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 the formula to estimate the number of data nodes (n):

**n= H/d**

where, d= disk space available per node

H= Hadoop storage

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

Hence final formula to calculate number of data nodes is

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

where d= disk space available per node

c = average compression ratio. It depends on the type of compression used and size of the data. When no compression is used then c=1.

r = replication factor. It is usually 3 in a production cluster. S = size of data to be moved to Hadoop.

i = intermediate factor. It is usually 1/3 or 1/4. Hadoop's working space dedicated to storing intermediate results of Map phases.

Consider **c=1; r=3; i=1/4; S=600TB; d=7TB**

**Total no of nodes = 1\*3\*600/((1-1/4)\*7)**

**= 342.85**

**= 343**

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 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