Diagnóstico de un proceso ETL con fallas

Este documento presenta el diagnóstico de un job de integración de ventas implementado en AWS Glue, que presenta fallas en la consolidación de datos desde archivos CSV en S3 hacia una tabla en Redshift. Se identifican los problemas, sus causas probables y las acciones correctivas sugeridas.

Caso: Job de integración de ventas

La empresa de retail procesa diariamente archivos CSV cargados en Amazon S3 mediante AWS Glue, transformándolos y cargándolos en una tabla de análisis en Amazon Redshift. En la última semana se han detectado:

- Reportes con cifras incompletas.
- Registros faltantes y duplicados.
- Errores en logs relacionados con permisos y tipos de datos.
- Tiempos de ejecución más largos, con fallos antes de completar la carga en Redshift.

Diagnóstico y acciones correctivas

Problema detectado	Causa probable	Acción correctiva
Errores 'Access Denied' en logs de Glue	Permisos insuficientes en S3 (IAM o políticas de bucket mal configuradas)	Revisar y actualizar políticas IAM y permisos de acceso a objetos en S3
Advertencias 'Mismatch' en tipos de datos	Esquemas heterogéneos entre archivos CSV (columnas con tipos distintos)	Estandarizar esquemas y convertir datos a formato Parquet con tipificación uniforme
Registros faltantes o duplicados	Particiones mal configuradas en Glue Data Catalog	Revisar y ajustar las particiones en el catálogo de Glue
Tiempos de ejecución excesivos	Procesamiento de grandes volúmenes en CSV sin particionar	Optimizar formatos a Parquet/ORC y particionar datos en S3

Conclusión

El análisis sugiere que las fallas del proceso ETL están relacionadas con una combinación de problemas de permisos, inconsistencias de esquemas y deficiencias en el manejo de particiones y formatos de datos. La aplicación de las acciones correctivas propuestas permitirá mejorar la calidad de los reportes, reducir errores y optimizar el rendimiento del pipeline de integración.