

Proceso de ETL en AWS.

Paso 1. Creación del Bucket en S3: **pruebatecnicajosemqe**

► **Instantánea de la cuenta:** *actualizada cada 24 horas* Todas las regiones de AWS

[Ver panel de Storage Lens](#)

Storage Lens permite visualizar el uso del almacenamiento y las tendencias de la actividad. Las métricas no incluyen los buckets de directorio. [Más información](#)

Buckets de uso general

Buckets de directorio

Buckets de uso general (3) [Información](#) Todas las regiones de AWS

[Copiar ARN](#) [Vaciar](#) [Eliminar](#) [Crear bucket](#)

Los buckets son contenedores de datos almacenados en S3.

< 1 >

	Nombre	Región de AWS	Analizador de acceso de IAM	Fecha de creación
<input type="radio"/>	aws-glue-assets-796092240045-us-east-2	EE.UU. Este (Ohio) us-east-2	Ver analizador para us-east-2	6 Apr 2025 11:05:56 PM -05
<input type="radio"/>	destinopruebatecnica	EE.UU. Este (Ohio) us-east-2	Ver analizador para us-east-2	5 Apr 2025 4:34:00 PM -05
<input type="radio"/>	pruebatecnicajosemqe	EE.UU. Este (Ohio) us-east-2	Ver analizador para us-east-2	5 Apr 2025 3:48:32 PM -05

Paso 2. En el Bucket creado se cargan las tablas en formato CSV.

pruebatecnicajosemqe [Información](#)

Objetos

Metadatos

Propiedades

Permisos

Métricas

Administración

Puntos de acceso

Objetos (5)

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

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](#)

< 1 >

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	customer.csv	csv	5 Apr 2025 4:14:04 PM -05	167.5 KB	Estándar
<input type="checkbox"/>	Films.csv	csv	5 Apr 2025 4:14:05 PM -05	189.1 KB	Estándar
<input type="checkbox"/>	inventory.csv	csv	5 Apr 2025 4:14:06 PM -05	146.2 KB	Estándar
<input type="checkbox"/>	rental.csv	csv	5 Apr 2025 4:14:08 PM -05	1.2 MB	Estándar
<input type="checkbox"/>	store.csv	csv	5 Apr 2025 4:14:08 PM -05	108.0 B	Estándar

Paso 3. Se realizar la creación de Crawler: crawlerprueba este se encargará de identificar el formato de los datos para cargar en el Glue Data Catalog.

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 (1) [Info](#)

Last updated (UTC)
April 7, 2025 at 16:23:13

[Action](#) [Run](#) [Create crawler](#)

View and manage all available crawlers.

< 1 >

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run times...	Log	Table changes fr...
<input type="checkbox"/>	crawlerprueba	Ready		Succeeded	April 5, 2025 at 2...	View log	5 created

crawlerprueba

Last updated (UTC)

April 7, 2025 at 16:27:43

Run crawler

Edit

Delete

Crawler properties

<div>Name</div> <div>crawlerprueba</div>	<div>IAM role</div> <div> AWSGlueServiceRole-GlueS3 <div></div> </div>	<div>Database</div> <div>bdprueba</div>	<div>State</div> <div>READY</div>
<div>Description</div> <div>Identificar la estructura de fuente de datos</div>	<div>Security configuration</div> <div>-</div>	<div>Lake Formation configuration</div> <div>-</div>	<div>Table prefix</div> <div>-</div>
<div>Maximum table threshold</div> <div>-</div>			

► Advanced settings

Crawler runs

Schedule

Data sources

Classifiers

Tags

Data sources (1)

Info

Edit

Remove

Add a data source

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

Type	Data source	Parameters
<div></div> <div>S3</div>	s3://pruebatecnicajosemqe	Recrawl all

Databases (1)

A database is a set of associated table definitions, organized into a logical group.

<input type="checkbox"/>	Name	Description	Location URI	Created on (UTC)
<input type="checkbox"/>	bdprueba	Base de datos en el catalogo de Glue	-	April 5, 2025 at 21:20:56

Last updated (UTC)

April 7, 2025 at 16:30:37

⌚

Edit

Delete

Add database

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 (1)
[Info](#)

Last updated (UTC)

April 7, 2025 at 16:53:26

Action ▾

Run

Create crawler

View and manage all available crawlers.

<input type="checkbox"/>	Name ▾	State ▾	Schedule	Last run ▾	Last run times... ▾	Log	Table changes fr...
<input type="checkbox"/>	crawlerprueba	Ready		Succeeded	April 5, 2025 at 2...	View log	5 created

bdprueba

Last updated (UTC)

April 7, 2025 at 16:34:32

⌂

Edit

Delete

Database properties

Name	Description	Location	Created on (UTC)
bdprueba	Base de datos en el catalogo de Glue	-	April 5, 2025 at 21:20:56

Tables (5)

Last updated (UTC)

April 7, 2025 at 16:34:33

⌂

Delete

Add tables using crawler

Add table

View and manage all available tables.

Q

Filter tables

<

1

>

⚙

<input type="checkbox"/>	Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statis...
<input type="checkbox"/>	customer_csv	bdprueba	s3://pruebatecnica	CSV	-	Table data	View data quality	View statistics
<input type="checkbox"/>	films_csv	bdprueba	s3://pruebatecnica	CSV	-	Table data	View data quality	View statistics
<input type="checkbox"/>	inventory_csv	bdprueba	s3://pruebatecnica	CSV	-	Table data	View data quality	View statistics
<input type="checkbox"/>	rental_csv	bdprueba	s3://pruebatecnica	CSV	-	Table data	View data quality	View statistics
<input type="checkbox"/>	store_csv	bdprueba	s3://pruebatecnica	CSV	-	Table data	View data quality	View statistics

Paso 8. Se realizar la creación del Script en el ETL Job, allí copiamos el código realizado en Google Colab utilizando Python como lenguaje de programación y framework de procesamiento de datos PySpark, El código de adecuarse para cargar las tablas que fueron creadas en la base de datos de S3.

CodigoPruebaJob

Last modified on 7/4/2025, 11:05:50

ActionsSaveRun

ScriptJob detailsRunsData qualitySchedulesVersion Control

ScriptInfo

```
1 import sys
2 from pyspark.context import SparkContext
3 from awsglue.context import GlueContext
4 from awsglue.utils import getResolvedOptions
5 from awsglue.job import Job
6
7 # Definimos los argumentos requeridos por Glue
8 args = getResolvedOptions(sys.argv, ['JOB_NAME'])
9
10 # Inicializamos Spark y Glue
11 sc = SparkContext()
12 glueContext = GlueContext(sc)
13 spark = glueContext.spark_session
14
15 # Inicializamos el Job
16 job = Job(glueContext)
17 job.init(args['JOB_NAME'], args)
```

PythonLn 1, Col 1Errors: 0Warnings: 0

Paso 9. Se compila el código, comprobando que no existan errores de compilación.

CodigoPruebaJob

Last modified on 7/4/2025, 11:05:50

ActionsSaveRun

ScriptJob detailsRunsData qualitySchedulesVersion Control

Job runs (1/15)Info

Last updated (UTC)
April 7, 2025 at 16:42:20

View detailsStop job runTroubleshoot with AI

Table ViewCard View

Filter job runs by property

Run status	Retries	Start time (Local)	End time (Local)	Duration	Capacity (...)	Worker type	Glue
Succeeded	0	04/07/2025 11:05:53	04/07/2025 11:09:33	3 m 33 s	10 DPUs	G.1X	5.0
Succeeded	0	04/07/2025 00:15:53	04/07/2025 00:19:21	3 m 22 s	10 DPUs	G.1X	5.0

Run detailsInput arguments (10)Continuous logsRun insightsMetricsTroubleshooting analysis - previewSpark UI

Job name	Start time (Local)	Glue version	Last modified on (Local)
CodigoPruebaJob	04/07/2025 11:05:53	5.0	04/07/2025 11:09:33
Id	End time (Local)	Worker type	Log group name
jr_3829e1b2bec648e0288e7f7127c5a7e86de1400a3357d8f4e0bbfe7163a8d9b	04/07/2025 11:09:33	G.1X	/aws-glue/jobs
Run status	Start-up time	Max capacity	Number of workers
Succeeded	6 seconds	10 DPUs	10
Retry attempt number	Execution time	Execution class	Timeout

Paso 10. Se obtienen y analizan los resultados.

CloudWatch > Grupos de registros > /aws-glue/jobs/output > jr_3829e1b2bec648e0288e7f7127c5a7e86de1400a3357d8f4e0bbfe7163a8d9b

CloudWatch

Favoritos y recientes

Paneles

- Alarmas
- Registros
 - Grupos de registros [Nuevo](#)
 - Anomalías de registros
 - Live Tail
 - Logs Insights [Nuevo](#)
 - Contributor Insights
- Métricas
- Rastros de X-Ray [Nuevo](#)
- Eventos
- Señales de aplicaciones
- Monitorización de redes

Eventos de registro

Puede utilizar la barra de filtros a continuación para buscar y hacer coincidir términos, frases o valores en sus eventos de registro. [Más información sobre los patrones de filtro](#)

Filtrar eventos: pulse Intro para buscar

Borrar 1m 30m 1h 12h Personalizado Zona horaria UTC

Mostrar

▼ Marca temporal | Mensaje

2025-04-07T16:09:03.793Z

columna	Q1	Q2	Q3	Limite_Inferior	Limite_Superior	Outliers
length	80.0	113.0	147.0	-20.5	247.5	0
num_voted_users	18400.0	39200.0	58100.0	-41150.0	117650.0	0
release_year	2006.0	2006.0	2006.0	2006.0	2006.0	0
rental_duration	4.0	5.0	6.0	1.0	9.0	0
rental_rate	0.99	2.99	4.99	-5.01	10.99	0
replacement_cost	14.99	20.99	24.99	-0.01	39.99	0

Número total de películas, clientes, tiendas y alquileres:

2025-04-07T16:09:05.106Z Películas: 958

Películas: 958

[Volver arriba](#)

Diagrama de flujo proceso en AWS.

