We have a cluster of one namenode and 10 datanodes all in one rack. Each one of the data nodes have a disk of 10 terabytes. HDFS is configured with the default replication factor of 3 and block size of 128 MB.

1. What is the total capacity of this setup of HDFS in terabytes? In other words, how much data can this HDFS cluster store?
   Non-HDFS size = 10 TB * 10 Machines = 100 TB
   HDFS size = 100 TB / 3 replication factor = 33.33 TB

2. What is the storage overhead of this HDFS configuration?
   Overhead = replication factor – 1 = 2

3. If we upload a file of size 2GB from a driver node that is not one of the data nodes, how much is the total network IO (incurred on all the machines) required to upload the data file in gigabytes? Note: Do not double count the network IO as output from the sender and input to the recipient. Only count it once.
   2 GB * 3 replicas = 6 GB of network IO

   For the same case mentioned above, how much of the network IO (in gigabytes) is incurred on the driver node that writes the file?
   2 GB
   How much network IO (in gigabytes) is required to download the file back from HDFS to the **namenode**?
   2GB

6. How much is the estimated network IO (in gigabytes) to download the 2GB file back from HDFS to one of the **data nodes**?
   2GB = 16 blocks
   Total number of replicas = 16 * 3 = 48 block replicas
   Assuming balanced load, number of local replicas = 48 / 10 = 4.8 blocks
   Number of remote replicas = 16 – 4.8 = 11.2 blocks
   Total remote size = 11.2 * 128 MB = 1433.6 MB = 1.4 GB