AWS Glue

https://aws.amazon.com/es/glue/

Bloques de construcción de AWS Glue

AWS Glue Crawler permite rastrear datos de varias fuentes de datos, incluido S3, y bases de datos como Redshift, MySQL, MSSQL, Oracle y MongoDB, entre otras. Una vez completado el proceso de rastreo, AWS Glue Crawler crea o actualiza automáticamente las tablas en el catálogo de datos de AWS Glue. Se Encuentra una lista de las fuentes de datos compatibles en https://docs.aws.amazon.com/glue/latest/dg/crawler-data-stores.html:

Catálogo de datos de AWS Glue (AWS Glue Data Catalog): AWS Glue Data Catalog es un catálogo centralizado que almacena metadatos de varios almacenes de datos. Estos pueden ser utilizados por otros servicios de AWS como Amazon Athena, Amazon QuickSight, Amazon EMR y Amazon Redshift.

Activadores de AWS Glue (**AWS Glue Triggers**): Los activadores de AWS Glue permiten crear programaciones que pueden iniciar manual o automáticamente uno o varios rastreadores de AWS Glue o trabajos ETL de AWS Glue. Los activadores se pueden configurar para que se activen bajo demanda, en función de una programación o en función de una combinación de eventos.

Flujos de trabajo de AWS Glue (AWS Glue Workflows): Los flujos de trabajo de AWS Glue permiten orquestar la ejecución de Glue ETL y Glue Crawler mediante un activador de Glue.

AWS Glue ETL: AWS Glue ETL permite a los usuarios crear canalizaciones de extracción, transformación y carga (ETL) con Python, PySpark o Scala. Con Glue ETL, se pueden procesar datos de varias fuentes de datos, realizar transformaciones y, finalmente, escribir en diferentes fuentes de datos (como Amazon Redshift, Oracle, OpenSearch y Amazon S3). Tiene la opción de elegir entre diferentes motores de procesamiento: para el procesamiento distribuido, puede elegir entre PySpark y Scala, mientras que para el procesamiento no distribuido, puede seleccionar Python:

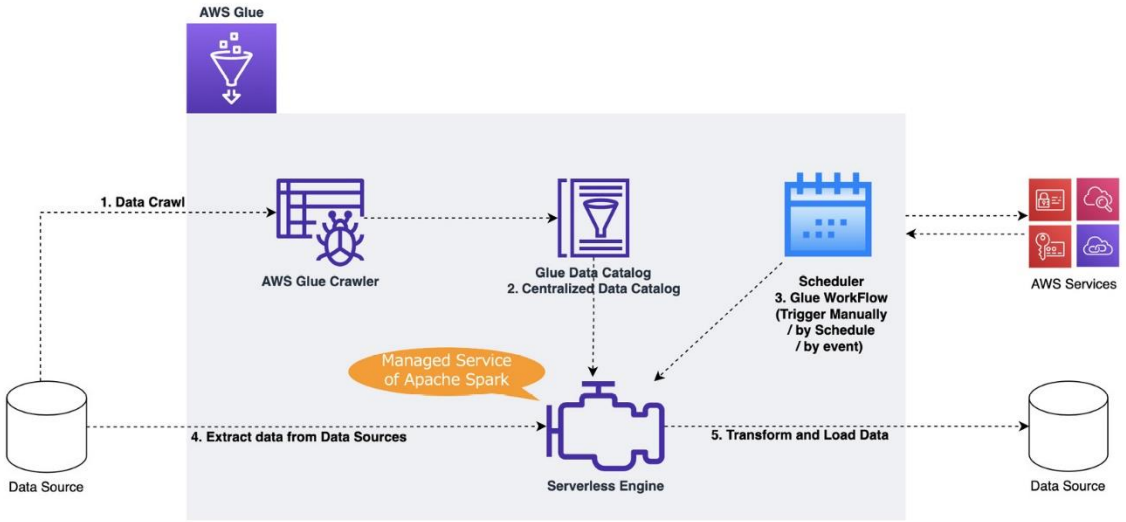


Figura 1 - Canalización de datos sencilla con AWS Glue

Sesiones interactivas de AWS Glue: Las sesiones interactivas de AWS Glue proporcionan un entorno de desarrollo integrado (IDE) basado en cuaderno Jupyter administrado para desarrollar y probar scripts ETL de AWS Glue.

Registro de esquemas de AWS Glue: AWS Glue Schema Registry es una herramienta que permite a los usuarios tener un control centralizado sobre los esquemas de transmisión de datos. Funciona a la perfección con plataformas de transmisión de datos populares como Apache Kafka, Amazon Kinesis y AWS Lambda. Con *AWS Glue Schema Registry*, se puede administrar y gobernar fácilmente los esquemas utilizados en los flujos de datos, garantizando la coherencia y la compatibilidad entre los diferentes componentes de una canalización de procesamiento de datos.

Conexiones de AWS Glue: *AWS Glue Connections* son objetos de catálogo que proporcionan una forma práctica para que los usuarios creen y almacenen información de conexión para diferentes almacenes de datos. Estas conexiones actúan como puente entre AWS Glue y varias fuentes de datos. Con *AWS Glue Connections*, se puede configurar y administrar fácilmente los detalles de conexión necesarios para nuestros almacenes de datos.

Además de los almacenes de datos estándar, AWS Glue Connections también admite Marketplace AWS Glue Connectors. Estos conectores permiten la integración con almacenes de datos de terceros como Apache Hudi, Google Big Query y Elasticsearch. Esto significa que se puede conectar e interactuar sin problemas con estas fuentes de datos externas utilizando AWS Glue, lo que facilita el acceso y el procesamiento de datos de diversos sistemas de manera unificada.

Descubrimiento de datos con AWS Glue

Una de las características exclusivas que diferencia a AWS Glue de otras herramientas ETL es su capacidad para crear un catálogo de datos centralizado. Este catálogo es crucial para realizar la detección de datos y se basa en dos componentes importantes de Glue:

- Catálogo de datos de Glue
- Rastreador de datos de Glue

Catálogo de datos de AWS Glue

Un catálogo de datos es un almacenamiento centralizado de metadatos para datos almacenados en diferentes almacenes de datos, como lagos de datos, almacenes de datos, bases de datos relacionales y bases de datos no relacionales. Los metadatos contienen información sobre columnas, formatos de datos, ubicaciones y mecanismos de serialización/deserialización. *Hive Metastore* es uno de los productos de metadatos más utilizados en el sector. Sin embargo, utiliza sistemas de gestión de bases de datos relacionales (RDBMS) como MySQL y PostgreSQL. El problema de utilizar un RDBMS para los metadatos de Hive es su gestión y mantenimiento, especialmente para cargas de trabajo de producción en las que hay que tener en cuenta la alta disponibilidad, el escalado y la redundancia. Esto aumenta la complejidad y el coste de la solución. AWS Glue Data Catalog resuelve este problema con su naturaleza sin servidor, eliminando la necesidad de una sobrecarga administrativa adicional en la administración de la infraestructura.

AWS Glue Data Catalog se compone de lo siguiente:

- Bases de datos
- Tablas
- Particiones
- Índices de partición

Bases de datos

En AWS Glue, una base de datos es una agrupación de tablas de metadatos. Cada vez que se crea una tabla en el catálogo de datos de AWS Glue, debe asociarse a una base de datos específica. No es posible tener la misma tabla presente en varias bases de datos:

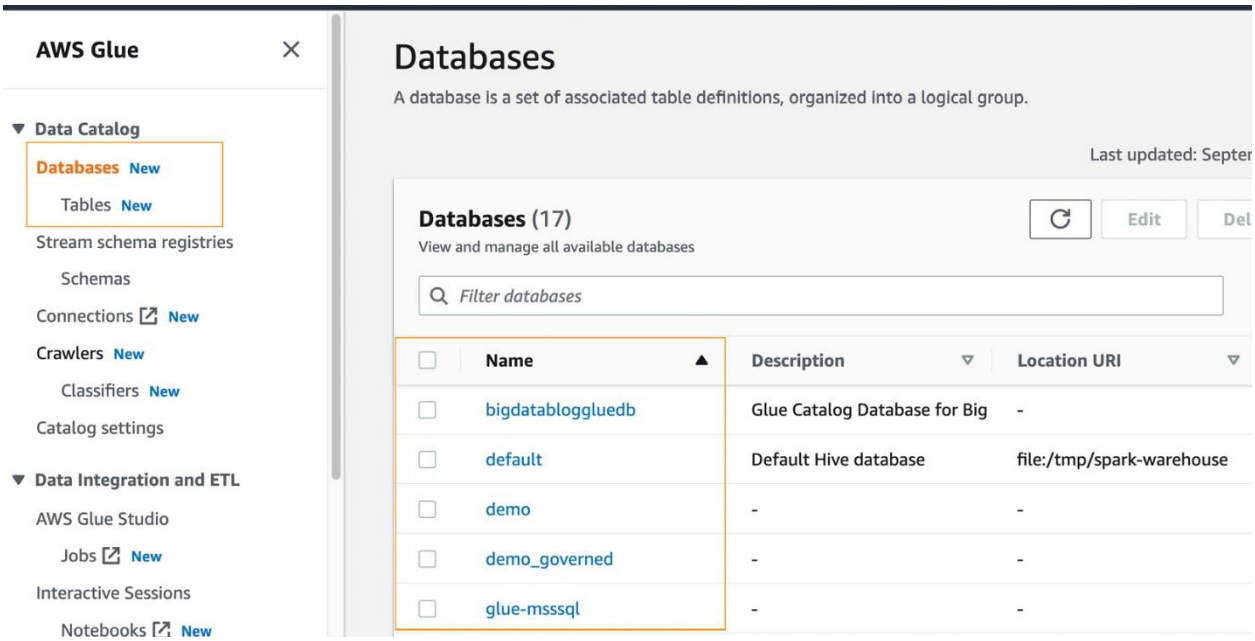


Figura 2 - Bases de datos en el catálogo de datos de AWS Glue

Tablas

Una tabla del Catálogo de Datos de AWS Glue es una representación de metadatos de sus datos. Contiene diversa información sobre sus datos, como su nombre, descripción, ubicación, formato de entrada/salida, bibliotecas de serialización, esquema, particiones, índices, etc.

AWS Glue Data Catalog admite el versionado de tablas, lo que significa que cada vez que se actualiza una tabla, se crea una nueva versión de la tabla de forma predeterminada. Se trata de una característica crucial para identificar cualquier cambio de esquema en el origen y evitar fallos de aplicaciones posteriores debidos a dichos cambios. En esta entrada de blog, que explica cómo detectar cualquier cambio en las estructuras de una fuente de datos utilizando AWS Glue: https://aws.amazon.com/blogs/big-data/identify-source-schema-changes-using-aws-glue/

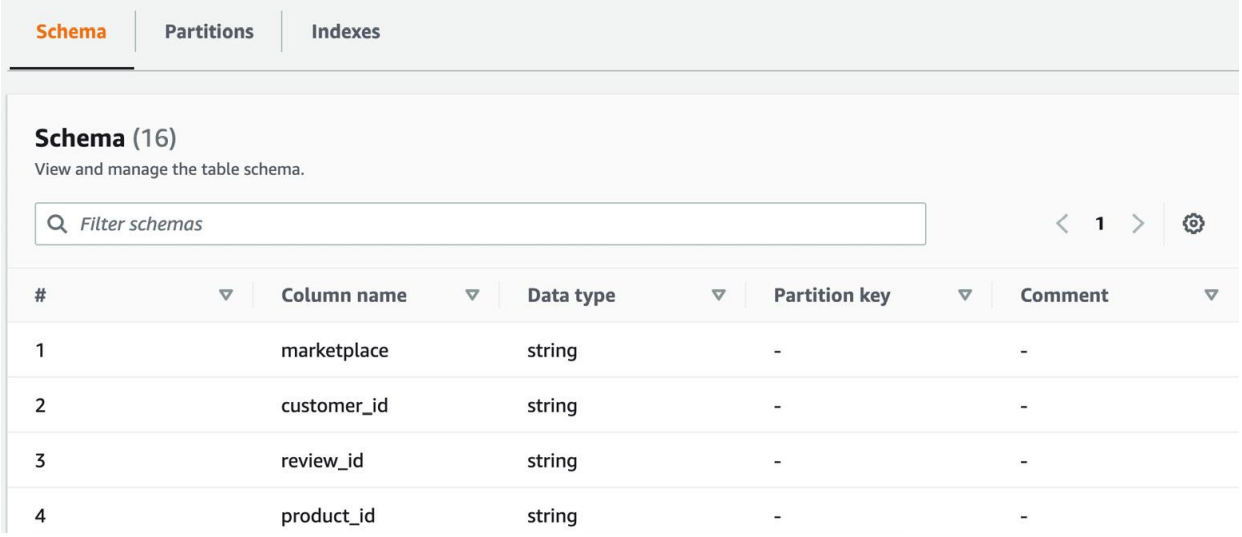


Figura 3 – Propiedades de una Tabla en el catálogo de datos de AWS Glue

Particiones

El particionamiento es una estrategia que se utiliza para optimizar el rendimiento de las consultas dividiendo una tabla en función de los valores de columnas específicas. Una tabla puede tener varias claves de partición que definen una partición específica (también conocida como especificación de partición). Esta técnica mejora el rendimiento de la consulta al reducir la cantidad de datos que deben procesarse para responder a una consulta específica:

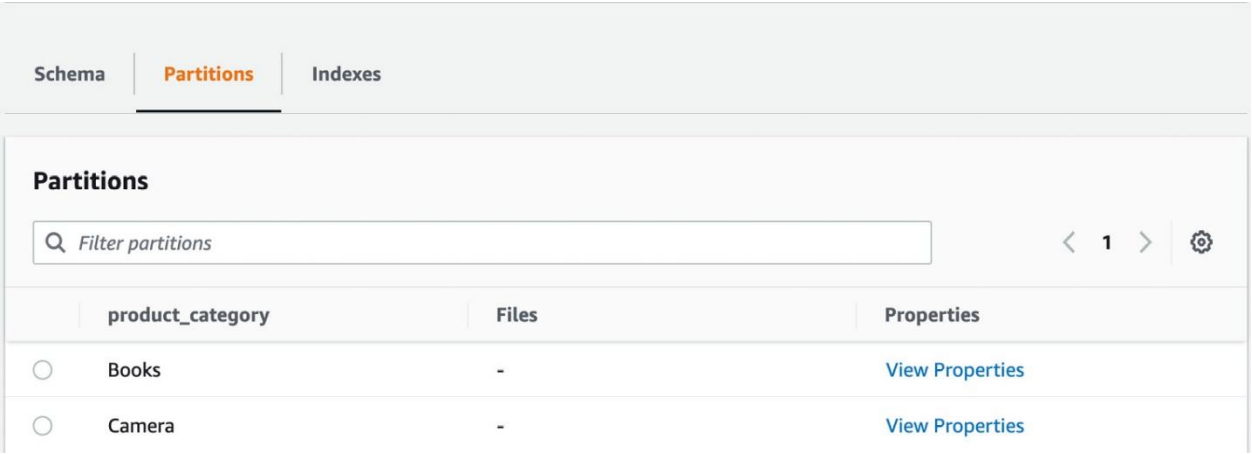


Figura 4 - Partición en una tabla del catálogo de datos de AWS Glue

Conexiones Glue

AWS Glue proporciona *Glue Connections* para permitirle conectarse a diferentes fuentes de datos. Glue Connections almacena información importante como credenciales de inicio de sesión, cadenas de conexión (URI) y detalles de configuración de la VPC (subred de la VPC e información del grupo de seguridad), todo lo cual es necesario para establecer una conexión con una fuente de datos:

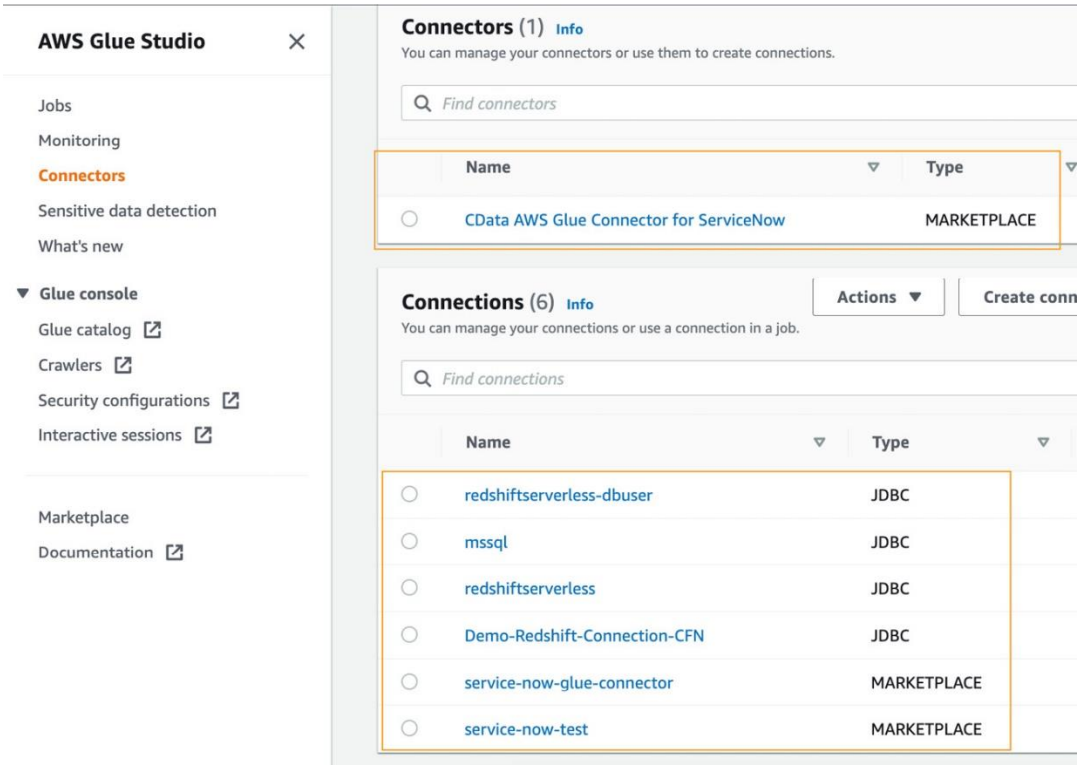


Figura 5 - Conexiones de AWS Glue

Rastreadores de AWS Glue (*AWS Glue crawlers*)

Un rastreador de Glue nos permite escrutar los datos de diferentes tipos de almacenes de datos, inferir el esquema y rellenar AWS Glue Data Catalog con los metadatos del conjunto de datos rastreado:

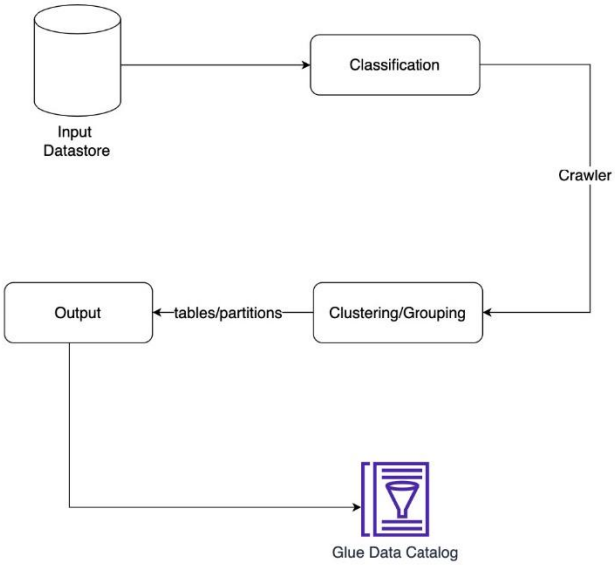


Figura 6 Flujo de trabajo de un rastreador Glue

El flujo de trabajo de Glue Crawler consta de varias etapas:

Clasificación:

En AWS Glue, un clasificador se encarga de examinar los datos de un almacén de datos y determinar su formato. Una vez reconocido el formato, el clasificador genera el esquema apropiado para los datos.

Para realizar esta tarea, Glue proporciona clasificadores integrados y personalizados. Cuando se crea y ejecuta un rastreador Glue, éste aplica en primer lugar los clasificadores personalizados que se especifiquen en el orden definido en la configuración del rastreador. Estos clasificadores personalizados tienen prioridad durante el proceso de reconocimiento de esquemas.

El rastreador también asigna un número de certeza para indicar el nivel de confianza en el reconocimiento del formato. Un valor de certeza de 1,0 significa que el rastreador está 100% seguro de crear el esquema correcto.

Si ninguno de los clasificadores devuelve un valor de certeza de 1,0, el rastreador selecciona el clasificador con el valor de certeza más alto. Sin embargo, si ninguno de los clasificadores devuelve un valor de certeza superior a 0,0, AWS Glue devuelve una cadena de clasificación predeterminada de UNKNOWN, lo que indica que no se ha podido determinar el formato con confianza.

Agrupación/clustering:

La salida de la etapa de clasificación es utilizada por el rastreador y los datos se agrupan en función del esquema del rastreador, la clasificación y otras propiedades.

Salida:

En esta etapa, la API Glue Data Catalog Glue escribe los objetos de tabla o partición que se crearon en la etapa de agrupación. Para evitar la duplicación de tablas, el rastreador crea una tabla con un sufijo de cadena *hash* si la tabla o tablas ya existen desde la ejecución anterior.

Funcionamiento interno de un rastreador para Amazon S3

El flujo de trabajo de Glue Crawler se divide en varias etapas. El rastreador se utiliza principalmente para rastrear datos de Amazon S3. Por defecto, el rastreador lee todos los datos de la ruta S3 especificada. Clasifica cada archivo disponible en S3 y mantiene los metadatos en el almacenamiento del lado del servicio del rastreador, que es diferente del *Catálogo de datos* de AWS Glue y no está expuesto al cliente. En ejecuciones posteriores, el rastreador reutiliza los metadatos almacenados y solo rastrea los archivos recién añadidos. Esto reduce el tiempo de ejecución del rastreador en ejecuciones posteriores. Si se carga una nueva versión del mismo archivo, el rastreador lo trata como un archivo nuevo y lo incluye en el nuevo rastreo.

El rastreador muestrea los datos en función de su tamaño y lee los primeros de 1 a 10 MB de datos. Si el tamaño del archivo es superior a 10 MB, se asegura de leer al menos un registro.

Para JDBC, Amazon DynamoDB y Amazon DocumentDB, el flujo de trabajo del rastreador es el mismo, pero las tablas lógicas se deciden en función del motor de base de datos. Para JDBC, los rastreadores de Glue se conectan al servidor de base de datos y extraen el esquema de la base de datos que coincide con el valor de ruta incluido en la configuración del rastreador.

Veamos un ejemplo de rastreo de datos y acceso a los datos rastreados mediante Amazon Athena. En este ejemplo, haremos lo siguiente:

- Explorar el conjunto de datos de código abierto Covid-19.
- Crear una base de datos Glue Catalog.
- Crear un rastreador Glue para rastrear el conjunto de datos Covid-19.
- Utilizar Athena para consultar el conjunto de datos.

Para ilustrar este ejemplo, utilizaremos el conjunto de datos Covid-19, que puede encontrarse en Open Data en AWS:

Vamos a la página del Registro de Datos Abiertos en AWS (https://registry.opendata.aws/). En el cuadro de búsqueda *Search datasets*, buscamos *COVID-19 Data Lake* como cláusula de búsqueda. Se deberán ver un conjunto de datos, COVID-19 Data Lake, como resultado.

Haz clic en el enlace Detalles (https://registry.opendata.aws/aws-covid19-lake/) de la página; verás la información de recursos de AWS sobre cómo acceder al conjunto de datos

Se puede acceder a los datos utilizando la CLI de AWS (https://aws.amazon.com/cli/) desde la ubicación de Amazon S3:

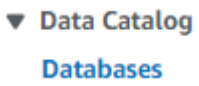
```
aws s3 ls --no-sign-request s3://covid19-lake/
```

Para nuestro ejemplo, queremos utilizar un subconjunto del conjunto de datos Covid-19:

```
aws s3 ls --no-sign-request s3://covid19-lake/rearc-covid-19-testing-data/
```

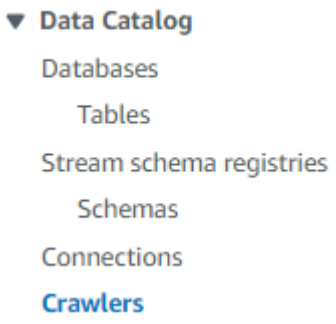
Para continuar con este ejemplo, el siguiente paso es crear una base de datos Glue Catalog.

Vamos a la consola de AWS Glue (https://console.aws.amazon.com/glue/) y hacemos clic en *Databases* a la izquierda.



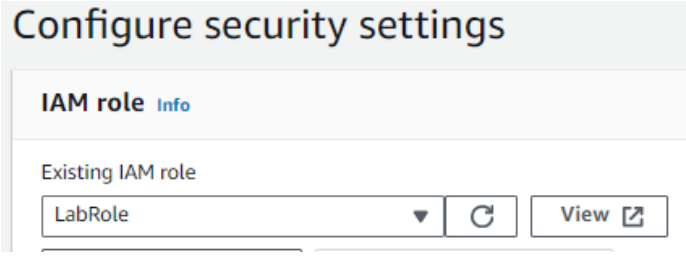
Creamos una base de datos llamada **covid19** haciendo clic en *Add Database*, y luego en *Create*

Para continuar con este ejemplo, el siguiente paso es crear un rastreador Glue.



Hacemos clic en *Crawlers* a la izquierda y luego en *Crear rastreador*.

- En la página Establecer propiedades del rastreador, proporcionamos un nombre para el nuevo rastreador, como *covid19-gluecrawler*, y hacemos clic en *Siguiente*.
- En la página *Elegir orígenes de datos y clasificadores*, seleccionamos *Todavía no* en Configuración de orígenes de datos
- Hacer clic en *ADD data source*, seleccionamos *En una cuenta diferente en Ubicación de los datos de S3*, establece s3://covid19-lake/rearc-covid-19-testing-data/ en *Ruta de S3* y hacemos clic en *Añadir un origen de datos de S3*. A continuación, hacemos clic en Siguiente y mantenemos el resto de las opciones en sus valores predeterminados.
- En la página *Configurar ajustes de seguridad*, seleccionamos el rol ***Labrole*** *(esto es específico para AWS Academy, en una cuenta normal crearíamos un rol específico para ello)*. A continuación, hacemos clic en *Crear* y luego en *Siguiente*.



- En la página *Set output and scheduling*, en la sección *Output configuration*, seleccionanmos ***covid19*** de la lista desplegable Database. Introducimos ***covid19_*** como prefijo que se añade a las tablas (opcional) y, en Programación del rastreador, mantenemos la frecuencia establecida en bajo demanda.

Revisamos todas las configuraciones y hacemos clic en Crear rastreador:

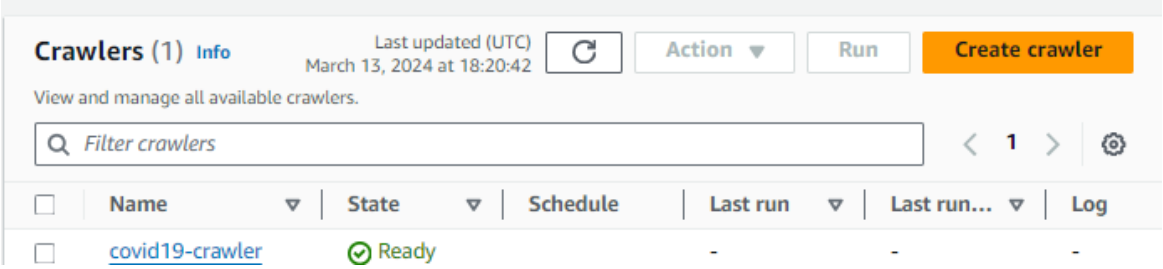


Figura 7 - AWS Glue - Crawlers

Ahora que hemos creado el rastreador, hacemos clic en la casilla junto a él y se elegimos ejecutar el rastreador haciendo clic en el botón *Ejecutar* en la parte superior de la página. El rastreador tardará uno o dos minutos en finalizar.

Cuando el rastreador se complete con éxito, listará cuatro tablas. Vayamos a la base de datos y exploremos la definición de la tabla.

Observando la ubicación y clasificación de la tabla, se ve que apunta al conjunto de datos Covid-19 y extrae la definición de la base de datos para diferentes tipos de archivo, CSV y JSON. Como hemos visto, Glue Crawler te permite rastrear datos de diferentes cuentas de AWS, lo que es muy útil cuando se construye un catálogo de datos centralizado donde los datos se almacenan en diferentes cuentas de AWS:

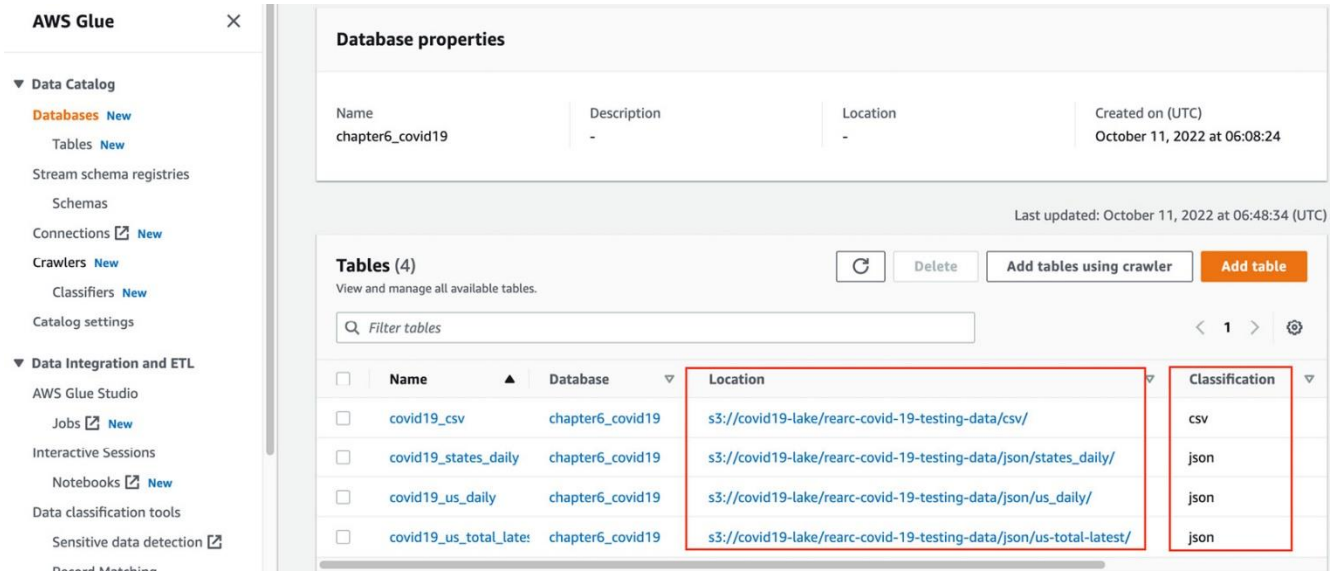


Figura 9 - Tablas del catálogo de datos de AWS Glue

Haciendo clic en el nombre de la tabla - en este caso, 'covid19_csv' - puedes explorar la definición de la tabla; esto proporciona información valiosa sobre la tabla, incluyendo estadísticas y metadatos:

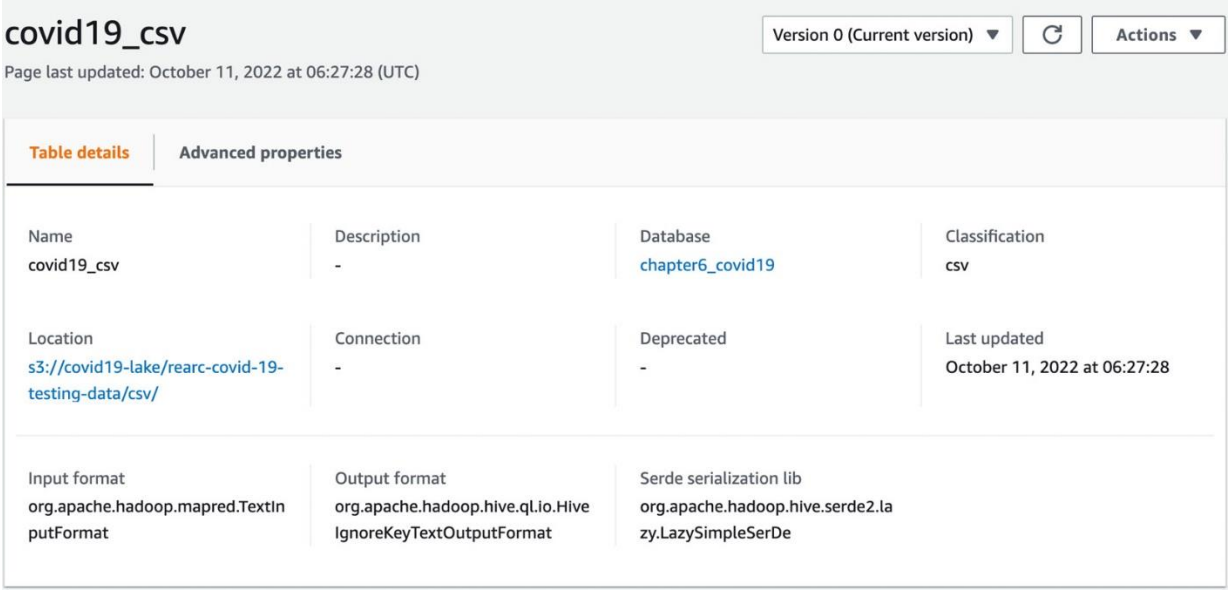


Figura 10 - Propiedades de la tabla AWS Glue Catalog

Estadísticas de tabla

En AWS Glue, las estadísticas de tabla se refieren a la información estadística sobre una tabla en el *Catálogo de Datos Glue*. Estas estadísticas incluyen el número de valores totales y distintos en cada columna, los valores mínimos y máximos, y la longitud media de los valores. Esta información es importante para la optimización de consultas y el ajuste del rendimiento:

Table details | Advanced properties

Serde parameters (1)

Key	Value
field.delim	,

Table properties (14)

Key	Value
CrawlerSchemaDeserializerVersion	1.0
CrawlerSchemaSerializerVersion	1.0
UPDATED_BY_CRAWLER	chapter-gluecrawler
areColumnsQuoted	false
averageRecordSize	482
classification	csv
columnsOrdered	true
compressionType	none
delimiter	,
objectCount	3
recordCount	11920
sizeKey	5763710
skip.header.line.count	1

Figura 11 - AWS Glue Catalog - Propiedades avanzadas

Las estadísticas de tabla pueden verse y actualizarse a través de la consola de AWS Glue o la API de Glue. También pueden ser recopiladas automáticamente por AWS Glue Crawler durante el proceso de descubrimiento de datos.

Las estadísticas de tabla en el catálogo de datos de Glue proporcionan metadatos importantes sobre las tablas almacenadas en su catálogo de datos. Estas estadísticas incluyen información como el número de filas de una tabla, el tamaño de la tabla y las estadísticas de columna, como los valores mínimo y máximo, la distribución de datos y el tipo de datos. Estas estadísticas desempeñan un papel crucial en la optimización de consultas y el ajuste del rendimiento en AWS Glue.

AWS Glue y los motores de consulta como Amazon Athena o Amazon Redshift Spectrum utilizan estas estadísticas de tabla para tomar decisiones informadas sobre los planes de ejecución de consultas. A continuación, se explica por qué las estadísticas de tabla son importantes y pueden aprovecharse para la optimización de consultas y el ajuste del rendimiento:

Planificación de consultas: El optimizador de consultas utiliza las estadísticas de tabla para estimar la cardinalidad (número de filas) y la distribución de datos de las tablas implicadas en una consulta. Esta estimación ayuda al optimizador a elegir el plan de ejecución de consultas más eficiente. Unas estadísticas precisas permiten al optimizador tomar mejores decisiones, lo que se traduce en un procesamiento más rápido y eficaz de las consultas.

Optimización de uniones: Cuando se unen varias tablas, el optimizador de consultas utiliza las estadísticas de las tablas para determinar la estrategia de unión óptima. Analizando las estadísticas de las

columnas, como la distribución de los datos y los rangos de valores, el optimizador puede decidir si realizar una unión *hash*, una unión *merge* u otros algoritmos de unión. Esta información ayuda a minimizar la mezcla de datos y a mejorar el rendimiento general de las uniones.

Despliegue de predicados: Las estadísticas de tabla también se utilizan para el *pushdown* de predicados, que consiste en enviar las operaciones de filtrado a la fuente de datos antes de introducir los datos en el motor de consultas. Al conocer la distribución de los datos y los rangos de valores de las columnas, el optimizador puede determinar si el filtrado en la fuente de datos reduciría significativamente el volumen de datos que hay que procesar, lo que aceleraría la ejecución de la consulta.

Detección de datos sesgados: Las estadísticas de tabla pueden ayudar a identificar la inclinación de los datos, cuando la distribución de los datos entre particiones o columnas es desigual. La desviación de los datos puede afectar negativamente al rendimiento de las consultas, haciendo que algunas tareas tarden mucho más que otras. Mediante el análisis de las estadísticas de tabla, puede detectar la inclinación de los datos y tomar medidas correctivas, como redistribuir los datos o utilizar diferentes estrategias de partición para lograr un mejor equilibrio de la carga y mejorar el rendimiento de la consulta.

Para garantizar la precisión de las estadísticas de las tablas, es importante actualizarlas periódicamente a medida que cambian los datos. AWS Glue proporciona una funcionalidad para recopilar automáticamente las estadísticas de las tablas mediante el uso de rastreadores.

Schema | Partitions | Indexes

Schema (60)

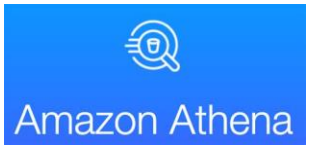
View and manage the table schema.

Filter schemas

#	Column name	Data type	Partition key	Comment
1	date	bigint	-	-
2	state	string	-	-
3	positive	bigint	-	-
4	probablecases	bigint	-	-
5	negative	bigint	-	-
6	pending	bigint	-	-
7	totaltestresultssource	string	-	-
8	totaltestresults	bigint	-	-
9	hospitalizedcurrently	bigint	-	-
10	hospitalizedcumulative	bigint	-	-
11	inicucurrently	bigint	-	-
12	inicucumulative	bigint	-	-
13	onventilatorcurrently	bigint	-	-
14	onventilatorcumulative	bigint	-	-
15	recovered	bigint	-	-

Figura 12 - Definición del esquema de tabla del catálogo de AWS Glue

Amazon Athena



Una vez explorada la definición de la tabla, ya podemos revisar los datos utilizando Amazon Athena.

Haciendo clic en *Table data* nos redirigirá a la consola de Athena:

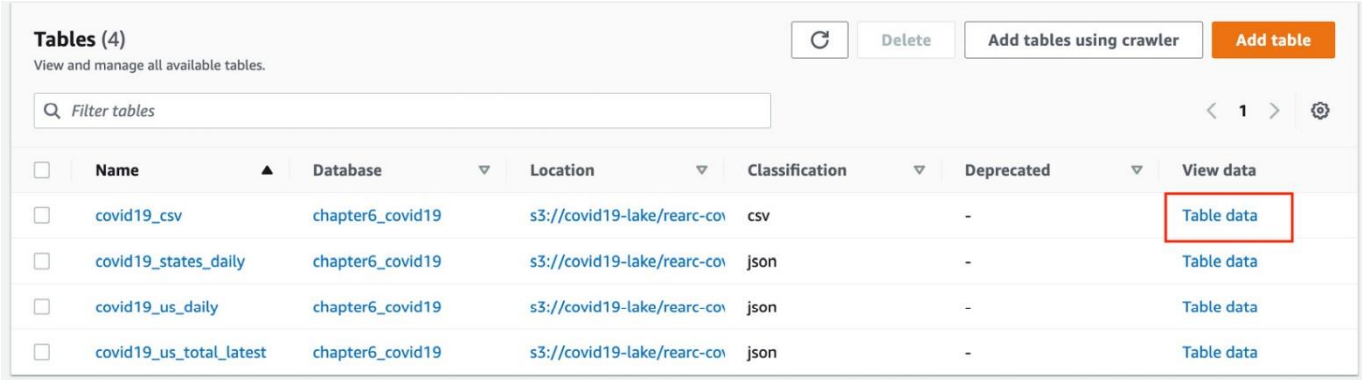


Figura 13 - Lista de tablas en AWS Glue Data Catalog

Aquí, accedimos a los datos de la tabla *covid19_csv* utilizando Amazon Athena. Al hacer clic en *Table data*, fuimos redirigidos a la consola de Athena, donde podemos ver los datos de la tabla. Otras tablas también aparecen en Athena.

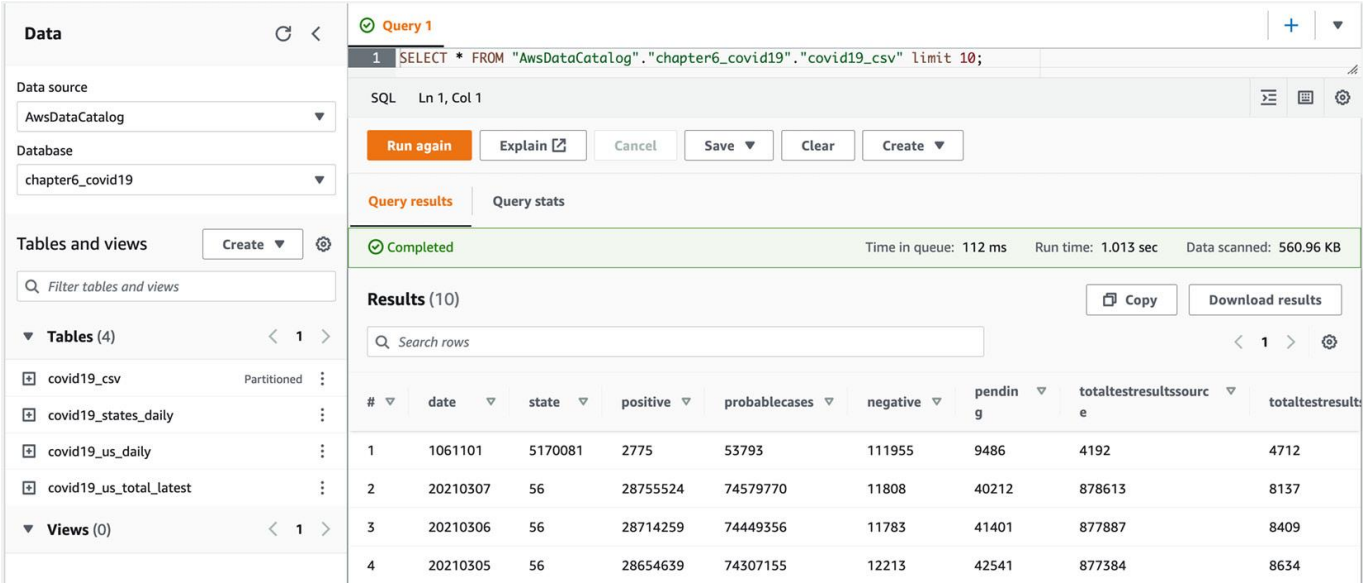


Figura 14 - Consola de consultas de Amazon Athena

En este ejemplo, hemos demostrado cómo utilizar AWS Glue Catalog para catalogar el conjunto de datos Covid-19 y ejecutar consultas SQL en él utilizando Amazon Athena. En la siguiente sección, nos centraremos en cómo realizar la ingesta de datos utilizando AWS Glue ETL.

NOTA

Previamente al uso de Athena hay que configurar un *bucket* de almacenamiento para la salida de las consultas:

Editor | Consultas recientes | Consultas guardadas | Configuración de consultas

Grupo de trabajo primary

Configuración de consultas

La configuración de consultas anula temporalmente los ajustes del grupo de trabajo. Sin embargo, para aplicar dichos ajustes, debe borrar toda la configuración.

Cifrado de resultados de consultas

Administrar

Ubicación del resultado de la consulta	Cifrado de los resultados de la consulta	Propietario previsto del bucket	Asigna al propietario del bucket el control total sobre los resultados de la consulta
s3://mpbprueba/	-	-	Desactivado