Big Data e Cloud Computing, 2020/21
Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
Exame de época normal – 21/06/2021
Duração: 2:00

PART A

Answer parts A and B in separate sheets.

1. In what concerns the cloud computing paradigm:

(a) Identify 3 essencial characteristics associated to the cloud computing paradigm.

(b) What is the distinction between public clouds and private/community clouds? State an advantage and a disadvantage of using a public cloud.

(c) Provide an example of cloud service for each of the IaaS, PaaS, e SaaS service models. Explain your choices.

2. Regarding data storage, state 2 fundamental differences between the use of:

(a) “Object stores” (e.g. buckets in Google Cloud) vs. file systems.

(b) Data warehouses like BigQuery vs. database systems.

3. Regarding cloud computation:

(a) Suppose you wish to implement a cloud application and that you have an option between using 1) Google AppEngine and 2) dedicated virtual machines using Google Compute Engine. Explain the difference from the perspective of resource management and billing.

(b) What is the difference between sole-tenant and multi-tenant nodes in a cloud data center? What kind of benefits may a a sole-tenant node bring?

4. Consider the following Pyspark code fragment, a variant of the classic “word count” example.

```
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()
```

(a) Explain what the code does in terms of data processing and the final results obtained. Relate your explanation to the concepts of RDD, transformation, and action.

(b) Considering the nature of Spark transformations, which can be *narrow* ou *wide*, and the eventual necessity of data *reshuffling*, explain what Spark execution *stages* we may have for the processing in the example.

PART B

Answer parts A and B in separate sheets.

1. Discuss about alternatives for handling **big data** in the Data Sources layer. In other words, what kind of operations, functions and implementations would guarantee a proper handling of data *volume*? Discuss about data loading and data preprocessing.

2. With respect to current systems, tailored for big data, studied in class, what are their limitations regarding data characteristics such as *velocity*, *value*, *veracity* and *variety*?

3. Explain the meaning of each line of Python code below:

```
def run(model_dir, feature_extraction, sink, beam_options=None):
  print(’Listening...’)
  with beam.Pipeline(options=beam_options) as p:
    _ = (p
        | ’Feature extraction’ >> feature_extraction
        | ’Predict’ >> beam.ParDo(Predict(model_dir, ’ID’))
        | ’Format as JSON’ >> beam.Map(json.dumps)
        | ’Write predictions’ >> sink)
```

4. Give two advantages and two disadvantages of the utilization of GPUs for data processing and analysis.

5. Given the Python code below, explain the function of lines: 6, 7, 10, 12, 14 and 16.

```
1  def parallel_map(function, iterable):
2    if FORCE_DISABLE_MULTIPROCESSING:
3      return [function(*args) for args in iterable]
4
5    original_sigint_handler = signal.signal(signal.SIGINT, signal.
         SIG_IGN)
6    num_threads = mp.cpu_count() * 2
7    pool = mp.Pool(processes=num_threads)
8    signal.signal(signal.SIGINT, original_sigint_handler)
9
10   p = pool.map_async(_function_wrapper, ((function, args) for args in
         iterable))
11   try:
12     results = p.get(0xFFFFFFFF)
13   except KeyboardInterrupt:
14     pool.terminate()
15     raise
16   pool.close()
17   return results
```

6. Figure 1 presents a set of data for binary classification. The problem is to build a predictive model that discriminates trains that go east from trains that go west. Represent this problem with a single bidimensional table. What are the disadvantages of representing this problem in this format? (Note: this problem could be solved using image processing machine learning algorithms, but we are interested in “interpretable” and “explainable” models.)

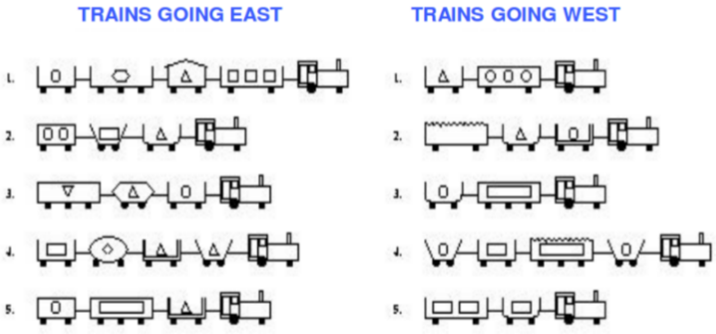

Figure 1: Figure for question 6

7. Given the slice of code below, what would be the problem of allocating multiple threads to execute this loop?

```
do I=2,9
   X[I] = Y[I] + Z[I]
   A[I] = X[I-1] + 1
enddo
```

8. Explain the differences between the two script codes (CODE 1 and CODE 2) below. **Do not omit details.**

```
               CODE 1

WORK_DIR=/tmp/cloudml-samples/molecules
python predict.py \
  --work-dir $WORK_DIR \
  --model-dir $MODEL_DIR \
  batch \
  --inputs-dir $WORK_DIR/data \
  --outputs-dir $WORK_DIR/predictions
```

```
               CODE 2

PROJECT=$(gcloud config get-value project)
WORK_DIR=/tmp/cloudml-samples/molecules
python predict.py \
  --work-dir $WORK_DIR \
  --model-dir $MODEL_DIR \
  stream \
  --project $PROJECT \
  --inputs-topic molecules-inputs \
  --outputs-topic molecules-predictions
```