Alumno: Javier Martínez Arias Máster MIMO - UPSA

Documentación - Tecnologías del lado del servidor. Cloud computing

1. Introducción

En este documento se van a explicar las operaciones realizadas para configurar el entorno Cloud con Amazon Web Services (AWS), visto durante la asignatura y desde el que se ha podido distribuir la aplicación de API REST, desarrollada con Play Framework.

También se van a exponer los motivos por los que se han tomado ciertas decisiones a la hora de configurar los servicios Cloud, con tal de mejorar aspectos que afectan a nuestra aplicación, tan importantes como la disponibilidad, escalabilidad, tolerancia a errores, etc.

2. Principales operaciones realizadas

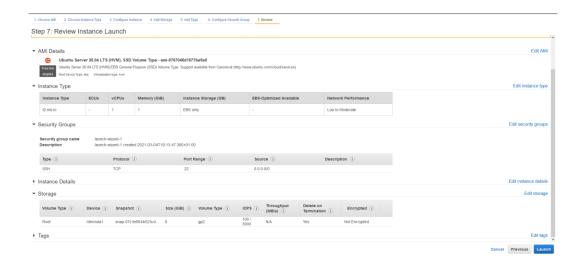
2.1. Despliegue de instancia EC2

En primer lugar, para la preparación del entorno que va a dar servicio a nuestra aplicación, vamos a repasar los pasos seguidos para su puesta en marcha de forma manual.

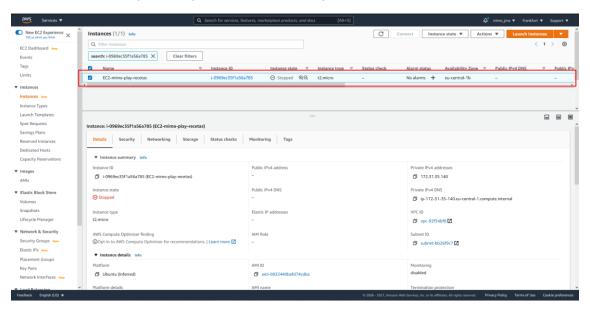
Empezamos creando una instancia utilizando el servicio EC2 que nos ofrece AWS. La configuración y prestaciones de la de la máquina son:

- Ubuntu Server 20.04 LTS 64bits
- Tipo de instancia: t2.micro. Dispone de 1 core y 1 gb de RAM
- 8gb de almacenamiento
- Con un security group: que nos permite configurar los permisos de conexión
- El paquete equivale al *ami*: ami-0767046d1677be5a0
- Al final de proceso, generamos un par de claves y lo descargamos en un fichero ".pem" que nos permitirá conectarnos más adelante por ssh.

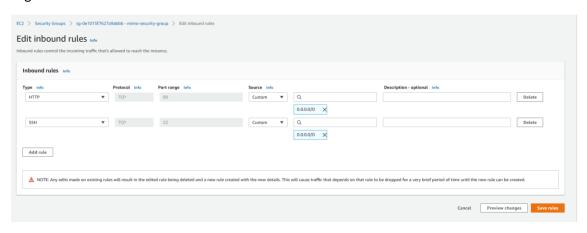
Estas prestaciones básicas son las que nos ofrece el paquete gratuito de Amazon, motivo por el que lo hemos elegido. Sin embargo, como veremos a continuación, será posible replicar la misma configuración en diferentes instancias.



Una vez finalizado el proceso, podemos ver nuestra primera instancia creada en el listado:



Antes de iniciarla y conectarnos vamos a habilitar el puerto HTTP 80 donde va a funcionar nuestra aplicación (por defecto), para no tener problemas de conexión en las pruebas más adelante. En este caso, habilitamos el tráfico sin restricción de IP (0.0.0.0/0), aunque hay que tener en cuenta que en un entorno de producción no sería la configuración correcta por seguridad.



Hecho esto, pasamos a iniciarla desde el botón Start de "Instance state". Transcurrido un tiempo, el estado pasará a conectado.

En este momento obtendremos la dirección DNS pública para conectarnos y configurar la máquina. Esta dirección es dinámica, por lo que cada vez que rearranquemos la instancia, nos generará una diferente.

Antes de conectarnos, debemos establecer permisos al fichero.pem de claves, descargado anteriormente. Abrimos una consola e introducimos el comando:

```
chmod 400 FICHERO.pem
```

Hecho esto, para conectarnos por ssh, tecleamos el comando con el formato:

```
ssh -i FICHERO.pem ubuntu@DNS_PUBLICA
```

En este caso es ubuntu@, porque es el sistema operativo seleccionado anteriormente.

Con esto ya deberíamos estar conectados a la máquina.

2.2. Configuración de aplicación y RDS

Pasamos entonces a instalar las dependencias necesarias para que nuestra aplicación pueda ser ejecutada. En nuestro caso es una aplicación que funciona con java, así que actualizamos los paquetes del Ubuntu e instalamos las dependencias Java mediante:

```
sudo apt-get update
sudo apt install -y default-jdk
```

Una vez finalizado subimos nuestra aplicación generada con versión de *Release*. Para generar la *Release* en Play Framework, revisamos la configuración del fichero application.conf y ejecutamos en la carpeta raíz del proyecto, el comando: sbt dist

En este caso nuestro fichero de configuración tiene las siguientes propiedades definidas:

```
* application.conf ×

D: > Usuarios > Javier > Escritorio > * application.conf

1  # This is the main configuration file for the application.

2  # https://www.playframework.com/documentation/latest/ConfigFile

3

4  ## Models

5  ebean.default = ["models.*"]

6

7  ## Internationalisation

8  play.i18n.langs = [ "en", "es" ]
```

Las configuraciones aquí definidas son aquellas que he considerado esenciales para el funcionamiento de la lógica de aplicación y cuyos valores no conviene parametrizar. Aquellos ajustes parametrizables los veremos más adelante.

El zip resultante de la Release lo subimos a nuestra instancia con el comando:

```
scp -i FICHERO.pem APLICACION.zip ubuntu@DNS_PUBLICA:/tmp
```

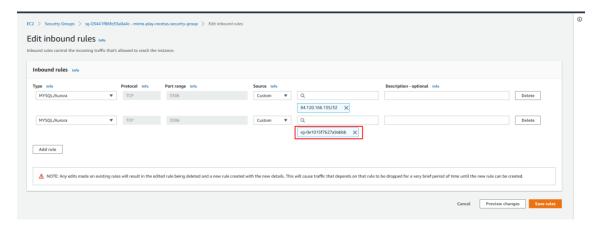
Volvemos a la consola con la conexión a la instancia y copiamos el fichero comprimido a la carpeta opt. Lo descomprimimos con: unzip APLICACION.zip, obteniendo la carpeta raíz con el contenido de la aplicación. En este caso hay que instalar previamente el gestor de descompresión: sudo apt-get install unzip

Antes de ponernos a configurar los parámetros de arranque, hay que crear la base de datos a la que vamos a conectar nuestra aplicación. Para ello, vamos al servicio RDS que nos ofrece AWS y pulsamos en Create Database. La configuración utilizada en este caso es:

- Base de datos Mysql 8.0.20 Edición Community.
 - Se ha seleccionado este motor de bases de datos por tener más soltura y experiencia que con el resto, aunque AWS nos ofrece otras opciones.
- Template: Free Tier
- Deshabilitar el autoescalado
- Crear y asociar un nuevo security group
- Deshabilitar el acceso público
- Ajustes de conexión (en este caso):
 - DB instance identifier: mimo_play_recetas

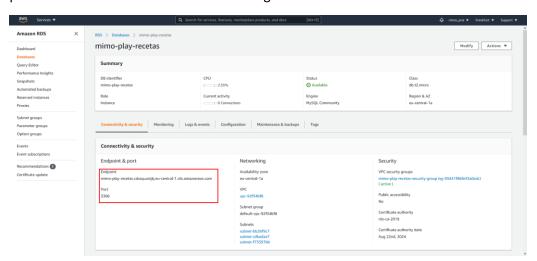
Username: adminPassword: admin123

Una vez creada, vamos a modificar los permisos de su security group, con tal de habilitar la comunicación entre nuestra instancia y la base de datos:



Buscamos el identificador del security group de la instancia y la añadimos. En este caso, establecemos el puerto por defecto de MySQL, el 3306.

Hecho esto, ya podemos iniciar la base de datos y obtener el Endpoint y puerto de conexión para añadirla más adelante a nuestra configuración:



Pasamos por tanto a realizar a la configuración de arranque de la aplicación.

Accedemos de nuevo a la máquina, más concretamente a nuestra carpeta raíz de proyecto. Allí creamos el fichero que lanzará nuestra aplicación Play, con la configuración deseada. Se trata de un fichero .sh. Al editarlo, los parámetros de configuración aquí definidos han sido:

```
GNU nano 4.8

#!/bin/sh

export JDBC_DATABASE_DRIVER=com.mysql.cj.jdbc.Driver
export JDBC_DATABASE_URL=jdbc:mysql://mimo-play-recetas.cdxzquuisjkj.eu-central-1.rds.amazonaws.com:3306/mimo_play_recetas
export JDBC_DATABASE_USERNAME_admin
export JDBC_DATABASE_PASSWORD=admin123

./bin/play-practica-final
-Dhttp.port=80
-Dplay.http.secret.key='AmmvuVYIXRMLJ|<SIw7QrswkCXyika'HC7ejgwAQplH>q?Bi;Qv0'8eosUca2D?A'
-Dplay.evolutions.db.default.enabled=true \
-Dplay.evolutions.db.default.autoApply=true \
-Dplay.evolutions.db.default.autoApplyDowns=true \
-Dplay.evolutions.db.default.autoApplyDowns=true \
-Dplay.filters.hosts.allowed.o='.'
-Ddb.default.url=$[JDBC_DATABASE_DRIVER] \
-Ddb.default.url=$[JDBC_DATABASE_USER] \
-Ddb.default.username=${JDBC_DATABASE_USERNAME} \
-J-Xms128M \
-J-Xms256m \
-J-server \
```

Pasamos a describir que realizan estos comandos:

- El primer párrafo con los exports representa las variables de conexión a la base de datos, que hemos visto anteriormente y se utilizarán más adelante. Se ha decidido colocar aquí pare centralizar los ajustes de conexión, de una forma parametrizable.
- En el siguiente párrafo se configura en el siguiente orden:
 - o El ejecutable de la aplicación
 - o El puerto en el que va a arrancar la aplicación
 - El secret key asociado a la aplicación de Play, generado para funcionar en la Release.
 - Las 3 siguientes líneas se refieren a la autoaplicación y configuración de los evolutions de Play, es decir, la creación de las tablas en base de datos, a partir de los modelos de datos definidos en la aplicación, con todos sus atributos.
 - La siguiente línea hace referencia a los hosts habilitados para ejecutar la aplicación. En este caso el '.' representa que se pueda ejecutar en todos. Inicialmente se había pensado en dejar '.amazonaws.com', pero además de generar algunos warnings de conexión no permitida, limitaba la migración de la aplicación, si se cambiara a otro servidor cloud que no fuera de Amazon.
 - Las siguientes 4 líneas son la asignación de las variables de conexión JDBC a la DB.
 - Los últimos 3 son parámetros de configuración de memoria reservada y ejecución de la aplicación Java.

Si realizando estos pasos no hemos cometido errores, podemos lanzar este script de ejecución con la configuración necesaria y debe funcionar correctamente mostrando:

```
ubuntu@ip-172-31-35-140: /opt/recetas
```

```
ubuntu@ip-172-31-35-140:/opt/recetas$ sudo ./start.sh
[info] p.a.d.DefaultDBApi - Database [default] initialized
[info] p.a.d.HikariCPConnectionPool - Creating Pool for datasource 'default'
[info] play.api.Play - Application started (Prod) (no global state)
[info] p.c.s.AkkaHttpServer - Listening for HTTP on /0:0:0:0:0:0:0:80
```

Llegados a este punto, nos falta crear el fichero de sistema, para que nuestro script se ejecute como servicio en cuanto la instancia esté arrancada. Para ello vamos a la ruta de sistema:

/etc/systemd/system

En este caso lo vamos a crear como: recetas.service con el contenido:

```
[Unit]
Description="Recetas Application"

[Service]
WorkingDirectory=/opt/recetas
ExecStart=/opt/recetas/start.sh
ExecStop=/bin/kill -TERM $MAINPID
Type=simple
Restart=always

[Install]
WantedBy=multi-user.target
```

Siendo en este caso "recetas" el directorio raíz de nuestra aplicación y "start.sh" el script creado anteriormente para lanzar la apliación.

Solo nos queda ponerlo en funcionamiento con los comandos:

```
sudo systemctl daemon-reload
sudo systemctl enable recetas.service
sudo systemctl start recetas
```

Y para comprobar si funciona, hacemos un:

sudo systemctl status recetas

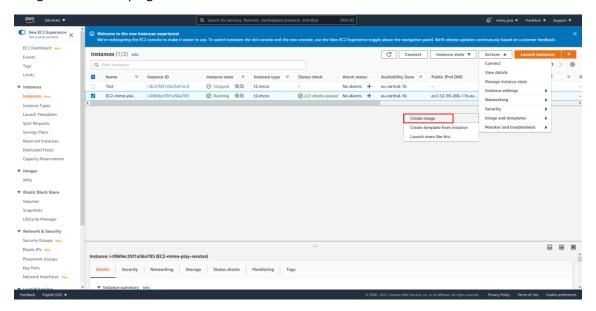
Con esta configuración como servicio y con opción de reiniciarse ante errores, aseguramos en la medida de lo posible que nuestra instancia mantenga en ejecución la aplicación desde que se inicia, siempre que esté disponible.

Por tanto, una vez realizado esto, ya tenemos nuestra aplicación funcionando correctamente y conectada a la base de datos.

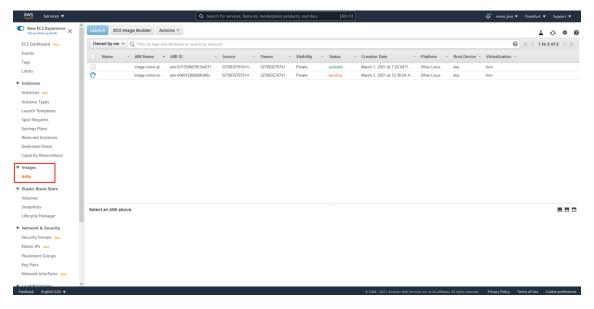
2.3. AMIs y LoadBalancers

Sin embargo, para facilitar la replicación y mejorar la escalabilidad horizontal de la aplicación, AWS nos ofrece algunos servicios muy sencillos de configurar y que nos aportan justo lo que necesitamos. Se trata de los AMIs y los Load Balancers, que comentaremos a continuación.

En primer lugar, vamos a ver los AMIs. Se trata de una foto o captura del estado de nuestra máquina/instancia creada. La foto incluye todas las configuraciones que hemos realizado anteriormente de forma manual, para replicarla tantas veces como queramos. Para ello, seleccionamos desde EC2 nuestra instancia creada anteriormente y vamos a la opción de create image desde el desplegable Actions:



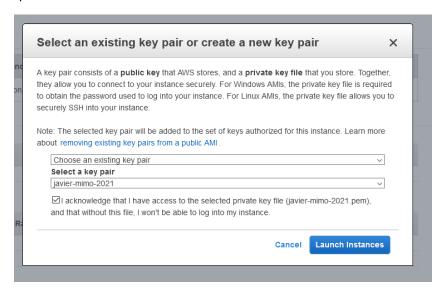
Le indicamos un nombre y pulsamos en crear con los ajustes predeterminados. Hecho esto podemos consultarla desde el apartado Images > AMIs:



Cuando el estado pase a disponible, podremos crear instancias en base a esta imagen, es decir, nuevas instancias con la misma configuración que la original en el momento en el que se creó el AMI. Simplemente pulsamos en Launch y volveremos al mismo proceso de alta que hicimos al

comienzo de isntancias. Aquí simplemente cabe recalcar que debemos asignarles a las nuevas, el mismo security group que a la primera, evitando así volver a configurar la conexión a la base de datos y abrir el puerto 80.

También es importante indicarle a AWS que ya tenemos un par de claves para esta conexión, no queremos crear otro:



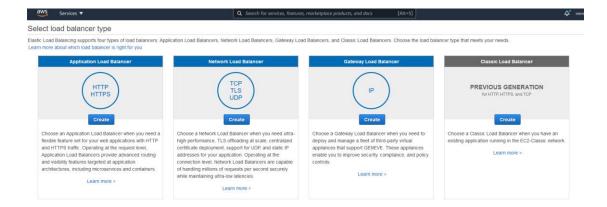
Unos minutos más tarde, ya tendremos las dos instancias corriendo a la vez dando servicio a nuestra aplicación.



Como vemos es una forma muy útil y rápida de escalar nuestras aplicaciones.

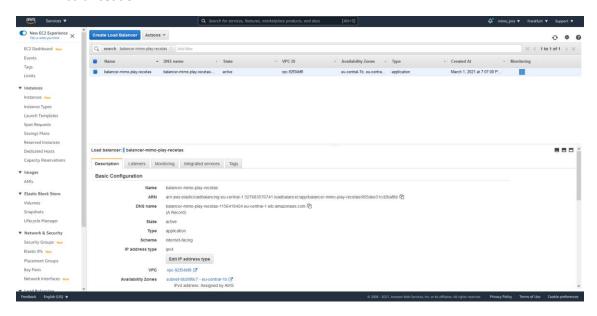
Sin embargo, cada una de ellas tiene una dirección DNS diferente (y dinámica), por lo que necesitamos algún servicio que nos centralice las peticiones en un único sitio, para no tener que ir cambiándolo manualmente. Aquí es donde entra en juego el Load Balancer, cuya función, además de centralizar las peticiones en una única URL estática, se encarga de derivar automáticamente las peticiones a la máquina cuya carga sea inferior en cada momento. Es por tanto, un servicio muy útil para el proceso de escalado y tolerancia a errores de la aplicación, como explicaremos más en detalle.

El proceso para crearlo se realiza desde el servicio de EC2: Load Balancing > Load Balancers > Create Load Balancer.

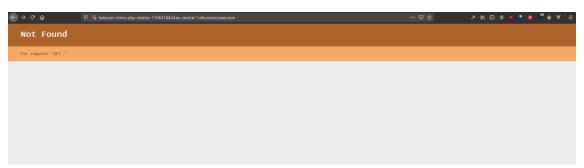


En nuestro caso, las peticiones a la aplicación se realizan a través del protocolo HTTP en el puerto 80, por lo que vamos a seleccionar a primera opción. La configuración de creación es:

- Indicamos el nombre del Load Balancer.
- Seleccionamos todas las zonas de disponibilidad que aparecen.
- Creamos un nuevo security group.
- Configuramos el routing en este caso indicando un nombre y la configuración por defecto
- Finalmente seleccionamos el Target Group, con las instancias que va a incluir en el balanceador.

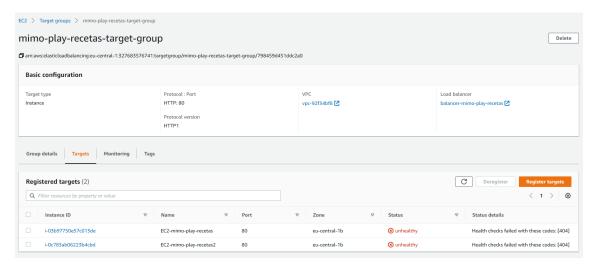


Si accedemos a la URL del balancer en el navegador, podemos ver que nos da una respuesta de nuestra aplicación:

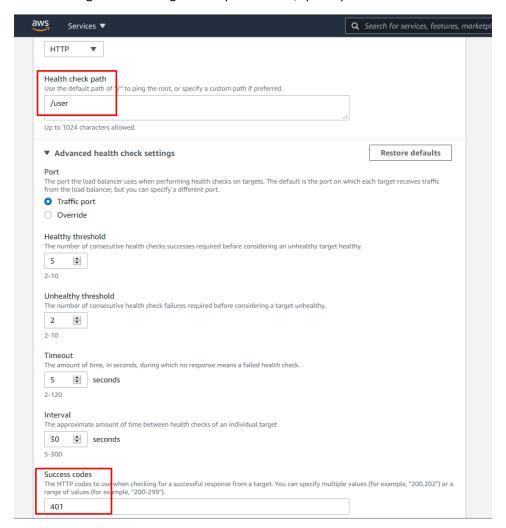


En este caso, nuestra aplicación está funcionando correctamente, pero nos devuelve un 404 por no estar definido el router / en la aplicación.

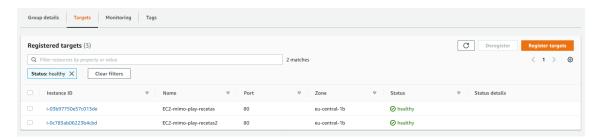
Podemos consultarlo también desde el Target Group del Balancer:



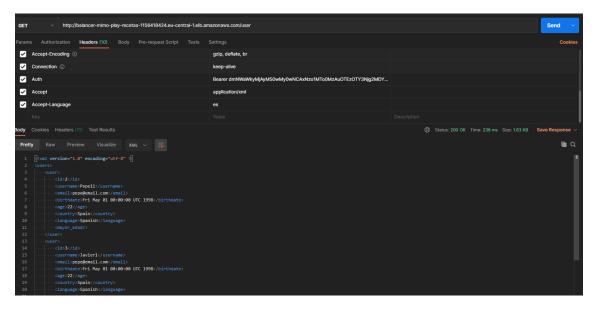
Al igual que en el caso anterior, al probar la ruta /, las peticiones AWS lo detectan como status Unhealthy, por el motivo que ya comentado. Para solucionarlo, podemos cambiar el "Health check settings" de la configuración por defecto, que esperaba un 200.



Vamos a cambiar la ruta a /user que sí está definida por la aplicación. La respuesta esperada en este caso es un 401, ya que se requiere pasar un Auth Token válido en la cabecera. Si tras hacer el cambio actualizamos el Health Status:



La misma petición GET realizada desde un cliente HTTP como Postman y con el Token correctamente definido en la cabecera, nos devuelve un 200:



Por tanto, se están centralizando en una única URL la peticiones que realizamos. Por detrás el Load balancer se encarga de derivar la carga a las instancias disponibles. En caso de que se cayera alguna de las máquinas seguiríamos dando servicio con la otra. Por tanto, este servicio mejora la disponibilidad y la tolerancia a errores.

En principio hemos asociado dos instancias, ya que la carga no se espera que sea muy alta. Sin embargo, si por algún motivo fuera necesario añadir una instancia más, simplemente crearíamos otra instancia con el AMI anterior y lo añadiríamos desde el Target Group al Load Balancer. Está sería una forma de escalado horizontal.

Sería conveniente definir reglas para indicar el mínimo y máximo de instancias que deben funcionar en un balancer, para que AWS se encargara de gestionar y crear nuevas instancias si fuera el caso que las habituales estuvieran caídas, para llegar a los mínimos definidos. Esto puede realizarse desde el servicio de AUTO SCALING que nos ofrece EC2:



Esto sin duda, mejoraría muchísimo la disponibilidad y la tolerancia a errores, asegurando que siempre se va a dar servicio a nuestra aplicación, en base a las reglas definidas de mínimo y máximo.

Hasta aquí las herramientas vistas y su puesta en práctica a nuestra aplicación de API REST desarrollada. Cabe decir que el uso de herramientas de infraestructura como código (por ejemplo: packer y terraform), facilitan mucho el trabajo, ya que permiten la automatización de todas las operaciones que hemos realizado a lo largo de este proyecto. Además, facilita tener un control de versiones del propio generador de instancias y servicios, por si alguno de los requerimientos cambiara en algún momento. Sería la mejor forma de automatizar los despliegues de las diferentes versiones de nuestra aplicación en un futuro.

3. Conclusiones

Con esta práctica y la aplicación de todos los conceptos vistos sobre IAAS y sus diferencias notables frente a los PAAS, han sido de gran utilidad para tener una visión general del funcionamiento y despliegue de las aplicaciones Cloud. Sin duda es una tendencia de desarrollo al alza por todas las ventajas que supone respecto a los métodos más tradicionales.

Tanto por las facilidades que aporta como por la fiabilidad de que la aplicación va a estar dando servicio en todo momento.

Ha sido también una buena introducción a las herramientas principales de AWS así como al scripting en relación con el despliegue de los servicios. Será de gran utilidad a la hora de aplicarlo en un futuro y conocer más funcionalidades interesantes.