CMPT 732 – Assignment 1 Ibrahim Ali

1. How did the output change when you submitted with -D mapreduce.job.reduces=3? Why would this be necessary if your job produced large output sets?

The output went from having one single output file, with mapreduce.job.reduces=3 had three output files, part-r-00000, part-r-00001, and part-r-00002

Main benefit of using multiple reducers is performance improvement for large data sets by distributing the workload across reducers. This is accomplished by speeding up the job and distributing the load across different nodes in the cluster. Additionally, each reducer gets a partition of the data, which prevents memory overload on a single node and ensures that the workload is evenly distributed.

1. How was the -D mapreduce.job.reduces=0 output different?

The output is directly written by the mapper, with each mapper writing a sperate file. When running mapreduce.job.reduces=0 , the output had 19 files, one for each input in directory wordcount-1.

1. Was there any noticeable difference in the running time of your RedditAverage with and without the combiner optimization?

No there was not any noticeable difference in the running time, this is due to the sample data being small. However, with bigdata there will be a noticeable difference. Without combiners, it can take a longer time due to the data being shuffled across the cluster, in comparison to with having a combiner, the data operates locally, removing the need for data to talk on the network.