**Assignment 4.6**

**Problem Statement 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)?

**Solution:**

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)?

The formula to calculate number of data nodes (n) is:

n= H/d

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

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

Where,

H- Hadoop Storage

d- Disk space available per node

c- Average compression ratio

The average compression ratio depends upon the type of compression used and the size of the data to be stored when no compression used c=1. Usually the compression ratio is 2.3

r- Replication factor

The replication factor for most of the Hadoop clusters is 3

S- Size of data to be moved to Hadoop

This data can be consists of historical data as well as incremental data (when data is added dynamically)

i- Intermediate factor

It is Hadoop’s workspace dedicated for temporary storage of intermediate results of map tasks.

Usually it is 1/3 or 1/4.

Assuming,

c=1

r=3

i=1/3

We have,

S=600 TB

d=7

Therefore,

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

H=1\*3\*600/(1-1/3)

H=2700

n=H/d

n=2700/7

n=385.714

Now,

If we consider compression factor=2.3 and disk space utilization =0.65

Then

n=385.714/2.3/0.65

n=258.0026 =259 machines

Total number of data nodes required will be 259 approximately.

**Problem Statement 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?

**Solution:**

100MB data is already uploaded successfully.If the block size is 100MB then next data will be written to the next block. In this case if the next data is being written in next block then the client can read the data which is already uploaded in the 1st block.

The client cannot read the data from the block which is being written.

The blocks which are already written can be read by the client.