Filtering Large Data Sets

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Filter datasets based on a condition

Find a distinct set of values within a dataset

Find the top N records in input data

Filtering Data

Consider a dataset with user purchases

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

Which users spent >300?

How many of them bought Books?

Selecting a **specific** set of records from a dataset

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

Selecting a specific set of records

If this were a database table

An SQL query

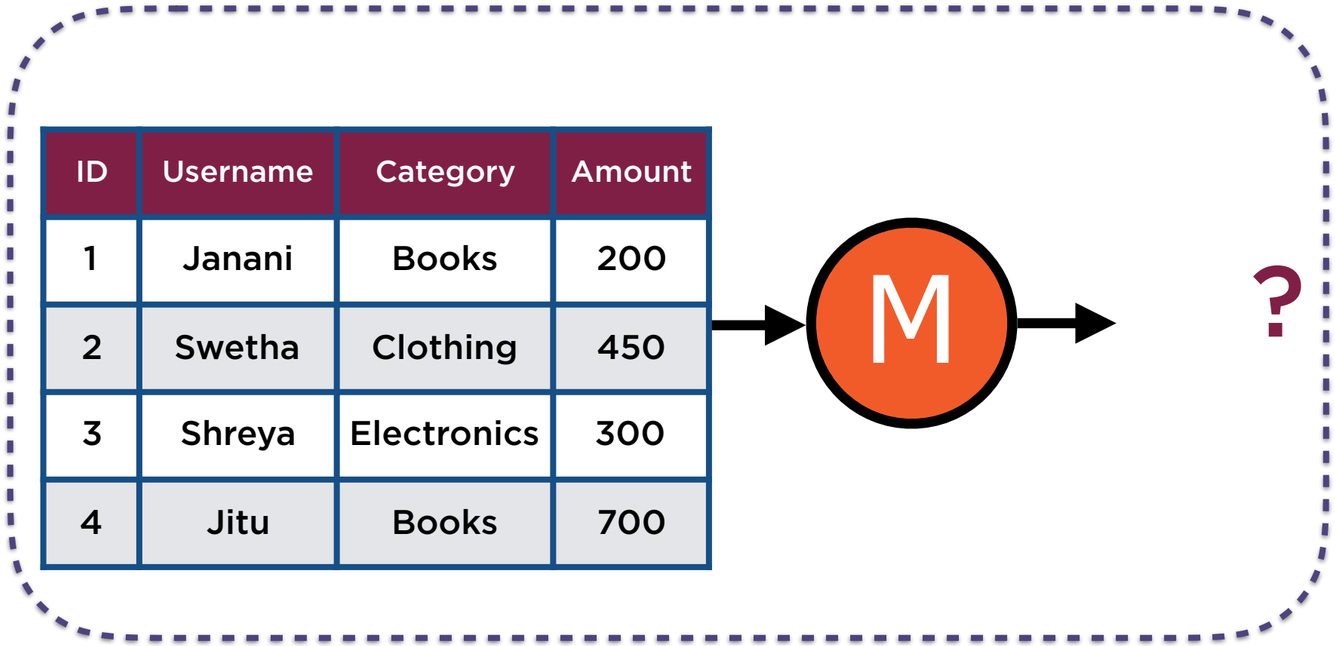
```
select * from <table name>
where <condition>
```

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

In Hadoop/Distributed Computing setups

MapReduce

Filtering Data



Filter users who spent >300

Map Step

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

?

Output should contain the filtered rows

Map Step

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

Key	Value		
Null	2	Swetha	Clothing	450
Null	4	Jitu	Books	700

Output should contain the filtered rows

Map Step

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

Key	Value		
Null	2	Swetha	Clothing	450
Null	4	Jitu	Books	700

The output of the Map is the final output!

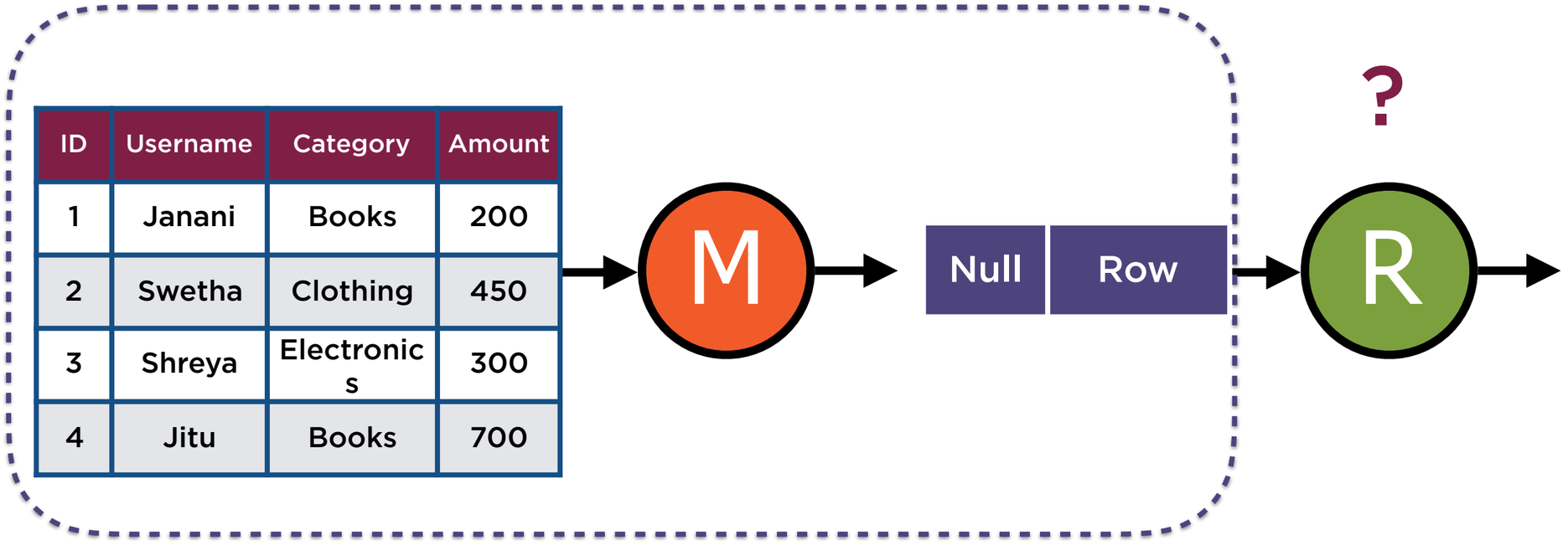
Map Step

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700

Key	Value		
Null	2	Swetha	Clothing	450
Null	4	Jitu	Books	700

The Reducer doesn't need to do anything!

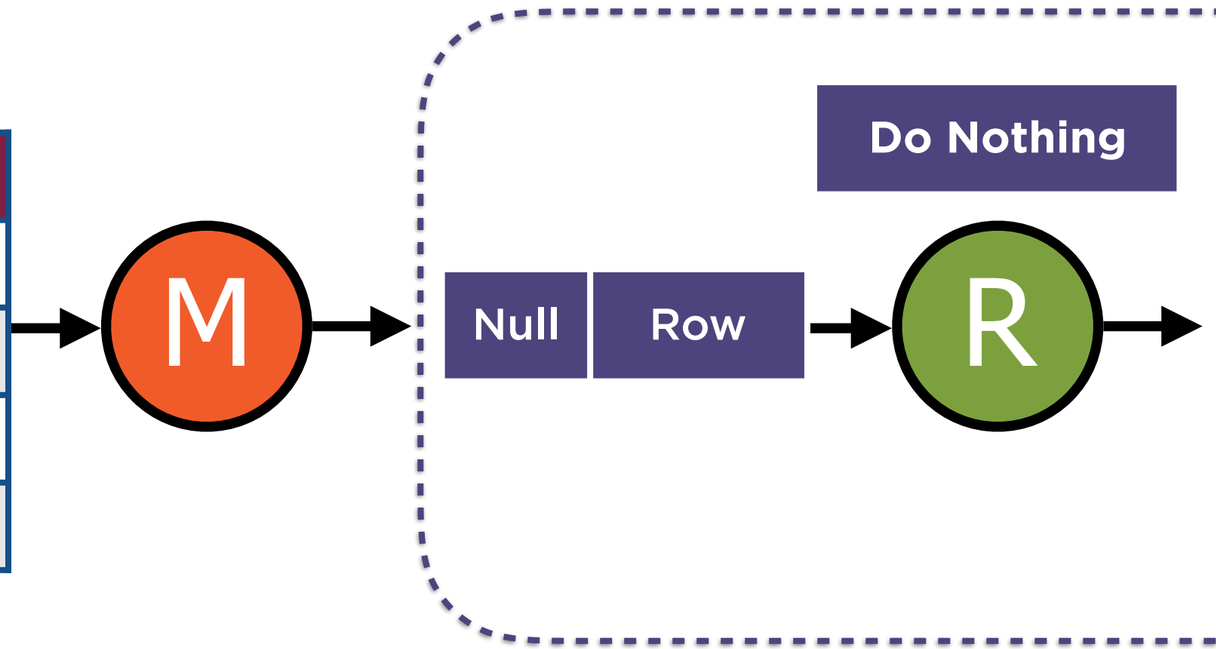
Filtering Data



ID	Username	Category	Amount
2	Swetha	Clothing	450
4	Jitu	Books	700

Filtering Data

ID	Username	Category	Amount
1	Janani	Books	200
2	Swetha	Clothing	450
3	Shreya	Electronics	300
4	Jitu	Books	700



ID	Username	Category	Amount
2	Swetha	Clothing	450
4	Jitu	Books	700

Demo

Implement a basic filter

Finding Distinct Values

Consider internal search data for a website

Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants

Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants

What are the distinct search terms from all searches?

Unique values from the keyword column

Finding Distinct Values

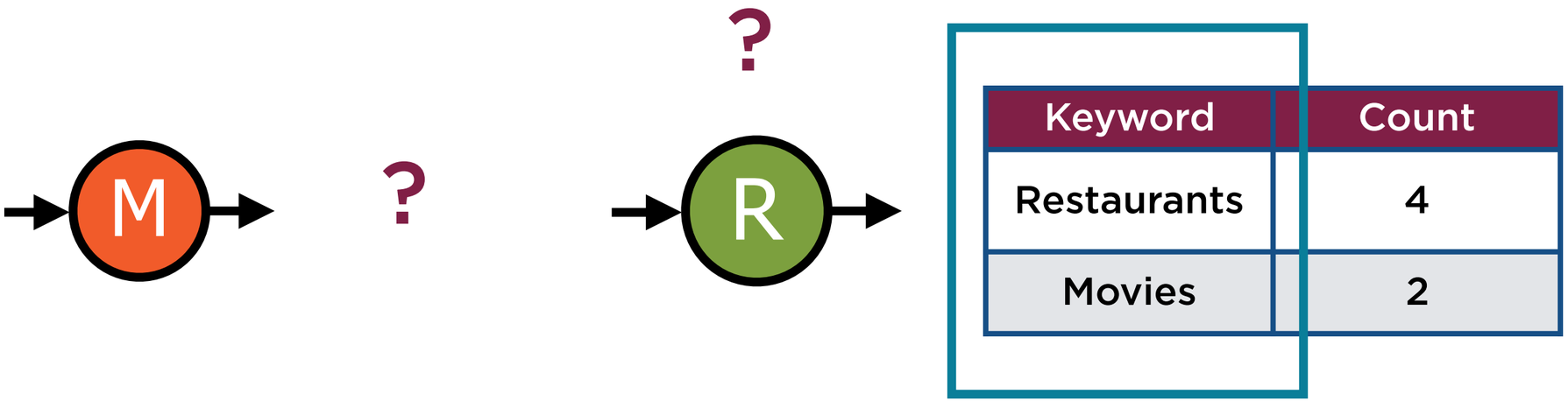
Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants

Keyword
Restaurants
Movies

This is similar to the summarization pattern

Finding Distinct Values

Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants



We need the group levels from the summarization output

Finding Distinct Values

Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants

Keyword	Count
Restaurants	4
Movies	2

And are not interested in the summary

Finding Distinct Values

Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants

?

?

Keyword
Restaurants
Movies

Map Step

Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants

Keyword	Count
Restaurants	1
Movies	1
Restaurants	1
Restaurants	1
Movies	1
Restaurants	1

Grouping Column	Summary Column

Just like the summarization pattern

Map Step

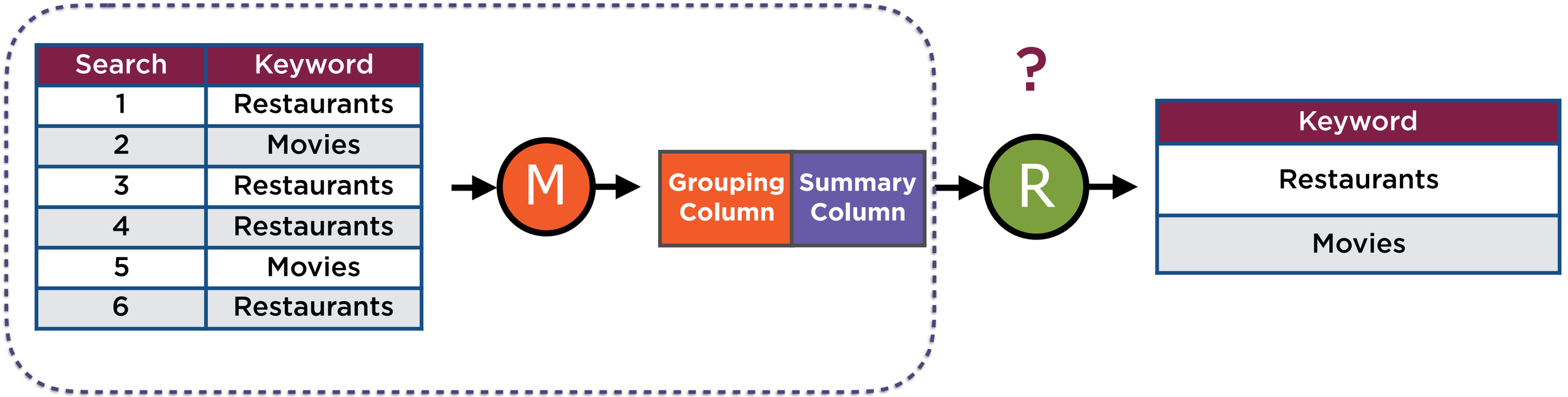
Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants

Keyword	Count
Restaurants	1
Movies	1
Restaurants	1
Restaurants	1
Movies	1
Restaurants	1

Grouping Column | Summary Column

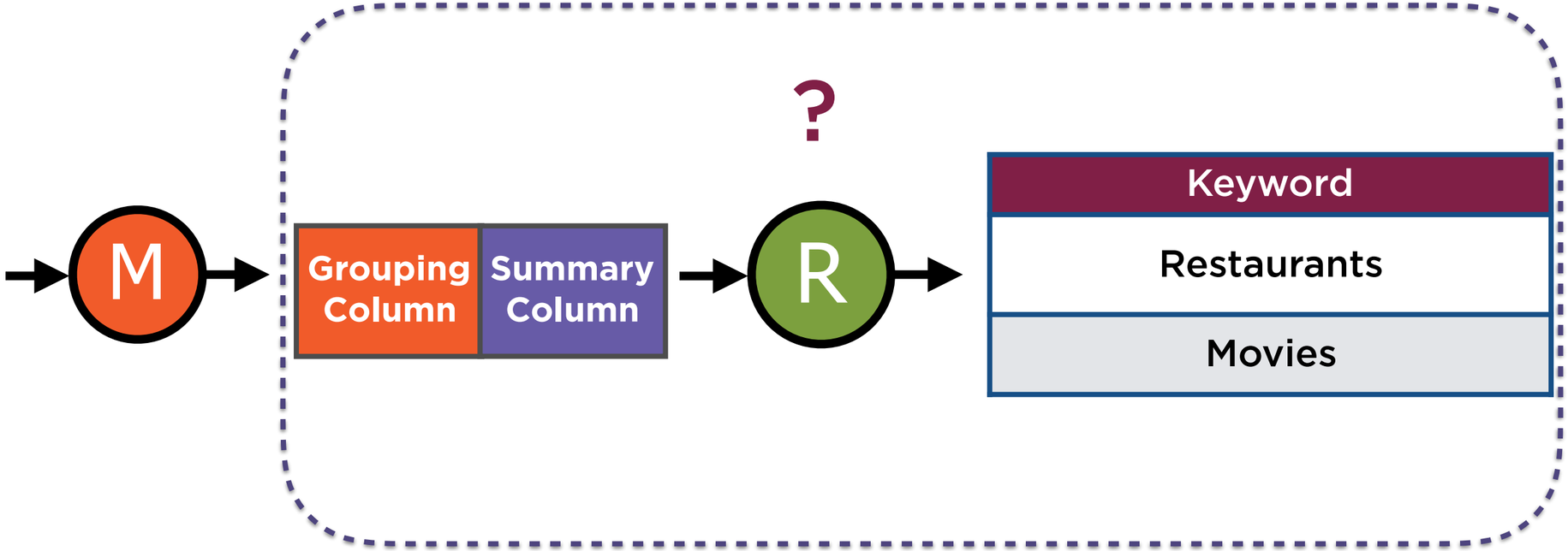
grouping column = column for which we want distinct values

Finding Distinct Values

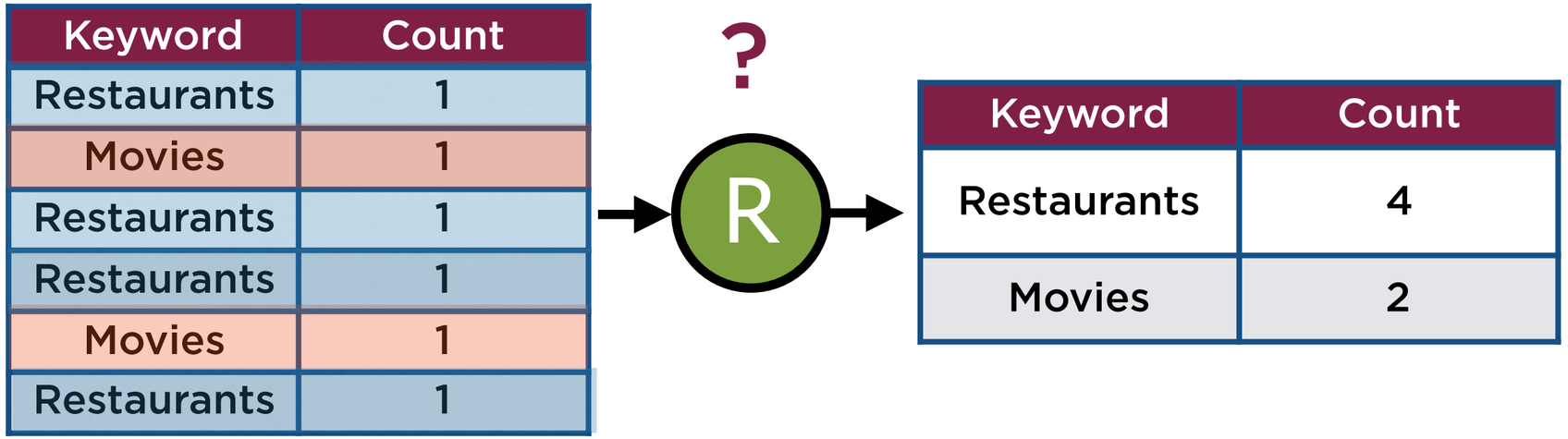


Finding Distinct Values

Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants

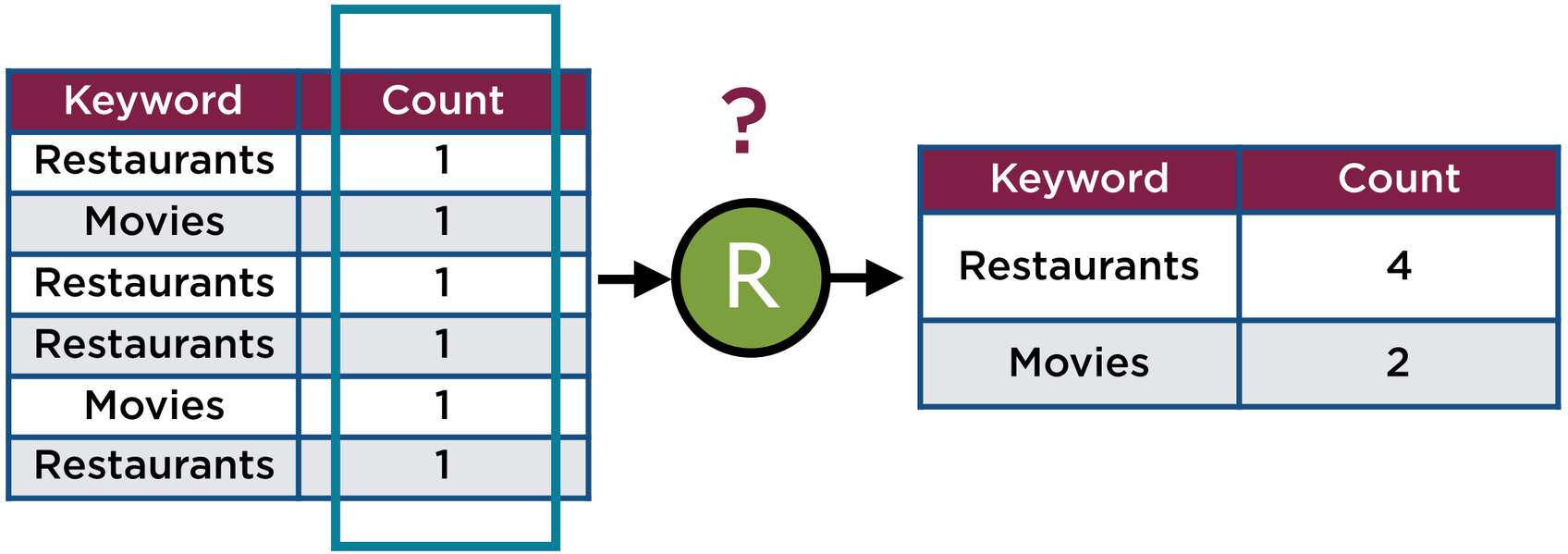


Reduce Step



The reduce step combines values with the same key

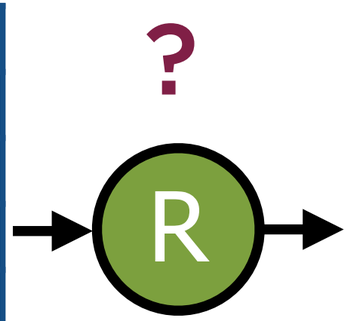
Reduce Step



The combining logic depends on the summary metric chosen

Reduce Step

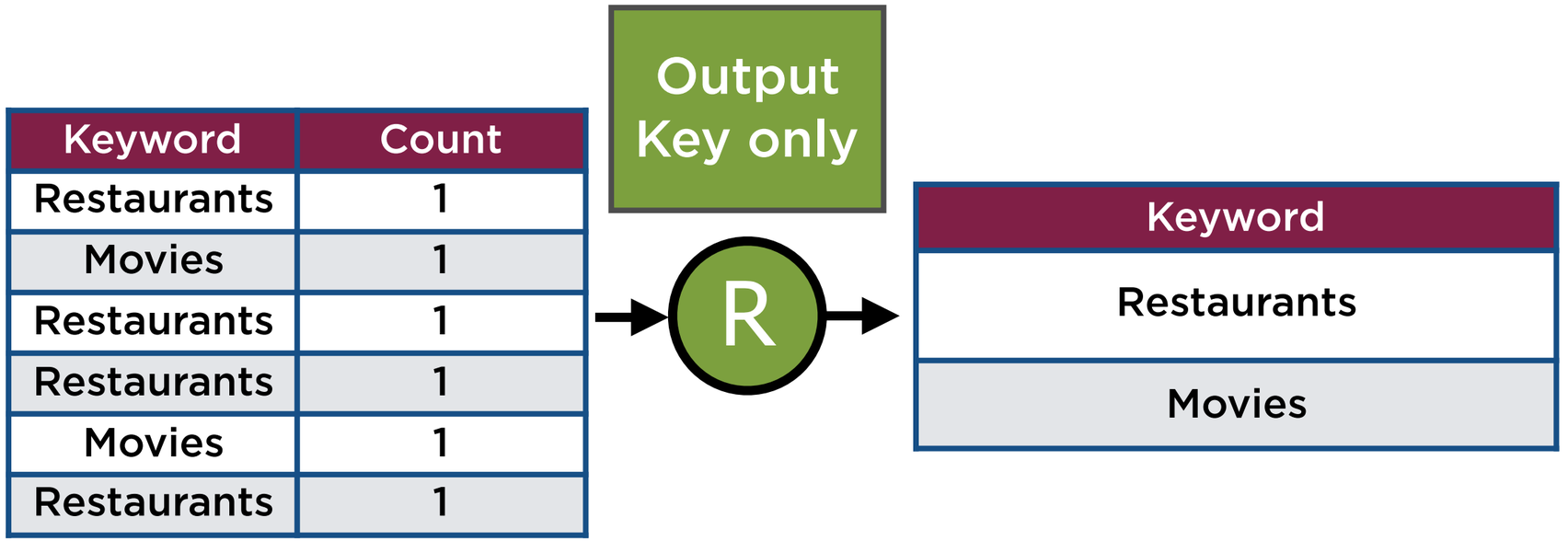
Keyword	Count
Restaurants	1
Movies	1
Restaurants	1
Restaurants	1
Movies	1
Restaurants	1



Keyword	Count
Restaurants	4
Movies	2

In this case actual summarization can be ignored!

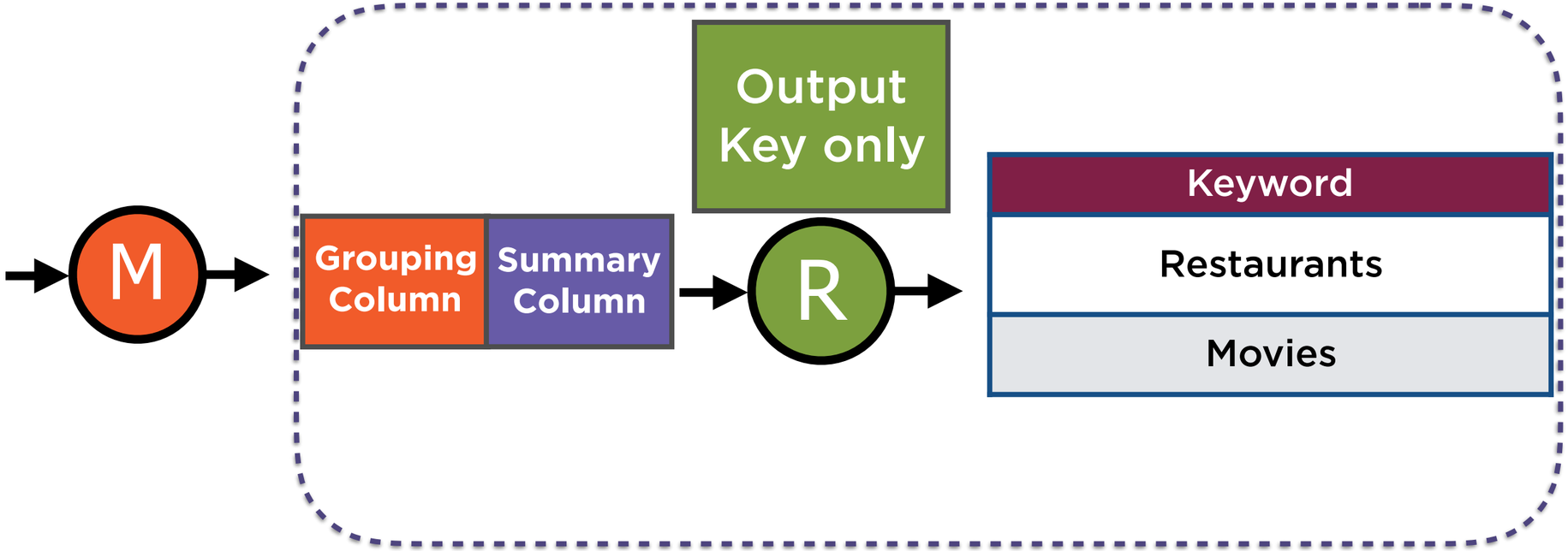
Reduce Step



Output only the keys

Finding Distinct Values

Search	Keyword
1	Restaurants
2	Movies
3	Restaurants
4	Restaurants
5	Movies
6	Restaurants



Demo

Filter distinct set of search terms from a given search data set

Filtering Data

Consider a dataset with users of a social network

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

Which users are the most influential?

Get the top N records

Get the top N records

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

This is a sorting problem

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

Get the top N records

If this were a database table

An SQL query order by

```
select * from <table name>
where <condition>
order by <column name>
```

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

Sort in parallel with MapReduce

Top N

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

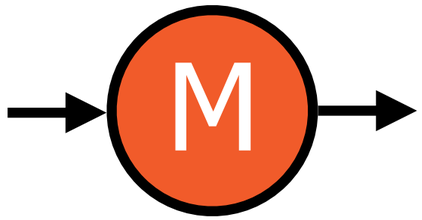
?

?

Top 3 Users
2
6
5

Map Step

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

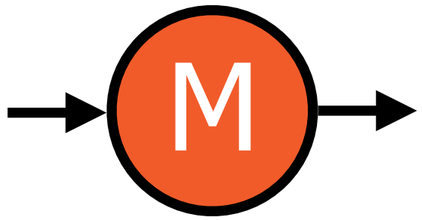


?

Each mapper works on a subset of the data

Map Step

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

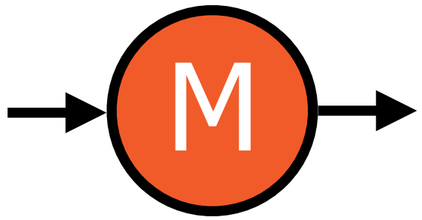


?

And can pick the top N only for that subset!

Map Step

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

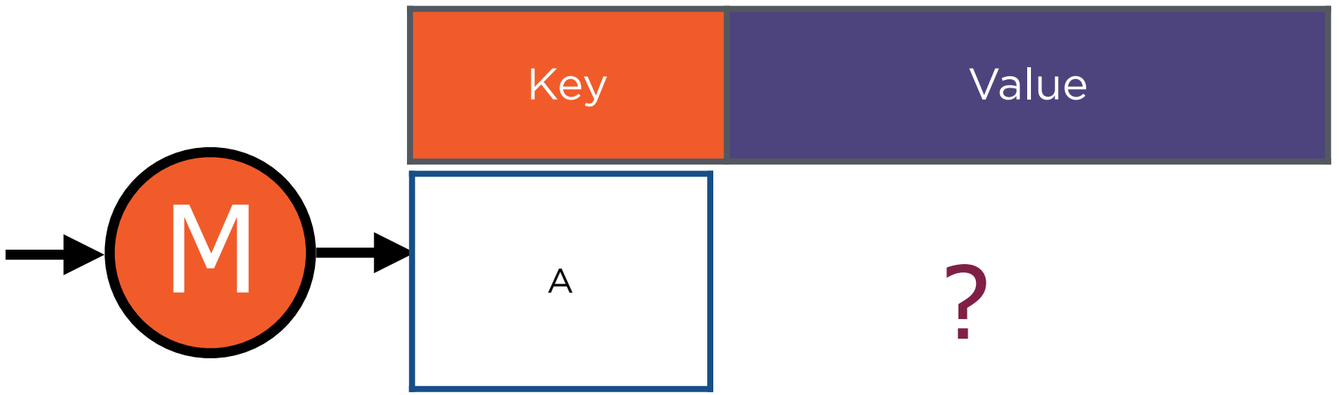


?

A mapper on one node has no idea what data exists on other nodes in the cluster

Map Step

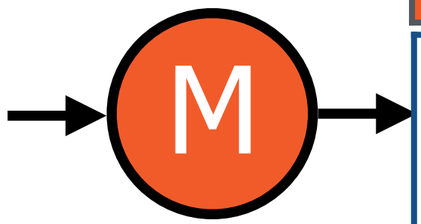
User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000



All mappers should output the same key

Map Step

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

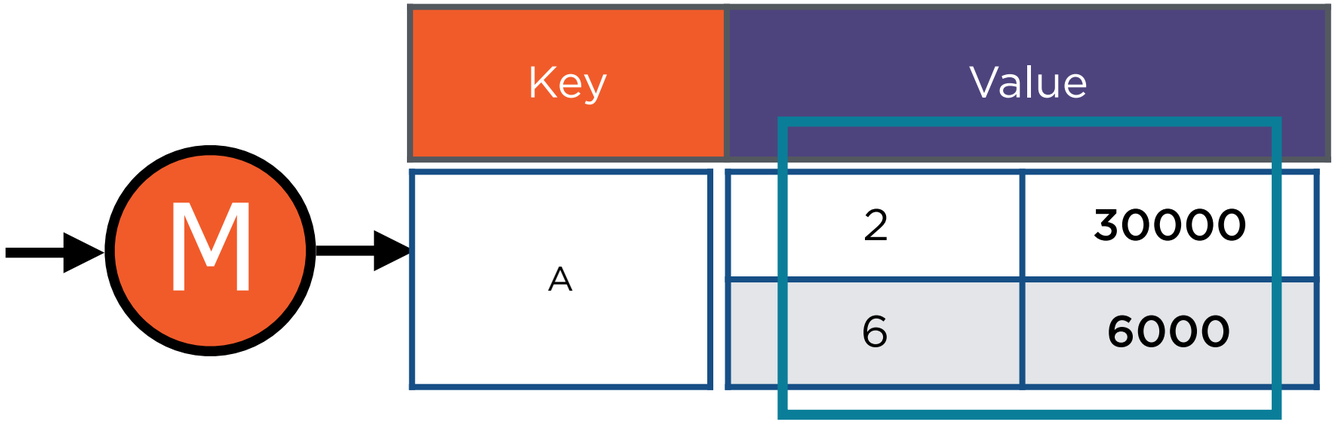


Key	Value	
A	2	30000
	6	6000

Assume “A” is the common key output by every Mapper process

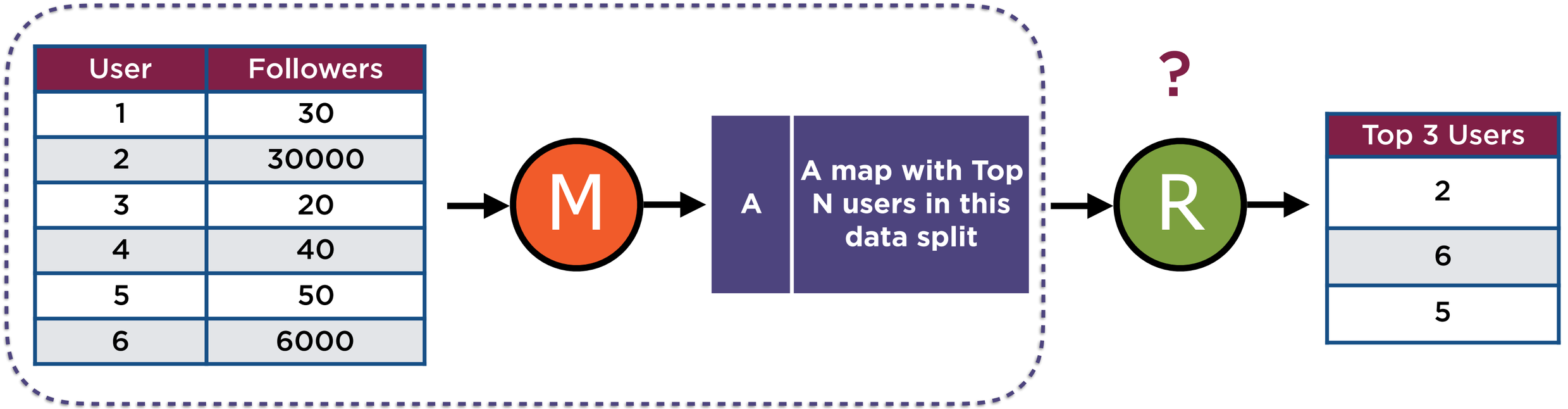
Map Step

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000



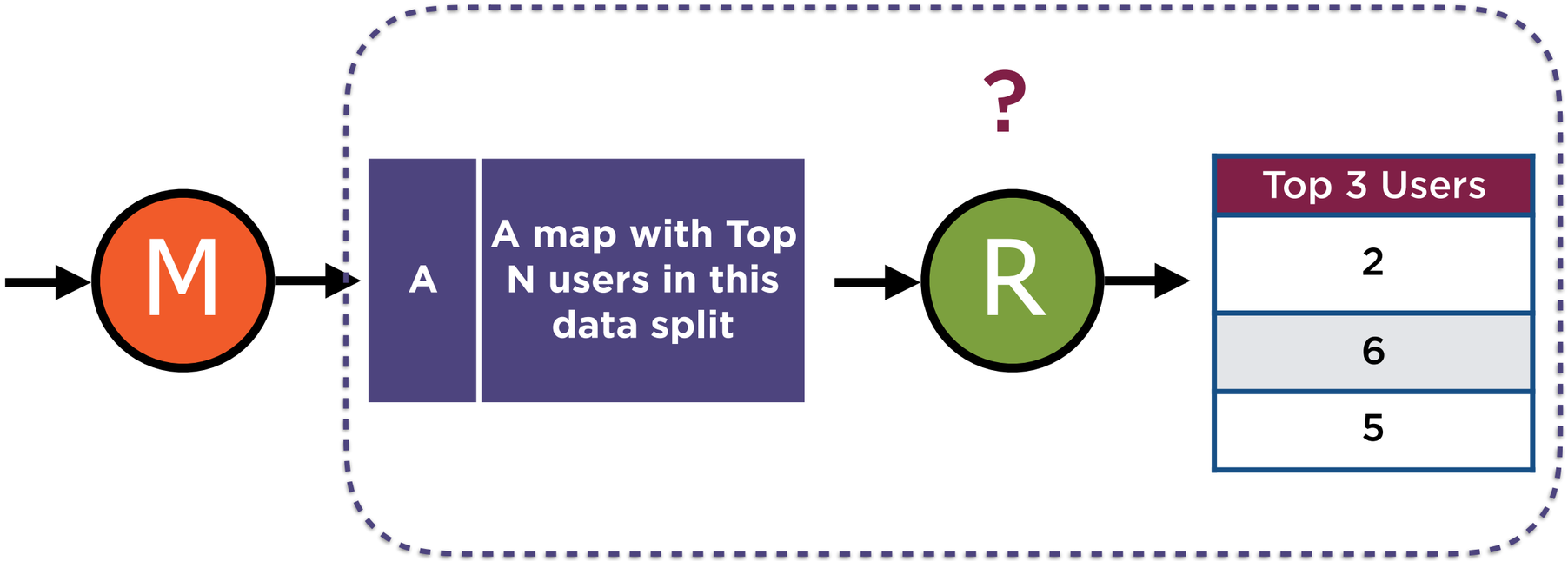
The value is a list of records of users who have the most followers

Top N

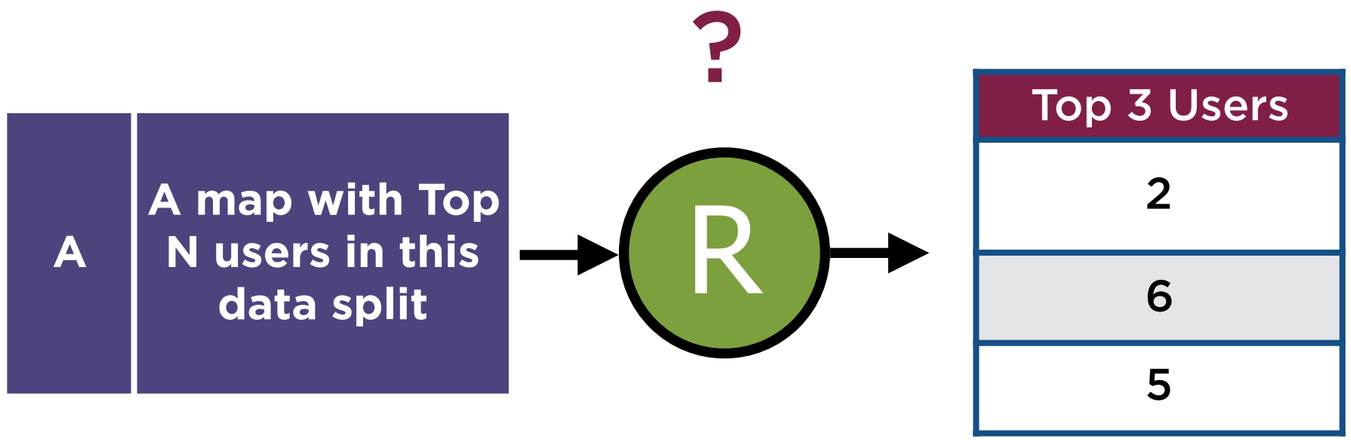

User
Followers
1
30
2
30000
3
20
4
40
5
50
6
6000
M
A
A map with Top N users in this data split
?
R
Top 3 Users
2
6
5

Top N

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000

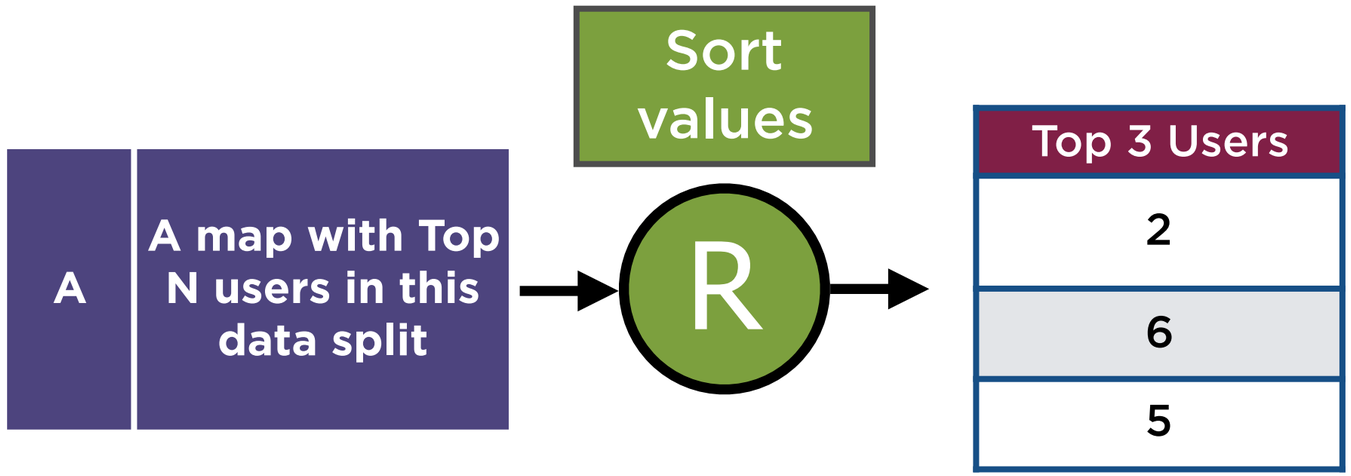


Top N



The reducer will collect the top N records from each mapper into one list

Top N

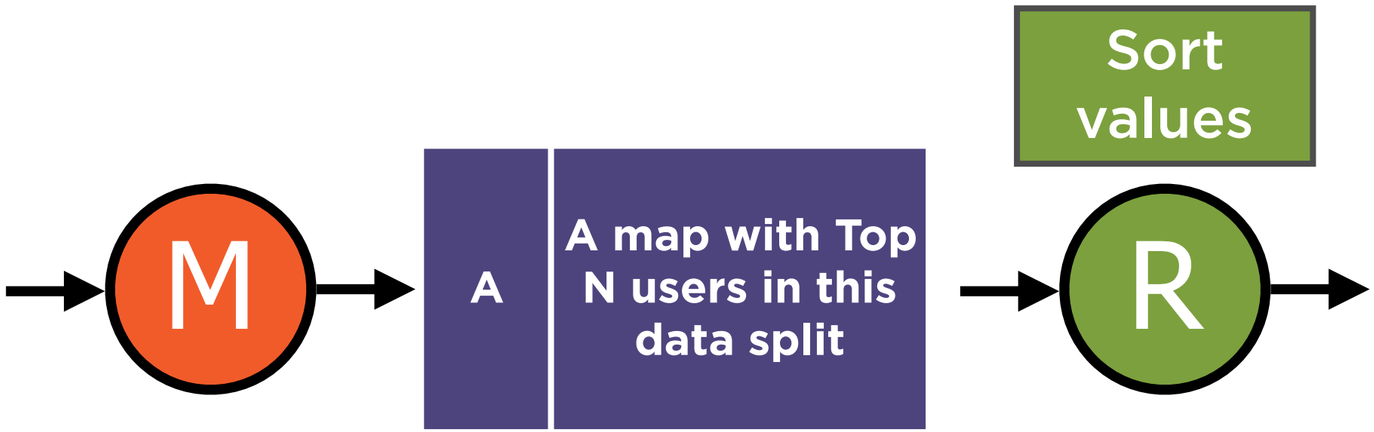


Sort the collection and pick the overall top N records

Top N for the whole dataset

Top N

User	Followers
1	30
2	30000
3	20
4	40
5	50
6	6000



Top 3 Users
2
6
5

A Caution with Top N

Use a single reducer

A Caution with Top N

Multiple reducers will not result in a global sort

The output will be top N records within each reducer

A Caution with Top N

The use of multiple reducers requires custom partitioning logic

Total Order Partitioning

Not covered in this class

Demo

Find the most influential users, in a social network

Summary

Filter datasets based on a condition

Find a distinct set of values within a dataset

Find the top N records in input data