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

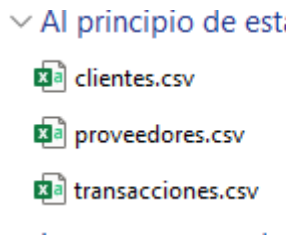
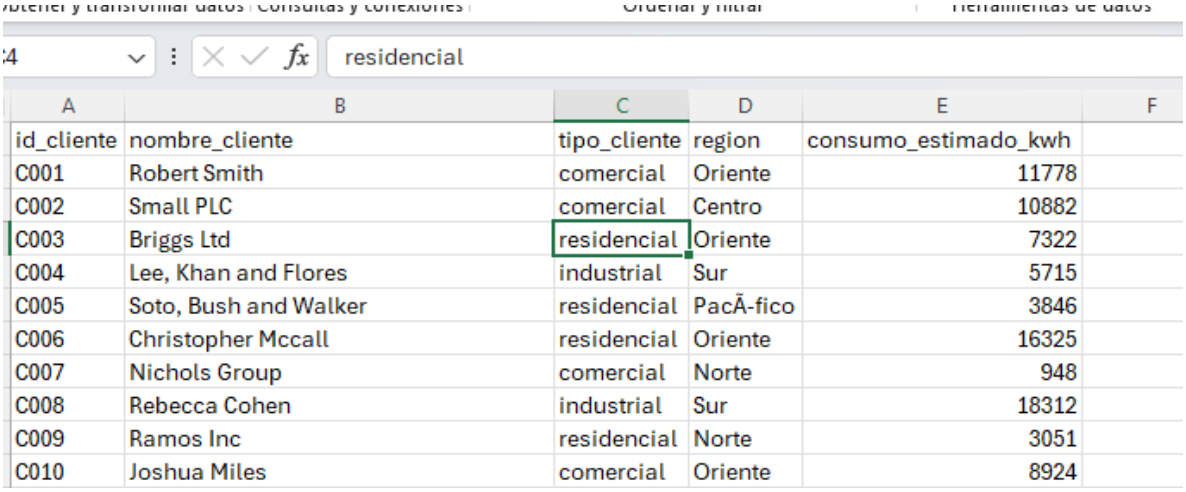


TABLA CLIENTES



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

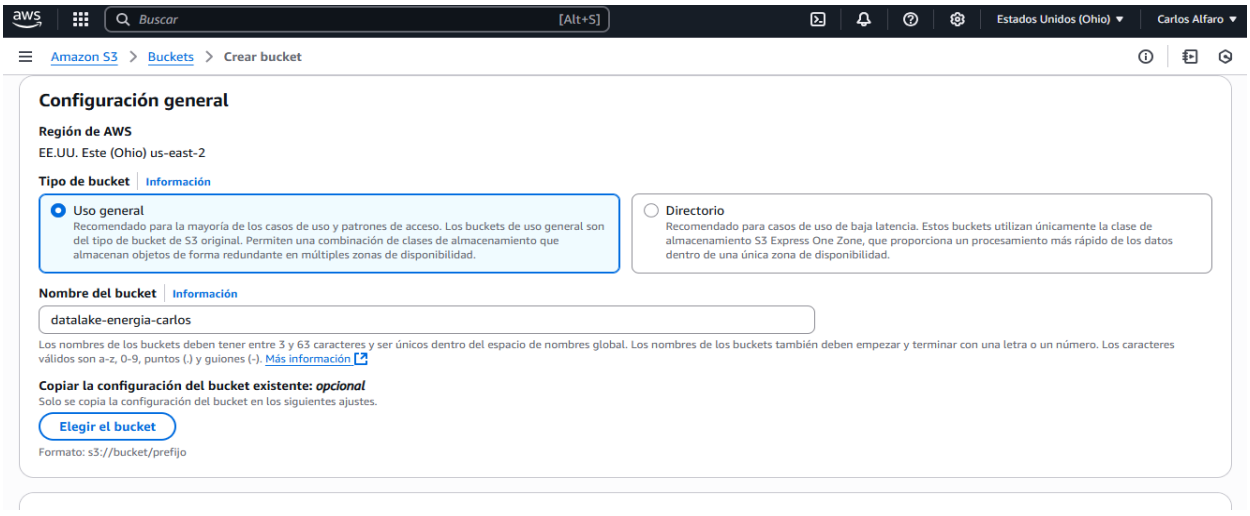
	A	B	C	D	E	F
1	id_proveedor	nombre_proveedor	tipo_energia	region	capacidad_mw	
2	P001	Reed-Rodriguez Energi-a	nuclear	Centro	76	
3	P002	Rogers Group Energi-a	hidraulica	PacÃ-fico	903	
4	P003	Gordon and Sons Energi-a	nuclear	Norte	351	
5	P004	Turner-Wilson Energi-a	termica	Norte	737	
6	P005	Anderson-Medina 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	

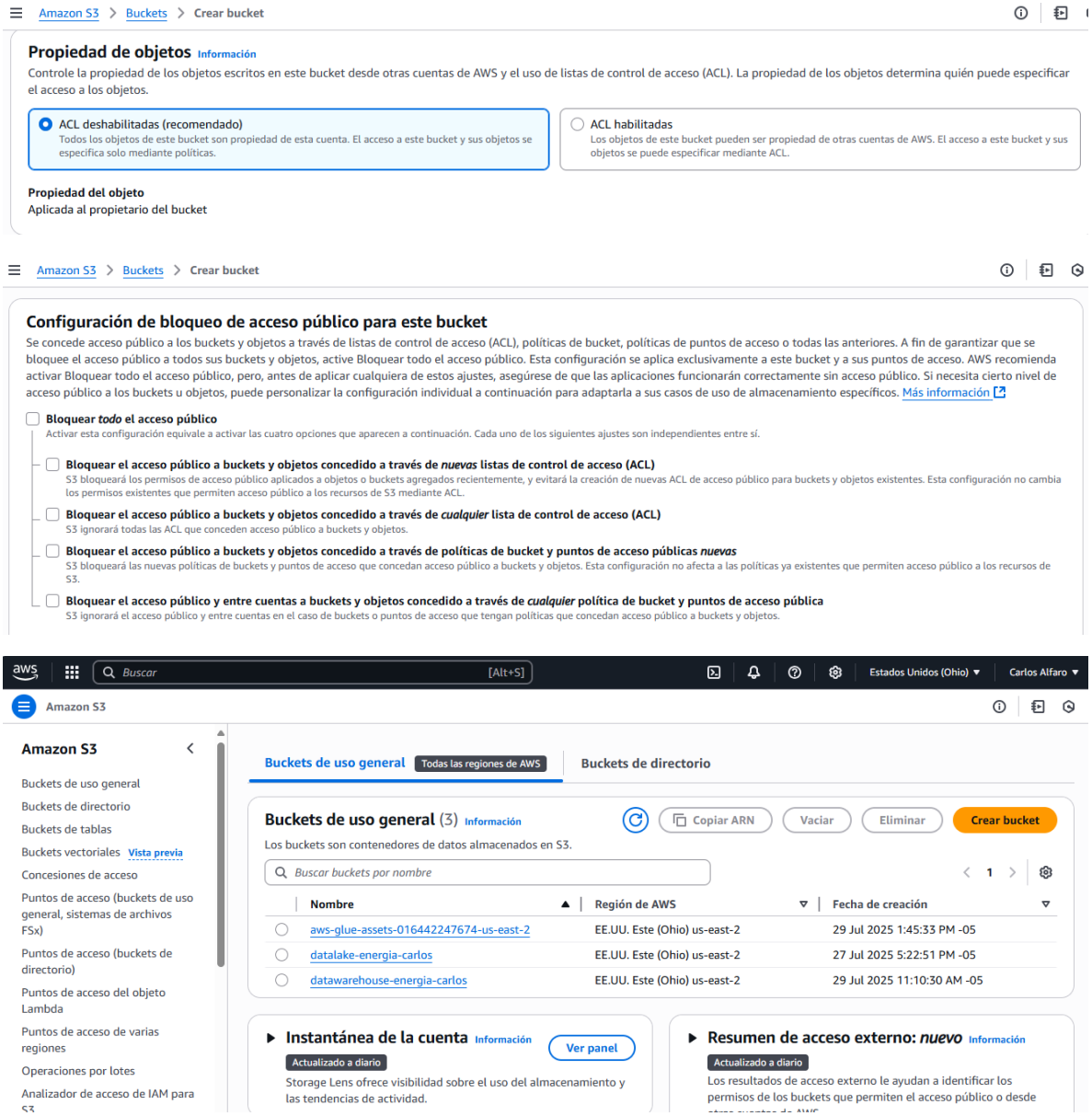
TABLA TRANSACCIONES

	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	
	T009	13/07/2025	C043	P048	3994	479.28	venta	
	T010	13/07/2025	C084	P007	12008	1801.2	compra	

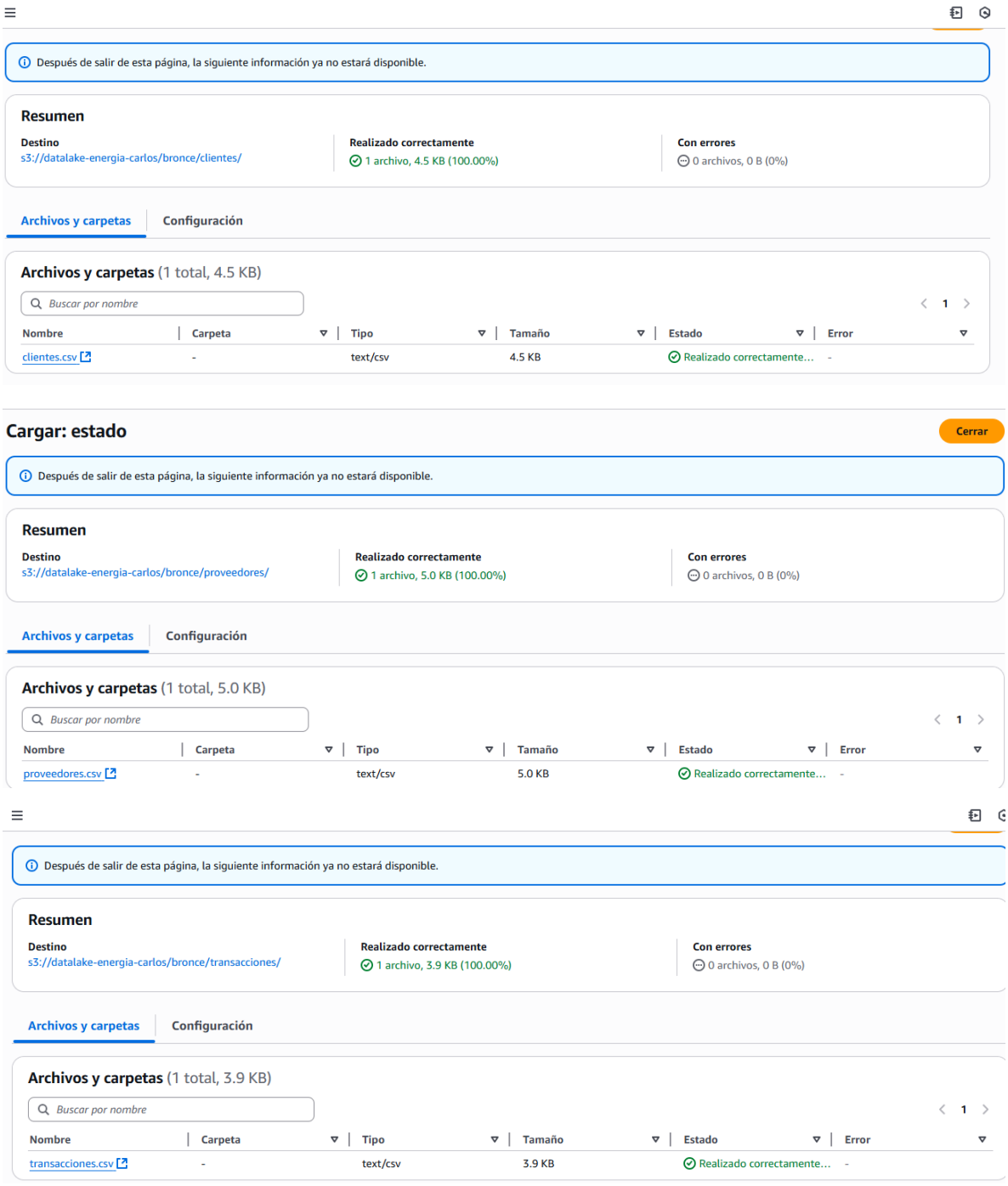
Estrategia de Data Lake en AWS S3

Inicialmente se crea los Bucket como se relaciona a continuación

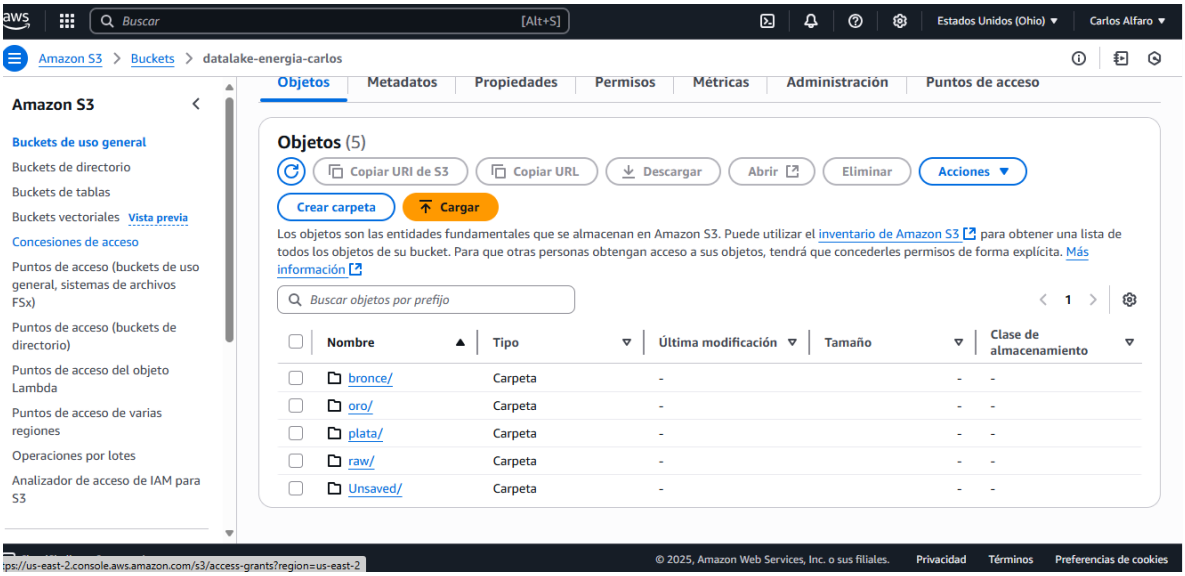


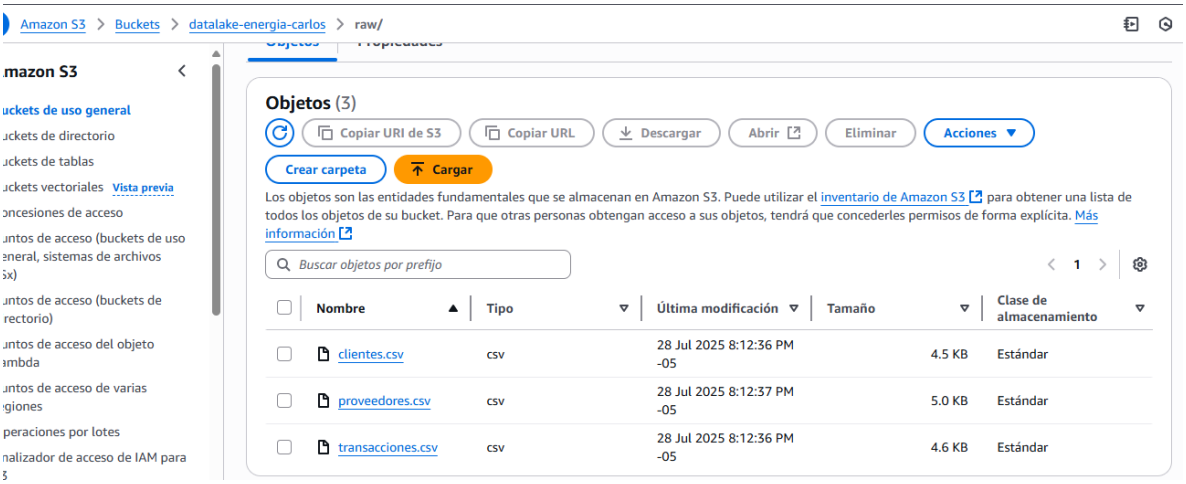


Cargue de información en repositorios

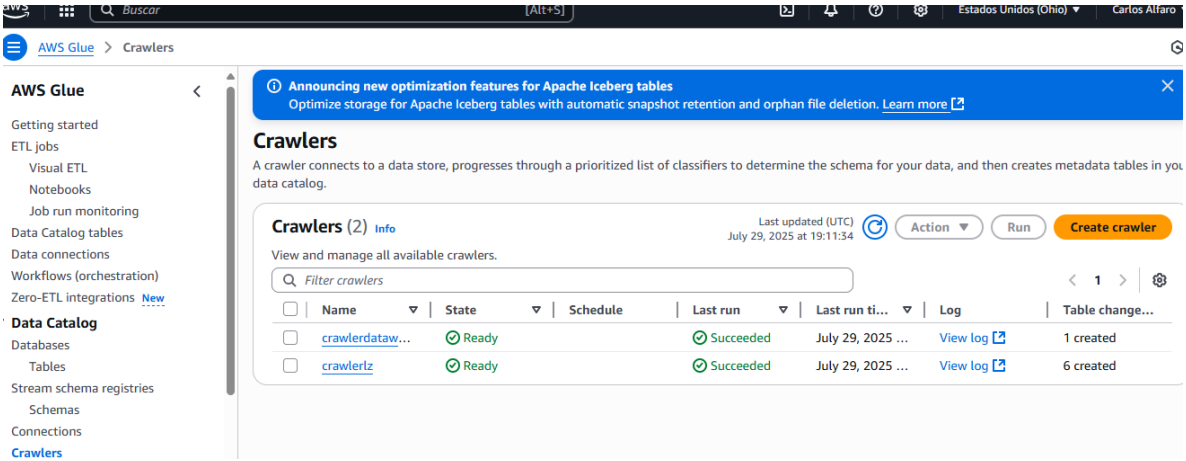


Creo carpetas en buckets **datalake-energia-carlos**

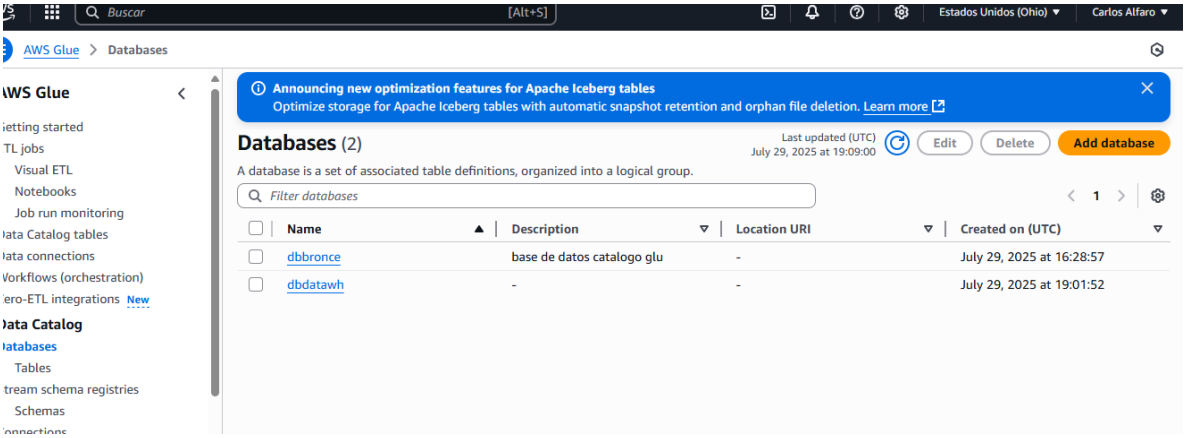




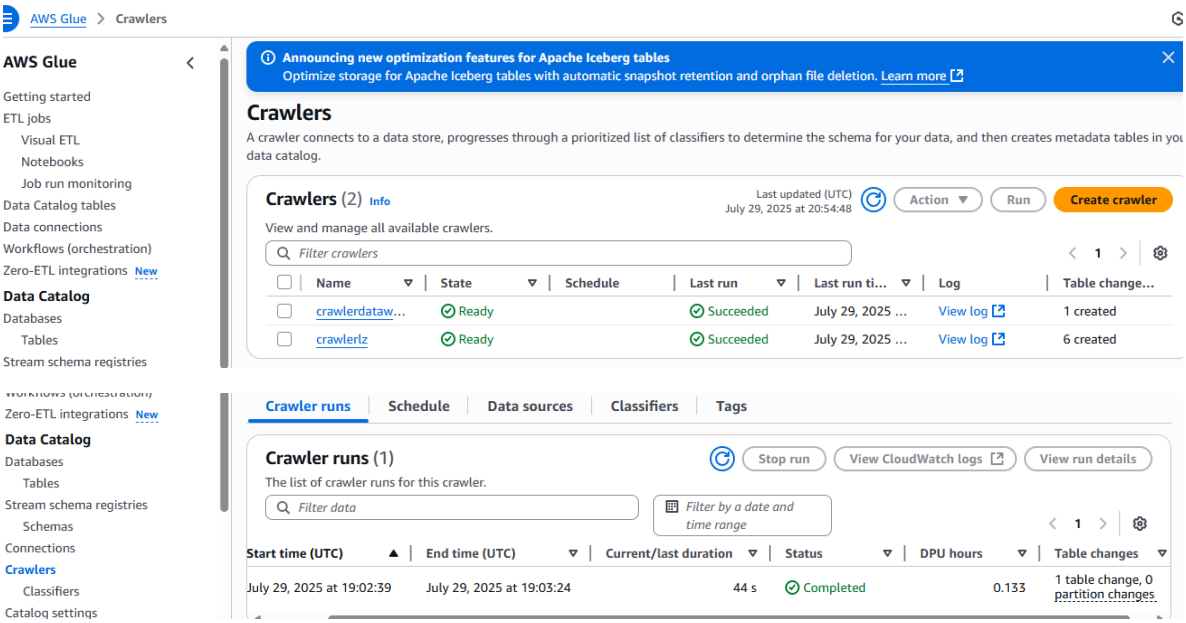
Me dispongo a la creación de crawler para identificar el servicio origen y destino



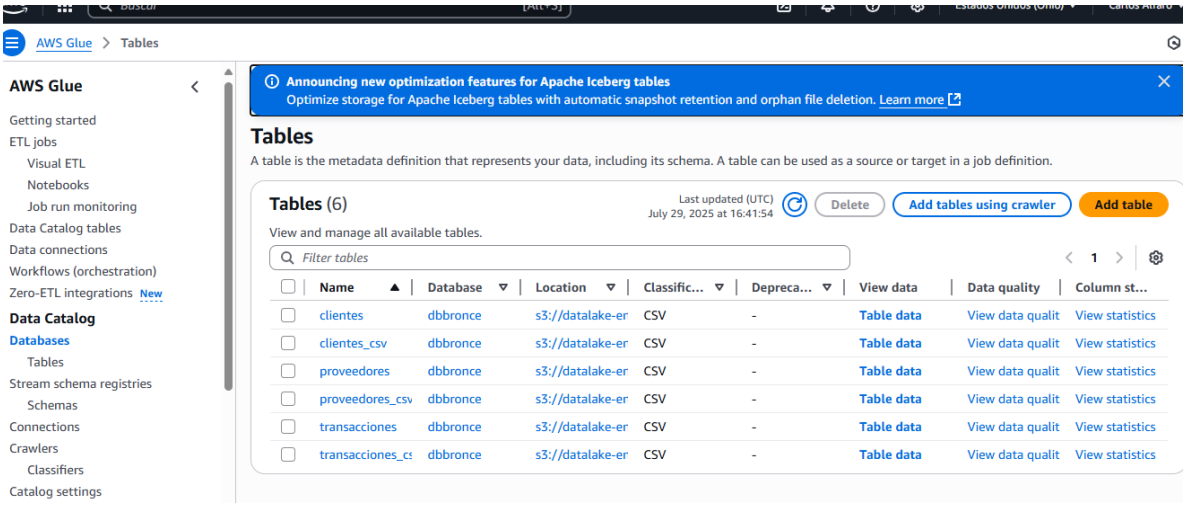
Creación bases de datos destino en catalogo glu



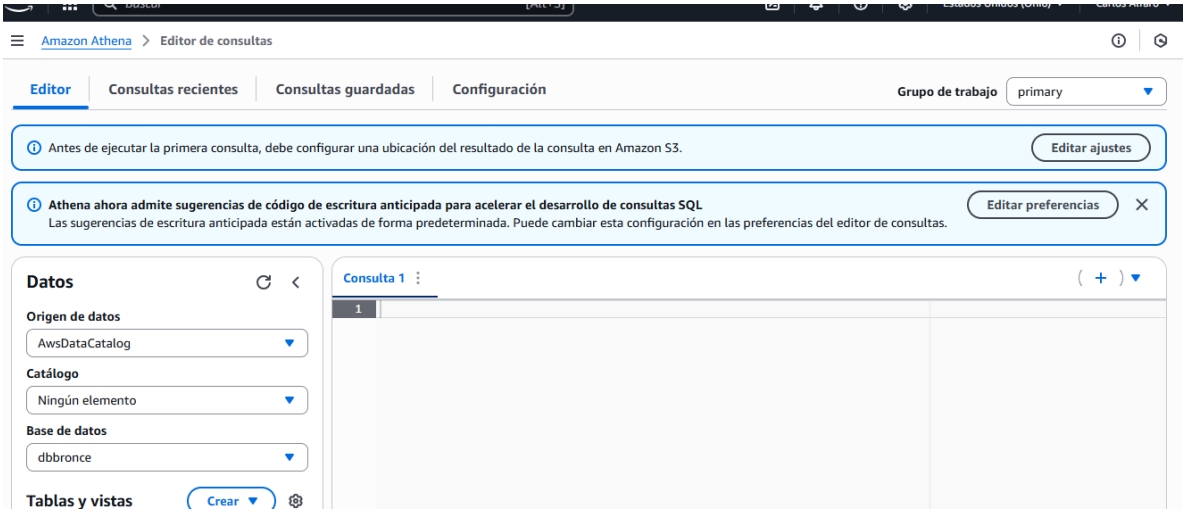
Se procede a ejecutar el crawler de origen

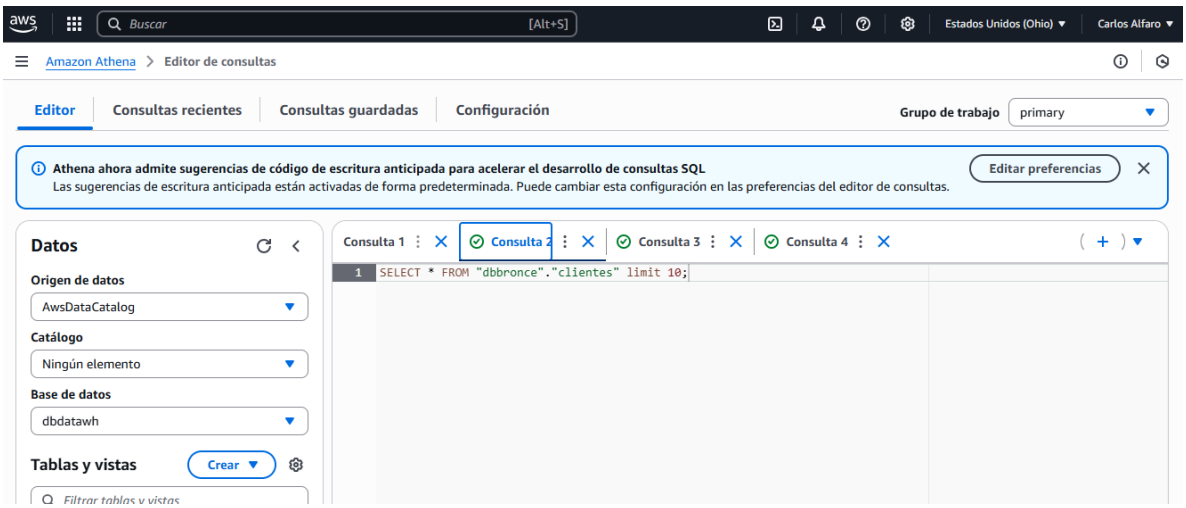


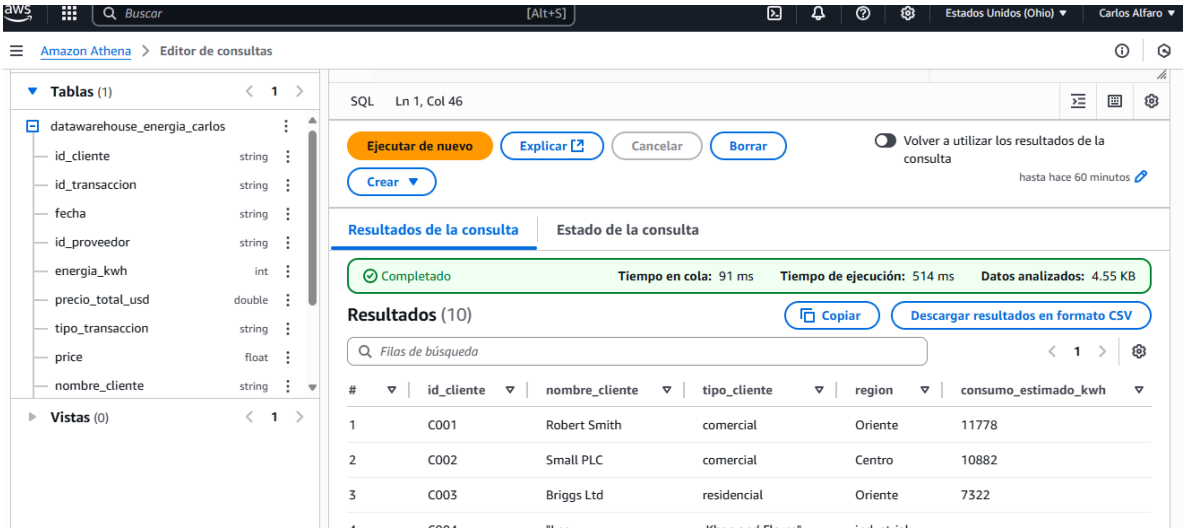
Se consulto en base de datos creada

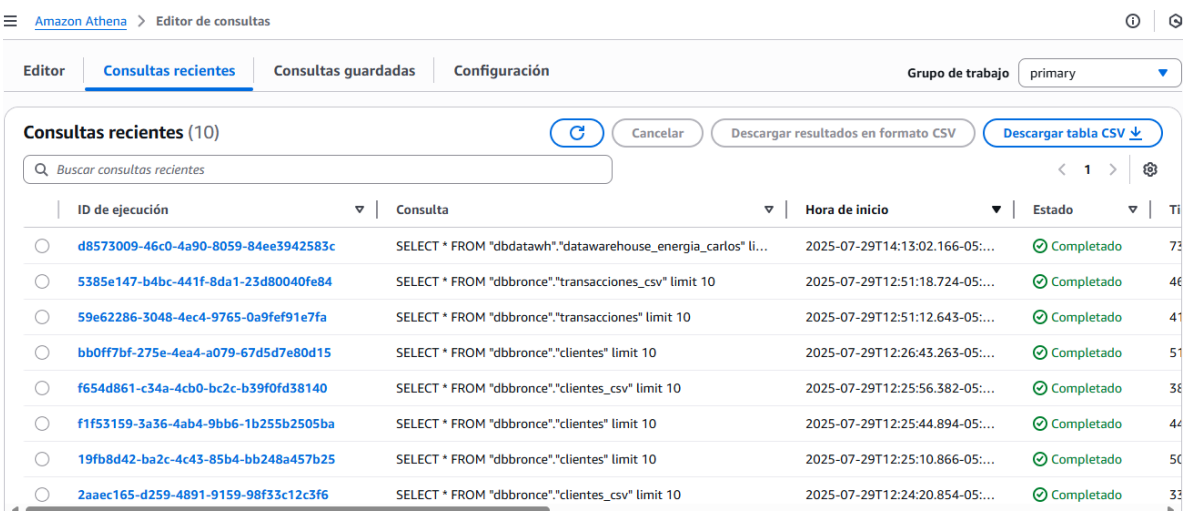


Consulta en athena

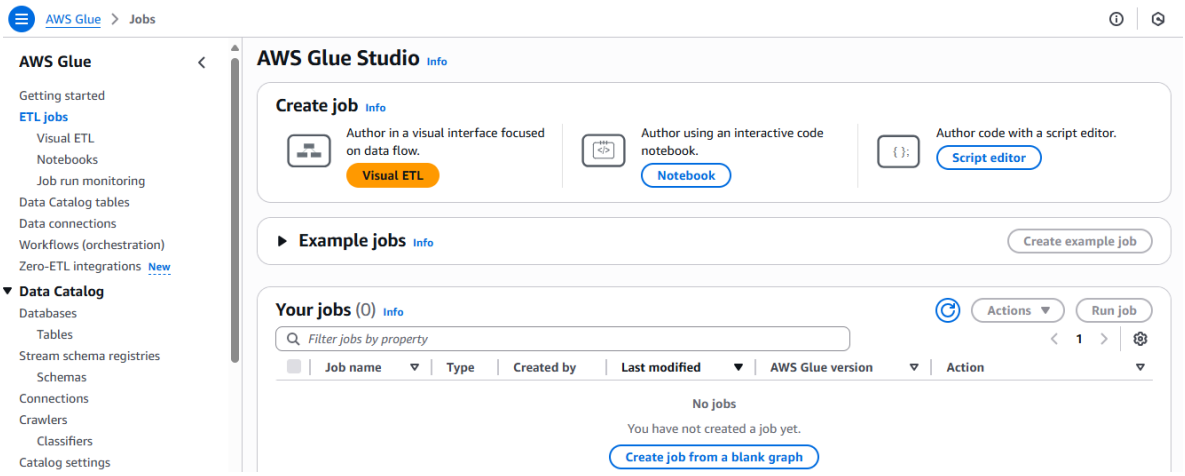


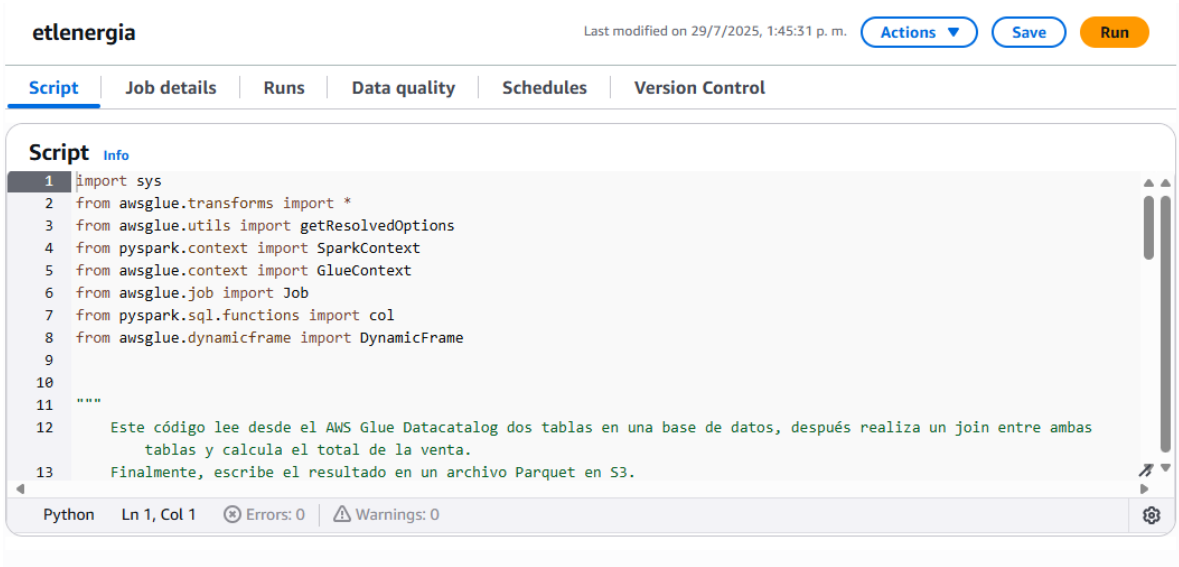




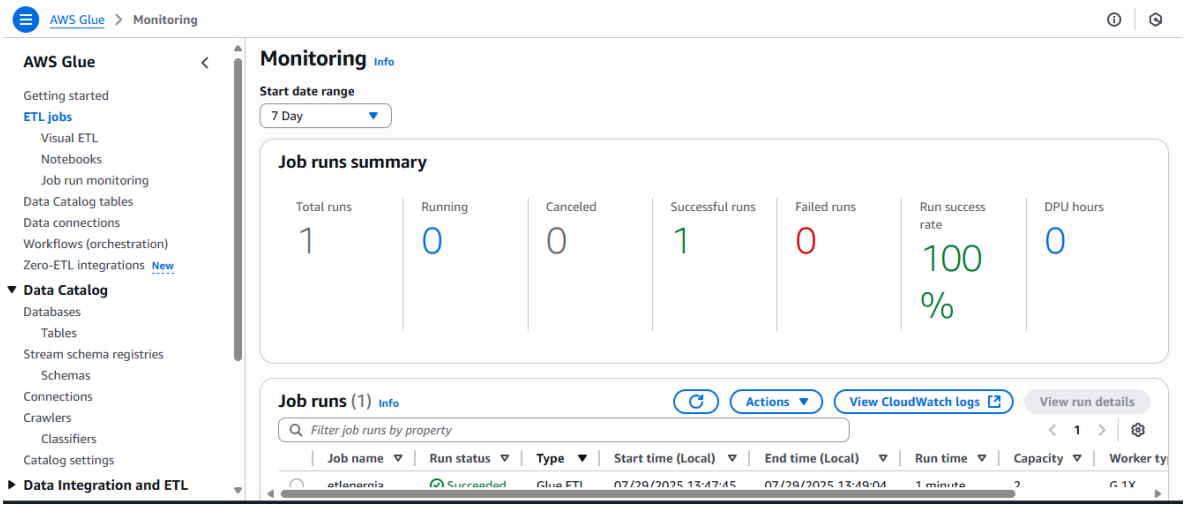


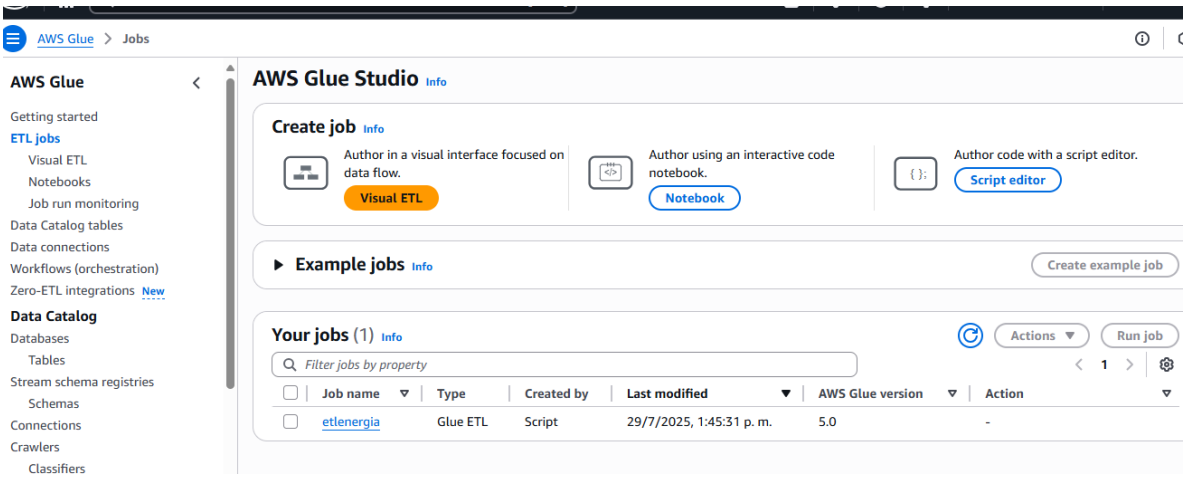
Creación ETL



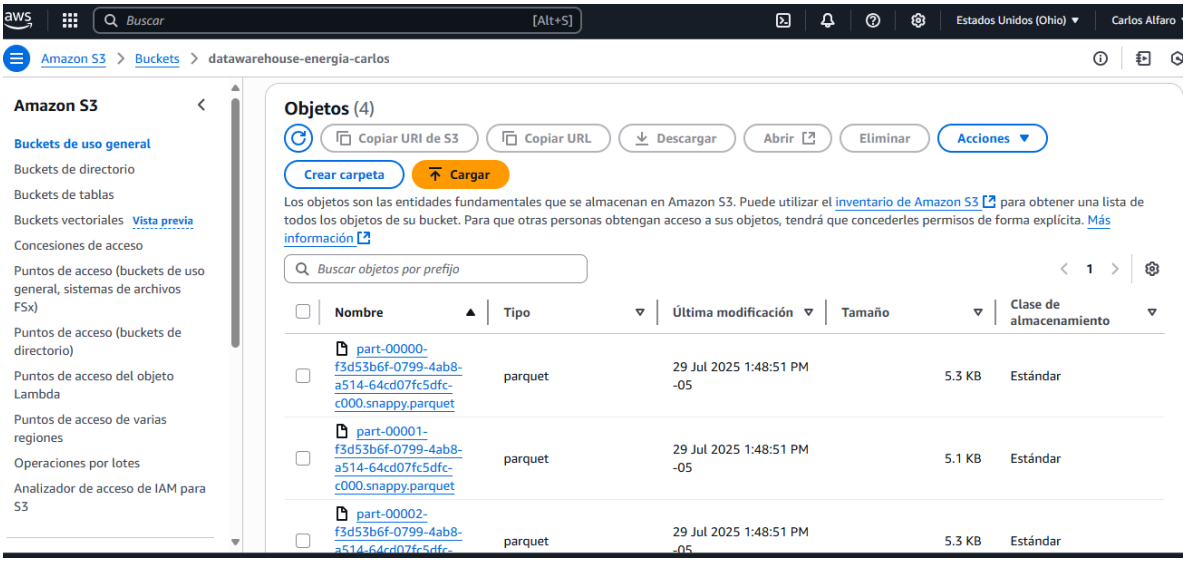


Ejecución de script

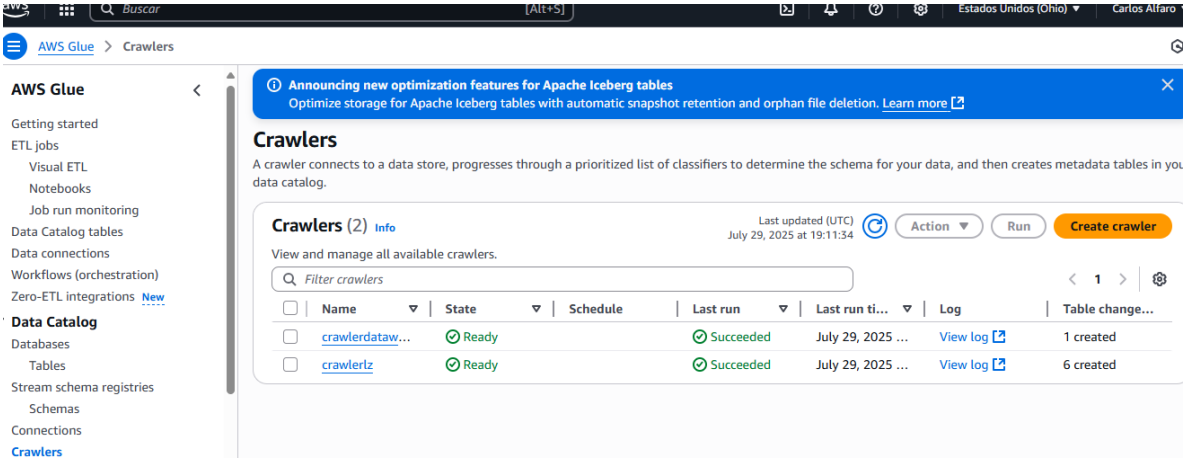




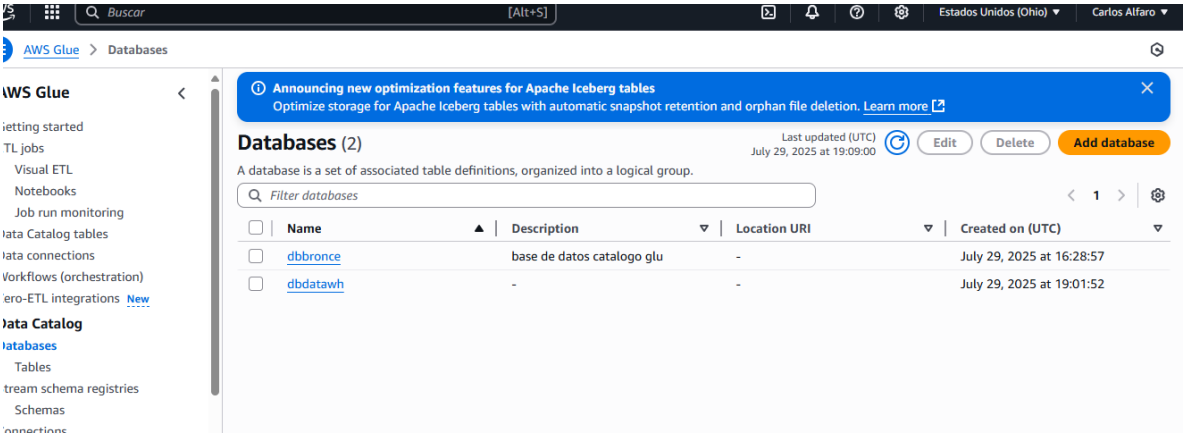
Validación ejecución ETL exitosa en buckets



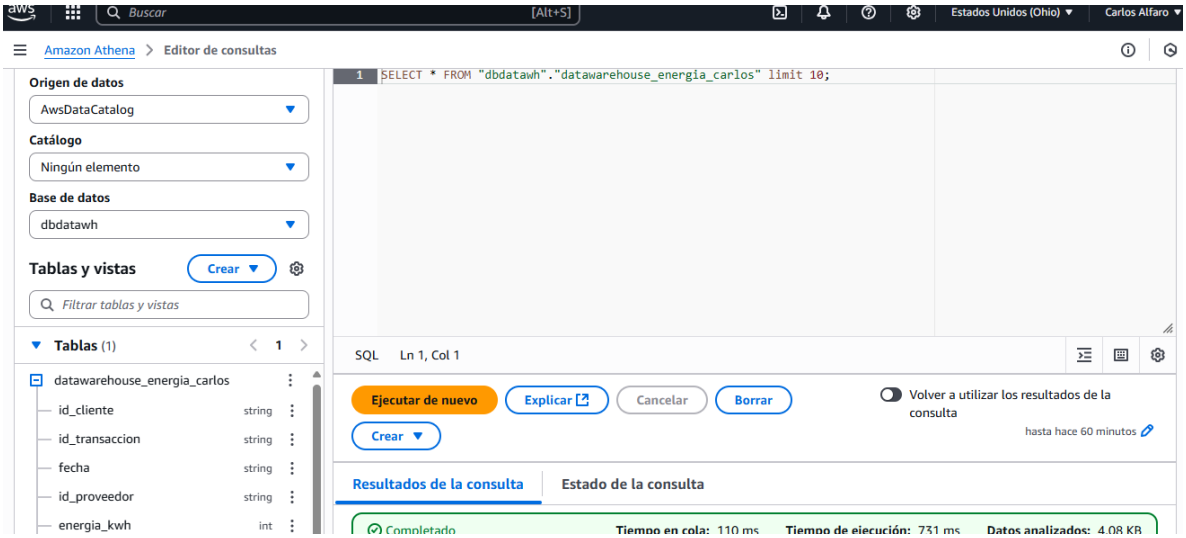
Creación de crawler para detectar datos de destino



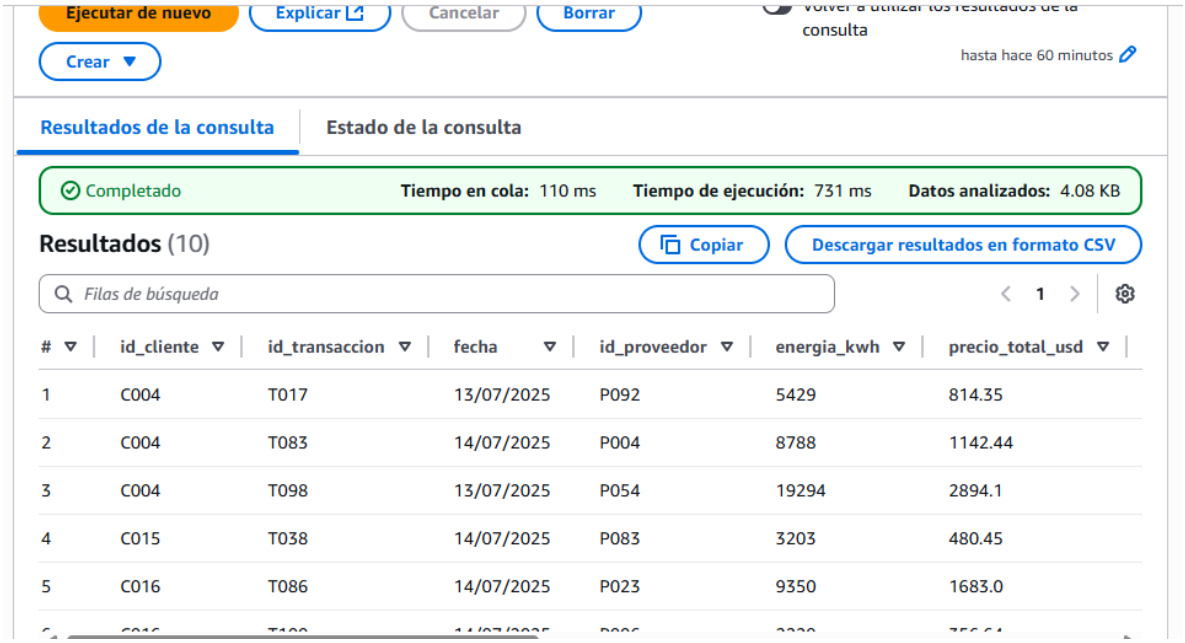
Crecion db destino



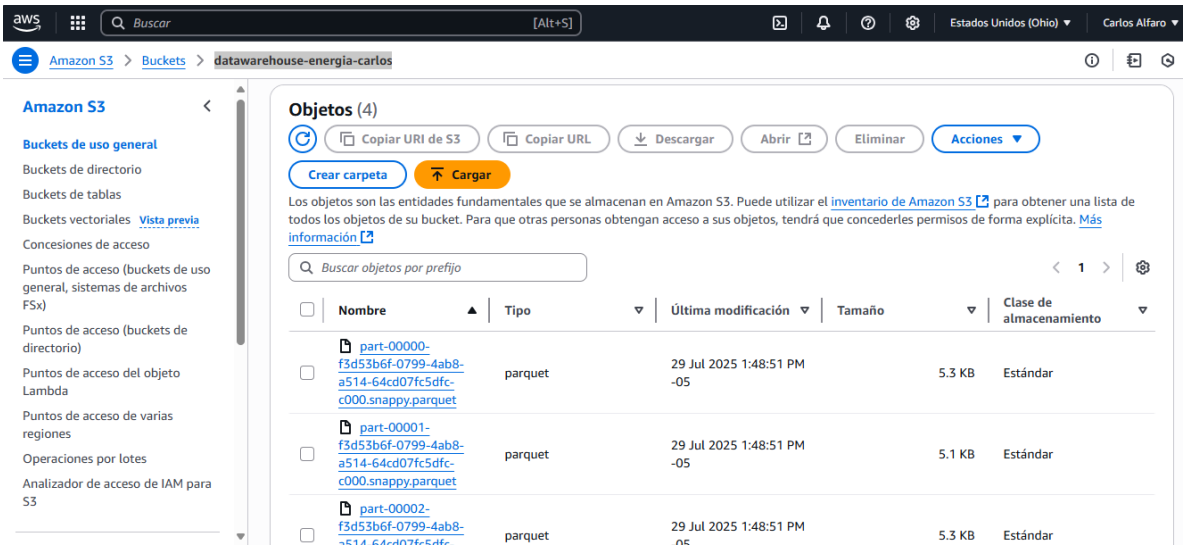
Consulto la base destino



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



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