Techniques for large-scale data (DIT871/DAT345)

CSE, Gothenburg University | Chalmers, Period 4, 2018
Graham Kemp, Alexander Schliep (alexander.schliep@cse.gu.se)

Problem set 4 from May 21, 2018 · Due on May 25, 2018

Problem 1 (3pt)**.** Implement a solution for the histogram computation, part (a) of Problem 1 from Problem set 2, using PySpark. You may use the example solution `mrjob-summary-statistics-hist.py` as a starting point. See also `spark-duplicates.py` for motivation.

Hint: It is not necessary to use `lambda`*-expressions. Defining a function (e.g.,* `def myFun(...)` *makes more sense for more complex operations*

Problem 2 (2pt)**.** A Map-Reduce job for computing summary statistics runs in 11 min for a 18 TB data set on your cluster. Detailed benchmarking shows that 73% of the time is spent on reading data from HDFS. Reads are in parallel.

Solid state drives (SDD) are 4 times faster reading data from HDFS than the hard disk drives (HDD) currently in the cluster.

Questions:

a. What is the total speed-up you can achieve by buying SDD for all nodes in the cluster?

b. How many additional nodes with HDDs would you need to add to the cluster to achieve the same speed-up?

Submission: Please submit a PDF with your answers to the questions, and, additionally, Python code. Jupyter notebooks are only accepted in addition to a PDF. We will not accept files in other formats.