1. A datanode has 50 data blocks. What is the maximum number of map tasks map-reduce can run?
2. A map-reduce job is running 10 map tasks and 5 reduce tasks. Three nodes running three reduce tasks die immediately after finishing the task. How many reduce tasks needs to be redone?
3. The map phase of map-reduce job produces 50 distinct keys. If 100 reducers are run, what is the maximum number of files the map-reduce job is going to produce?
4. How does Hadoop determine a datanode failure?
5. What is the main problem with storing too many small files in Hadoop cluster?
6. What is the correct sequence of data flow in Map-Reduce? Assume M for map, R reduce, P for partition and C for combiner
   a) MRCP, b) MCPR, c) MPCR, d) MCRP
7. For a given data, the mappers produce 1000 distinct keys. How many files the reducers can produce? Choose the correct option.
   a) 1000 b) less than 1000, c) more than 1000, d) it can be more than 1000 or less than 1000 depending upon the number of reducers.

8. A computing cluster has 1024 nodes (servers) where each rack has 8 servers. The cluster is organized hierarchically with switches. The cluster has two types of switches: base switch and the backbone switch. The base switches are for each rack only and there is one base switch for each rack. The backbone switches can connect at most two switches (base or backbone). What is the number of backbone switches the cluster need in the optimal case?

9. A map-reduce (MR) cluster has 1000 datanodes and it stores a 1 GB data with block of size 1 MB each. A map-reduce job on the data recognized that only 50 blocks are being used repeatedly. What is the amount of additional cluster DRAM being used by the MR job? Mention amount of memory usage separately for the datanodes and the master node.

10. A company has two Hadoop datacentre (say d1 and d2) each having two racks (r1 and r2). Each rack has 1000 datanodes. We use the notation i/rj/dk to denote the i-th datanode of the j-the rack of k-th datacentre. Compute the network distance between the following nodes:
   i) 1/r1/d1 and 2/r1/d1
   ii) 5/r1/d1 and 10/r2/d2
   iii) 1/r2/d2 and 2/r2/d1

11. The map task of a map-reduce job produces 100 key-value pairs with 10 distinct keys. Assuming that load-balancing is ensured, what is the maximum number of record that a reduce task has to process in the worst case?

12. What is replica pipeline in Hadoop? How it is created (explain with an example).

13. The reduce phase of a map-reduce job requires access to a read only shared variable of 100 MB. Each data node of the Hadoop cluster has 32 GB DRAM and the block size it operates on is 2 GB. The reduce phase needs the shared data for each key-value aggregation. As a programmer suggest a strategy to minimize data communication during the reduce phase. Also mention the amount of additional cluster DRAM your approach is going to consume.