Building Indexes

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

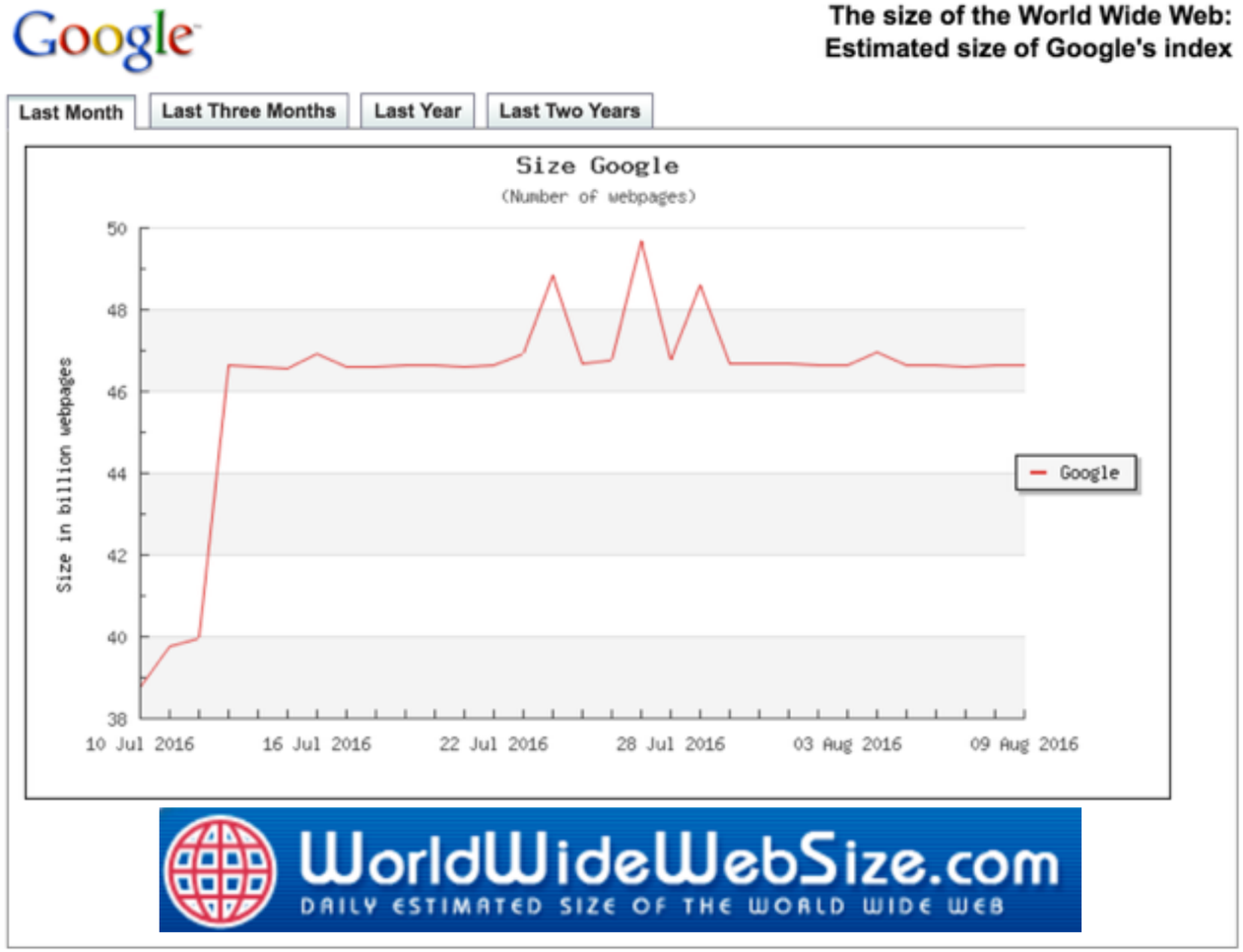
Overview

Understand the role of an inverted index in building search engines

Implement a MapReduce for an inverted index

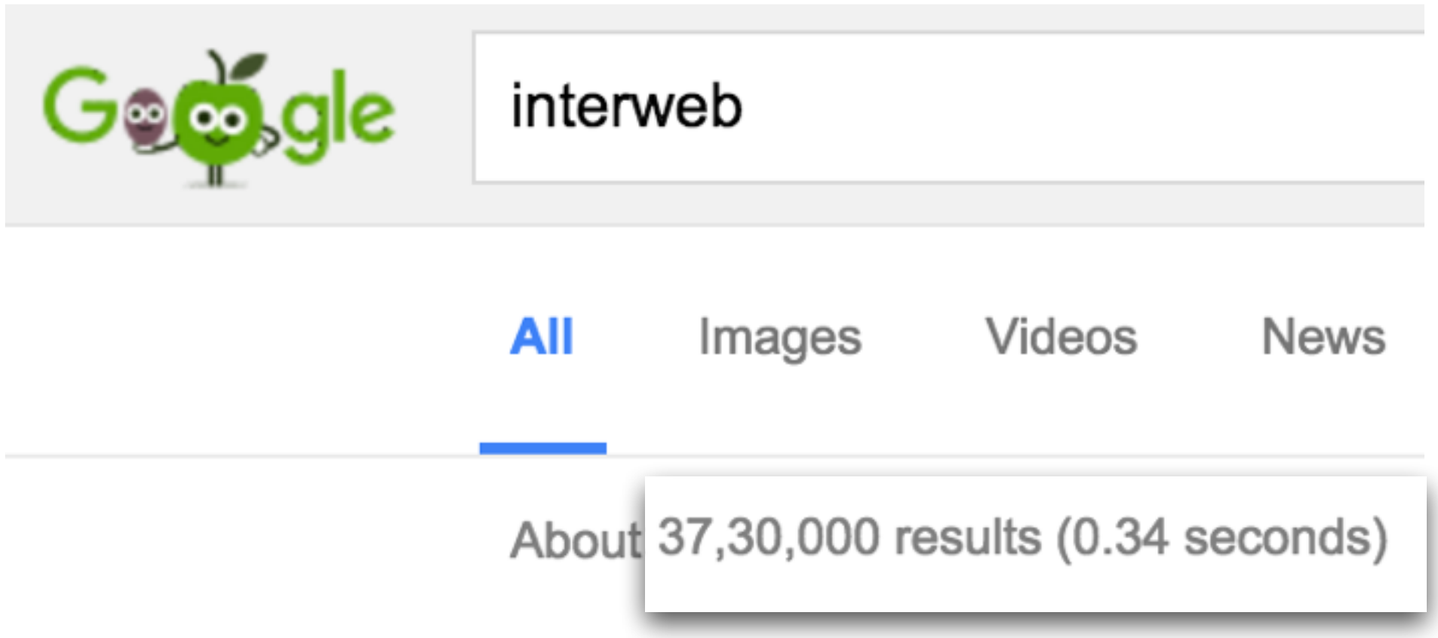
Google Search

The Internet has more than 40 billion webpages



Google Search

Yet, Google can return a search result in less than a second

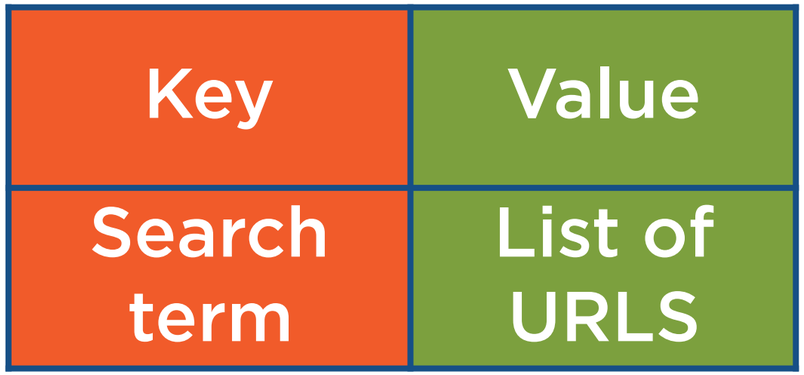


Google Search

Google Search is powered by a massive index

Building an Index

An Index is a Map with a really fast way to look up values



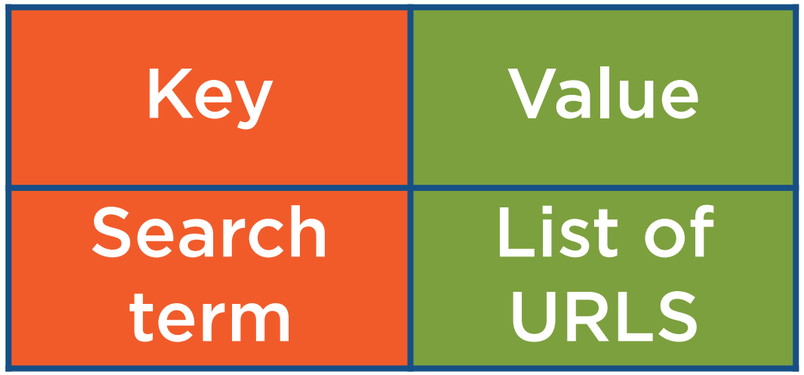
Building an Index

Given a search term, find all pages where that term occurs

Key	Value
Search term	List of URLS

Building an Index

Search is just a lookup into an index of all the webpages in the world



Building an Index

Hugely data intensive

Process the complete set of webpages known to the search engine

Repetitive

Rebuild the index periodically

- Hourly, daily
- Scan the entire data set each time

Building an Index

Hugely data intensive

Repetitive

A great use case for MapReduce

The index used in search engines is an inverted index

Inverted Index

An inverted index is like the index found at the back of a dense textbook

Inverted Index

Let’s first define a book in big data terms

Books and Indexes

A book = an index

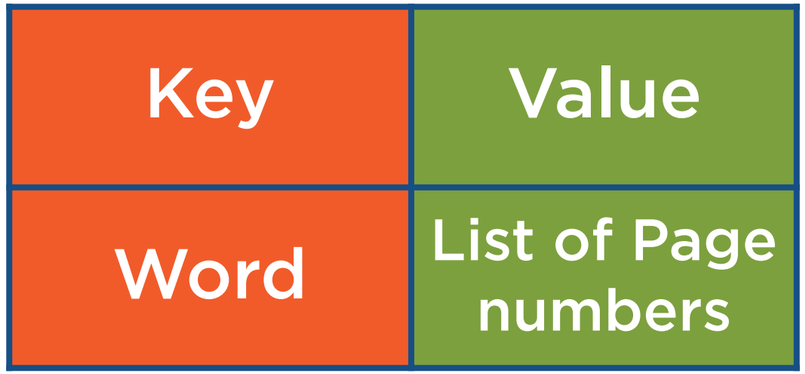
Books and Indexes

Key	Value
Page Number	Words on the Page

An index where words can be looked up by page numbers

Books and Indexes

Most books have an index at the back of the book



Word Index

Books and Indexes

Book

Key	Value
Page Number	Words on the Page

Word Index

Key	Value
Word	List of Page numbers

Books and Indexes

Book

Key	Value
Page Number	Words on the Page

Word Index

Key	Value
Word	List of Page numbers

Books and Indexes

Book

Key	Value
Page Number	Words on the Page

Word Index

Key	Value
Word	List of Page numbers

Books and Indexes

Book

Key	Value
Page Number	Words on the Page

Word Index

Key	Value
Word	List of Page numbers

The word index is an inverted index of the book

Inverted Index

Key	Value
Word	List of Page numbers

The word index helps us find things in the book

Inverted Index

The word index helps us find things in the book

But what we're really interested in is the internet

The World Wide Web

The internet = an index

The World Wide Web

Key	Value
Page URL	Words on the Page

An index where web page contents can be looked up by URLs

The World Wide Web

Key	Value
Page URL	Words on the Page

Invert this index

Inverted Index

www

Key	Value
Page URL	Words on the Page

Key	Value
Word	List of Page URLS

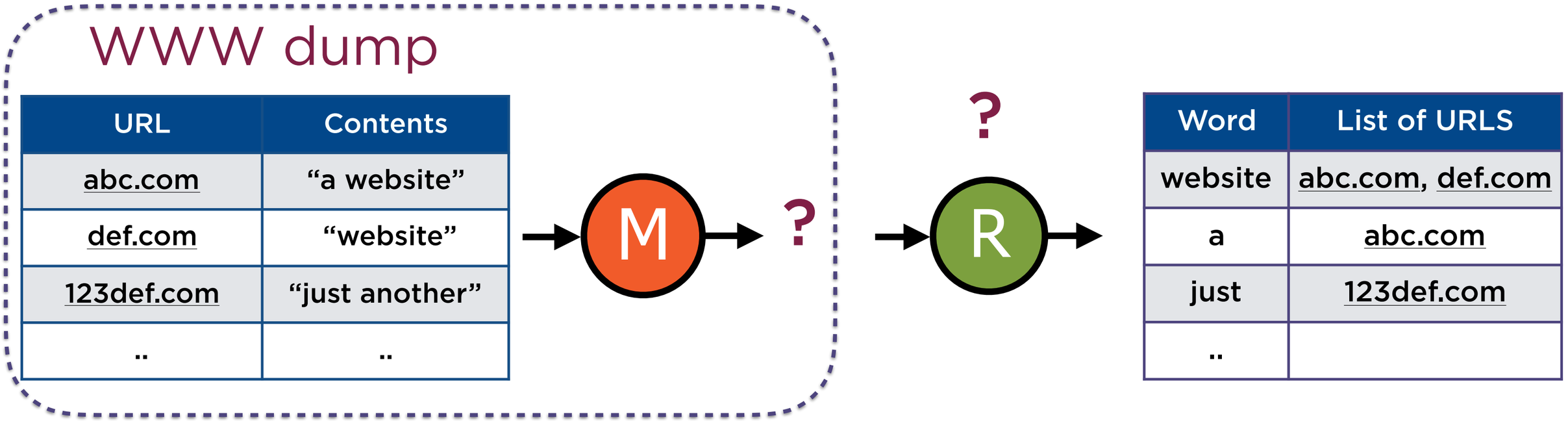
Inverted Index

Key	Value
Word	List of Page URLS

This index helps us find things on the internet

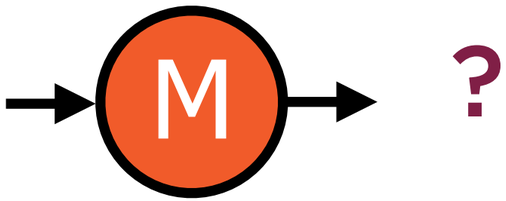
This is exactly what Search Engines use!

MapReduce Inverted Index



Map Step

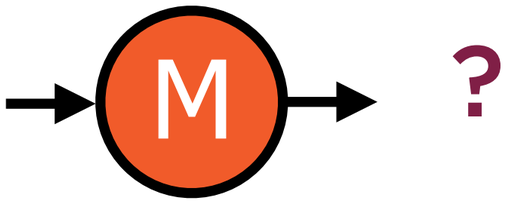
URL	Contents
abc.com	“a website”
def.com	“website”
123def.com	“just another”
..	..



The key has to be a word

Map Step

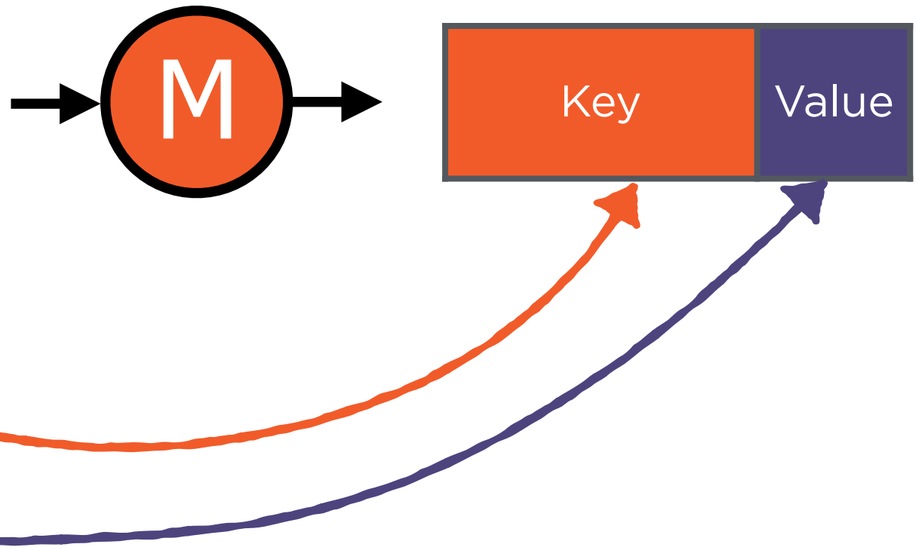
URL	Contents
abc.com	“a website”
def.com	“website”
123def.com	“just another”
..	..



The value has to contain URLs

Map Step

URL	Contents
abc.com	“a website”
def.com	“website”
123def.com	“just another”
..	..

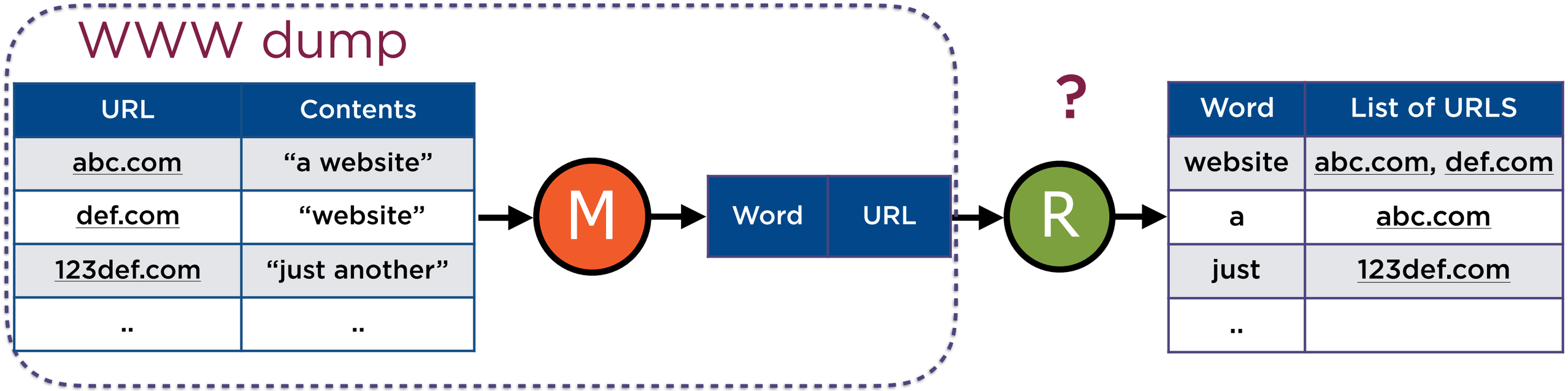


Map Step

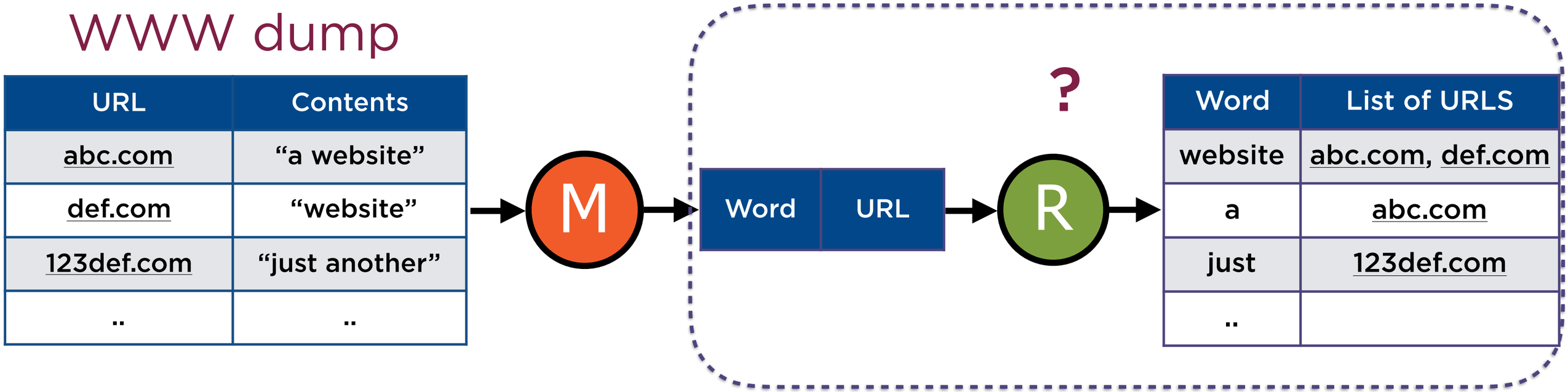
URL	Contents
abc.com	“a website”
def.com	“website”
123def.com	“just another”
..	..

Key	Value
Word	URL
a	abc.com
website	abc.com
website	def.com
just	123def.com

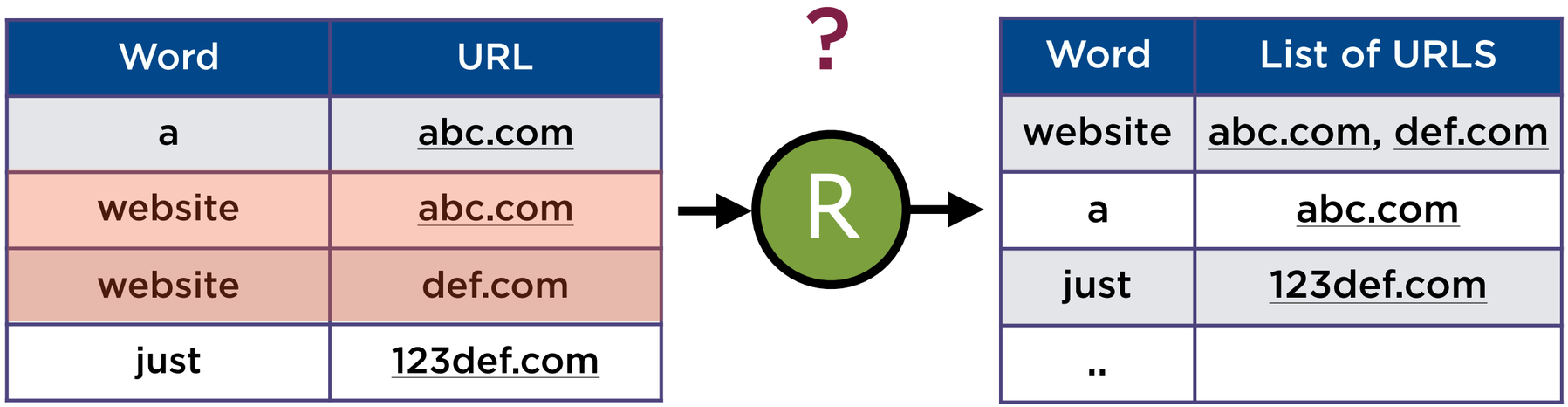
MapReduce Inverted Index



MapReduce Inverted Index



Reduce Step



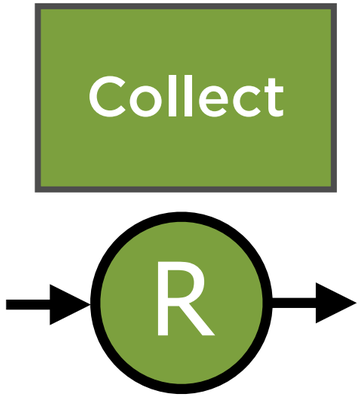
Word	URL
a	abc.com
website	abc.com
website	def.com
just	123def.com

Word	List of URLS
website	abc.com, def.com
a	abc.com
just	123def.com
..	

The reduce step combines values with the same key

Reduce Step

Word	URL
a	abc.com
website	abc.com
website	def.com
just	123def.com



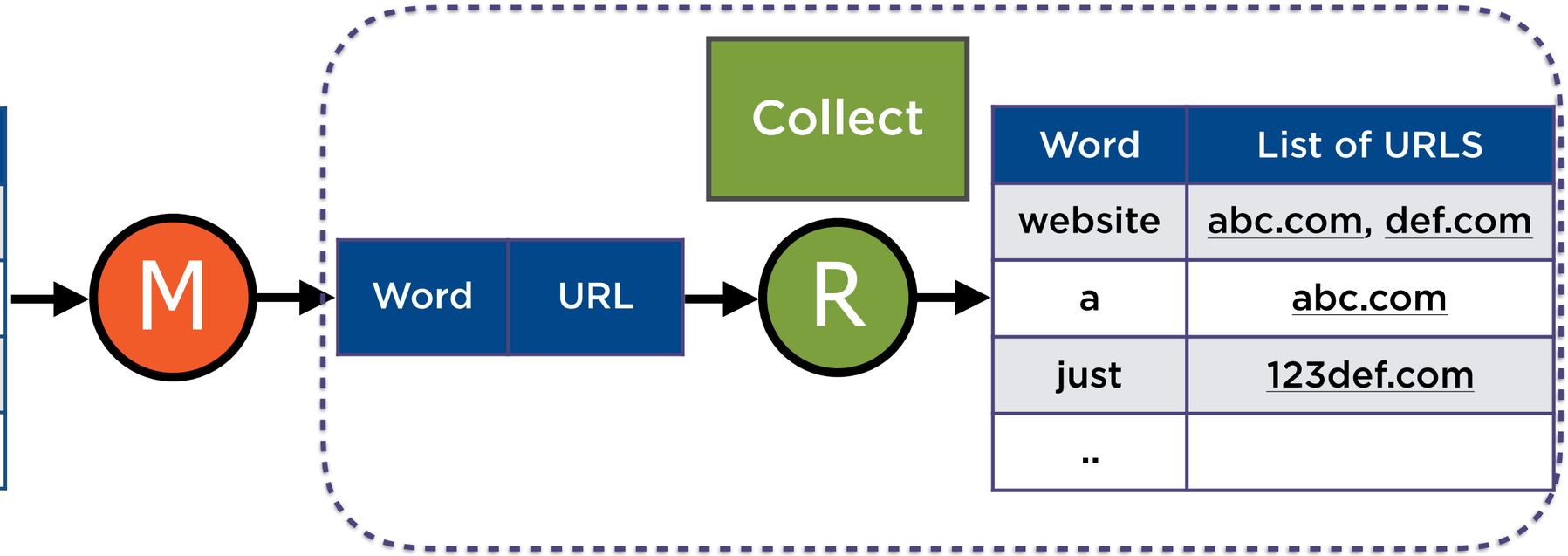
Word	List of URLS
website	abc.com, def.com
a	abc.com
just	123def.com
..	

The combining logic is simply to collect the values into a list

MapReduce Inverted Index

WWW dump

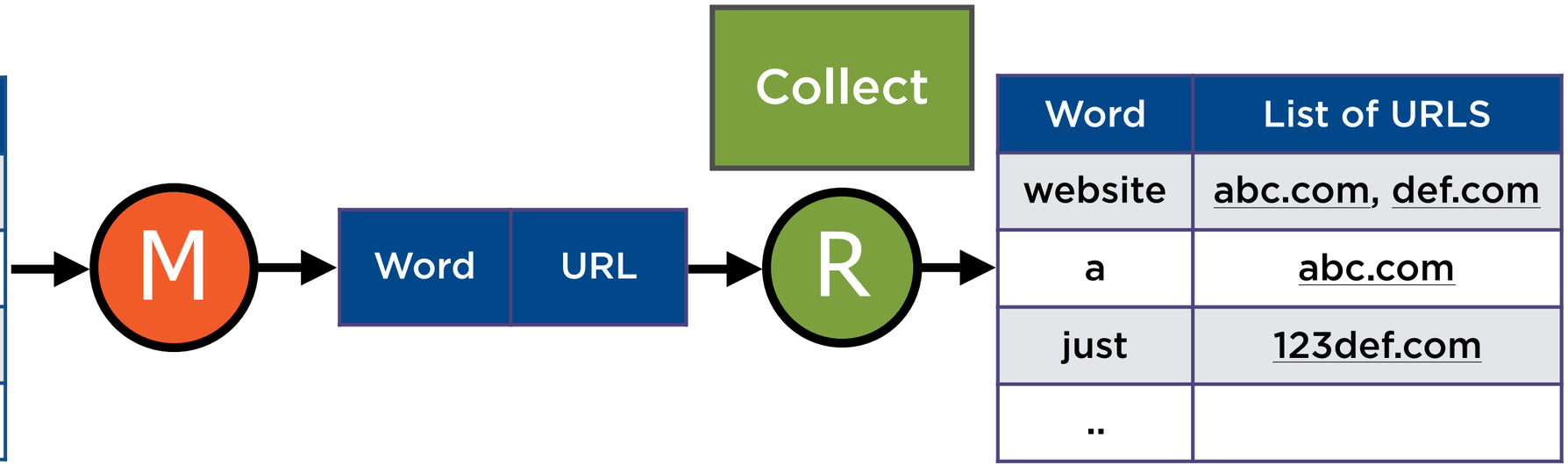
URL	Contents
abc.com	“a website”
def.com	“website”
123def.com	“just another”
..	..



MapReduce Inverted Index

WWW dump

URL	Contents
abc.com	“a website”
def.com	“website”
123def.com	“just another”
..	..



Word	List of URLS
website	abc.com, def.com
a	abc.com
just	123def.com
..	

Demo

Implementing an Inverted Index

Summary

Understood what is an inverted index and its role in building search engines

Implemented a MapReduce for an inverted index