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)?

ANS:

We have the formula for calculating HDFS storage(H) as :

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

Where,

C is the Compression Ratio.

-If C=1 then nothing is mentioned with respect to compression.

R is the Replication Factor.

-The default value of R is 3.

-So, R=3.

S is the initial size of data.

Given: S=600 TB

i is the Intermediate Data Factor.

-It is used to store results of mapping tasks and any other processing.

Cloudera has recommended 25 % for intermediate data.Thus, i=1/4

Some space (20%) should be reserved for file system underlying the HDFS.

Generally, it will be ext3 or ext4.

It is already mentioned that 2 disks are used for OS.

Therefore, this factor can be ignored in the calculation.

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

H=4\*S

The number of data nodes(n) can be found out by the formula:

n=H/d

d=available disk space per node

So, d=7 (as given)

n=4\*S/d

n=4\*600/7

n=342.85

Therefore,

n=343

So,the number of data nodes(n) 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?

ANS:

-The default size of the block in HDFS is 128 MB.

- So, the number of blocks here in this problem will be :

N= 500/128

N= 4

-Here, when 100 MB of data is uploaded, HDFS will continue to upload the data till the entire block of 128 MB is uploaded.

-Simultaneous read and write operations is not allowed on the same data as it can compromise data integrity.

-Hence, the data will not be displayed.