TIE-22306 Data-Intensive Programming

Exam 16.10.2015
Timo Aaltonen

It is forbidden to use any written material such as cheatsheets, lecture notes etc. besides an English dictionary. Electrical devices (calculators, cell phones, computers, etc.) may not be used during the exam.

Make sure you have answered all questions.
Answer shortly and clearly – the answers are not graded based on their length. Incorrect answers don't normally reduce points. However, the examiner reserves the right to make point reductions, if answer is completely irrational or in clear contradiction with itself, i.e. clearly a guess

1. Explain shortly the following concepts or terms
   i) three Vs of big data (3 p)
   ii) design principles of Hadoop (4 p)
   iii) Combiner (2 p)
   iv) Shared-nothing architecture (2 p)

-hardware fails, manage it
-name nodes, data nodes
-hide the structure from user
-map-reduce
-data storaged as blocks of same size
- tasks are run independently...

2. Explain how a HDFS client reads data from the Hadoop Distributed File System. Illustrate the architecture of the system and include the communication between entities into the figure. (7 p)

3. Below is a table of the data regarding the electrical consumption of an organization.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2012	12	10	10	11	6	5	5	6	8	10	13	14
2013	11	10	12	13	7	9	8	7	9	11	13	13
2014	10	13	11	9	8	8	7	7	10	12	15	16
2015	9	11	11	12	8	5	6	6	8	13	16	17

Sketch a Hadoop application for calculating maximum monthly electricity usage for each year (14 for 2012, 13 for 2013, and so on). Cover all phases from the data preparation to the final output. (10 p)

4. Are the following Hadoop-related claims *true* or *false*? Grading: correct answer gives +1 p, no answer 0 p , wrong answer -1 p. (7 p)

i) Map and reduce can happen in parallel.
ii) FileInputFormat splits the file by newline (i.e. it is line-oriented format).
iii) Pig Latin provides standard relational transforms.
iv) Hive offers a relational view to HDFS files.
v) Checksum can be used for fixing corrupted data.
vi) XML is suitable format for Hadoop applications.
vii) Compressed input files are newer splittable.