Intro to Hadoop : MapReduce

What is MapReduce?

MapReduce is a computing paradigm for processing data that resides on many computers

MapReduce

- **Definition:** MapReduce is a programming model for processing and generating large data sets with a parallel, distributed algorithm.
- **Developed by:** Google
- **Key Components:**
  - **Map function:** Processes and filters data
  - **Reduce function:** Aggregates results

Simpler terms : Mapper (The Sorter)

- **What it Does**: Think of the Mapper as a sorter or organizer. It takes in a big pile of unsorted items (like a bunch of words from a book) and sorts them into little piles based on some criteria (like sorting all the same words together).
- **Example**: Imagine you have a list of sentences, and your task is to count how many times each word appears. The Mapper looks at each word and says, "Here's a word, and I found it one time!" So if the sentence is "apple banana apple," the Mapper will output something like:
  - "apple": 1
  - "banana": 1
  - "apple": 1

Simpler terms : Reducer (The aggregator)

- **What it Does**: The Reducer is like a aggregator or calculator. It takes all those little piles of sorted items from the Mapper and agg them together to get a final total.
- **Example**: Continuing from the previous example, the Reducer gets the little piles of "apple" and "banana" from the Mapper. It then counts & agg (adds - as per problem statement)how many "apples" and how many "bananas" there are:
  - "apple": 2 (because there were two "apple" piles)
  - "banana": 1 (because there was one "banana" pile)

MapReduce-Features

- **Scalability:** Easily handles large data sets
- **Parallelism:** Processes data concurrently across multiple nodes
- **Fault Tolerance:** Automatically handles node failures

MapReduce - Data Processing

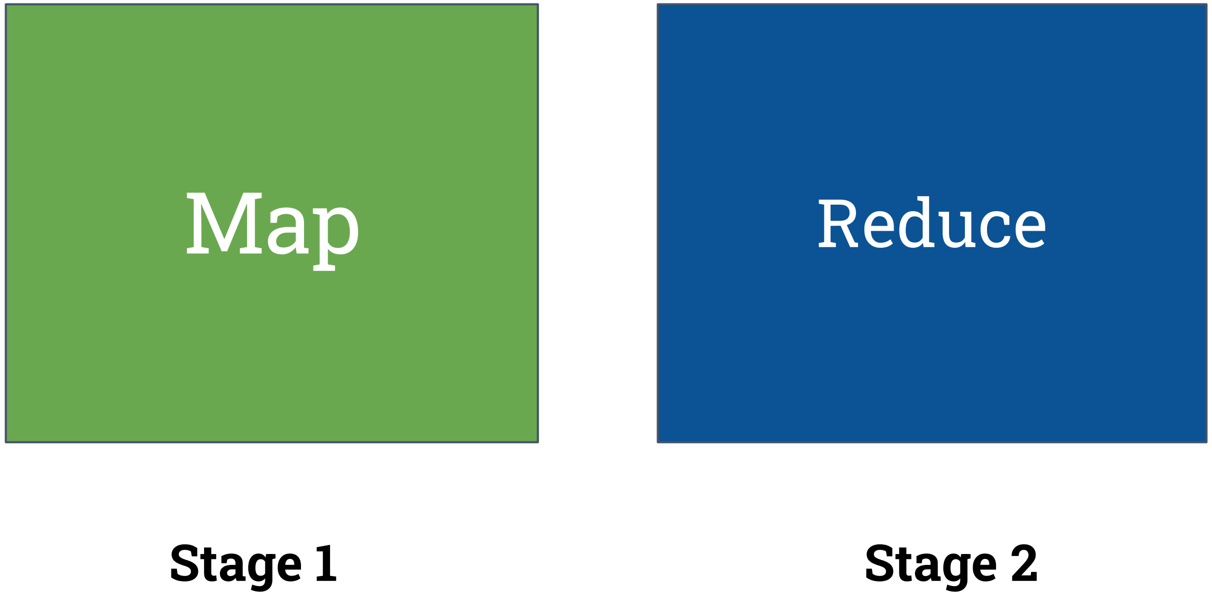
- Batch Processing
- High Latency Jobs (MapReduce jobs take time to complete)
- No live stream processing capabilities
- MapReduce jobs read data from a stable storage (Ex. HDFS)

MapReduce - Phases

- MapReduce Program (2 phases)
  - Map
    - Example: If you have a list of words, the Mapper would count how many times each word appears.
  - Reduce
    - Example: If the word "apple" showed up twice, the Reducer would add those counts together to tell you "apple" appeared 2 times.
- *Mantra*
  - *Use Transformation Logic in Map*
  - *Use Aggregation Logic in Reduce*

What is MapReduce?

There are 2 stages in MapReduce



Map

Reduce

Stage 1

Stage 2

What is MapReduce?

There are 2 stages in MapReduce

Both Map and Reduce only works on (Key, Value) pair

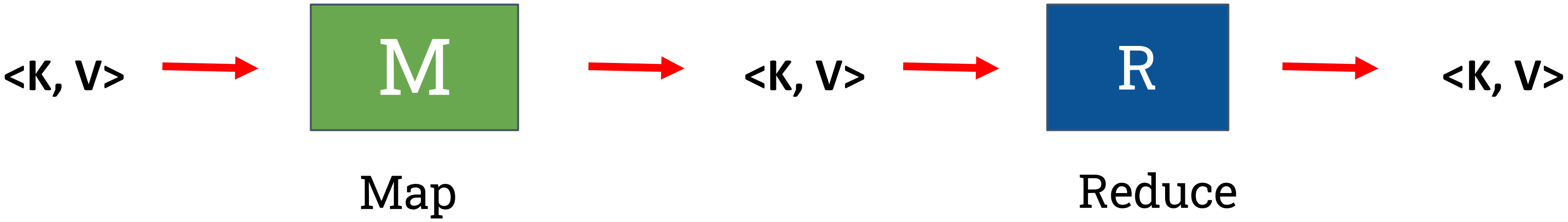
What is MapReduce?

What is (Key, Value)?

Key	Value
ID	101
Name	Ram
Designation	Developer

What is MapReduce?

Both Map and reduce only works on (Key, value) pair



Example to Understand Map and Reduce

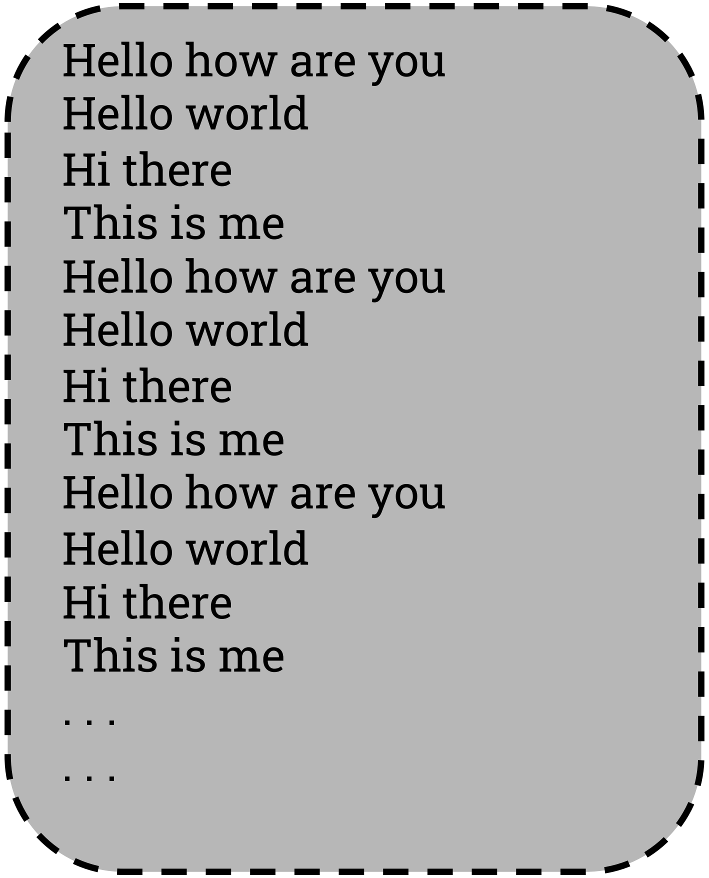
What is MapReduce?

Suppose we have a large file (file1.txt) with millions of records

```
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
. . .
. . .
```

file1.txt
(500mb)

What is MapReduce?

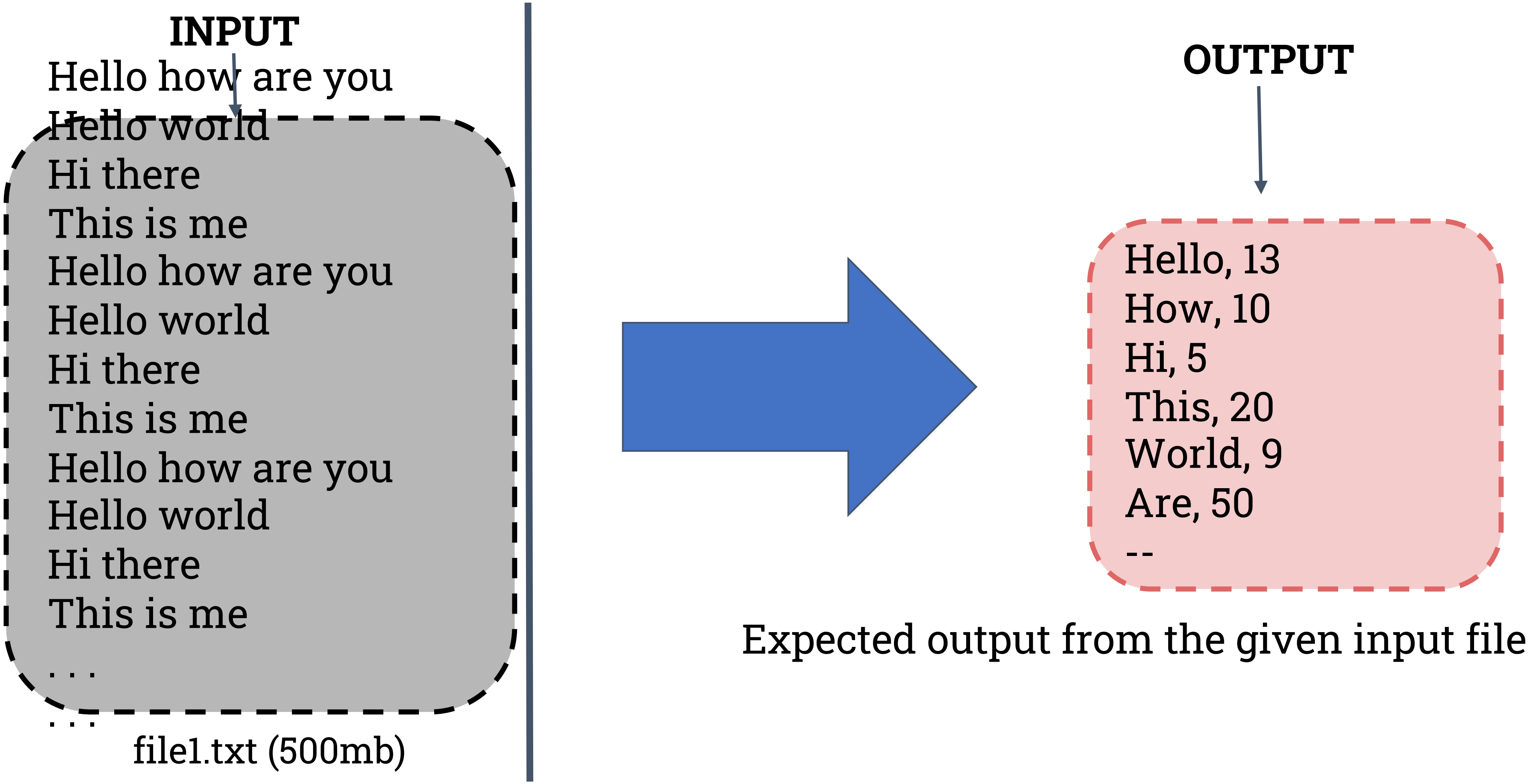


```
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
. . .
. . .
```

file1.txt (500mb)

We need to find out frequency of each word

What is MapReduce?



INPUT

Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
. . .
. . .

file1.txt (500mb)

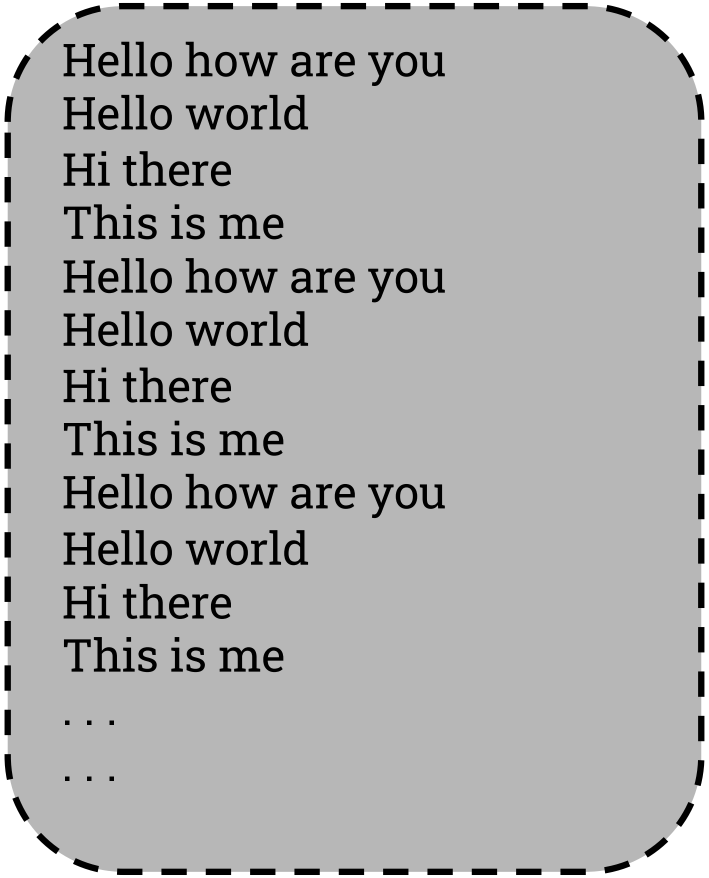
OUTPUT

Hello, 13
How, 10
Hi, 5
This, 20
World, 9
Are, 50
--

Expected output from the given input file

Now, how to solve this problem using Map & Reduce

What is MapReduce?



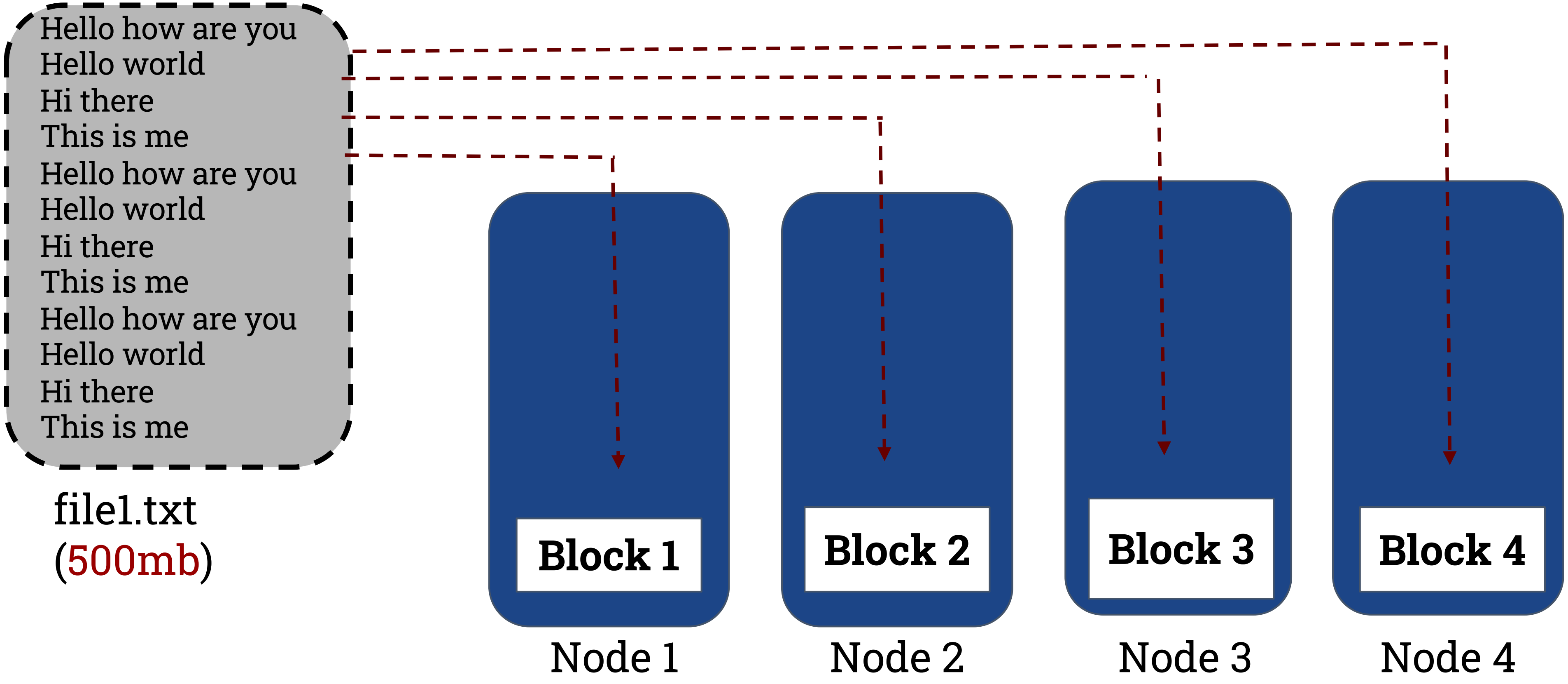
```
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
. . .
. . .
```

file1.txt (500mb)

In Hadoop default block size is 128 mb

So this file will be divided into 4 Blocks

What is MapReduce?



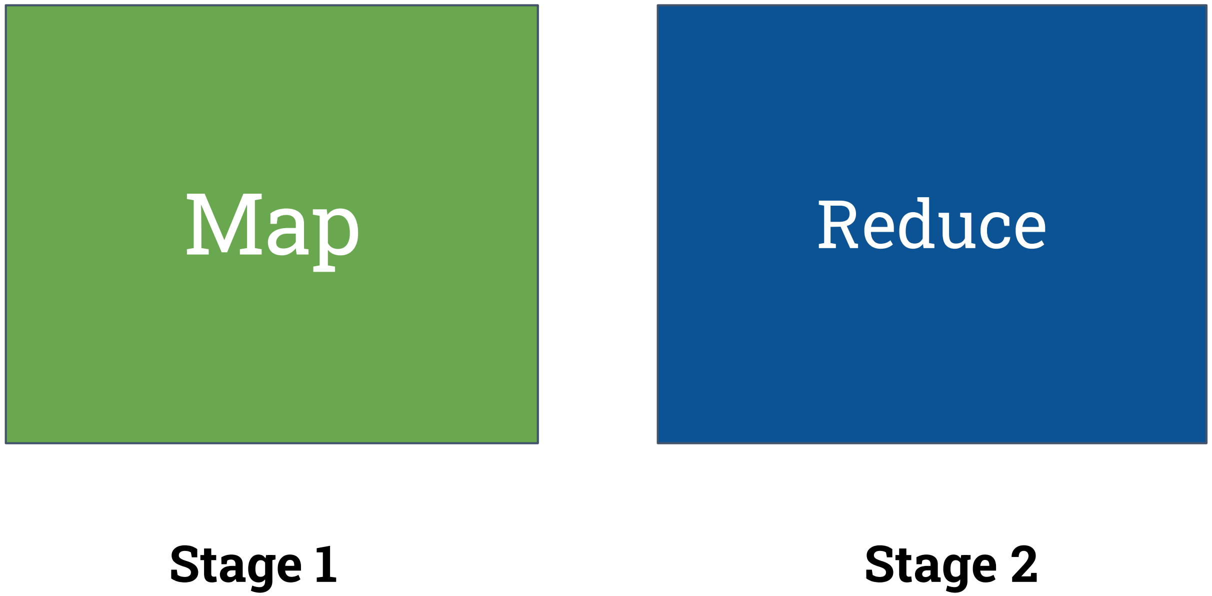
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me

file1.txt
(500mb)

File1.txt divided into 4 blocks

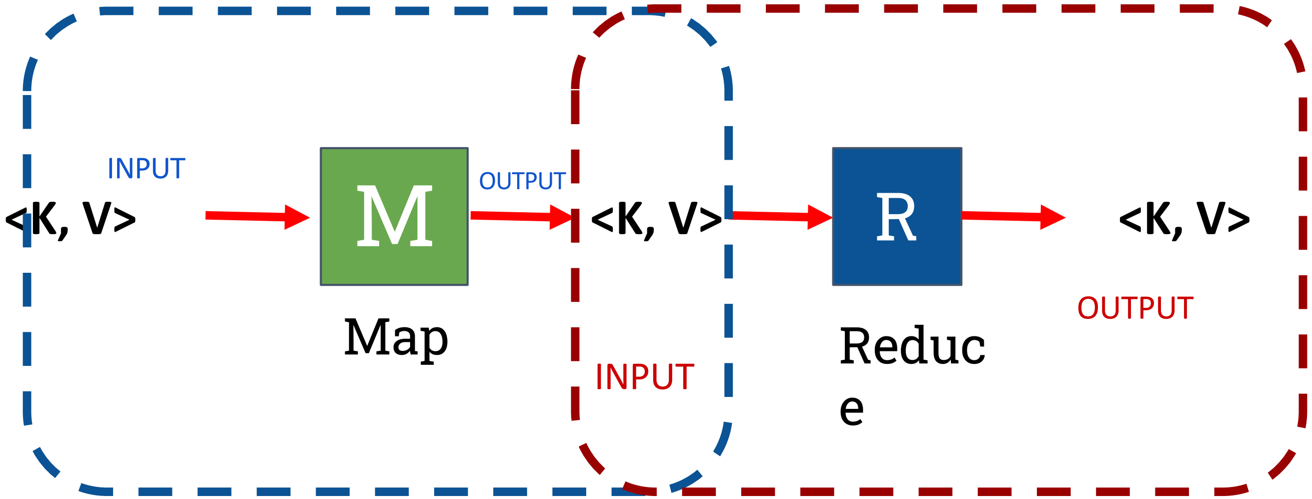
What is MapReduce?

As we know that there are 2 stages in MapReduce



What is MapReduce?

We also know that both Map and Reduce only works on (key, Value) pair



What is MapReduce?

But, in our example We have input records which are like:

Input Records

Hello how are you
Hello world
. . .

What is MapReduce?

Input Records

Hello how are you
Hello world
. . .
. . .

Are these input records are (Key, Value) pairs ?

What is MapReduce?

Input Records

Hello how are you
Hello world
. . .
. . .

Are these input records are (Key, Value) pairs ?

NO

What is MapReduce?

Input Records

Hello how are you
Hello world
. . .
. . .

Are these input records are (Key, Value) pairs ?

NO

↓

How to solve this problem?

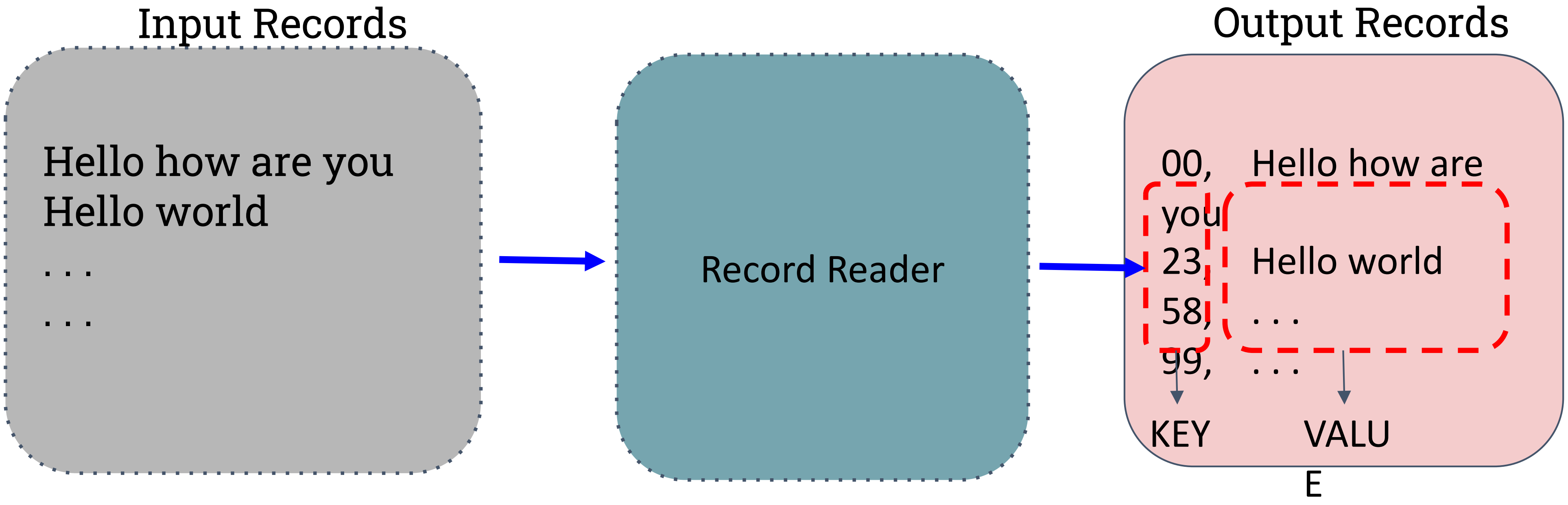
Here Record Reader comes into picture

What is MapReduce?

Record Reader

The role of Record Reader is to convert each input line into (Key, Value) pair suitable for reading by the Mapper

What is MapReduce?

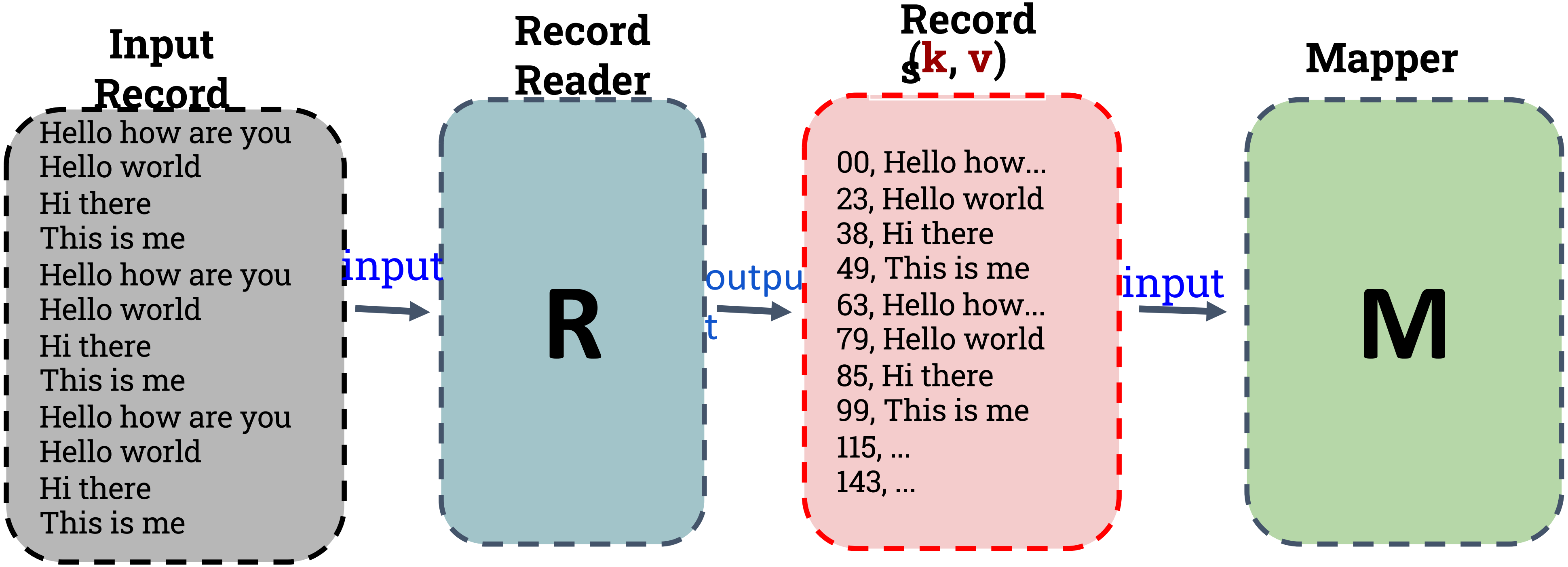


Record Reader converting input record to (Key, Value) pair

MAPPER

A Timesgroup Initiative

What is MapReduce?



Input Record

Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me
Hello how are you
Hello world
Hi there
This is me

input

Record Reader

R

output

Records (**k**, **v**)

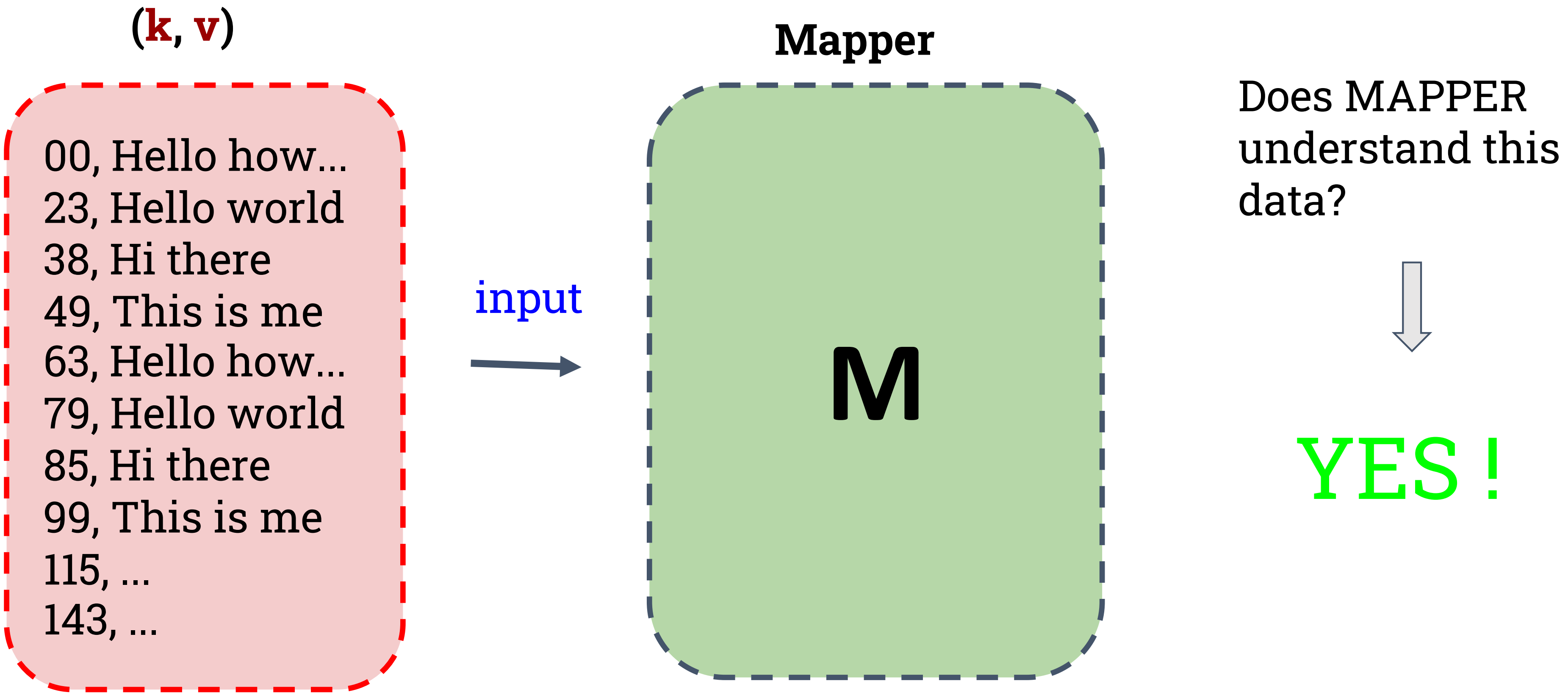
00, Hello how...
23, Hello world
38, Hi there
49, This is me
63, Hello how...
79, Hello world
85, Hi there
99, This is me
115, ...
143, ...

input

Mapper

M

What is MapReduce?



(k, v)

00, Hello how...
23, Hello world
38, Hi there
49, This is me
63, Hello how...
79, Hello world
85, Hi there
99, This is me
115, ...
143, ...

input

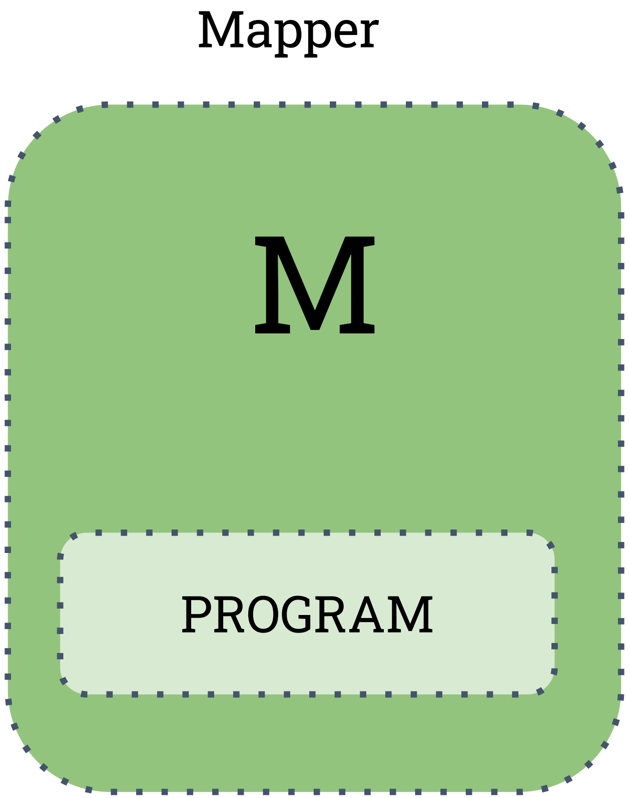
Mapper

M

Does MAPPER understand this data?

YES !

What is MapReduce?

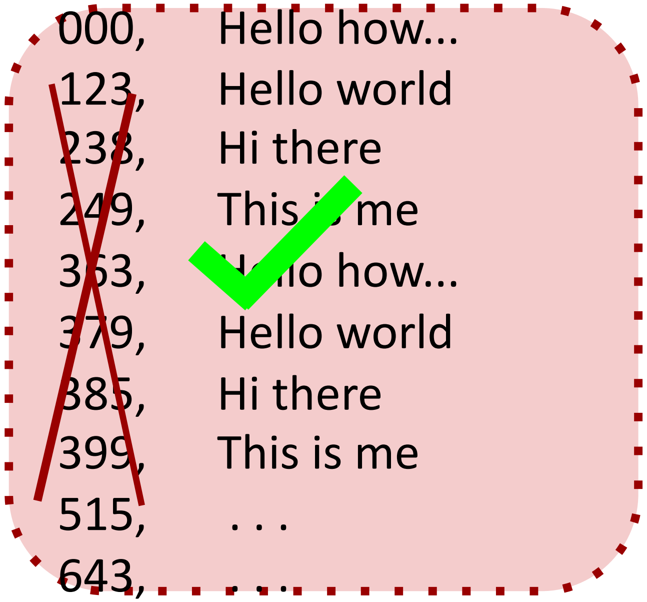


Now, what should be the Mapper logic?

This logic/program has to be written by the developer

What is MapReduce?

(k, v) Pairs

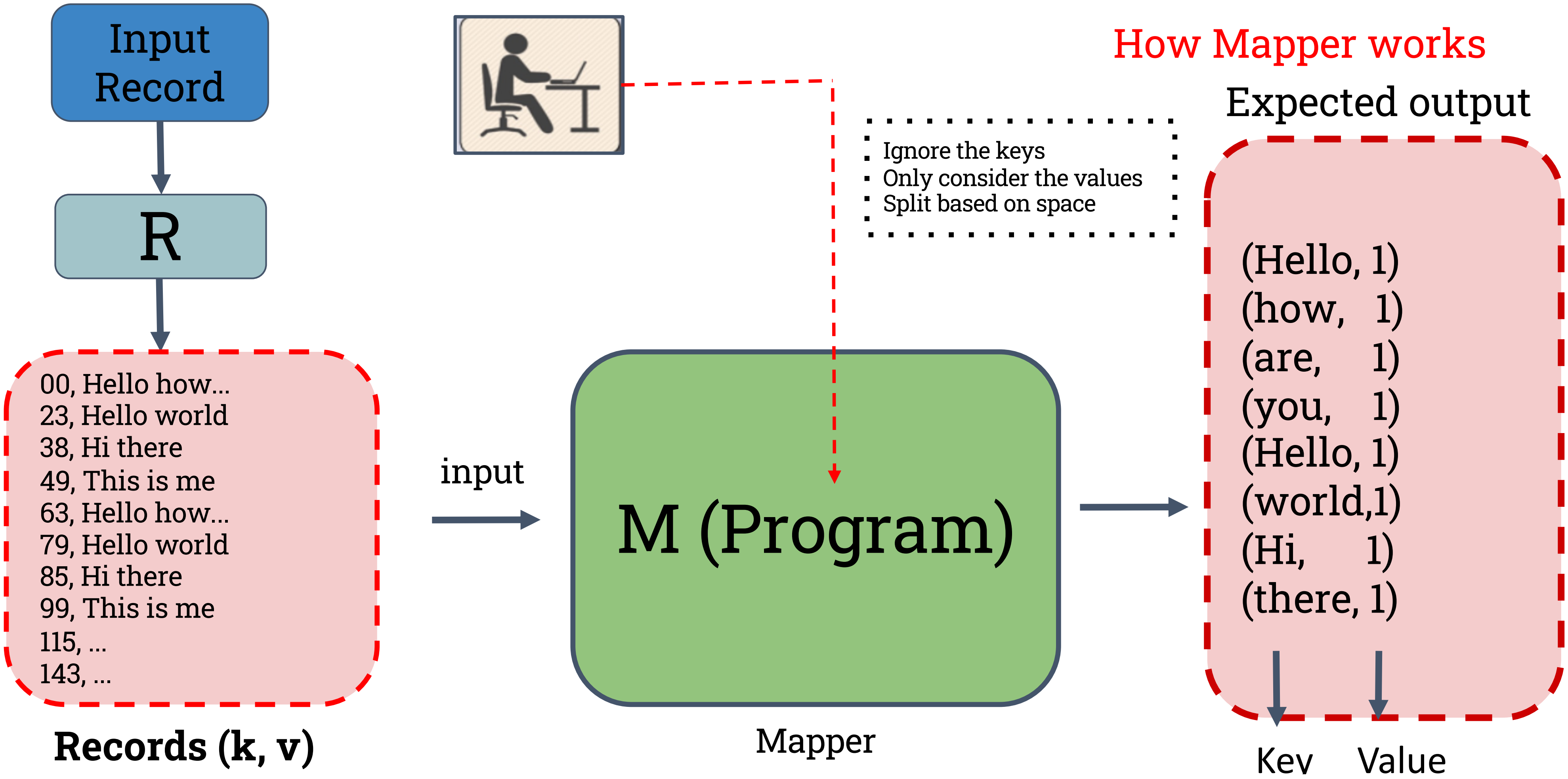


000,	Hello how...
123,	Hello world
238,	Hi there
249,	This is me
353,	Hello how...
379,	Hello world
385,	Hi there
399,	This is me
515,	. . .
643,	. . .

Here the keys are not relevant for us

We only consider the values

What is MapReduce?



Input Record

R

00, Hello how...
23, Hello world
38, Hi there
49, This is me
63, Hello how...
79, Hello world
85, Hi there
99, This is me
115, ...
143, ...

Records (k, v)

input

M (Program)

Mapper

Ignore the keys
Only consider the values
Split based on space

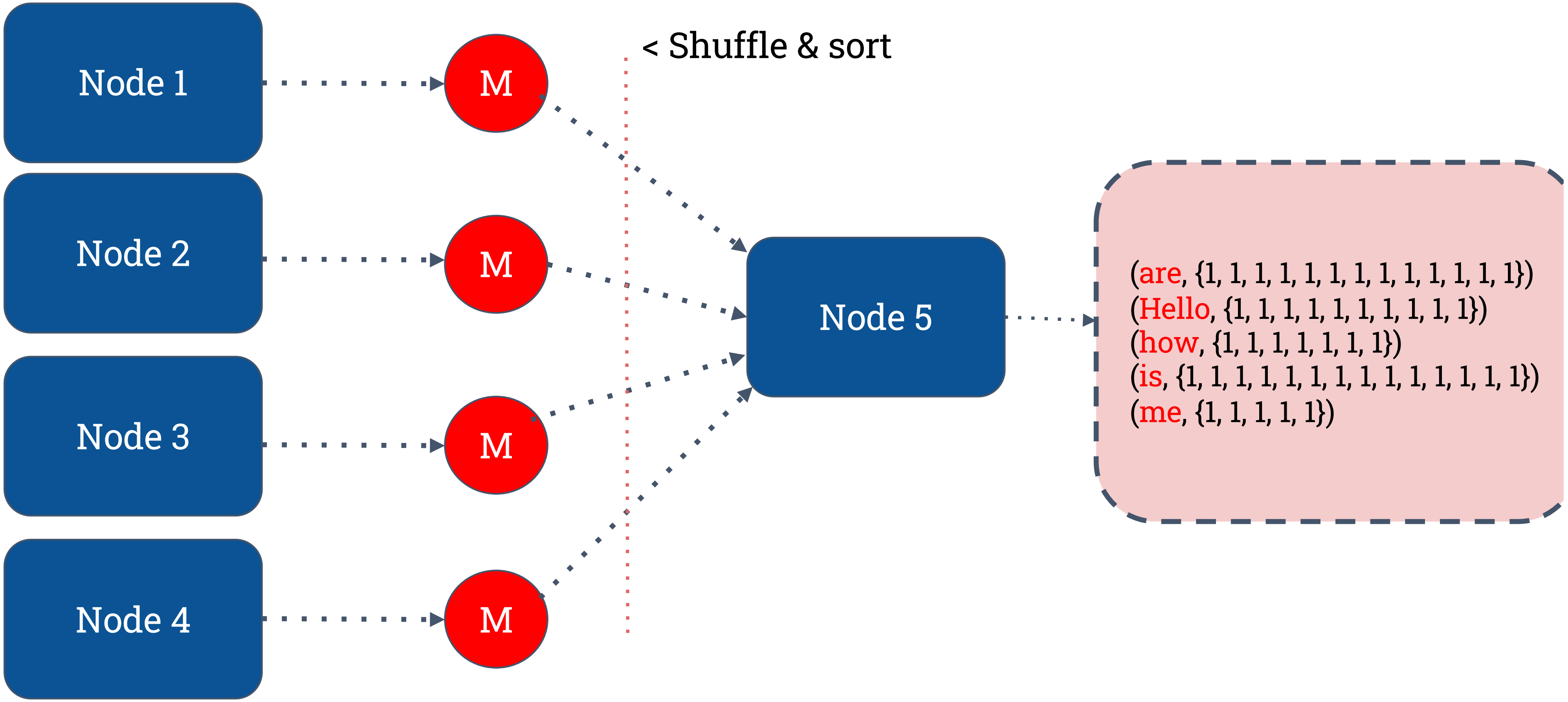
How Mapper works

Expected output

(Hello, 1)
(how, 1)
(are, 1)
(you, 1)
(Hello, 1)
(world,1)
(Hi, 1)
(there, 1)

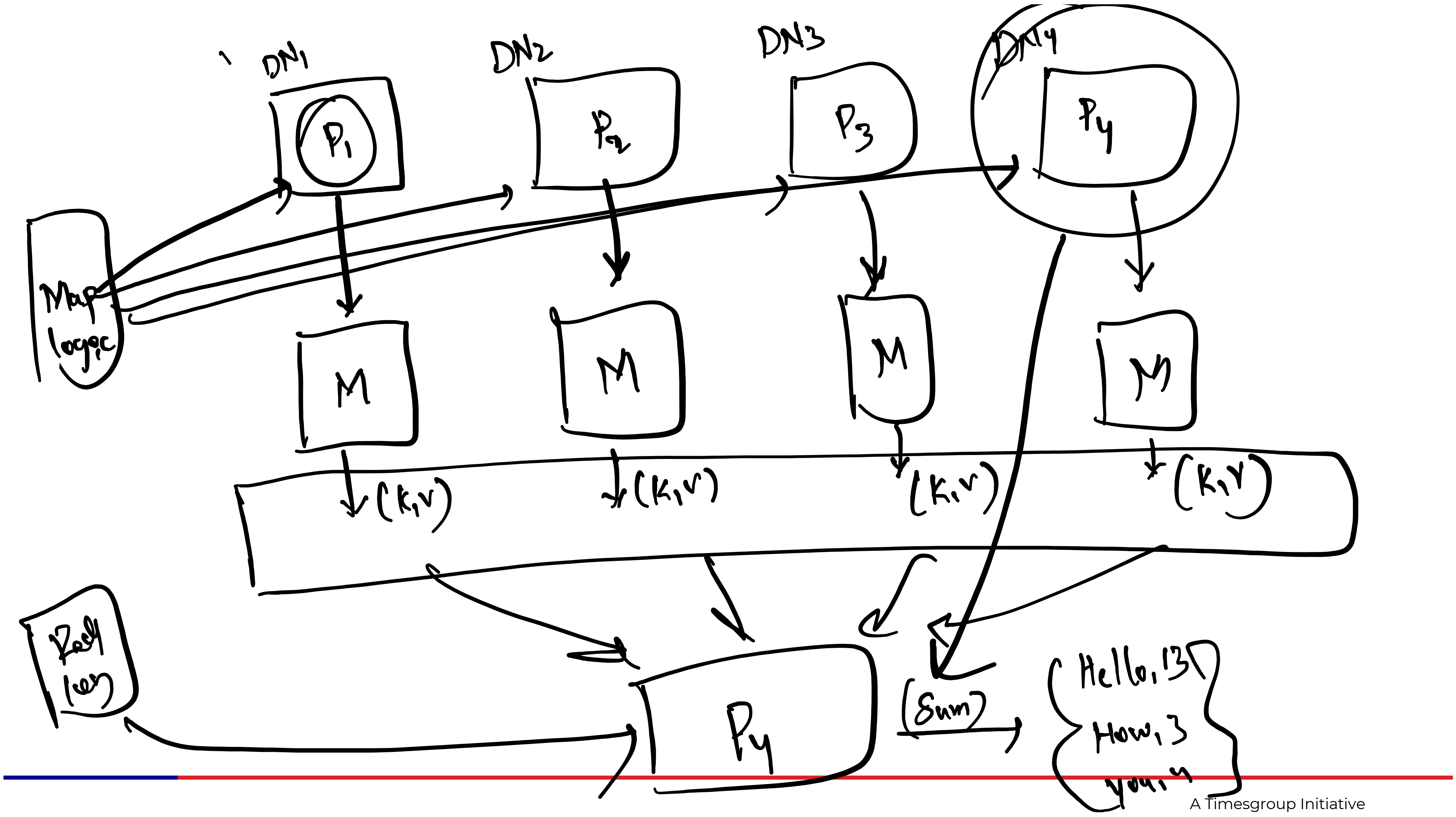
Key Value

What is MapReduce?

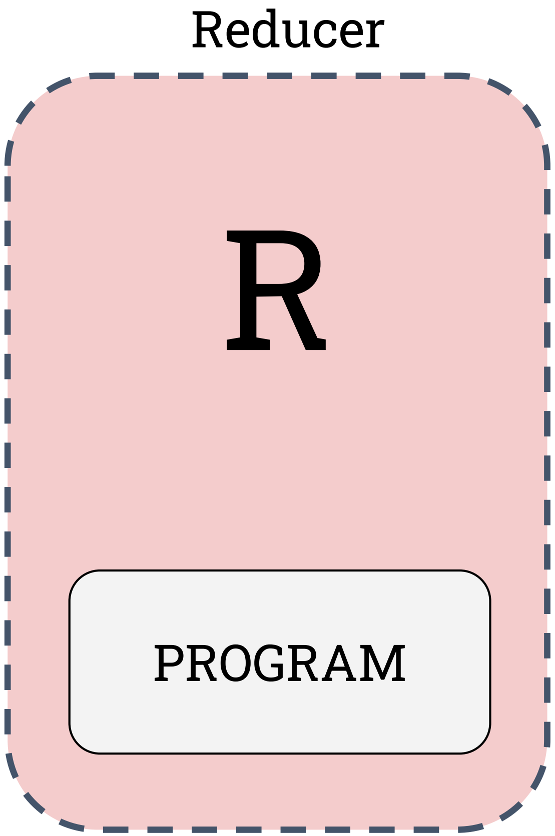


(are, {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1})
(Hello, {1, 1, 1, 1, 1, 1, 1, 1, 1, 1})
(how, {1, 1, 1, 1, 1, 1, 1})
(is, {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1})
(me, {1, 1, 1, 1, 1})

REDUCER


DN1
DN2
DN3
DN4
P1
P2
P3
P4
Map logic
M
M
M
M
(K,V)
(K,V)
(K,V)
(K,V)
Redlcer
P4
(Sum)
Hello,13
How,3
you,4

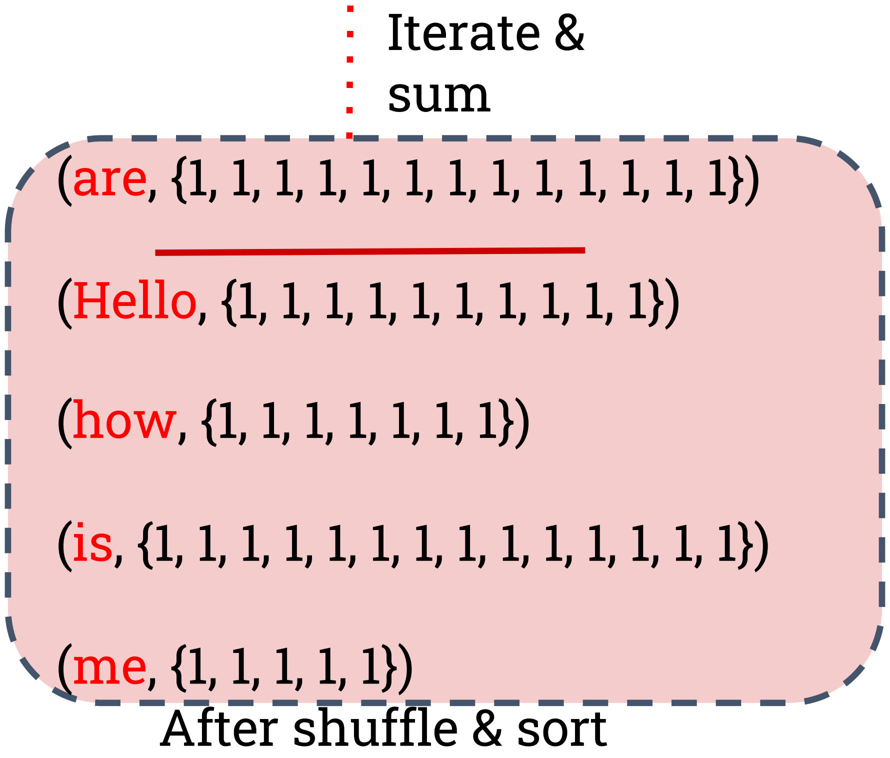
What is MapReduce?



Now, What should be the Reducer logic?

The Reducer logic/program has to be written by the developer

What is MapReduce?



Iterate & sum

(are, {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1})

(Hello, {1, 1, 1, 1, 1, 1, 1, 1, 1, 1})

(how, {1, 1, 1, 1, 1, 1, 1})

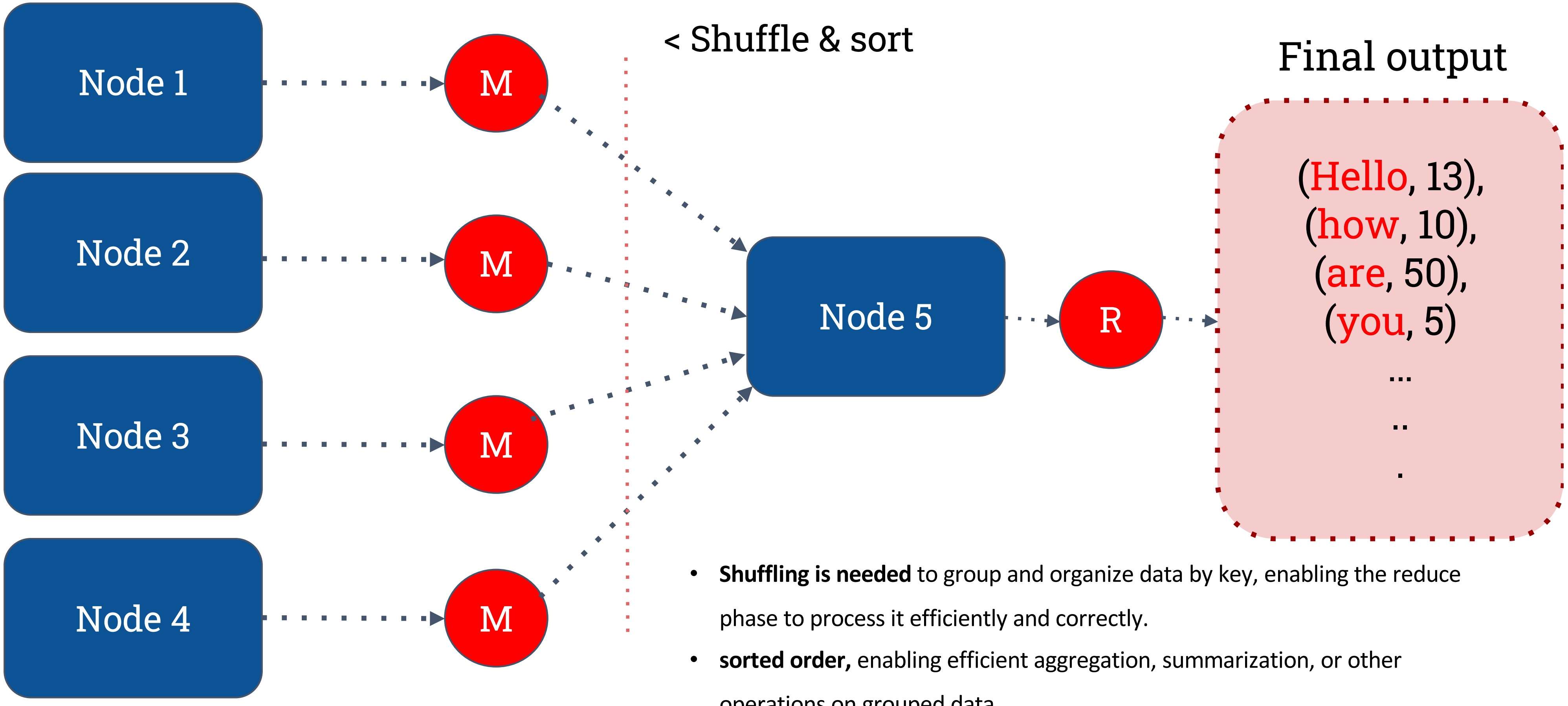
(is, {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1})

(me, {1, 1, 1, 1, 1})

After shuffle & sort

The Reducer logic should be - we iterate over the list of values and sum it up

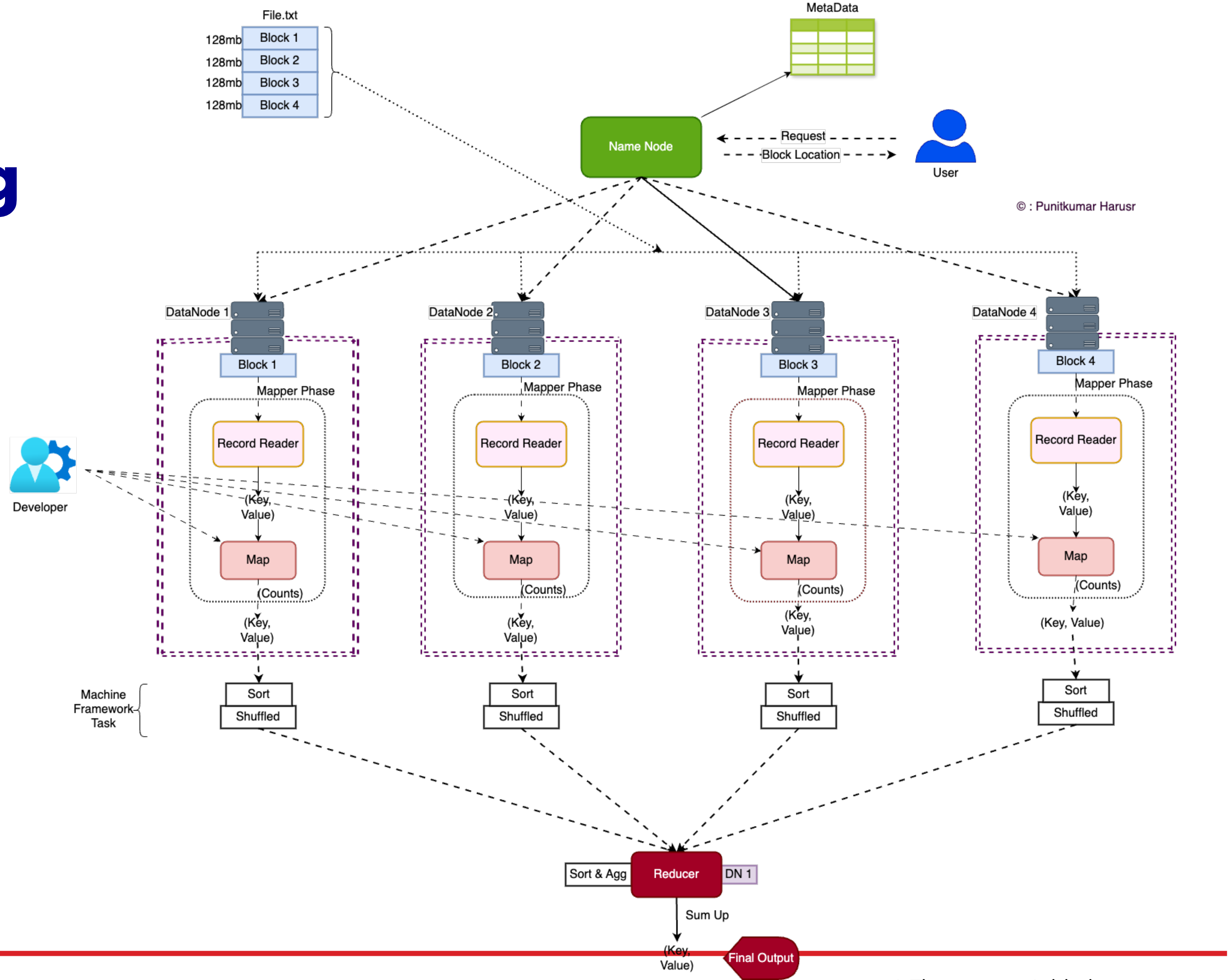
What is MapReduce?



- **Shuffling is needed** to group and organize data by key, enabling the reduce phase to process it efficiently and correctly.
- **sorted order,** enabling efficient aggregation, summarization, or other operations on grouped data.

Understanding MapReduce Workflow Detailed

Blog Link: understanding-mapreduce-workflow-detailed



© : Punitkumar Harusr

Summary : Stages of MapReduce

- **Data Split:** Input data is split into smaller chunks
- **Map Phase:** Each chunk is processed by a Map function in parallel
- **Shuffle and Sort:** Intermediate data is shuffled and sorted
- **Reduce Phase:** Aggregated and summarized results are processed by the Reduce function
- **Output:** Final result is written to the output file system

Assignment - Blog

Write a LinkedIn blog on MapReduce Workflow with MapReduce diagram(use draw.io) and tag.

Zoom Quiz

Today's topic revision

Challenges of MapReduce

1. Less Performant due to many IO disk seeks.
2. Need to write many lines of Code to accomplish even a simple task.
3. MapReduce Supports only Batch Processing
4. Learning curve is high
5. Constrained to always think in a Map-Reduce perspective.
6. No Interactive mode

Reads:

Useful reads

References

- **Books:** "Mining of Massive Datasets" by Jure Leskovec, Anand Rajaraman, Jeff Ullman
- **Papers:** "MapReduce: Simplified Data Processing on Large Clusters" by Jeffrey Dean and Sanjay Ghemawat
- **Websites:** Apache Hadoop, Google Research Publications