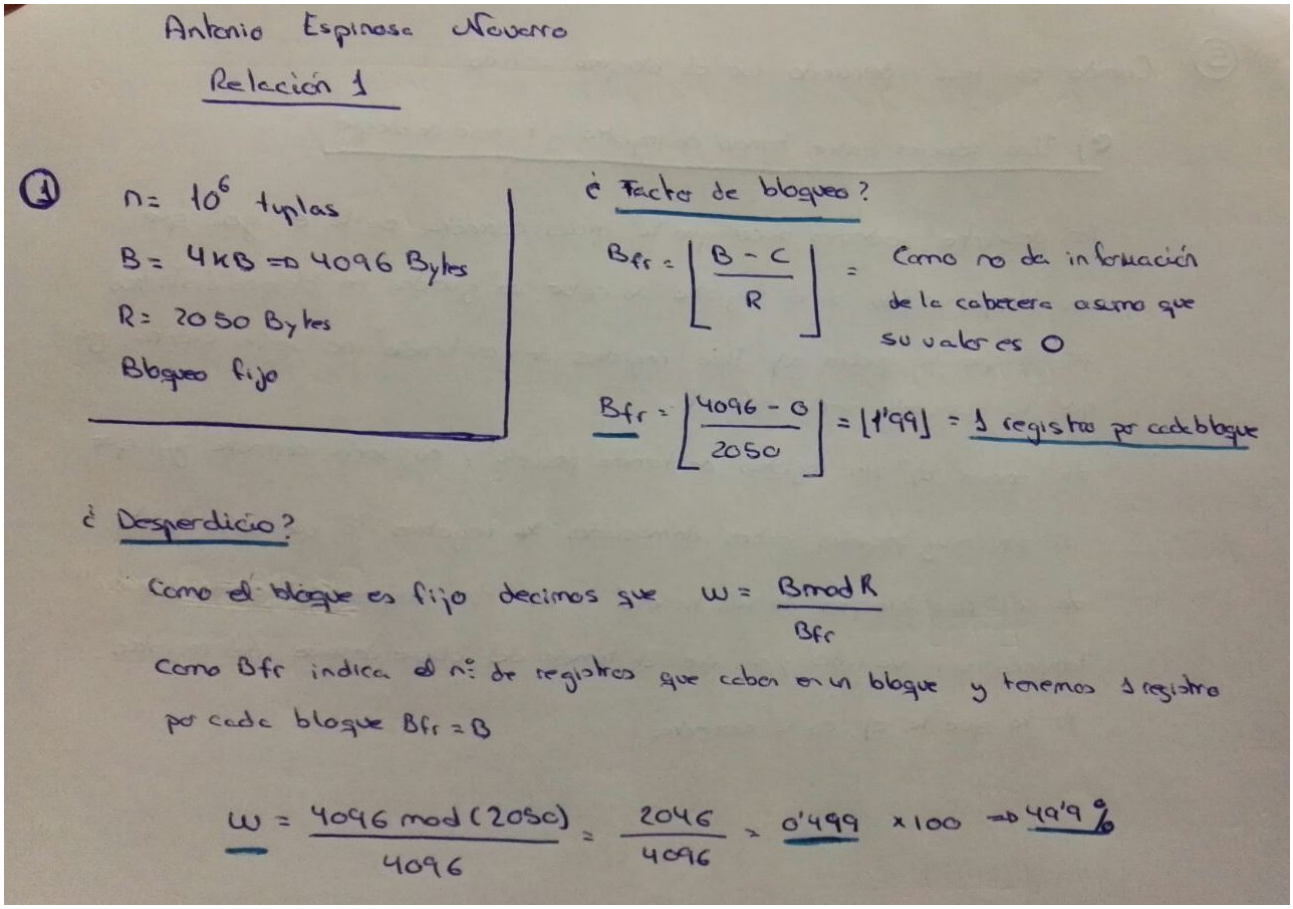
Relación de problemas 1

1. Sea una relación con n=10^6 tuplas, B=4KB, R=2050b y bloqueo fijo. Calcula el factor de bloqueo así como el desperdicio y el porcentaje de utilización de los bloques.

Resolución de Antonio Espinosa Navarro:

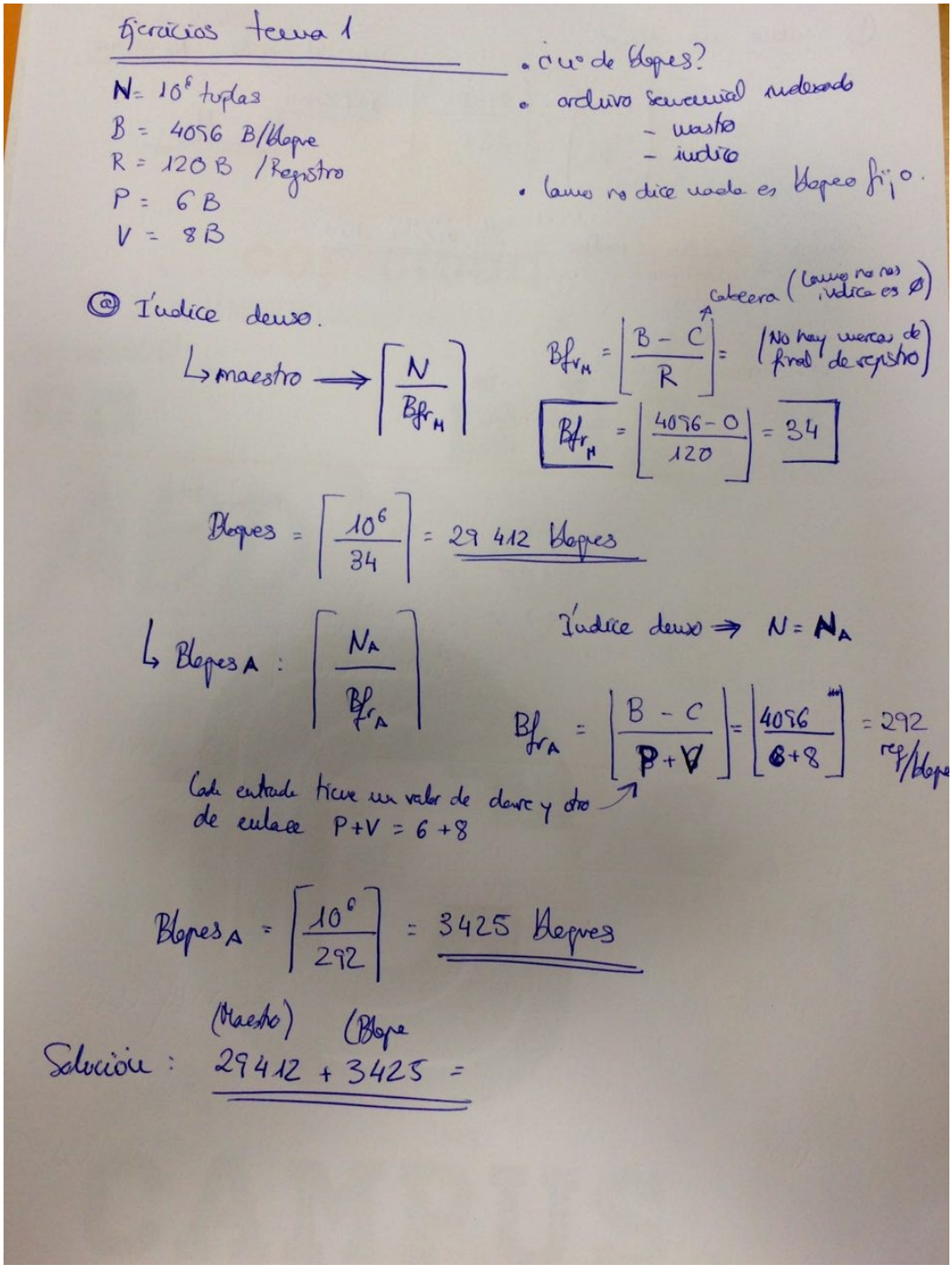


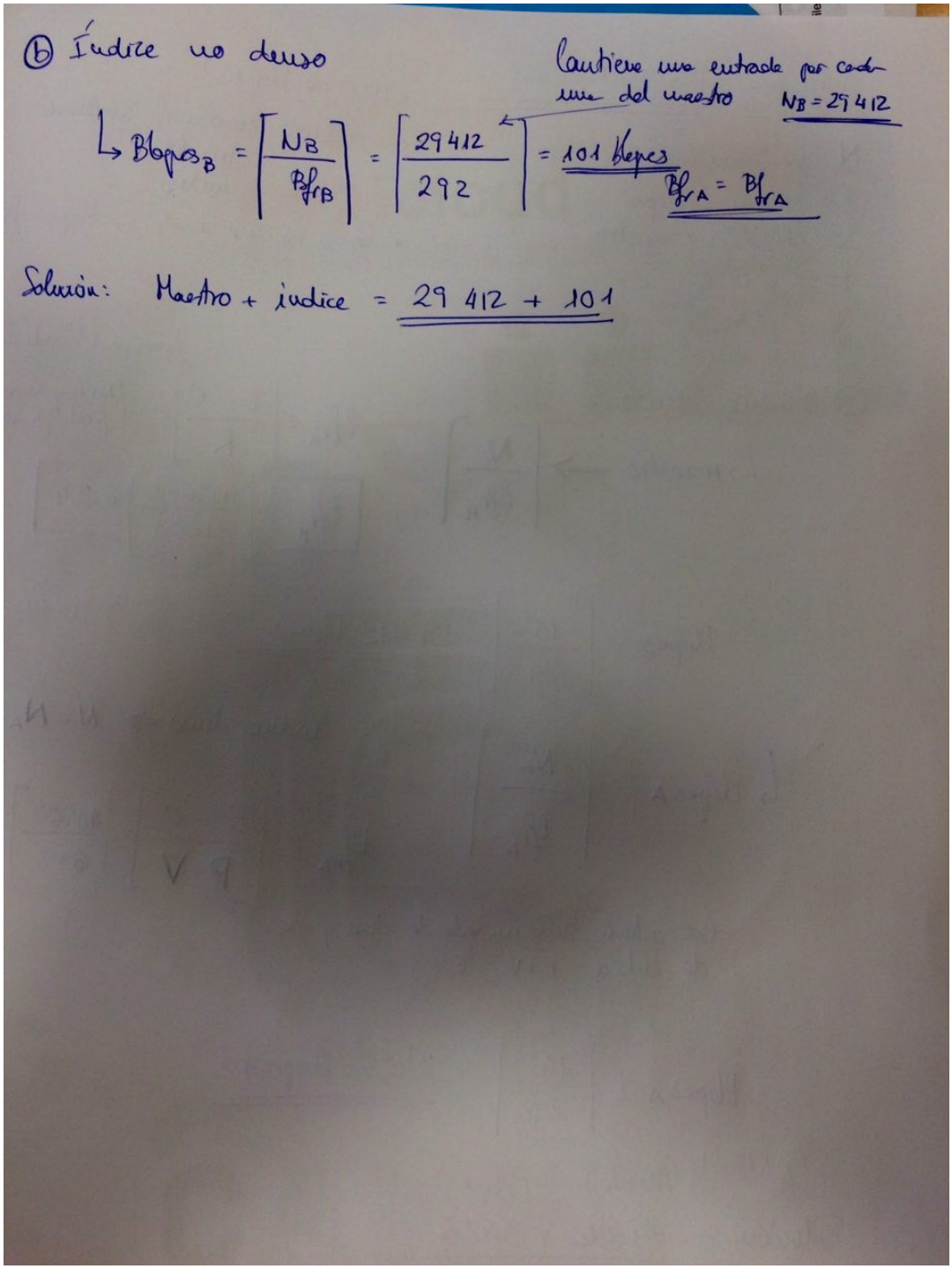
Compleción de la información del ejercicio:

$$Desperdicio = B - (Bfr * R) = 4096B - (1 * 2050B) = 2046B$$

$$\%Desperdicio = \frac{B - (Bfr * R)}{B} = \frac{2046\,B}{4096\,B} = 0{,}49951171875 \approx 49{,}9\%$$

2. Sea una relación con n=10^6 tuplas, B=4KB, R=120b, P=6b y V=8b. Calcula el número de bloques necesarios para almacenar los datos organizados mediante un archivo secuencial indexado en caso de tratarse de:
   a) un índice denso
   b) un índice no denso




⑥ Índice no denso
Contiene una entrada por cada una del maestro
$N_B = 29412$
$\text{Bloques}_B = \left\lceil \frac{N_B}{Bf_{rB}} \right\rceil = \left\lceil \frac{29412}{292} \right\rceil = 101$ bloques
$Bf_{rA} = Bf_{rA}$
Solución: Maestro + índice = 29 412 + 101

3. Indica las ventajas e inconvenientes de tener registros de longitud variable y razona las respuestas.

Resolución parcial de Jose Ángel Díaz García:

Ventajas

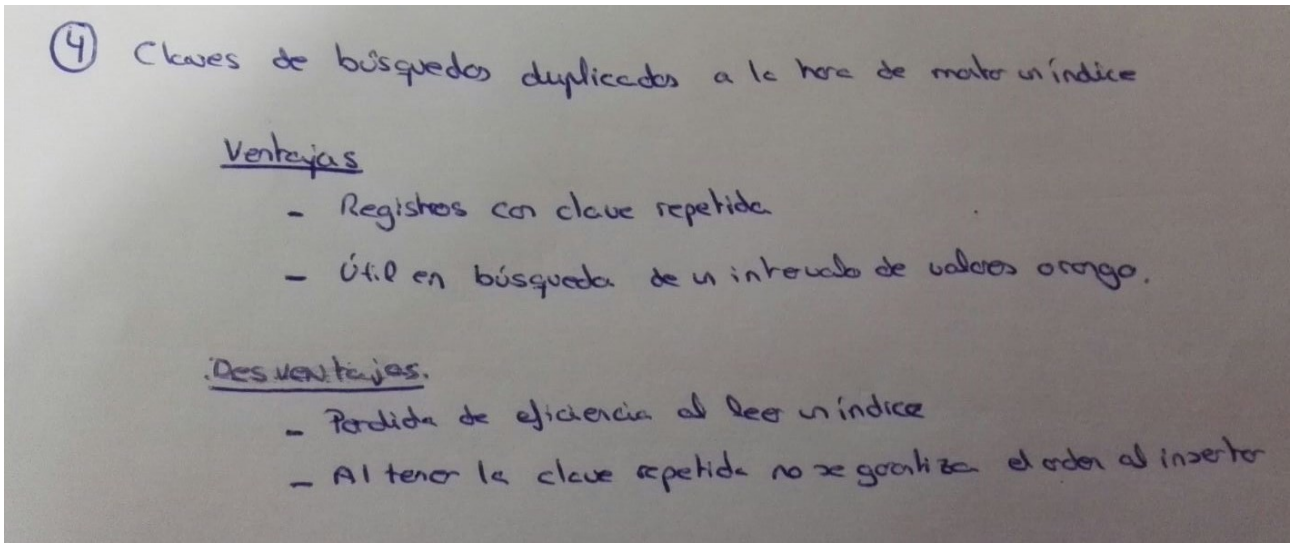
- Permiten almacenar varios tipos de registro en un mismo archivo.
- Permite tener tipos de datos con longitud variable como varchar.
- Permite un mayor aprovechamiento de los bloques ya que se pueden ajustar los registros mejor al tamaño del bloque e incluso dividir el último si no cabe en un bloque.
- **Es la única posibilidad cuando el tamaño del registro excede el tamaño del bloque.**

Inconvenientes

- Se ha de almacenar la estructura de los registros en **las cabeceras de** los bloques **o en cada uno de los registros** con lo que se pierde espacio de almacenamiento.
- Algunos registros quedarán divididos en varios bloques al no poderse introducir un registro en el espacio libre del bloque donde deba insertarse.
- Requiere el uso de marcas (delimitadores de campos) para indicar donde termina un campo y empieza el siguiente para delimitarlos.
- Presenta problemas en la modificación si el registro resultante es de mayor tamaño que el original pudiendo obligar a relocalizar el registro.
- No permiten una ubicación inmediata (búsquedas) dentro de los archivos (y de los bloques que lo componen) de almacenamiento debido a que las longitudes de los registros no son siempre similares.
- Por la misma razón, la extracción de datos no es inmediata.

4. Indica las ventajas y/o los inconvenientes de claves de búsquedas duplicadas a la hora de montar un índice, y razona las respuestas.

Resolución de Antonio Espinosa Navarro:



Compleción de la información del ejercicio:

Ventajas:

- Índice de poco tamaño y ayuda a la búsqueda del primer elemento con dicha clave.
- Permite la búsqueda por varias claves que no sean la clave primaria.
- Va muy bien para búsquedas por rangos, es decir, la búsqueda de un intervalo de valores en un atributo.

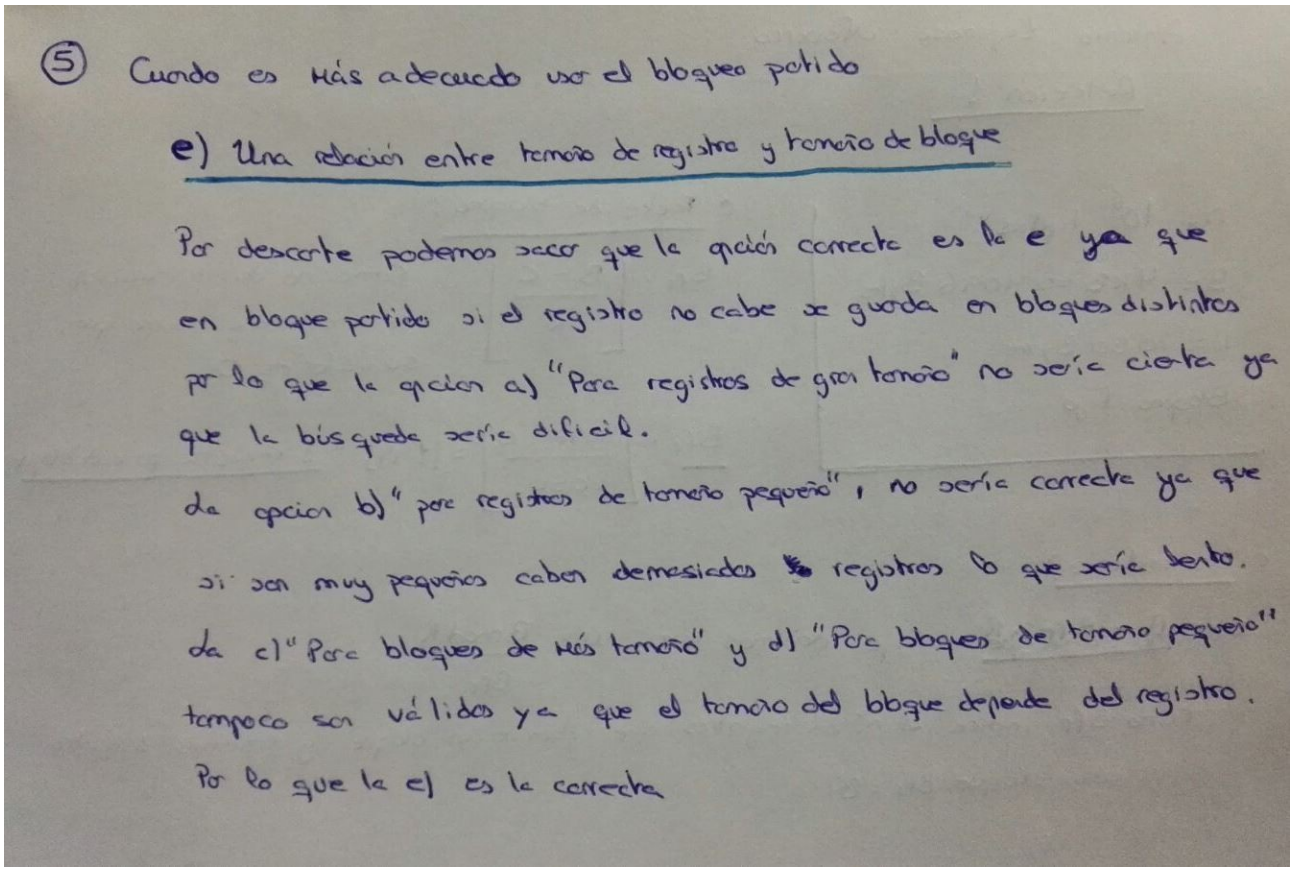
Inconvenientes:

- El primer dato es de acceso directo (acceso mediante el índice) y el resto de los datos del índice se acceden de forma secuencial. Al estar éstos duplicados, se accede al primero de los índices y al resto se van recorriendo de forma secuencial.
- No se garantiza el orden al insertar, al tener tanta clave repetida uno no sabe si ese registro se insertará antes o después de otro registro con la misma clave.

5. Indica cuándo crees que es más adecuado usar el bloqueo partido:
   a) para registros de gran tamaño
   b) para registros de tamaño pequeño
   c) para bloques de más tamaño
   d) para bloques de tamaño pequeño
   e) una relación entre tamaño de registro y tamaño de bloque

---

Resolución de Antonio Espinosa Navarro:



⑤ Cuando es más adecuado usar el bloqueo partido

e) Una relación entre tamaño de registro y tamaño de bloque

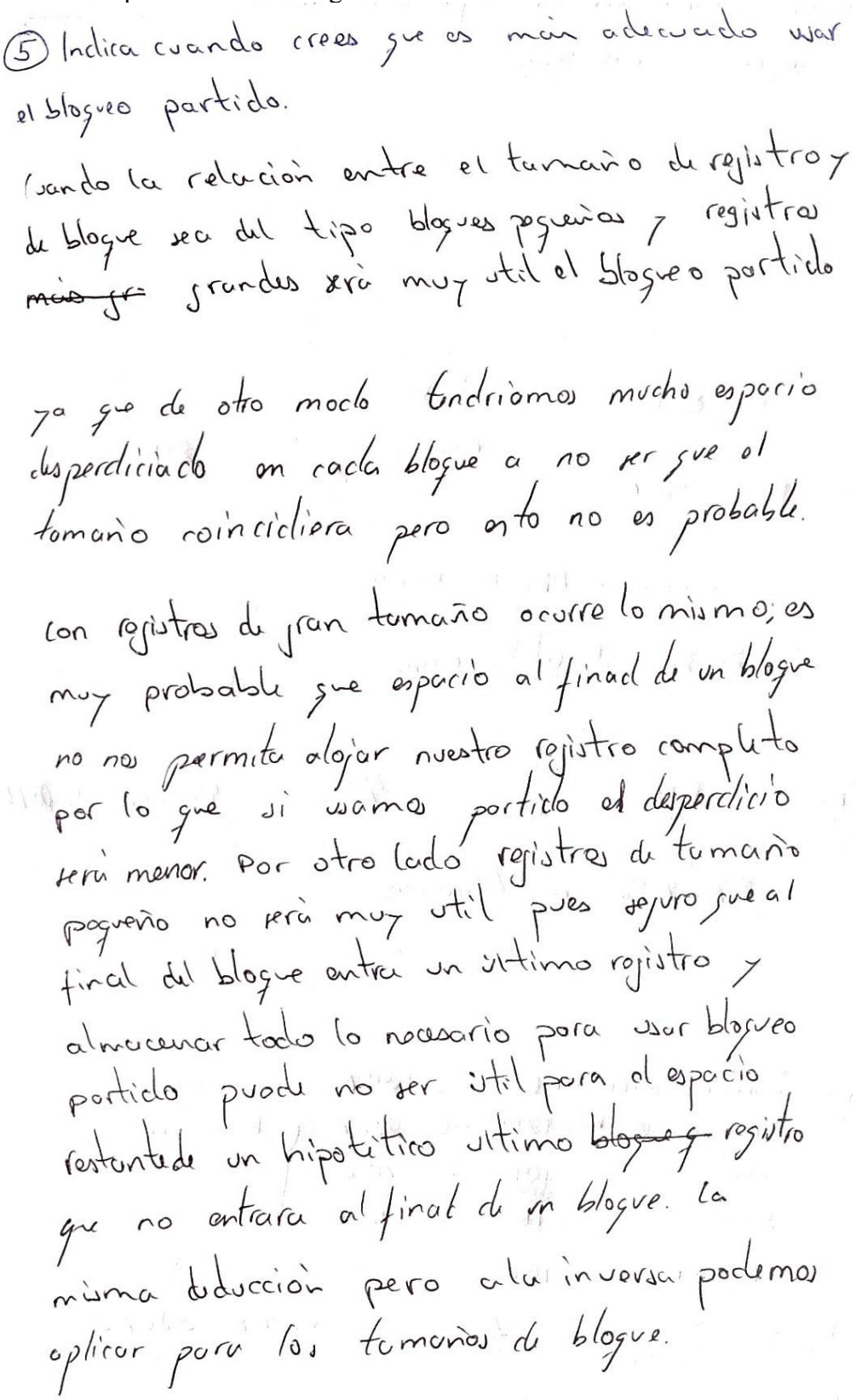
Por descarte podemos sacar que la opción correcta es la e ya que en bloque partido si el registro no cabe se guarda en bloques distintos por lo que la opción a) "Para registros de gran tamaño" no sería cierta ya que la búsqueda sería difícil.

La opción b) "para registros de tamaño pequeño", no sería correcta ya que si son muy pequeños caben demasiados registros lo que sería lento.

La c) "Para bloques de más tamaño" y d) "Para bloques de tamaño pequeño" tampoco son válidos ya que el tamaño del bloque depende del registro.

Por lo que la e) es la correcta

Una resolución más explicada de Jose Ángel Díaz García:



(5) Indica cuando crees que es más adecuado usar el bloqueo partido.

Cuando la relación entre el tamaño de registro y de bloque sea del tipo bloques pequeños y registros ~~más grandes~~ grandes será muy útil el bloqueo partido

ya que de otro modo tendríamos mucho espacio desperdiciado en cada bloque a no ser que el tamaño coincidiera pero esto no es probable.

Con registros de gran tamaño ocurre lo mismo; es muy probable que espacio al final de un bloque no nos permita alojar nuestro registro completo por lo que si usamos partido el desperdicio será menor. Por otro lado registros de tamaño pequeño no será muy útil pues seguro que al final del bloque entra un último registro y almacenar todo lo necesario para usar bloqueo partido puede no ser útil para el espacio restante de un hipotético último ~~bloque~~ registro que no entrara al final de un bloque. La misma deducción pero a la inversa podemos aplicar para los tamaños de bloque.

6. Indica por qué mejoran las consultas mediante los índices:
   a) el número de bloques del índice es menor
   b) las claves están ordenadas por valor de clave en el índice
   c) si son suficientemente pequeños están en memoria

7. Supón una tabla con nombre_paciente, que es un varchar(55) que ocupa 56B, una fecha que ocupa 10B, un peso de tipo real que ocupa 8B, un número_intervenciones que es un entero y ocupa 4B, un número_hijos que es un entero y ocupa 4B, un atributo fumador que es lógico y ocupa 1B, y un R de 83B. Calcula el factor de bloqueo y el porcentaje de utilización en caso de tratarse de bloqueo fijo en los casos de:
   a) bloque de 2KB
   b) bloque de 4KB

Resolución de Jose Ángel Díaz García:

Supon una tabla con nombre_paciente, que es un varchar(55) que ocupa 56B, una fecha que ocupa 10B, un peso de tipo real que ocupa 8B, un número_intervenciones que es un entero y ocupa 4B, un número_hijos que es un entero y ocupa 4B, un atributo fumador que es lógico y ocupa 1B, y un R de 83B. Calcula el factor de bloqueo y el porcentaje de utilización en caso de tratarse de **bloqueo fijo** en los casos de:

a) bloque de 2KB
   R= 83B
   B=2 KB
   Bfr = [(B-C)/R] = 2048/83 = parte entera(24.67) = 24
   % de utilización = R·BFr/B = 83·24/2048 = 97,27%
b) bloque de 4KB
   R= 83B
   B=4 KB
   Bfr = [(B-C)/R] = 4096/83 = parte entera(49.34) = 49
   % de utilización = R·BFr/B = 83·49/4096 = 99,29%

Como puede observarse en esta resolución, se considera que el tipo VARCHAR es de tamaño fijo y se contabilizar para él un tamaño de 56 B.

Sin embargo, podemos considerar también que, al haber un campo de tipo VARCHAR, que es de tipo variable, eso hace que todo el registro sea de tamaño variable, por lo que el cálculo del tamaño de registro (R) que nos proporcionan no es correcto. Rehagamos los cálculos:

$$R = a' \cdot (A + V + 1) = 5 \cdot (0 + \frac{55+10+8+4+4+1}{5} + 1) = 87$$

La justificación de este valor de R se debe a que consideramos un bloque homogéneo (con todos los registros iguales) por lo que carece de sentido incluir los identificadores de campo en cada registro y, por tanto, el tamaño de dichos identificadores en cada registro es de cero. Además, todo el bloque contiene registros iguales por lo que el número medio de campos por registro es el número de campos de un sólo registro (o sea, cinco). Por último, al ser un registro de longitud variable, cada campo termina con un terminador que ocupa un byte.

Considerando este cálculo, podemos decir que el porcentaje de utilización en cada caso es de:

a) $Bfr = parte\,entera(\frac{B-C}{R+M}) = parte\,entera(\frac{2048-0}{87+1}) = 23$

Porcentaje de utilización = $\frac{Bfr \cdot R}{B-C} = \frac{23 \cdot 87}{2048-0} = 0{,}977050781 = 97{,}7\,\%$

b) $Bfr = parte\,entera\left(\frac{B-C}{R+M}\right) = parte\,entera\left(\frac{4096-0}{87+1}\right) = 46$

Porcentaje de utilización = $\frac{Bfr \cdot R}{B-C} = \frac{46 \cdot 87}{4096-0} = 0{,}977050781 = 97{,}7\,\%$

Sin embargo, podría argumentarse que R tiene en cuenta los terminadores de campo, pero es este cálculo no tenemos en cuenta los terminadores de registro. Al considerar que los terminadores de campo son necesarios, tendríamos que tener en cuenta también los de registro, para lo cual, el porcentaje de ocupación de cada tamaño de bloque sería de:

a) Porcentaje de utilización = $\frac{Bfr \cdot (R+1)}{B-C} = \frac{23 \cdot 88}{2048-0} = 0{,}98828125 = 98{,}83\,\%$

b) Porcentaje de utilización = $\frac{Bfr \cdot (R+1)}{B-C} = \frac{46 \cdot 88}{4096-0} = 0{,}98828125 = 98{,}83\,\%$

8. Se tienen registros con un nombre que es un varchar(29), una dirección que es un varchar(255), una fecha que ocupa 10B, un valor para sexo que es un lógico y ocupa 1B, y un tamaño de bloque B=4KB. Calcula el factor de bloqueo y el porcentaje de utilización en caso de tratarse de bloque fijo, si el bloque contiene 10B de cabecera y un directorio de entradas en el bloque.

Resolución del profesor:

Registros:

nombre	varchar(29)	30 Bytes
dirección	varchar(255)	256 Bytes
fecha		10 Bytes
sexo		1 Byte

Calcular la longitud del registro:

La presencia de *varchar* obliga al uso de registros de longitud variable en bloques homogéneos, por lo que el tamaño medio de registro se calcula como:

$$R = a' \cdot (A + V + 1) = 4 \cdot (0 + \frac{29 + 255 + 10 + 1}{4} + 1) = 299$$ **Bytes**

Es necesario tener en cuenta que, en un bloque homogéneo, la estructura del registro se puede almacenar en la cabecera, por lo que no es necesario incluir identificadores de atributos ni separadores identificador-valor en cada uno de los atributos del registro (de ahí, el 0 en el tamaño medio de los identificadores de atributo). Como tal, cada valor irá acompañado de un terminador de valor (de ahí, el 1 que acompaña a cada valor).

B = 4 KB = 4096 Bytes

C = 10 Bytes

El directorio estará formado por el valor del campo del registro que actúa como identificador (en nuestro caso, el campo *nombre*, que ocupa 30B). De modo que, parte del espacio del bloque tendrá que emplearse en almacenar dicho directorio, que debe tener espacio para tantas entradas como registros quepan en el bloque (es decir, del *Bfr*). El cálculo del factor de bloqueo debería tener en cuenta el tamaño usable para registros, que depende del tamaño del bloque, del tamaño de la cabecera, del tamaño del directorio (30B por el número de registros del bloque; o sea, del propio *Bfr*) y del tamaño medio de cada registro, además del terminador de registro (por tratarse de registros de longitud variable).

Entrada de directorio = 30 Bytes => Tamaño del registro “nombre” (Debe ser la clave)

$$Bfr = (\frac{B - C - 30 \cdot Bfr}{R + M}) = (\frac{4096 - 10 - 30 \cdot Bfr}{299 + 1}) = (\frac{4086 - 30 \cdot Bfr}{300})$$

$300 \cdot Bfr = 4086 - 30 \cdot Bfr \Rightarrow 330 \cdot Bfr = 4086 \Rightarrow Bfr = 12{,}3818181818$, luego Bfr=12

Se pueden almacenar 12 registros por bloque.

Porcentaje de utilización = $\frac{Bfr \cdot R}{B - C - Bfr \cdot 30} = \frac{12 \cdot 299}{4096 - 10 - 12 \cdot 30} = 0{,}962962962963 \Rightarrow 96{,}3\,\%$

Teniendo en cuenta que se trata de registros de longitud variable, que implican el uso de

terminadores de campo (incluidos en el cálculo de R) y terminadores de registro (incluidos en el cálculo de *Bfr*), y que dichos registros están dentro de bloques que incluyen una cabecera y un directorio que no puede ser ocupado por registros, el espacio ocupado por estos elementos para el control no pueden considerarse espacio desperdiciado.

El espacio ocupado por los registros depende del factor de bloqueo y el tamaño de cada registro. El espacio disponible para registros es el total de bloque menos el espacio necesario para la cabecera y el espacio necesario para el directorio. El porcentaje de ocupación supone la relación entre el espacio ocupado y el disponible.

Si, además, consideramos que las marcas de terminación de registro ocupan espacio que no está disponible para otro registro, habrán de considerarse, puesto que las marcas de terminación de campo ya están incluidas en R. De este modo, el cálculo de porcentaje usado podría ser:

Porcentaje de utilización = $\frac{Bfr \cdot (R+1)}{B-C-Bfr \cdot 30} = \frac{12 \cdot 300}{4096-10-12 \cdot 30} = 0{,}966183575 \Rightarrow 96{,}62\,\%$

9. Se tienen registros con: char(215), integer -2B-, fecha -10B-, real -8B-, R=235B, B=4KB. Supuesta la estructura de longitud variable, una cabecera con 2 punteros -de 4B- más un carácter, calcula el factor de bloqueo para:
   a) bloqueo fijo
   b) bloqueo encadenado

Resolución parcial de Jose Ángel Díaz García y completada por el profesor:

En primer lugar, habría que notar que ninguno de los campos incluidos en el registro es de longitud variable, por lo que no tiene sentido el uso de tales registros.

Como tal, nos ceñimos al cálculo del factor de bloqueo para registros de longitud fija y bloqueo fijo:

$$Bfr = parte\ entera\left(\frac{B-C}{R}\right) = parte\ entera\left(\frac{4096-9}{235}\right) = parte\ entera\left(17{,}3914893617\right) = 17$$

Sin embargo, para el cálculo del factor de bloqueo para bloqueo partido, hay que tener en cuenta la existencia de un puntero adicional de 4B para referenciar al siguiente bloque:

$$Bfr = parte\ entera\left(\frac{B-C-P}{R}\right) = parte\ entera\left(\frac{4096-9-4}{235}\right) = parte\ entera\left(17{,}3744680851\right) = 17$$

En ambos casos, el factor de bloqueo es de 17.

10. Supongamos una relación con 10^6 tuplas, un tamaño de bloque B=4KB, un factor de bloqueo BFr=10, V=10B y P=8B. Calcula el tiempo de búsqueda T_F en un índice denso primario si estuviese en memoria. Y supuesto que no cabe en memoria y se monta un índice de segundo nivel, calcula el espacio adicional ocupado.

11. Supongamos BFr=30 del fichero de datos, y BFr_I=100 del índice. Sean *n* los registros de datos. Indica cuántos bloques se necesitan para:

    a) un índice denso

    b) un índice no denso

    Supuesta una ocupación del 100%, vuelve a calcular los casos a) y b) suponiendo que los bloques se ocupan inicialmente al 80%.

12. Sea BFr=30 y BFr_I=100, y sea *n* el número de registros. Montar tantos niveles de índices como sea necesario hasta llegar a un índice de un único bloque.

$$\frac{1}{2}\log_{10}\frac{n}{30}=x$$

13. Supón el siguiente esquema: dpto (d#, nombre, extension, dir) y prof (NRP, nombre, categoría, dpto)

    Indique la organización que favorecería más la siguiente consulta:

    ```
    SELECT depto categoria, count(*) FROM prof GROUP BY dpto, categoría;
    ```

    a) un índice denso sobre dpto de prof
    b) un índice no denso sobre dpto de prof
    c) un índice denso sobre dpto y categoría de prof
    d) un hashing sobre dpto y categoría de prof
    e) un índice denso sobre d# de dpto
    f) ninguno

14. Suponiendo que se han definido como claves primarias d# y NRP, revisa la consulta anterior y justifica la respuesta.

15. Sea R(A, B, C, D) el esquema de una relación. A continuación, se presentan tres planes lógicos de una consulta:

    $(a)\sigma_{A=a \wedge B=b}(R)$
    $(b)\sigma_{A=a}(\sigma_{B=b}(R))$
    $(c)\sigma_{B=b}(\sigma_{A=a}(R))$

    Indica cuál de los tres escogería un optimizador para ejecutar y justifica tu respuesta.

Resolución en clase:

Para nosotros (a) y (b),son equivalentes entre si debido a las propiedades de asociatividad del producto, reunión y unión, y por tanto el planificador cogería de los tres planes propuestos, los planes, (a) y (c), o bien los planes (b) y (c) para evaluarlos.

Si no nos da mas información podrían ser válidos cualquiera de ellos, para saber por cual se decidiría finalmente, habría que contemplar más datos de los que se nos suministra en el ejercicio como el número de tuplas, bloques de cada selección, proyección o estadísticas para evaluar la mejor opción. Razonamos sobre ésto:

Por tanto, (a) la descartaríamos en todos los casos, con lo cual el optimizador elegirá entre el plan (b) y el (c) dependiendo principalmente de la variabilidad del número de tuplas en los distintos casos, teniendo:

- para (b): n1 número de tuplas resultado de la selección sobre b, siendo ésta la relación división del numero total de tuplas (n) entre la variabilidad de la relación (R) sobre B.

  $$n1=n/V(R,B)$$

- para (a): n2 número de tuplas resultado de la selección sobre a, siendo ésta relación división del numero total de tuplas (n) entre la variabilidad de la relación (R) sobre A.

  $$n2=n/V(R,A)$$

Se nos pueden dar 2 casos de elección para el planificador:

- PLAN A : sí suponemos n1<n2 entonces V(R,B) > V(R,A), nos quedaremos con la opción (b)
- PLAN B: elegiremos la opción (c) en otro caso, por una razón análoga.

16. Para los planes lógicos del ejercicio anterior y si se dispone de las siguientes estadísticas sobre los atributos A, B y C: T(R)=1000, V(R,A)=75, V(R,B)=20, V(R,C)=80, indica cuál de ellos escogería un optimizador para ejecutar la consulta. Justifica tu respuesta.

Resolución de Daniel López García:

Como hay 75 valores distintos de A y 20 de B, la consulta sobre el atributo A es más restrictiva que la consulta sobre el atributo B. Por tanto, el plan que se lleva a cabo es el (c) que realiza, en primer lugar, la consulta sobre el atributo más restrictivo (con un mayor número de valores).

17. Indica las diferencias entre plan lógico y plan físico.

18. Suponiendo que se tiene la relación prof (NRP, nombre, categoría, dpto) y se tienen dos índices I_1 (sobre NRP) e I_2 (sobre categoría), y dada la consulta:

    ```
    SELECT dpto, categoría, NRP FROM prof WHERE categoría='AS1';
    ```

    Indica:

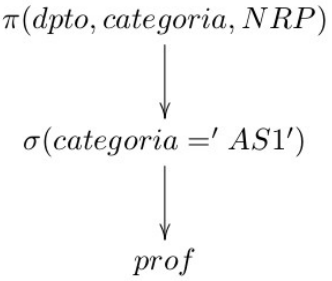
    a) algún plan lógico

    b) un plan de ejecución que podría generar el optimizador de consultas

---

Resolución:

a) Resolución de de Daniel López García:



b)