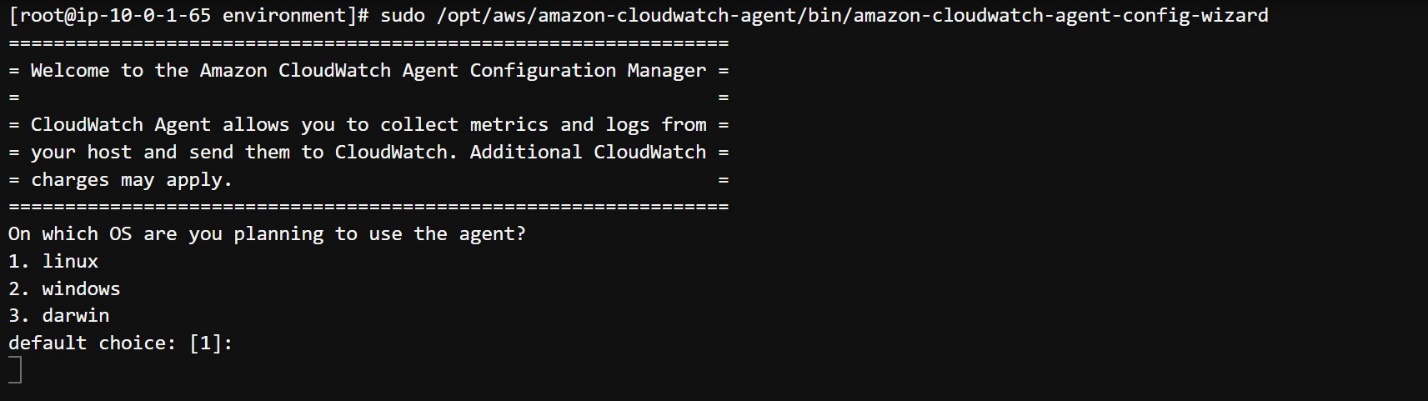
Primeramente instalaremos el servicio de agente de CloudWatch en la instancia del servidor para que este monitoree los logs tanto de acceso como logs de error.

Para ello primeramente usaremos el siguiente comando:



## 2.2 Creación del archivo de configuración

Para ello debemos usar un servicio que se instala junto al propio agente de cloudwatch, como se muestra a continuación tras usar el comando para la creación del archivo de configuración, se nos harán una serie de preguntas para especificar los parámetros.



En caso que descarguemos de manera externa el archivo de configuración tendremos que llevarlo al directorio de configuración del agente de cloudwatch **/opt/aws/amazon-cloudwatch-agent/bin/**

C:\Users\Usuario\AppData\Local\Packages\Microsoft.Windows.Photos_8wekyb3d8bbwe\TempState\ShareServiceTempFolder\Captura de pantalla 2024-05-05 180724.jpeg

## 2.3 Configurar el archivo httpd.conf para que los log se registren en formato JSON

Ahora pasaremos a configurar el archivo httpd.conf de configuración del servicio de apache2 para que los registros se guarden en formato JSON que es el tipo de formato de logs que maneja el servicio de CloudWatch.

Primero localizaremos el archivo httpd.conf, y para mayor seguridad crearemos un archivo de back up de configuración.

Si queremos utilizarlo desde, en este caso, cloud9 podemos crear un enlace simbólico en el directorio de cloud9 y le damos permisos para poder modificar el archivo con los siguientes comandos.

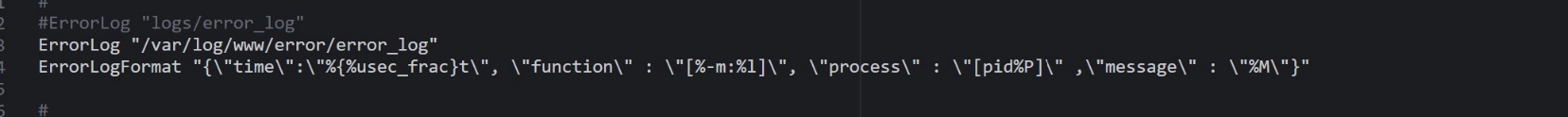




Una vez accedemos al archivo realizaremos las siguientes configuraciones:

Nos moveremos hasta la línea entorno a la 182 donde se lee **ErrorLog "logs/error\_log"** y comentaremos la línea,escribiendo seguidamente **ErrorLog “/var/log/www/error/error\_log”**  donde indicaremos el archivo en el que se van a guardar los registros de error.

Tras ello indicaremos el formato de los logs de error.

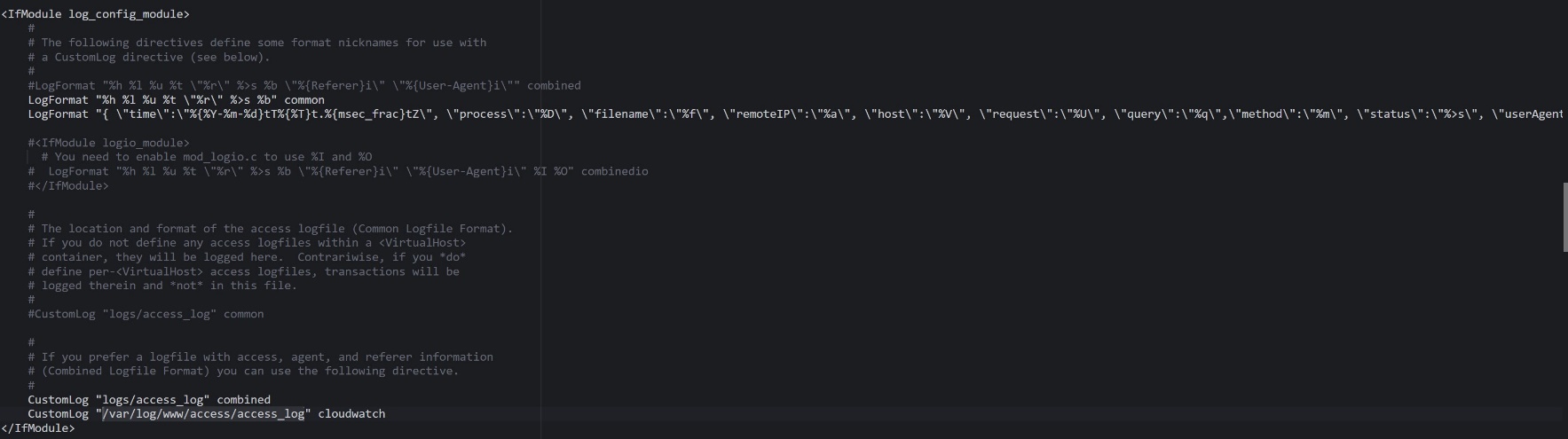


Tras eso nos moveremos a la línea 191 donde se lee **<IfModule log\_config\_module>** y comentaremos la línea inferior a esa **LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\"" combined.**

Un poco más abajo tras la línea **LogFormat "%h %l %u %t \"%r\" %>s %b" common** escribiremos

**LogFormat "{ \"time\":\"%{%Y-%m-%d}tT%{%T}t.%{msec\_frac}tZ\", \"process\":\"%D\", \"filename\":\"%f\", \"remoteIP\":\"%a\", \"host\":\"%V\", \"request\":\"%U\", \"query\":\"%q\",\"method\":\"%m\", \"status\":\"%>s\", \"userAgent\":\"%{User-agent}i\",\"referer\":\"%{Referer}i\"}" cloudwatch**

También comentaremos el modulo entero de **<IfModule logio\_module>** que consta de tres líneas y para terminar nos moveremos hasta la línea 219 y justo la siguiente línea a la que se lee **CustomLog "logs/access\_log" combined** escribiremos **CustomLog "/var/log/www/access/access\_log" cloudwatch** indicando donde se guardaran los registros de acceso.

Finalmente el archivo quedara así:

Tras eso guardaremos y saldremos.

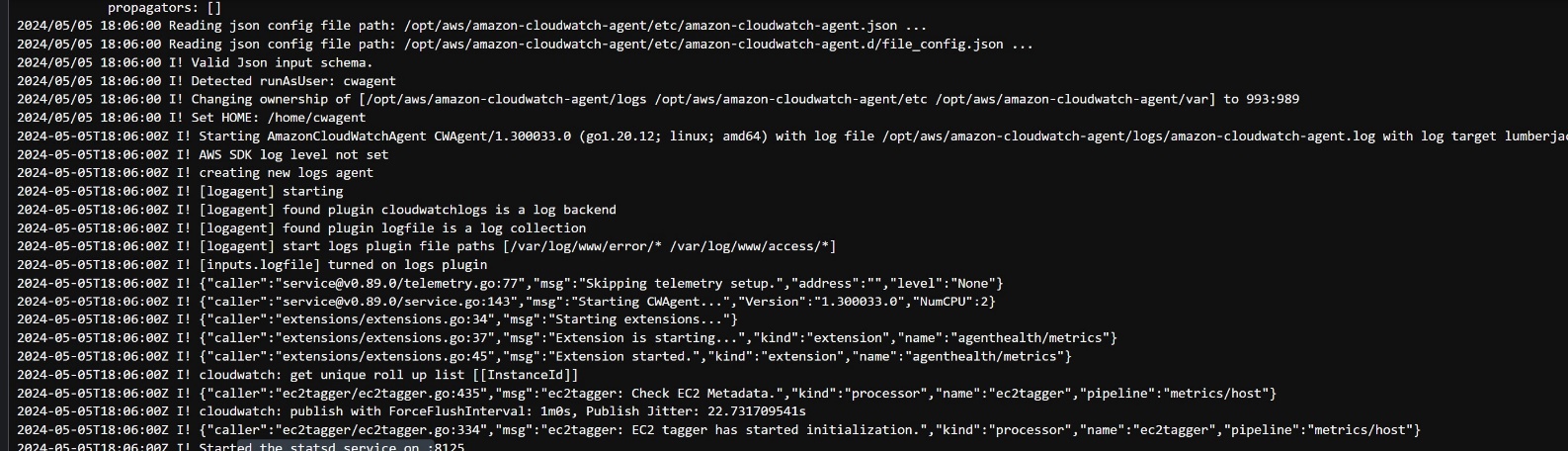
## 2.4 Usar el archivo de configuración actualizado con el CloudWatch Agent

Primero tendremos que crear los directorios para los archivos de registros de acceso y de error que hemos especificado en el archivo httpd.conf, cabe aclarar que los archivos se crearan automáticamente por el servicio de apache2 en el momento en el que se creen los directorios especificados

Reiniciaremos el servicio de apache2 para que los cambios se apliquen, e iniciaremos el agente de CloudWatch para que utilice el archivo de configuración que creamos config.json como referencia.

Para ello utilizaremos el siguiente comando.

C:\Users\Usuario\AppData\Local\Packages\Microsoft.Windows.Photos_8wekyb3d8bbwe\TempState\ShareServiceTempFolder\Captura de pantalla 2024-05-05 200618.jpeg

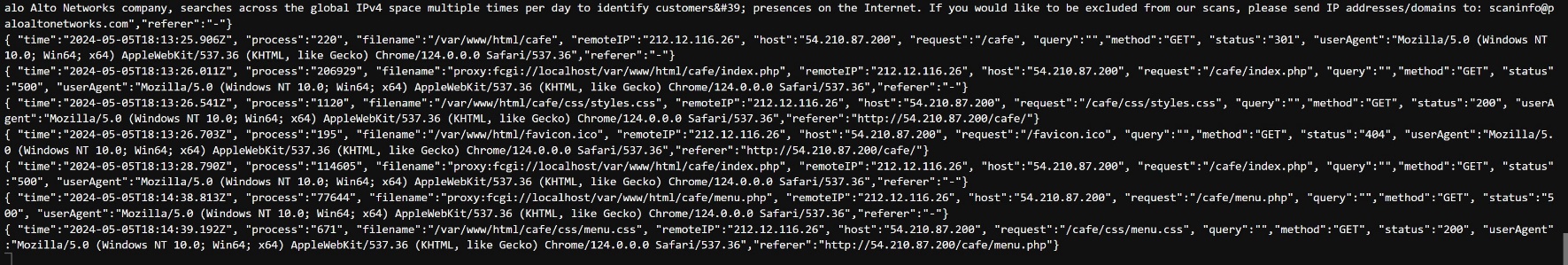


Y comprobaremos que el servicio está funcionando correctamente



# 3. Comprobación del funcionamiento del agente de CloudWatch

Realizaremos acciones de acceso a la página web de manera que podamos comprobar que los nuevos registros se muestran en formato json en los archivos que especificamos previamente.



Observaremos que el agente de CloudWatch está funcionando y transmitiendo, para ello comprobaremos el siguiente archivo con la ayuda del comando cat

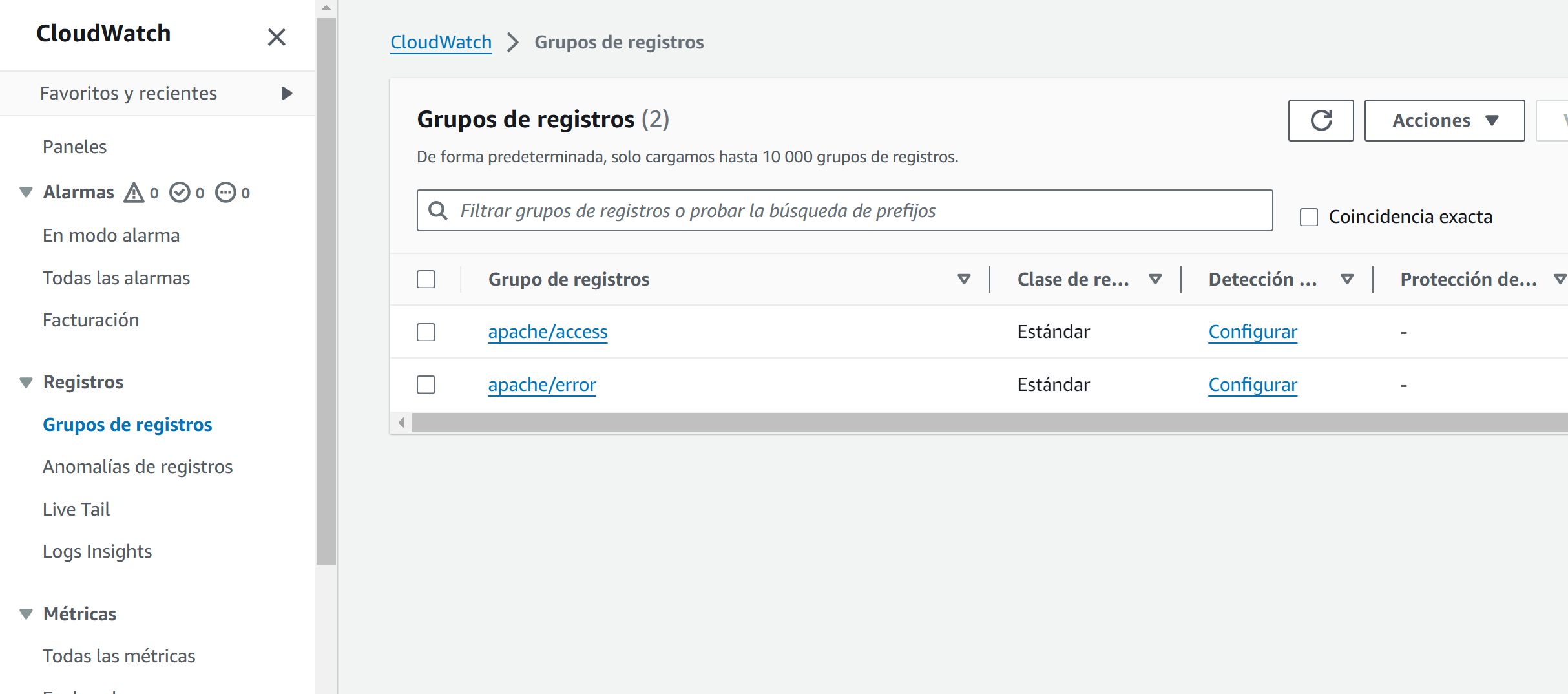


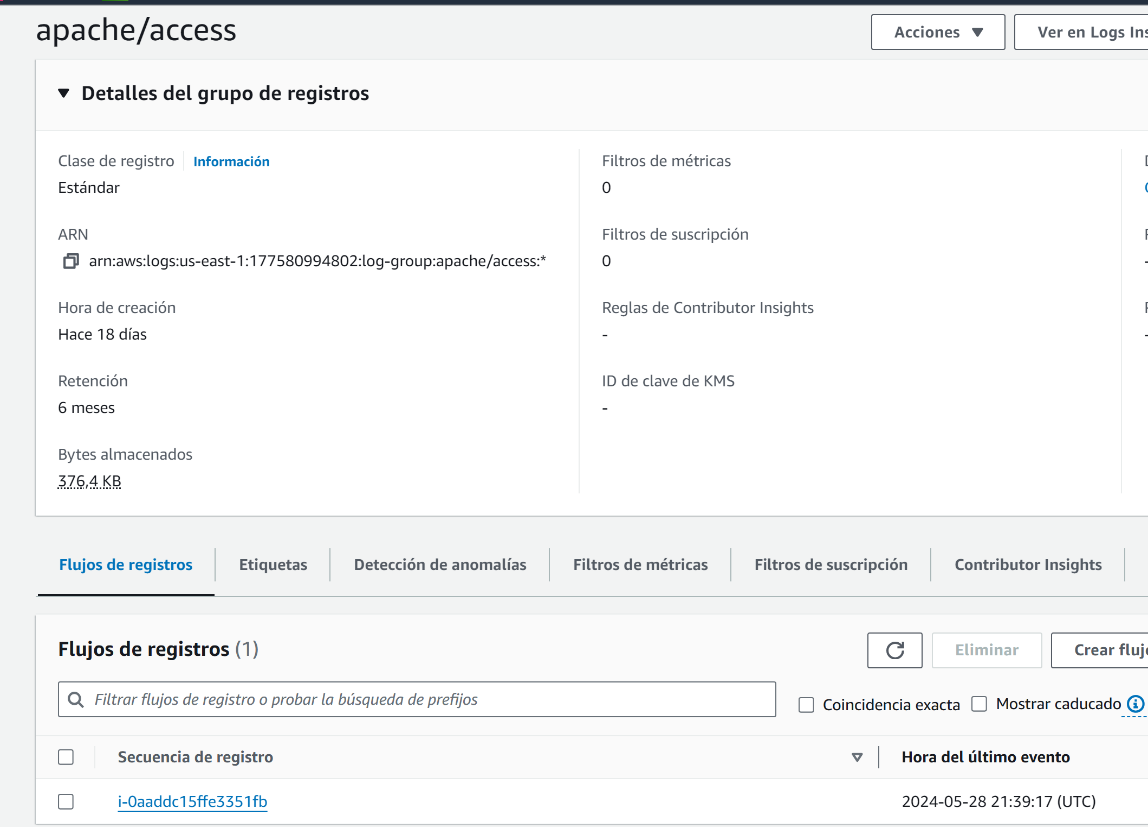
Donde buscaremos líneas parecidas a las siguientes.

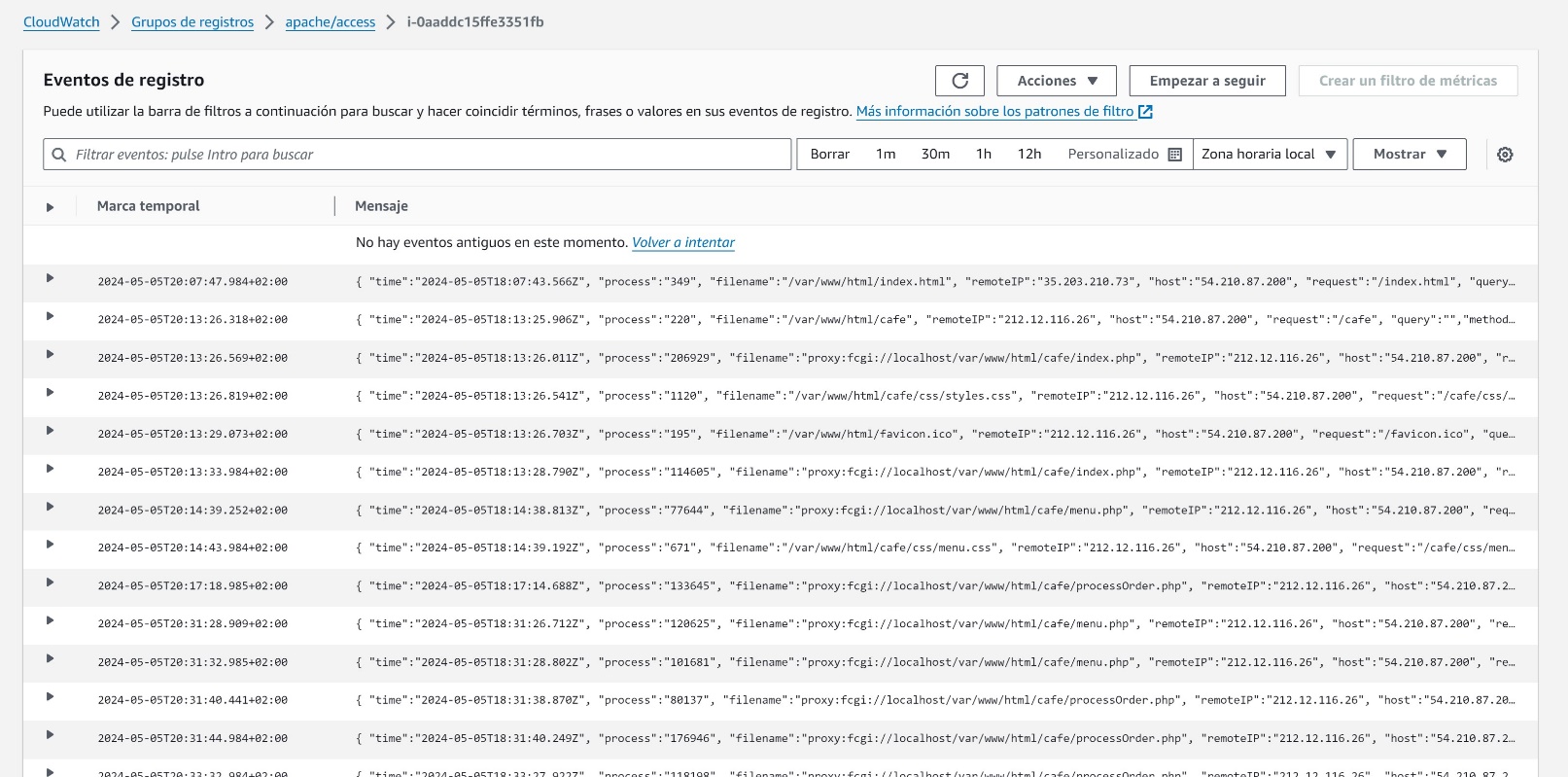


Donde se nos indica que el agente de CloudWatch está leyendo de los dos archivos de logs y enviando dichos registros al servicio de CloudWatch

Ahora podremos verificar que los registros se visualizan en el servicio de CloudWatch, para ello nos desplazaremos en la consola de AWS al Grupo de Registros, y haremos click encima de los grupos, posteriormente nos dirigiremos a flujo de registros y comprobaremos que efectivamente se visualizan las últimas acciones realizadas en la web.





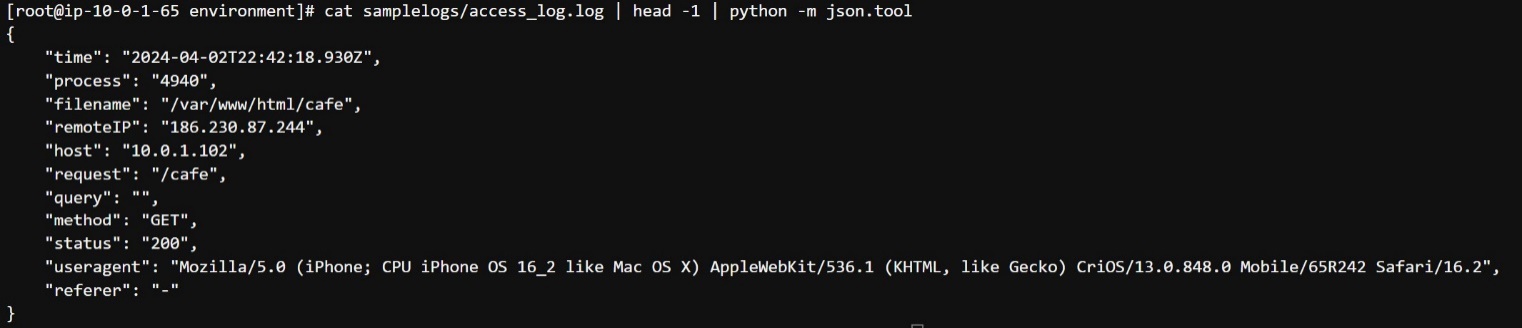


## 3.1 Uso de logs de ejemplo generados manualmente

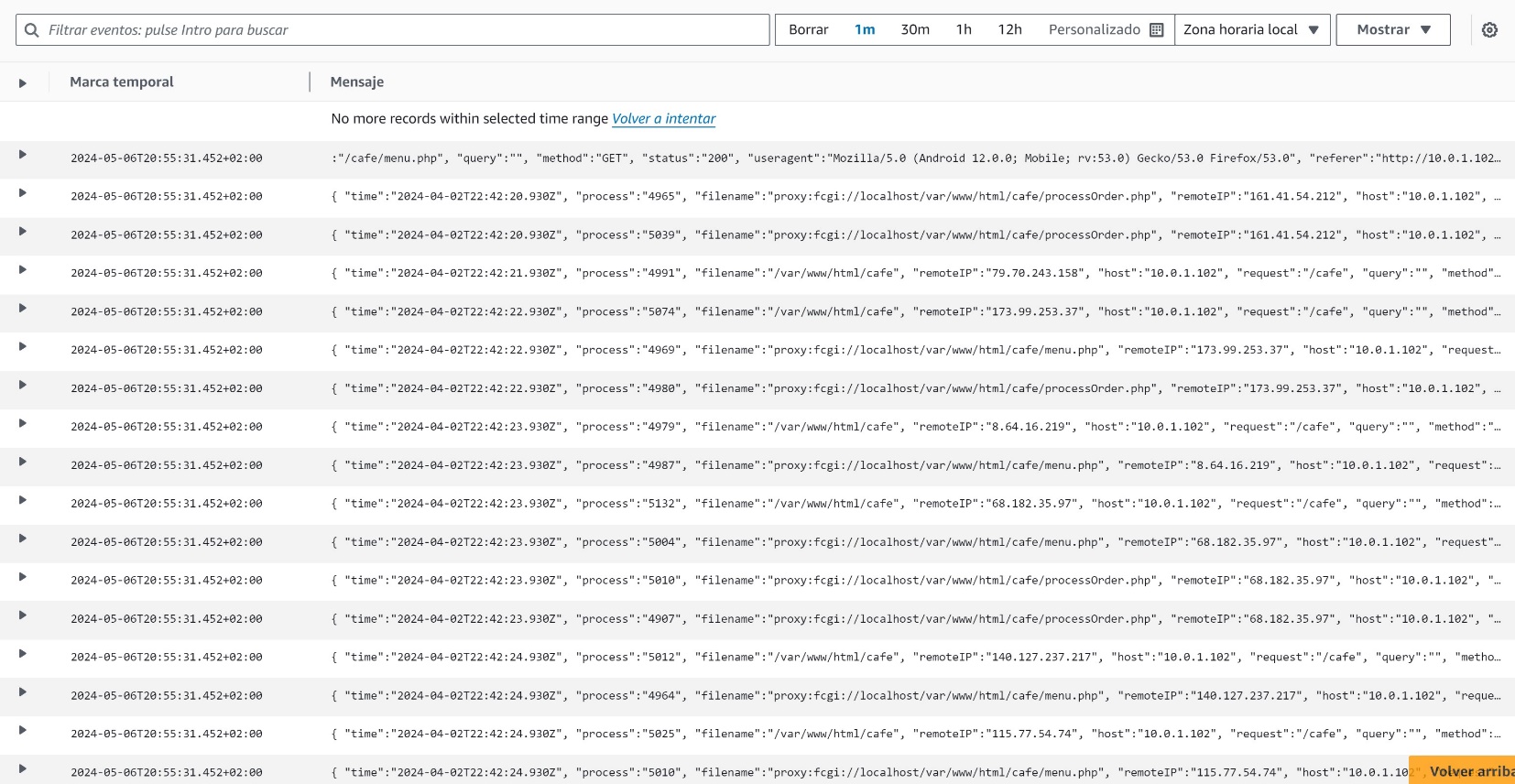
Adicionalmente, para comprobar que el servicio funciona en casos mayores con cantidades grandes de registros, como es el que se nos presenta entre manos, podemos agregar manualmente un archivo de registros, suplantando el existente en las rutas de los logs de apache2.

Para hacer eso primeramente debemos detener el servicio del agente de CloudWatch antes de actualizar el archivo.





Como podemos comprobar los registros se mostraran en los grupos de registros del servicio de cloudwatch

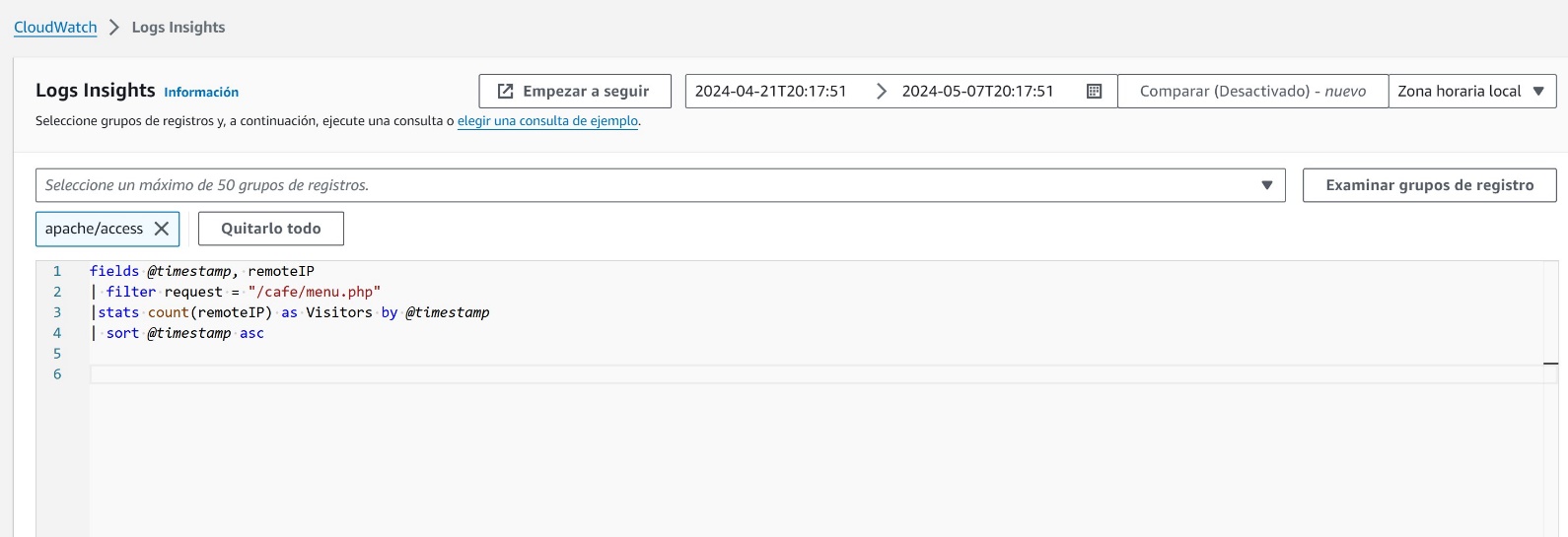


# 4. Usar la función de CloudWatch Log Insight para el Análisis

La función de Log Insight nos permite utilizar consultas para obtener información de los registros almacenados en el servicio de CloudWatch, para comprobar su funcionamiento lo usaremos para obtener el número de visitantes que accedieron al menú de la pagina web.

## 4.1 Determinar el número de visitantes con log Insight al menu

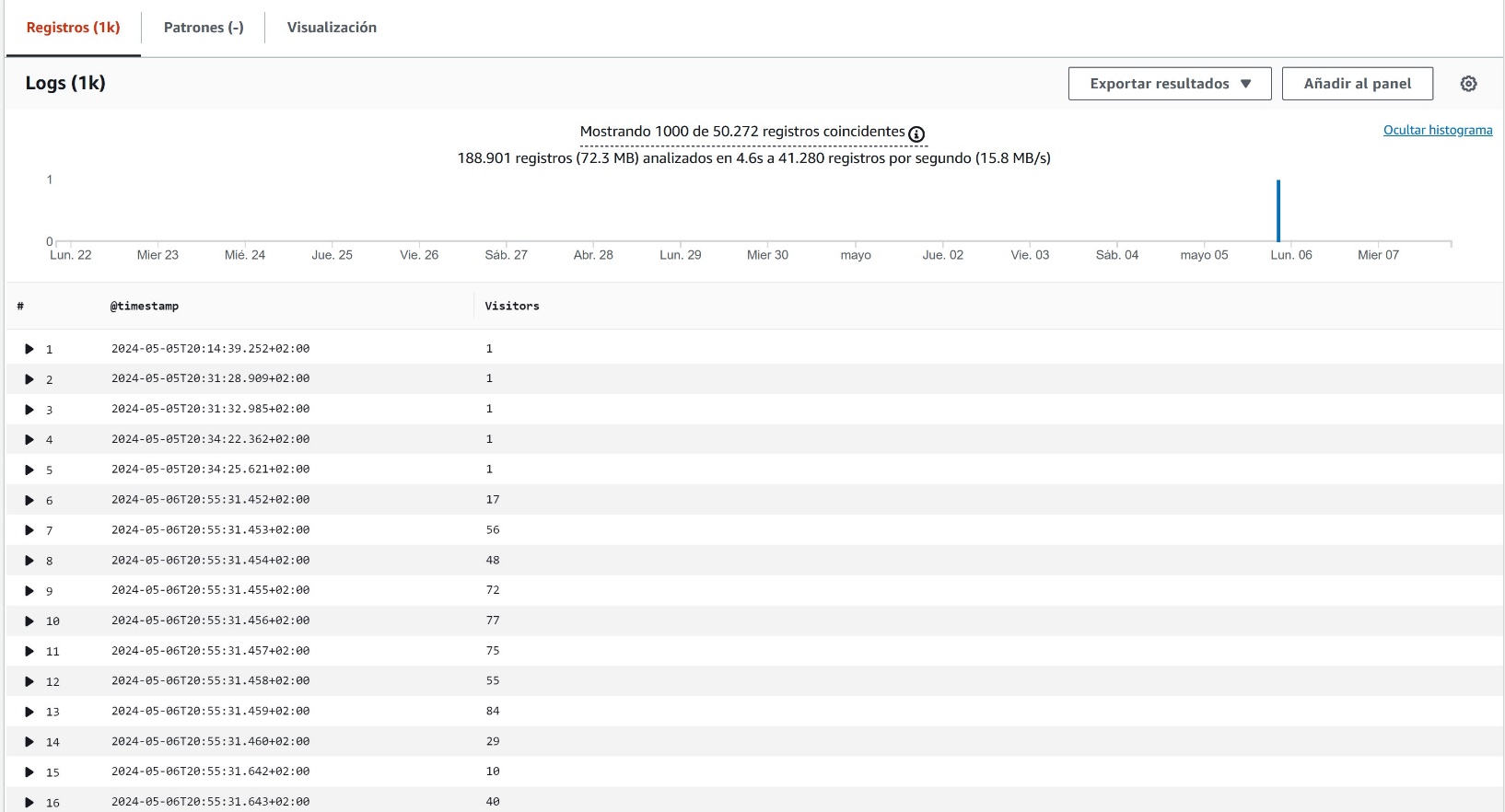
Para obtener la información haremos uso de consultas, en este caso de ejemplo la consulta que usaremos será la siguiente:



En esta se filtrara por la solicitud a la dirección de **/café/menú.php** y contara los resultados por fecha y hora.

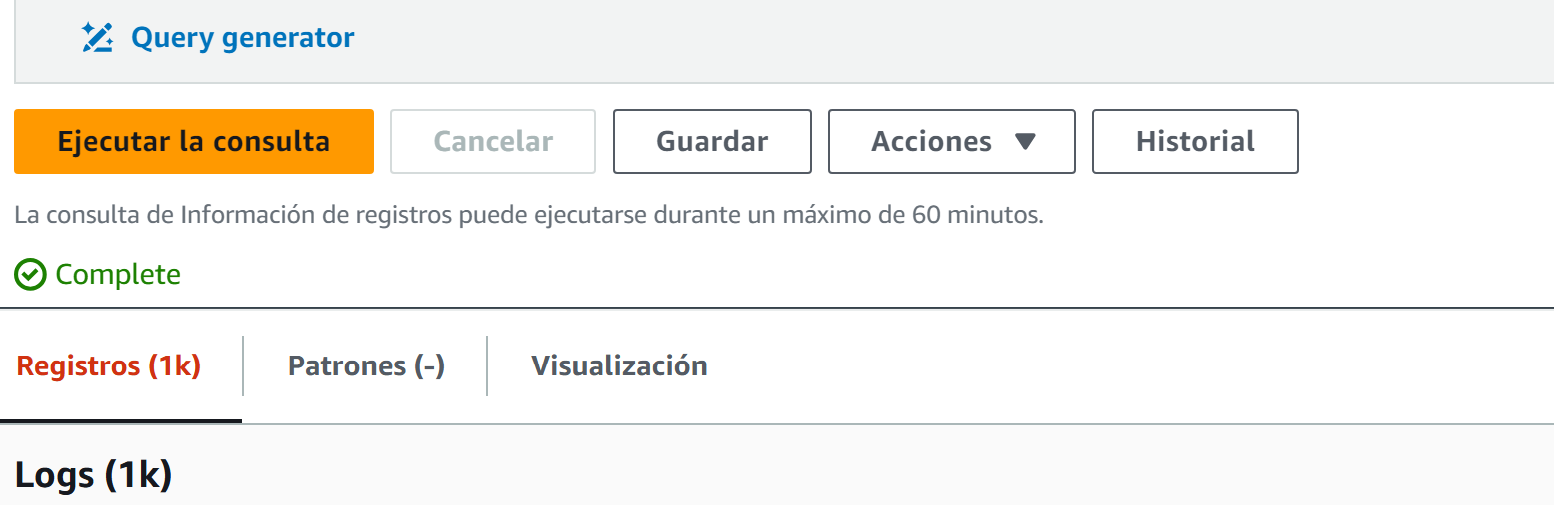
Asignaremos el rango de fechas que queremos analizar con dicha consulta y el grupo de logs que queremos analizar.

El resultado es el siguiente:



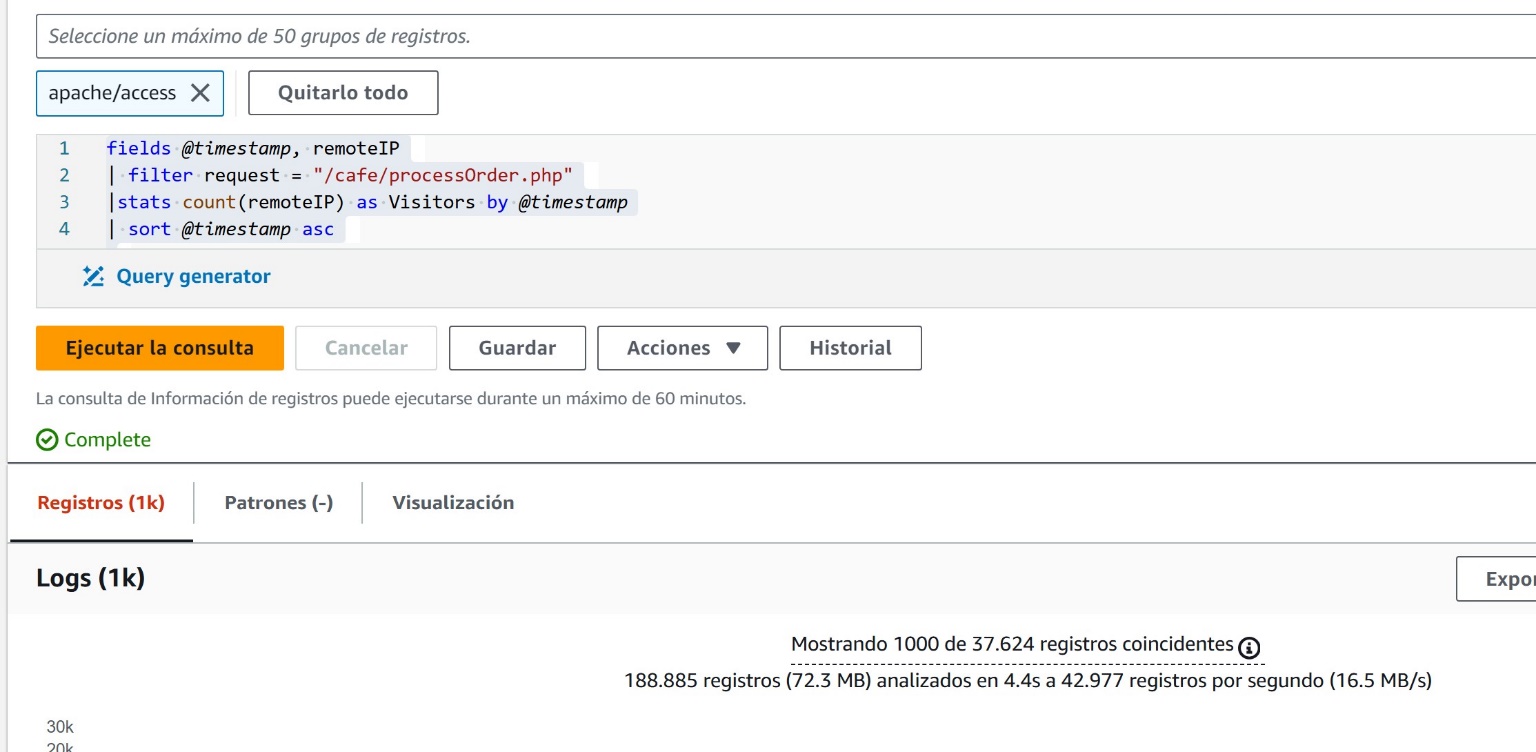
Como vemos solo se muestran 1000 de los 50.272 resultados.

Justo debajo podemos guardar la consulta con un nombre determinado como por ejemplo “menu-visitors”, en la opción “guardar”.

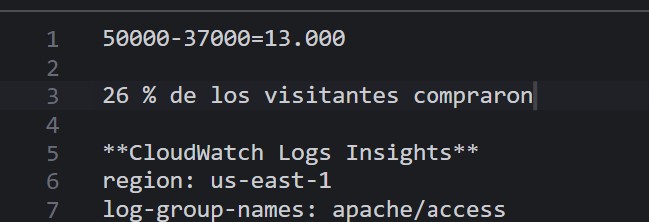


## 4.2 Determinar el número de visitantes que han realizado una compra con log Insight

También podemos comprobar los visitantes que han realizado compras en la página web, cambiando la dirección de la solicitud a **/café/processOrder.php**



Y con esto podríamos incluso calcular porcentajes, como el porcentaje de personas que visitaron la web pero no compraron nada.



# 5. Ajustar las consultas y paneles de CloudWatch para obtener información más detallada

El servicio de CloudWatch permite la creación de diferentes tipos de gráficos para la visualización de manera sencilla y estructurada la información de los registros almacenados.

En esta implementación añadiremos 4 tipos de tablas para ofrecer una información lo más amplia posible.

Antes de realizar las tablas introduciremos nuevos registros, en este caso estos contendrán información de geolocalización de los visitantes de la web. Cabe aclarar que estos registros están introducidos a mano al igual que los anteriores, para proporcionar una cantidad más lógica con el escenario presentado, pero para que el servicio apache2 capturase la información de geolocalizacion sería necesario instalar el módulo **Geo2IP** que contrasta la información de los registros con una base de datos de localizaciones.

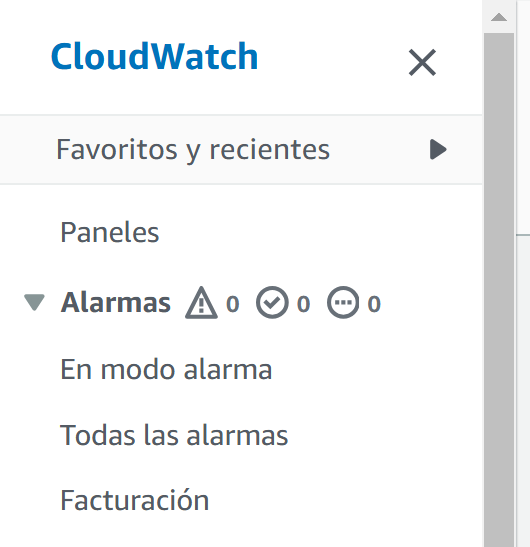
Sin embargo en esta implementación no desarrollaremos el funcionamiento del módulo en cuestión por limitaciones del entorno utilizado.

Una vez introducidos los registros plantearemos las tablas con la información que nos interesa:

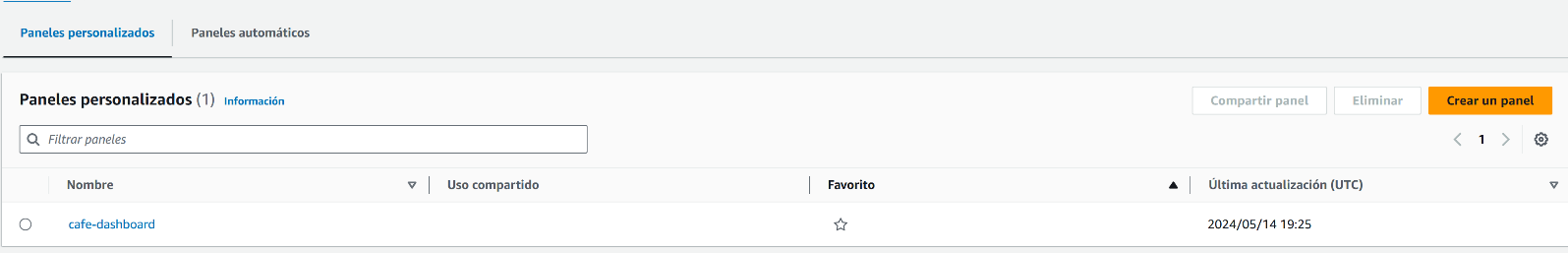
* Un gráfico circular que muestra las 10 ciudades que tuvieron la mayor cantidad de visitantes al sitio web que accedieron a la página del menú.
* Una tabla de registro que muestra las 10 ciudades que tuvieron la mayor cantidad de visitantes al sitio web que realizaron una compra.
* Un gráfico circular que muestra las 10 regiones que tuvieron la mayor cantidad de visitantes que accedieron a la página principal del sitio web.
* Un gráfico de barras que muestra las 10 regiones que tuvieron la mayor cantidad de visitantes al sitio web que realizaron una compra.

## 5.1 Crear las tablas para observar la información.

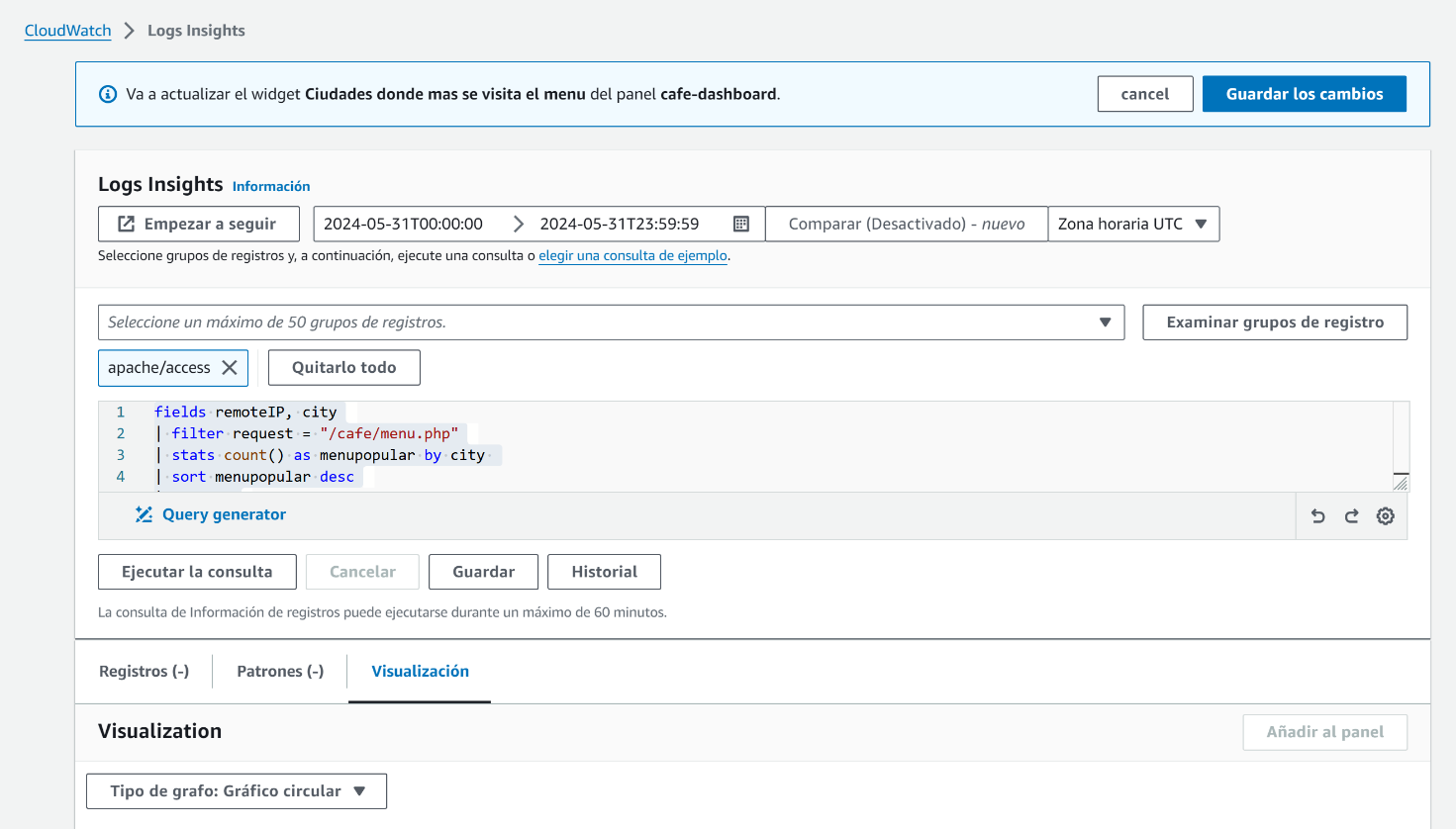
Para crear las tablas debemos acceder a la opción de tablas en la parte izquierda del servicio de cloudwatch

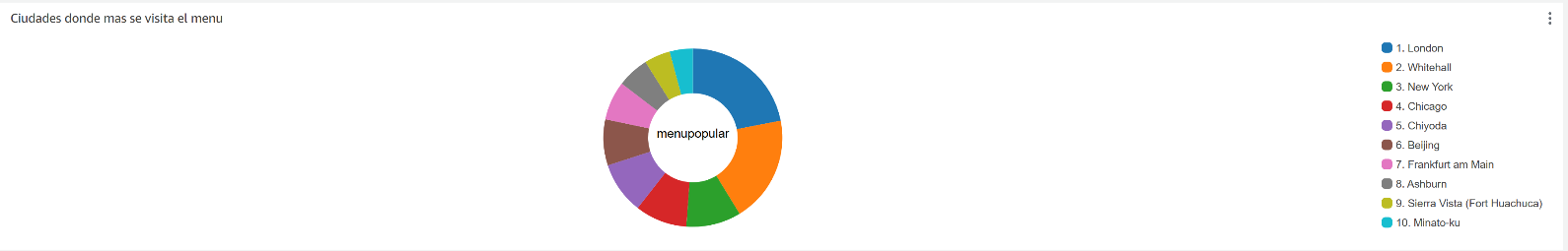


Clicaremos en crear un panel y especificaremos el nombre del mismo

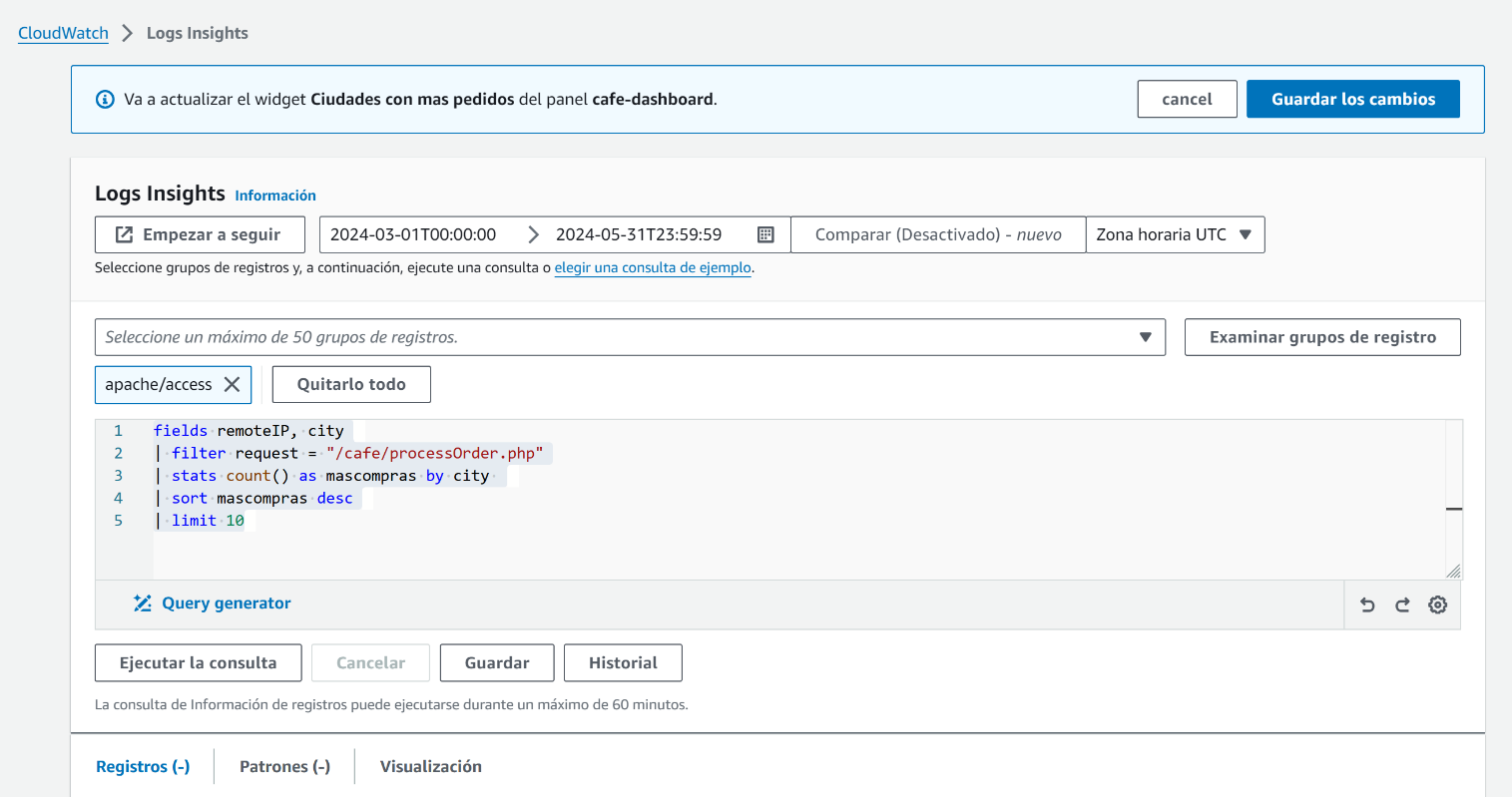


Una vez hecho eso accederemos al panel y crearemos un widget, donde indicaremos tanto el tipo de widget, el origen de los datos, el periodo de tiempo y la consulta.

* Un gráfico circular que muestra las 10 ciudades que tuvieron la mayor cantidad de visitantes al sitio web que accedieron a la página del menú.

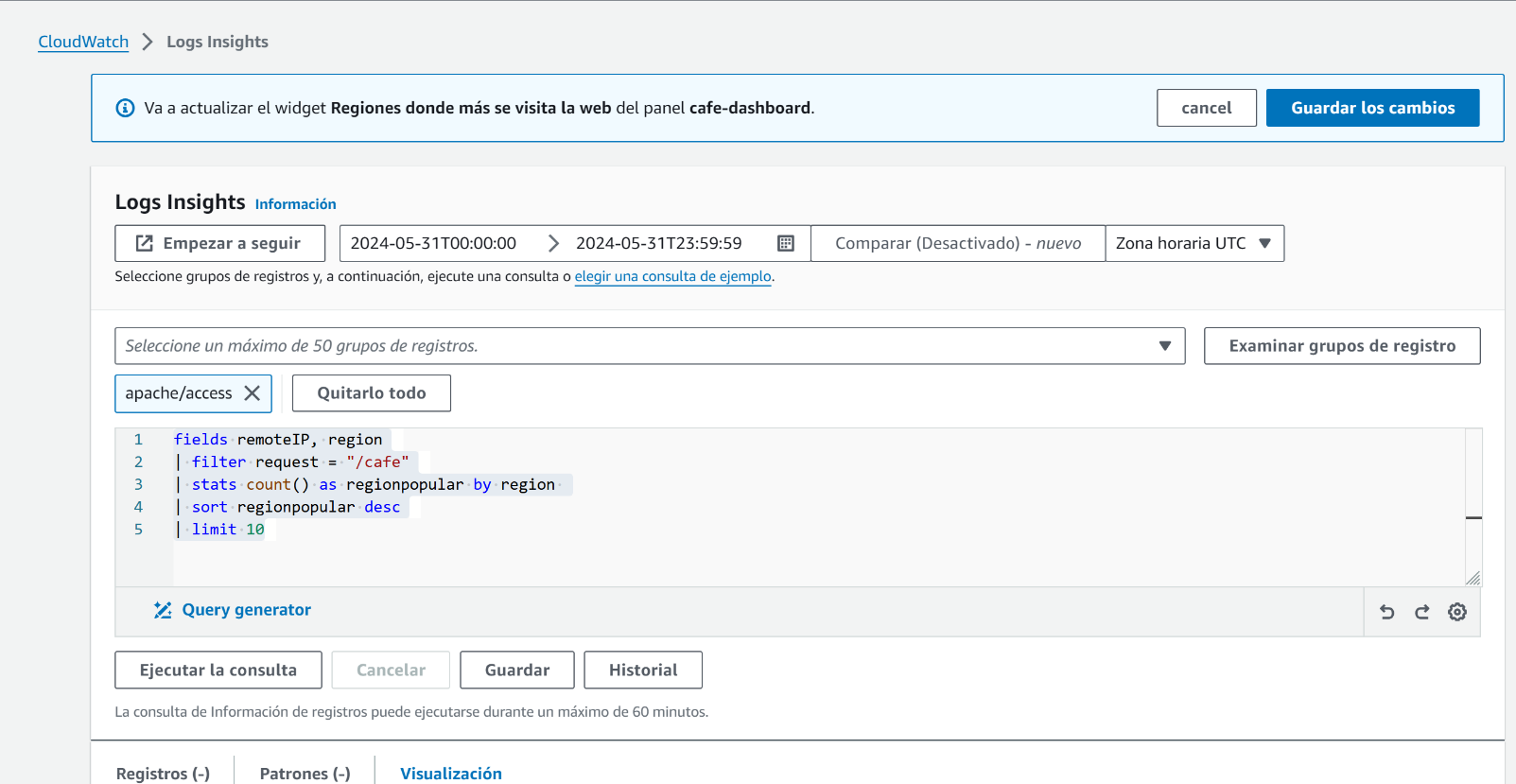


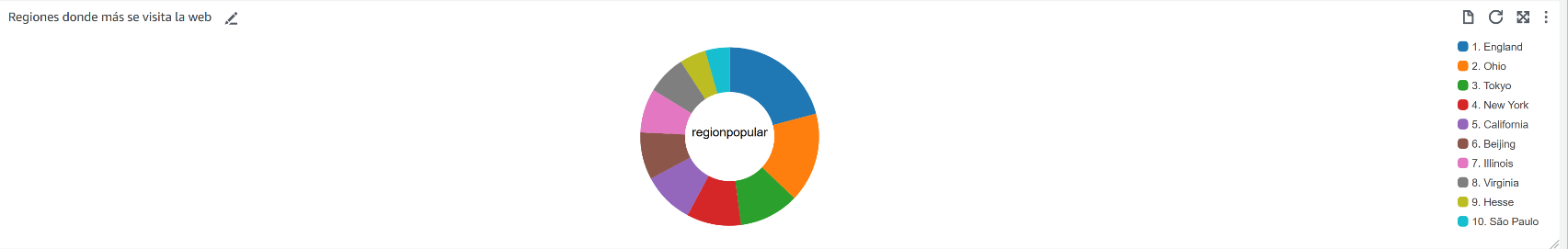
* Una tabla de registro que muestra las 10 ciudades que tuvieron la mayor cantidad de visitantes al sitio web que realizaron una compra.



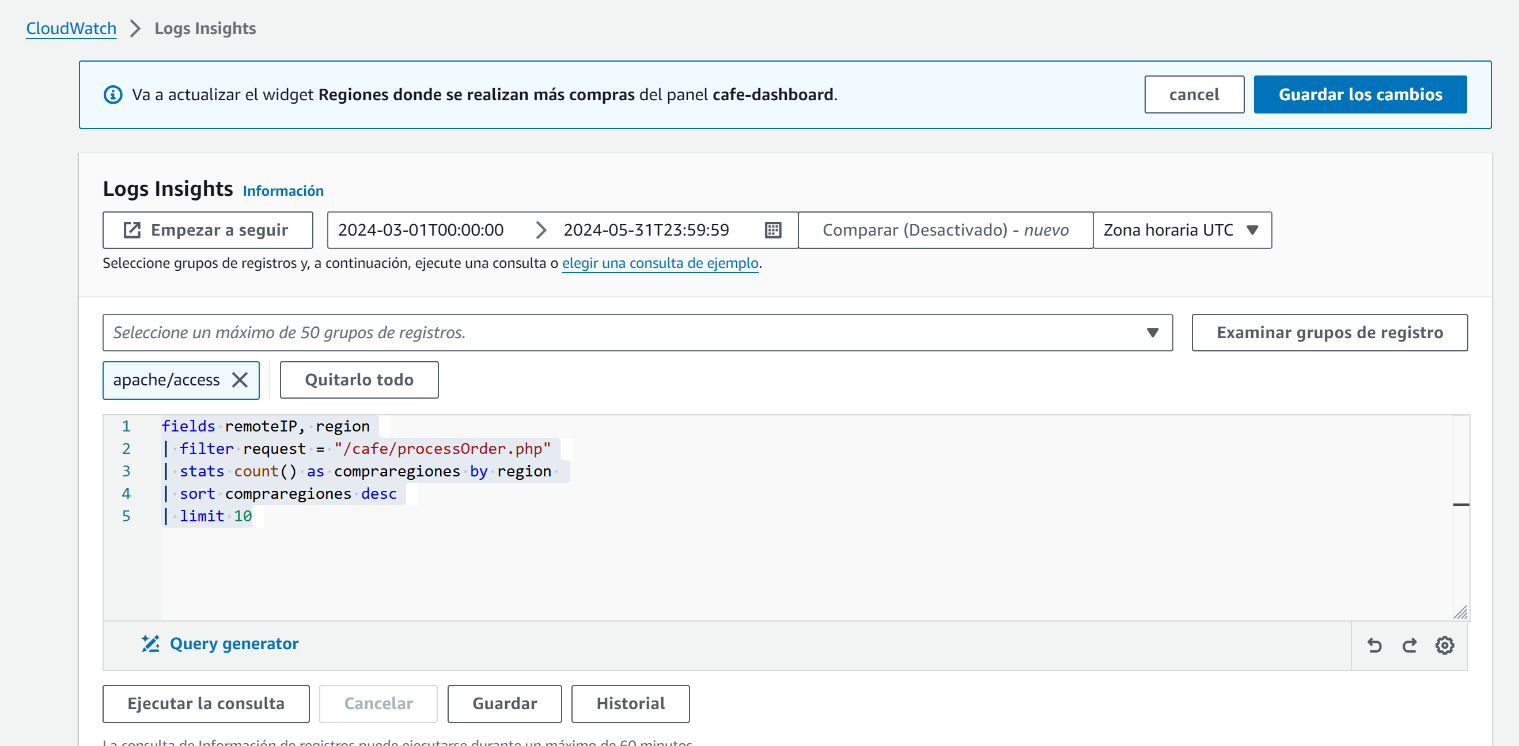


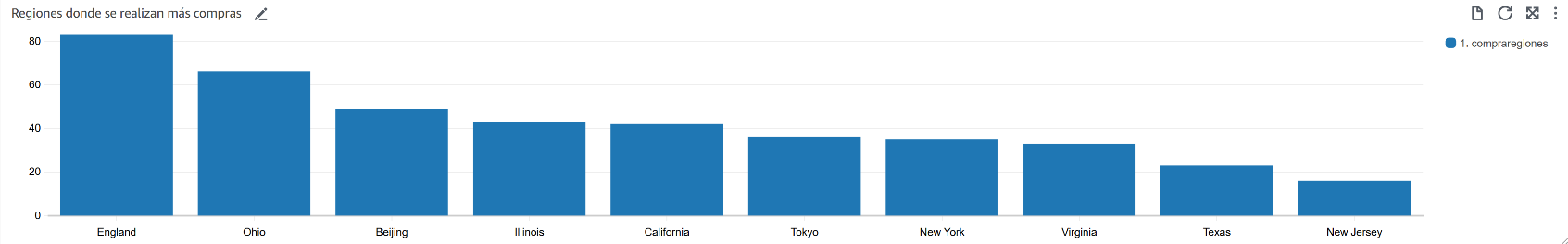
* Un gráfico circular que muestra las 10 regiones que tuvieron la mayor cantidad de visitantes que accedieron a la página principal del sitio web.





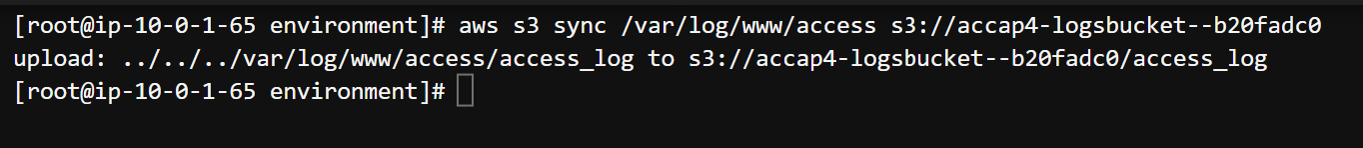
* Un gráfico de barras que muestra las 10 regiones que tuvieron la mayor cantidad de visitantes al sitio web que realizaron una compra.



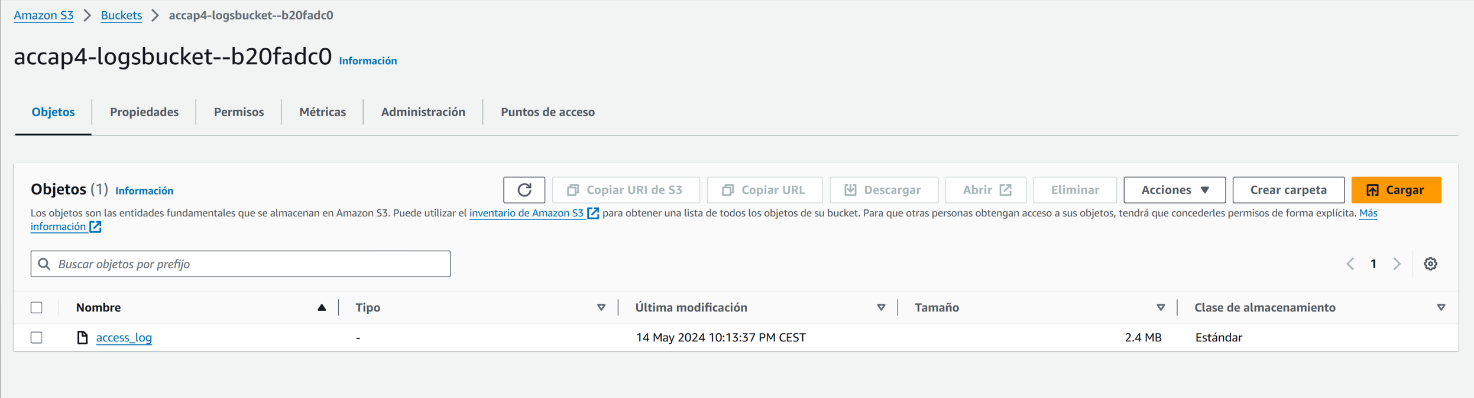


# 6. Guardar los registros en servicio S3 para posteriores ampliaciones y otorgar flexibilidad con otros servicios.

Para terminar, con objetivo de tener tanto un backup, y expandir la flexibilidad de servicios guardaremos los registros en un servicio de almacenamiento S3, para ello usaremos el siguiente comando que sincronizara los datos a la hora de traspasarlos, haciendo un almacenamiento completo de todos los registros presentes en el servidor web.



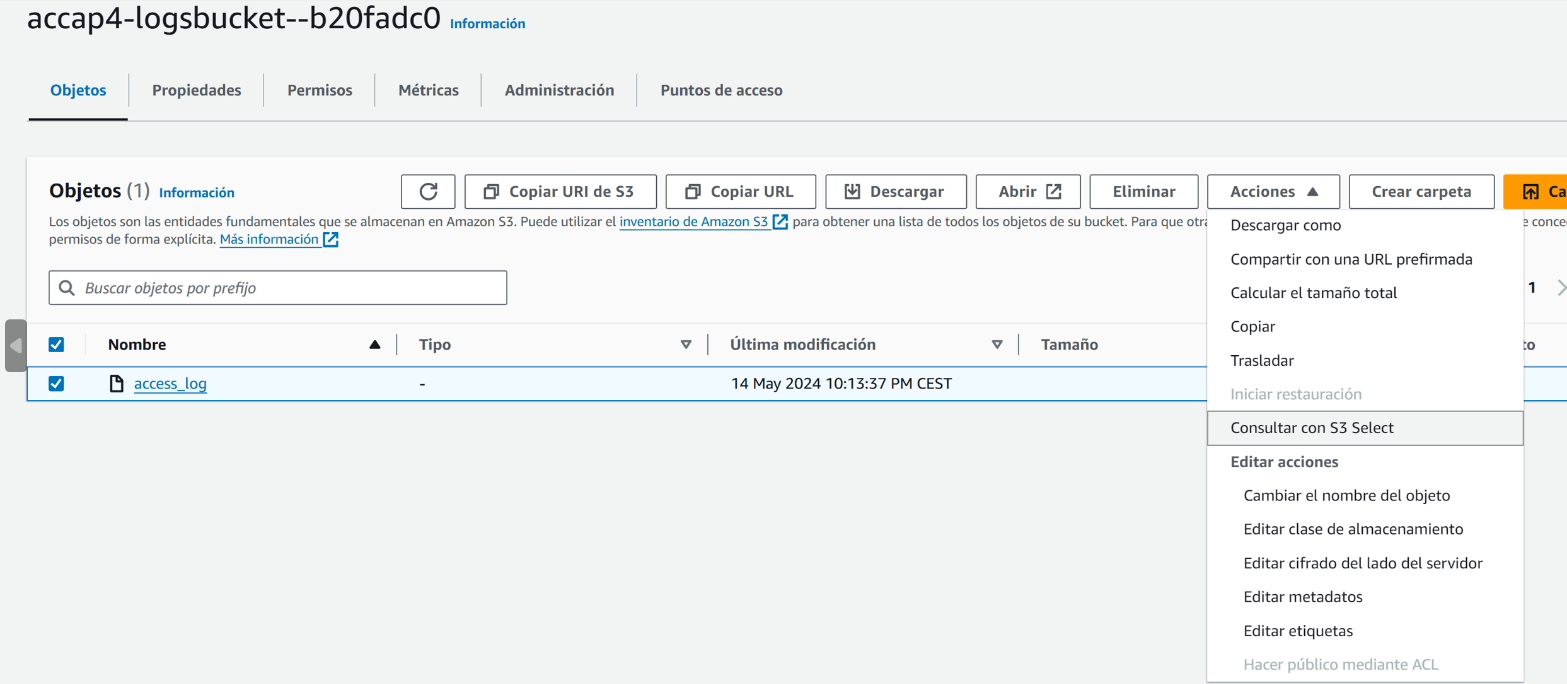
Una vez hecho eso nos dirigiremos a la instancia de S3 creada y veremos en los objetos que efectivamente se han traspasado los registros.



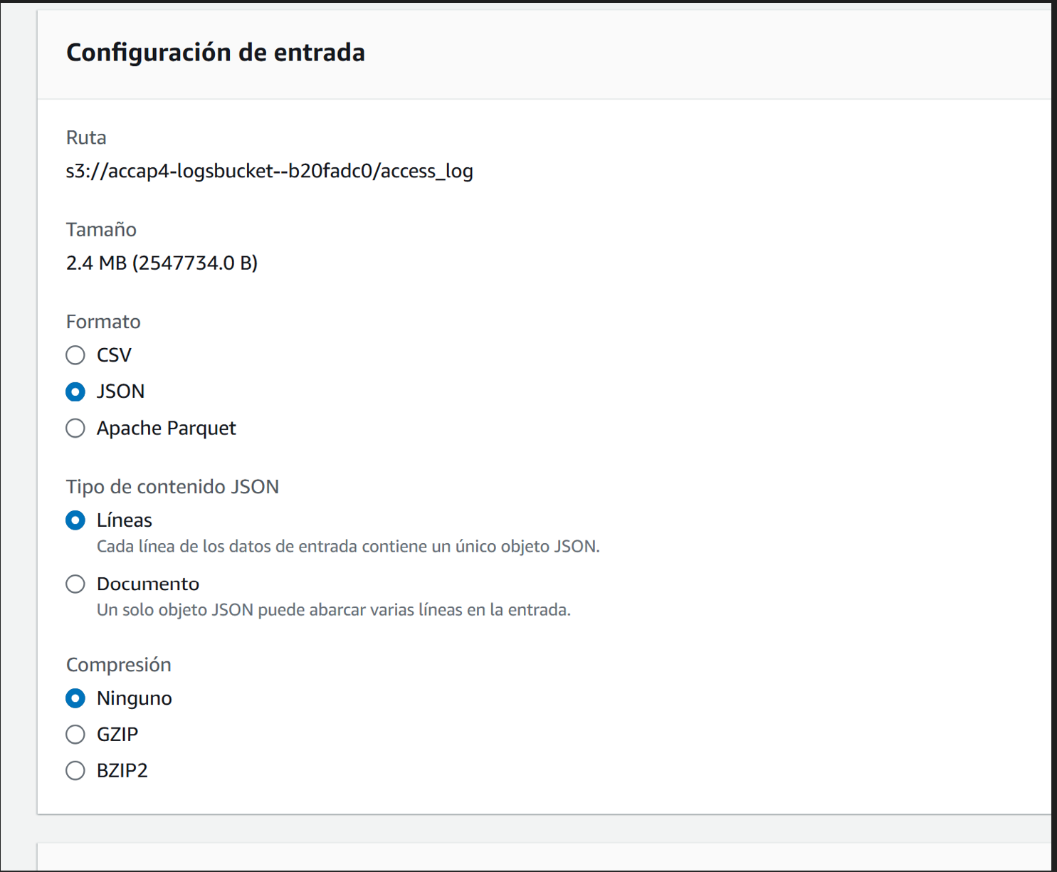
## 6.1 Comprobar que podemos realizar consultas con el servicio de S3 Select

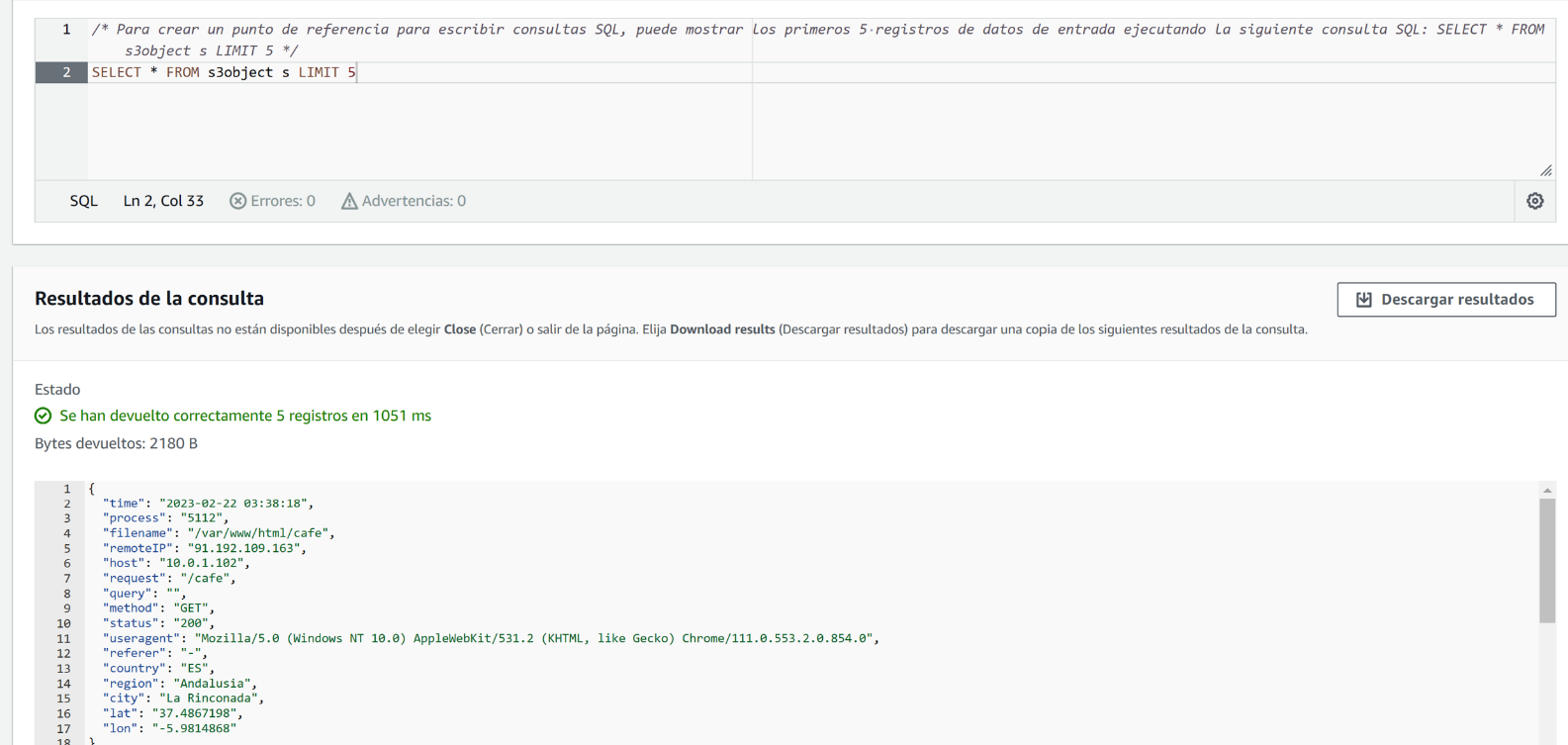
Como último paso comprobaremos que efectivamente los registros se han traspasado de manera adecuada, para ello usaremos el servicio de S3 Select que permite realizar consultas utilizando el lenguaje de SQL.

Para ello en Acciones accederemos a la opción de “Consultar con S3 Select”



Podremos especificar el formato en el que se encuentran los registros además de algunos parámetros más, en nuestro caso indicaremos que estamos trabajando con el formato JSON y En lineas.



Y como vemos podemos comprobar con una simple consulta que efectivamente los registros se han traspasado adecuadamente.