Prueba Corta #5/6 Bases de datos II

Isaac Araya Solano 2018151703

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)**

Un clustered index es un índice que se almacena dentro del archivo de los datos, es decir, está incrustado en los datos. Esto quiere decir que los datos están ordenados de acuerdo con la clave primaria o la clave definida del índice.

El índice non-clustered funciona en una estructura aparte de la tabla de los datos y los datos no necesariamente están en el mismo orden que el índice.

La diferencia entre un índice clustered y un non-clustered que utiliza INCLUDE radica en la forma en la que se almacenan y acceden los datos. El non-clustered agrega columnas no llave a la estructura del índice haciendo que se pueda hacer consultas más rápidas a través de columnas no llave. Por lo tanto, el índice clustered es más rápido para acceder datos en una tabla, pero no puede contener tantas columnas como el non-clustered, por lo que el non-clustered permite más velocidad con las demás columnas.

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

El memory footprint consiste en la cantidad de memoria necesaria para que una aplicación mantenga los datos y los índices en memoria principal para su acceso rápido. La creación de índices se relaciona con el memory footprint ya que los índices deben ser cargados a memoria principal, lo que hace que el memory footprint suba. Esto hace que deba de tenerse muy en cuenta el memory footprint a la hora de crear índices, ya que, aunque crear muchos índices podrían teóricamente aumentar la velocidad, el memory footprint subiría muchísimo haciendo que se pueda perder rendimiento en lugar de ganarlo. Un alto memory footprint podría hacer que no se puedan cargar los datos a memoria principal y la base tenga que hacer constantemente paginación a disco disminuyendo considerablemente el rendimiento y haciendo que el propósito de los índices de mejorar la velocidad de las consultas no solo no se logre, sino que se consiga el efecto contrario.

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 de búsqueda, actualmente se hace un scan sobre todos los datos. Asuma que no existe una base de datos mencione estructuras de datos que utilizará. ¿Qué tipo de base de datos recomendaría a TooSlow para almacenar sus datos? (50 pts)

Para la empresa TooSlow recomendaría usar una base de datos NoSQL como MongoDB o Cassandra. Estas bases permiten el escalado horizontal y que se realice un almacenamiento distribuido de datos. Además, estas bases permiten la creación de índices en columnas relevantes que funcionan de forma bastante eficiente. Por otro lado, haría un índice para la columna country, city y date. Para el índice de date se usaría una estructura de árbol B ya que este permite una búsqueda eficiente por rango y el árbol B permite tener más de dos hijos por lo que es un árbol con altura baja que lo hace eficiente para búsquedas e inserción de datos y es más eficiente para cantidades de datos grandes. Para los índices de city y country usaría un hash index ya que estas columnas posiblemente tengan un conjunto limitado de valores distintos.