

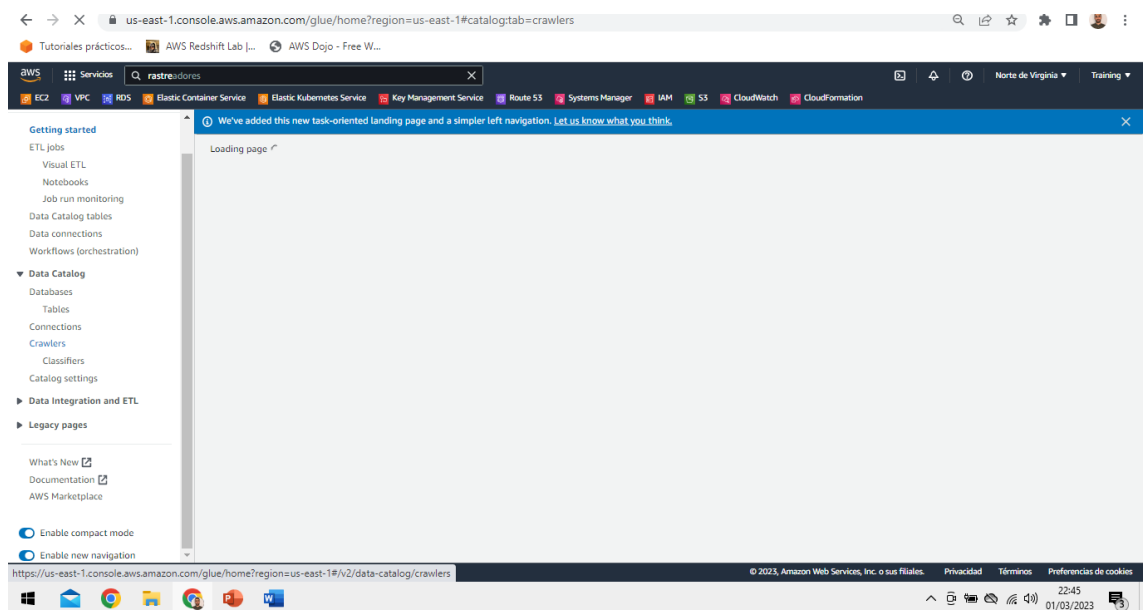
Consulta de datos dentro de un depósito S3 con AWS Glue y Amazon Athena

En este laboratorio, vamos a usar un conjunto de datos de vuelos de muestra que se almacenan en un depósito público de S3 para crear una tabla de AWS Glue.

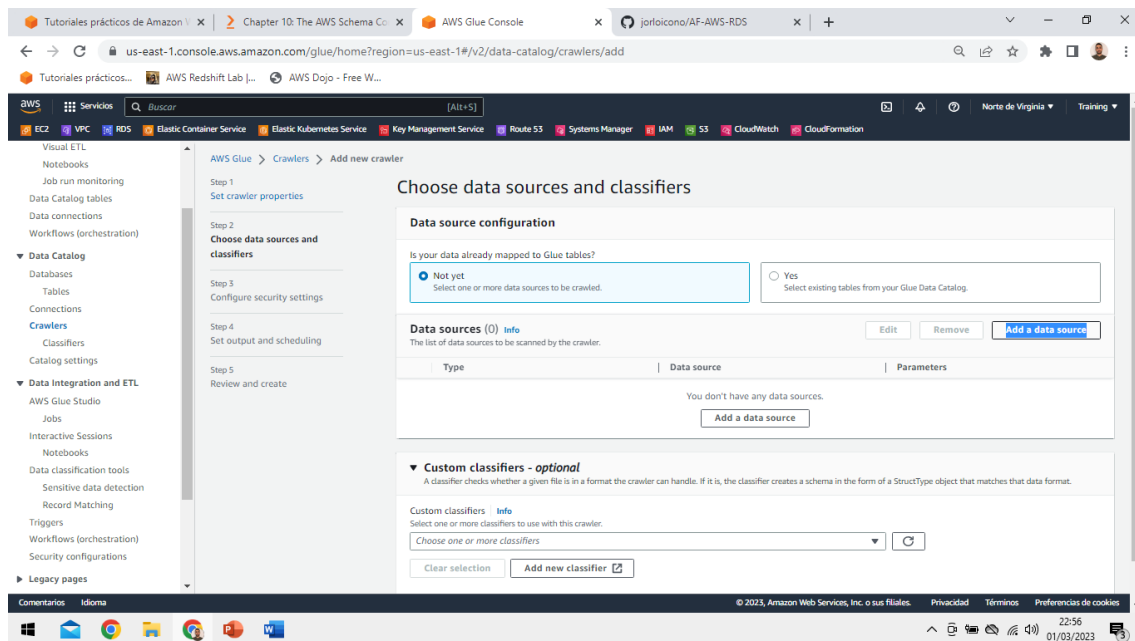
Luego, ejecutaremos consultas en esa tabla de AWS Glue para obtener información sobre vuelos. Empecemos:

Inicie sesión en la consola de AWS y navegue hasta **AWS Glue**.

Haga clic en **Rastreadores (Crawler)** en el menú principal de la izquierda y luego haga clic en **Agregar rastreador**



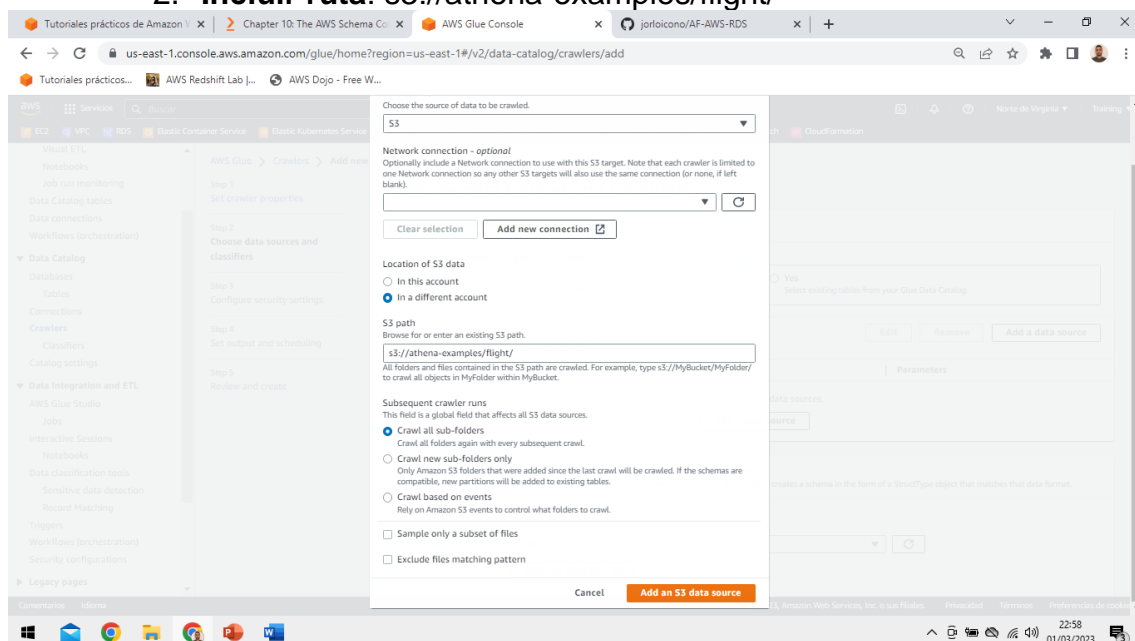
Introduzca su nombre como **Nombre del rastreador** y haga clic en **Siguiente**



Pulsa en Add data source

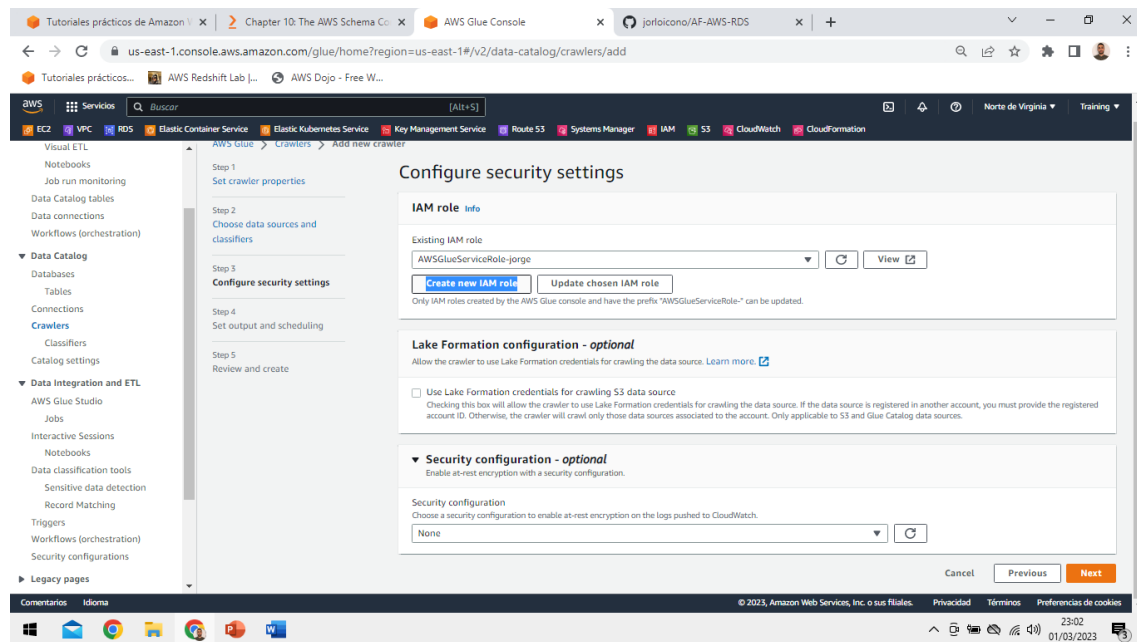
Complete la ventana emergente utilizando los siguientes detalles:

1. **Nombre:** DBCertFlight
2. **Incluir ruta:** s3://athena-examples/flight/

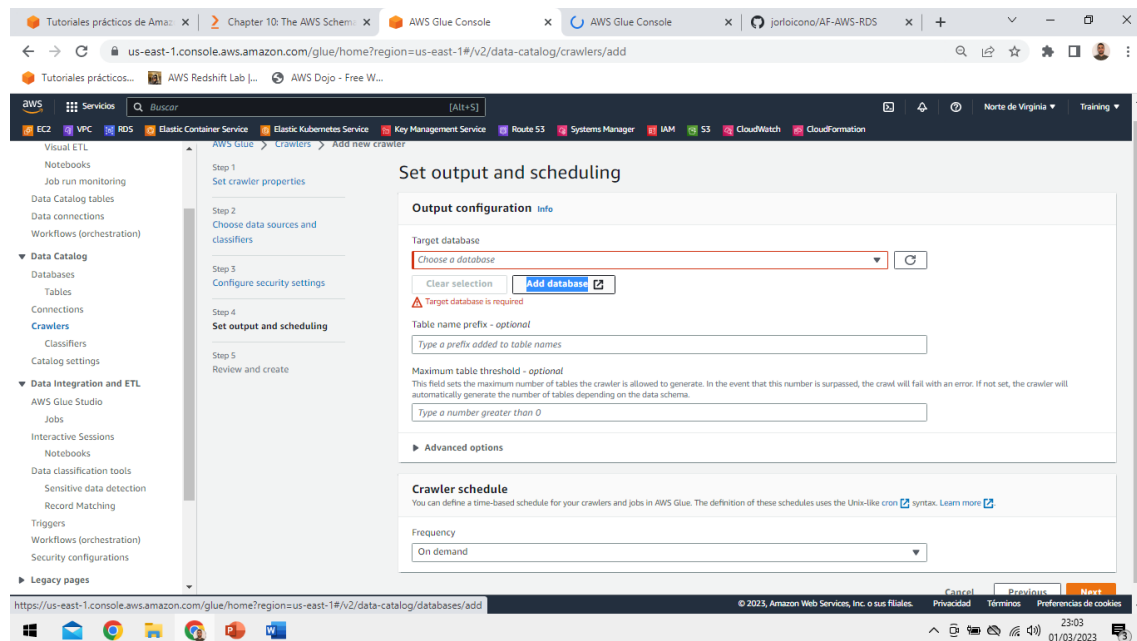


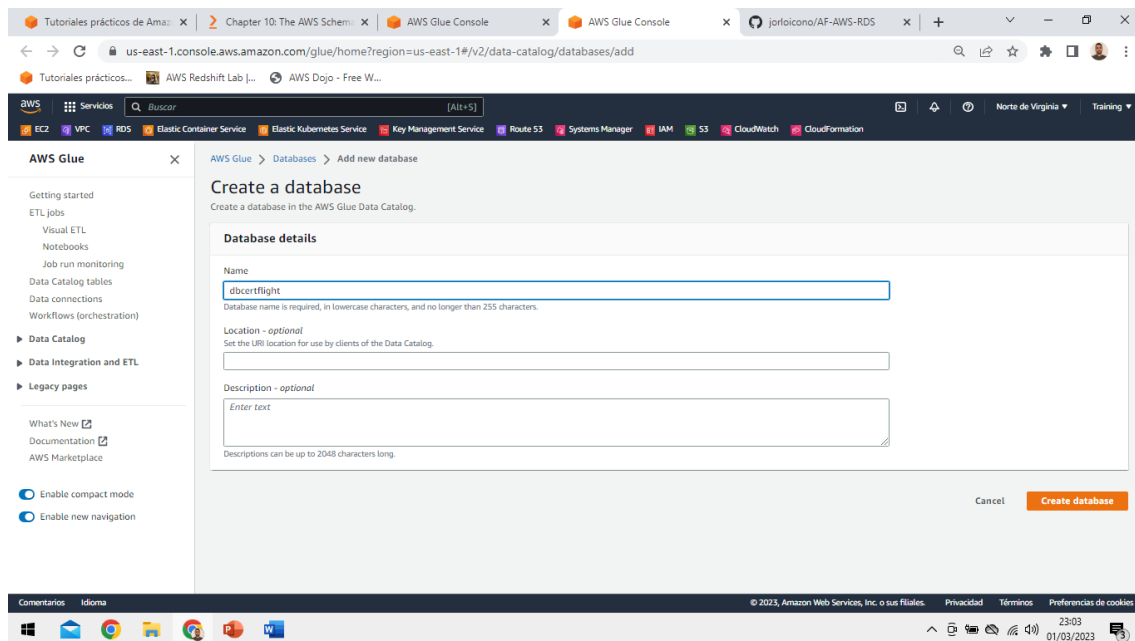
La siguiente captura de pantalla muestra cómo se debe completar el formulario:

Ingrese `AWSGlueServiceRole-jorge` para el rol de IAM y haga clic en **Siguiente**



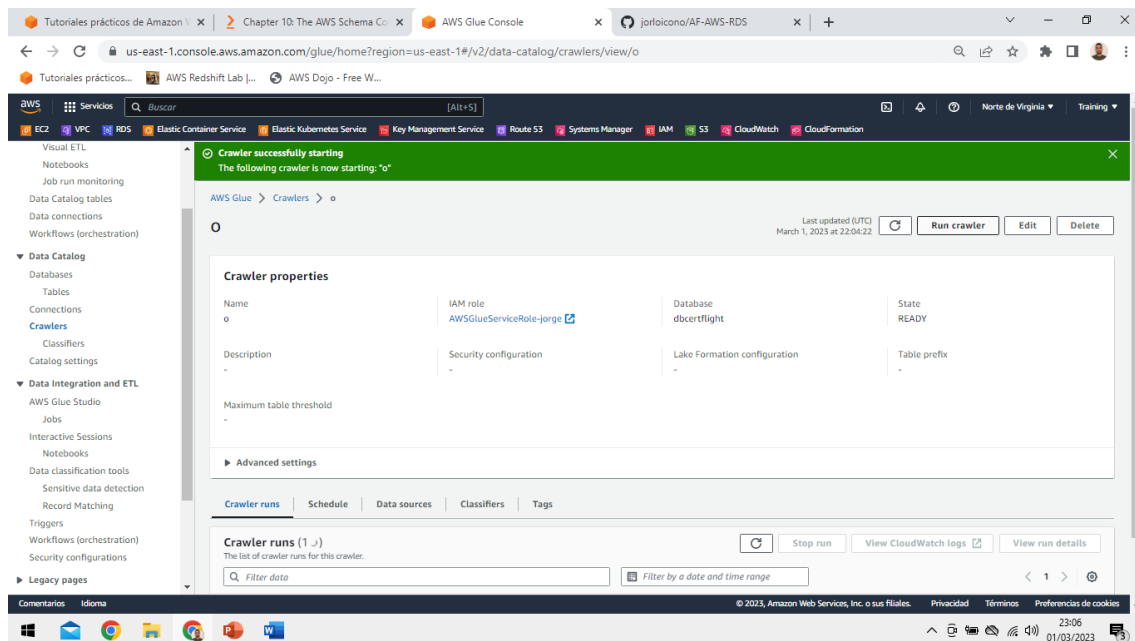
Deje la frecuencia establecida para ejecutar bajo demanda y establezca el nombre de la base de datos en `dbcertflight`. Luego, haz clic en **Crear**





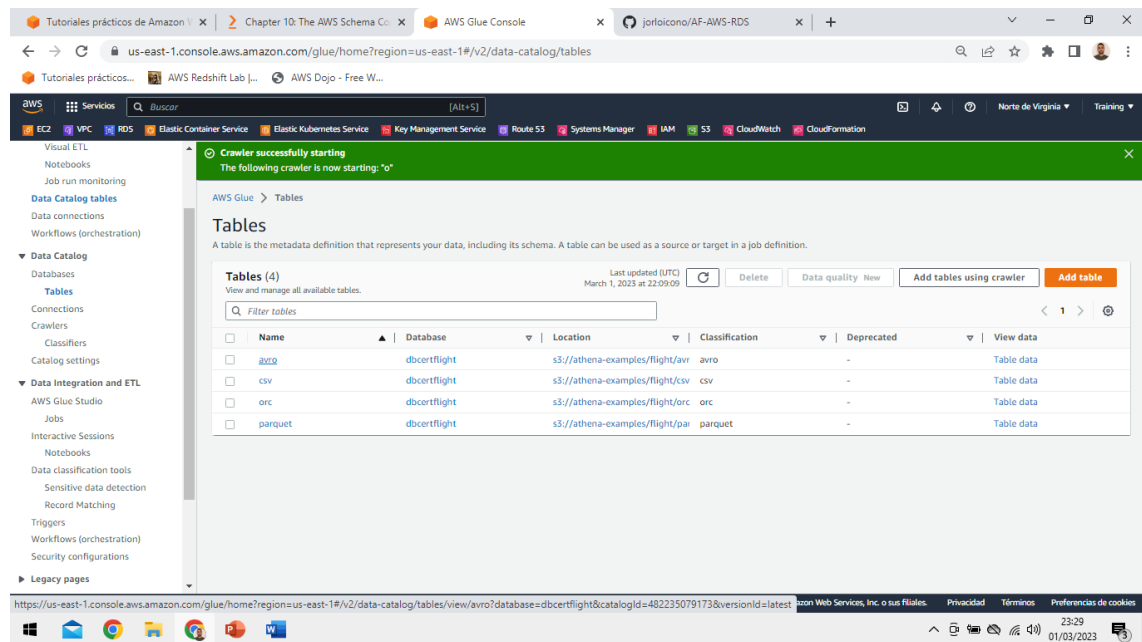
Haga clic en **Finalizar** para crear la tabla. Volver al panel de control de los rastreadores

Dadle a run crawler, le llevará 1 minuto para completar.



Cuando su estado sea **Listo**, haga clic en **Tablas** en el menú de la izquierda. Verá que se han creado cuatro nuevas tablas.

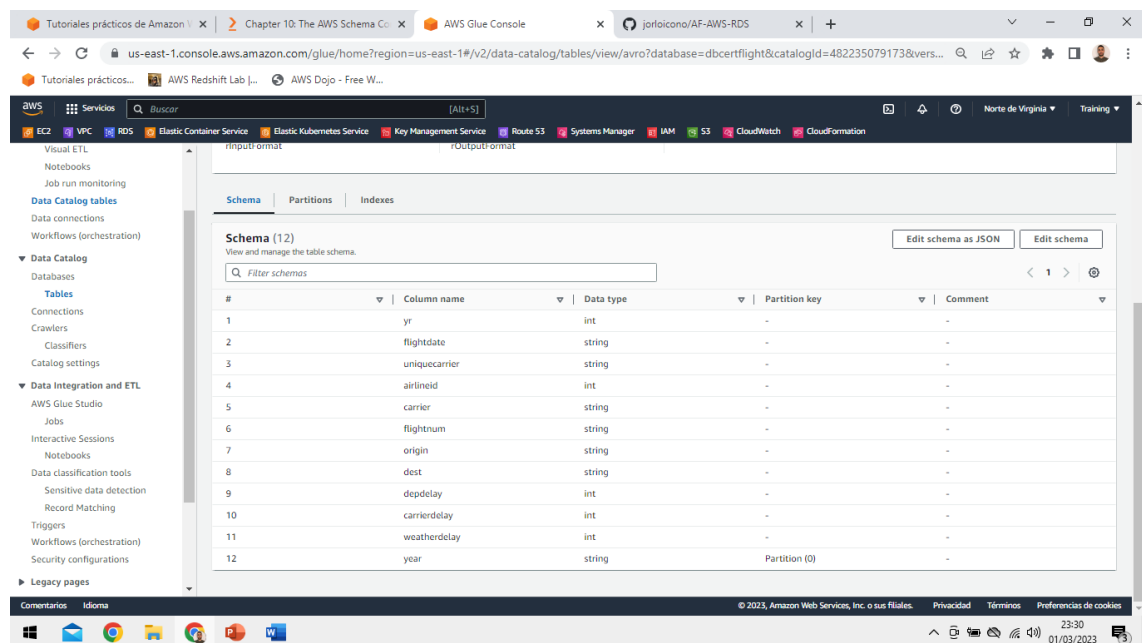
Haga clic en Avro



The screenshot shows the AWS Glue Console interface. A green banner at the top indicates 'Crawler successfully starting'. The main content area is titled 'Tables' and shows a list of tables created by the crawler. The table 'avro' is selected, and its details are shown below.

Name	Database	Location	Classification	Deprecated	View data
avro	dbcertflight	s3://athena-examples/flight/avro	avro	-	Table data
csv	dbcertflight	s3://athena-examples/flight/csv	csv	-	Table data
orc	dbcertflight	s3://athena-examples/flight/orc	orc	-	Table data
parquet	dbcertflight	s3://athena-examples/flight/parquet	parquet	-	Table data

Podrá ver el esquema que AWS Glue ha creado para nosotros en función del archivo **avro** en S3. Ahora podemos usar Athena para consultarlo.

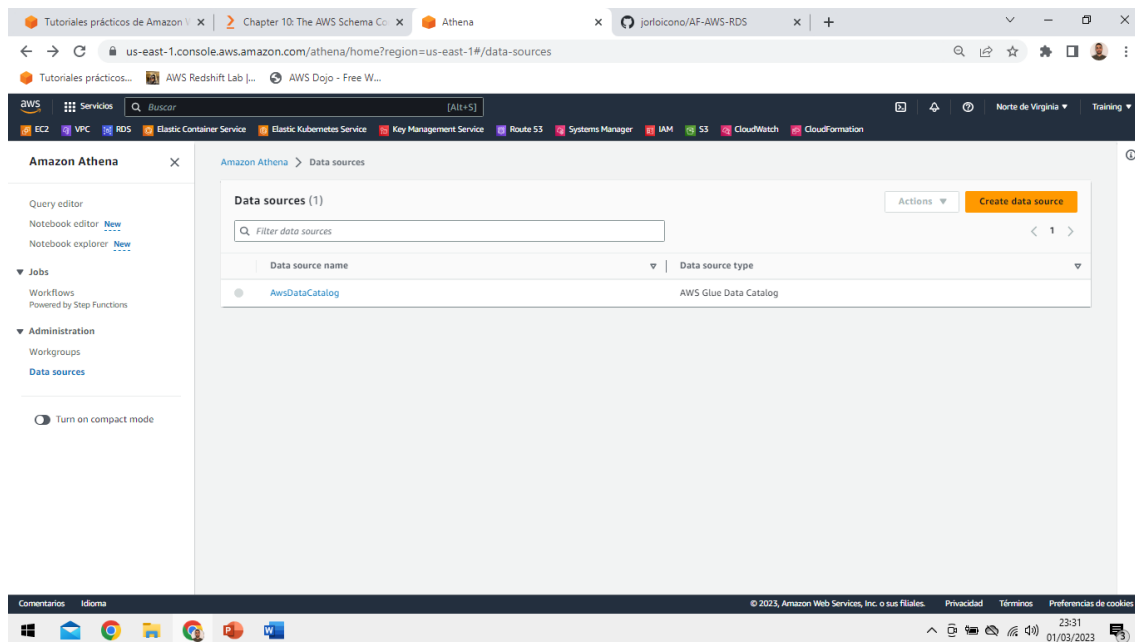
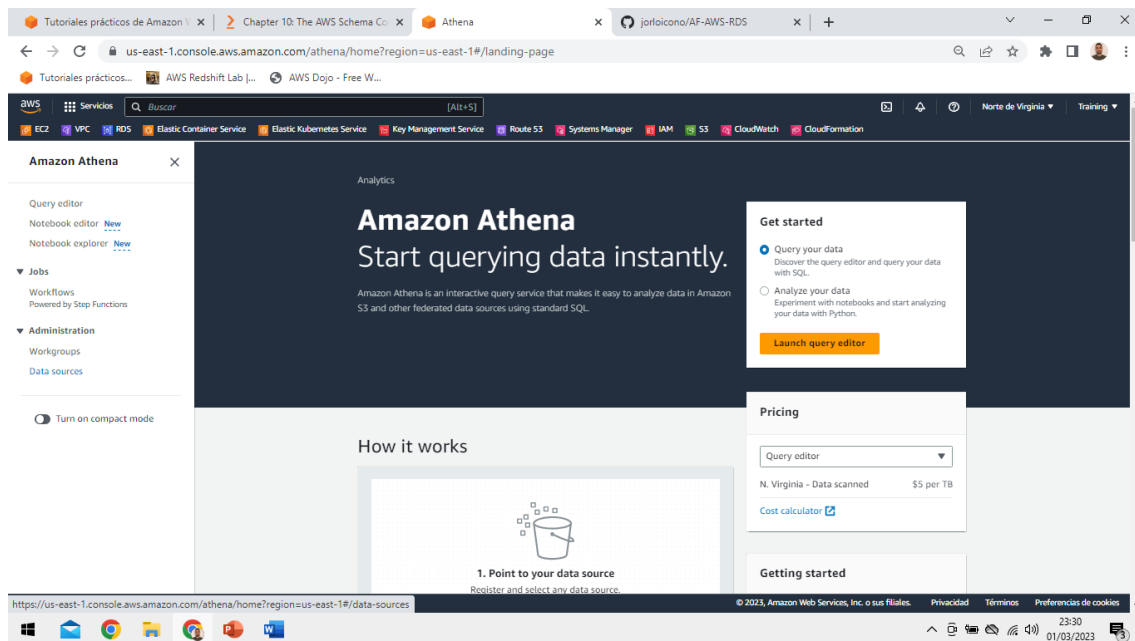


The screenshot shows the AWS Glue Console interface, specifically the 'Schema' page for the 'avro' table. The schema is displayed as a table with 12 columns. The 'Schema' tab is selected, and the 'Schema (12)' section shows the table structure.

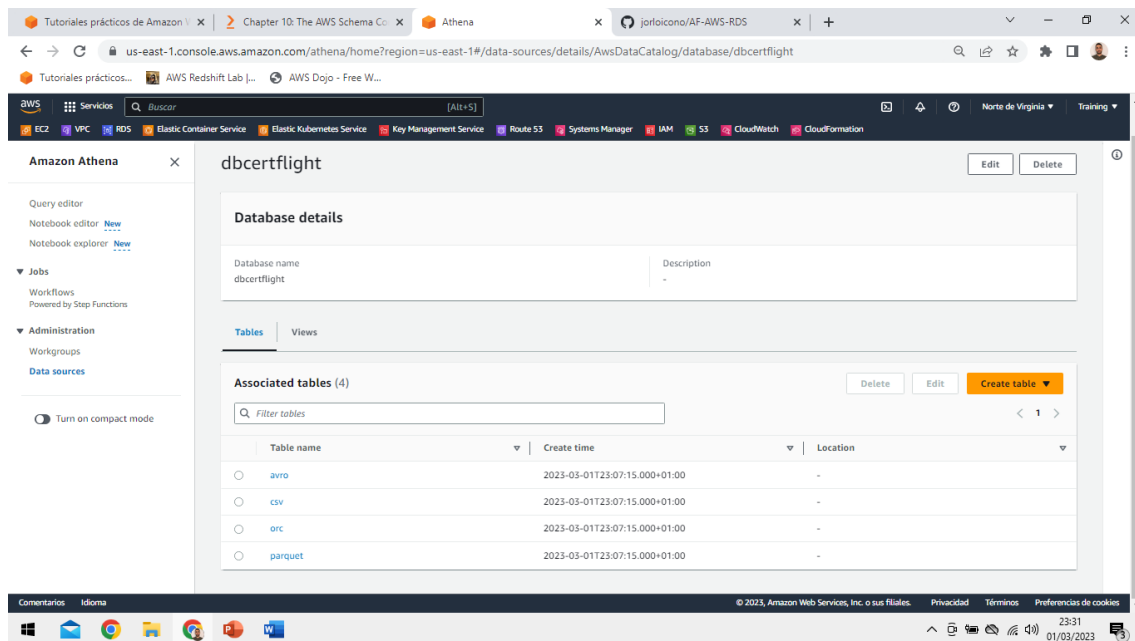
#	Column name	Data type	Partition key	Comment
1	yr	int	-	-
2	flightdate	string	-	-
3	uniquecarrier	string	-	-
4	airlineid	int	-	-
5	carrier	string	-	-
6	flightnum	string	-	-
7	origin	string	-	-
8	dest	string	-	-
9	depdelay	int	-	-
10	carrierdelay	int	-	-
11	weatherdelay	int	-	-
12	year	string	Partition (0)	-

Navegue a **Amazon Athena** desde el menú principal.

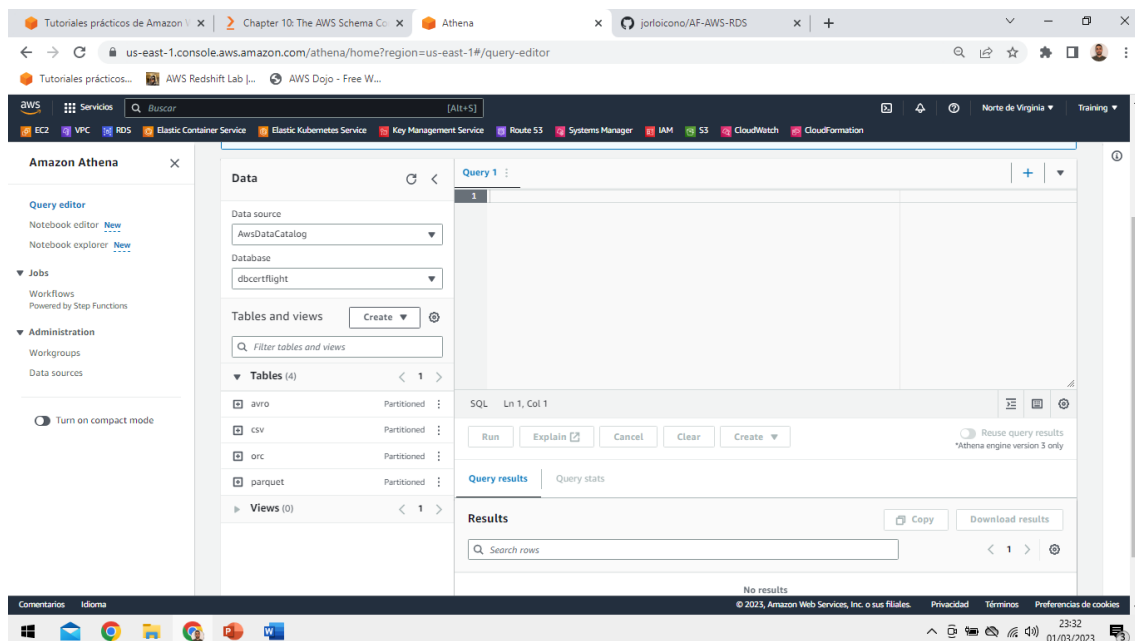
Haga clic en **Fuentes de datos** en el menú de la izquierda. Debería ver una fuente de datos llamada **AwsDataCatalog**.



Al hacer clic aquí, verá la dbcertflightbase de datos que creamos en AWS Glue.

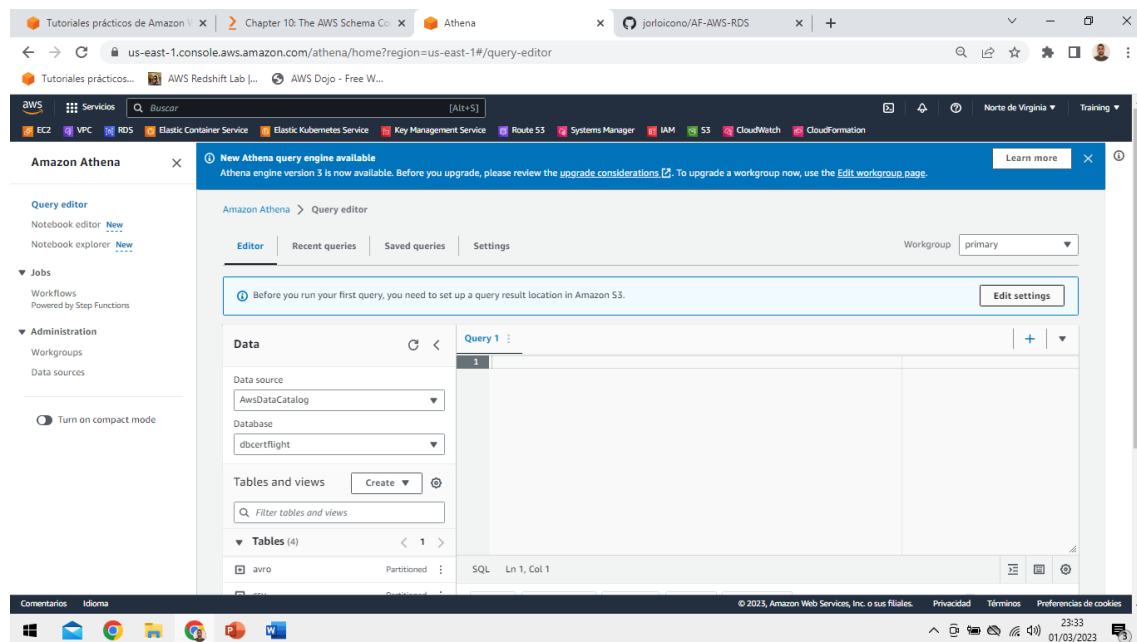


Haga clic en **Editor de consultas** en el menú de la izquierda. Verifique que **AwsDataCatalog** se haya seleccionado para **Fuente de datos** y elija **dbcertflight** del menú desplegable **Base de datos**



Antes de poder correr una consulta, necesitaremos configurar un depósito S3 para los resultados de nuestra consulta. Haga clic en la pestaña **Configuración** y luego haga clic en **Administrar**.

Ingrese una ruta de depósito S3 en el cuadro. También puede **examinar S3** para encontrar un depósito adecuado si es necesario.



Una vez que haya elegido su depósito S3, haga clic en **Guardar**

Regrese a la pestaña **Editor** para que podamos ejecutar nuestras consultas SQL. Si expande la tabla avro haciendo clic en el símbolo (+) junto a ella, verá todas las columnas que puede consultar. Ingrese la siguiente consulta en el cuadro para encontrar todos los vuelos que se retrasaron más de 15 minutos:

```
SELECT *  
FROM avro  
WHERE depdelay > 15;
```

Esta consulta tardará alrededor de 90 segundos en completarse y devolverá una salida similar a esta

Data

Data Source: AwsDataCatalog

Database: dbcertflight

Tables and views: Create

Filter tables and views

Tables (4)

- avro
- yr
- flightdate
- uniquecarrier
- airlineid
- carrier
- flightnum
- origin
- dest
- depdelay
- carrierdelay

Query 1

```
1 SELECT *
2 FROM avro
3 WHERE depdelay > 15;
4
```

SQL Ln 3, Col 21

Run again Cancel Save as Clear Create

Completed Time in queue: 0.167 sec Run time: 1 min 32.196 sec Data scanned: 8.85 GB

Results (100+)

Copy Download results

Search rows

yr	flightdate	uniquecarrier	airlineid	carrier	flightnum	origin	dest	depdelay
2000	2000-06-02	"UA"	19977	"UA"	"2014"	"LAX"	"SFO"	88
2000	2000-06-02	"UA"	19977	"UA"	"2016"	"LAX"	"SFO"	81
2000	2000-06-03	"UA"	19977	"UA"	"2016"	"LAX"	"SFO"	62
2000	2000-06-04	"UA"	19977	"UA"	"2016"	"LAX"	"SFO"	35

Ahora puede ejecutar otras consultas para obtener más información sobre las consultas con Athena y las limitaciones de SQL.

Athena guarda todos los resultados de la consulta en el bucket de S3 que especificó anteriormente.

Si lo desea, puede navegar a S3 para encontrar los archivos de salida. Los archivos de salida ahora pueden ser utilizados por otro servicio si lo desea, como una herramienta gráfica como AWS Quicksight, pero eso está más allá del alcance de este laboratorio.