In practice, a Hadoop cluster consists from tens to thousands of slave nodes. These slave nodes are put in server racks in a data center. Network traffic through different racks is known to be more expensive compared to the traffic within a rack.

HDFS (Hadoop Distributed File System), the storage layer of Hadoop, stores files in the unit of blocks. The default block size is 128MB and it can be configured. A file is broken into blocks and stored in data nodes. By default, each block is replicated to three separate physical machines for fault tolerance. Considering the large number of servers and hard disks used in a Hadoop cluster, hard disk failure is very common. If a block becomes unavailable in a server that data can be copied from a different server. HDFS is designed to select different data nodes in separate racks when data is written to HDFS. The first replica is placed at a data node in the local rack as the client. The second replica is placed at a different rack from the first replica. The last replica is placed at a different datanode in the same rack as the second replica. This means that in one rack there are two replicas of a given block.

When a HDFS client wants to read a file from HDFS, the client contacts the namenode to get the addresses of the datanodes that have a copy of the first few blocks of the file. P.94

When a job runs and data locality is not available, blocks are read from HDFS to the server running that task.

Job scheduler assigns tasks to take the nodes where input data is available. This is called data locality. One data block is replicated on three data nodes by default. Job scheduler assigns a task randomly if all the nodes containing the input data for that task are already busy with other tasks. In the current implementation of Hadoop, the node that was assigned the task gets the data from only the closest node that contains the input data. System administrator predefines the distance between nodes/racks manually. This is called rack awareness in Hadoop. When the task assigned node reads data from the source, it only copies from only one node, where the same data is available in other two nodes. Network distance from the task assigned node for these two nodes might be longer than the nearest node, but it is also possible to utilize these other two nodes to transfer the data needed for the task assigned. Finding the ratio of the data is the problem.