

Parcial II

Desarrollar los ETL y un Workflow en AWS Glue para la descarga de información de periódicos (eltiempo y el espectador):

a) Crear un Job en AWS Glue(con un trigger) que descargue cada día la página principal de el Tiempo y El Espectador(o publímetro).

La información debe quedar en S3 con la estructura:

- s3://bucket/headlines/raw/contenido-yyyy-mm-dd.html

b) Una vez llega el archivo a la carpeta raw, se debe activar un segundo job que procese los datos que llegaron utilizando BeautifulSoup. Este proceso debe extraer la categoría, el titular y el enlace para cada noticia. Estos datos se deben guardar en un csv en la siguiente ruta:

- s3://bucket/headlines/final/periodico=xxx/year=xxx/month=xxx/day=xxx

Para usar paquetes externos revisar:

<https://aws.amazon.com/es/premiumsupport/knowledge-center/glue-version2-external-python-libraries/>

c) Una vez terminados estos jobs, se debe activar un crawler que actualice el catálogo de AWS Glue y permita visualizar los datos en AWS Athena.

d) Crear un Job que inserte la información en una base de datos MYSQL(usando aws glue connectors y aws job). Para esto se debe crear la BD de MYSQL en RDS con la respectiva tabla. Luego se debe mapear con un crawler al catálogo del glue. Finalmente crear el job con la interfaz que copie de tabla a tabla(la que representa s3 y la que representa RDS en el catálogo).

- Activar la opción "job bookmarks" cuando se cree el job por interfaz. Esto permite que glue lleve una trazabilidad de los datos insertados y evita que se vuelvan a insertar datos ya insertados.

Nota:

** Se debe entregar el código en github(utilizar ramas, commits, código limpio, código comentado, pruebas unitarias si aplica).

*** Crear un pipeline de despliegue continuo en github para los scripts de los jobs

Amazon S3 > Buckets > periodicospc2

periodicospc2

Info

Objetos

Propiedades

Permisos

Métricas

Administración

Puntos de acceso

Objetos (1)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Copiar URI de S3

Copiar URL

Descargar

Abrir

Eliminar

Acciones

Crear carpeta

Cargar

Q

Buscar objetos por prefijo

< 1 >

	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	headlines/	Carpeta	-	-	-

periodicos job

Last modified on 24/4/2023, 18:01:07

Actions

Save

Run

Script

Job details

Runs

Data quality New

Schedules

Version Control

Script

Info

1

import sys

2

import boto3

3

import urllib.request

4

import datetime

5

6

def download_page(url, bucket,nombre):

7

response = urllib.request.urlopen(url)

8

content = response.read().decode('utf-8')

9

today = datetime.date.today().strftime('%Y-%m-%d')

10

s3 = boto3.resource('s3')

11

s3.Object(bucket, 'headlines/raw/{}-{}.html'.format(nombre,today)).put(Body=content)

12

13

download_page('https://www.eltiempo.com', 'periodicospc2','El_Tiempo')

14

download_page('https://www.elspectador.com', 'periodicospc2','El_Espectador')

Python

Ln 1, Col 1

Errors: 0

Warnings: 0

periodicos job

Last modified on 24/4/2023, 18:01:07

Actions

Save

Run

Script

Job details

Runs

Data quality New

Schedules

Version Control

Job runs (1/1) Info

Last updated (UTC)
April 24, 2023 at 23:03:32



View details

Stop job run

Table View

Card View

Filter job runs by property

< 1 > ⚙

Run status	Retry	Start time	End time	Duration	Capacity
🟢 Succeeded	0	04/24/2023 18:01:12	04/24/2023 18:02:56	16 s	0.0625 DPU's

Amazon S3 > Buckets > periodicoscpc2 > headlines/ > raw/ > EL_Tiempo-2023-04-24.html

EL_Tiempo-2023-04-24.html Info

Copiar URI de S3

Descargar

Abrir

Acciones de objetos

Propiedades

Permisos

Versiones

Información general sobre el objeto

Propietario

awslabsc0w5046548t1671494192

Región de AWS

EE. UU. Este (Norte de Virginia) us-east-1

Última modificación

24 Apr 2023 6:02:46 PM -05

Tamaño

472.7 KB

Tipo

html

Clave

headlines/raw/EL_Tiempo-2023-04-24.html

URI DE S3

s3://periodicospc2/headlines/raw/EL_Tiempo-2023-04-24.html

Nombre de recurso de Amazon (ARN)

arn:aws:s3::periodicospc2/headlines/raw/EL_Tiempo-2023-04-24.html

Etiqueta de entidad (Etag)

d74ad8a7df33b005dc156db464033848

URL del objeto

https://periodicospc2.s3.amazonaws.com/headlines/raw/EL_Tiempo-2023-04-24.html

```
File Edit View Navigate Code Refactor Build Run Tools Git Window Help learn2grow (CORRECCIONES) C:\Users\malej\Downloads\EL_Tiempo-2023-04-24.html
C:\Users\malej\Downloads\EL_Tiempo-2023-04-24.html Learn2growApplication
Project
  learn2grow (CORRECCIONES)
    idea
    learn2grow
      src
        main
          java
            com.example.learn2grow
              controller
              entity
              repository
              service
              Learn2growApplication
            resources
            test
              java
                com.example.learn2grow
                  Learn2growApplicationTests
          target
            gitignore
            HELP.md
            mvnw
            mvnw.cmd
            pom.xml
      External Libraries
      Scratches and Consoles
Download grammar and spelling checker for Spanish? Download Spanish Never suggest Spanish
This document contains very long lines. Soft wraps were enabled to improve editor performance.
1 <!DOCTYPE html>
2 <html lang="es">
3 <head>
4 <script async type="text/javascript">var _sf_startpt=(new Date()).getTime()/</script>
5 <script type="text/javascript">
6 window.NREUM||(NREUM={});__nr_require=function(t,e,n){function r(n){if(!e[n]){var
o=e[n]={exports:{}};t[n][0].call(o.exports,function(e){var o=t[n][1][e];return r(o||e)},o,
o.exports)}return e[n].exports}if("function"==typeof __nr_require)return __nr_require;for(var
o=0;o<n.length;o++)r(n[o]);return r}({1:[function(t,e,n){function r(t){try{s.console&&console.log
(t)}catch(e){}var o,i=t("ee"),a=t(23),s={};try{o=localStorage.getItem("__nr_flags").split(",")
console&&"function"==typeof console.log&&(s.console=!0,o.indexOf("dev")!==-1&&(s.dev=!0))o
.indexOf("nr_dev")!==-1&&(s.nrDev=!0)}catch(c){s.nrDev&&i.on("internal-error",function(t){r(t
.stack)}),s.dev&&i.on("fn-err",function(t,e,n){r(n.stack)}),s.dev&&(r("NR AGENT IN DEVELOPMENT
MODE"))r("flags: "+a(s,function(t,e){return t}).join(", "));},2:[function(t,e,n){function
r(t,e,n,r,s){try{l?l=1:o(s||new UncaughtException(t,e,n),!0)}catch(f){try{i("ierr",[f,c.now(),
!0])}catch(d){}return"function"==typeof u&&u.apply(this,a(arguments))}function UncaughtException
(t,e,n){this.message=t||"Uncaught error with no additional information",this.sourceURL=e,this
.line=n}function o(t,e){var n=e?null:c.now();i("err",[t,n])var i=t("handle"),a=t(24),s=t("ee"),
c=t("loader"),f=t("gos"),u=window.onerror,d=!1,p="nr@seenError",l=0,c.features.err=!0,t(1),window
.onerror=r;try{throw new Error}catch(h){"stack"in h&&(t(13),t(12),"addEventListener" in window&&t
(6),c.xhrWrappable&&t(14),d=!0)}s.on("fn-start",function(t,e,n){d&&(l+=1)},s.on("fn-err",
```

Se ha realizado la carga correctamente. Consulte los detalles a continuación.

Amazon S3 > Buckets > periodicospc2

periodicospc2

Objetos Propiedades Permisos Métricas Administración Puntos de acceso

Objetos (2)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

[Actualizar](#) [Copiar URI de S3](#) [Copiar URL](#) [Descargar](#) [Abrir](#) [Eliminar](#) [Acciones](#) [Crear carpeta](#) [Cargar](#)

Buscar objetos por prefijo

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	beautifulsoup4-4.12.2-py3-none-any.whl	whl	24 Apr 2023 6:13:12 PM -05	139.6 KB	Estándar
<input type="checkbox"/>	headlines/	Carpeta	-	-	-

Amazon S3 > Buckets > periodicospc2 > beautifulsoup4-4.12.2-py3-none-any.whl

beautifulsoup4-4.12.2-py3-none-any.whl

[Copiar URI de S3](#) [Descargar](#) [Abrir](#) [Acciones de objetos](#)

Propiedades Permisos Versiones

Información general sobre el objeto

Propietario aws-labs-c0w5046548t1671494192	URI DE S3 s3://periodicospc2/beautifulsoup4-4.12.2-py3-none-any.whl
Región de AWS EE. UU. Este (Norte de Virginia) us-east-1	Nombre de recurso de Amazon (ARN) arn:aws:s3:::periodicospc2/beautifulsoup4-4.12.2-py3-none-any.whl
Última modificación 24 Apr 2023 6:13:12 PM -05	Etiqueta de entidad (Etag) 25b3ffb957a8b2b0c5762d6ec9498a5b
Tamaño 139.6 KB	URL del objeto https://periodicospc2.s3.amazonaws.com/beautifulsoup4-4.12.2-py3-none-any.whl
Tipo whl	
Clave	

periodicosbeutifulsoup job

Last modified on 24/4/2023, 18:27:45 [Actions](#) [Save](#) [Run](#)

Successfully updated job
Successfully updated job periodicosbeutifulsoup job. To run the job choose the Run Job button.

Script Job details Runs Data quality New Schedules Version Control

Script

```
24 try:
25     link="" + links["href"] + ""
26     #print(link)
27     categoria="" + link.split('/')[1] + ""
28     #print(categoria)
29     titulo="" + (links.text) + ""
30     headers+=f'{titulo},{categoria},{link}\n'
31 except:
32     ...
33
34
35 url="headlines/final/periodico=eltiempo/year="+str(date_day.year)+"/month="+str(date_day.strftime('%m'))+"/day="+str(date_day.strftime('%d'))+"/elTiempo"
```

Python Ln 1, Col 1 0 Errors: 0 0 Warnings: 0

aws

Servicios

Buscar

[Alt+S]

Norte de Virginia

voclabs/user2374338=maria.rodriguez11.usa@gmail.com @ 3021-97...

periodicosbeutifulsoup job

Last modified on 24/4/2023, 18:30:21

Actions

Save

Run

Successfully started job

Successfully started job periodicosbeutifulsoup job. Navigate to [Run details](#) for more details.

Script

Job details

Runs

Data quality New

Schedules

Version Control

Job runs (1/2)

Info

Last updated (UTC)
April 24, 2023 at 23:30:55

View details

Stop job run

Table View

Card View

Filter job runs by property

	Run status	Retry	Start time	End time	Duration	Capacity	Worker type	Glue
<input type="radio"/>	✔ Succeeded	0	04/24/2023 18:30:25	04/24/2023 18:30:48	18 s	0.0625 DPU	-	1.0

☐

headlines/

Carpeta

-

-

-

☐

final/

Carpeta

-

-

-

Amazon S3

>

Buckets

>

periodicospc2

>

headlines/

>

final/

final/

Copiar URI de S3

Objetos

Propiedades

Objetos (2)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Copiar URI de S3

Copiar URL

Descargar

Abrir

Eliminar

Acciones

Crear carpeta

Cargar

Buscar objetos por prefijo

	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	periodico=espectador/	Carpeta	-	-	-
<input type="checkbox"/>	periodico=eltiempo/	Carpeta	-	-	-

Amazon S3 > Buckets > periodicospc2 > headlines/ > final/ > periodico=elespectador/

periodico=elespectador/

Copiar URI de S3

ObjetosPropiedades

Objetos (1)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

🔄

Copiar URI de S3

Copiar URL

⬇️ Descargar

🔗 Abrir

🗑️ Eliminar

Acciones ▼

Crear carpeta

Cargar

🔍

Buscar objetos por prefijo

< 1 > ⚙️

<input type="checkbox"/>	Nombre ▲	Tipo ▼	Última modificación ▼	Tamaño ▼	Clase de almacenamiento ▼
<input type="checkbox"/>	📁 year=2023/	Carpeta	-	-	-

☐

📁 [month=04/](#)

Carpeta

-

-

-

☐

📁 [day=24/](#)

Carpeta

-

-

-

☐

📄 [elEspectador2023-04-24.csv](#)

csv

24 Apr 2023 6:30:38 PM -05

2.6 KB

Estándar

Acciones de objetos ▼

Info

 Descargar

Acciones de objetos ▲

Compartir con una URL prefirmada

Copiar

Iniciar restauración

Editar acciones

Editar clase de almacenamiento

Editar cifrado del lado del servidor

Editar metadatos

Editar etiquetas

Hacer público mediante ACL

URL del objeto

☐ **Excluir la primera línea de CSV datos**
Habilite esta configuración si CSV contiene una fila de encabezado.

Compresión

☒ Ninguno

☐ GZIP

☐ BZIP2

Configuración de salida

Formato

☐ JSON

CSV delimitador

☐ Tabulad

Consulta SQL

Amazon S3 Select solo admite el comando `SELECT SQL`. Con la consola de S3, puede extraer hasta 40 MB de registros de un objeto con un tamaño de hasta 128 MB. Para trabajar con archivos más grandes o más registros, utilice la CLI de AWS, el SDK de AWS o la API REST de Amazon S3. Para consultas SQL más complejas, utilice [Amazon Athena](#).

Agregar SQL desde plantillas

Ejecutar consulta SQL

```
1 /* Para crear un punto de referencia para escribir consultas SQL, puede mostrar los primeros 5 registros de datos de entrada ejecutando la siguiente consulta SQL: SELECT * FROM s3object s LIMIT 5 */
2 SELECT * FROM s3object s LIMIT 5
```

Resultados de la consulta

Los resultados de las consultas no están disponibles después de elegir **Close** (Cerrar) o salir de la página. Elija **Download results** (Descargar resultados) para descargar una copia de los siguientes resultados de la consulta

 Descargar resultados

Estado

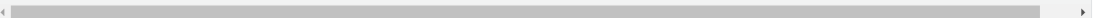
✔ Se han devuelto correctamente 5 registros en 566 ms

Bytes devueltos: 864 B

Sin procesar

Formateado

titulo,categoria,link



Resultados de la consulta

Los resultados de las consultas no están disponibles después de elegir **Cerrar** o salir de la página. Elija **Download results** (Descargar resultados) para descargar una copia de los siguientes resultados de la consulta.

Estado

Se han devuelto correctamente 5 registros en 566 ms

Bytes devueltos: 864 B

Sin procesar

Formateado

< 1 >

titulo	categoría	link
futbol colombiano las propuestas de la dimayor para frenar la violencia en el futbol colombiano	deportes	/deportes/futbol-colombiano/las-propuestas-de-la-dimayor-para-frenar-la-violencia-en-el-futbol-colombiano/
procuraduria solicita audiencia publica por presuntas irregularidades en relleno dona juana noticias hoy	bogota	/bogota/procuraduria-solicito-audiencia-publica-por-presuntas-irregularidades-en-relleno-dona-juana-noticias-hoy/
finanzas personales el dolar abrio la semana a la baja y quedo en 4468 este 24 de abril de 2023	economia	/economia/finanzas-personales/el-dolar-abrio-la-semana-a-la-baja-y-quedo-en-4468-este-24-de-abril-de-2023/
mocion de censura contra canciller alvaro leyva se cayo en la camara de representantes	politica	/politica/mocion-de-censura-contra-canciller-alvaro-leyva-se-cayo-en-la-camara-de-representantes/

AWS

Servicios

Buscar

[Alt+S]

1

the following 25 data source data source is now deleted: "s3://periodicospc2"

AWS Glue > Crawlers > Edit crawler

Step 1
Set crawler properties

Step 2
Choose data sources and classifiers

Step 3
Configure security settings

Step 4
Set output and scheduling

Step 5
Review and update

Review and update

Step 1: Set crawler properties

Set crawler properties

Name	Description	Tags
periodicos_crawler	-	-

Step 2: Choose data sources and classifiers

Data sources (2) Info

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

Type	Data source	Parameters
S3	s3://periodicospc2/headlin...	Recrawl all
S3	s3://periodicospc2/headlin...	Recrawl all

Step 3: Configure security settings

Configure security settings

IAM role	Security configuration	Lake Formation configuration
LabRole	-	-

Step 4: Set output and scheduling

Set output and scheduling

Database	Table prefix - optional	Maximum table threshold - optional	Schedule
periodicos	-	-	At 12:00 AM, only on Monday, Tuesday, Wednesday, Thursday, and Friday

Cancel

Previous

Update

se crea el crawler y se corre

Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (1)

The list of crawler runs for this crawler.

Filter data

Filter by a date and time range

< 1 > ⚙

	Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours
<input type="radio"/>	April 24, 2023 at 23:53:59	-	12 s	Running	

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (1)

The list of crawler runs for this crawler.

< 1 > ⚙

	Start time (UTC) ▲	End time (UTC) ▼	Current/last duration ▼	Status ▼	DPU hours ▼
○	April 24, 2023 at 23:53:59	April 24, 2023 at 23:55:00	01 min 01 s	✔ Completed	

Amazon Athena > Query editor

Editor

Recent queries

Saved queries

Settings

Workgroup primary ▼

Data

Data source

AwsDataCatalog ▼

Database

periodicos ▼

Tables and views

Create ▼ ⚙

▼ Tables (2) < 1 >

periodico_espectador

Partitioned

⋮

periodico_el tiempo

Partitioned

⋮

► Views (0) < 1 >

Query 19 ⋮

1

SQL Ln 1, Col 1

Run

Explain ⚡

Cancel

Clear

Create ▼

☐ Reuse query results
*Athena engine version 3 only

RDS > Databases

ⓘ

Considere la posibilidad de crear una implementación azul-verde para minimizar el tiempo de inactividad durante las actualizaciones.

Es posible que desee considerar el uso de las implementaciones azul-verde de Amazon RDS y minimizar el tiempo de inactividad durante las actualizaciones. Una implementación azul-verde proporciona un entorno de ensayo para los cambios en las bases de datos de producción. [Guía del usuario de RDS](#) [Guía del usuario de Aurora](#) ⚡

Bases de datos

Recursos del grupo

⌂

Modificar

Acciones ▼

Restaurar desde S3

Crear base de datos

< 1 > ⚙

Identificador de base de datos	Rol ▲	Motor ▼	Región y AZ ▼	Tamaño ▼	Estado ▼	Acción
database-1	Instancia	MySQL Community	us-east-1a	db.t3.micro	⌛ Detención temporal	2 Acc
database-2	Instancia	MySQL Community	us-east-1f	db.t3.micro	⌛ Detención temporal	2 Acc
database-3	Instancia	MySQL Community	us-east-1d	db.t3.micro	✔ Disponible	2 Acc

```
Immediate (Javascript (br x  ejercicio4/funciones.py - ξ x  ejercicio4/test_funciones. | x  mysql - "ip-172-31-20-73" x  +)
```

```
voclabs:~/environment $ mkdir periodicosPC2
voclabs:~/environment $ cd periodicosPC2
voclabs:~/environment/periodicosPC2 $ mysql -u admin -p -h database-3.cl6kuwtng9aj.us-east-1.rds.amazonaws.com
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 16
Server version: 8.0.32 Source distribution

Copyright (c) 2000, 2023, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> █
```

```
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> create database periodicos_p
-> ;
Query OK, 1 row affected (0.01 sec)

mysql> █
```

```
Type 'help;' or '\h' for help. Type '\c' to clear the current

mysql> create database periodicos_p
-> ;
Query OK, 1 row affected (0.01 sec)

mysql> use periodicos_p;
Database changed
mysql> █
```

```
Query OK, 1 row affected (0.01 sec)

mysql> use periodicos_p;
Database changed
mysql> create table El_Espectador (titulo varchar(255), categoria varchar(255), link varchar(255));
Query OK, 0 rows affected (0.02 sec)

mysql> create table El_Tiempo (titulo varchar(255), categoria varchar(255), link varchar(255));
Query OK, 0 rows affected (0.03 sec)

mysql> █
```

Punto de enlace de la VPC creado correctamente

vpce-05fb4733e0216df57

Puntos de enlace (1) [Información](#)

Acciones

Crear punto de enlace

Find resources by attribute or tag

ID de punto de enlace de la VPC : vpce-05fb4733e0216df57

Clear filters

<input type="checkbox"/>	Name	ID de punto de enlace de la ...	ID de la VPC	Nombre del servicio
<input type="checkbox"/>	endpoint2	vpce-05fb4733e0216df57	vpc-0a5e72d81ee26451f	com.amazonaws.us-east-1

AWS Glue

Connectors

Create connection

Create connection [Info](#)

Connection properties [Info](#)

Name

Enter a unique name for your connection.

periodicos_prueba

Connection type

JDBC

☐ Require SSL connection

The connection will fail if it's unable to connect over SSL.

Description - optional

Descriptions can be up to 2048 characters long.

Test Connection

Testing connection periodicos_prueba access to data store.

Cancel

Connectors [Info](#)

can manage your conn

Filter connections

Name

connections [Info](#)

can manage your conn

Actions

Create connection

Create job

Filter connections by property

Name	Type	Last modified
periodicos_prueba	JDBC	Apr 24, 2023
conexion1	JDBC	Apr 21, 2023

Test Connection

✔ Successfully connected to the data store with connection **periodicos_prueba**.

View log

Cancel

Create connectionCreate job

Add data source

Data source

Choose the source of data to be crawled.

JDBC

Connection

Select a connection to access the data sources below.

periodicos_prueba

Clear selectionAdd new connection

Include path

periodicos_p/%

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).

Additional metadata - optional

Select additional metadata properties for the crawler to crawl.

☐ Exclude tables matching pattern

Cancel

Add a JDBC data source

Set output and scheduling

Output configuration [Info](#)

Target database

periodicos_cloud9



Clear selection

Add database [↗](#)

Table name prefix - *optional*

Type a prefix added to table names

► Advanced options

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



Cancel

Previous

Next

Crawler properties

Name
periodico2_crawler

IAM role
[LabRole](#) [↗](#)

Database
periodicos_cloud9

State
READY

Description
-

Security configuration
-

Table prefix
-

► Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Crawler runs (1)

The list of crawler runs for this crawler.



Stop run

View CloudWatch logs [↗](#)

View run de

🔍 Filter data

📅 Filter by a date and time range

< 1 >

Start time (UTC)



End time (UTC)



Current/last duration



Status



DPU h



April 25, 2023 at 01:23:40

April 25, 2023 at 01:25:43

02 min 02 s

🟢 Completed

AWS Glue > Jobs

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.

Source

AWS Glue Data Catalog
AWS Glue Data Catalog table as the data source.

Target

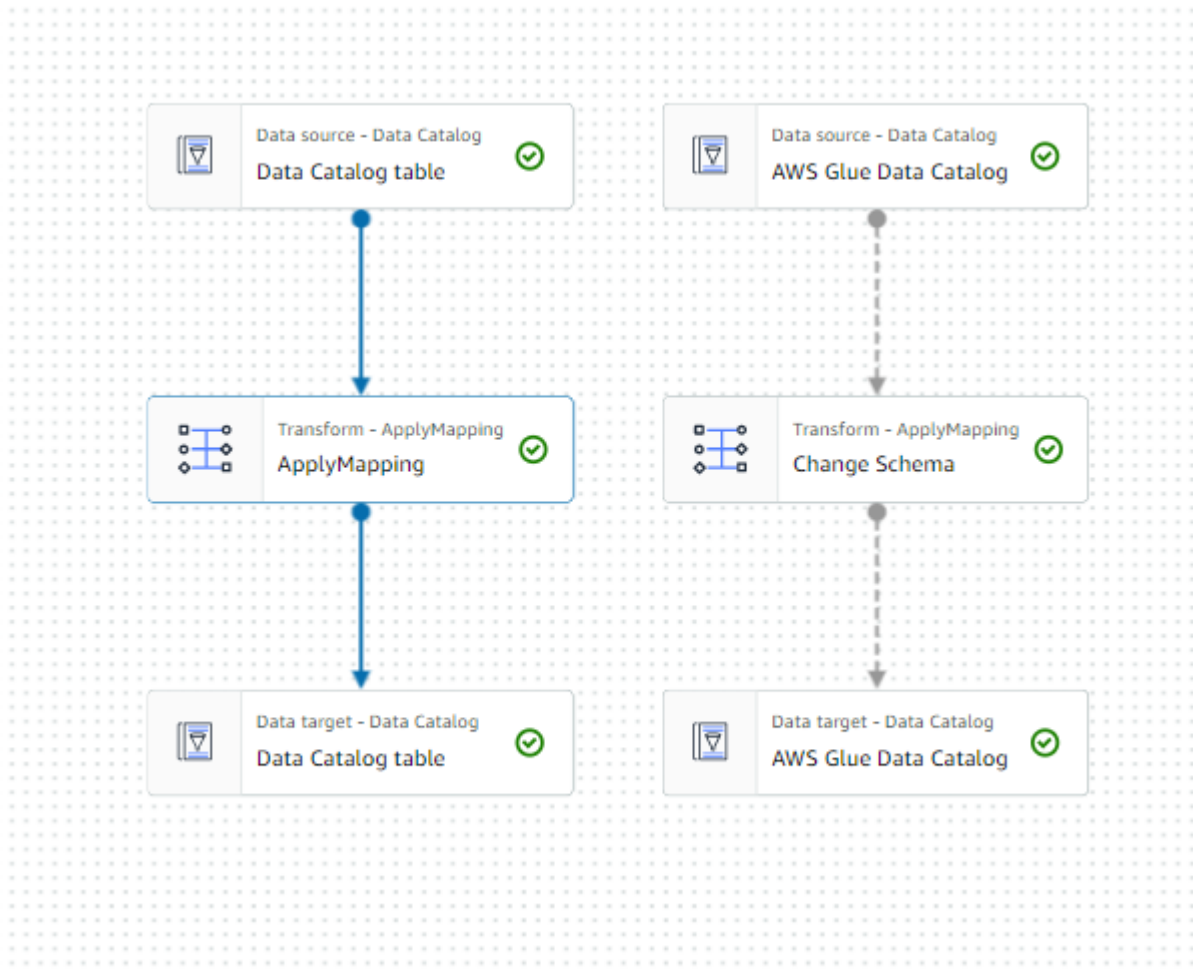
AWS Glue Data Catalog
AWS Glue Data Catalog table as the data target.

Your jobs (3) [Info](#)

Filter jobs
< 1 >
⚙️

<input type="checkbox"/>	Job name	Type	Last modified	AWS Glue version
--------------------------	----------	------	---------------	------------------

se realizó uno para el tiempo y otro para el espectador



periodicos_conex job Last modified on 24/4/2023, 20:38:34 Actions Save Run

Visual | Script | Job details | **Runs** | Data quality New | Schedules | Version Control

Job runs (1/1) Info Last updated (UTC) April 25, 2023 at 01:40:04 View details Stop job run Table View Card View

Run status	Retry	Start time	End time	Duration	Capacity	Worker type	GL
Running	0	04/24/2023 20:38:39	-	1 m 8 s	10 DPUs	G.1X	3.0

periodicos_conex job Last modified on 24/4/2023, 20:38:34 Actions Save Run

Visual | Script | Job details | **Runs** | Data quality New | Schedules | Version Control

Job runs (1/1) Info Last updated (UTC) April 25, 2023 at 01:40:34 View details Stop job run Table View Card View

Run status	Retry	Start time	End time	Duration	Capacity	Worker type	GL
Succeeded	0	04/24/2023 20:38:39	04/24/2023 20:40:07	1 m 11 s	10 DPUs	G.1X	3.0

Realizamos consultas

```
mysql> use periodicos_p;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> select * from El_Tiempo
-> ;

+-----+-----+-----+
| titulo | categoria | link |
+-----+-----+-----+
| titulo | categoria | link |
| Los pecados del PAE: ¿por qué 240.000 niños están sin alimentación escolar? | vida | /vida/educacion/los-pecados-del-pae-por-que-240-000-ninos-estan-sin-alimentacion-escolar-762257 |
| Fórmula 1 en Barranquilla: ¿de dónde saldrá la plata para carrera? | colombia | /colombia/barranquilla/carrera-de-formula-1-en-barranquilla-cuanto-costara-hacerla-762112 |
```

```
Immediate (Javascript (br x ejercicio4/funciones.py - $x ejercicio4/test_funciones.py) x mysql - "ip-172-31-20-73" x +
| Fórmula 1 en Barranquilla: ¿de dónde saldrá la plata para carrera? | colombia | /colombia/barranquilla/carrera-de-formula-1-en-barranquilla-cuanto-costara-hacerla-762112 |
| ¿Por qué Guaidó no puede participar en la cumbre de Bogotá? | politica | /politica/gobierno/por-que-guaido-no-puede-participar-en-cumbre-de-bogota-analisis-762136 |
| Acribillan a 4 colombianos en México | unidad-investigativa | /unidad-investigativa/mexico-los-cuatro-colombianos-acribillados-en-carretera-de-zacatecas-762100 |
| Fórmula 1 en Barranquilla: ¿de dónde saldrá la plata para carrera? | colombia | /colombia/barranquilla/carrera-de-formula-1-en-barranquilla-cuanto-costara-hacerla-762112 |
| Mauricio lizcano saldría de la Presidencia antes de una semana | politica | /politica/gobierno/mauricio-lizcano-saldria-de-la-presidencia-antes-de-una-semana-762278 |
| Nacional, alcaldía y barra brava: detalles del aplazamiento de reunión | deportes | /deportes/futbol-colombiano/atletico-nacional-alcaldia-y-barras-no-se-reunen-estos-son-los-motivos-762288 |
| ¿De qué debería hablarse en la Cumbre Internacional sobre Venezuela? | mundo | /mundo/venezuela/de-que-deberia-hablar-se-en-la-cumbre-internacional-sobre-venezuela-762261 |
| Los otros pasajeros VIP de avioneta que ha movilizó al Ministro de Transporte | unidad-investigativa | /unidad-investigativa/ministro-reyes-y-otros-pasajeros-de-avioneta-del-contratista-pedro-contecha-762269 |
| Este es perfil criminal del hombre capturado por abusar mujeres en Picap | bogota | /bogota/abusador-de-picap-en-bogota-esta-es-su-identidad-y-su-historial-criminal-762239 |
| Sismo de magnitud 7,3 sacude Indonesia: activan alerta de tsunami | mundo | /mundo/asia/terremoto-de-magnitud-7-3-sacude-indonesia-activan-alerta-de-tsunami-762282 |
| Se hunde la moción de censura contra el canciller Alvaro Levva en la Cámara | politica | /politica/congreso/alvaro-levva-se-hunde-la-mocion-de-censura-contra-el-canciller-alvaro-levva-en-la-camara-762282 |
```

```
mysql> select * from El_Espectador;
+-----+-----+-----+
| titulo | categoria | link |
+-----+-----+-----+
| futbol colombiano las propuestas de la dimayor para frenar la violencia en el futbol colombiano | deportes | /deportes/futbol-c |
| olombiano/las-propuestas-de-la-dimayor-para-frenar-la-violencia-en-el-futbol-colombiano/ | | |
| procuraduria solicito audiencia publica por presuntas irregularidades en relleno dona juana noticias hoy | bogota | /bogota/procuradur |
| ia-solicito-audiencia-publica-por-presuntas-irregularidades-en-relleno-dona-juana-noticias-hoy/ | | |
| finanzas personales el dolar abrio la semana a la baja y quedo en 4468 este 24 de abril de 2023 | economia | /economia/finanzas |
+-----+-----+-----+
```

```
mysql - "ip-172-31-20-73" x
+-----+-----+-----+
| titulo | categoria | link |
+-----+-----+-----+
| ia-solicito-audiencia-publica-por-presuntas-irregularidades-en-relleno-dona-juana-noticias-hoy/ | bogota | /bogota/procuradur |
| finanzas personales el dolar abrio la semana a la baja y quedo en 4468 este 24 de abril de 2023 | economia | /economia/finanzas |
| -personales/el-dolar-abrio-la-semana-a-la-baja-y-quedo-en-4468-este-24-de-abril-de-2023/ | | |
| mocion de censura contra canciller alvaro leyva se cayo en la camara de representantes | politica | /politica/mocion-d |
| e-censura-contra-canciller-alvaro-leyva-se-cayo-en-la-camara-de-representantes/ | | |
| mas paises 47 personas murieron de hambre para seguir a jesucristo en una secta en africa | mundo | /mundo/mas-paises/ |
| 47-personas-murieron-de-hambre-para-seguir-a-jesucristo-en-una-secta-en-africa/ | | |
| yo vi de cerca la cara de la violencia luz janeth forero directora de la ubpd | judicial | /judicial/yo-vi-de |
| -cerca-la-cara-de-la-violencia-luz-janeth-forero-directora-de-la-ubpd/ | | |
| mas paises sudan y la comunidad internacional ante una guerra civil en ciernes noticias hoy | mundo | /mundo/mas-paises/ |
| sudan-y-la-comunidad-internacional-ante-una-guerra-civil-en-ciernes-noticias-hoy/ | | |
| caso juan sebastian arismendi familiares senalan que policia encubre a culpables noticias hoy | bogota | /bogota/caso-juan- |
| sebastian-arismendi-familiares-senalan-que-policia-encubre-a-culpables-bogota-noticias-hoy/ | | |
| contenido patrocinado guardianes de la selva asi funciona la tecnologia que salva ecosistemas | contenido-patrocinado | /contenido-patroci |
| nado/guardianes-de-la-selva-asi-funciona-la-tecnologia-que-salva-ecosistemas/ | | |
| columnistas leopoldo villar borda cazador cazado | opinion | /opinion/columnist |
| as/leopoldo-villar-borda-cazador-cazado/ | | |
| columnistas lorenzo madrigal el nuevo estilo | opinion | /opinion/columnist |
| as/lorenzo-madrigal-el-nuevo-estilo/ | | |
+-----+-----+-----+
```

se crean los triggers

Trigger successfully created

The following trigger was created: "trigg1"

AWS Glue > Triggers > trigg1

trigg1

Last updated (UTC)
April 25, 2023 at 02:49:16

Edit trigger

Trigger properties

Name

trigg1

Description

-

Trigger type

On demand

Status

Created

Associated workflow

-

Target resources

Tags

Resources to trigger (1)

List of resources to start once the trigger activates

Type	Name	Parameters
------	------	------------

Step 1

Set trigger properties

Step 2

Choose jobs or crawlers to activate

Step 3

Review and create

Set trigger properties

Trigger details

Name

trigg2

Name may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and can be up to 255 characters long.

Description - optional

Enter a description

Descriptions can be up to 2048 characters long.

Trigger type

☐ On demand

Fire the trigger immediately when started.

☐ Schedule

Fire the trigger on a timer.

☒ Job or crawler event

Fire the trigger when job or crawler events match your watched list.

Conditional logic

How to combine conditions to determine when the trigger fires.

☐ All

All watched crawler/job event are checked before firing.

☒ Any

Any watched crawler/job event will fire the trigger.

Watched resources (1)

List of conditions that will start the trigger

Remove

Add a watched resource

Type	Name	Status
<input type="radio"/> Job	periodicosimp job	<input checked="" type="checkbox"/> Succeeded

Step 1

Set trigger properties

Step 2

Choose jobs or crawlers to activate

Step 3

Review and create

Review and create

Trigger successfully created

The following trigger was created: "trigg1"

X

Step 1: Set trigger properties

Edit

Trigger details

Name	Description	Tags
trigg2	-	-

Watched resources (1)

List of conditions that will start the trigger

Type	Name	Status
Job	periodicosimp job	<input checked="" type="checkbox"/> Succeeded

Step 2: Choose jobs or crawlers to activate

Edit

Resources to trigger (1)

List of resources to start once the trigger activates


Type	Name	Parameters
Job	periodicosbeutifulsoup job	-

☐ Enable trigger on creation


Cancel

Previous

Create

 **Triggers successfully created**

The following triggers were created: "trigg3", "trigg2", "trigg1"



[AWS Glue](#) > [Triggers](#) > **trigg3**


trigg3

Last updated (UTC)
April 25, 2023 at 02:59:40



Edit trigger


Trigger properties

Name	Description
trigg3	-
Trigger type	Status
Conditional	 Created
Associated workflow	
-	

- Target resources
- Watched resources
- Tags

Resources to trigger (1)

List of resources to start once the trigger activates

Type	Name	Parameters
Crawler	periodicos_crawler 	-

✔ Triggers successfully created

The following triggers were created: "trigg4", "trigg3", "trigg2", "trigg1"

✕

[AWS Glue](#) > [Triggers](#) > **trigg4**

trigg4

Last updated (UTC)
April 25, 2023 at 03:01:22



Edit trigger

Trigger properties

Name	Description
trigg4	-
Trigger type	Status
Conditional	✔ Created
Associated workflow	
-	

Target resources


Watched resources

Tags


Resources to trigger (1)

List of resources to start once the trigger activates

Type	Name	Parameters
Crawler	periodico2_crawler	-

 **Triggers successfully created**

The following triggers were created: "trigg5", "trigg4", "trigg3", "trigg2", "trigg1"



[AWS Glue](#) > [Triggers](#) > [trigg5](#)


trigg5

Last updated (UTC)
April 25, 2023 at 03:02:55




Edit trigger

Trigger properties

Name	Description
trigg5	-
Trigger type	Status
Conditional	 Created
Associated workflow	
-	

[Target resources](#) | Watched resources | Tags

Resources to trigger (1)
List of resources to start once the trigger activates

Type	Name	Parameters
Job	periodicos_conex job 	-

We've redesigned the AWS Glue Workflows console to make it easier to use. [Let us know what you think.](#) Continue to use the new console, or use the [old console](#).

AWS Glue > Workflows > workflowperiodicos

workflowperiodicos

Last updated (UTC)
April 25, 2023 at 03:04:40



Run workflow

Edit

Delete

Workflow details

Advanced properties

Name workflowperiodicos	Description -	Max concurrency -	Last run status -
Last run -	Last modified April 25, 2023 at 03:04:20	Blueprint name -	Blueprint run Id -

Graph

History

Tags

Legend:

● Start ◆ Trigger 📄 Job 🕸 Crawler ✓ Incomplete ✖ Error ⏸

Remove

Action ▼

The workflow is empty

Add trigger

Last run
-

Last modified
April 25, 2023 at 03:04:20

Blueprint name
-

Blueprint run Id
-

Graph

History

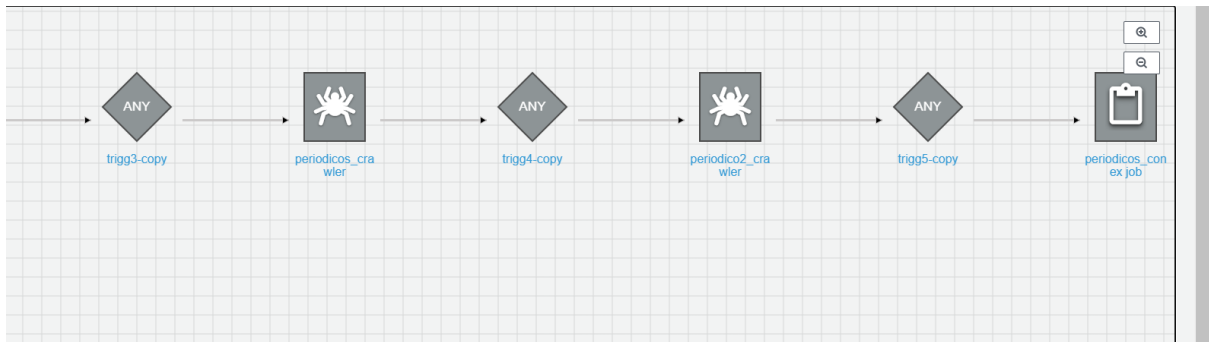
Tags

Legend: ● Start ◆ Trigger 📄 Job 🕸 Crawler ✓ Incomplete ✖ Error ⏸ Deleting

Remove

Action ▼





Introducing the new AWS Glue Workflows console experience. We've redesigned the AWS Glue Workflows console to make it easier to use. [Let us know what you think.](#) Continue to use the new console, or use the [old console](#).

AWS Glue > Workflows > workflowperiodicos

workflowperiodicos Last updated (UTC) April 25, 2023 at 03:27:31 Run workflow Edit Delete

Workflow details | Advanced properties

Name workflowperiodicos	Description -	Max concurrency -	Last run status Completed
Last run April 25, 2023 at 03:27:30	Last modified April 25, 2023 at 03:04:20	Blueprint name -	Blueprint run Id -

Graph | History | Tags

Legend: Start Trigger Job Crawler Incomplete Error Deleting

```

graph LR
    Start(( )) --> trigg1-copy{trigg1-copy}
    trigg1-copy --> periodicosimp_job[periodicosimp job]
    periodicosimp_job --> trigg2-copy{trigg2-copy}
    trigg2-copy --> periodicosbeuti_fulsoup_job[periodicosbeuti fulsoup job]
    periodicosbeuti_fulsoup_job --> trigg3-copy{trigg3-copy}
    trigg3-copy --> periodicos_crawler[periodicos_crawler]
    
```

se comprueba de que si funcionó

Amazon S3 > Buckets > periodicosp2 > headlines/ > final/ > periodo=eltiempo/ > year=2023/ > month=04/

month=04/ Copiar URI de S3

Objetos | Propiedades

Objetos (2)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Recargar Copiar URI de S3 Copiar URL Descargar Abrir Eliminar Acciones

Crear carpeta Cargar

Buscar objetos por prefijo

	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	day=24/	Carpeta	-	-	-
<input type="checkbox"/>	day=25/	Carpeta	-	-	-

Objetos

Propiedades

Objetos (2)

Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

🔄

Copiar URI de S3

Copiar URL

⬇️ Descargar

Abrir ↗️

Eliminar

Acciones ▼

Crear carpeta

Cargar

🔍 Buscar objetos por prefijo

< 1 > ⚙️

<input type="checkbox"/>	Nombre ▲	Tipo ▼	Última modificación ▼	Tamaño ▼	Clase de almacenamiento ▼
<input type="checkbox"/>	📁 day=24/	Carpeta	-	-	-
<input type="checkbox"/>	📁 day=25/	Carpeta	-	-	-