Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

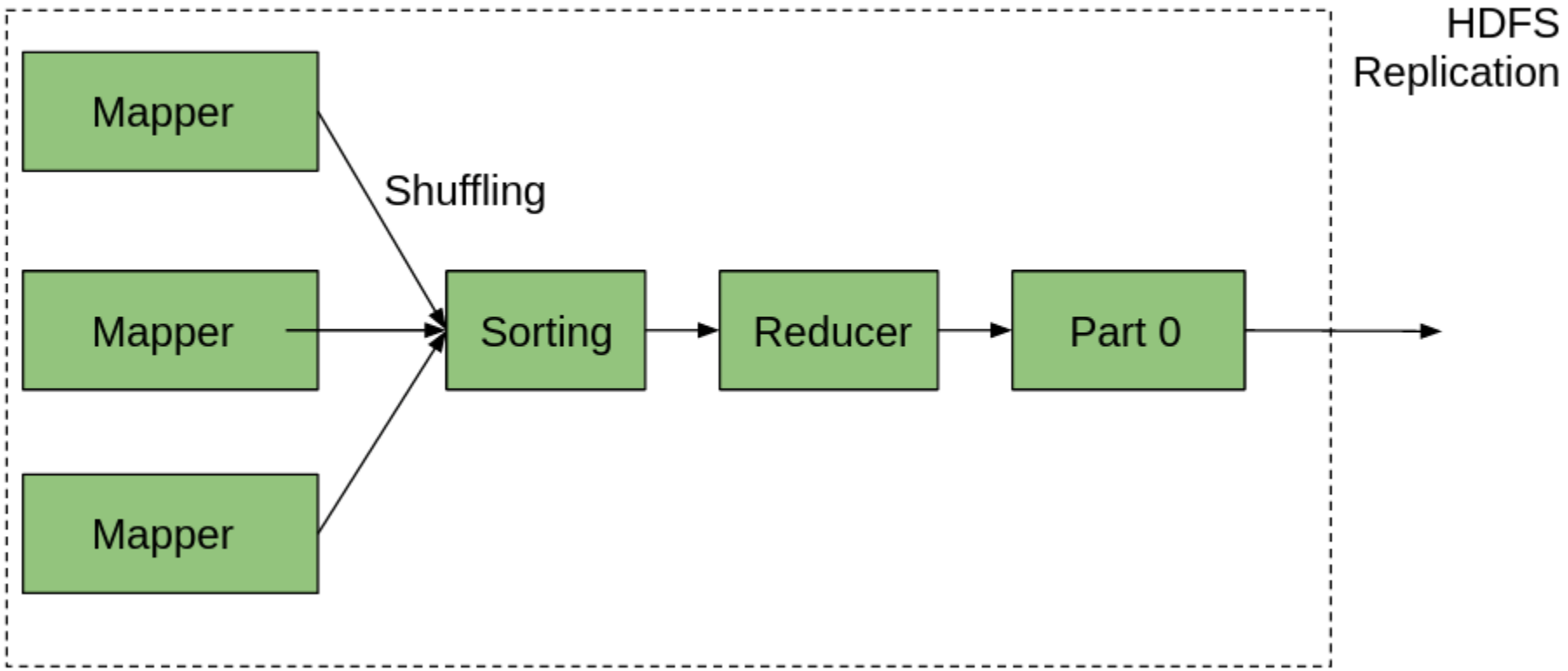
Associate Professor

Learning Outcomes

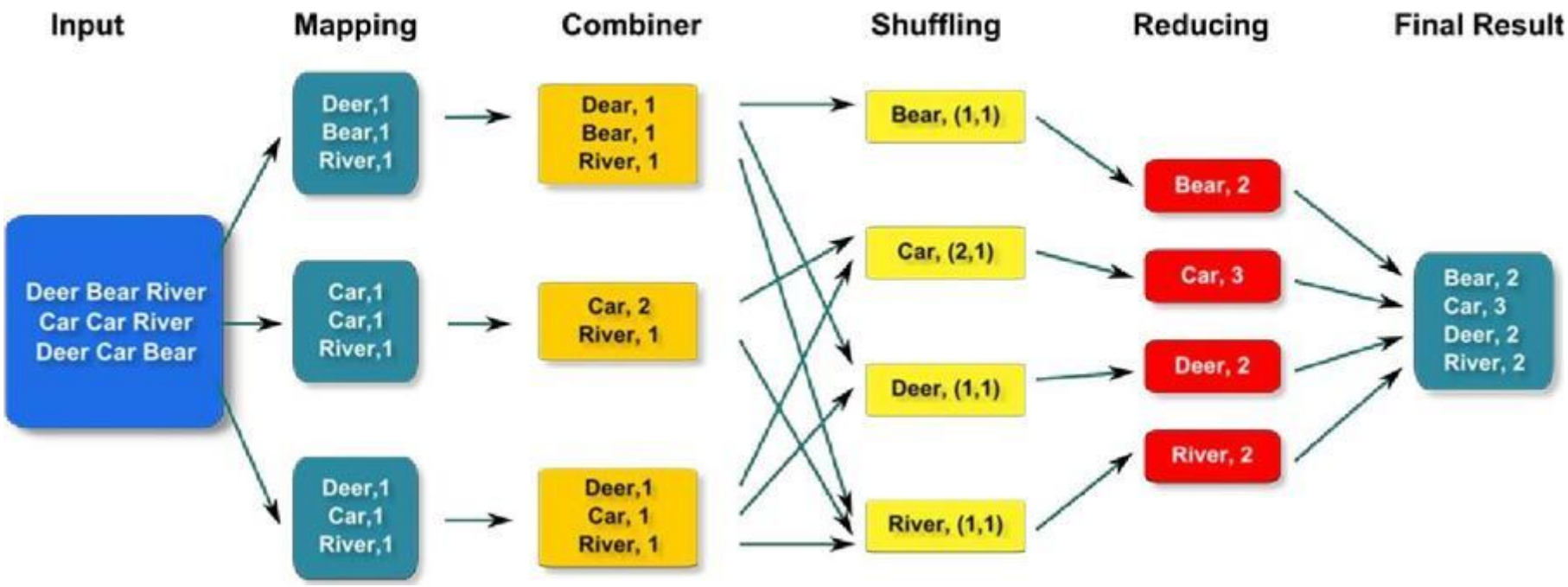
After this lecture, you will be able to

- explore and learn the concepts of Shuffle and Sort

Introduction



Introduction




What is Shuffling and Sorting in Hadoop MapReduce?

Phases of MapReduce

Mapper

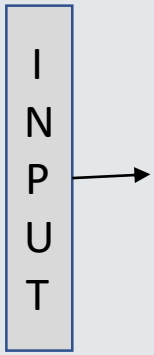
Reducer in MapReduce

Combiner

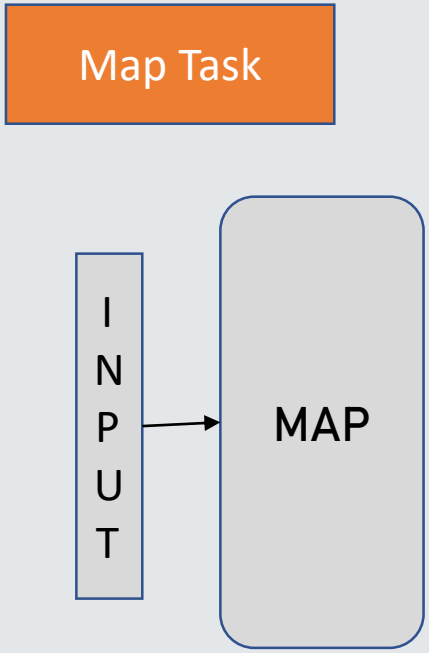
Partitioner in MapReduce

inputFormat
in MapReduce

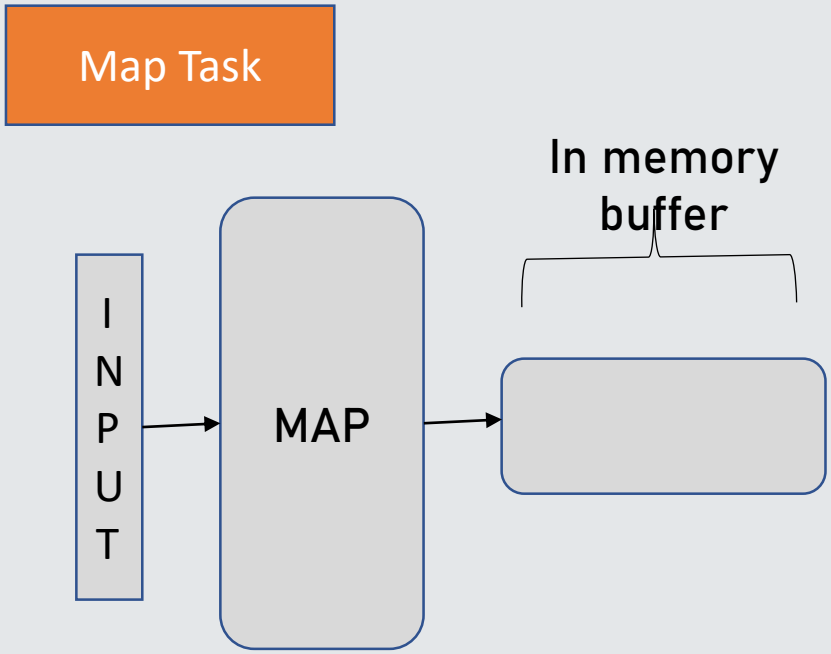
Shuffle and Sort



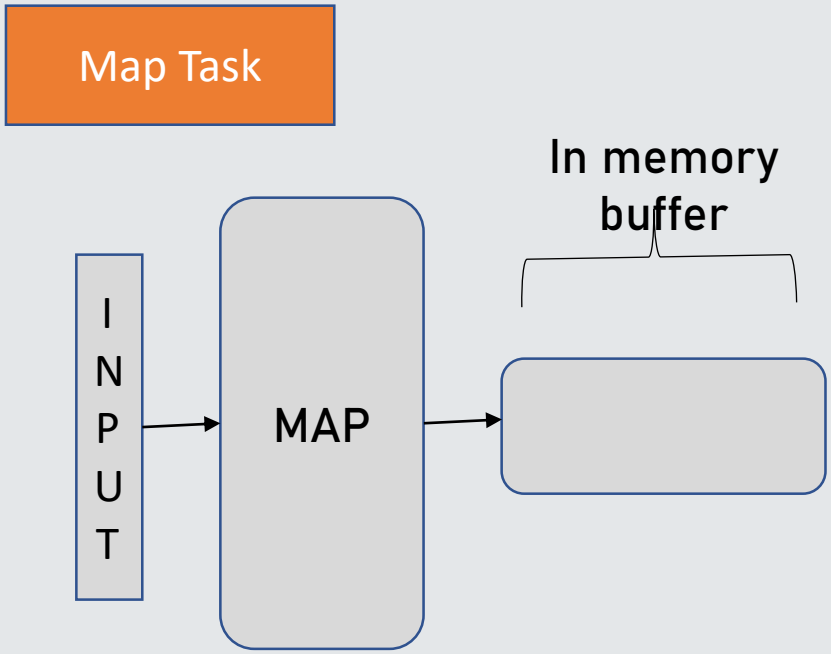
Shuffle and Sort



Shuffle and Sort

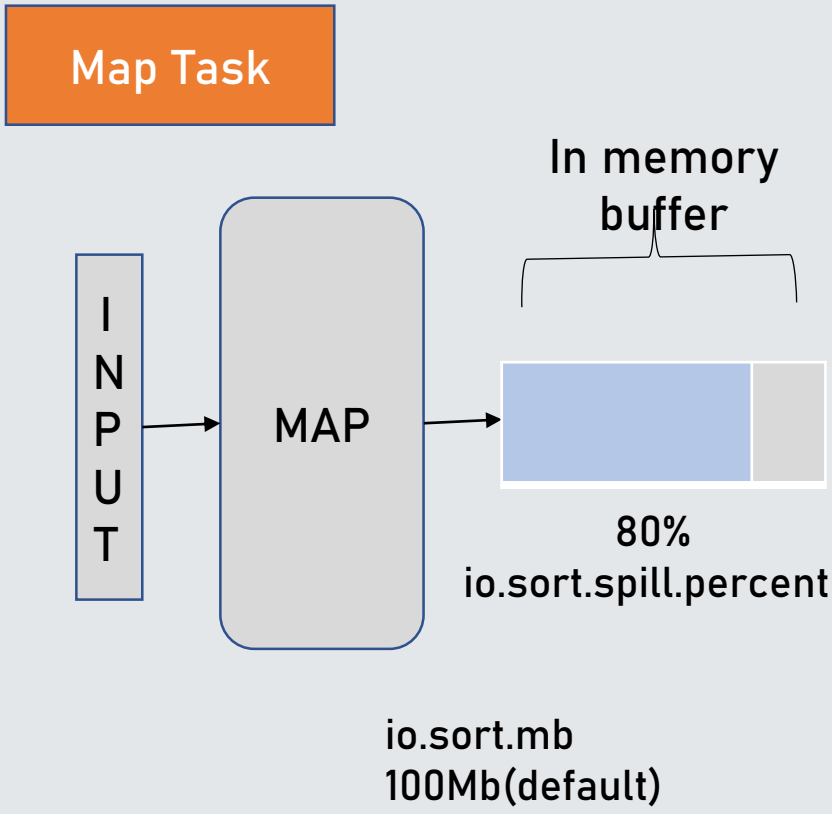


Shuffle and Sort

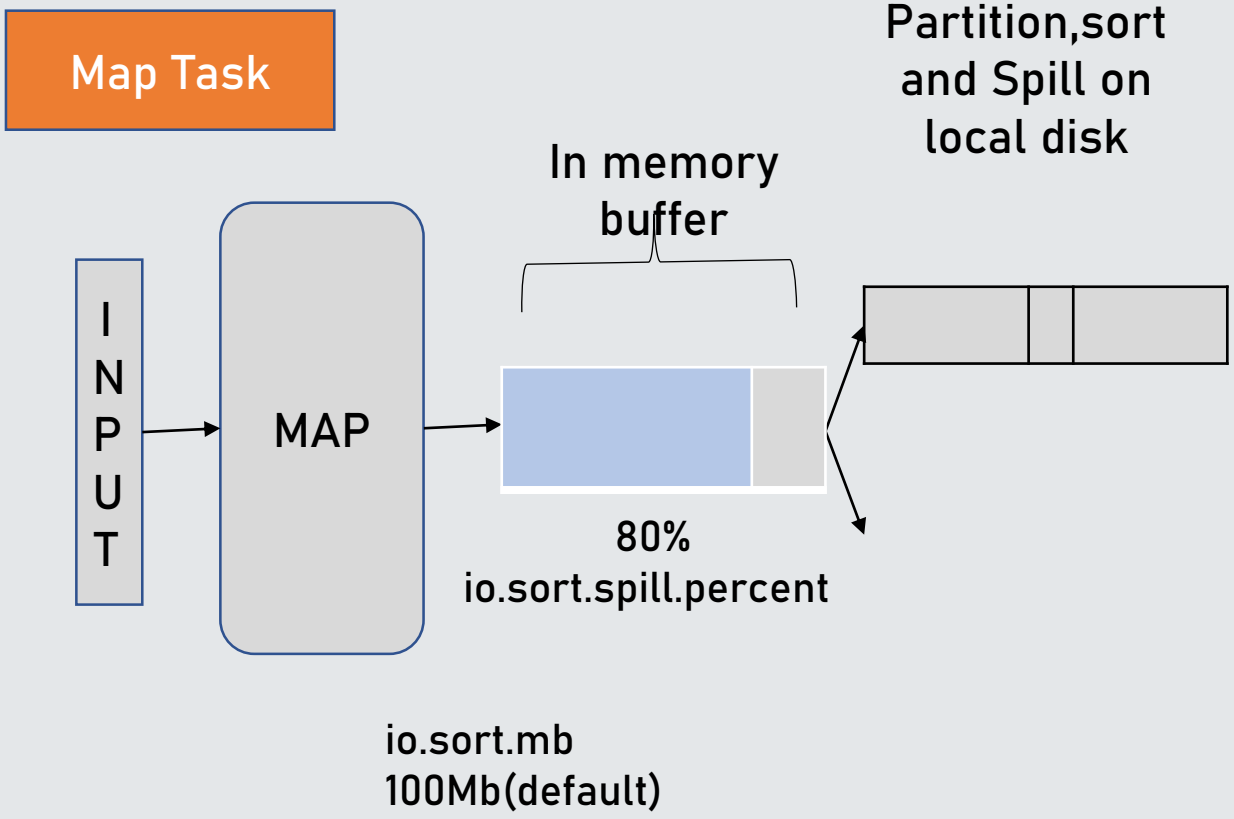


Partition,sort and Spill on local disk

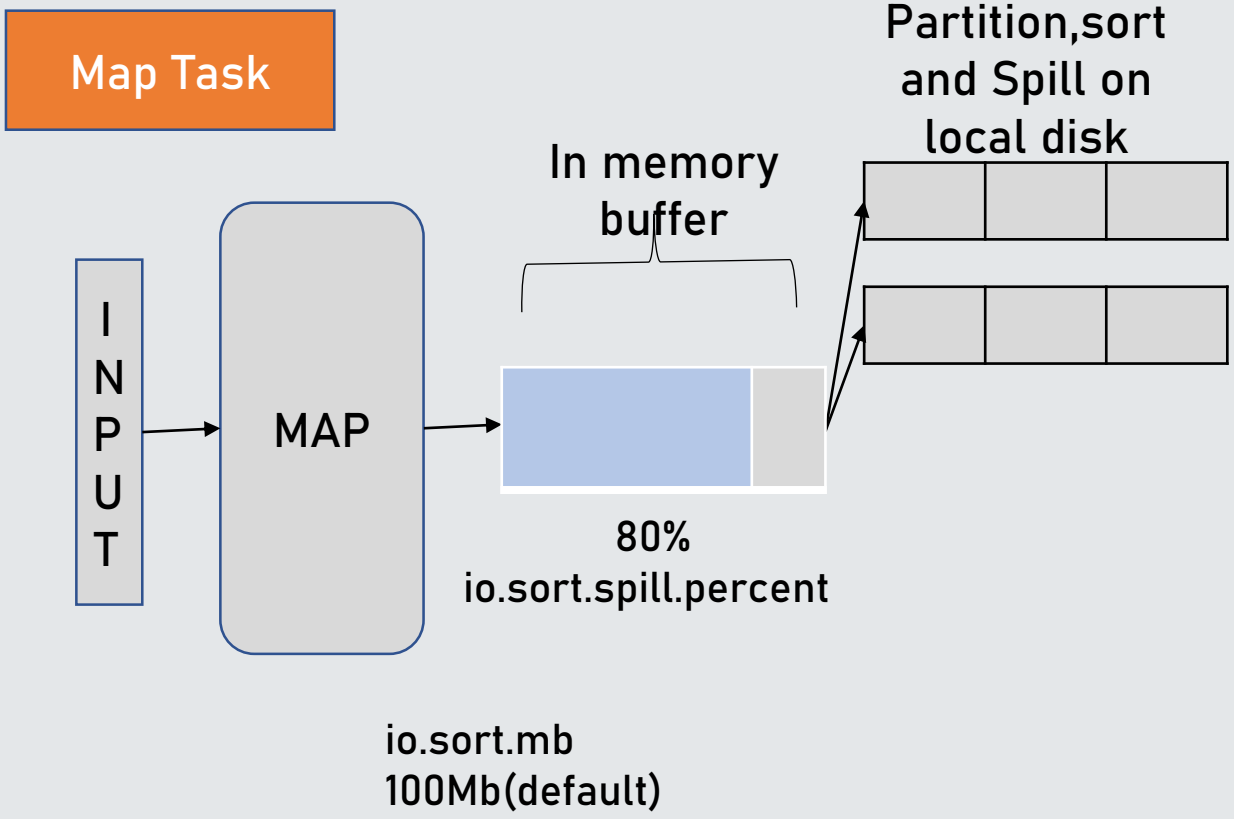
Shuffle and Sort



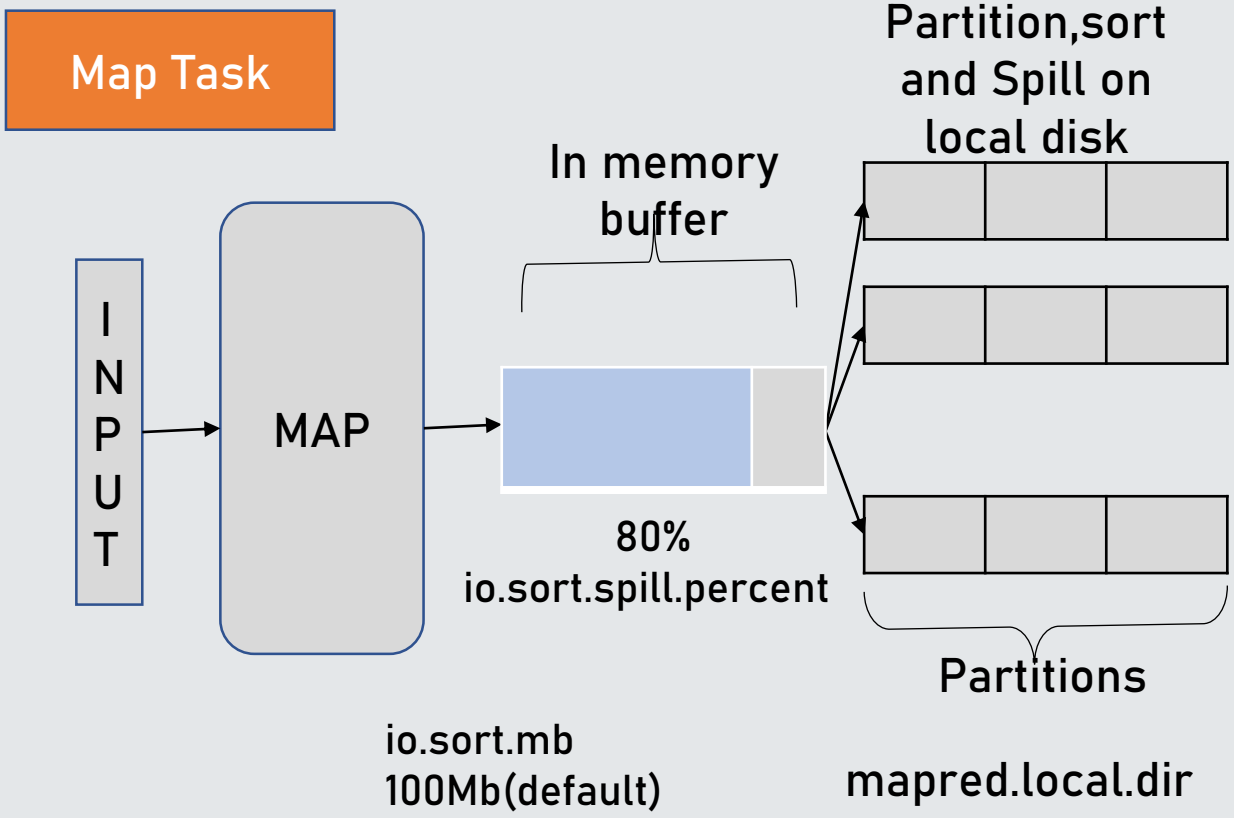
Shuffle and Sort



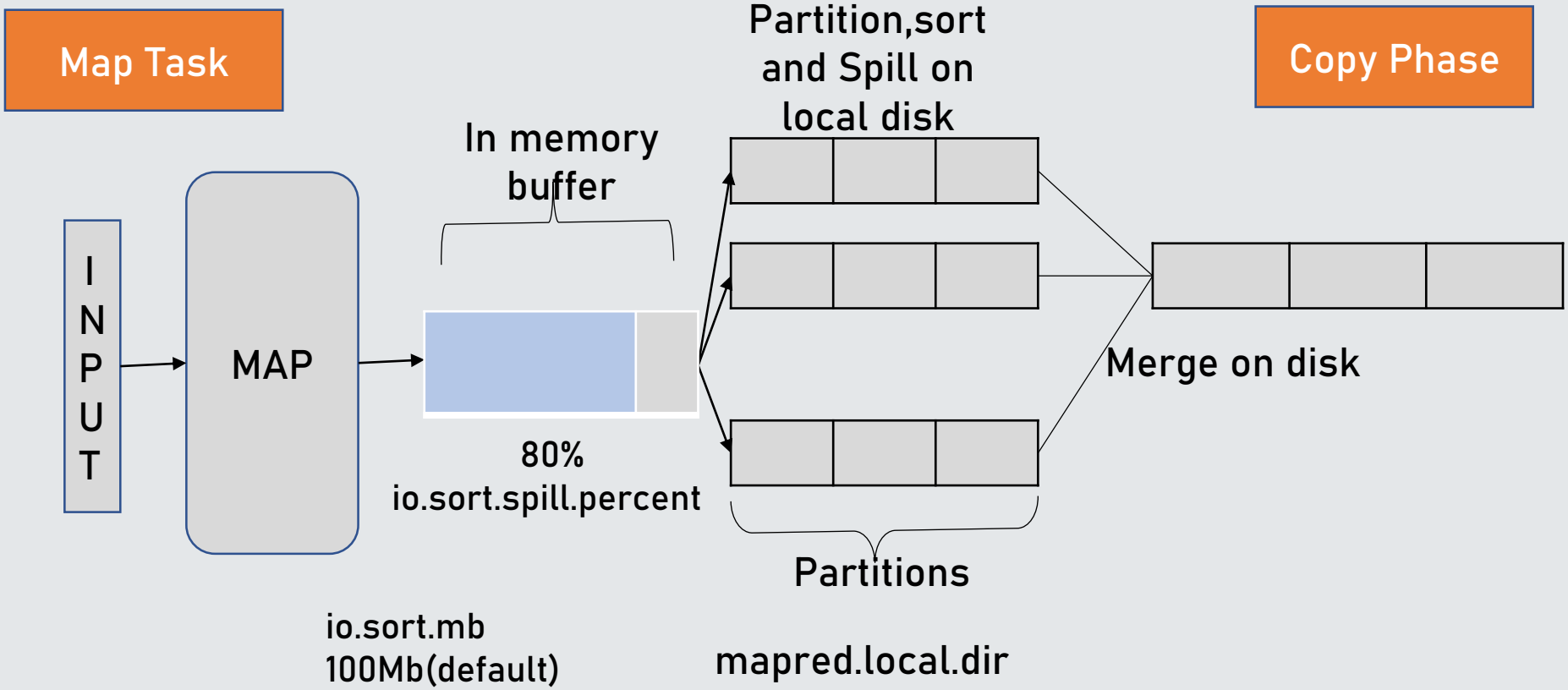
Shuffle and Sort



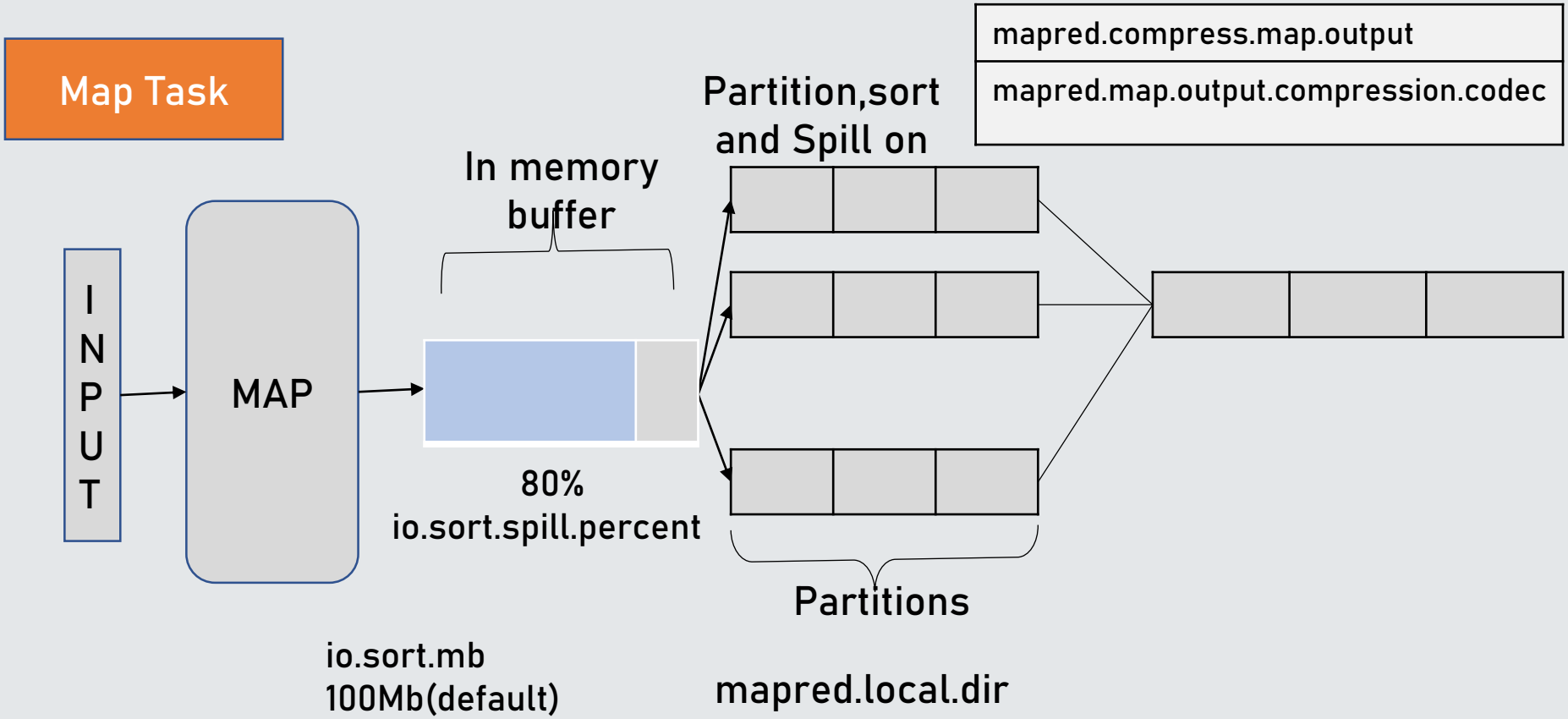
Shuffle and Sort



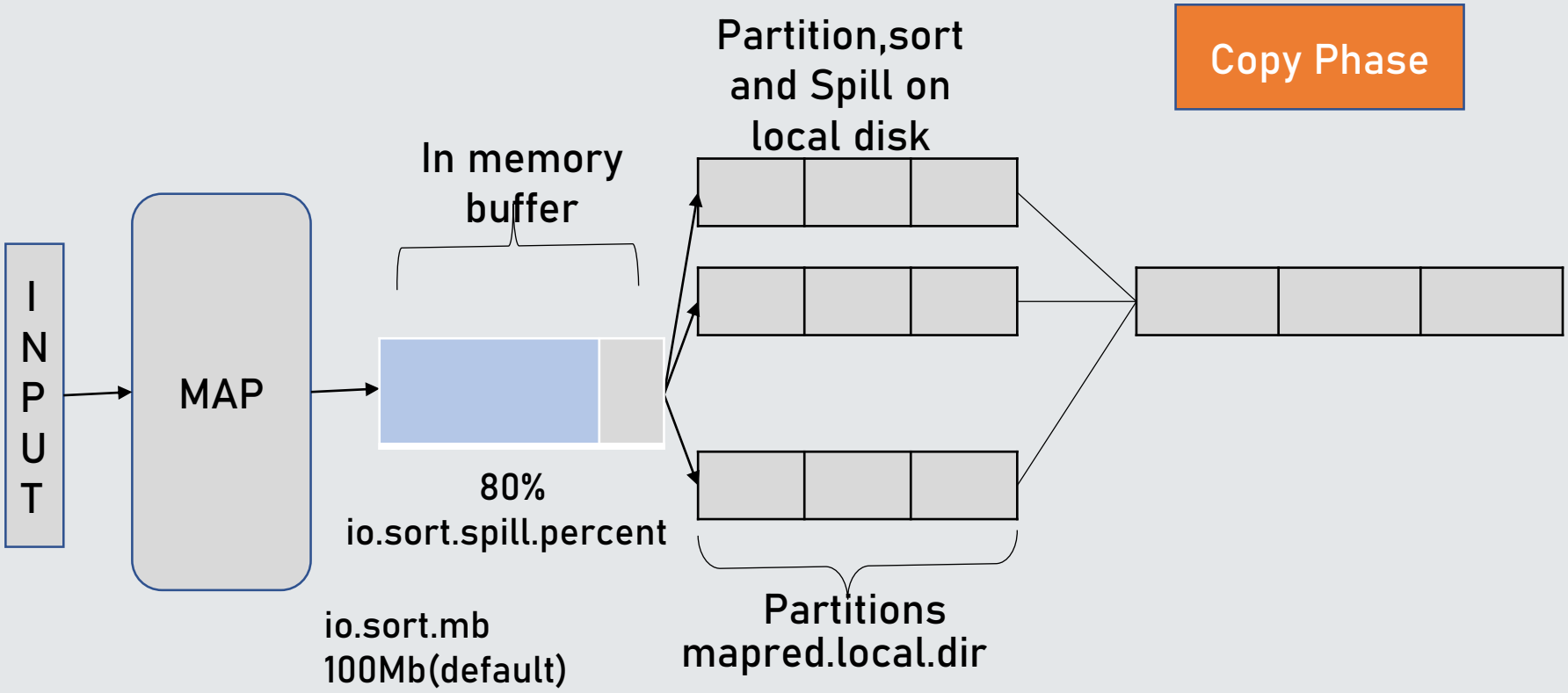
Shuffle and Sort



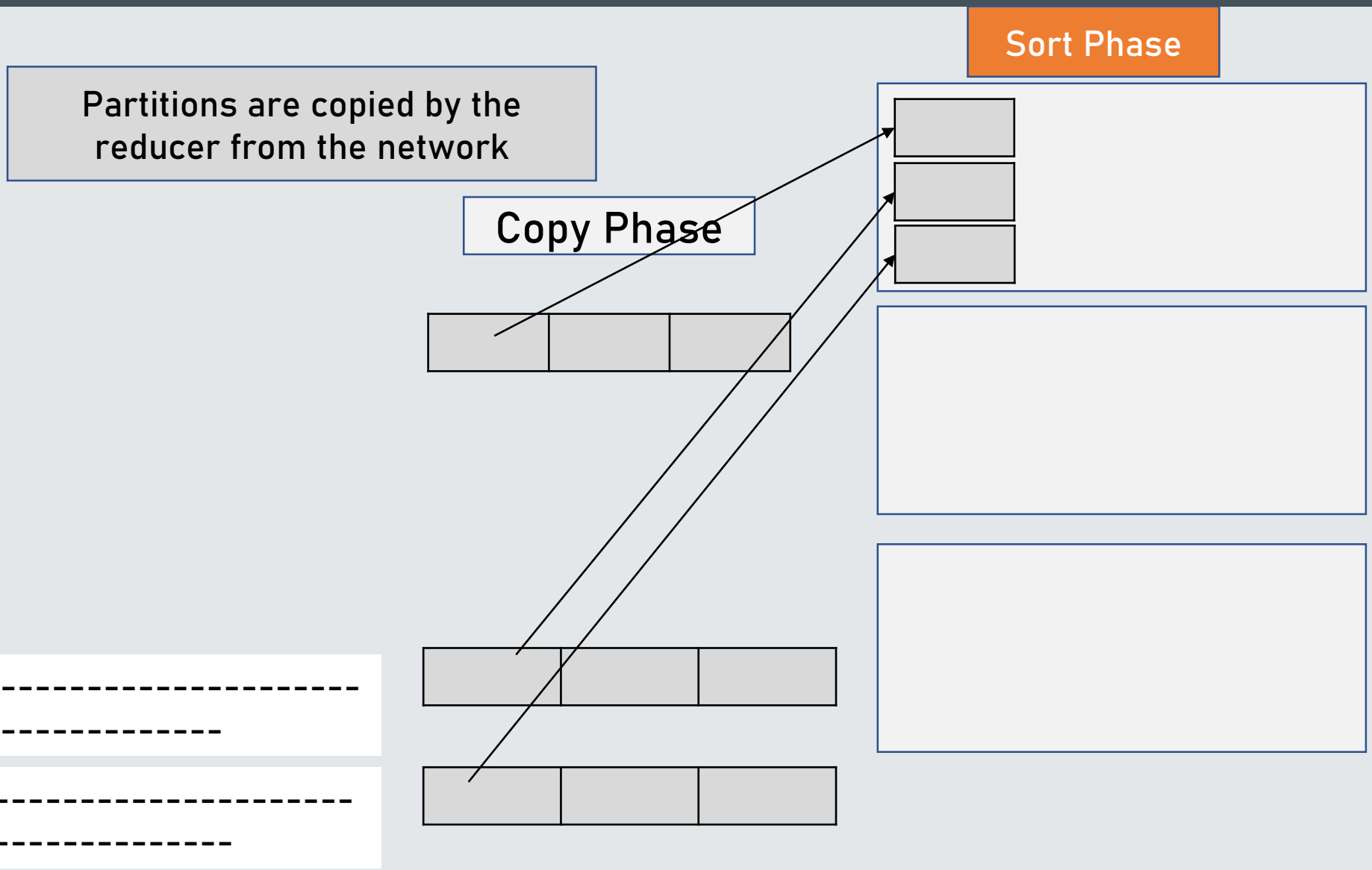
Shuffle and Sort



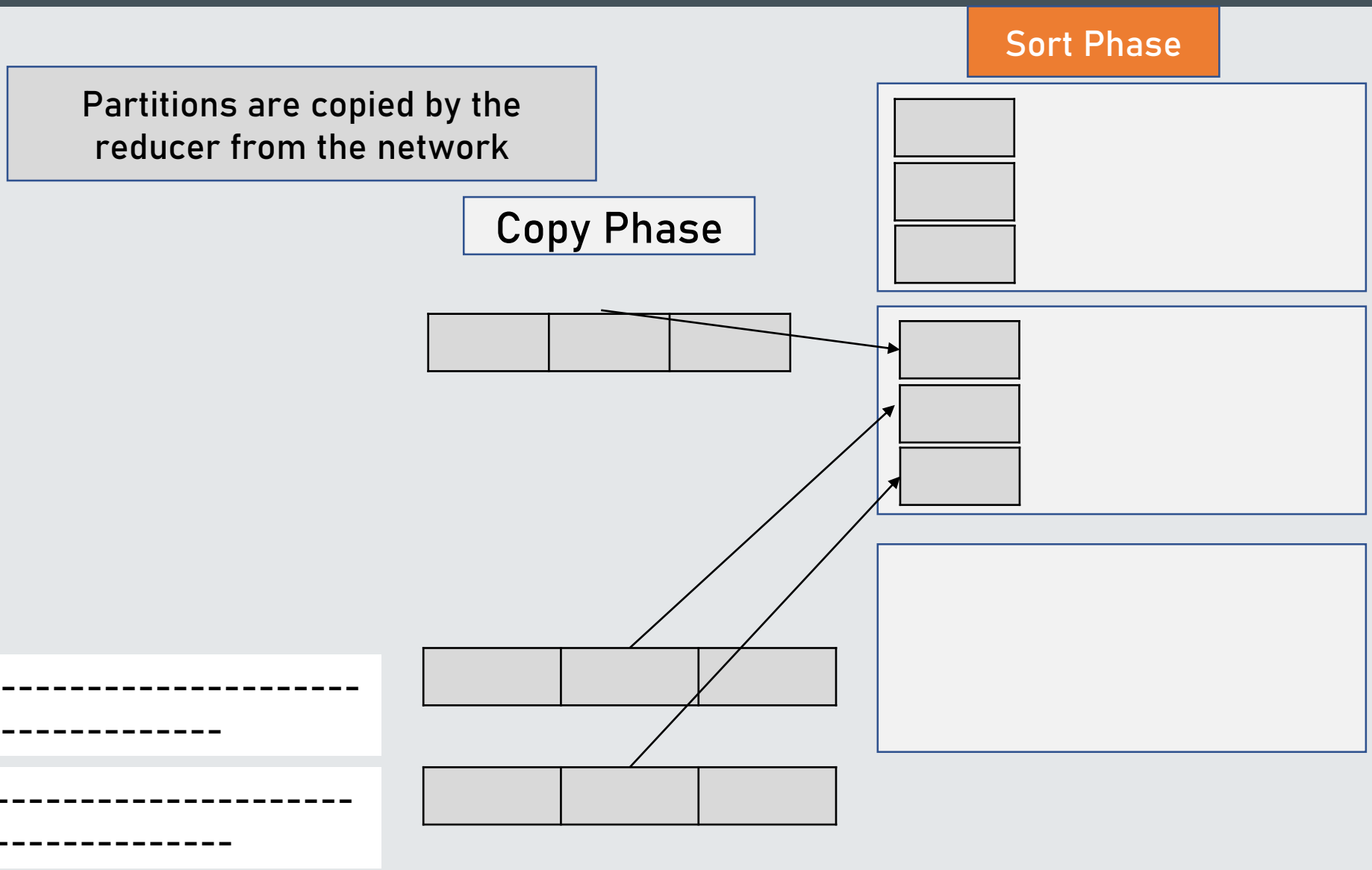
Shuffle and Sort



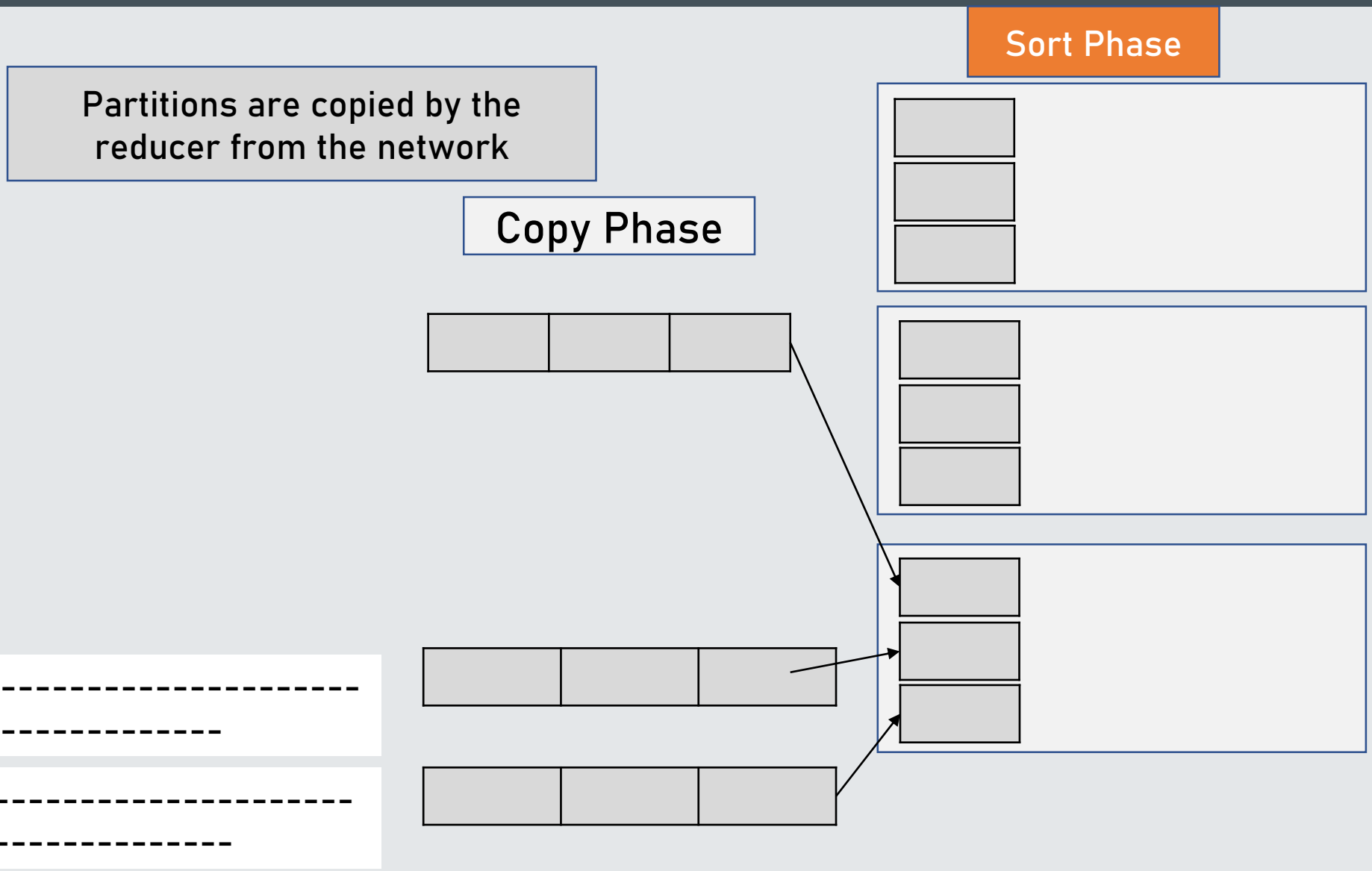
Shuffle and Sort



Shuffle and Sort

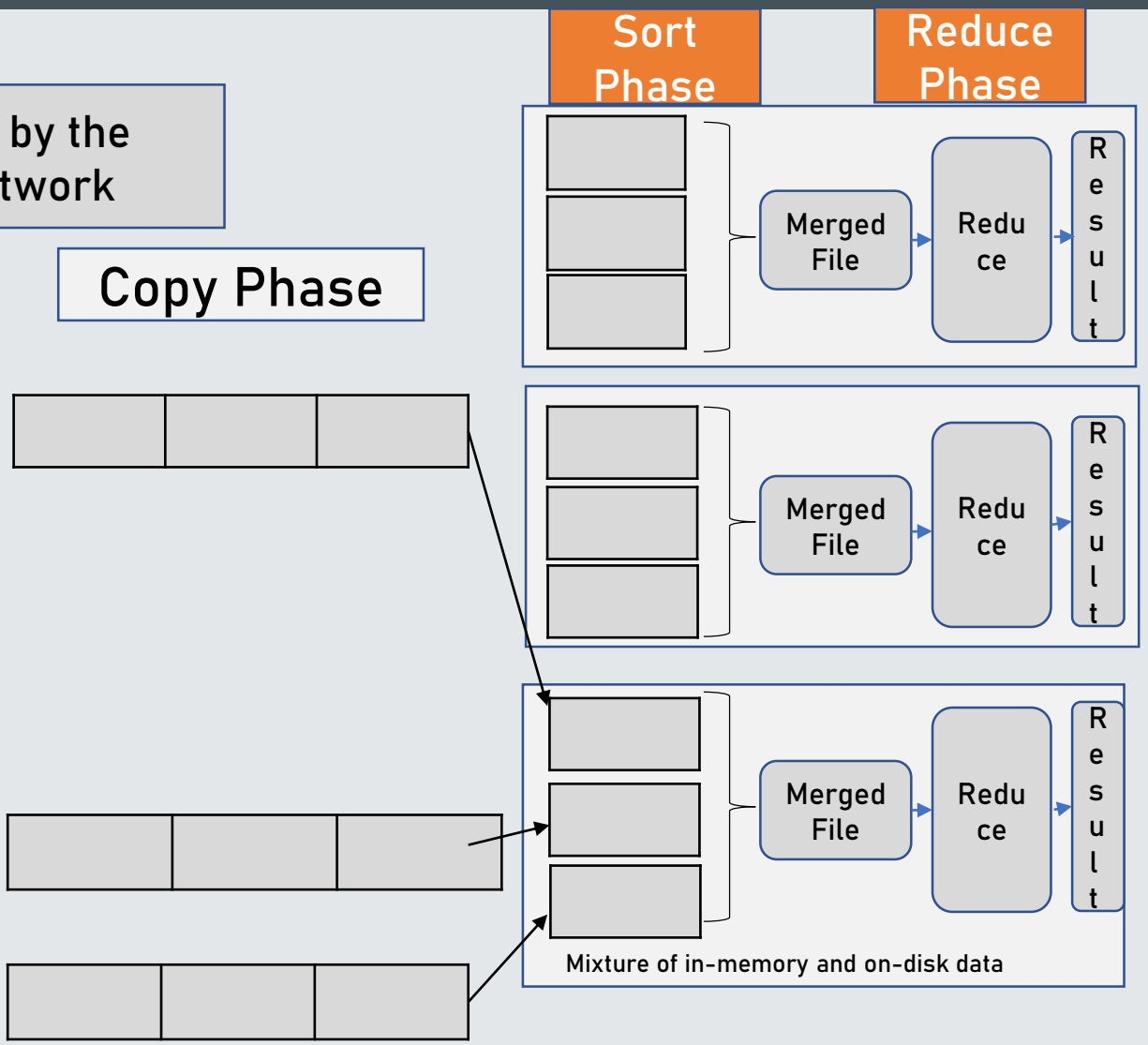


Shuffle and Sort



Shuffle and Sort

Partitions are copied by the reducer from the network

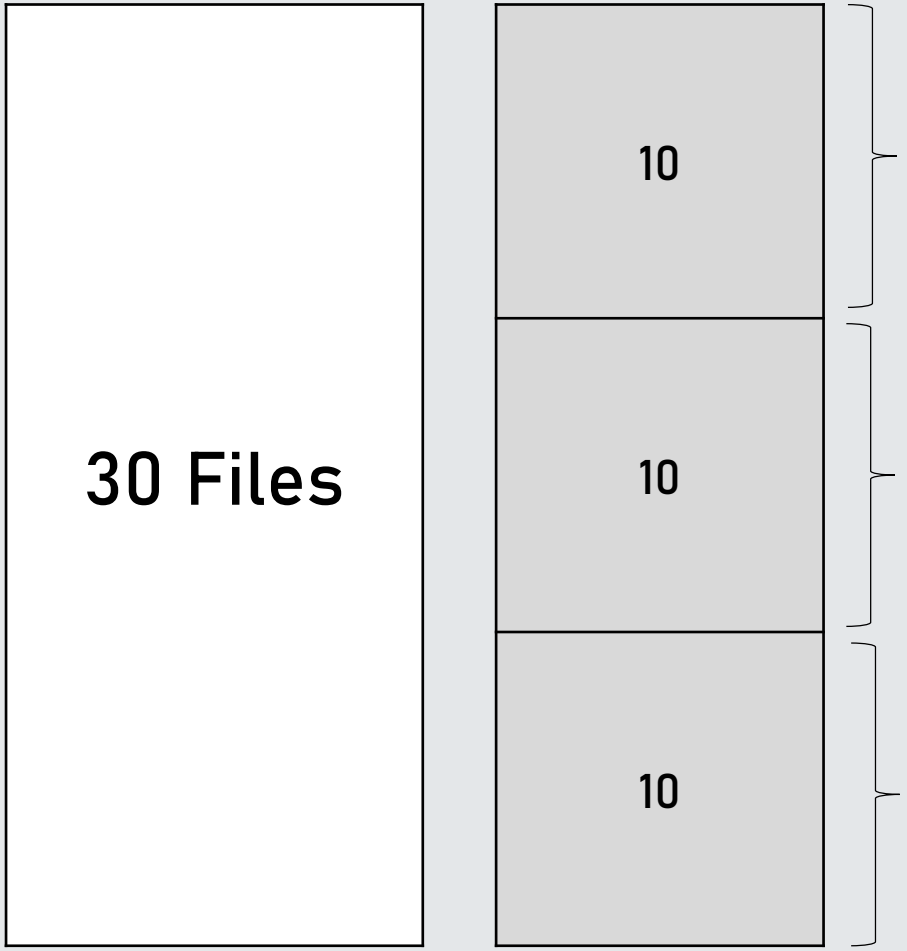


Shuffle and Sort

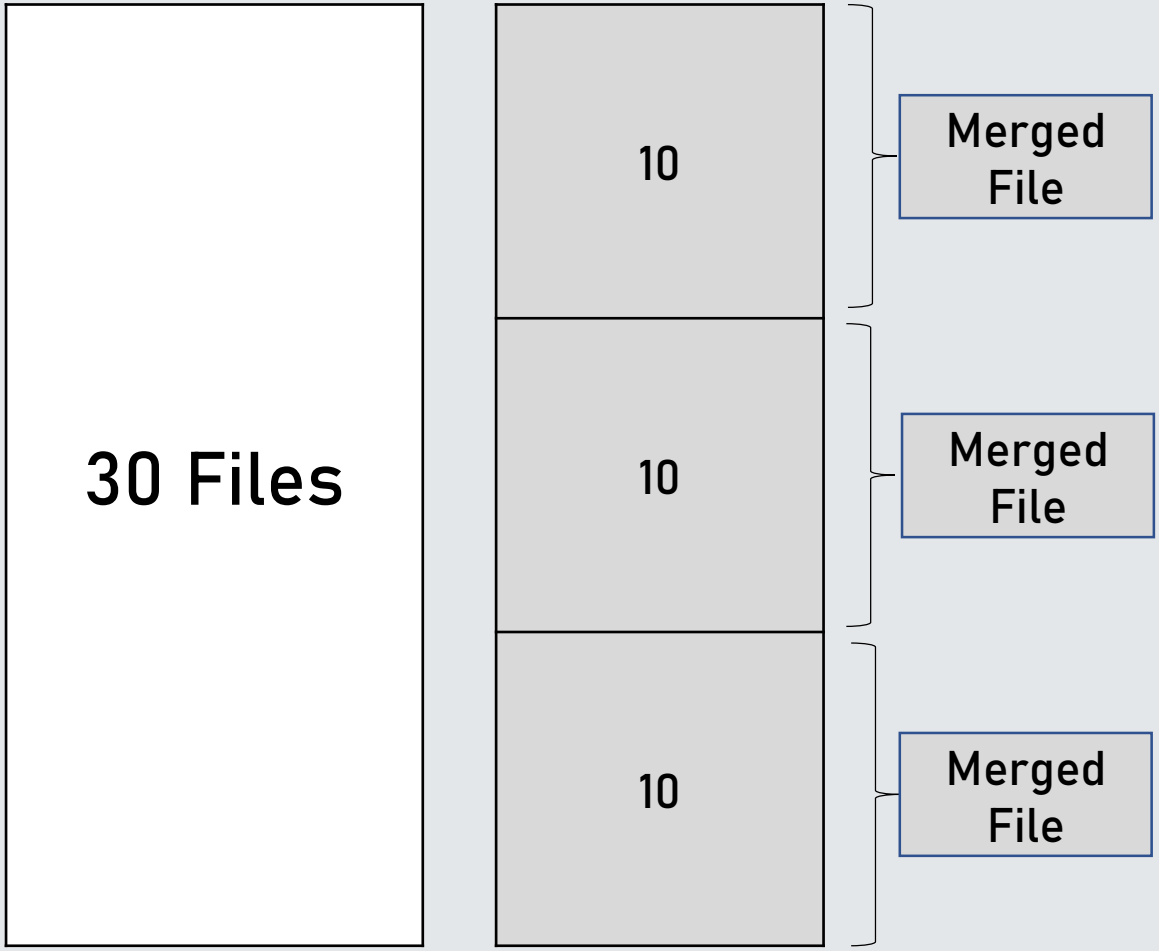
- Merge factor => io.sort.factor
- Its default value is 10

Shuffle and Sort

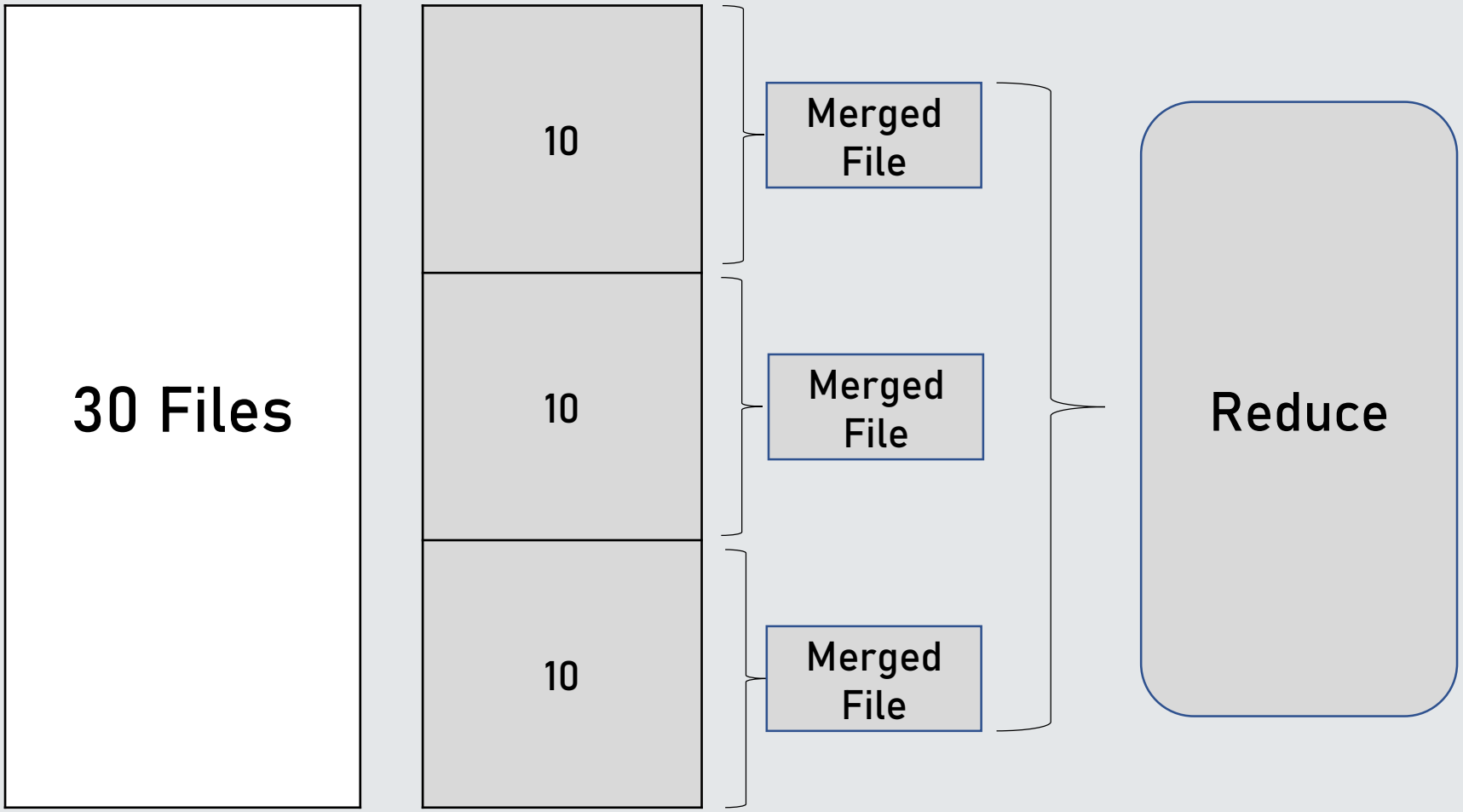
Shuffle and Sort



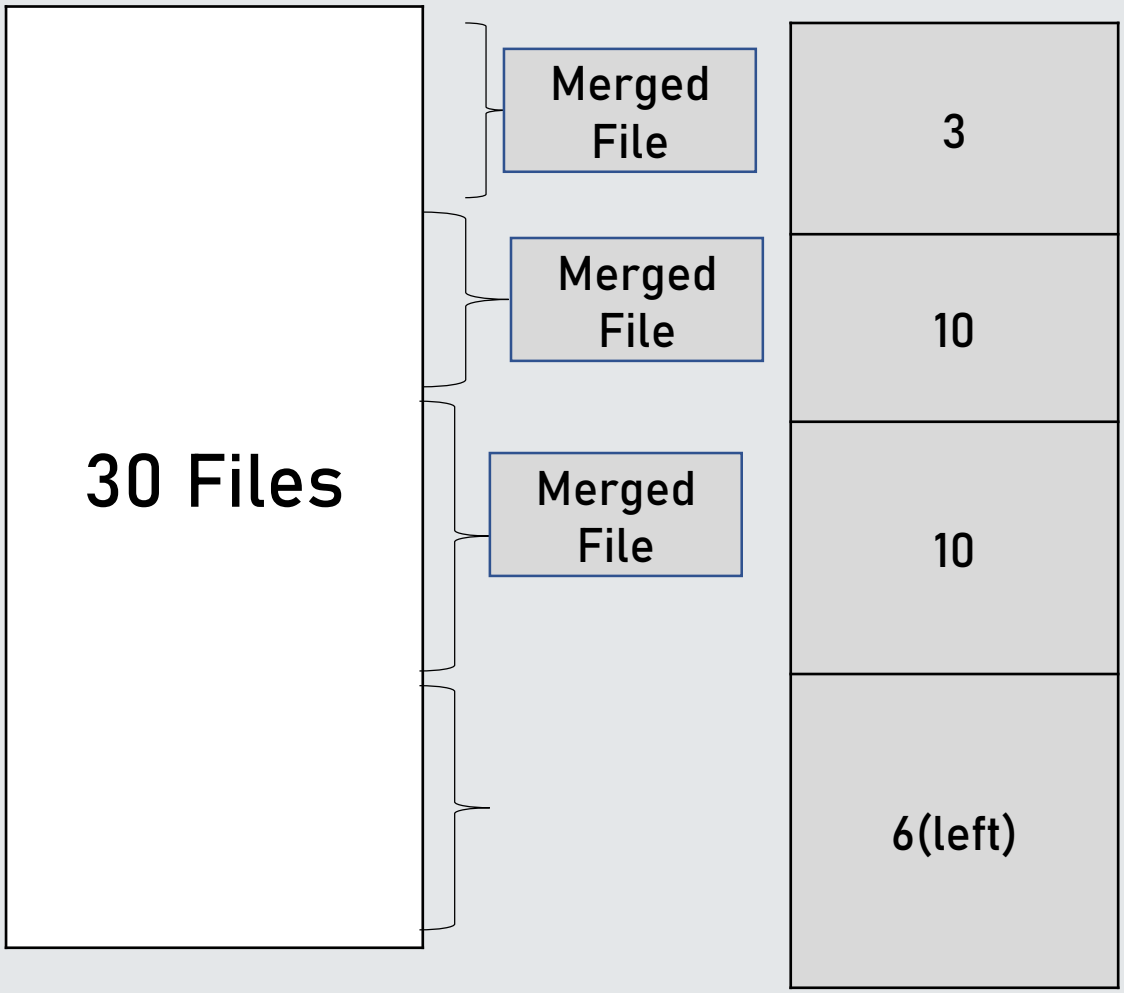
Shuffle and Sort



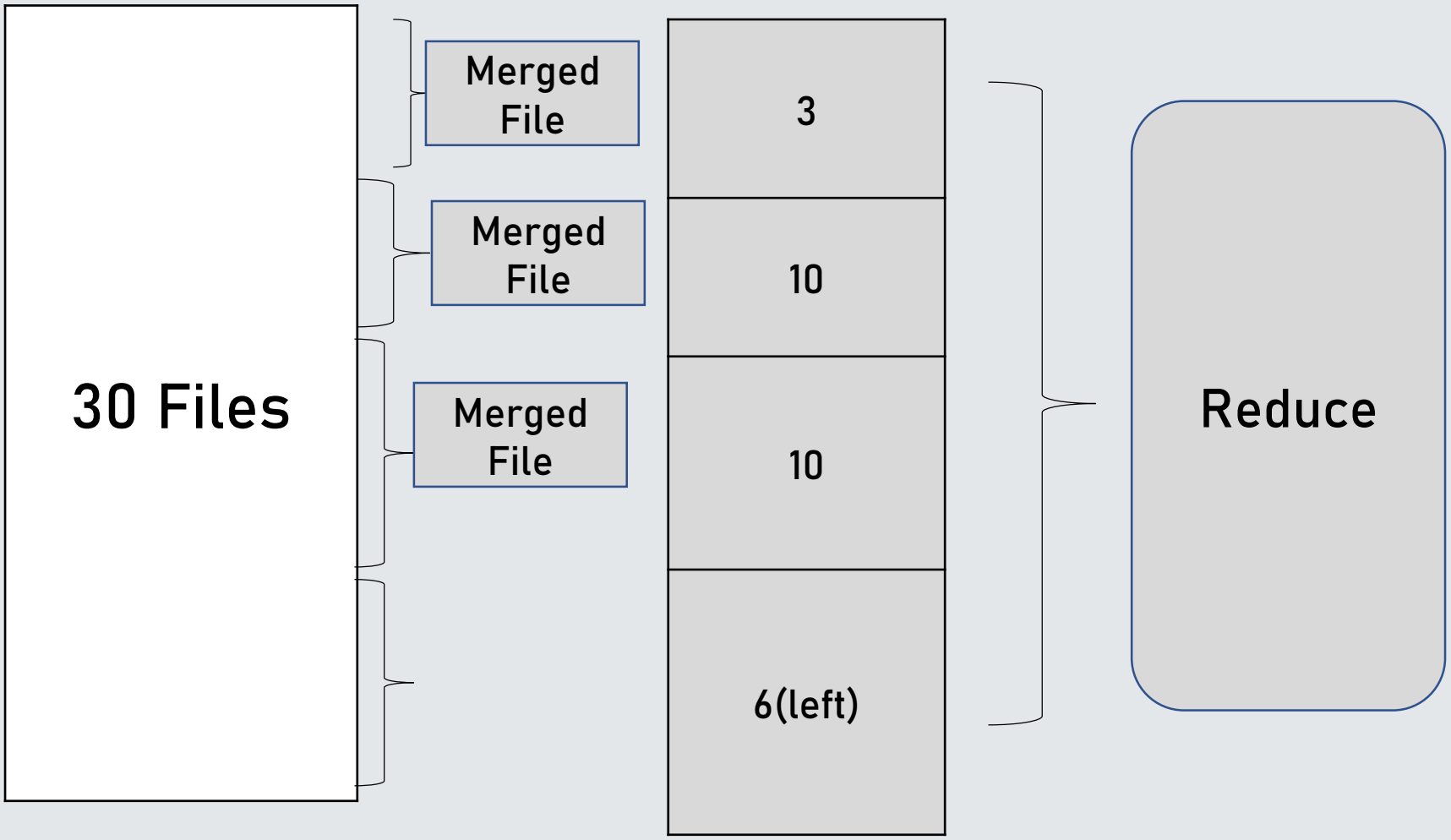
Shuffle and Sort



Shuffle and Sort



Shuffle and Sort

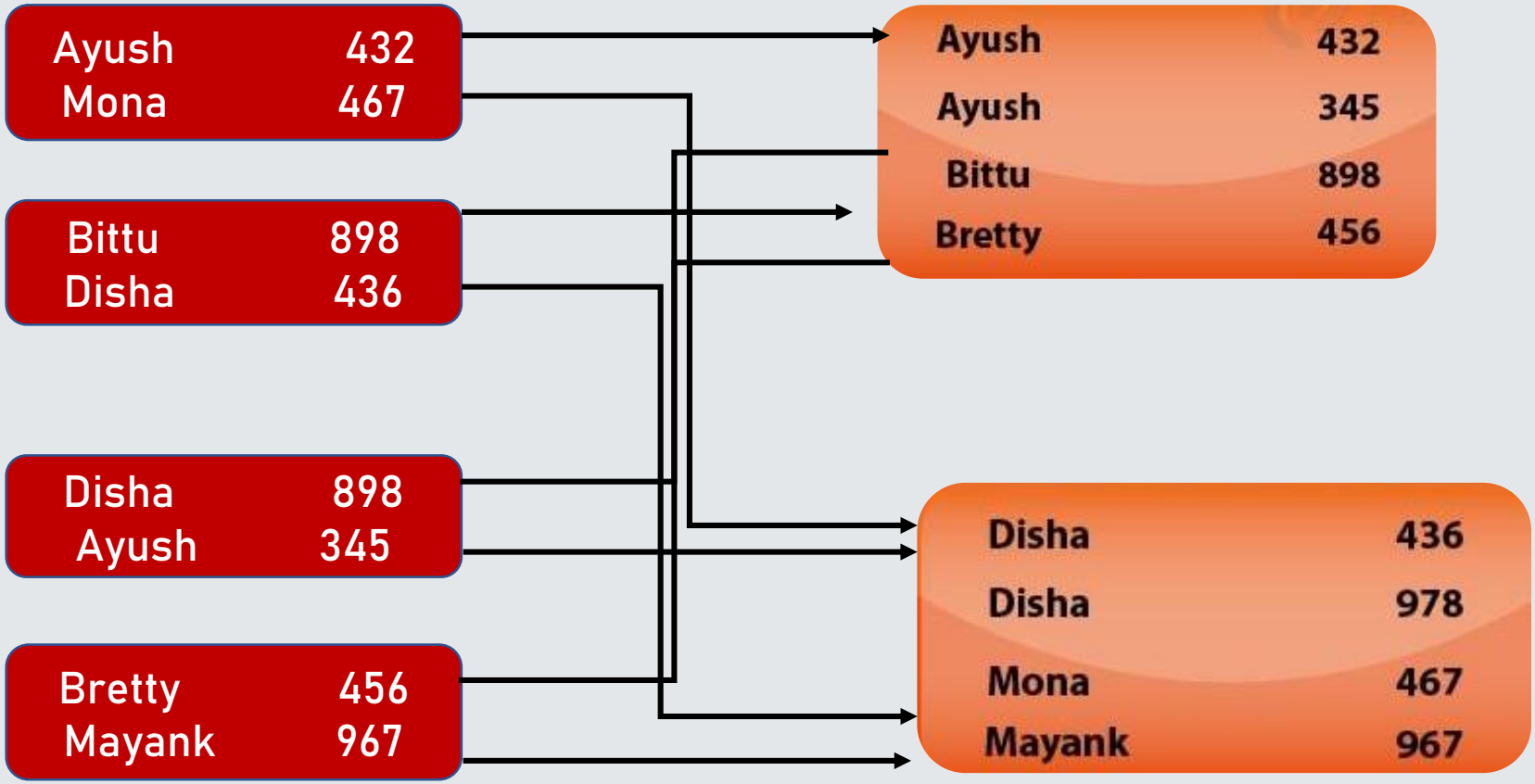


Shuffle and Sort

Objective

- Intermediate output from mappers is transferred to the reducer is called Shuffling.
- Intermediated key-value generated by mapper is sorted automatically by key.

Shuffle and Sort



Shuffling and Sorting in Hadoop MapReduce

Shuffle and Sort

- Shuffle phase.
- Sort phase
- Data from the mapper are grouped by the key, split among reducers and sorted by the key.
- All values associated with the same key.
- Shuffle and sort phase in Hadoop occur simultaneously and are done by the MapReduce framework.

Shuffle and Sort

Shuffling in MapReduce


Shuffle and Sort
Sorting in MapReduce

Shuffle and Sort

Secondary Sorting in MapReduce


That's all for now...