Ans-5

MapReduce is processing framework used to process data over a large no. of machines. Hadoop uses MapReduce to process the data distributed in Hadoop Cluster. MapReduce is not similar to other regular processing framework like Hibernate, JDK, .NET. All these previous frameworks are designed to use traditional sys. where the data is stored at single location like NFS, Oracle. But when processing big data the data is located on multiple commodity mac. with helps of HDFS

[Input] → [Map] → [Shuffling & Sorting] → [Reduce] → [O/P]

MapReduce is terminology that comes with Map phase & Reducer phase. map is used for transformation while reducer is used for aggregation kind of operation. The terminology for Map & Reduce is derived from some functional program lang. like Lisp, Scala etc. MapReduce processing framework prog. comes with 3 main components i.e. our Driver code, Mapper, & Reducer

PAGE NO.:
DATE: / /

Working of Mapper

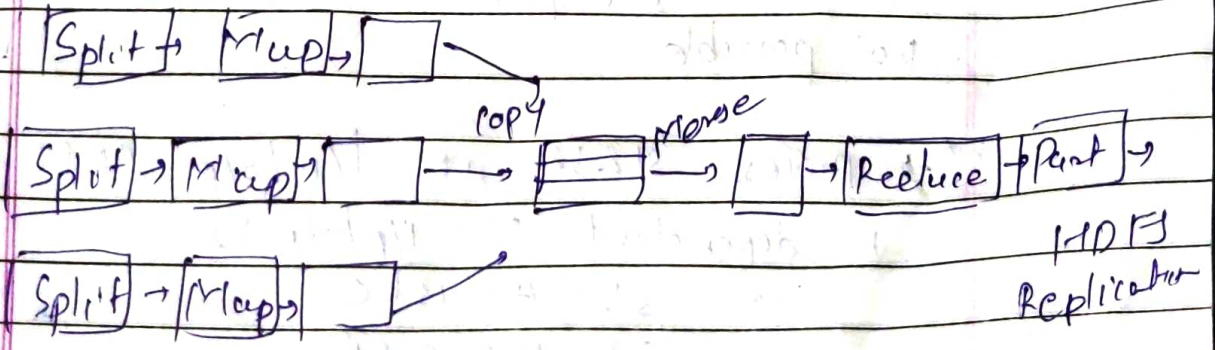
→ Mapper is initial line of code that initially interacts with i/p data sets. suppose if we have 120 data Blocks of dataset we are applying then the in that case, will be 120 Mapper program. that runs in parallel on machine & produce own output known as intermediate o/p which stored in local Disk not on HDFS. The o/p of Mapper act as i/p for Reducer which perform Sorting & aggregation o/p on data & produce final o/p

Working of Reducer

→ Mapper produces o/p in form key-value pairs which works as i/p for Reducer. But before sending this intermediate key-value pairs directly to Reducer some process will be done which shuffle & sort key-value pairs according to key values.

→ o/p of Reducer is stored on HDFS.

Reducer mainly perform some computation +, x.



MapReduce Data flow