

# MAP REDUCE (Word count Job)

Hi how are you	64
How is your job	
How is your family	64
How is your sister	
How is your brother	64
What is the time now	
What is the strength of hadoop	8

If I use in HDFS, 200MB should be divided into 4 parts & maximum of 64MB

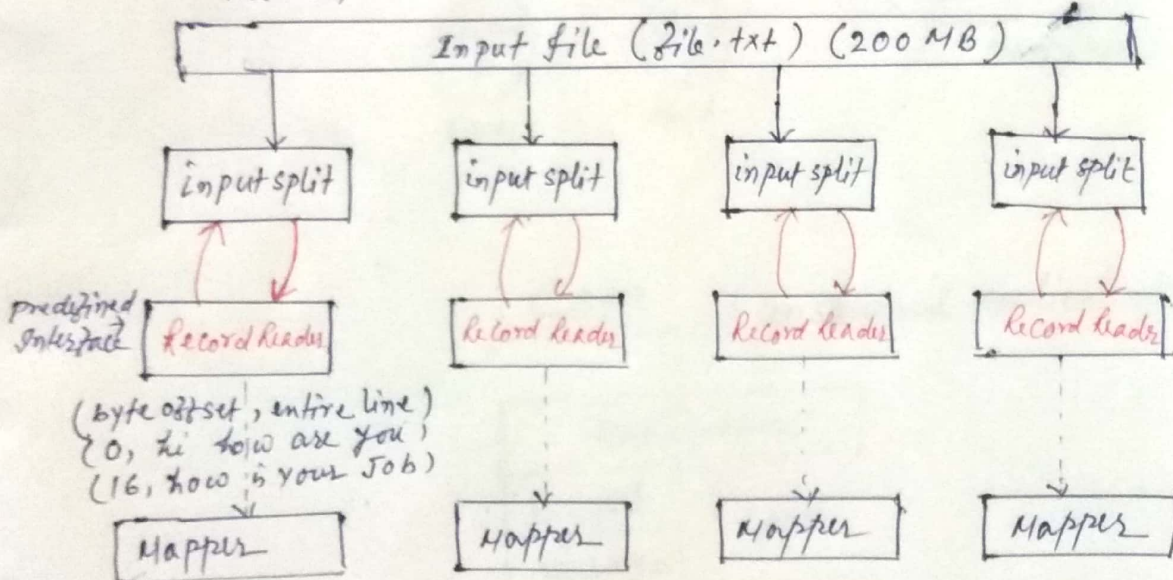
## File Format

- ① Text input format
- ② keyvalues Text Input Format
- ③ Sequence File Input format
- ④ Sequence File As Text Input format

By default → text input format

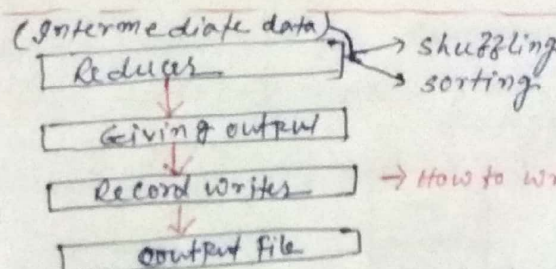
- \* Convert every line to (key, value) for every i/p
- \* Mapper only know (key, value)

hadoop jar test.jar Drives code file.txt testoutput



(hi, 1)	(how, 1)	(how, 1)	(how, 1)	(what, 1)	(what, 1)
(how, 1)	(is, 1)	(is, 1)	(is, 1)	(is, 1)	(is, 1)
(are, 1)	(your, 1)	(your, 1)	(your, 1)	(the, 1)	(the, 1)
(you, 1)	(job, 1)	(family, 1)	(sister, 1)	(brother, 1)	(time, 1)
				(now, 1)	(strength, 1)
					(of, 1)
					(hadoop, 1)

\* (how, 1, 1, 1, 1, 1)  
all the 1's associated to single identical key  
shuffling (how, 1, 1, 1, 1, 1)  
(is, 1, 1, 1, 1, 1)



## shuffling

\* Mapper does not know how to sort the data & shuffle data

\* Identical Reducers know how to sort the data not shuffle.

\* Own Reducer knows both sort & shuffle. { part-00000 }

\* Duplicate key not possible but value may be duplicate. { success → directory }

Log →