Exercise set 1

Return your solutions via Moodle page **no later than 22.03.2017 by 23:59 strictly**. We do not count belated returnings. You have to use .pdf file type for any written answers, and .scala for your code. Do not return any project or object files, etc. Mark your **full name and student number** clearly to all your solution files. Prepare to explain and discuss your solutions on Friday exercise session. Each exercise will be graded pass/fail.

Exercises 1-3 are warm ups to the course. There can be more than only one correct answers. Feel free to use references, figures, etc. in your solutions.

Exercise 1

a. What does it mean when we discuss Big Data? What makes it “Big”? Give some explanations you can find from literature/Internet as well as your own opinion as a short essay. Use references where needed.
b. Describe a practical use case or application for each of these datasets:
    i. All flights by commercial airlines from year 2014 (*Addition:* You can assume that the data have all the departures and arrivals, with times and airlines, from all the airports.)
    ii. Applications and system settings of 750.000 mobile devices from three years
    iii. Entire English Wikipedia text dump with editing history

Exercise 2

Describe a data mining / machine learning algorithm of your choice, e.g. from courses you have passed in your earlier studies (Introduction to Machine Learning, Tietorakenteet ja Algoritmit, etc.). Give an example, for which purpose your algorithm works, and what kind of problems you could face when implementing it in the distributed environment. You are not supposed to implement anything, but discuss the problems and possible solutions. Also mention that what kind of dataset you would like to have and whether you would like to have some project on it or not (Don’t worry about this special project).

Exercise 3

Describe problems and possible solutions you can face when managing, storing, and analyzing:

a. A 10TB text data set in the cloud, where each line represents one element of the data and each file contains from 10.000 to 100.000 lines
b. A data stream of 1000 elements per second, without a need to store everything but collect some useful information

Exercise 4:

The purpose of this exercise is to have pre-understanding of some Spark design primitives. You will read the following paper and write short notes on the following terms. You should use only seven sentences to explain these terms.

https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final138.pdf

(1) Resilient Distributed Datasets (RDDs).
(2) Advantages of RDD over Distributed Shared Memory Model.
(3) Directed Acyclic Graphs (DAGs).
(4) Checkpointing.
(5) Fault-tolerance.
(6) IndexedRDD.