TDDE31 Big Data Analytics

Exam
Part 1

August 20, 2020

8:00 – **10:30**

(Part 2 becomes available at 9:30)

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

Grades: You can get up to 14 points for this first part of the exam and another 15 points for the second part, which together may give you an overall of max 29 points. To pass the exam (grade 3 or E) you have to meet both of the following two conditions: First, you need to achieve at least 7 of the 14 points that can be achieved in the first part of the exam. Second, for both parts together, you need to achieve at least 14.5 of the 29 points that can be achieved overall. If you do not meet the first condition, your second part will *not* be considered for grading.

After fulfilling the aforementioned requirements to pass the exam, then for grade 4, you need at least 20 points (for both parts together), and for grade 5, you need at least 26 points.

Questions: If you have clarification questions regarding some of the exercises in the exam, please do the following depending on the exercise.

If you need clarifications on Questions 6–8, then email christoph.kessler@liu.se

If you need clarifications on Question 9, then email jose.m.pena@liu.se

If you need clarifications on Questions 1–5, or about something more general related to the exam, the examiner will be available in the following Zoom meeting room throughout the whole time of the exam.

https://liu-se.zoom.us/j/61055128986?pwd=SDlpOE04U2tEbllMHRGMWx3Vi9IZz09
Meeting ID: 610 5512 8986
Password: 588119

Notice that this Zoom meeting room has been set up using the waiting room feature of Zoom. Hence, when you enter, you will be put into the waiting room and, from there, you will then be admitted to the meeting room to ask your question.

Question 1 (1p)

Consider the following key-value database which contains four key-value pairs where the keys are user IDs and the values consist of a user name, the user's year of birth, and an array of IDs of users that the current user likes (for instance, Bob likes Charlie).

```
"alice_in_se" → "Alice, 1987, [bob95 charlie]"
      "bob95" → "Bob, 1995, [charlie]"
    "charlie" → "Charlie, 1996, []"
     "selaya" → "Alice, 1974, [charlie]"
```

Describe how the types of queries typically implemented in a key-value store can be used to retrieve the names of all users liked by a given user, specified by his/her ID (i.e., the ID of that user is given as input).

To answer this question write a maximum of 200 words.

Question 2 (1p)

Describe one advantage that using a document store has, in comparison to using a key-value store.

To answer this question write a maximum of 100 words.

Question 3 (1p)

Describe in two to four sentences how write scalability can be achieved by scaling vertically (scale up).

Question 4 (1p)

Consider the following claim:

> *In master-slave replication, a replica (copy) of a database object at a slave node never changes.*

Is this claim correct or wrong? Justify your answer in about one to three sentences.

Question 5 (1p)

Assume a (multi-master) system in which each database object is replicated at 4 nodes. In order to achieve strong consistency for writes ("write consistency"), how many nodes have to confirm a write of a database object for the write to be considered successful? Justify your answer in about one to three sentences.

Question 6 (1.5p)

The execution of a MapReduce operation involves seven phases. Which of these phases of MapReduce may involve *network* I/O, and for what purpose?

To answer this question write a maximum of 200 words.

Question 7 (3p)

Write pseudocode for a MapReduce program that reads floating point numbers x_i from a HDFS input file (in a text format of your choice, such as one value per line) and computes their geometric mean ($\sqrt{\sum_i x_i^2}$). Explain your code!

Hint: Make sure to follow the MapReduce programming model and clearly identify its different program parts in your code. — If you are unable to write MapReduce pseudocode, you may, at reduced points, write Spark pseudocode instead, or just try and describe the MapReduce program in plain English as precisely as you possibly can.

To answer this question write a maximum of 50 lines of commented pseudocode and a maximum of 200 words of explanation.

Question 8 (0.5p)

Where are the data elements of a Spark RDD stored when evaluating a lineage of RDDs?

Question 9 (4p)

You are asked to implement in Spark (PySpark) an initialization method for the K-means algorithm. In slide 10 of lecture 11, you can find a PySpark implementation of the K-means algorithm. In that implementation, the initial centroids are K randomly chosen data points. Now, you are asked to implement a slightly more advance initialization. Specifically, you are asked to assign the data points to clusters at random and, then, compute the average of each such randomly generated cluster. These averages are the initial centroids for the K-means algorithm.

To get full points you need to comment your code.