

Infraestructura para la Computación de Altas

Prestaciones

Proyecto Final: AquaSenseCloud

Francisco Javier Mercader Martínez

Javier Moreno

Pablo Meseguer

Curso: 2024/2025

Índice

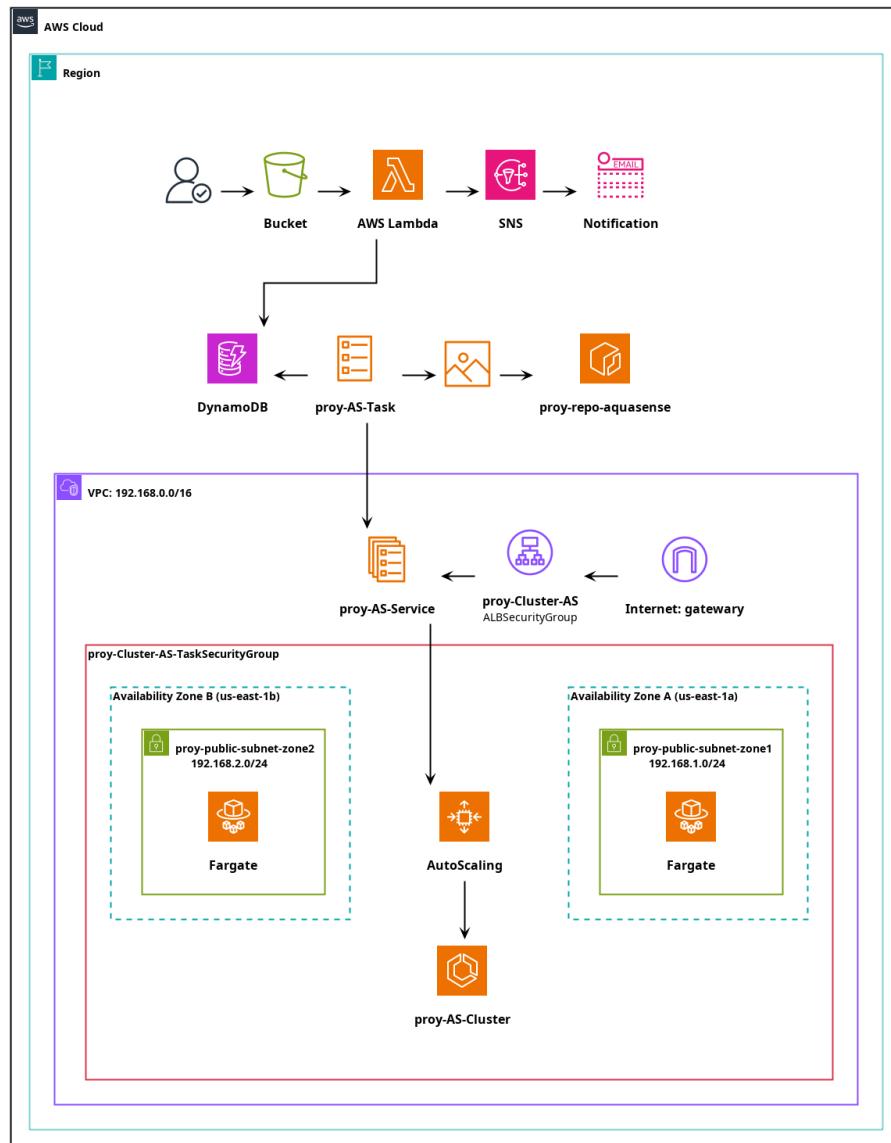
1	Cronograma	2
2	Diagrama Arquitectura de la solución	4
3	Listado de Recursos y Servicios con su funcionalidad	4
4	Ejemplos de demostración de correcta funcionalidad	8

1 Cronograma

Fecha	Tarea	Responsable	Duración
11 de noviembre	Inicio y Diseño de Arquitectura: Definición de requisitos, análisis del dataset CSV y diseño de la arquitectura completa en AWS y endpoints de la API.	Todo el Equipo	1 día
12-13 de noviembre	Infraestructura de Red Base: Creación de VPC, subnets, Security Groups, bucket S3 para datos crudos y tablas en DynamoDB.	Javier Moreno	2 días
14-15 de noviembre	Lógica de Negocio y Pipeline: Desarrollo de funciones Lambda para procesamiento de CSV y cálculos estadísticos (maxdiff, temp mensual).	Pablo Meseguer	2 días
16-17 de noviembre	Configuración API Base: Configuración del API Gateway y desarrollo de Lambdas para endpoints REST básicos (/sd, /temp).	Fco. Jav. Mercader	2 días
18-19 de noviembre	Seguridad y Automatización: Configuración de roles IAM, políticas de seguridad y creación de plantillas CloudFormation base (IaC).	Javier Moreno	2 días
20-21 de noviembre	Sistema de Alarms: Configuración de SNS Topic, integración de alertas por desviación de datos y pruebas de envío de emails.	Pablo Meseguer	2 días
22-24 de noviembre	Integración y Pruebas: Integración de todas las Lambdas con DynamoDB, manejo de errores en API y creación de suite de pruebas en Postman.	Fco. Jav. Mercader	3 días
25-26 de noviembre	Dockerización: Creación de Dockerfile optimizado para la API, construcción de imagen y subida al repositorio ECR.	Pablo Meseguer	2 días
27-28 de noviembre	Infraestructura Avanzada (ECS): Configuración de Application Load Balancer (ALB), Target Groups y creación del clúster ECS Fargate.	Javier Moreno	2 días
29 de nov - 1 de dic	Escalabilidad: Configuración de Auto Scaling Groups, políticas de escalado basadas en métricas y ejecución de pruebas de carga.	Fco. Jav. Mercader	3 días

Fecha	Tarea	Responsable	Duración
2 de diciembre	Documentación Técnica: Redacción de la memoria técnica, diagramas finales de arquitectura y justificación de decisiones.	Javier Moreno	1 día
3 de diciembre	Manual de Replicación: Elaboración de guía paso a paso para replicar el despliegue y scripts de limpieza de recursos.	Pablo Meseguer	1 día
4 de diciembre	Pruebas Finales y Despliegue: Despliegue final limpio en cuentas AWS Academy y validación "End-to-End" de todo el flujo.	Todo el Equipo	1 día
5 de diciembre	Entrega Final: Recopilación de entregables, preparación de material de defensa y envío del proyecto.	Todo el Equipo	1 día

2 Diagrama Arquitectura de la solución



3 Listado de Recursos y Servicios con su funcionalidad

1. Ingesta y almacenamiento crudo

- **Amazon S3 (AWS::S3::Bucket):** Almacena los archivos CSV de datos brutos con las medias y desviaciones típicas de temperaturas, se utilizan como entrada para el pipeline de datos. La ingestión de nuevos datos provoca el desencadenamiento de la función Lambda desarrollada, por lo tanto, nuestro bucket es asignado como desencadenador.

2. Pipeline de datos

- **AWS Lambda (AWS::Lambda::Function):** Ejecuta funciones que procesan los archivos almacenados en S3, transformando dichos datos y cargándolos en una tabla

DynamoDB. Esos datos almacenados en la tabla ya están preparados para ser consultados.

- **Amazon SNS (AWS::SNS::Topic):** Se utilizó para enviar notificaciones cuando la desviación estándar semanal de las temperaturas supera el umbral de 0.5. Esta modificación se publica en el tema SNS, informando a los suscriptores (en este caso los analistas) sobre la situación crítica en el monitoreo de temperaturas.

3. Almacenamiento de datos procesados/transformados

- **DynamoDB (AWS::DynamoDB::Table):** Almacena todos los datos transformados y procesados, listos para ser consultados, proporcionando así un acceso rápido y escalable a los datos.

4. Recursos para el desarrollo y pruebas iniciales

- **Instancia EC2 (AWS::EC2::Instance):** Utilizada para desarrollar, probar y ajustar el contenedor con la aplicación AquaSense. En esta instancia se configuró el entorno de Docker y se realizaron pruebas locales antes de enviar la imagen al repositorio de ECR.

5. Infraestructura de red y conectividad

- **VPC (AWS::EC2::VPC):** Proporciona un entorno de red aislado donde se despliegan los recursos de la infraestructura.
- **Subnets (AWS::EC2::Subnet):** Crea subredes públicas que permiten el acceso a Internet para los recursos dentro de la VPC.
- **Internet Gateway (AWS::EC2::InternetGateway):** Facilita la conexión de la VPC a Internet, permitiendo que los analistas accedan a los servicios configurados y desplegados.
- **Route Table (AWS::EC2::RouteTable):** Define las rutas de tráfico de red dentro de la VPC, asegurando que el tráfico fluya adecuadamente entre las subredes y el Internet Gateway.

6. Almacenamiento de imágenes

- **ECR (Amazon Elastic Container Registry):** Repositorio para almacenar y versionar la imagen del contenedor de AquaSense después de desarrollarla en EC2. Este recurso nos facilita la integración con ECS para el despliegue en producción de nuestra aplicación.

7. Recursos para el despliegue de la aplicación

- **ECSCluster (AWS::ECS::Cluster):** Agrupa, administra y organiza todas las tareas y servicios relacionados con la ejecución del contenedor de la aplicación. Actúa como punto central de administración para la ejecución de tareas, asegurando que estas puedan desplegarse de manera distribuida y eficiente.
- **ALB (AWS::ElasticLoadBalancingV2::LoadBalancer):** Balanceador de carga que gestiona/balancea el tráfico de los analistas hacia las tareas ECS. Redirige las solicitudes desde el puerto 80 (tráfico HTTP) al puerto 5000 del contenedor, además, mejora la disponibilidad y la tolerancia a fallos de nuestra aplicación.
- **ALBListener (AWS::ElasticLoadBalancingV2::Listener):** Configura reglas en el balanceador de carga para redirigir las solicitudes hacia el grupo de destino, escucha el tráfico de los analistas a través del puerto 80.
- **ECSTargetGroup (AWS::ElasticLoadBalancingV2::TargetGroup):** Asocia las tareas ECS con el balanceador de carga y configura verificaciones de estado mediante el endpoint `/health` que ha ido configurado en nuestra aplicación. Este permite conocer el estado en que se encuentran las tareas para que, en caso de que alguna falle, se marca la tarea como no saludable y eso informa al ECS Service de que hay que lanzar nuevas tareas para mantener su número deseado.

8. Definición y gestión de tareas

- **ECSTaskDefinition (AWS::ECS::TaskDefinition):** Este recurso describe las configuraciones necesarias para ejecutar contenedores en ECS. Especifica detalles como la imagen del contenedor a usar, la cantidad de CPU y memoria asignada, y la compatibilidad con el modo Fargate, que permite ejecutar contenedores sin gestionar servidores subyacentes. Además, define un mapeo de puertos, asignando el puerto anterior del contenedor (5000) al tráfico entrante. Es el punto central de la especificación para las tareas ECS.

Describe los contenedores que ejecutan la aplicación AquaSense, incluyendo la configuración de recursos (CPU, memoria) y el puerto utilizado (5000).

- **ECSService (AWS::ECS::Service):** Administra la ejecución de las tareas ECS, asegura la ejecución continua de un número deseado de tareas definidas en la configuración de tarea. En este caso, se garantiza que siempre haya dos tareas activas para manejar las solicitudes entrantes. Se utiliza el modo de lanzamiento Fargate para simplificar la gestión de la infraestructura. El servicio monitorea las tareas y las reinicia automáticamente si fallan, garantizando alta disponibilidad y resiliencia para nuestra aplicación.

9. Escalado automático

- **ECS Scalable Target (AWS::ApplicationAutoScaling::ScalableTarget):** Define el rango de capacidad del servicio ECS, asegurando que el número de tareas

FARGATE activas esté entre un mínimo de 2 y máximo de 10. Este recurso actúa como el objetivo escalable que la política de escalado monitorea y ajusta según las métricas configuradas. Está asociado al recurso ECSService y permite mantener un nivel óptimo de disponibilidad y desempeño de nuestra aplicación.

- **Scaling Policy (AWS::ApplicationAutoScaling::ScalingPolicy)**
 - **High CPU:** Implementa una política de escalado automático que aumenta la cantidad de tareas activas cuando la utilización promedio de CPU en el clúster ECS supera el 75%. Esto asegura que la capacidad del servicio crezca dinámicamente para manejar un aumento en la carga de trabajo. Incluye un tiempo de enfriamiento (*cooldown*) de 60 segundos para evitar ajustes excesivamente frecuentes.
 - **Low CPU:** Establece una política de reducción de escala que disminuye el número de tareas activas cuando el uso de CPU cae por debajo del 25%. Esto permite optimizar costos al liberar recursos durante períodos de baja demanda, manteniendo solo las tareas necesarias para sorportar la carga en ese momento. Similar a la política de escalado ascendente, aplica un enfriamiento de 60 segundos.

10. Seguridad

- **ALB Security Group (AWS::EC2::SecurityGroup):** Permite el acceso al balanceador de carga desde Internet (a través del puerto 80-HTTP), asegurando que el tráfico pueda alcanzar la aplicación.
- **Task Security Group (AWS::EC2::SecurityGroup):** Permite el tráfico del ALB a las tareas de ECS (a través del puerto 5000), asegurando que las solicitudes lleguen al contenedor de la aplicación de manera distribuida y segura.

11. Gestión e implementación

- **CloudFormation:** Automatiza el despliegue de la infraestructura mediante plantillas asegurando la coherencia y la reproducibilidad en la implementación.
- **AWS CLI:** Utilizada para el desarrollo de la imagen y autenticación de recursos a la hora de subir la imagen del contenedor de nuestra aplicación a AWS ECR.
- **Docker:** Facilita el desarrollo y la contenedorización de la aplicación, asegurando un entorno de ejecución consistente.
- **IAM Roles (AWS::IAM::Role):** Proporcionan los permisos necesarios para que las tareas de ECS accedan a DynamoDB y otros recursos de AWS. Permite que los distintos recursos de la arquitectura desplegada se puedan comunicar entre sí con seguridad.

4 Ejemplos de demostración de correcta funcionalidad

- Función Lambda:
- ECR:
- Automatización del aprovisionamiento de la infraestructura mediante CloudFormation:
- Auto Escalado con métricas de utilización de recursos:
- Funcionamiento de la aplicación:

Distintas consultas que pueden realizar los analistas a través del servidor web desplegado por las tareas. En la propia URL definen el tipo de dato a consultar con su mes y año específico, además del correcto funcionamiento de la aplicación con el endpoint /health.