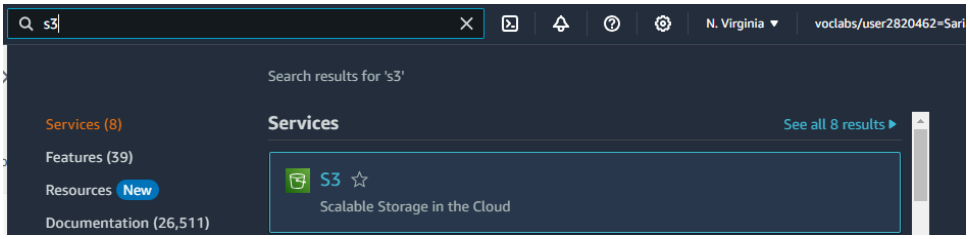
Sebastián Arias Usma.
C.C 1017932811
Evidencias laboratorios 3 big data
st0263-241 Tópicos de telemática

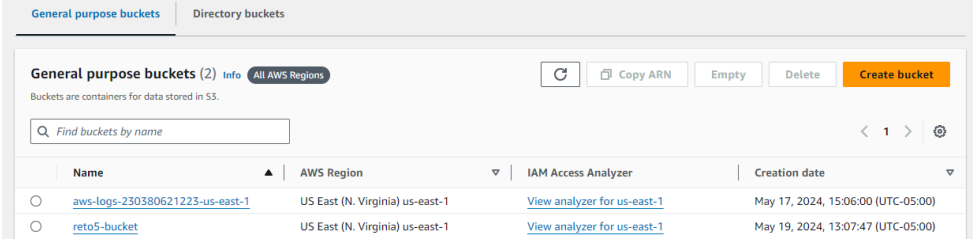
Github: https://github.com/sebasarias17/sariasu-st0263/tree/main/Reto%206
Reto#6 IMPLEMENTACIÓN DE UN DATA WAREHOUSE SENCILLO CON AWS S3, GLUE y ATHENA.

Paso # 1: Creamos un Bucket de S3 en nuestro AWS de la siguiente manera.

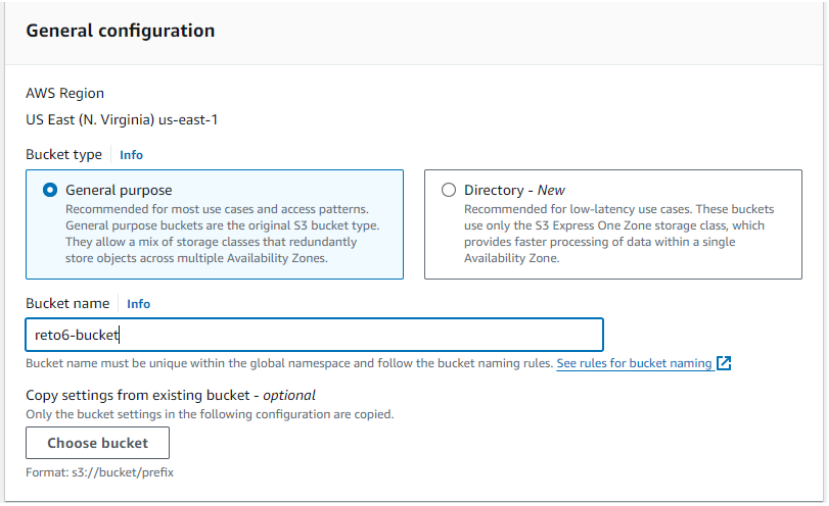
- Nos vamos al apartado de S3 en AWS



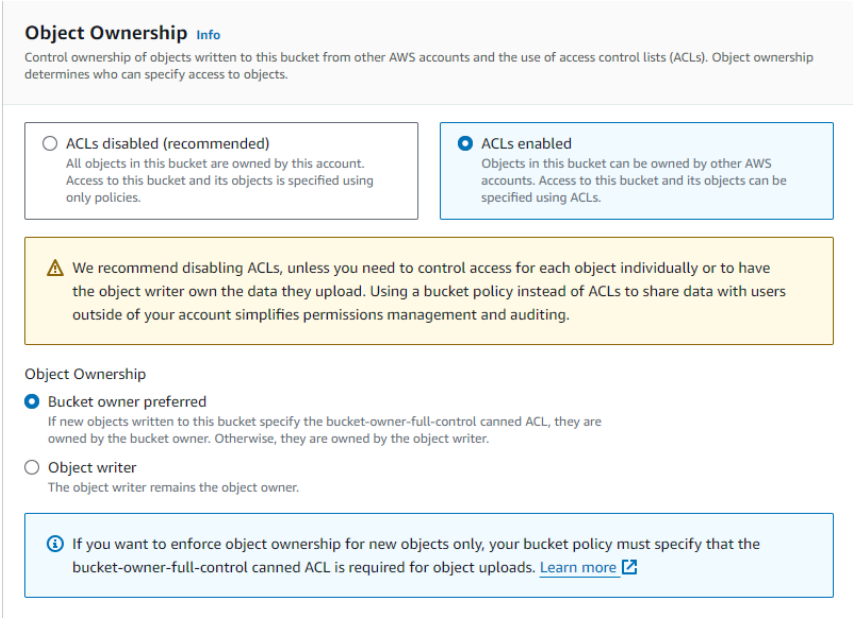
- Ahora le damos click al botón de crear bucket



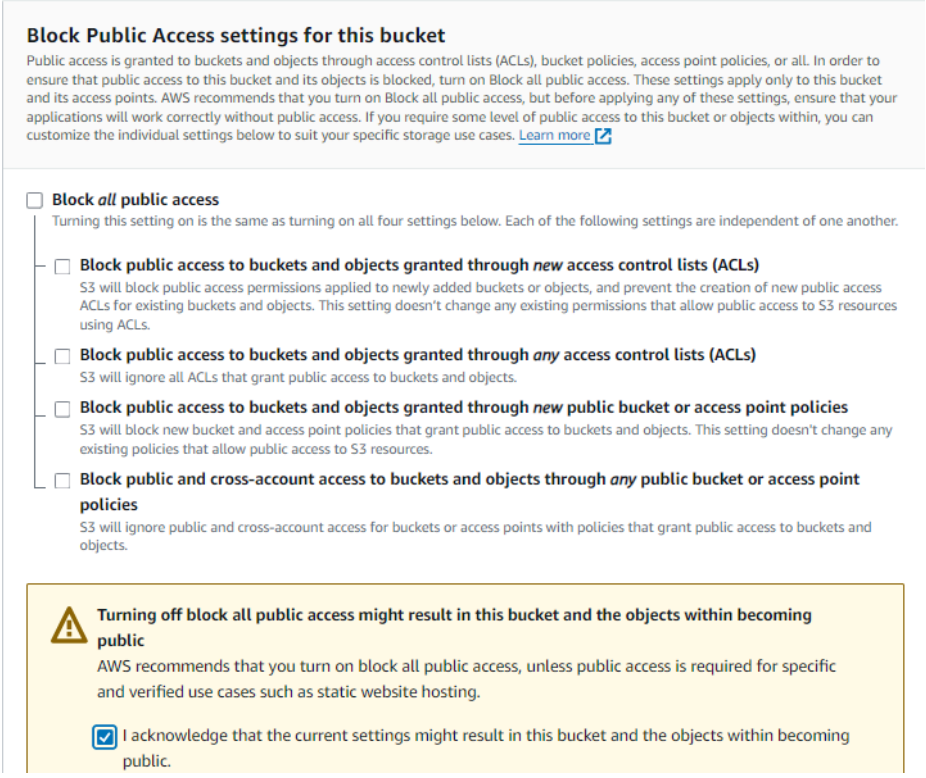
- Creamos el Bucket de S3 Con la siguiente configuración.



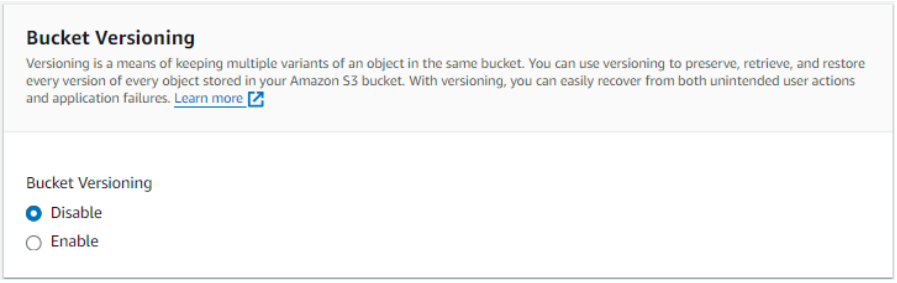
- Modificamos el Object Ownership y lo habilitamos.

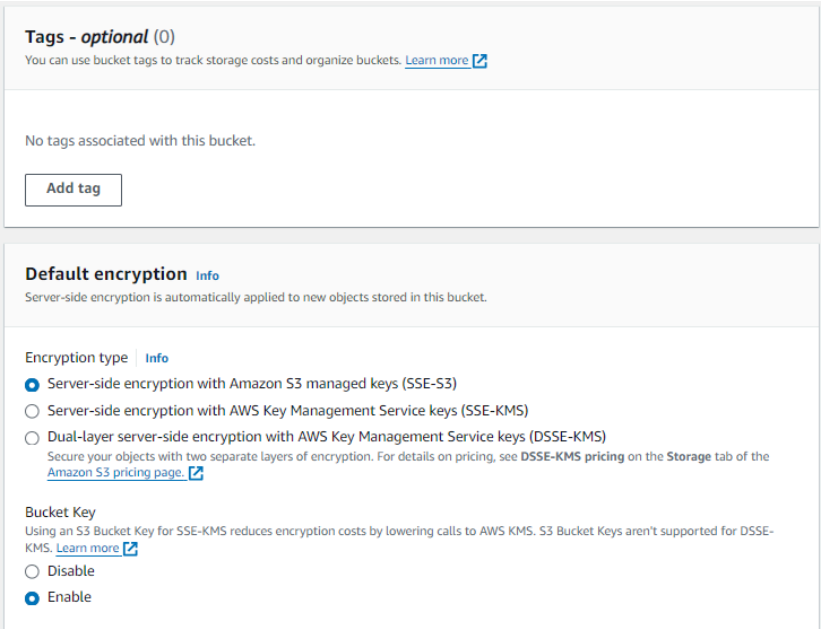


- Ahora modificamos el acceso para que sea público.

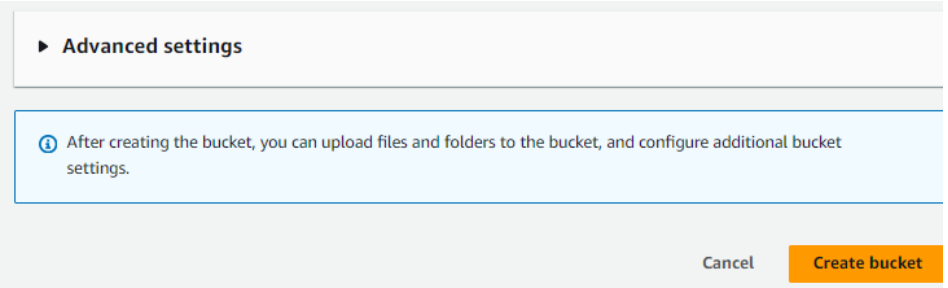


- Y estos ultimos ajustes los dejamos por defecto

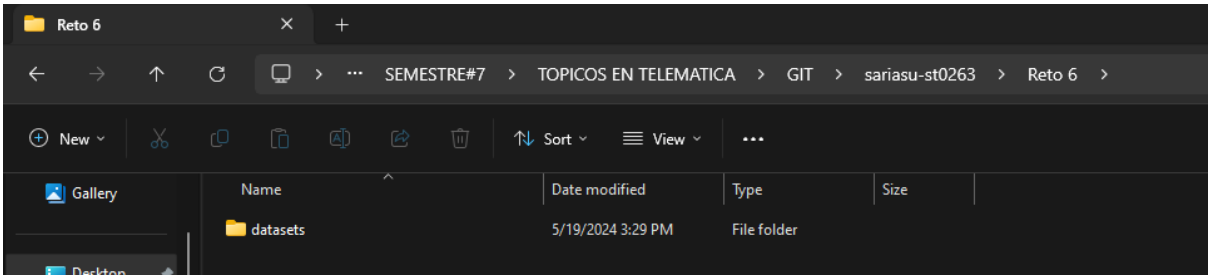


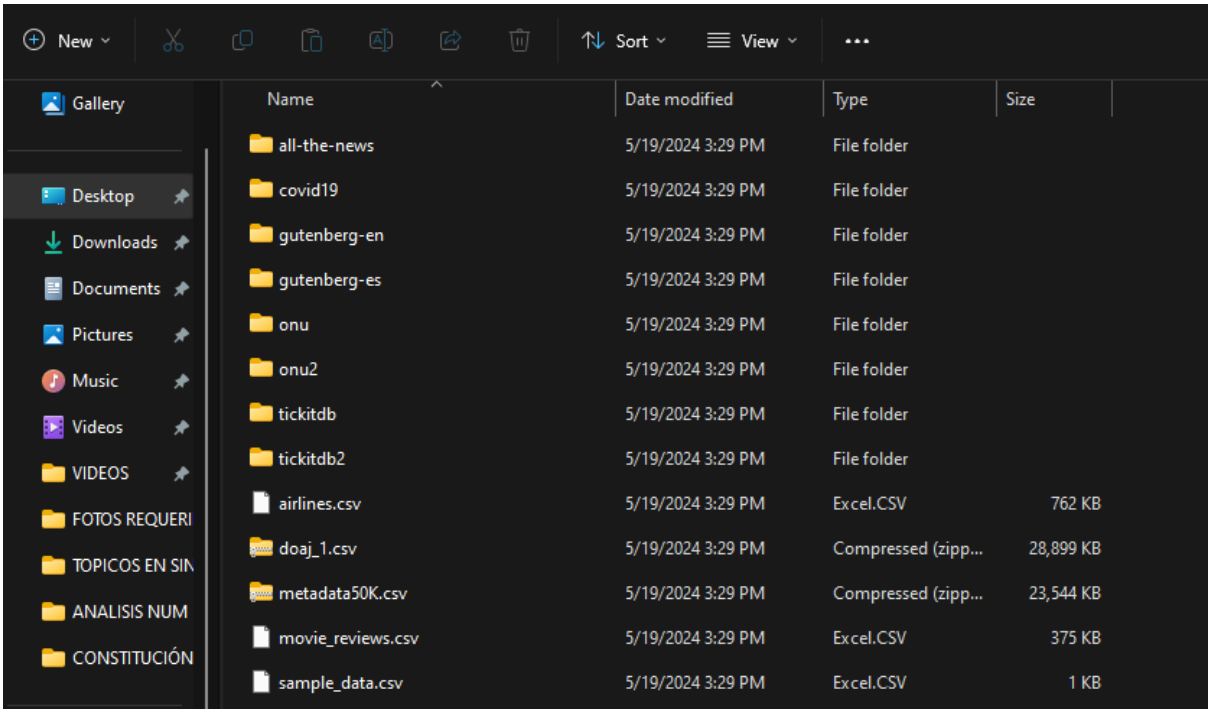


- Ahora le damos click en crear el cluster.



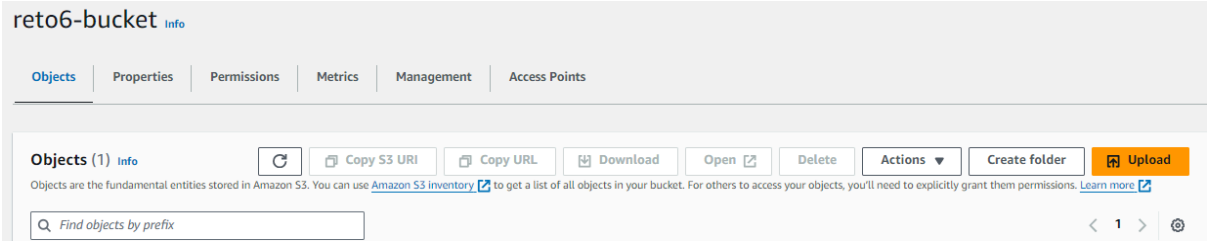
Paso # 2: Ahora clonaremos este github, el cual contiene los datasets necesarios para este laboratorio (https://github.com/sebasarias17/sariasu-st0263/tree/main/Reto%206) .



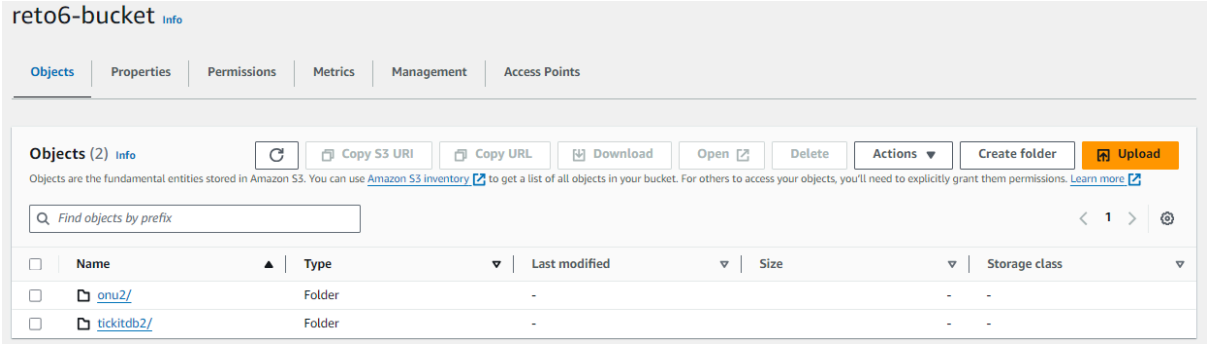


Paso #3: Ahora subiremos de manera manualmente a nuestro bucket las carpetas (onu2 y tickitdb2)

- Damos click en upload.

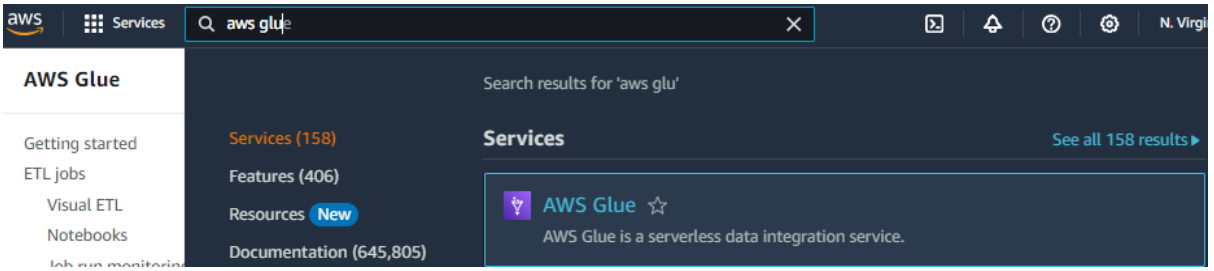


- Ahora subiremos las dos carpetas anteriormente mencionadas, primero una y luego la otra para obtener algo así

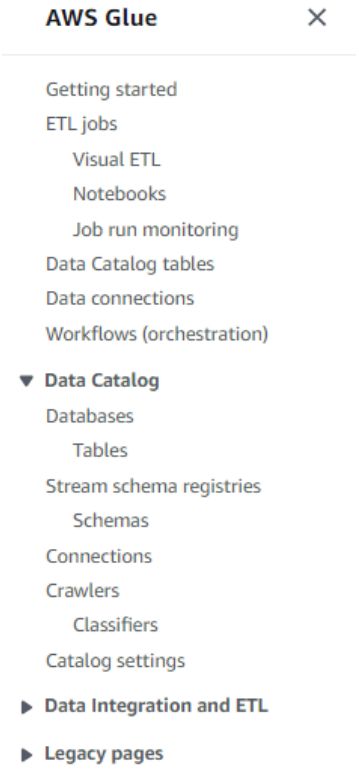


Paso #4: Ahora pasaremos a AWS Glue para catalogar y crear las tablas.

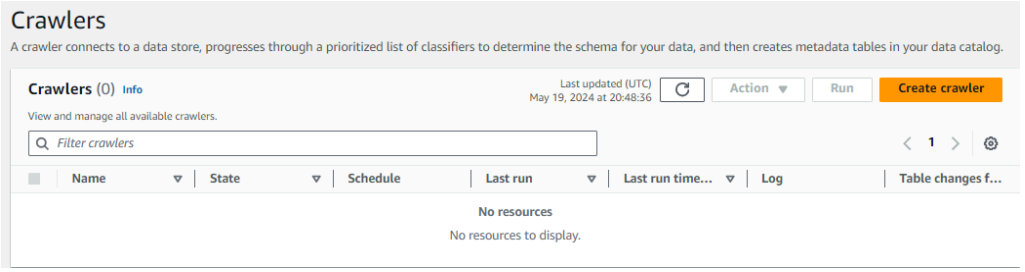
- Accedemos al apartado de AWS Glue en AWS



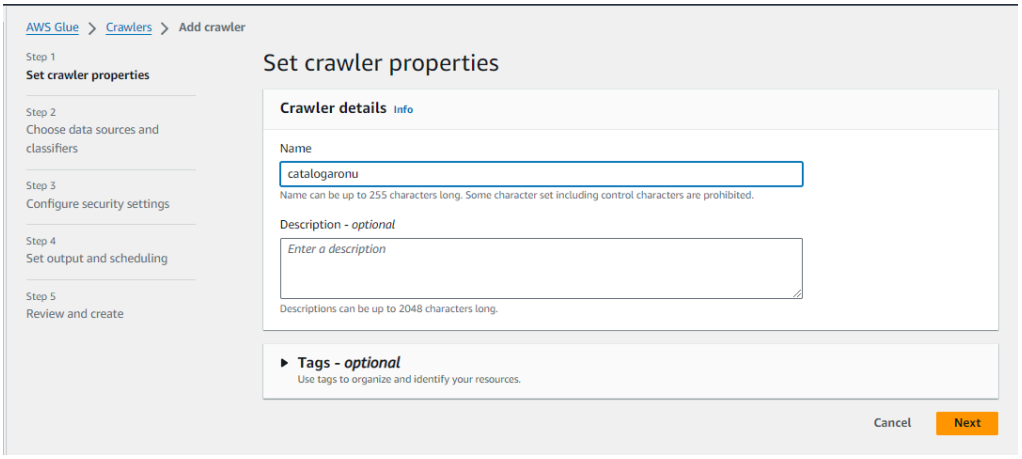
- Ahora en el menú de la izquierda iremos al apartado de Data Catalog y haremos la configuración en la opción de (crawlers).



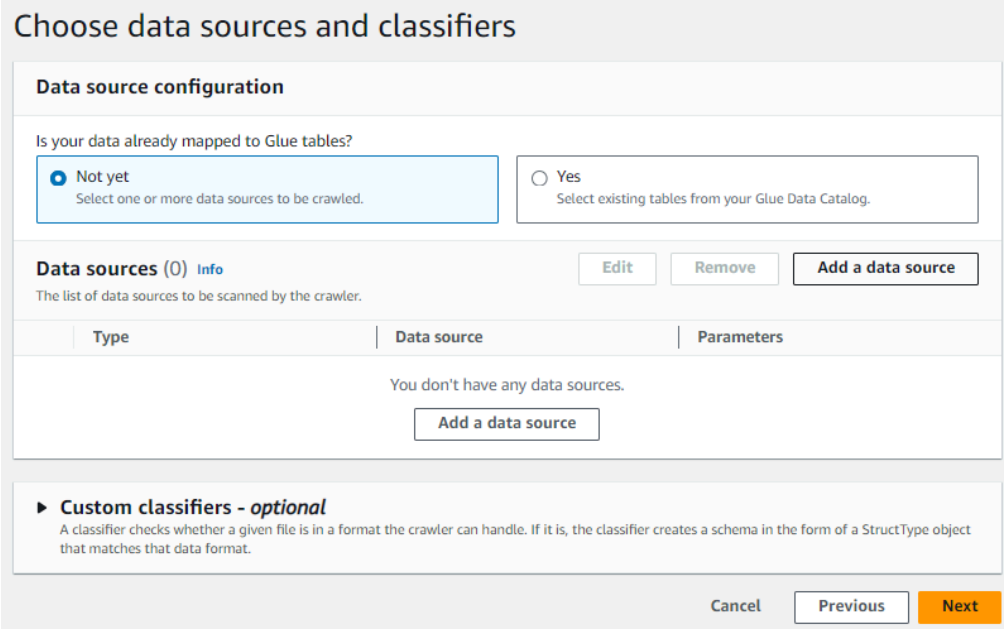
- Ahora crearemos el crawler y lo configuraremos de la siguiente manera.



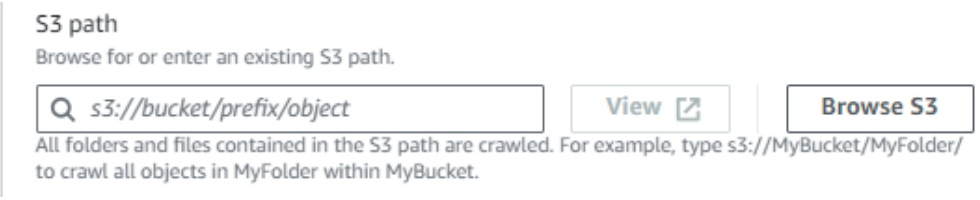
- Le daremos el nombre que deseemos



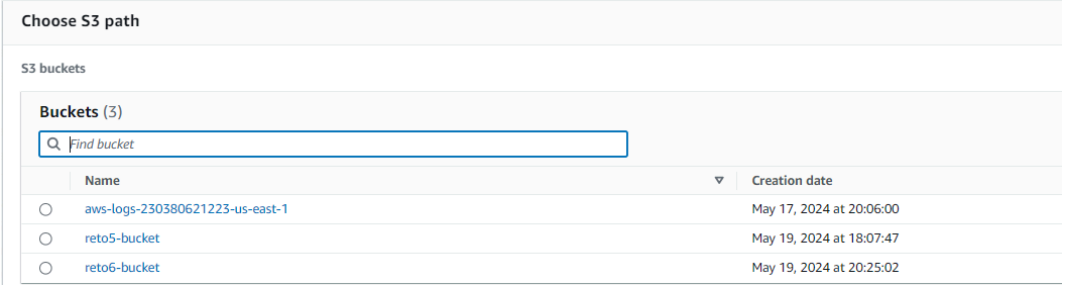
- Adicionamos una fuente de datos en el apartado de Add data source



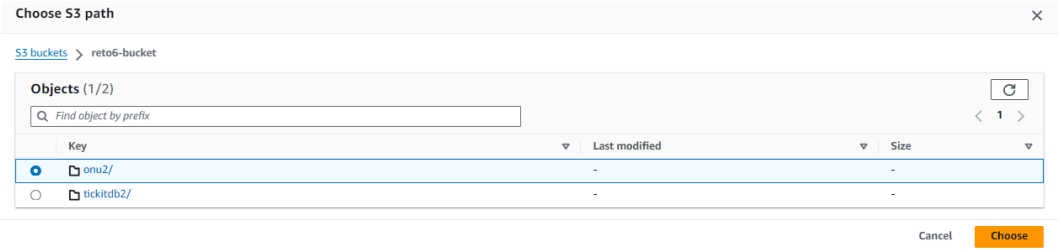
- Nos iremos a modificar el path donde buscará los archivos, haremos click en browse S3



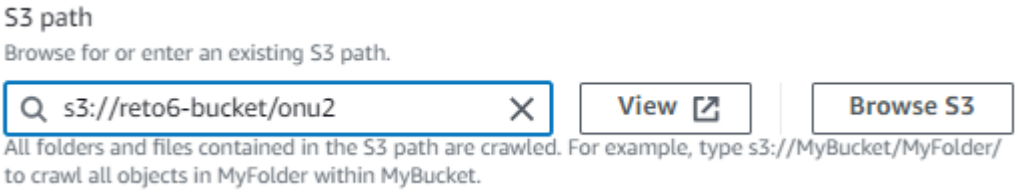
- Ahora ingresamos el bucket que hemos creado, en mi caso es reto6-bucket, y daremos click en su nombre



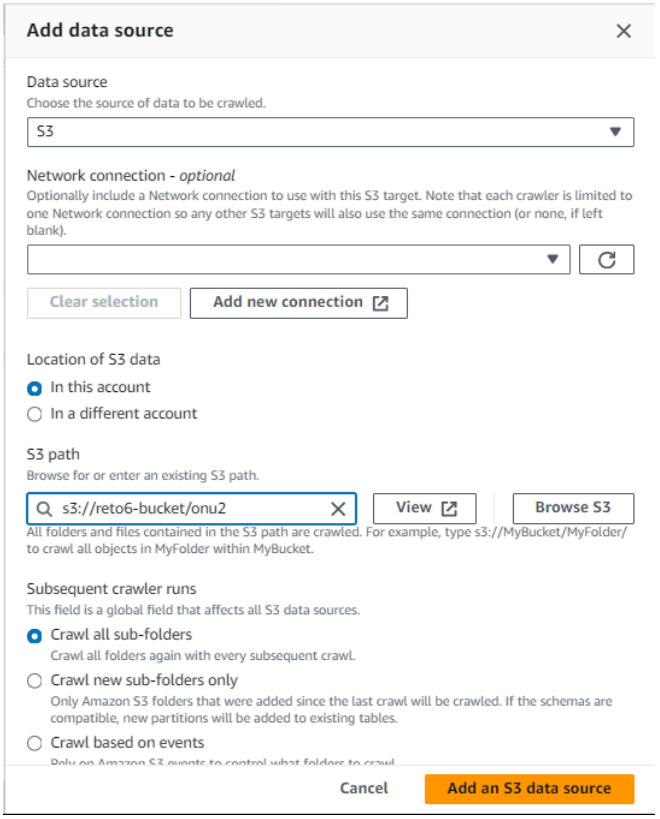
- Ahora seleccionaremos la carpeta de “onu2” y le daremos choose



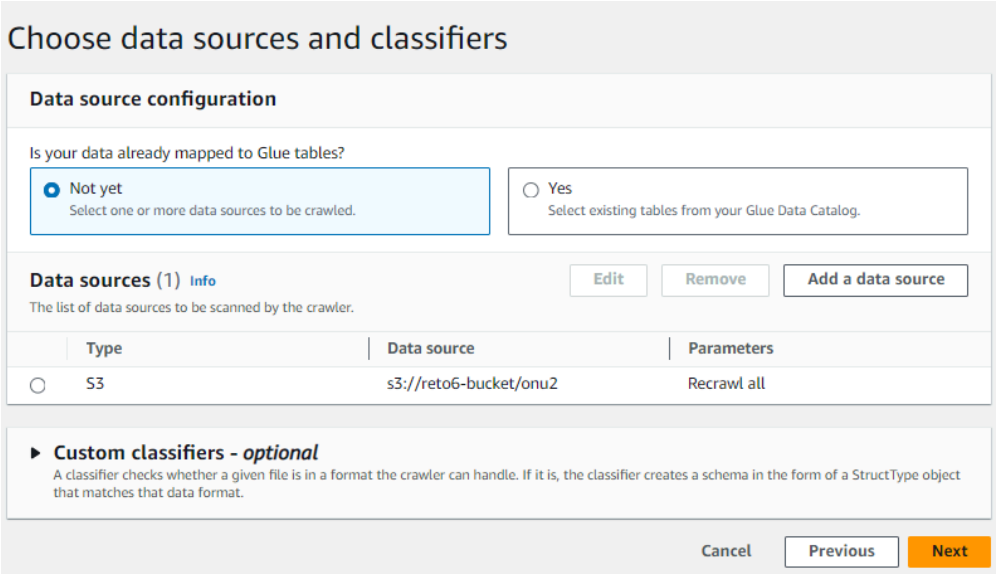
- Borraremos el ultimo slash del path para obtener algo así



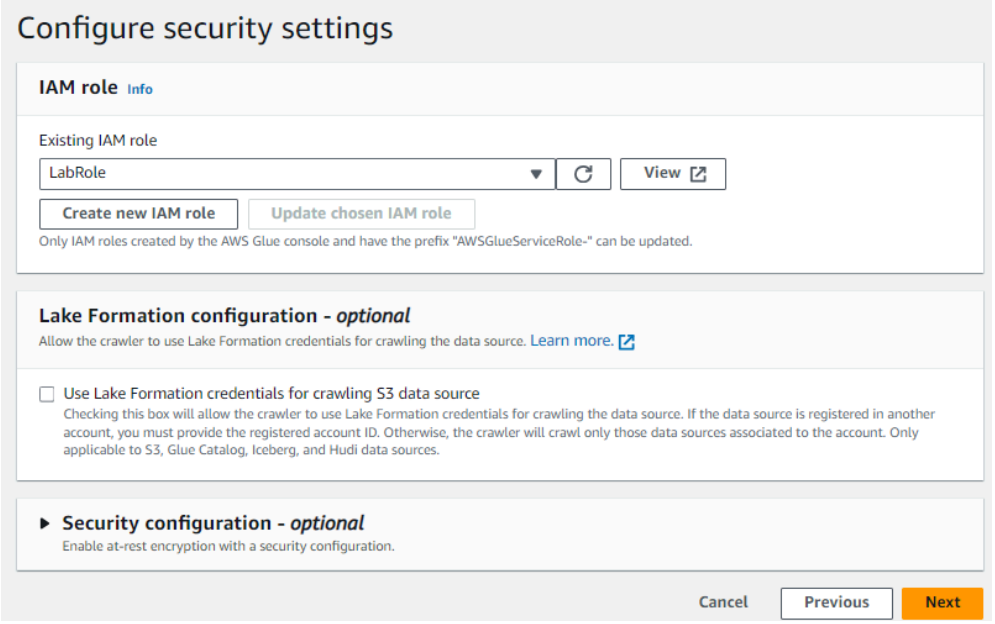
- Ya teniendo la configuración de esta manera le daremos click al boton de “Add an S3 data source”



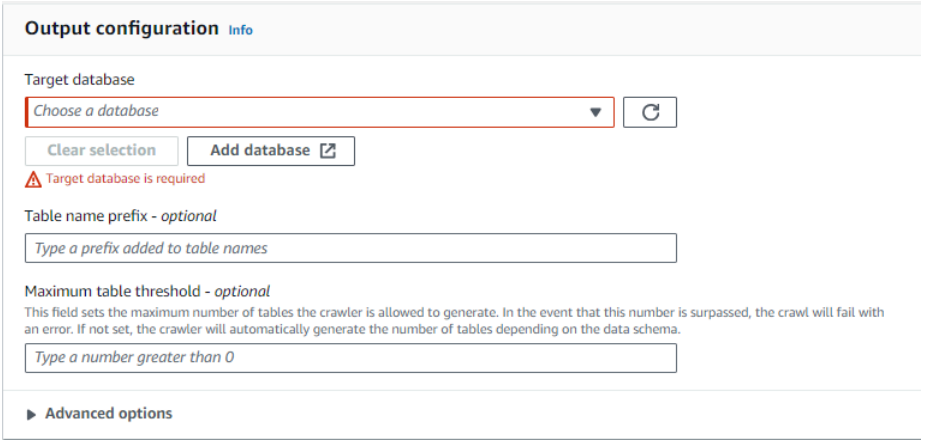
- Ahora le daremos next a este paso



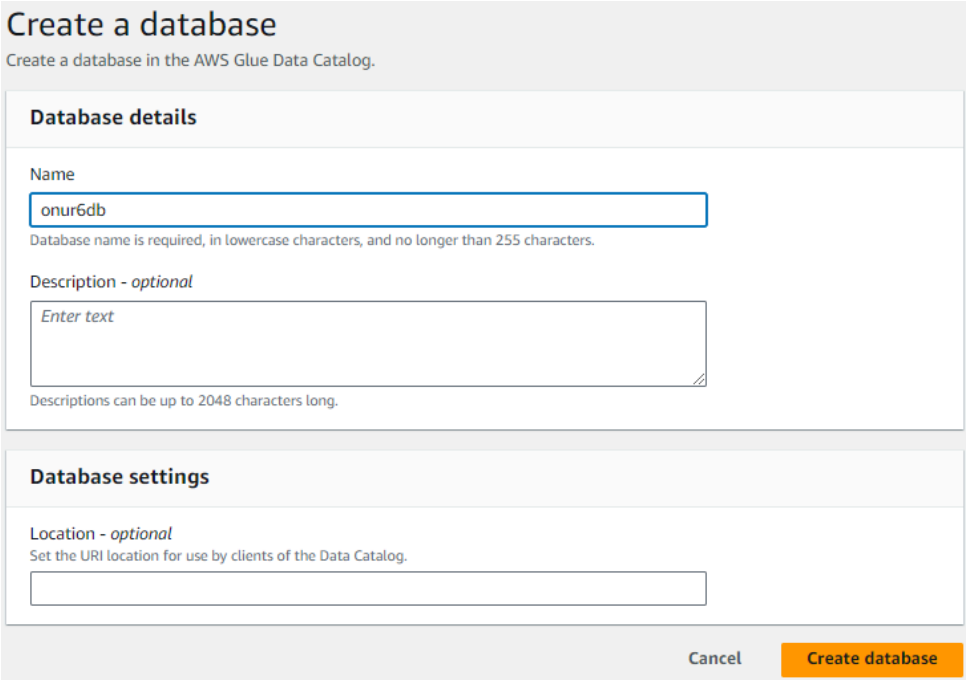
- Configuraremos las opciones de seguridad y damos next



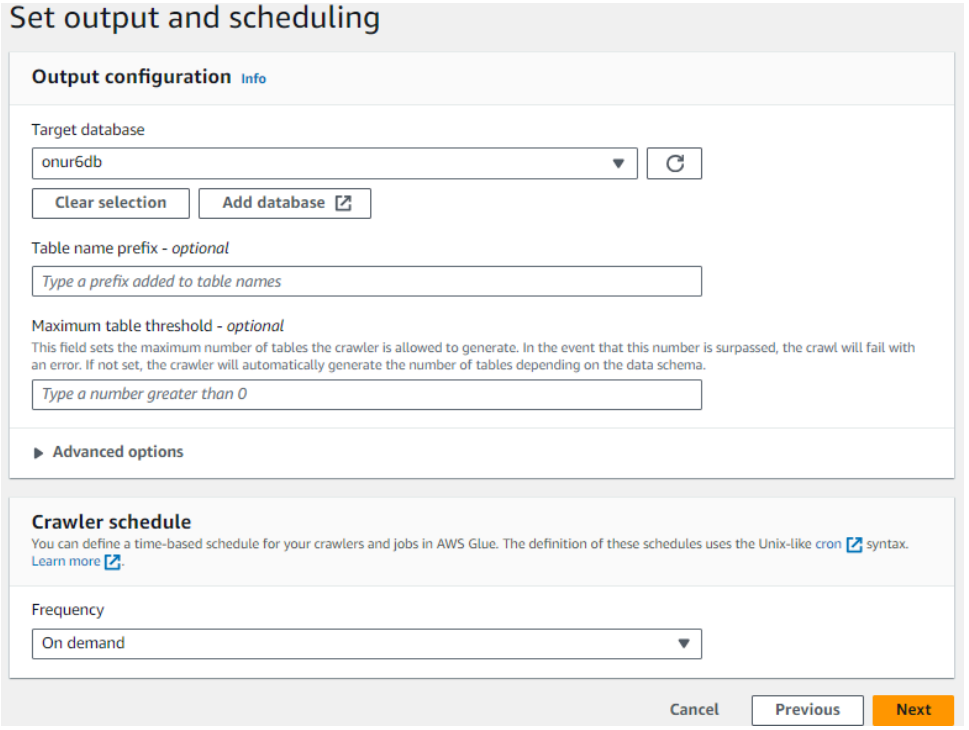
- Ahora para crear una nueva base de datos la adicionamos en el botón “Add database”



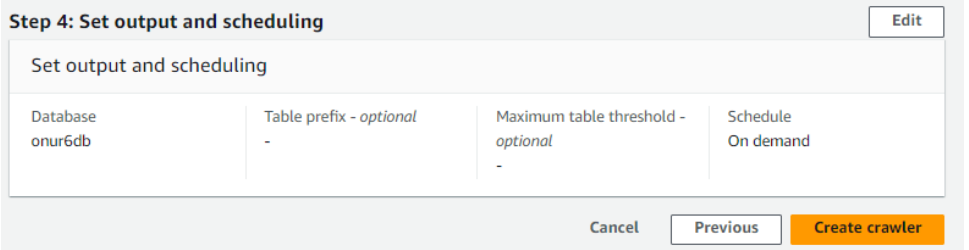
- La crearemos así y le daremos “Next”



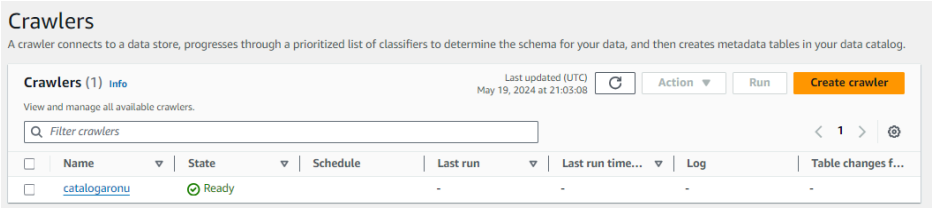
- Habiendo creado la base de datos la asignaremos y daremos “Next”



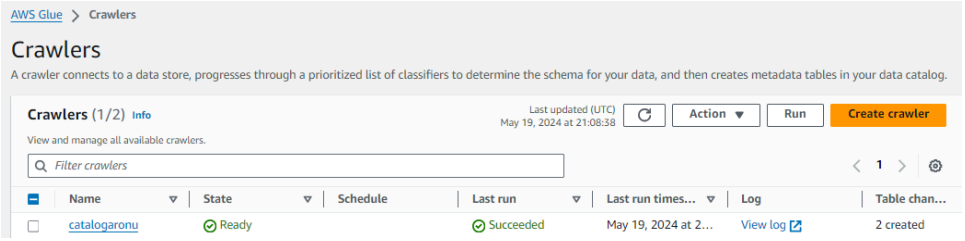
- Por ultimo le daremos a “create crawler”



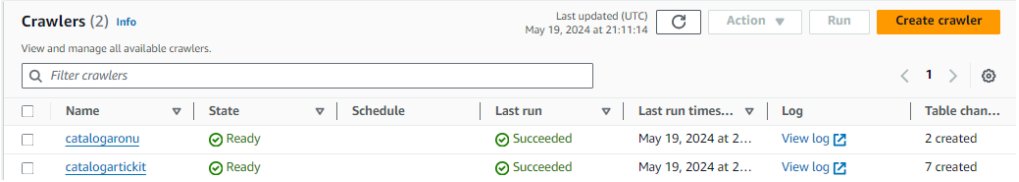
- Así nos debería salir el crawler que hicimos



- Ahora lo ejecutamos seleccionando el crawler y seleccionando “Run”

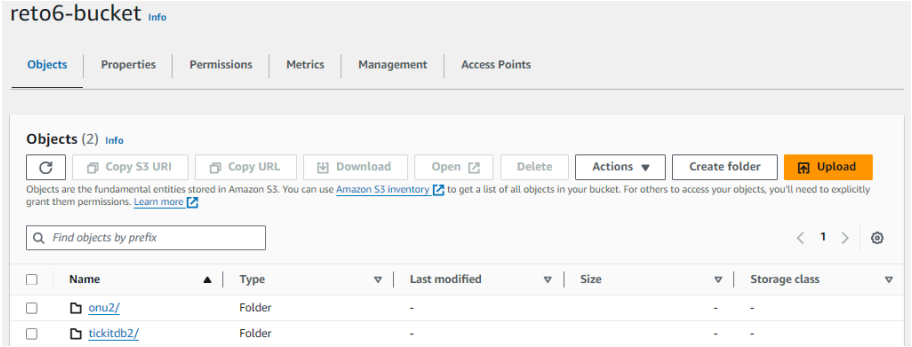


- Por ultimo seguiremos los mismos pasos para crear el crawler de tickitdb, para obtener esto en los crawlers.

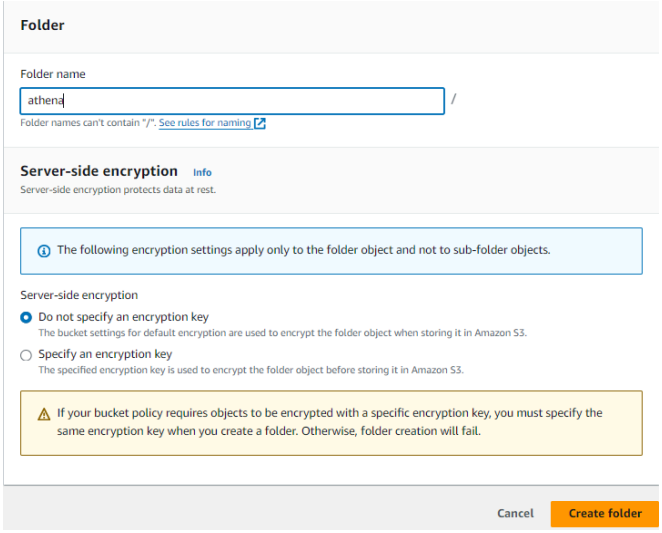


Paso #5: Ahora crearemos el directorio de salida de Athena en S3

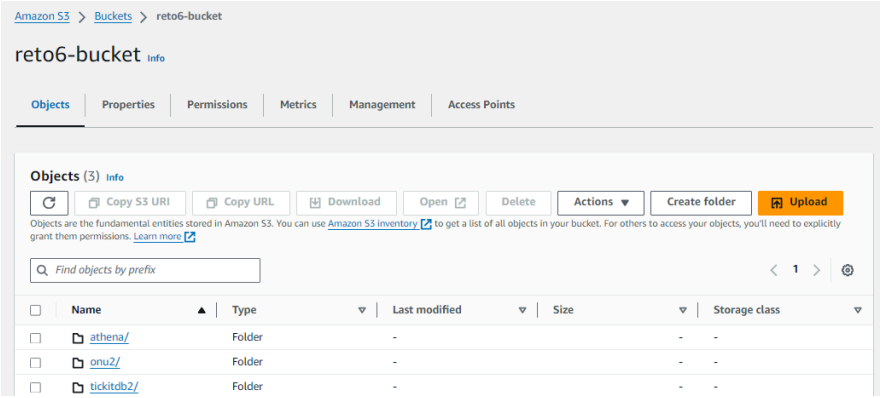
- Haremos click en create Folder



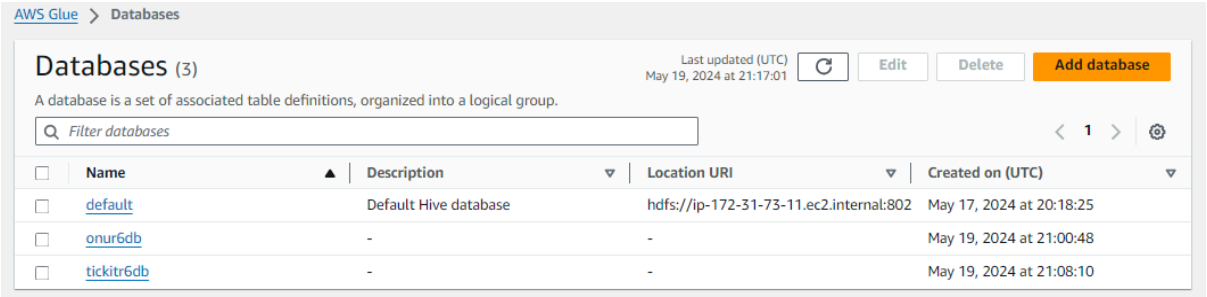
- Creamos la carpeta con estas credenciales y daremos click en “create folder”



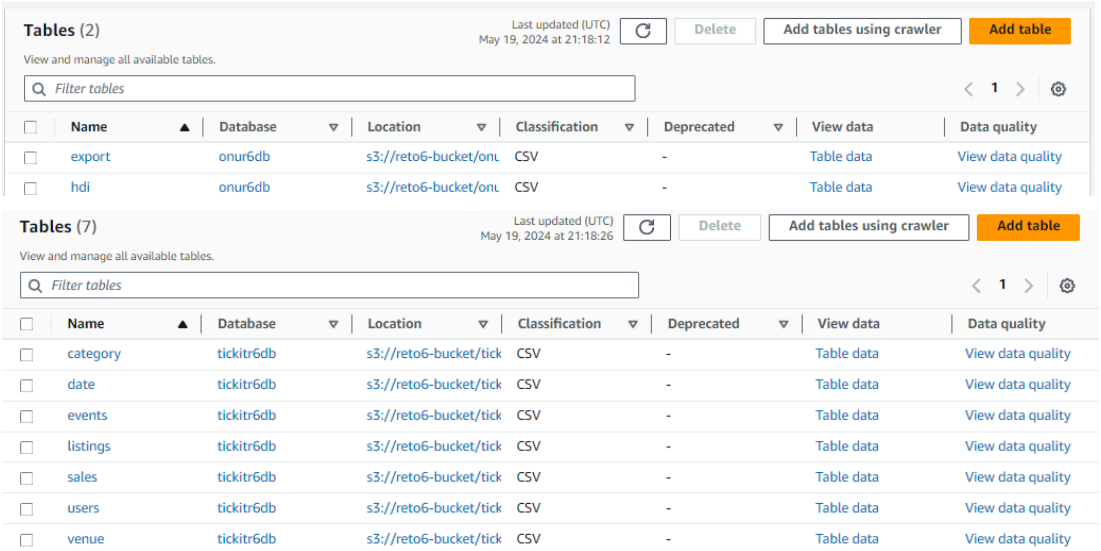
- Asi obtendremos estas 3 carpetas.



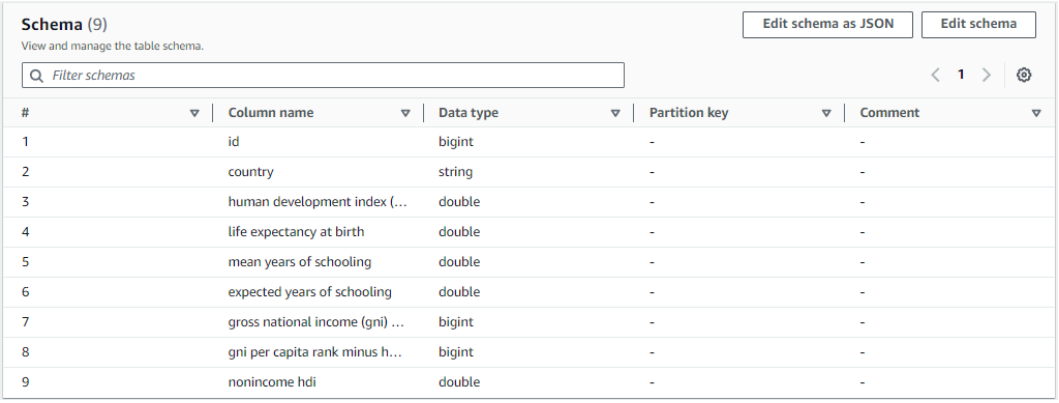
Paso #6: Ahora en nuestro AWS Glue iremos al apartado de bases de datos



- Ahora cuando accedemos a cualquier base de datos que creamos anteriormente nos deberan salir las tablas que se crearon, de esta manera



- De la misma manera si accedemos a cualquier tabla podremos ver los datos y columnas que tiene cada una, en este caso yo accedí a la carpeta “hdi” en la db de onu2



Schema (9)

View and manage the table schema.

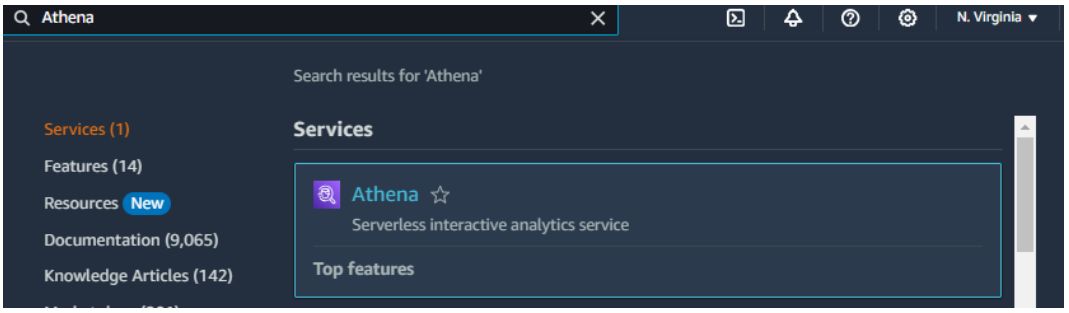
Edit schema as JSON | Edit schema

Filter schemas

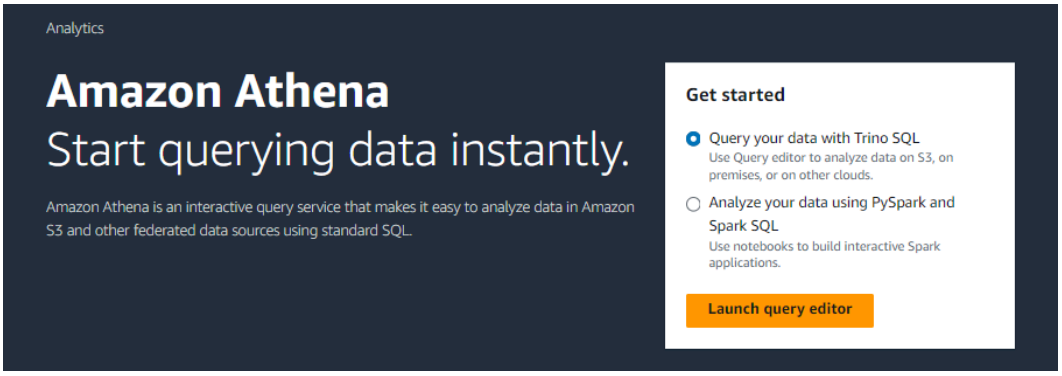
#	Column name	Data type	Partition key	Comment
1	id	bigint	-	-
2	country	string	-	-
3	human development index (...	double	-	-
4	life expectancy at birth	double	-	-
5	mean years of schooling	double	-	-
6	expected years of schooling	double	-	-
7	gross national income (gni) ...	bigint	-	-
8	gni per capita rank minus h...	bigint	-	-
9	nonincome hdi	double	-	-

Paso #7: Ahora podremos realizar consultas SQL mediante athena de la siguiente manera

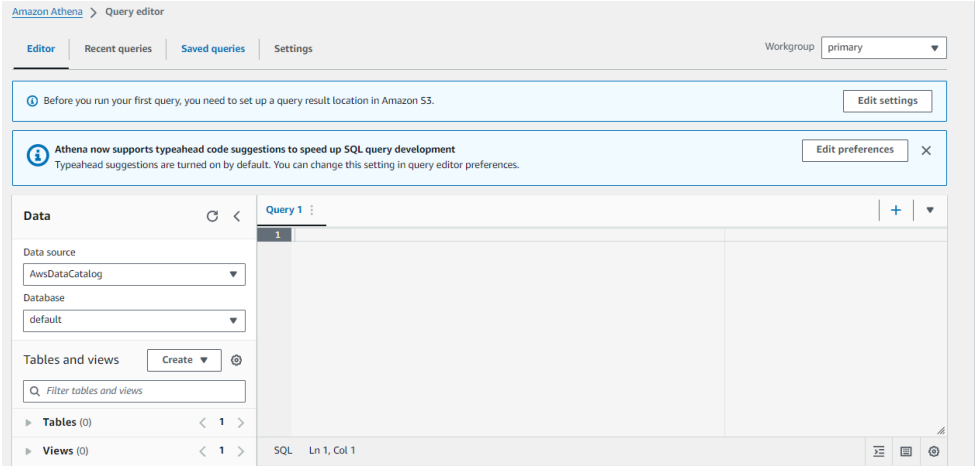
- Accedemos al servicio de Athena en AWS de la siguiente manera.



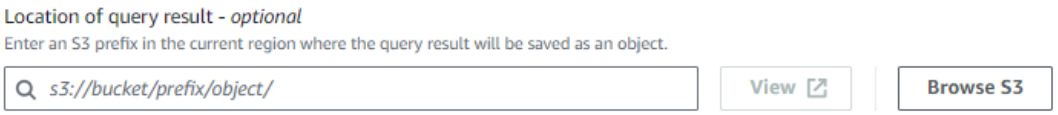
- Accederemos dado click a “Launch query editor”



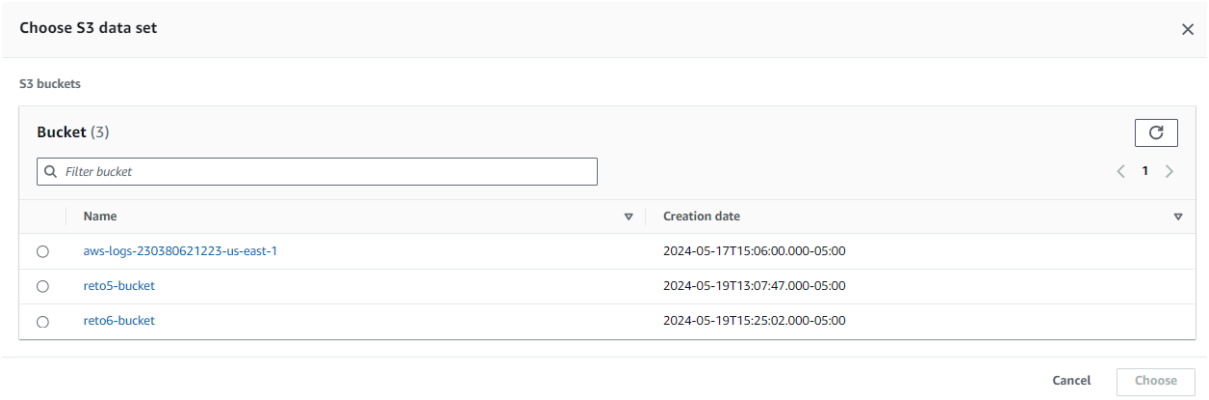
- Ahora debemos configurar esto por primera vez en el botón de “Edit Settings”



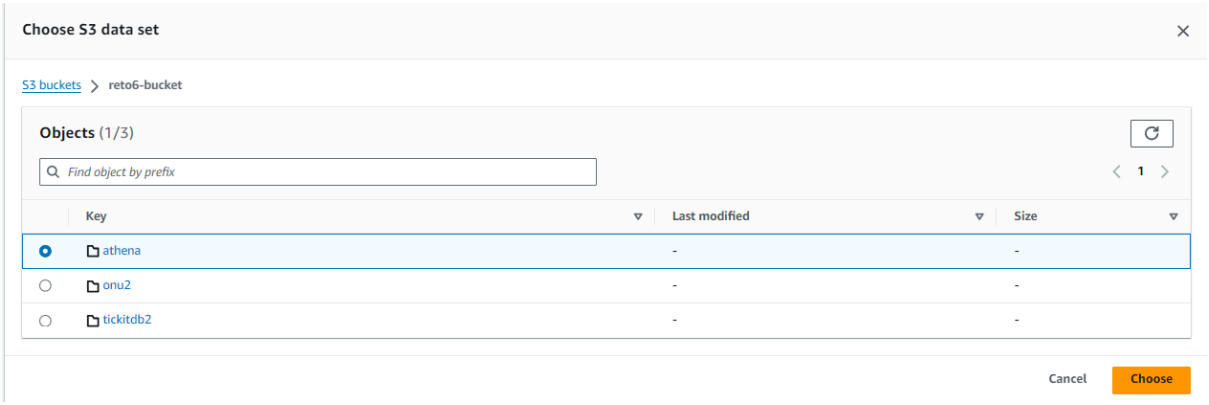
- Ahora seleccionamos el directorio de salida que creamos anteriormente en nuestro bucket llamado “Athena”



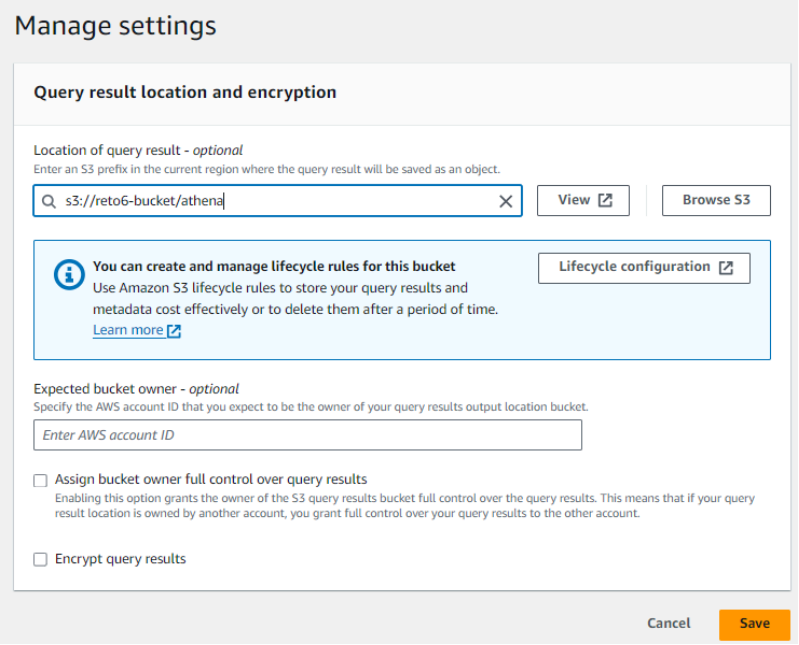
- Aca seleccionamos nuestro bucket con nuestro nombre en este caso “reto6-bucket”



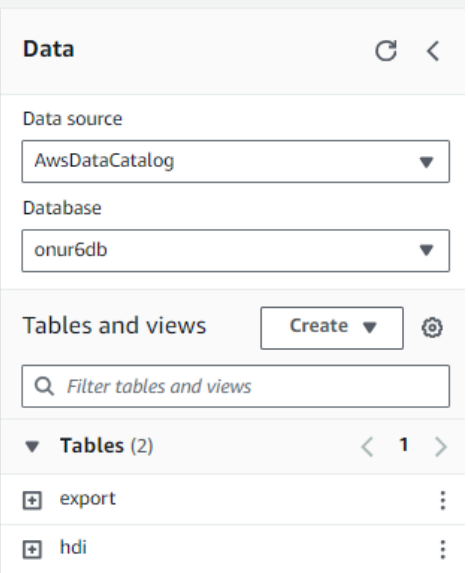
- Ahora seleccionamos la carpeta que creamos con el nombre “Athena” y daremos click en “choose”

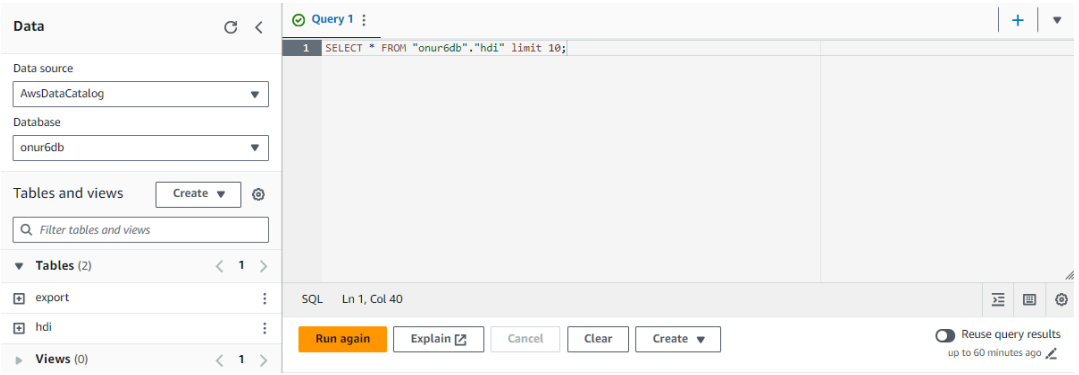


- Por último daremos click en “Save” cuando nuestras configuraciones de vean así

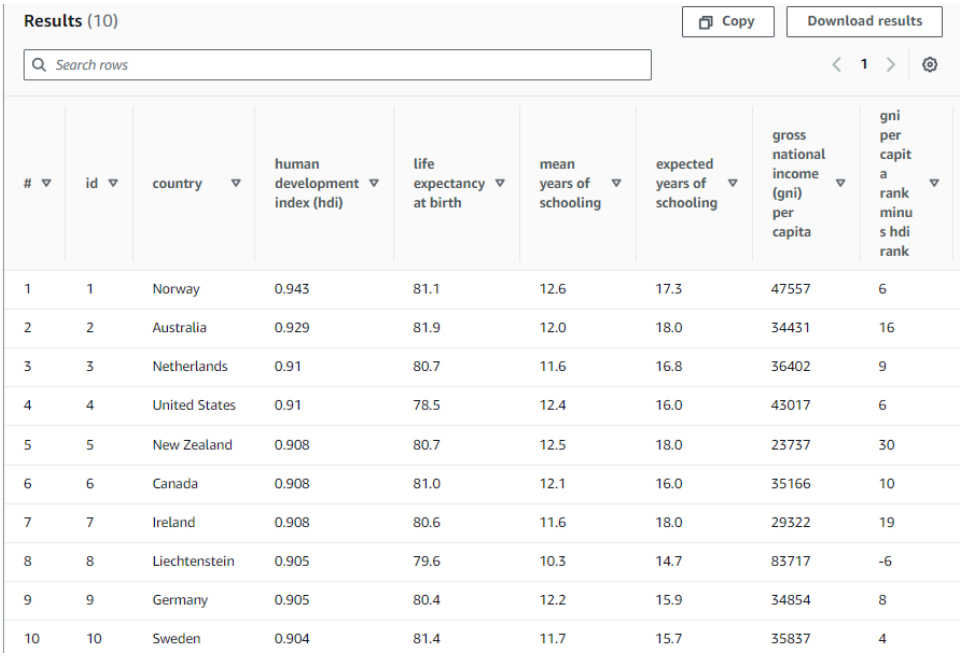


- Ahora volvemos a nuestro editor y seleccionamos la base de datos en la que haremos consultas y podremos empezar a realizar consultas con comandos SQL





El comando que utilice fue “SELECT * FROM "onur6db"."hdi" limit 10;”



Results (10)

Copy | Download results

Search rows

#	id	country	human development index (hdi)	life expectancy at birth	mean years of schooling	expected years of schooling	gross national income (gni) per capita	gni per capita rank minus hdi rank
1	1	Norway	0.943	81.1	12.6	17.3	47557	6
2	2	Australia	0.929	81.9	12.0	18.0	34431	16
3	3	Netherlands	0.91	80.7	11.6	16.8	36402	9
4	4	United States	0.91	78.5	12.4	16.0	43017	6
5	5	New Zealand	0.908	80.7	12.5	18.0	23737	30
6	6	Canada	0.908	81.0	12.1	16.0	35166	10
7	7	Ireland	0.908	80.6	11.6	18.0	29322	19
8	8	Liechtenstein	0.905	79.6	10.3	14.7	83717	-6
9	9	Germany	0.905	80.4	12.2	15.9	34854	8
10	10	Sweden	0.904	81.4	11.7	15.7	35837	4

Y podremos hacer lo mismo con nuestra otra base de datos de tickit