CSE587 Data Intensive Computing

Report of Programming Assignment 1

Name : Sagar Bansilal Shinde
UBIT Name: sshinde4
UB Person Number: 50134022

The execution time in different cases is as follows:

Problem size	Execution time: 1 node (12 cores)	Execution time: 2 node (24 cores)	Execution time: 4 node (48 cores)
Small	00:42:00	00:21:30	00:12:45
Medium	02:08:50	01:02:35	00:35:08
Large	06:57:16	03:07:45	01:48:22

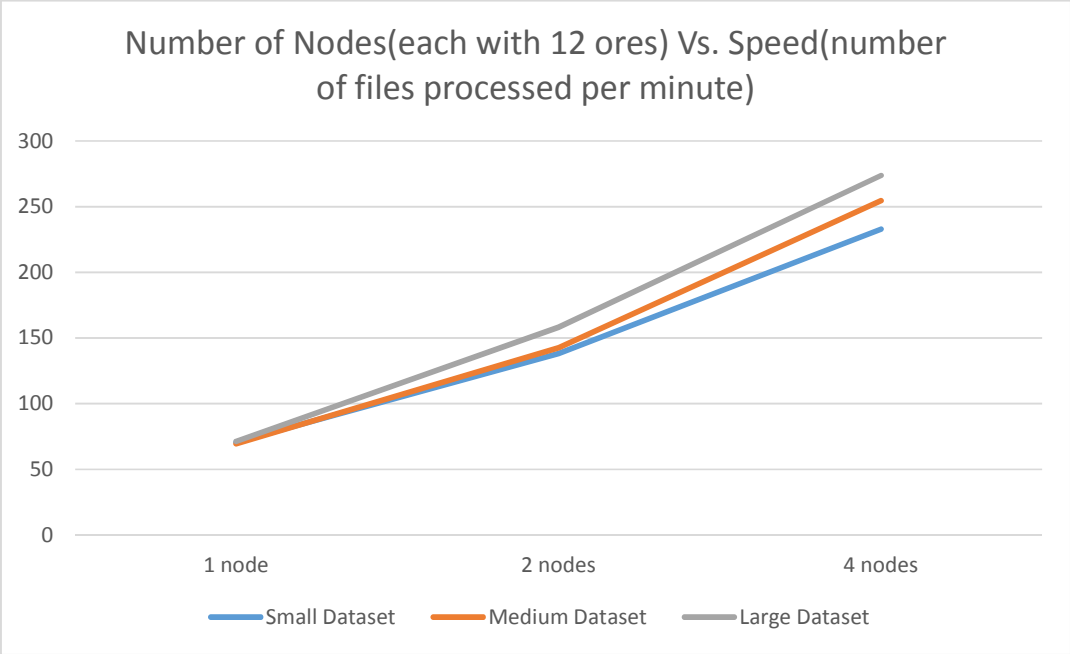
To calculate speed, we will use the formula:

Speed = number of files processed/Time taken in minutes

For small dataset,
Speed on 1 node = 2970/42 = 70.71 files/minute
Speed on 2 nodes= 2970/21.5 = 138.14 files/minute
Speed on 4 nodes= 2970/12.75= 232.94 files/minute

For medium dataset,
Speed on 1 node = 8910/128.5= 69.33 files/minute
Speed on 2 nodes= 8910/62.5 = 142.56 files/minute
Speed on 4 nodes= 8910/35 = 254.57 files/minute

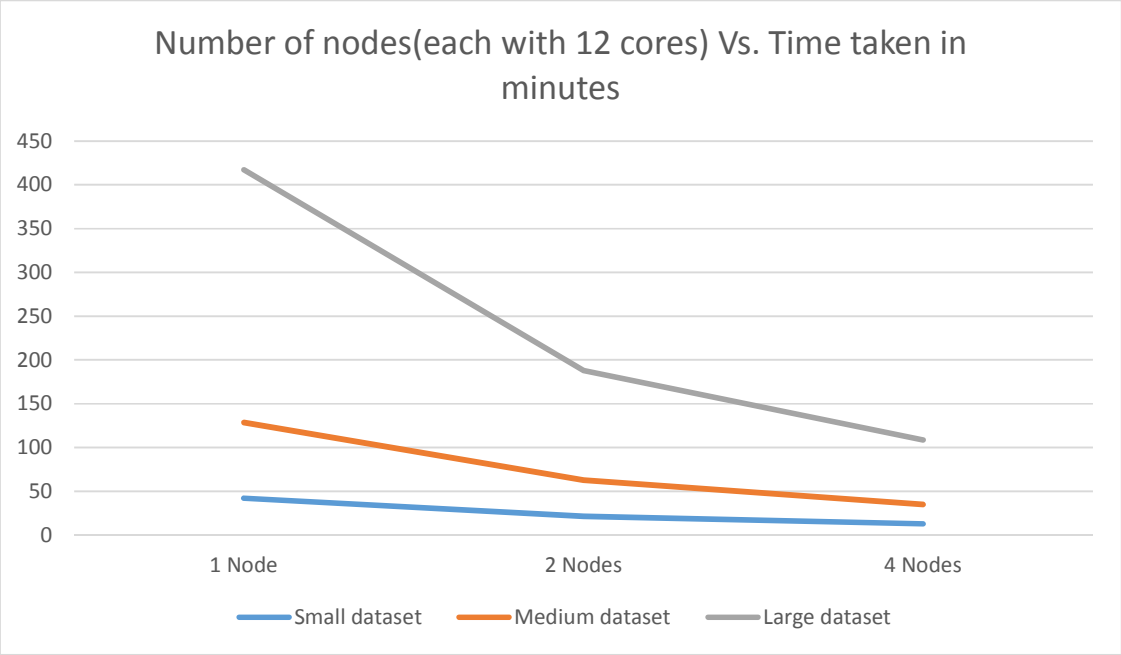
For large dataset,
Speed on 1 node = 29701/417 = 71.22 files/minute
Speed on 2 nodes= 29701/187.75= 158.19 files/minute
Speed on 4 nodes= 29701/108.5 = 273.74 files/minute



Discussion:
As seen from the above graph, the speed of the Hadoop mapreduce program depends the number of cores it is running on and the number of input files. The speed of the program for large dataset is more than the speed for medium and small dataset. The speed of the Hadoop mapreduce program for medium dataset is more than the speed for small dataset. The speed for small dataset is the less than the speed of both large dataset and medium dataset.

When we run the program on a single node with 12 cores, the speed for the three datasets is almost the same. As the number of nodes increases, the speed of the larger data becomes more than the speed of the smaller data. This shows that Hadoop mapreduce programs works well for larger data compared to the smaller data. So Hadoop programs are mainly useful for large amount of data and they can process that data very fast.

We can also see the performance of the system under different number of cores on the basis of time the system has taken. Such a graph is shown below. The graph indicates that the time required reduces which is shown with the declining graph and this decline is most with large data than medium and small dataset. This graph also proves that Hadoop mapreduce programs are effective for larger datasets.


Number of nodes(each with 12 cores) Vs. Time taken in minutes
450
400
350
300
250
200
150
100
50
0
1 Node
2 Nodes
4 Nodes
Small dataset
Medium dataset
Large dataset