e-Fólio B

U.C. 21103

Sistemas de Gestão de Bases de Dados

2019-2020

INSTRUÇÕES

1) O e-fólio é constituído por 5 perguntas. A cotação global é de 5 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 10% a 100%.

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

4) O nome do ficheiro deve seguir a normal "eFolioB" + <nº estudante> + <nome estudante com o máximo de 3 palavras>. Penalização de 10% a 100%.

5) Na primeira página do e-fólio deve constar o nome completo do estudante bem como o seu número. Penalização de 10% a 100%.

6) 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.

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

8**)** A legibilidade, a objectividade e a clareza nas respostas serão valorizadas, pelo que, a falta destas qualidades será penalizada.

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

1) (1 valor) Capítulo 15, Concorrency Control
1.a) Defina o protocolo 2-PL. Quais as vantagens e desvantagens? Justifique a resposta.

1.b) Considere o protocolo 2-PL e explique detalhadamente a execução das seguintes transações, usando os operadores X-lock(_), S-lock(_) e Unlock(_). Como classifica a concorrência destas duas transações? Justifique a resposta.

	T1	T2
1	Read A	
2		Read B
3	Write A	
4		Read A
5		Write A
6		Write B
7	Read B	
8	Write B	

2) (1 valor) Capítulo 16, Recovery System
2.a) Quais as principais fases que devem ser consideradas na recuperação? Justifique a resposta.

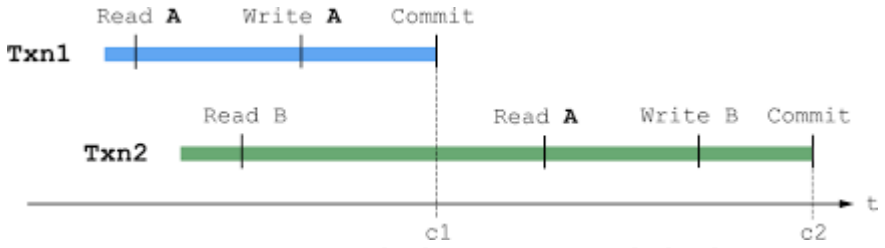
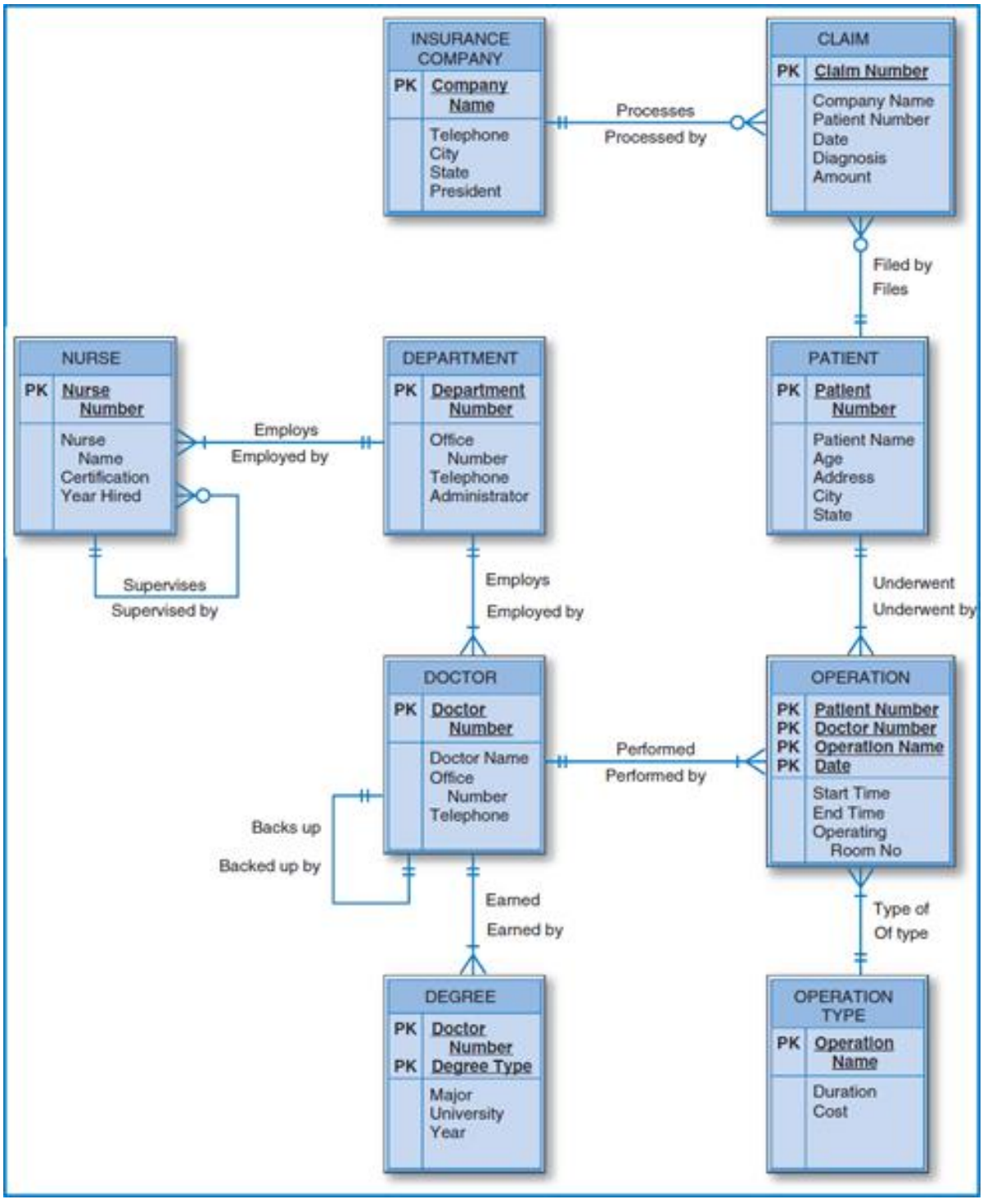


Figura: exemplo de transações na linha do tempo

2.b) Considere a seguinte sequência de log de duas transações em uma conta bancária, com saldo inicial de 12.000, que transfere 2.000 para um pagamento e recebe juros de 5%. Aplique o algoritmo de recuperação ao seguinte log. Represente as transações na linha do tempo como na figura em cima e acrescente os registos na recuperação. Justifique a resposta.

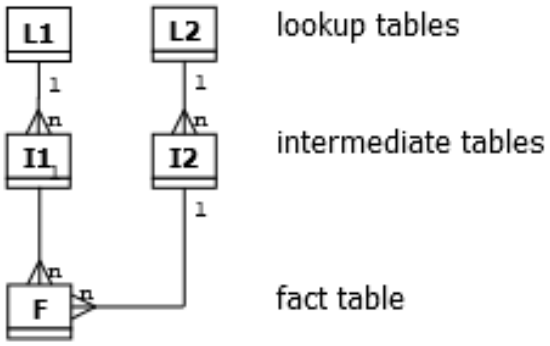
300. Checkpoint
310. T1 start
320. T1 B old=12000 new=10000
330. T1 M old=0 new=2000
340. T1 commit
350. T2 start
360. T2 B old=10000 new=10500
FAIL

Para as perguntas **3) e 4) Desnormalização e Data Warehousing**, considere a seguinte base de dados:



3) (1 valor) Desnormalização

3.a) Reutilize a base de dados transacional na 3ª forma normal. Faça o carregamento de dados. Represente graficamente as ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo. Depois de representar as tabelas classifique-as segundo a tipologia indicada.



3.b). Encontre a 1FD (1ª forma desnormalizada) e a 2FD (1ª forma desnormalizada). Justifique a resposta.

4) (1 valor) Data Warehouse

4.a) Pretendemos desenhar um "Data Warehouse" relacional em estrela ou em constelação, i.e. com duas ou mais estrelas com a maior granularidade possível. Defina a(s) tabela(s) de factos e mostre a tabela depois da desnormalização dos dados. Defina as dimensões com os níveis de agregação para o "Data Warehouse" relacional. Apresente a(s) tabela(s) de factos associada às dimensões. Quantas tabelas de factos encontrou? Preencha a 'bus matrix' (ou business matriz) apresentada em baixo. Justifique a resposta.

		dimension	dimension	dimension	dimension	
business process	fact table	1	2	3	4	
Process X	A	X			X	
	B		X	X		
	C				X	

4.b) Crie duas perguntas e traduza para SQL com Pivot Tables utilizando pelo menos duas dimensões (OLAP). Justifique a resposta.

5. (1 valor) Information Retrieval

Seja n(d) o número de termos num documento "d" e n(d,t) o número de termos "t" num documento "d",

em que a Frequência de um Termo "t" num documento "d" é dado no manual por:

$$TF(d, t) = \log_e\left(1 + \frac{n(d, t)}{n(d)}\right)$$

Seja n(t) o número de documentos que contêm o termo "t" e N o número total de documentos,

onde o *Inverse Document Frequency* (IDF) de Salton & Buckley 1988 é dado por:

$$IDF(t) = \log_{10}\left(\frac{N}{n(t)}\right)$$

Assim, a relevância de um termo "t" num documento "d" é dado por:

$$TF\text{-}IDF(d,t) = TF(d,t).IDF(t)$$

Para a seguinte tabela de frequências de termos versus documentos, encontre para cada documento o termo mais relevante.

Termos \ Documentos	1	2	3	4	5	6	7	8	9	10	11	12	13	14
universidade	81	70	0	72	0	0	60	224	200	0	240	112	96	40
aberta	0	40	100	135	48	0	16	54	90	90	0	6	30	0
informática	90	24	64	224	0	180	0	72	48	0	120	4	70	243
gestão	0	36	0	98	300	630	48	36	72	0	40	0	90	240
humanidades	360	189	80	216	0	120	160	150	98	20	180	72	36	120
matemática	360	10	175	0	540	120	560	9	160	9	420	160	80	0
ambiente	350	147	16	144	0	0	0	0	504	240	0	0	0	324
educação	0	216	0	60	105	648	0	96	240	0	40	49	2	0
história	56	0	250	0	0	720	120	72	81	96	147	24	180	20

Para um determinado termo "t" será possível encontrar os documentos mais relevantes? Discuta a abordagem TF-IDF. Justifique a resposta.