CESTE
Escuela Internacional de Negocios

AWS SDK Data Wrangler

Sesión 3

AWS Lambda

Es un servicio de computación sin servidor ofrecido por AWS.

¿Qué significa esto?

Que te permite ejecutar código sin tener que preocuparte por la administración de servidores. Tú simplemente escribes tu código y Lambda se encarga de todo lo demás

CESTE

AWS Lambda: ¿Cómo funciona?

- **Eventos:** Lambda se activa en respuesta a eventos.
- **Código:** Escribes el código en tu lenguaje favorito (Python, Node.js, Java, C# y otros) y lo subes a Lambda.
- **Ejecución:** Cuando se produce un evento, Lambda ejecuta tu código en un entorno aislado, proporcionando los recursos necesarios para que se ejecute correctamente.
- **Escalado automático:** Lambda escala automáticamente la cantidad de recursos para manejar la carga de trabajo, lo que significa que puedes manejar picos de tráfico sin problemas.

AWS Glue

Es un servicio totalmente administrado de **extracción, transformación y carga (ETL)**. Su propósito principal es **preparar y procesar datos** para análisis, machine learning y otros casos de uso relacionados con los datos.
Funciona como una herramienta para preparar los datos antes de ser analizados, ofreciendo componentes que permiten descubrir, ***catalogar***, limpiar, enriquecer y mover datos.

AWS Glue: Componentes principales

- Catálogo de datos (AWS Glue Data Catalog)
- **Crawlers**: Exploradores de datos
- **Triggers**: Código desencadenante
- Scripts de ETL (Jobs)
- Conexiones externas con recursos AWS
- AWS Glue Studio

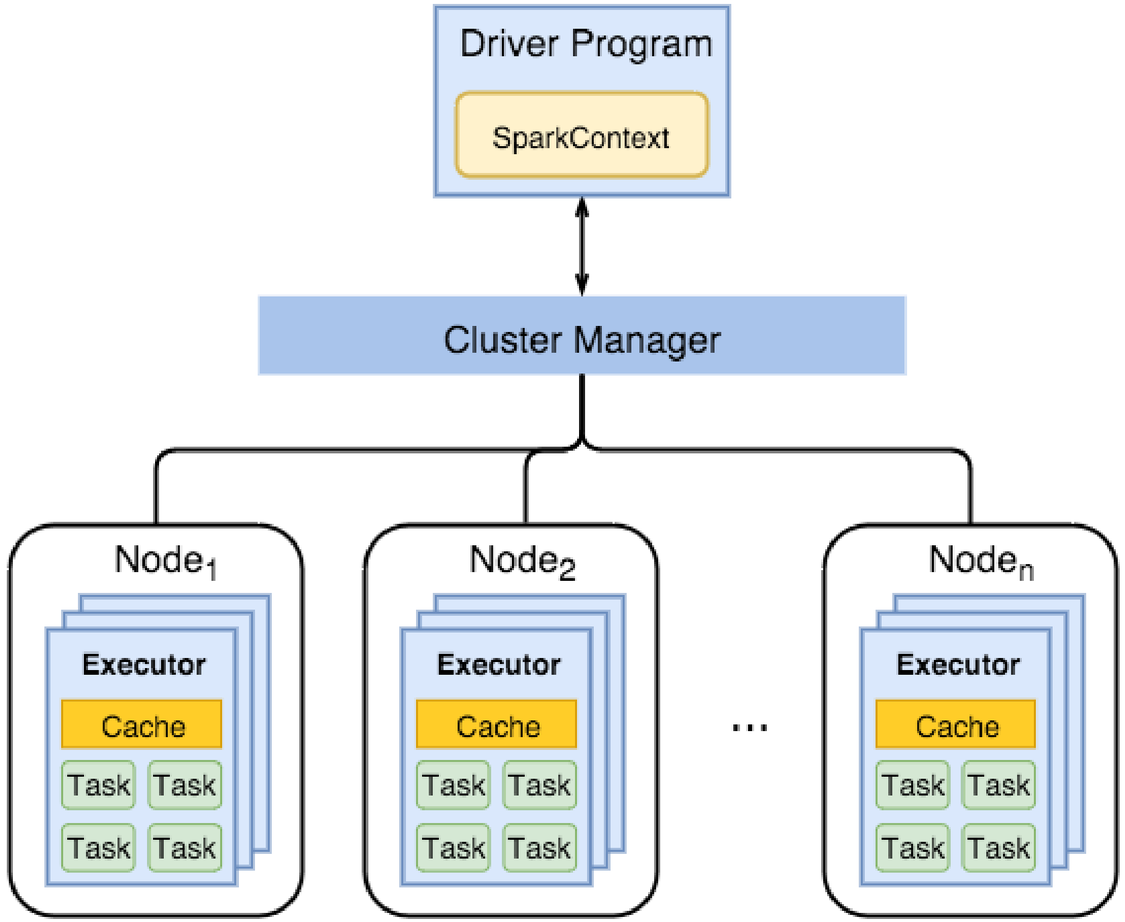
CESTE

AWS Glue: Data Catalog

Es su almacén de metadatos técnicos persistentes. Se trata de un servicio administrado que puede usar para **almacenar, comentar y compartir metadatos** en la nube de AWS.

AWS Glue Data Catalog está organizado en **bases de datos** y **tablas** para proporcionar una estructura lógica para almacenar y administrar los metadatos. Esta estructura permite un control preciso del acceso a los datos a nivel de tabla o base de datos mediante AWS IAM.

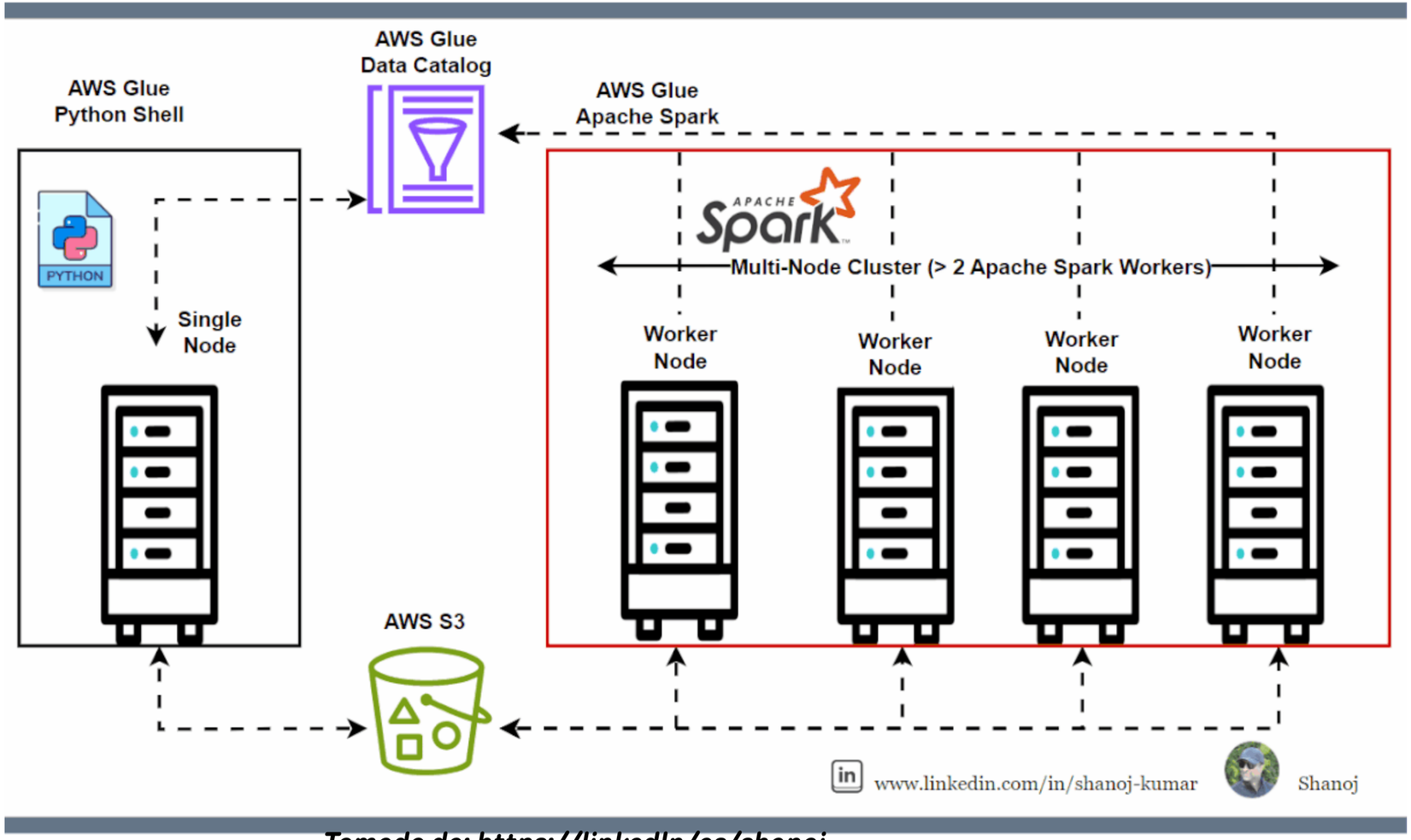
AWS Glue: Data Catalog – Spark



Tomado de: https://runawayhorse001.github.io/LearningApacheSpark/introduction.html

CESTE

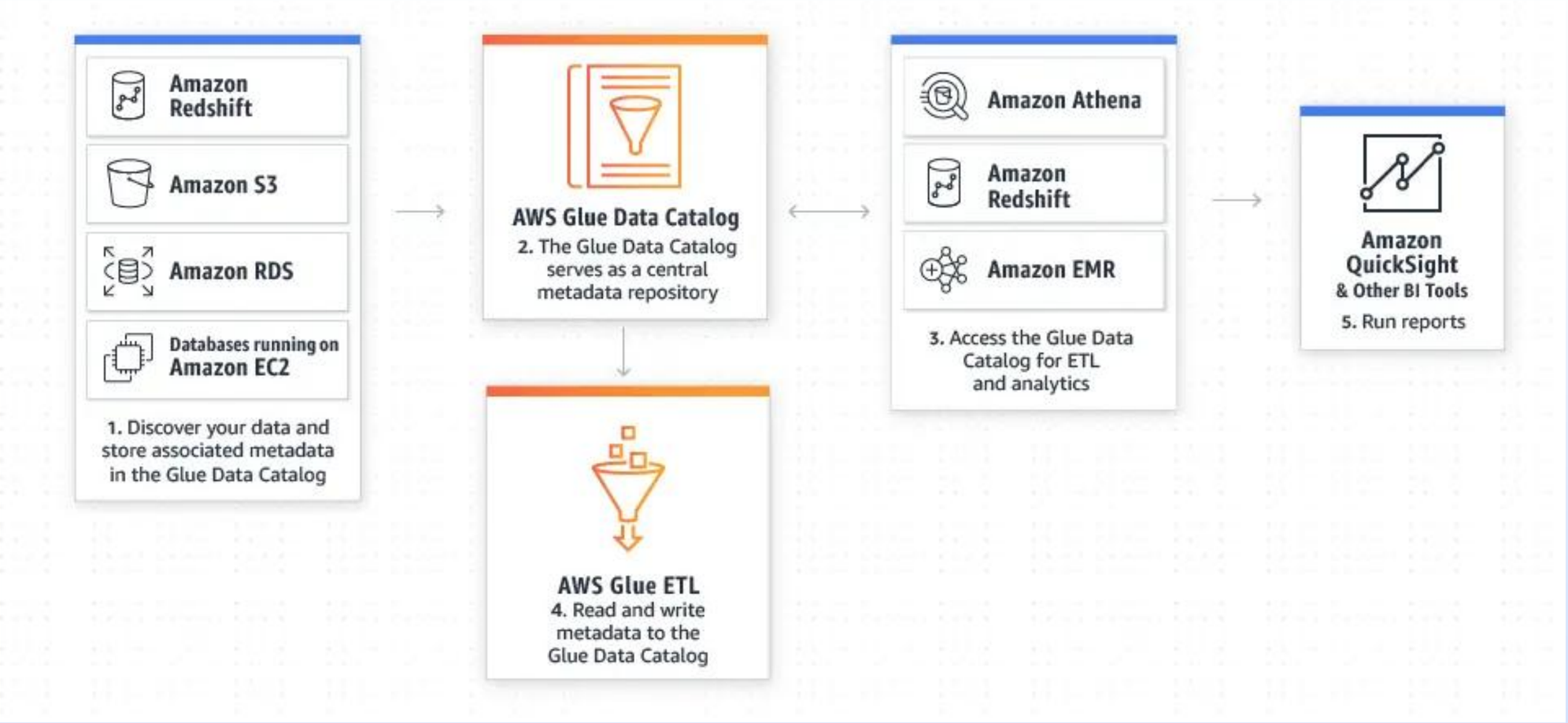
AWS Glue: Data Catalog – Spark



Tomado de: https://linkedIn/es/shanoj

AWS Glue

AWS Glue: Data Catalog



AWS Glue

Tomado de: https://aws.amazon.com/es/glue-catalog

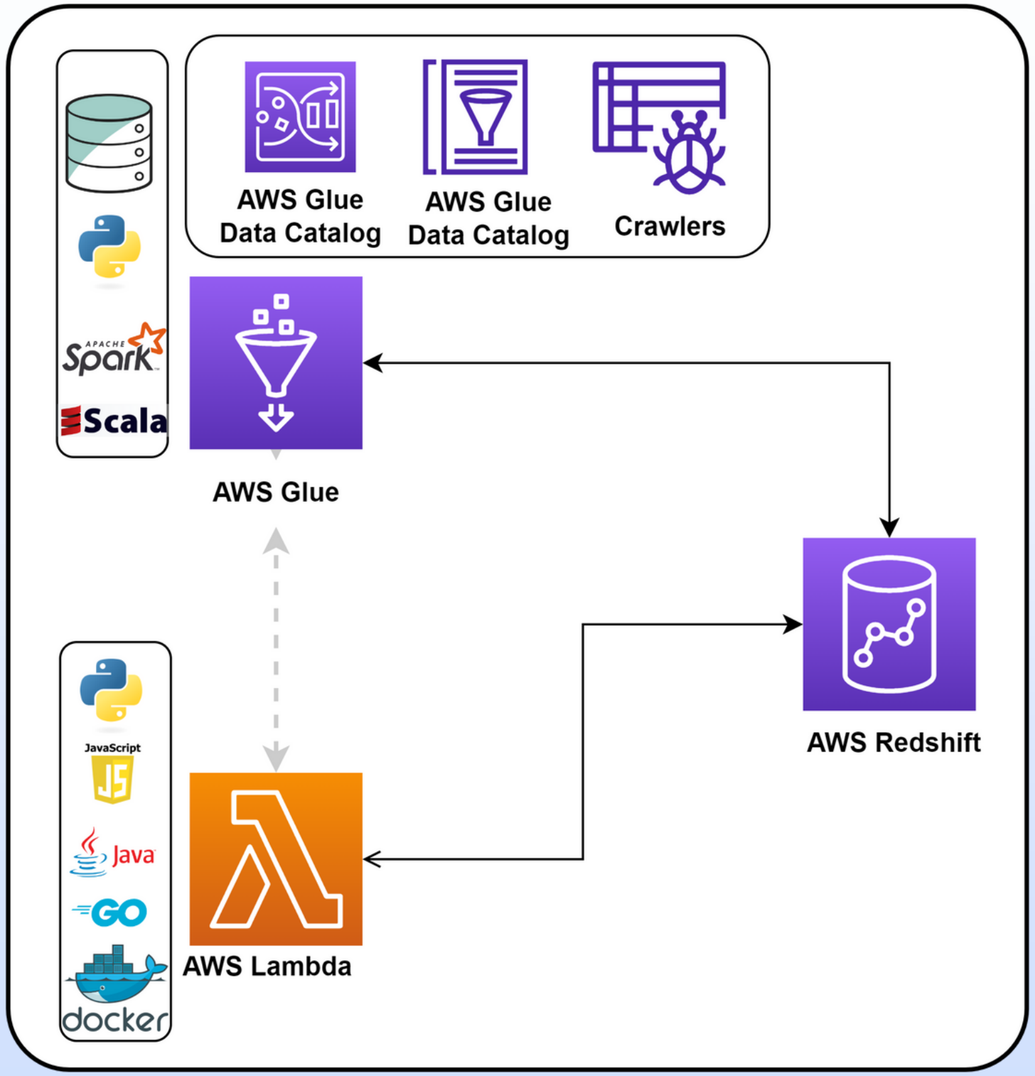
CESTE

AWS Lambda Vs. Glue

AWS Lambda	AWS Glue
Procesar archivos pequeños en S3 en tiempo real.	Procesar grandes volúmenes de datos en S3 o Redshift.
Activar funciones en respuesta a eventos (ej: SQS, DynamoDB, API Gateway).	Ejecutar pipelines ETL complejos para análisis de datos.
Automatizar tareas pequeñas (ej: reescalar imágenes).	Transformar y preparar datos para análisis con Spark.
Procesar eventos en streaming con Kinesis.	Catalogar datos con Glue Crawlers y Glue Catalog.

CESTE

AWS Lambda Vs. Glue



AWS Lambda

AWS Glue

¿Cuándo usar AWS Lambda vs Glue?

Escenario	Escenario
Procesamiento en tiempo real de datos pequeños.	AWS Lambda
ETL complejo sobre grandes volúmenes de datos.	AWS Glue
Tareas rápidas (máximo 15 minutos).	AWS Lambda
Tareas largas o procesamiento masivo.	AWS Glue
Automatización simple y basada en eventos.	AWS Lambda
Preparación y catalogación de datos.	AWS Glue

CESTE

AWS Redshift

Es un **data warehouse** en la nube administrado por AWS. Permite analizar grandes volúmenes de datos con alta velocidad y rendimiento, utilizando SQL estándar. Está diseñado para ejecutar consultas analíticas complejas y es ideal para casos de **business intelligence** (BI) y **análisis de datos**.

AWS Redshift: ¿Cómo funciona?

Utiliza una arquitectura de procesamiento **MPP (Massively Parallel Processing)**, lo que significa que divide los datos y las tareas entre múltiples nodos para procesarlas en paralelo.

- **Cluster:** El clúster es la unidad principal de Redshift.
  - **Nodo líder**: Recibe las consultas SQL y las distribuye a los nodos de cómputo.
  - **Nodos de cómputo**: Ejecutan las consultas y procesan los datos.
- **Almacenamiento columnar:** Los datos se almacenan en columnas en lugar de filas, lo que acelera las consultas analíticas que solo necesitan leer unas pocas columnas.
- **Integración con Amazon S3 y SQL:** Redshift puede cargar y exportar datos directamente desde Amazon S3.

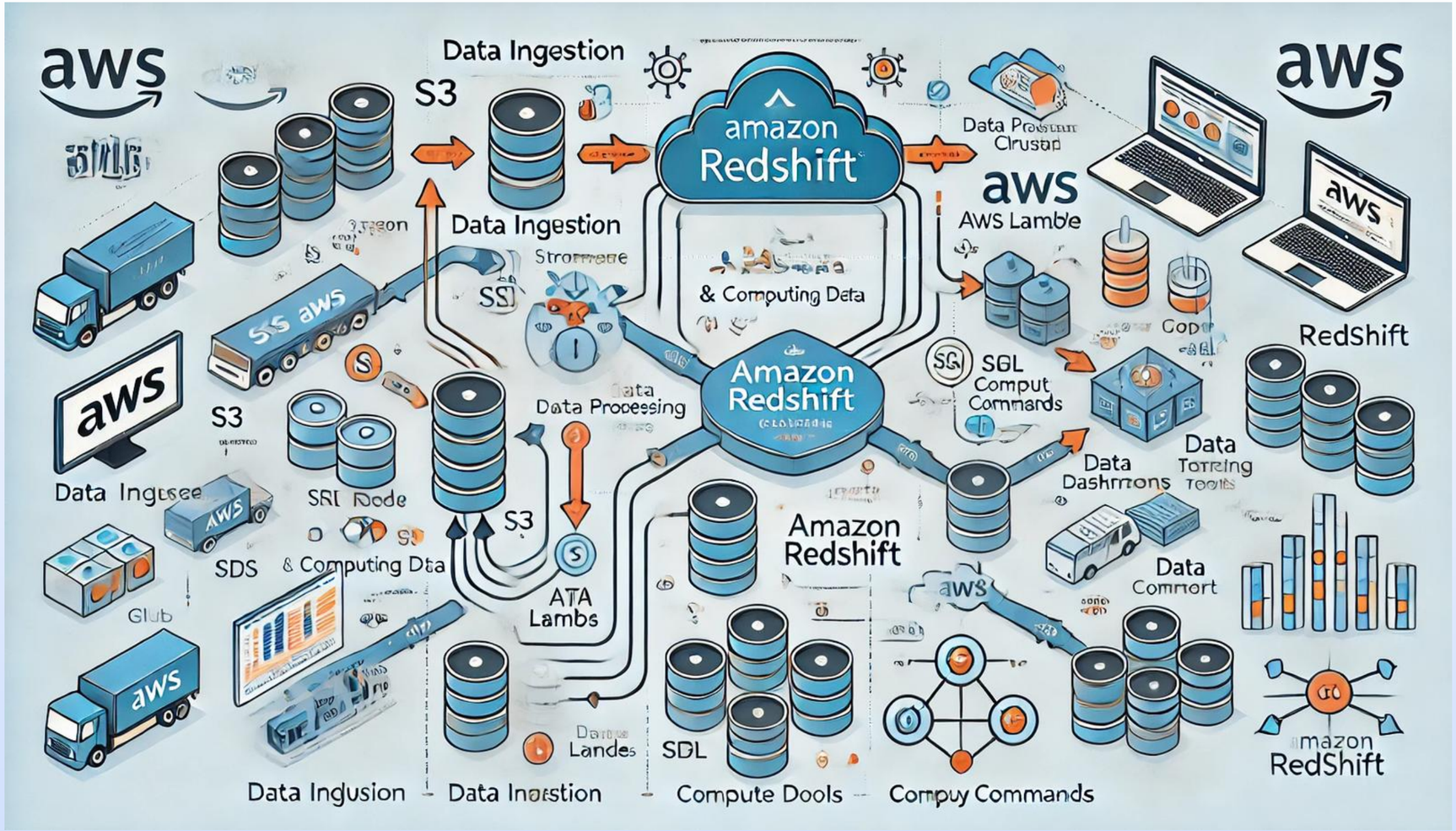
CESTE

AWS Redshift Vs. AWS RDS

Criterio	Amazon Redshift	Amazon RDS
Uso Principal	Procesamiento analítico (OLAP)	Procesamiento transaccional (OLTP)
Modelo de Almacenamiento	Almacenamiento columnar	Almacenamiento en filas
Escalabilidad	Escalable hasta petabytes	Escalabilidad limitada por el tamaño del motor de BBDD
Procesamiento	Procesamiento Paralelo Masivo (MPP)	Procesamiento en un solo nodo, sin distribución de carga.
Optimización	Diseñado para análisis de big data.	Optimizado para transacciones pequeñas y frecuentes.
Integración con AWS	Integración nativa con Glue, S3, QuickSight, y herramientas de ETL/ELT para análisis de datos.	Compatible con servicios AWS como Lambda, pero menos optimizado para flujos de análisis complejos.
Herramientas BI	Altamente compatible con Power BI.	Compatible con herramientas BI.
Costos	Mejor relación costo/rendimiento	Más económico para bases de datos pequeñas y aplicaciones transaccionales.
Tamaño de Datos	Ideal para conjuntos de datos grandes (> 1 TB).	Diseñado para bases de datos pequeñas a medianas (< 1 TB).

CESTE

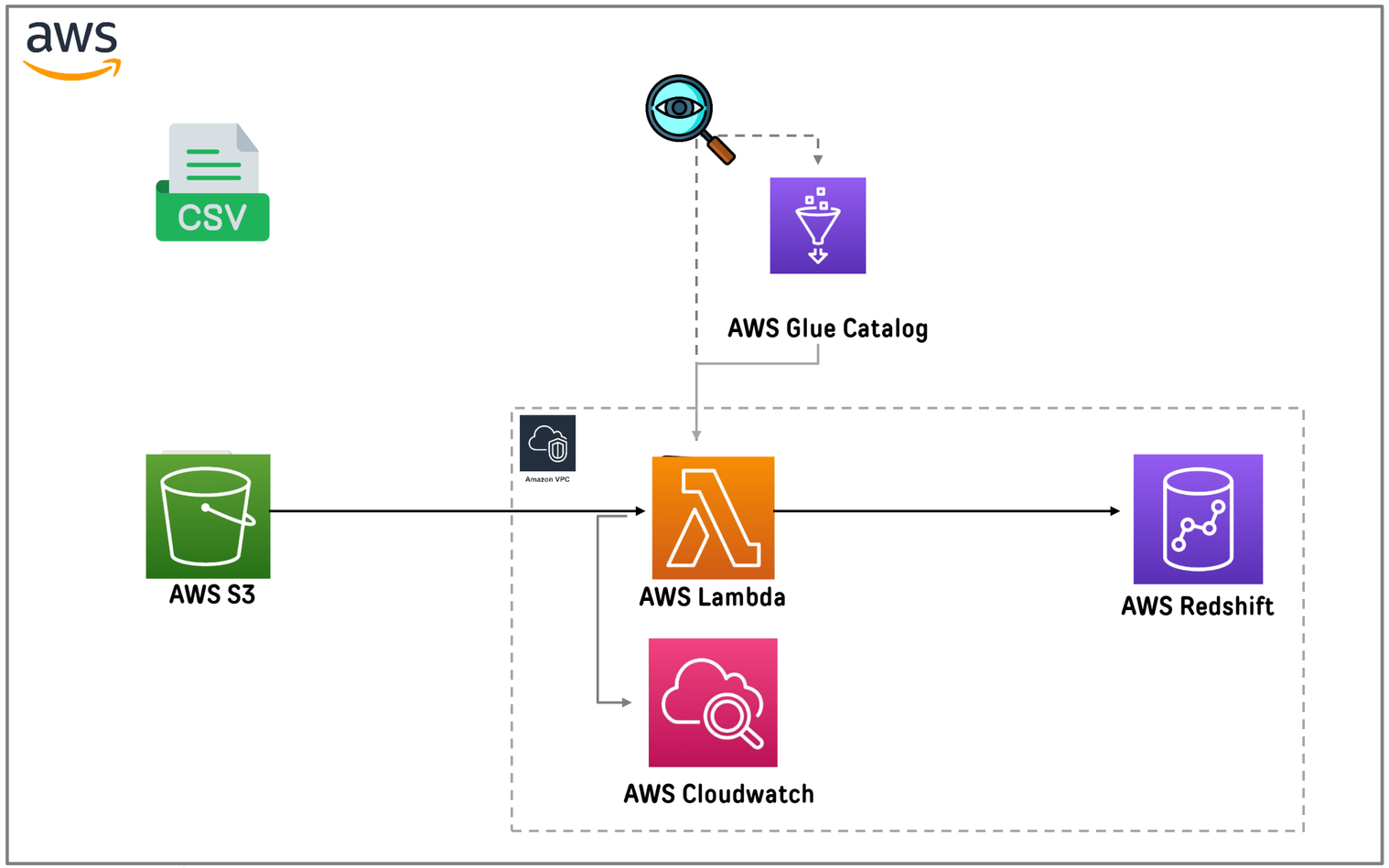
AWS Redshift: Dentro de un pipeline



Tomado de: https://aws.amazon.com/es/redshift

AWS Redshift

AWS Lambda Vs. Glue



AWS Lambda

AWS Glue