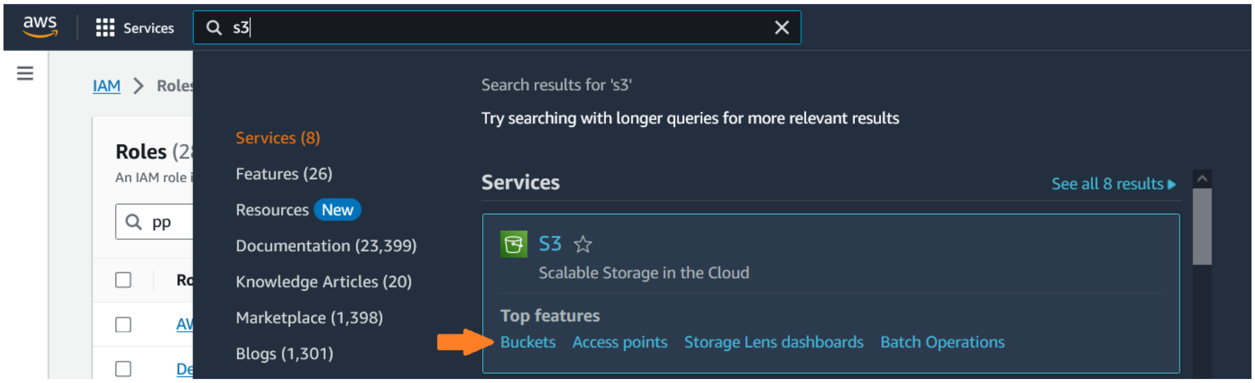
Guía para configurar Servicios de Amazon Web (AWS) y Automatizar el proceso carga al Data Warehouse

Configuración de los servicios

Almacenamiento en s3

Dirigirse al servicio de s3 en la opción de Bucket

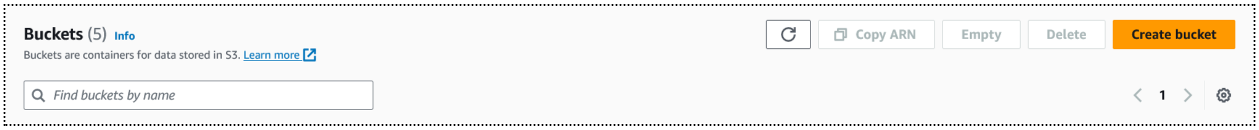


Search results for 's3'

Try searching with longer queries for more relevant results

S3 — Scalable Storage in the Cloud

Crear un nuevo depósito (bucket)



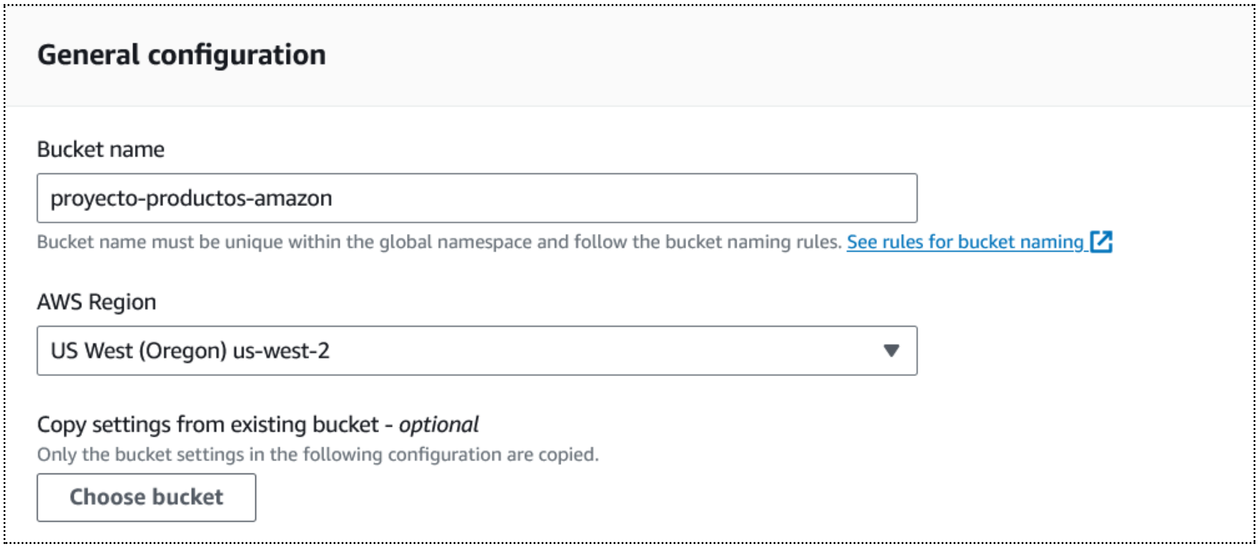
Buckets are containers for data stored in S3. Learn more

Colocar nombre y región

Como más adelante se va a crear un cluster para ser utilizado en el servicio de Redshift, este tiene ciertos requerimientos, por lo que se debe escoger la región de Oregon (us-west-2).

El nombre debe ser único entre los depósitos ya creados, y no acepta guión bajo, solo guión medio.

No se requiere modificar nada más, a parte de los mostrado en la imagen a continuación.



General configuration

Bucket name

proyecto-productos-amazon

Bucket name must be unique within the global namespace and follow the bucket naming rules. See rules for bucket naming

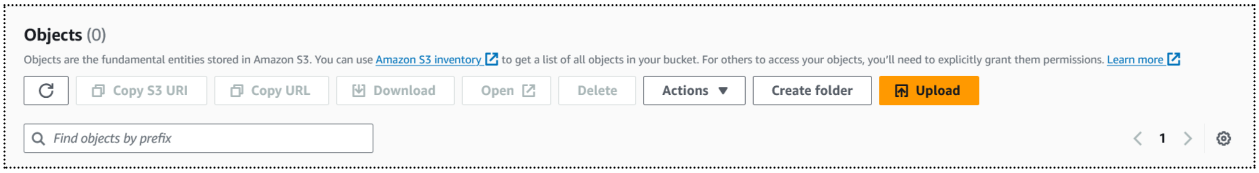
AWS Region

US West (Oregon) us-west-2

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

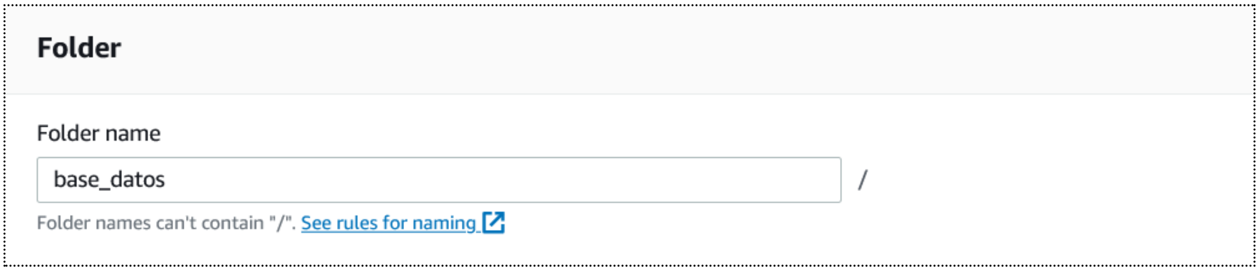
Crear las carpetas necesarias
Ingresar al depósito creado
Hacer clic en “Crear carpeta”



Objects (0)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Colocar el nombre y crear



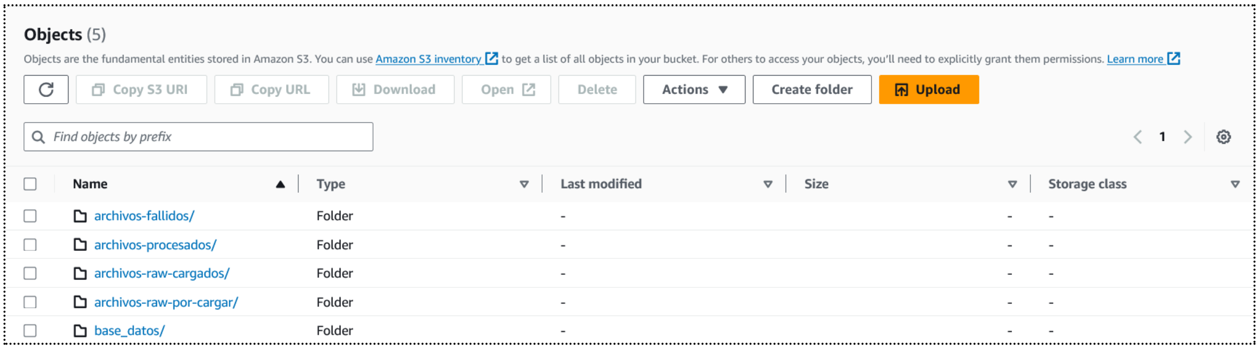
Folder

Folder name

base_datos /

Folder names can't contain "/". See rules for naming

Estructura de carpetas para el proyecto



Objects (5)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	archivos-fallidos/	Folder	-	-	-
□	archivos-procesados/	Folder	-	-	-
□	archivos-raw-cargados/	Folder	-	-	-
□	archivos-raw-por-cargar/	Folder	-	-	-
□	base_datos/	Folder	-	-	-

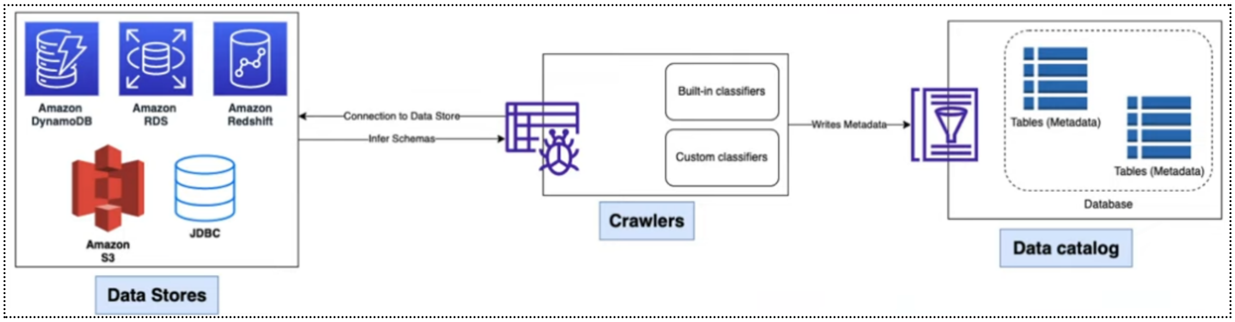
Carpeta	**Descripción**
base_datos	Contiene archivos que

archivos-fallidos	Si el proceso de ETL es fallido, los archivos de la carpeta "archivos-raw-por-cargar" se mueven a esta carpeta.
archivos-raw-cargados	Si el proceso de ETL es exitoso, los archivos de la carpeta "archivos-raw-por-cargar" se mueven a esta carpeta. Contiene las subcarpetas: -- dataset_meta -- dataset_review
archivos-raw-por-cargar	Se suben los archivos a los que se les va a realizar el proceso de ETL. Contiene las subcarpetas: -- dataset_meta -- dataset_review
archivos-procesados	Después del proceso de ETL, se genera un nuevo archivo con los datos del dataset procesados, el cual se guarda en esta carpeta. Contiene las subcarpetas: -- dataset_meta -- dataset_review

Conexión s3 - Glue

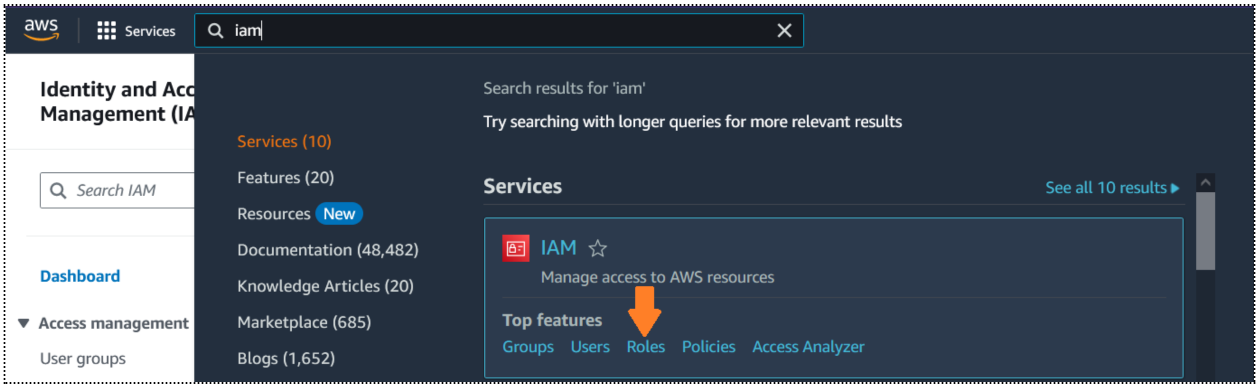
Para realizar esta conexión se vio conveniente el uso de un Crawler y un Data Catalog puesto que proporciona una forma eficiente y automatizada de administrar y utilizar los datos, lo que puede ahorrar tiempo, reducir errores y facilitar la colaboración en proyectos de datos complejos.

La imagen a continuación muestra una estructura general de su conexión.



Creación Rol de Glue

Dirigirse al servicio de IAM en la opción de Rol

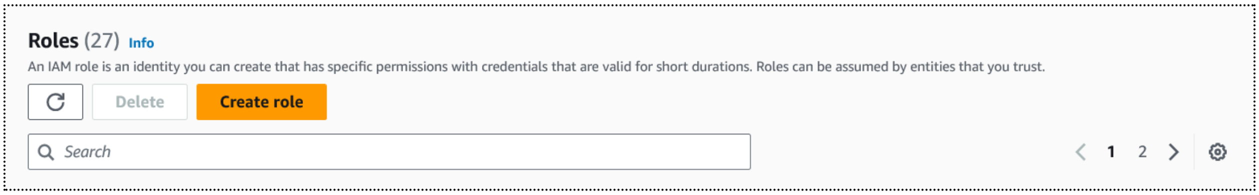


Search results for 'iam'

Try searching with longer queries for more relevant results

IAM ☆

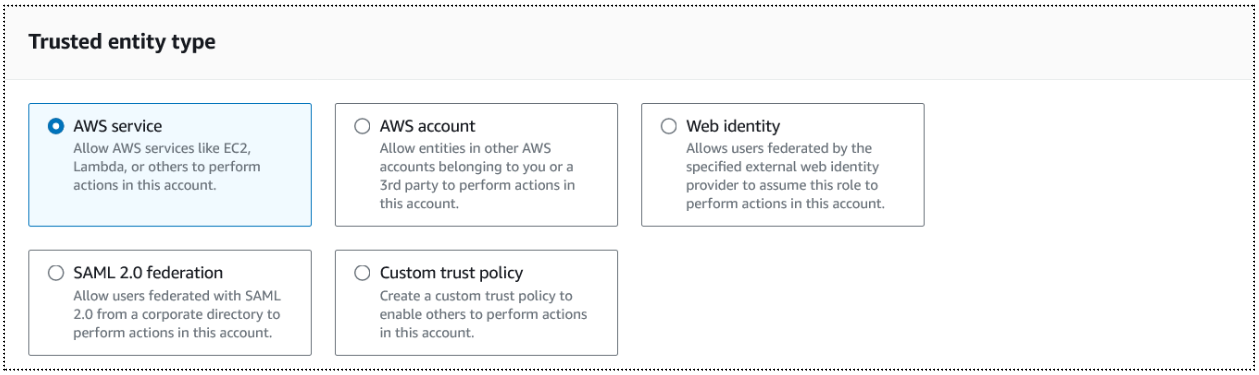
Manage access to AWS resources



Roles (27)

An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.

Crear un nuevo Rol, seleccionando el servicio de Glue



Trusted entity type

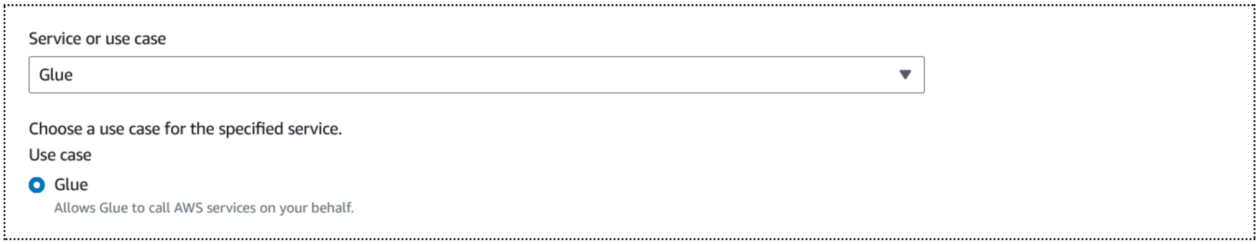
● AWS service
Allow AWS services like EC2, Lambda, or others to perform actions in this account.

○ AWS account
Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.

○ Web identity
Allows users federated by the specified external web identity provider to assume this role to perform actions in this account.

○ SAML 2.0 federation
Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.

○ Custom trust policy
Create a custom trust policy to enable others to perform actions in this account.



Service or use case

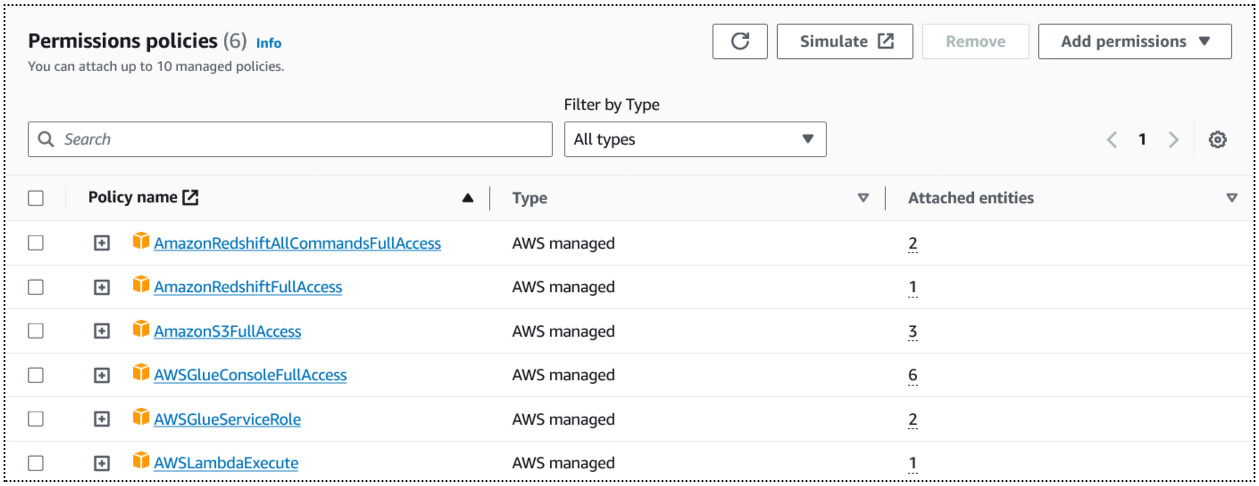
Glue

Choose a use case for the specified service.

Use case

● Glue
Allows Glue to call AWS services on your behalf.

Añadir los siguientes permisos

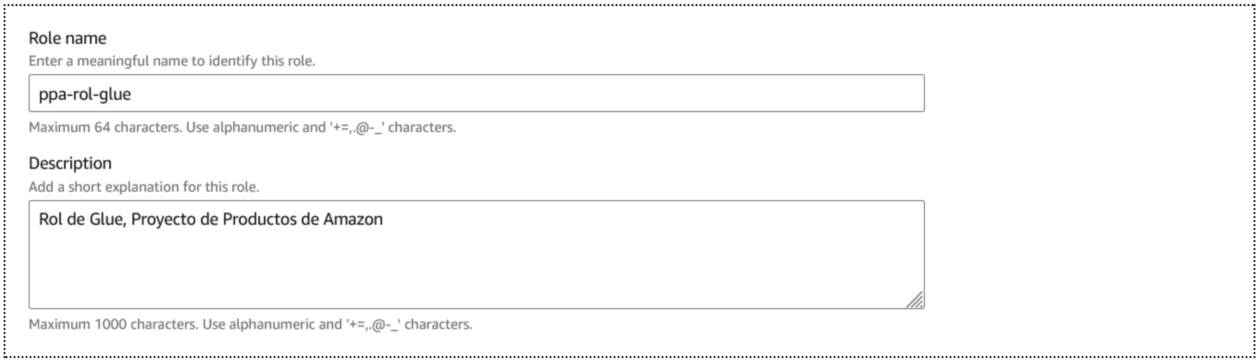


Permissions policies (6) Info

You can attach up to 10 managed policies.

☐	Policy name	Type	Attached entities
☐	AmazonRedshiftAllCommandsFullAccess	AWS managed	2
☐	AmazonRedshiftFullAccess	AWS managed	1
☐	AmazonS3FullAccess	AWS managed	3
☐	AWSGlueConsoleFullAccess	AWS managed	6
☐	AWSGlueServiceRole	AWS managed	2
☐	AWSLambdaExecute	AWS managed	1

Ingresar el nombre del rol y crear



Role name

Enter a meaningful name to identify this role.

ppa-rol-glue

Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.

Description

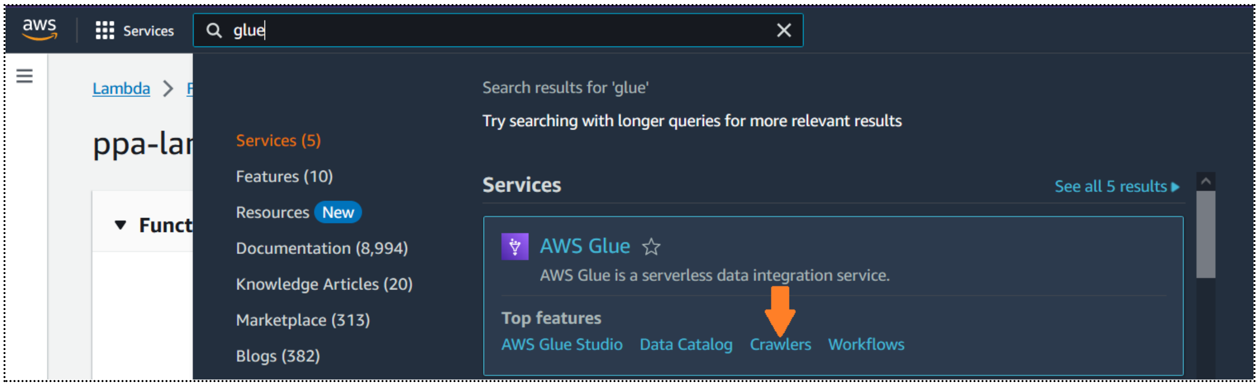
Add a short explanation for this role.

Rol de Glue, Proyecto de Productos de Amazon

Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.

Creación de Classifier

Dirigirse al servicio de Glue la opción de Crawlers



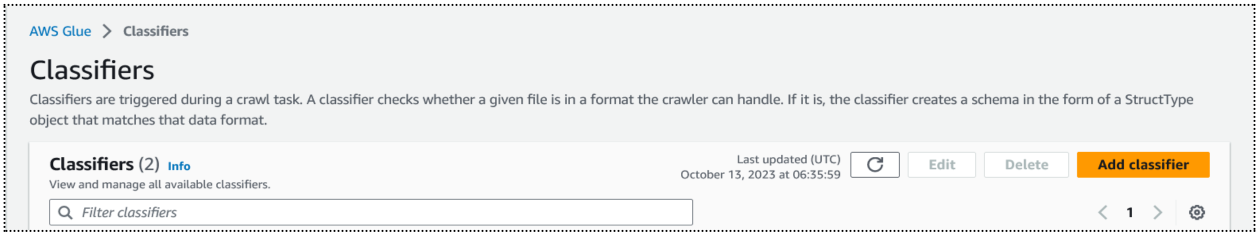
Search results for 'glue'

Try searching with longer queries for more relevant results

AWS Glue is a serverless data integration service.

En el menú izquierdo, seleccionar Classifier

Añadir nuevo Classifier para archivos tipo json



Classifiers

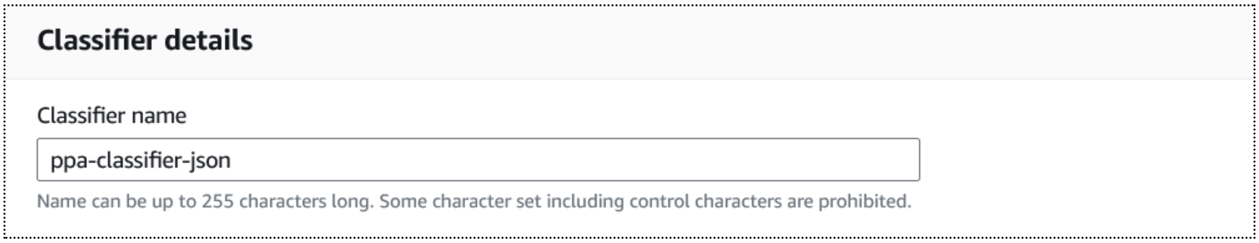
Classifiers are triggered during a crawl task. A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

Classifiers (2) Info

View and manage all available classifiers.

Last updated (UTC) October 13, 2023 at 06:35:59

Colocar un nombre



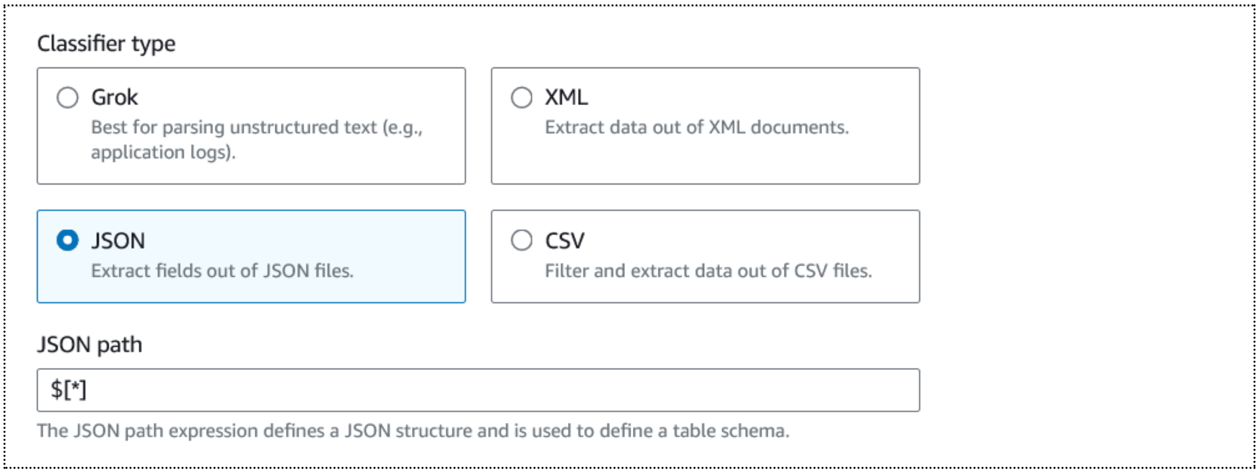
Classifier details

Classifier name

ppa-classifier-json

Name can be up to 255 characters long. Some character set including control characters are prohibited.

Configurar de la siguiente manera y crear



Classifier type

○ Grok
Best for parsing unstructured text (e.g., application logs).

○ XML
Extract data out of XML documents.

● JSON
Extract fields out of JSON files.

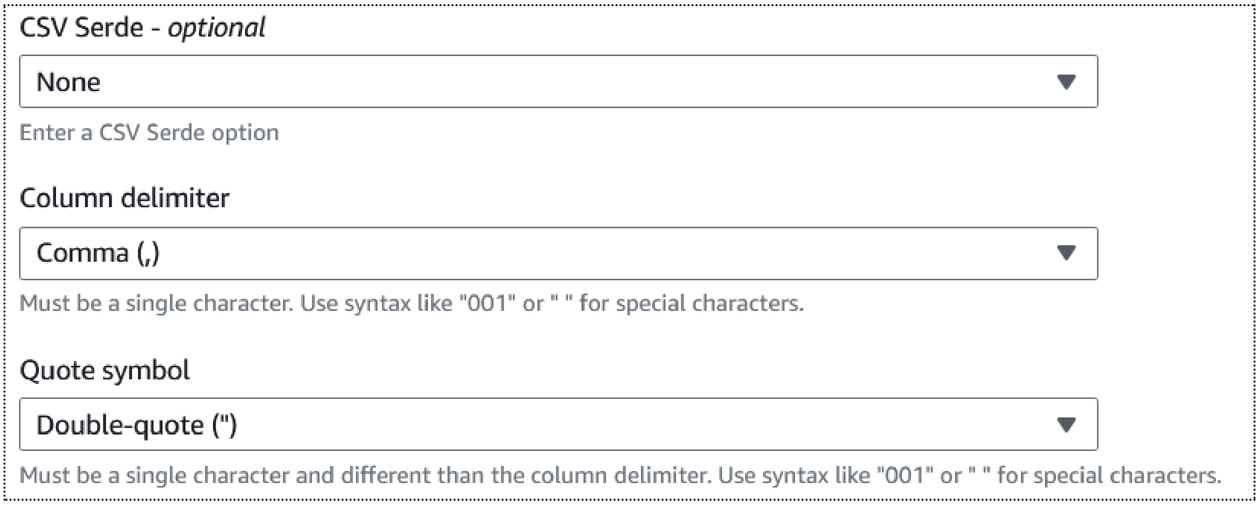
○ CSV
Filter and extract data out of CSV files.

JSON path

$[*]

The JSON path expression defines a JSON structure and is used to define a table schema.

Repetir el proceso para tipo csv
Configurar de la siguiente manera



CSV Serde - *optional*

None ▼

Enter a CSV Serde option

Column delimiter

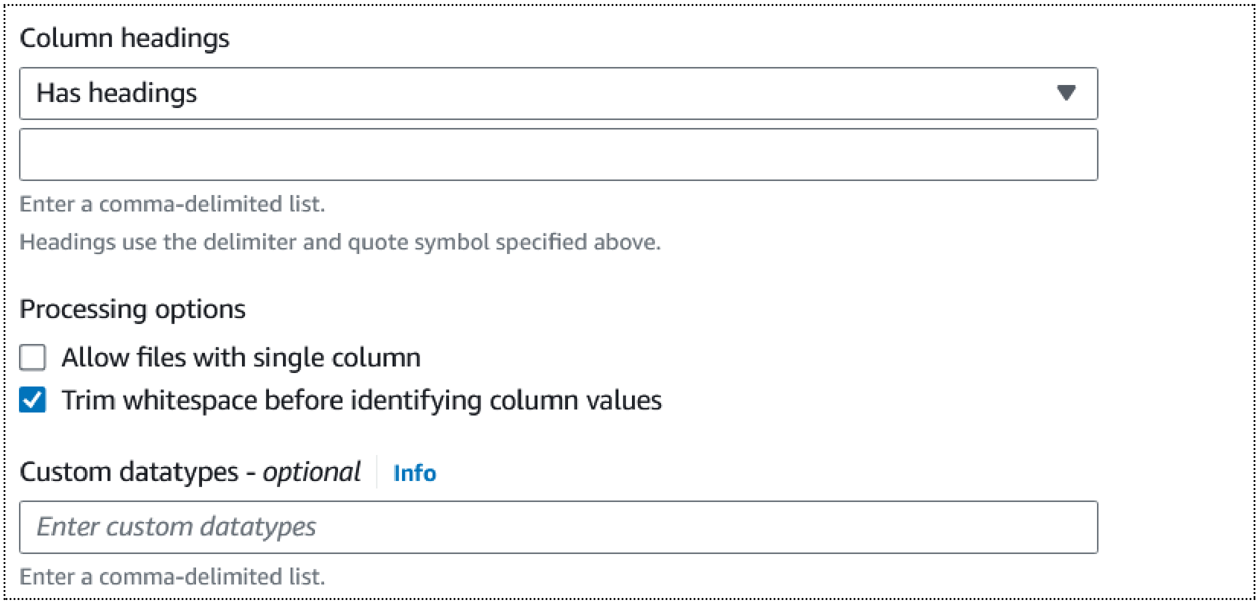
Comma (,) ▼

Must be a single character. Use syntax like "001" or " " for special characters.

Quote symbol

Double-quote (") ▼

Must be a single character and different than the column delimiter. Use syntax like "001" or " " for special characters.



Column headings

Has headings ▼

Enter a comma-delimited list.

Headings use the delimiter and quote symbol specified above.

Processing options

□ Allow files with single column

☑ Trim whitespace before identifying column values

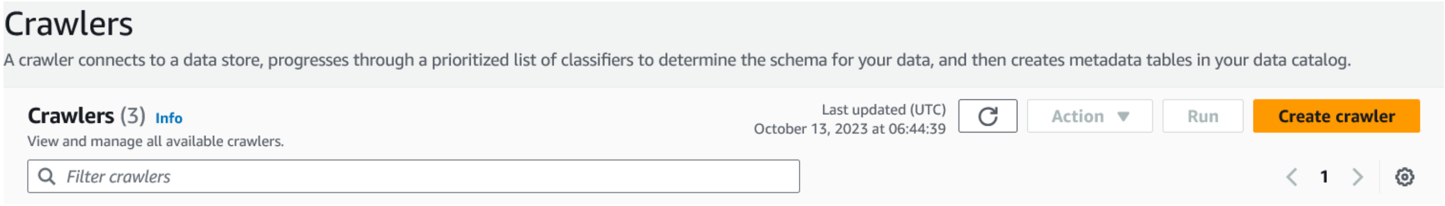
Custom datatypes - *optional* Info

Enter custom datatypes

Enter a comma-delimited list.

Creación Crawler

Regresar a la opción de Crawlers y crear un nuevo



Crawlers

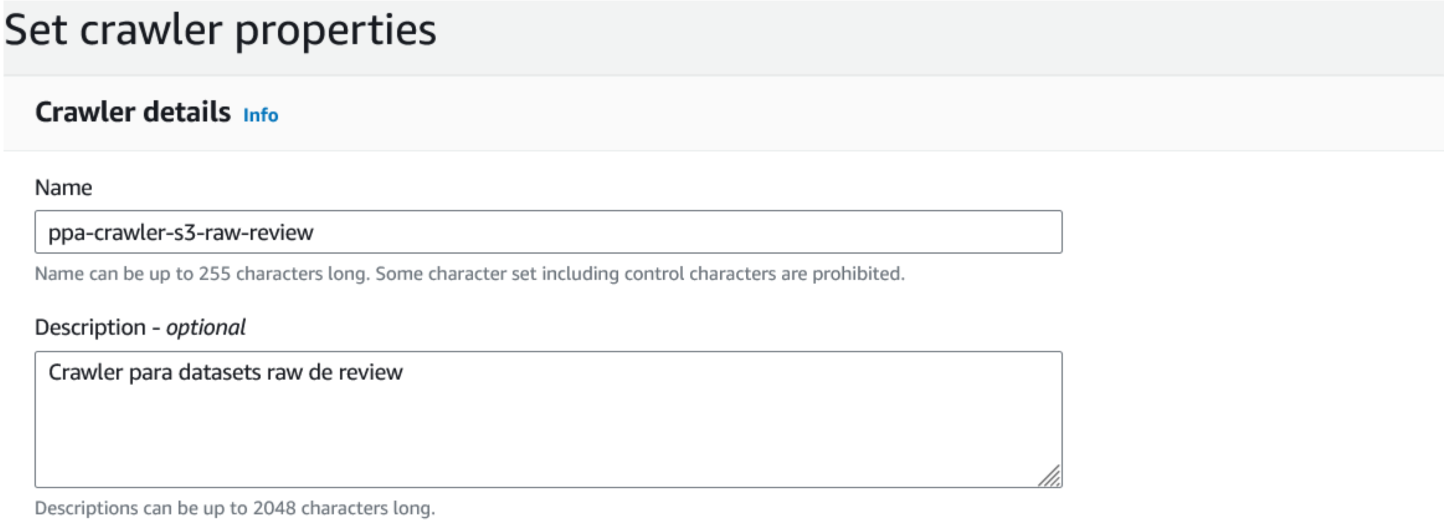
A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Crawlers (3) Info

View and manage all available crawlers.

Last updated (UTC)
October 13, 2023 at 06:44:39

Ingresar el nombre



Set crawler properties

Crawler details Info

Name

ppa-crawler-s3-raw-review

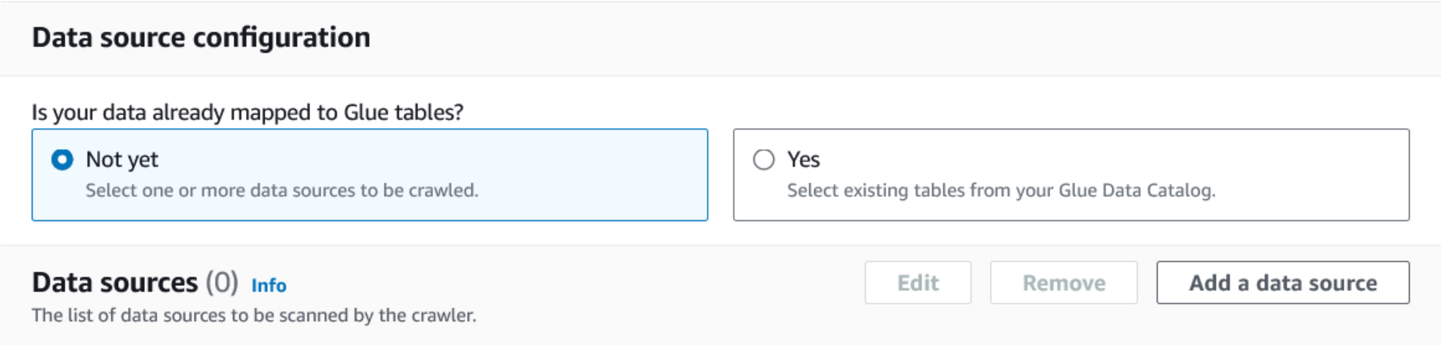
Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - *optional*

Crawler para datasets raw de review

Descriptions can be up to 2048 characters long.

Añadir una nueva fuente de datos



Data source configuration

Is your data already mapped to Glue tables?

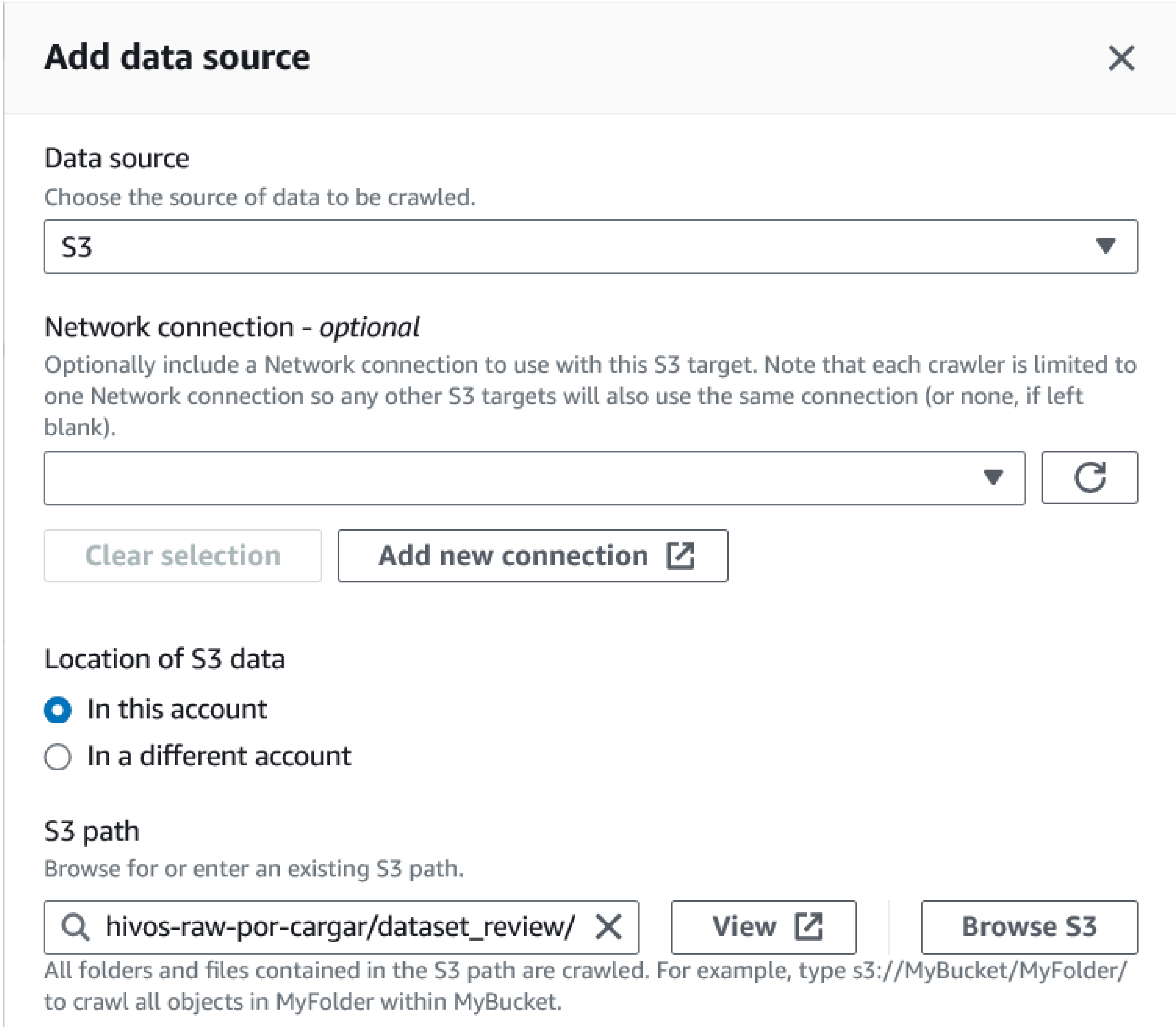
● Not yet
Select one or more data sources to be crawled.

○ Yes
Select existing tables from your Glue Data Catalog.

Data sources (0) Info

The list of data sources to be scanned by the crawler.

Configuración de la fuente de datos
La dirección del almacenamiento es: archivos raw por cargar, dataset review



Add data source

Data source
Choose the source of data to be crawled.

S3

Network connection - *optional*
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Location of S3 data

● In this account

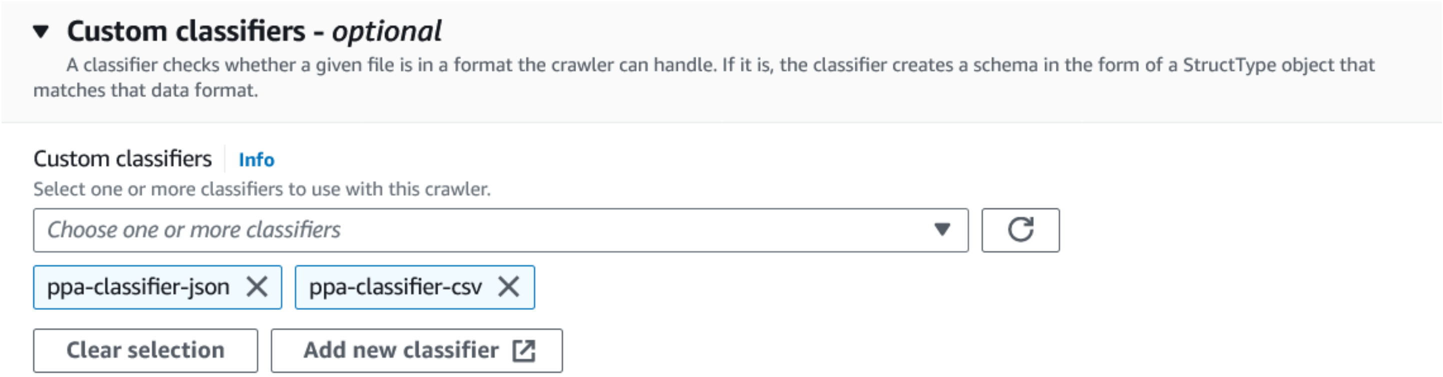
○ In a different account

S3 path
Browse for or enter an existing S3 path.

hivos-raw-por-cargar/dataset_review/

All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Se añaden los classifiers



▼ Custom classifiers - *optional*

A classifier checks whether a given file is in a format the crawler can handle. If it is, the classifier creates a schema in the form of a StructType object that matches that data format.

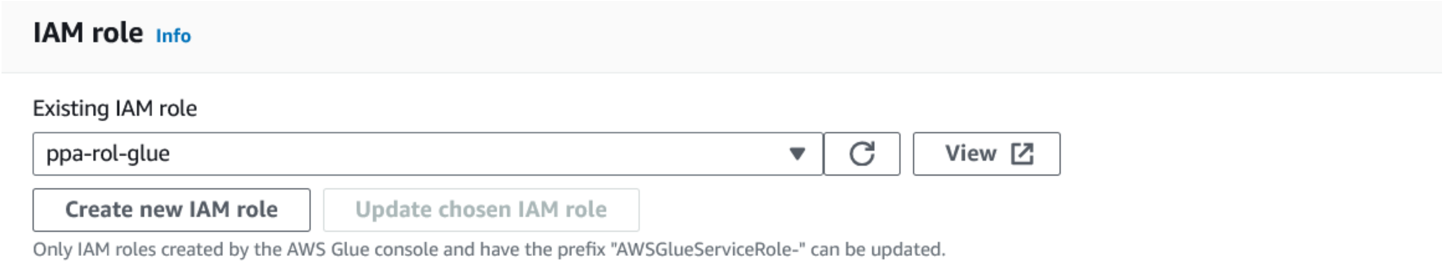
Custom classifiers

Select one or more classifiers to use with this crawler.

ppa-classifier-json

ppa-classifier-csv

Se selecciona el rol, previamente creado

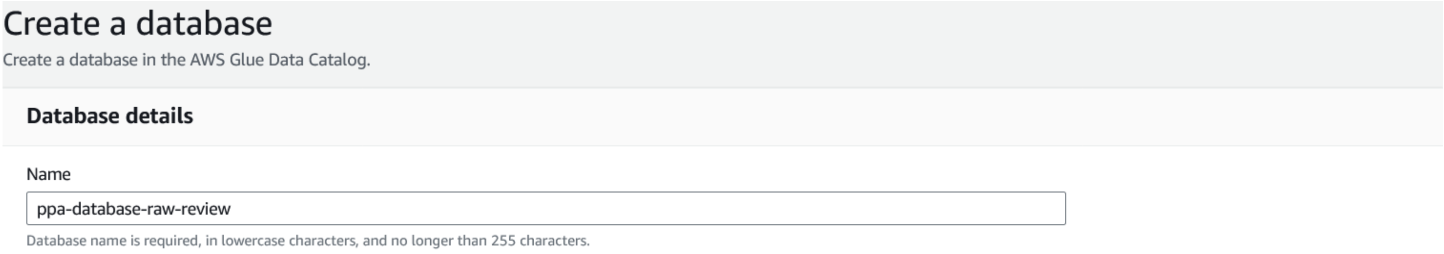


IAM role

Existing IAM role: ppa-rol-glue

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Se crea una base de datos



Create a database

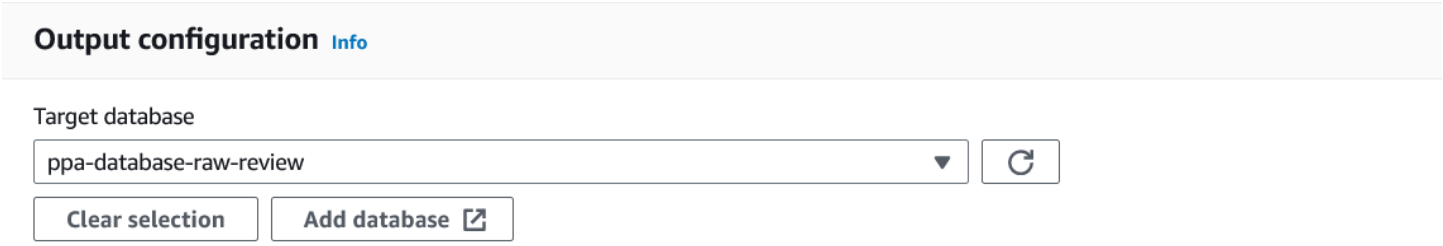
Create a database in the AWS Glue Data Catalog.

Database details

Name: ppa-database-raw-review

Database name is required, in lowercase characters, and no longer than 255 characters.

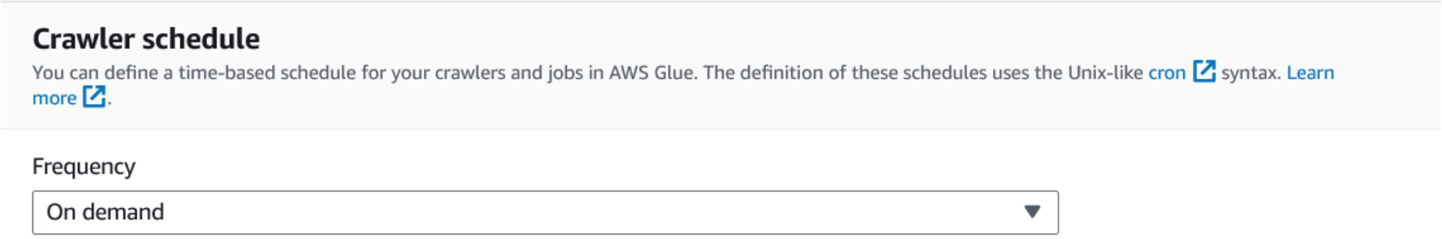
Se selecciona la base de datos creada



Output configuration

Target database: ppa-database-raw-review

Se selecciona la opción On-demand y se crea el crawler



Crawler schedule

You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. Learn more.

Frequency

On demand ▼

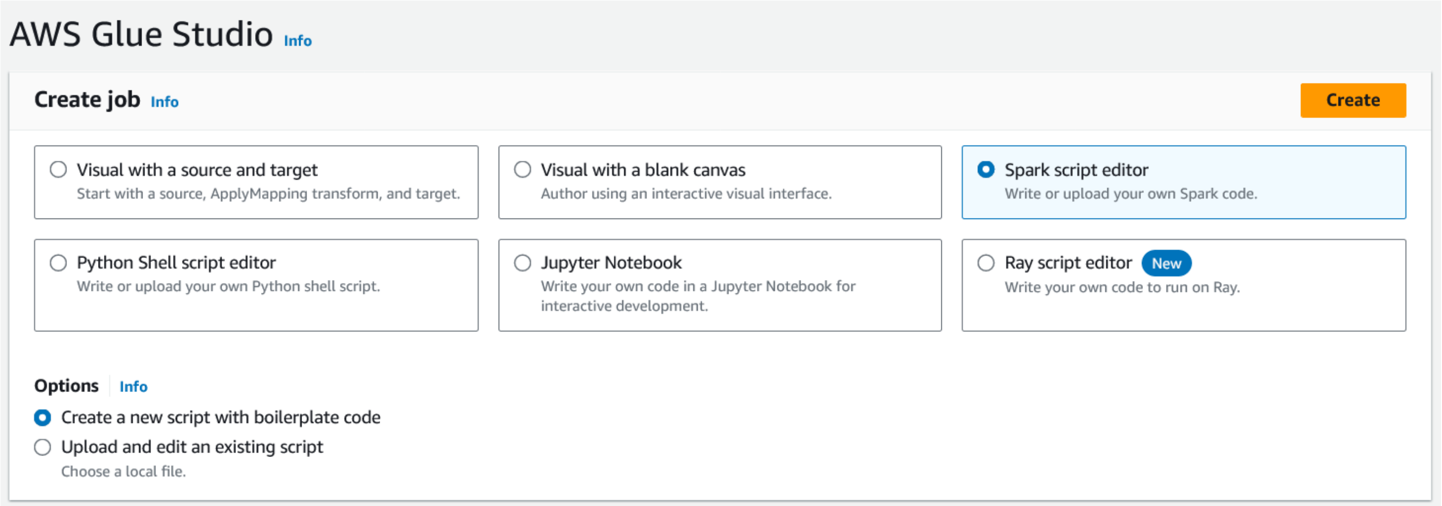
Crear un nuevo Crawler para los datasets raw de meta, repitiendo los pasos anteriores.

Se puede crear la tabla del Data catalog de manera manual, haciendo correr los crawlers creados. Sin embargo, esto se puede lograr haciendo la primera parte de la automatización del proceso.
En caso de querer hacerlo utilizando la forma de creación de manera automática, dirigirse al apartado de automatización.

Crear crawlers para nuevos archivos ya procesados, repitiendo el procedimiento recién realizado.

Crear trabajo ETL en Glue

Dirigirse a trabajos ETL ubicado en el menú izquierdo
Seleccionar “Spark script editor” y crear



AWS Glue Studio Info

Create job Info

Create

○ Visual with a source and target
Start with a source, ApplyMapping transform, and target.

○ Visual with a blank canvas
Author using an interactive visual interface.

● Spark script editor
Write or upload your own Spark code.

○ Python Shell script editor
Write or upload your own Python shell script.

○ Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

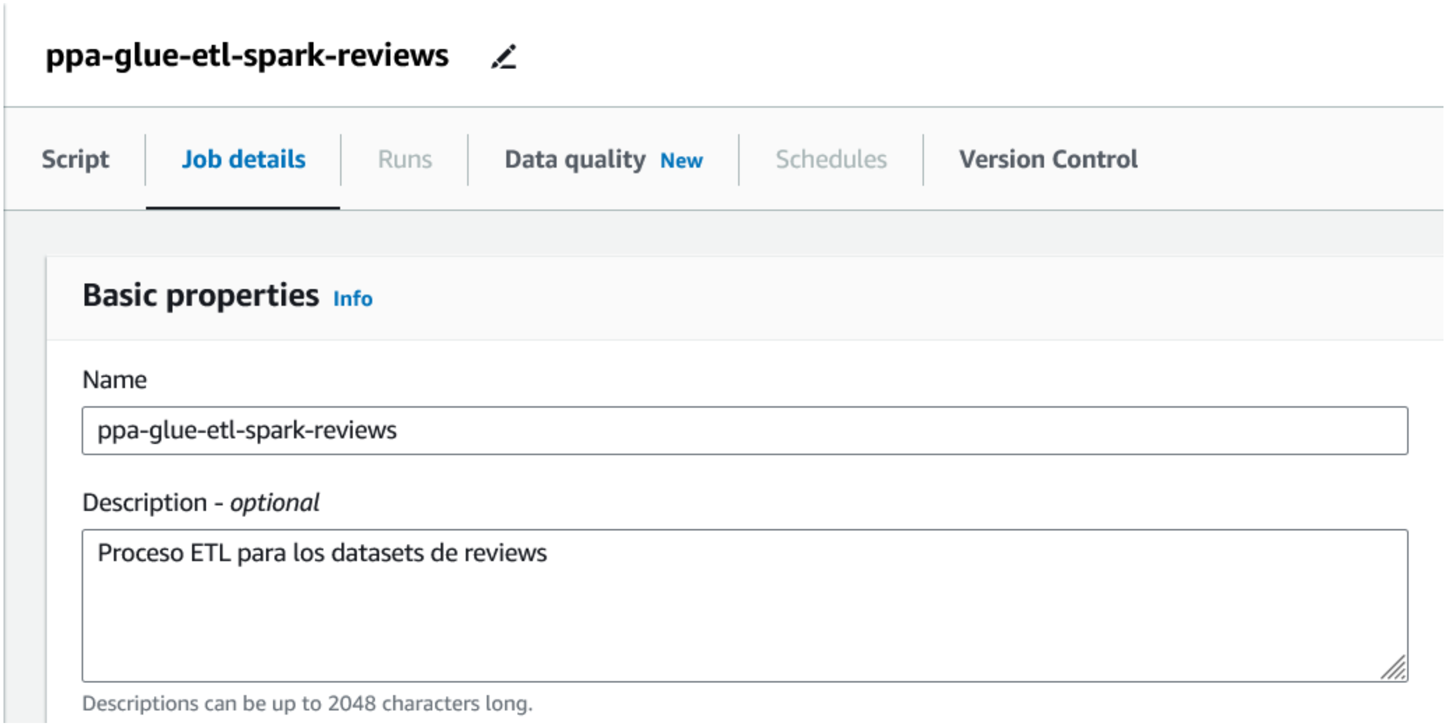
○ Ray script editor New
Write your own code to run on Ray.

Options Info

● Create a new script with boilerplate code

○ Upload and edit an existing script
Choose a local file.

En la pestaña de detalles de trabajo configurar lo siguiente:
Ingresar el nombre del trabajo



ppa-glue-etl-spark-reviews

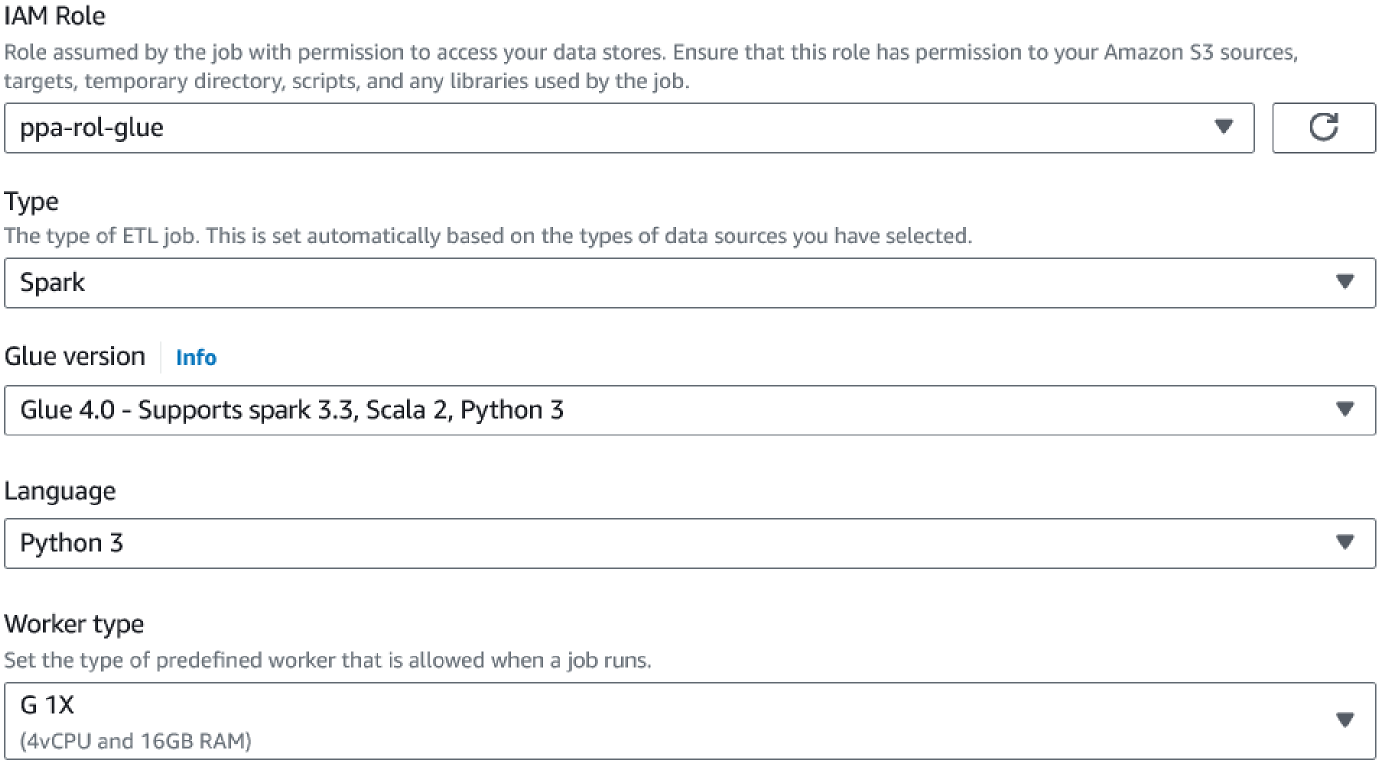
Basic properties

Name: ppa-glue-etl-spark-reviews

Description - *optional*: Proceso ETL para los datasets de reviews

Descriptions can be up to 2048 characters long.

Seleccionar el rol



IAM Role
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.
ppa-rol-glue

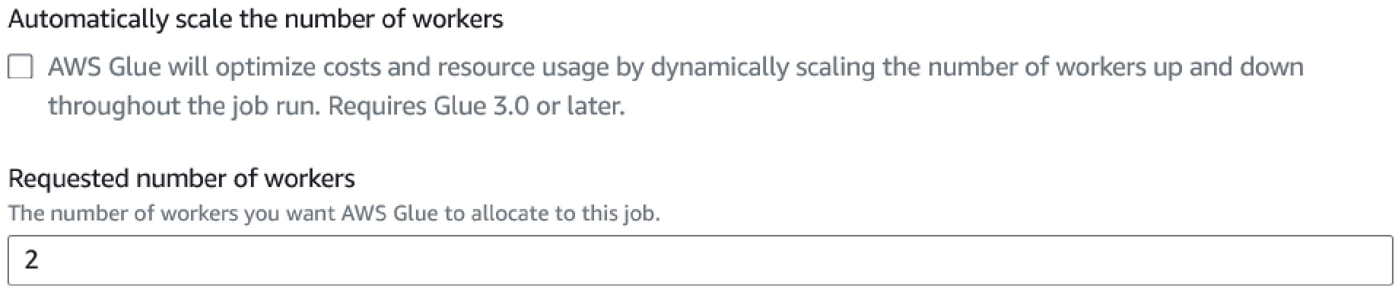
Type
The type of ETL job. This is set automatically based on the types of data sources you have selected.
Spark

Glue version
Glue 4.0 - Supports spark 3.3, Scala 2, Python 3

Language
Python 3

Worker type
Set the type of predefined worker that is allowed when a job runs.
G 1X
(4vCPU and 16GB RAM)

Seleccionar el número de trabajadores



Automatically scale the number of workers

□ AWS Glue will optimize costs and resource usage by dynamically scaling the number of workers up and down throughout the job run. Requires Glue 3.0 or later.

Requested number of workers

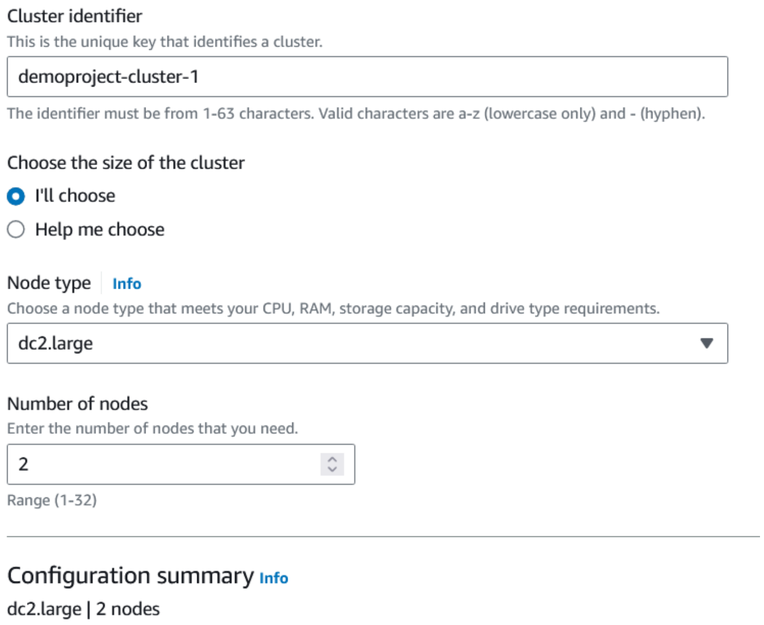
The number of workers you want AWS Glue to allocate to this job.

2

Presionar el botón “Guardar”, para evitar pérdidas en caso de desconexión
Modificar el código utilizando el script “glue_etl_spark_review”
Realizar el mismo procedimiento para los dataset de meta

Conexión Glue - Redshift

Crear un Cluster



Cluster identifier

This is the unique key that identifies a cluster.

demoproject-cluster-1

The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).

Choose the size of the cluster

● I'll choose

○ Help me choose

Node type

Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.

dc2.large

Number of nodes

Enter the number of nodes that you need.

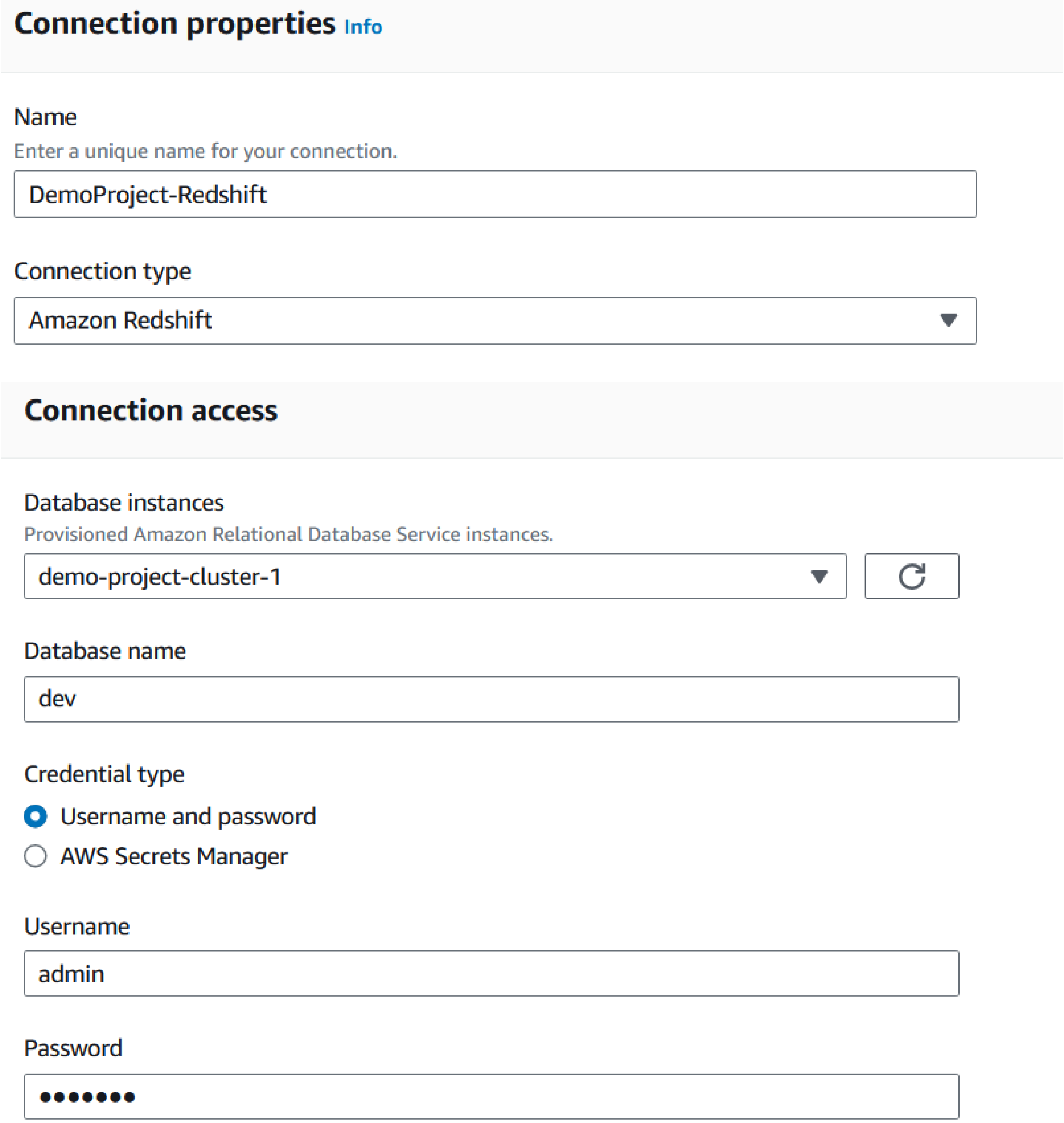
2

Range (1-32)

Configuration summary

dc2.large | 2 nodes

Crear Conexión a Redshift



Connection properties

Name

Enter a unique name for your connection.

DemoProject-Redshift

Connection type

Amazon Redshift

Connection access

Database instances

Provisioned Amazon Relational Database Service instances.

demo-project-cluster-1

Database name

dev

Credential type

● Username and password

○ AWS Secrets Manager

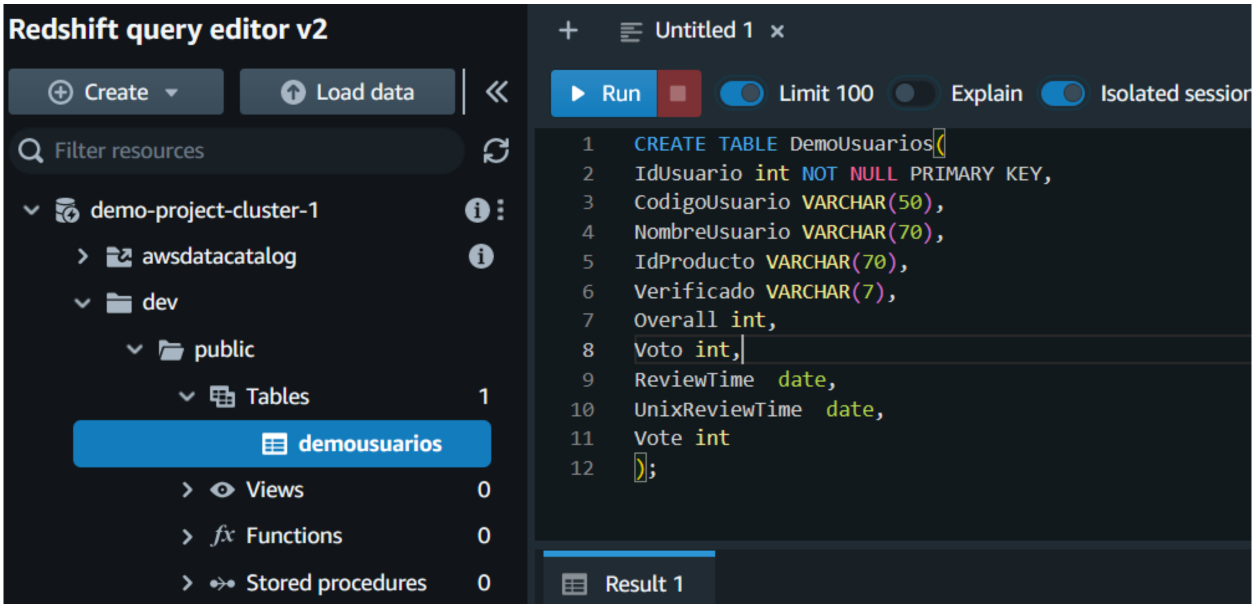
Username

admin

Password

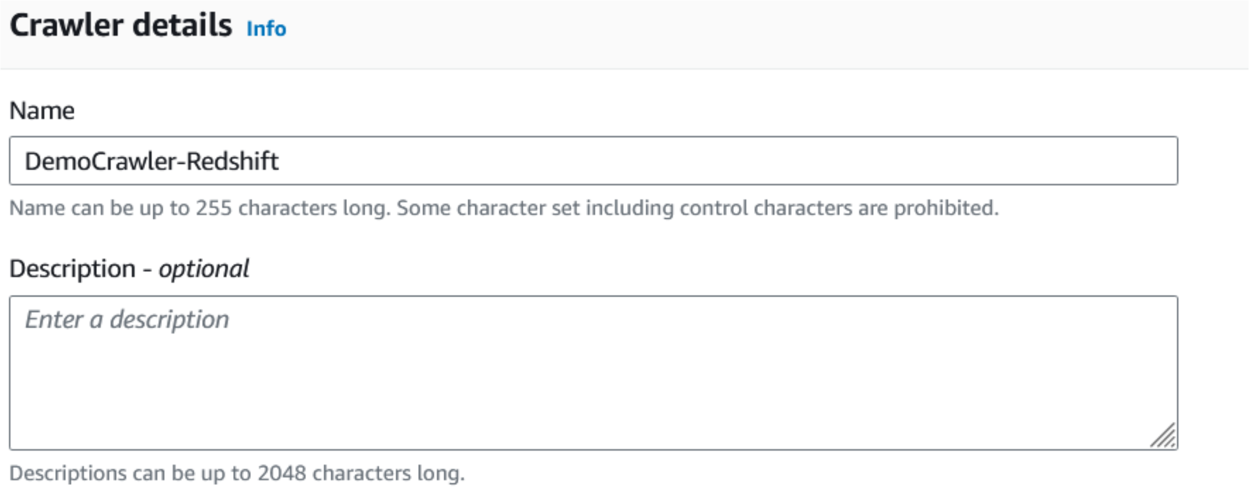
●●●●●●●

Crear tabla en Query



```sql
CREATE TABLE DemoUsuarios(
IdUsuario int NOT NULL PRIMARY KEY,
CodigoUsuario VARCHAR(50),
NombreUsuario VARCHAR(70),
IdProducto VARCHAR(70),
Verificado VARCHAR(7),
Overall int,
Voto int,
ReviewTime  date,
UnixReviewTime  date,
Vote int
);
```

Crear Crawler para Redshift



Crawler details Info

Name

DemoCrawler-Redshift

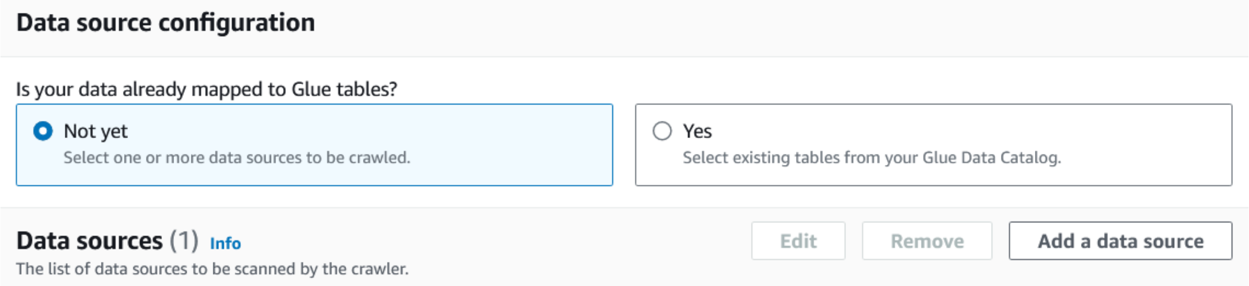
Name can be up to 255 characters long. Some character set including control characters are prohibited.

Description - *optional*

Enter a description

Descriptions can be up to 2048 characters long.

Añadir la fuente de la información



Data source configuration

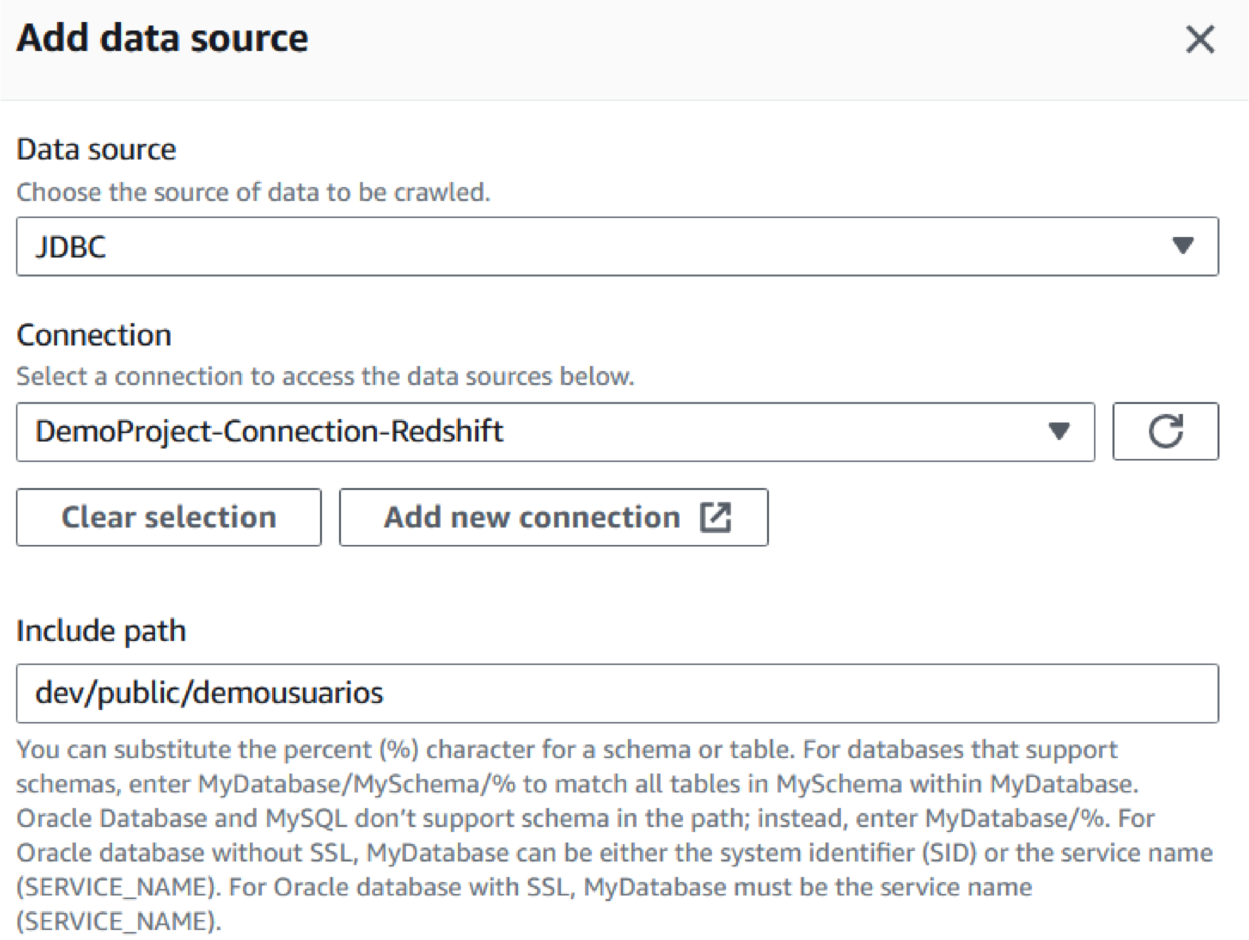
Is your data already mapped to Glue tables?

● Not yet
Select one or more data sources to be crawled.

○ Yes
Select existing tables from your Glue Data Catalog.

Data sources (1) Info

The list of data sources to be scanned by the crawler.



Add data source

Data source

Choose the source of data to be crawled.

JDBC

Connection

Select a connection to access the data sources below.

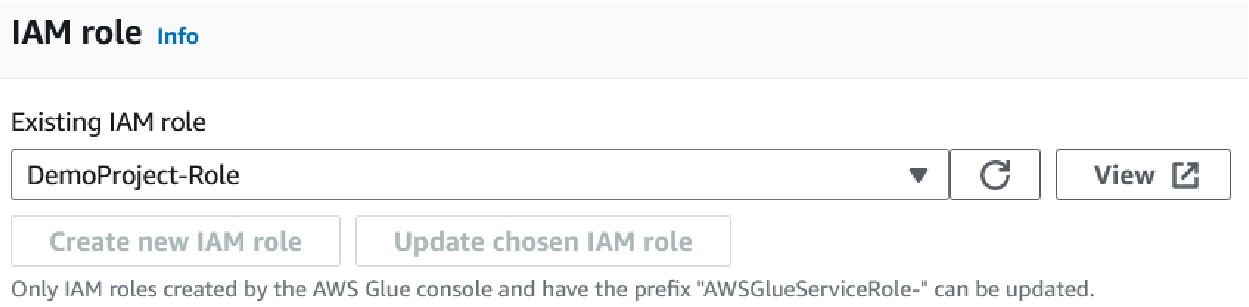
DemoProject-Connection-Redshift

Include path

dev/public/demousuarios

You can substitute the percent (%) character for a schema or table. For databases that support schemas, enter MyDatabase/MySchema/% to match all tables in MySchema within MyDatabase. Oracle Database and MySQL don't support schema in the path; instead, enter MyDatabase/%. For Oracle database without SSL, MyDatabase can be either the system identifier (SID) or the service name (SERVICE_NAME). For Oracle database with SSL, MyDatabase must be the service name (SERVICE_NAME).

Colocar el rol creado anteriormente



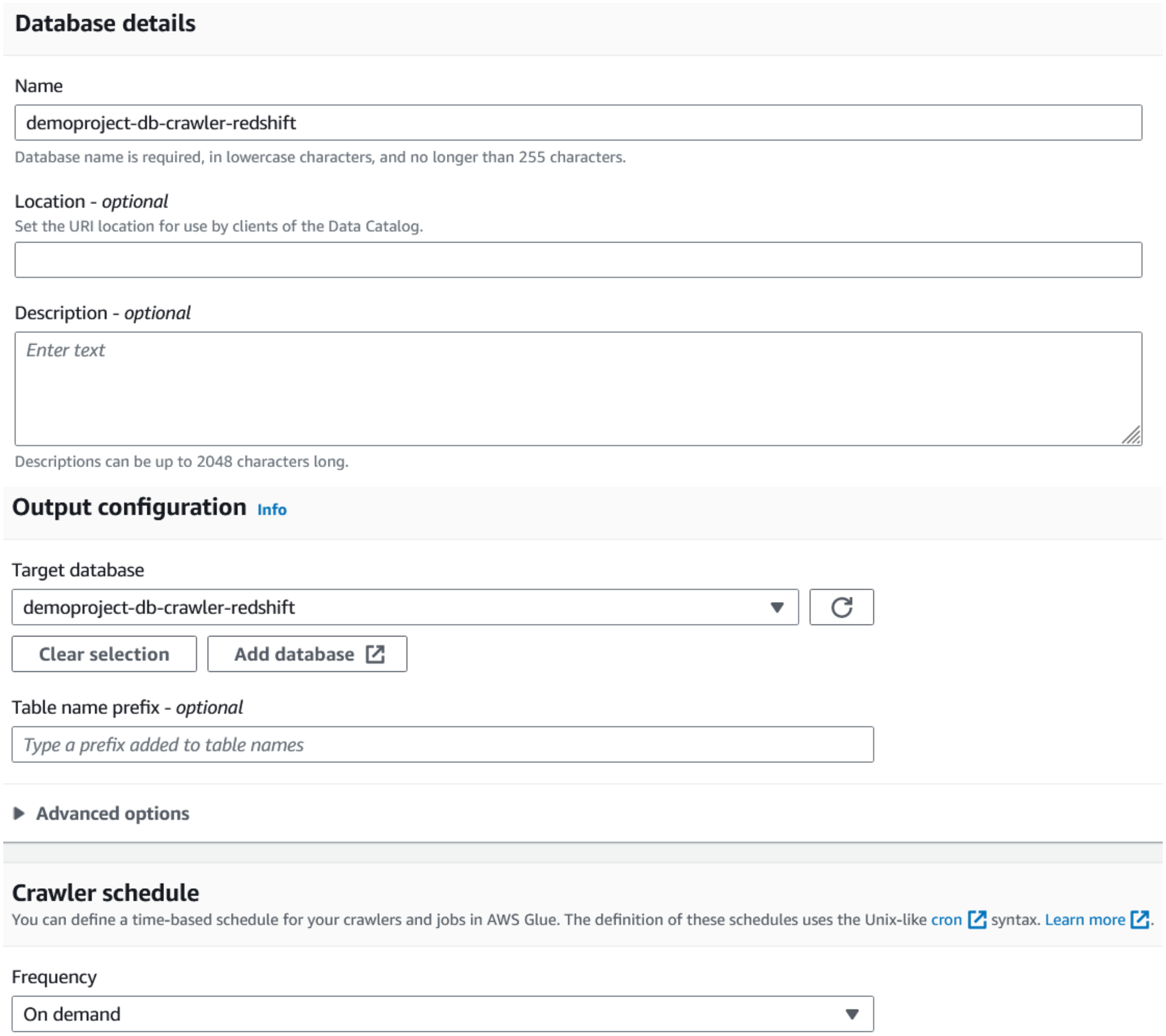
IAM role Info

Existing IAM role

DemoProject-Role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Crear la tabla temporal que se crea con el crawler



Database details

Name

demoproject-db-crawler-redshift

Database name is required, in lowercase characters, and no longer than 255 characters.

Location - *optional*

Set the URI location for use by clients of the Data Catalog.

Description - *optional*

Enter text

Descriptions can be up to 2048 characters long.

Output configuration Info

Target database

demoproject-db-crawler-redshift

Table name prefix - *optional*

Type a prefix added to table names

Crawler schedule

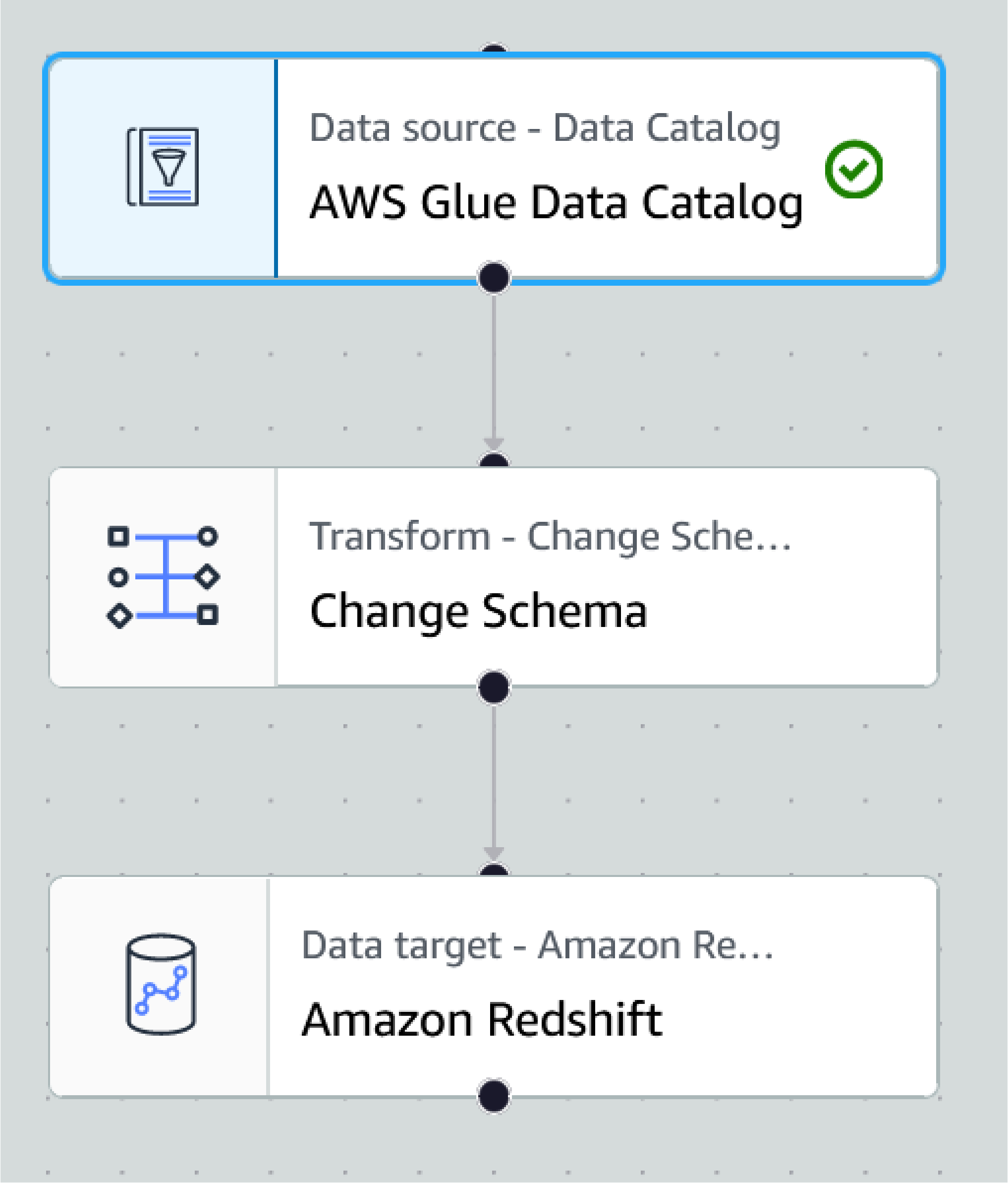
You can define a time-based schedule for your crawlers and jobs in AWS Glue. The definition of these schedules uses the Unix-like cron syntax. Learn more.

Frequency

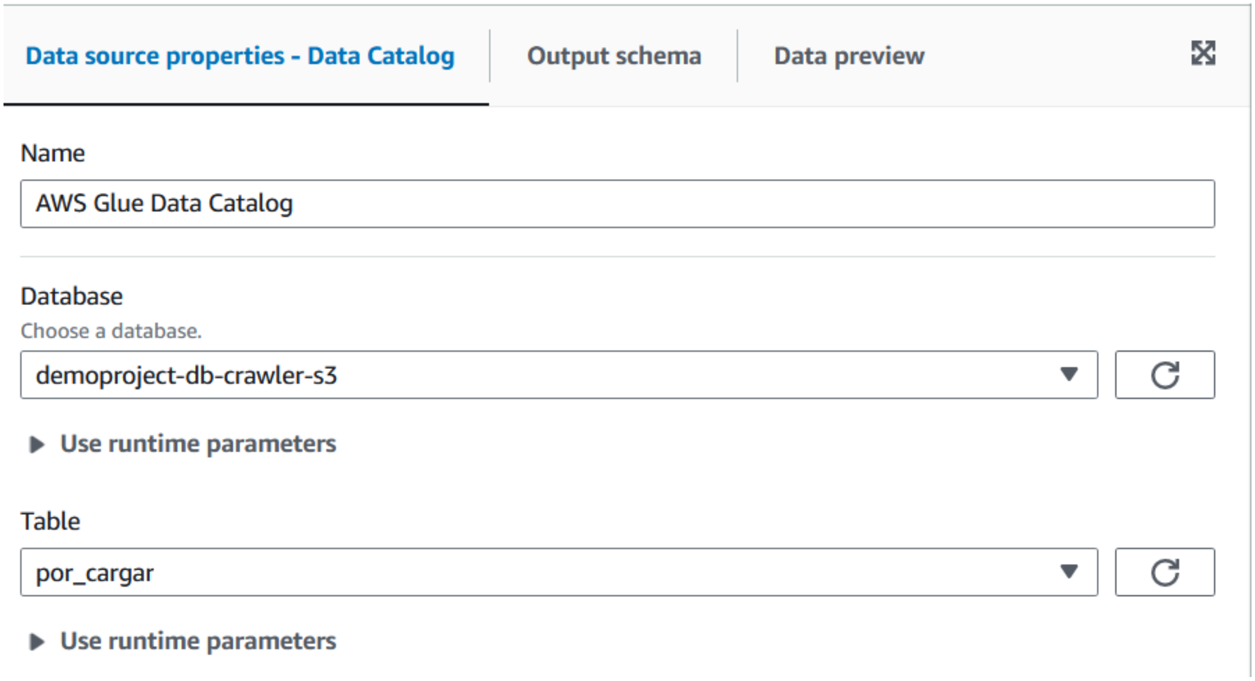
On demand

Crear el Trabajo de ETL

Utilizando Visual ETL Jobs - blank canvas
Insertar “Source” - ”Transformations” - “Target”



Configuración de “Source”

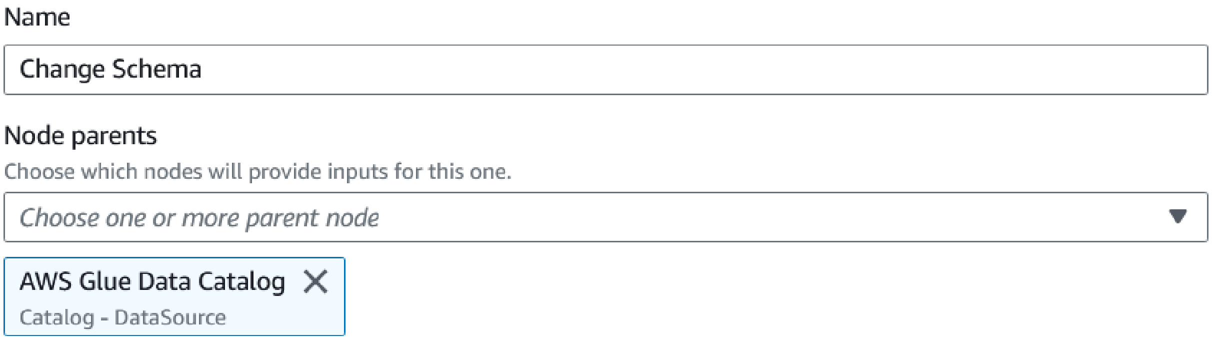


Name: AWS Glue Data Catalog

Database
Choose a database.
demoproject-db-crawler-s3

Table: por_cargar

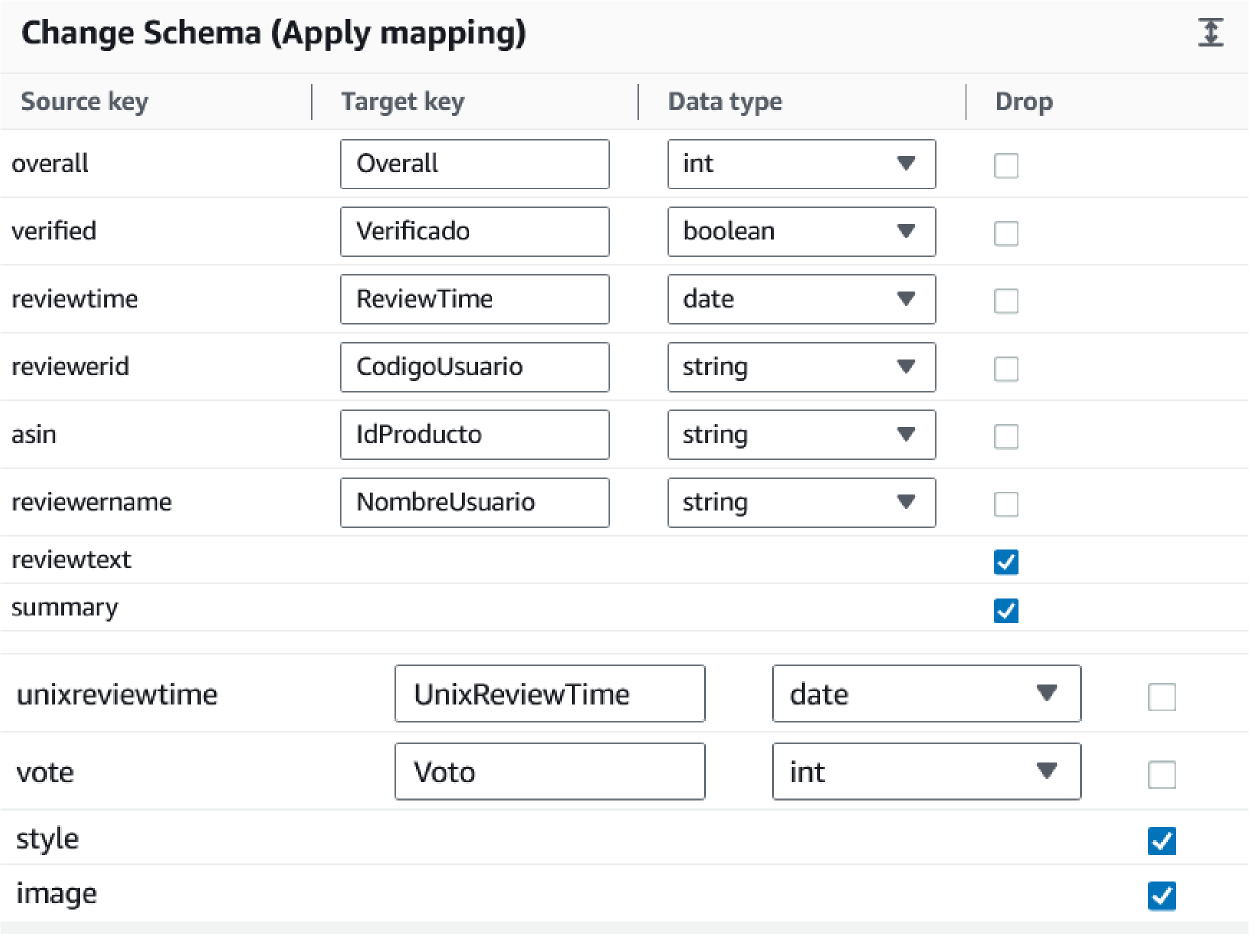
Transformaciones



Name: Change Schema

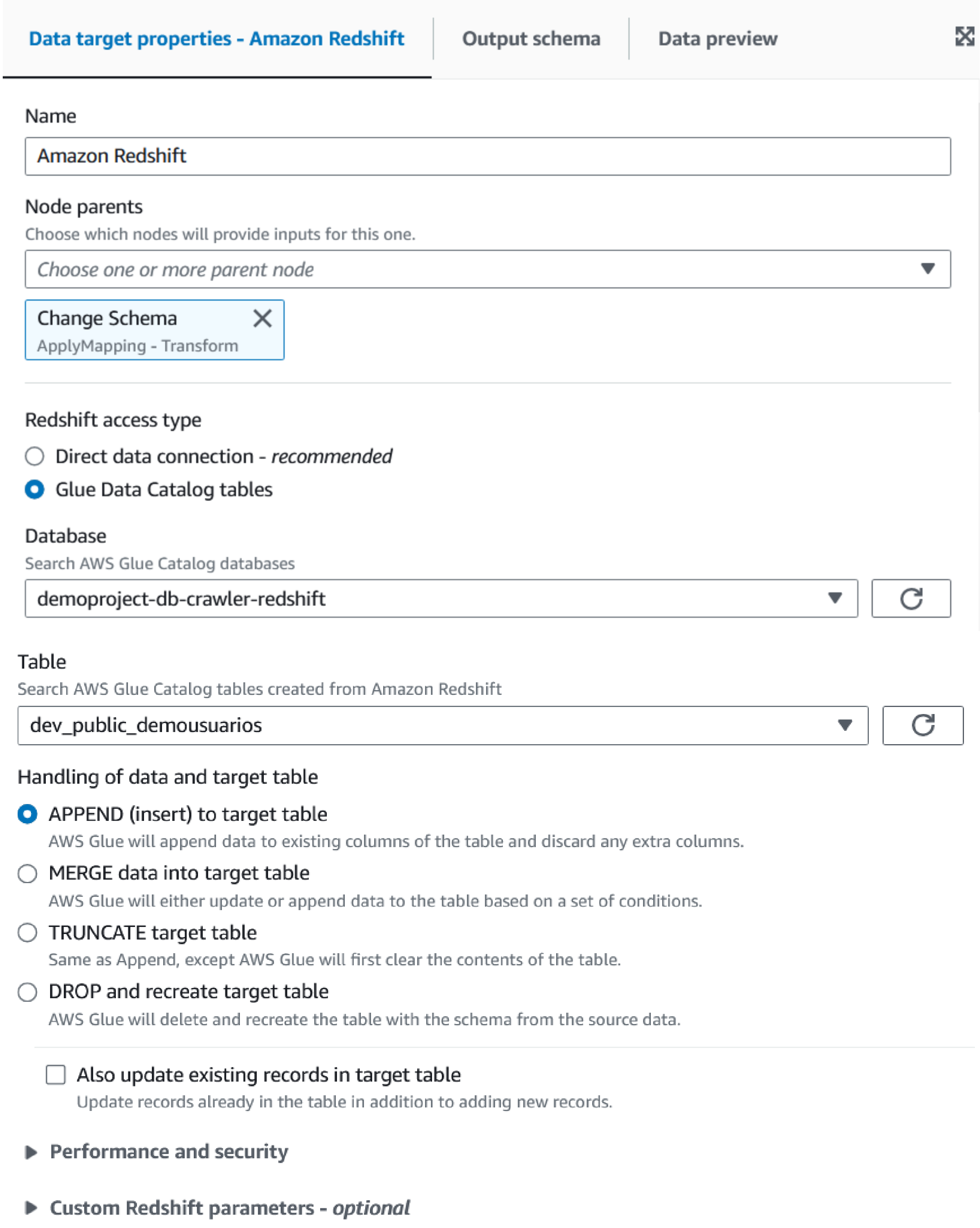
Node parents
Choose which nodes will provide inputs for this one.
Choose one or more parent node

AWS Glue Data Catalog
Catalog - DataSource



Change Schema (Apply mapping)

Source key	Target key	Data type	Drop
overall	Overall	int	☐
verified	Verificado	boolean	☐
reviewtime	ReviewTime	date	☐
reviewerid	CodigoUsuario	string	☐
asin	IdProducto	string	☐
reviewername	NombreUsuario	string	☐
reviewtext			☑
summary			☑
unixreviewtime	UnixReviewTime	date	☐
vote	Voto	int	☐
style			☑
image			☑



Name

Amazon Redshift

Node parents

Choose which nodes will provide inputs for this one.

Choose one or more parent node

Change Schema ✕
ApplyMapping - Transform

Redshift access type

○ Direct data connection - *recommended*

● Glue Data Catalog tables

Database

Search AWS Glue Catalog databases

demoproject-db-crawler-redshift

Table

Search AWS Glue Catalog tables created from Amazon Redshift

dev_public_demousuarios

Handling of data and target table

● APPEND (insert) to target table
AWS Glue will append data to existing columns of the table and discard any extra columns.

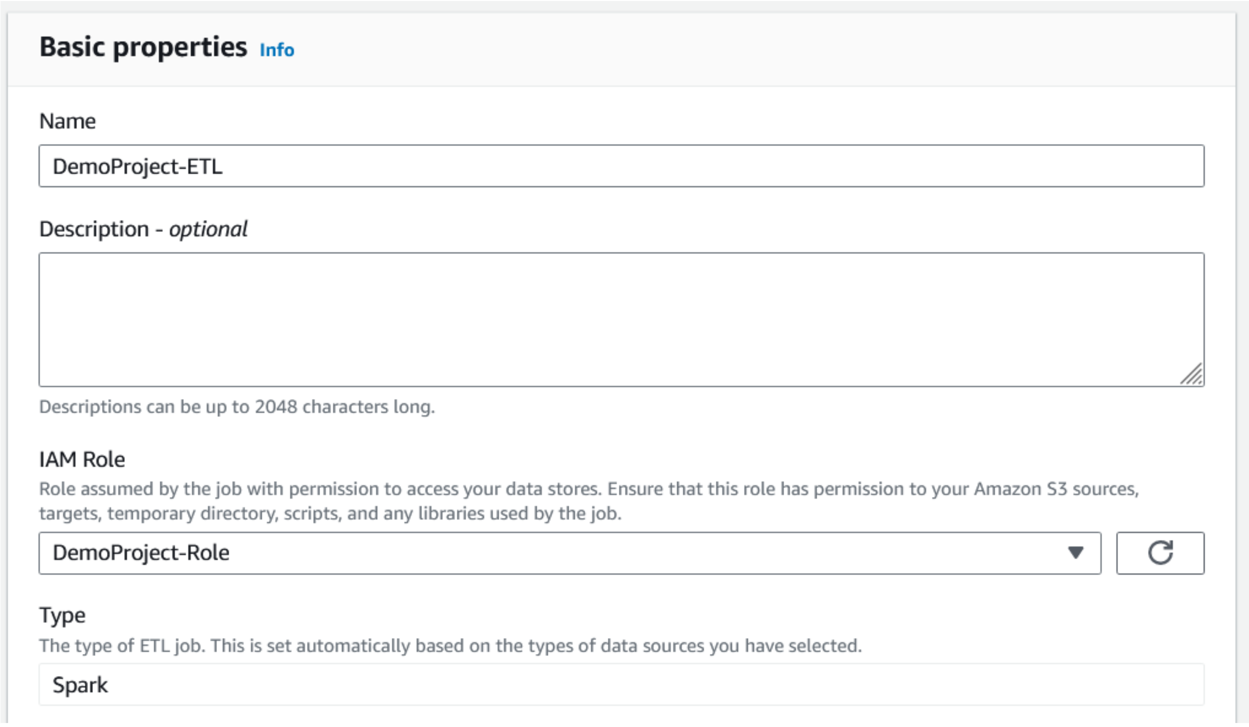
○ MERGE data into target table
AWS Glue will either update or append data to the table based on a set of conditions.

○ TRUNCATE target table
Same as Append, except AWS Glue will first clear the contents of the table.

○ DROP and recreate target table
AWS Glue will delete and recreate the table with the schema from the source data.

□ Also update existing records in target table
Update records already in the table in addition to adding new records.

▶ Performance and security
▶ Custom Redshift parameters - *optional*



Basic properties

Name

DemoProject-ETL

Description - *optional*

Descriptions can be up to 2048 characters long.

IAM Role

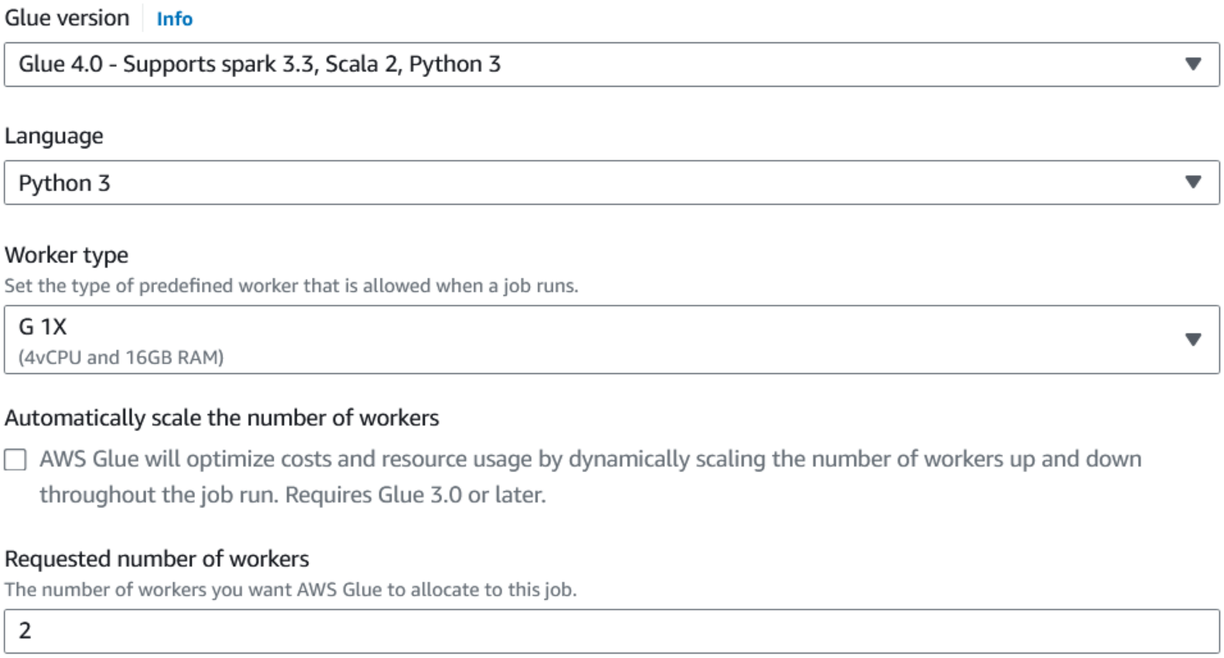
Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

DemoProject-Role

Type

The type of ETL job. This is set automatically based on the types of data sources you have selected.

Spark



Glue version

Glue 4.0 - Supports spark 3.3, Scala 2, Python 3

Language

Python 3

Worker type

Set the type of predefined worker that is allowed when a job runs.

G 1X
(4vCPU and 16GB RAM)

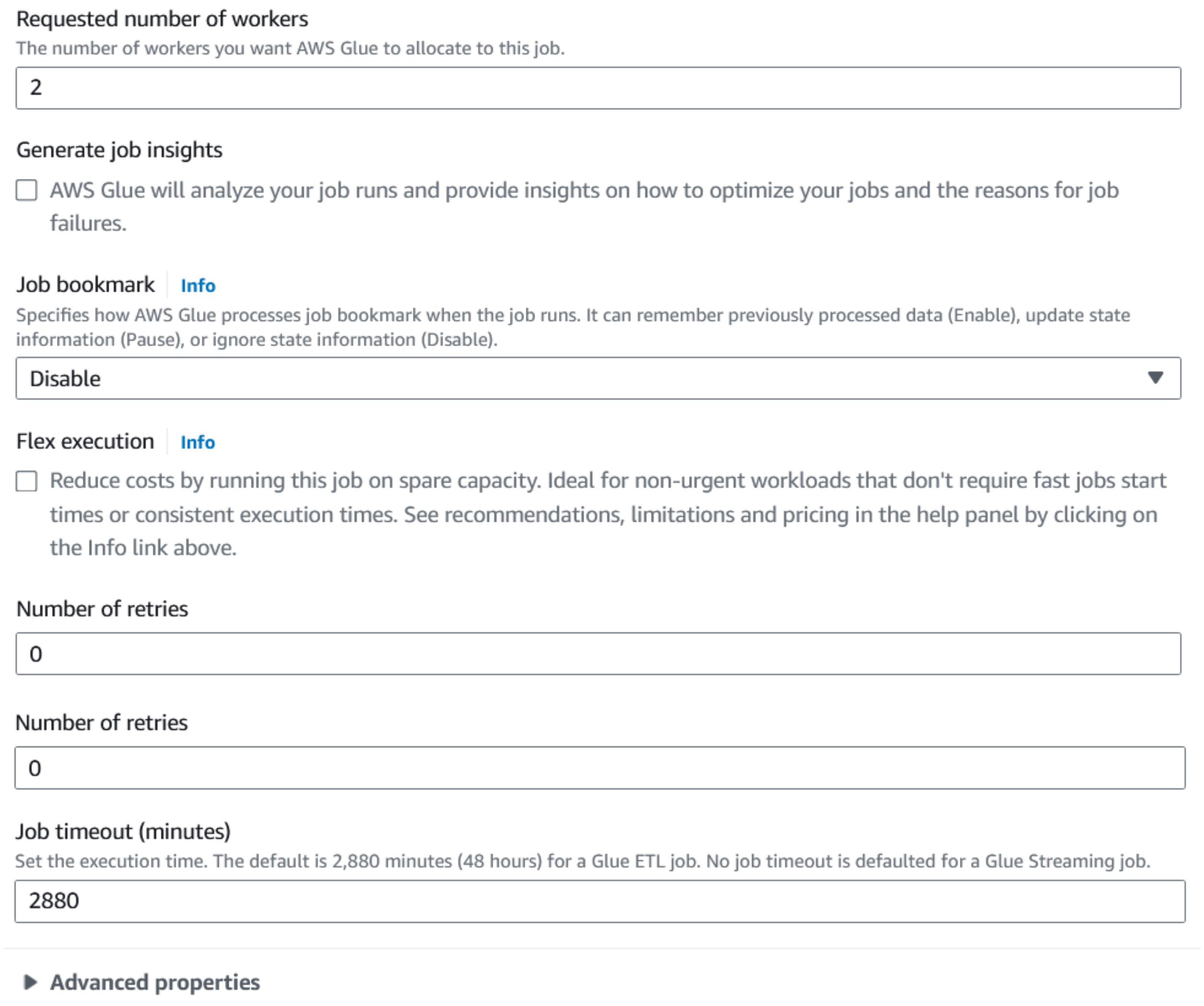
Automatically scale the number of workers

☐ AWS Glue will optimize costs and resource usage by dynamically scaling the number of workers up and down throughout the job run. Requires Glue 3.0 or later.

Requested number of workers

The number of workers you want AWS Glue to allocate to this job.

2



Requested number of workers

The number of workers you want AWS Glue to allocate to this job.

2

Generate job insights

□ AWS Glue will analyze your job runs and provide insights on how to optimize your jobs and the reasons for job failures.

Job bookmark | Info

Specifies how AWS Glue processes job bookmark when the job runs. It can remember previously processed data (Enable), update state information (Pause), or ignore state information (Disable).

Disable ▼

Flex execution | Info

□ Reduce costs by running this job on spare capacity. Ideal for non-urgent workloads that don't require fast jobs start times or consistent execution times. See recommendations, limitations and pricing in the help panel by clicking on the Info link above.

Number of retries

0

Number of retries

0

Job timeout (minutes)

Set the execution time. The default is 2,880 minutes (48 hours) for a Glue ETL job. No job timeout is defaulted for a Glue Streaming job.

2880

▶ Advanced properties

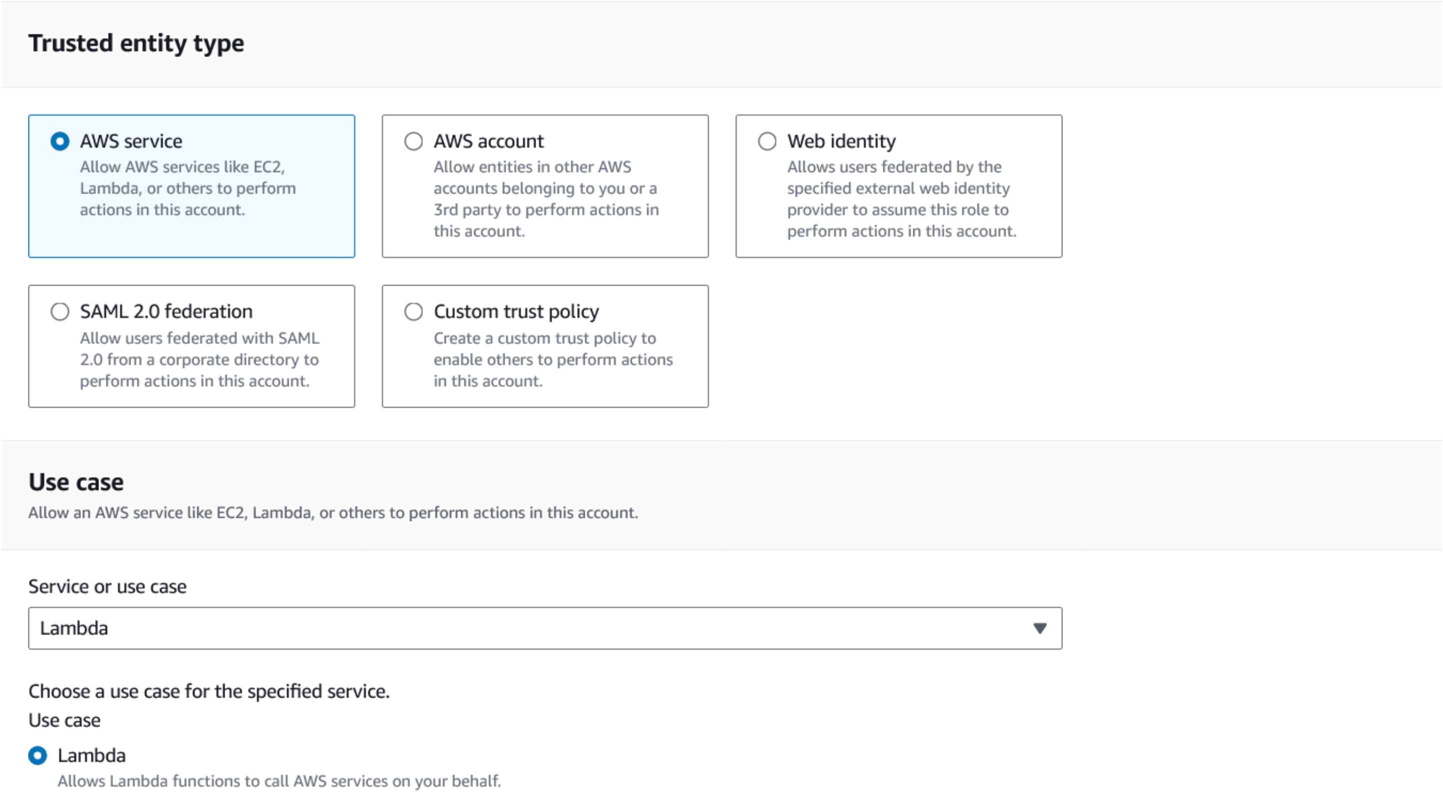
Automatización carga al Data warehouse

Detectar nuevo documento en almacenamiento

Creación Rol para servicio Lambda

Ingresar al servicio de IAM, opción Roles

Crear nuevo rol, con servicio de lambda



Trusted entity type

● AWS service
Allow AWS services like EC2, Lambda, or others to perform actions in this account.

○ AWS account
Allow entities in other AWS accounts belonging to you or a 3rd party to perform actions in this account.

○ Web identity
Allows users federated by the specified external web identity provider to assume this role to perform actions in this account.

○ SAML 2.0 federation
Allow users federated with SAML 2.0 from a corporate directory to perform actions in this account.

○ Custom trust policy
Create a custom trust policy to enable others to perform actions in this account.

Use case

Allow an AWS service like EC2, Lambda, or others to perform actions in this account.

Service or use case

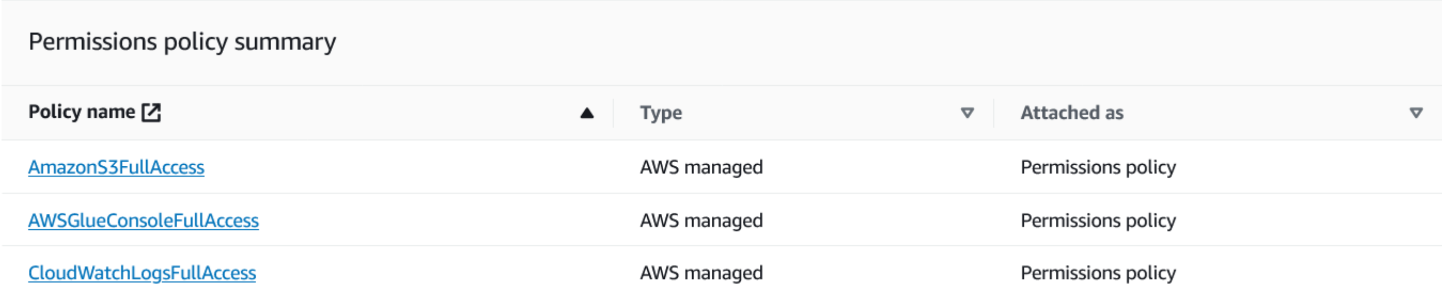
Lambda ▼

Choose a use case for the specified service.

Use case

● Lambda
Allows Lambda functions to call AWS services on your behalf.

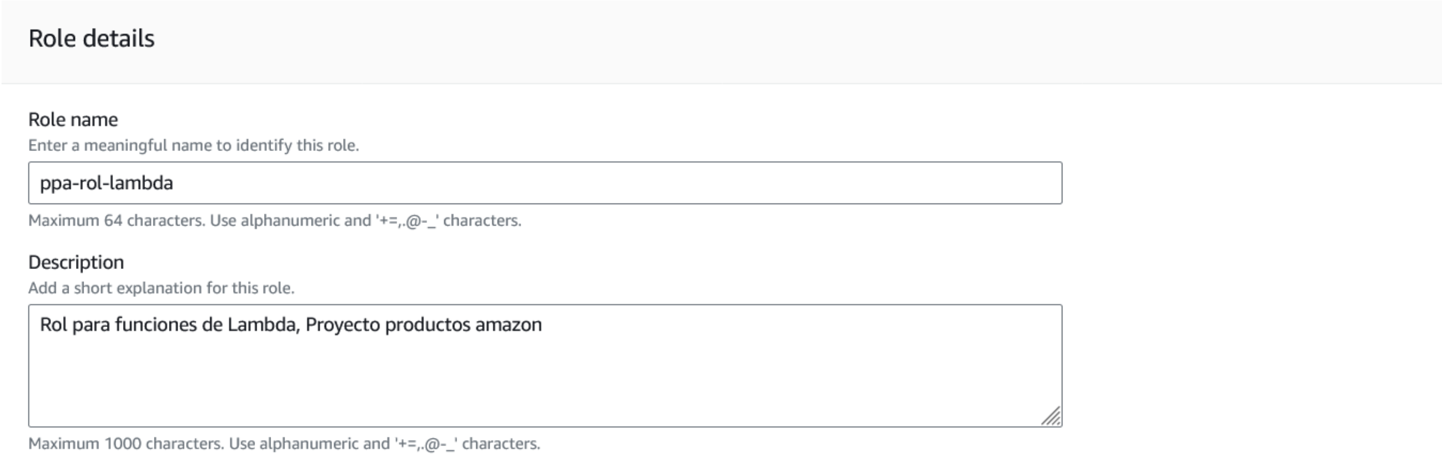
Seleccionar los permisos



Permissions policy summary

Policy name ▲	Type	Attached as
AmazonS3FullAccess	AWS managed	Permissions policy
AWSGlueConsoleFullAccess	AWS managed	Permissions policy
CloudWatchLogsFullAccess	AWS managed	Permissions policy

Ingresar el nombre del rol



Role details

Role name
Enter a meaningful name to identify this role.

ppa-rol-lambda

Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.

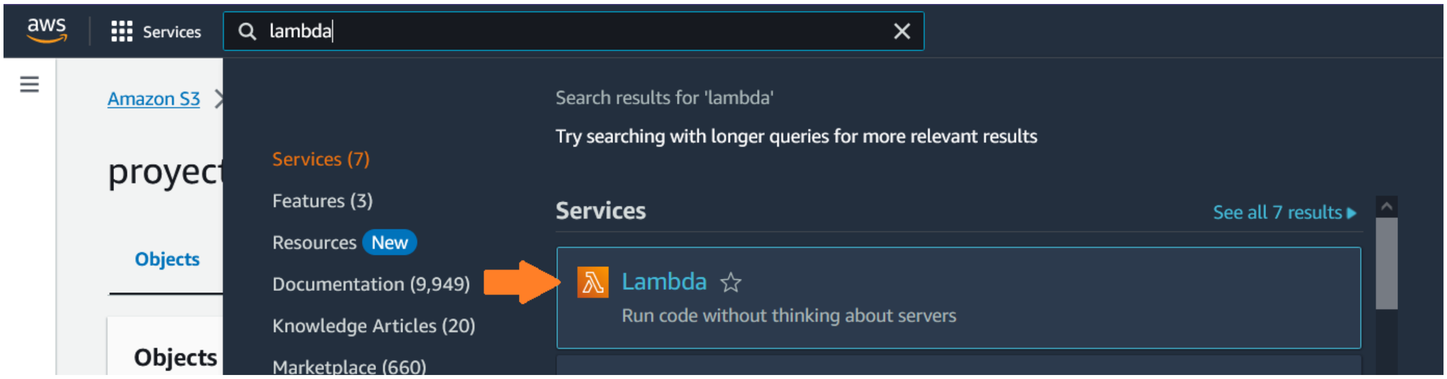
Description
Add a short explanation for this role.

Rol para funciones de Lambda, Proyecto productos amazon

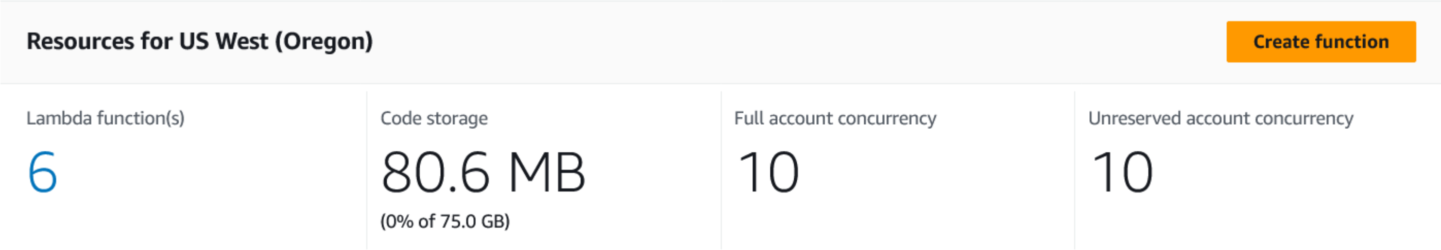
Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.

Crear la 1ra función Lambda, para empezar proceso de carga

Dirigirse al servicio de Lambda



Crear nueva función

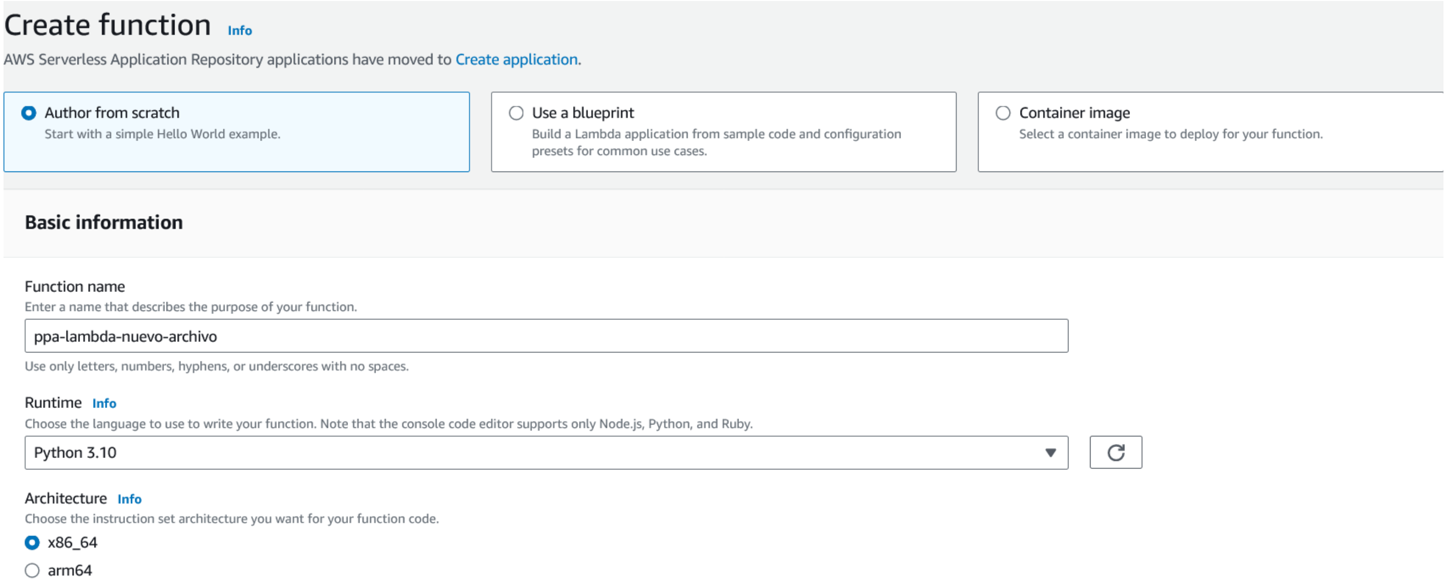


Resources for US West (Oregon)

Create function

Lambda function(s)	Code storage	Full account concurrency	Unreserved account concurrency
6	80.6 MB (0% of 75.0 GB)	10	10

Ingresar nombre, seleccionar lenguaje de programación y guardar



Create function Info

AWS Serverless Application Repository applications have moved to Create application.

● Author from scratch — Start with a simple Hello World example.
○ Use a blueprint — Build a Lambda application from sample code and configuration presets for common use cases.
○ Container image — Select a container image to deploy for your function.

Basic information

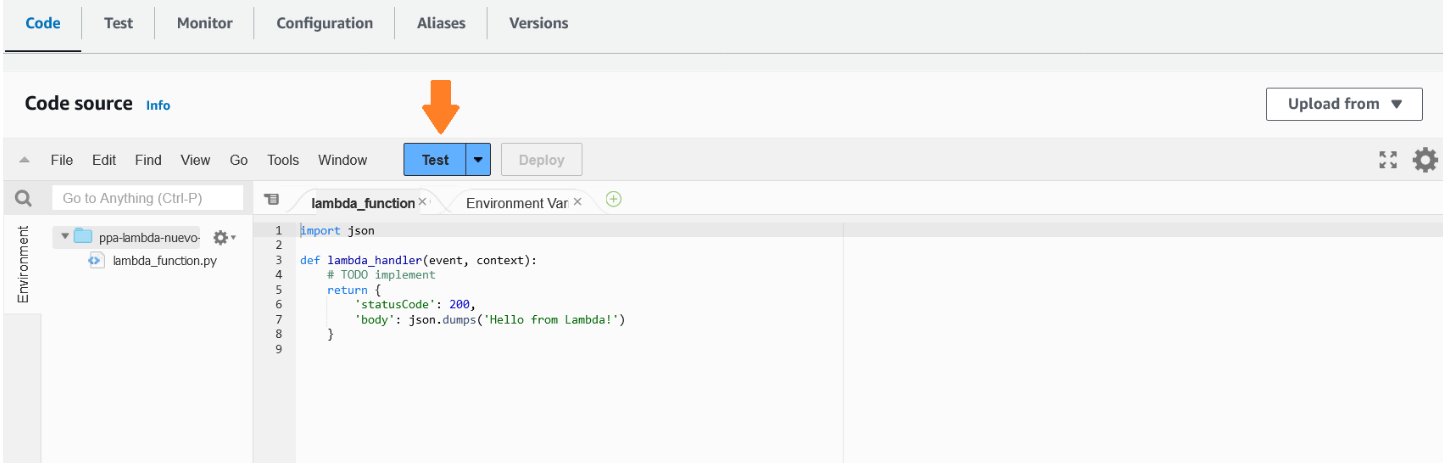
Function name
Enter a name that describes the purpose of your function.
ppa-lambda-nuevo-archivo
Use only letters, numbers, hyphens, or underscores with no spaces.

Runtime Info
Choose the language to use to write your function. Note that the console code editor supports only Node.js, Python, and Ruby.
Python 3.10

Architecture Info
Choose the instruction set architecture you want for your function code.
● x86_64
○ arm64

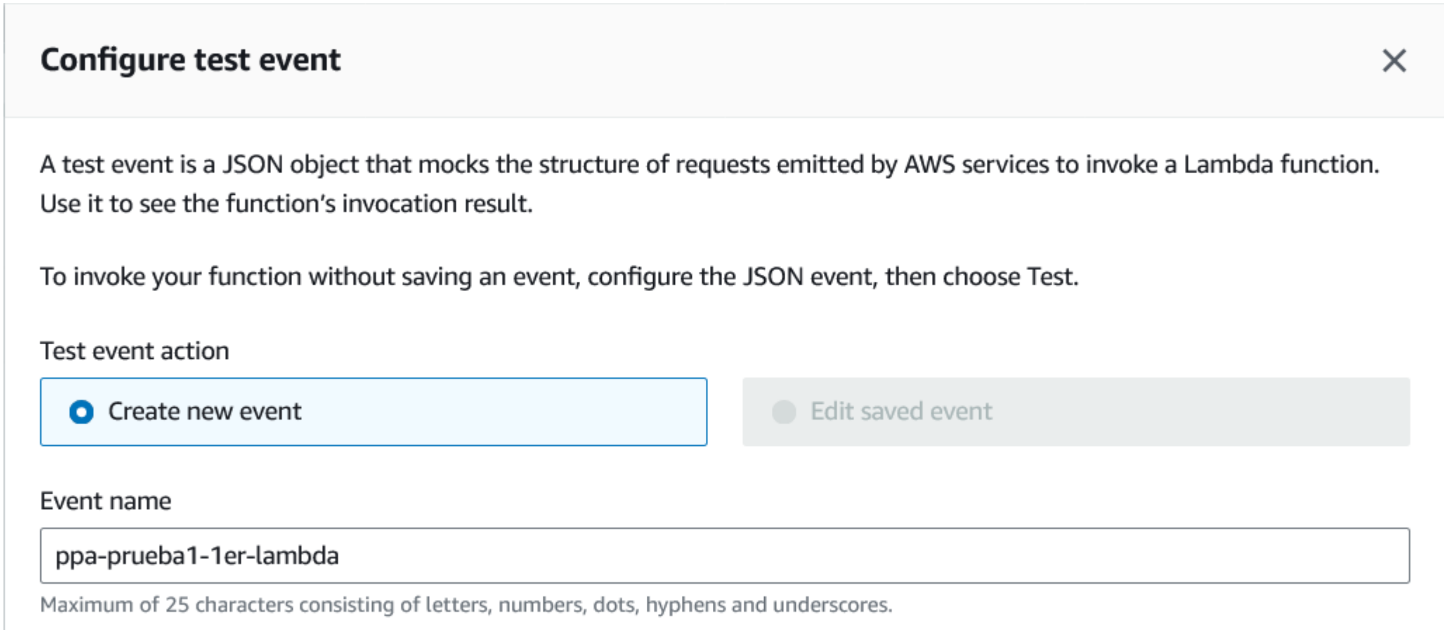
Activar funcionalidad de Cloudwatch

Para ello se realiza una prueba de la función.
Presionar el botón “Test”



```
import json

def lambda_handler(event, context):
    # TODO implement
    return {
        'statusCode': 200,
        'body': json.dumps('Hello from Lambda!')
    }
```

Pide crear una nueva prueba, Ingresar el nombre y guardar


Configure test event ✕

A test event is a JSON object that mocks the structure of requests emitted by AWS services to invoke a Lambda function. Use it to see the function's invocation result.

To invoke your function without saving an event, configure the JSON event, then choose Test.

Test event action

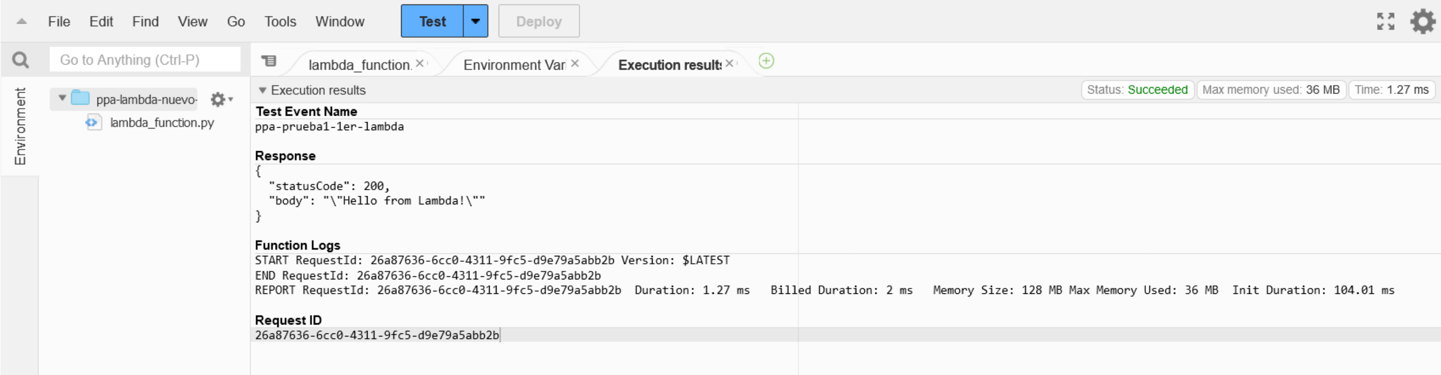
● Create new event ○ Edit saved event

Event name

ppa-prueba1-1er-lambda

Maximum of 25 characters consisting of letters, numbers, dots, hyphens and underscores.

Ejecutar la prueba, presionando el botón "Test"



▼ Execution results | Status: Succeeded | Max memory used: 36 MB | Time: 1.27 ms

Test Event Name

```
ppa-prueba1-1er-lambda
```

Response

```
{
  "statusCode": 200,
  "body": "\"Hello from Lambda!\""
}
```

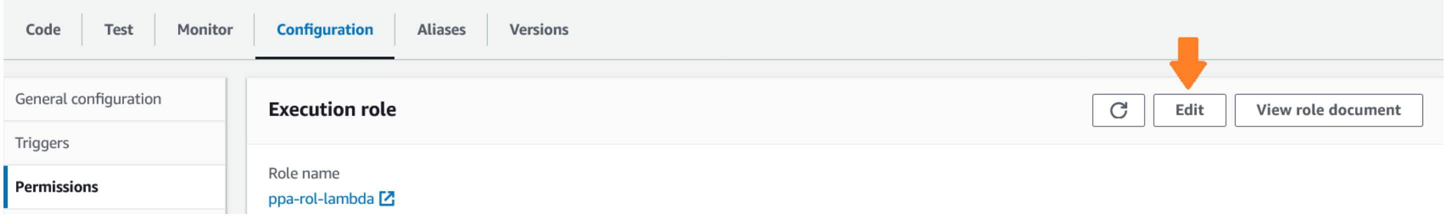
Function Logs

```
START RequestId: 26a87636-6cc0-4311-9fc5-d9e79a5abb2b Version: $LATEST
END RequestId: 26a87636-6cc0-4311-9fc5-d9e79a5abb2b
REPORT RequestId: 26a87636-6cc0-4311-9fc5-d9e79a5abb2b  Duration: 1.27 ms   Billed Duration: 2 ms   Memory Size: 128 MB Max Memory Used: 36 MB  Init Duration: 104.01 ms
```

Request ID

```
26a87636-6cc0-4311-9fc5-d9e79a5abb2b
```

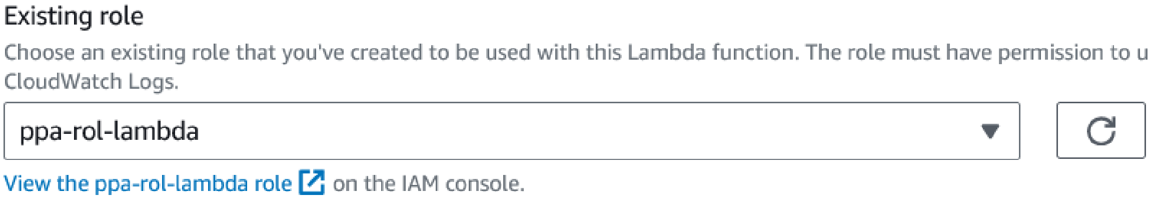
Cambiar el rol de la función por el creado previamente
Ir a la opción de Permisos en la pestaña de Configuración
Presionar el botón editar



Execution role

Role name
ppa-rol-lambda

Al final se encuentra la opción para seleccionar el rol, escoger el que se creó y guardar



Existing role

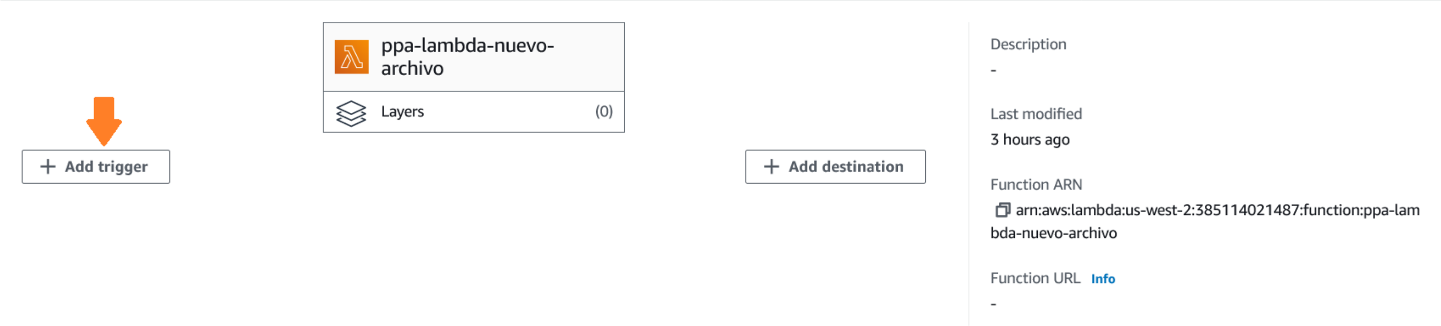
Choose an existing role that you've created to be used with this Lambda function. The role must have permission to upload logs to Amazon CloudWatch Logs.

ppa-rol-lambda

View the ppa-rol-lambda role on the IAM console.

Crear trigger para detectar nuevo documento cargado a s3

Seleccionar añadir trigger



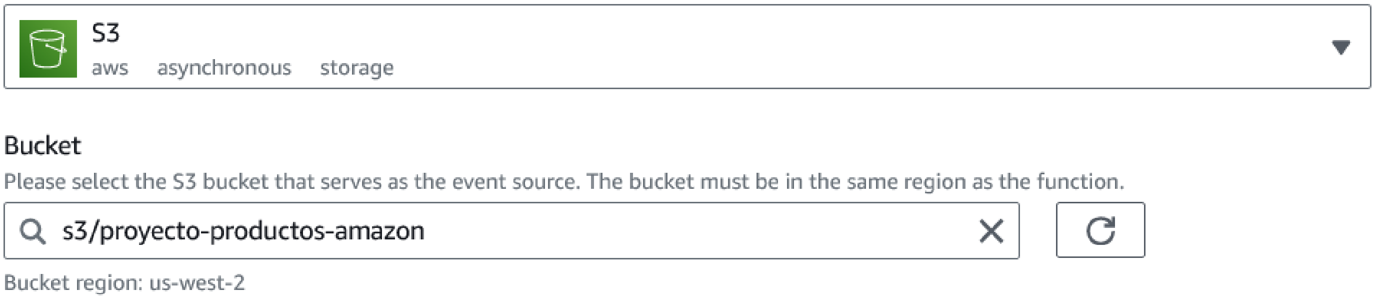
Description
-

Last modified
3 hours ago

Function ARN
arn:aws:lambda:us-west-2:385114021487:function:ppa-lambda-nuevo-archivo

Function URL Info
-

Seleccionar tipo de evento s3, configurar como se observa en las imágenes a continuación y hacer clic en añadir:

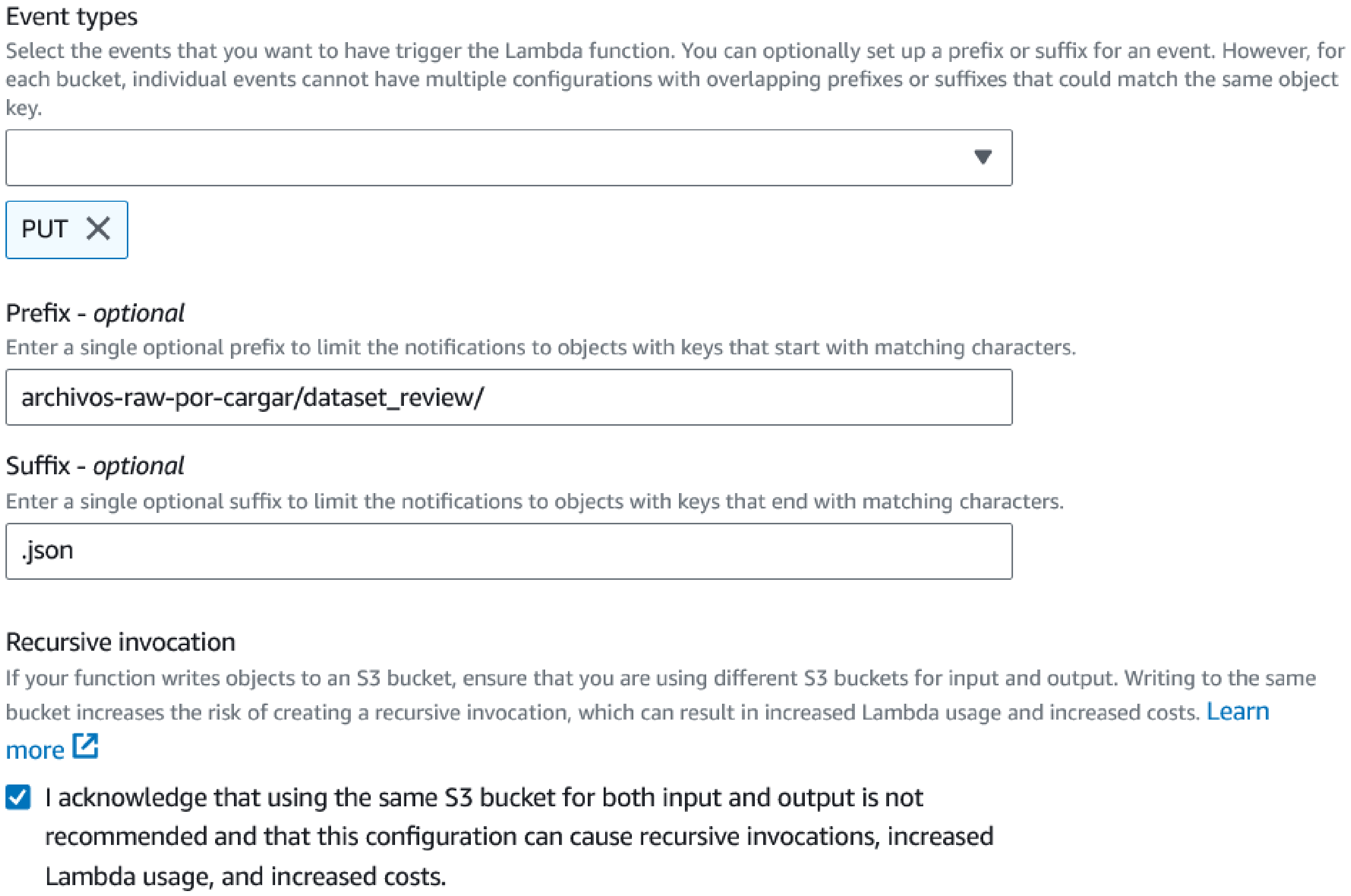


S3
aws asynchronous storage

Bucket
Please select the S3 bucket that serves as the event source. The bucket must be in the same region as the function.

s3/proyecto-productos-amazon

Bucket region: us-west-2



Event types

Select the events that you want to have trigger the Lambda function. You can optionally set up a prefix or suffix for an event. However, for each bucket, individual events cannot have multiple configurations with overlapping prefixes or suffixes that could match the same object key.

PUT ✕

Prefix - *optional*

Enter a single optional prefix to limit the notifications to objects with keys that start with matching characters.

archivos-raw-por-cargar/dataset_review/

Suffix - *optional*

Enter a single optional suffix to limit the notifications to objects with keys that end with matching characters.

.json

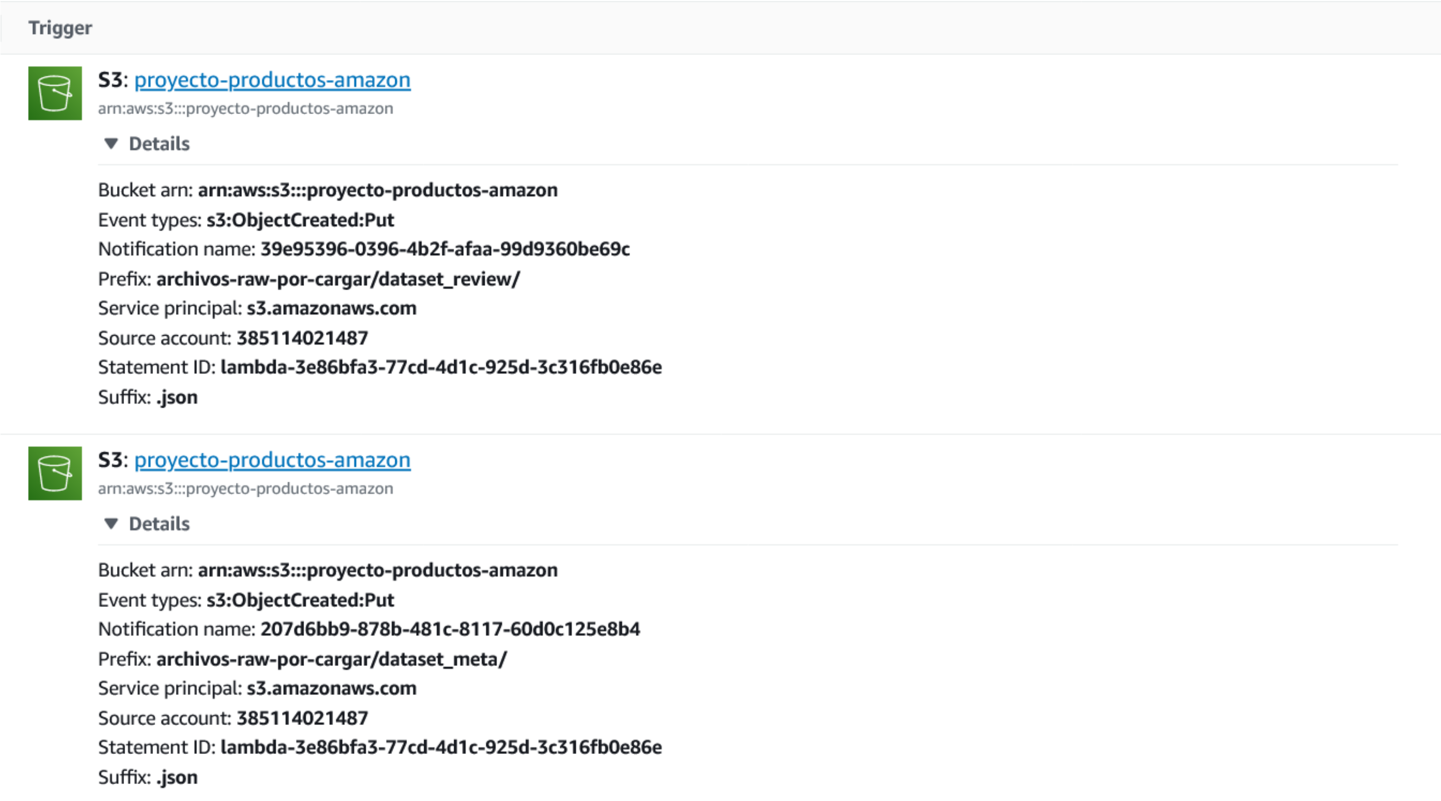
Recursive invocation

If your function writes objects to an S3 bucket, ensure that you are using different S3 buckets for input and output. Writing to the same bucket increases the risk of creating a recursive invocation, which can result in increased Lambda usage and increased costs. Learn more

☑ I acknowledge that using the same S3 bucket for both input and output is not recommended and that this configuration can cause recursive invocations, increased Lambda usage, and increased costs.

Crear un nuevo trigger para la carpeta de “meta”, siguiendo el mismo proceso anterior

En la opción de Triggers en la pestaña de Configuración se puede observar los triggers creados



Trigger

S3: proyecto-productos-amazon
arn:aws:s3:::proyecto-productos-amazon

▼ Details

Bucket arn: **arn:aws:s3:::proyecto-productos-amazon**
Event types: **s3:ObjectCreated:Put**
Notification name: **39e95396-0396-4b2f-afaa-99d9360be69c**
Prefix: **archivos-raw-por-cargar/dataset_review/**
Service principal: **s3.amazonaws.com**
Source account: **385114021487**
Statement ID: **lambda-3e86bfa3-77cd-4d1c-925d-3c316fb0e86e**
Suffix: **.json**

S3: proyecto-productos-amazon
arn:aws:s3:::proyecto-productos-amazon

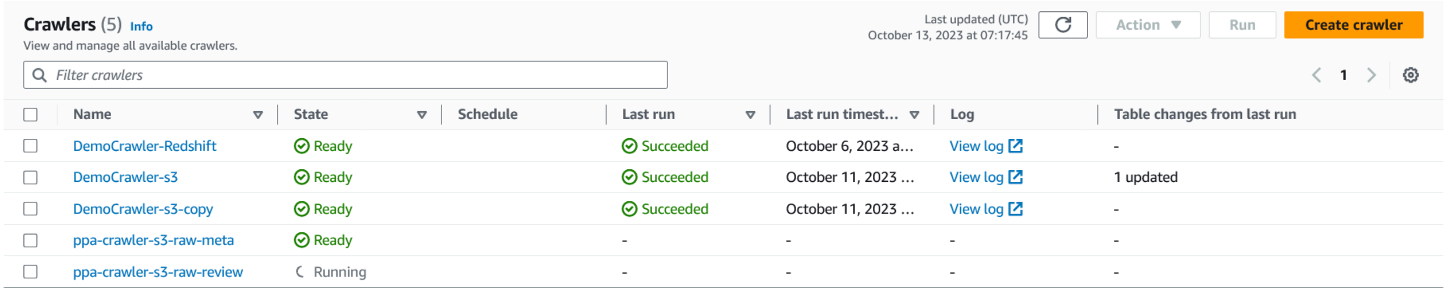
▼ Details

Bucket arn: **arn:aws:s3:::proyecto-productos-amazon**
Event types: **s3:ObjectCreated:Put**
Notification name: **207d6bb9-878b-481c-8117-60d0c125e8b4**
Prefix: **archivos-raw-por-cargar/dataset_meta/**
Service principal: **s3.amazonaws.com**
Source account: **385114021487**
Statement ID: **lambda-3e86bfa3-77cd-4d1c-925d-3c316fb0e86e**
Suffix: **.json**

Modificar la función lambda, utilizando el script de “lambda _1_ new_file”

Probar el funcionamiento

Cargar un archivo en la carpeta review o meta de la carpeta archivos por cargar
Dirigirse a Crawlers en el servicio de Glue
El crawler respectivo se debe estar ejecutando de forma automática



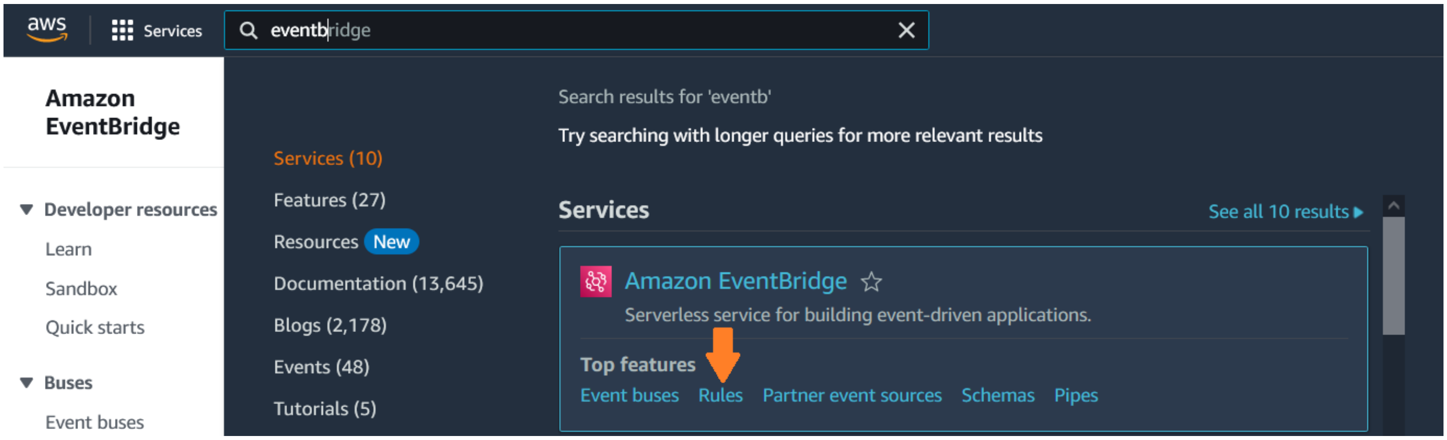
Crawlers (5) Info
View and manage all available crawlers.

Last updated (UTC)
October 13, 2023 at 07:17:45

□	Name	State	Schedule	Last run	Last run timest...	Log	Table changes from last run
□	DemoCrawler-Redshift	✓ Ready		✓ Succeeded	October 6, 2023 a...	View log	-
□	DemoCrawler-s3	✓ Ready		✓ Succeeded	October 11, 2023 ...	View log	1 updated
□	DemoCrawler-s3-copy	✓ Ready		✓ Succeeded	October 11, 2023 ...	View log	-
□	ppa-crawler-s3-raw-meta	✓ Ready		-	-	-	-
□	ppa-crawler-s3-raw-review	Running		-	-	-	-

Regla de Eventbridge

Para nuevos archivos raw de reviews
Dirigirse al servicio de EventBridge



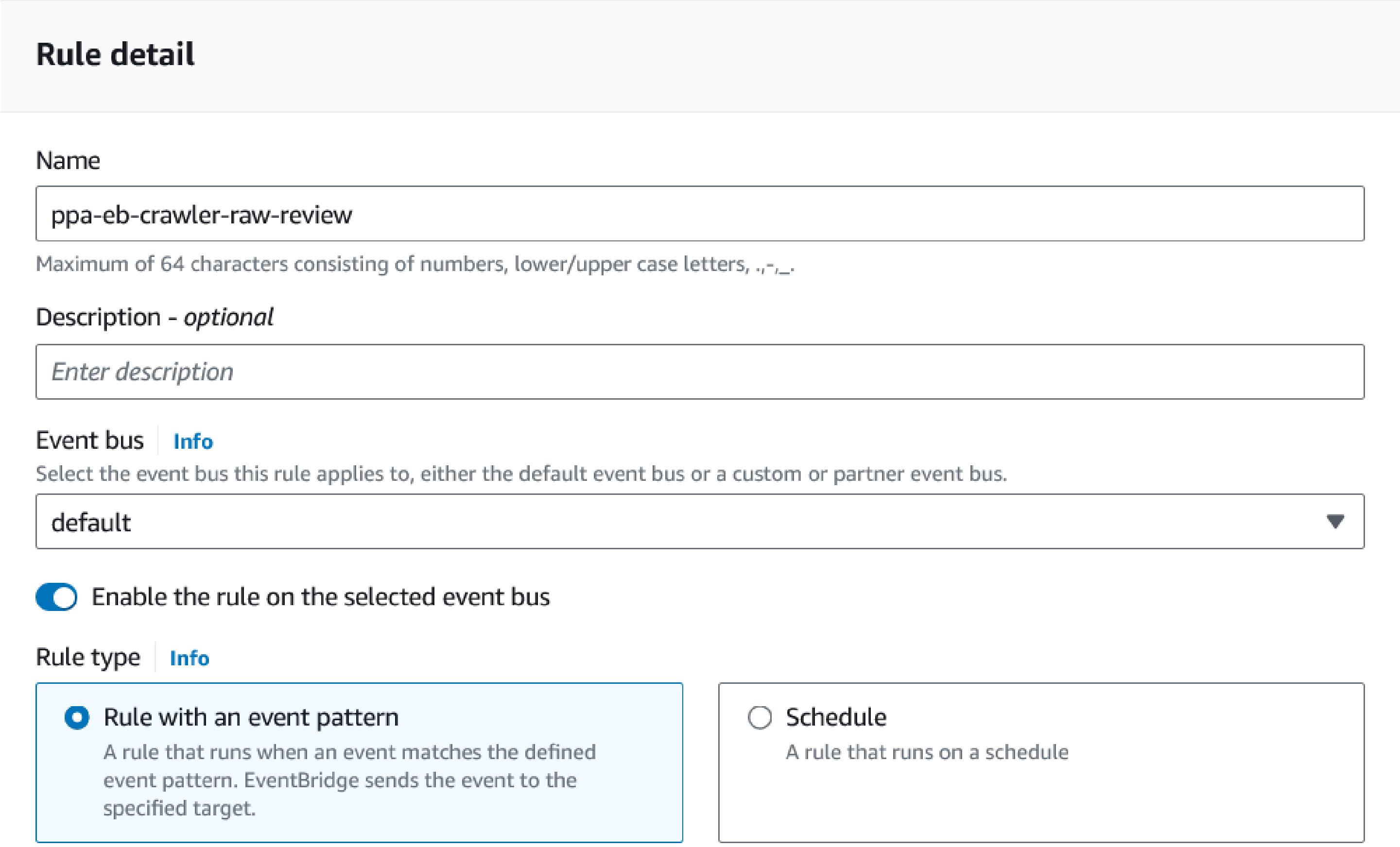
eventbridge

Search results for 'eventb'

Try searching with longer queries for more relevant results

Amazon EventBridge: Serverless service for building event-driven applications.

Crear nueva regla



Rule detail

Name

ppa-eb-crawler-raw-review

Maximum of 64 characters consisting of numbers, lower/upper case letters, .,-,_.

Description - *optional*

Enter description

Event bus Info

Select the event bus this rule applies to, either the default event bus or a custom or partner event bus.

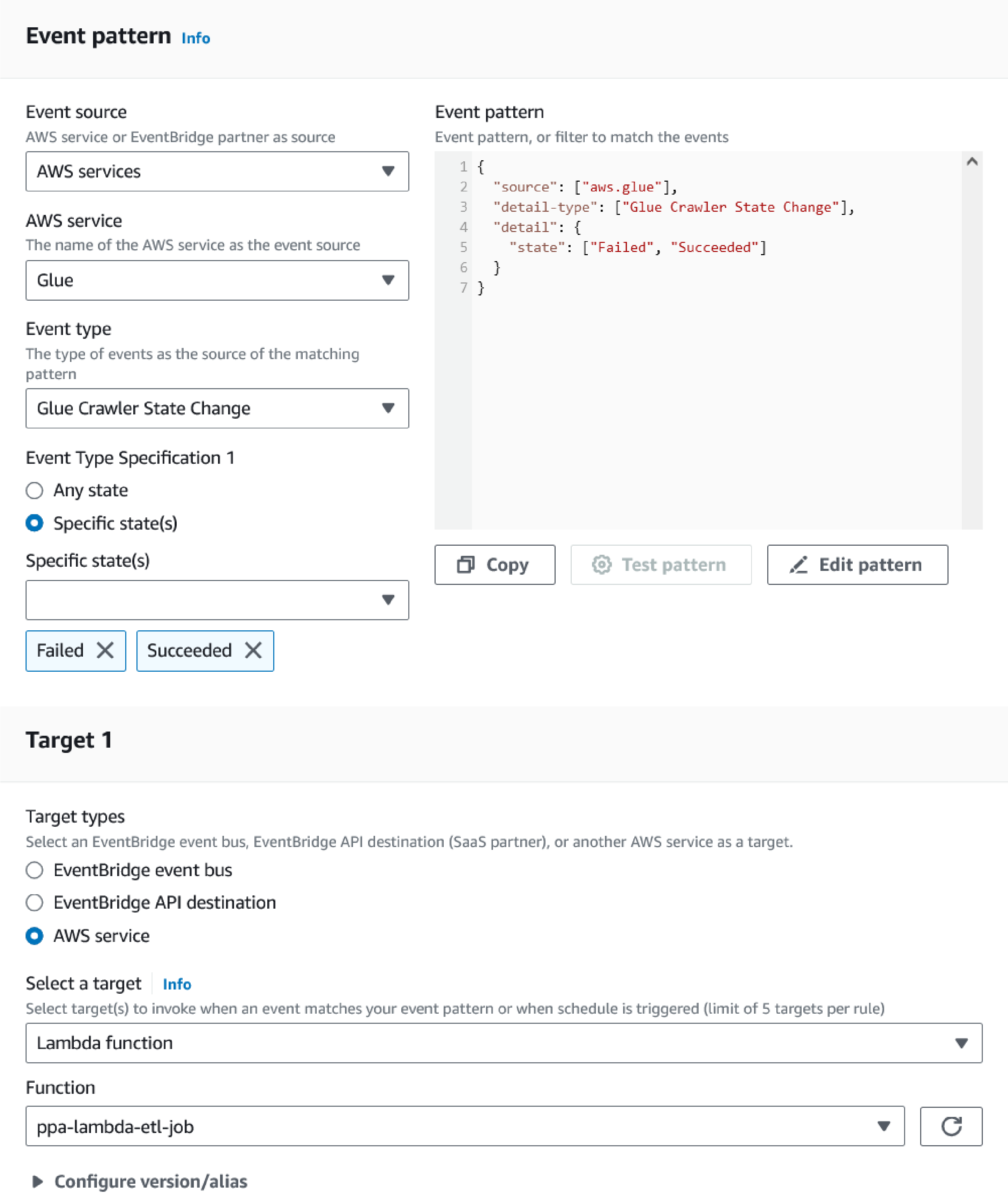
default

Enable the rule on the selected event bus

Rule type Info

● Rule with an event pattern
A rule that runs when an event matches the defined event pattern. EventBridge sends the event to the specified target.

○ Schedule
A rule that runs on a schedule



Event pattern

Event source

AWS service or EventBridge partner as source

AWS services

AWS service

The name of the AWS service as the event source

Glue

Event type

The type of events as the source of the matching pattern

Glue Crawler State Change

Event Type Specification 1

○ Any state

● Specific state(s)

Specific state(s)

Failed ✕ Succeeded ✕

Event pattern

Event pattern, or filter to match the events

```
{
  "source": ["aws.glue"],
  "detail-type": ["Glue Crawler State Change"],
  "detail": {
    "state": ["Failed", "Succeeded"]
  }
}
```

Target 1

Target types

Select an EventBridge event bus, EventBridge API destination (SaaS partner), or another AWS service as a target.

○ EventBridge event bus

○ EventBridge API destination

● AWS service

Select a target

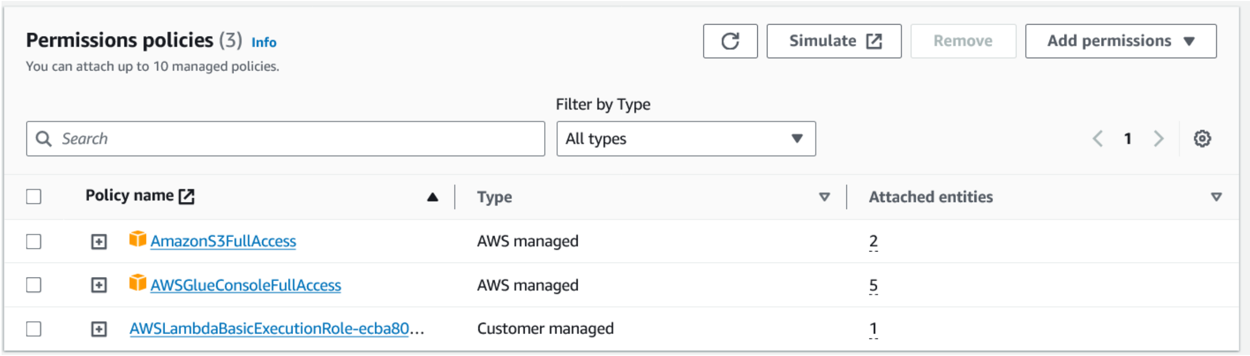
Select target(s) to invoke when an event matches your event pattern or when schedule is triggered (limit of 5 targets per rule)

Lambda function

Function

ppa-lambda-etl-job

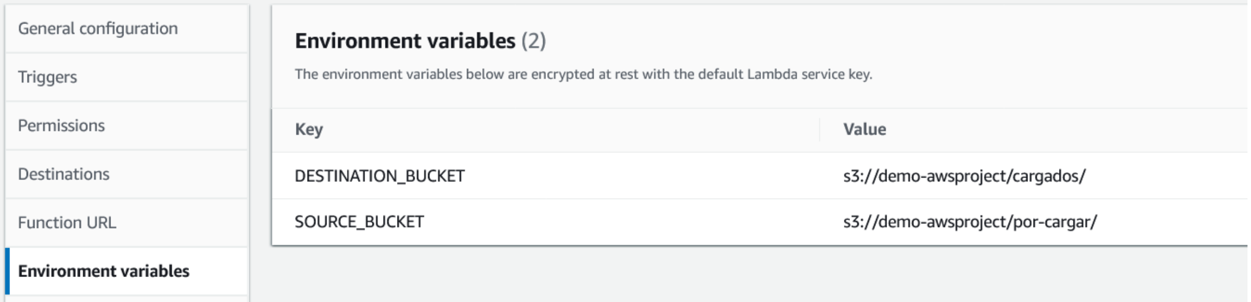
Crear función lambda para inicio de proceso ETL



Permissions policies (3)

You can attach up to 10 managed policies.

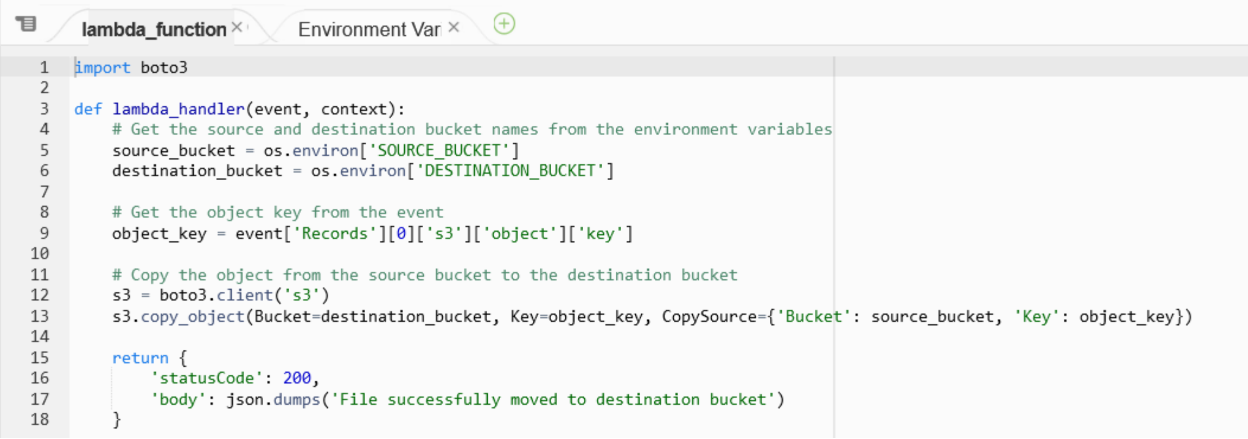
□	Policy name	Type	Attached entities
□	AmazonS3FullAccess	AWS managed	2
□	AWSGlueConsoleFullAccess	AWS managed	5
□	AWSLambdaBasicExecutionRole-ecba80...	Customer managed	1



Environment variables (2)

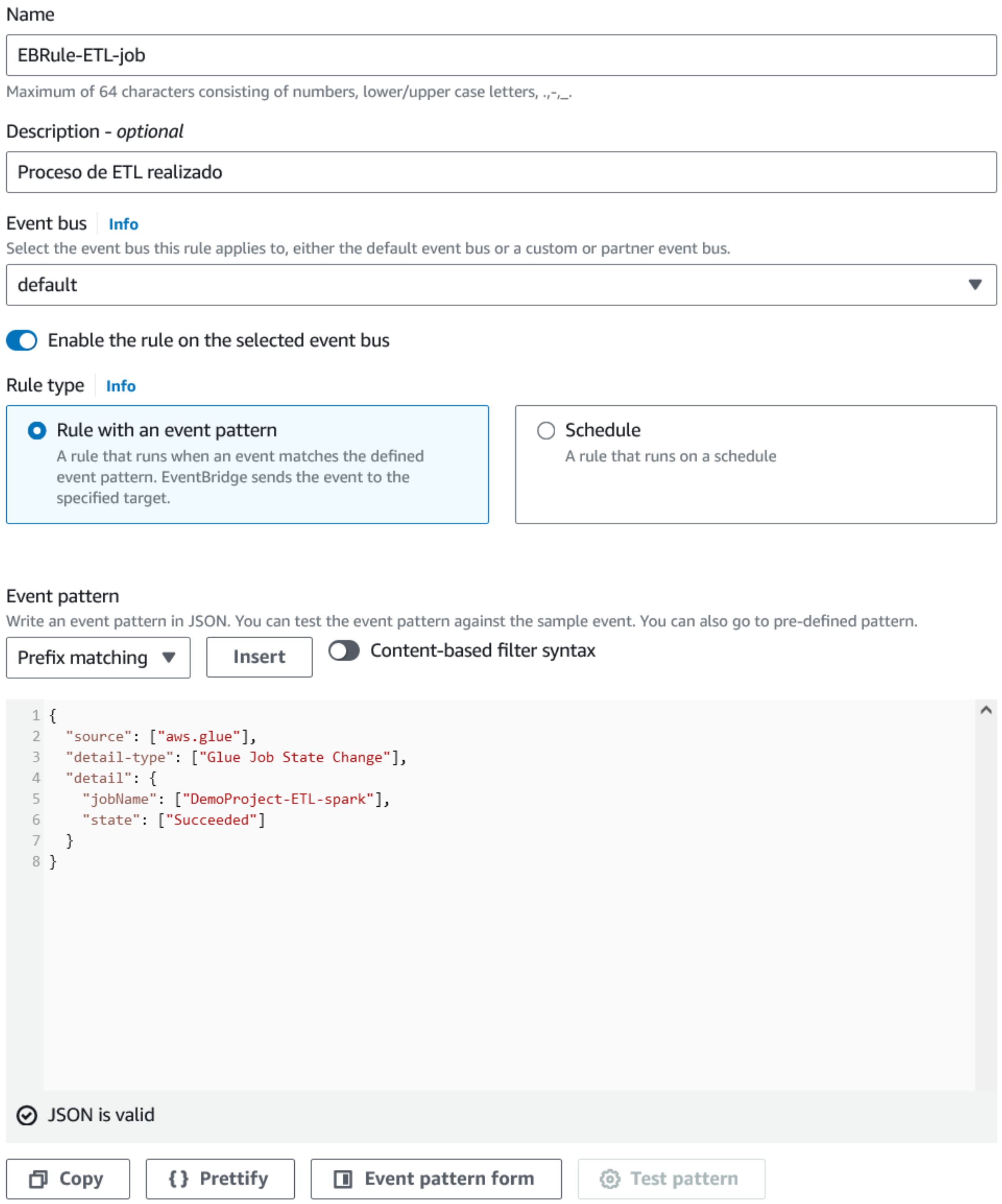
The environment variables below are encrypted at rest with the default Lambda service key.

Key	Value
DESTINATION_BUCKET	s3://demo-awsproject/cargados/
SOURCE_BUCKET	s3://demo-awsproject/por-cargar/



```python
import boto3

def lambda_handler(event, context):
    # Get the source and destination bucket names from the environment variables
    source_bucket = os.environ['SOURCE_BUCKET']
    destination_bucket = os.environ['DESTINATION_BUCKET']

    # Get the object key from the event
    object_key = event['Records'][0]['s3']['object']['key']

    # Copy the object from the source bucket to the destination bucket
    s3 = boto3.client('s3')
    s3.copy_object(Bucket=destination_bucket, Key=object_key, CopySource={'Bucket': source_bucket, 'Key': object_key})

    return {
        'statusCode': 200,
        'body': json.dumps('File successfully moved to destination bucket')
    }
```

Crear regla en EventBridge de proceso ETL exitoso



Name

EBRule-ETL-job

Maximum of 64 characters consisting of numbers, lower/upper case letters, .,-,_.

Description - *optional*

Proceso de ETL realizado

Event bus | Info

Select the event bus this rule applies to, either the default event bus or a custom or partner event bus.

default ▼

Enable the rule on the selected event bus

Rule type | Info

● Rule with an event pattern
A rule that runs when an event matches the defined event pattern. EventBridge sends the event to the specified target.

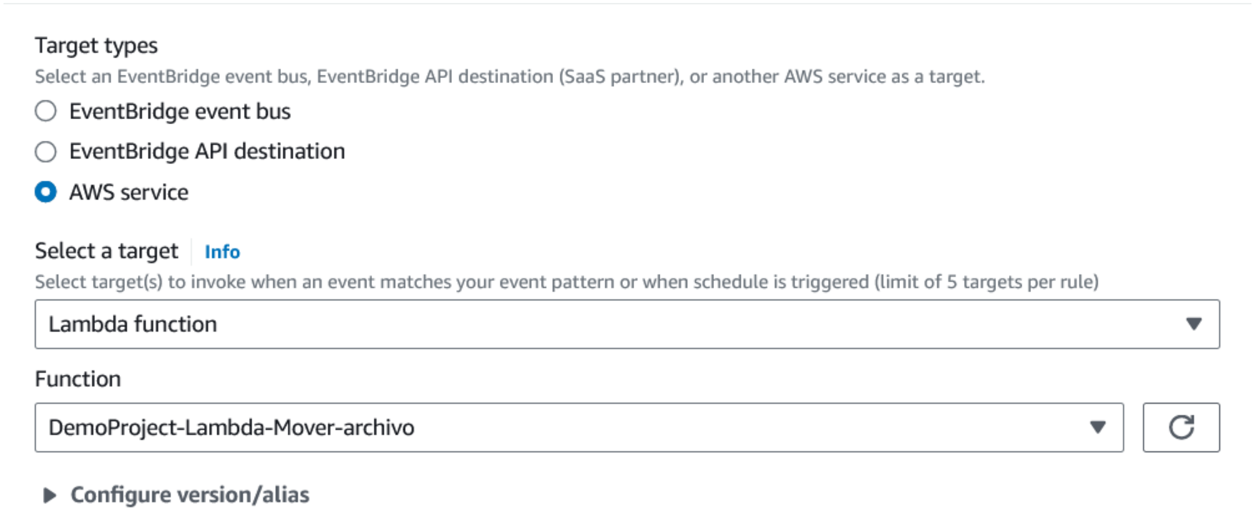
○ Schedule
A rule that runs on a schedule

Event pattern

Write an event pattern in JSON. You can test the event pattern against the sample event. You can also go to pre-defined pattern.

```
{
  "source": ["aws.glue"],
  "detail-type": ["Glue Job State Change"],
  "detail": {
    "jobName": ["DemoProject-ETL-spark"],
    "state": ["Succeeded"]
  }
}
```

JSON is valid



Target types

Select an EventBridge event bus, EventBridge API destination (SaaS partner), or another AWS service as a target.

○ EventBridge event bus

○ EventBridge API destination

● AWS service

Select a target Info

Select target(s) to invoke when an event matches your event pattern or when schedule is triggered (limit of 5 targets per rule)

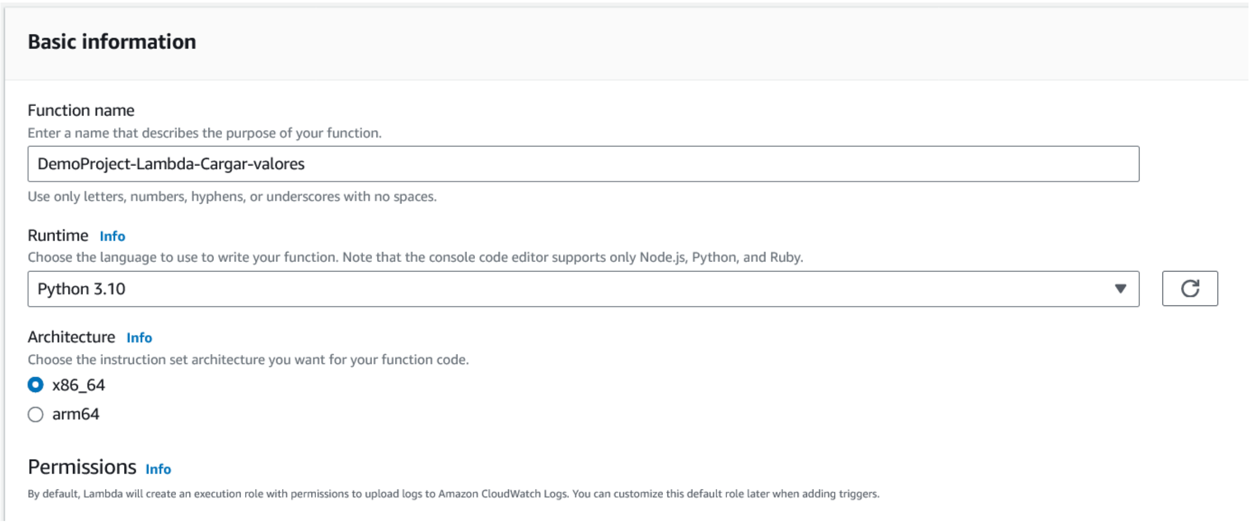
Lambda function

Function

DemoProject-Lambda-Mover-archivo

▶ Configure version/alias

Crear función Lambda para cargar valores a redshift



Basic information

Function name

Enter a name that describes the purpose of your function.

DemoProject-Lambda-Cargar-valores

Use only letters, numbers, hyphens, or underscores with no spaces.

Runtime Info

Choose the language to use to write your function. Note that the console code editor supports only Node.js, Python, and Ruby.

Python 3.10

Architecture Info

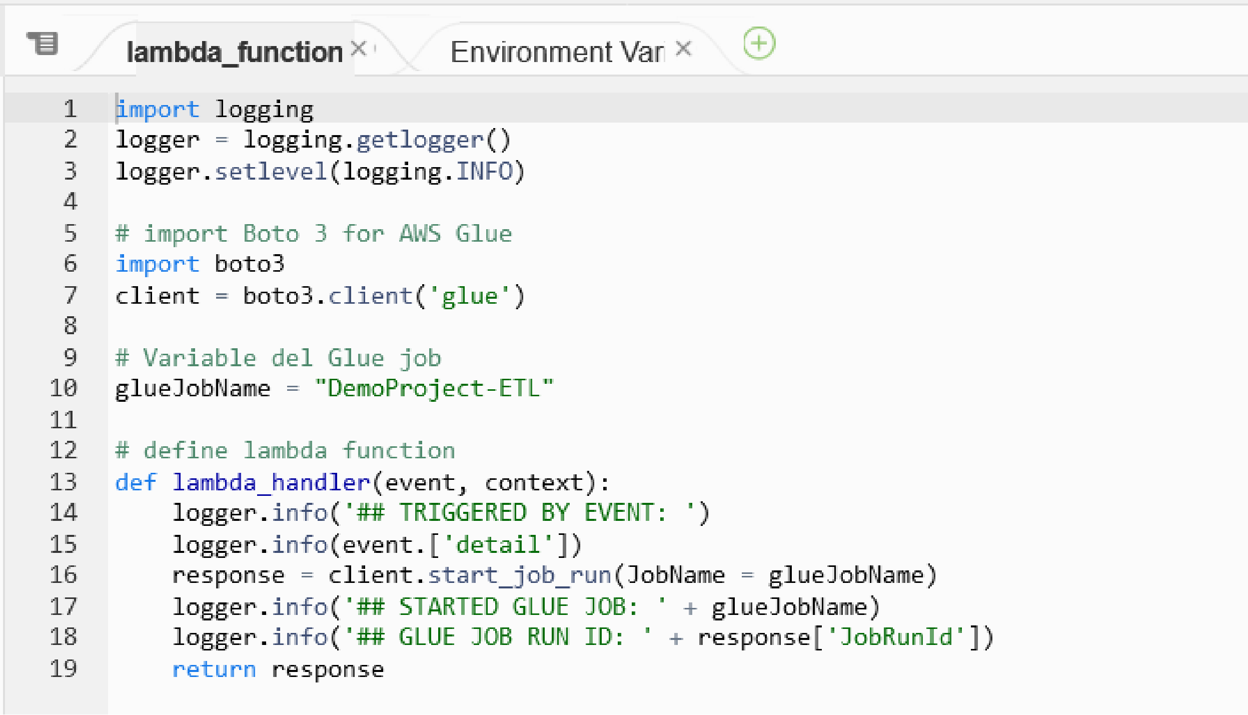
Choose the instruction set architecture you want for your function code.

● x86_64

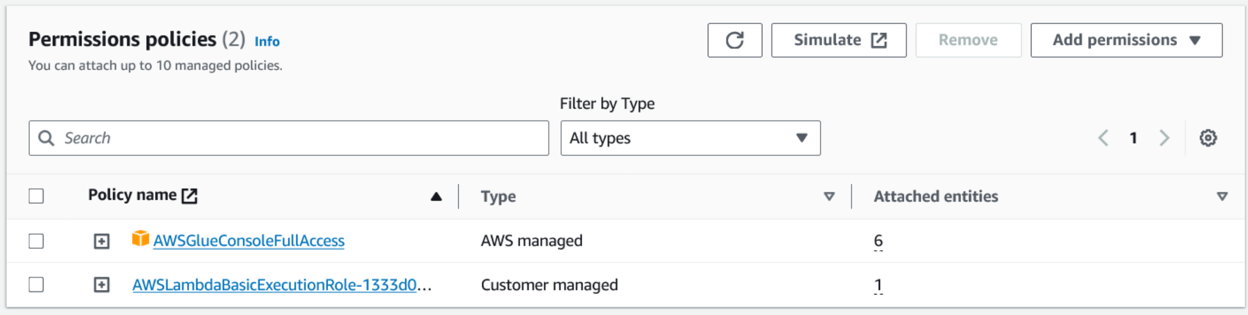
○ arm64

Permissions Info

By default, Lambda will create an execution role with permissions to upload logs to Amazon CloudWatch Logs. You can customize this default role later when adding triggers.



```
import logging
logger = logging.getlogger()
logger.setlevel(logging.INFO)

# import Boto 3 for AWS Glue
import boto3
client = boto3.client('glue')

# Variable del Glue job
glueJobName = "DemoProject-ETL"

# define lambda function
def lambda_handler(event, context):
    logger.info('## TRIGGERED BY EVENT: ')
    logger.info(event.['detail'])
    response = client.start_job_run(JobName = glueJobName)
    logger.info('## STARTED GLUE JOB: ' + glueJobName)
    logger.info('## GLUE JOB RUN ID: ' + response['JobRunId'])
    return response
```



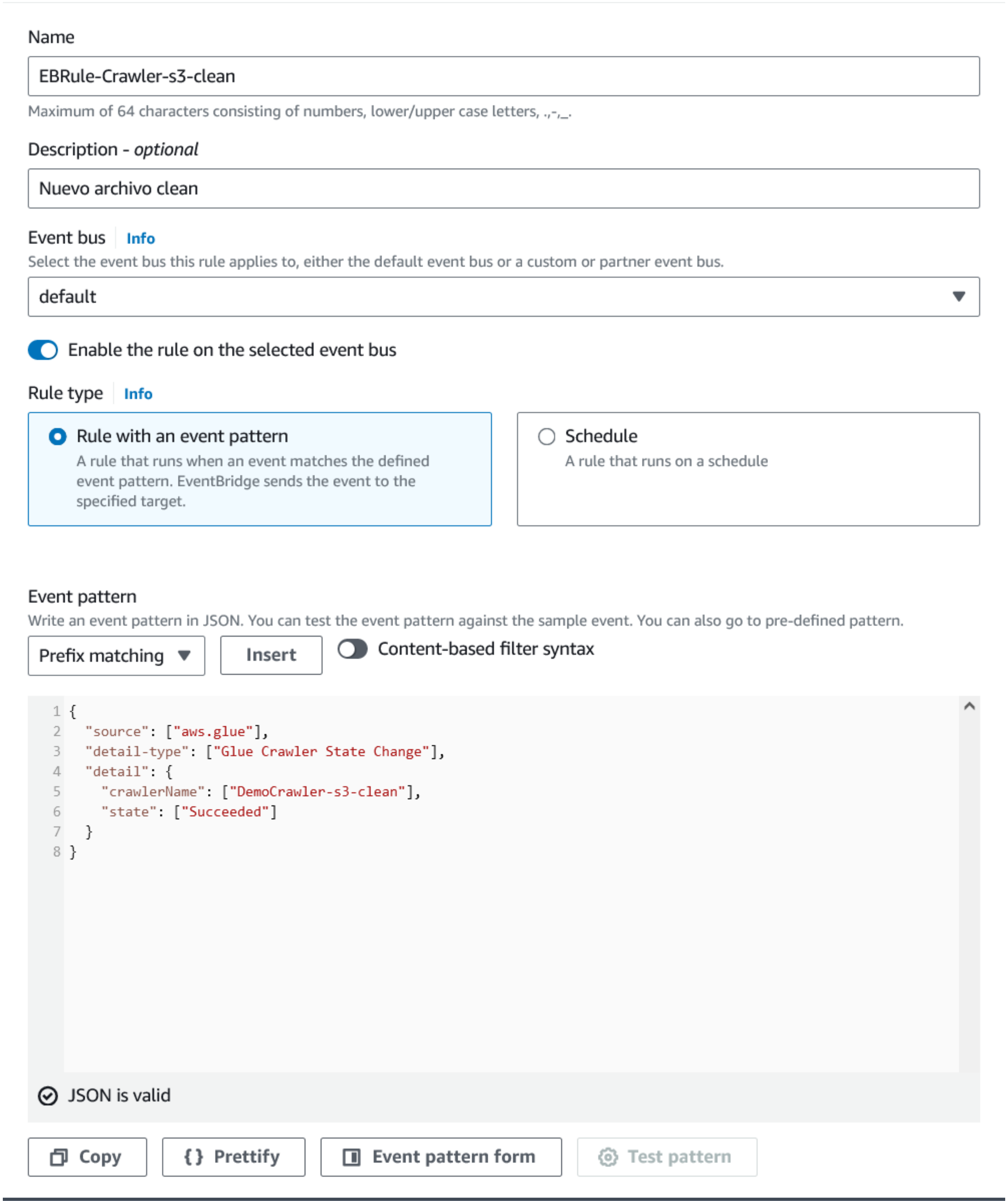
Permissions policies (2) Info

You can attach up to 10 managed policies.

Filter by Type: All types

□	Policy name	Type	Attached entities
□	AWSGlueConsoleFullAccess	AWS managed	6
□	AWSLambdaBasicExecutionRole-1333d0...	Customer managed	1

Crear EventBridge que detecte nuevo archivo clean



Name

EBRule-Crawler-s3-clean

Maximum of 64 characters consisting of numbers, lower/upper case letters, .,-,_.

Description - *optional*

Nuevo archivo clean

Event bus | Info

Select the event bus this rule applies to, either the default event bus or a custom or partner event bus.

default ▼

Enable the rule on the selected event bus

Rule type | Info

● Rule with an event pattern
A rule that runs when an event matches the defined event pattern. EventBridge sends the event to the specified target.

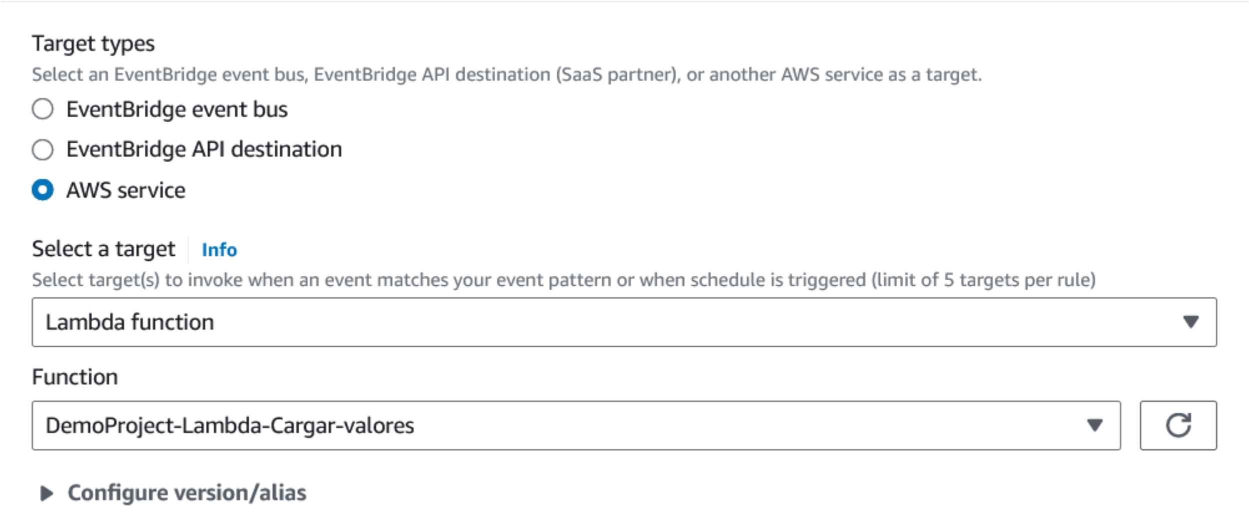
○ Schedule
A rule that runs on a schedule

Event pattern

Write an event pattern in JSON. You can test the event pattern against the sample event. You can also go to pre-defined pattern.

```
{
  "source": ["aws.glue"],
  "detail-type": ["Glue Crawler State Change"],
  "detail": {
    "crawlerName": ["DemoCrawler-s3-clean"],
    "state": ["Succeeded"]
  }
}
```

JSON is valid



Target types

Select an EventBridge event bus, EventBridge API destination (SaaS partner), or another AWS service as a target.

○ EventBridge event bus

○ EventBridge API destination

● AWS service

Select a target | Info

Select target(s) to invoke when an event matches your event pattern or when schedule is triggered (limit of 5 targets per rule)

Lambda function ▼

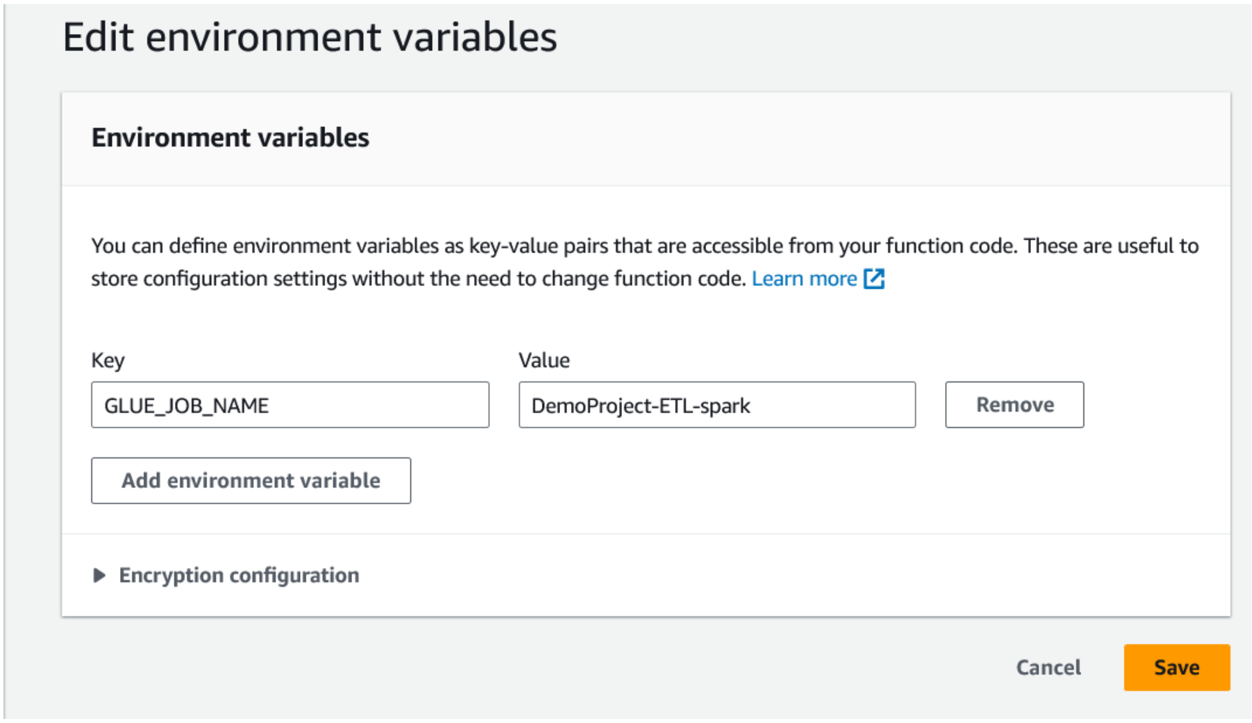
Function

DemoProject-Lambda-Cargar-valores ▼

▶ Configure version/alias

—---

Trigger Iniciar ETL job

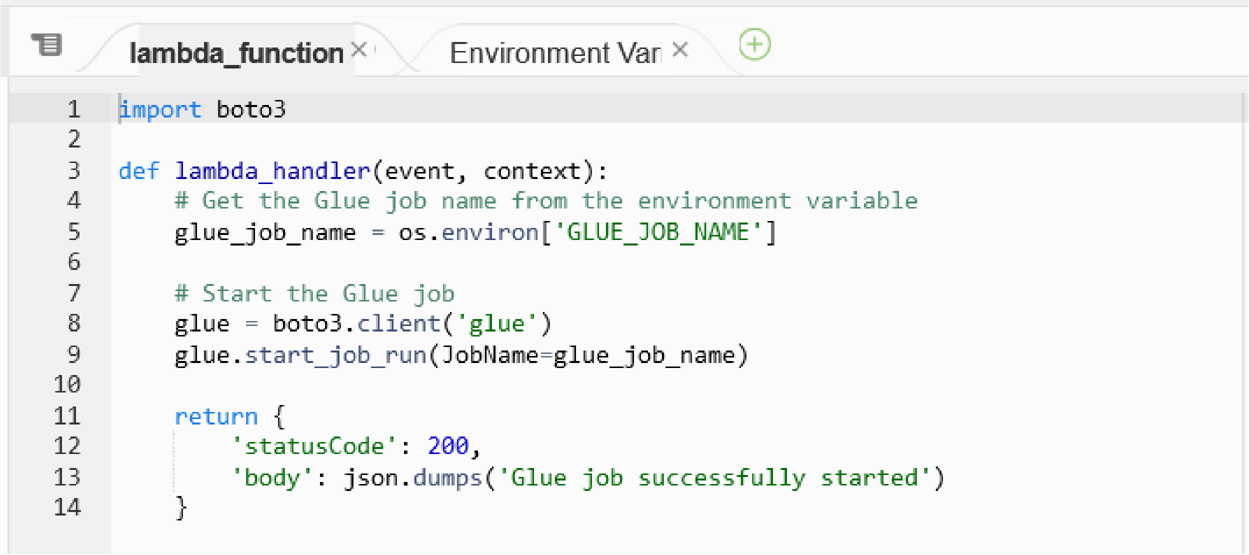


Edit environment variables

Environment variables

You can define environment variables as key-value pairs that are accessible from your function code. These are useful to store configuration settings without the need to change function code. Learn more

Key	Value
GLUE_JOB_NAME	DemoProject-ETL-spark



```python
import boto3

def lambda_handler(event, context):
    # Get the Glue job name from the environment variable
    glue_job_name = os.environ['GLUE_JOB_NAME']

    # Start the Glue job
    glue = boto3.client('glue')
    glue.start_job_run(JobName=glue_job_name)

    return {
        'statusCode': 200,
        'body': json.dumps('Glue job successfully started')
    }
```