ETL Engineer Case

Hi there!

Welcome to our ETL Engineer Test!

Please read each question thoroughly, don't rush, you have time :)

We recommend the following playlist for you to listen to while performing the test:

Cyberpunk 2077 Radio by NightmareOwl

Here we go!

Cornershop operates in more than 25 cities, reaching tens of thousands of orders and customers per day. It is for this reason that the volume of data we generate day by day is huge.

In this case, we ask you to help us in the development of our Data Lake, where we need to store data structures with "pre-calculated" KPIs, in order to facilitate analysis to the different areas of the company.

A tiny sample of our data is included, and it is expected that you can solve the requested using it. Please take the following into consideration:

- The data is anonymized to protect the privacy of customers and shoppers.
- It is a relatively small sample of data, however your answers should be thinking about scalability. The idea is that you show off your talents regarding Data Extraction, Transformation and Loading (ETL) always targeting large volumes of data.
- We expect you to use SQL for querying the data, MySQL or Postgres will do :)

  (You can use any ETL tool you want, but we'd love you if you use Airflow and SQL)

File description and data fields

1. **order_products.csv:**
   - order_id: ID of the order
   - product_id: ID of the product
   - quantity: The quantity ordered of this product
   - buy_unit: The unit of the product (KG/UN)

2. **orders.csv:**
    - order_id: ID of the order
    - lat: The latitude of the delivery location
    - lng: The longitude of the delivery location
    - promised_time: The delivery time promised to the user
    - on_demand: If true, the order was promised to be delivered in less than X minutes
    - shopper_id: ID representing the shopper completed the order.
    - store_branch_id: ID of the store branch
    - total_minutes: The total minutes it took to complete the order (label)

3. **storebranch.csv:**
    - store_branch_id: ID of the store branch
    - store: ID representing the store
    - lat: Latitude of the branch location
    - lng: Longitude of the branch location

Questions:

A. The operations team is concerned because distances between branch locations and delivery locations (customers homes) have increased a lot. It's for that reason that they've requested a data source in order to make a report about it.

   By using the available tables (described above) we need to build a new table, where the "Distance between branch location and delivery location" KPI is consolidated.

   Pro tip: You should use Manhattan distance.

   What we ask of you is a new table with the following structure:

   a. Date in format 'YYYY-MM-DD'
   b. Week of the year (ISO Format)
   c. Store
   d. Store Branch ID
   e. Average traveled distance
   f. Median of the traveled distance
   g. For the following intervals in KMs:
      $[0,10] -]10,15] -]15,20] -]20,\infty[$
      Find the total percentage of shoppers that fall into each travelling interval for each day. A single shopper cannot be in two intervals at once on the same day.
      (This should output 4 columns)

B. The Capacity team needs a consolidated data source to visualize the relationship between order times (total minutes it took to complete an order) and the unique quantity of products and the total quantity of products (total items) of each order.

Notice: When I request 3 pineapples, the unique product quantity is 1, but the total quantity of products (or units) is 3.

We ask you to generate a table containing the following structure:

a. Date in format 'YYYY-MM-DD'
b. Week of the year (ISO Format)
c. Store
d. Store Branch ID
e. Median of Total Minutes
f. Average of Total Minutes
g. Average quantity of unique products and average quantity of total products (total units) for orders that take on average the following time intervals (in minutes): $[0, 25] - \ ]25{,}40] -]40{,}60] -]60, \infty[$ (This should output 8 columns)

The following questions don't have right or wrong answers, we just want to know how you approach problems :)

C. According to your experience, describe what you understand by 'ETL Pipeline' and explain each step of the process.

Una "ETL Pipeline" es un flujo de trabajo que consiste en 3 etapas:

- Extracción de datos desde un almacenamiento principal o un conjunto de archivos con el objetivo de realizar alguna tarea para un o unos objetivos particulares
- Transformación de los datos, ya sea para manejar la presencia de valores nulos, ajustes de escala o el cálculo de datos basado en los extraídos, esto entre otras posibles acciones a tomar dependiendo de las características de los datos extraídos. La transformación siempre debe ser realizada considerando los requerimientos que establece el objetivo que se desea alcanzar con la data.
- Por último, la etapa de carga donde se almacenan los datos transformados para que sean consumidos y de esta forma se pueda trabajar en cumplir el objetivo establecido previamente.
-

El pipeline debe considerar también que estas tres etapas pueden requerir su ejecución periódica y en horarios específicos para no interferir con la operación y apuntando a la eficiencia tanto en velocidad de ejecución como también en costos.

D. Describe a problem you were able to solve by using an ETL or one of its variants.

Un problema que fui capaz de resolver mediante ETL está relacionado con la extracción de datos dentro de la red interna de la empresa en la que trabajo, donde gracias a una máquina virtual dentro de la red pude crear un script en python que se encargará de extraer los datos desde la base, realizar limpieza de ellos eliminando columnas que no eran importantes para la solución y cargándolas a un almacenamiento en la nube.

La solución que requería los datos era un dashboard para monitorear las métricas de los ejecutivos en el canal de Whatsapp. Parte del procesamiento de los datos, correspondiente a cálculos de promedios y cálculos de métricas agrupadas por horas o días, fue desarrollado por el lado de la solución más que por el lado del script en la máquina virtual. Esto podría modificarse, para que la maquina suba los datos agrupados, pero dado que se requería información especifica de cada ejecutivo, se optó por hacer la carga sin mayor trabajo previo.

Aparte de esto otros trabajos que he realizado con ETL son más del lado académico, en la universidad en las actividades de un electivo de Inteligencia de Negocios, utilizando la herramienta Talend.

E. As you might know, Cornershop has presence in 8 countries, so we span across many time zones.

   Knowing that the last city to finish its operation is San José, Costa Rica, **and** that we need to have the previous day's data processed as soon as possible, **and** that we can not start processing it before the operation of that day finishes:

   At what time in UTC Time would it be reasonable for a daily ETL that processes data from the previous day's operation to start? Justify your answer

Dado que es necesario que el procesamiento se inicie luego de haber terminado las operaciones del día, y tomando como supuesto que las operaciones terminan a las 22:00 en San José, considero que el proceso ETL debería comenzar a 40 minutos después, que corresponde a las 4:40 UTC, este margen de 40 minutos es para que los pedidos realizados al borde del cierre de operaciones puedan ser concluidos y considerados dentro de la data del día anterior.

Please attach results and all files used.