gelt data facts

Prueba Data Engineer

Tiempo estimado: 8/10 horas

En Gelt Data Facts recopilamos muchos tipos de datos de distintas fuentes: datos de los usuarios, de los tickets que suben a la aplicación, de lo que compran, etc. Todos estos datos necesitan un paso de procesamiento y limpieza inicial para que los equipos de Analytics y Science puedan hacer uso de ellos. Aquí es donde entras tú.

Adjuntos a esta prueba recibirás **3 ficheros CSV:**

- **users.csv**: Contiene información relativa a los usuarios, como su año de nacimiento, su género, y diversos datos sobre la configuración de su hogar.

- **tickets.csv**: Contiene información sobre una muestra de tickets de estos usuarios (asociado por user_id), como el supermercado donde se efectuó la compra, el modo de pago y la fecha.

- **ticket_lines.csv**: Contiene información detallada sobre las líneas de producto de cada uno de los tickets presentes en tickets.csv (asociado por ticket_id). Se puede encontrar el nombre del producto, las unidades, el precio total pagado y una categorización del producto en dos niveles.

Como podrás comprobar, esta información está en un estado bruto, por lo que tu labor consistirá en ingerir, limpiar, procesar y guardar los datos en un formato más adecuado para su explotación.

Para ello utilizarás Python y SQL para manejar las tablas y las transformaciones correspondientes, para finalmente guardar el resultado de las transformaciones en un almacén persistente.

Ejercicio 1. **Carga y limpieza de los datos.**

Tu primer objetivo será estudiar los ficheros, familiarizarte con los datos que contienen, sus tipos y cómo se relacionan entre ellos. Tras esto, podrás efectuar una primera limpieza de los datos, rellenando valores nulos cuando corresponda, arreglando tipos mal asignados y reemplazando valores incorrectos.

Objetivos:
- Analizar los datos que se te presentan, entendiendo qué representan, qué tipos de datos contienen y el grado de limpieza de los mismos.
- Rellenar valores nulos en los campos que corresponda. (Por ejemplo, el campo users.kids_at_home tiene valores nulos, pero quizás es mejor representar la misma información con un valor 0)
- Corregir tipos de datos. (Por ejemplo, el campo users.gender se define como 0, 1, 2 ó 3, pero quizás sería mejor definirlo como ‘Masculino’, ‘Femenino’, ‘No binario’ y ‘Desconocido’)
- Reemplazar valores incorrectos. (Por ejemplo, el campo users.pet tiene valores ‘Perro’, ‘Gato’, ‘Otros’ y ‘0’. Este último valor no concuerda con los demás en cuanto a nomenclatura)

Tras este ejercicio, tendrás una data en bruto más confiable con la que poder trabajar.

***Ejercicio 2*. Generar nuevos campos analíticos.**

En este ejercicio, tu objetivo es transformar la data cruda en información de interés, que permita a los analistas extraer insights de la misma.

Algunas de las transformaciones que puedes hacer:

- Campo age_bracket para los usuarios. Basándote en users.birth_year, puedes calcular la edad de los usuarios y asignarles uno de los siguientes rangos de edad:

Edad:	**Grupo:**
Entre 15 y 24 años	15-24 Gen Z
Entre 25 y 39 años	25-39 Millenials
Entre 40 y 54 años	40-54 Gen X
Entre 55 y 75 años	55-75 Boomers
Otros	Resto

- Campo home_type para los usuarios. En base a la configuración de su hogar (número de adultos y de niños), puedes asignarles un tipo de hogar en base a la siguiente tabla:

adults_at_home	kids_at_home	home_type
1	0	Singles
1	1 o más	Singles with kids
2	0	Couples
2 a 5	1 ó 2	Families
2 a 5	3 o más	Large families
3 o más	0	3 or more adults
Otros	Otros	Other

- Campo total_amount para los tickets. Se puede obtener agrupando las líneas de cada ticket y sumando sus precios totales, de manera que tengamos el importe total del ticket.
- Campo preferred_payment_type para los usuarios. Se puede calcular como el tipo de pago (TARJ, EFE) más recurrente para los tickets de cada usuario (En caso de empate, es desconocido)

***Ejercicio 3*. Nuevo modelo de tablas.**

En este ejercicio, el objetivo es generar las tablas finales que el equipo de Analytics utilizará para sus estudios. Tienes libertad total para elegir el esquema de tablas que mejor te parezca. Puede ser un esquema relacional, puedes agregar y cruzar las tablas entre sí como mejor te parezca de cara a disponer de la data de la forma más cómoda posible. Tomes la decisión que tomes, debes justificar por qué decidiste ese formato.

Algunas opciones son:
- Una tabla para cada entidad.
- Tablas que agrupen diferentes entidades relacionadas.
- Una sola tabla con toda la información agrupada.

Tras este paso, deberías tener una o más tablas preparadas para ser almacenadas en un sistema persistente, como una base de datos.

***Ejercicio 4*. Wrapping it up.**

Este ejercicio es el paso final, en el que todo lo que hiciste hasta ahora se junta para construir un pipeline de datos. El objetivo es generar un script en Python que reciba los ficheros CSV, aplique los pasos de limpieza y transformación que has diseñado, y escriba los resultados en una persistencia que el equipo de Analytics pueda acceder y consultar.
Lo ideal sería armar una estructura basada en Docker, que levante una base de datos (del sabor que elijas, mySQL, Postgres, Mongo, etc.), ejecute el script de extracción, transformación y carga, y escriba las tablas finales en esa base de datos persistente.

***Ejercicio BONUS*. Analizando los datos.**

En este ejercicio bonus, puedes diseñar una serie de queries para responder a las siguientes preguntas:

- *Cuál es el ranking de supermercados más frecuentes?*
- *Cual es la canasta básica? (Es decir, las categorías más frecuentes presentes en los tickets)*
- *Cuál es el importe medio de ticket por tipo de hogar?*
- *Cuál es el método de pago favorito por rango de edad?*

Consideraciones:
- Se entiende que muchas partes de esta prueba están sujetas a interpretación, por lo que tienes total libertad para tomar las decisiones que te parezcan oportunas en función de lo que consideres que sea la mejor solución al problema.
- En caso de duda o ambigüedad en algún punto, documenta la decisión que tomes y avanza con ella.
- Esta prueba asume ciertos conocimientos sobre una serie de tecnologías. Si no dispones de ellos, siéntete libre de simplificar los requisitos con los que menos confianza sientas. Las tecnologías se aprenden, pero lo que importa es intentarlo.
- Cualquier duda que te surja y sientas que no te permite continuar, no dudes en consultarnos y te guiaremos en la dirección correcta.