Techniques for Large Scale Data - Group 23

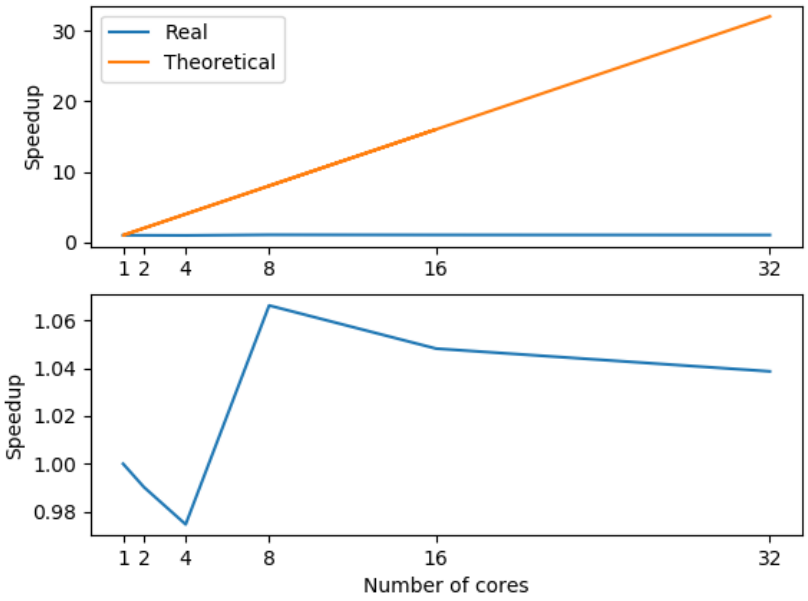
Julia Szulc - juliasz@student.chalmers.se
Ahmed Groshar - gusgroah@student.gu.se

May 4, 2020

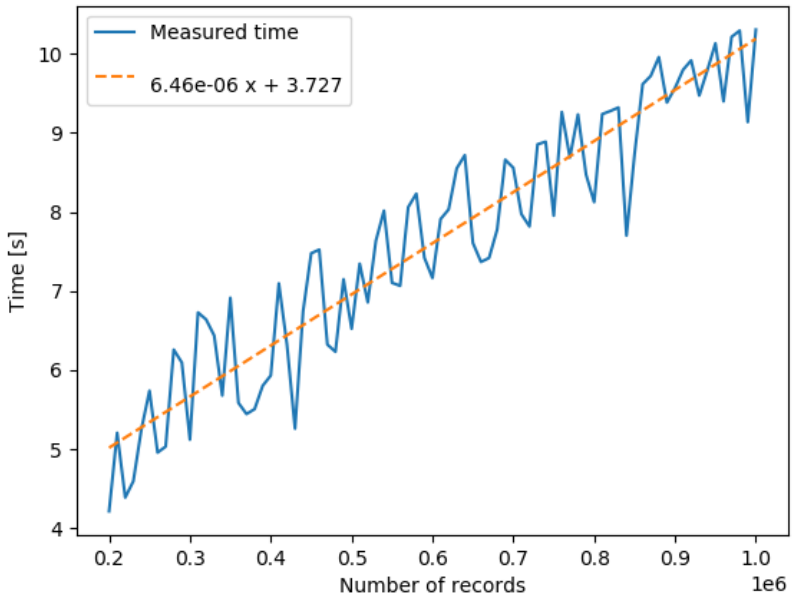
Assignment 3

Problem 1

b)



c)



Problem 3

a)

The data we are interested in is in only 0.667391% of the data set which means that the vast majority of the records is not relevant. What's more, over 90% of the set consists of values that appear only once. This leads to the fact that the reducer part still needs to be called over 900 million times. Paralleling the task over 16 nodes should visibly improve overall performance as all the non-complicated parts of the algorithm can always be split. Increasing the number of nodes should only improve this effect.