UNIVERSIDADE ABERTA
www.uab.pt

Ministério da Educação e Ciência

p-Fólio

Resolução e Critérios de Correção

U.C. 21103

Sistemas de Gestão de Bases de Dados

12 de fevereiro de 2015

INSTRUÇÕES

- O tempo de duração da prova de p-fólio é de 90 minutos.
- O estudante deverá responder à prova na folha de ponto e preencher o cabeçalho e todos os espaços reservados à sua identificação, com letra legível.
- Visto que o enunciado da prova não é utilizado para resposta, poderá ficar na posse do mesmo.
- Verifique no momento da entrega das folhas de ponto se todas as páginas estão rubricadas pelo vigilante. Caso necessite de mais do que uma folha de ponto, deverá numerá-las no canto superior direito.
- Em hipótese alguma serão aceites folhas de ponto dobradas ou danificadas.
- Exclui-se, para efeitos de classificação, toda e qualquer resposta apresentada em folhas de rascunho.
- Os telemóveis deverão ser desligados durante toda a prova e os objectos pessoais deixados em local próprio da sala das provas presenciais.
- O enunciado da prova é constituído por **2** páginas e termina com a palavra **FIM**. Verifique o seu exemplar do enunciado e, caso encontre alguma anomalia, dirija-se ao professor vigilante nos primeiros 15 minutos da mesma, pois qualquer reclamação sobre defeitos de formatação e/ou de impressão que dificultem a leitura não será aceite depois deste período.
- Utilize unicamente tinta azul ou preta.
- O p-fólio é sem consulta. A interpretação das perguntas também faz parte da sua resolução, se encontrar alguma ambiguidade deve indicar claramente como foi resolvida.

A informação da avaliação do estudante está contida no vetor das cotações:

Questão:	1	2	3	4	5	
C:	25	25	25	25	20	décimas

Grupo A – Sistemas de Bases de Dados

1. (2,5 valores) Qual o número de acessos a disco na operação de álgebra relacional de divisão r ÷ s.

(Resposta: 1 página)

Suponha r(A,B) e s(B). No exemplo temos A=pilotos e B=aviões; pretende-se saber se existe algum piloto que voa todos os aviões referidos na relação s:

r	A	B
	Xavier	B52
	Yam	A380
	Yam	B747
	Zito	A380
	Zito	B52
	Zito	B747

÷

s	B
	A380
	B52
	B747

=

r÷s	A
	Zito

a) O algoritmo segue os seguintes passos:
- Começamos por ordenar a relação s por B. De seguida ordenamos a relação r por (A,B).
- Varrendo a relação r considere o atributo A. Em simultâneo, varrendo a relação s verifique se os atributos B são iguais na relação r e s. Se for o caso, o resultado da divisão é o atributo A e continua a pesquisa.

A relação r é varrida uma única vez. A relação s é varrida N vezes, sendo N o número de tuplos distintos de A.

O número total de acessos a disco, depois das tabelas ordenadas, é de |r|+N∗|s|, onde N corresponde ao número de tuplos distintos de A.

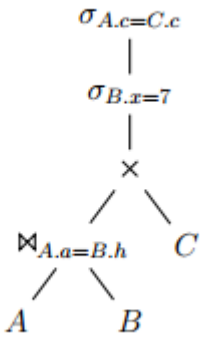
b) Foi ainda considerada correta a resposta obtidas através da implementação possível em SQL, que recorre ao produto cartesiano.

Sabendo que $r \div s = \Pi_{R-S}(r) - \Pi_{R-S}((\Pi_{R-S}(r) \times s) - r)$, sendo r e s quaisquer relações, R e S o conjunto de atributos daquelas relações r e s, e $S \subseteq R$ (Maier, 1983).

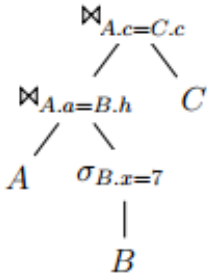
Critério de correção:
- (1,5) explicar a operação de divisão,
- (1,0) valor de r|+N∗|s|,

2) (2,5 valores) Otimize a consulta e apresente a justificação. De seguida, esclareça qual o papel da otimização de consultas na determinação dos planos.

(Resposta: 1 página)



árvore original



Resposta: árvore otimizada

i) A seleção de B.x=7 deve ser realizada na base da árvore. O produto cartesiano associado à seleção A.c=C.c é equivalente a uma junção com a condição A.c=C.c.

ii) o papel da otimização de consultas na determinação dos planos das consultas.

O plano de execução da consulta pretende ser o mais eficiente possível, com vista a consumir o menor número de recursos computacionais.

A otimização da consulta é realizada através de árvores equivalentes, regras de equivalência e heurísticas. As estimativas de custo de cada solução têm em conta as operações e as dimensões das tabelas.

Depois de encontrar uma solução ótima, ou quase-ótima, a consulta é implementada conforme o plano (ou estratégia) previamente definida.

Critério de correção:
- (1,5) árvore ótima: produto cartesiano e B.x=7;
- (1,0) plano da consulta

3. (2,5 valores) O que entende por protocolos “2-phase locking” e “timestamp protocol”? Que aspetos levam à utilização de cada um deles?

(Resposta: 1 página)

No âmbito dos algoritmos para controlo da concorrência existem duas escolas ditas Pessimistas e Otimistas. As pessimistas partem do princípio que vão existir muito conflitos, pelo contrário os otimistas julgam que o conflito vai ser raro. Se vão existir muitos conflitos (pessimista) será melhor utilizar um sistema apertado com várias restrições e assim nasce o algoritmo “two-phase locking” (2PL). Um exemplo de algoritmo otimista (com poucos conflitos) é o algoritmo baseado em “timestamp”.

Num extremo temos os métodos Pessimistas, que acreditam que existem muitos conflitos que geram várias faltas de integridade nas bases de dados, por outro lado temos os métodos Otimistas, que acham pouco provável que um registo seja utilizado em simultâneo, pelo que a falta de integridade é muito rara.

Assim, a família dos métodos 2PL, utiliza "locks" dos registos para leitura e escrita, com a desvantagem de originar "deadlocks". A resposta ao "deadlock" terá de ser resolvido com um "abort"/"rollback" de uma das transações.

Os métodos de Time Stamp, regista o instante de leitura ou escrita dos registos, e no caso de conflito originam o "abort"/"rollback" de uma das transações. De notar que com TS não existem “deadlocks”.

Protocolo	**A favor**	**Contra**
2-phase locking	- transações de leitura (*read only*)	- o *deadlock* deve ser evitado
2-phase locking with multiple granularity locking	- aplicações com mistura de transações leves (que acedem a registos individuais) com transações pesadas (que produzem relatórios)	- o *deadlock* deve ser evitado
Timestamp ordering protocol	- livre de *deadlocks*	- não adaptado para aplicações com mistura de transações leves com pesadas - risco de *roll-back*

Critério de correção:
- (1,50) definições
- (1,00) prós e contras

4. (2,5 valores) Em "Information Retrieval" o que entende por "PageRank"? Qual a forma de calcular esta métrica?

(Resposta: 1 página)

O *PageRank* é o algoritmo que permite calcular o "valor" de uma página na Web. O valor da página não depende apenas da quantidade de *links* apontados para ela, mas do "valor" das páginas que apontam para ela.

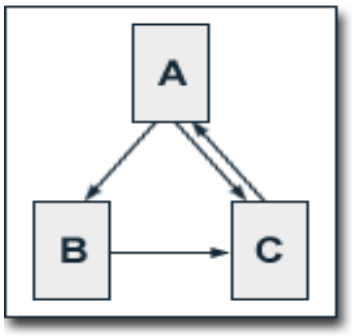
O algoritmo original de PageRank descrito por Lawrence Page and Sergey Brin em 1995 é dado por:

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

onde

- PR(A) é o PageRank da página A,
- PR(Ti) é o PageRank das páginas Ti que estão ligadas (apontam) para a página A,
- C(Ti) é o número de apontadores ("outbound links") na página Ti
- d é o fator de amortecimento que varia em 0 e 1.

Exemplo:



Seja d=0.5,
PR(A) = 0.5 + 0.5 (PR(C) /1)
PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B) /1)

Resolvendo o sistema de 3 equações e 3 incógnitas obtemos os seguintes PR:
PR(A) = 14/13 = 1.07692308
PR(B) = 10/13 = 0.76923077
PR(C) = 15/13 = 1.15384615

Critério de correção:
- (1,00) definição
- (1,50) forma de cálculo

Grupo B – Prática em "Data Warehousing"

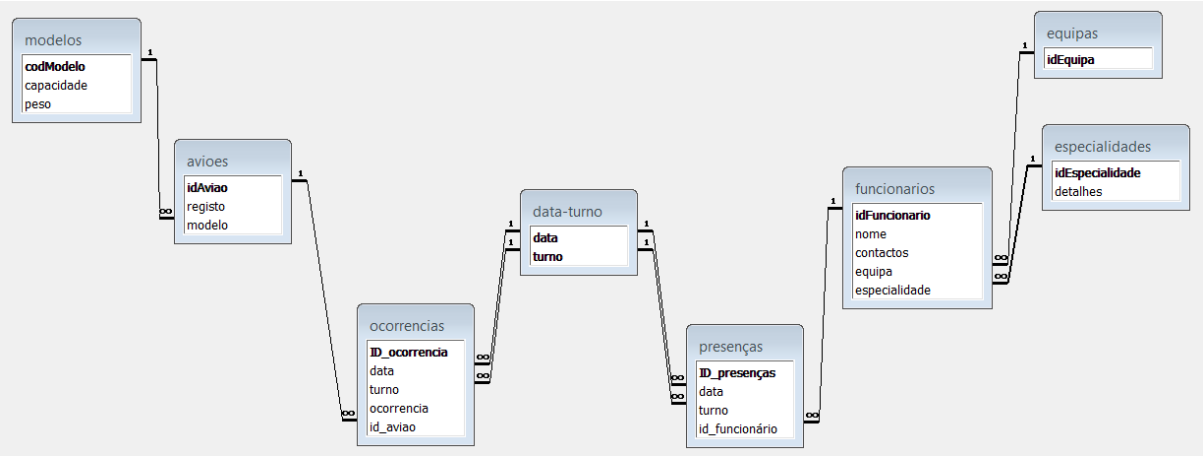
5. (2 valores) Pretendemos desenhar um "Data Warehouse" do seguinte sistema. Defina a tabela de factos em primeiro lugar. De seguida, defina três dimensões para o "Data Warehouse" e apresente a tabela de factos associada às três dimensões.

O aeroporto da Portela resolveu organizar a sua informação num sistema de bases de dados para registar as presenças dos membros das equipas de manutenção das aeronaves.

- Cada avião tem um número de registo e cada avião é de um modelo específico. O aeroporto pode acolher um certo número de modelos de aviões e cada modelo tem um código de modelo (ex. Airbus320, Boeing747), bem como uma capacidade e um peso.
- Cada equipa da manutenção tem um ou mais chefes, vários "aviónicos" (verificação de peças), vários mecânicos, vários técnicos de manutenção (combustível, etc). Cada funcionário da manutenção deve estar registado com nome e contactos.
- As equipas de manutenção trabalham por turnos. É importante registar a presença de cada elemento e as eventuais ocorrências (paragem, falta recursos) em cada turno.

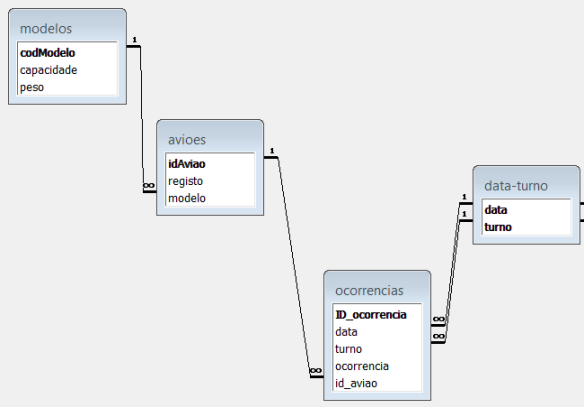
(Resposta: 1 página)

A base de dados correspondente aos requisitos definidos terá o seguinte aspeto, onde se distinguem o registo das ocorrências nos aviões e das presenças dos funcionários.

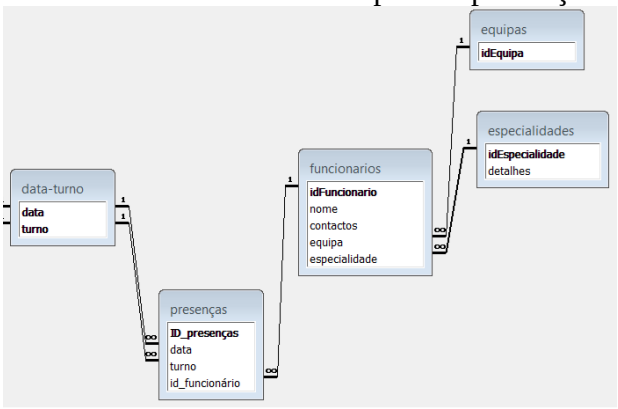


Dado que pretendemos evitar "connection traps", teremos duas tabelas de factos: uma para as ocorrências e outra para as presenças dos funcionários.

Estrutura da tabela de factos para as ocorrências:



Estrutura da tabela de factos para as presenças:



Critérios de correção:
- criar DW com 2 tabelas de factos
- penalização até 50% para esquema mal desenhado
- penalização até 50% atributos desadequados na tabela factos
- penalização até 50% dimensões desadequadas
- penalização até 50% ligações mal estabelecidas

FIM