LECTURA SESIÓN MOVIMIENTOS DE DATOS EN LA NUBE MEDIANTE ETL

AWS Glue es un servicio de Extract, Transform, and Load (ETL) completamente administrado ofrecido por Amazon Web Services (AWS). Su principal objetivo es facilitar la preparación, transformación y movimiento de datos para su posterior análisis y procesamiento en entornos de big data o inteligencia empresarial.

R **por Kibernum Capacitación S.A.**

Introducción a AWS Glue

A diferencia de las soluciones tradicionales de ETL que requieren la provisión, configuración y mantenimiento de servidores o infraestructura, AWS Glue proporciona un entorno sin servidor (serverless), lo que permite a los usuarios centrarse exclusivamente en el flujo de datos y las transformaciones, sin preocuparse por la gestión subyacente de hardware o software.

AWS Glue automatiza tareas comunes asociadas a los flujos de datos, como el descubrimiento de fuentes de datos, el mapeo de esquemas, la generación de código ETL y la ejecución de cargas. Además, se integra de forma nativa con otros servicios del ecosistema AWS, lo que permite construir soluciones de datos escalables, ágiles y altamente disponibles.

En resumen, AWS Glue proporciona todas las herramientas necesarias para extraer datos de múltiples fuentes, transformarlos según necesidades específicas, y cargarlos en destinos de almacenamiento o análisis, todo ello de manera ágil, segura y escalable.

Características clave del servicio

Servicio sin servidor

AWS Glue opera en un modelo de computación serverless, lo que significa que los usuarios no necesitan provisionar, configurar ni administrar servidores físicos o virtuales. El servicio automáticamente escala los recursos de cómputo y memoria necesarios para ejecutar los trabajos ETL, optimizando costos y simplificando la operación.

Catálogo de datos centralizado

AWS Glue incluye un Data Catalog que actúa como una única fuente de referencia para los metadatos de la organización. Este catálogo almacena información estructurada sobre fuentes de datos, como tablas, bases de datos, ubicaciones de almacenamiento y esquemas de los datos. Así, facilita la búsqueda, el descubrimiento y la gestión de activos de datos de forma centralizada, promoviendo su reutilización en múltiples aplicaciones y flujos de trabajo.

Automatización del proceso ETL

AWS Glue automatiza tareas clave del ciclo de vida de los datos:

- Crawlers: Exploran fuentes de datos y generan automáticamente esquemas para el catálogo.
- Clasificadores: Detectan y clasifican el formato y estructura de los datos.
- Desencadenadores: Permiten la ejecución automática de trabajos ETL basados en horarios predefinidos o eventos (por ejemplo, la llegada de nuevos archivos a un bucket de S3).

Compatibilidad con múltiples formatos y fuentes de datos

Formatos soportados

- JSON
- Parquet
- Avro
- ORC
- CSV

Asimismo, puede integrarse con diversas fuentes y destinos de datos tales como Amazon S3, Amazon RDS, Amazon Redshift, Amazon DynamoDB, y también sistemas de bases de datos externos a AWS

Transformaciones potentes

El motor de procesamiento de AWS Glue está construido sobre Apache Spark, lo que permite a los desarrolladores aprovechar un entorno de procesamiento distribuido de alto rendimiento para ejecutar transformaciones de datos complejas. Además, AWS Glue ofrece una opción de programación visual mediante el Glue Studio, permitiendo diseñar flujos de trabajo ETL mediante una interfaz gráfica intuitiva.

Beneficios y casos de uso

Reducción del tiempo de desarrollo ETL

AWS Glue proporciona herramientas automáticas de generación de scripts, asistentes de configuración, y motores de inferencia de esquemas que aceleran significativamente la construcción de pipelines de datos. Esto permite que las organizaciones pasen más tiempo analizando datos en lugar de gestionando infraestructuras o desarrollando soluciones ETL desde cero.

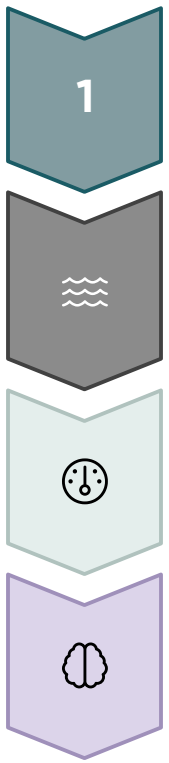
Integración nativa con el ecosistema AWS

Glue se integra de manera fluida con otros servicios esenciales de AWS, como:

- Amazon S3: Almacenamiento de datos de entrada y salida.
- Amazon Redshift: Carga de datos optimizada para análisis de grandes volúmenes.
- Amazon Athena: Consultas serverless directamente sobre datos almacenados en S3 usando SQL.
- AWS Lake Formation: Para la creación y gestión de data lakes seguros.

Esta integración profunda facilita la construcción de arquitecturas de datos modernas, permitiendo un acceso seguro, ágil y escalable a los datos.

Casos de uso más comunes



Migraciones de bases de datos

Ayuda en la extracción, transformación y carga de datos desde sistemas antiguos hacia soluciones modernas en la nube.

Creación de data lakes

Automatiza la ingesta y catalogación de grandes volúmenes de datos de diferentes fuentes para su centralización en un data lake sobre Amazon S3.

Cargas para análisis en tiempo real

Mueve y transforma datos de forma eficiente para habilitar dashboards en tiempo real y otras soluciones analíticas.

Preparación de datos para Machine Learning

Limpia, transforma y organiza datos para su uso en pipelines de Machine Learning o en servicios como Amazon SageMaker.

Elementos Fundamentales de AWS Glue

Catálogo de Metadatos

El AWS Glue Data Catalog es un componente esencial dentro de la arquitectura de AWS Glue. Funciona como un repositorio centralizado de metadatos, proporcionando una estructura organizada y accesible para describir, descubrir y gestionar las diferentes fuentes de datos disponibles dentro de la organización.

Cada vez que se crea una tabla o base de datos en el catálogo, se almacenan metadatos tales como:

- Nombre y ubicación de los conjuntos de datos (por ejemplo, rutas en Amazon S3).
- Formatos de archivo (CSV, JSON, Parquet, Avro, etc.).
- Esquema de los datos (nombres de columnas, tipos de datos, particiones).
- Propiedades adicionales como marcas de tiempo, etiquetas de clasificación o políticas de acceso.

El catálogo de metadatos permite a distintos servicios de AWS (como Athena, Redshift Spectrum y EMR) consultar los datos sin necesidad de conocer su ubicación física o su estructura interna, promoviendo así una federación y reutilización eficiente de los datos.

Además, el Data Catalog puede integrarse con AWS Lake Formation para proporcionar capacidades avanzadas de control de acceso basado en políticas, garantizando así la seguridad y gobernanza de los datos a escala.

Trabajos y Desencadenadores

Trabajos (Jobs)

Los trabajos en AWS Glue representan los procesos ETL que transforman los datos brutos en información utilizable para análisis, almacenamiento o procesamiento posterior.

Características principales:

- Lenguajes compatibles: Se pueden escribir en Python o Scala, utilizando el motor de procesamiento de datos basado en Apache Spark.
- Transformaciones: Pueden realizar operaciones como limpieza de datos, normalización, agregaciones, combinaciones (joins) entre conjuntos de datos, y más.
- Creación de scripts automática: Glue puede generar automáticamente scripts ETL a partir de plantillas, basándose en las fuentes y destinos de datos seleccionados.
- Modos de ejecución: Los trabajos pueden ejecutarse de forma manual, programada o en respuesta a eventos específicos.

Adicionalmente, AWS Glue ofrece el Glue Studio, una herramienta visual que permite diseñar, construir y monitorear trabajos ETL a través de una interfaz gráfica intuitiva, sin necesidad de codificar manualmente.

Desencadenadores (Triggers)

Los desencadenadores son mecanismos que inician la ejecución de trabajos ETL en AWS Glue basándose en:

- Eventos programados: Definición de horarios específicos mediante reglas de tipo cron o programación simple.
- Eventos de cambios: Como la llegada de nuevos archivos a un bucket de Amazon S3 o la actualización de un registro en una base de datos.
- Dependencias de otros trabajos: Permiten diseñar flujos de trabajo complejos donde la ejecución de un trabajo depende de la finalización exitosa de otros.

Gracias a los desencadenadores, AWS Glue facilita la automatización de flujos de datos, asegurando la orquestación adecuada de múltiples tareas ETL sin intervención manual.

Crawlers y Clasificadores

Crawlers

Los crawlers son procesos automatizados que se encargan de explorar diversas fuentes de datos para inferir esquemas y actualizar el catálogo de metadatos.

Funciones principales:

- Analizan los datos almacenados en S3, bases de datos relacionales, NoSQL, entre otras fuentes.
- Detectan automáticamente el tipo de archivos, la estructura de las tablas y los tipos de columnas.
- Pueden particionar datos basados en carpetas o patrones de nombres, optimizando consultas posteriores.
- Actualizan de forma periódica el catálogo de datos, manteniendo siempre la información de metadatos sincronizada con las fuentes reales.

La programación de crawlers permite que el catálogo de datos se mantenga actualizado sin intervención manual, lo cual es crítico en entornos donde los datos cambian o crecen dinámicamente.

Clasificadores

Los clasificadores son componentes que AWS Glue utiliza dentro de los crawlers para interpretar el formato y el esquema de los archivos analizados.

Tipos de clasificadores:

- Clasificadores de archivos estructurados: Para identificar formatos como JSON, CSV, XML, Parquet, Avro, ORC, entre otros.
- Clasificadores personalizados: Permiten definir expresiones regulares o patrones específicos para analizar fuentes de datos no tradicionales o altamente personalizadas.

Durante el proceso de crawling, Glue utiliza estos clasificadores para determinar:

- El tipo de datos.
- Cómo debe representarse su esquema en el catálogo.
- Cómo deben procesarse posteriormente en trabajos ETL.

Esto garantiza una detección precisa y una automatización eficiente en ambientes de datos heterogéneos.

Configuración y Uso de AWS Glue

AWS Glue es un servicio totalmente administrado de ETL (Extract, Transform, Load) diseñado para facilitar el descubrimiento, la preparación y la integración de datos. Para aprovechar su potencial, es necesario realizar una configuración adecuada y entender cómo usarlo de manera efectiva. A continuación, se detallan los pasos clave para la configuración, el uso y el monitoreo de trabajos en AWS Glue.

Crear un catálogo de datos

El Catálogo de Datos de AWS Glue es un repositorio centralizado que almacena metadatos sobre los datos que se encuentran en diversas fuentes, como bases de datos, archivos en S3, etc. Para empezar, se debe crear un catálogo de datos que contenga la información relevante sobre los datos que se van a procesar.

Un catálogo de datos define las tablas, las columnas y los esquemas de las fuentes de datos, lo que facilita la búsqueda y el procesamiento posterior.

Configurar un crawler para explorar la fuente de datos

Un crawler (rastreador) es una herramienta que se utiliza para explorar automáticamente las fuentes de datos y definir las tablas en el catálogo de datos. AWS Glue proporciona un servicio de crawler que puede conectarse a diferentes fuentes de datos, como bases de datos relacionales, archivos en S3, y otros almacenes de datos.

Configurar un crawler implica definir las fuentes de datos, las credenciales de acceso y los esquemas. El crawler escaneará los datos, inferirá su estructura y actualizará el catálogo de datos con las tablas necesarias.

Crear un trabajo ETL

Un trabajo ETL es el proceso que extrae datos de una fuente, los transforma según las necesidades del negocio, y los carga en el destino adecuado. AWS Glue permite crear trabajos ETL de dos formas:

- Consola Visual: AWS Glue proporciona una interfaz gráfica en la consola que permite diseñar y arrastrar transformaciones, simplificando el proceso de creación de trabajos ETL sin necesidad de escribir código.
- Escribir el script: Si se requiere mayor flexibilidad, se puede escribir un script ETL utilizando el lenguaje de programación Python o Scala, aprovechando el entorno de Apache Spark para procesar los datos de manera distribuida.

Definir desencadenadores si es necesario

Los desencadenadores permiten automatizar la ejecución de trabajos ETL. Pueden ser configurados para ejecutar un trabajo de manera regular (por ejemplo, diariamente a una hora específica) o en respuesta a ciertos eventos (por ejemplo, cuando se agregan nuevos archivos a un bucket de S3). Los desencadenadores pueden ser definidos desde la consola de AWS Glue, y pueden ser programados o activados por eventos.

Ejecutar el trabajo y almacenar los resultados

Una vez configurado el trabajo ETL, se puede ejecutar. AWS Glue proporciona herramientas para iniciar manualmente la ejecución de trabajos o hacerlo de manera automatizada a través de desencadenadores.

Los resultados del procesamiento pueden ser almacenados en diferentes destinos, como bases de datos Amazon Redshift, Amazon S3, o incluso otros sistemas de almacenamiento que AWS Glue soporta.

Calendarización y Ejecución de Trabajos

Ejecuciones bajo demanda

Los trabajos ETL en AWS Glue pueden ejecutarse bajo demanda, lo que significa que el usuario puede iniciar el trabajo en cualquier momento según sea necesario. Esto es útil para situaciones en las que no se requiere una ejecución regular, sino que se necesita procesar datos en momentos específicos, como una actualización de datos urgente.

Ejecuciones programadas

Los trabajos también pueden ser ejecutados de manera programada, lo que permite automatizar la ejecución en intervalos regulares. AWS Glue permite programar trabajos de acuerdo a horarios específicos, como ejecutar un trabajo todos los días a las 3 a.m. o en cualquier otra frecuencia personalizada. Esto es ideal para procesos ETL que deben ejecutarse periódicamente, como la actualización de informes o la consolidación de datos.

Ejecuciones basadas en eventos

Además de las ejecuciones programadas, AWS Glue permite la ejecución de trabajos en función de eventos. Por ejemplo, se puede configurar un trabajo para que se ejecute automáticamente cuando nuevos archivos se carguen en un bucket de Amazon S3 o cuando se produzcan ciertos cambios en una base de datos. Esta capacidad es útil para escenarios en los que los datos cambian frecuentemente y requieren procesamiento en tiempo real o casi en tiempo real.

Monitoreo y Diagnóstico

WS Glue proporciona herramientas de monitoreo y diagnóstico para asegurar que los trabajos ETL se ejecuten sin problemas y para ayudar a detectar problemas rápidamente en caso de que surjan. AWS Glue se integra con Amazon CloudWatch, un servicio de monitoreo de AWS que permite visualizar métricas y logs de los trabajos en tiempo real.

Integración con Amazon CloudWatch

AWS Glue se integra con Amazon CloudWatch para ofrecer métricas detalladas sobre los trabajos ETL, como el tiempo de ejecución, el uso de recursos (CPU, memoria), el número de registros procesados y otros indicadores clave de rendimiento. Las métricas pueden ser visualizadas en el panel de CloudWatch para monitorear la salud de los trabajos y asegurarse de que los datos se procesen de acuerdo con lo esperado.

Logs de ejecución

AWS Glue también genera logs detallados de cada ejecución de trabajo, los cuales pueden ser accesados desde la consola de CloudWatch Logs. Estos logs incluyen información sobre los pasos del trabajo, errores, advertencias y cualquier otro evento relevante. Esto facilita la depuración de problemas y permite identificar rápidamente las causas raíz de cualquier fallo en el proceso ETL.

Métricas de rendimiento, errores y advertencias

AWS Glue ofrece métricas sobre el rendimiento de los trabajos, como la velocidad de procesamiento, la cantidad de datos procesados y el uso de recursos. Además, se pueden configurar alertas basadas en métricas específicas, como cuando un trabajo falla o si hay errores en la ejecución. Las advertencias permiten a los administradores tomar medidas preventivas antes de que ocurran fallos graves, lo que mejora la fiabilidad de los trabajos ETL.

Con estas herramientas de monitoreo y diagnóstico, los usuarios pueden gestionar de manera efectiva los trabajos ETL, asegurando que los datos se procesen correctamente y se tomen acciones rápidamente en caso de errores.

Optimización del Rendimiento en AWS Glue

AWS Glue es una herramienta poderosa para procesar y transformar grandes volúmenes de datos. Sin embargo, para obtener el máximo rendimiento y reducir los costos, es fundamental optimizar tanto la configuración de los trabajos como el uso de recursos. A continuación, se detallan las mejores prácticas y enfoques para optimizar el rendimiento en AWS Glue, tanto en términos de velocidad de procesamiento como de costos operativos.

Mejores Prácticas para Optimización

El rendimiento en AWS Glue depende de cómo se gestionan los datos, las configuraciones de los trabajos ETL, y los recursos de cómputo disponibles. Algunas de las mejores prácticas incluyen:

Optimizar particiones de datos para mejorar la eficiencia

Una de las formas más efectivas de mejorar el rendimiento es particionar los datos adecuadamente. La partición de datos divide grandes conjuntos de datos en fragmentos más pequeños y manejables, lo que permite a AWS Glue leer solo las particiones necesarias en lugar de procesar todos los datos.

Al particionar los datos por columnas como fechas o regiones, puedes asegurarte de que solo se acceda a los datos relevantes durante el procesamiento. Esto reduce el volumen de datos leídos y, en consecuencia, mejora la velocidad de las transformaciones.

La partición eficiente también mejora el rendimiento de las consultas, ya que las particiones permiten la lectura de un subconjunto de los datos, optimizando el tiempo de acceso a las filas necesarias.

Usar formatos columnar (Parquet, ORC) que ofrecen mejor compresión y velocidad

Los formatos de almacenamiento columnar, como Parquet y ORC, son altamente eficientes para el procesamiento de grandes volúmenes de datos. Estos formatos permiten una mayor compresión, lo que reduce el almacenamiento necesario, y mejoran el rendimiento al leer solo las columnas que se necesitan.

A diferencia de los formatos tradicionales de filas como CSV o JSON, los formatos columnar permiten una lectura más eficiente y rápida de los datos, lo que se traduce en una reducción significativa en los tiempos de procesamiento de los trabajos ETL.

Además, estos formatos permiten que las operaciones de lectura y escritura sean más eficientes, lo que también reduce el costo de las operaciones de E/S (entrada/salida) en los trabajos ETL.

Además, estos formatos permiten que las operaciones de lectura y escritura sean más eficientes, lo que también reduce el costo de las operaciones de E/S (entrada/salida) en los trabajos ETL.

Más prácticas de optimización

Configurar correctamente las unidades de cómputo (DPU) según el volumen de datos

AWS Glue utiliza unidades de procesamiento (DPU) para ejecutar trabajos ETL. Es fundamental configurar correctamente el número de DPUs necesarias para el trabajo en función del volumen de datos que se va a procesar.

Si el trabajo ETL tiene una gran cantidad de datos, puedes aumentar el número de DPUs para obtener más capacidad de procesamiento y reducir el tiempo de ejecución. Sin embargo, si el volumen de datos es bajo, utilizar más DPUs de las necesarias sólo aumentará el costo innecesariamente.

AWS Glue también permite configurar los trabajos para que utilicen el número adecuado de DPUs dinámicamente, lo que optimiza los costos y mejora la eficiencia.

Aplicar pushdown predicates para reducir la cantidad de datos leídos

Los pushdown predicates son filtros que se aplican directamente en la base de datos o en la fuente de datos antes de que los datos se transfieran a AWS Glue para su procesamiento. En lugar de leer todos los datos y luego filtrar los que no son necesarios, el predicado se empuja a la fuente de datos para que solo se lean los datos que cumplen con los criterios especificados.

Esta técnica reduce significativamente el volumen de datos que se transfieren y procesan, lo que mejora el rendimiento y reduce tanto el tiempo de ejecución como el costo asociado al procesamiento de datos.

Reducción de Costos y Mejora de Eficiencia

Además de optimizar el rendimiento de los trabajos, AWS Glue ofrece diversas técnicas para reducir los costos operativos y mejorar la eficiencia de los procesos ETL. Algunas de las estrategias clave incluyen:

Detener crawlers innecesarios después de la exploración inicial

Los crawlers son procesos que escanean las fuentes de datos para extraer metadatos y actualizar el catálogo de datos de AWS Glue. Sin embargo, una vez que un crawler ha realizado su trabajo inicial y ha registrado toda la información necesaria, dejarlo en funcionamiento continuo puede ser innecesario y costoso.

Es recomendable detener los crawlers después de la exploración inicial o cuando ya no sean necesarios. Esto ayuda a evitar costos adicionales por la ejecución innecesaria de crawlers y permite que los recursos se utilicen de manera más eficiente.

Optimizar el tamaño de las cargas para aprovechar mejor las DPU

El tamaño de las cargas de trabajo puede afectar la eficiencia de las DPUs. Si el tamaño de los datos es demasiado pequeño, las DPUs podrían no aprovecharse al máximo, lo que resultaría en un uso ineficiente de los recursos y, en consecuencia, mayores costos.

Para optimizar el rendimiento y los costos, se debe ajustar el tamaño de las cargas de trabajo de manera que las DPUs puedan manejar de manera eficiente los datos. Esto puede implicar la combinación de múltiples archivos pequeños en archivos más grandes antes de procesarlos, lo que reduce el número de tareas de procesamiento y mejora la eficiencia.

Reutilizar catálogos existentes para minimizar duplicación de esfuerzos

Al utilizar un catálogo de datos de AWS Glue, se pueden almacenar metadatos de múltiples fuentes de datos. En lugar de crear nuevos catálogos para cada trabajo ETL, es posible reutilizar catálogos existentes, lo que ahorra tiempo y esfuerzo en la configuración de los trabajos.

Además, reutilizar catálogos ayuda a evitar la duplicación de los esfuerzos de exploración de datos, lo que se traduce en menos costos por ejecución de crawlers y mayor eficiencia en el uso de recursos.

Utilizar particiones de datos para reducir el procesamiento innecesario

Al almacenar los datos de manera particionada y optimizada (por ejemplo, particionándolos por fecha), se puede reducir el volumen de datos que se procesan en cada ejecución de trabajo ETL. Esto mejora tanto la eficiencia como el costo, ya que solo se procesan las particiones relevantes en lugar de todo el conjunto de datos.

Aprovechar el almacenamiento en Amazon S3 con optimización de costos

AWS Glue se integra perfectamente con Amazon S3, lo que permite aprovechar las políticas de almacenamiento en capas de S3 para reducir los costos de almacenamiento a largo plazo. Por ejemplo, se pueden mover los datos procesados a almacenamiento de baja frecuencia de acceso (S3 Glacier o S3 Intelligent-Tiering) después de que hayan sido procesados y almacenados en S3. Esto no solo optimiza los costos de almacenamiento, sino que también reduce la latencia de los trabajos futuros.

Mediante la implementación de estas mejores prácticas de optimización en AWS Glue, las organizaciones pueden mejorar el rendimiento de sus trabajos ETL, reducir costos operativos y asegurar que los recursos sean utilizados de manera eficiente. Esto se traduce en una mayor agilidad en el procesamiento de datos y en un uso más rentable de los servicios de AWS Glue.

Casos de Uso y Ejemplos Prácticos

Aplicaciones reales de AWS Glue en la industria

Sector financiero: Integración de datos para cumplimiento normativo

En la industria financiera, AWS Glue se utiliza para facilitar la integración de datos de diversas fuentes y asegurar que las instituciones cumplan con las normativas de reguladores financieros. Por ejemplo, muchas entidades financieras operan en un entorno con estrictas regulaciones de privacidad y protección de datos, como la Ley de Protección de Datos Personales o las normativas relacionadas con el cumplimiento de la normativa financiera internacional, como la Ley de Secreto Bancario. AWS Glue puede automatizar la integración de datos de múltiples sistemas, realizar transformaciones necesarias para asegurar la calidad de los datos y permitir su almacenamiento en una infraestructura centralizada.

E-commerce: Unificación de datos de clientes para análisis de comportamiento

En el ámbito del comercio electrónico (e-commerce), AWS Glue se utiliza para unificar grandes volúmenes de datos generados por los clientes durante su interacción con los sistemas. Las empresas de e-commerce suelen obtener datos de diversas fuentes, como registros de transacciones, interacciones en sitios web, interacciones a través de aplicaciones móviles, redes sociales y más. Estos datos a menudo están dispersos en diferentes sistemas, formatos y bases de datos. AWS Glue ayuda a consolidar toda esta información en un solo lugar y permite transformarla para facilitar un análisis más profundo del comportamiento del cliente.

Salud: Procesamiento de registros médicos electrónicos para analítica avanzada

En el sector salud, el uso de AWS Glue se ha expandido para la integración y procesamiento de grandes volúmenes de datos provenientes de registros médicos electrónicos (EHR, por sus siglas en inglés), datos de investigación clínica, resultados de pruebas y otros sistemas de salud. Los hospitales y las organizaciones de salud suelen manejar información de pacientes que proviene de distintas fuentes, sistemas y formatos, lo que hace que la integración y el análisis de estos datos sea una tarea compleja.

Comparación con otras herramientas ETL en la nube

Característica	AWS Glue	Azure Data Factory	Google Cloud Dataflow
Administración de infraestructura	Totalmente administrado	Parcialmente administrado	Totalmente administrado
Motor de procesamiento	Apache Spark	Data Flows (Spark)	Apache Beam
Soporte de catalogación	Integrado	Limitado	No nativo
Modelo de precios	Por DPU por hora	Por actividades	Por procesamiento