Task 1

Initial data Size =600TB

Available disk space per node= 7TB

c = average compression ratio.

r = replication factor. It is usually 3 in a production cluster.

S = size of data to be moved to Hadoop. This could be a combination of historical data and

incremental data. The incremental data can be daily for example and projected over a period of time (1 years for example).

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

d = disk space available per node. All other parameters remain the same as in 1.

The formula to estimate the number of data nodes (n): n= H/d = c\*r\*S/(1-i)\*d

here, H=600TB, d =7TB

therefore,

n=H/d

n=600/7

n=85.71

So, Total Number of data Nodes=86

Task 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?

the default blocks size is 64 MB in Hadoop 1x and 128 MB in Hadoop 2x whereas in such a scenario let

us consider block size to be 100 MB which means that we are going to have 5 blocks replicated 3 times (default replication factor).

Let, We have 5 blocks (P/Q/R/S/T) for a file, a client, a name node and a data node. So, first the client will take Block P and

will approach name node for data node location to store this block and the replicated copies. Once client is aware about the data node

information, it will directly reach out to data node and start copying Block A which will be simultaneously replicated to other 2 data nodes.

Once the block is copied and replicated to the data nodes, client will get the confirmation about the Block P storage and then, it

will initiate the same process for next block “Block Q”.

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.