S3, Glue, Athena

Introducción

Este documento servirá como representación de las características generales y consideraciones de la api de nóminas en desarrollo apartir de los conocimientos y aplicación práctica de bases de datos NoSQL, visualizadores de datos y procesamiento de datos.

En la actualidad la cantidad de información en el mundo crece con gran velocidad, es por esto que su almacenamiento se ha visto limitado en cuanto a características como el volumen, velocidad y escalabilidad de la información, siendo esto el motivo principal para problemas relevantes como la pérdida de información y sobrecostos.

Teniendo en cuenta lo anterior, se puede afirmar que la información tiene un enorme poder dentro de muchas áreas, específicamente en la toma de decisiones en cualquier organización, puesto que proporciona una ventaja competitiva o una oportunidad de negocio.

Una posible solución al problema de almacenamiento para grandes volúmenes de datos y su consulta de manera rápida y eficiente es la incorporación de nuevos enfoques tecnológicos en cuanto al almacenamiento, tal enfoque es conocido como bases de datos NoSQL.

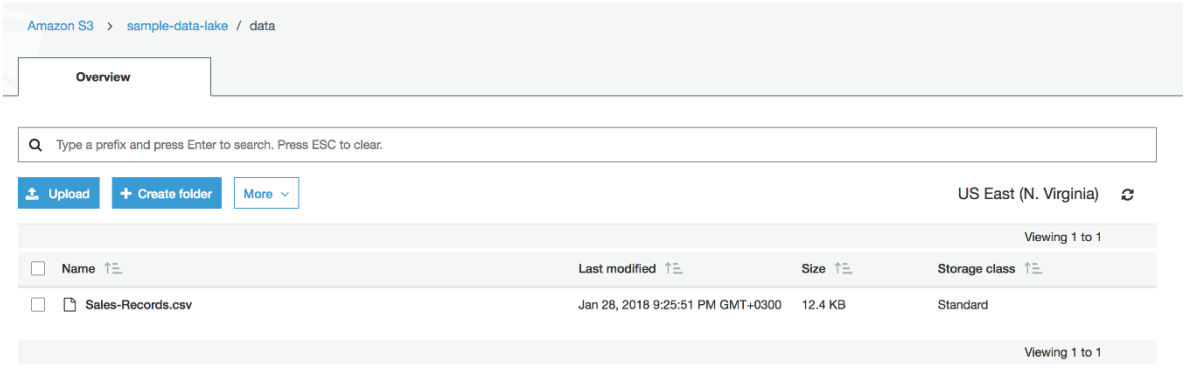
AWS

1. Crear una cuenta de AWS

Se debe tener una cuenta de AWS activa, es fácil de crear.

2. Cargar un archivo csv al S3

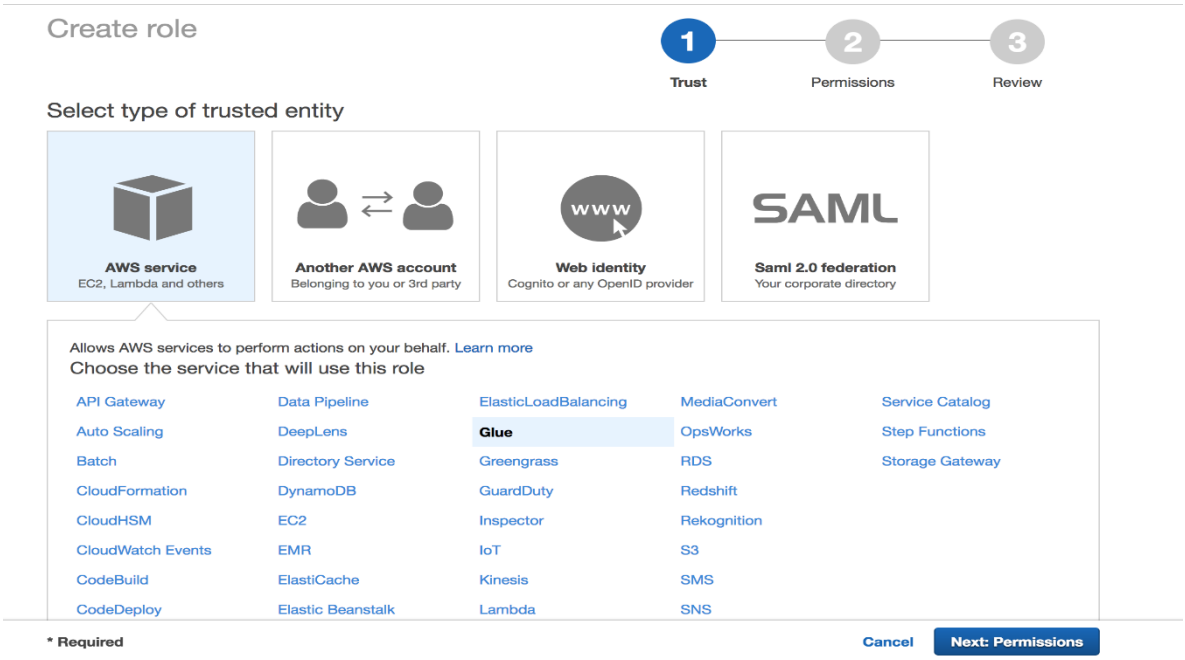
Encontré un csv de registros de ventas, se puede descargar y usar. (100 registros de ventas). Sube el archivo al S3 y crea un bucket. Presione el botón "Siguiente" para todos los pasos.



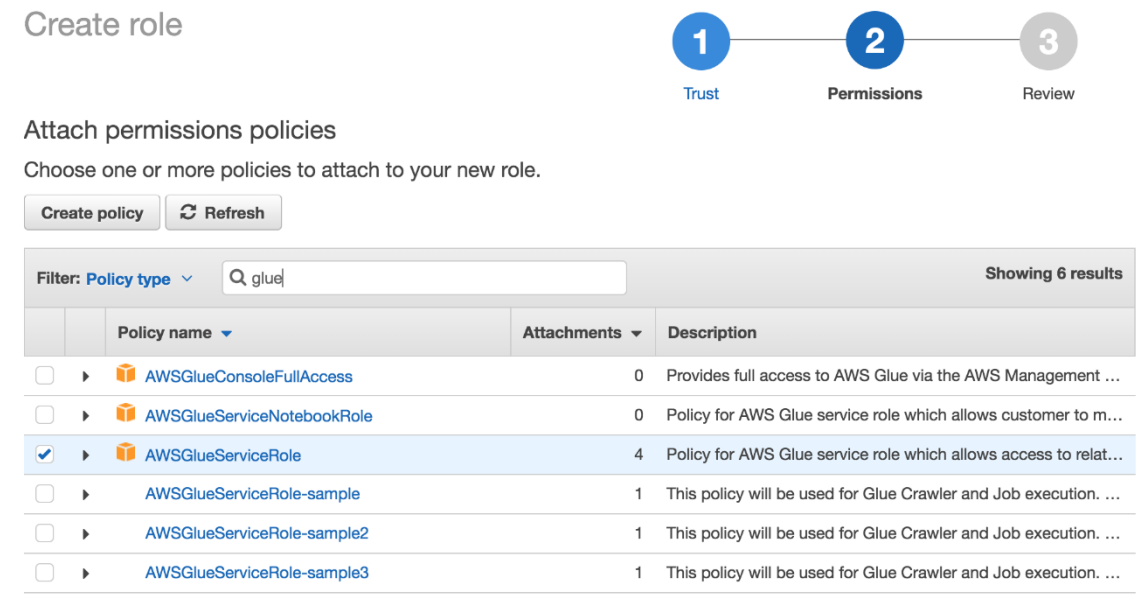
3. Crea un rol

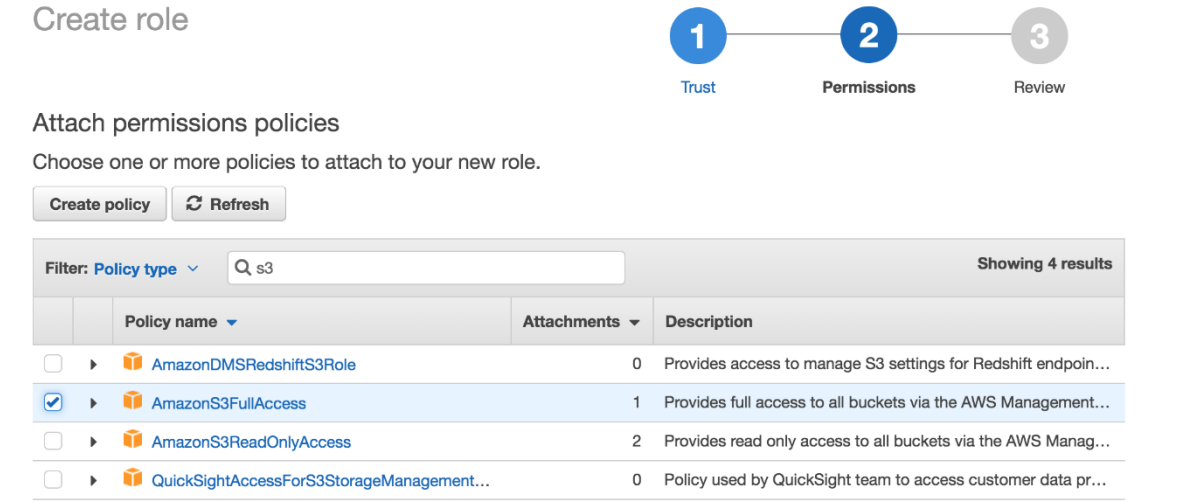
Ingresa a https://console.aws.amazon.com/iam/, click en "Roles" y luego en el botón azul "Crear rol".

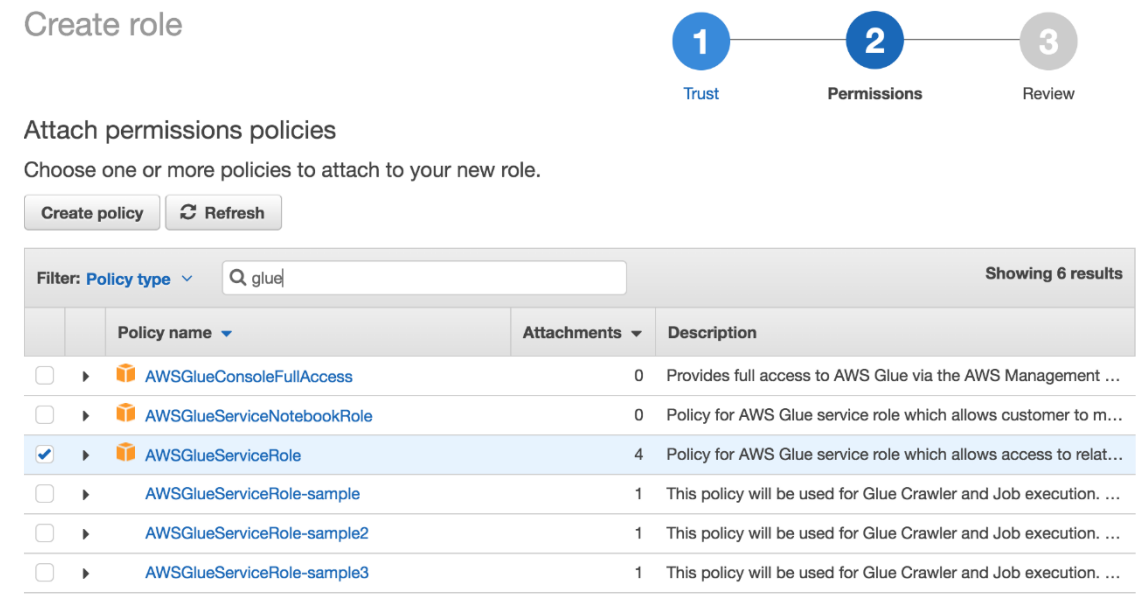
Elige "Glue" y luego.

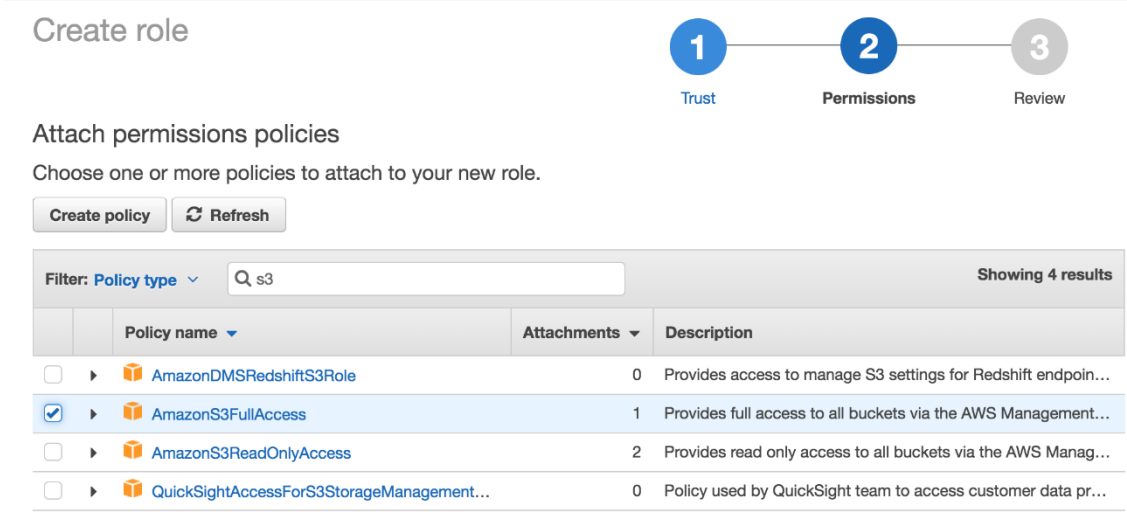


Se debe dar permisos a Glue y S3.









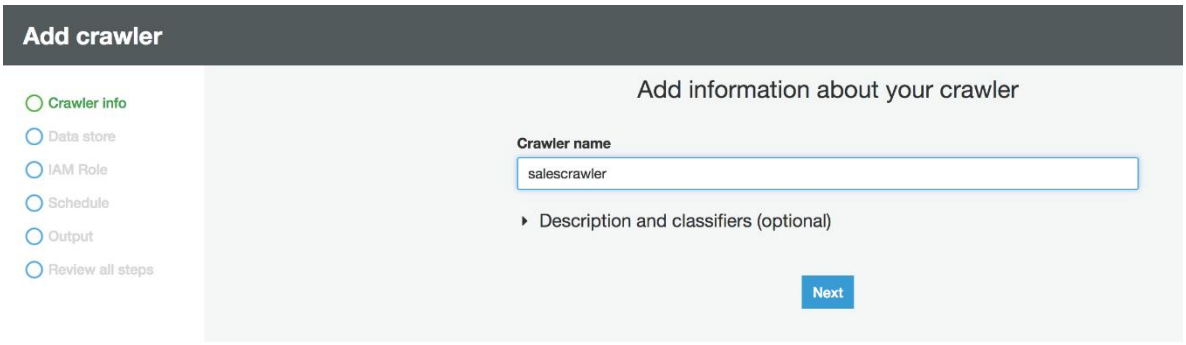
Ingrese un nombre de rol (AWSGlueServiceRoleDefault) y haz click en el botón "Crear rol".

4. Funcionamiento de Glue

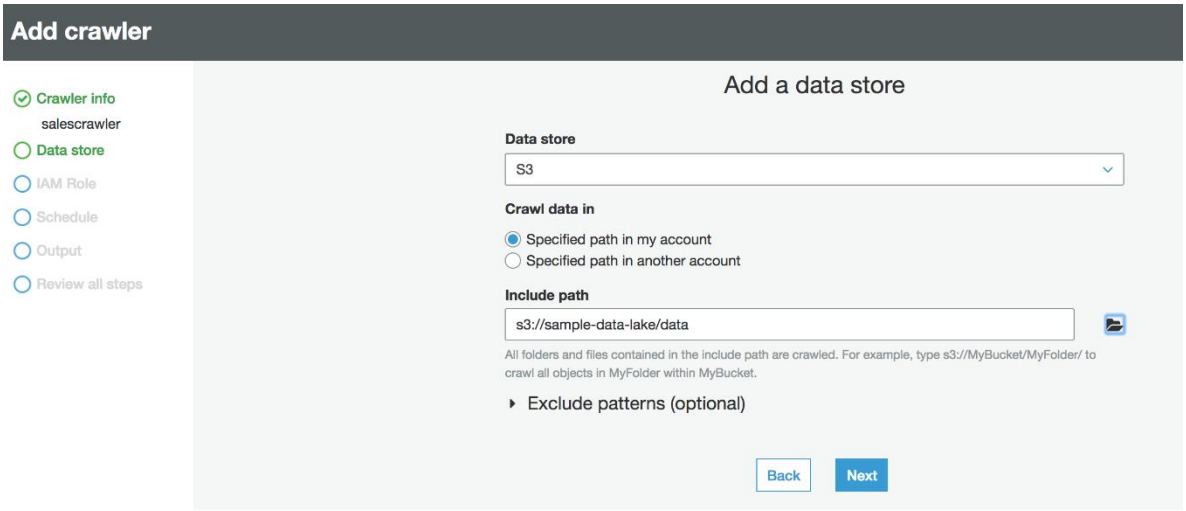
Agregar una nueva base de datos (sampledb)

Agregar una nueva tabla mediante la opción "Agregar tablas mediante un rastreador"

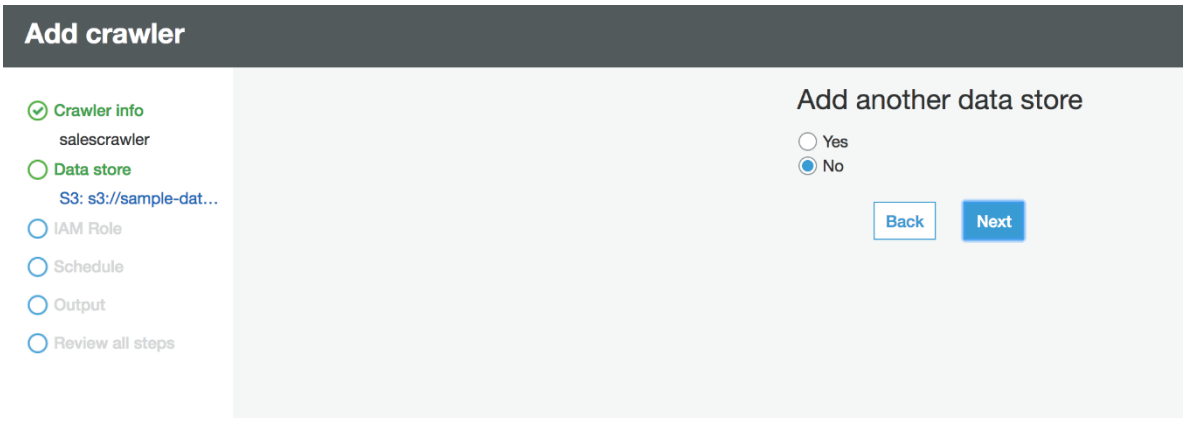
Ingrese el nombre del rastreador (rastreador de ventas) y haga clic en el botón Siguiente



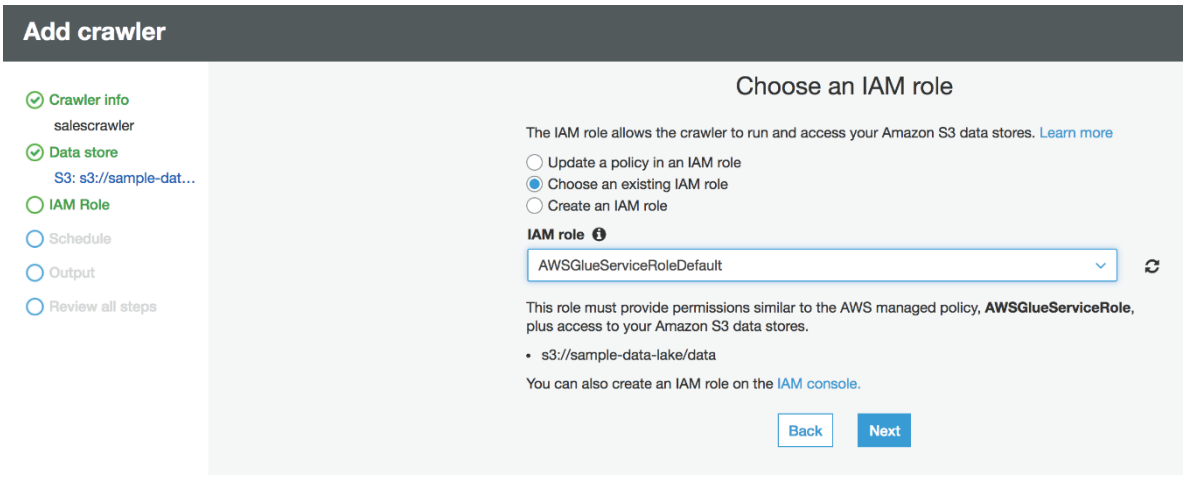
Ingresa la ruta del archivo csv que se almacenó en S3 y click en el botón Siguiente



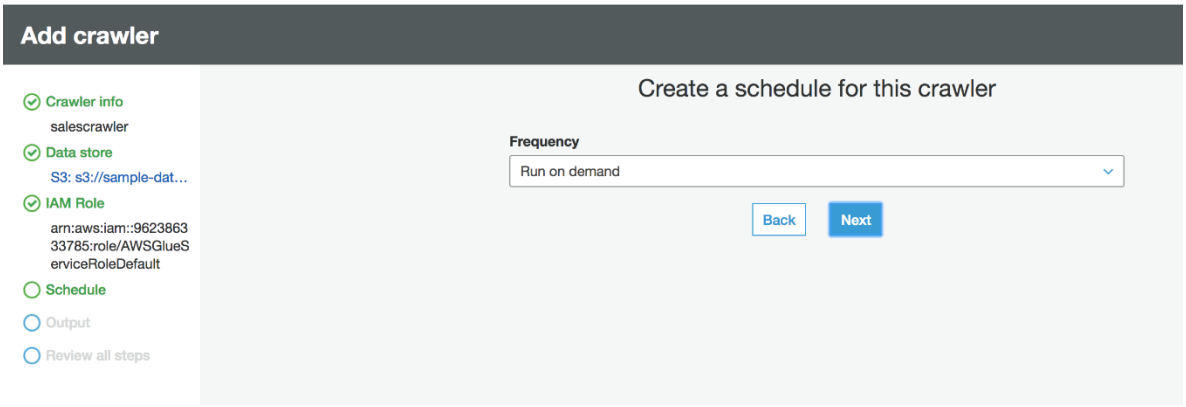
No necesitamos agregar un nuevo almacén de datos, selecciona "No", click en Siguiente



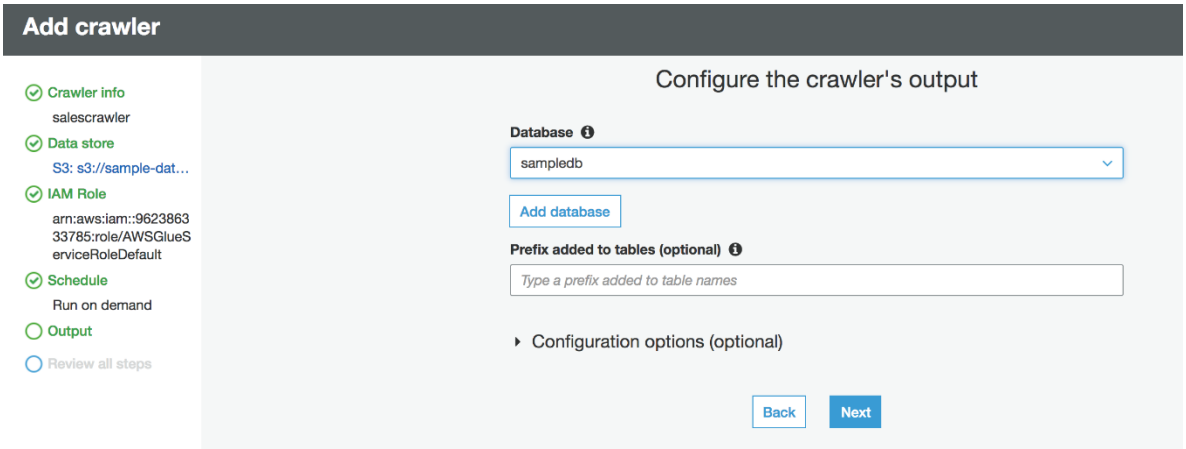
Selecciona el rol que se creó antes, click en Siguiente.

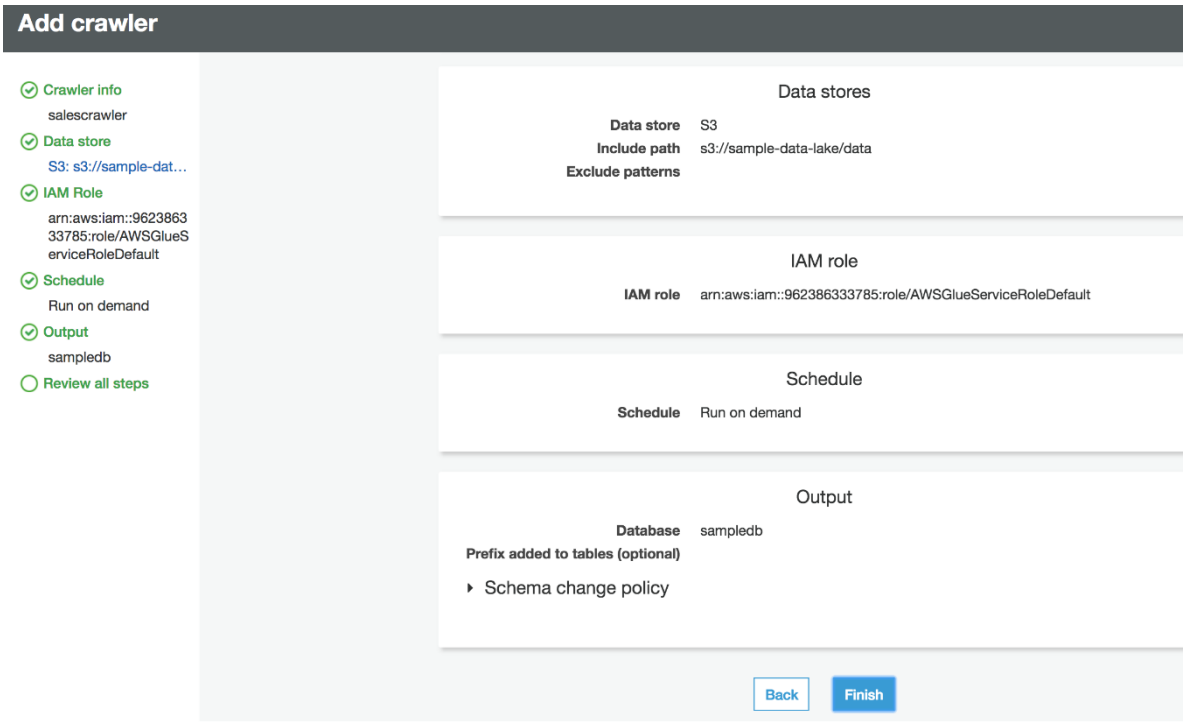


Ejecutar bajo demanda, Siguiente.

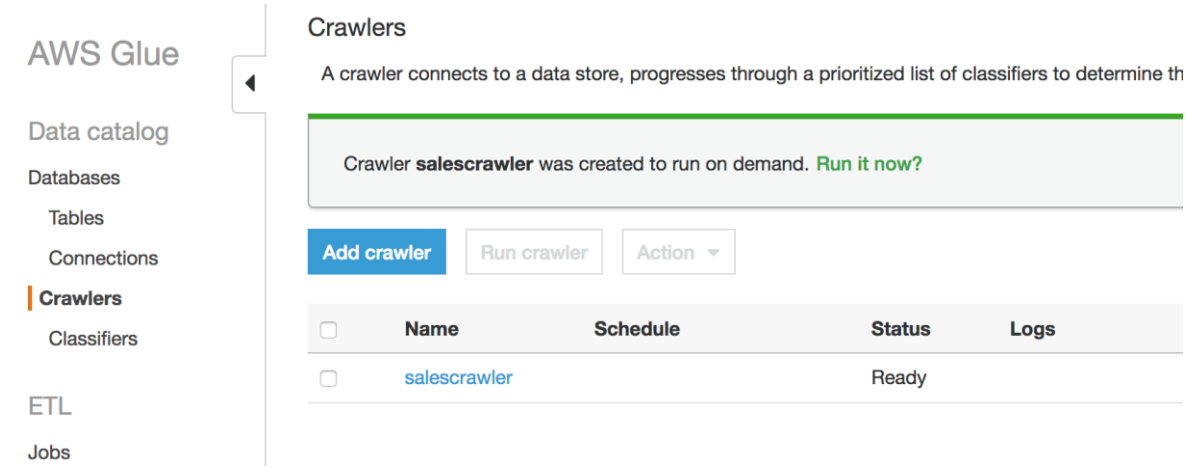


Selecciona "sampledb" y luego click en Siguiente

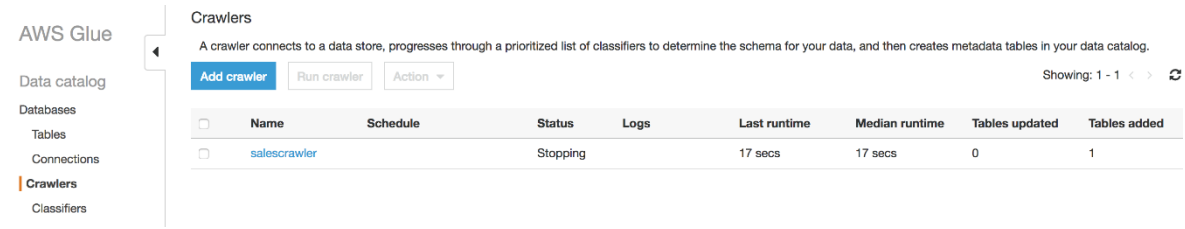




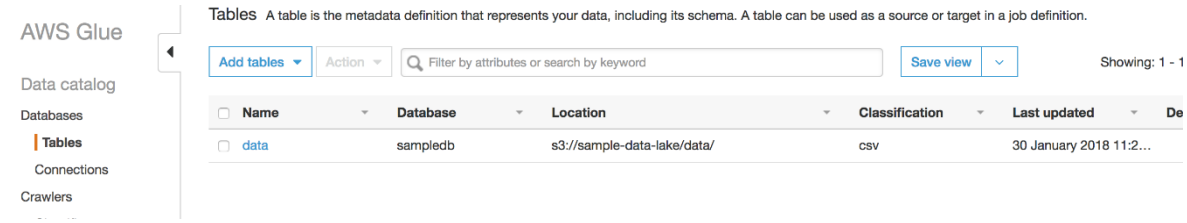
Se creó el crawler. Posteriormente se debe dar click en “Ejecutar ahora”. Al ejecutar el crawler, incorporaremos la información de Sales-Records.csv a Glue.



Una vez completado el crawler, en la columna "Tablas agregadas". Sales-Records.csv ya está en la plataforma Glue.



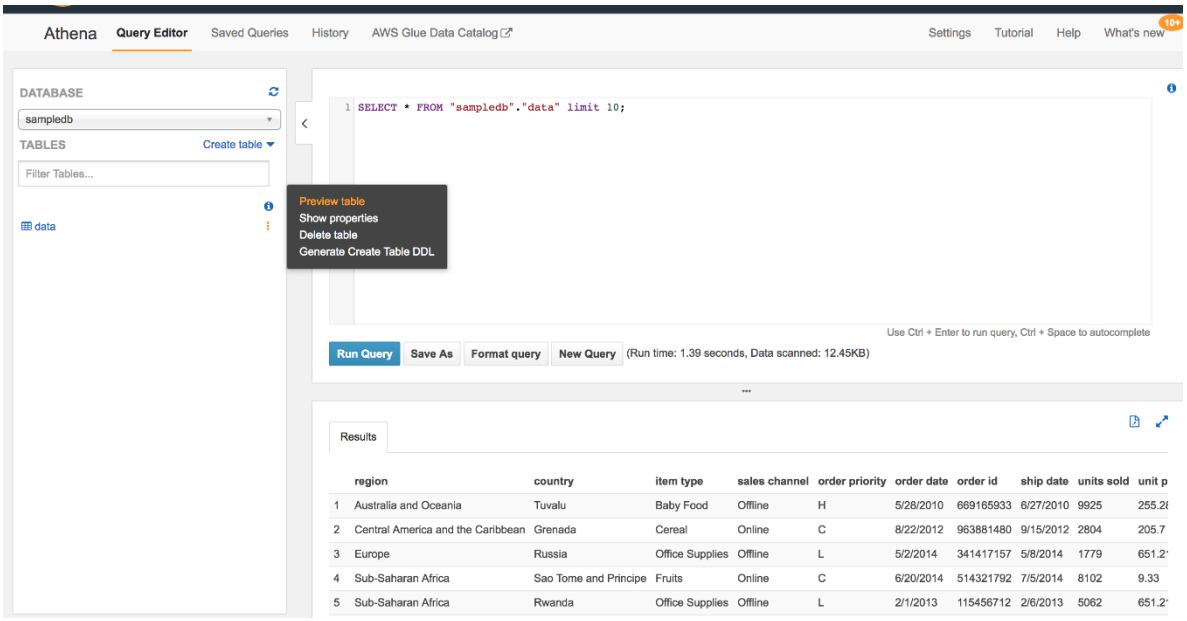
Puedes ver la tabla de "datos" en la sección Tablas.



5. Athena

Podemos consultar este csv usando Athena.

Con este ejemplo se listan 10 registros usando la opción "Vista previa de la tabla".



Conclusión

Podemos consultar un archivo csv usando S3, Glue y Athena. Debido a la arquitectura escalable de AWS, los costos de consulta son muy bajos.

Como cualquier desarrollo, este tipo de arquitectura y diseño tiene sus beneficios e inconvenientes, pero podría destacar:

1. No es necesario realizar mantenimiento de los servidores donde se tienen instalados programas y aplicaciones. El código se ejecuta en un contenedor temporal por lo que ya no es necesario instalar software, gestionar puertos de acceso o estar pendiente de las actualizaciones.

2. Se puede realizar un escalamiento de manera horizontal tanto como se requiera. Es posible añadir todos los clusters, balanceo de cargas etc, conforme se necesite.

3. Relacionado a los costos, solamente se va a pagar por el tiempo que se estén utilizando los procesos.

4. Las funciones que se utilizen, es posible integrarlas con el resto de servicios que ofrece la plataforma, como son logging, virtualización o endpoints.