Exam

TDDE31 and 732A54
Big Data Analytics

November 2, 2018, 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)

Give and explain 4 V's (big data properties) and give an example for each.

Question 2 (2p)

Give and explain the CAP theorem. Explain the notions in the CAP theorem.

Question 3 (3p)

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)
```

Draw the temporal relationships between the events of the processes. 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 K (0,0,1)
- I (2,3,3) and E (4,0,2).

Question 4 (3p)

(a) Explain how consistent hashing for replication works (including the idea, structure, what happens when a machine fails and when a machine is added). (2p)
(b) Exemplify what happens when a machine fails for the following ring: A $\rightarrow$ 2 $\rightarrow$ B $\rightarrow$ 3 $\rightarrow$ C $\rightarrow$ 4 $\rightarrow$ 1 $\rightarrow$ A (draw the ring for clarity), where A, B, C are machines and 1, 2, 3 and 4 are hash values. Assume that machine C fails. (1p)

Question 5 (3p)

Cluster computing
(a) How does a distributed file (in a distributed file system like HDFS) differ from a traditional file, and what is the advantage for the processing of a big-data computations over a distributed file compared to a traditional file? (1p)
(b) Why is it important to consider (operand) data locality when scheduling tasks (e.g., mapper tasks of a MapReduce program) to the nodes of a cluster? (1p)
(c) Define the following technical terms:
(i) (Parallel) *work* performed by a (parallel) algorithm (0.5p)
(ii) *Data parallelism* (0.5p)

Question 6 (6p)

MapReduce and Spark
(a) What (mathematical) properties do functions need to fulfill that are to be used in *Combine* or *Reduce* steps of MapReduce, and why? (1p)
(b) Which steps of MapReduce involve disk I/O, and for what purpose? (1p)
(c) Describe the fault tolerance mechanisms in MapReduce. (1p)
(d) How can the MapReduce user control (i) the grain size of work and (ii) the degree of parallelism? (1p)
(e) Spark classifies its functions on RDDs into two main categories: "Transformations" and "Actions". Describe the main difference between those, and give one example operation for each category. (1p)
(f) How can Spark be used with input data that arrives in a continuous stream (e.g., from external sensors over the network)? In particular, how can such a data stream be structured for processing by Spark? (1p)

Question 7 (1p)

Cluster Resource Management Systems
Which main benefits do systems such as YARN and Mesos bring to the owner/operator of a cluster computer in a data center? Explain your answer.

Question 8 (5 p)

Implement in Spark (PySpark) logistic regression. Recall from the lectures that we consider a binary classification problem with class labels $t \in \{-1, +1\}$ and a model $y(x) = w^T x$ so that the posterior class distribution is

$$p(t = +1|x) = \sigma(y(x)) = \frac{1}{1 + \exp(-y(x))}$$

$$p(t = -1|x) = 1 - \sigma(y(x)) = \frac{1}{1 + \exp(y(x))}.$$

Thus, given some training data $\{(x_1, t_1), \ldots, (x_N, t_N)\}$, the negative log-likelihood becomes

$$L(w) = \sum_{n=1}^{N} \log(1 + \exp(-t_n y(x_n)))$$

whose gradient is given by

$$-\sum_{n=1}^{N} t_n(1 - 1/(1 + \exp(-t_n w^T x_n)))x_n.$$

Specifically, you are asked to implement gradient descent to find the maximum likelihood estimates of w.

Question 9 (5 p)

Cross-validation is a technique to estimate the error of a classifier. It works as follows:

1	Split the training data into K folds of roughly equal size
2	For i in $1:K$
3	Train the classifier on all the folds but the i-th
4	Test the classifier on the i-th fold
5	Report the average of the K test errors

You are asked to implement in Spark (PySpark) cross-validation to estimate the error of the classifier built in the previous question. To run the classifier, simply call the function $LR(x, i)$ where x is the point to classify and i is the index of the fold not used to train the classifier, i.e. the index of the fold to which the point x belongs.