Laboratorio 7 - Índices

Profesores: Claudio Gutiérrez
Matías Toro
Auxiliares: Scarlett Plaza
Daniel Radrigán
Cristian Salazar
Fran Zautzik

La entrega de este laboratorio deberá ser un informe en pdf, el cual está descrito al final de este archivo./

P1. Sea la relacion $R(a, b, c, d)$ cuyo tamaño es de 1 millón de tuplas, y cada página/bloque contiene $B > 2$ tuplas. Las tuplas de R están ordenados de manera aleatoria (pila). El atributo a es llave, cuyos valores son enteros que van del 0 al 999.999 (distribuídos uniformemente). Para cada una de las consultas a continuación, diga el número de I/O (lecturas/escrituras) que se harán en cada uno de los siguientes casos:

- Analizar R sin ningún índice.
- Usar un *B+Tree unclustered* sobre el atributo a. El árbol es de altura h y cada página contiene P punteros ($P > B$).
- Usar un *B+Tree clustered* sobre el atributo a (asuma que el archivo está ordenado por a). El árbol es de altura h y cada página de hoja está ocupada al 60%.
- Usar un *Hash Index unclustered*. Cada página del indice contiene P punteros ($P > B$).
- Usar un *Hash Index clustered*.

Las consultas son:

(a) Encontrar todas las tuplas de R.
(b) Encontrar todas las tuplas de R tal que $a < 50$.
(c) Encontrar todas las tuplas de R tal que $a = 50$.

P2. Se cuenta con varias tablas que son instancias del siguiente esquema abstracto:

- **pelicula**(**nombre**, **anho**,**calificacion**, **votos**)
- **actor**(**nombre**,**genero**)
- **personaje**(**a_nombre**, **p_nombre**,**p_anho**, **personaje**)

En la base de datos hay dos esquemas: uno con datos indexados (`opti`) y otro sin índices (`opt`). En cada esquema está la misma estructura tres veces con los datos para películas con más de 10.000 votos, más de 1.000 votos y más de 100 votos. Note que las tablas de más de 10.000 votos tienen **menos** tuplas que las 1.000 y muchas **menos** que las 100. En esta pregunta, usted medirá el efecto de utilizar índices en los tiempos de consultas, entregando un informe (en PDF) con sus respuestas.

Use el siguiente comando para obtener los planes de consulta y tiempos de ejecución de alguna consulta:

```
EXPLAIN ANALYZE CONSULTA SQL;
```

Por ejemplo:

```
EXPLAIN ANALYZE SELECT * FROM opti.personaje100 WHERE p_nombre='Up' AND p_anho=2009;
```

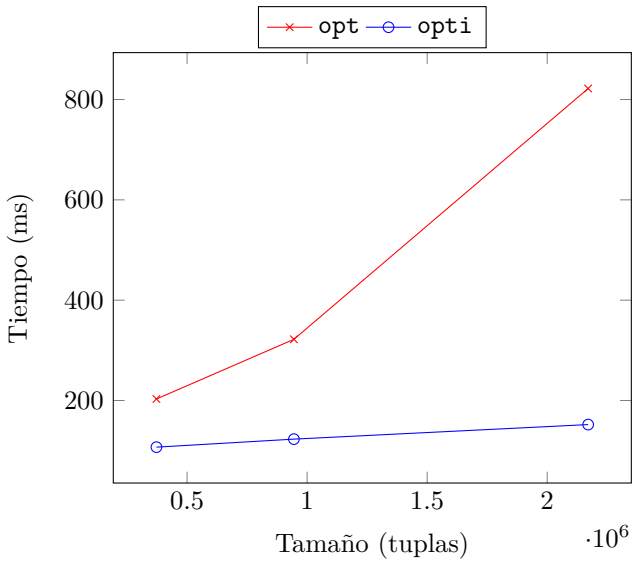

Figure 1: Gráfico de ejemplo: se muestran las curvas para la misma consulta con y sin índices y cómo varía el tiempo de ejecución respecto al tamaño (número de tuplas) de la tabla `personaje*`.

(a) Indique que índices existen para las tres tablas con el prefijo `10000` en ambos esquemas usando los dos comandos `\d+ opt.`TABLA`10000` y `\d+ opti.`TABLA`10000`. Recuerde que Postgres agrega un índice para la llave primaria por defecto, entonces `opt` solo tiene esos índices. (Se configuran los índices de las tablas `1000` y `100` de la misma forma.)

(b) Consulte por las **otras** películas (nombre y año) en las cuales los actores/actrices de PELÍCULA FAVORITA han participado (que la película aparezca en la tabla `pelicula10000`).
Para ello, escriba dos versiones equivalentes de la misma consulta: una versión sin anidación y otra versión con anidación (usando `IN`).

Para las dos consultas resultantes, debe:

i. Ejecutar las consultas en el esquema `opt` usando las tablas terminadas en `100`, `1000` y `10000` usando `EXPLAIN ANALYZE` y registrar los tiempos totales (planificación más ejecución).

ii. Ejecutar las consultas en el esquema `opti` usando las tablas terminadas en `100`, `1000` y `10000` usando `EXPLAIN ANALYZE` y registrar los tiempos totales (planificación más ejecución).

iii. Mostrar gráficamente (con la herramienta que estime conveniente) cómo varía el tiempo tomado (la suma del tiempo de ejecución y el tiempo de planificación) respecto al tamaño de las tablas, tanto en la versión sin índices, como en la versión indexada, como se ejemplifica en la Figura 1. Debe incluir un gráfico por consulta. Considere que el tamaño de la tabla es el número de tuplas en la tabla `personaje*`.

Informe

En su informe debe incluir las respuestas de **P1** junto a su justificación y una descripción de los índices de la **P2**. Finalmente, para cada consulta en la **P2**, debe incluir:

- La consulta misma (*solo* para la escala `100` en `opt`).
- La planificación de la consulta con y sin índices (con `EXPLAIN ANALYZE`, solo para la escala `100`).
- Los gráficos de comparación debidamente detallado (mirar ejemplo de Figura 1).