

FUNDAMENTOS DE BASES DE DATOS

Examen Diciembre 2010 - SOLUCION

La duración del examen es de 3 horas y ½.

Presentar la resolución del examen:

- Con las hojas numeradas y escritas de un solo lado.
- Con la cantidad de hojas entregadas en la primer hoja.
- Con cédula de identidad y nombre en cada hoja.
- Escrita a lápiz y en forma prolija.
- Comenzando cada ejercicio en una nueva hoja

Ejercicio 1 (puntos).

Se desea modelar la realidad correspondiente a la organización de encuentros de motociclistas. Se realizan diferentes encuentros cada año en diferentes lugares, por lo que tanto el año como el lugar son necesarios para identificar a cada uno de los encuentros.

De los motociclistas que asisten a los encuentros se conoce su cédula de identidad y su nombre. Además, cada motociclista pertenece a un grupo ("Los Lamas", "Moderando", "Los Harleys", etc.). Por otro lado, los encuentros son organizados por los grupos, es decir, varios grupos se unen para realizar un encuentro.

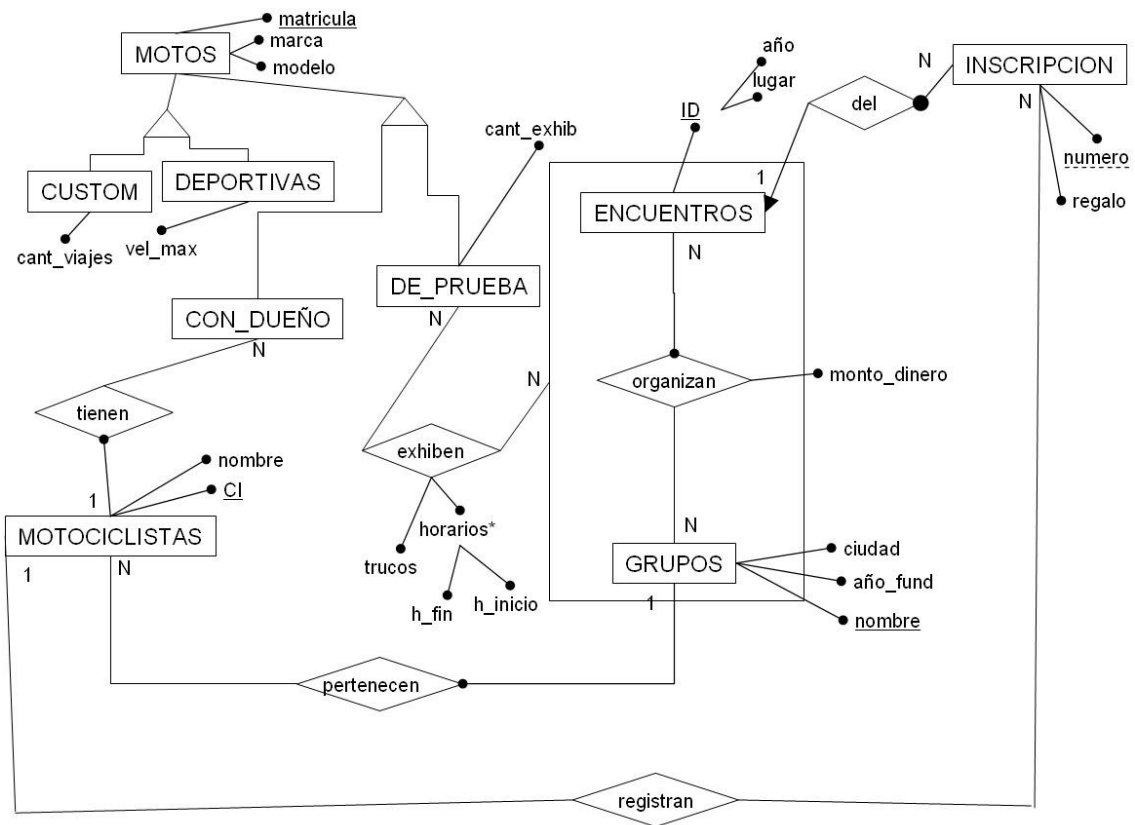
De los grupos se conoce la ciudad a la que pertenecen, el año de su fundación y el nombre que los identifica. Se necesita mantener un registro del dinero necesario para cada encuentro. Por lo tanto, cada vez que se organiza un encuentro, interesa saber el monto de dinero que necesitó cada grupo organizador para realizar dicho encuentro.

De las motos utilizadas por los motociclistas se conoce la matrícula que las identifica, la marca y el modelo. Además se sabe que pueden ser de dos tipos. Las "custom" que son usadas para viajes, por lo que interesa saber la cantidad máxima de kilómetros que puede recorrer cada moto, y las "deportivas", que son especiales para velocidad, por lo que interesa saber la velocidad máxima alcanzada por este tipo de motos. Por otro lado, las motos pueden ser con dueño (uno de los motociclistas) o de prueba (no pertenecen a los motociclistas). De las motos de prueba se guarda la cantidad de exhibiciones en las que fueron usadas.

Un motociclista tiene al menos una moto y las motos tienen un único dueño. Un motociclista se puede inscribir a todos los encuentros que desee. Cada inscripción a un encuentro tiene un regalo asociado (una remera, un cambio de aceite, etc., que lo elige el motociclista) y un número que la identifica en ese encuentro. Si un motociclista pertenece a uno de los grupos organizadores del encuentro, entonces el motociclista no se inscribe a dicho encuentro.

Cada uno de los grupos organizadores del encuentro realiza varias exhibiciones, con una moto de prueba, en dicho encuentro. Una moto de prueba puede ser utilizada por varios grupos en una misma exhibición, a su vez, un grupo utiliza varias motos de pruebas en un mismo encuentro. De la exhibición interesa la hora de comienzo y de fin, y la cantidad de trucos. Se debe tener en cuenta que cada exhibición se repite durante todo el encuentro.

Se pide: Modelo Entidad Relación completo.



RNE:

- Custom \cap Deportivas = \emptyset
- Custom \cup Deportivas = Motos
- Con_Dueño \cap De_Prueba = \emptyset
- Con_Dueño \cup De_Prueba = Motos

- Si un motociclista pertenece al grupo organizador de un encuentro, el motociclista no se inscribe a dicho encuentro:

$(\forall m \in \text{Motociclista}) (\forall g \in \text{Grupos}) (\text{Pertenece}(m, g) \rightarrow$

$\neg((\exists e \in \text{Encuentro}) \wedge (\exists i \in \text{Inscripción}) \wedge \text{Organiza}(g, e) \wedge \text{Registran}(m, i) \wedge \text{Del}(i, e)))$

- Un motociclista no tiene más de una inscripción en un mismo encuentro:

$(\forall m \in \text{Motociclista}) ((\exists i \in \text{Inscripción}) \wedge (\exists e \in \text{Encuentro}) \wedge \text{Registran}(m, i) \wedge \text{Del}(i, e) \rightarrow \neg(\exists i2 \in \text{Inscripción}) \wedge \text{Registran}(m, i2) \wedge \text{Del}(i2, e)))$

Ejercicio 2 (puntos).

Parte a)

Demostrar que toda relación cuyo esquema tiene exactamente dos atributos, se encuentra en Forma Normal de Boyce-Codd.

Parte b)

Sea el esquema $R(A,B,C,D,E,G)$, y el conjunto de Dependencias :

$F = \{ AB \rightarrow C; C \rightarrow A; DE \rightarrow G; G \rightarrow E \}$

- (a) Indicar TODAS las claves candidatas. Justificar.
- (b) En que FN se encuentra el esquema? Justificar la respuesta.
- (c) Suponga el subesquema $R_1(C,D,E,G)$. Contestar las preguntas (a) y (b) para este subesquema.
- (d) Suponga que R satisface las DMs $A \twoheadrightarrow B$ y $B \twoheadrightarrow D$. Si una instancia r de R contiene las tuplas $(0,1,2,3,4,5)$ y $(0,6,7,8,9,10)$, qué otras tuplas debe tener necesariamente r ?

Ejercicio 3 (puntos).

Se mantiene información sobre obras, sus empresas a cargo, trabajadores y maquinaria en una base de datos con el siguiente esquema:

EMPRESAS (NombreEmpresa, Dirección, Categoría)
 OBREROS(NroObrero, NombreObrero, Especialidad)
 OBRAS(CodObra, SupervisorObra, DirecciónObra, FechaComienzo, EmpresaEncargada)
 MAQUINARIA(NroMaquina, Tipo, CodObra)
 TRABAJA (NroObrero, CodObra, Salario)

En este esquema no existen tablas vacías y se cumplen las siguientes dependencias de inclusión.

- $\Pi_{\text{EmpresaEncargada}}(\text{OBRAS}) \subseteq \Pi_{\text{NombreEmpresa}}(\text{EMPRESAS})$
- $\Pi_{\text{CodObra}}(\text{MAQUINARIA}) \subseteq \Pi_{\text{CodObra}}(\text{OBRAS})$
- $\Pi_{\text{CodObra}}(\text{TRABAJA}) \subseteq \Pi_{\text{CodObra}}(\text{OBRAS})$
- $\Pi_{\text{NroObrero}}(\text{TRABAJA}) \subseteq \Pi_{\text{NroObrero}}(\text{OBREROS})$
- $\Pi_{\text{SupervisorObra}}(\text{OBRAS}) \subseteq \Pi_{\text{NroObrero}}(\text{OBREROS})$

Resolver en álgebra relacional las siguientes consultas:

- Nombre y especialidad de los supervisores de las obras que tienen máquinas de todos los tipos conocidos.

$$A = \Pi_{\text{CodObra, tipo}}(\text{MAQUINARIA}) \quad \Pi_{\text{tipo}} \text{MAQUINARIA}$$

$$B = \Pi_{\text{SupervisorObra}}(A * \text{OBRAS})$$

$$\text{SOL} = \Pi_{\$3, \$4} (B \underset{\$1=\$2}{\text{OBREROS}})$$

- Nombre de los obreros que solo trabajan en obras que tienen una única máquina. Para esta consulta suponer que todas las obras tienen por lo menos una máquina.

$$A = \Pi_{\$3} (\text{MAQUINARIA} \underset{\$3=\$6 \wedge \$1 \neq \$4}{\text{MAQUINARIA}})$$

Obras con por lo menos dos máquinas.

$$B = (\Pi_{\text{CodObra}} \text{MAQUINARIA}) - A$$

Obras con 1 máquina

$$C = \Pi_{\$1} (\text{TRABAJA} \underset{\$2=\$4}{B})$$

Obreros que trabajan en obras que tienen 1 máquina.

$$D = \Pi_{\$1} (\text{TRABAJA} \underset{\$2=\$4}{A})$$

Obreros que trabajan en obras que tienen más de 1 máquina.

$$\text{SOL} = \Pi_{\text{NombreObrero}} ((C - D) * \text{OBREROS})$$

Resolver en cálculo relacional las siguientes consultas:

- 3) Nombre de los obreros que son supervisores en alguna obra y trabajan en otras pero no son supervisores en ellas.

```
{ o.NombreObrero / OBREROS(o) ∧
  (∃ ob)(OBRAS(ob) ∧ ob.SupervisorObra = o.NroObrero ) ∧
  (∃ ob2)(OBRAS(ob2) ∧
    (∃ t)(TRABAJA(t) ∧ t.CodObra = ob2.CodObra ∧
      t.NroObrero = o.NroObrero ∧
      ob2.SupervisorObra ≠ o.NroObrero
    )
  )
}
```

- 4) Tipos de máquinas tal que las máquinas de ese tipo solo se encuentran en obras en las que las empresas encargadas son de categoría "AAA".

```
{ m.tipo / MAQUINARIA(m) ∧
  (∃ o)(OBRAS(o) ∧
    (∃ e)(EMPRESAS(e) ∧ e.NombreEmpresa = o.NombreEmpresa
      ∧ e.Categoria = AAA ∧ m.CodObra = o.CodObra
    )
  )
} ∧
¬ ( (∃ e)(EMPRESAS(e) ∧
  (∃ o)(OBRAS(o) ∧
    (∃ m1)(MAQUINARIA(m1) ∧ m1.tipo = m.tipo ∧
      m1.CodObra = o.CodObra ∧
      o.NombreEmpresa = e.NombreEmpresa ∧
      e.Categoria ≠ AAA
    )
  )
)
}
```

Resolver en SQL, sin utilizar vistas ni sub-consultas en el FROM, las siguientes consultas:

- 5) Categoría de las empresas que están encargadas de más de 10 obras en las cuales hay por lo menos una máquina de tipo "hormigonera".

```
SELECT E.Categoria
FROM Empresas E
WHERE E.NombreEmpresa IN
  (SELECT E2.NombreEmpresas
   FROM EMPRESAS E2 NATURAL JOIN OBRAS O
   WHERE O.CodObra IN
     (SELECT M.CodObra
      FROM MAQUINARIA M
      WHERE M.tipo = HORMIGONERA
     )
  )
GROUP BY E.NombreEmpresa
HAVING count(*) > 10
)
```

- 6) Nombre y dirección de las empresas que están encargadas de por lo menos una obra donde se paga el mayor salario de todas las empresas/obras a por lo menos 20 obreros.

```
SELECT E.NombreEmpresas, E.dirección
FROM EMPRESAS E NATURAL JOIN OBRAS O
WHERE O.CodObra IN
      (SELECT O2.CodObra
       FROM OBRAS O2 NATURAL JOIN TRABAJA T
       WHERE T.salario =
              (SELECT max(Salario)
               FROM Trabaja
              )
      )
GROUP BY O2.CodObra
HAVING count(*) >= 20
```

Ejercicio 4 (puntos).

El siguiente es parte de un esquema de gestión de ventas de una empresa:

CLIENTES (idCliente, nomCliente, rubro)

Representa información de los clientes, su identificador, nombre y rubro en el que trabaja.

PRODUCTOS (idProducto, nomProducto, categProducto)

Representa información de los productos, su identificador, nombre y su categoría.

VENTAS (idCliente, categProducto, mes, año, monto)

Registra las ventas, para cada cliente, de determinada categoría de productos en determinado mes y año, así como el monto de la venta.

Considere la siguiente consulta sobre el esquema dado:

```
SELECT P.idProducto, P.nomProducto
FROM PRODUCTOS P, VENTAS V, CLIENTES C
WHERE C.idCliente = V.idCliente AND
      P.categProducto = V.categProducto AND
      V.mes = 'diciembre' AND
      V.monto > 20000 AND
      C.rubro = 'minorista'
```

En este esquema se cumplen las siguientes restricciones de inclusión:

$$\prod_{\text{categProducto}}(\text{VENTAS}) \subseteq \prod_{\text{categProducto}}(\text{PRODUCTOS})$$

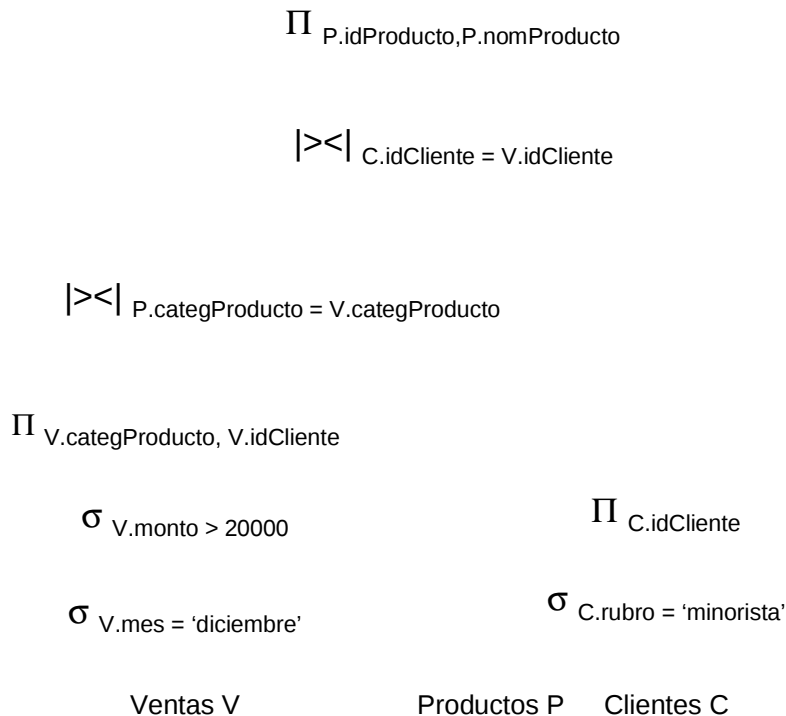
$$\prod_{\text{idCliente}}(\text{VENTAS}) \subseteq \prod_{\text{idCliente}}(\text{CLIENTES})$$

Datos sobre las tablas:

	Cantidad tuplas	Factor de bloqueo	Observaciones
Ventas	30000	6	Solo hay registros para todos los meses del último año. Puede asumirse distribución uniforme en las ventas por mes. De las ventas de diciembre se sabe que un 5% tienen monto mayor a \$20.000 y que estas corresponden a las categorías "vestimenta" y "comestibles".
Productos	150	5	Los productos se distribuyen uniformemente en 5 categorías.
Clientes	100	5	Un 75% de los clientes son minoristas
Productos >< Ventas		4	

Se pide:

- a) Dibuje el árbol canónico para la consulta SQL anterior.
- b) Luego de aplicar heurísticas de optimización vistas en el curso, se obtuvo el siguiente árbol:



- 1- Indique que pasos se siguieron para obtenerlo.
 - 2- ¿Considera que el árbol está completamente optimizado según las heurísticas vistas en el curso? Justifique y en caso contrario explique que debería realizarse y cuál es el árbol resultante.
- c) Dado el árbol presentado en la parte b), ¿qué índices sugeriría y por qué? Para cada uno, decir que tipo de índice sería y sobre que atributo/s se definiría. Asuma que existe un índice primario por la clave primaria de cada tabla.
- d) Dar un plan físico para el árbol presentado en la parte b) que considere los índices propuestos en la parte c).
- e) Calcular los costos de las operaciones de selección y de la operación de join entre *Productos* y *Ventas*, considerando las grabaciones de los resultados de cada una.

Nota para la parte e:

- Para los índices asuma cantidad de niveles = 2.
- Suponga que se cuenta con 3 buffers en memoria.

SOLUCION:

Parte a):

Canónico

$\Pi_{P.idProducto, P.nomProducto}$

$\sigma_{C.idCliente = V.idCliente \wedge P.categProducto = V.categProducto$
 $\wedge V.mes = 'diciembre' \wedge V.monto > 20000 \wedge C.rubro = 'minorista'}$

X

X

Productos P Ventas V Clientes C

Parte b) 1 :

Para obtener el árbol de la parte b se aplicaron las siguientes heurísticas vistas en el curso

H1: Cambiar las selecciones conjuntivas por una “**cascada**” de selecciones simples.

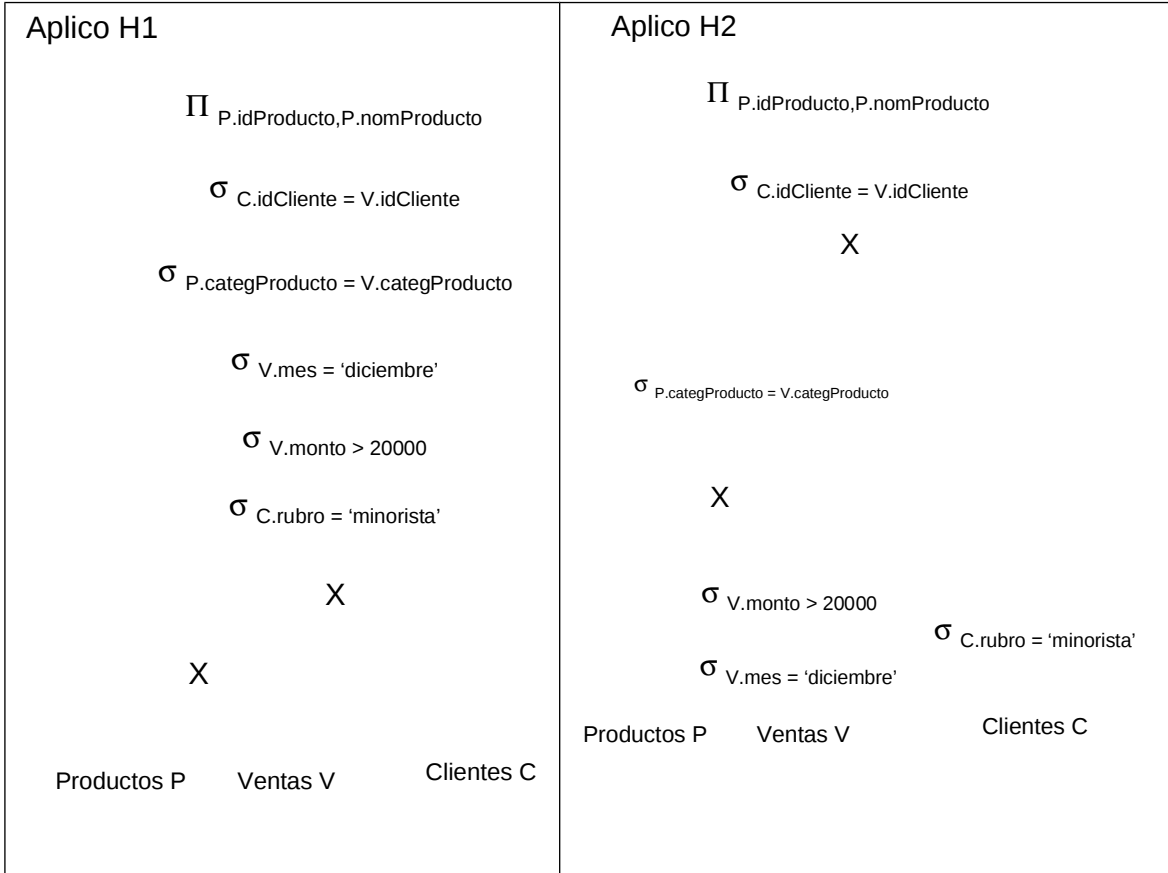
H2: Mover las **selecciones lo más abajo** que se pueda en el árbol.

H3: Poner a la **izquierda** de los productos las hojas que generen menos tuplas.

H4: **Cambiar** secuencias de selecciones y productos por join's.

H5: Mover las **proyecciones lo más abajo** posible en el árbol, agregando las proyecciones que sean necesarias.

A continuación se muestra el resultado de aplicar cada una de ellas, partiendo del árbol canónico solución de la parte a)



Aplico H3

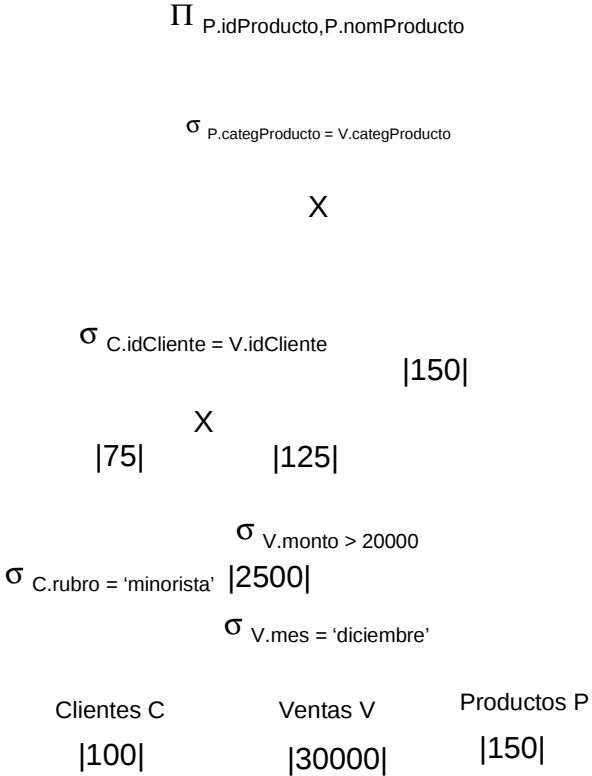
Para aplicar H3 es necesario calcular los tamaños de las hojas luego de aplicar las selecciones:

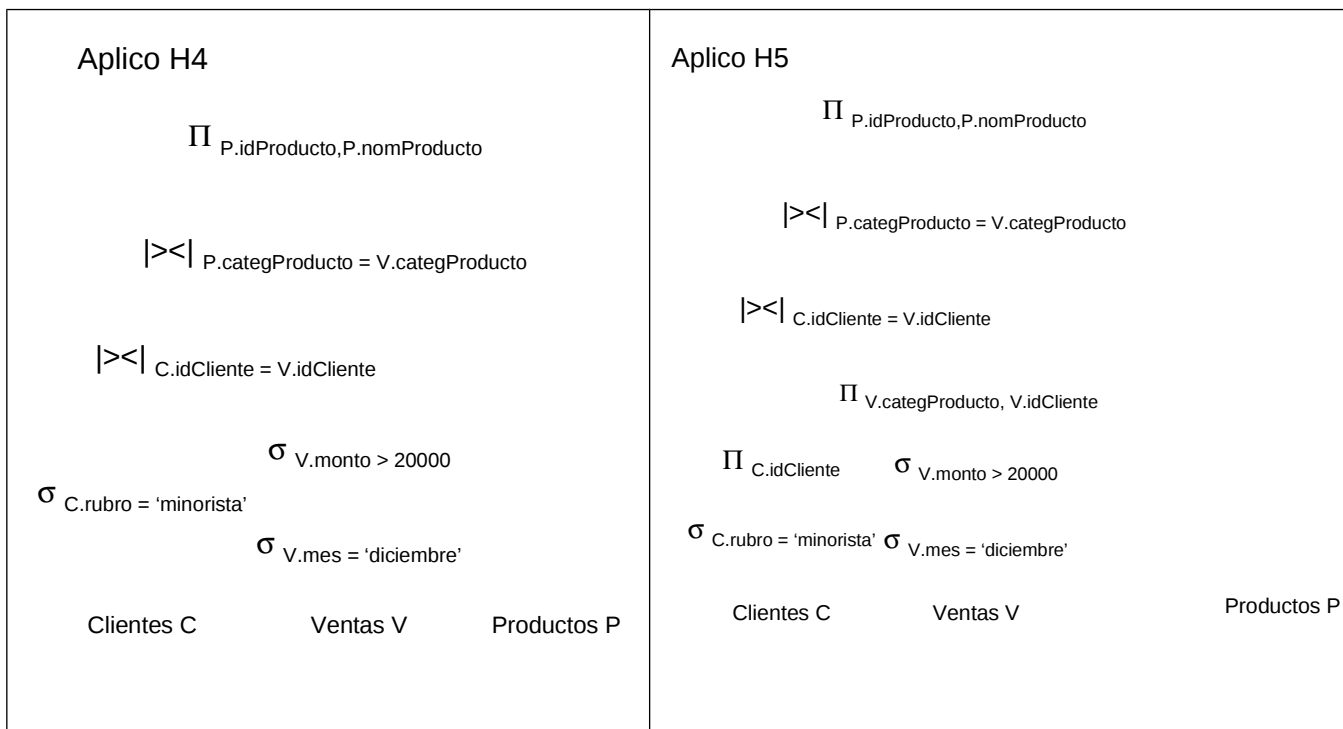
Productos: No tiene selección, por lo tanto su tamaño corresponde a **150 tuplas**.

Ventas: Si hay registros de todos los meses del último años hay 30000/12 tuplas por cada mes. De esas 2500 un %5 tienen monto mayor a \$20.000 o sea **125 tuplas**.

Cientes: Un 75% de los clientes son minoristas, por lo tanto **75 tuplas**.

En orden de tamaño debería poner Clientes, Ventas y Productos. Como el árbol debe estar recostado a la izquierda esto nos lleva a tener que hacer primero el join entre clientes y ventas y luego el join con productos





En conclusión el árbol presentado **no** aplica por completo las heurísticas ya que las hojas no están ordenadas de forma que las mas selectivas estén a la izquierda. Tampoco respeta el orden en que deberían estar en el canónico.

Parte c):

Ventas	Indice Secundario por Mes	Arbol B+
Productos	Indice Secundario por Categoria	Arbol B+
Clientes	Indice Secundario por Rubro	Arbol B+

Parte d):

Plan Físico Loop Join

Index Join

Selección por
Búsqueda Lineal

Selección por Índice
Sec. en Rubro

Selección por Índice
Sec. en Mes

Ventas V

Productos P

Clientes C

Parte e):

	operacion	Costo leer	Costo grabar
T1	σ V.mes = 'diciembre' (VENTAS)	$x + s = 2 + 2500$ (mes)	$s/bf_{ventas} = 2500/6 = 416$
T2	σ V.monto > 20000 (T1)	$ T1 /bf_{ventas} = 2500/6 = 416$	$ T2 /bf_{ventas} = 125/6 = 21$
T3	σ C.rubro = 'minorista' (CLIENTES)	$x + s = 2 + 75 = 77$	$s/bf_{clientes} = 75/5 = 15$
T4	T2 >< PRODUCTOS P.categProducto = T2.categProducto	$b_{T2} + (n_{T2} * Z)$ con $Z = x + s_{PRODUCTOS}$ n_{T2} es la cantidad de tuplas de ventas luego de las selecciones = 125 b_{T2} es la cantidad de bloques para las tuplas de ventas luego de las selecciones = 21 $Z = 21 + (125 * (2 + 15)) = 19521$	$125 * 30 / 4$ [1]

[1] Lo que tenemos que estimar es cuántas tuplas se generan luego del join entre productos y el resultado de T2 (las selecciones sobre ventas)

Del lado de T2 lo que sabemos es que son 125 tuplas que corresponden a 2 categorías de la tabla productos: "vestimenta" y "comestibles" (esta restricción se cumple para las ventas de diciembre con monto menor a 20.000) pero no sé su distribución por categoría.

Del lado de productos tenemos a todas las tuplas (que son 150) y sabemos que éstas se distribuyen uniformemente en 5 categorías, por lo tanto hay $150/5 = 30$ productos por categoría.

Ahora bien, el join entre T2 (ventas) y productos se hace por la categoría del producto, que no es clave en ninguna de las dos tablas. Lo que sabemos es que cada venta de categoría "vestimenta" va a joinear con 30 productos de esa categoría, y que cada venta de categoría "comestibles" va a joinear con 30 productos de esa categoría.

Sea la cantidad de ventas con cat comestibles V_c y la cantidad de ventas con cat vestimenta V_v , entonces el tamaño del join es $V_c * 30 + V_v * 30$. No conocemos el valor de V_c y V_v , pero sabemos que $V_c + V_v = 125$, por lo tanto la cantidad de tuplas luego del join es $125 * 30$