Computing Numeric Summary Metrics

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

See patterns in calculating numeric summary metrics using MapReduce

Use a Combiner to calculate numeric summaries

Implement a MapReduce to calculate averages using a Custom Writable class

Collecting Data

Consider the United States Census

A massive data collection activity

Undertaken every 10 years

Collecting data is only half the battle!

Snapshot of 1990 US Census

Age	Work	Education	Marital status	Occupation	Gender	Hrs/Wk	Income
39	State-gov	Bachelors	Never-married	Adm-clerical	Male	40	<=50K
50	Self-emp-	Bachelors	Married-civ-	Exec-managerial	Male	13	<=50K
38	Private	HS-grad	Divorced	Handlers-cleaners	Male	40	<=50K
53	Private	11th	Married-civ-	Handlers-cleaners	Male	40	<=50K
28	Private	Bachelors	Married-civ-	Prof-specialty	Female	40	<=50K
37	Private	Masters	Married-civ-	Exec-managerial	Female	40	<=50K
49	Private	9th	Married-	Other-service	Female	16	<=50K
52	Self-emp-	HS-grad	Married-civ-	Exec-managerial	Male	45	>50K
31	Private	Masters	Never-married	Prof-specialty	Female	50	>50K
42	Private	Bachelors	Married-civ-	Exec-managerial	Male	40	>50K
37	Private	Some-college	Married-civ-	Exec-managerial	Male	80	>50K
30	State-gov	Bachelors	Married-civ-	Prof-specialty	Male	40	>50K
23	Private	Bachelors	Never-married	Adm-clerical	Female	30	<=50K
32	Private	Assoc-acdm	Never-married	Sales	Male	50	<=50K
[illegible]	[illegible]	[illegible]	Married-civ-	[illegible]	[illegible]	[illegible]	[illegible]

Snapshot of 1990 US Census

Age	Work	Education	Marital status	Occupation	Gender	Hrs/Wk	Income
39	State-gov	Bachelors	Never-married	Adm-clerical	Male	40	<=50K
50	Self-emp-	Bachelors	Married-civ-	Exec-managerial	Male	13	<=50K
38	Private	HS-grad	Divorced	Handlers-cleaners	Male	40	<=50K
53	Private	11th	Married-civ-	Handlers-cleaners	Male	40	<=50K
28	Private	Bachelors	Married-civ-	Prof-specialty	Female	40	<=50K
37	Private	Masters	Married-civ-	Exec-managerial	Female	40	<=50K
49	Private	9th	Married-	Other-service	Female	16	<=50K
52	Self-emp-	HS-grad	Married-civ-	Exec-managerial	Male	45	>50K
31	Private	Masters	Never-married	Prof-specialty	Female	50	>50K
42	Private	Bachelors	Married-civ-	Exec-managerial	Male	40	>50K
37	Private	Some-college	Married-civ-	Exec-managerial	Male	80	>50K
30	State-gov	Bachelors	Married-civ-	Prof-specialty	Male	40	>50K
23	Private	Bachelors	Never-married	Adm-clerical	Female	30	<=50K
32	Private	Assoc-acdm	Never-married	Sales	Male	50	<=50K
[illegible]	[illegible]	[illegible]	Married-civ-	[illegible]	[illegible]	[illegible]	[illegible]

This is simply a data dump

Data Dump -> Insight

Age	Work	Education	Marital	Occupation	Gender	Hrs/	Incom
39	State-	Bachelors	Never-	Adm-clerical	Male	40	<=50K
50	Self-	Bachelors	Married-	Exec-	Male	13	<=50K
38	Private	HS-grad	Divorced	Handlers-	Male	40	<=50K
53	Private	11th	Married-	Handlers-	Male	40	<=50K
28	Private	Bachelors	Married-	Prof-specialty	Female	40	<=50K
37	Private	Masters	Married-	Exec-	Female	40	<=50K
49	Private	9th	Married-	Other-service	Female	16	<=50K
52	Self-	HS-grad	Married-	Exec-	Male	45	>50K
31	Private	Masters	Never-	Prof-specialty	Female	50	>50K
42	Private	Bachelors	Married-	Exec-	Male	40	>50K
37	Private	Some-	Married-	Exec-	Male	80	>50K
30	State-	Bachelors	Married-	Prof-specialty	Male	40	>50K
23	Private	Bachelors	Never-	Adm-clerical	Female	30	<=50K
32	Private	Assoc-	Never-	Sales	Male	50	<=50K
40	Private	Assoc-voc	Married-	Craft-repair	Male	40	>50K
34	Private	7th-8th	Married-	Transport-	Male	45	<=50K
25	Self-	HS-grad	Never-	Farming-	Male	35	<=50K
32	Private	HS-grad	Never-	Machine-op-	Male	40	<=50K
38	Private	11th	Married-	Sales	Male	50	<=50K
43	Self-	Masters	Divorced	Exec-	Female	45	>50K
40	Private	Doctorate	Married-	Prof-specialty	Male	60	>50K
54	Private	HS-grad	Separated	Other-service	Female	20	<=50K
35	Federal-	9th	Married-	Farming-	Male	40	<=50K
43	Private	11th	Married-	Transport-	Male	40	<=50K
59	Private	HS-grad	Divorced	Tech-support	Female	40	<=50K
56	Local-	Bachelors	Married-	Tech-support	Male	40	>50K
19	Private	HS-grad	Never-	Craft-repair	Male	40	<=50K
54	?	Some-	Married-	?	Male	60	>50K

Ask questions that transform raw data to insights

Data Dump -> Insight

Age	Work	Education	Marital	Occupation	Gender	Hrs/	Incom
39	State-	Bachelors	Never-	Adm-clerical	Male	40	<=50K
50	Self-	Bachelors	Married-	Exec-	Male	13	<=50K
38	Private	HS-grad	Divorced	Handlers-	Male	40	<=50K
53	Private	11th	Married-	Handlers-	Male	40	<=50K
28	Private	Bachelors	Married-	Prof-specialty	Female	40	<=50K
37	Private	Masters	Married-	Exec-	Female	40	<=50K
49	Private	9th	Married-	Other-service	Female	16	<=50K
52	Self-	HS-grad	Married-	Exec-	Male	45	>50K
31	Private	Masters	Never-	Prof-specialty	Female	50	>50K
42	Private	Bachelors	Married-	Exec-	Male	40	>50K
37	Private	Some-	Married-	Exec-	Male	80	>50K
30	State-	Bachelors	Married-	Prof-specialty	Male	40	>50K
23	Private	Bachelors	Never-	Adm-clerical	Female	30	<=50K
32	Private	Assoc-	Never-	Sales	Male	50	<=50K
40	Private	Assoc-voc	Married-	Craft-repair	Male	40	>50K
34	Private	7th-8th	Married-	Transport-	Male	45	<=50K
25	Self-	HS-grad	Never-	Farming-	Male	35	<=50K
32	Private	HS-grad	Never-	Machine-op-	Male	40	<=50K
38	Private	11th	Married-	Sales	Male	50	<=50K
43	Self-	Masters	Divorced	Exec-	Female	45	>50K
40	Private	Doctorate	Married-	Prof-specialty	Male	60	>50K
54	Private	HS-grad	Separated	Other-service	Female	20	<=50K
35	Federal-	9th	Married-	Farming-	Male	40	<=50K
43	Private	11th	Married-	Transport-	Male	40	<=50K
59	Private	HS-grad	Divorced	Tech-support	Female	40	<=50K
56	Local-	Bachelors	Married-	Tech-support	Male	40	>50K
19	Private	HS-grad	Never-	Craft-repair	Male	40	<=50K
54	?	Some-	Married-	?	Male	60	>50K

How well is the population educated?

Education	# People
HS-grad	?
Bachelors	?
Masters	?
Doctorate	?

Counts by education level

Data Dump -> Insight

Age	Work	Education	Marital	Occupation	Gender	Hrs/	Incom
39	State-	Bachelors	Never-	Adm-clerical	Male	40	<=50K
50	Self-	Bachelors	Married-	Exec-	Male	13	<=50K
38	Private	HS-grad	Divorced	Handlers-	Male	40	<=50K
53	Private	11th	Married-	Handlers-	Male	40	<=50K
28	Private	Bachelors	Married-	Prof-specialty	Female	40	<=50K
37	Private	Masters	Married-	Exec-	Female	40	<=50K
49	Private	9th	Married-	Other-service	Female	16	<=50K
52	Self-	HS-grad	Married-	Exec-	Male	45	>50K
31	Private	Masters	Never-	Prof-specialty	Female	50	>50K
42	Private	Bachelors	Married-	Exec-	Male	40	>50K
37	Private	Some-	Married-	Exec-	Male	80	>50K
30	State-	Bachelors	Married-	Prof-specialty	Male	40	>50K
23	Private	Bachelors	Never-	Adm-clerical	Female	30	<=50K
32	Private	Assoc-	Never-	Sales	Male	50	<=50K
40	Private	Assoc-voc	Married-	Craft-repair	Male	40	>50K
34	Private	7th-8th	Married-	Transport-	Male	45	<=50K
25	Self-	HS-grad	Never-	Farming-	Male	35	<=50K
32	Private	HS-grad	Never-	Machine-op-	Male	40	<=50K
38	Private	11th	Married-	Sales	Male	50	<=50K
43	Self-	Masters	Divorced	Exec-	Female	45	>50K
40	Private	Doctorate	Married-	Prof-specialty	Male	60	>50K
54	Private	HS-grad	Separated	Other-service	Female	20	<=50K
35	Federal-	9th	Married-	Farming-	Male	40	<=50K
43	Private	11th	Married-	Transport-	Male	40	<=50K
59	Private	HS-grad	Divorced	Tech-support	Female	40	<=50K
56	Local-	Bachelors	Married-	Tech-support	Male	40	>50K
19	Private	HS-grad	Never-	Craft-repair	Male	40	<=50K
54	?	Some-	Married-	?	Male	60	>50K

How do working hours vary based on occupation?

Occupation	# Hrs/Week
Adm-clerical	?
Prof-specialty	?
Craft-repair	?
Sales	?

Sum totals by occupation

Data Dump -> Insight

Age	Work	Education	Marital	Occupation	Gender	Hrs/	Incom
39	State-	Bachelors	Never-	Adm-clerical	Male	40	<=50K
50	Self-	Bachelors	Married-	Exec-	Male	13	<=50K
38	Private	HS-grad	Divorced	Handlers-	Male	40	<=50K
53	Private	11th	Married-	Handlers-	Male	40	<=50K
28	Private	Bachelors	Married-	Prof-specialty	Female	40	<=50K
37	Private	Masters	Married-	Exec-	Female	40	<=50K
49	Private	9th	Married-	Other-service	Female	16	<=50K
52	Self-	HS-grad	Married-	Exec-	Male	45	>50K
31	Private	Masters	Never-	Prof-specialty	Female	50	>50K
42	Private	Bachelors	Married-	Exec-	Male	40	>50K
37	Private	Some-	Married-	Exec-	Male	80	>50K
30	State-	Bachelors	Married-	Prof-specialty	Male	40	>50K
23	Private	Bachelors	Never-	Adm-clerical	Female	30	<=50K
32	Private	Assoc-	Never-	Sales	Male	50	<=50K
40	Private	Assoc-voc	Married-	Craft-repair	Male	40	>50K
34	Private	7th-8th	Married-	Transport-	Male	45	<=50K
25	Self-	HS-grad	Never-	Farming-	Male	35	<=50K
32	Private	HS-grad	Never-	Machine-op-	Male	40	<=50K
38	Private	11th	Married-	Sales	Male	50	<=50K
43	Self-	Masters	Divorced	Exec-	Female	45	>50K
40	Private	Doctorate	Married-	Prof-specialty	Male	60	>50K
54	Private	HS-grad	Separated	Other-service	Female	20	<=50K
35	Federal-	9th	Married-	Farming-	Male	40	<=50K
43	Private	11th	Married-	Transport-	Male	40	<=50K
59	Private	HS-grad	Divorced	Tech-support	Female	40	<=50K
56	Local-	Bachelors	Married-	Tech-support	Male	40	>50K
19	Private	HS-grad	Never-	Craft-repair	Male	40	<=50K
54	?	Some-	Married-	?	Male	60	>50K

Roughly how many hours do people work in a week?

Hrs/Week	
Minimum	?
Maximum	?
Average	?

Aggregates at the overall level

Data Dump -> Insight

Age	Work	Education	Marital	Occupation	Gender	Hrs/	Incom
39	State-	Bachelors	Never-	Adm-clerical	Male	40	<=50K
50	Self-	Bachelors	Married-	Exec-	Male	13	<=50K
38	Private	HS-grad	Divorced	Handlers-	Male	40	<=50K
53	Private	11th	Married-	Handlers-	Male	40	<=50K
28	Private	Bachelors	Married-	Prof-specialty	Female	40	<=50K
37	Private	Masters	Married-	Exec-	Female	40	<=50K
49	Private	9th	Married-	Other-service	Female	16	<=50K
52	Self-	HS-grad	Married-	Exec-	Male	45	>50K
31	Private	Masters	Never-	Prof-specialty	Female	50	>50K
42	Private	Bachelors	Married-	Exec-	Male	40	>50K
37	Private	Some-	Married-	Exec-	Male	80	>50K
30	State-	Bachelors	Married-	Prof-specialty	Male	40	>50K
23	Private	Bachelors	Never-	Adm-clerical	Female	30	<=50K
32	Private	Assoc-	Never-	Sales	Male	50	<=50K
40	Private	Assoc-voc	Married-	Craft-repair	Male	40	>50K
34	Private	7th-8th	Married-	Transport-	Male	45	<=50K
25	Self-	HS-grad	Never-	Farming-	Male	35	<=50K
32	Private	HS-grad	Never-	Machine-op-	Male	40	<=50K
38	Private	11th	Married-	Sales	Male	50	<=50K
43	Self-	Masters	Divorced	Exec-	Female	45	>50K
40	Private	Doctorate	Married-	Prof-specialty	Male	60	>50K
54	Private	HS-grad	Separated	Other-service	Female	20	<=50K
35	Federal-	9th	Married-	Farming-	Male	40	<=50K
43	Private	11th	Married-	Transport-	Male	40	<=50K
59	Private	HS-grad	Divorced	Tech-support	Female	40	<=50K
56	Local-	Bachelors	Married-	Tech-support	Male	40	>50K
19	Private	HS-grad	Never-	Craft-repair	Male	40	<=50K
54	?	Some-	Married-	?	Male	60	>50K

Do working hours vary based on family circumstances?

	Hrs/Week		
Marital Status	**Min**	**Max**	**Avg**
Never-Married	?	?	?
Married	?	?	?
Divorced	?	?	?
Separated	?	?	?

Aggregates at a group level

Summaries such as Count, Sum, Average

Aggregates at an overall level

Aggregates at a group level

Numeric Summarizations have a distinct pattern

Summarizing Data

Summary metric	**Pick from sum, min, max, average etc** Computed for a specific column containing numeric values
Aggregation level	**Pick from overall level, group level** Groups are distinct values from a specific column

Summarizing Data

Summary metric

Aggregation level

	Hrs/Week		
Marital Status	Min	Max	Avg
Never-Married	?	?	?
Married	?	?	?
Divorced	?	?	?
Separated	?	?	?

Summarizing Data

Summary metric

Aggregation level

	Hrs/Week		
Marital Status	Min	Max	Avg
Never-Married	?	?	?
Married	?	?	?
Divorced	?	?	?
Separated	?	?	?

MapReduce Summarization

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40

?

?

Summary

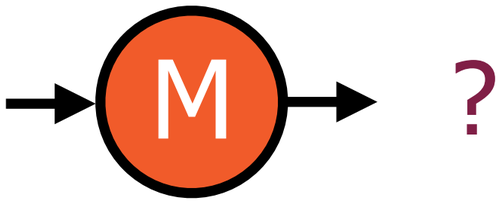
Marital Status	Avg Hrs/Wk

Remember the 2 questions to ask when we set up a MapReduce

Map Step

Raw data dump

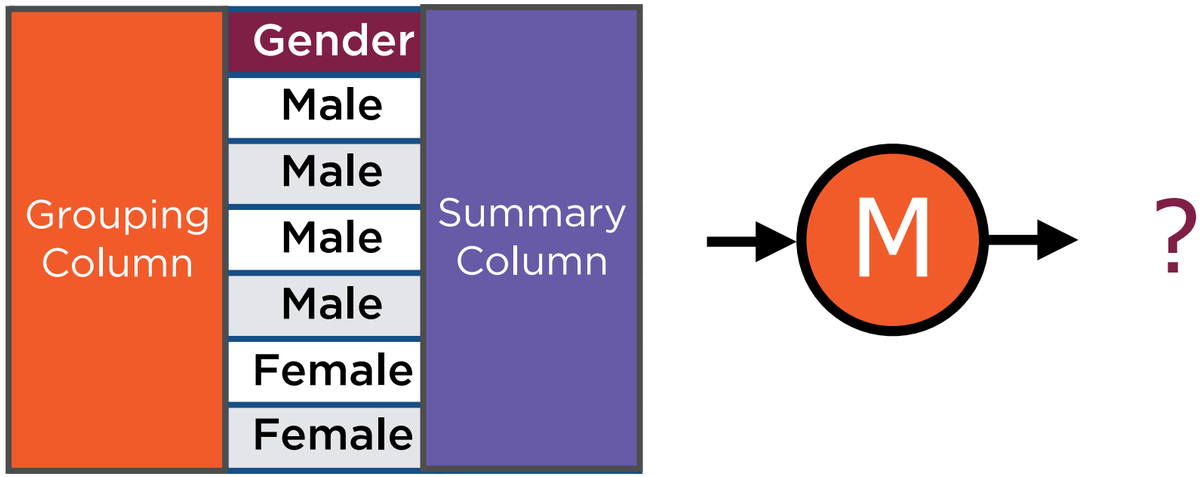
Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40



The output depends on the columns chosen for summarizing, grouping

Map Step

Raw data dump



Map Step

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40

Grouping Column

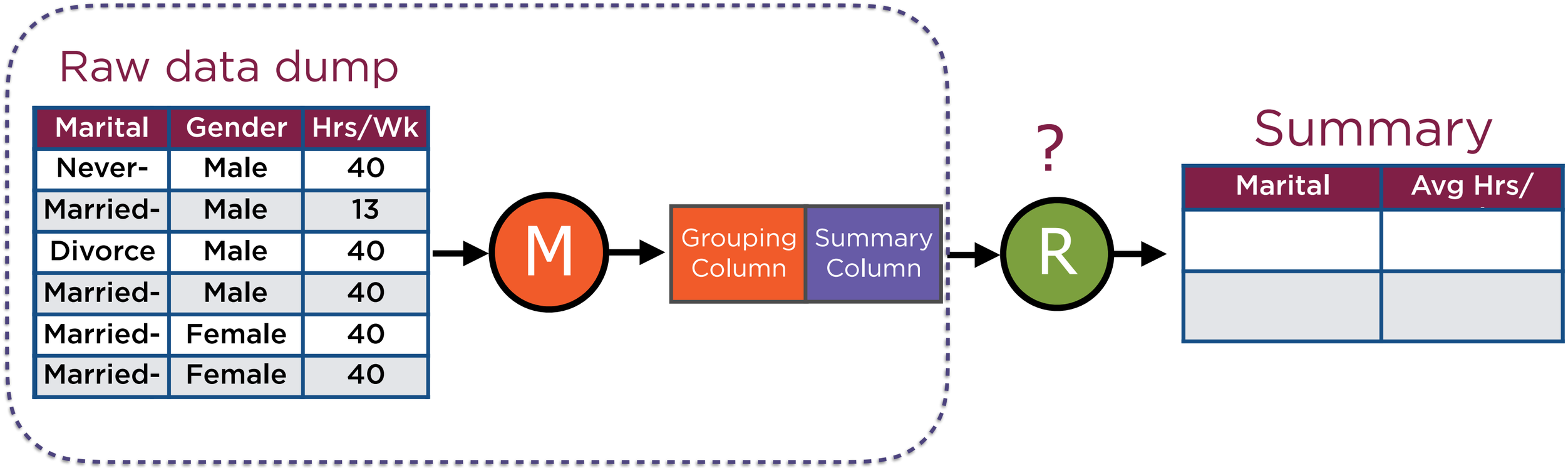
Summary Column

Never-	40
Married-civ-	13
Divorced	40
Married-civ-	40
Married-civ-	40
Married-civ-	40

Key

Value

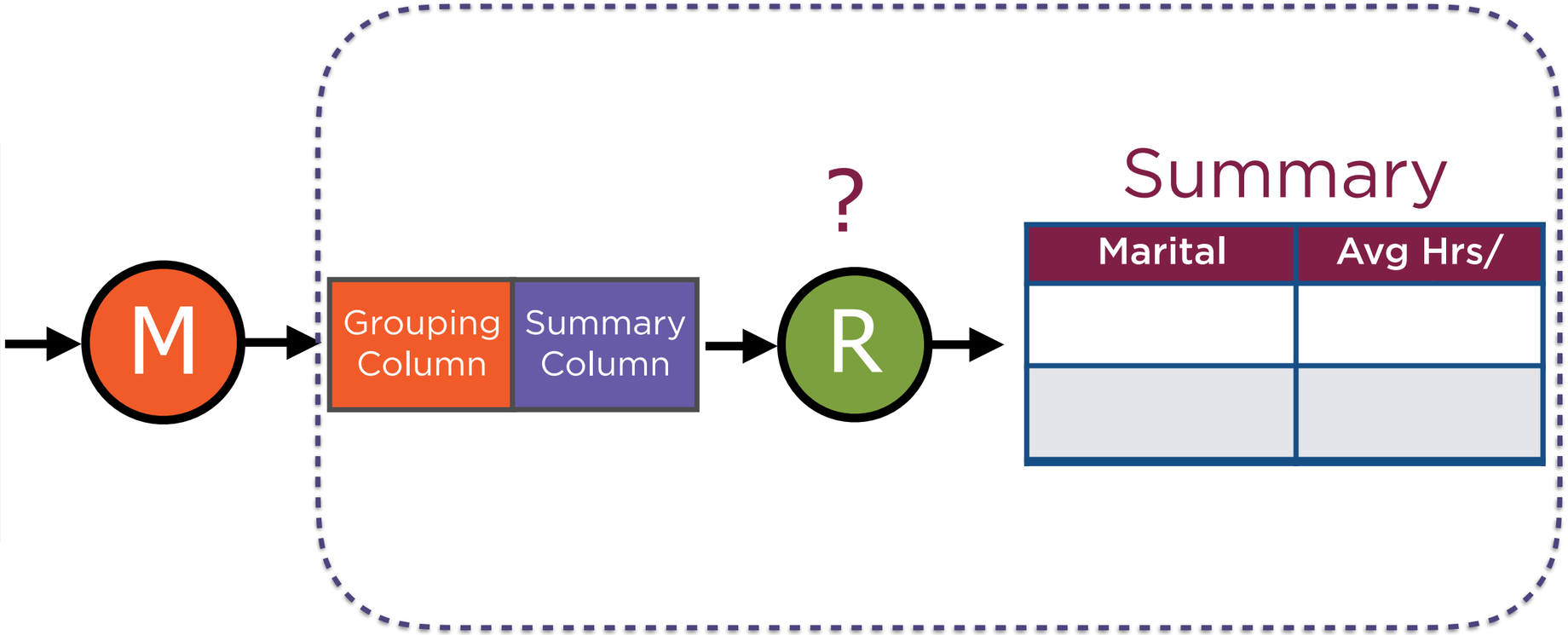
MapReduce Summarization



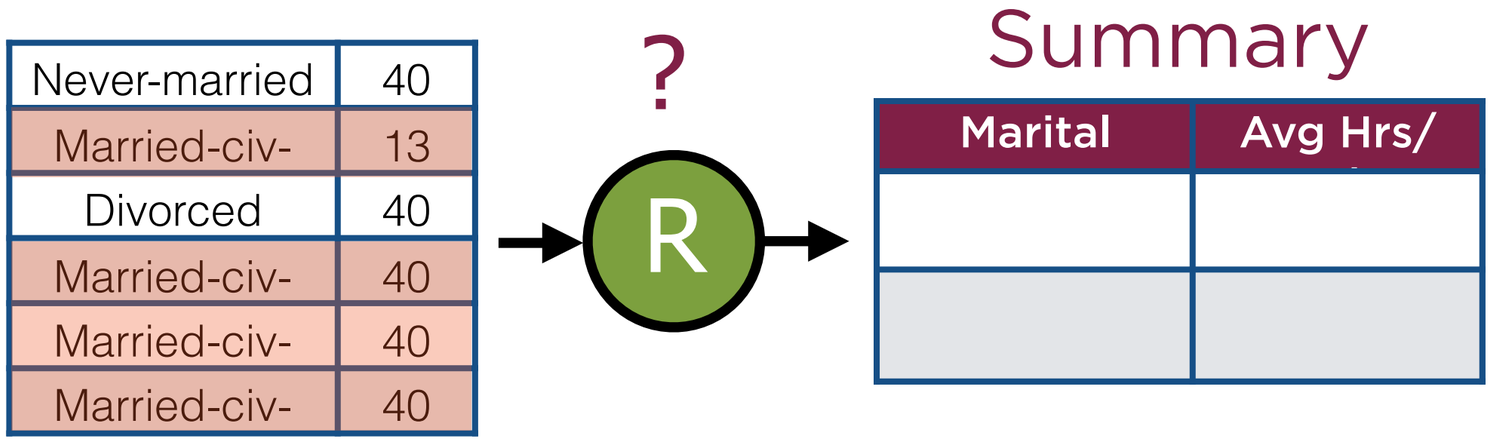
MapReduce Summarization

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40

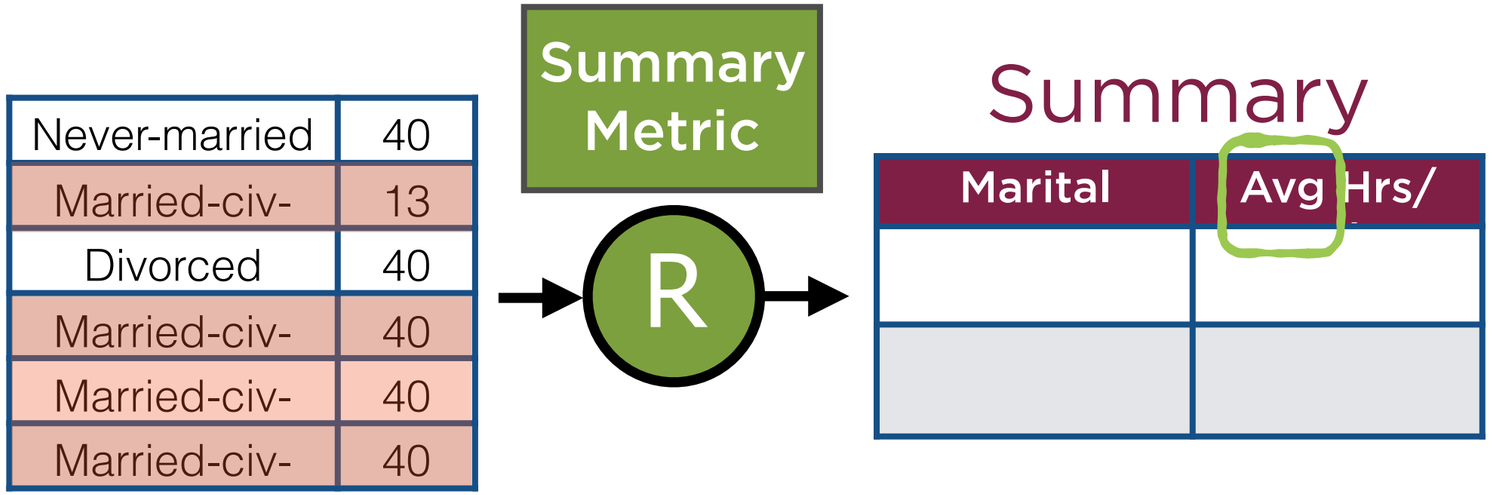


Reduce Step



The reduce step combines values with the same key

Reduce Step

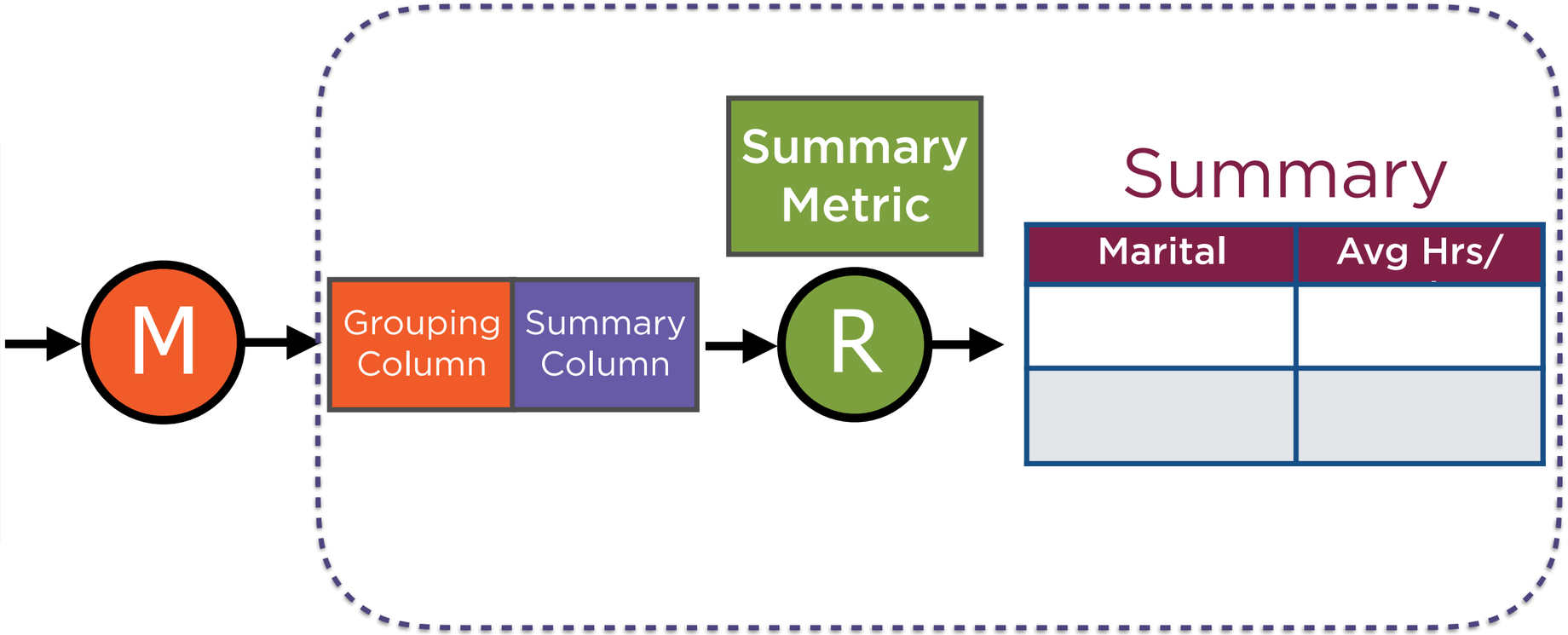


The combining logic depends on the summary metric chosen

MapReduce Summarization

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40



Summarizing Data

Summary metric

Determines the reducer logic

Compute this metric for values with the same key

Aggregation level

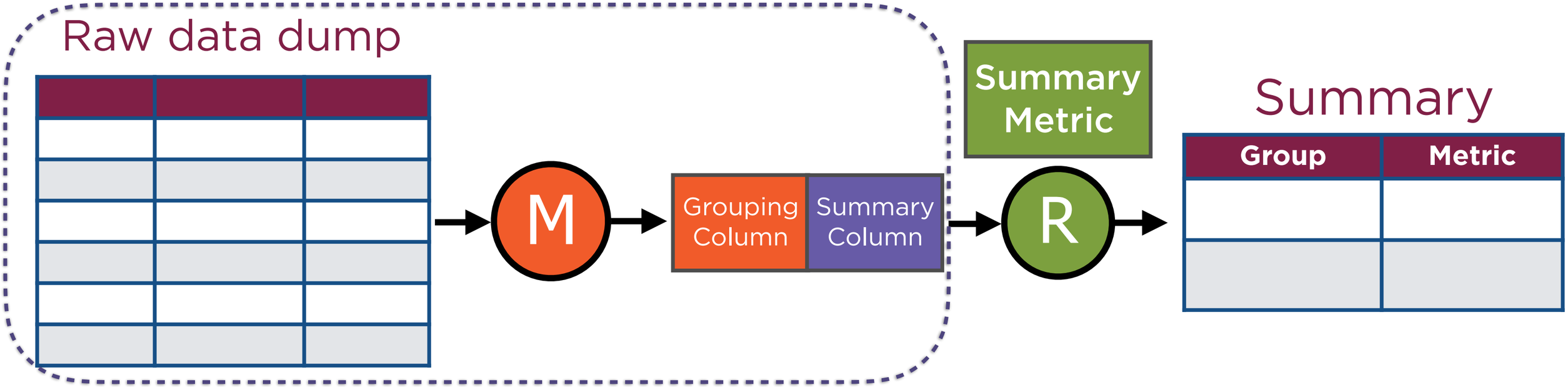
Determines the key of the map output

Key = Grouping Column Value

Demo

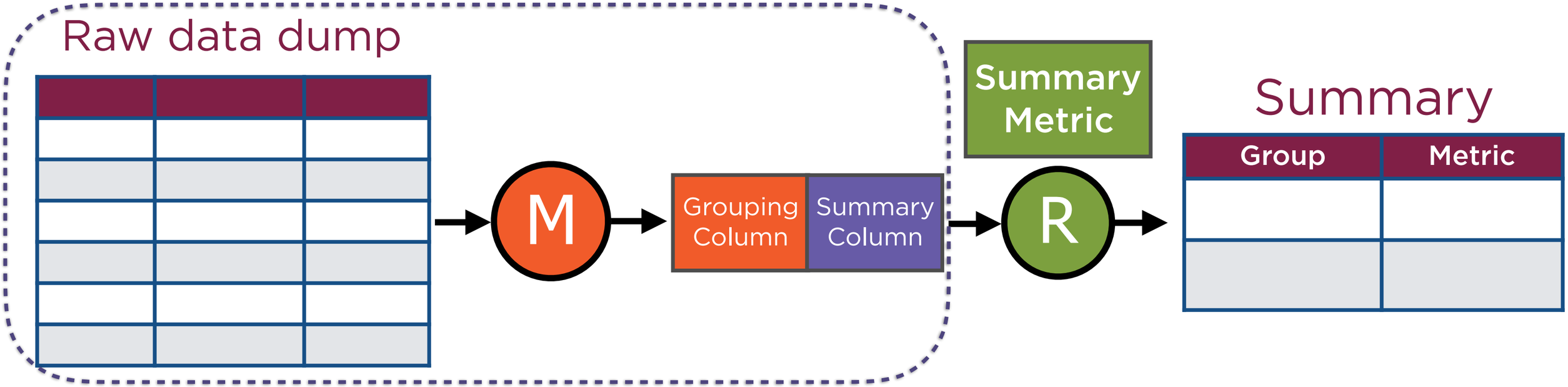
Implement a MapReduce to calculate, on average, how many hours the population works, depending on their marital status

MapReduce Summarization



What happens if we introduce a combiner step?

Summary with Combiner



What happens if we introduce a combiner step?

Combiner Function

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40

Never-	40
Married-civ-	13
Divorced	40
Married-civ-	40
Married-civ-	40
Married-civ-	40

Combiner Function

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40

Never-	40
Married-civ-	13
Divorced	40
Married-civ-	40
Married-civ-	40
Married-civ-	40

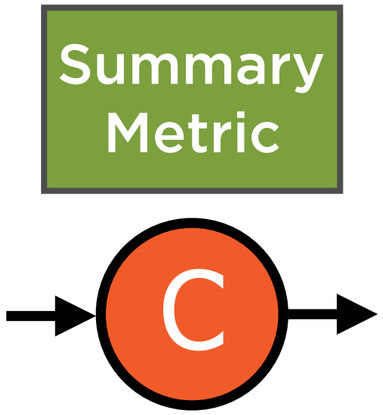
Marital	Max
Never-	40
Married-civ-	40
Divorced	40

Combine values with the same key before they are copied over to the reducer

Combiner Function

If the summary metric is max, min, sum, count

Never-	40
Married-civ-	13
Divorced	40
Married-civ-	40
Married-civ-	40
Married-civ-	40



Marital	Max
Never-	40
Married-civ-	40
Divorced	40

The Combiner logic is the same as the reducer!

No impact on the final result

MapReduce result with and without combiners should be the same

Reducers as Combiners

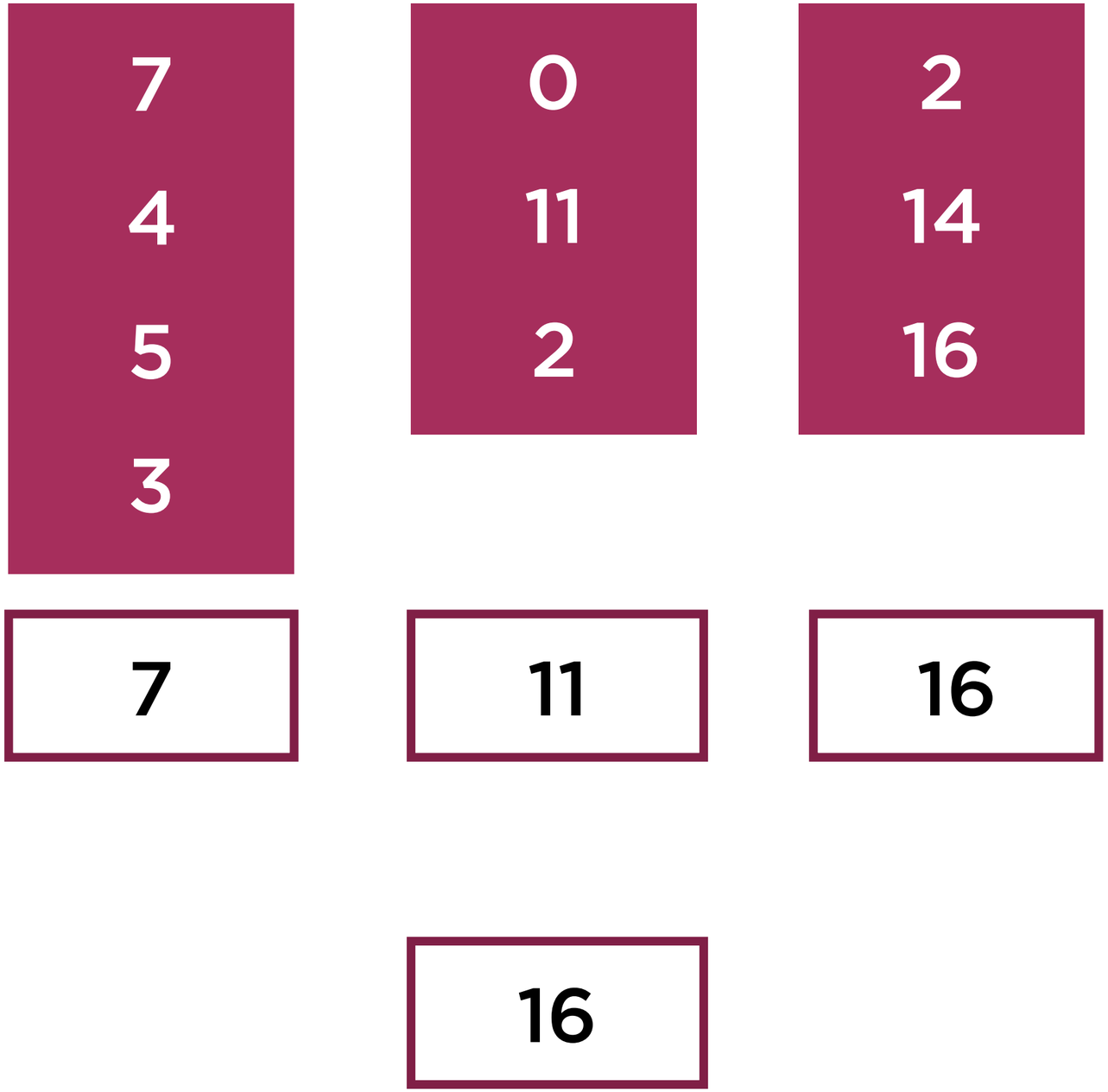
Maximum

Minimum

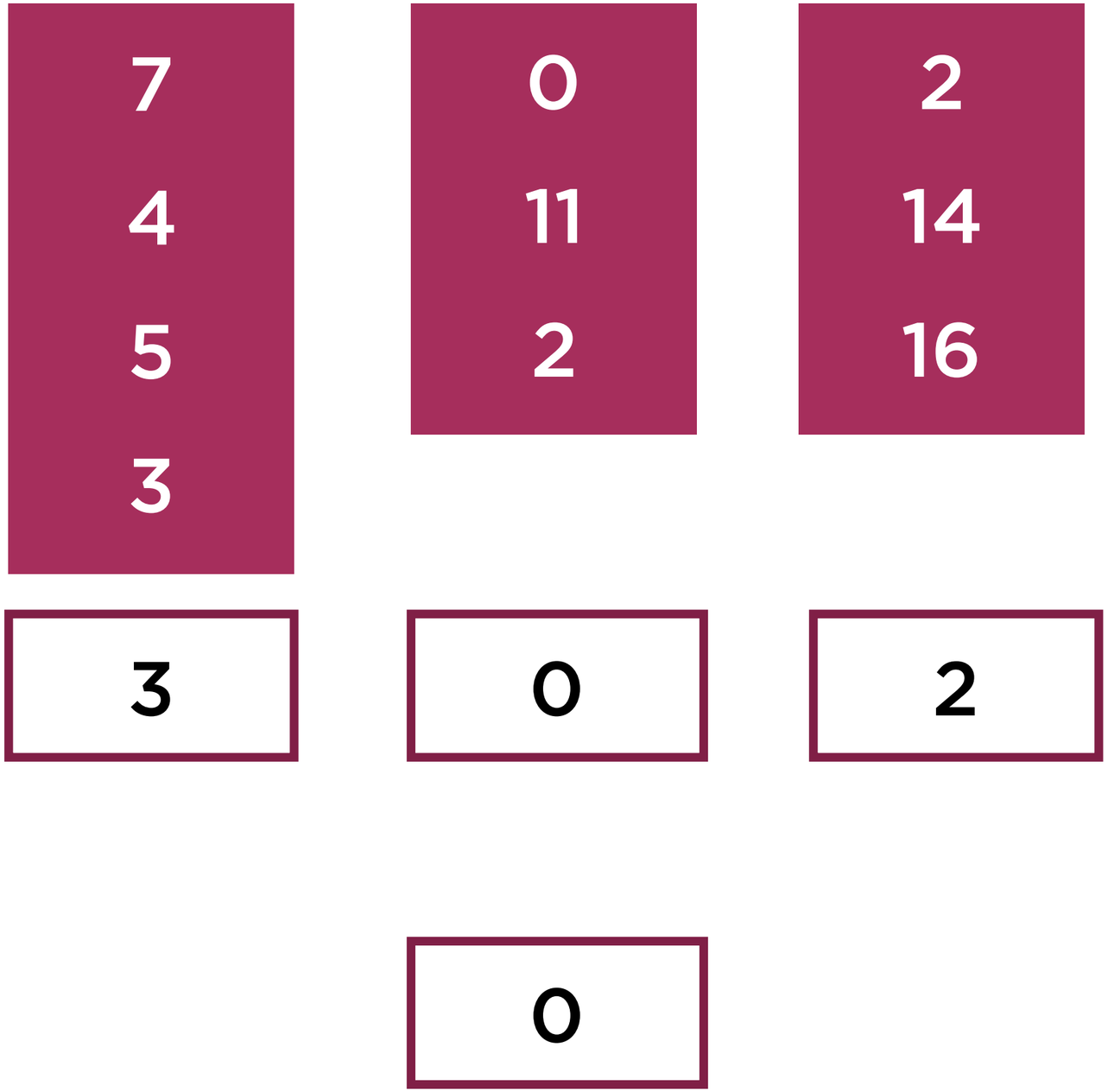
Sum

The combiner performs the same operation as the reducer

Maximum



Minimum



Sum

7	0	2
4	11	14
5	2	16
3		
19	13	32
	64	

Reducers as Combiners

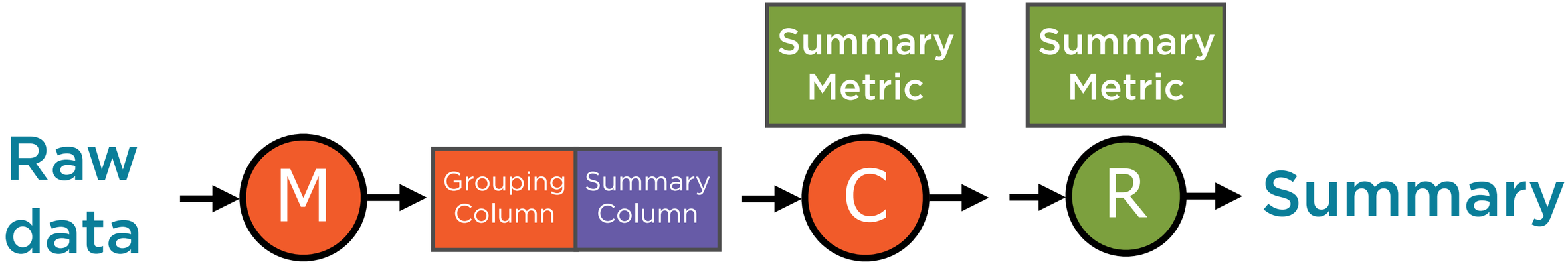
Maximum

Minimum

Sum

The combiner performs the same operation as the reducer

Summary with Combiner



This pattern will not work for average

Reducers as Combiners

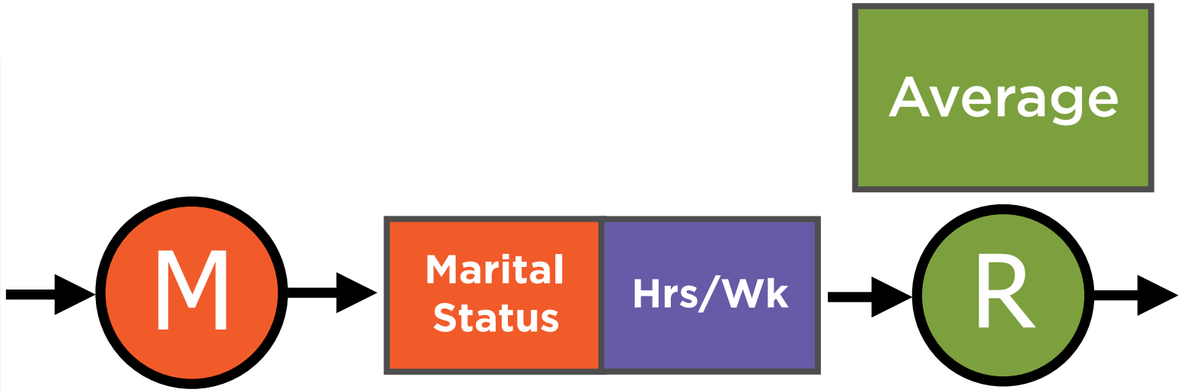
Average

Combiner and reducer functions have to be different!

MapReduce Average

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40



Summary

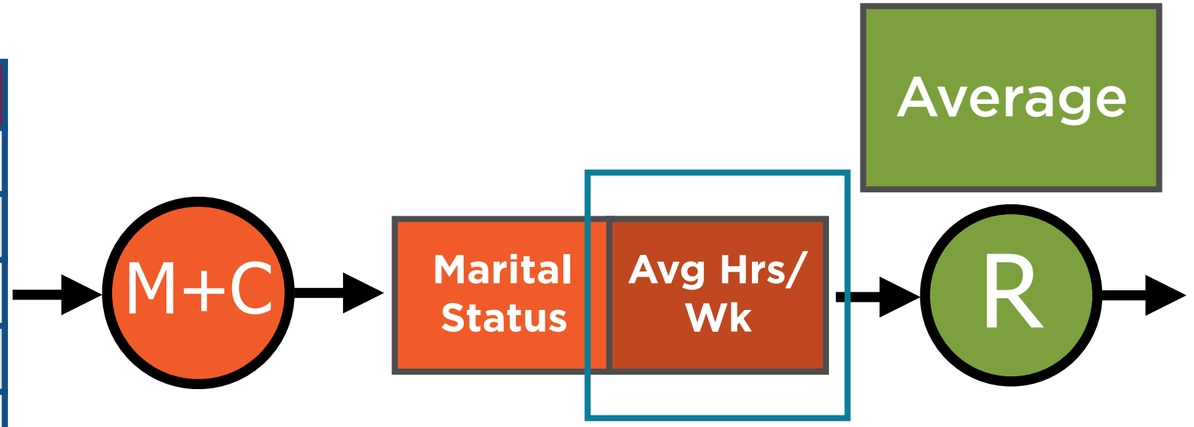
Marital Status	Avg Hrs/Wk

Without a combiner

MapReduce Average

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40



Summary

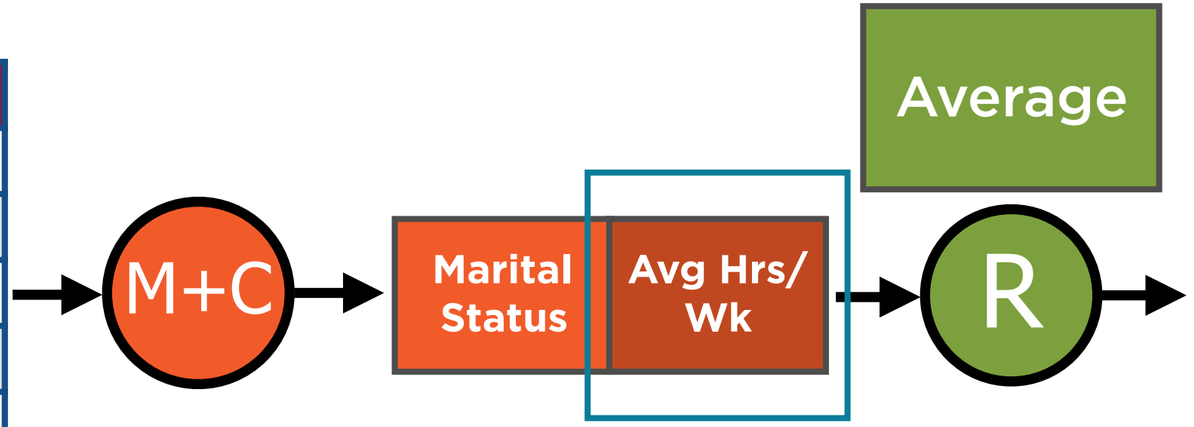
Marital Status	Avg Hrs/Wk

What if we used the Reducer as the combiner function?

MapReduce Average

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40

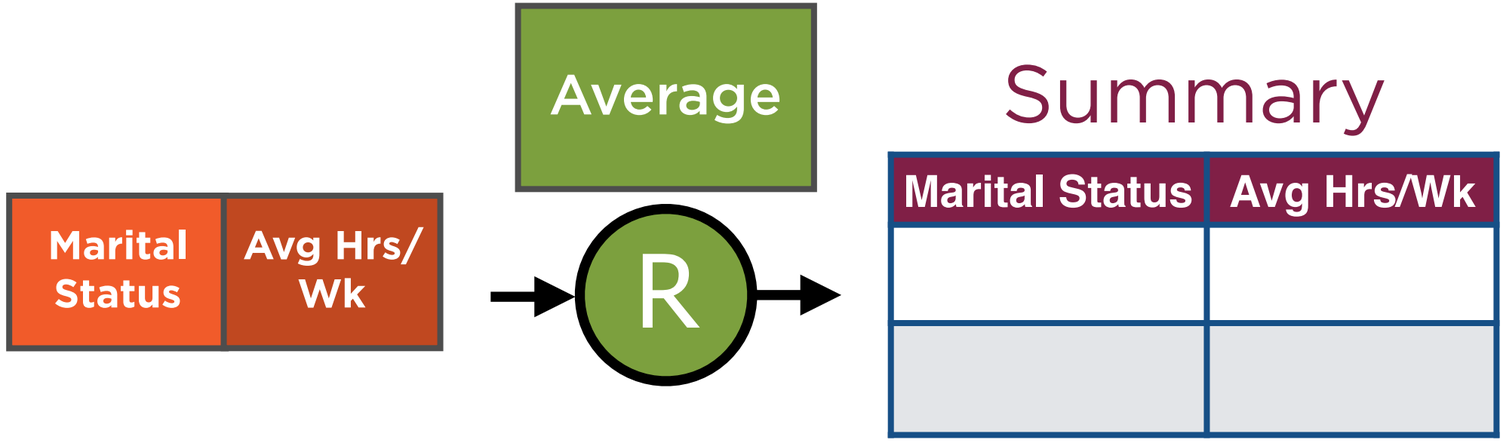


Summary

Marital Status	Avg Hrs/Wk

This is wrong!

MapReduce Average



Average of a set of numbers <> Average (Averages of subsets)

Reducers as Combiners

Average

Combiner and reducer functions have to be different!

Average

7	3	3
4	2	11
6	4	16
3		
5	3	10
	6	Wrong!

Average

7	3	3
4	2	11
6	4	16
3		

Correct answer is

5.9

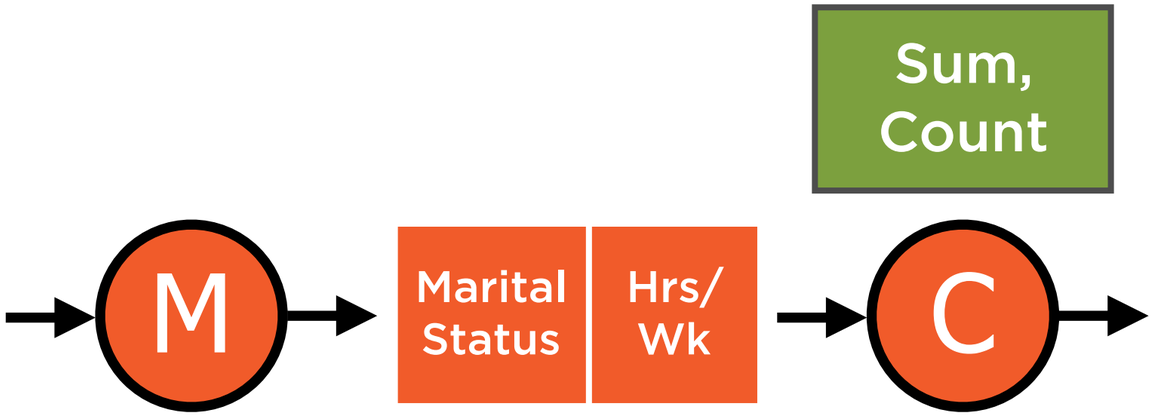
Average

How do we implement MapReduce average using a combiner?

Average with Combiner

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40



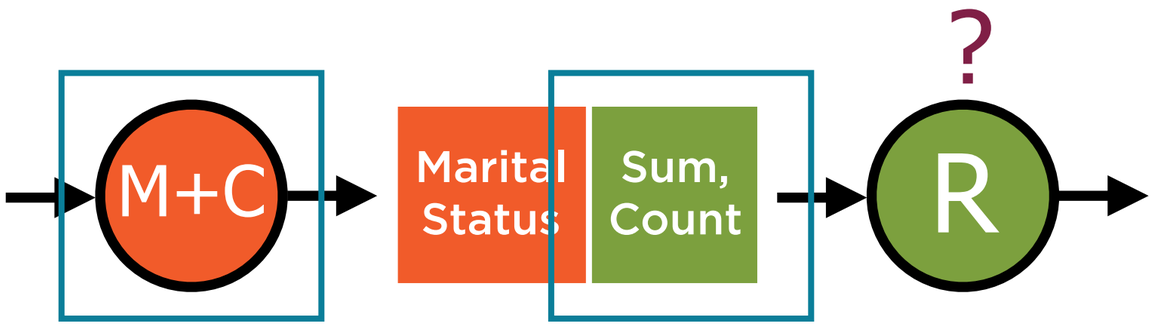
Marital	Sum,Count
Never-	40,1
Married-civ-	133, 4
Divorced	40, 1

Output a tuple from each combiner

Average with Combiner

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40

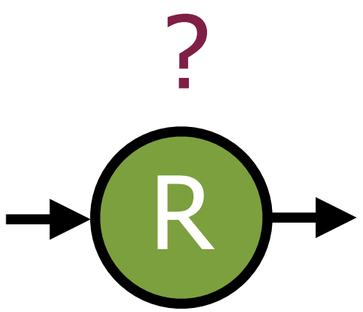


Summary

Marital	Avg Hrs/Wk

Average with Combiner

Marital Status	Sum,Count
Never-married	40,1
Married-civ-	133, 4
Divorced	40, 1
Married-civ-	30, 1



Summary

Marital	Avg Hrs/

Combine tuples with the same key

Average with Combiner

Marital	Sum,Count
Married-civ-	133, 4
Married-civ-	30, 1

Marital Status	Sum,Count
Married-civ-spouse	163, 5

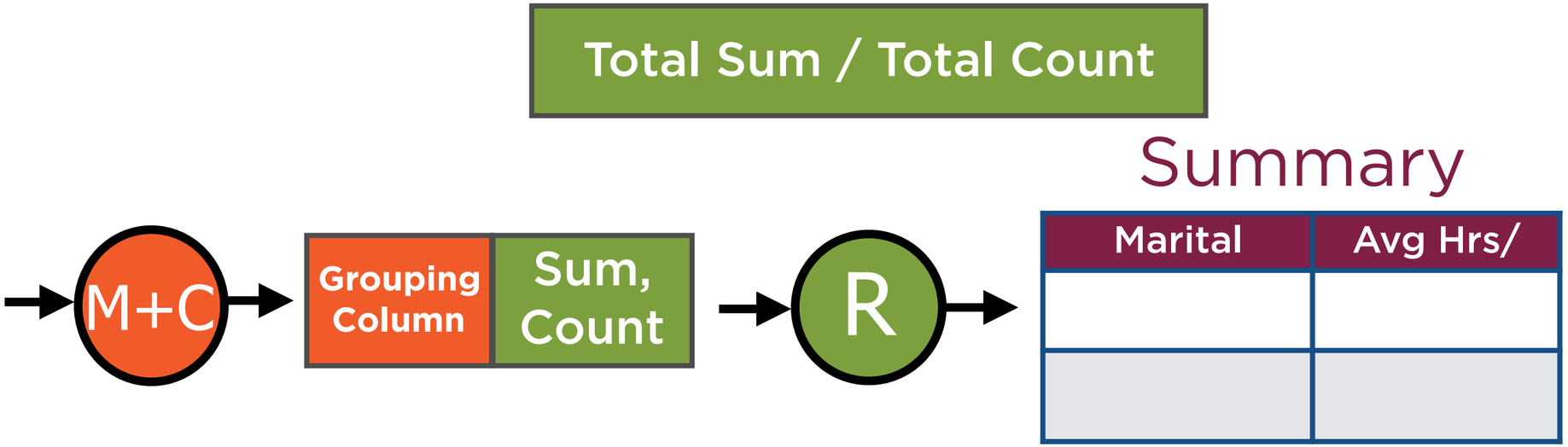
Marital Status	Avg Hrs/Wk
Married-civ-spouse	32.6

1. Sum the sums
2. Sum the counts
3. Compute a ratio

Average with Combiner

Raw data dump

Marital	Gender	Hrs/Wk
Never-	Male	40
Married-	Male	13
Divorce	Male	40
Married-	Male	40
Married-	Female	40
Married-	Female	40



Average with Combiner

Output of the mapper/combiner

Input to the reducer

Writable Interface

(Sum, Count)
of Summary
Column

The value is a tuple of integers (Sum, Count)

This needs a custom Writable type

Writable Interface

Hadoop has a bunch of Writable classes

Text

IntWritable

LongWritable

Writable Interface

Text

IntWritable

LongWritable

These act as wrappers around the regular Java primitives

Writable Interface

Text

IntWritable

LongWritable

They implement the Writable Interface

```java
public interface Writable {
    void write(DataOutput var1) throws IOException;

    void readFields(DataInput var1) throws IOException;
}
```

Writable Interface

All the Writable classes implement the write() and readFields() methods

Writable Interface

Text

IntWritable

LongWritable

The Writable classes we know actually inherit from a subInterface of Writable and java.lang.Comparable

WritableComparable

```
public interface WritableComparable<T>
extends Writable, Comparable<T> {
}
```

WritableComparable

In addition to the readFields() and write() methods, these classes also implement compareTo()

Demo

Implementing a Combiner with a Custom Writable

Summary

Understood patterns in calculating numeric summaries using MapReduce

Used a Combiner correctly based on the kind of numeric summaries

Implemented a MapReduce to calculate averages using a Custom Writable class