Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación
Primer Semestre de 2024

IIC 2440 – Procesamiento de Datos Masivos
Tarea 1

1. Enunciado

En esta tarea van a trabajar sobre datos relacionados a un *e-commerce* que vende frutas en distintas regiones. En la primera parte vas a tener que crear un modelo de datos para una base de datos transaccional, pensando que esta base de datos es la que soporta a una aplicación utilizada por usuarios del *e-commerce*. En la segunda parte vas a tener que crear un *dataset* en BigQuery y hacer ciertas consultas de analítica utilizando SQL. En base a los resultados de las consultas, también deberás hacer algunas visualizaciones. Finalmente, debes exponer los principales resultados en un video que debes subir a internet.

2. Parte 1 [1.5 pts] - Creando un modelo de datos y sus índices

Tienes un amigo que tiene un emprendimiento de venta de frutas a domicilio a través de una aplicación web. Lamentablemente, tu amigo ha tenido bastantes problemas con su base de datos porque las consultas funcionan muy lento y el modelo de datos se ha vuelto muy engorroso, por lo mismo, te pide ayuda para volver a crear el modelo de datos junto con los índices necesarios para que las consultas funcionen rápido. Para esto, tienes que tener en consideración la siguiente descripción de la aplicación.

La aplicación tiene que vender los productos de la tienda. Estos productos son frutas que tienen un nombre, una descripción y un precio unitario. Estos productos son comprados por clientes que tienen un nombre, correo, entre otros atributos típicos de una tabla de usuarios. Estos clientes pueden comprar frutas en un pedido, y un pedido puede tener varias frutas en distinta cantidad. Un pedido tiene una fecha de despacho y una fecha de entrega. Los pedidos son despachado a alguna de las direcciones del cliente, y un cliente puede tener una o más direcciones. Cada dirección incluye la calle, número, región, código postal y detalles adicionales (ej. dpto, oficina, etc.). Algunas de las consultas que nos interesa hacer son:

- Entrega la fruta con identificador i junto a toda su información.
- Para un usuario, entrega todos sus pedidos, junto con la dirección, fecha de despacho, fecha de entrega y total en pesos ($) del pedido.
- Para una compra de un usuario, entrega cada fruta junto a la cantidad que se compró, y el total en pesos ($) gastado en la compra separado por fruta.
- Para una fruta en particular, entrega los usuarios que la compraron en un mes m.

Además, puedes considerar alguna consulta que debe soportar la aplicación que no esté en las descritas más arriba.

Para esta parte se pide que:

1. Entregues el esquema, junto con las llaves primarias y foráneas.
2. Para cada tabla, indica cuáles son los índices que crearías.
3. Para cada consulta de interés, indica el plan de consultas y haz énfasis en cómo los índices ayudan a responder la consulta de forma eficiente.

3. Parte 2 - [4 pts] - Haciendo analítica en un Data Warehouse

Parte 2.1 (0.5 pts). Aunque tu amigo va a renovar su modelo de datos en base a lo que le digas, necesita proyectar la demanda de frutas para el año que viene en base a los datos que obtuvo del año anterior. Su idea era cargar los datos en la misma base de datos de la aplicación y correr ahí consultas de analítica.

1. Para esta parte se pide que le respondas a tu amigo por qué lo quiere hacer es mala idea y justifiques por qué hay que cargar los datos en un Data Warehouse, y hacer la analítica ahí.

Parte 2.2 (1 pto). Para ayudar a convencer a tu amigo de cargar los datos en un Data Warehouse, vas a adelantar trabajo y cargar los datos en BigQuery para mostrarle lo bien que funciona y lo mucho que puede ganar. Tu amigo tiene presencia en dos regiones y te entregó los datos separados por región. Los archivos se llaman:

- `ordenes_r<x>.json`: las órdenes de compra de la región `<x>`. Por ejemplo, los datos de la región 1 se llaman `ordenes_r1.json`. Tiene la información por cada mes de qué fruta se compró en qué orden.
- `usuarios_ordenes_r<x>.json`: la información de qué usuario hizo qué orden en qué fecha. También hay un archivo por región.

Una vez teniendo estos archivos, necesitas hacer lo siguiente.

1. Como los datos no presentan un formato amigable, lo primero que tienes que hacer es cargar los datos a un esquema normalizado que diseñes tú. Tú diseño debe ser elegante, por lo que no deben haber tablas separadas por región, y debe estar pensado en que se agregarán más regiones en el futuro.
2. Luego, corre algunas consultas para validar los datos. Lo mínimo que se espera es que compruebes que cada orden tiene un usuario asociado, pero la idea es que también propongas algún test a los datos que te haga sentido.

Parte 2.3 (1.5 pts). En esta parte tienes que hacer consultas de analítica sobre los datos. Piensa que estas consultas las vas a guardar como vistas para que tu amigo pueda usarlas cuando quiera, por lo que tienen que ser eficientes (además, no vas a convencer a tu amigo de usar BigQuery si las consultas funcionan lento o no son elegantes). En concreto se pide que hagas las siguientes consultas:

1. Dada una región, entrega el número de unidades vendidas al mes para cada fruta.
2. Dado un cliente, entrega como ha evolucionado el dinero gastado por el cliente en el tiempo.
3. Para cada región y mes, entrega la fruta más vendida.
4. Calcula para cada semana el dinero que ha entrado a la tienda dicha semana. Luego entrega el dinero entrante acumulado.

Parte 2.4 (1 pto). En esta parte tienes que hacer visualizaciones que te permitan extraer conocimiento de los datos. Se espera que al menos entregues una visualización para:

1. El número de unidades vendidas el mes de cada fruta, separado por región.
2. El dinero acumulado que ha entrado a la tienda.
3. La distribución del dinero que gastan los clientes en la tienda. Hazlo de forma que sea un gráfico que se entienda (por ejemplo, algunos boxplot para distintos meses o grupos de meses).

En base a tus análisis se te pide que ayudes a sacar valor de los datos. En concreto, ayuda a responder las siguientes preguntas:

4. ¿Que región crees que va en crecimiento y va a aumentar su demanda el siguiente año?
5. ¿Puedes ver alguna relación entre la venta de las frutas?
6. Tu amigo tiene una predicción de la demanda del kiwi para el próximo año en distintas regiones. ¿Cómo lo harías para predecir la venta de una o más frutas en base a esa predicción?

4. El video que debes entregar

Para explicar los resultados de la tarea debes hacer un video. Este video debería dedicar cerca de 5 minutos para la parte 1 y cerca de 10 minutos para la parte 2. Nos interesa que nos respondas todos los puntos que enumeramos en este enunciado. En concreto:

- Para la parte 1 esperamos algún diagrama del modelo y ver explicaciones de alto nivel de los índices y de los planes de consulta. No esperamos ver consultas SQL detalladas, pero sí queremos que nos expliques bien por qué tu modelo y tus índices cumplen con lo solicitado.
- Para la parte 2.1 esperamos una explicación con justificaciones técnicas.
- Para la parte 2.2 esperamos que nos enseñes el modelo de datos que utilizaste y nos cuentes las validaciones que hiciste.
- Para la parte 2.3 no nos interesa que nos muestres SQL detallado, solo que nos expliques en un alto nivel como funcionan las consultas. Haz énfasis en las consultas que usan funciones de ventana y explica como decidiste la partición, el rango a considerar para cada ventana y la función utilizada.
- Para la parte 2.4 nos interesa ver las visualizaciones, y que estas sean de calidad. Además nos interesa que nos muestres tus análisis apoyado en los datos y las mismas visualizaciones.

5. Detalles académicos

Esta tarea debe resolverse en grupos de dos personas. El formato de entrega consta de los siguientes archivos:

- Un archivo `.pdf` que contenga el esquema de los datos, junto a las llaves y los índices.
- Un archivo `.pdf` que contenga las consultas para validar tus datos de la parte **2.2** y las consultas de la parte **2.3**. Para cada consulta, entrega un pantallazo de la consulta en BigQuery en donde se vea la respuesta.
- Un link al video. Puedes subirlo a YouTube o a un Drive, pero debes asegurarte que al recibir el link podamos ver el video.

Importante. El enunciado es bastante abierto en algunas partes a propósito. Si tienes que tomar alguna decisión hazlo con confianza mientras la justificación técnica sea razonable.

Fechas. La fecha de entrega de la tarea es el 19 de abril, a las 20:00 hrs.