Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

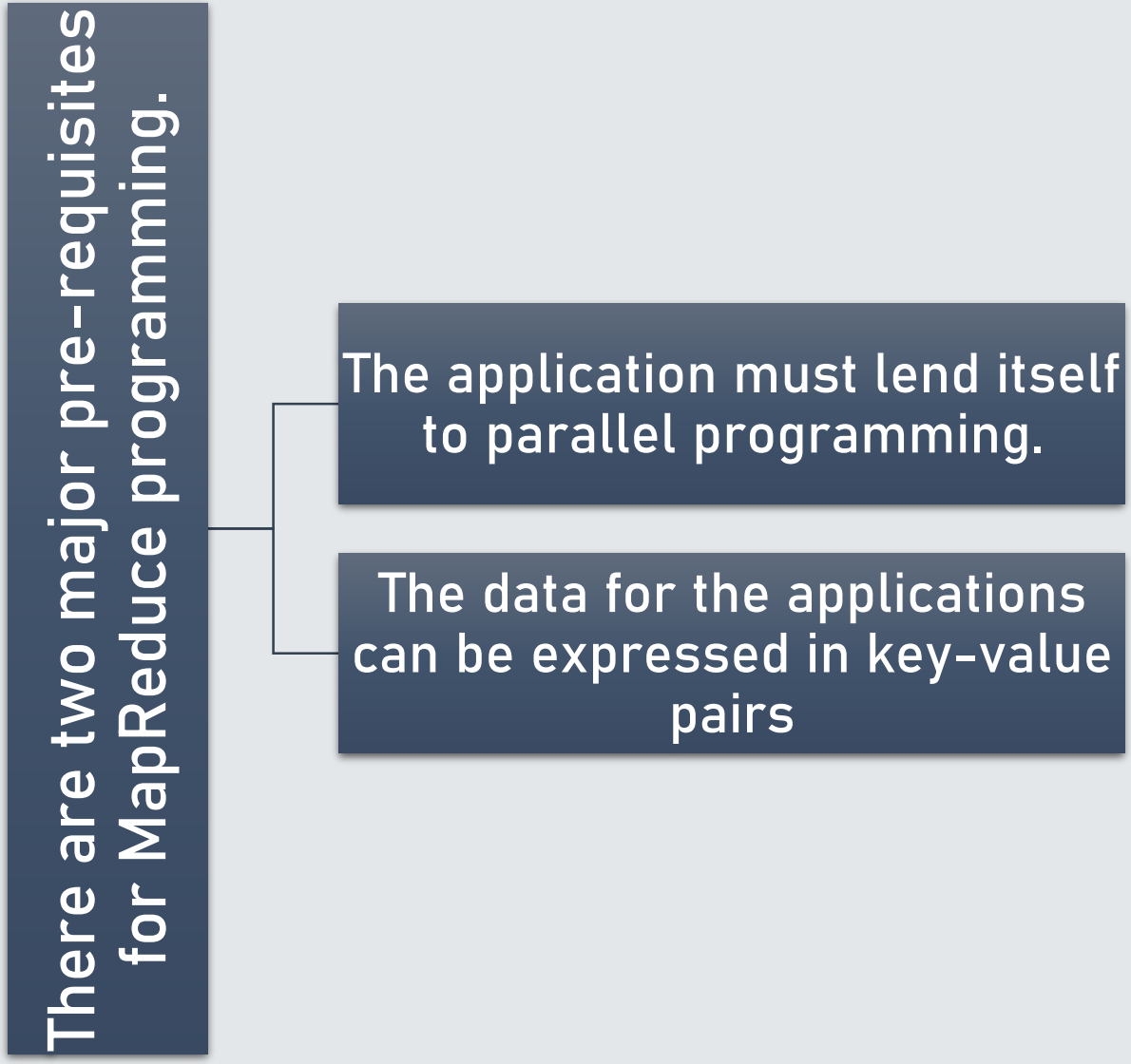
Learning Outcomes

After this lecture, you will be able to

- learn Hadoop-MapReduce
- learn Hadoop – Streaming

Introduction

Introduction



Introduction

MapReduce processing is similar to UNIX sequence (also called pipe) structure

e.g. the UNIX command:

grep | sort | count myfile.txt

will produce a wordcount in the text document called file1.txt.

Example

For example: Suppose file1.txt contains the following text:

file1: We are going to a picnic near our house. Many of our friends are coming.

You are welcome to join us. We will have fun.

Example

The outputs of Grep, Sort and Wordcount will be as follows:

Example

The outputs of Grep, Sort and Wordcount will be as follows:

Example

The outputs of Grep, Sort and Wordcount will be as follows:

Word Count	
A	1
Are	3
Coming	1
Friends	1
Fun	1
Going	1
Have	1
House	1
Join	1
Many	1
Near	1
Of	1

WordCount	
Our	2
Picnic	1
To	2
Us	1
We	2
Welcome	1
Will	1
you	1

MAPREDUCE

- If the file is very large, then _____
- _____ can help here.
- Speeds up the computation.

Example: Thus if a file can be broken down into 100 small chunks.

The total time taken = 1/100

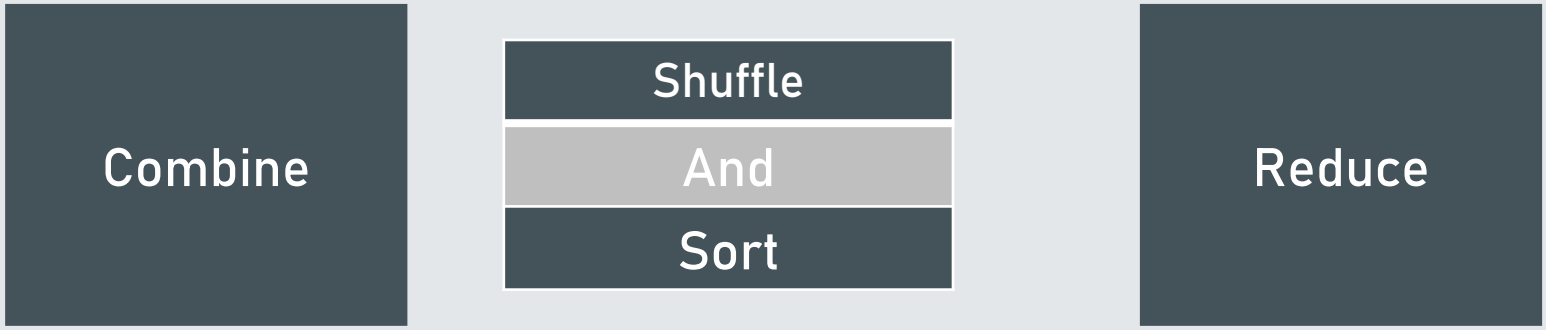
Results of the computation on small chunks are residing in a 100 different places.

MAPREDUCE

Combine

MAPREDUCE

MAPREDUCE



MAPREDUCE

Input

Dear bear river
Car Car river
Dear Car Bear

References: https://www.edupristine.com/blog/hadoop-mapreduce-framework

MAPREDUCE

Input

Dear bear river
Car Car river
Dear Car Bear

Splitting

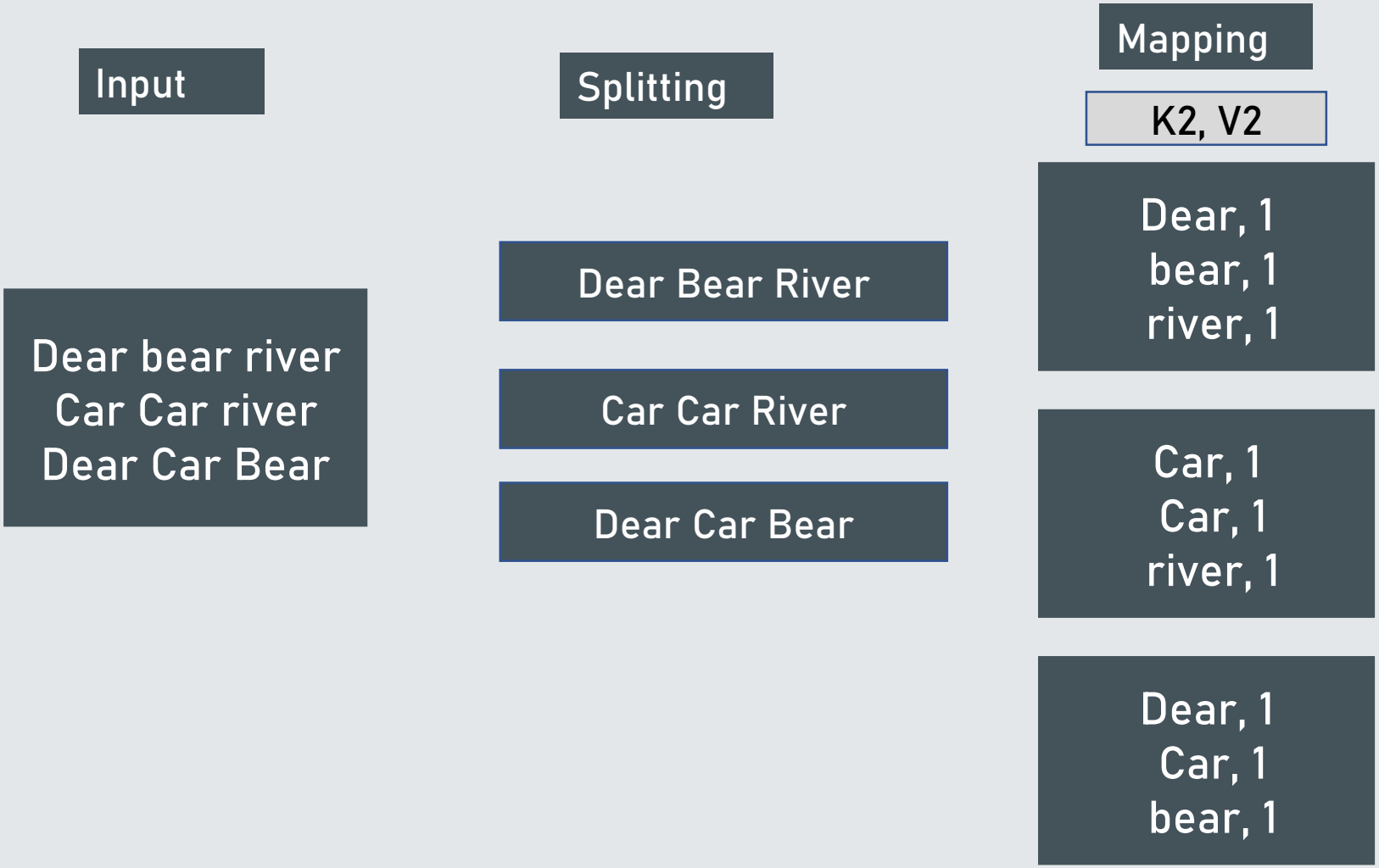
K1, V1

Dear Bear River

Car Car River

Dear Car Bear

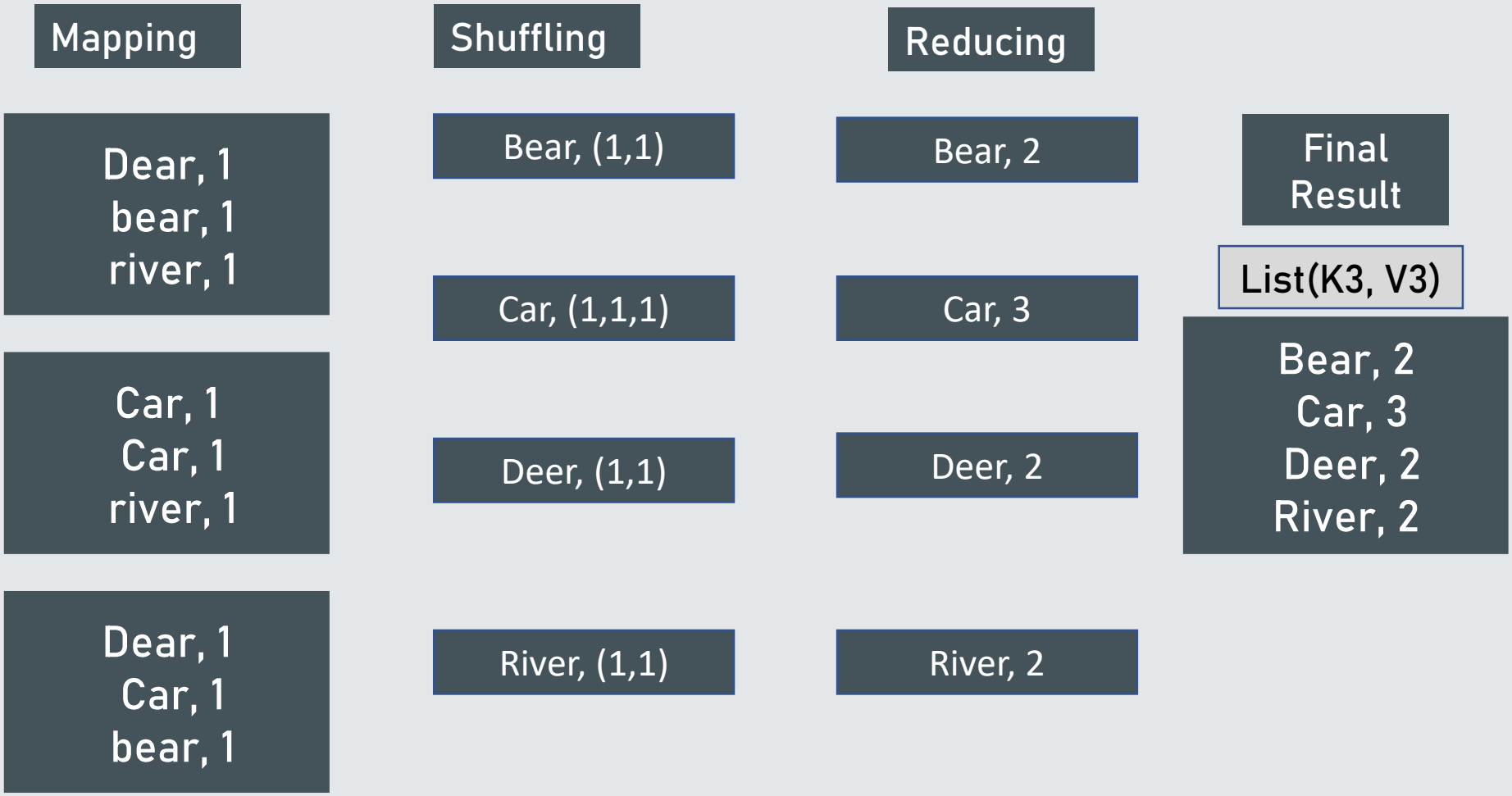
MAPREDUCE


Input
Splitting
Mapping
K2, V2
Dear bear river
Car Car river
Dear Car Bear
Dear Bear River
Car Car River
Dear Car Bear
Dear, 1
bear, 1
river, 1
Car, 1
Car, 1
river, 1
Dear, 1
Car, 1
bear, 1

MAPREDUCE


Mapping
Shuffling
K2, List(V2)
Reducing
Dear, 1
bear, 1
river, 1
Car, 1
Car, 1
river, 1
Dear, 1
Car, 1
bear, 1
Bear, (1,1)
Car, (1,1,1)
Deer, (1,1)
River, (1,1)
Bear, 2
Car, 3
Deer, 2
River, 2

MAPREDUCE


Mapping
Shuffling
Reducing
Dear, 1
bear, 1
river, 1
Car, 1
Car, 1
river, 1
Dear, 1
Car, 1
bear, 1
Bear, (1,1)
Car, (1,1,1)
Deer, (1,1)
River, (1,1)
Bear, 2
Car, 3
Deer, 2
River, 2
Final Result
List(K3, V3)
Bear, 2
Car, 3
Deer, 2
River, 2

MAP Stage

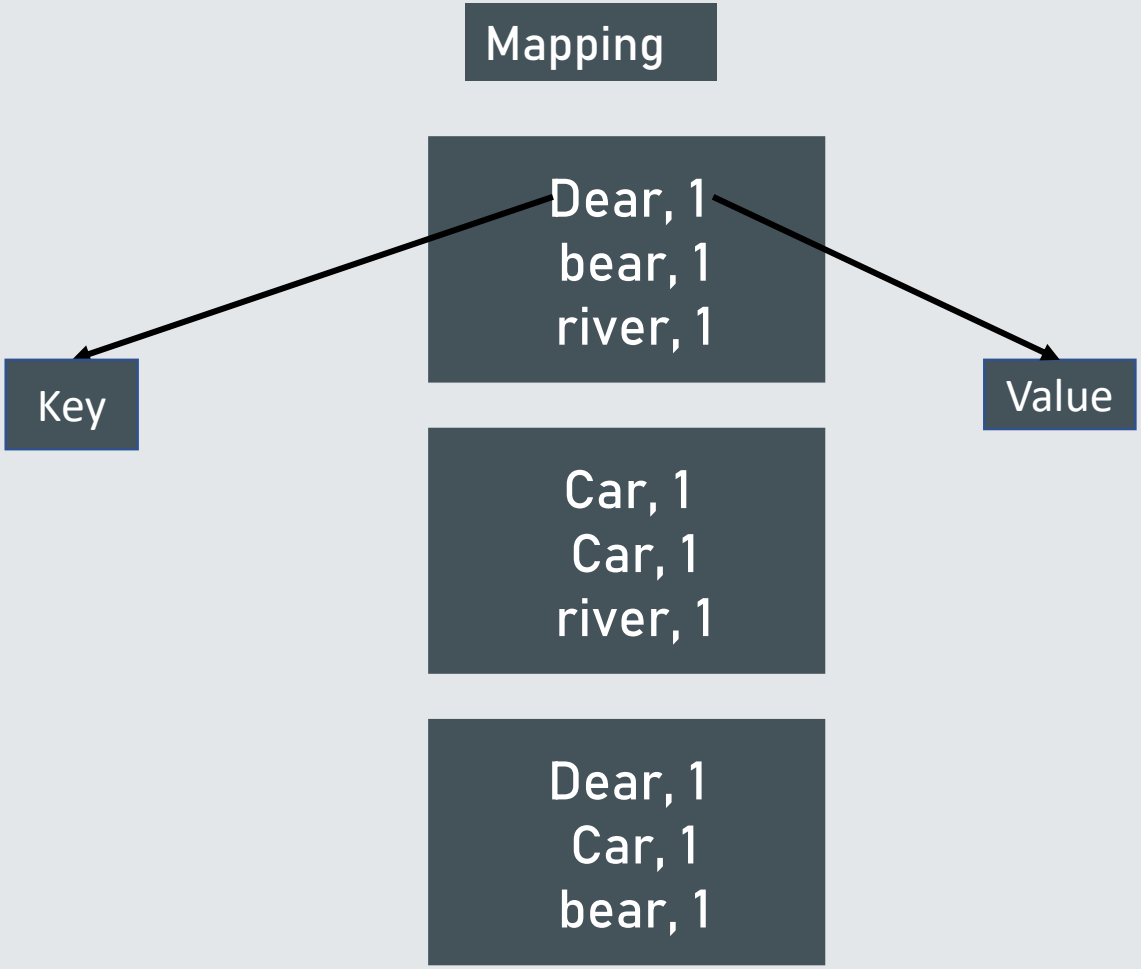
Mapping

Dear, 1
bear, 1
river, 1

Car, 1
Car, 1
river, 1

Dear, 1
Car, 1
bear, 1

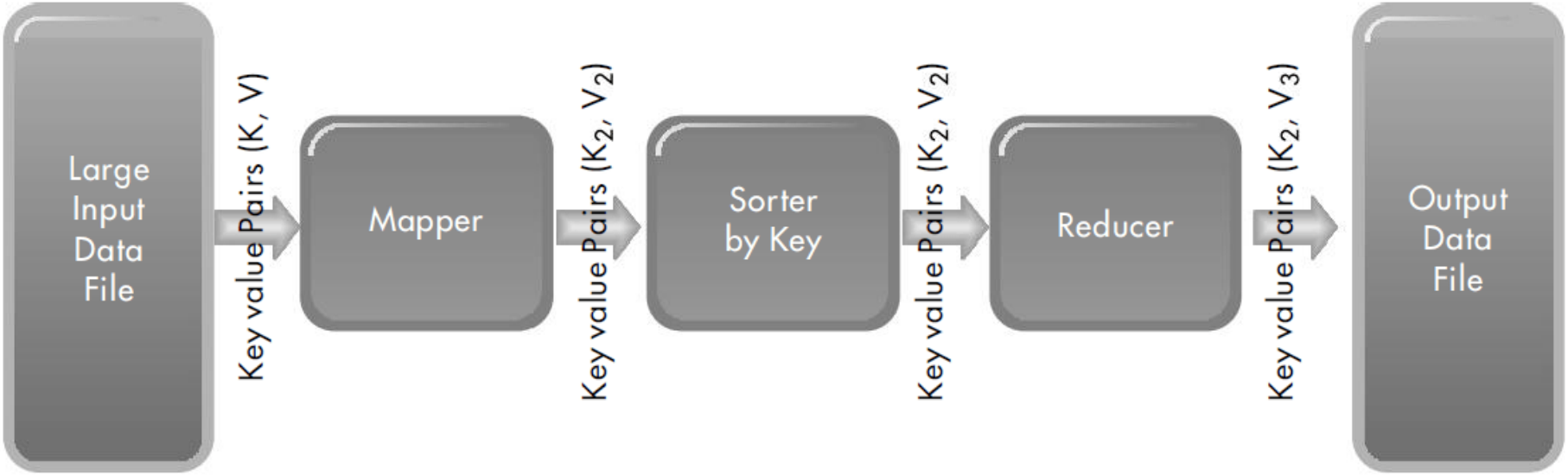
MAP Stage


Mapping
Dear, 1
bear, 1
river, 1
Key
Value
Car, 1
Car, 1
river, 1
Dear, 1
Car, 1
bear, 1

Reduce Stage

The Reducer will take all the key-value pairs from the mapper and check the association of all keys with value. All the values associated with a single key will be taken and it will provide an output of any number of key-value pairs.

MapReduce Architecture



MAP-REDUCE STAGE

- MapReduce is a sequential computation.
- Mapper must have completed the execution
- Reducer will have access to all the values
- Reducers are working on different keys
- Work simultaneously and parallelism is achieved.

MAPREDUCE

How to manage variety of data structures in the file system?

- Data is stored as one key and one non-key attribute value.
- Thus the data is represented as a key-value pair.

MAPREDUCE

How to manage variety of data structures in the file system?

- The intermediate results, and the results all will also be in key-pair format.
- Thus a key requirement for the use of MapReduce parallel processing system is that the input data and output data must both be represented in key-values formats

Why MapReduce for Big DATA?

Hadoop-Streaming

- Uses standard Unix streams
- It is an ideal application for text processing.
- Map input data is passed over standard input to your map function
- Key-value pair is written as a single tab-delimited line.

Hadoop-Streaming

- A tab-separated key-value pair—passed over standard input.
- The Reduce function reads lines from standard input


That's all for now...