Sistemas de Bases de Dados
Exame de Recurso
Duração: 3 horas (sem consulta)

Grupo 1

Considere parte de uma base de dados registando informação sobre contribuintes identificados por um NIF, podendo ser individuais ou colectivos (empresas), empresas e seus acionistas e respectivo capital social (montante em dinheiro), e atividade económica realizada pela empresa classificada de acordo com códigos de atividade (CAEs) tendo obrigatoriamente um código primário e no máximo 3 secundários (não primários). Essa base de dados inclui as seguintes tabelas, não sendo permitidos valores nulos em nenhum atributo (onde se sublinham os atributos que constituem a chave primária):

contribuintes({NIF,Rua,Local,CodPostal,...}
empresas({NIFEmpr,Nome,Tipo,DataInício,...})
atividades({NIFEmpr,CAE,Primário})
pessoas({NIFPess,Nome,Sexo,DataNasc,...})
acionistas({NIFEmpr,NIFContribuinte,Capital})
codigosAtividade({CAE,Designação,...})

Todas as chaves são formadas por atributos inteiros. O atributo Tipo da tabela empresas tem 4 valores possíveis e o atributo Primário da tabela atividades é booleano. Para cada uma das tabelas existe um índice **clustered** de árvore B+ sobre o(s) atributo(s) da chave primária. Além disso são definidas na base de dados as seguintes *chaves estrangeiras*: de NIFPess em pessoas e NIFEmpr em empresas para contribuintes(NIF), de NIFEmpr em acionistas e atividades para empresas, NIFContribuinte para contribuintes(NIF), e CAE em atividades para codigosAtividade. Os CAE encontram-se numerados sequencialmente a partir de 1. A tabela empresas regista informação de empresas dos últimos 50 anos.

Tendo em conta o sistema de gestão de base de dados usado, tipicamente cabem num bloco 20 tuplos da tabela contribuintes, pessoas e empresas, 50 tuplos da tabela codigosAtividade e 100 tuplos das tabelas atividades e acionistas. Sabemos ainda que num dado momento existem 500.000 contribuintes, dos quais 100.000 empresas e 400.000 pessoas, a tabela acionistas tem 250.000 tuplos e a de atividades 350.000. A tabela de codigosAtividade tem 4000 tuplos. Um nó da árvore B^+ pode conter cerca de 100 chaves de pesquisa e sabe-se que o tempo de um seek é dez vezes superior ao da transferência de um bloco ($t_S = 10 * t_T$). A memória RAM disponível tem capacidade para albergar 100 blocos.

Nota: Neste grupo, sempre que se solicitarem exemplos, estes devem ser **exclusivamente** sobre esta base de dados. Adicionalmente, **todas** as respostas deverão conter uma **breve justificação**.

1a) Apresente dois planos de execução para a query SQL seguinte (obter o NIF e nomes da empresas que têm um código de actividade 3725), indicando qual deles é o mais eficiente.

```
SELECT DISTINCT NifEmpr, Nome
FROM empresas NATURAL JOIN atividades NATURAL JOIN codigosAtividade
WHERE CAE = 3725;
```

1b) Para efeitos estatísticos é necessário saber frequentemente o número de empresas de cada tipo criadas em cada ano. Apresente os comandos para criar o(s) índice(s) para otimizar a resposta a este tipo de consultas. Assuma a existência da função YEAR(DATE *d*) que devolve o ano de *d*.

1c) Justifique matematicamente se compensa a utilização de um índice secundário em árvore B+ sobre o atributo **datainicio** da tabela **empresas** para obter as empresas criadas numa determinada data, quando comparado com uma pesquisa linear (*linear scan*).

1d) Considere que o sistema só implementa o algoritmo de junção *block nested-loop join*, eventualmente utilizando índices quando disponíveis. Para a junção `empresas` ⋈ `atividades` ⋈ `codigosAtividade` indique qual lhe parece ser a melhor ordem para fazer as junções e, em cada operação, qual deveria ser a "inner table" e qual deveria ser a "outer table".

1e) Explique como o facto da tabela acionistas ter um índice primário pode ser vantajoso para responder a consultas da seguinte forma, em que **A** <= **B** são valores conhecidos:

```
SELECT NIFEmpr, SUM(Capital) AS capital_social
FROM acionistas
WHERE NIFEmpr BETWEEN A AND B
GROUP BY NIFEmpr
```

1f) Considere a seguinte lista de eventos envolvendo três transações (as únicas a executar no sistema). Apresente o registo de log de acordo com o algoritmo estudado, sabendo que as empresas com NIF 1, 2 e 3 têm tipo 0.

	T1	**T2**	**T3**	**sistema**
1	begin transaction UPDATE empresas SET Tipo = 1 WHERE NIFEmpr IN (1,2);			
2		begin transaction UPDATE empresas SET Tipo = 2 WHERE NIFEmpr=3;		
3				**CHECKPOINT**
4	ROLLBACK;			
5			begin transaction UPDATE empresas SET Tipo = 1 WHERE NIFEmpr=1;	
6			COMMIT**;**	
7				**CRASH!**

1g) Apresente um escalonamento com apenas duas transações que origine uma exceção quando executado no sistema ORACLE 11g no modo de isolamento SERIALIZABLE.

1h) Os sistemas de gestão de base de dados que estudou permitem fazer locks com diversos níveis de granularidade (e.g. ao nível da tabela ou do tuplo). Apresente uma operação sobre a base de dados em que seja potencialmente mais vantajoso efectuar o lock ao nível da tabela em vez de ao nível dos tuplos.

Grupo 2

Nota: Dê respostas breves mas justifique adequadamente.

2 a) Indique como o algoritmo de hash-join deve ser modificado para calcular junções externas do tipo R LEFT NATURAL JOIN S, assumindo que não é necessário particionamento recursivo. Distinga os casos em que R é a relação *build* ou *probe*.

2 b) O modo de isolamento SNAPSHOT ISOLATION pode introduzir anomalias conhecidas como *write skews*. Explique em que consiste este problema, mencionando como pode surgir e concretizando com um exemplo simples.

2 c) Os protocolos de maioria (*majority*) e enviesado (*biased*) são utilizados para controlo de concorrência a dados replicados. Compare ambos os protocolos, analisando as suas vantagens e inconvenientes.