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)?

N= no of data nodes needed

H = Hadoop Storage

D = disk space per node

**H=c\*r\*S/(1-i)**

c = average compression ratio

r = replication factor

S = size of data to be moved to Hadoop

 i = Intermediate factor

**So considering the ideal conditions like replication factor to be 3 , average compression ratio as 1, intermediate factor as ¼. We get H = 4\*S.**

Hence in our problem **4\*600= 2400.**

N= H/D = 2400TB/7TB= **342.85 ≈ 343** nodes will be needed approximately.

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?

Answer –

Here as we know that default block size of each block is 128mb in Hadoop 2.x . Explaining the concept with respect to Hadoop 2.x still 28 mb of data can be written into the block. So hdfs will continue to write the next 28MB. Simultaneously replication of the data is also taking place. Suppose that if the block has completed 128 MB of writing the data then it goes on writing the next block. That time one can read the already written 128mb of data from another machine. Simultaneous read and write of the same data is not possible.