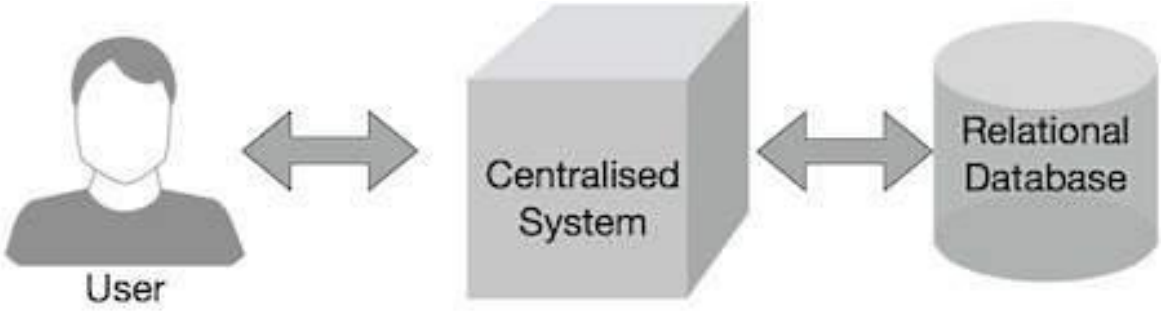
MAP-REDUCE

Map-Reduce

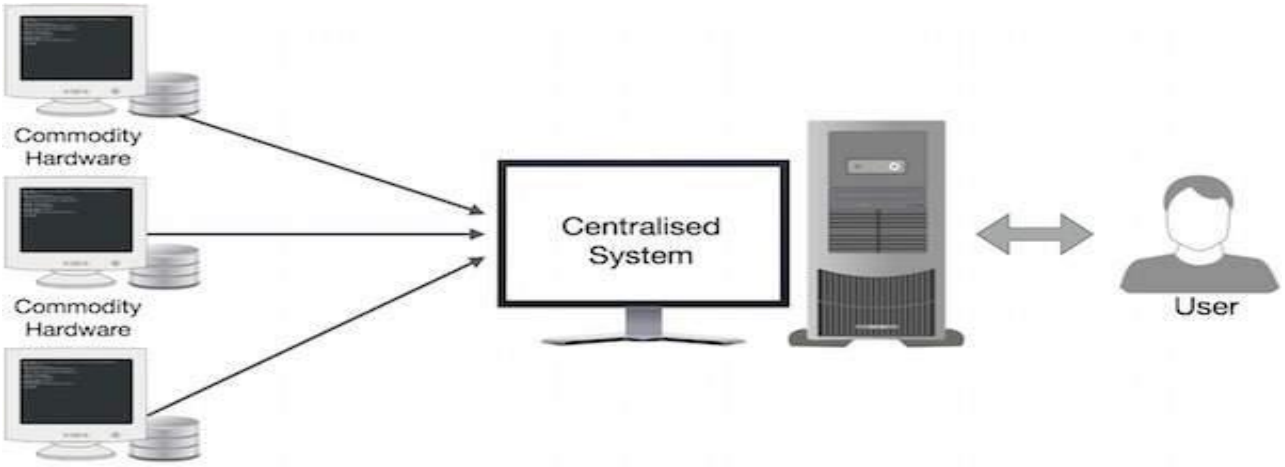
- MapReduce is the system used to process data in the Hadoopcluster
- MapReduce is a method for distributing a task across multiple nodes
- It is a programming paradigm that runs in the background of Hadoop to provide scalability and easy data-processing solutions
- Consists of two phases:
  - Map
  - Reduce

Why MapReduce?

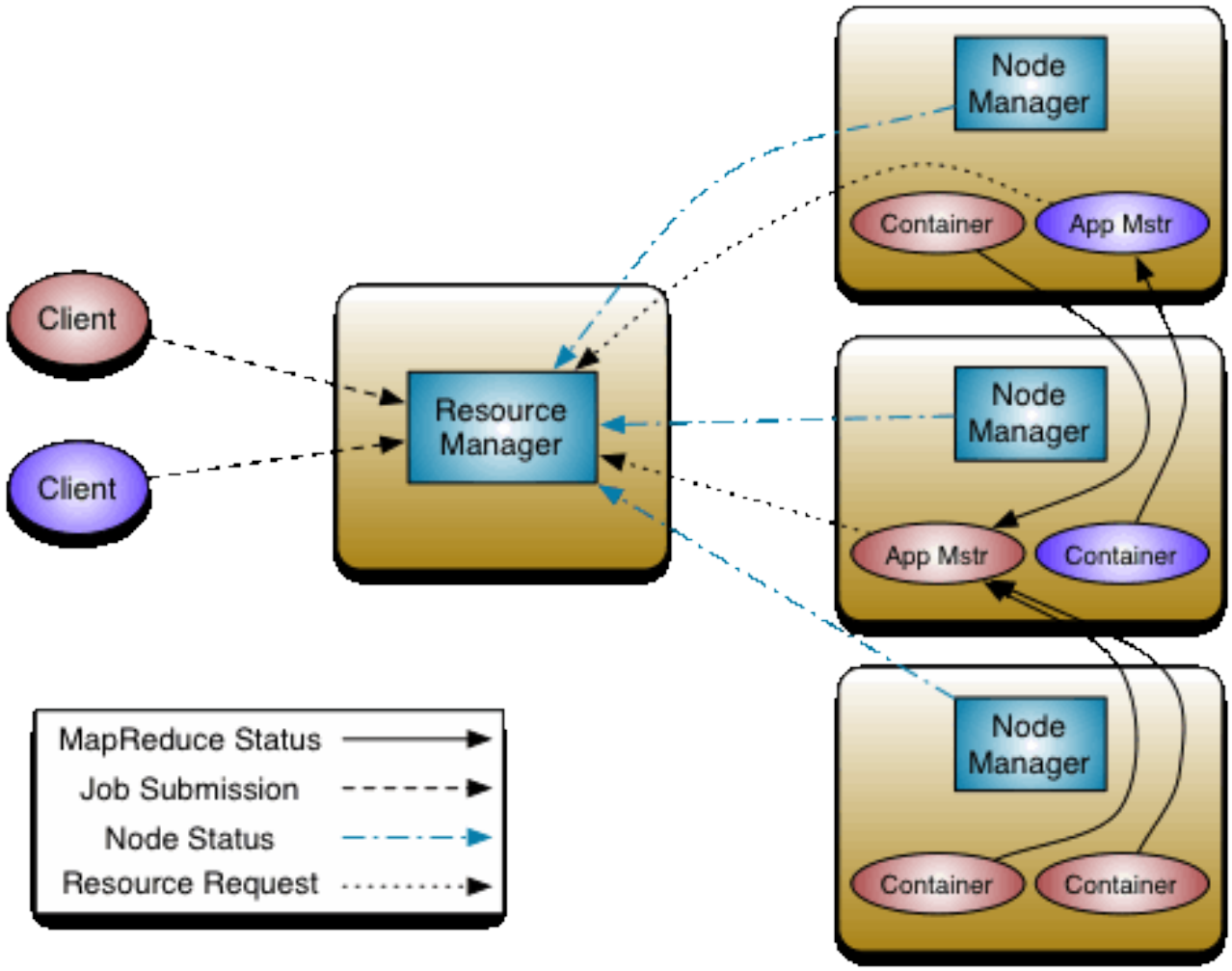
Traditional Enterprise System



Map-Reduce System



YARN – Yet Another Resource Negotiator



How Map-Reduce Works

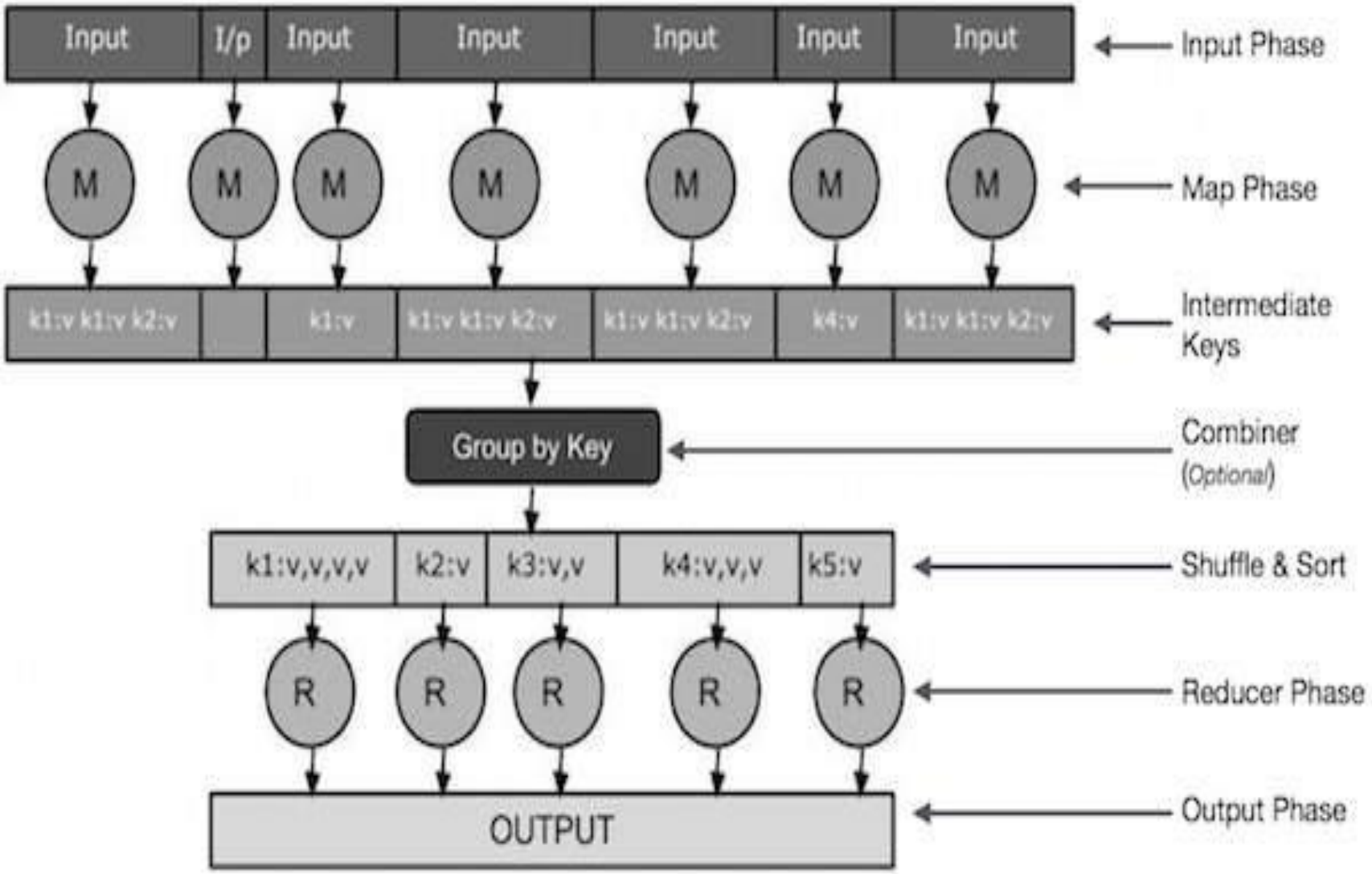
❖The MapReduce algorithm contains two important tasks, namely Map and Reduce.

❖The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).

❖The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.

❖The reduce task is always performed after the map job

Continued…


Input
I/p
Input
Input
Input
Input
Input
Input Phase
M
Map Phase
k1:v k1:v k2:v
k1:v
k1:v k1:v k2:v
k1:v k1:v k2:v
k4:v
k1:v k1:v k2:v
Intermediate Keys
Group by Key
Combiner
(Optional)
k1:v,v,v,v
k2:v
k3:v,v
k4:v,v,v
k5:v
Shuffle & Sort
R
Reducer Phase
OUTPUT
Output Phase

Input Phase

A Record Reader that translates each record in an input file and sends the parsed data to the mapper in the form of key-value pairs.

Map

Map is a user-defined function, which takes a series of key-value pairs and processes each one of them to generate zero or more key-value pairs.

Intermediate Keys

They key-value pairs generated by the mapper are known as intermediate keys.

Combiner

A combiner is a type of local Reducer that groups similar data from the map phase into identifiable sets

Combiner

- ❖ It takes the intermediate keys from the mapper as input and applies a user-defined code to aggregate the values in a small scope of one mapper
- ❖ It is not a part of the main MapReduce algorithm; it is optional

Shuffle and Sort

- ❖ The Reducer task starts with the Shuffle and Sort step.
- ❖ It downloads the grouped key-value pairs onto the local machine, where the Reducer is running.
- ❖ The individual key-value pairs are sorted by key into a larger data list.
- ❖ The data list groups the equivalent keys together so that their values can be iterated easily in the Reducer task.

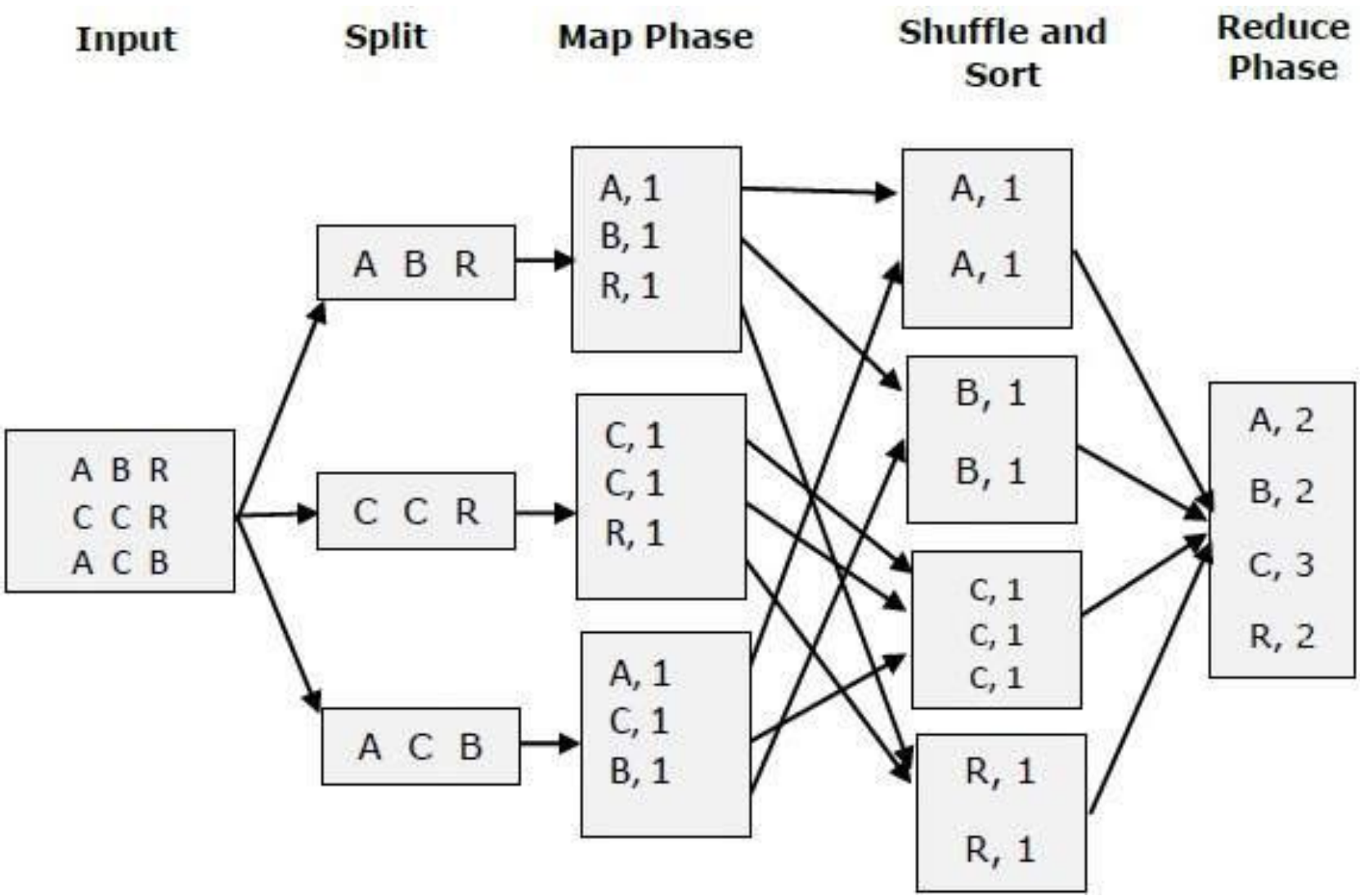
Reducer

- The Reducer takes the grouped key-value paired data as input and runs a Reducer function on each one of them.
- Here, the data can be aggregated, filtered, and combined in a number of ways, and it requires a wide range of processing.
- Once the execution is over, it gives zero or more key-value pairs to the final step.

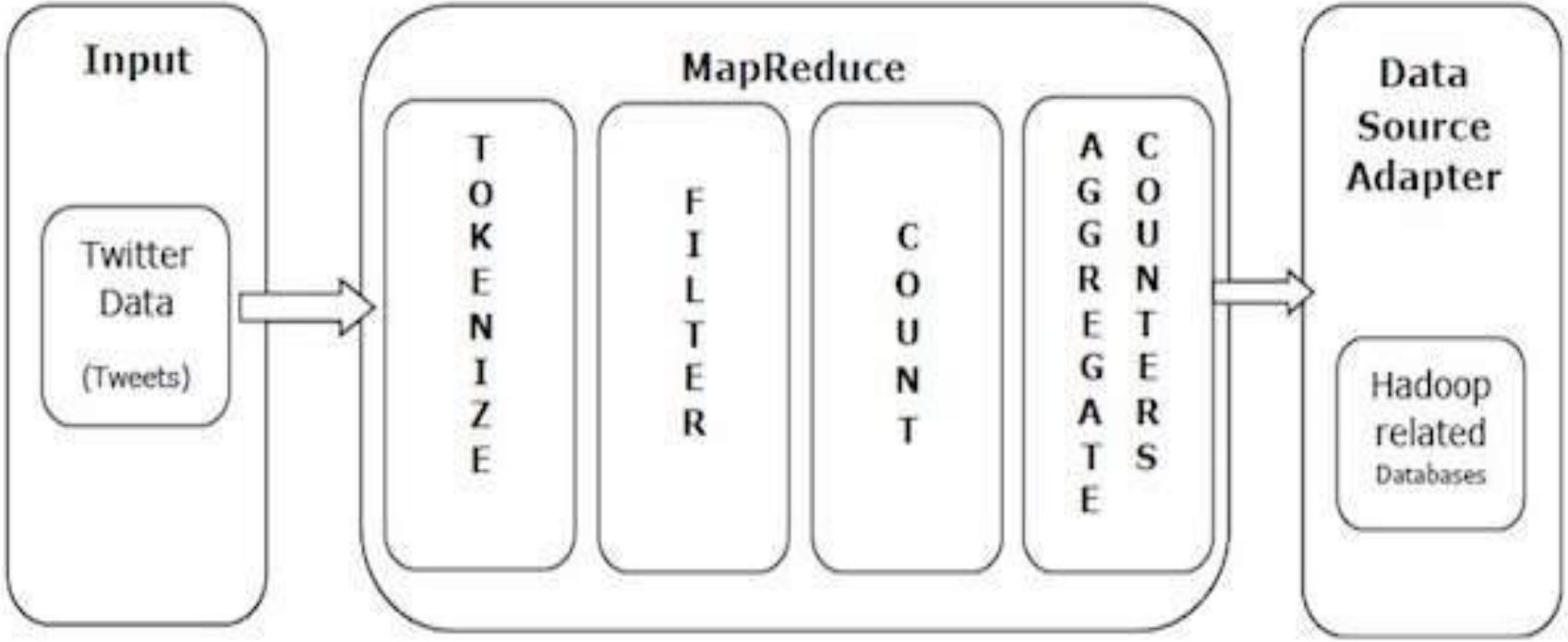
Output Phase

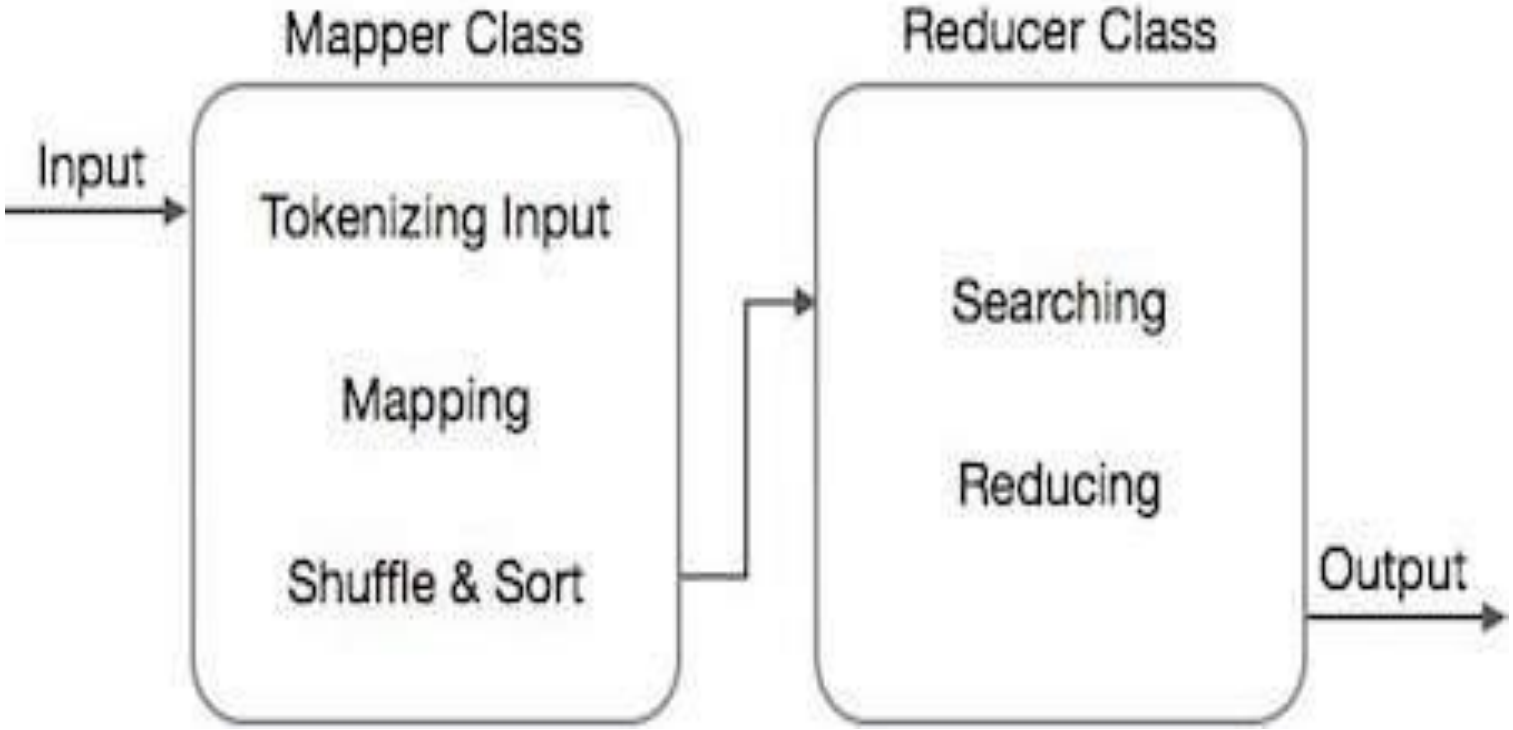
- In the output phase, we have an output formatter that translates the final key-value pairs from the Reducer function and writes them onto a file using a record writer

Map – Reduce Problem



Twitter receives around 500 million tweets per day, which is nearly 3000 tweets per second



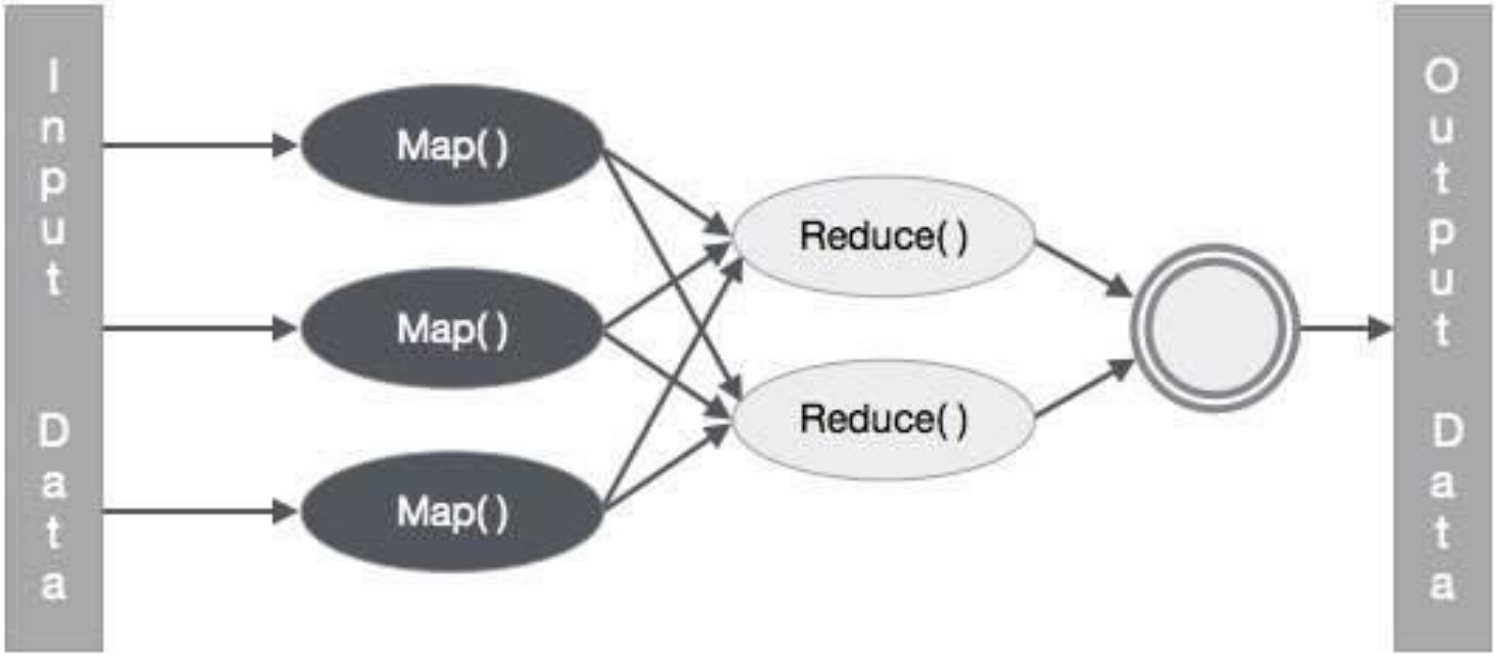

Mapper Class
Reducer Class
Input
Tokenizing Input
Mapping
Shuffle & Sort
Searching
Reducing
Output

Classes and Methods

Map Reduce Programming

- JobContext Interface
  - It defines different jobs in MapReduce
  - Sub interface
    - MapContext<KEYIN,VALIN,KEYOUT,VALOUT>
    - ReduceContext<KEYIN,VALIN,KEYOUT,VALOUT>
- Job Class
  - It allows the user to configure the job, submit it, control its execution, and query the state
  - Methods
    - getJobName()
    - getJobState()
    - isComplete()
    - setMapperClass(Class)
- Mapper Class
  - It defines the map job
  - map(KEYIN key, VALUEIN value, org.apache.hadoop.mapreduce.Mapper.Context context)
- Reducer Class
  - It defines the reduce job in map reduce
  - **reduce**(KEYIN key, Iterable<VALUEIN> values, org.apache.hadoop.mapreduce.Reducer.Context context)

Summary – Map Reduce



THANK YOU