Universidad de Chile
Departamento de Ciencias de la Computación
CC3201 - Bases de Datos

Laboratorio 7 - Optimización de Consultas

Profesores: Sebastián Ferrada
Claudio Gutiérrez
Auxiliares: Marco Caballero
Florencia Yáñez

Se cuenta con la siguiente estructura:

- **Pelicula**(nombre:string, anho:int, calificacion:float, votos:int)
- **Actor**(nombre:string, genero:char)
- **Personaje**(a_nombre:string, p_nombre:string, p_anho:int, personaje:string)

En la base de datos hay dos esquemas en el servidor del curso: uno con los datos indexados (`lab7_i`) y otro sin índices (`lab7`). En cada esquema está la estructura tres veces: en la primera se encuentran los datos para películas con más de 10.000 votos, en la segunda los datos de películas con más de 1.000 votos y en la tercera para las películas con más de 100 votos.

En el laboratorio, usted deberá medir el efecto de utilizar índices en consultas complejas. Ud. debe entregar un breve reporte (en pdf) con sus respuestas. Al final del enunciado puede revisar lo que se pide agregar en dicho reporte.

P1. [6 Puntos] Usando el esquema `lab7` cuente las tuplas de las tablas presentes. Debe notar que las tablas `10000` tienen **menos** tuplas que las `1000` y muchas **menos** que las `100`. Registre sus resultados. Usando la siguiente consulta, contar cuántos bloques (`relpages`: se llaman páginas en Postgres) hay en cada tabla:
`SELECT DISTINCT relname, relpages FROM pg_class WHERE relname = 'TABLA-NOMBRE';`
Calcule el número promedio de tuplas por bloque para cada tabla. (Observe que estos conteos no cambian en el caso de `lab7_i` entonces no hay que contar las tuplas y los bloques dos veces.)

P2. [Sin Puntos] Compare los índices disponibles para las tablas `10000` en ambos esquemas usando los dos comandos `\d+ lab7.TABLA10000` y `\d+ lab7_i.TABLA10000`. Recuerde que Postgres agrega un índice para la llave primaria por defecto, entonces `lab7` solo tiene esos índices. (Se configuran los índices de las tablas `1000` y `100` de la misma forma.)

P3. [9 Puntos] Use la consulta detallada más adelante con `EXPLAIN ANALYZE` para obtener los planes de consulta y tiempos de ejecución. Ejecútela en los esquemas `lab7` y `lab7_i`.
`EXPLAIN ANALYZE SELECT * FROM ESQUEMA.personaje100 WHERE p_nombre='Pulp Fiction'`
Registre el plan de consulta y el tiempo de ejecución. Calcule y registre la cantidad de consultas por segundo (según el tiempo de ejecución) que pueden realizarse.

P4. [9 Puntos] De la pregunta anterior, calcule y registre una estimación de la cantidad de tuplas leídas y la cantidad de bloques leídos. Para esto considere el plan, el número de tuplas en el resultado, el número de tuplas en la tabla, y el número promedio de tuplas por bloque, en cada uno de los dos esquemas.

P5. [12 Puntos] Consulte por las *otras* películas en las cuales los actores/actrices de `Pulp Fiction` han participado. Escribe dos versiones equivalentes de la misma consulta: una versión sin anidación y otra versión con una consulta anidada usando `IN`. Utilizando `EXPLAIN ANALYZE` obtenga los planes de ambas versiones de la consulta y tiempos de ejecución en ambos esquemas considerando las tablas `100`.

P6. [24 Puntos] Seleccione tres consultas complejas: una que use una consulta por rango (mientras más pequeño el rango, más se beneficia la consulta del indexamiento), una que requiera joins entre (al menos) dos tablas distintas y una que utilice consultas anidadas. Para cada consulta ud. debe:

- Ejecutar las consultas en el esquema `lab7` usando las tablas terminadas en `100`, `1000` y `10000` usando `EXPLAIN ANALYZE` y registre los tiempos.

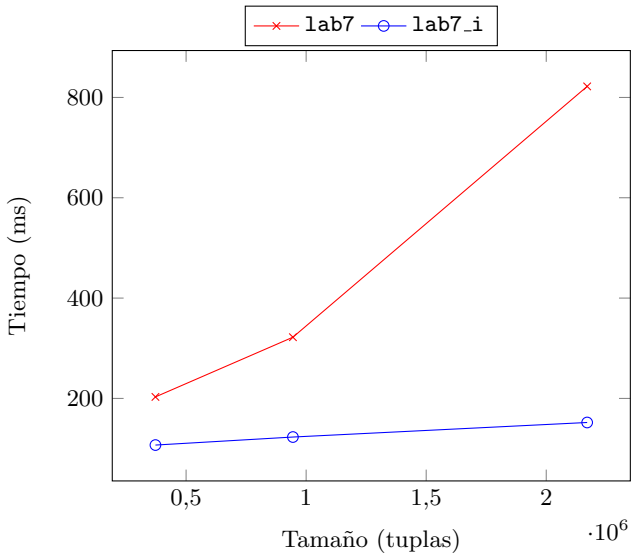

Figura 1: Gráfico de ejemplo: se muestran las curvas para las consultas con y sin índices y cómo varia el tiempo de ejecución respecto al tamaño de las tablas.

- Ejecutar las consultas en el esquema `lab7_i` usando las tablas terminadas en `100`, `1000` y `10000` usando `EXPLAIN ANALYZE` y registre los tiempos. Note que es imprescindible que su consulta **utilice** alguno de los índices **proporcionados** solo por el esquema `lab7_i` (es decir, no los índices por defecto de las llaves primarias), sino deberá seleccionar otra consulta.
- Mostrar gráficamente (con la herramienta que estime conveniente) cómo varía el tiempo de ejecución respecto al tamaño de las tablas, tanto en la versión sin índices, como en la versión indexada. Ejemplo del gráfico esperado en la Fig. 1. El tamaño de la tabla es el número de tuplas en la tabla base; si una consulta usa múltiples tablas puede indicar el tamaño de la tabla más grande.

Reporte

En su reporte entonces debe mostrar los datos recolectados en **P1**, **P3**, **P4** y **P5**. Además, debe mostrar las consultas usadas en la **P5**. Finalmente, para cada consulta en la **P6**, debe mostrar:

- El SQL de la consulta (*sin* duplicar la consulta para diferentes esquemas/escalas).
- La planificación de la consulta con y sin índices (para la escala `100`).
- El gráfico de comparación debidamente detallados (mirar ejemplo de Fig. 1).