732A54 Big Data Analytics

Exam
Part 1

June 1, 2021
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 requirements to pass the exam, then for grade D, you need at least 18 points (for both parts together); for grade C, you need at least 21 points; for grade B, you need at least 24 points; for grade A, you need at least 27 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 6–9, then email christoph.kessler@liu.se

If you need clarifications on Question 10, 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/69035369731?pwd=U3h1N2taRk95d2JkY1JnMUttTmwvZz09
Meeting ID: 690 3536 9731
Passcode: 790863

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 claim:

> *Write scalability can be achieved not only by scaling horizontally (scale out), but also 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 project in the relation Report is a foreign key that references the primary key (attribute name) in the relation Project. Notice also that multiple reports may be for the same project.

Project

name	budget
UsMis	1,000,000
AMee3	3,700,000
Bee	1,300,000

Report

id	project	filename
121	Bee	beerep.pdf
391	UsMis	rep391.pdf
699	Bee	OldRep.pdf

Capture *all* the data in this relational database as a key-value database.

Question 3 (1p)

Remember that values in a key-value database are opaque to the key-value store. What exactly does this mean both

i) from the perspective of the key-value store and

ii) for a user of such a system?

Answer this question in two to five sentences *by using your key-value database from the previous question as an example*.

Question 4 (1p)

Data warehouses are usually much bigger in size than operational databases. Explain, in two to five sentences, why that is.

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 2 nodes when we *write* a database object (i.e., for the write to be considered successful, 2 nodes have to confirm that they have completed the write). Then, in order to achieve strong consistency (for reads), how many nodes have to agree on the value to be returned in response to a read request for a database object? Justify your answer in about two to four sentences.

Question 6 (1p)

We learned about the concept of high-level parallel programming using *algorithmic skeletons*. Give one example of such an algorithmic skeleton that occurs in the Spark API. (Hint: it is *not* MapReduce.)

Explain shortly what the given algorithmic skeleton does, describe its input-output dependence structure and whether Spark evaluates it lazily or not.

To answer this question write a maximum of 100 words.

Question 7 (1 + 1 = 2p)

The block size used in the Hadoop distributed file system is, by default, many Megabytes large (typically, 64MB). Considering the way how cluster architectures and parallel computations atop HDFS are organized, give the technical reasons why this size is usually a reasonable choice from a performance point of view:

(a) why should one not use much smaller block sizes (e.g. just a few dozen bytes)?

(b) why should one not use much larger ones (e.g. many Gigabytes) either?

For each of these two questions, write a maximum of 100 words, respectively.

Question 8 (1p)

We learned about different node interconnection networks that can be used in clusters, and the implications of their topology for network cost, throughput, and scalability. Give the name and describe these properties (be thorough) for a very cheap interconnection network which has very limited throughput and which does not scale up to many nodes.

To answer this question write a maximum of 100 words.

Question 9 (1p)

Given a Spark program that consists of a chain of $K > 1$ dependent *transformations* $T_1, ..., T_K$, where each transformation T_i uses the result of the previous transformation T_{i-1} as its input. The input (of T_1) is a HDFS file with B blocks. Assume that each transformation takes time t_B per block.

First, draw the *task (dependence) graph* for this computation for $K = 3$ and $B = 4$, where each transformation over each block constitutes a task.

If $P \geq B$ workers are available for this computation, what is the fastest possible parallel time (general formula, using the above symbols) for the entire computation? Explain your calculation.

To answer this question write a maximum of 100 words.

Question 10 (4p)

In this course, we have seen examples of classification (e.g., logistic regression), regression (e.g., kernel methods) and clustering (e.g., k-means). Yet another important task in machine learning is that of density estimation. The goal in this task is estimating the density function $f(X = x)$. This task can be solved via kernel methods as

$$f(X = x) = \frac{1}{N} \sum_n k\left(\frac{x - x_n}{h}\right)$$

assuming that the kernel function $k(u)$ satisfies the following two conditions:$^{(*)}$ i) $k(u) \geq 0$ for all u and ii) $\int k(u)du = 1$.

Assume that we have access to some training data of the form

$$\{(x_n, y_n) | n = 1, \ldots, N\},$$

where X_n is a unidimensional continuous predictor random variable and Y_n is a binary class random variable. You are asked to implement in PySpark the kernel method above to estimate $f(X = 1|Y = 0)$ and $f(X = 2|Y = 0)$. Hence, you have to estimate the density function values for the points $X = 1, 2$ for class $Y = 0$. The training data is in the text file `data.csv`. You can call the PySpark function `kernel(u)`, which implements a kernel function that satisfies the conditions $(*)$ mentioned above.

To get full points you need to comment your code.