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:

Number of data nodes: Data size available/disk space per node

=600/7

=85

Number of data nodes=85

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:

\*For this 500 mb data there will be 5 blocks and 3 replicas for every block.

\*The client will take this block (100 mb) of data and will approach namenode for datanode location to store this block and the replicated copies.

\* Once client is aware about the datanode information, it will directly reach out to datanode and start copying this block of data which will be simultaneously replicated to other 2 datanodes.

\*Once the block is copied and replicated to the datanodes, client will get the confirmation about the Block storage and then, it will initiate the same process for next block.

\*So,in this process it will make that 100 mb of data available to the client but it will not display the block of data that it is currently writing.