

INTRODUCTION TO BIG DATA

ECAP456

Dr. Rajni Bhalla
Associate Professor

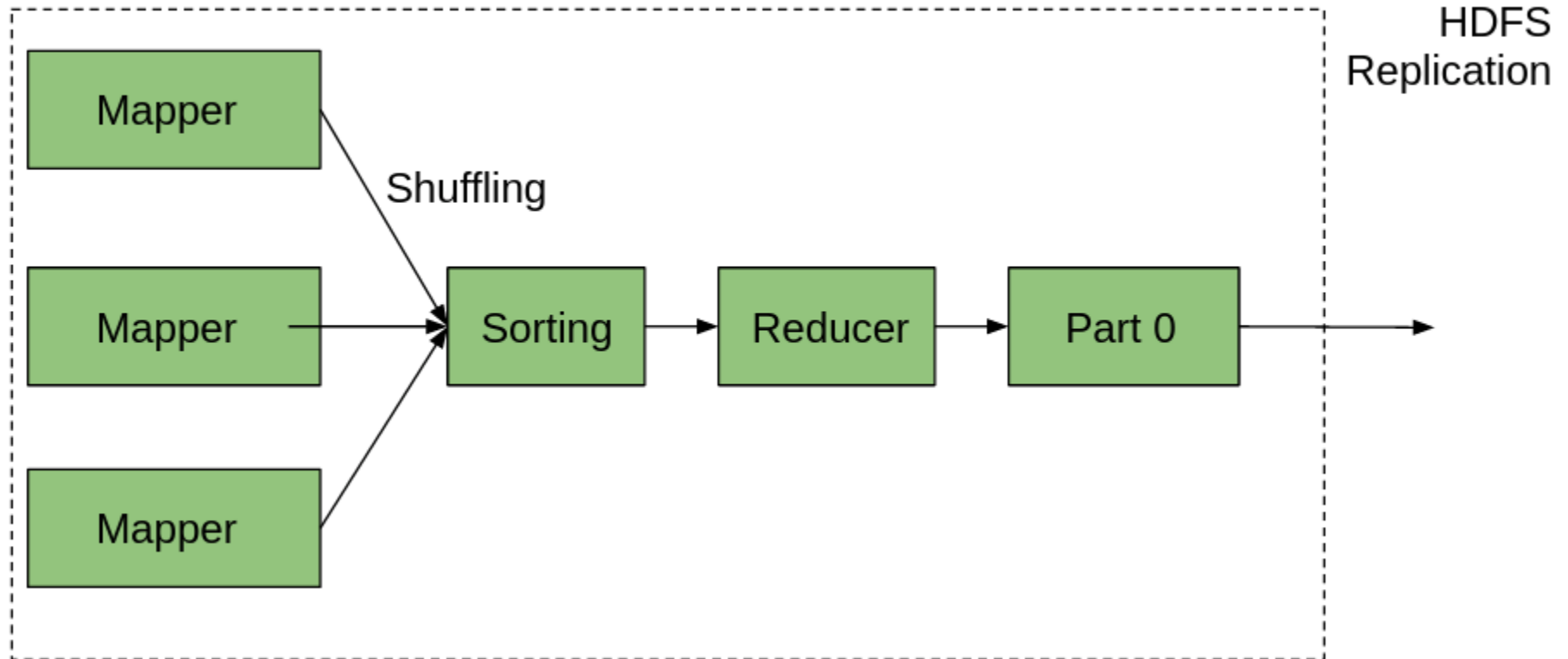
Learning Outcomes



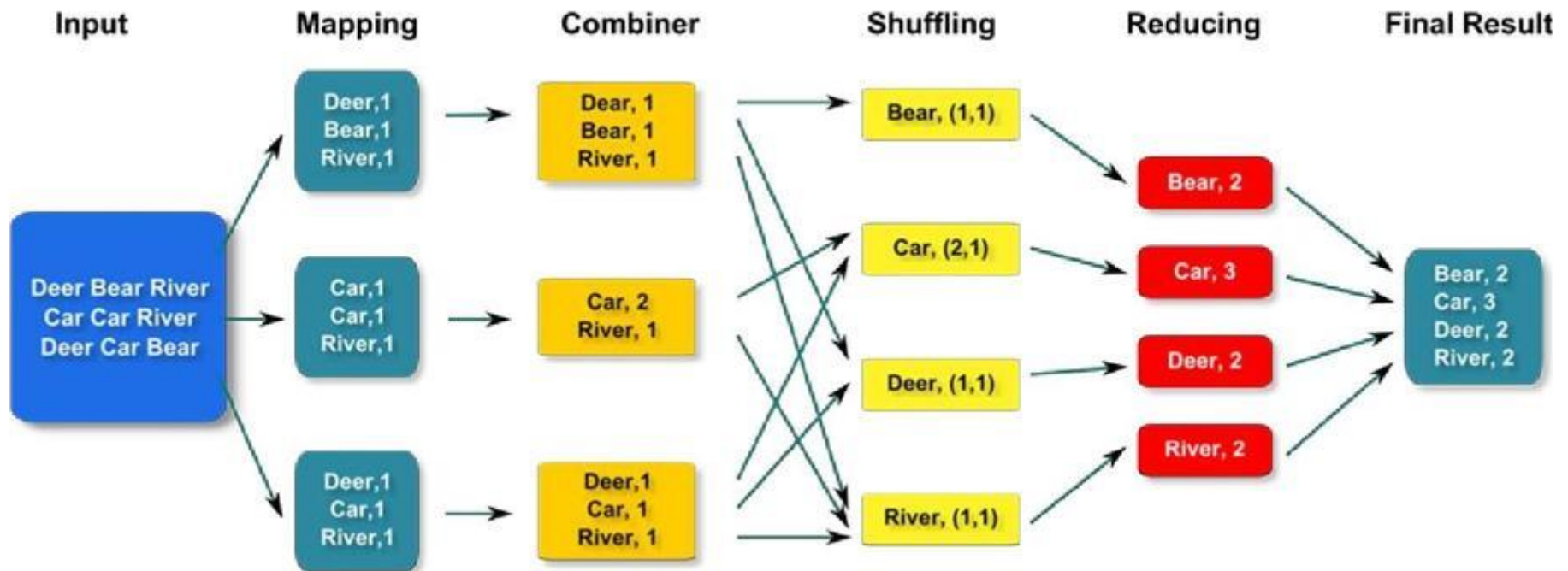
After this lecture, you will be able to

- explore and learn the concepts of Shuffle and Sort

Introduction



Introduction



A photograph of a warehouse conveyor belt system. Several cardboard boxes are in motion on the belt, which is flanked by green safety rails. In the background, there are high industrial shelving units filled with more boxes. The scene is dimly lit, with the primary light source coming from above, creating a professional, industrial atmosphere.

What is Shuffling and Sorting in Hadoop MapReduce?

Phases of MapReduce

Mapper

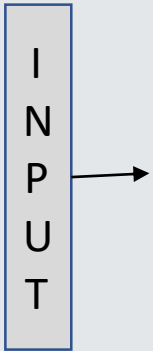
Reducer in MapReduce

Combiner

Partitioner in MapReduce

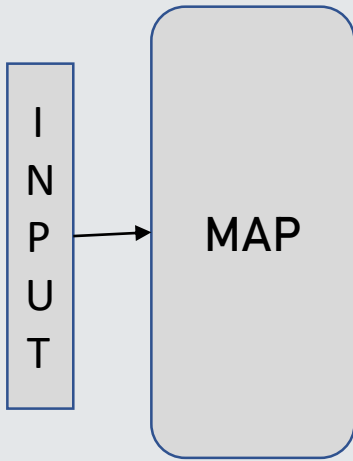
inputFormat
in MapReduce

Shuffle and Sort

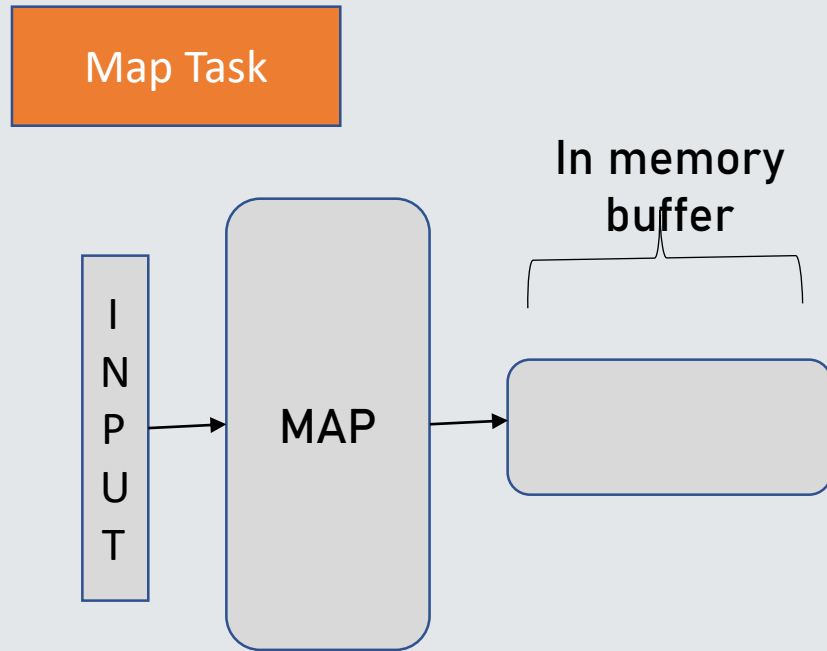


Shuffle and Sort

Map Task

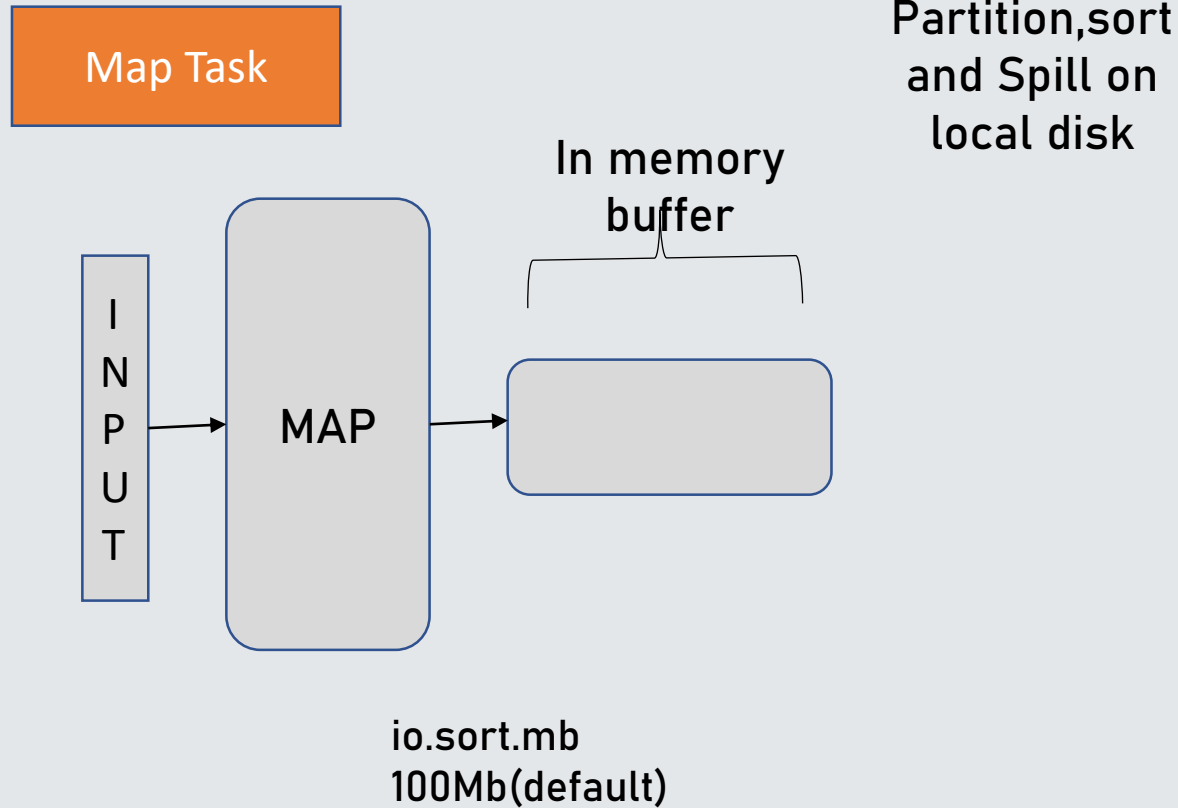


Shuffle and Sort



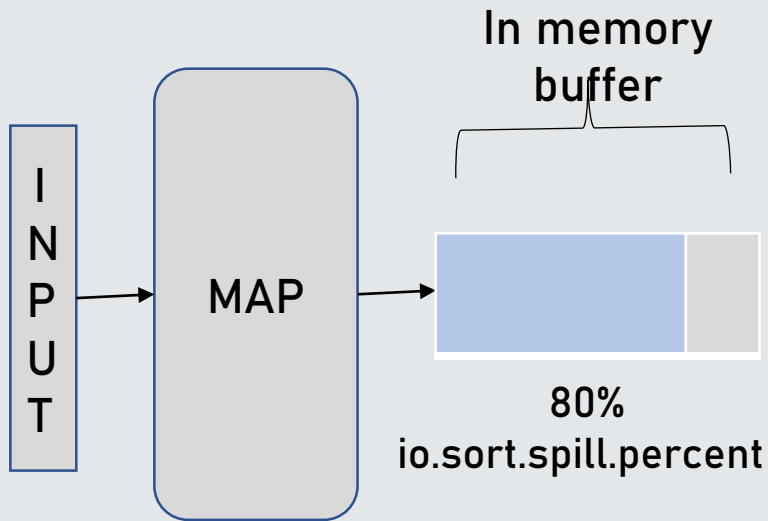
Partition, sort
and Spill on
local disk

Shuffle and Sort



Shuffle and Sort

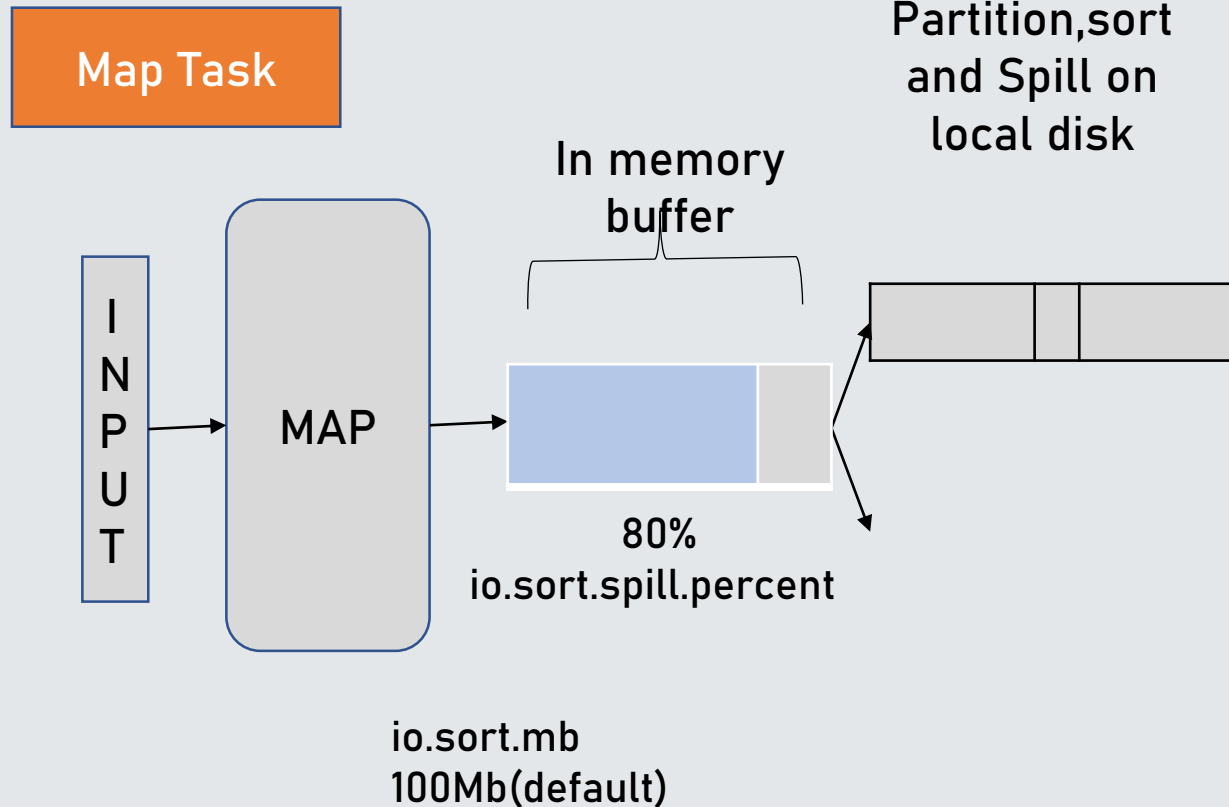
Map Task



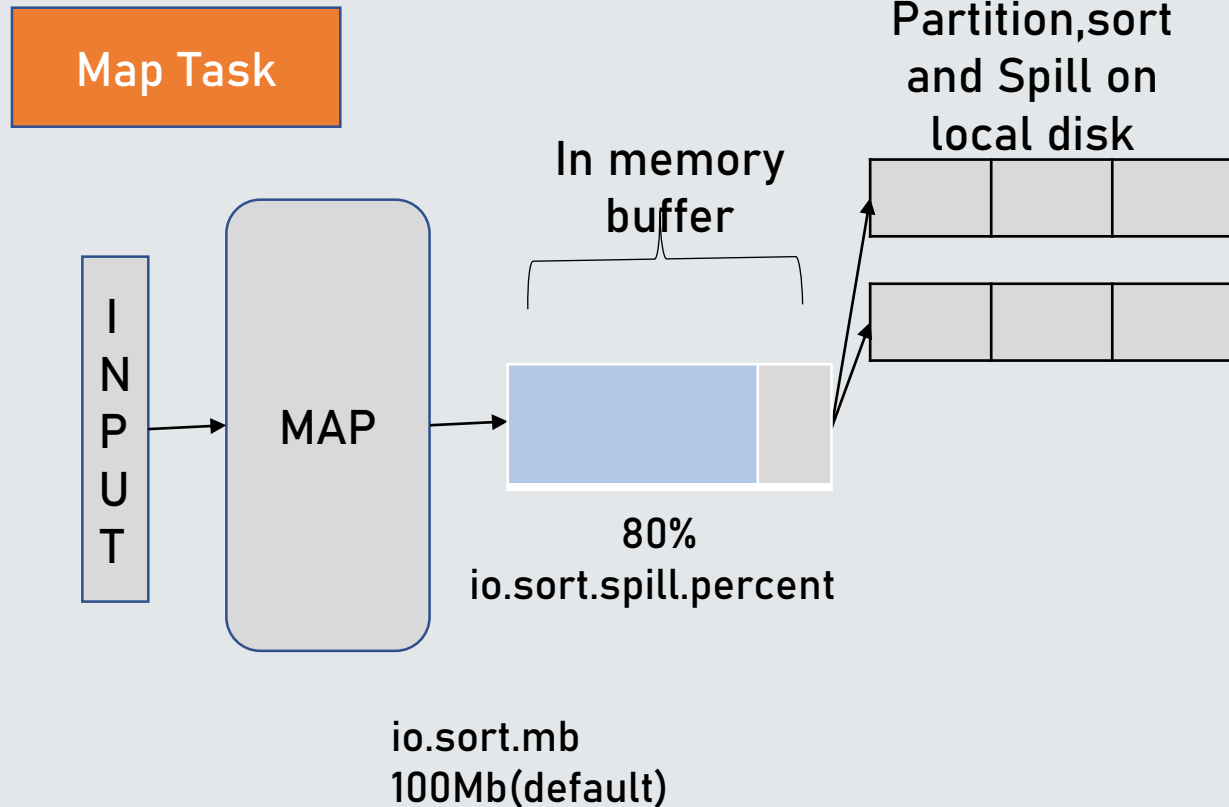
Partition, sort
and Spill on
local disk

io.sort.mb
100Mb(default)

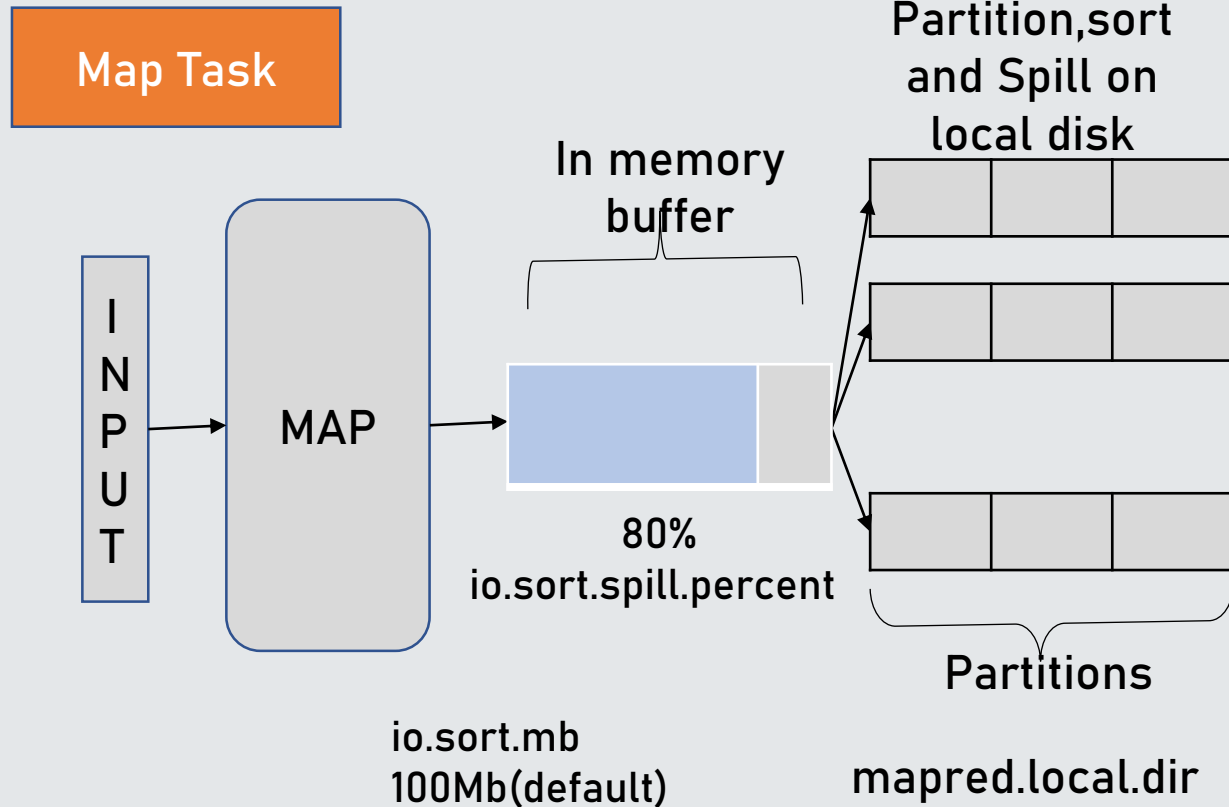
Shuffle and Sort



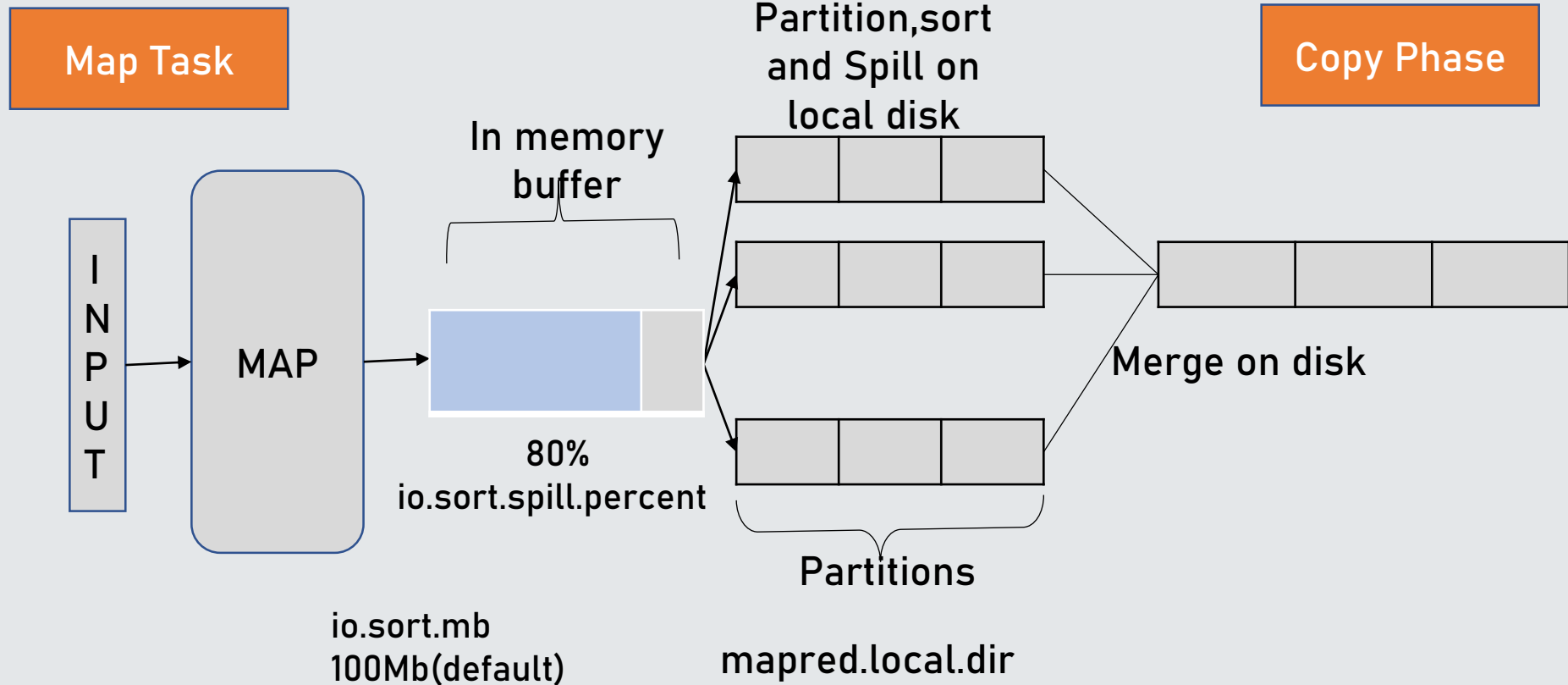
Shuffle and Sort



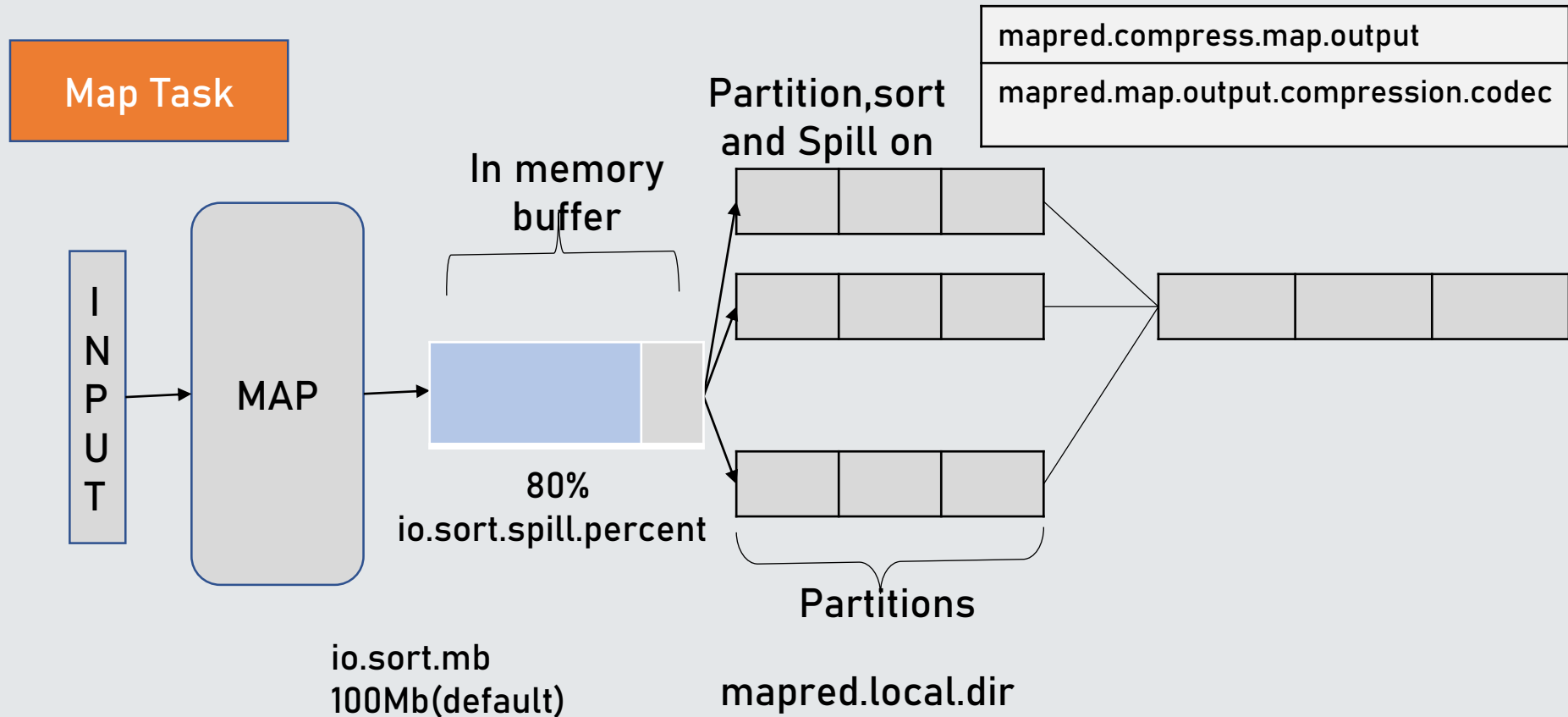
Shuffle and Sort



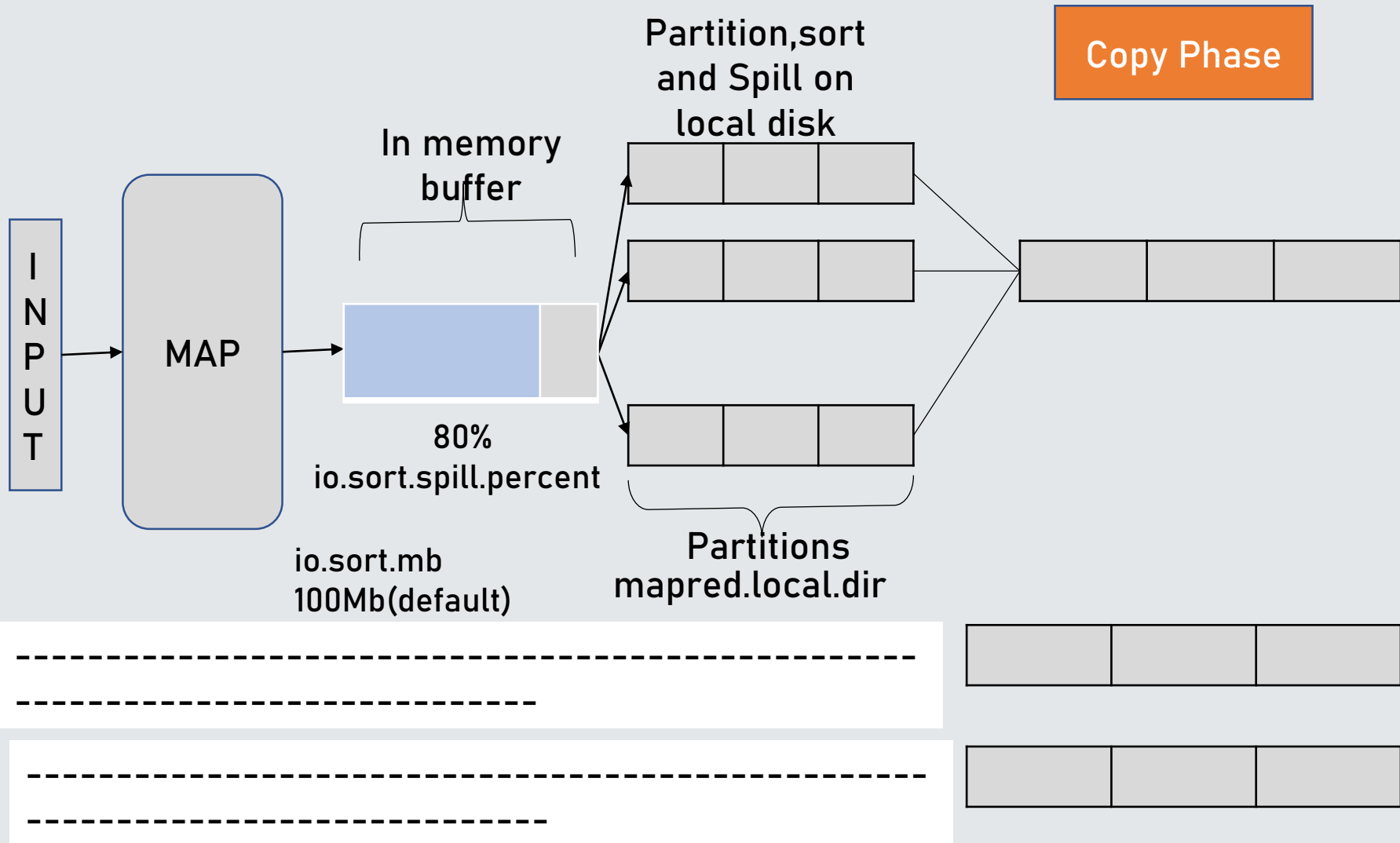
Shuffle and Sort



Shuffle and Sort



Shuffle and Sort

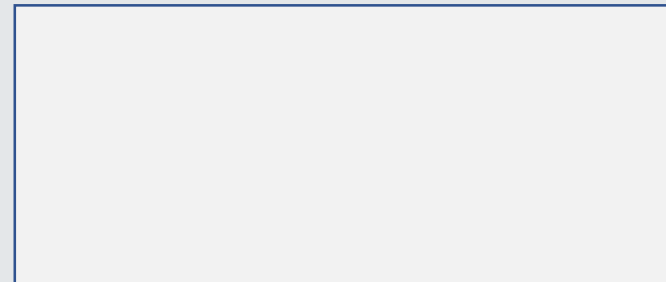
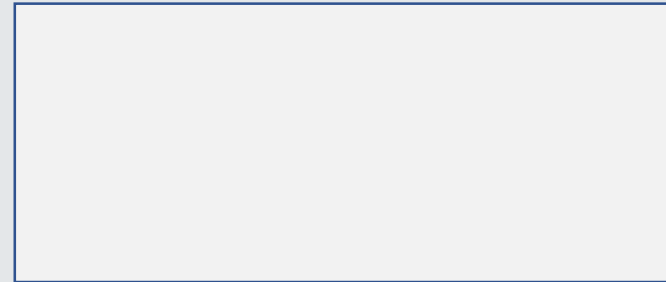
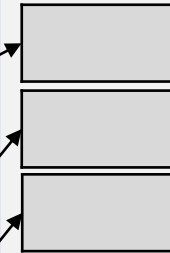
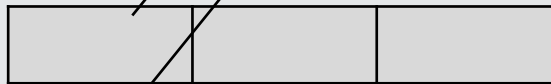
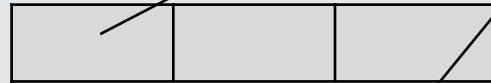


Shuffle and Sort

Partitions are copied by the reducer from the network

Copy Phase

Sort Phase

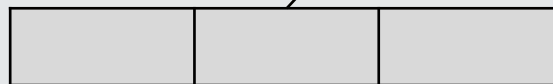
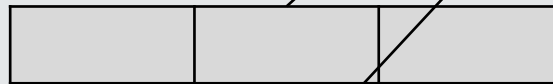
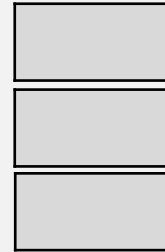
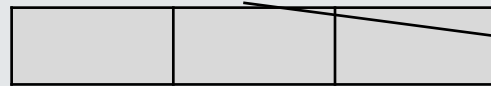


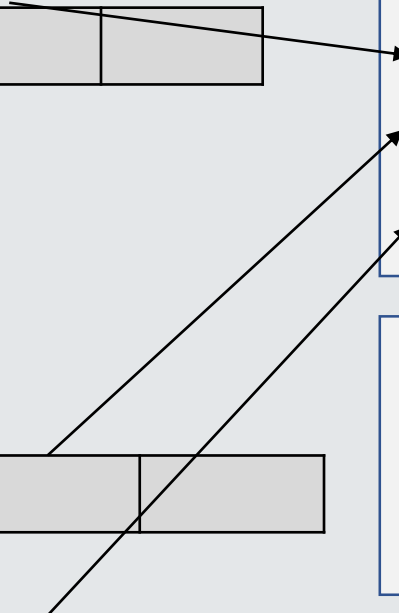
Shuffle and Sort

Partitions are copied by the reducer from the network

Copy Phase

Sort Phase



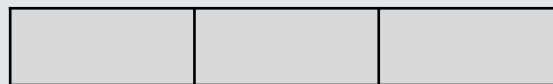
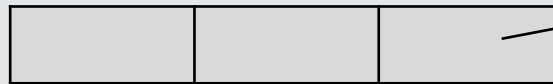
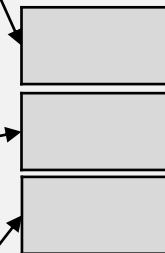
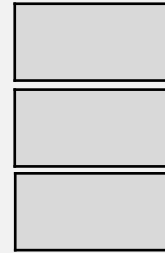
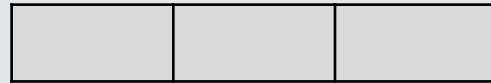


Shuffle and Sort

Partitions are copied by the reducer from the network

Copy Phase

Sort Phase



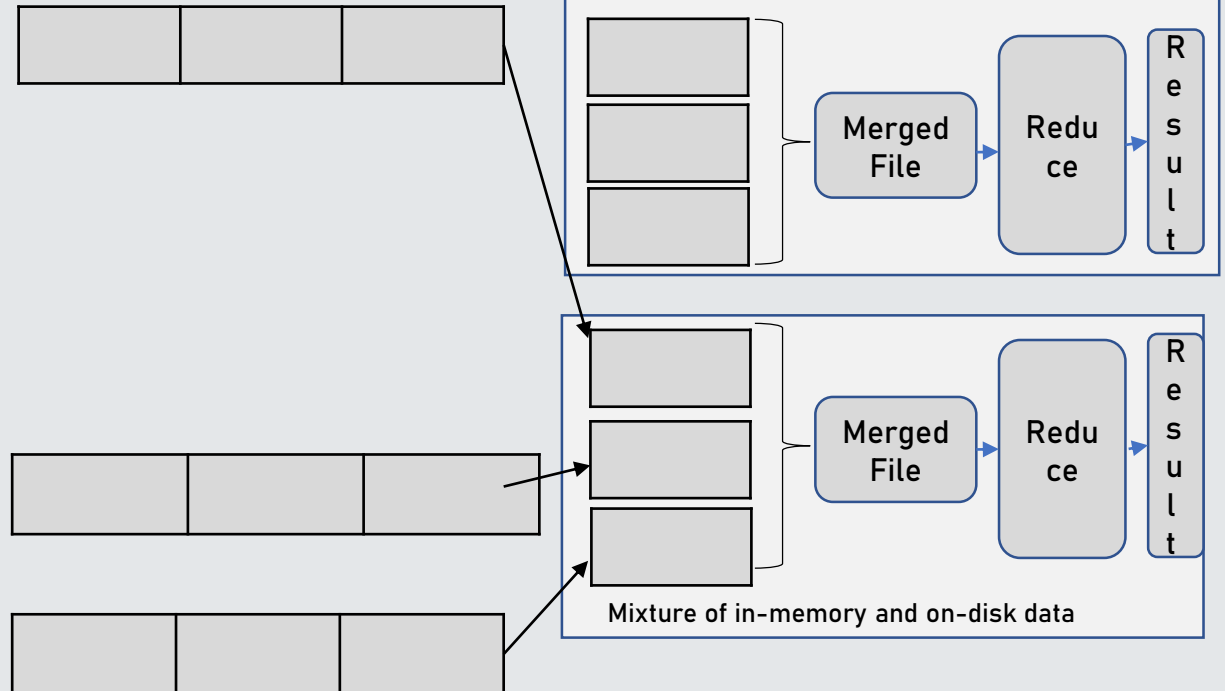
Shuffle and Sort

Partitions are copied by the reducer from the network

Copy Phase

Sort
Phase

Reduce
Phase



Shuffle and Sort

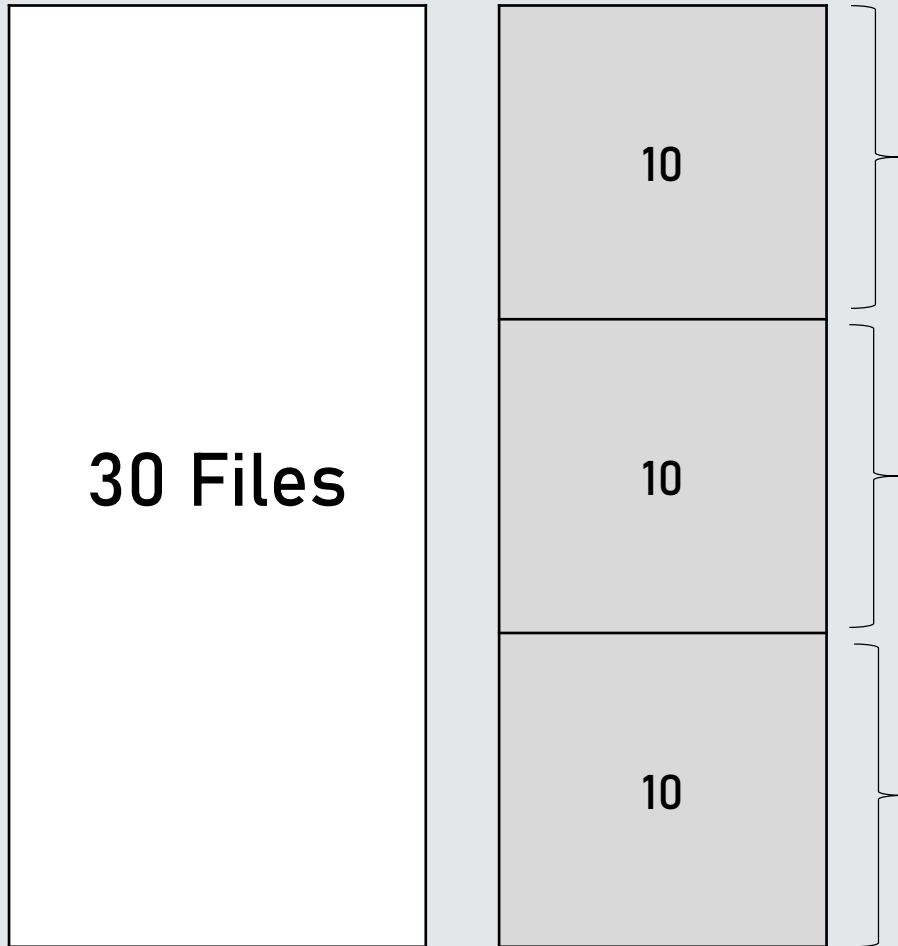
- Merge factor => `io.sort.factor`
- Its default value is 10

Shuffle and Sort

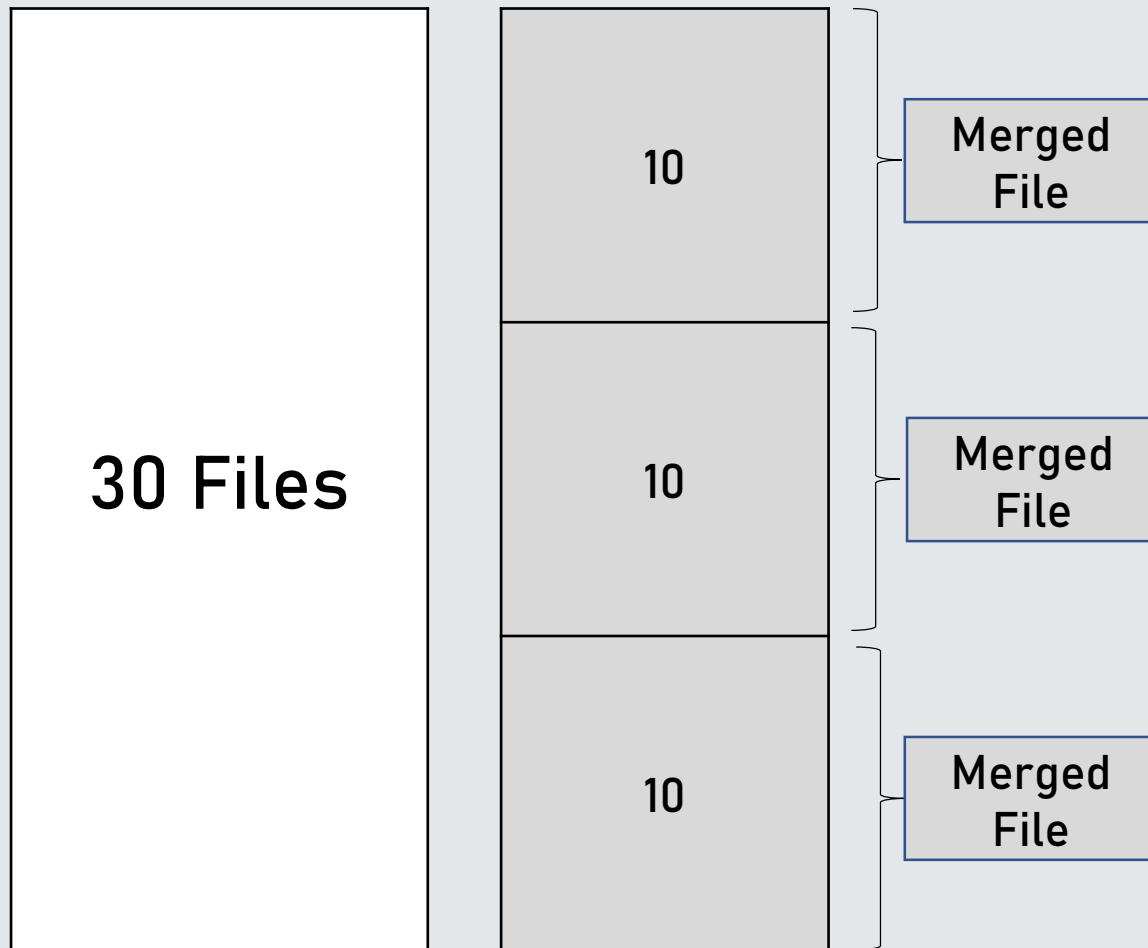


30 Files

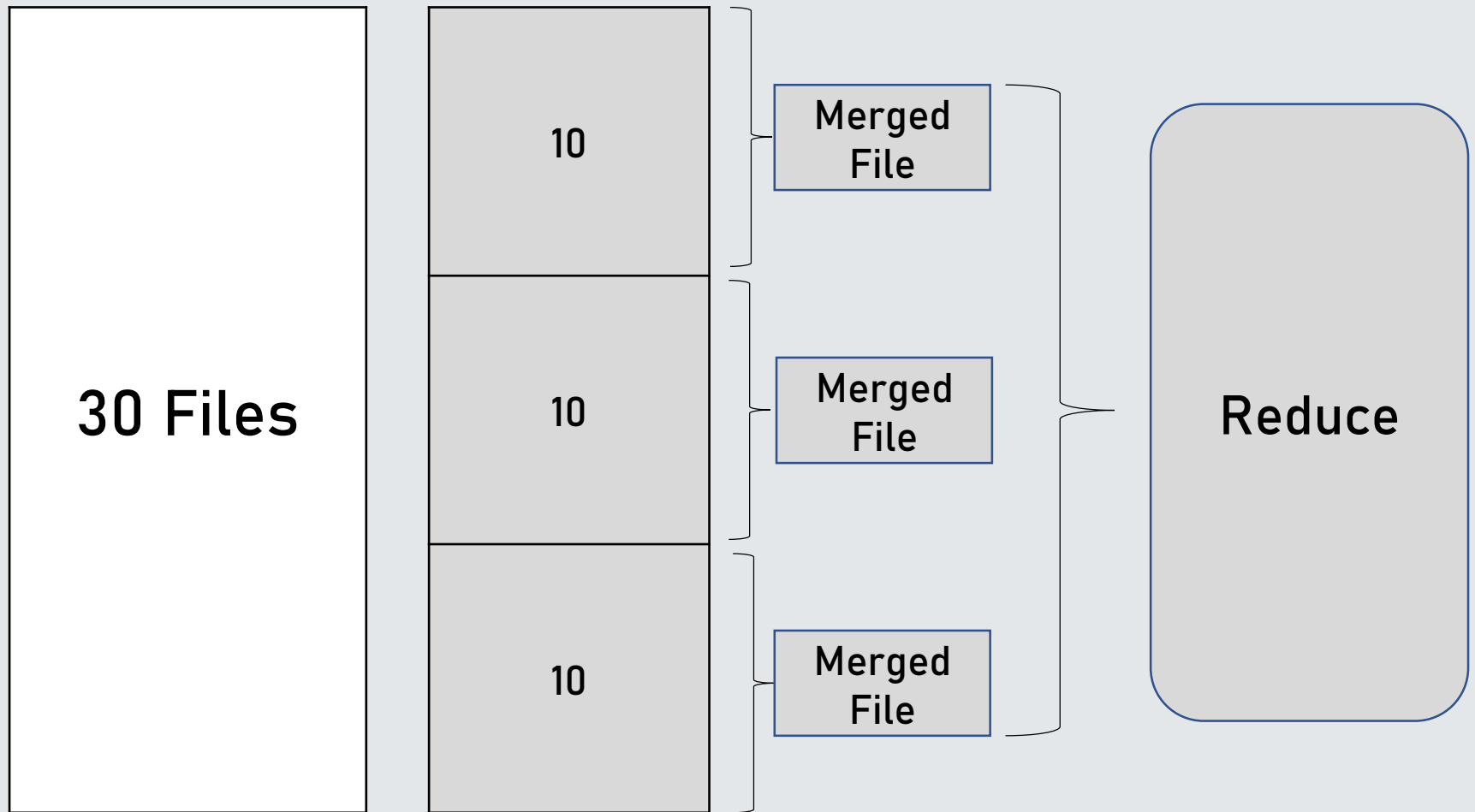
Shuffle and Sort



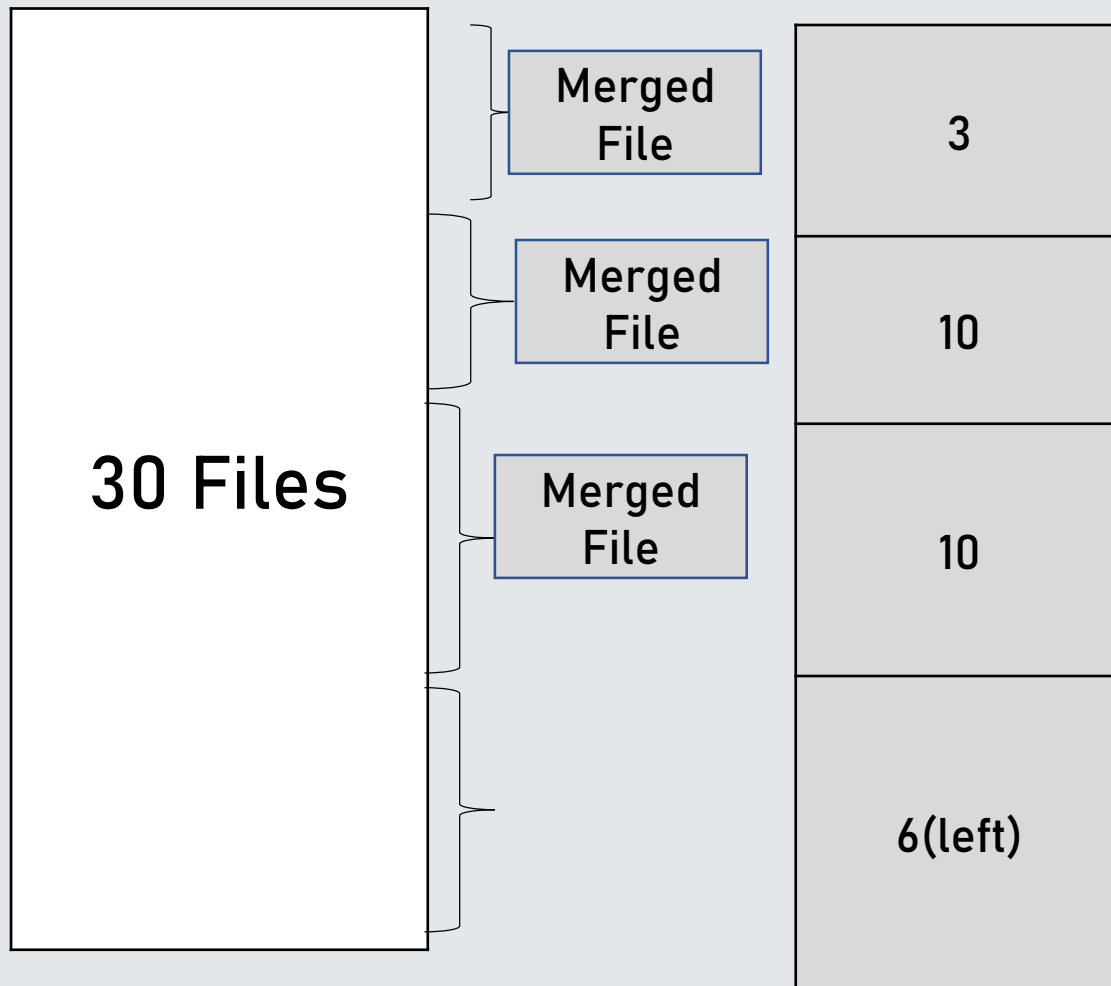
Shuffle and Sort



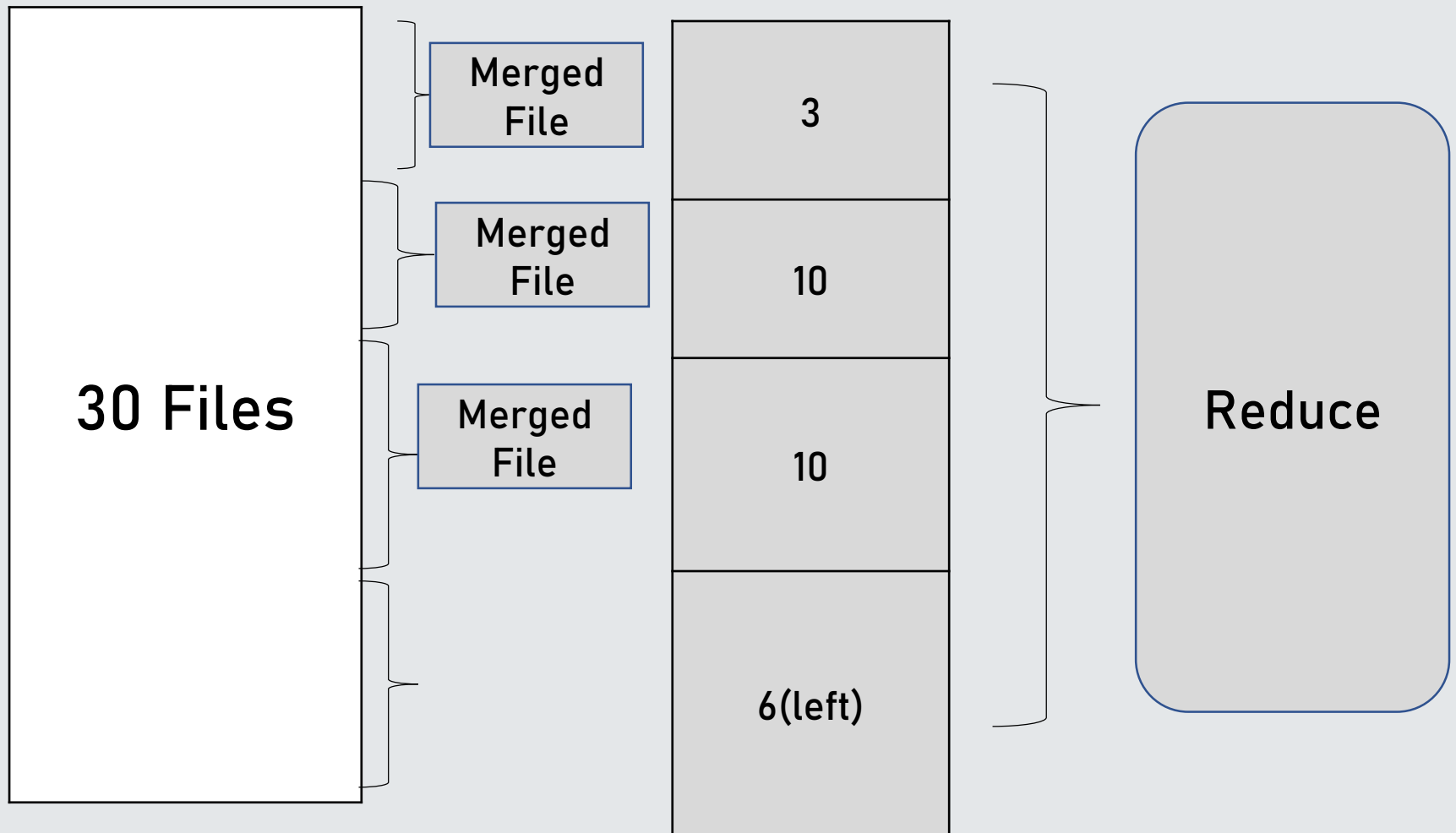
Shuffle and Sort



Shuffle and Sort



Shuffle and Sort

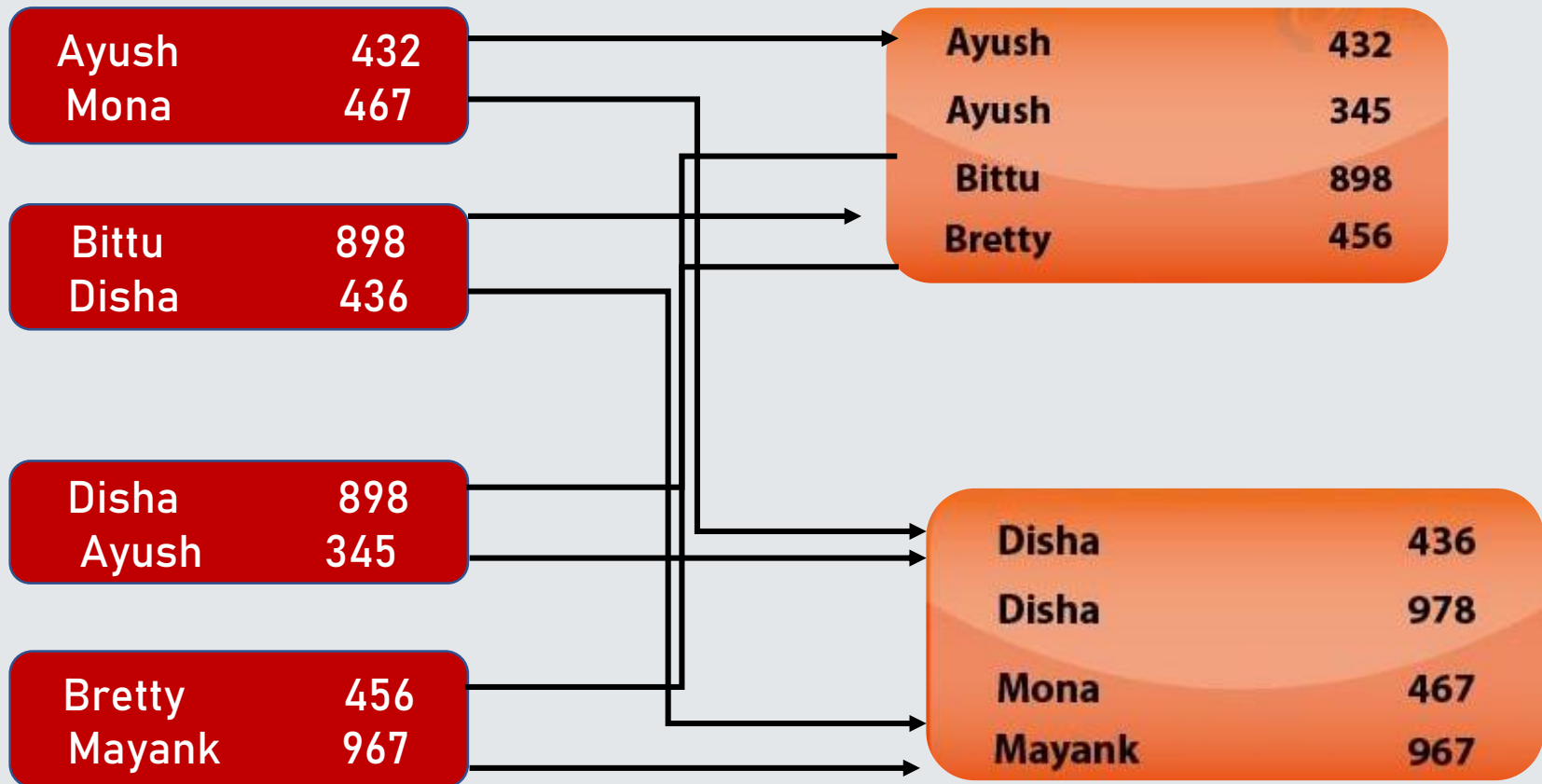


Shuffle and Sort

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...