21103 - Sistemas de Gestão de Bases de Dados
2015-2016
e-fólio A

Resolução e Critérios de Correção

PARA A RESOLUÇÃO DO E-FÓLIO, ACONSELHA-SE QUE LEIA ATENTAMENTE O SEGUINTE:

1) O e-fólio é constituído por 4 perguntas. A cotação global é de 2 valores.

2) O e-fólio deve ser entregue num único ficheiro PDF, não zipado, com fundo branco, com perguntas numeradas e sem necessidade de rodar o texto para o ler. Penalização de 1 a 2 valores.

3) Não são aceites e-fólios manuscritos, i.e. tem penalização de 100%.

4) O nome do ficheiro deve seguir a normal "eFolioA" + <nº estudante> + <nome estudante com o máximo de 3 palavras>

5) Durante a realização do e-fólio, os estudantes devem concentrar-se na resolução do seu trabalho individual, não sendo permitida a colocação de perguntas ao professor ou entre colegas.

6) A interpretação das perguntas também faz parte da sua resolução, se encontrar alguma ambiguidade deve indicar claramente como foi resolvida.

7) A legibilidade, a objectividade e a clareza nas respostas serão valorizadas, pelo que, a falta destas qualidades serão penalizadas.

Vetor Cotações
1 2 3 4 pergunta
5 5 5 5 décimas

Critérios de correção gerais: todas as respostas devem ser justificadas, incluir imagens e exemplos com vista a clarificar os argumentos expostos.

1) Relativo ao Cap.10 - Armazenamento e Estrutura dos Ficheiros
Numa organização de SGBD existem 3 tipos de administradores: DA (data dictionary), DB (database) e DC (data communication). Os administradores DA gerem o dicionário de dados. Qual a informação que deve conter o dicionário de dados de um SGDB? Quais as tarefas dos administradores DA?

Resposta:

Dicionário de Dados:
O Dicionário de Dados ou Catálogo do Sistemas regista os meta-dados i.e. os "dados sobre os dados". Num SGDB o dicionário regista o nome das tabelas, atributos, índices, vistas e utilizadores. Num departamento de informática o dicionário pode ainda incluir os nomes das aplicações, dos programas e das variáveis dos programas, por forma a criar uma linguagem comum a todos os elementos da comunidade.

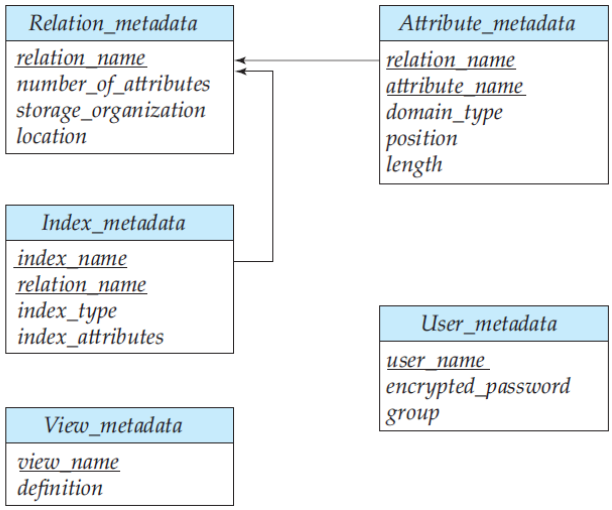


Figura: modelo relacional de Dicionário de Dados

Tarefas dos administrados do Dicionário de Dados (DA):
As tarefas dos administrados do Dicionário de Dados prendem-se com criação e manutenção das regras para dar os nomes às diversas entidades deste as tabelas e os atributos até aos programas e respetivas variáveis.

Exemplo 1 - Tabelas com nomes no plural e indentificadores <'Id'>+<nome_tabela>
- clientes (IdCliente -> nome, data_nascimento, género, contacto)
- restaurantes (IdRestaurante -> nome, morada, cod_postal)
- menus (IdMenu -> nome, preço, tipo)
- transações (IdCliente, data, hora -> IdRestaurante, IdMenu)

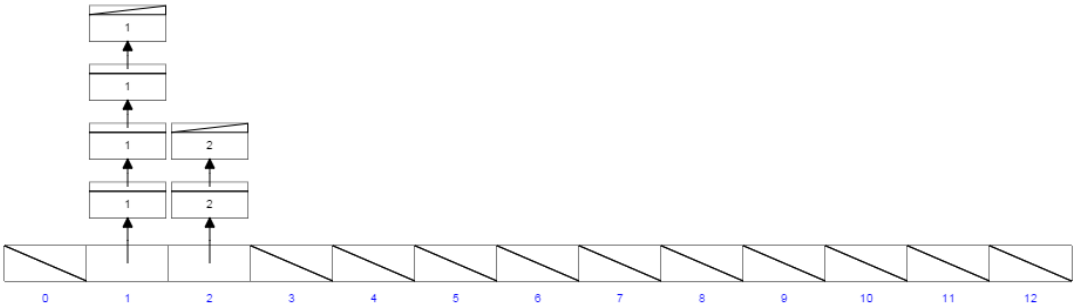
Exemplo 2 - Nome tabela 3 letras, atributos 3 letras, variável 3 letras
- Tabela cliente: TCLI ('T'+<nome>)
- Atributo data_nascimento: ADNA ('A'+<nome>)
- Variável do programa relativa ao cliente: YCLI ('Y'+<nome>)

Critério de correção:
- (0,3) dicionário de dados
- (0,2) tarefas DA

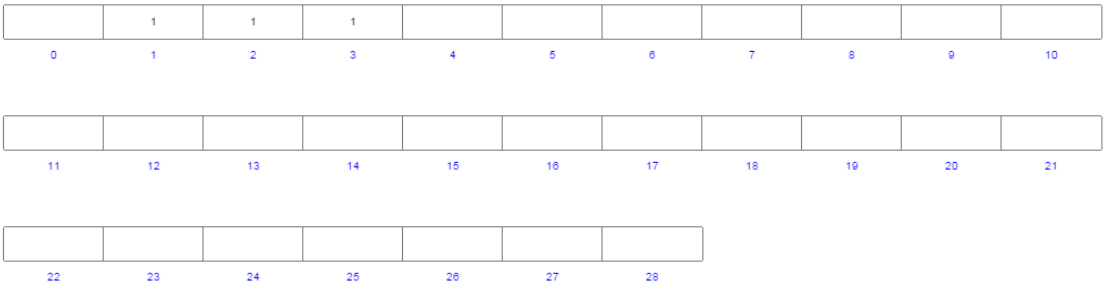
2) Relativo ao Cap. 11 - Indexing and Hashing
Considere os algoritmos os três algoritmos de Hash Tables definidos em https://www.cs.usfca.edu/~galles/visualization/Algorithms.html. Especifique em pseudo-código cada um dos três algoritmos.

Resposta:

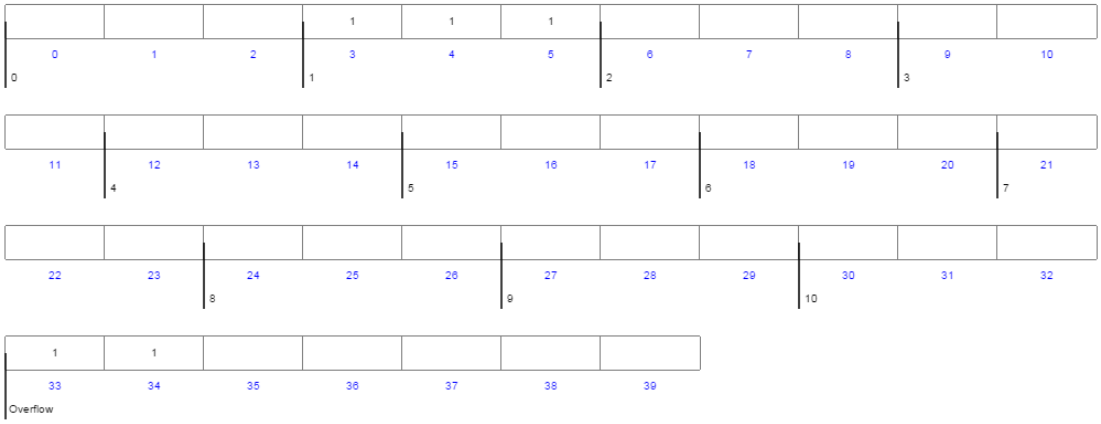
Open_Hashing (int X)
int loc = = X % Table_Size;
if (Table[loc]==Empty) Table[loc]=X;
else append(X, Table[loc]);



Closed_Hashing (int X)
int loc = X % Table_Size;
if (Table[loc]==Empty) Table[loc]=X;
else do
loc=(loc+1)% Table_Size;
while (Table[loc]==Full);
Table[loc]=X;



```
Close_Hashing _with_Buckets (int X)
int initLoc=loc = X % Table_Size;
if (Table[loc]==Empty) Table[loc]=X;
else do
        loc=(loc+1)% Table_Size;
    while (loc<=InitLoc+Buckets and Table[loc]==Full);
    if  (loc>InitLoc+Buckets) Insert_Overflow(X);
    else Table[loc]=X;
```



Critério de correção:

- (0,1) Inserção com Open_Hashing (int X)
- (0,2) Inserção com Closed_Hashing (int X)
- (0,2) Inserção com Close_Hashing _with_Buckets (int X)

3) Relativo ao Cap. 12 - Query Processing; Cap. 6 – Consultas, Feliz Gouveia
Considere os números de blocos B_R=B_S=10.000. Usando o algoritmo de junção de blocos em ciclo, qual o valor dos blocos em memória M, para o qual não seja necessário realizar mais do que 50.000 leituras. Explique detalhadamente o seu raciocínio.

Resposta:

i) No pior caso de apenas houver espaço em memória para guardar um bloco de R e um bloco de S, teremos o custo de:
#Leituras = $B_R + B_R \times B_S$

ii) Assumindo que R (a tabela/relação mais pequena) cabe na memória com M blocos/páginas, e que há um bloco disponível para ler S, obtemos o seguinte algoritmo:

```
Ler todos os blocos b_R de R do disco
Para cada bloco b_S de S
      Ler bloco b_s do disco
      Para cada tuplo s de b_s
            Para cada tuplo r de todos os b_r
                  Se os tuplos (r, s) respeitam a condição:
                        (r, s) são adicionados ao resultado
```

iii) Dividimos a memória disponível a seguinte forma: M-2 blocos para R, 1 bloco para S e 1 bloco para guardar os resultados parciais da junção.

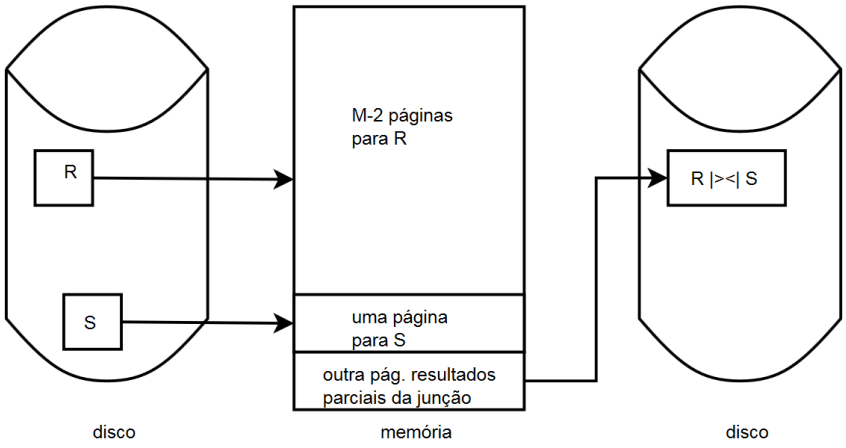


Figura: junção de blocos em ciclo

Como podemos ter simultaneamente M-2 blocos de R em memória, terá de ser lida $\lceil B_R/(M-2) \rceil$ vezes e o custo total será:

#Leituras = $B_R + (\lceil B_R / (M-2) \rceil) \times B_S$

$50.000 = 10.000 + (\lceil 10.000 / (M-2) \rceil) \times 10.000$
logo M=2.502

Assim, com 2.502 blocos em memória, não é necessário realizar mais do que 50.000 leituras.

Para este (iii) caso o algoritmo será:

```
Para cada bloco b_R de R
Ler os R/(M-2) blocos b_R de do disco
     Para cada bloco b_S de S
          Ler 1 bloco b_s do disco
          Para cada tuplo s de b_s
               Para cada tuplo r de todos os b_r
                    Se os tuplos (r, s) respeitam a condição:
                         (r, s) são adicionados ao resultado
```

Critério de correção:
- (0,3) fórmula e resultado
- (0,2) explicação de detalhada

4) Relativo ao Cap. 13 - Query Optimization; Cap. 7 – Otimiz. Consultas, Feliz Gouveia
Na otimização de consultas de um SGDB quais as principais técnicas de estimação de resultados? Quais os tipos de histogramas mais comuns?

Resposta:

A escolha de um "bom" plano é essencial na execução de uma consulta SQL, que tem as seguintes fases: análise sintática -> escolha do plano - > execução.

A otimização do plano de execução baseada em custos tem duas tarefas essenciais:
- estimar a cardinalidade do resultado da aplicação de um operador, i.e. o número tuplos (linhas) do resultado;
- escolher a combinação de operadores (seleção, projeção e junção) de menor custo.

As principais técnicas de estimação de resultados de um operador são: amostragem, técnicas paramétricas e histogramas.
- amostragem: obriga a várias leituras, contudo, fornecem geralmente bons resultados

- técnicas paramétricas: obriga que a distribuição dos dados tenha funções conhecidas, ex: Normal (média, desvio padrão), Poisson (lambda)

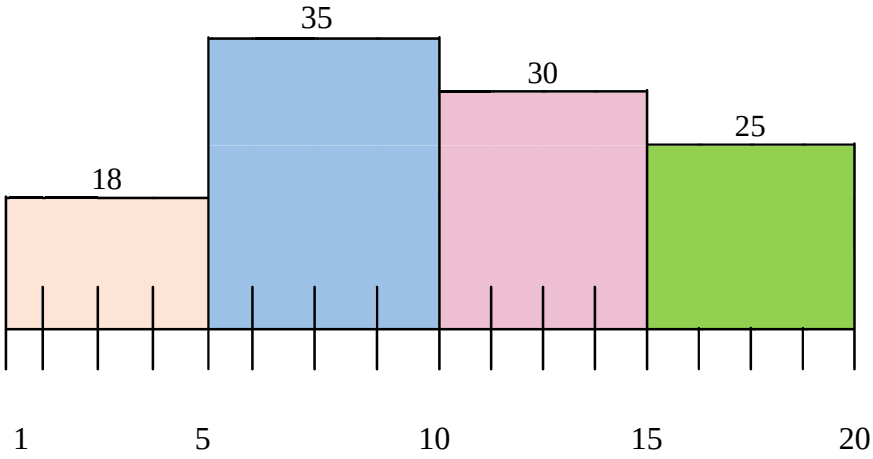
- histogramas: fornece um resumo dos dados com um grau de aproximação passível de configuração

Existem dois tipos de histogramas: equi-largos e equi-profundos.

Histograma Equi-largo – divide o intervalo total de valores em intervalos com igual amplitude.

No exemplo seguinte está a tabela de distribuição de valores de um atributo numérico e o respetivo histograma Equi-largo. Está dividido em 4 blocos e cada bloco dividido em 5 valores. Em cima esta a contagem dos valores de cada bloco.

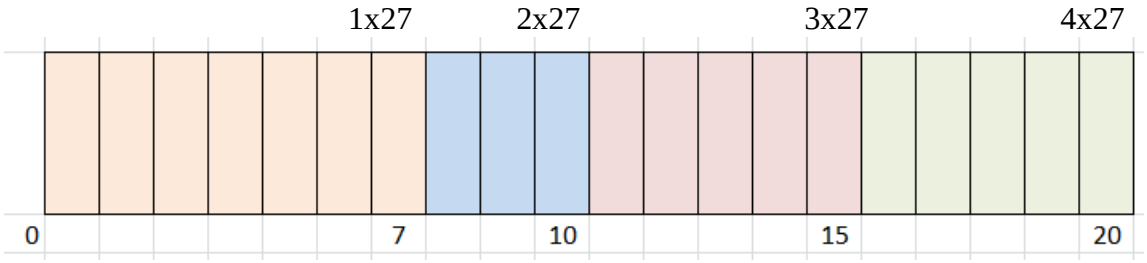
3	5	1	7	2	6	10	3	5	11	7	10	2	6	5	3	5	8	2	7
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*



Histograma Equi-profundo – ajusta os limites dos intervalos para que a todos os intervalos tenham a mesma frequência; neste histograma a frequência é dada pela seguinte expressão: *Frequência = Frequência total / número de blocos*

Usando a mesma tabela do exemplo anterior, a soma dos valores é igual a 108 a dividir por 4 blocos é igual a 27. Assim sendo o 1º bloco tem o limite (1, 7) o 2º bloco (8, 10), o 3ª bloco (11, 15) e o 4º bloco (16, 20).

3	5	1	7	2	6	10	3	5	11	7	10	2	6	5	3	5	8	2	7
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20



Critério de correção:
- (0,3) técnicas
- (0,2) histogramas