⑨ Its data model is strict schema based and it is static.	⑨ Its data model is a flat schema based and it is dynamic.
⑩ It is easy to manage and manipulate data	⑩ It is difficult to manage and manipulate data.
⑪ Traditional data is in manageable volume	⑪ Big data is in huge volume which becomes unmanageable

a) What is mapreduce

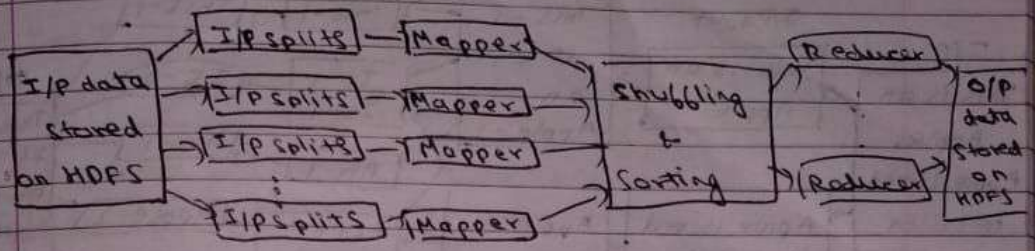
Ans ① Map Reduce is a programming model and framework within the Hadoop ecosystem that enables efficient processing of big data by automatically distributing and parallelizing the computation. It consists of two fundamental tasks - Map and Reduce.

② In Map phase, the input data is divided into smaller chunks and processed independently in parallel across multiple nodes in a distributed computing environment. Each chunk is transformed or "mapped" into a key value pair by applying a user defined function. The output of the Map phase is set of intermediate key-value pair

③ The Reduce phase follows the Map phase. It gathers the intermediate key value pair generated by Map tasks performs data shuffling to group together pairs with the same key, and then applies a user-defined reduction function to aggregate and process the data.

④ Architecture

(i) Input Split
(ii) Mapping
(iii) Shuffling
(iv) Sorting
(v) Reducing.



(i) Input splits :- Map Reduce split the input into smaller chunks called input split, representing a block of work with a single mapper task.

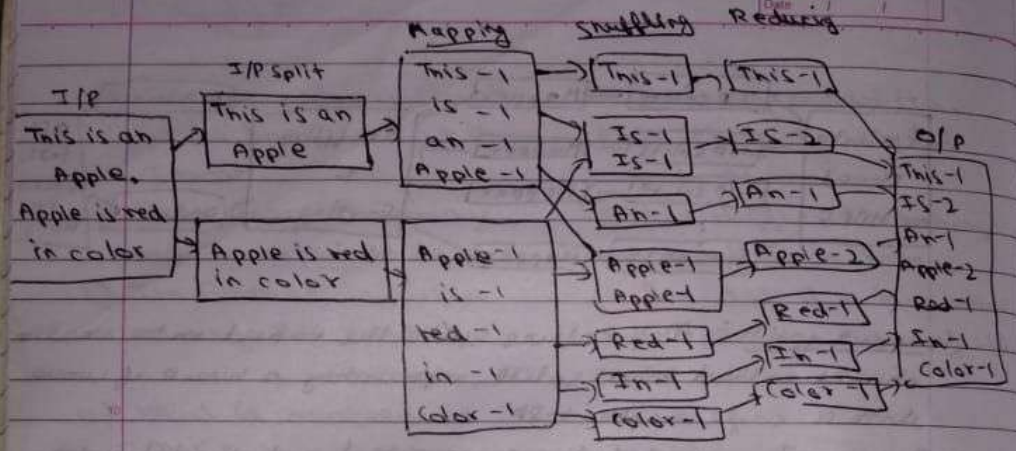
(ii) Mapping :- The input data is processed and divided into smaller segments in mapper phase, where the number of mapper is equal to the number of input split

(iii) Shuffling :- In the shuffling phase, the out of the mapper phase is passed to the reducer phase by removing duplicate values and grouping the values.

(iv) Sorting :- ~~Sorted~~ Sorting is performed simultaneously with shuffling. The sorting phase involves merging and sorting the output generated by the mapper.

(v) Reducing :- In the reducer phase, the intermediate values from the shuffling phase are reduced to produce a single output value ~~of~~ that summarizes the entire dataset. HDFS is then used to store final output.

Eg :- "This is an Apple. Apple is red in color."



(i) The input data is divided into multiple segments, then processed in parallel to reduce processing time. In this case, the input data will be divided into two input splits so that the work can be distributed over all the map nodes.

(ii) The mapper counts the number of times each word occurs from input splits in form of key-value pair when the key is the word, and the value is the frequency.

(iii) For the first I/p split, it generated 4 key-value pair This:1, is:1, an:1, apple:1, and for second it generated 5 key-value pair: Apple:1, is:1, red:1, in:1, color:1

(iv) It is followed by shuffle phase, in which the values are grouped by keys in form of key-value pair. Here we get a total of 6 groups of key-value pair.

(v) The same reducer is used for all key-value pair with the same key.

(vi) All the words present in the data are combined into a single output in reducer phase. The output shows frequency of each word.

(vii) Final output key value pair This:1, is:2, an:1, Apple:2, Red:1, In:1, cole:1.