Consideraciones:

Este challenge se puede resolver de diversas maneras y con diferentes tecnologías.
Se valorarán las soluciones que utilicen algunas de las siguientes tecnologías:

1. Servicios de AWS
2. Spark
3. Python
4. Docker
5. IaC (Cloudformation o SAM)

Puntos a evaluar:

- Orden y comentarios del código.
- Decisiones de diseño tomadas.
- Calidad del entregable.
- Cualquier detalle adicional que consideres que pueda aportar valor y/o esté alineado con buenas prácticas.

Objetivo:

Se requiere la construcción de un data pipeline encargado de leer un dataset, ingestar los datos y transformarlos por las distintas capas de un mini-datalake que deberás diseñar. No importa cuantas capas elijas, mientras puedas justificar tu elección y las características de cada una.
Finalmente, se deberá generar y alimentar un modelo dimensional en la última capa del datalake.

Descripción del pipeline

1. El pipeline debe ser capaz de manejar tanto cargas iniciales como incrementales con nuevos datos.
2. Se requiere que todas los seteos del pipeline estén en un archivo de configuración (en el formato que prefieras).
3. Se requiere que la información en el datalake esté comprimida.
4. Se requiere que los datos estén particionados con el formato year=xxxx, month=xx, day=xx.
5. Se requiere una función que elimine duplicados para ser reutilizada en donde sea necesario.
6. La última capa del datalake tiene que estar lista para ser consumida de acuerdo al modelo dimensional.

Carga inicial

1. La fuente de datos para la carga inicial está en formato csv.. Se puede descargar desde el siguiente link: dataset_ini.zip

Carga incremental

1. La fuente de datos para la carga incremental está en formato csv.. Se puede descargar desde el siguiente link: dataset_inc.zip.
2. Puede incluir registros que no se hayan cargado previamente en la carga inicial (inserts) así como también registros que sí ya se cargaron (mismo ***id***) pero que tienen algunos valores actualizados para algunas de sus columnas en un momento dado (**updated_date)**.
3. Es requisito ingestar los datos en modo ***append only***, pero luego deberán manejarse de alguna forma las actualizaciones para representar el último ***snapshot*** de los datos en la siguiente capa.

Modelo dimensional

El modelo a cargar estará conformado por las entidades descritas a continuación (los campos en *italic* deberán generarse):

Category:
PK, category_code, category_name, category_color, *loaded_date*

Manufacter: *PK*, manufacturer_id, manufacturer_name, *loaded_date*

Product: *PK,* sku, unit, title, subtitle, details, ean, brand, flavor, variant, category_code, *product_type*, *loaded_date*

product_type = [“**Coca cola sin azúcar”, “Coca cola original”, “Other”**]

Donde:

1. **Coca cola sin azúcar:** filtro sobre el campo ***subtitle*** que permita obtener los siguientes valores:
    - Coca-Cola Sin Azúcar Retornable
    - COCA-COLA SIN AZÚCAR SIN AZÚCARES
    - Bebida Coca-Cola Sin Azúcar Mediana
    - COCA-COLA SIN AZÚCARES
    - Coca-Cola Sin Azúcar
2. **Coca cola original:** filtro sobre el campo ***subtitle*** que permita obtener los siguientes valores:
    - Coca-Cola Original
    - Bebida sin Alcohol Coca-Cola Sabor Original
    - Coca-Cola SABOR ORIGINAL
    - COCA-COLA SABOR ORIGINAL SABOR ORIGINAL
    - Bebida Coca Cola Sabor Original
3. **Other:** todos los demás valores de ***subtitle.***

Sales: *PK, FK_product, FK_manufacturer, FK_category*, delivery_id, amount, currency, quantity, retail_amount, created_date, updated_date, *loaded_date*