Prueba Corta # 5 y #6

Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Bases de datos II (IC 4302)
Primer Semestre 2023
Fecha de entrega: **02/05/23 antes de las 11:59 pm**
Forma de entrega: **Email al profesor siguiendo los lineamientos del programa de curso, adjuntando documento y link al repositorio.**
Formato: **Markdown**
Nombre Archivo: **pc56.md**

Gerald Núñez Chavarría - 2021023226

1. **Explique en que consiste un clustered index y cuál es la diferencia entre este y un índice non-clustered que utiliza INCLUDE para agregar columnas al índice. (25 pts)**

   Los índices agrupados (clustered index) son índices que permiten almacenar los datos de una tabla de acuerdo con un orden especifico. Los datos solo pueden estar ordenados de una forma, es decir, solo puede haber un índice agrupado por tabla. Normalmente el índice se agrupa por la primary key, pero se pueden crear índices personalizados y agruparlos en base a otras columnas de interés.

   La diferencia entre un índice agrupado y un índice no agrupado (non-clustered) con columnas incluidas (INCLUDE) es que el primero ordena físicamente las filas de la tabla según los valores de las claves del índice, mientras que el segundo no lo hace, pero incluye algunas columnas adicionales en el índice para mejorar el rendimiento de las consultas.

2. **Explique el concepto de memory footprint y como afecta este la creación de índices.**
   ¿Cuál es la relación entre un memory footprint alto y la paginación a disco? (25 pts)

   Memory footprint hace referencia al uso de memoria que se utiliza para realizar proceso de almacenamiento y procesamiento de datos en la base de datos. La creación de índices afecta ya que los índices ocupan espacio adicional en memoria para almacenar las estructuras de datos necesarias para realizar las consultas de manera eficiente. Si se crean muchos índices en una tabla, el memory footprint de la base de datos aumentará significativamente.

   Cuando se tiene un alto uso de la memoria en el contexto de un “memory footprint” podría relacionarse con la paginación ya que si la paginación a disco se utiliza con frecuencia, puede ralentizar el acceso a los datos y los índices, lo que puede llevar a una disminución en el rendimiento de las consultas.

3. **FASTantic Inc es una empresa especializada en optimización de búsquedas sobre datos, está a sido contratada por la empresa TooSlow para ayudarle a organizar 40 billones de registros, los registros tienen las siguientes columnas:**
   a. **country: este es un código de país**
   b. **city: está es una ciudad en un país específico.**
   c. **date: está es la fecha en que el registro fue agregado a los datos.**
   d. **payload: es un documento JSON que contiene el evento.**

   FASTantic Inc debe optimizar la búsqueda sobre las columnas country, city y date. Explique la mejor forma de organizar los datos para incrementar la velocidad debúsqueda, actualmente se hace un scan sobre todos los datos.

Asuma que no existeuna base de datos mencione estructuras de datos que utilizará.
¿Que tipo de base de datos recomendaría a TooSlow para almacenar sus datos? (50pts)

Lo primero sería que, para optimizar las consultas de esos tres datos se podría realizar un índice compuesto, combinando estas tres columnas con el objetivo de realizar búsquedas más rápidas que utilizando la técnica de scan.

Se podría utilizar estructuras como los árboles B que permiten realizar búsquedas en tiempos logarítmicos, o bien utilizar hash tables que todavía son más veloces y realizan búsquedas en tiempos contantes, sin embargo, depende de la memoria, porqué por ejemplo utilizar hash tables implica más memoria que los árboles B.

Respecto a la base de datos que se podría utilizar, mi opinión se inclina sobre una base de datos No SQL, porqué son ideales para manejar grandes cantidades de datos que se encuentran semiestructurados como es el caso, que se incluye un archivo JSON por ejemplo. Mi recomendación sería Mongo DB.