Universidad Nacional de La Matanza

Catedra de Base de Datos

Clase Teórica de Costos – Parte 2

Jueves 11/06/2020

Procesamiento de Consultas

Repaso de algunos conceptos:

El objetivo del Procesamiento de Consultas es minimizar el costo que tiene asociado la ejecución de una sentencia SQL.

El Costo sobre el que se trabaja es:
Costo de Acceso a Memoria Secundaria (disco rígido)

Procesamiento de Consultas

Cómo se mide el Costo?

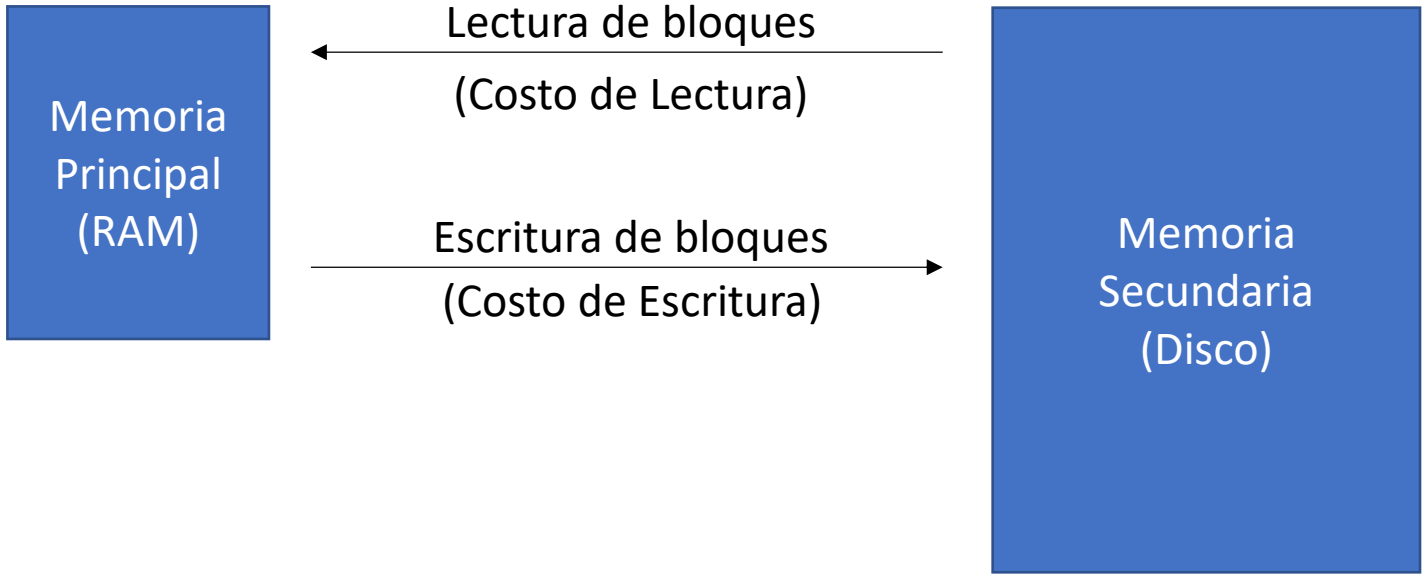
Se mide en cantidad de Accesos a Disco.

Procesamiento de Consultas

Cómo se mide el Costo?

Se mide en cantidad de Accesos a Disco.

Cada vez que se lee o escribe un bloque de datos en el disco, se cuenta como un Acceso.



Procesamiento de Consultas

Etapas del Procesamiento de Consultas:

1) **Análisis léxico:** Identifica los componentes del lenguaje en el texto de la consulta.
   Análisis sintáctico: Revisa la sintaxis de la consulta para determinar si está bien formulada de acuerdo con las reglas sintácticas del lenguaje.
   Análisis semántico: Se validan los nombres de columna y de las tablas/vistas, funciones, etc. Para ver que correspondan con objetos de la base de datos.
2) **Optimización Algebraica:** Se arma el árbol optimizado que determina el orden de las operaciones.
3) **Generación del Plan de Acceso:** Define cómo se va a acceder a los datos (con índice, recorriendo toda la tabla, el método de junta, etc.)

Procesamiento de Consultas

Plan de Acceso

Variables que tiene en cuenta (Metadata)

Tr = Cantidad de tuplas de la instancia de relación r de R.

Br = Cantidad de bloques de la instancia de relación r de R.

B = Tamaño de un bloque de disco (en bytes). Ej: 512 bytes.

FBR = Factor de Bloqueo. Indica la cantidad de tuplas de la relación R que entran en un bloque de disco.

I(A, r) = Imagen del atributo A en la instancia de relación r. Es la cantidad de valores distintos que tiene A en r. Puede ir cambiando según las tuplas que tengamos en cada momento.

CL = Costo de Lectura (en cantidad de accesos)

CE = Costo de Escritura (en cantidad de accesos)

CT = Costo Total (CL + CE)

Mem = Cantidad de bloques disponibles en la Memoria Principal

Procesamiento de Consultas

Tablas empaquetadas

Una tabla se encuentra empaquetada cuando ocupa la menor cantidad de bloques posibles.

Ejemplo:

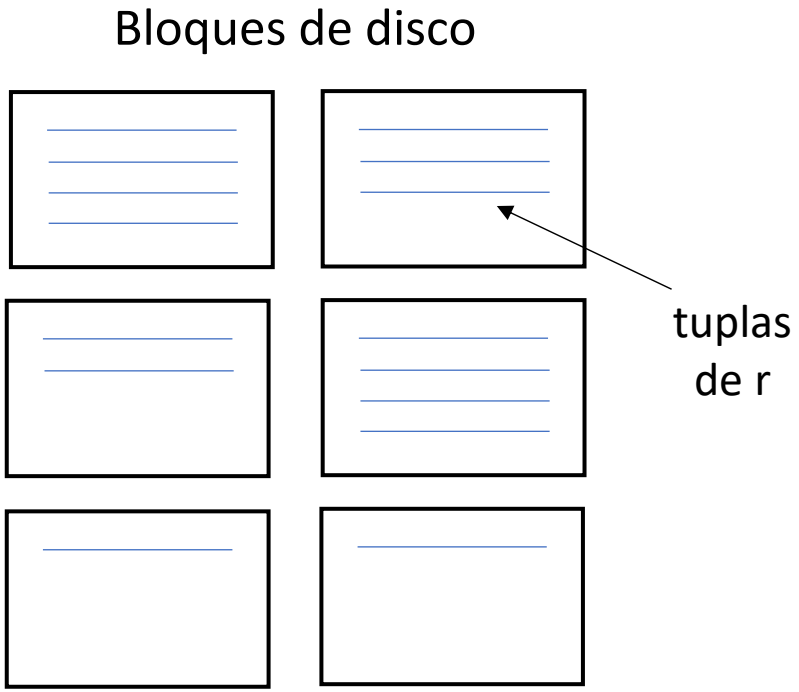
Tabla R

Tr = 15 tuplas

FBR = 4 tuplas por bloque

Br = 6 bloques

No está empaquetada



Procesamiento de Consultas

Tablas empaquetadas

Una tabla se encuentra empaquetada cuando ocupa la menor cantidad de bloques posibles.

Ejemplo:

Tabla R

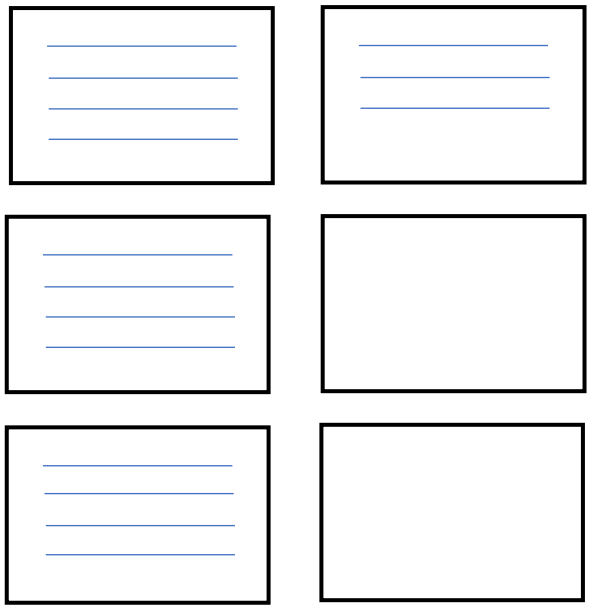
Tr = 15 tuplas

FB_R = 4 tuplas por bloque

Br = 4 bloques

Está empaquetada

Bloques de disco



Procesamiento de Consultas

INDICES

Los índices son estructuras de datos adicionales que pueden crearse sobre una o varias columnas de una tabla.

Esta estructura permite conocer la ubicación de las filas para los distintos valores de la columna sobre la cual se creó el índice.

Procesamiento de Consultas

INDICES

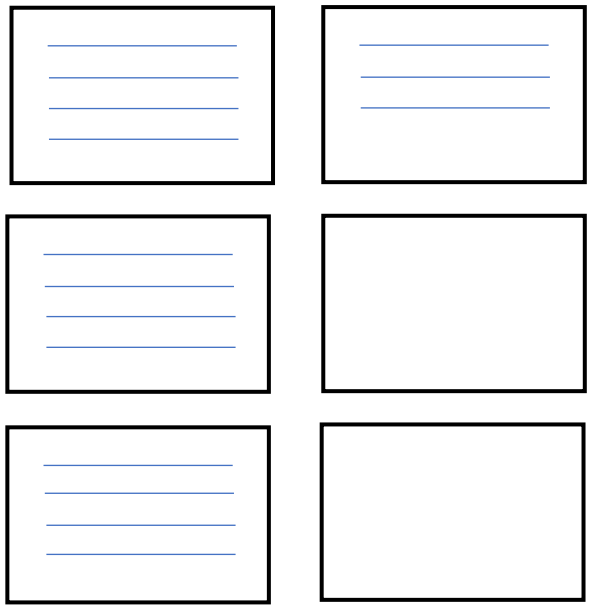
Ejemplo:
EMPLEADO (legajo, nom, ape, categoria, ciudad, sexo)

Supongamos que tenemos una instancia de relación de Empleado con 15 empleados.

I(categoria, empleado) = 5
Los valores de categoría son A, B, C, D y E

Supondremos siempre que los valores tiene una **distribución uniforme**.
Es decir, que tendremos la misma cantidad de tuplas para cada valor.
En este caso, 15 / 5 = 3 filas para cada valor

Bloques de disco



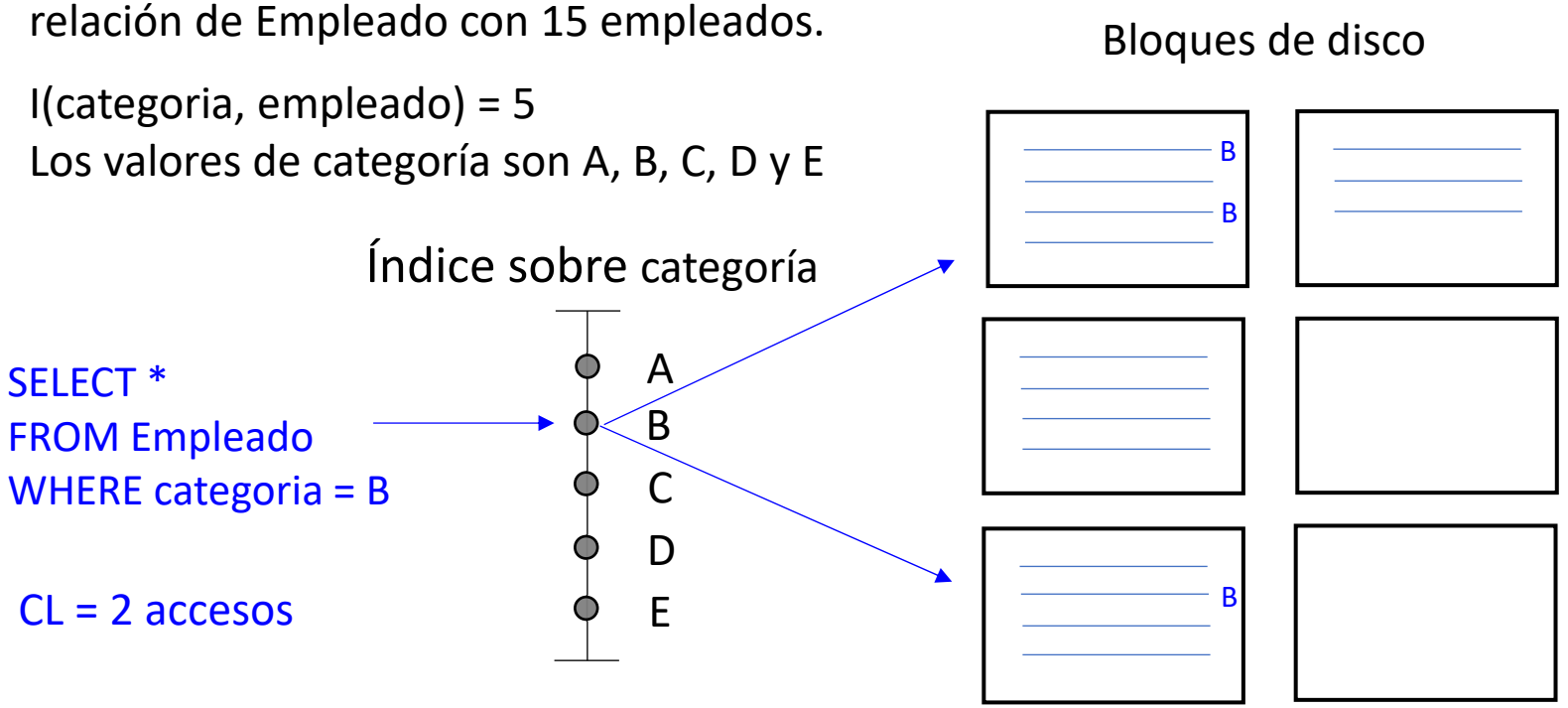
Procesamiento de Consultas

INDICES

Ejemplo:
EMPLEADO (legajo, nom, ape, categoria, ciudad, sexo)

Supongamos que tenemos una instancia de relación de Empleado con 15 empleados.

I(categoria, empleado) = 5
Los valores de categoría son A, B, C, D y E



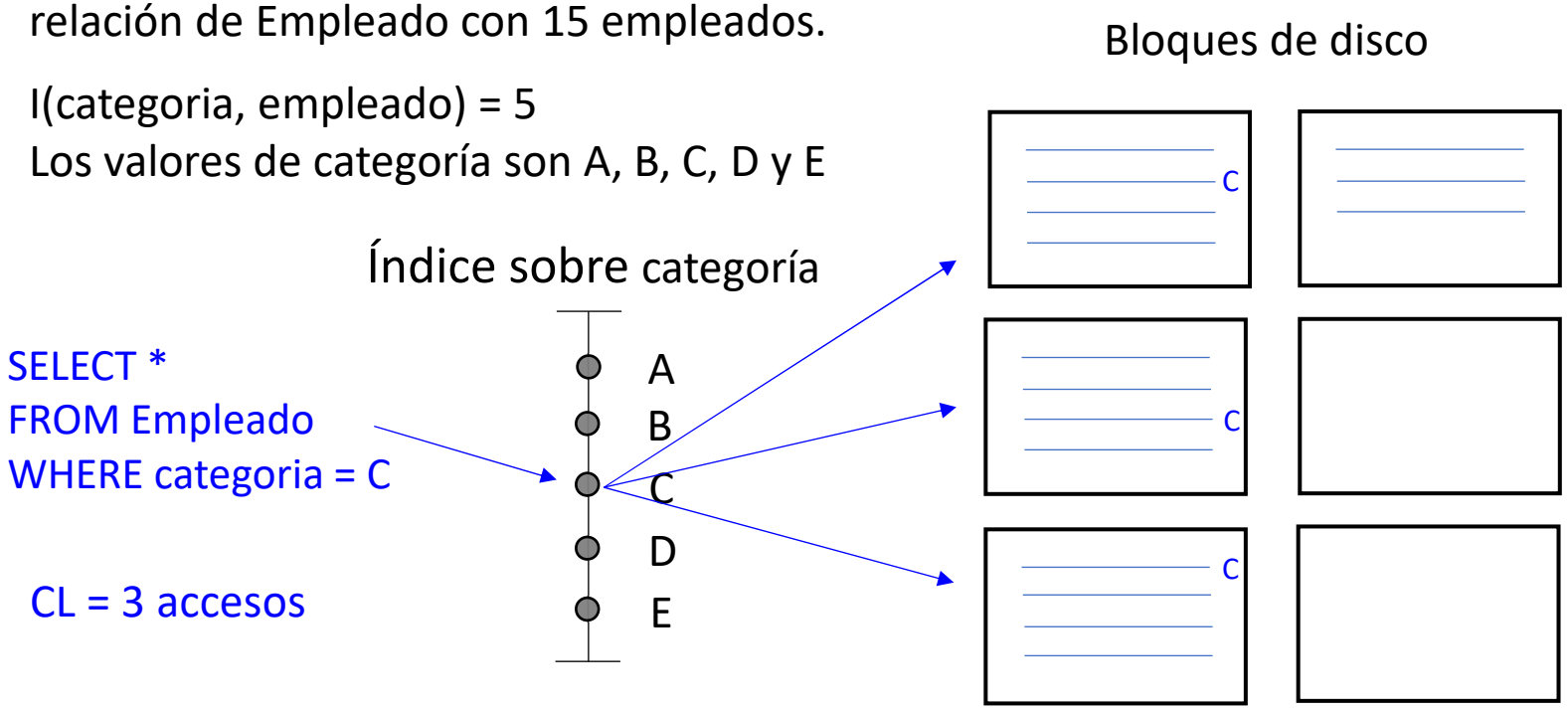
Procesamiento de Consultas

INDICES

Ejemplo:
EMPLEADO (legajo, nom, ape, categoria, ciudad, sexo)

Supongamos que tenemos una instancia de relación de Empleado con 15 empleados.

I(categoria, empleado) = 5
Los valores de categoría son A, B, C, D y E



Procesamiento de Consultas

INDICES

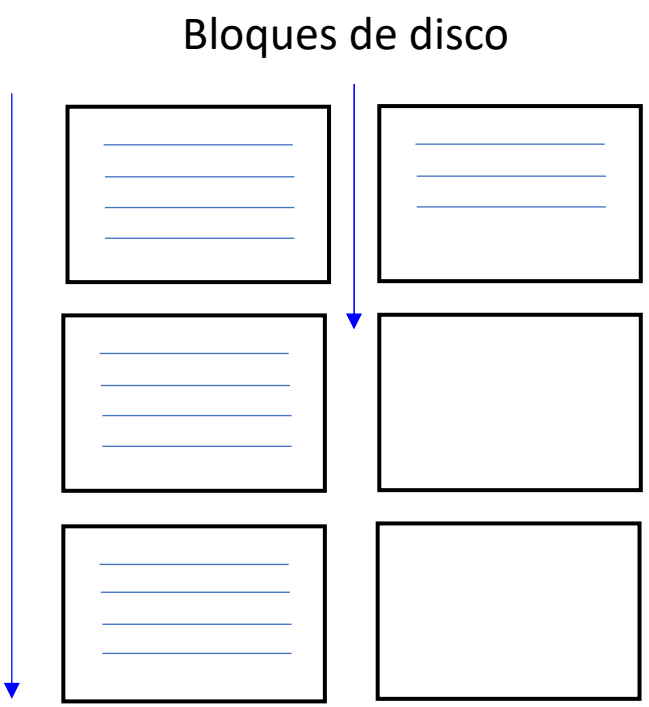
Ejemplo:
EMPLEADO (legajo, nom, ape, categoria, ciudad, sexo)

Supongamos que tenemos una instancia de relación de Empleado con 15 empleados.

I(categoria, empleado) = 5
Los valores de categoría son A, B, C, D y E

```
SELECT *
FROM Empleado
WHERE categoria = C
```

Si no tuviésemos un índice en la columna categoría, tendríamos que recorrer toda la tabla para encontrar las filas que tienen ese valor.
CL = 4 accesos



Procesamiento de Consultas

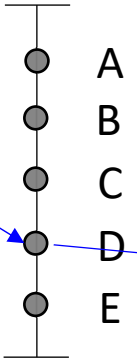
INDICES

Veamos que pasaría si las filas de la tabla estuviesen ordenadas por el valor que buscamos

Supongamos que tenemos una instancia de relación de Empleado con 15 empleados.

I(categoria, empleado) = 5
Los valores de categoría son A, B, C, D y E

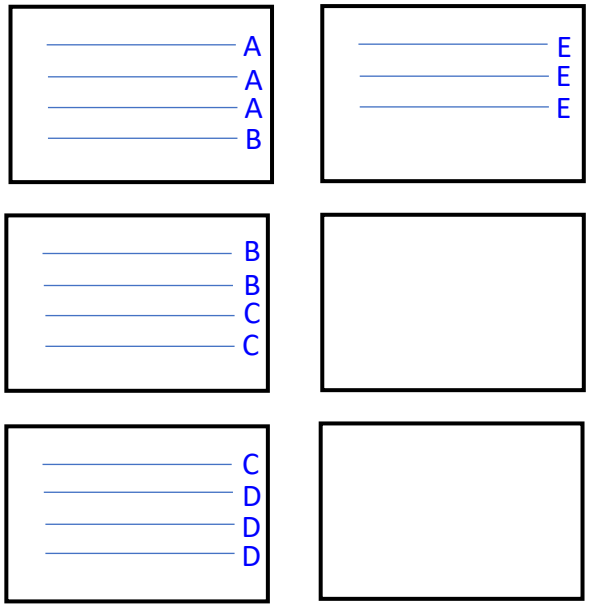
Índice sobre categoría



SELECT *
FROM Empleado
WHERE categoria = D

CL = 1 acceso

Bloques de disco



Al estar ordenado se reduce el costo de lectura

Procesamiento de Consultas

INDICES

Los índices se implementan internamente en forma de arboles B+ (u otras formas), pero no vamos a detenernos en este aspecto ya que puede variar mucho en cada motor.

El objetivo principal de los índices es acelerar los tiempos de respuesta en el procesamiento de consultas, ya que permiten agilizar la búsqueda de las filas que debe dar como resultado una consulta.

El concepto es el muy similar al índice de un libro.

Procesamiento de Consultas

INDICES

VENTAJAS:

- Acelera el tiempo de respuesta de las consultas

DESVENTAJAS:

- Ocupan espacio físico
- Retardan las operaciones de actualización de datos (delete, insert y update) porque se debe regenerar el índice
- Tienen un Costo de Lectura asociado. Es decir, primero hay que leer el índice y luego leer la tabla. Es muy pequeño, pero es un costo adicional que se suma al costo total.

Procesamiento de Consultas

Tipos de Índices

Cluster (o de agrupamiento)
Cuando la tabla se encuentra ordenada físicamente por ese índice.
Solo puede existir 1 índice de este tipo por tabla (generalmente en la PK)

No Cluster (o de no agrupamiento)
Es un índice lógico (no físico).
Podemos tener varios índices de este tipo en una tabla.

Procesamiento de Consultas

Es conveniente crear un índice en todas las columnas de cada tabla?

No, porque como vimos, los índices tienen desventajas

Entonces en qué columnas es conveniente crear índices?

Esto requiere un análisis y es algo que debe decidir el diseñador de la base de datos, el desarrollador o el administrador de la base de datos (DBA).
Por defecto, el motor de base de datos crea automáticamente un índice cluster en la PK de cada tabla.

Procesamiento de Consultas

En términos generales podemos decir que es conveniente crear índices sobre aquellas columnas que:

- Se utilizan frecuentemente en las búsquedas.
- Sus valores no se modifican demasiado, o en todo caso, se consultan mucho mas de lo que se modifican.
- Tienen una imagen muy grande (muchos valores distintos).

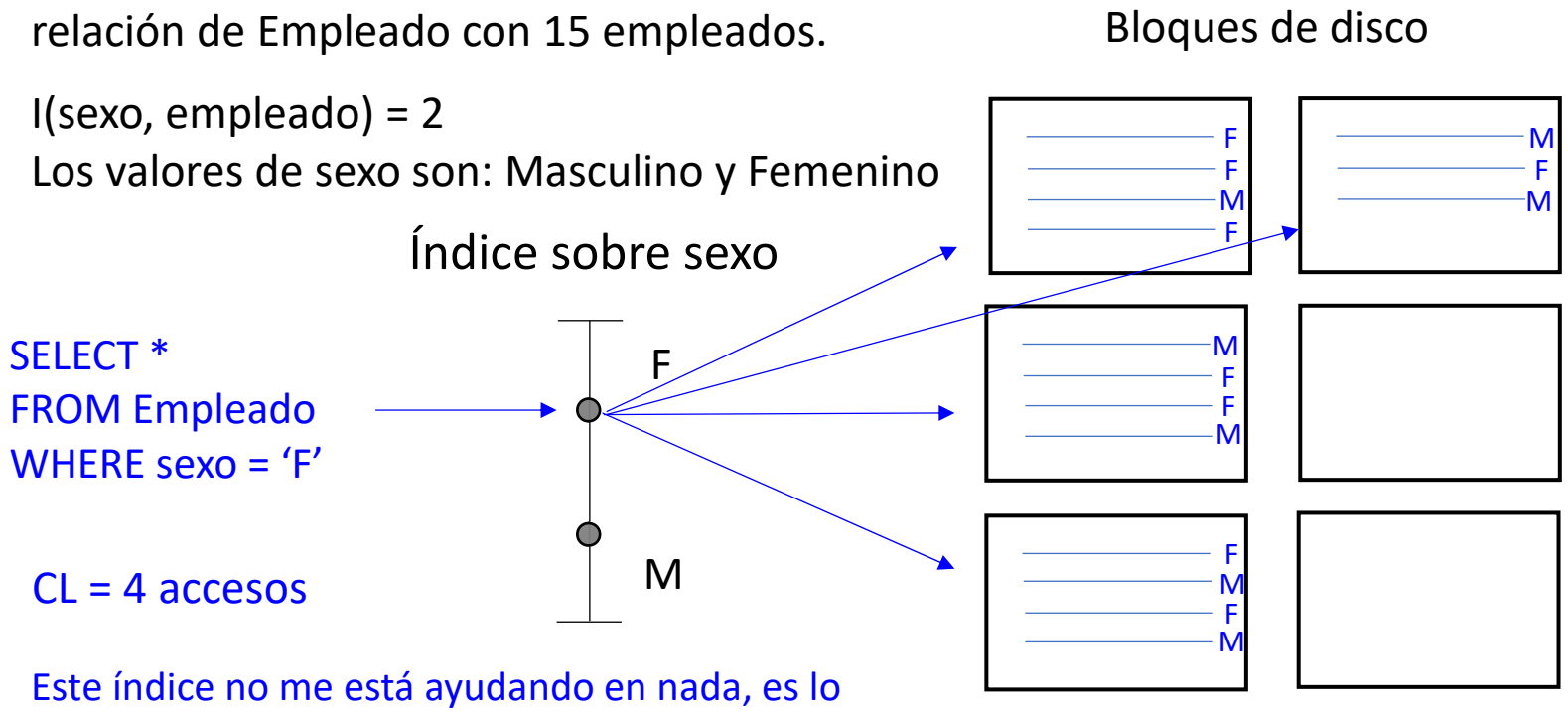
Procesamiento de Consultas

Por qué NO es conveniente crear un índice en una columna con imagen muy pequeña?

EMPLEADO (legajo, nom, ape, categoria, ciudad, sexo)

Supongamos que tenemos una instancia de relación de Empleado con 15 empleados.

I(sexo, empleado) = 2
Los valores de sexo son: Masculino y Femenino



Este índice no me está ayudando en nada, es lo mismo que leer toda la tabla