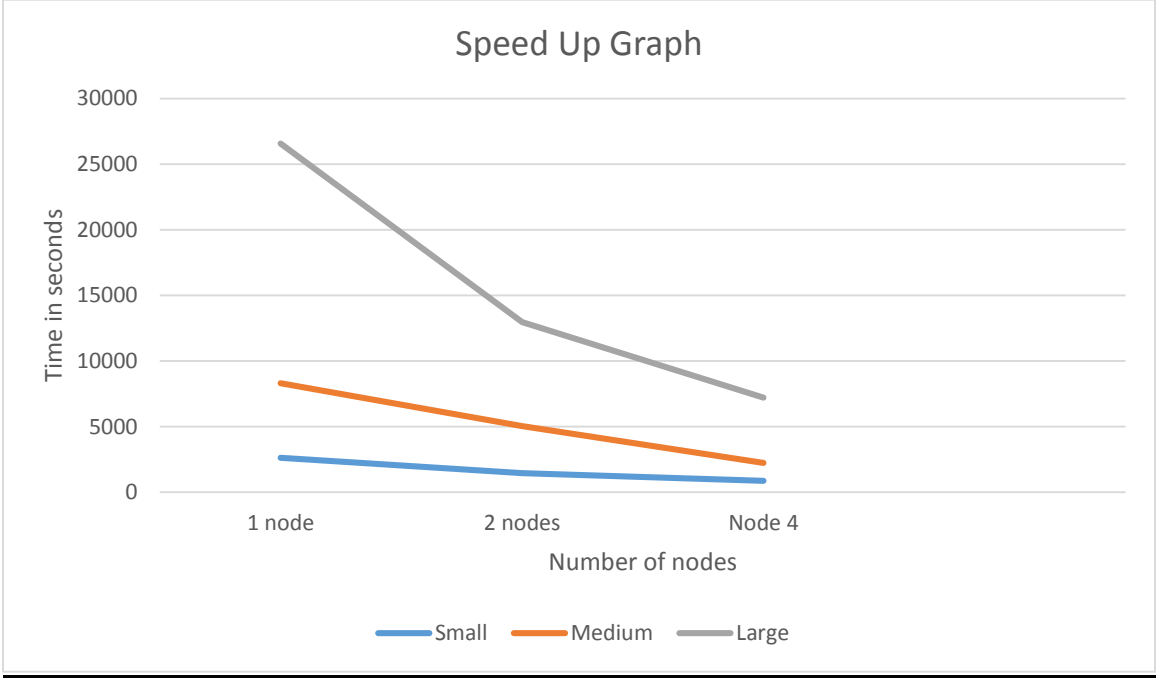
CSE 587

ASSIGNMENT1 REPORT

The below table provides the run time for the three datasets overs the CCR cluster with1, 2 and 4 nodes respectively.

Node scaling

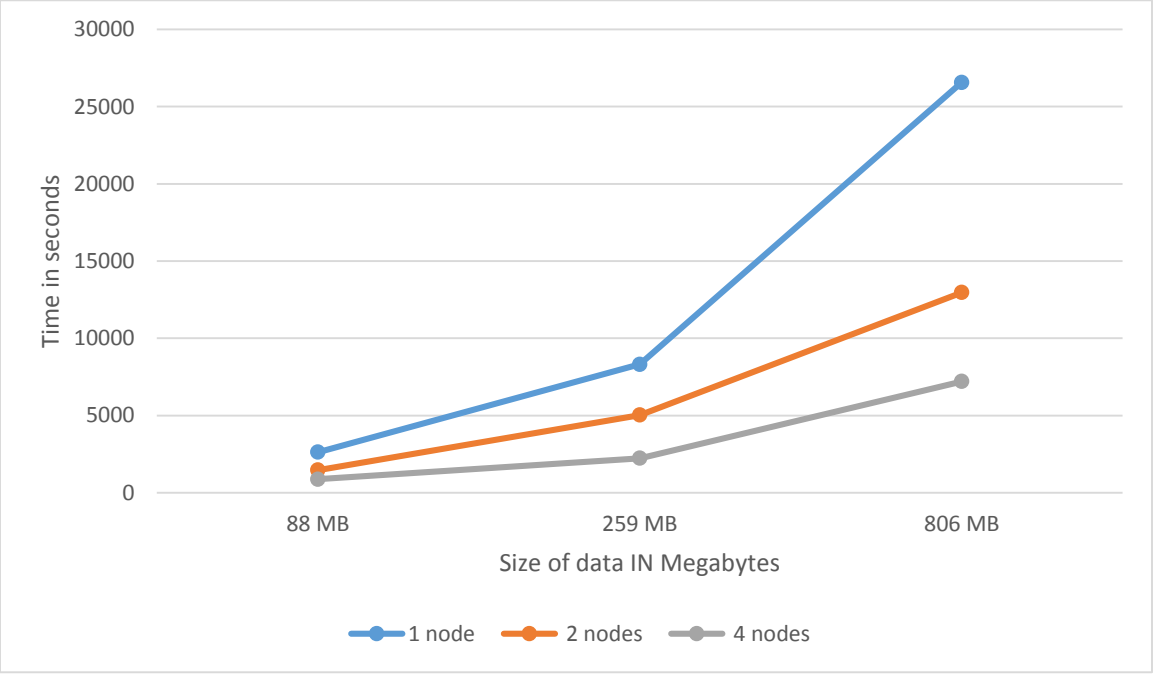
	1 NODE(12 cores)	2 NODES(24 cores)	4 NODES(48 cores)
Small dataset	2623 seconds	1461 seconds	875 seconds
Medium dataset	8313 seconds	5032 seconds	2231 seconds
Large dataset	26568 seconds	12967 seconds	7199 seconds



In the above graph, a plot of number of execution time vs number of nodes is made for different data sets. As the number of nodes is doubled, the execution time is approximately halved. Maximum time is taken for file I/O operations and hence it forms the bottle neck for performance. However by scaling out, the load is divided and hence performance is improved as the files are retrieved in a distributed fashion and processed simultaneously.

Dependency of Execution time on the Data Size

The size of the data in the correspond to the size of the data for small, medium and large datasets given in the assignment



The above graph shows us how the execution time is dependent on size of the data for different node configurations. We can observe that for a single node, as the size of the data increases, the execution time of the increases drastically .When data is processed by scaling out, the size of the data has lesser impact on the execution time.