Teste #1 de Ciência de Dados em Larga Escala, 2023/24
Data: 10/04/2024, Duração: 1h:50min
Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto

1. Com relação ao paradigma de "cloud computing":

(a) Descreva 3 características essenciais deste paradigma.

(b) Dê um exemplo de serviço concreto para cada um dos modelos de serviço "cloud" IaaS, PaaS, e SaaS. **Justifique a sua escolha.**

(c) Qual é a distinção entre "clouds" públicas e privadas/comunitárias? Indique uma vantagem e uma desvantagem de usar uma "cloud" pública.

2. Da perspectiva de armazenamento de dados na "cloud", indique 2 diferenças fundamentais entre o uso de "Object stores" (ex. "buckets" no Google Cloud Storage) e sistemas de ficheiros ("file systems").

3. Para computação na "cloud", suponha que pretende implementar uma determinada aplicação "cloud" acessível através da Internet e que tem a opção de fazê-lo usando máquinas virtuais dedicadas via Google Compute Engine. Explique as diferenças entre utilizar esta opção ou uma implementação num servidor web local, dedicado, em termos do uso de recursos, gestão do serviço, e faturação.

4. Da perspectiva de processamento de dados segundo o modelo MapReduce:

(a) O que se quer dizer por "map" e "reduce"?

(b) Num "cluster" de máquinas, como é feito o processamento de dados com estágios "map" e "reduce"?

(c) Considere o seguinte fragmento de código PySpark, uma variante do exemplo clássico de contagem de palavras. Explique o significado do código: que processamento dos dados é feito e que resultados finais são obtidos? Na explicação faça a relação com os conceitos de RDD, transformação e ação (apresente claramente estes conceitos).

```
someFile = ...
rdd =
  sc.textFile(someFile)\
    .flatMap(lambda line:
      [(word,1) for word in line.split()])\
    .reduceByKey(lambda x,y: x + y) \
    .filter(lambda pair: pair[1] >= 10)\
    .sortByKey(ascending=False)\
    .map(lambda pair: (pair[1], pair[0]))
data = rdd.collect()
```

5. Sobre o modelo *Publisher-Subscriber*, responda:

(a) Qual é o significado do comando:

```
gcloud pubsub topics publish messages --message=''message10''
```

(b) É possível criar múltiplos tópicos para a mesma subscrição?

(c) Qual é o resultado de `gcloud pubsub subscriptions pull my-sub --auto-ack` quando my-sub tem vários tópicos associados?

6. O trecho de programa da Figura 1 mostra um exemplo de programação paralela em Python. Assuma que a variável `data` é uma matriz 16x16 e que temos 4 processadores (`num_cpus`), responda:

(a) Qual é a parte da matriz que cabe a cada processador ao executar a função `how_many_within_range`?

(b) Responda à mesma pergunta e indique a diferença em execução se substituirmos o `apply` pelo `apply_async`.

```
3 import multiprocessing as mp
4
5 # Is this actually running with multiple cpus?
6 num_cpus = mp.cpu_count()
7 print('Num cpus = ', num_cpus)
8
9 # begin timing
10 start_time = time()
11
12 # Step 1: Init multiprocessing.Pool()
13 pool = mp.Pool(mp.cpu_count())
14 # end timing init
15 print('Time to create pool: ',round(time() - start_time,8), 'seconds')
16
17 # Step 2: `pool.apply` the `howmany_within_range()`
18 results = [pool.apply(howmany_within_range, args=(row, 4, 8)) for row in data]
19
20 # Step 3: Don't forget to close
21 pool.close()
```

Figure 1: Exemplo async

7. Indique 1 (uma) vantagem e 1 (uma) desvantagem para cada uma das arquiteturas paralelas:

(a) memória centralizada

(b) memória distribuída

8. No pseudo-código abaixo, implementado utilizando um paradigma de passagem de mensagens, responda:

```
define submatriz local
for num_iters
    if pid != 0
        send first line to process pid-1
        receive last line from process pid-1
    if pid != P-1
        send last line to pid+1
        receive first line from pid+1
    for num_linhas
        compute
```

(a) Faz sentido trocar a ordem das operações `send` e `receive`? **Justifique sua resposta.**

(b) Se a função `compute` for implementada para resolver o problema de Successive-Over-Relaxation (SOR), onde cada nova célula da matriz é calculada a partir dos seus vizinhos nas linhas e colunas, a distribuição de dados usada neste código é adequada? **Justifique sua resposta.**

9. Qual é a função de um *Virtual Machine Monitor* (VMM)?

10. Fostes contratado para otimizar o tratamento de dados na empresa OptData. A empresa recebe diariamente dezenas de terabytes de dados que são armazenados em disco numa cloud externa e que precisam ser usados para calcular várias estatísticas. Qual seria a melhor forma de lidar com estes dados e alcançar a tua tão esperada promoção (discuta sobre formas de armazenamento, distribuição, infraestrutura e processamento)?