Arquitectura propuesta

gperezb1

February 19, 2025

Ingesta de datos

- Todos los datos se reciben en S3 una vez al día en archives Parquet.
- Por organización, se recomienda que ambos *labels* y *geo* estén en el mismo bucket

Proceso de ETL

- Herramienta usada: **AWS GLUE**
- Se lanza un job de Glue que lea los datos en S3
- Usando pyspark, se implementa el ETL diseñado.
- Ventajas por ser serverless, por integrarse directamente con pyspark.
- Salida del job $\implies$ se guarda en S3 en una carpeta designada para analítica

Orquestración del Pipeline

- Herramienta usada: **AWS step functions**
- Flujo: Se activa la ingesta de datos en S3 $\implies$ step function lanza el job de Glue $\implies$ Se monitorea el estado $\implies$ Se integra con SNS para notificar cualquier alerta sobre el estado.
- Ventajas: Todo dentro del mismo ambiente (AWS).
- Desventaja: Poca escalabilidad

Sistema de alertas

- Servicio de mensajería que envía notificaciones en respuesta a eventos o fallos en el pipeline (en forma de email, o el necesario).
- Ventajas: Bajo costo y sencillez de integración, centralización de alerta a multiples canales.

Herramientas para DATA SCIENTISTS

- Herramienta utilizada: **Amazon SageMaker**
- Oferce entornos preconfigurados y esalables
- Conecta con otros servicios de AWS, si es necesario.
- Incluye notebooks gestionados

Herramientas para BI

Se plantean dos herramientas: **AWS Athena** y **AWS Redshift**

Athena: Ofrece busquedas muy rápidas, siendo económica y cobrando por query.

RedShift: Optimizado para consultas complejas y alta concurrencia, integración con erramientas de BI

Estructura gráfica del flujo

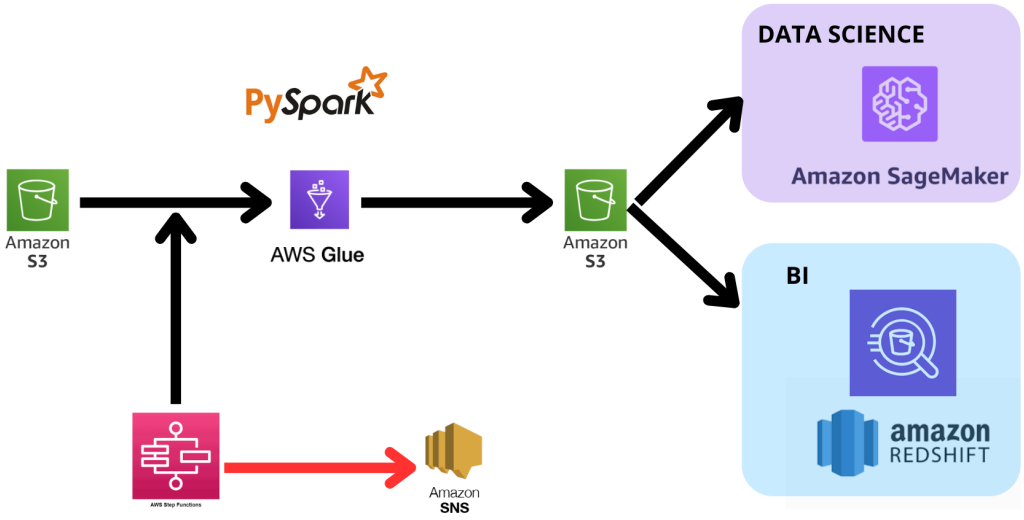

Figure 1: Estructura gráfica a seguir