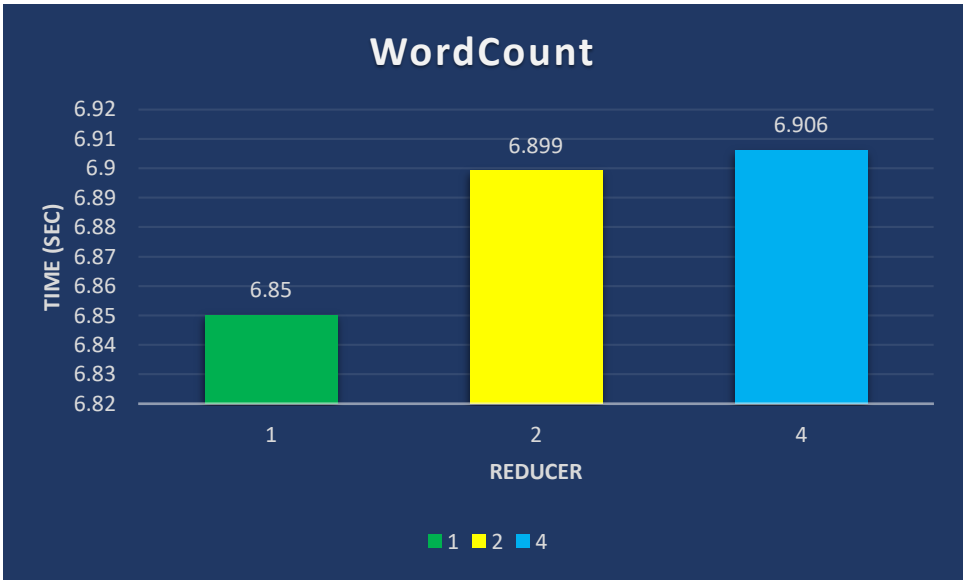
//JATIN LAMBA //A20362075

Times:

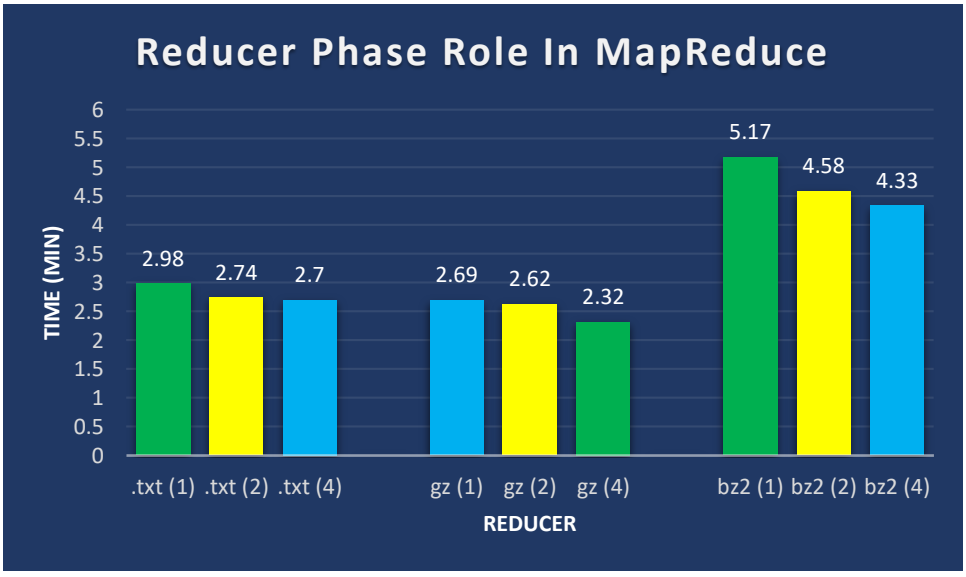
Task 1:



Ans: The execution times increase as you add more reducers in this case because as the file size plays a major role during the MapReduce job. As from the above timing obtained with the increase in number of reducer is getting increased because the size of data is less than 20GB. Reduces come into play when the data is generally big data in more Gigabytes and reduce the time. Hence in this case the reducer doesn't play much role and increased time is obtained with the increase in the number of reducers.

The & **of** are the 2 words with the most number of occurrences for wordcount 1 and wordcount 2 examples except the patterns.txt file.

Task 2:



//JATIN LAMBA //A20362075

Ans: In the above graph we can say that the compressed gz format has a great advantage over basic text files and bzip2 files. We can see that bzip2 files took the maximum top to complete the job even by using the increase number of reducers though it compresses the file efficiently but it is a slow process. Whereas the gz file took less time for the job when compared to bzip2 but gzip does not support splitting as bzip2 compression so it is one of the reason for less timing of job. Increasing the number of reducers only makes the reduce phase go shorter, since we get more of parallelism. However, if we take this too far, we can have end up with lots of small files, which is suboptimal.