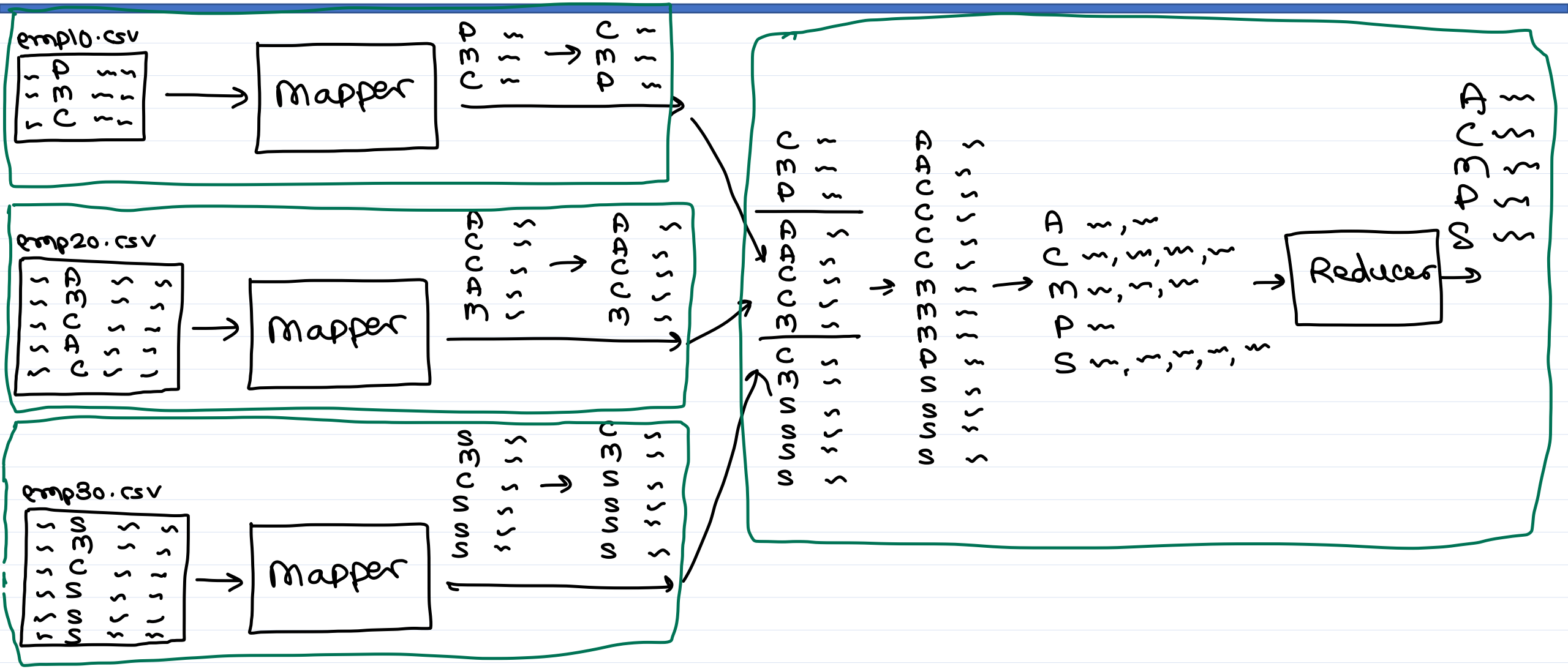
Big Data Technologies

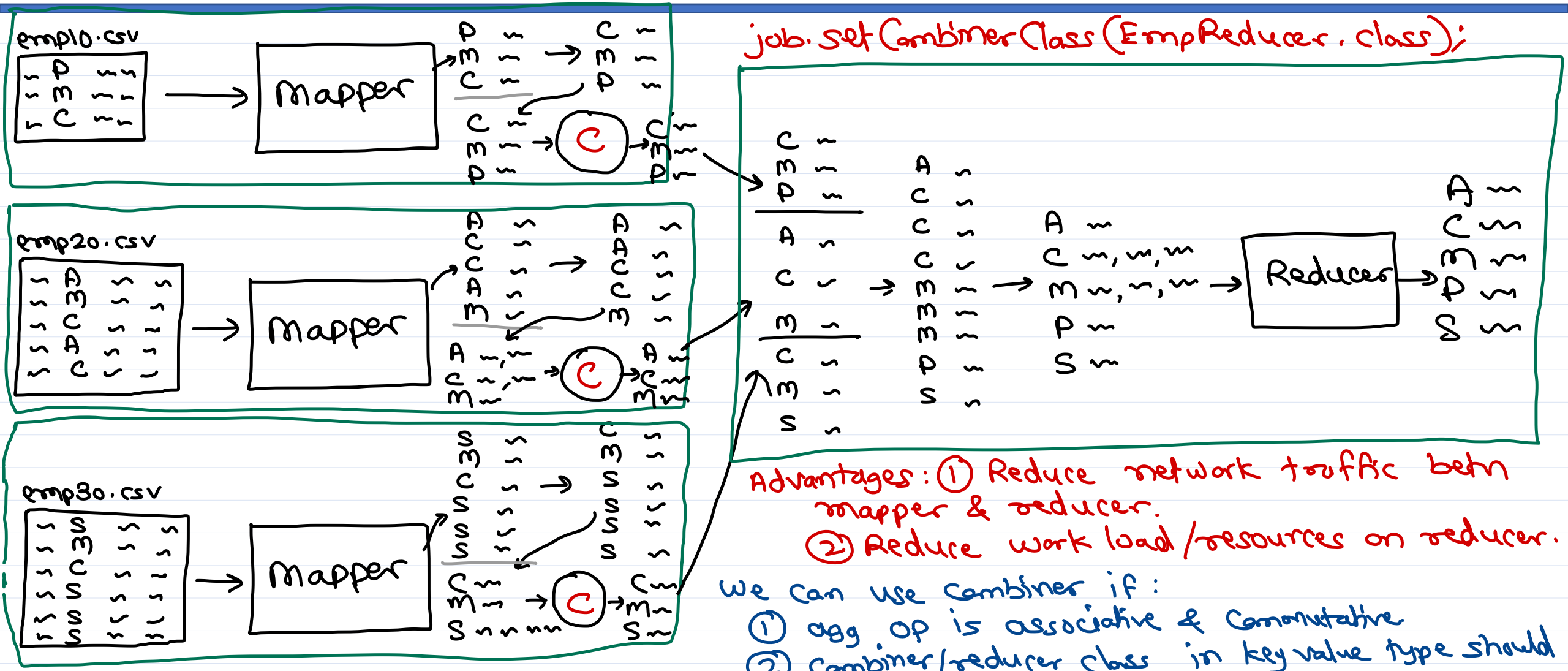
Trainer: Mr. Nilesh Ghule.

MR execution flow Multiple Mappers and Single Reducer



MR execution flow

Combiner a.k.a Local Reducer

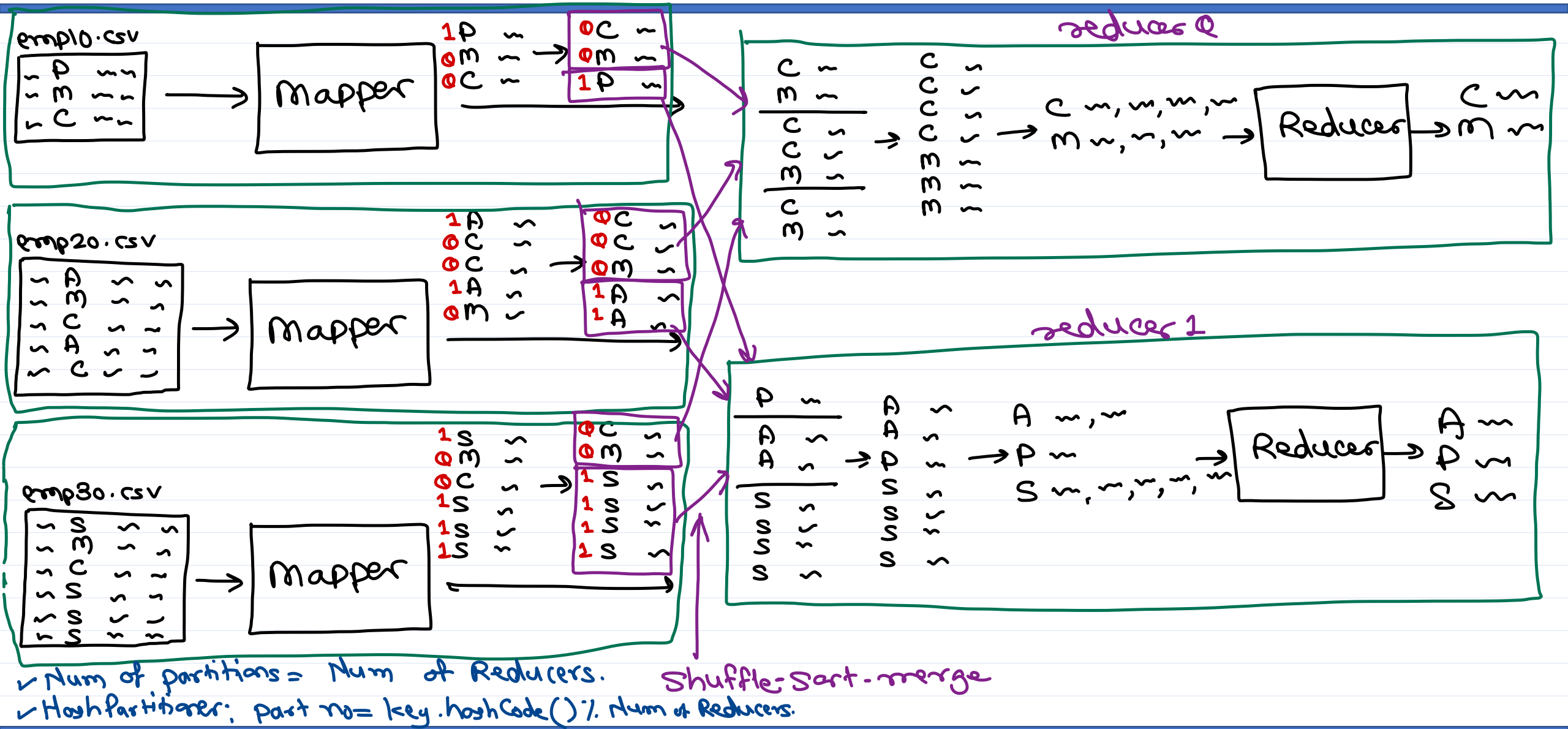


Advantages : ① Reduce network traffic betn mapper & reducer.
② Reduce work load/resources on reducer.

We can use combiner if :
① agg op is associative & commutative
② Combiner/reducer class in key value type should be same as out key value type.

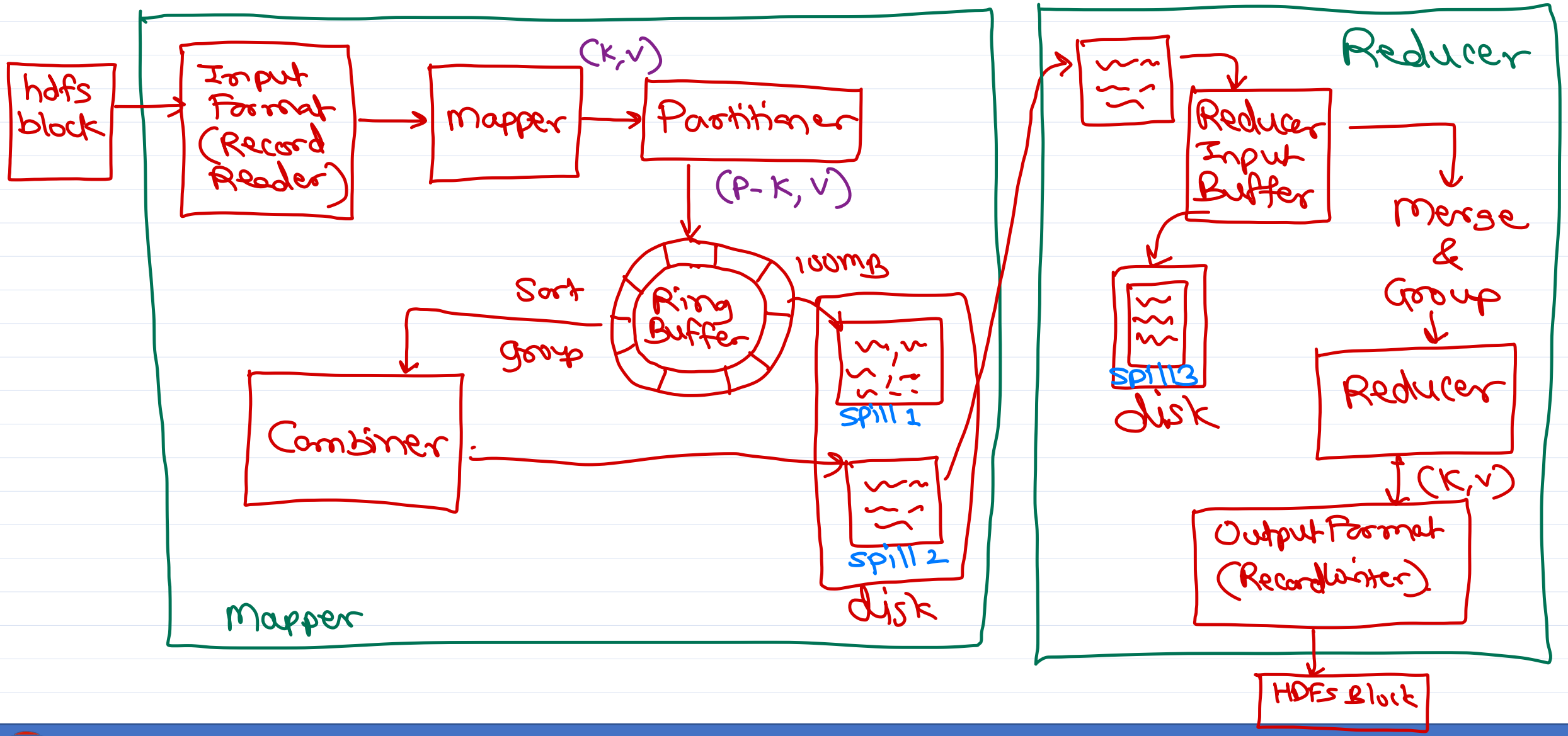
MR execution flow

Multiple Mappers and Multiple Reducer

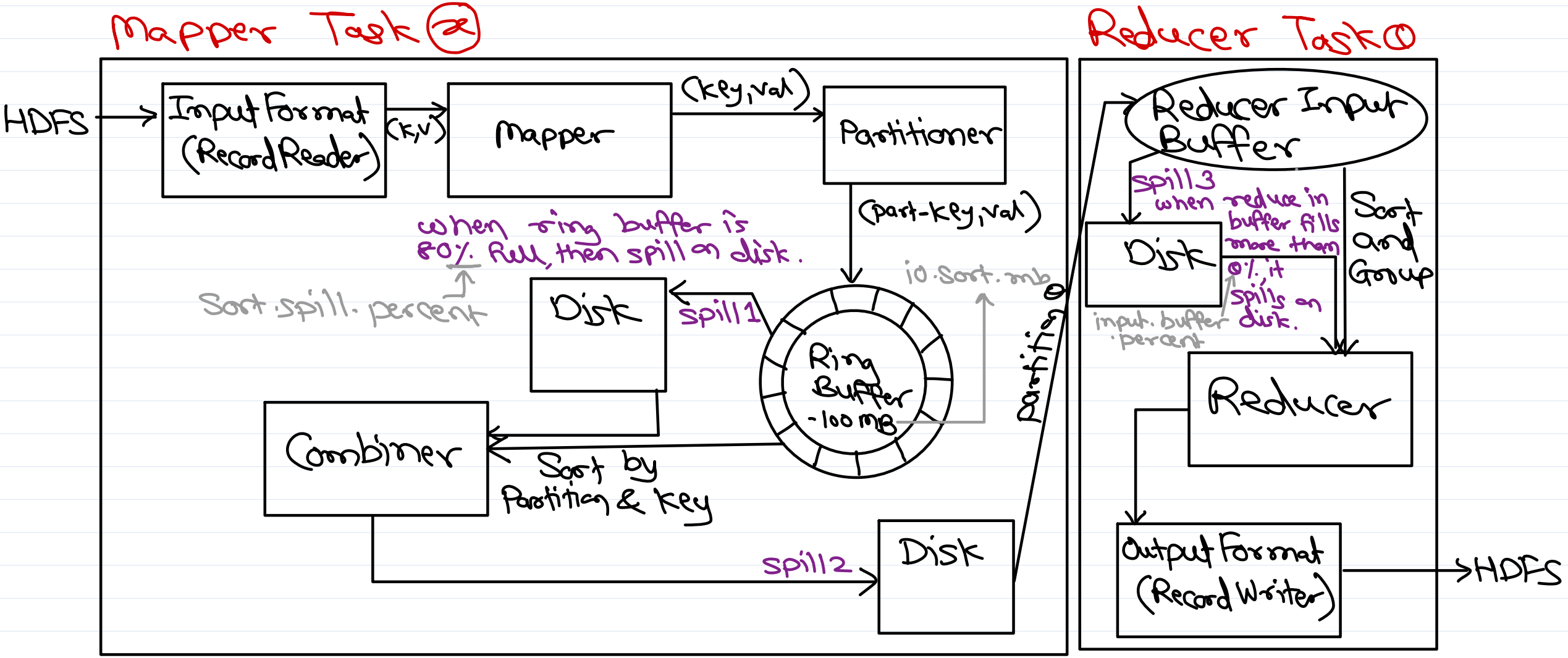


- Num of partitions = Num of Reducers.
- HashPartitioner: part no = key.hashCode() % Num of Reducers.

Map Reduce Flow



Map Reduce Data Flow



Input Splits

Number of mappers = Number of input splits

Number of Input splits ≈ Number of HDFS blocks

Ex1: 20 files each of 1MB approx
- → Num of HDFS blocks = 20
- → Num of input splits = 20
- → Num of mappers = 20.

Ex2: 2 files – f1 = 200MB, f2 = 300MB
- → Num of HDFS blocks = 2+3 = 5
- → Num of input splits = 5
- → Num of mappers = 5

Input format is responsible for reading records one by one. It has an internal RecordReader.

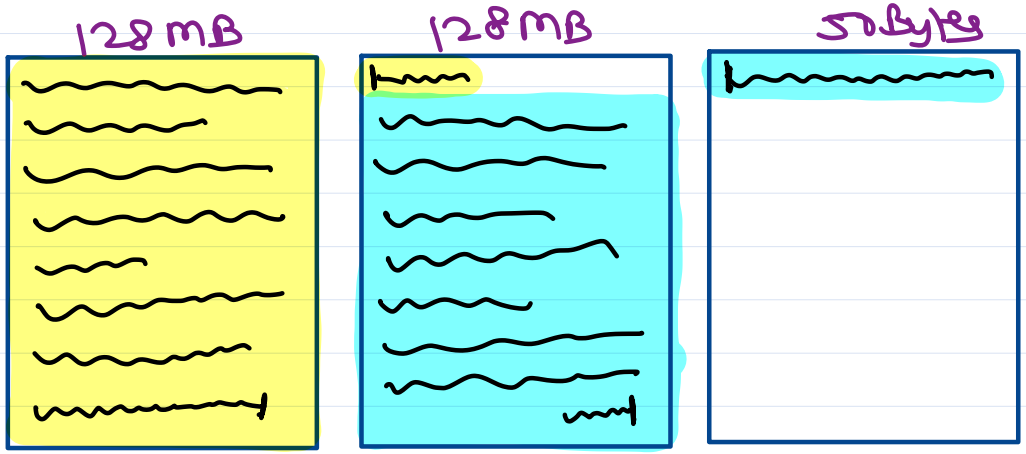
- TextInputFormat → 1 record = 1 line → (offset, line).
- KeyValueTextInputFormat → 1 record = 1 line → (key, value)

key \t value

- DbInputFormat → 1 record = 1 row (RDBMS)
- CombinedTextInputFormat → Combine multiple text files into single input split.

f1 = 256 MB + 50 bytes.

HDFS blocks = 3.



Input splits = 2

Input split is logical part of data that is to be processed by a single mapper.

Mostly 1 split = 1 hdfs block.

Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>