**QN-1.1.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)?**

This is a formula to estimate Hadoop storage (H): H=crS/(1-i)

where: c = average compression ratio. It depends on the type of compression used (Snappy, LZOP, ...) and size of the data. When no compression is used, c=1. r = replication factor. It is usually 3 in a production cluster. S = size of data to be moved to Hadoop. This could be a combination of historical data and incremental data. The incremental data can be daily for example and projected over a period of time (3 years for example). i = intermediate factor. It is usually 1/3 or 1/4. Hadoop's working space dedicated to storing intermediate results of Map phase Data Size – 600 TB Replication factor – 3 Intermediate data – 1/4 Total Storage requirement – 13600/(1-1/4) = 2400 TB Available disk size for storage – 7 TB Total number of required data nodes (approx.): 2400/7 = 350 machines

Actual Calculation(Including compression ratio and disk space utilisation) Disk space utilization factor– 65 % (differ business to business) Compression ratio – 2.3 Total Storage requirement – 2400/2.3 = 1043.5 TB Available disk size for storage – 70.65 = 4.55 TB Total number of required data nodes (approx.): 1043.5/4.55 = 230 machines Actual usable cluster size (100 %): (2307\*2.3)/4 = 1058 TB QN-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? 1.We assume that we are using a Hadoop 2.x and we have configured the block size to be 100 MB and replication factor of 3 2.So we have 5 blocks 1,2,3,4,5. 2.First client will ask Namenode and namenode after checking permission give the datanode where first bock(100Mb) is written and Datanode simultaneously makes replicas of blocks after getting permission from namenode. 3. Once the block is copied and replicated to the datanodes, client will get the confirmation about the Block 1 storage and then, it will initiate the same process for next block “Block 2”.Now block 2 will be write on DataNode as per permission given by namenode 4.So, during this process if 1st block of 100 MB is written to HDFS and the next block has been started by the client to store then 1st block will be visible to readers. Only the current block being written will not be visible by the readers.