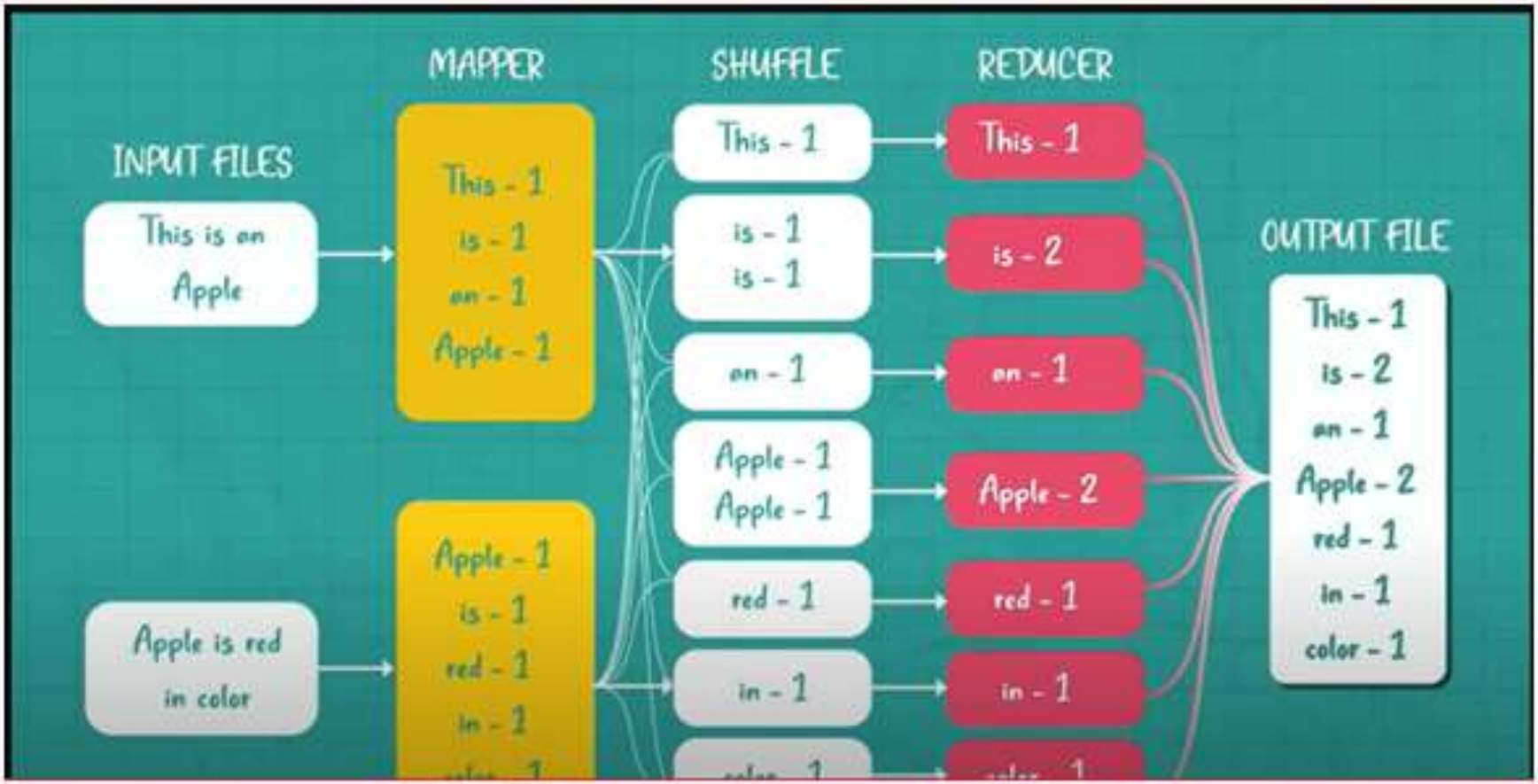
# Map Reduce – Shuffle & Sort



- In a MapReduce job, the input data is divided into smaller chunks, which are processed by the mappers in parallel.
- The mappers output key-value pairs, which are shuffled and sorted by the framework.
- The reducers then process the sorted key-value pairs and generate the final output.

Shuffle and Sort Phase

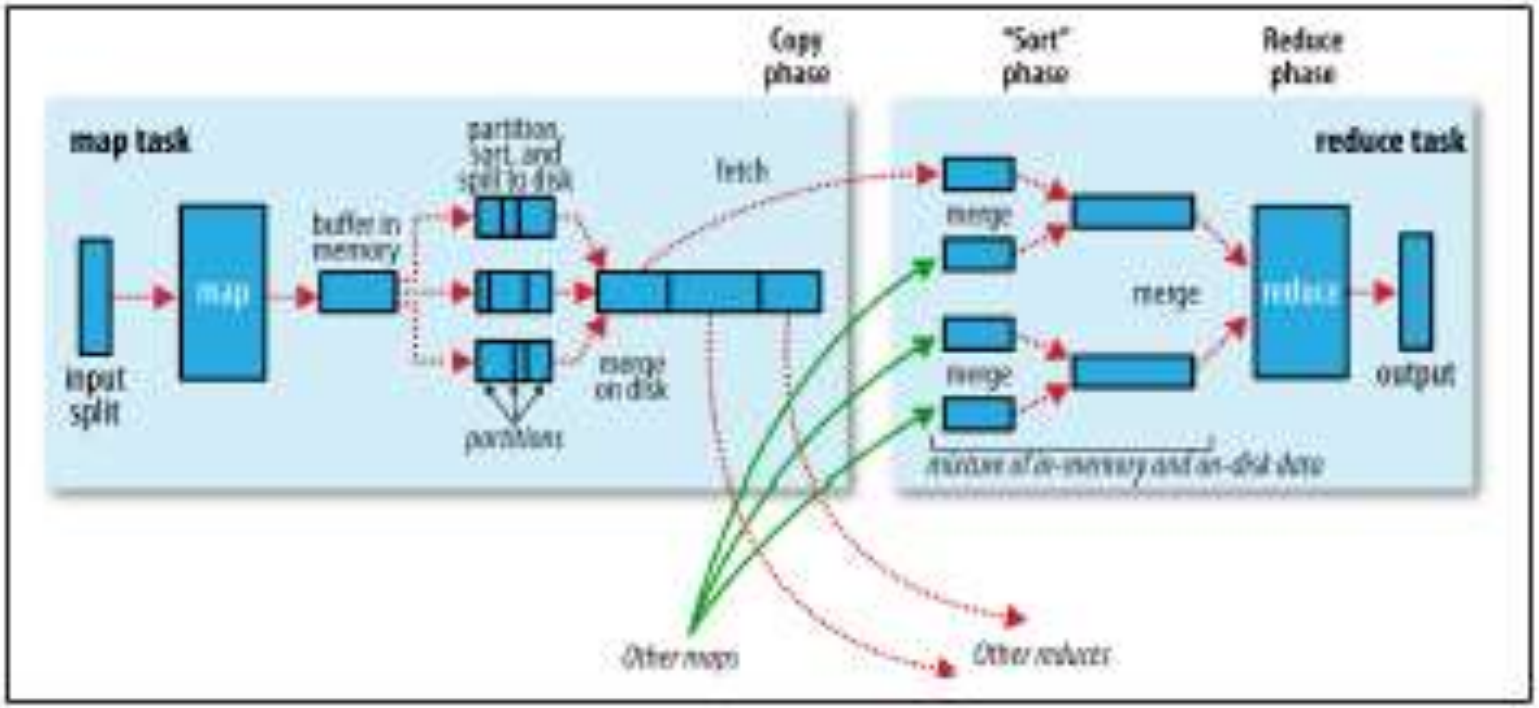
- The shuffle and sort phase is a critical component of MapReduce.
- It is responsible for transferring the Shuffled intermediate outputs of the mappers to the reducers and sorting them by key.

## Map Side

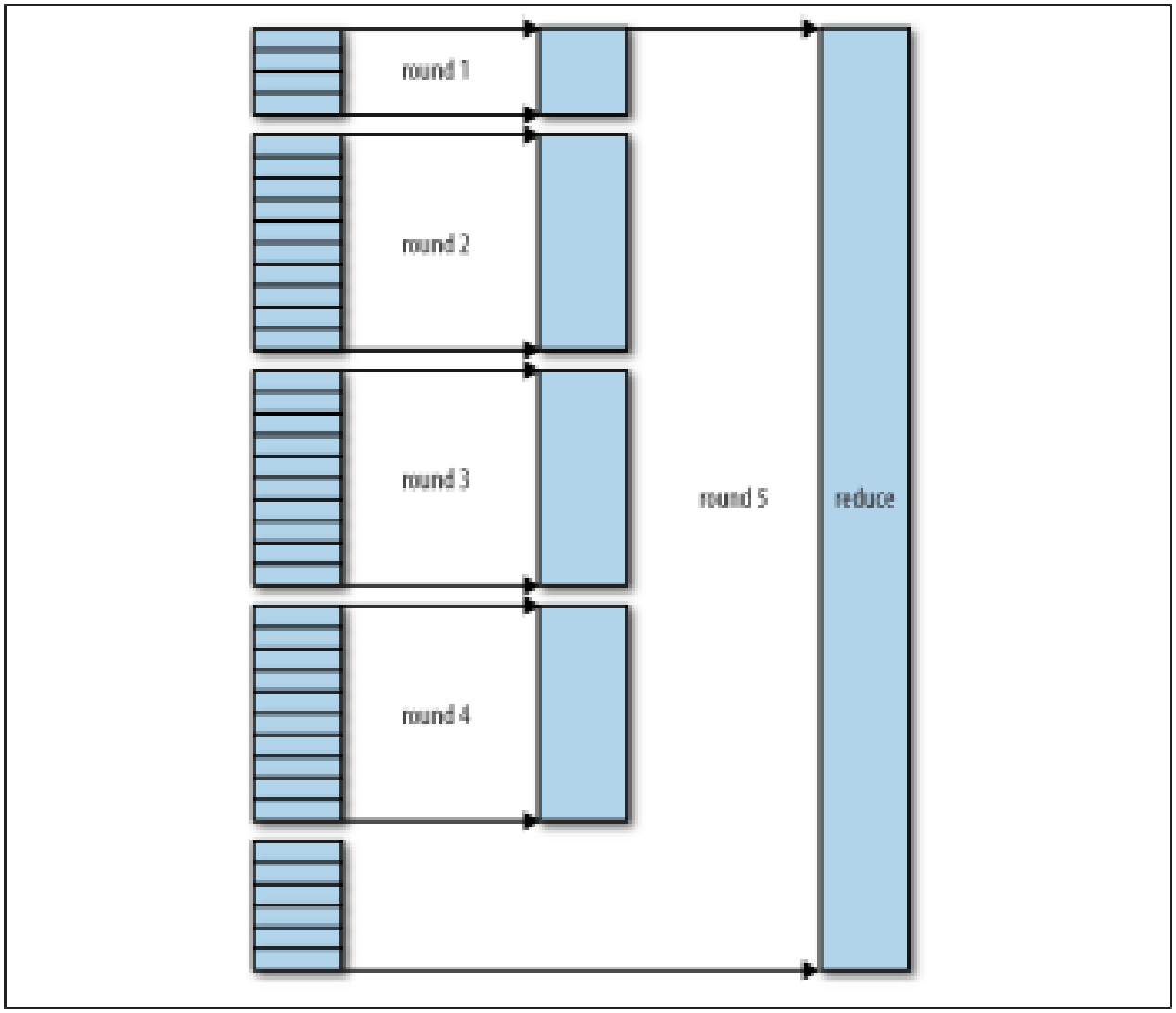
- On the map side, the shuffle process begins by partitioning the intermediate key-value pairs into buckets based on their keys.
- The buckets are then serialized and written to a local disk.

## Reduce Side

- On the reduce side, the shuffle process begins by reading the serialized intermediate key-value pairs from the local disk of each mapper.
- The pairs are then sorted by key and merged into a single stream.




# Shuffle and sort in MapReduce




**Efficiently merging 40 file segments with a merge factor of 10**

# Configuration Tuning – Adjusting Parameters

The performance of the shuffle and sort phase can be tuned by **Adjusting A Number Of Configuration Parameters**, such as:

**io.sort.mb**: The maximum amount of memory to use for sorting intermediate key-value pairs in memory.

**io.sort.factor**: The number of intermediate files to merge into a single file before writing to disk.

**io.sort.spill.percent**: The percentage of the memory used for sorting that triggers spilling to disk.

| Property name | Type | Default value | Description |
|---|---|---|---|
| mapreduce.task.io.sort.mb | int | 100 | The size, in megabytes, of the memory buffer to use while sorting map output. |
| mapreduce.map.sort.spill.percent | float | 0.80 | The threshold usage proportion for both the map output memory buffer and the record boundaries index to start the process of spilling to disk. |
| mapreduce.task.io.sort.factor | int | 10 | The maximum number of streams to merge at once when sorting files. This property is also used in the reduce. It's fairly common to increase this to 100. |
| mapreduce.map.combine.minspills | int | 3 | The minimum number of spill files needed for the combiner to run (if a combiner is specified). |
| mapreduce.map.output.compress | boolean | false | Whether to compress map outputs. |
| mapreduce.map.output.compress.codec | Class name | org.apache.hadoop.io.compress.DefaultCodec | The compression codec to use for map outputs. |
| mapreduce.shuffle.max.threads | int | 0 | The number of worker threads per node manager for serving the map outputs to reducers. This is a cluster-wide setting and cannot be set by individual jobs. 0 means use the Netty default of twice the number of available processors. |

## Map Side Tuning Properties

# Reduce-side Tuning Properties

| Property name | Type | Default value | Description |
|---|---|---|---|
| mapreduce.reduce.shuffle.parallelcopies | int | 5 | The number of threads used to copy map outputs to the reducer. |
| mapreduce.reduce.shuffle.maxfetchfailures | int | 10 | The number of times a reducer tries to fetch a map output before reporting the error. |
| mapreduce.task.io.sort.factor | int | 10 | The maximum number of streams to merge at once when sorting files. This property is also used in the map. |
| mapreduce.reduce.shuffle.input.buffer.percent | float | 0.70 | The proportion of total heap size to be allocated to the map outputs buffer during the copy phase of the shuffle. |
| mapreduce.reduce.shuffle.merge.percent | float | 0.66 | The threshold usage proportion for the map outputs buffer (defined by mapred.job.shuffle.input.buffer.percent) for starting the process of merging the outputs and spilling to disk. |
| mapreduce.reduce.merge.inmem.threshold | int | 1000 | The threshold number of map outputs for starting the process of merging the outputs and spilling to disk. A value of 0 or less means there is no threshold, and the spill behavior is governed solely by mapreduce.reduce.shuffle.merge.percent. |
| mapreduce.reduce.input.buffer.percent | float | 0.0 | The proportion of total heap size to be used for retaining map outputs in memory during the reduce. For the reduce phase to begin, the size of map outputs in |