Introduction to Big Data

ECAP456

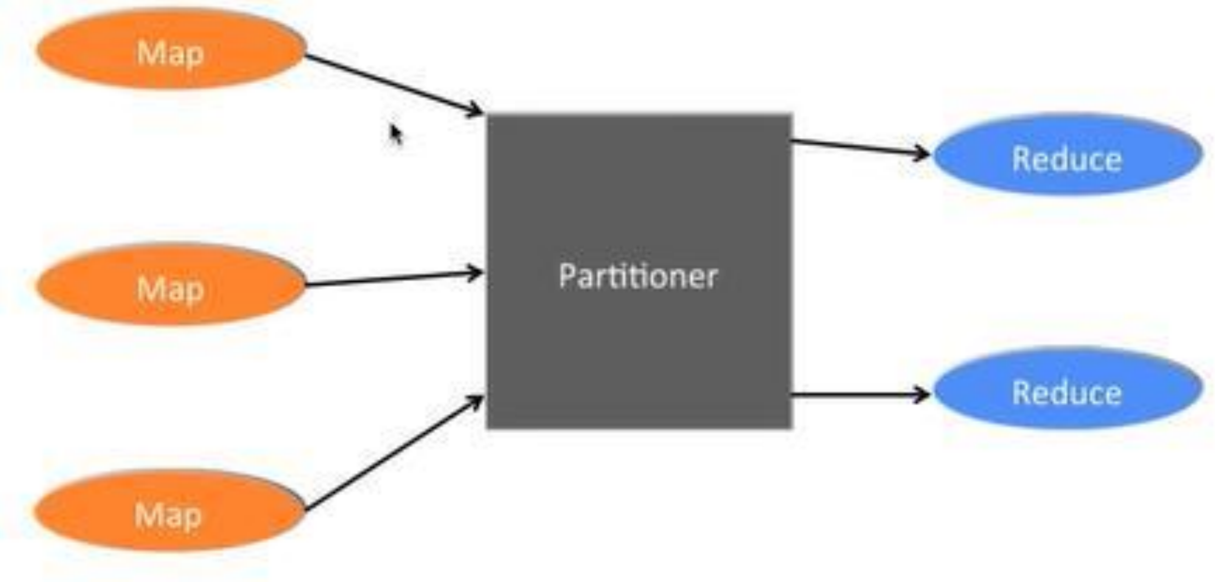
Dr. Rajni Bhalla
Associate Professor

Learning Outcomes

After this lecture, you will be able to

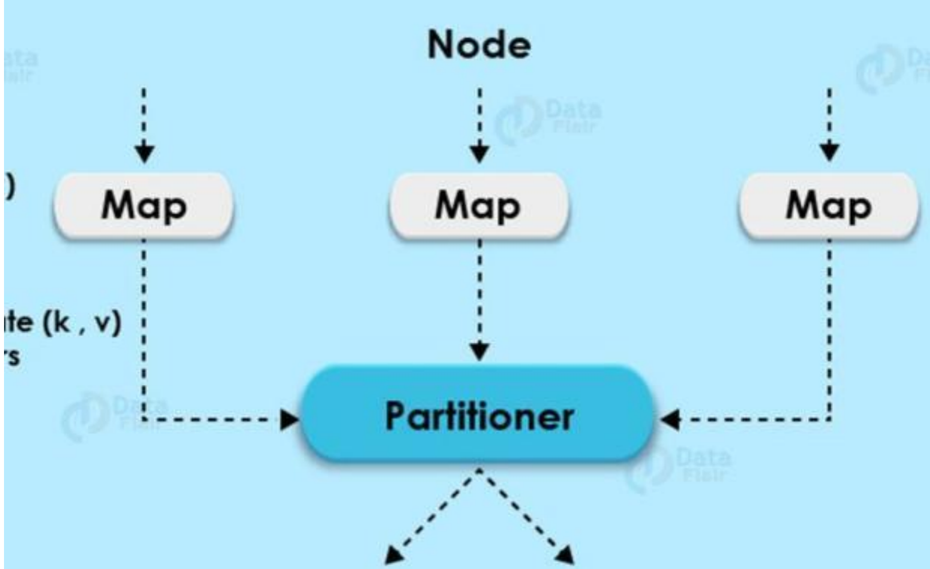
- learn what is hadoop partitioner.
- learn what is the need of partitioner in hadoop
- learn what is the default partitioner in mapreduce,
- learn how many mapreduce partitioner are used in hadoop?
- learn what is mapreduce combiner

Introduction



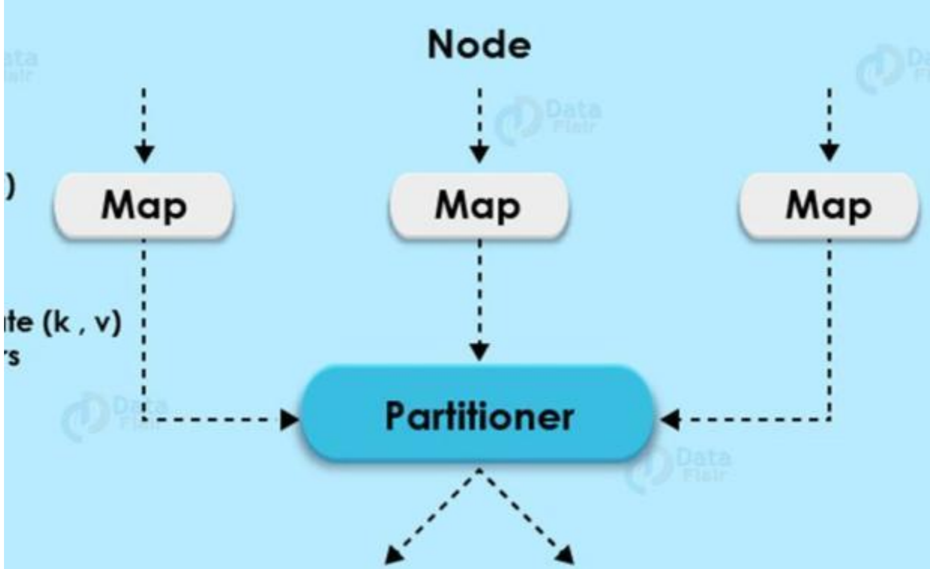
What is Hadoop Partitioner

- Controls the partitioning of the keys
- Keys derive the partition
- Total number of partitions == Number of reduce tasks
- It runs on the same machine
- Entire mapper output sent to partitioner.



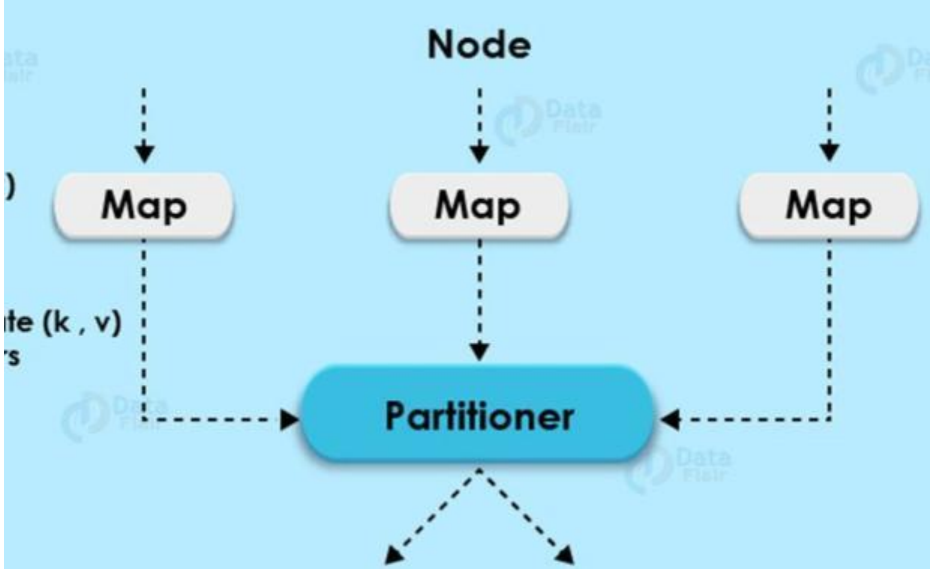
What is Hadoop Partitioner

- Controls the partitioning of the keys
- Keys derive the partition
- Total number of partitions == Number of reduce tasks
- It runs on the same machine
- Entire mapper output sent to partitioner.



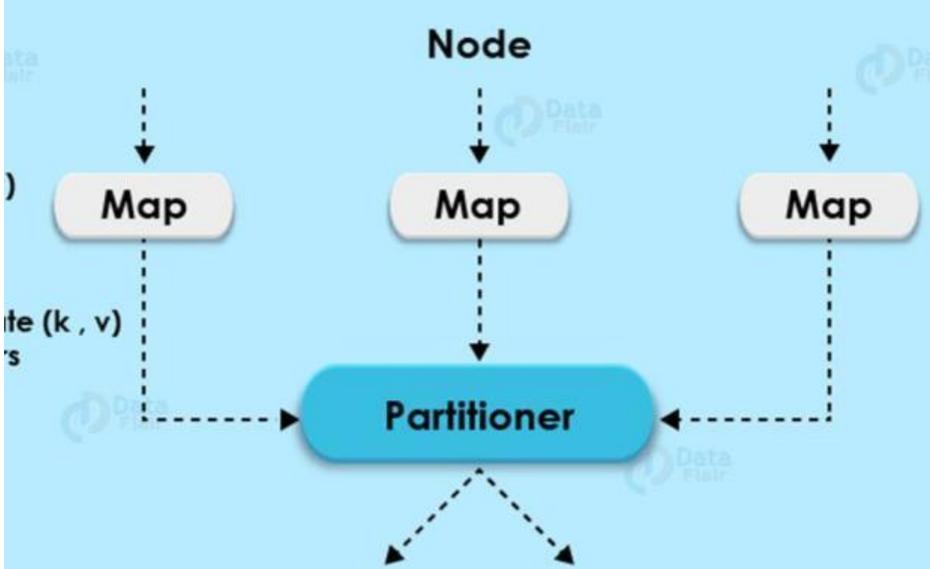
What is Hadoop Partitioner

- Controls the partitioning of the keys
- Keys derive the partition
- Total number of partitions == Number of reduce tasks
- It runs on the same machine
- Entire mapper output sent to partitioner.



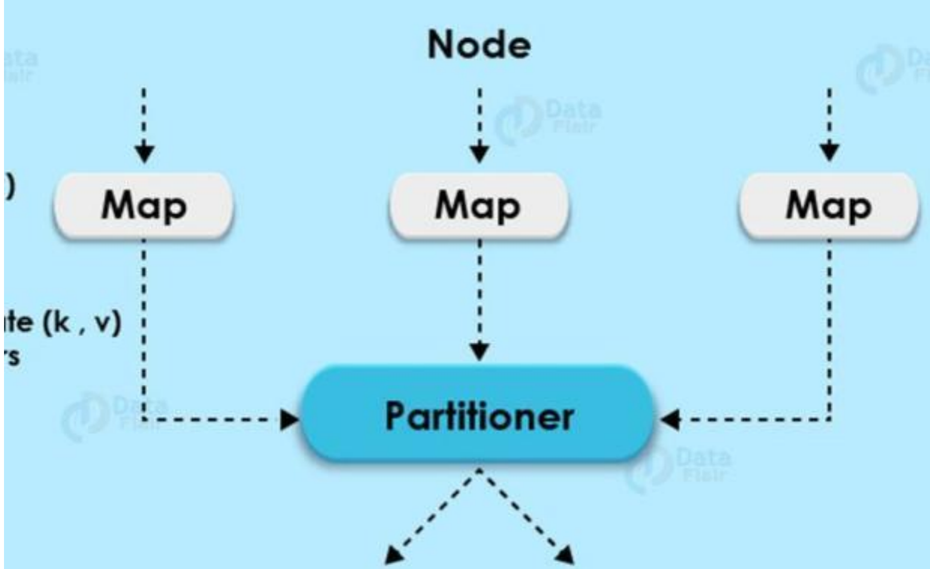
What is Hadoop Partitioner

- Controls the partitioning of the keys
- Keys derive the partition
- Total number of partitions == Number of reduce tasks
- It runs on the same machine
- Entire mapper output sent to partitioner.



What is Hadoop Partitioner

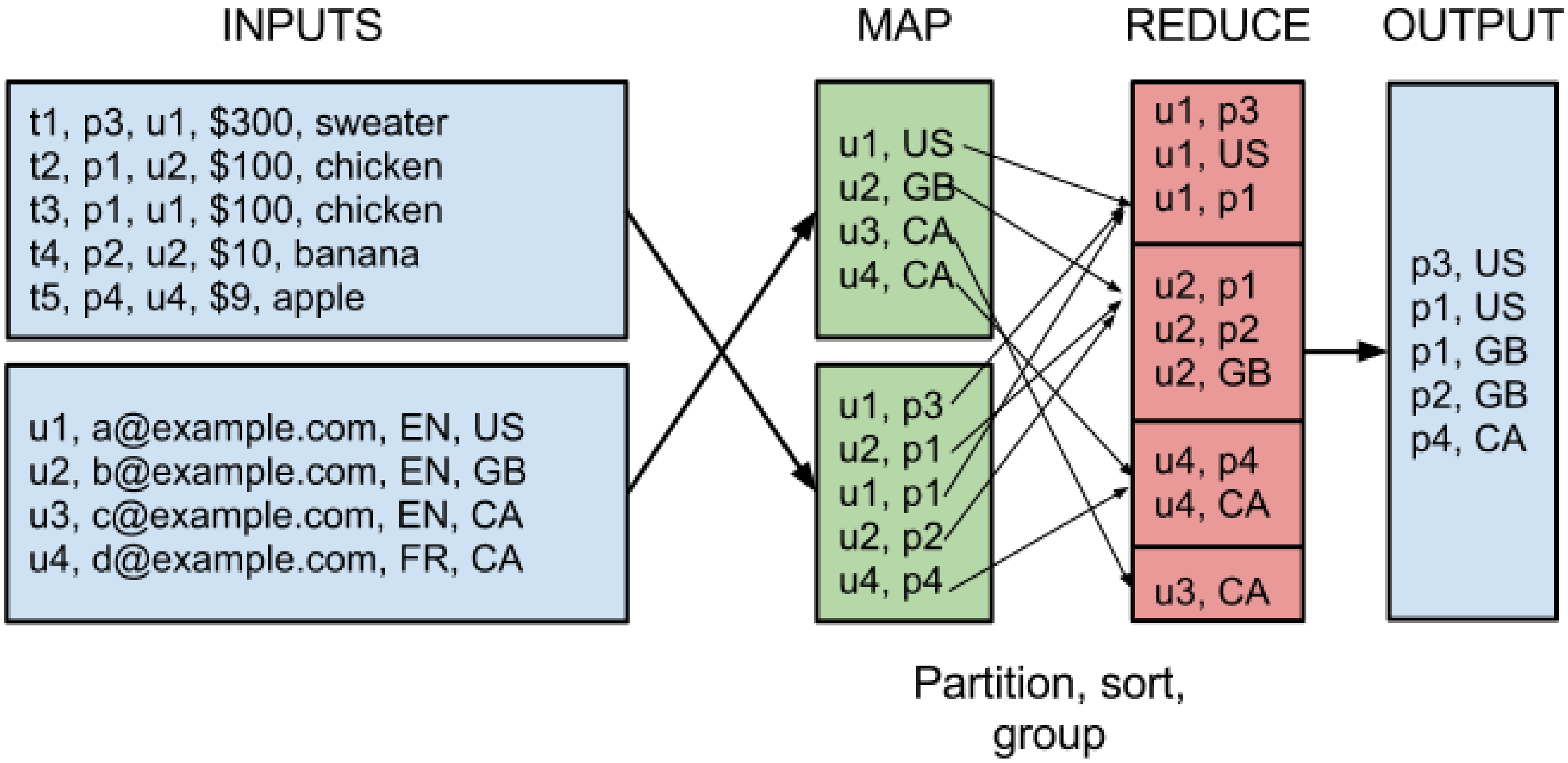
- Controls the partitioning of the keys
- Keys derive the partition
- Total number of partitions == Number of reduce tasks
- It runs on the same machine
- Entire mapper output sent to partitioner.



What is Hadoop Partitioner.

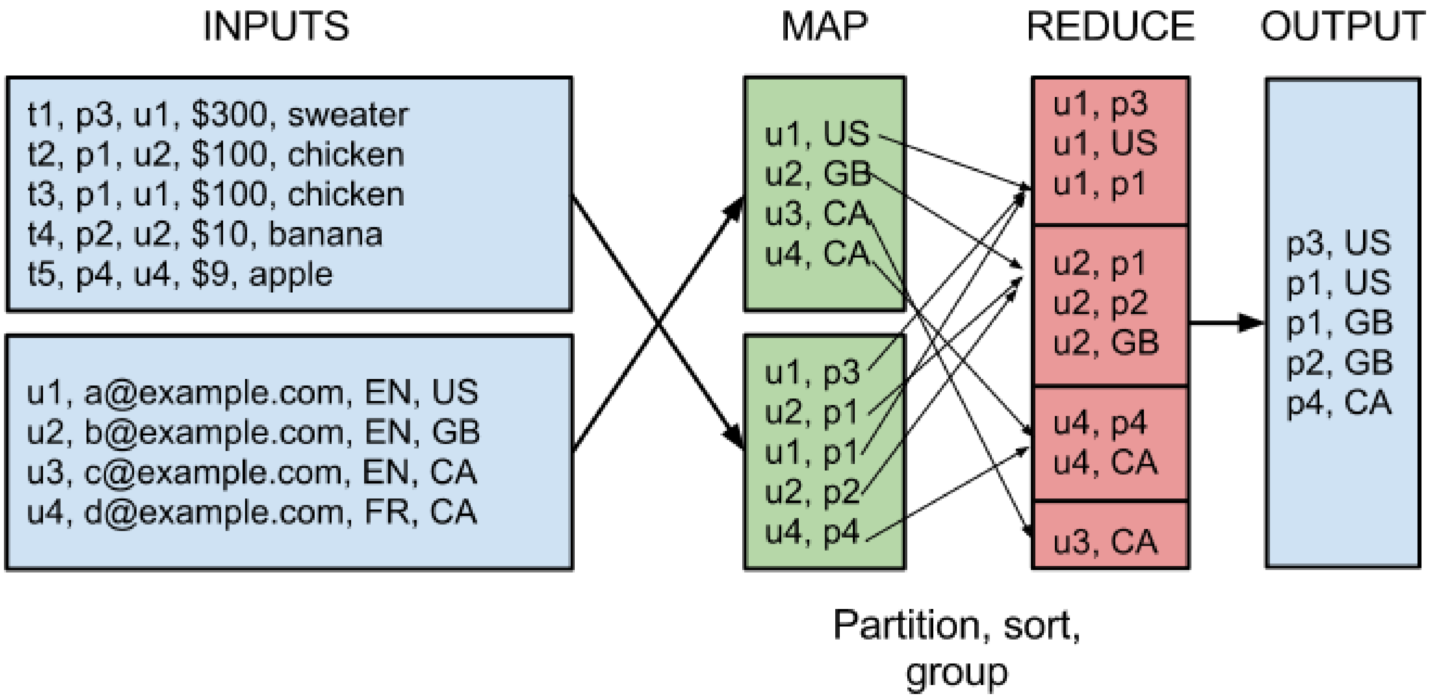
- By default. Hadoop framework is hash based partitioner.
- Hash partitioner partitions the key space by using the hash code.

What is Hadoop Partitioner.



What is Hadoop Partitioner.

Which partition a given (key, value) pair will go ?



INPUTS

t1, p3, u1, $300, sweater
t2, p1, u2, $100, chicken
t3, p1, u1, $100, chicken
t4, p2, u2, $10, banana
t5, p4, u4, $9, apple

u1, a@example.com, EN, US
u2, b@example.com, EN, GB
u3, c@example.com, EN, CA
u4, d@example.com, FR, CA

MAP

u1, US
u2, GB
u3, CA
u4, CA

u1, p3
u2, p1
u1, p1
u2, p2
u4, p4

REDUCE

u1, p3
u1, US
u1, p1

u2, p1
u2, p2
u2, GB

u4, p4
u4, CA

u3, CA

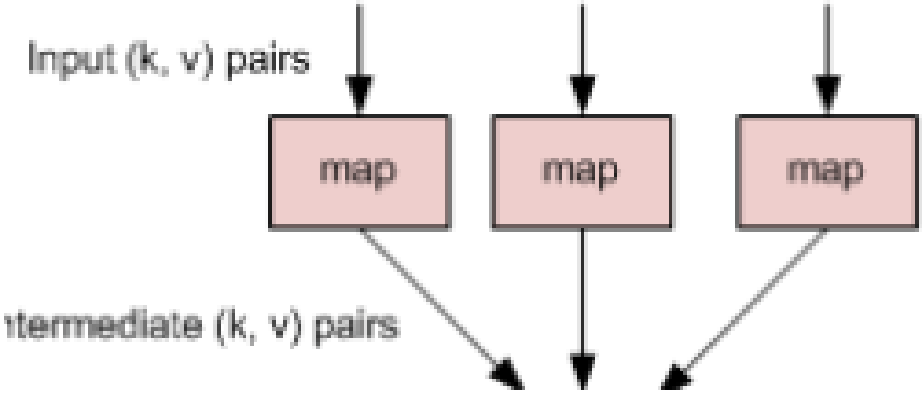
OUTPUT

p3, US
p1, US
p1, GB
p2, GB
p4, CA

Partition, sort, group

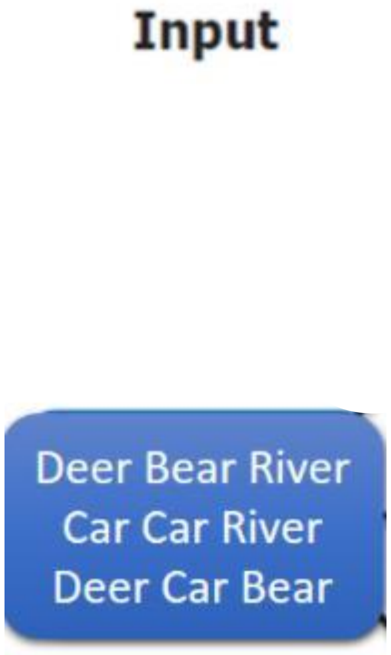
Key,value pair
Example (u1,US) (u2,GB) (u3,CA) (u4,CA)……….

What is Hadoop Partitioner.



Data flow takes place after map phase

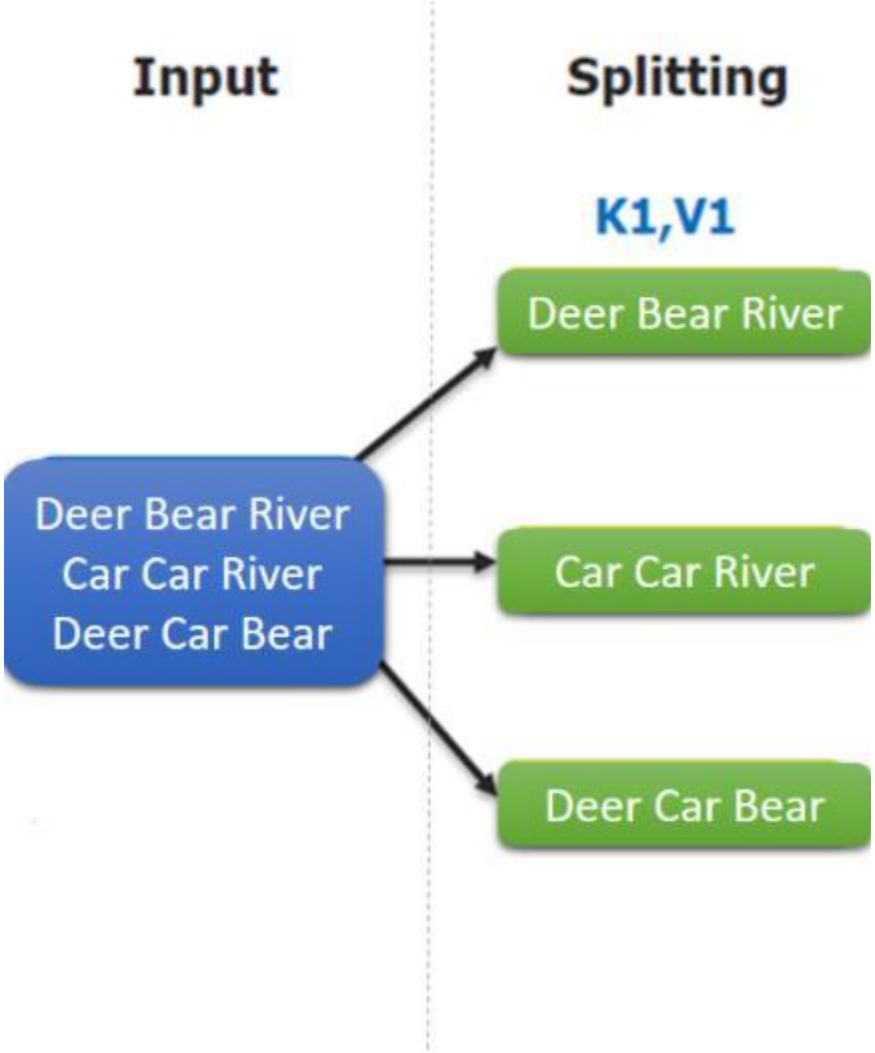
Need of MapReduce Partitioner in Hadoop



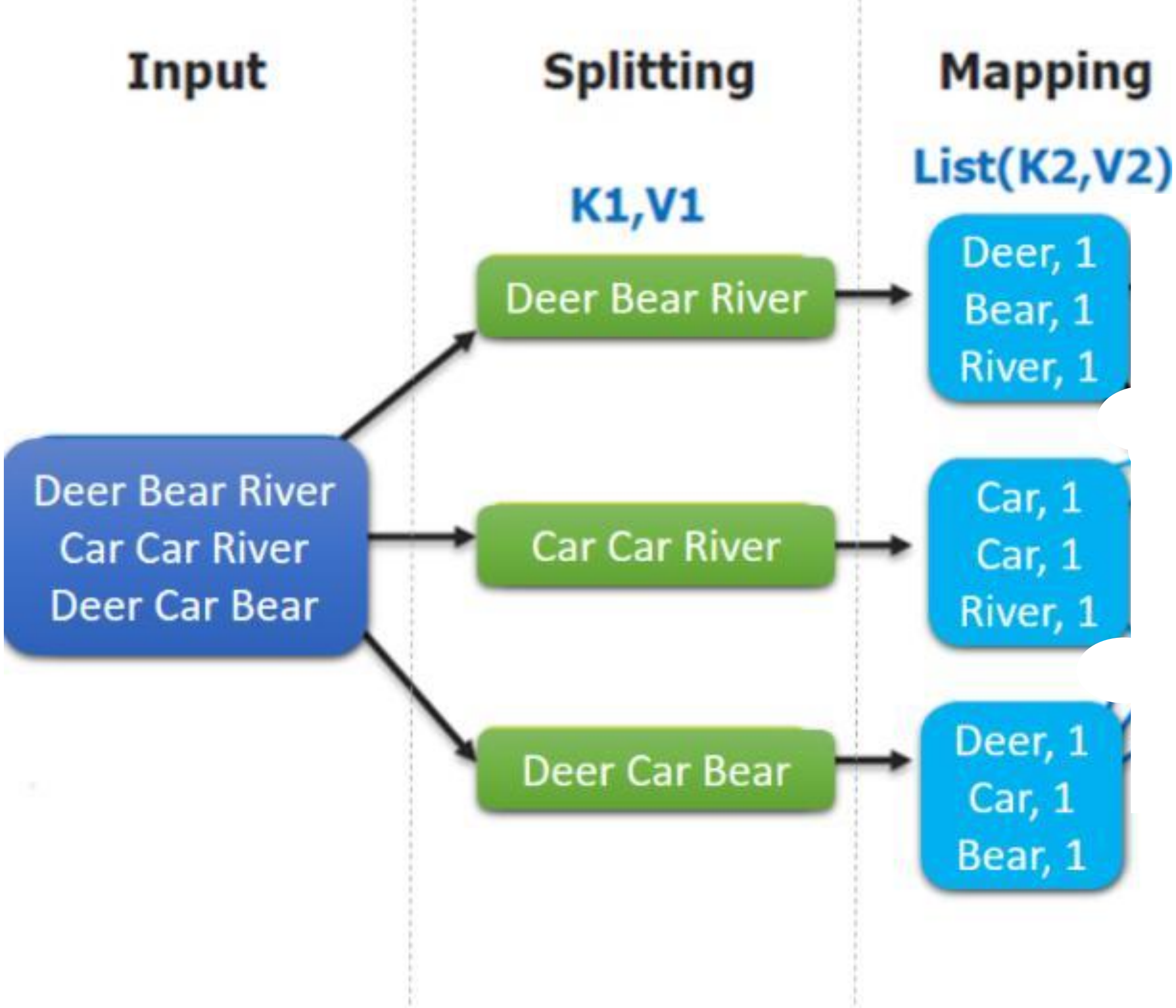
Input

Deer Bear River
Car Car River
Deer Car Bear

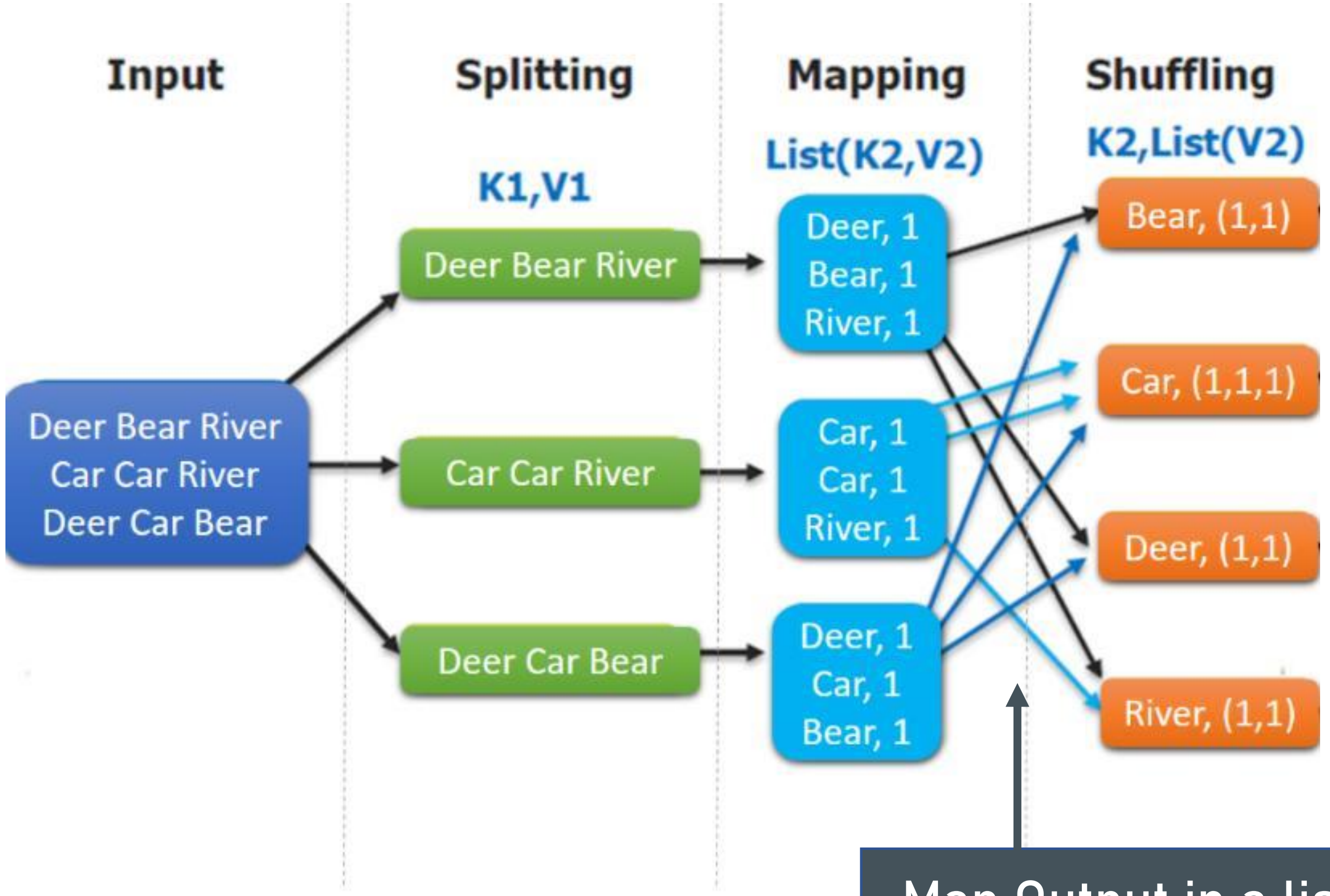
Need of MapReduce Partitioner in Hadoop



Need of MapReduce Partitioner in Hadoop



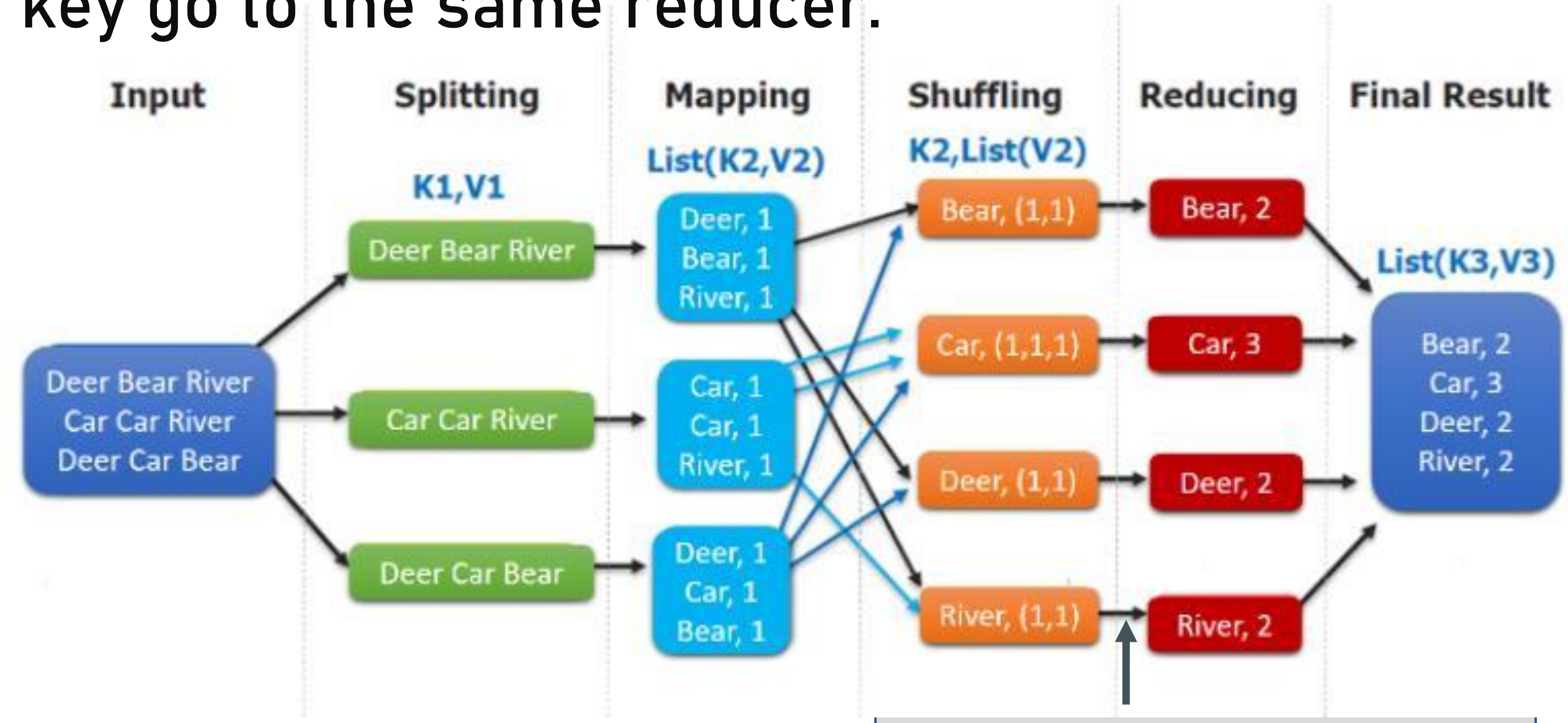
Need of MapReduce Partitioner in Hadoop



Map Output in a list of (key,value) in the intermediate file

Need of MapReduce Partitioner in Hadoop

It also makes sure that all the values of a single key go to the same reducer.



Ouput in a list of (Key, List of Values) in the intermediate file

Need of MapReduce Partitioner in Hadoop

Partitioner makes sure that same key goes to the same reducer!

Hadoop Default Partitioner

- Hash Partitioner is the default Partitioner.
- It computes a hash value for the key.
- It also assigns the partition based on this result.

How many Partitioner in Hadoop?

- Depends on the number of reducers.
- Partitioner divides the data.
- It is set by JobConf.setNumReduceTasks() method.
- Single reducer processes the data.
- Framework creates partitioner only when there are many reducers.

Poor Partitioning in Hadoop MapReduce

If in data input in MapReduce job one key appears more than any other key.

The key appearing more number of times will be sent to one partition.

Poor Partitioning in Hadoop MapReduce

If in data input in MapReduce job one key appears more than any other key.

The key appearing more number of times will be sent to one partition.

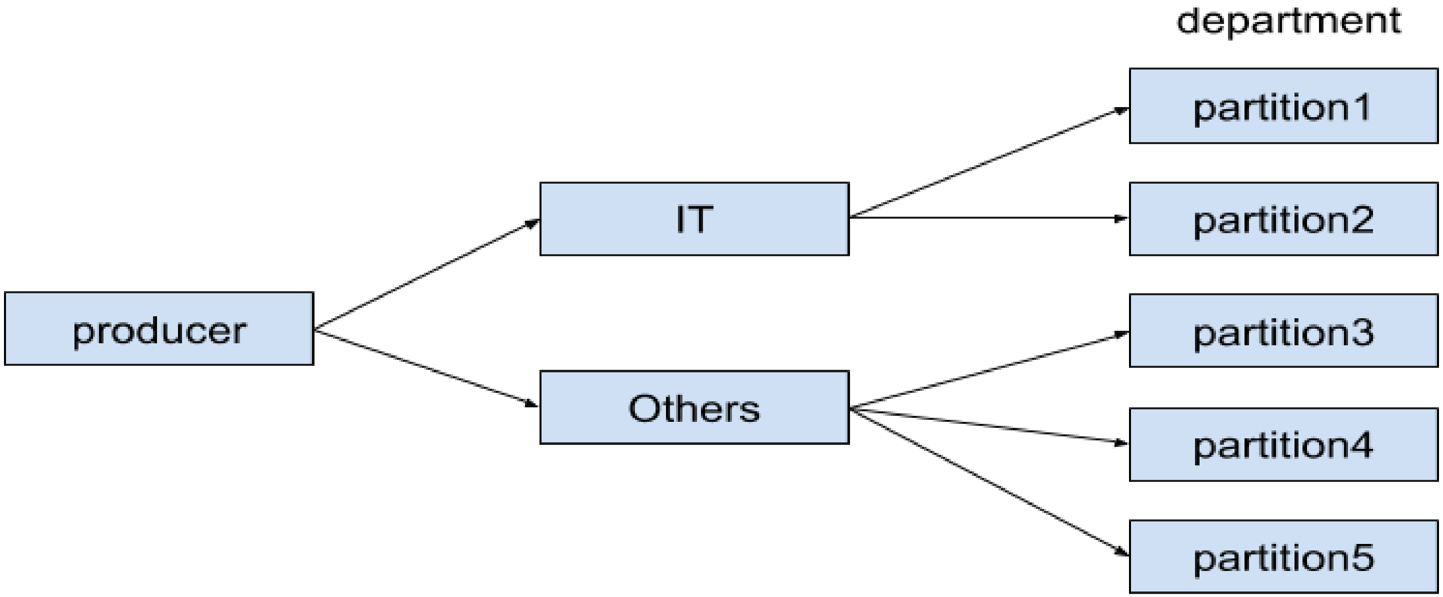
All the other key will be sent to partitions on the basis of their hashCode().

Poor Partitioning in Hadoop MapReduce

If hashCode() method does not distribute other key data over the partition range.

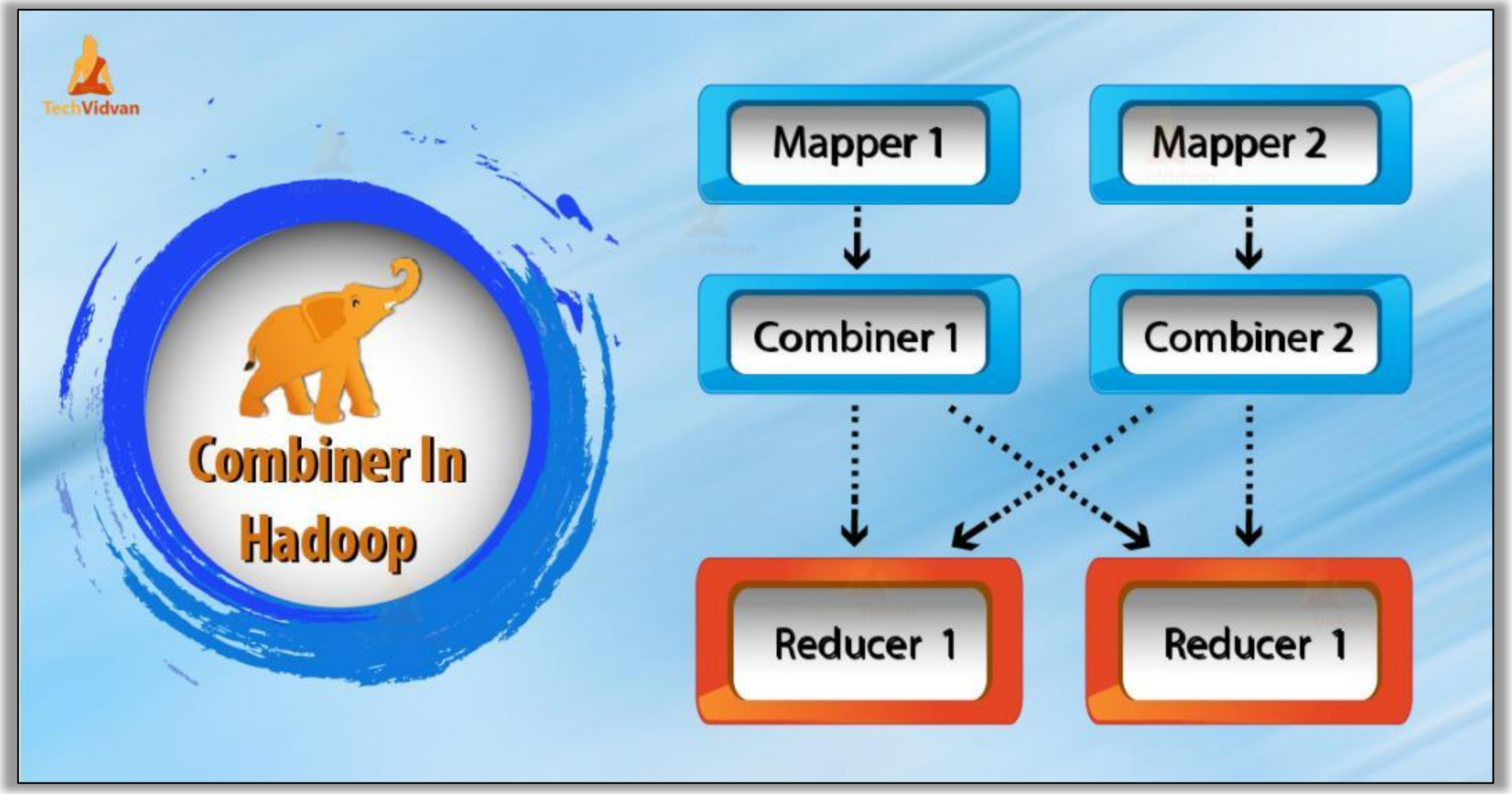
Data will not be sent to the reducers.

How to overcome poor partitioning in MapReduce?



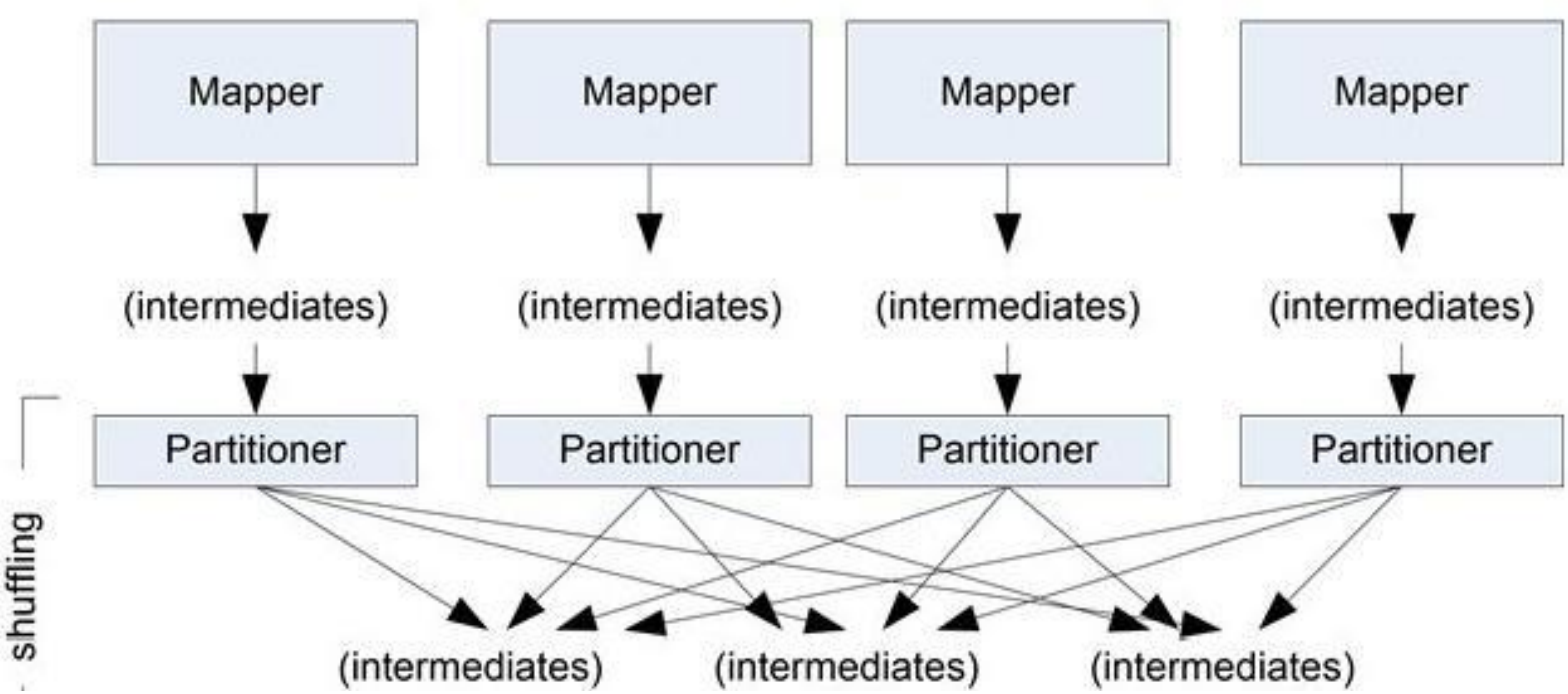
Hadoop Combiner

Introduction



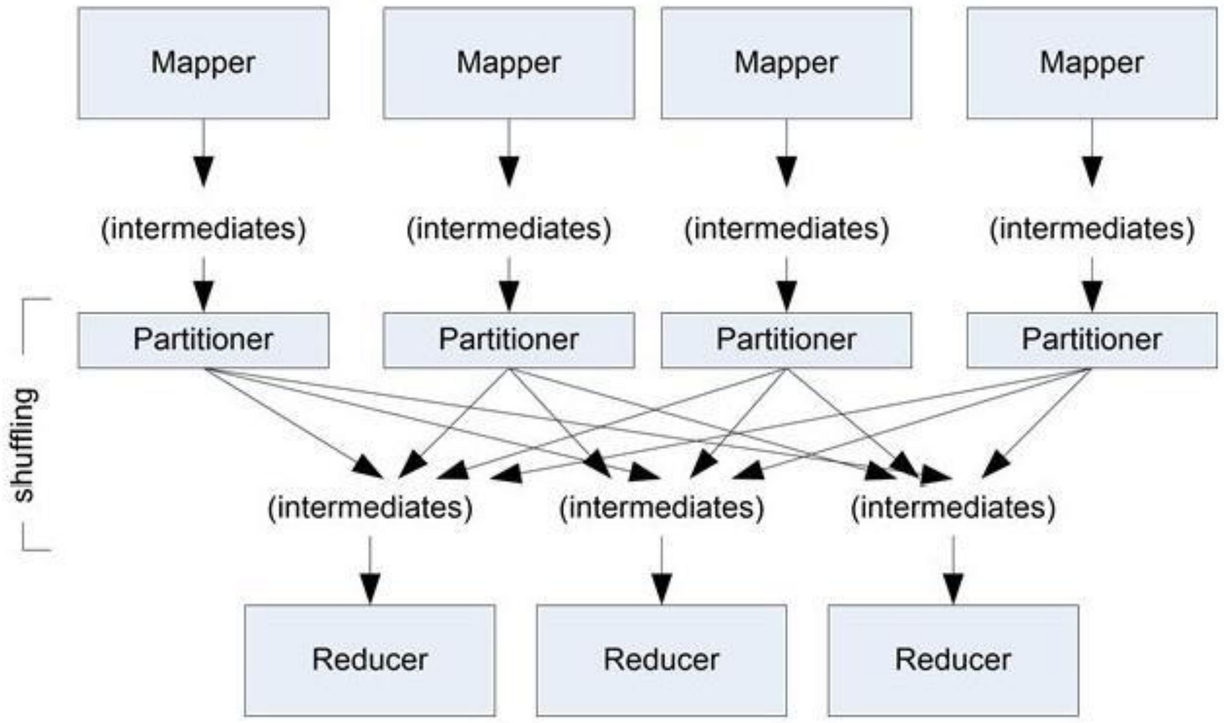
Hadoop Combiner

So Mapper generates large chunks of intermediate data.



Hadoop Combiner

So Mapper generates large chunks of intermediate data.



Hadoop Combiner

Network Congestion

Hadoop Combiner

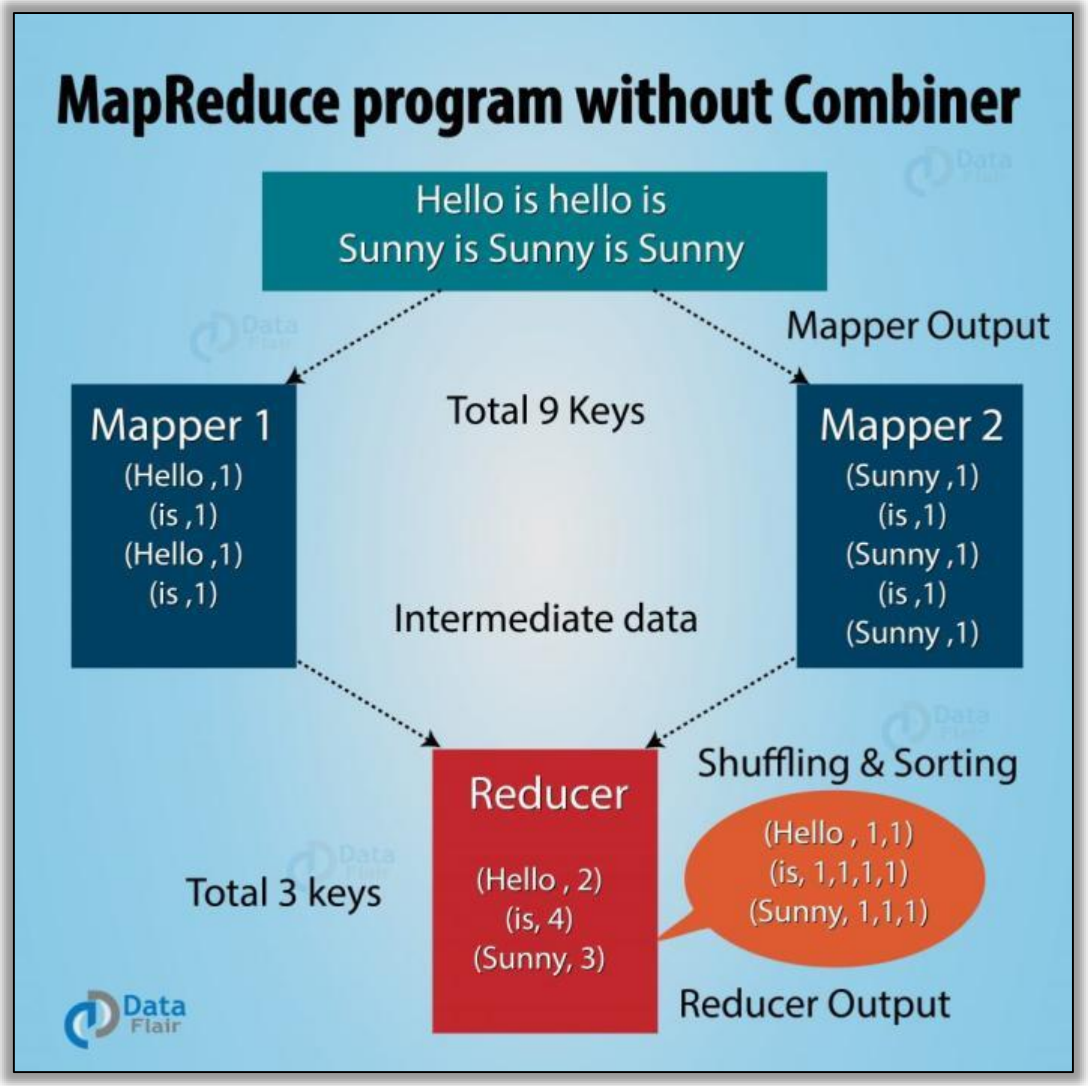
Hadoop Combiner

MapReduce

program

without

Combiner



MapReduce program without Combiner

Hello is hello is
Sunny is Sunny is Sunny

Mapper Output

Mapper 1
(Hello ,1)
(is ,1)
(Hello ,1)
(is ,1)

Total 9 Keys

Mapper 2
(Sunny ,1)
(is ,1)
(Sunny ,1)
(is ,1)
(Sunny ,1)

Intermediate data

Shuffling & Sorting

Reducer
(Hello , 2)
(is, 4)
(Sunny, 3)

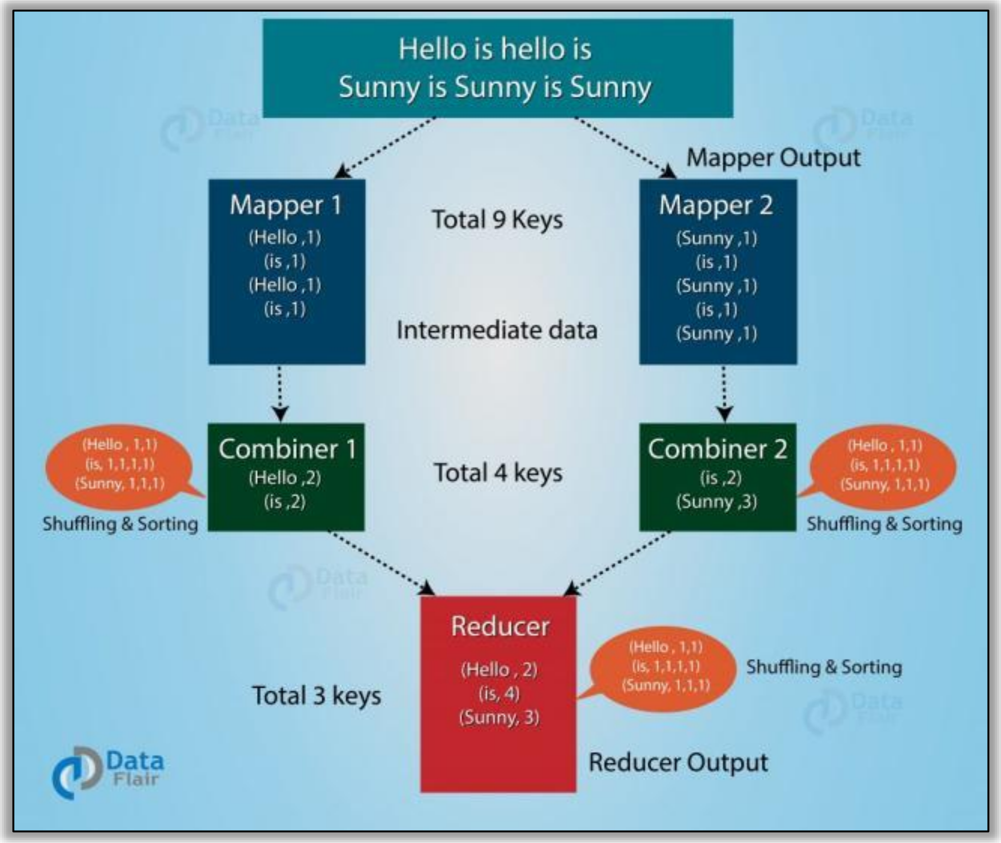
Total 3 keys

(Hello , 1,1)
(is, 1,1,1,1)
(Sunny, 1,1,1)

Reducer Output

Hadoop Combiner

MapReduce program with Combiner in between Mapper and Reducer



References

https://data-flair.training/blogs/hadoop-combiner-tutorial/

Advantages of MapReduce Combiner

- Reduces the time taken for data

- Decreases the amount of data.

- Improves the overall performance of the reducer.

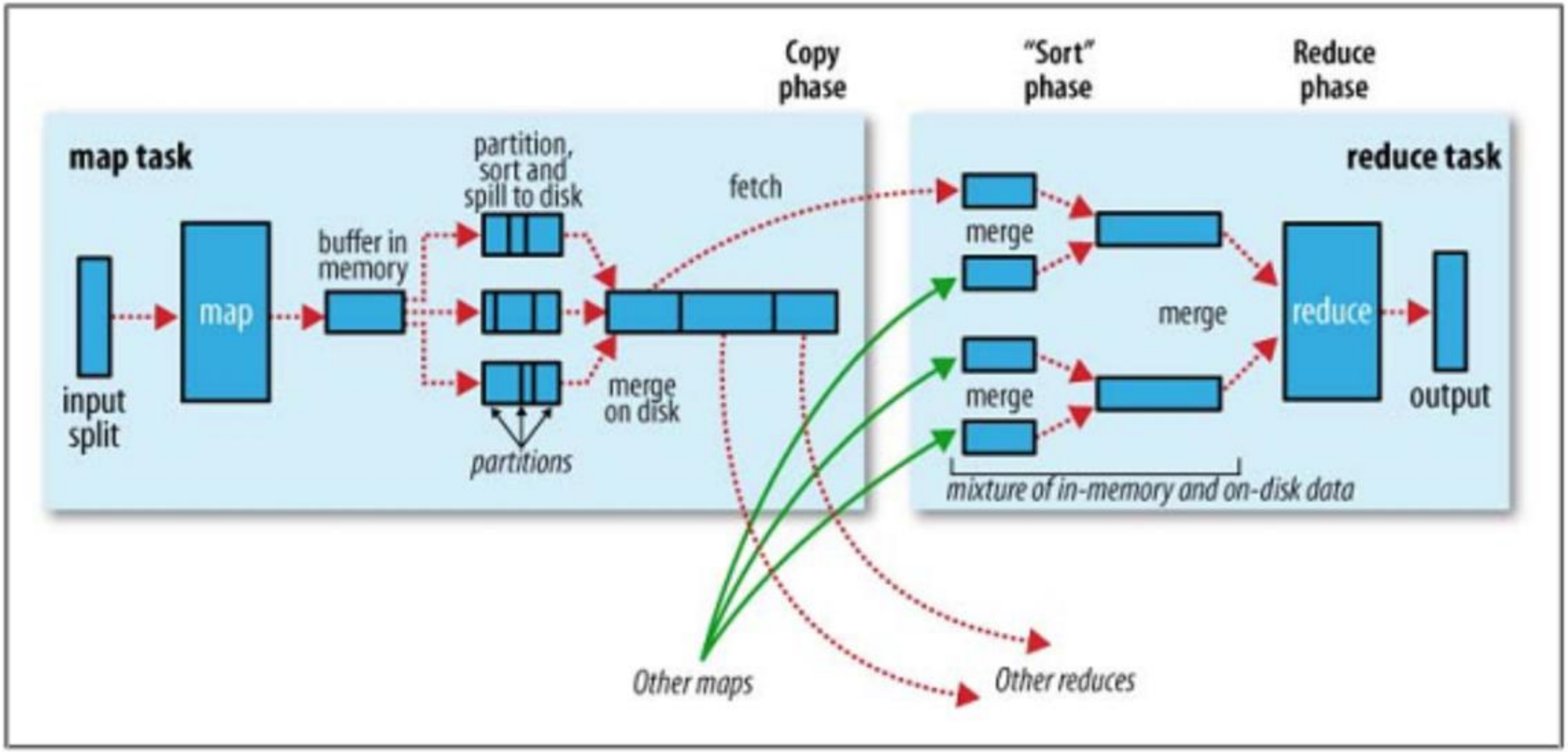
Disadvantages of Hadoop combiner in MapReduce

- There are also some disadvantages of hadoop Combiner. Let’s discuss them one by one-
- MapReduce jobs cannot depend on the Hadoop combiner execution because there is no guarantee in its execution.
- In the local filesystem, the key-value pairs are stored in the Hadoop and run the combiner later which will cause expensive disk IO

Disadvantages of Hadoop combiner in MapReduce

- MapReduce jobs cannot depend on the Hadoop combiner execution.
- Expensive disk IO

Shuffle and Sort in MapReduce



That’s all for now…