Information page for written examinations at Linköping University

Examination date	2017-03-01
Room (1)	TER1(19)
Time	8-12
Course code	732A54
Exam code	TENT
Course name **Exam name**	Big Data Analytics (Analys av Big data) Examination (Tentamen)
Department	IDA
Number of questions in the examination	13
Teacher responsible/contact person during the exam time	Christoph Kessler, Jose Pena, Valentina Ivanova (Patrick Lambrix)
Contact number during the exam time	2406 / 1651 / 2605
Visit to the examination room approximately	10:00
Name and contact details to the course administrator (name + phone nr + mail)	Annelie Almquist, 2934, annelie.almquist@liu.se
Equipment permitted	dictionary
Other important information	
Number of exams in the bag	

Exam

732A54 Big Data Analytics

March 1, 2017, 8-12

Grades: For a pass grade you need to obtain 50% of the total points.

Instructions:

In addition to the instructions on the cover page:

- Write clearly.
- Start the answers to a question on a new page.
- If you make assumptions that are not given in a question, then clearly describe these assumptions. (Of course, these assumptions cannot change the exercise.)
- Give relevant answers to the questions. Points can be deducted for answers that are not answers to the question.
- Answer in English.

Question 1 (2p)

Describe the graph data model and explain what type of applications it is good for.

Question 2 (2p)

What is HDFS and what is its role in the Big Data Analytics stack?

Question 3 (2p)

Explain the ACID properties.

Question 4 (2p)

P1, P2 and P3 are three distributed processes. The events following below have occurred during the processes and the values for their vector clocks are given:

```
P1: A (0,0,0); B (1,0,0); C (2,0,0); D (3,0,0); E (4,0,2)
P2: F (0,0,0); G (1,1,0); H (2,2,0); I (2,3,3)
P3: J (0,0,0); K (0,0,1); L (0,0,2); M (0,0,3)
```

Name the relationships between the following two pairs of events and explain (with the help of the respective formal rules) how you have determined them:
- B (1,0,0) and I (2,3,3)
- L (0,0,2) and E (4,0,2)

Question 5 (2p)

What is the Dymano data model and what are the type of applications is it ***NOT*** good for?

Question 6 (2p)

Describe the memory organization of a modern cluster computer (Hint: shared? distributed? ...?). In particular, explain which memory can be directly accessed with load/store instructions from a core and from a node, respectively, and how other memory can be accessed if necessary.

Question 7 (1p)

Assume you should run a huge analytics job on a cluster with thousands of nodes, with considerable communication load on the network. Name and shortly describe one *interconnection network topology* for clusters that would be appropriate in this scenario, and why.

Question 8 (4p)

Describe the MapReduce programming model. Be thorough! For example, what do the different internal steps of a MapReduce call do? What kind of data items are input and output of each of these steps? Which steps can be parameterized by user code, and how? How is parallelism used in each step? Where does I/O happen, and where happens non-local communication over the interconnection network?

Question 9 (1p)

What is a *Resilient Distributed Dataset (RDD)* in *Spark*?

Question 10 (1p)

For iterative machine learning jobs, Spark usually outperforms MapReduce by an order of magnitude. Explain why.

Question 11 (1p)

What do systems such as YARN and Mesos do?

Question 12 (5p)

Implement in Spark (PySpark) the moving window classifier. This is a classifier that only considers the training points that are at most at a Euclidean distance h from the point to classify and, then, issues the majority class label. In other words, in a two class domain, the class label assigned to a point $\boldsymbol{x}$ is given by

$$y(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \sum_{n=1}^{N} 1_{\{t_n=1, \boldsymbol{x}_n \in S(\boldsymbol{x},h)\}} \leq \sum_{n=1}^{N} 1_{\{t_n=0, \boldsymbol{x}_n \in S(\boldsymbol{x},h)\}} \\ 1 & \text{otherwise} \end{cases}$$

where $\{\boldsymbol{x}_n, t_n\}_{n=1}^{N}$ is the training data, t_n is the class label for the n-th training point, $S(\boldsymbol{x}, h)$ is a D-dimensional closed ball of radius h centered at $\boldsymbol{x}$, and the function $1_{\{condition\}}$ returns 1 if the condition is satisfied and 0 otherwise.

Question 13 (5p)

Implement in Spark (PySpark) the following k-means algorithm.

```
1  Assign each point to a cluster at random
2  Compute the cluster centroids as the averages of the points assigned to each cluster
3  Repeat the following lines l times
4     Assign each point to the cluster with the closest centroid
5     Update the centroids as the averages of the points assigned to each cluster
```

You can use the functions `randint(A,B)` which produces a random integer in the given interval, and `distance(A,B)` which returns the distance between two points.