Implantación de Sistemas Canalización de Datos para análisis de Información

Contenido

[1. Introducción 3](#_Toc168944115)

[2. Objetivos y Requisitos 4](#_Toc168944116)

[- OBJETIVO DEL PROYECTO 4](#_Toc168944117)

[- REQUISITOS DE LA SOLUCIÓN 5](#_Toc168944118)

[3. Estudio Previo 6](#_Toc168944119)

[- ESTADO ACTUAL 6](#_Toc168944120)

[- POSIBLES SOLUCIONES 6](#_Toc168944121)

[4. Plan de Trabajo 7](#_Toc168944122)

[5. Diseño 8](#_Toc168944123)

[- DISEÑO GENERAL 8](#_Toc168944124)

[- DISEÑO DETALLADO 8](#_Toc168944125)

[6. Implantación 10](#_Toc168944126)

[7. Recursos 11](#_Toc168944127)

[- HERRAMIENTAS HARDWARE 11](#_Toc168944128)

[- HERRAMIENTAS SOFTWARE 12](#_Toc168944129)

[- PERSONAL 16](#_Toc168944130)

[- PRESUPUESTO 18](#_Toc168944131)

[8. Conclusiones 19](#_Toc168944132)

# 1. Introducción

La información sobre las preferencias de los clientes es uno de los recursos más necesarios e importante a la hora de emprender prácticamente cualquier proyecto de negocio, para obtener dicha información, a lo largo del tiempo se han llevado a cabo diferentes técnicas para atraer la atención de los clientes a posibles productos de su interés, recogiendo los resultados a posteriori y analizándolos con intención de desarrollar tácticas de mercado sobre dicha información.

Hoy en día las técnicas de recolección de información se han adaptado a las nuevas tecnologías, aprovechando la popularidad de las páginas web y, la compra online de productos y servicios.

En este proyecto se trabajara sobre un escenario ficticio, en el cual se llevara a cabo la implementación de un sistema de canalización de datos sobre los clientes que acceden a la página web de una franquicia de cafeterías llamada AnyCompany Café, con intención de que la empresa pueda utilizar dicha información para tomar decisiones sobre dónde invertir a futuro.

# 2. Objetivos y Requisitos

## - OBJETIVO DEL PROYECTO

La situación planteada sería la siguiente:

La empresa AnyCompany Café vende postres y artículos de café a través de su página web. Esta posee múltiples cafeterías en varias ciudades alrededor del mundo.

Y la empresa está interesada en obtener información sobre su negocio mediante el uso de datos sobre como las personas interactúan con su sitio web. Con ello la empresa planea analizar las tendencias de los datos del flujo de clicks, con intención de tomar decisiones inteligentes sobre dónde invertir en el futuro.

Por ello AnyCompany Café contrataría a mi empresa de consultoría para que llevara a cabo este proyecto.

Los datos de registro del flujo de clics del sitio web de la cafetería incluyen una entrada por cada clic que hace un cliente potencial mientras navega por el sitio web. El objetivo del proyecto por ende consiste en diseñar y crear un canal de análisis de datos para capturar la información del flujo de clics. Además se creará un panel de análisis de manera que el dueño de las cafeterías pueda observar rápidamente el comportamiento de los clientes.

De esta manera se busca utilizar la información obtenida sobre el comportamiento de los usuarios en la pagina web para centrar los esfuerzos publicitarios eficientemente, e incluso decidir donde abrir nuevos locales.

## - REQUISITOS DE LA SOLUCIÓN

Conociendo el objetivo, los requisitos necesarios para la solución son los siguientes:

* Diseñar y optimizar los costos de la solución antes de la implantación de la misma
* Asegurarnos que la solución sea funcional y rentable
* Transformar los datos recogidos para su posterior analisis
* Registrar los datos del Access\_log del servidor web de la cafetería
* Usar datos de registro simulados
* Analizar y visualizar los datos
* Generar conocimiento e información a partir de los datos recogidos

# 3. Estudio Previo

## - ESTADO ACTUAL

En la situación desde la que partimos tenemos que contamos con una pagina web implementada por la propia empresa de AnyCompany Café, así como un nodo EC3, el cual en este caso supuesto se usará como host de la misma página web, además que podrá acceder a los servicios de AWS Systems Manager, Amazon S3, and CloudWatch para enviar los datos.

En el equipo utilizado se utilizará Apache2 y el servicio de Cloud9 para la manipulación de archivos.

## - POSIBLES SOLUCIONES

Una de las soluciones más viables que podemos encontrar para el análisis de datos de esta clase, es la implantación y utilización del servicio de cloudwatch proporcionado por AWS para la canalización, almacenamiento y análisis de datos mediante consultas programadas en el tiempo, de manera que se pueda llevar una monitorización clara y concisa de las visitas a la web y las ventas realizadas a lo largo del tiempo.

También convendría permitir el almacenamiento de los logs recogidos, en un servicio s3 para su posterior utilización mediante diferentes posibles servicios, esto otorgaría más flexibilidad para posibles cambios que puedan producirse como vista a futuro para la empresa.

# 4. Plan de Trabajo

Conociendo los objetivos de la solución tenemos que las fases del proyecto se pueden dividir en las siguientes:

| **Fases del Proyecto** | **Detalles** |
| --- | --- |
| 1 | Creación de un diagrama de arquitectura y coste estimado para la solución |
| 2 | Usar AWS Cloud9 para acceder al server web. Observar los registros de acceso que se generan cuando se navega en la web. |
| 3 | Instalar y configurar el agente de Amazon CloudWatch y el archivo de configuracion (httpd.conf) del servidor web de apache de manera que los logs de acceso y error puedan ser recogidos y enviados a CloudWatch. |
| 4 | Confirmar que el servidor web genera logs que el agente de CloudWatch puede recoger y enviar a CloudWatch. |
| 5 | Usar log simulados para verificar que CloudWatch recibe los logs correctamente. |
| 6 | Usar CloudWatch Logs Insights para hacer consultas al grupo de logs de acceso y generar visualizaciones. |
| 7 | Ajustar el proceso de analisis de datos para ofrecer nueva información con los datos de geolocalizacion de los visitantes. Crear un panel de visualizacion para que el dueño del Cafe pueda tomar decisiones en funcion de la información recogida |

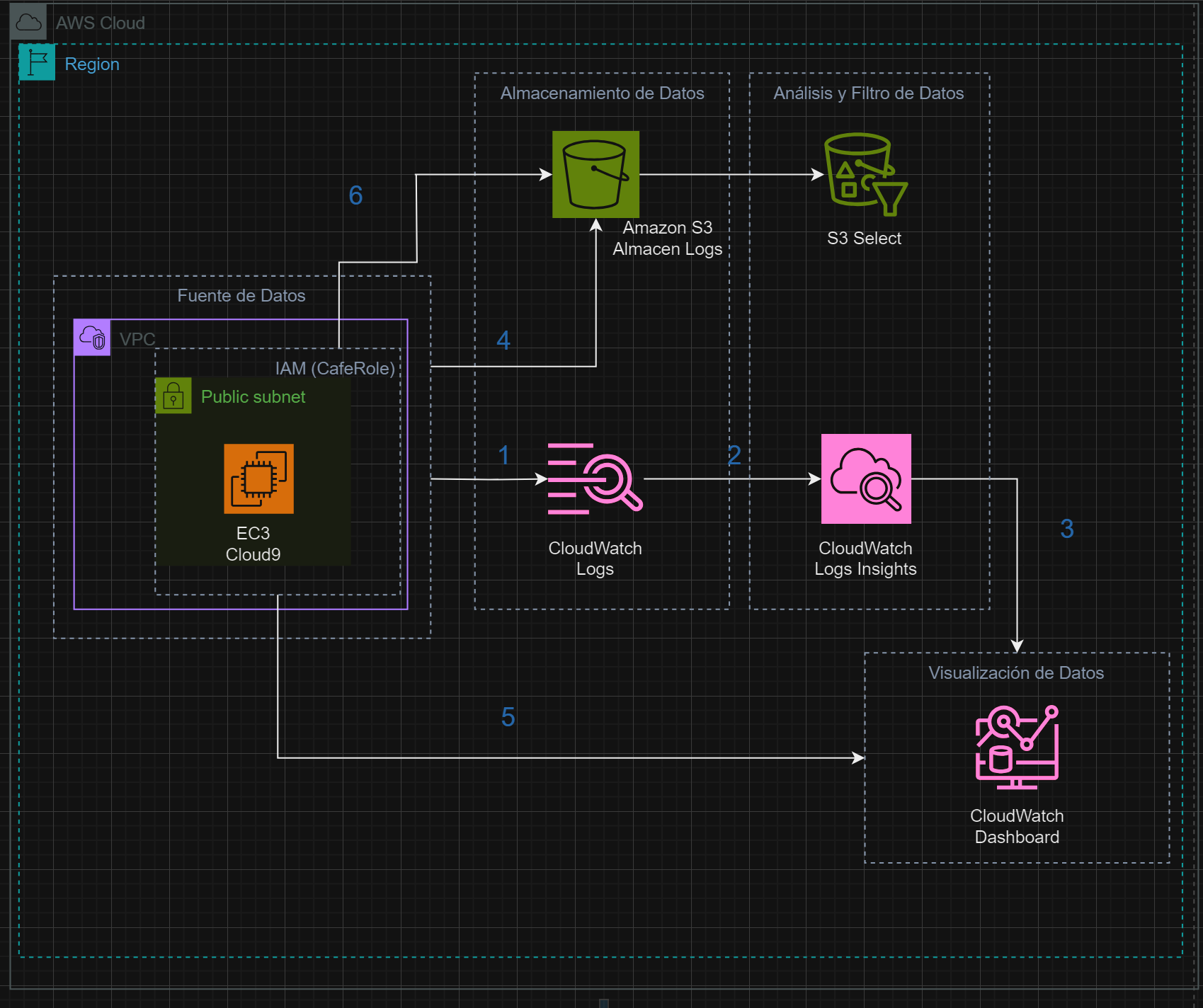
# 5. Diseño

## - DISEÑO GENERAL

La solución propuesta consistiría en la implantación y utilización del servicio de cloudwatch proporcionado por AWS para la canalización, almacenamiento y análisis de datos, de manera que se pueda llevar una monitorización clara y concisa de las visitas a la web y las ventas realizadas a lo largo del tiempo.

## - DISEÑO DETALLADO

La solución propuesta se puede ver ilustrada de mejor manera una vez finalizada la primera fase del proyecto, una vez creado el diagrama de la arquitectura del sistema que se planteará:



1. Los registros se recogen desde los logs de los servidores web y se envían a Cloudwatch almacenándose en el servicio
2. Posteriormente dichos registros almacenados son procesados por el servicio de CloudWatch Logs Insights donde podrán filtrados mediante consultas.
3. Se obtiene información sobre los movimientos a lo largo del tiempo en la web mediante consultas con los parámetros requeridos utilizando el servicio de CloudWatch Logs Insights, posteriormente dicha información se muestra actualizada en el panel de muestra de CloudWatch Dashboard.
4. Los registros de igual manera se recogen desde los logs del servidor web para ser enviados a un almacenamiento Bucket S3, de manera que se pueda acceder a ellos utilizando consultas SQL con el servicio S3 Select, de esta forma nos aseguramos que en el futuro se pueda acceder y filtrar los datos utilizando otra clase de servicios, con visión a posibles expansiones de la estructura.
5. Se contara con un rol IAM CafeRole el cual estará vinculado a una EC3 y que permitirá el envío de datos de datos a CloudWatch
6. Igualmente el rol IAM CafeRole permitirá el envió de datos a instancias S3

# 6. Implantación

En este paso vamos a proceder a la instalación e implementación de la solución que hemos propuesto, para ello utilizaremos las herramientas software de aws mencionadas anteriormente, instancias EC3, Cloud9, AIM, CloudWatch y Amazon S3.

La implementación llevada a cabo, con sus pasos explicados más a profundidad son documentados en el manual de implementación que presento a continuación:

# 7. Recursos

Para la elaboración de este proyecto el grueso de los recursos se concentra en el apartado de software puesto que los servicios son proporcionados por Amazon Web Services, a continuación se enumerara los recursos necesitados para la puesta en marcha del proyecto y su funcionamiento adecuado.

## HERRAMIENTAS HARDWARE

En este proyecto los recursos hardware necesarios son básicos, **contando principalmente con la necesidad de un equipo informático preparado para poder realizar la administración y configuración necesaria por parte del soporte técnico, en los servicios** de aws.

Además, también pueden ser necesario equipos adicionales para la monitorización de los datos, si es que el cliente así lo solicita.

Por lo demás, **también es imprescindible contar con una conexión red estable** puesto que, tanto el almacenamiento, como las aplicaciones utilizadas y demás infraestructura se encuentran en los servidores de Amazon, de manera que si existieran cortes en la red, no se podrían acceder a estos correctamente.

## HERRAMIENTAS SOFTWARE

Donde se encuentra la mayor carga de recursos en este proyecto es en las herramientas software utilizadas, esto debido a que el mismo se desarrolla sobre la plataforma de Amazon Web Services, la cual contiene, no solo los servicios y aplicaciones que se usaran en la implantación de la canalización de datos, sino también las propias instancias donde se levantan los servidores web y el servicio para el almacenamiento de seguridad de los registros.

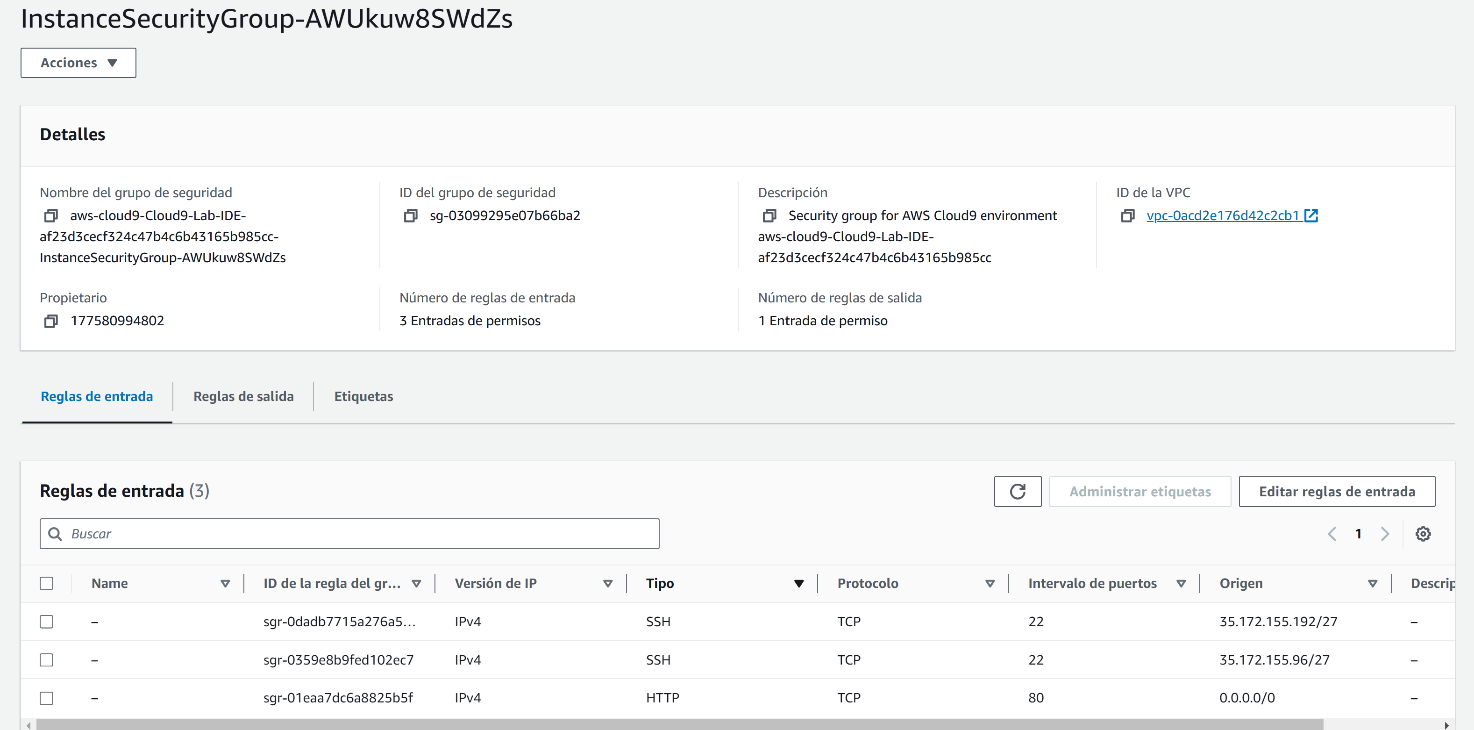
De esta manera el proyecto se encuentra desarrollado en su gran parte en la nube de los servidores de Amazon, esto permite que una empresa no deba contar necesariamente con infraestructura de servidores físicos, ahorrando dinero en la cara implantación que ello supondría, sobre todo para empresas medianas o pequeñas.

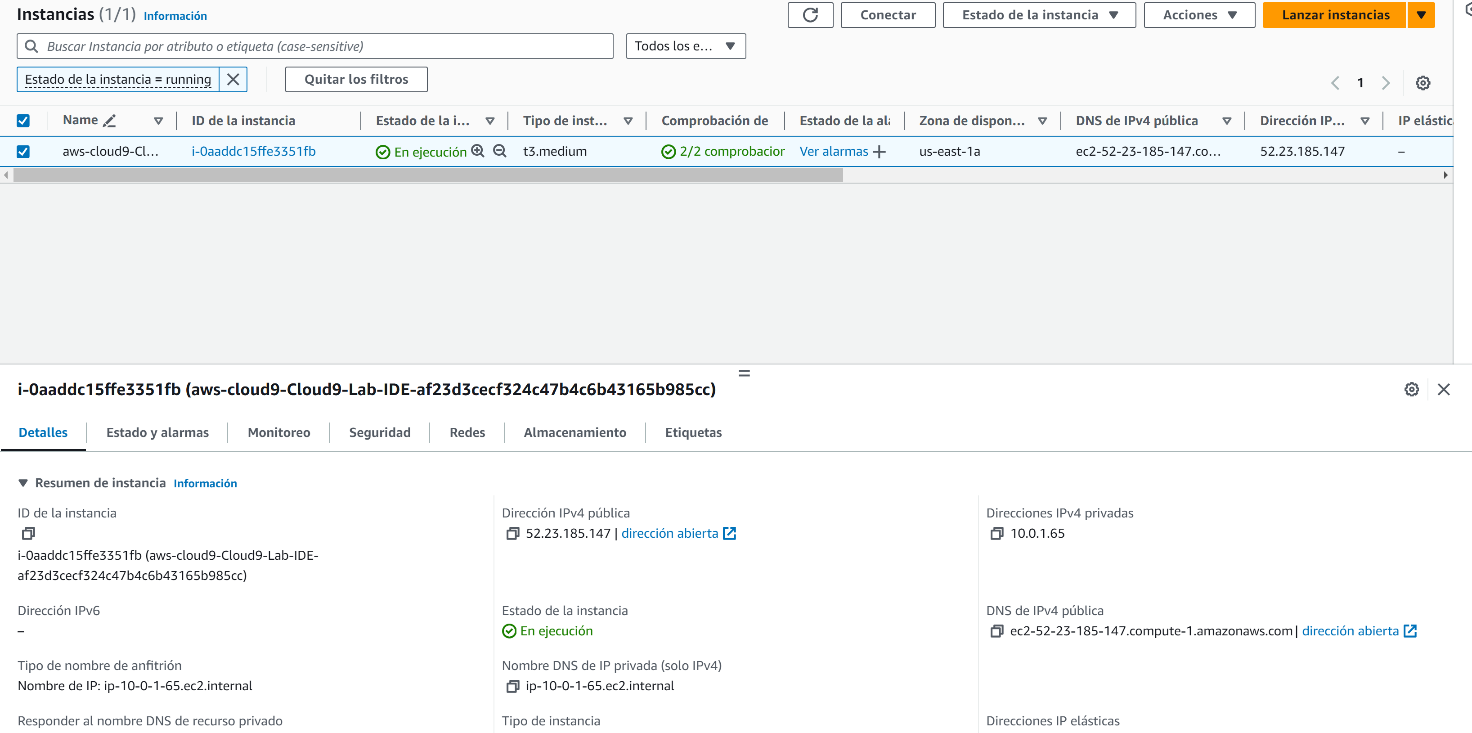
Las herramientas software utilizadas para este proyecto se enumeran como las siguientes:

1. **EC2**: El servicio de Amazon Elastic Compute Cloud o por sus siglas Amazon EC2, se trata de un servicio que permite a los usuarios crear instancias virtuales, las cuales se pueden utilizar como si se tratara de un equipo informático común utilizando los servidores de amazon.

En estas instancias serán donde se encuentra alojado tanto el servidor web, como la base de datos de la propia página web de la franquicia de cafeterías.

En este caso se puede ver como la instancia cuenta con IP publica para el acceso y se encuentra securizada limitando el tipo de protocolos y direcciones IP a los que responde.





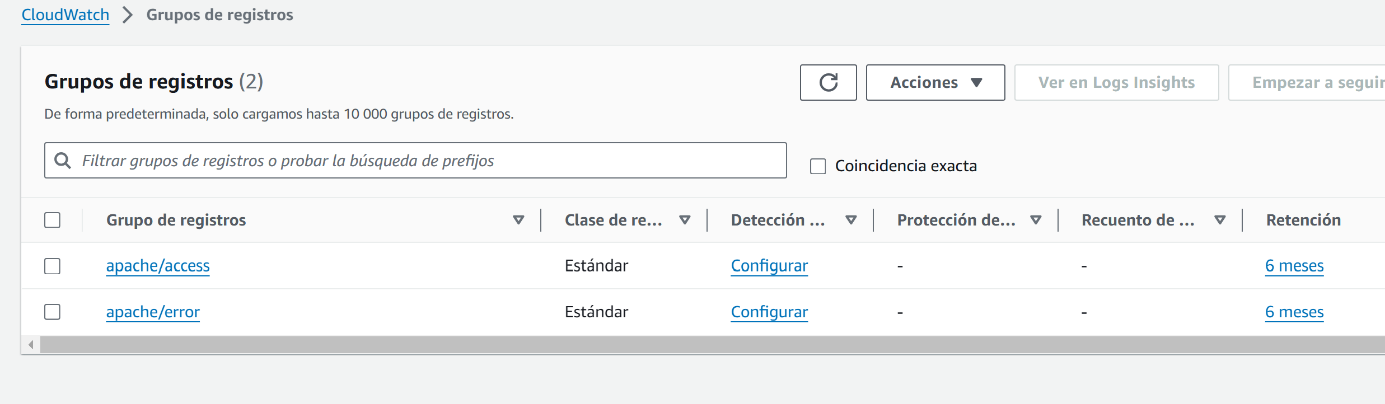
1. **Cloud9**:Es un entorno de desarrollo integrado (IDE) basado en la nube que permite a los desarrolladores escribir, ejecutar y depurar código utilizando solo un navegador web. Ofrece un editor de código completo, un terminal y un depurador potente.

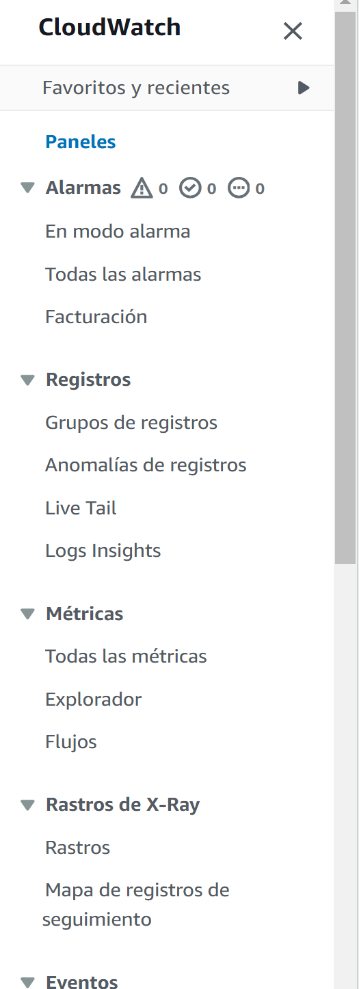
En este caso se utilizara para mayor comodidad a la hora de la configuración e integración de aplicaciones en la instancia EC2.

1. **VPC**: Es un servicio de amazon de computación en la nube que brinda a los usuarios una nube privada virtual, en este caso se utilizara para asignar una red a las instancias EC2, además también se creara una puerta de enlace a internet para dar acceso a la máquina hacia el exterior.
2. **CloudWatch**: Es el principal servicio que usaremos para poder monitorear los registros e información generada con las visitas de los clientes al sitio web.

Se trata de un servicio el cual permite monitorear y recopilar métricas, que se tratan de variables que se usan para medir recursos y aplicaciones. Además permite crear paneles personalizados para mostrar métricas específicas de manera clara y sencilla sobre sus aplicaciones.

También puede crear alarmas que observen métricas y envíen notificaciones o realicen cambios automáticamente en los recursos que está monitoreando cuando se supera un umbral. Por ejemplo, puede monitorear el uso de la CPU y las lecturas y escrituras en disco de sus instancias Amazon EC2 y luego usar esos datos para determinar si debe lanzar instancias adicionales para manejar una mayor carga.





1. **S3**: Es un servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento.

En este caso se utilizará para poder almacenar los registros previamente enviados al servicio de cloudwatch, de esta manera podemos conseguir elasticidad para expandir el proyecto con otros tipo de servicios, como por ejemplo S3 Select.

## PERSONAL

En el apartado se detalla el equipo humano necesario para llevar a cabo el proyecto, incluyendo los roles, responsabilidades.

Que en este caso consta de un equipo bastante reducido debido a la automatización del proyecto, y a las limitaciones, sin embargo esto puede variar dependiendo de las circunstancias y las dimensiones de la empresa.

**- Roles y Responsabilidades**

1. Administrador e Integración de Servicios

* **Responsabilidades**:
  + Planificación e implantación de Servicios.
  + Coordinación del equipo.
  + Gestión de riesgos y resolución de problemas.

Sera quien se encargue principalmente del seguimiento del proyecto y la integración de los servicios y aplicaciones necesarias, tendrá acceso de administrador y recursos requeridos para la correcta implementación del proyecto, además también ayuda a solventar los posibles problemas que puedan surgir**.**

2. Monitoreo de Datos y Errores

* **Responsabilidades**:
  + Monitorización de Funcionamiento
  + Identificación de Errores
  + Solución de Errores, y mantenimiento de Integridad y Mantenimiento de Datos

Sera el grupo de personas que se encargue de monitorizar que el flujo de datos está funcionando correctamente y transmitiendo datos de manera adecuada, además de verificar que se realizan copias de seguridad de la información.

Identificaran los Errores o posibles errores que puedan surgir con los servicios de datos, ayudando a resolverlos lo antes posible.

El tamaño del grupo es variable dependiendo del tamaño y las sedes de la empresa.

3. Análisis de Datos

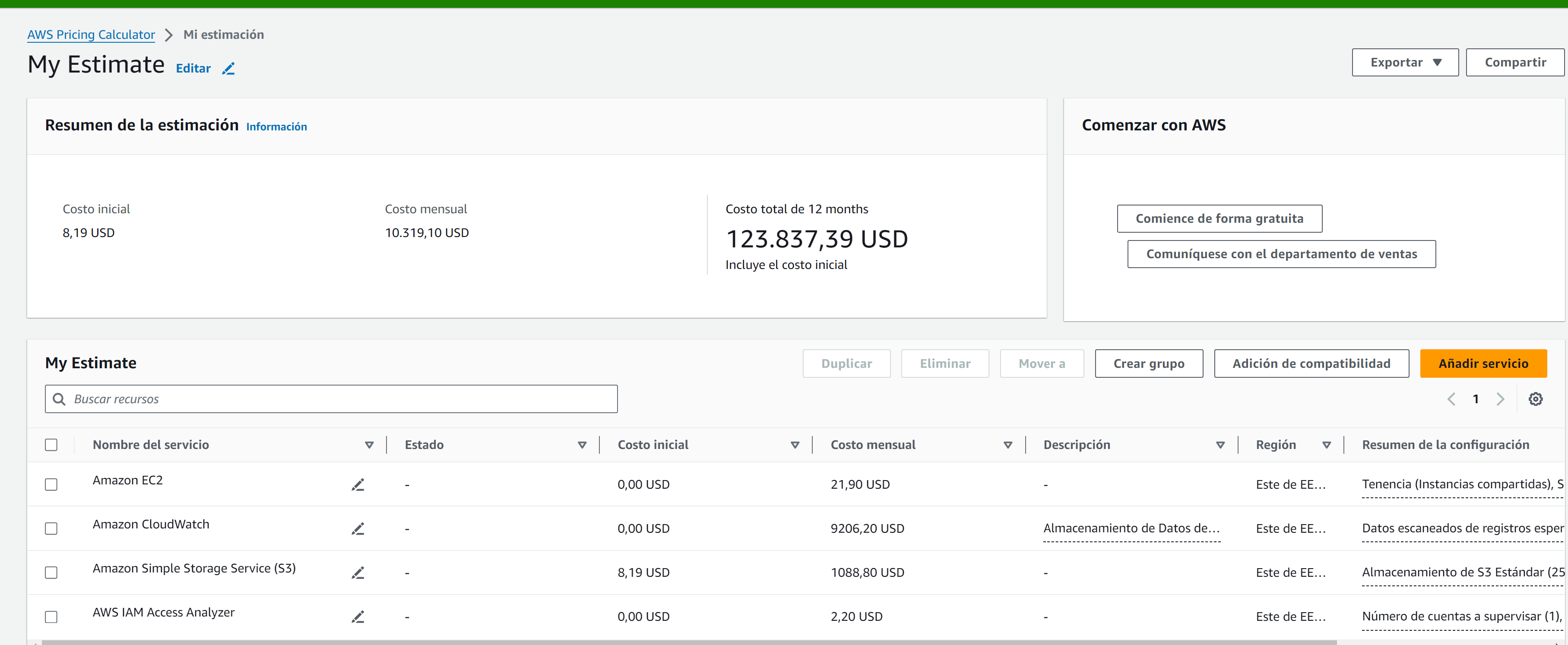
* **Responsabilidades**:
  + Análisis de la tendencia de los datos registrados
  + Asesorar sobre los tipos de datos en los que concentrar la atencion

Como ayuda adicional, un grupo de personas que se encargara de analizar la tendencia de los datos registrados, ayudando a determinar en qué clase de datos concentrar la atención y monitorear.

**Cabe destacar que la última palabra en este aspecto siempre la tendrá el cliente**, pues este grupo solo se encargara de aconsejar y asesorar de la mejor manera posible.

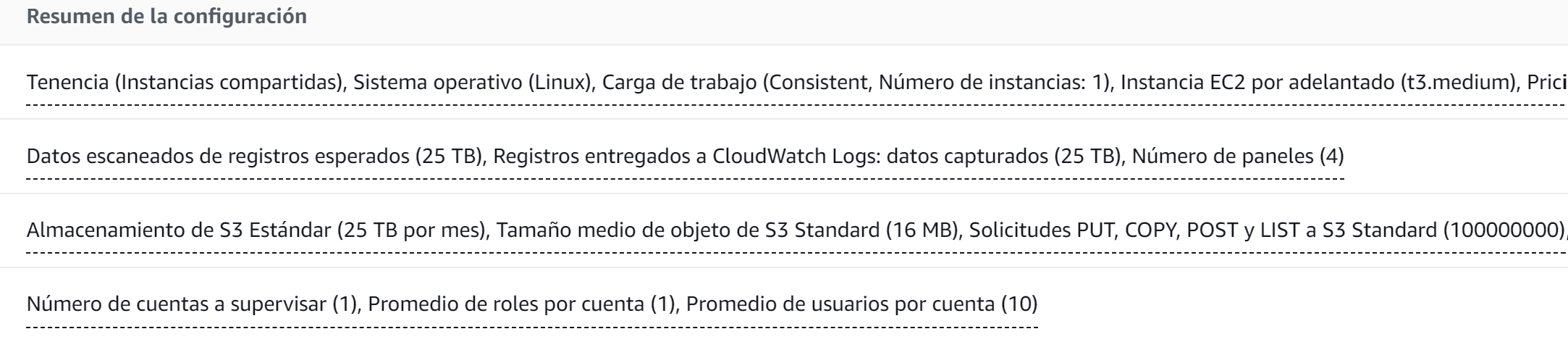
## PRESUPUESTO

El presupuesto necesario cumpliendo todos los requerimientos de servicios ascendería a 10.319 $ al més lo que haría unos 123.837 $ al año.



El precio asciende de tal manera debido a las dimensiones de la empresa, se ha tenido en cuenta que se trata de una supuesta empresa a escala internacional, que cuenta con múltiples cafeterías múltiples ciudades a lo largo de todo el globo, esto hace que la necesidad por capacidad de almacenamiento de registros de clientes potenciales ascienda exponencialmente

Como resultado, la capacidad media inicial de almacenamiento de datos tanto en Cloudwatch como S3 por la que se ha optado para realizar el aproximado, es de unos 25 TB de almacenamiento los cuales se refrescaran mensualmente.

Sin embargo esta capacidad puede variar dependiendo del tamaño real de la empresa y la carga de datos.

# 8. Conclusiones

## GRADO DE CONSECUCION DE OBJETIVOS

1. **Creación de un diagrama de arquitectura y coste estimado para la solución – TOTALMENTE COMPLETADO**
2. **Usar AWS Cloud9 para acceder al server web. Observar los registros de acceso que se generan cuando se navega en la web – TOTALMENTE COMPLETADO**
3. **Instalar y configurar el agente de Amazon CloudWatch y el archivo de configuracion (httpd.conf) del servidor web de apache de manera que los logs de acceso y error puedan ser recogidos y enviados a CloudWatch – TOTALMENTE COMPLETADO**
4. **Confirmar que el servidor web genera logs que el agente de CloudWatch puede recoger y enviar a CloudWatch – TOTALMENTE COMPLETADO**
5. **Usar log simulados para verificar que CloudWatch recibe los logs correctamente – TOTALMENTE COMPLETADO**
6. **Usar CloudWatch Logs Insights para hacer consultas al grupo de logs de acceso y generar visualizaciones - TOTALMENTE COMPLETADO**
7. **Ajustar el proceso de analisis de datos para ofrecer nueva información con los datos de geolocalizacion de los visitantes. Crear un panel de visualizacion para que el dueño del Cafe pueda tomar decisiones en funcion de la información recogida – MAYORMENTE COMPLETADO:**

**Si bien el objetivo esta completado casi en su totalidad, este se podría redondear totalmente si se pudiera aplicar un módulo como GeoIP2 al servicio apache2, de manera que este pudiera recoger datos de geolocalización en tiempo real para posteriormente enviarlos al servicio de Cloudwatch, en lugar de introducir dichos registros de geolocalizacion manualmente, como se ha realizado.**

**Sin embargo debido a las limitaciones de seguridad del laboratorio en el que se ha realizado el proyecto, no se ha podido implementar dicho modulo.**

## PROBLEMAS ENCONTRADOS

Los principales problemas encontrados durante el proyecto han tenido que ver con los permisos limitados de los laboratorios utilizados, por ejemplo, debido a que es necesario contar con ciertos roles de acceso para que el agente pueda trasmitir datos al servicio de Cloudwatch, es imprescindible poder crear nuevos roles y asignar permisos, sin embargo estas funciones están completamente bloqueadas en laboratorios comunes Learner Lab, de manera que es imposible realizar la implementación correctamente en un laboratorio normal de enseñanza.

Por otro lado, el laboratorio facilitado para realizar el proyecto, si bien cuenta con los roles necesarios, a la hora de **ampliar las posibilidades**, como por ejemplo, añadiendo el módulo GeoIP2 al servicio apache2 que permite el registro de datos de geolocalización, este se encuentra limitado por restricciones de seguridad y también debido a tener que utilizar el SO de Amazon Linux 2 de manera forzada.

Aparte de esos inconvenientes el proyecto no ha presentado mayor problema en la mayor parte de su implementación.

## FUTURAS MEJORAS

Como futuras mejoras podemos pensar en la aplicación de otros servicios sobre los registros almacenados en S3 de manera que se puedan usar aplicaciones externas para el acceso a la información.

También una posible mejora consistiría en aplicar un módulo para la detección en tiempo real de la información de geolocalización sobre los visitantes a la página web.

Y por último otra posibilidad interesante a considerar seria, utilizar el servicio de alertas que ofrece cloudwatch para mantener controladas las fluctuaciones de las compras en determinadas zonas o ciudades, de manera que se pueda realizar un control de los recursos necesarios en cada lugar, y se pueda realizar la entrega de los mismos con antelación.

# 9. Bibliografía

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/Stop_Start.html>

<https://docs.aws.amazon.com/cloud9/latest/user-guide/tutorials-basic.html>

<https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/install-CloudWatch-Agent-commandline-fleet.html#start-CloudWatch-Agent-EC2-commands-fleet>

<https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/troubleshooting-CloudWatch-Agent.html#CloudWatch-Agent-troubleshooting-stopping-restarting>

<https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/AnalyzingLogData.html>

<https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/CloudWatch_Dashboards.html>

<https://docs.aws.amazon.com/cli/latest/reference/s3/#cli-aws-s3>

<https://docs.aws.amazon.com/AmazonS3/latest/userguide/selecting-content-from-objects.html>

<https://github.com/johnjyin/Install-LAMP-Magento-on-AWS-EC2/blob/master/Install%20GeoIP%20on%20AWS%20EC2.txt>

<https://www.ip2location.com/development-libraries/ip2location/apache>