Automatización de un flujo de datos con AWS Lambda

Este documento presenta una solución basada en AWS Lambda para automatizar el procesamiento de archivos CSV cuando son cargados en un bucket S3. La función Lambda se encarga de transformar los datos y almacenarlos en otro bucket en formato optimizado para análisis.

Escenario

Una empresa necesita procesar automáticamente archivos CSV de clientes que se cargan en un bucket S3. Cada archivo debe limpiarse (eliminar columnas irrelevantes y filtrar registros inválidos) y luego guardarse en otro bucket en formato Parquet, optimizado para consultas posteriores.

Flujo propuesto

1. Evento: carga de archivo en S3 (bucket-raw) → dispara Lambda.
2. Lambda (función Python):
   - Lee archivo CSV.
   - Aplica transformaciones: elimina columnas innecesarias, filtra duplicados y registros nulos.
   - Convierte el archivo a formato Parquet.
   - Guarda el resultado en S3 (bucket-processed).
3. Destino: datos listos para análisis con Athena o Redshift Spectrum.

Permisos necesarios

- Rol IAM para Lambda con permisos:
  - s3:GetObject sobre bucket-raw.
  - s3:PutObject sobre bucket-processed.
  - Logs en CloudWatch.

Ejemplo de código Lambda (Python)

```
import boto3
import pandas as pd
import io

s3 = boto3.client('s3')

def lambda_handler(event, context):
    bucket = event['Records'][0]['s3']['bucket']['name']
    key = event['Records'][0]['s3']['object']['key']

    obj = s3.get_object(Bucket=bucket, Key=key)
    df = pd.read_csv(io.BytesIO(obj['Body'].read()))

    df = df.drop(columns=['extra'], errors='ignore')
    df = df.dropna()

    out_buffer = io.BytesIO()
    df.to_parquet(out_buffer, index=False)

    dest_bucket = 'bucket-processed'
    dest_key = key.replace('.csv', '.parquet')
    s3.put_object(Bucket=dest_bucket, Key=dest_key, Body=out_buffer.getvalue())

    return {"status": "ok", "file": dest_key}
```

Tabla del flujo de datos

Evento	Transformación	Destino
Archivo CSV cargado en S3 (bucket-raw)	Lambda limpia columnas, elimina duplicados y convierte a Parquet	Archivo transformado en S3 (bucket-processed), listo para Athena/Redshift

Conclusión

El flujo automatizado con AWS Lambda permite procesar datos sin servidores, reduciendo costos y escalando automáticamente según la carga. Los archivos resultantes en formato Parquet facilitan el análisis en Athena o Redshift.