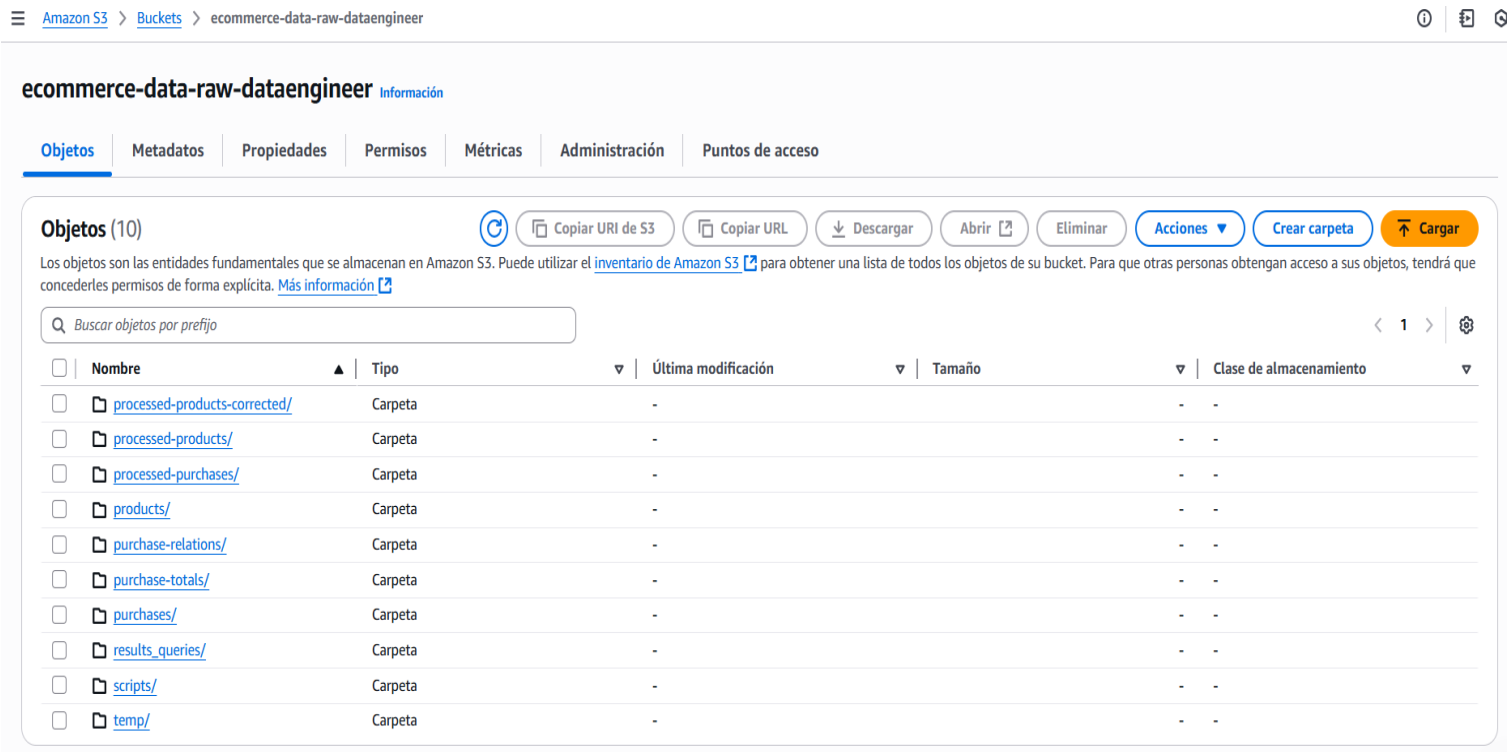
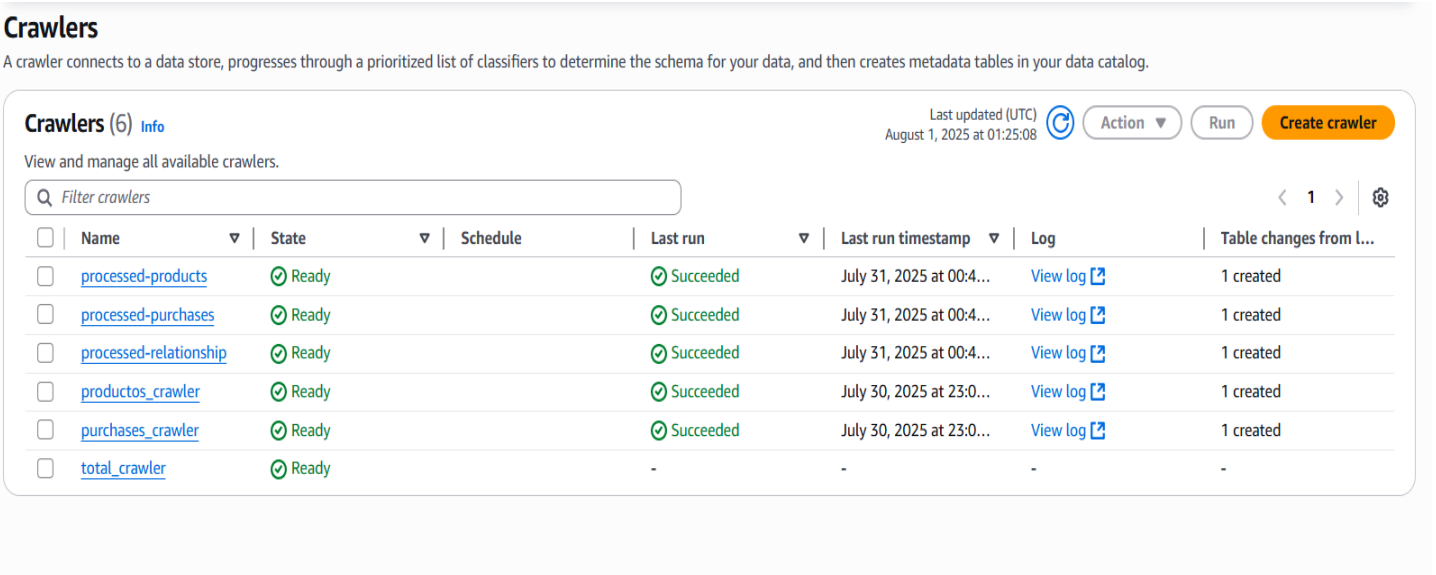
Creación de bucket en s3 para carga y transformación de datos



Crawlers para la inferencia de los esquemas



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.

Filter tables

1

Name	Database	Location	Classification	Deprecated	View data	Data quality	Column statistics
ecommerceproducts	ecommerce_central_db	s3://ecommerce-data-r	JSON	-	Table data	View data quality	View statistics
processed_products	ecommerce_central_db	s3://ecommerce-data-r	Parquet	-	Table data	View data quality	View statistics
processed_purchases	ecommerce_central_db	s3://ecommerce-data-r	Parquet	-	Table data	View data quality	View statistics
products_corrected	ecommerce_central_db	s3://ecommerce-data-r	-	-	Table data	View data quality	View statistics
purchase_relations	ecommerce_central_db	s3://ecommerce-data-r	Parquet	-	Table data	View data quality	View statistics
purchasespurchases	ecommerce_central_db	s3://ecommerce-data-r	JSON	-	Table data	View data quality	View statistics

Glue Jobs

AWS Glue Studio Info

Create job Info

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 Info — Create example job

Your jobs (7) Info

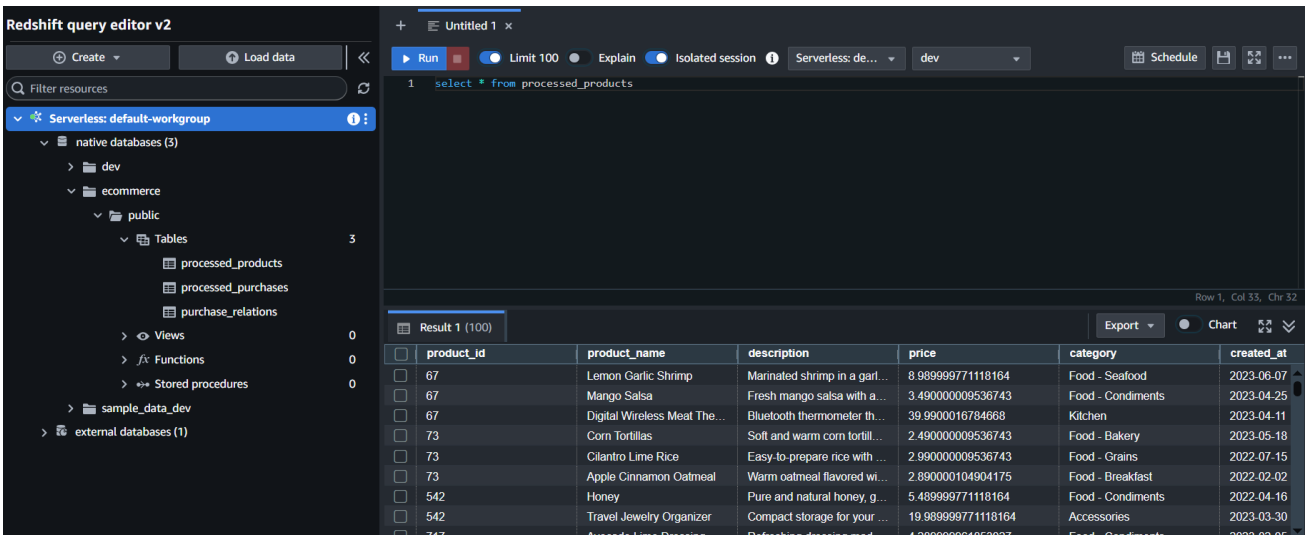
Filter jobs by property

Actions | Run job

1

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

Base de datos creada en AWS redshift



Cálculo del total Total, con descuentos

Result 1 (100) Export Chart

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

Cantidad de producto por compra

Result 1 (100)

purchase_id	product_id	quantity	discount
01K18SECQGPKQTRM1...	707	1	0
01K18SECQGPKQTRM1...	174	1	0
01K18SECQHJP1JJBS1...	633	1	0
01K18SECQKZVF67TS1...	345	1	0
01K18SECQKZVF67TS1...	320	1	0
01K18SECQN3CBP09C...	13	2	0
01K18SECQPQN16770N...	98	1	0
01K18SECQRCG904J4N...	56	2	0
01K18SECQV23952V31T...	81	1	0

Query ID 616287

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