DI/FCT/NOVA
Mestrado Integrado em Engenharia Informática

Cloud Computing Systems
1st Semester, 2020/2021

Final Test (8/January/2021)

IMPORTANT NOTE: you should answer with a reasonable letter size. The idea of boxes is not to see how small you can write ☺ As an indication, the box for the first question should not hold more than 4-5 lines (and concise and precise answers will be highly appreciated).

1) In the context of the first project of the course, it would be possible to replicate images to a different region by using an Azure function that, whenever an image was added to the local blob store, would write the image also in the remote blob store. What are the interesting properties of this approach? Justify (contrasting this approach with having the REST method to write the image in both the local and remote blob store) .

1. Using Azure functions to replicate blobs guarantees that images are replicated without the need to add the code for replication in every place where images are added to the blob store.
2. By writing only on the local storage, the method can return faster.

2) "In Map-Reduce, a reducer can only start executing after all mappers have finished". State if this statement is true or false and justify.
Suggestion: consider the logical execution steps of a map-reduce task, in particular, the steps between map and reduce.

True, because… / False, because…

True, because between the mapper and the reducer the system needs to sort the data before passing it to the reducers. For sorting the data, it is necessary to have all data from the mappers.

3) Assume a data set including a log of transfer transactions in a banking system, with the following format, where: **Date** is the date of the transfer, **Source** and **Destination** are identifiers of the source and destination accounts (with the first three digits in an identifier identifying the bank of the account), **Amount** is the amount transferred, **Commission** is the commission paid for the transaction and **State** is either 0 for failed transfers and 1 for successful transactions. The element are separated by a tab ('\t'). Example:

```
Date Source Destination Amount Commission State
2016-12-06T08:58:35.318+0000 032001234 034007890 100.00 1.25 1
...
```

The following Python program processes information from the log of transfer transactions using Spark RDD interface.

```
result = sc.textFile('log.log')
           .map( lambda line: line.split('\t'))
           .map( lambda row: (row[1], float(row[4]) ) )
           .reduceByKey( lambda v1, v2: v1 + v2)
           .filter( lambda pair : pair[0].startWith( '032'))
```

a) Explain the problem that the presented program is solving.

Compute, for every account of bank 032, the sum of commissions of transfers from that account.
(the sum includes commisions from transfers that have failed)

b) Do you think it would be possible to optimize this program, making it more efficient, by reordering the operators of the program? Justify.

Yes, by executing the filter after the first map, the following operators would process a much smaller number of tuples.
This is specially relevant for the reduceByKey, where data needs to be transferred among machines.

4) Computing whether a road is congested or not is a typical example of a computation that is performed using stream processing systems. Can this computation be performed using a mini-batch model or should it be performed using a continuous processing model? Justify.

Can use mini-batch model because… / Need continuous model because…

Can use mini-batch model because the period of each batch can be (and typically is) so small (in the order of seconds) that the result of of the traffic will not change significantly in a way that is relevant for the drivers.

5) Briefly explain what is paravirtualization and why it was necessary in x86 architectures.

Paravirtualization is a virtualization technique in which the guest OS is modified to invoke special hypervisor calls to execute privileged operations.
This technique is necessary the VMM needs to control the privileged operation executed by the guest OS kernel, and x86 originally did not have hardware support for this. Alternative solutions, such as trapping and interpreting these operations was very slow.

6) A VM can run with a managed disk (or virtual hard disk) stored in a blob store service. Present two techniques used by IaaS services to make access to these virtual disks efficient?

1. 1. Use the local disk in the machine where the VM is running as a big cache of the virtual hard disk.Thus, most accesses to the virtual hard disk will end up being local.

   2. Use reserved network bandwidth.
2. 3. Make some directories, such as /tmp, to be local-only, thus not incurring in the overhead of getting the data from a remote server.

7) Consider you want to run some given software (e.g. database) in a given computer. Present reasons to use a solution based on containers and, alternatively, on Virtual Machines.

Possible reasons to use containers:

The key reason for using containers (instead of VMs) is that containers are more lightweight.

Other reasons may be the fact that typically there is a container with all the database software needs and there is no need to install it in the VM, the fact that containers have version management that helps keeping the software up-to-date.

Possible reasons to use VMs:

The key reason for using VMs (instead of containers) is when you want to run a software with an OS different from the OS of the node where the VM is running.

Note: in this question, many students just copied here the general properties of containers and VMs without considering the specific question.

8) Consider a container service where users can run their containers. Suppose this service is implemented by running one (or a small number of) VMs in each physical computer, with each VM running multiple containers at the same time. Explain why the use of Copy-on-write File systems helps in making the service running efficiently.

Copy-on-write file systems, like UnionFS, help by:

1. allowing different containers to share some of the layers that compose the container (e.g. language support)

2. by creating a new layer when writes are performed, this allows any number of instances to use the same container image without sharing the created files.

9) Kubernetes is often used to support micro-service platforms. Discuss why this is the case (consider the properties of Kubernetes that could be used for supporting micro-services).

Kubernetes provides solutions for the key challenges of micro-services, by having, e.g., each micro-service in a Pod:

* How to manage micro-services?
• How to make micro-services reliable?

Kubernetes manages the deployment and guarantees that the specified pods (micro-services) are running at all time. It also provides support for having multiple replicas of a micro-service.

• How to find a micro-service?
• How to communicate with a micro-service?

By associating a Kubernetes service with a given pod (or set of pods), Kubernetes provides a simple way for a micro-service to reach some other micro-service, providing the necessary service discovery and managing the communication to the micro-services.

10) Consider that a company wants to deploy a new multi-user online game. The company has a (small) private cloud (data center) where it will host the servers of the game. In which conditions do you think it would be interesting to consider using an hybrid cloud solution? Note: consider both hybrid monocloud and hybrid multicloud approaches in your reply.

In this case, the main reason would be able to address peaks of load, by using resources of the public cloud.
Other reasons that could be mention include keeping sensitive data in the private cloud and run non-sensitive services in the public cloud and to use the resources of the public cloud to run the service closers to users.

These are generic reasons. As for the hybrid monocloud vs. hybrid multicloud, the former would make it simpler to devlop and manage the software while the latter would minimize the vendor lock-up and could probably allow for a more wide geographic spread of the application (if this is needed).