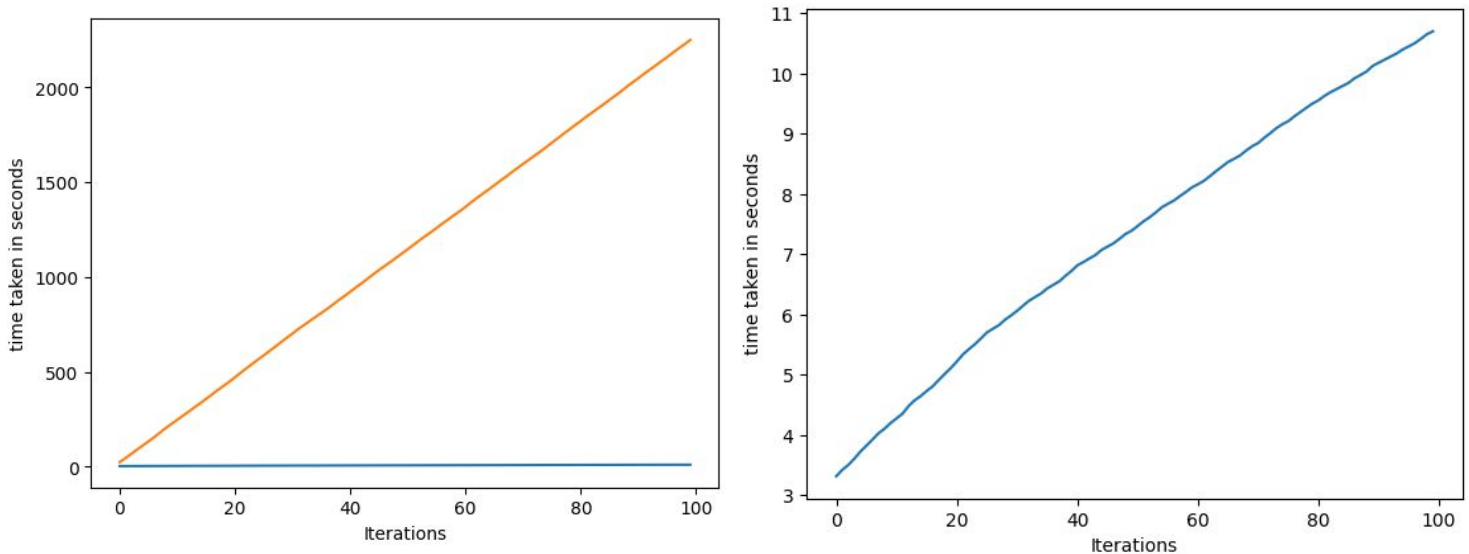


Distributed Systems

Assignment 5 Technical Report

Gautam Jagdish.M - 170010031



On the left, we have Hadoop(orange) vs Spark(blue). We can observe that Hadoop takes 2250 seconds to finish 100 iterations whereas Spark takes only 11 seconds. On the right, we have the same graph, but only for spark so that we can see the difference in scale. We see this observation because of Hadoop scheduling jobs on the worker nodes for each iteration. At the end of each iteration, the data on the nodes are deleted. On spark, we essentially create a session in which the data is cached in memory across different nodes, so the map-reduce can be done without erasing memory. The speedup is because of deleting data from the nodes after each iteration.