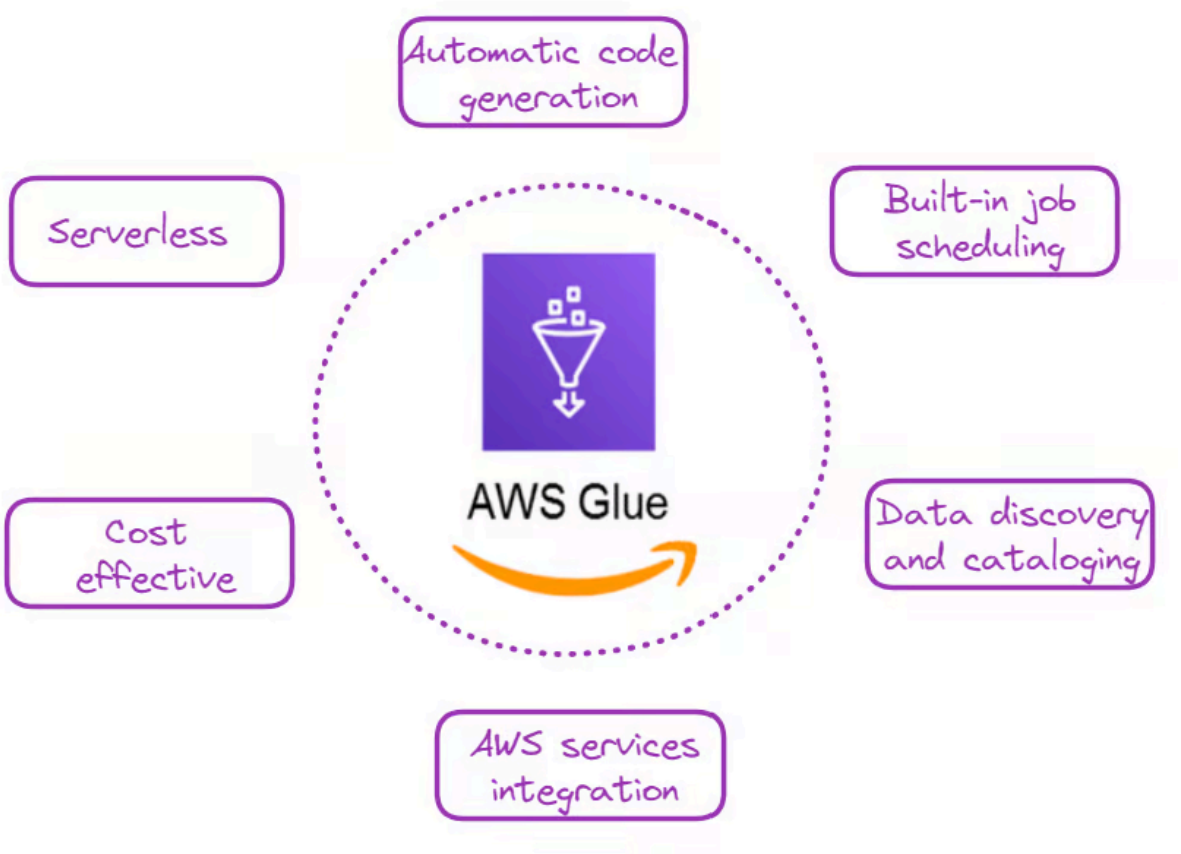
Introducción a AWS Glue

Created by	Juan Alejandro Carrillo Jaimes
Created time	@March 13, 2024 1:13 PM
Tags	AWS, AWS Glue, Cloud, ETL, Fundamentals

¿Qué es AWS Glue?



AWS Glue es un servicio de integración de datos sin servidor que facilita a los usuarios de análisis descubrir, preparar, migrar e integrar datos de varios orígenes. Puede utilizarlo para análisis, machine learning y desarrollo de aplicaciones. También incluye herramientas adicionales de productividad y operaciones de datos para la creación, la ejecución de trabajos y la implementación de flujos de trabajo empresariales.

Con AWS Glue, puede descubrir y conectarse a más de 70 orígenes de datos diversos y administrar sus datos en un catálogo de datos centralizado. Puede crear, ejecutar y supervisar visualmente canalizaciones de extracción, transformación y carga (ETL) para cargar datos en los lagos de datos. Además,

puede buscar y consultar datos catalogados de forma inmediata mediante Amazon Athena, Amazon EMR y Amazon Redshift Spectrum.

Casos de Uso 📕

Simplifique la administración de canalizaciones ETL

Elimine la administración de infraestructura con aprovisionamiento automático y administración de trabajadores, y consolide todas sus necesidades de integración de datos en un único servicio.

Explore, experimente y procese datos de forma interactiva

Mediante las sesiones interactivas de AWS Glue, los ingenieros de datos pueden explorar y preparar datos de forma interactiva utilizando el entorno de desarrollo integrado (IDE) o el bloc de notas de su elección.

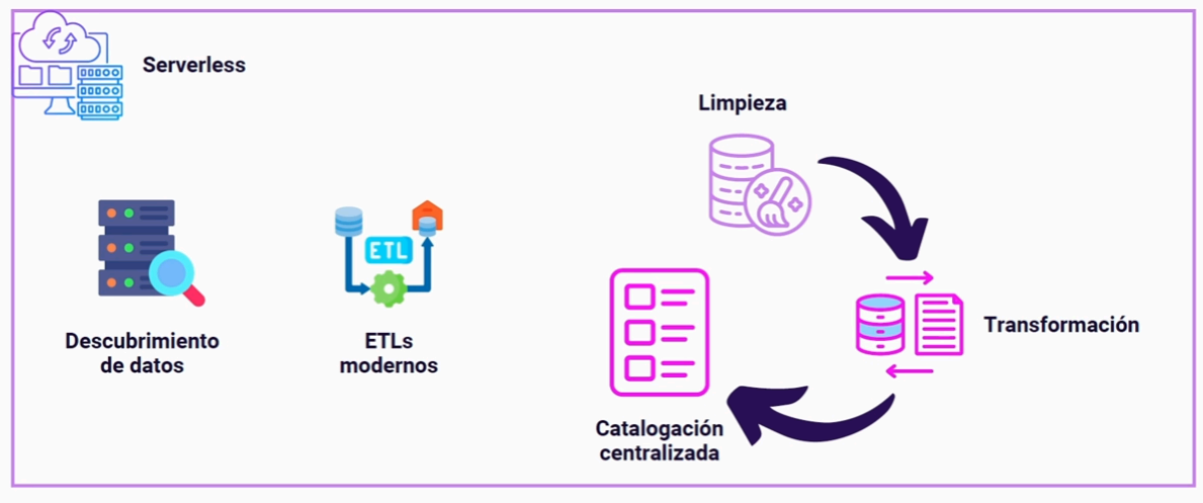
Descubra datos de manera eficiente

Identifique datos rápidamente en AWS, en las instalaciones y en otras nubes y, a continuación, póngalos a disposición al instante para realizar consultas y transformaciones.

Soporta varios marcos de procesamiento y cargas de trabajo

Soporta más fácilmente varios marcos de procesamiento de datos, como ETL y ELT, y varias cargas de trabajo, incluidos lotes, microlotes y streaming.

Beneficios de AWS Glue



1. **Compatible con todas las cargas de trabajo**: Soporte flexible para ETL, ELT, batch, streaming y más, sin lock-in.

2. **Escala bajo demanda**: Escala de petabytes, facturación por uso y tamaño de datos.

3. **Herramientas a medida**: Soporte a todos los usuarios de datos, desde desarrolladores a usuarios de negocio.

4. **Asistente de IA generativa**: Obtenga ayuda impulsada por IA a lo largo de su viaje de integración de datos, desde la generación automática de código ETL hasta la modernización de sus trabajos de Spark. AWS Glue proporciona generación de código inteligente, actualizaciones de Spark asistidas por IA (vista previa) y solución de problemas de Spark integrada.

5. **Todo en uno**: Capacidades completas de integración de datos en un servicio sin servidor.

Serverless: No hay infraestructura para administrar

Data Integration Options

🔧 Opciones de Integración de Datos en AWS Glue

1. **AWS Glue for Apache Spark**
   - Ideal para cargas de trabajo grandes y transformaciones complejas.
   - Procesamiento distribuido y escalable.
   - Soporta `DynamicFrame`, PySpark y particionamiento automático.
2. **AWS Glue for Python Shell**
   - Usado para tareas ligeras o scripts rápidos.
   - Ideal para validación de archivos, llamadas a APIs, procesamiento de metadatos.
   - Usa solo Python estándar (no Spark).
3. **AWS Glue for Ray**
   - Motor más nuevo orientado a procesamiento paralelo con Python puro.
   - Bueno para tareas que no requieren Spark pero sí paralelismo.
   - Compatible con librerías como Pandas, NumPy y Dask.

🧠 Interfaces para crear y ejecutar estos jobs:

- **AWS Glue Studio**: interfaz gráfica para diseñar y correr ETLs.
- **Amazon SageMaker Notebooks**: desarrollo en notebooks Jupyter.
- **Notebooks e IDEs**: integración con VS Code, Jupyter local, etc.

📥 Orígenes de datos:

- Amazon S3, DynamoDB, RDS, EC2, SaaS, etc.

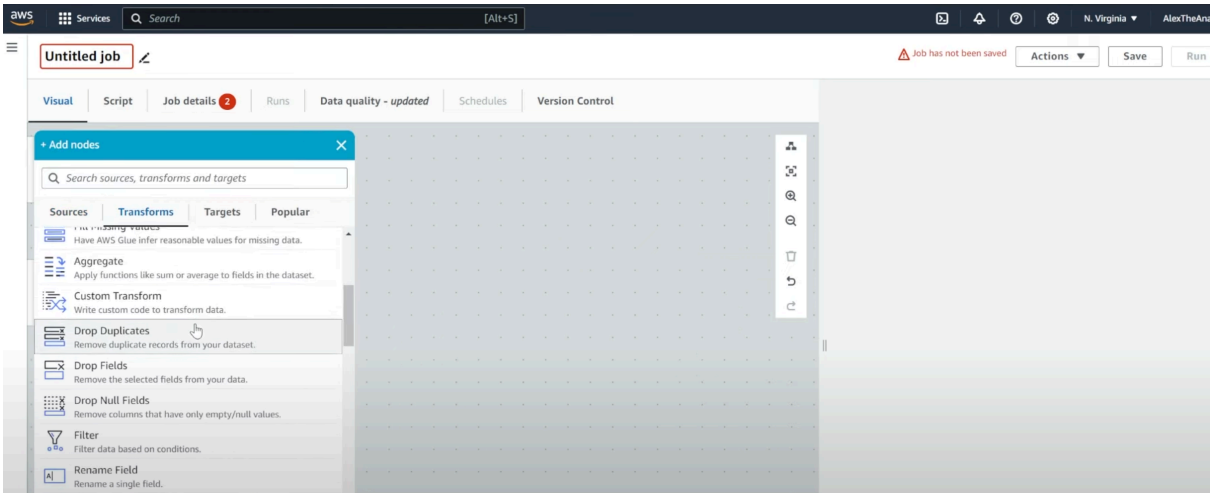
📤 Destinos típicos:

- Amazon Redshift (data warehouse), Data Lakes.

Data Preparation

Con AWS Glue DataBrew, puede explorar y experimentar con datos directamente desde su lago de datos, almacenes de datos y bases de datos, incluidos Amazon S3, Amazon Redshift, AWS Lake Formation, Amazon Aurora y Amazon Relational Database Service (RDS). Puede elegir entre más de 250 transformaciones predefinidas en DataBrew para automatizar tareas de preparación de datos como filtrar anomalías, estandarizar formatos y corregir valores no válidos.

¿Qué es AWS Glue DataBrew?



AWS Glue DataBrew es una herramienta visual de preparación de datos que permite a analistas y científicos de datos limpiar y transformar datos **sin escribir código**. Está diseñada para trabajar directamente con fuentes de datos como Amazon S3, Redshift, RDS, Aurora y más.

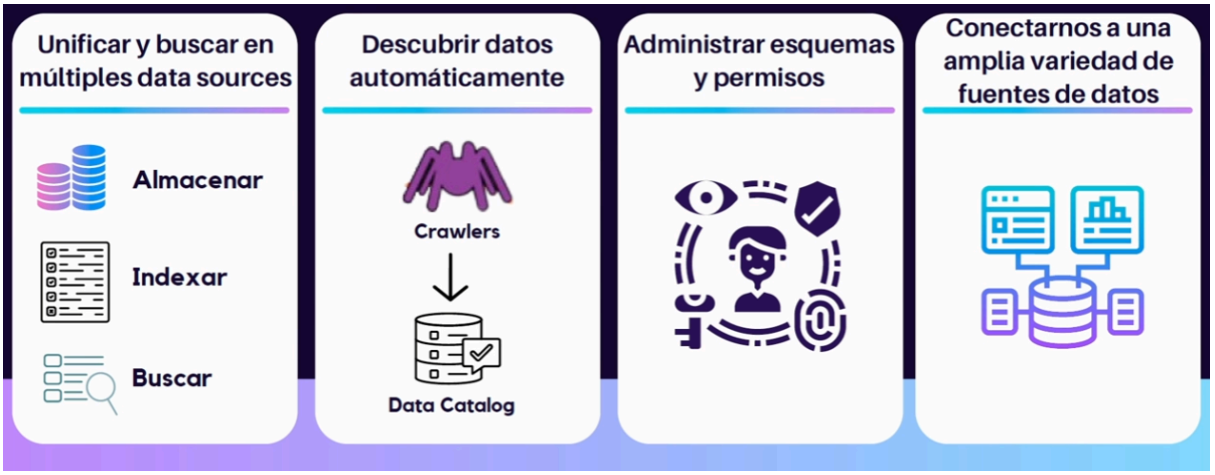
Funcionalidades Clave:

1. 🔌 **Conectividad con fuentes de datos**
   - Se conecta directamente a data lakes, data warehouses y bases de datos.
2. ⚙️ **+250 transformaciones predefinidas**
   - Puedes limpiar, normalizar, transformar y enriquecer datos sin necesidad de programar.
3. 📊 **Evaluación de calidad de datos**
   - Perfilado de datos para detectar patrones, valores nulos, outliers y anomalías.
4. 🔁 **Automatización escalable**
   - Reutiliza pasos de transformación automáticamente cuando llegan nuevos datos.
5. 📤 **Publicación en Amazon S3**
   - Guarda los datos transformados directamente en S3, listos para usarse en análisis o ML.
6. ⚡ **Obtención rápida de insights**

- Datos listos para alimentar dashboards, sistemas de BI o modelos de Machine Learning.

Características Principales de AWS Glue

Descubrir y Organizar datos



AWS Glue facilita la **unificación y organización de datos dispersos** provenientes de diversas fuentes tanto en la nube como on-premise. Este proceso se apoya en varias capacidades clave:

- 🔍 **Descubrimiento Automático con Crawlers**: Los *crawlers* examinan automáticamente tus datos, infieren los esquemas y tipos de datos, y los registran en el **AWS Glue Data Catalog**. Esto elimina la necesidad de definir manualmente estructuras complejas.
- 🗃️ **Catálogo Centralizado de Datos (Glue Data Catalog)**: Funciona como un metastore que organiza tus datasets en bases de datos y tablas accesibles para herramientas como Amazon Athena, Redshift Spectrum y EMR.
- 🔐 **Gestión de Accesos y Permisos**: Se integra con **AWS Lake Formation** y **IAM** para establecer permisos granulares a nivel de base de datos, tabla, columna o fila.
- 🔗 **Conectividad Multifuente**: Glue puede conectarse a datos almacenados en Amazon S3, Amazon RDS, Redshift, DynamoDB, y también a fuentes

externas mediante conectores JDBC y conectores personalizados para SaaS o entornos on-premise.

- 🔄 **Soporte para Esquemas Evolutivos**: Permite gestionar cambios en los esquemas a lo largo del tiempo, versión tras versión, sin romper los pipelines existentes.

Transformar, preparar y limpiar los dato para su análisis

AWS Glue permite crear flujos de transformación y pipelines ETL de manera intuitiva y potente, con las siguientes capacidades:

- 🎛️ **Interfaz Visual Drag-and-Drop (AWS Glue Studio):**

  Diseña flujos ETL complejos sin necesidad de escribir código. Ideal para usuarios técnicos y no técnicos que desean transformar datos desde diversas fuentes como Amazon S3, RDS, Redshift, DynamoDB y más.

- 📅 **Orquestación Flexible de Pipelines de Datos:**

  Crea pipelines escalables con ejecuciones programadas (basadas en calendario), bajo demanda o disparadas por eventos (por ejemplo, carga de archivos en S3).

- ⚡ **Procesamiento en Streaming:**

Transforma datos en tiempo real a medida que llegan, lo que permite casos de uso como monitoreo de eventos, análisis en vivo y limpieza sobre datos en tránsito.

- 🧠 **Limpieza de Datos con IA y Machine Learning (ML):**

  Utiliza funcionalidades avanzadas como **FindMatches ML Transform** para realizar deduplicación inteligente y limpieza automática con algoritmos de aprendizaje automático.

- 📓 **Notebooks Serverless Integrados:**

  Experimenta, depura y ajusta tus scripts ETL en entornos interactivos como Jupyter Notebooks sin tener que aprovisionar servidores, gracias a la integración con Amazon SageMaker y Glue Notebooks.

- 🧪 **Edición y Depuración en IDEs o Notebooks:**

  Desarrolla tus ETLs directamente en entornos conocidos como Visual Studio Code, notebooks o Glue Studio, con ejecución de secciones interactivas que facilitan la depuración paso a paso.

- 🔐 **Detección y Protección de Datos Sensibles:**

  Identifica y remedia automáticamente datos confidenciales (como PII) en tus pipelines y lagos de datos, ayudando a cumplir normativas como GDPR, HIPAA o CCPA.

Crear y supervisar pipelines de datos

AWS Glue permite **construir y orquestar pipelines de datos escalables y totalmente administrados**, diseñados para manejar desde simples transformaciones hasta flujos de datos complejos y dependientes. A continuación se detallan las capacidades clave:

🧱 Diseño de Pipelines Complejos

- Puedes **encadenar múltiples jobs, crawlers y triggers**, creando flujos de trabajo ETL bien definidos y modulares.
- Los *triggers* permiten iniciar procesos basados en eventos específicos (por ejemplo, carga de archivos en S3) o en horarios programados (*cron*).
- Los flujos de trabajo (*Workflows*) permiten visualizar la ejecución completa del pipeline en forma de grafo.

⚙️ Ejecución y Automatización

- AWS Glue **escala automáticamente los recursos subyacentes** en función del tamaño del dataset y complejidad del job, eliminando la necesidad de configurar clústeres manualmente.
- Puedes automatizar la ejecución de jobs mediante **triggers condicionales**, activando procesos de manera inteligente según el éxito, fallo o finalización de tareas previas.

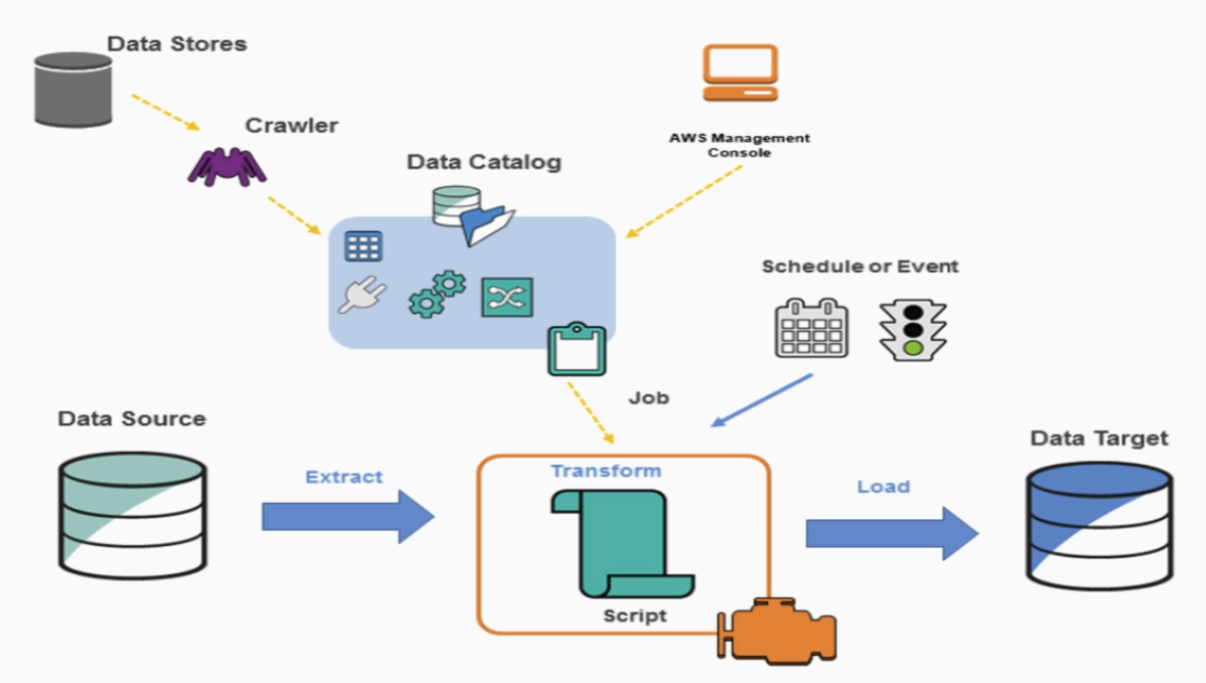
🔍 Monitoreo y Auditoría

- La ejecución de jobs puede **monitorearse en tiempo real** desde AWS Glue Studio o mediante APIs.
- Integración con **AWS CloudTrail** para auditar todas las acciones realizadas en Glue: ejecución de jobs, activación de triggers, uso de crawlers, etc.
- Glue registra métricas clave en **Amazon CloudWatch**, permitiendo configurar alertas o dashboards para detectar cuellos de botella, errores o tareas fallidas.

🔄 Trazabilidad y Mantenimiento

- Cada componente del pipeline (jobs, crawlers, triggers) está versionado y puede gestionarse de forma independiente.
- Las dependencias y relaciones entre tareas se gestionan desde **Glue Workflows**, lo que facilita el mantenimiento y la evolución del sistema ETL.

Conceptos de AWS Glue



🔁 Flujo de AWS Glue explicado paso a paso

1. 🗃️ Data Stores

- Representan las **fuentes de datos** y **destinos** como buckets S3, bases de datos relacionales (RDS, Aurora, etc.), NoSQL, etc.
- Estos datos pueden estar **estructurados, semi-estructurados o no estructurados**.
- Se conectan mediante **AWS Glue Connections**, que son objetos en el **Glue Data Catalog** usados para acceder de forma segura a estas fuentes o destinos.

---

2. 🕷️ Crawler

- El **Crawler** escanea un Data Store (por ejemplo, un bucket S3 con archivos CSV).
- Usa **clasificadores** para identificar el formato (CSV, JSON, Parquet, etc.).
- Automáticamente infiere el esquema (nombres de columnas, tipos de datos) y lo **registra como tablas** en el **Glue Data Catalog**.
- Se puede ejecutar de forma manual o automática (por eventos o triggers).

---

3. 📒 AWS Glue Data Catalog

- Es un almacén o **repositorio centralizado de metadatos**.
- Contiene:
  - **Bases de datos lógicas** (agrupan tablas).
  - **Tablas** (esquemas de archivos o bases de datos reales).
  - **Conexiones** (credenciales + ubicación).
- Cada vez que un crawler corre, actualiza el catálogo.

---

4. 💻 AWS Management Console

- Es la **interfaz para configurar** crawlers, jobs, triggers, conexiones, etc.
- También puedes lanzar trabajos manualmente desde aquí.
- Alternativamente, puedes usar **AWS CLI** o **SDKs**.

---

5. 🧠 Job

- Un **Job de Glue** es una unidad de procesamiento que ejecuta un **script de transformación** de datos.
- Puede ser de tipo:
  - **Spark** (PySpark o Scala): procesamiento distribuido.
  - **Python Shell**: scripts ligeros como validación, pequeñas tareas.
- Puede usar:
  - **DynamicFrame**: estructura optimizada de Glue (más flexible que DataFrame).
  - **DataFrame** de Spark para transformaciones más avanzadas.

---

6. 🧾 Script (Transformación)

- Aquí defines tu lógica ETL:
  - Limpieza, filtrado, mapeo, enriquecimiento.
  - En tu caso: validaciones, flags por FICO, ratios, etc.
- Se puede desarrollar usando:
  - **Glue Studio (visual)**
  - **Development Endpoints + Jupyter Notebook Server** (interactivo)
  - **VS Code + CLI Toolkit** (como lo haces tú)
- Se ejecuta en una infraestructura administrada por AWS usando **Workers (DPU)**.

---

7. ⏰ Trigger / Schedule or Event

- Un Job puede ejecutarse:
  - Por programación (**Schedule**).
  - Por evento (**por ejemplo, cuando se sube un archivo a S3**).
  - En cadena (cuando termina otro job, **Trigger**).
- Los **Triggers** permiten definir workflows complejos.

---

8. 🏁 Transform → Load

- El script toma datos del Data Source, aplica transformaciones y los carga en el **Data Target**.
- El destino puede ser:
  - S3 (en formatos optimizados como Parquet).
  - Amazon Redshift (usando JDBC Sink).
  - Bases de datos relacionales.
- Este paso es típicamente el resultado final de un Job exitoso.

🧠 Términos clave conectados al flujo:

Término	Descripción	Rol en el flujo
Glue Data Catalog	Catálogo de metadatos	Centraliza info de fuentes/destinos
Clasificador	Identifica formato de archivo	Usado por el Crawler
Conexión	Objeto del Catálogo con credenciales	Para acceder a RDS, Redshift, etc.
Crawler	Escanea fuentes y crea tablas	Detecta y registra datos automáticamente
Base de Datos (Glue)	Agrupador lógico de tablas	Organiza tus datasets
Development Endpoint	Entorno para desarrollo de scripts	Útil para Jupyter Notebooks
Dynamic Frame	Estructura de Glue optimizada	Más robusta que Spark DF
Job	Proceso ETL que ejecuta scripts	Core de la lógica empresarial
Notebook Server	Entorno interactivo	Desarrollo exploratorio
Script	Código PySpark/Scala/Python	Contiene la lógica de transformación
Tabla	Representación del esquema	Se almacena en el Data Catalog
Transformación	Proceso de cambio en los datos	Parte del script de ETL
Trigger	Desencadenador de Jobs	Automatización
Worker (DPU)	Unidad de procesamiento	Se configura por job

Sources 🔗

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

https://docs.aws.amazon.com/glue/latest/dg/machine-learning-teaching.html