CSC 555 Assignment 4
Sarah Cummings

1) Input data file of size equal to 25 disk blocks, with replication factor set to one. The mapper in this job requires 2 minutes to read and fully process a single block of data, and reducer takes two seconds per key there are a total of 500 distinct keys.

   a. How long will it take to complete the job if you only had one Hadoop worker node?

      Total of 50 minutes for the mapping (2 mins * 25 blocks) and 1000 seconds for the reducing (500 keys * 2 seconds)

      Answer: Approximately 66 and 2/3 minutes— 1 hour six minutes and 40 seconds.

   b.10 hadoop worker nodes

      First, the ten nodes each process two blocks. This takes four minutes for 20 of the blocks to be processes. There are five blocks left which can all be processed at once in 2 minutes. Total of six minutes for the mapping.

      Then, the ten worker nodes must each run the reducing task on 50 keys, for the total of 500 keys among the ten nodes. Two seconds for each of the 50 keys a node processes gives us 100 seconds, or total of 1 minute 40 seconds for the reducing.

      Answer: 7 minutes 40 seconds.

   c. 50 hadoop worker nodes

      The mapping will only take the time of one node being processed because between the 50 nodes, all 25 blocks can be processed at once. Total of two minutes for mapping.

      Then the worker nodes each must run the reducing task for 10 of the 500 keys. Total of 20 seconds for mapping (10 keys * 2 seconds).

      Answer 2 minutes 20 seconds.

   d. Would the introduction of a combiner affect this job?

      No. Since we already know that the 500 keys are distinct, the combiner would have no affect on this job. The combiner gets rid of repeated keys to streamline the process, but all of our keys are already distinct.

2) Write and execute MapReduce python code that will count the frequencies for all individual characters (i.e., character count, similarly to word count). Run it against http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/big.txt

```
import urllib.request as urllib

file=  urllib.urlopen("http://rasinsrv07.cstcis.cti.depaul.edu/CSC555/big.txt")

def letterFreq(file):
    letterdict={}
    for letter in file:
        letterdict[letter] = 0
    for letter in file:
        letterdict[letter] += 1
    return letterdict
```

---

wget

3) Set up a 3 node cluster