ASSIGNMENT 3.6

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)?

A) The number of nodes (data nodes) required can be calculated using the formula,

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

c= average compression ratio(c=1)

r= replication factor (usually 3)

S= size of data to be moved

i= intermediate factor (1/3 or 1/4)

as per the given data,

n= no.of nodes present

d= available amount of disk space per node (initialize by 600TB)

H= estimated hadoop storage

n = H/d

= 600/7 = 85

Thus no.of data nodes required = 85 datanodes.

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?

A) Yes , the 100MB of data that is uploaded is displayed successfully .

This block of 100MB is copied and replicated into 2 desirable datanodes.During this scenario of process the 100MB of data is passed on to the HDFS

In this case we are trying to upload 500MB of data out of which 100MB of data is uploaded successfully.Here we have 5 blocks of data that are being replicated for 3 times.According to this case 1 block of data is uploaded and so can be displayed.

|  |
| --- |
|  |
|  |
|  |  |
|  |  |
|  |  |
|  |  |