

# Informe:

## Práctica 5: Ingesta, almacenamiento y Serveless

Jordi Blasco Lozano  
Infraestructuras y Servicios Cloud  
Universidad de Alicante

19 de noviembre de 2025

### Resumen

## Índice

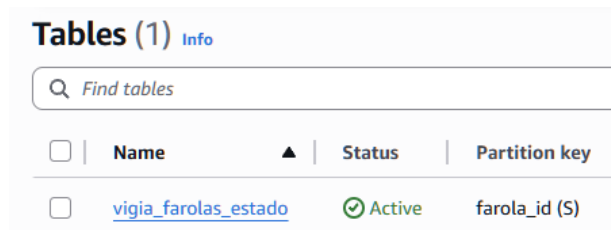
<b>1. Flujo IoT: Ingesta de Telemetría y Almacenamiento Dual</b>	<b>2</b>
1.1. Preparación de Destinos y Configuración IoT	2
1.1.1. Creación de la Tabla DynamoDB	2
1.1.2. Configuración de AWS IoT Core	2
1.2. Creación de la Regla IoT	3
1.2.1. Consulta SQL	3
1.2.2. Acción S3 bucket	3
1.2.3. Accion DynamoDB	4
1.3. Simulación de ingesta con Node-RED	4
1.3.1. Pruebas	5
<b>2. Flujo de Contexto: Web Scraping con Lambda Serverless (RA2)</b>	<b>7</b>
2.1. Creación de la Función Lambda	7
2.2. Orquestación Serverless (EventBridge)	7
<b>3. Flujo Masivo: Transformación y Análisis Serverless</b>	<b>9</b>
3.1. Carga y Catalogación (RA 1)	9
3.1.1. Carga Masiva	9
3.1.2. AWS Glue Crawler	9
3.2. Optimización del Data Lake (AWS Glue ETL) (RA 2)	10
3.2.1. Creación y Propiedades del Job	11
3.2.2. Configuración del Script de Transformación	11
3.2.3. Ejecución del Job y Almacenamiento de Datos Procesados	11
3.3. Consultas y Análisis con Amazon Athena	11
3.3.1. Configuración de Athena	11
3.3.2. Consultas SQL sobre los Datos Originales	11
3.3.3. Consultas SQL sobre los Datos Transformados	11
3.3.4. Análisis de Resultados y Optimización	11

# 1 Flujo IoT: Ingesta de Telemetría y Almacenamiento Dual

## 1.1 Preparación de Destinos y Configuración IoT

### 1.1.1 Creación de la Tabla DynamoDB

Se ha creado una tabla en DynamoDB llamada `vigia_farolas_estado` para almacenar el último estado reportado por cada farola junto con métricas del consumo y el momento de medir estas métricas. La clave de la tabla será la id de cada farola. Solamente indicando como queremos la clave de partición nuestra tabla será funcional.



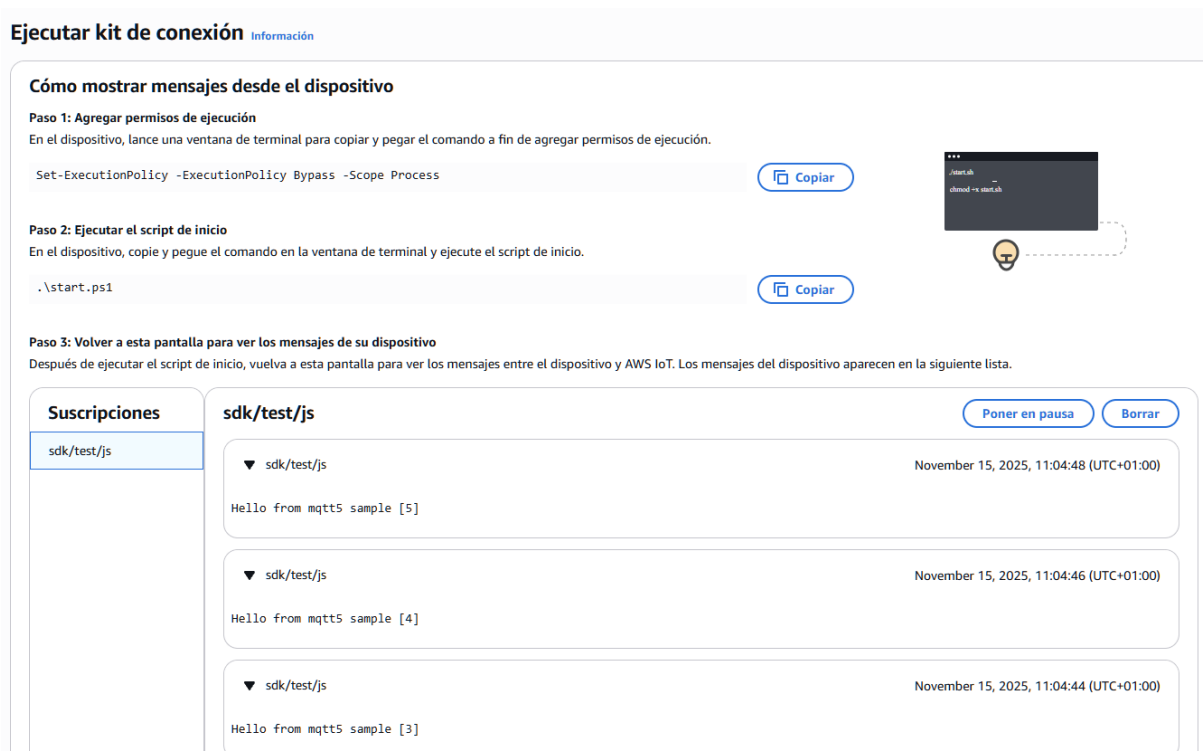
The screenshot shows the AWS DynamoDB console interface. At the top, it says 'Tables (1) Info'. Below that is a search bar labeled 'Find tables'. A table lists the details of the 'vigia\_farolas\_estado' table. It has a checkbox, the name 'vigia\_farolas\_estado', a status of 'Active' with a green checkmark, and a partition key of 'farola\_id (S)'.

<input type="checkbox"/>	Name	Status	Partition key
<input type="checkbox"/>	<a href="#">vigia_farolas_estado</a>	Active	farola_id (S)

Figura 1: Tablas DynamoDB

### 1.1.2 Configuración de AWS IoT Core

- **Conexión del dispositivo:** Mediante el asistente de AWS IoT Core registré un dispositivo nuevo. Durante este proceso, se generaron y descargaron los certificados de seguridad (`farolasestado.cert.pem`, la clave privada correspondiente y publica correspondiente). Esto nos servirá para realizar la prueba de conexión y para conectarnos mediante Node-RED posteriormente.
- **Prueba de conexión:** Para verificar la conectividad, se ejecutó el script de prueba `start.ps1` proporcionado por AWS. El script instaló las dependencias necesarias y envió con éxito cinco mensajes de prueba, que fueron validados tanto en la terminal local como en el cliente de pruebas MQTT en el tema `sdk/test/js` en el mismo panel de conexión de dispositivos IoT.



The screenshot shows the 'Ejecutar kit de conexión' page in the AWS IoT console. It has three steps: 1. 'Agregar permisos de ejecución' with a terminal command and a 'Copiar' button. 2. 'Ejecutar el script de inicio' with a terminal command and a 'Copiar' button. 3. 'Volver a esta pantalla para ver los mensajes de su dispositivo'. Below the steps is a 'Suscripciones' section with a list of messages for the 'sdk/test/js' topic. Each message is a 'Hello from mqtt5 sample' with a count and a timestamp.

**Ejecutar kit de conexión** Información

**Cómo mostrar mensajes desde el dispositivo**

**Paso 1: Agregar permisos de ejecución**  
En el dispositivo, lance una ventana de terminal para copiar y pegar el comando a fin de agregar permisos de ejecución.

`Set-ExecutionPolicy -ExecutionPolicy Bypass -Scope Process` Copiar

**Paso 2: Ejecutar el script de inicio**  
En el dispositivo, copie y pegue el comando en la ventana de terminal y ejecute el script de inicio.

`.\start.ps1` Copiar

**Paso 3: Volver a esta pantalla para ver los mensajes de su dispositivo**  
Después de ejecutar el script de inicio, vuelva a esta pantalla para ver los mensajes entre el dispositivo y AWS IoT. Los mensajes del dispositivo aparecen en la siguiente lista.

**Suscripciones**

**sdk/test/js** Poner en pausa Borrar

- ▼ sdk/test/js November 15, 2025, 11:04:48 (UTC+01:00)  
Hello from mqtt5 sample [5]
- ▼ sdk/test/js November 15, 2025, 11:04:46 (UTC+01:00)  
Hello from mqtt5 sample [4]
- ▼ sdk/test/js November 15, 2025, 11:04:44 (UTC+01:00)  
Hello from mqtt5 sample [3]

Figura 2: Test start.ps1

### ■ Cambio en la Política:

Automáticamente se asocia una política de seguridad, esta política es bastante restrictiva la cambiaremos para que se permitan la conexiones y publicaciones a cualquier tema, yo lo haré en el tema "smartcity/consumo/test". Usamos 'test' porque en ningún momento serán farolas reales por lo que en un supuesto despliegue de la aplicación deberíamos de cambiar el tema. Cabe recalcar que para activar la política debemos de crear una nueva versión seleccionarla como activa y eliminar la versión anterior si no la vamos a volver a usar.

Versión activa: 2 [Información](#)

Efecto de la política	Acción de la política
Allow	iot:Connect
Allow	iot:Publish
Allow	iot:Receive
Allow	iot:Subscribe

Figura 3: Política activa

## 1.2 Creación de la Regla IoT

### 1.2.1 Consulta SQL

Antes de definir la regla, se creó un bucket en S3 para el almacenamiento histórico de la telemetría. La regla de IoT se configuró para procesar los mensajes que llegan al tema `smartcity/consumo/test`. Para la consulta SQL tuve que cambiar un par de instrucciones ya que no me funcionaba correctamente el `processed_time`. Luego veremos lo que use en la función de Node-RED pero por ahora analizaremos la consulta SQL que he utilizado.

Bloque 1: Consulta SQL para la regla IoT

```

1 SELECT
2     consumption_kw,
3     timestamp,
4     concat(day, '-', month, '-', year, ' ', hour, ':', minute, ':', seconds) AS
5     processed_time
6 FROM
7     'smartcity/consumo/test'
```

Como vemos en el SQL hemos quitado el `home_id` y hemos cambiado el `processed_time`, esto se debe a que AWS no me dejaba utilizar DynamoDBv2 con plantilla y al usar DynamoDB legacy la `clave_id` la he mapeado directamente en el apartado correspondiente de DynamoDB, de esta forma ya no me apareciera doble en la tabla de Dynamo (como clave y como atributo). Posteriormente cambie el `processed_time` porque no hubo forma de procesar el tiempo con la función propuesta. Esta ha sido la forma más sencilla de obtener el tiempo procesado. Podría haberlo procesado al crear el mensaje cambiando la función del Node-RED o después al recibir el mensaje, y he decidido hacerlo después para que las carpetas del S3 pudieran mapear mucho mas facil el tiempo, y así crear las carpetas sin errores.

### 1.2.2 Acción S3 bucket

En esta acción también cambiamos un apartado, la "clave", en el enunciado se pedía utilizar un almacenamiento en carpetas clasificadas por año por mes por día, el problema que le vi principalmente fue que en un sistema real podríamos tener diferentes farolas por lo que utilice una ultima clasificación que en la carpeta días tuviera carpetas por "farola\_id". Dentro de esta ultima carpeta ya tenemos cada archivo ordenado por la hora, minuto y segundo usando "hh:mm:ss-data.json" en vez de con timestamp (que a mi se me ponía en milisegundos). He usado la siguiente clave con los parámetros obtenidos directamente desde sus variables del mensaje (en vez de desde el timestamp).

Bloque 2: Clave del S3

```

1 farolas/year=${year}/month=${month}/day=${day}/${home_id}/${hour}:${minute}:${seconds}-data.json
```

### 1.2.3 Accion DynamoDB

En esta otra acción se pedía usar DynamoDBv2, despues de probar mil formas cambie a la versión a la de DynamoDB normal en la cual nosotros mismos debemos de vincular la variable clave para usarla como clave de particion de nuestra tabla. Lo hice manualmente porque no sabia como insertar la plantilla de atributos que proporciona el enunciado. Y no conseguí que las claves se vincularan automáticamente, por lo que lo hice manualmente con DynamoDB de la siguiente forma.

The screenshot shows the configuration interface for a DynamoDB action in the AWS IoT console. The form is titled 'DynamoDB' and includes an 'Eliminar' button. The configuration steps are as follows:

- Nombre de la tabla:** 'vigia\_farolas\_estado'. There are buttons for 'Ver' and 'Crear una tabla de DynamoDB'.
- Clave de partición:** 'farola\_id'. A note states: 'La clave de partición (también denominada clave hash) debe coincidir con la clave de partición de la tabla de DynamoDB que ha creado.'
- Tipo de clave de partición:** 'STRING'. A note states: 'El tipo de clave de partición (también denominada clave hash) puede ser STRING o NUMBER. El valor predeterminado es STRING.'
- Valor de clave de partición:** '\${home\_id}'. A note states: 'El valor de clave de partición (también denominada clave hash) admite plantillas de sustitución que proporcionan datos en tiempo de ejecución.'
- Clave de ordenación:** 'MySortKey'. A note states: 'La clave de ordenación (también denominada clave de rango) debe coincidir con la clave de ordenación de la tabla de DynamoDB que ha creado.'
- Tipo de clave de rango:** (Empty). A note states: 'El tipo de clave de ordenación (también denominada clave de rango) puede ser STRING o NUMBER. El valor predeterminado es STRING.'
- Valor de clave de rango:** 'MysortKeyValue'. A note states: 'El valor de clave de ordenación (también denominada clave de rango) admite plantillas de sustitución que proporcionan datos en tiempo de ejecución.'
- Escritura de datos del mensaje en esta columna:** 'payload'. A note states: 'Escritura de datos del mensaje en esta columna: opcional'.
- Operación:** 'INSERT'. A note states: 'La operación puede ser INSERT, UPDATE o DELETE. El valor predeterminado es INSERT.'
- Rol de IAM:** 'LabRole'. There are buttons for 'Ver' and 'Crear un nuevo rol'. A note states: 'AWS IoT creará automáticamente una política con el prefijo "aws-iot-rule" bajo el rol de IAM seleccionado.'

Figura 4: Configuración de la acción DynamoDB

### 1.3 Simulación de ingesta con Node-RED

Node-RED permite conectarse mediante MQTT a nuestro tema de 'thing' de IoT para simular el envío de datos de las farolas. El flujo consta de tres nodos:

- **Inject:** Un disparador manual para iniciar el flujo, cada 10 segundos se envia una señal al siguiente nodo para que mande un mensaje.
- **Function:** Un nodo que construye el mensaje JSON con los datos de la farola (ID, consumo y el tiempo) y establece el tema de destino en `smartcity/consumo/test`. Este nodo consta de una serie de variables que hemos randomizado para simular la farola. He utilizado 6 variables de más para el tiempo, de forma que tengamos en variables separadas el año, el mes, el día, la hora, los minutos y los segundos. Consiguiendo que los pasos anteriores hayan sido más sencillos de implementar sabiendo que no tenemos que parsear ni usar funciones diferentes para obtener los datos del tiempo.

### Bloque 3: Funcion constructora del mensaje

```
1 const now = new Date();
2 msg.topic = "smartcity/consumo/test";
3 msg.payload = {
4   home_id: "farola-001",
5   consumption_kw: +(Math.random() * 2).toFixed(2),
6   timestamp: now.getTime(),
7   year: now.getFullYear(),
8   month: String(now.getMonth() + 1).padStart(2, "0"),
9   day: String(now.getDate()).padStart(2, "0"),
10  hour: String(now.getHours()).padStart(2, "0"),
11  minute: String(now.getMinutes()).padStart(2, "0"),
12  seconds: String(now.getSeconds()).padStart(2, "0"),
13 };
14 return msg;
```

- **MQTT Out:** Este nodo está configurado para publicar el mensaje en el endpoint de AWS IoT (a1paa0c5brn8mp-ats.iot.us-east-1.amazonaws.com) a través del puerto 8883 con TLS. La autenticación se realizó utilizando los mismos certificados de dispositivo que en la prueba de conexión inicial. Usamos TLS el cual contiene todos los certificados necesarios para conectarnos al endpoint que anteriormente hemos introducido.

Con estos tres nodos configurados tenemos ya el flujo funcional, y en darle a instanciar tendremos el sistema enviando mensajes cada 10 segundos al tema de MQTT de nuestro objeto. Para parar el sistema es tan simple como deshabilitar el nodo del disparador y volver a darle a instanciar.

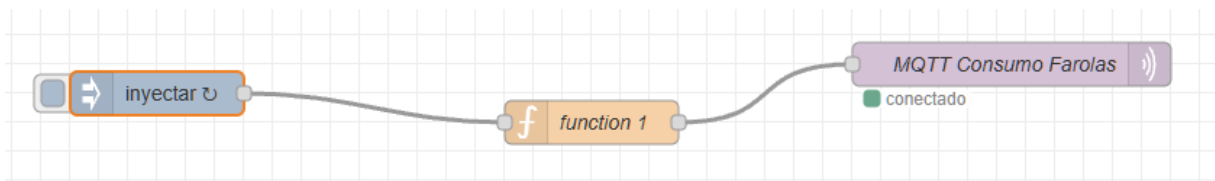


Figura 5: Flujo Node-RED

#### 1.3.1 Pruebas

- **Cliente de prueba MQTT:** Para comprobar que cada una de las acciones anteriores de la regla funcionen debemos de comprobarlo lanzando el Node-RED. El primer paso es comprobar que los mensajes son enviados al tema correspondiente, para esto nos vamos al cliente de prueba de MQTT dentro de la pestaña de IoT de AWS y nos subscribimos a `smartcity/consumo/#` cuando nos subscribamos nos empezaran a salir mensajes como estos.

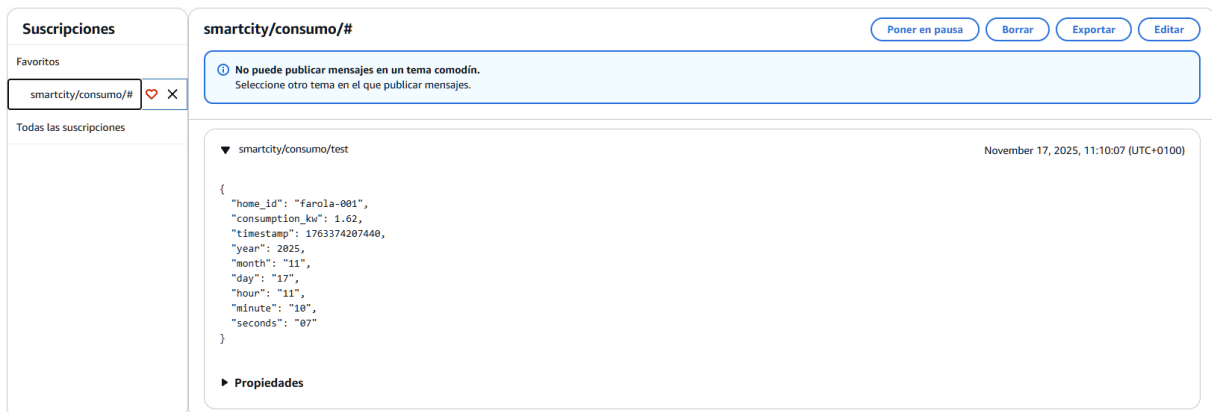


Figura 6: Cliente de prueba de MQTT

- **S3:** Respecto al guardado de mensajes historicos debemos de comprobarlo dentro del S3 de la practica, nos dirigimos a la sucesiones de carpetas que tenemos configurada y dentro de la ultima vemos como tenemos estos datos. Los datos de dentro de cada archivo tambien corresponden con el .json definido en el SQL, ej:

```
1 {"consumption_kw":0.3,"timestamp":1763374077358,"processed_time":"17-11-2025 11:07:57"}
```



- **DynamoDB** Finalmente debemos de comprobar la tabla de DynamoDB, dentro de `vigia_farolas_estado`, y en explorar elementos de la tabla vemos como nuestra tabla contiene una `farola001` con el json correspondiente. Este json se va actualizando cada 10 segundos y muestra siempre el estado actual de la farola. Si tuviéramos mas farolas se mostraría una tabla con el estado actual de cada una de las farolas.



## 2 Flujo de Contexto: Web Scraping con Lambda Serverless (RA2)

### 2.1 Creación de la Función Lambda

Para crear la función Lambda debemos de ingresar al servicio Lambda y crear una función de python desde cero con los permisos de Lab-Role para poder ejecutar correctamente el proceso y realizar el PutObject al S3. Una vez dentro se nos abrirá un editor de código, que en si es un fork web de visual Studio Code. Debemos de pasar la variable por entorno, para esto dentro del panel de configuración ingresamos a variables de entorno y copiamos el nombre del bucket con el nombre de la variable del código proporcionado.

The screenshot shows the 'Editar las variables de entorno' (Edit environment variables) page in the AWS Lambda console. At the top, there's a breadcrumb trail: 'Lambda > Funciones > scraper-tarifas-luz > Editar variables de entorno'. The main section is titled 'Variables de entorno' and explains that these are key-value pairs for the function code. Below this, there's a table with two columns: 'Clave' (Key) and 'Valor' (Value). The first row has 'S3BUCKETNAME' as the key and 'p5-buquet' as the value. To the right of the value is an 'Eliminar' (Delete) button. Below the table is an 'Agregar variable de entorno' (Add environment variable) button. At the bottom of the table, there's a 'Configuración de cifrado' (Encryption configuration) link. At the very bottom of the page are 'Cancelar' (Cancel) and 'Guardar' (Save) buttons.

Figura 9: Variable de entorno

Cuando tengamos la variable de entorno copiamos el código que simulará nuestra ingesta de datos de contexto aletorizando la tarifa de luz del día `tarifakwh`. Tras copiar el código ya podemos desplegarlo.

### 2.2 Orquestación Serverless (EventBridge)

Al igual que hemos hecho en el apartado anterior con la ingesta de telemetría mediante la simulación por Node-RED, en este apartado también tenemos que simular una entrada programada cada cierto tiempo. En este caso mandaremos una señal a nuestra función Lambda cada 1 minuto. Esto lo haremos mediante el servicio de Amazon EventBridge, este nos permitirá crear un evento programado al destino de AWS que nosotros queramos. Lo configuraremos de la siguiente forma eligiendo nuestra función Lambda y poniendo una programación que se repita cada minuto.

The screenshot shows the 'Patrón de programación' (Pattern) section of the AWS EventBridge 'Create Event Rule' page. It asks the user to 'Elija el tipo de programación que mejor se adapte a sus necesidades' (Choose the type of scheduling that best fits your needs). There are two radio button options: 'Una programación detallada que se ejecuta a una hora específica, como las 8.00 h PST del primer lunes de cada mes.' (A detailed scheduling that runs at a specific time, such as 8:00 h PST on the first Monday of each month.) and 'Un horario que se ejecuta con una frecuencia periódica; p. ej., cada 10 minutos.' (A recurring frequency; e.g., every 10 minutes.). The second option is selected. Below this is the 'Expresión de frecuencia' (Frequency expression) section, which says 'Especifique un valor y la unidad de tiempo para ejecutar la programación.' (Specify a value and the unit of time to execute the scheduling.). There's a 'frecuencia' (frequency) input field with '1' and a unit dropdown menu currently showing 'Minu...' (Minutes).

Figura 10: Frecuencia Event Bridge

The screenshot shows the 'Destino 1' (Destination 1) section of the AWS EventBridge 'Create Event Rule' page. It asks the user to 'Seleccione un bus de eventos de EventBridge, un destino de la API de EventBridge (socio de SaaS) u otro servicio de AWS como destino.' (Select an EventBridge event bus, an EventBridge API destination (SaaS partner), or another AWS service as a destination.). There are three radio button options: 'Bus de eventos de EventBridge', 'Destino de la API de EventBridge', and 'Servicio de AWS'. The third option is selected. Below this is the 'Seleccione un destino' (Select a destination) section, which says 'Seleccione los destinos que desee invocar cuando un evento coincida con el patrón de eventos o cuando se active la programación (hay un límite de 5 destinos por regla).' (Select the destinations you want to invoke when an event matches the event pattern or when the scheduling is activated (there is a limit of 5 destinations per rule)). There's a dropdown menu showing 'Función Lambda'. Below this is the 'Ubicación de destino' (Destination location) section, with two radio button options: 'Destino en esta cuenta' (Destination in this account) and 'Destino en otra cuenta de AWS'. The first option is selected. Below this is the 'Función' (Function) section, with a dropdown menu showing 'scraper-tarifas-luz'. Below this is the 'Configurar la versión o el alias' (Configure the version or alias) section. Below this is the 'Permisos' (Permissions) section, with a checkbox 'Usar el rol de ejecución (opción recomendada)' (Use execution role (recommended)) checked. Below this is the 'Rol de ejecución' (Execution role) section, which says 'EventBridge necesita permiso para enviar eventos al destino seleccionado. Al continuar, nos autoriza a hacerlo. EventBridge y AWS Identity and Access Management' (EventBridge needs permission to send events to the selected destination. By continuing, we authorize it to do so. EventBridge and AWS Identity and Access Management). There are two radio button options: 'Crear un nuevo rol para el recurso especificado' (Create a new role for the specified resource) and 'Usar el rol existente' (Use the existing role). The second option is selected. Below this is the 'Nombre del rol' (Role name) section, with a dropdown menu showing 'LabRole'.

Figura 11: Destino Event Bridge

Si nos fijamos en el código de nuestra función, el guardado de datos se sobrescribiera al poner como nombre DD-MM-YYYY-data.json. Pero podremos comprobar que se sobrescriba correctamente comprobando el contenido del archivo del S3 cada minuto. Nuestros archivos `-data.json` tendrán la estructura de este ejemplo.

```
1 {"timestampscraper": "2025-11-17T18:20:11.307175", "tarifakwh": 0.1833, "ciudad": "Neo-Tech"}
```

Nuestra carpeta dentro del bucket dentro esta otra forma.

**tarifas/**

**Objetos** | Propiedades

**Objetos (1)** Copiar URI de S3

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para concederles permisos de forma explícita. [Más información](#)

🔍 *Buscar objetos por prefijo*

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	<a href="#">2025-11-17-data.json</a>	json	17 Nov 2025 7:40:11 PM CET	93.0 B	Estándar

Figura 12: S3 tarifas

Finalmente en los registros de cloudWatch podemos ver las invocaciones a la función mas recientes.

Invocaciones recientes								
#	Timestamp	RequestId	LogStream	DurationInMS	BilledDurationInMS	MemorySetInMB	MemoryUsedInMB	
▶ 1	2025-11-17T18:58:11.013Z	575619ea-a83b-471c-9d6f-37f6058545f0	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	208.84	209.0	128.0	86.0	
▶ 2	2025-11-17T18:57:10.973Z	c6ac8c9f-68dd-4933-af51-a44ff7ae7fe7	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	212.12	213.0	128.0	86.0	
▶ 3	2025-11-17T18:56:10.998Z	227f0914-51f4-486a-86bd-a33f9685c66f	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	234.47	235.0	128.0	86.0	
▶ 4	2025-11-17T18:55:11.153Z	3d3c76bb-5ebc-4971-818a-1d2d424088bd	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	228.2	229.0	128.0	86.0	
▶ 5	2025-11-17T18:54:10.952Z	75f3bae5-0c4a-45cd-905f-4ffc487708f2	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	278.49	271.0	128.0	86.0	
▶ 6	2025-11-17T18:53:10.897Z	3dcace58-048f-499c-8faa-8381f8bd7b81	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	242.78	243.0	128.0	86.0	
▶ 7	2025-11-17T18:52:11.173Z	0ce625f9-53d6-4f0f-b4eb-089ca92fae0f	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	229.41	230.0	128.0	86.0	
▶ 8	2025-11-17T18:51:10.889Z	b116b073-292c-48fb-8908-d2524f95e4ec	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	221.17	222.0	128.0	86.0	
▶ 9	2025-11-17T18:50:11.634Z	78b36c11-ff32-4986-846c-72811370e9b3	2025/11/17/[SLATEST]d6aae628ca1c4256887b1a572de74437	212.55	213.0	128.0	86.0	

Figura 13: S3 tarifas



## 3 Flujo Masivo: Transformación y Análisis Serverless

### 3.1 Carga y Catalogación (RA 1)

Para cargar el dataset historico de datos de movilidad hice uso de una carpeta nueva en nuestro S3 de la practica, la llame /historicomovilidad e introduje el .csv proporcionado en la practica, utilice el de 10 filas para evitar consumos. Antes de subirlo lo abrí en visual studio code y me di cuenta de que me saltaba el linter porque al haber eliminado la categoria repetida `Airport_fee` no me había fijado que también debía eliminar las comas de esta columna.

#### 3.1.1 Carga Masiva

Con el .csv listo y la carpeta creada cargué el archivo.



Figura 14: csv historico movilidad

#### 3.1.2 AWS Glue Crawler

Mediante AWS Glue Crawler, automatizamos el proceso de descubrimiento y catalogación de los datos almacenados en el S3. El Crawler escanea los archivos en la carpeta, infiere su esquema (nombres de columnas, tipos de datos y particiones) y crea las Glue Tables dentro de nuestra base de datos.

Es importante destacar que las tablas creadas en el Data Catalog no contienen los datos en sí, sino metadatos: apuntan a la ubicación física de los archivos en S3 y definen su estructura. Esto actúa como una capa de abstracción que permite a servicios como Amazon Athena consultar los archivos CSV (u otros formatos) directamente usando SQL estándar, como si se tratara de una base de datos relacional tradicional, sin necesidad de mover o cargar los datos previamente. El crawler lo configuramos de la siguiente forma:

- **Creación del Crawler:** en este apartado debemos de definir el nombre del crawler (el mio vigia-crawler) y dentro de su panel de configuración debemos de añadir nuestra fuente de datos (la dirección S3 de nuestra carpeta /historicomovilidad que creamos anteriormente). Posteriormente asignamos nuestro IAM role (el LabRole) y finalmente creamos una base de datos vacia para que el Crawler genere las tablas dentro de ella. Al acabar estos pasos debemos de tener un Crawler como este:

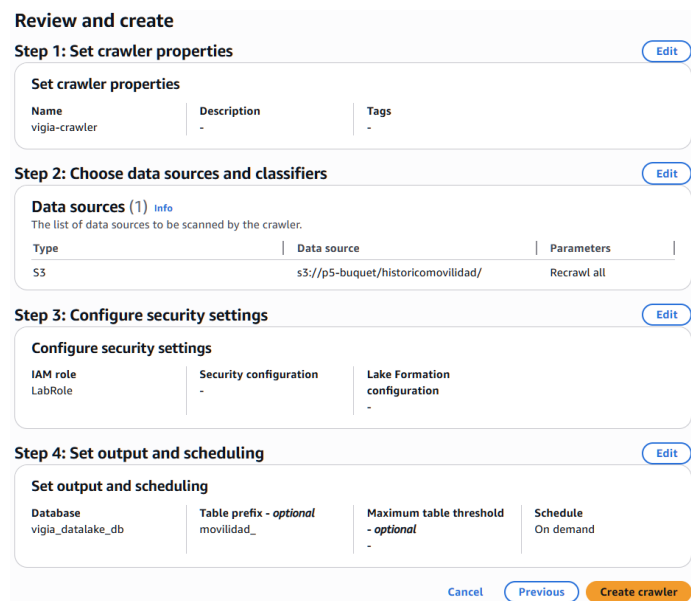


Figura 15: Crawler

- **Ejecución del Crawler:** tras terminar la configuración del Crawler debemos de ejecutarlo. Durante este proceso de ejecución el crawler lo podemos observar en los logs de ejecución: el crawler **vigia-movilidad-crawler** inicia el proceso de clasificación de los datos. Una vez analizada la estructura de los archivos en S3, escribe los resultados en la base de datos **vigia\_datalake\_db**, creando exitosamente la tabla **historicomovilidad**. Finalmente, actualiza el Data Catalog (indicando ADD: 1 en los logs) y vuelve al estado **READY**, dejando los datos listos para ser consultados. Para comprobar si ha el crawler ha conseguido analizar correctamente la estructura accedemos a las tablas de nuestro Data Catalog y dentro de la tabla **historicomovilidad** deberá de aparecernos este esquema:

**Schema (20)**  
View and manage the table schema.

Q Filter schemas

#	Column name	Data type
1	rowid	bigint
2	vendorid	bigint
3	tpep_pickup_datetime	string
4	tpep_dropoff_datetime	string
5	passenger_count	double
6	trip_distance	double
7	ratecodeid	double

Figura 16: Table Schema

- **Consulta SQL:** Al tener la tabla ya creada, podemos consultar el archivo csv directamente con SQL mediante el servicio de Athena que proporciona amazon. Para las guardar las consultas lo optimo será crear un S3 nuevo que en mi caso he llamado **resultados-p5-bucket** que guardará tanto las consultas que hagamos con esta tabla (de datos en bruto) como las que hagamos con la tabla (analítica). He obtenido estos resultados al hacer la consulta:

**Resultados (10)**

Q Filas de búsqueda

#	rowid	vendorid	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance
1	0	2	2023-06-30 23:59:59	2023-07-01 00:47:49	2.0	17.62
2	1	2	2023-06-30 23:59:57	2023-07-01 00:17:36	1.0	3.32
3	2	1	2023-06-30 23:59:55	2023-07-01 00:14:20	1.0	2.8
4	3	2	2023-06-30 23:59:55	2023-07-01 00:05:52	1.0	0.89
5	4	2	2023-06-30 23:59:55	2023-07-01 00:07:08	4.0	1.56
6	5	2	2023-06-30 23:59:54	2023-07-01 00:14:58	1.0	3.85
7	6	2	2023-06-30 23:59:53	2023-07-01 00:05:52	2.0	1.05

Figura 17: Table Schema

### 3.2 Optimización del Data Lake (AWS Glue ETL) (RA 2)

En este paso, implementaremos un proceso ETL (Extract, Transform, Load) utilizando un Job de AWS Glue. El objetivo principal es transformar los datos crudos que tenemos en formato CSV a un formato optimizado para analítica (Parquet) y particionarlos adecuadamente. Esto nos servirá para mejorar drásticamente el rendimiento de las consultas en Athena y reducir los costos, ya que Parquet es un formato columnar comprimido que permite leer solo las columnas necesarias, y el particionado permite escanear solo los datos relevantes para una consulta (por ejemplo, filtrar solo por un año y mes específicos).

El Job realizará las siguientes acciones clave:

- **Extracción:** Leerá los datos de la tabla original catalogada en el Data Lake.
- **Transformación:** Convertirá los tipos de datos (casteo de strings a double/timestamp), limpiará el formato de fecha y seleccionará las columnas críticas para el análisis.
- **Carga:** Escribirá los datos procesados en una nueva carpeta */analitica* del S3 en formato Parquet, particionados por año y mes.

### 3.2.1 Creación y Propiedades del Job

### 3.2.2 Configuración del Script de Transformación

### 3.2.3 Ejecución del Job y Almacenamiento de Datos Procesados

## 3.3 Consultas y Análisis con Amazon Athena

### 3.3.1 Configuración de Athena

### 3.3.2 Consultas SQL sobre los Datos Originales

### 3.3.3 Consultas SQL sobre los Datos Transformados

### 3.3.4 Análisis de Resultados y Optimización