Your company needs to set up new 1000 warehouses for storing agricultural commodities which will constantly have inflow and outflow of orders and supply. Based on supply and demand your company wants real time price prediction on stored commodities so as to maximize the profit. To do so it is expected to process a lot of data every hour. Data processing will perform a lot of interim transformations. Hence:

1. You have a high CPU system
2. You are constrained on available RAM and persistent storage

Install a 2 node system, which can achieve a file processing SLA for wordcount job as follows:

1. for a single file, 6 GB shall be processed in under 2 minutes.
2. for a 5 file group present under your input directory(meaning 5 files of 6 GB each), 6G MB shall be processed in under 3 minutes.
3. Achieve the same or better througput with 10 files of same size.

While you achieve the above, ensure that MapReduce doesn't take more than 40% of your available hdfs space.

The output of above wordcount jobs will be used once a month. No point keeping hdfs occupied all the while. Find a way to store
the output such that it consumes less space.

Namenode and datanode shall be created under separate directories and must be separate than tmp directory of hadoop.

Sometime in future you are expecting a significant capacity addition to your cluster. Hence to decongest IO propose a solution to frequent datanode block reports which will be received by namenode.