DI/FCT/NOVA
Mestrado Integrado em Engenharia Informática
Mestrado em Engenharia Informática

Cloud Computing Systems
1st Semester, 2021/2022
Duration: 2 hours
Final Test (8/January/2022)

Num: _____ Name:_____

Wrong answers to V/F questions discount up to the equivalent of the corresponding right answer value. For multiple choice questions, the discount is $1/(n-1)$, with n the number of choices. The penalty only accumulates in the context of the same question. For each question, the first wrong answer does not count.

1) Answer the following questions.
a) _____ (V/F) Unlike Map-reduce that is designed to run batch processing jobs, Spark is designed primarily to perform stream processing jobs.
b) _____ (V/F) Spark includes an optimizer that, given a program written using the Dataframes API or SQL, optimizes the order of execution of operation to minimize execution time.
c) _____ (V/F) In Spark, one of the reasons why wide dependencies impact performance is because for computing a partition of a given Dataframe, it might be necessary to wait for a late partition of the previous Dataframe to complete.
d) _____ (A/B) Consider you have a virtualization system A based on paravirtualization and a virtualization system B based on binary translation. Which one can run more systems?
e) _____ (V/F) When using PCI passthrough approach to support GPU virtualization, a GPU can only be used by a single virtual machine (as the VM has direct access to the GPU).
f) _____ (V/F) Live migration is useful mainly for minimizing problems when computers running VMs fail unexpectedly.
g) _____ (V/F) In Azure, an application can improve security of resources created (e.g., Azure Storage service, SQL databases, etc.) by restricting the origin of connections to the resource to a virtual network of the application.
h) _____ (V/F) It is common that large data centers have (diesel) generators to be able to continue operation during power outages.
i) _____ (V/F) For supporting VLANs, switches need to maintain configuration tables specifying which VLANs are accessible through which interface.

2) Answer the following questions.
a) _____ (V/F) Cgroups is the technique that guarantees a container cannot access resources from other container even if a container is running malicious code.
b) _____ (A/B/C) Consider a machine running multiple containers. The images of these containers all share the same base (e.g., FROM tomcat) but include different application code (e.g., different war files). Copy-on-write filesystem can be used to improve space used in the machine: (A) only for containers of the same image; (B) both for containers of the same image and for containers with a different image (i.e., different war files); (C) only for container with different images.
c) _____ (V/F) In a Dockerfile, a RUN command generates a new layer in the copy-on-write file system used to store a container image.
d) _____ (V/F) Docker compose allows to define and deploy multi-container applications in multiple machines.

e) ____ (V/F) Kubernetes services can expose a different port for applications than the Pods/containers implementing that service.
f) ____ (V/F) After modifying the deployment (YAML) files of a Kubernetes application, it is necessary to explicitly stop the application (Pods, services, etc.) before applying the new configuration.
g) ____ (V/F) OpenStack is a service-level multi-cloud solution.
h) ____ (V/F) Applications that perform analytics on the edge typically are restricted to only execute classification on the edge, while learning is executed in the cloud.
i) ____ (V/F) For companies that manage some sensitive data (e.g., medical information), adopting a hybrid cloud computing approach is preferable to relying only on public cloud.

3) Consider a data processing program with a larger number of steps, leading to a large sequence of map-reduce jobs or a large workflow (or sequence) of Spark Dataframes. Compare the fault-tolerance approach of map-reduce and Spark in this context, briefly **explaining how they work** and their **advantages/problems**.

Map-reduce:

Spark:

4) Consider you have a log of food delivery services, with the following format, where: Date is the date of the food delivery, Restaurant is the id of the restaurant where the food was picked up, Driver is the id of the person delivering the food, Destination is the postcode of the delivery location, Price is the price of the food being delivered, and Tip is the tip given by the client. The following log entry represents a job performed on 2020-12-06 at 08:58:35, for restaurant with id 4356 by driver with id 87645, for destination 1495-712, with food price of 20.5€ and a tip of 3€.

Date,Restaurant,Driver,Destination,Price,Tip

2020-12-06,4356,87645,1495-712,20.5,3

Assuming that you have read the log of entries and registered it under the name "deliveries", answer the following questions.

a) What is being computed by the following Spark SQL statement:

SELECT Driver, SUM(Price+Tip) AS rev FROM deliveries GROUP BY Driver ORDER BY rev DESC LIMIT 10

b) Present the SQL statement to compute, for each day, the restaurant (or restaurants) that had the largest number of deliveries.

5) Sensitive instructions are those which behave differently when executed in user and supervisor modes. If all sensitive operations were privileged, would it be necessary or useful to resort to techniques like binary translation when performing virtualization? Justify.

Not necessary nor useful, because… / Not necessary, but useful, because… / Necessary and useful, because…

6) Consider the following screenshot, showing the information associated with the docker image of MongoDB version 4.2.16 available at DockerHub:

Given what you learned about how containers work, discuss why there are images for Linux and Windows, and images for amd64 and arm64.

7) A Pod is the basic execution unit of a Kubernetes application. Is there any advantage of this design, or using a (simple) container as the basic execution unit would lead to similar properties for the Kubernetes application? Justify.

Having Pods is advantageous because… / It would be the same because…

8) Consider that you want to create a service similar to discord (as the one developed in the SCC project this year). Could you use edge computing to provide a better experience for users? Justify explaining why not or which functionalities you could implement adopting edge computing (give concrete examples).

Yes, because… / No, because …