

Creación de bucket en s3 para carga y transformación de datos

Amazon S3 > Buckets > ecommerce-data-raw-dataengineer

ecommerce-data-raw-dataengineer

Información

Objetos

Metadatos

Propiedades

Permisos

Métricas

Administración

Puntos de acceso

Objetos (10)

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](#)

Buscar objetos por prefijo

< 1 >

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	<a href="#">processed-products-corrected/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">processed-products/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">processed-purchases/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">products/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">purchase-relations/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">purchase-totals/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">purchases/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">results_queries/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">scripts/</a>	Carpeta	-		-
<input type="checkbox"/>	<a href="#">temp/</a>	Carpeta	-		-

Crawlers para la inferencia de los esquemas

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 (6)

Info

Last updated (UTC)  
August 1, 2025 at 01:25:08

Action

Run

Create crawler

View and manage all available crawlers.

Filter crawlers

< 1 >

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run timestamp	Log	Table changes from l...
<input type="checkbox"/>	<a href="#">processed-products</a>	Ready		Succeeded	July 31, 2025 at 00:4...	<a href="#">View log</a>	1 created
<input type="checkbox"/>	<a href="#">processed-purchases</a>	Ready		Succeeded	July 31, 2025 at 00:4...	<a href="#">View log</a>	1 created
<input type="checkbox"/>	<a href="#">processed-relationship</a>	Ready		Succeeded	July 31, 2025 at 00:4...	<a href="#">View log</a>	1 created
<input type="checkbox"/>	<a href="#">productos_crawler</a>	Ready		Succeeded	July 30, 2025 at 23:0...	<a href="#">View log</a>	1 created
<input type="checkbox"/>	<a href="#">purchases_crawler</a>	Ready		Succeeded	July 30, 2025 at 23:0...	<a href="#">View log</a>	1 created
<input type="checkbox"/>	<a href="#">total_crawler</a>	Ready		-	-	-	-

# Tablas creadas en el data catalog

## Tables

A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Tables (6)

Last updated (UTC)  
August 1, 2025 at 01:25:06

Delete

Add tables using crawler

Add table

View and manage all available tables.

<input type="checkbox"/>	Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
<input type="checkbox"/>	ecommerceproducts	ecommerce_central_db	s3://ecommerce-data-r	JSON	-	Table data	View data quality	View statistics
<input type="checkbox"/>	processed_products	ecommerce_central_db	s3://ecommerce-data-r	Parquet	-	Table data	View data quality	View statistics
<input type="checkbox"/>	processed_purchases	ecommerce_central_db	s3://ecommerce-data-r	Parquet	-	Table data	View data quality	View statistics
<input type="checkbox"/>	products_corrected	ecommerce_central_db	s3://ecommerce-data-r	-	-	Table data	View data quality	View statistics
<input type="checkbox"/>	purchase_relations	ecommerce_central_db	s3://ecommerce-data-r	Parquet	-	Table data	View data quality	View statistics
<input type="checkbox"/>	purchasespurchases	ecommerce_central_db	s3://ecommerce-data-r	JSON	-	Table data	View data quality	View statistics

## Glue Jobs

Create job

Author in a visual interface focused on data flow.  
Visual ETL

Author using an interactive code notebook.  
Notebook

Author code with a script editor.  
Script editor

Example jobs

Create example job

Your jobs (7)

Filter jobs by property

<input type="checkbox"/>	Job name	Type	Created by	Last modified	AWS Glue version	Action
<input type="checkbox"/>	Load Data Ecommerce	Glue ETL	Notebook	31/7/2025, 23:19:02	5.0	-
<input type="checkbox"/>	Data Transformation ecommerce	Glue ETL	Notebook	31/7/2025, 23:17:08	5.0	-
<input type="checkbox"/>	Data Transformation ecommerce	Glue ETL	Script	31/7/2025, 23:08:54	3.0	Upgrade with AI
<input type="checkbox"/>	Load Data Ecommerce	Glue ETL	Script	31/7/2025, 21:24:17	3.0	Upgrade with AI
<input type="checkbox"/>	Load Data Ecommerce	Glue ETL	Script	31/7/2025, 17:32:49	3.0	Upgrade with AI
<input type="checkbox"/>	Data Transformation ecommerce	Glue ETL	Script	31/7/2025, 15:46:44	3.0	Upgrade with AI
<input type="checkbox"/>	1.Pyspark Tutorial	Glue ETL	Notebook	22/7/2025, 15:09:59	5.0	-

## Base de datos creada en AWS redshift

Redshift query editor v2

Create

Load data

Filter resources

Serverless: default-workgroup

native databases (3)

dev

ecommerce

public

Tables

processed\_products

processed\_purchases

purchase\_relations

Views

Functions

Stored procedures

sample\_data\_dev

external databases (1)

Untitled 1

Run

Limit 100

Explain

Isolated session

Serverless: de...

dev

Schedule

1

select \* from processed\_products

Result 1 (100)

Export

Chart

product_id	product_name	description	price	category	created_at
67	Lemon Garlic Shrimp	Marinated shrimp in a garl...	8.989999771118164	Food - Seafood	2023-06-07
67	Mango Salsa	Fresh mango salsa with a...	3.490000009536743	Food - Condiments	2023-04-25
67	Digital Wireless Meat Th...	Bluetooth thermometer th...	39.9800016784668	Kitchen	2023-04-11
73	Corn Tortillas	Soft and warm corn tortill...	2.490000009536743	Food - Bakery	2023-05-18
73	Cilantro Lime Rice	Easy-to-prepare rice with ...	2.990000009536743	Food - Grains	2022-07-15
73	Apple Cinnamon Oatmeal	Warm oatmeal flavored wi...	2.890000104904175	Food - Breakfast	2022-02-02
542	Honey	Pure and natural honey, g...	5.489999771118164	Food - Condiments	2022-04-16
542	Travel Jewelry Organizer	Compact storage for your ...	19.989999771118164	Accessories	2023-03-30

Cálculo del total Total, con descuentos

	purchase_id	status	credit_card_type	purchase_date	total
<input type="checkbox"/>	01K18SC931XARQEGD0...	pending	diners-club-carte-blanche	2022-09-25	55.254501378536226
<input type="checkbox"/>	01K18SDR1DW6FKBEX...	pending	bankcard	2022-09-25	148.27300295829772
<input type="checkbox"/>	01K18SDR46CX4QF86G...	pending	jcb	2022-09-25	1.5119999885559083
<input type="checkbox"/>	01K18SDRFGZPVRH6F...	cancelled	diners-club-enroute	2022-09-25	14.989999771118164
<input type="checkbox"/>	01K18SCJRP6EAZZ054...	completed	china-unionpay	2024-04-22	59.970001220703125
<input type="checkbox"/>	01K18SCJRWZ68WPR1...	cancelled	maestro	2024-04-22	56.92999887466431
<input type="checkbox"/>	01K18SDRSYWR93DB0...	completed	china-unionpay	2024-04-22	28.739999771118164
<input type="checkbox"/>	01K18SC8RACNF2PGC1...	completed	switch	2021-11-29	163.73999905586243
<input type="checkbox"/>	01K18SC99YW5WZBYH...	cancelled	mastercard	2021-11-29	5.989999771118164

Cantidad de producto por compra

	purchase_id	product_id	quantity	discount
<input type="checkbox"/>	01K18SECQGPKQTRM1...	707	1	0
<input type="checkbox"/>	01K18SECQGPKQTRM1...	174	1	0
<input type="checkbox"/>	01K18SECQHJP1JJBS1...	633	1	0
<input type="checkbox"/>	01K18SECQKZVF67TS1...	345	1	0
<input type="checkbox"/>	01K18SECQKZVF67TS1...	320	1	0
<input type="checkbox"/>	01K18SECQN3CBP09C...	13	2	0
<input type="checkbox"/>	01K18SECQPQN16770N...	98	1	0
<input type="checkbox"/>	01K18SECQRCG904J4N...	56	2	0
<input type="checkbox"/>	01K18SECQV23952V31T...	81	1	0

Carga de los datos a s3 desde airflow

Comando: docker-compose exec webserver airflow tasks test ecommerce\_etl extract\_data 2025-07-31

```
[2025-08-01 04:06:50,710] {taskinstance.py:1509} INFO - Exporting the following env vars:
AIRFLOW_CTX_DAG_OWNER=***
AIRFLOW_CTX_DAG_ID=ecommerce_etl
AIRFLOW_CTX_TASK_ID=extract_data
AIRFLOW_CTX_EXECUTION_DATE=2025-07-31T00:00:00+00:00
AIRFLOW_CTX_TRY_NUMBER=3
AIRFLOW_CTX_DAG_RUN_ID>manual__2025-07-31T00:00:00+00:00
[+] products.json guardado en S3 - 200
[+] purchases.json guardado en S3 - 200
Datos extraídos y guardados correctamente.
[2025-08-01 04:07:02,097] {python.py:177} INFO - Done. Returned value was: None
[2025-08-01 04:07:02,098] {taskinstance.py:1323} INFO - Marking task as SUCCESS. dag_id=ecommerce_etl, task_id=extract_data, execution_date=20250731T000000, start_date=20250801T003222, end_date=20250801T040702
```

## Transformación de los datos usando AWS Glue desde Airflow

```
erce current run state with status RUNNING
[2025-08-01 04:18:54,864] {glue.py:247} INFO - Polling for AWS Glue Job Data Transformation ecomm
erce current run state with status RUNNING
[2025-08-01 04:19:01,022] {glue.py:247} INFO - Polling for AWS Glue Job Data Transformation ecomm
erce current run state with status RUNNING
[2025-08-01 04:19:07,181] {glue.py:247} INFO - Polling for AWS Glue Job Data Transformation ecomm
erce current run state with status RUNNING
[2025-08-01 04:19:13,333] {glue.py:236} INFO - Exiting Job jr_ce1951a4e775a09b722644517dd2552faa2
58a5853b5e0863f81bcd1273ddcc Run State: SUCCEEDED
[2025-08-01 04:19:13,334] {glue.py:153} INFO - AWS Glue Job: Data Transformation ecommerce status
: SUCCEEDED. Run Id: jr_ce1951a4e775a09b722644517dd2552faa258a5853b5e0863f81bcd1273ddcc
[2025-08-01 04:19:13,785] {taskinstance.py:1323} INFO - Marking task as SUCCESS. dag_id=ecommerce
_etl, task_id=transform_data_with_glue, execution_date=20250731T000000, start_date=20250801T00322
1, end_date=20250801T041913
PS C:\Backup Santiago Moreno\Documents\Cursos\ecommerce-etl>
```