

TDDE31 / 732A54 Big Data Analytics

Exam **Part 1**

June 2, 2020

8:00 – 9:40

Instructions: See <https://www.ida.liu.se/~732A54/exam/distanceexam.en.shtml>

Question 1 (1p)

Consider the following claim:

While read scalability can be achieved by scaling horizontally (scale out), it cannot be achieved by scaling vertically (scale up).

Is this claim correct or wrong? Justify your answer in about two to four sentences.

Question 2 (1p)

Consider the following relational database which consists of two relations (Project and Report). Notice that the attribute finalreport in the relation Project is a foreign key that references the primary key (attribute id) in the relation Report. Notice also that multiple projects may have the same final report.

Project			Report		
<u>name</u>	budget	finalreport	<u>id</u>	pages	location
UsMis	1,000,000	391	121	70	http://acme.com/beerep
AMee3	3,700,000	391	391	350	http://acme.com/r391
Bee	1,300,000	121	699	100	http://acme.com/Other

Capture all the data in this relational database as a document database.

Question 3 (1p)

Identify two differences between the key-value database model and the document database model that was introduced in class.

To answer this question write a maximum of 200 words.

Question 4 (1p)

Consider the following claim:

In master-slave replication, one compute node is selected as the master node for all the database objects (e.g., all key-value pairs in a key-value database) and the other compute nodes become slave nodes.

Is this claim correct or wrong? Justify your answer in about two to four sentences.

Question 5 (1p)

Assume a (multi-master) system in which each database object is replicated at 4 nodes. We want to allow the system to require a quorum of only 1 node when we read a database object. Then, in order to still achieve strong consistency for these reads, how many nodes have to confirm a write of a database object for the write to be considered successful? Justify your answer in about two to four sentences.

Question 6 (1p)

Distributed file systems such as HDFS work on relatively large blocks of data (such as 64 MB) as the units of data sharing and transfer. Why is it not advisable to use much smaller blocks (e.g., 1 KB), or even individual key-value pairs? Give the technical reason(s) for this design choice. Be thorough.

To answer this question write a maximum of 200 words.

Question 7 (1p)

Execution of a MapReduce operation involves seven phases. Which of these phases of MapReduce may involve disk I/O

- (a) to/from HDFS,
- (b) not to/from HDFS,

and for what purpose?

To answer this question write a maximum of 200 words.

Question 8 (1.5p)

Spark classifies its functions on RDDs into two main categories: “Transformations” and “Actions”. Describe the main differences between these categories, and give one example operation for each category.

To answer this question write a maximum of 200 words.

Question 9 (1.5p)

Characterize the type and structure of applications that are expected to perform significantly better when expressed in Spark than in MapReduce, and explain why.

To answer this question write a maximum of 200 words.

Question 10 (4p)

You are asked to implement in Spark (PySpark) a very simple regression algorithm and estimate its generalization error. To do so, there is some training and test data available. The former takes the form $\{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)\}$ and the latter the form $\{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_M, t_M)\}$. Use the training data to learn the regressor, which will simply be the mean of the target values in the training data. Use the test data to estimate the generalization error of the learned regressor. The error is defined as the mean absolute error, i.e., the mean of the absolute value of the difference between the predicted target and the true target.

It is not required that your code actually compiles; nevertheless, it must be code rather than pseudocode, i.e., you have to use the transformations and actions properly.

To get full points you need to comment your code.