Big Data Frameworks mid-term Exam IIS3 **on** 12/04/2019 **Intructor:** Dr. Rim Moussa

First Name:..... **Last Name**: ..... **Group#**:.....

scoring:1.5, 1, 1, 1.5, 5

1/Assume a cluster has 10 DataNodes, each with a 1TB hard drive capacity dedicated for HDFS. Default replication settings i.e. 3 is implemented for all files. What is the storage capacity of the Hadoop cluster (assuming no compression) ? ….....

2/What does it mean when an algorithm is said to 'scale well'? ….....

….....

3/ What happens in a MapReduce job when you set the number of reducers to zero?
a) No reducer executes, and the output of each mapper is written to a separate file in HDFS.
b) No reducer executes, but the outputs of all the mappers are gathered together and written to a single file in HDFS.

4/Tell the type of the following scenarios : batch processing or real-time processing
a) A bank wants to aggregate the savings of all its customers across individual bank accounts to provide this information to the tax authorities: ….....
b) Every second a flood monitoring centre receives information from thousands of sensors embedded in dikes. Unusual readings are relayed to the authorities: ….....

Big Data Frameworks mid-term Exam IIS3 **on** 12/04/2019 **Intructor:** Dr. Rim Moussa

First Name:..... **Last Name**: ..... **Group#**:.....

scoring:1.5, 1, 1, 1.5, 5

1/Assume a cluster has 10 DataNodes, each with a 1TB hard drive capacity dedicated for HDFS. Default replication settings i.e. 3 is implemented for all files. What is the storage capacity of the Hadoop cluster (assuming no compression) ? ….....

2/What does it mean when an algorithm is said to 'scale well'? ….....

….....

3/ What happens in a MapReduce job when you set the number of reducers to zero?
a) No reducer executes, and the output of each mapper is written to a separate file in HDFS.
b) No reducer executes, but the outputs of all the mappers are gathered together and written to a single file in HDFS.

4/Tell the type of the following scenarios : batch processing or real-time processing
a) A bank wants to aggregate the savings of all its customers across individual bank accounts to provide this information to the tax authorities: ….....
b) Every second a flood monitoring centre receives information from thousands of sensors embedded in dikes. Unusual readings are relayed to the authorities: ….....

5/Assume you are given a list of [filename : string, md5hash : string] pairs. How would you find the names of duplicate files where you should report only distinct file names? FYI Two files are duplicate if their md5hash values are equal. Provide the pseudo-code for the map and reduce functions

Mapper	Reducer

5/Assume you are given a list of [filename : string, md5hash : string] pairs. How would you find the names of duplicate files where you should report only distinct file names? FYI Two files are duplicate if their md5hash values are equal. Provide the pseudo-code for the map and reduce functions

Mapper	Reducer