Q1-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)?

ANSWER:

The formula for calculating HDFS storage(H) is :

H=C\*R\*S/(1-i)\*120

C = Compression ratio. here C=1 as nothing is mentioned with respect to compression.

R = Replication factor. The default value of R is 3.Thus, R=3.

S=initial size of data. S=600 TB

i=intermediate data factor. This space is required to store results of mapping tasks and any other processing. Cloudera has recommended a 25 % for intermediate data.

i=1/4

120 % : some space ie. about 20% should be reserved for file system underlying the HDFS.Generally, it will be ext3 or ext4.Here, it is already mentioned that 2 disks are used for os.Therefore, this factor can be ignored in the calaculation.

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

H=4\*S

The number of data nodes(n) can be calculated by the following formula:

n=H/d

d=available disk space available per node(d=7 TB)

n=4\*S/d

n=4\*600/7

n=342.85

n=343

The number of data nodes will be 343.

Q2-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:

1. The default size of block in HDFS is 128 MB.

2. In the given scenario, the number of blocks will be :

N=500/128 = 4

3. As per the given scenario, when 100 MB of data is uploaded, HDFS will continue uploading the data till the entire block of 128 MB is uploaded.

4. Simultaneous read and write operations on same data is not permitted as it can compromise data integrity.

5.Therefore the data will not be displayed.