

Prueba técnica ingeniero de datos

Carlos Alberto alfaro

Ejercicio 1

Una compañía comercializadora de energía compra la electricidad a los generadores en el mercado mayoritario, donde después de una serie de contratos y control riesgos de precios esta se vende a los usuarios finales que pueden ser clientes residenciales, comerciales o industriales.

El sistema de la compañía que administra este producto tiene la capacidad de exportar la información de proveedores, clientes y transacciones en archivos CSV.

Requisitos técnicos:

1. Crear una estrategia de datalake en s3 con las capas que usted considere necesario tener y cargue esta información de manera automática y periódica. Los archivos deben particionarse por fecha de carga.
2. Realice 3 transformaciones básicas de datos utilizando AWS Glue y transforme la información para que esta sea almacenada en formato parquet en una zona procesada.
3. Utilizando AWS Glue, crea un proceso que detecte y catalogue automáticamente los esquemas de los datos almacenados en el datalake.
4. Utilizando Amazon Athena desde Python, realiza consultas SQL básicas sobre los datos que han sido transformados.

Documentación:

1. Realiza una descripción detallada del pipeline de datos construido.
2. Indica que proceso es necesario seguir para configurar permisos y políticas para los diferentes servicios de AWS utilizados.

Puntos adicionales (plus)

1. Crea la IaC (Infraestructura como código) necesaria para desplegar esta solución en AWS.
2. Configure AWS Lakeformation para centralizar el gobierno, la seguridad y compartir los datos alojados en el datalake creado.
3. Construya un pipeline de datos que permita cargar esta información desde el datalake en la zona procesada a un datawarehouse en redshift.

Utilice información ficticia en los archivos csv que se simula entregue el sistema transaccional. A continuación, tendrá una recomendación de columnas de estos archivos, pero podrá ajustarla a sus necesidades:

Archivo proveedores: nombre de proveedor, tipo de energía (eólica, hidroeléctrica, nuclear).

Archivo clientes: tipo de identificación, identificación, nombre, ciudad.

Archivo transacciones: tipo de transacción (venta o compra), nombre del cliente/proveedor, cantidad comprada, precio, tipo de energía.

Importante: El código fuente creado deberá ser desplegado en una herramienta de control de versiones como github, azure devops, gitlab o similares.

SOLUCION PLANTEADA

Descripción del desarrollo del Ejercicio 1

Para el desarrollo del ejercicio, el primer paso consistió en la **creación de tres fuentes de datos ficticias** (clientes, proveedores y transacciones), con la estructura definida previamente. Estas fuentes fueron generadas en formato **CSV**, tal como se evidencia en las imágenes adjuntas.

A continuación, se procedió a configurar el servicio **Amazon S3**, donde se crearon dos buckets:

- datalake-energia-carlos (zona de entrada del data lake)
- datawarehouse-energia-carlos (zona de salida o procesada)

Dentro de cada bucket, se organizaron carpetas correspondientes a cada tipo de dato: clientes/, proveedores/ y transacciones/, con el fin de almacenar los archivos de manera estructurada y particionada por fecha de carga.

El siguiente paso fue utilizar el servicio **AWS Glue**, específicamente la herramienta **Data Catalog**. Allí se creó un **crawler**, que se encargó de identificar automáticamente los esquemas de las fuentes de datos en S3, mapearlas y registrarlas en una base de datos del catálogo. Para su correcta ejecución, fue necesario:

- Crear y configurar una política de permisos (**IAM Role**) con acceso tanto a los buckets de S3 como a los recursos de Glue.
- Definir correctamente la ruta de origen del bucket y carpeta donde se encuentran los archivos CSV.
- Crear una base de datos en Glue que almacenara los metadatos generados por el crawler.

Una vez ejecutado el crawler, se validó que las tablas fueran correctamente creadas en la base de datos del Data Catalog. Posteriormente, se utilizó el servicio **Amazon Athena** para realizar consultas SQL y explorar los datos catalogados directamente desde S3.

Después de validar los datos, se procedió a la creación de un **ETL Job en AWS Glue**, el cual permitió transformar los datos de manera sencilla:

- Se unieron las tablas de clientes y transacciones a través del campo `id_cliente`.
- Se transformaron los archivos CSV al formato **Parquet**, lo cual optimiza el rendimiento de las consultas y reduce el almacenamiento.
- La salida de este proceso se almacenó en la zona procesada del bucket `datawarehouse-energia-carlos`.

Para esta etapa también fue necesario:

- Configurar la base de datos de destino que recibiría la información transformada.
- Asignar adecuadamente los permisos del rol IAM utilizado, asegurando que tuviera autorización para escribir en el bucket de salida.

FUENTES CREADAS SEGÚN ESTRUCTURA DEFINIDA

- ✓ Al principio de esta semana

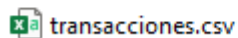
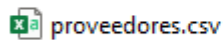
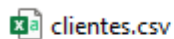


TABLA CLIENTES

A	B	C	D	E	F
id_cliente	nombre_cliente	tipo_cliente	region	consumo_estimado_kwh	
C001	Robert Smith	comercial	Oriente	11778	
C002	Small PLC	comercial	Centro	10882	
C003	Briggs Ltd	residencial	Oriente	7322	
C004	Lee, Khan and Flores	industrial	Sur	5715	
C005	Soto, Bush and Walker	residencial	PacÃ-fico	3846	
C006	Christopher Mccall	residencial	Oriente	16325	
C007	Nichols Group	comercial	Norte	948	
C008	Rebecca Cohen	industrial	Sur	18312	
C009	Ramos Inc	residencial	Norte	3051	
C010	Joshua Miles	comercial	Oriente	8924	

TABLE PROVEEDORES

	A	B	C	D	E	F
1	id_proveedor	nombre_proveedor	tipo_energia	region	capacidad_mw	
2	P001	Reed-Rodriguez Energi-a	nuclear	Centro	76	
3	P002	Rogers Group Energi-a	hidraulica	PacÃ-fico	903	
4	P003	Gordon and Sons Energi-a	nuclear	Norte	351	
5	P004	Turner-Wilson Energi-a	termica	Norte	737	
6	P005	Anderson-Medina Energi-a	termica	Norte	456	
7	P006	Barber, Smith and McClure Energi-a	termica	Oriente	774	
8	P007	Jones, Berry and Manning Energi-a	nuclear	PacÃ-fico	286	
9	P008	Young, Flynn and Dalton Energi-a	eolica	Oriente	817	
10	P009	Davis-Perez Energi-a	nuclear	PacÃ-fico	841	
11	P010	Carter LLC Energi-a	hidraulica	PacÃ-fico	619	

TABLA TRANSACCIONES

	A	B	C	D	E	F	G	H
	id_transaccion	fecha	id_cliente	id_proveedor	energia_kwh	precio_total_usd	tipo_transaccion	
	T001	14/07/2025	C081	P040	3096	557.28	venta	
	T002	14/07/2025	C092	P048	10805	1404.65	compra	
	T003	14/07/2025	C069	P066	10034	1505.1	venta	
	T004	13/07/2025	C025	P094	15043	2406.88	compra	
	T005	14/07/2025	C059	P100	15641	1876.92	venta	
	T006	14/07/2025	C064	P089	8245	1236.75	compra	
	T007	13/07/2025	C028	P091	16127	2580.32	venta	
	T008	13/07/2025	C079	P038	4591	688.65	compra	
0	T009	13/07/2025	C043	P048	3994	479.28	venta	
1	T010	13/07/2025	C084	P007	12008	1801.2	compra	

Estrategia de Data Lake en AWS S3

Inicialmente se crea los Bucket como se relaciona a continuación

aws

Buscar

[Alt+S]

Estados Unidos (Ohio)

Carlos Alfaro

Amazon S3

Buckets

Crear bucket

Configuración general

Región de AWS

EE.UU. Este (Ohio) us-east-2

Tipo de bucket

Información

☒ Uso general

Recomendado para la mayoría de los casos de uso y patrones de acceso. Los buckets de uso general son del tipo de bucket de S3 original. Permiten una combinación de clases de almacenamiento que almacenan objetos de forma redundante en múltiples zonas de disponibilidad.

☐ Directorio

Recomendado para casos de uso de baja latencia. Estos buckets utilizan únicamente la clase de almacenamiento S3 Express One Zone, que proporciona un procesamiento más rápido de los datos dentro de una única zona de disponibilidad.

Nombre del bucket

Información

datalake-energia-carlos

Los nombres de los buckets deben tener entre 3 y 63 caracteres y ser únicos dentro del espacio de nombres global. Los nombres de los buckets también deben empezar y terminar con una letra o un número. Los caracteres válidos son a-z, 0-9, puntos (.) y guiones (-). [Más información](#)

Copiar la configuración del bucket existente: *opcional*

Solo se copia la configuración del bucket en los siguientes ajustes.

Elegir el bucket

Formato: s3://bucket/prefijo

Amazon S3 > Buckets > Crear bucket

Propiedad de objetos Información

Controle la propiedad de los objetos escritos en este bucket desde otras cuentas de AWS y el uso de listas de control de acceso (ACL). La propiedad de los objetos determina quién puede especificar el acceso a los objetos.

ACL deshabilitadas (recomendado)

Todos los objetos de este bucket son propiedad de esta cuenta. El acceso a este bucket y sus objetos se especifica solo mediante políticas.

ACL habilitadas

Los objetos de este bucket pueden ser propiedad de otras cuentas de AWS. El acceso a este bucket y sus objetos se puede especificar mediante ACL.

Propiedad del objeto

Aplicada al propietario del bucket

Amazon S3 > Buckets > Crear bucket

Configuración de bloqueo de acceso público para este bucket

Se concede acceso público a los buckets y objetos a través de listas de control de acceso (ACL), políticas de bucket, políticas de puntos de acceso o todas las anteriores. A fin de garantizar que se bloquee el acceso público a todos sus buckets y objetos, active Bloquear todo el acceso público. Esta configuración se aplica exclusivamente a este bucket y a sus puntos de acceso. AWS recomienda activar Bloquear todo el acceso público, pero, antes de aplicar cualquiera de estos ajustes, asegúrese de que las aplicaciones funcionarán correctamente sin acceso público. Si necesita cierto nivel de acceso público a los buckets u objetos, puede personalizar la configuración individual a continuación para adaptarla a sus casos de uso de almacenamiento específicos. [Más información](#)

☐ Bloquear todo el acceso público

Activar esta configuración equivale a activar las cuatro opciones que aparecen a continuación. Cada uno de los siguientes ajustes son independientes entre sí.

☐ Bloquear el acceso público a buckets y objetos concedido a través de nuevas listas de control de acceso (ACL)

S3 bloqueará los permisos de acceso público aplicados a objetos o buckets agregados recientemente, y evitará la creación de nuevas ACL de acceso público para buckets y objetos existentes. Esta configuración no cambia los permisos existentes que permiten acceso público a los recursos de S3 mediante ACL.

☐ Bloquear el acceso público a buckets y objetos concedido a través de cualquier lista de control de acceso (ACL)

S3 ignorará todas las ACL que conceden acceso público a buckets y objetos.

☐ Bloquear el acceso público a buckets y objetos concedido a través de políticas de bucket y puntos de acceso públicas nuevas

S3 bloqueará las nuevas políticas de buckets y puntos de acceso que concedan acceso público a buckets y objetos. Esta configuración no afecta a las políticas ya existentes que permiten acceso público a los recursos de S3.

☐ Bloquear el acceso público y entre cuentas a buckets y objetos concedido a través de cualquier política de bucket y puntos de acceso pública

S3 ignorará el acceso público y entre cuentas en el caso de buckets o puntos de acceso que tengan políticas que concedan acceso público a buckets y objetos.

Amazon S3

Buckets de uso general

Todas las regiones de AWS

Buckets de directorio

Buckets de uso general (3) Información

Los buckets son contenedores de datos almacenados en S3.

Buscar buckets por nombre

< 1 >

	Nombre	Región de AWS	Fecha de creación
<input type="radio"/>	aws-glue-assets-016442247674-us-east-2	EE.UU. Este (Ohio) us-east-2	29 Jul 2025 1:45:33 PM -05
<input type="radio"/>	datalake-energia-carlos	EE.UU. Este (Ohio) us-east-2	27 Jul 2025 5:22:51 PM -05
<input type="radio"/>	datawarehouse-energia-carlos	EE.UU. Este (Ohio) us-east-2	29 Jul 2025 11:10:30 AM -05

Instantánea de la cuenta Información

Actualizado a diario

Storage Lens ofrece visibilidad sobre el uso del almacenamiento y las tendencias de actividad.

Ver panel

Resumen de acceso externo: nuevo Información

Actualizado a diario

Los resultados de acceso externo le ayudan a identificar los permisos de los buckets que permiten el acceso público o desde otras cuentas de AWS.

Cargue de información en repositorios

ⓘ Después de salir de esta página, la siguiente información ya no estará disponible.

Resumen

Destino

s3://datalake-energia-carlos/bronze/clientes/

Realizado correctamente

✔ 1 archivo, 4.5 KB (100.00%)

Con errores

⌚ 0 archivos, 0 B (0%)

Archivos y carpetas

Configuración

Archivos y carpetas (1 total, 4.5 KB)

🔍 Buscar por nombre

< 1 >

Nombre	Carpeta	Tipo	Tamaño	Estado	Error
clientes.csv	-	text/csv	4.5 KB	✔ Realizado correctamente...	-

Cerrar

ⓘ Después de salir de esta página, la siguiente información ya no estará disponible.

Resumen

Destino

s3://datalake-energia-carlos/bronze/proveedores/

Realizado correctamente

✔ 1 archivo, 5.0 KB (100.00%)

Con errores

⌚ 0 archivos, 0 B (0%)

Archivos y carpetas

Configuración

Archivos y carpetas (1 total, 5.0 KB)

🔍 Buscar por nombre

< 1 >

Nombre	Carpeta	Tipo	Tamaño	Estado	Error
proveedores.csv	-	text/csv	5.0 KB	✔ Realizado correctamente...	-

ⓘ Después de salir de esta página, la siguiente información ya no estará disponible.

Resumen

Destino

s3://datalake-energia-carlos/bronze/transacciones/

Realizado correctamente

✔ 1 archivo, 3.9 KB (100.00%)

Con errores

⌚ 0 archivos, 0 B (0%)

Archivos y carpetas

Configuración

Archivos y carpetas (1 total, 3.9 KB)

🔍 Buscar por nombre

< 1 >

Nombre	Carpeta	Tipo	Tamaño	Estado	Error
transacciones.csv	-	text/csv	3.9 KB	✔ Realizado correctamente...	-

Creo carpetas en buckets **datalake-energia-carlos**

Amazon S3 Buckets datalake-energia-carlos

Objetos Metadatos Propiedades Permisos Métricas Administración Puntos de acceso

Objetos (5)

Crear carpeta Cargar Copiar URI de S3 Copiar URL Descargar Abrir Eliminar Acciones

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

	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	bronze/	Carpeta	-	-	-
<input type="checkbox"/>	oro/	Carpeta	-	-	-
<input type="checkbox"/>	plata/	Carpeta	-	-	-
<input type="checkbox"/>	raw/	Carpeta	-	-	-
<input type="checkbox"/>	Unsaved/	Carpeta	-	-	-

© 2025, Amazon Web Services, Inc. o sus filiales. Privacidad Términos Preferencias de cookies

Amazon S3 Buckets datalake-energia-carlos raw/

Objetos Metadatos Propiedades Permisos Métricas Administración Puntos de acceso

Objetos (3)

Crear carpeta Cargar Copiar URI de S3 Copiar URL Descargar Abrir Eliminar Acciones

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

	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	clientes.csv	csv	28 Jul 2025 8:12:36 PM -05	4.5 KB	Estándar
<input type="checkbox"/>	proveedores.csv	csv	28 Jul 2025 8:12:37 PM -05	5.0 KB	Estándar
<input type="checkbox"/>	transacciones.csv	csv	28 Jul 2025 8:12:36 PM -05	4.6 KB	Estándar

Me dispongo a la creación de crawler para identificar el servicio origen y destino

AWS Glue Crawlers

Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)

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 (2) Info Last updated (UTC) July 29, 2025 at 19:11:34 Action Run Create crawler

View and manage all available crawlers.

Filter crawlers

	Name	State	Schedule	Last run	Last run ti...	Log	Table change...
<input type="checkbox"/>	crawlerdataw...	Ready		Succeeded	July 29, 2025 ...	View log	1 created
<input type="checkbox"/>	crawlerlz	Ready		Succeeded	July 29, 2025 ...	View log	6 created

Creación bases de datos destino en catalogo glu

The screenshot shows the AWS Glue Databases console. A notification banner at the top reads: "Announcing new optimization features for Apache Iceberg tables. Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)". Below the banner, the "Databases (2)" section is displayed. It includes a description: "A database is a set of associated table definitions, organized into a logical group." and a "Last updated (UTC)" timestamp of "July 29, 2025 at 19:09:00". There are buttons for "Edit", "Delete", and "Add database". A search bar labeled "Filter databases" is present. Below, a table lists the databases:

<input type="checkbox"/>	Name	Description	Location URI	Created on (UTC)
<input type="checkbox"/>	dbbronze	base de datos catalogo glu	-	July 29, 2025 at 16:28:57
<input type="checkbox"/>	dbdatawh	-	-	July 29, 2025 at 19:01:52

Se procede a ejecutar el crawler de origen

The screenshot shows the AWS Glue Crawlers console. A notification banner at the top reads: "Announcing new optimization features for Apache Iceberg tables. Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)". Below the banner, the "Crawlers (2)" section is displayed. It includes a description: "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." and a "Last updated (UTC)" timestamp of "July 29, 2025 at 20:54:48". There are buttons for "Action", "Run", and "Create crawler". A search bar labeled "Filter crawlers" is present. Below, a table lists the crawlers:

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run ti...	Log	Table change...
<input type="checkbox"/>	crawlerdataw...	Ready		Succeeded	July 29, 2025 ...	View log	1 created
<input type="checkbox"/>	crawlerlz	Ready		Succeeded	July 29, 2025 ...	View log	6 created

Below the crawlers table, there are tabs for "Crawler runs", "Schedule", "Data sources", "Classifiers", and "Tags". The "Crawler runs (1)" section is active, showing a description: "The list of crawler runs for this crawler." and buttons for "Stop run", "View CloudWatch logs", and "View run details". A search bar labeled "Filter data" and a filter button "Filter by a date and time range" are present. Below, a table lists the crawler runs:

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours	Table changes
July 29, 2025 at 19:02:39	July 29, 2025 at 19:03:24	44 s	Completed	0.133	1 table change, 0 partition changes

Se consulto en base de datos creada

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Announcing new optimization features for Apache Iceberg tables

Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)

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)
July 29, 2025 at 16:41:54

Delete

Add tables using crawler

Add table

View and manage all available tables.

Filter tables

< 1 >

<input type="checkbox"/>	Name	Database	Location	Classific...	Depreca...	View data	Data quality	Column st...
<input type="checkbox"/>	clientes	dbbronce	s3://datalake-er	CSV	-	Table data	View data qualit	View statistics
<input type="checkbox"/>	clientes_csv	dbbronce	s3://datalake-er	CSV	-	Table data	View data qualit	View statistics
<input type="checkbox"/>	proveedores	dbbronce	s3://datalake-er	CSV	-	Table data	View data qualit	View statistics
<input type="checkbox"/>	proveedores_csv	dbbronce	s3://datalake-er	CSV	-	Table data	View data qualit	View statistics
<input type="checkbox"/>	transacciones	dbbronce	s3://datalake-er	CSV	-	Table data	View data qualit	View statistics
<input type="checkbox"/>	transacciones_cs	dbbronce	s3://datalake-er	CSV	-	Table data	View data qualit	View statistics

Consulta en athena

Amazon Athena

Editor de consultas

Editor

Consultas recientes

Consultas guardadas

Configuración

Grupo de trabajo

primary

Antes de ejecutar la primera consulta, debe configurar una ubicación del resultado de la consulta en Amazon S3.

Editar ajustes

Athena ahora admite sugerencias de código de escritura anticipada para acelerar el desarrollo de consultas SQL

Las sugerencias de escritura anticipada están activadas de forma predeterminada. Puede cambiar esta configuración en las preferencias del editor de consultas.

Editar preferencias

Datos

Origen de datos

AwsDataCatalog

Catálogo

Ningún elemento

Base de datos

dbbronce

Tablas y vistas

Crear

Consulta 1

1

aws

Buscar

[Alt+S]

Estados Unidos (Ohio)

Carlos Alfaro

Amazon Athena

Editor de consultas

Editor

Consultas recientes

Consultas guardadas

Configuración

Grupo de trabajo

primary

Athena ahora admite sugerencias de código de escritura anticipada para acelerar el desarrollo de consultas SQL

Las sugerencias de escritura anticipada están activadas de forma predeterminada. Puede cambiar esta configuración en las preferencias del editor de consultas.

Editar preferencias

Datos

Origen de datos

AwsDataCatalog

Catálogo

Ningún elemento

Base de datos

dbdatawh

Tablas y vistas

Crear

Consulta 1

Consulta 2

Consulta 3

Consulta 4

1

SELECT * FROM "dbbronce"."clientes" limit 10;

Amazon Athena > Editor de consultas

Tablas (1)

- datawarehouse_energia_carlos
 - id_cliente: string
 - id_transaccion: string
 - fecha: string
 - id_proveedor: string
 - energia_kwh: int
 - precio_total_usd: double
 - tipo_transaccion: string
 - price: float
 - nombre_cliente: string

Vistas (0)

SQL Ln 1, Col 46

Ejecutar de nuevo Explicar Cancelar Borrar

Crear

Volver a utilizar los resultados de la consulta hasta 60 minutos

Resultados de la consulta Estado de la consulta

Completado Tiempo en cola: 91 ms Tiempo de ejecución: 514 ms Datos analizados: 4.55 KB

Resultados (10) Copiar Descargar resultados en formato CSV

Filas de búsqueda

#	id_cliente	nombre_cliente	tipo_cliente	region	consumo_estimado_kwh
1	C001	Robert Smith	comercial	Oriente	11778
2	C002	Small PLC	comercial	Centro	10882
3	C003	Briggs Ltd	residencial	Oriente	7322

Amazon Athena > Editor de consultas

Editor Consultas recientes Consultas guardadas Configuración

Grupo de trabajo primary

Consultas recientes (10) Cancelar Descargar resultados en formato CSV Descargar tabla CSV

Buscar consultas recientes

ID de ejecución	Consulta	Hora de inicio	Estado	Tiempo
d8573009-46c0-4a90-8059-84ee3942583c	SELECT * FROM "dbdatawh"."datawarehouse_energia_carlos" li...	2025-07-29T14:13:02.166-05:...	Completado	73
5385e147-b4bc-441f-8da1-23d80040fe84	SELECT * FROM "dbbronze"."transacciones_csv" limit 10	2025-07-29T12:51:18.724-05:...	Completado	46
59e62286-3048-4ec4-9765-0a9fef91e7fa	SELECT * FROM "dbbronze"."transacciones" limit 10	2025-07-29T12:51:12.643-05:...	Completado	41
bb0ff7bf-275e-4ea4-a079-67d5d7e80d15	SELECT * FROM "dbbronze"."clientes" limit 10	2025-07-29T12:26:43.263-05:...	Completado	51
f654d861-c34a-4cb0-bc2c-b39f0fd38140	SELECT * FROM "dbbronze"."clientes_csv" limit 10	2025-07-29T12:25:56.382-05:...	Completado	38
f1f53159-3a36-4ab4-9bb6-1b255b2505ba	SELECT * FROM "dbbronze"."clientes" limit 10	2025-07-29T12:25:44.894-05:...	Completado	44
19fb8d42-ba2c-4c43-85b4-bb248a457b25	SELECT * FROM "dbbronze"."clientes" limit 10	2025-07-29T12:25:10.866-05:...	Completado	50
2aaec165-d259-4891-9159-98f33c12c3f6	SELECT * FROM "dbbronze"."clientes_csv" limit 10	2025-07-29T12:24:20.854-05:...	Completado	33

Creación ETL

AWS Glue > Jobs

AWS Glue Studio

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 (0) Filter jobs by property

Job name	Type	Created by	Last modified	AWS Glue version	Action
No jobs					
You have not created a job yet.					
Create job from a blank graph					

etlenergia

Last modified on 29/7/2025, 1:45:31 p. m.

Actions

Save

Run

Script

Job details

Runs

Data quality

Schedules

Version Control

Script

Info

```
1 import sys
2 from awsglue.transforms import *
3 from awsglue.utils import getResolvedOptions
4 from pyspark.context import SparkContext
5 from awsglue.context import GlueContext
6 from awsglue.job import Job
7 from pyspark.sql.functions import col
8 from awsglue.dynamicframe import DynamicFrame
9
10
11 """
12 Este código lee desde el AWS Glue Datacatalog dos tablas en una base de datos, después realiza un join entre ambas
13 tablas y calcula el total de la venta.
14 Finalmente, escribe el resultado en un archivo Parquet en S3.
```

PythonLn 1, Col 1Errors: 0Warnings: 0

Ejecución de script

AWS Glue

Monitoring

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

Catalog settings

Data Integration and ETL

Monitoring

Info

Start date range

7 Day

Job runs summary

Total runs

1

Running

0

Canceled

0

Successful runs

1

Failed runs

0

Run success rate

100%

DPU hours

0

Job runs (1)

Info

Filter job runs by property

Job name

Run status

Type

Start time (Local)

End time (Local)

Run time

Capacity

Worker ty

etlenergia

Succeeded

Glue ETL

07/29/2025 13:47:45

07/29/2025 13:49:04

1 minute

?

G 1Y

AWS Glue

Jobs

AWS Glue

Getting started

ETL jobs

Visual ETL

Notebooks

Job run monitoring

Data Catalog tables

Data connections

Workflows (orchestration)

Zero-ETL integrations

Data Catalog

Databases

Tables

Stream schema registries

Schemas

Connections

Crawlers

Classifiers

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 (1)

Info

Filter jobs by property

Job name

Type

Created by

Last modified

AWS Glue version

Action

etlenergia

Glue ETL

Script

29/7/2025, 1:45:31 p. m.

5.0

-

Validación ejecución ETL exitosa en buckets

Amazon S3 > Buckets > datawarehouse-energia-carlos

Objetos (4)

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

<input type="checkbox"/>	Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
<input type="checkbox"/>	part-00000-f3d53b6f-0799-4ab8-a514-64cd07fc5dfc-c000.snappy.parquet	parquet	29 Jul 2025 1:48:51 PM -05	5.3 KB	Estándar
<input type="checkbox"/>	part-00001-f3d53b6f-0799-4ab8-a514-64cd07fc5dfc-c000.snappy.parquet	parquet	29 Jul 2025 1:48:51 PM -05	5.1 KB	Estándar
<input type="checkbox"/>	part-00002-f3d53b6f-0799-4ab8-a514-64cd07fc5dfc-c000.snappy.parquet	parquet	29 Jul 2025 1:48:51 PM -05	5.3 KB	Estándar

Creación de crawler para detectar datos de destino

Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)

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 (2) Info
Last updated (UTC) July 29, 2025 at 19:11:34

View and manage all available crawlers.

Filter crawlers

<input type="checkbox"/>	Name	State	Schedule	Last run	Last run ti...	Log	Table change...
<input type="checkbox"/>	crawlerdataw...	Ready		Succeeded	July 29, 2025 ...	View log	1 created
<input type="checkbox"/>	crawlerlz	Ready		Succeeded	July 29, 2025 ...	View log	6 created

Creación db destino

Announcing new optimization features for Apache Iceberg tables
Optimize storage for Apache Iceberg tables with automatic snapshot retention and orphan file deletion. [Learn more](#)

Databases (2)
Last updated (UTC) July 29, 2025 at 19:09:00

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

Filter databases

<input type="checkbox"/>	Name	Description	Location URI	Created on (UTC)
<input type="checkbox"/>	dbbronze	base de datos catalogo glu	-	July 29, 2025 at 16:28:57
<input type="checkbox"/>	dbdatawh	-	-	July 29, 2025 at 19:01:52

Consulta la base destino

Amazon Athena > Editor de consultas

Origen de datos: AwsDataCatalog

Catálogo: Ningún elemento

Base de datos: dbdatawh

Tablas y vistas: [Crear](#)

SQL: `SELECT * FROM "dbdatawh"."datawarehouse_energia_carlos" limit 10;`

[Ejecutar de nuevo](#) [Explicar](#) [Cancelar](#) [Borrar](#) [Crear](#)

☐ Volver a utilizar los resultados de la consulta hasta hace 60 minutos

Resultados de la consulta Estado de la consulta

Completado Tiempo en cola: 110 ms Tiempo de ejecución: 731 ms Datos analizados: 4.08 KB

Información consultada y migrada con éxito se cambia el tipo de formato CSV a parquet

[Ejecutar de nuevo](#) [Explicar](#) [Cancelar](#) [Borrar](#) [Crear](#)

☐ Volver a utilizar los resultados de la consulta hasta hace 60 minutos

Resultados de la consulta Estado de la consulta

Completado Tiempo en cola: 110 ms Tiempo de ejecución: 731 ms Datos analizados: 4.08 KB

Resultados (10) [Copiar](#) [Descargar resultados en formato CSV](#)

Filas de búsqueda

#	id_cliente	id_transaccion	fecha	id_proveedor	energia_kwh	precio_total_usd
1	C004	T017	13/07/2025	P092	5429	814.35
2	C004	T083	14/07/2025	P004	8788	1142.44
3	C004	T098	13/07/2025	P054	19294	2894.1
4	C015	T038	14/07/2025	P083	3203	480.45
5	C016	T086	14/07/2025	P023	9350	1683.0

Se consulta la data en buckets destino verificando la información y tipo formato

aws

Buscar

[Alt+S]

Estados Unidos (Ohio)

Carlos Alfaro

Amazon S3

Buckets

datawarehouse-energia-carlos

Amazon S3

Buckets de uso general

Buckets de directorio

Buckets de tablas

Buckets vectoriales [Vista previa](#)

Concesiones de acceso

Puntos de acceso (buckets de uso general, sistemas de archivos FSx)

Puntos de acceso (buckets de directorio)

Puntos de acceso del objeto Lambda

Puntos de acceso de varias regiones

Operaciones por lotes

Analizador de acceso de IAM para S3

Objetos (4)

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"/>	part-00000-f3d53b6f-0799-4ab8-a514-64cd07fc5dfc-c000.snappy.parquet	parquet	29 Jul 2025 1:48:51 PM -05	5.3 KB	Estándar
<input type="checkbox"/>	part-00001-f3d53b6f-0799-4ab8-a514-64cd07fc5dfc-c000.snappy.parquet	parquet	29 Jul 2025 1:48:51 PM -05	5.1 KB	Estándar
<input type="checkbox"/>	part-00002-f3d53b6f-0799-4ab8-a514-64cd07fc5dfc-c000.snappy.parquet	parquet	29 Jul 2025 1:48:51 PM -05	5.3 KB	Estándar

