MODULO 09 - EJERCICIO 05-B

ALEXIS YURI M.

Analice un escenario ficticio en el que un proceso ETL con AWS Glue no está funcionando correctamente. Su tarea será identificar posibles causas, proponer soluciones y sugerir una mejora en el monitoreo.

A continuación, se presenta el diagnóstico de los cuatro errores en el proceso ETL y las posibles soluciones:

1. Permiso.

El log de AWS Glue muestra errores de "access denied" en los objetos de S3. Esto significa que el rol IAM asociado al trabajo de Glue no tiene los permisos necesarios para leer los archivos de origen en el bucket de S3.

Solución: Se debe revisar la política de permisos del rol de servicio de AWS Glue. Se debe asegurar de que tenga permiso de lectura (s3:GetObject) en la ruta donde están los archivos CSV de ventas.

Monitoreo para evitar futuras fallas: Configurar alertas en tiempo real para fallas de autenticación o autorización (códigos 401, 403, "Access Denied") en los sistemas de origen y destino.

2. Tipo de Dato.

Los reportes muestran inconsistencias y los logs de Glue advierten sobre "mismatch" en los tipos de datos. Esto ocurre porque el esquema inferido por el crawler de Glue no coincide con la estructura de los datos que se están cargando. Por ejemplo, una columna que debería ser un número está siendo interpretada como texto.

Solución: Es necesario verificar y editar la definición de la tabla en el Catálogo de Datos de AWS Glue. Corrige los tipos de datos de las columnas para que coincidan con la información real de los archivos. También puedes configurar un clasificador de datos personalizado en AWS Glue.

Monitoreo para evitar futuras fallas: Establecer reglas de calidad (ej. rangos numéricos, formatos de fecha, valores nulos) y generar un informe de registros rechazados para cada ejecución.

3. Ir más Lento.

El trabajo de ETL tarda más de lo habitual en completarse. Esto puede deberse a un incremento en el volumen de datos, una configuración ineficiente del trabajo o la falta de optimización del formato de los archivos.

Solución: Para optimizar la velocidad del trabajo, puedes:

- Particionar los datos en el bucket de S3 por fecha. Esto permite a Glue procesar solo los datos nuevos, reduciendo el volumen de información.

- Convertir los archivos CSV a un formato columnar optimizado, como Parquet, en la etapa de transformación. Este formato es más eficiente para análisis y reduce el tiempo de procesamiento.

Monitoreo para evitar futuras fallas:Registrar y visualizar el tiempo de ejecución total y el tiempo consumido en cada etapa del proceso (ej. extracción, transformación, carga).

4. Se detiene antes de la carga final.

El trabajo ETL a veces falla y no llega a cargar los datos en Redshift. Este problema es una consecuencia directa de los errores de permiso y tipo de dato. El proceso se detiene cuando encuentra un registro que no puede leer o procesar por estas fallas.

Solución: La solución principal es resolver los dos primeros errores (Permiso y Tipo de Dato). Una vez que el trabajo de Glue tenga los permisos correctos y el esquema sea coherente, los errores de lectura y procesamiento desaparecerán, permitiendo que el trabajo se complete de principio a fin, cargando los datos correctamente en Redshift.

Monitoreo para evitar futuras fallas:Monitorear el registro de logs del scheduler o motor de orquestación (Airflow, Cron, etc.) para detectar errores durante la fase de inicialización o setup (ej. archivos de configuración faltantes, bloqueos).