Assignment 3.2

1.If 7TB is the available disk space per node (9 disks with 1 TB, 2 disk for operating system etc. were excluded.). Assuming initial data size is 600 TB. How will you estimate the number of data nodes (n)?

C – compression

R – replication factor

S – initial size of data moved to hadoop

I – intermediate data factor( usually 1/3 or ¼)

H – hdfs node storage

H = C \* R \* S (1 – 1/4)

If there is no compression c = 1, replication factor is 3 and S = 600 TB then

H = 1 \* 3 \* S (3/4)

H = 4 \* S

H = 4 \* 600 = 2400

Number of data nodes n = H / d

d = disk space available per node (7 TB)

n = 2400 / 7

n = 342.85

n = 343.

The number of data node is 343 or 342.

2. Imagine that you are uploading a file of 500MB into HDFS.100MB of data is successfully uploaded into HDFS and another client wants to read the uploaded data while the upload is still in progress. What will happen in such a scenario, will the 100 MB of data that is uploaded will it be displayed?

In the hadoop 1.x block size is 64MB. So the 1st block is written to the hdfs , this block is visible to the client but the second block writing is still running so it’s not visible to the client.

In hadoop 2.x the block size is 128MB ,So here the first block itself is not visible to the client because it’s still running.

Once the data is written the blocks will be visible. The running process are not visible to the client.