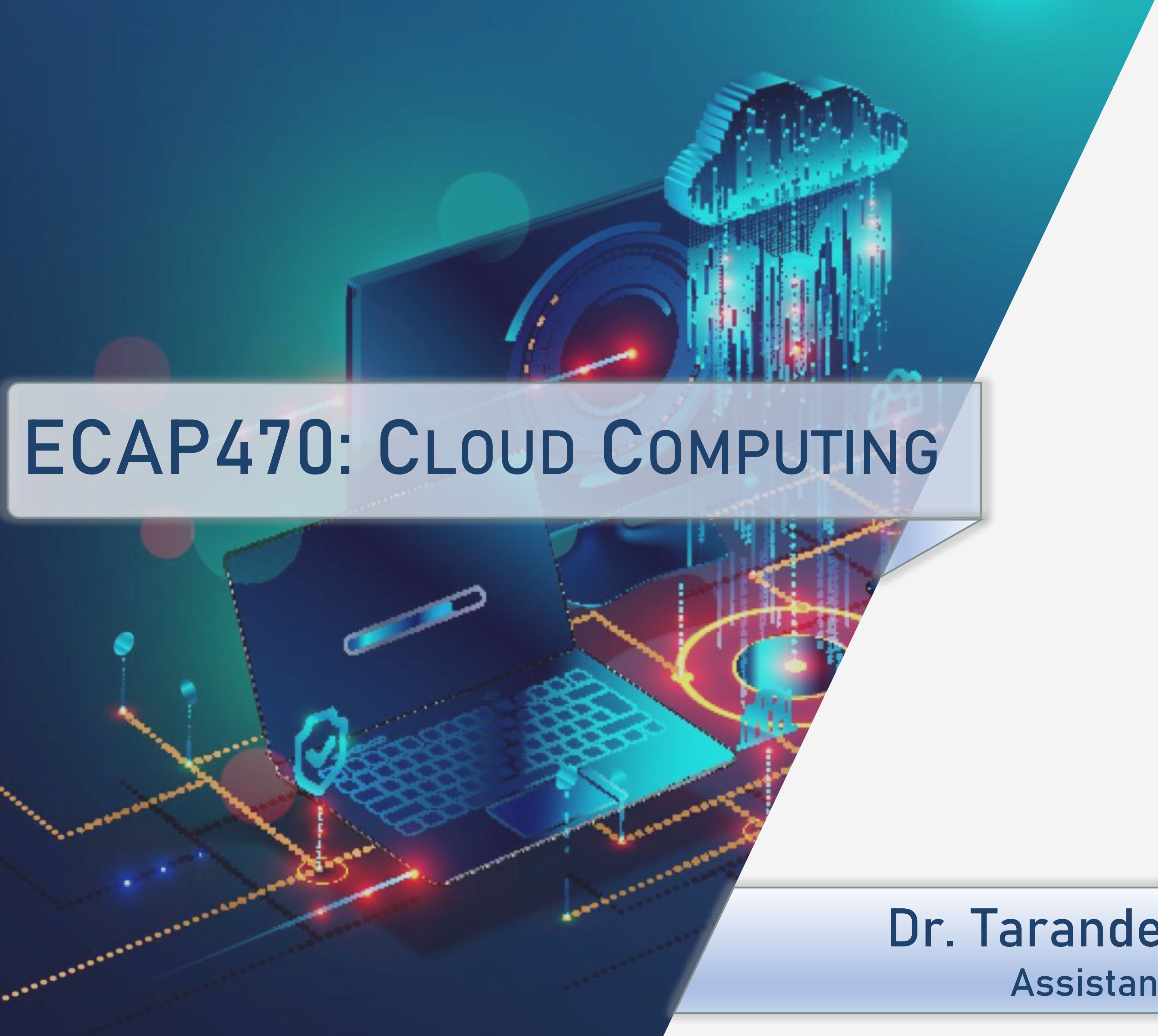


ECAP470: Cloud Computing

Dr. Tarandeep Kaur

Assistant Professor

Learning Outcomes

After this lecture, you will be able to,

✓ Learn about MapReduce and its phases.

✓ Know about the MapReduce Algorithm.

Hadoop MapReduce

- Framework of Hadoop used to process a huge amount of data parallelly on large clusters of commodity hardware in a reliable manner.
- Allows the application to store the data in the distributed form and process large dataset across groups of computers using simple programming models.

Hadoop MapReduce

How does MapReduce in Hadoop make working so easy?

Scaling up the data processing

MapReduce Architecture

- Hadoop cluster stores a large set of data which is parallelly processed mainly by MapReduce.
- Provides parallelism, fault-tolerance, and data distribution.
- Map Reduce provides API with features such as parallel processing of huge amounts of data, batch processing, and high availability.

MapReduce Architecture

- Hadoop cluster stores a large set of data which is parallelly processed mainly by MapReduce.
- Provides parallelism, fault-tolerance, and data distribution.
- Map Reduce provides API with features such as parallel processing of huge amounts of data, batch processing, and high availability.

MapReduce Architecture

- Hadoop cluster stores a large set of data which is parallelly processed mainly by MapReduce.
- Provides parallelism, fault-tolerance, and data distribution.
- Map Reduce provides API with features such as parallel processing of huge amounts of data, batch processing, and high availability.

MapReduce Architecture

- Map Reduce programs are written by programmers when there is a need for an application for business scenarios.
- Development of applications and deployment across Hadoop clusters is done by the programmers when they understand the flow pattern of MapReduce.

MapReduce Architecture

- Map Reduce programs are written by programmers when there is a need for an application for business scenarios.
- Development of applications and deployment across Hadoop clusters is done by the programmers when they understand the flow pattern of MapReduce.

MapReduce Architecture

- Architecture of MapReduce basically has two main processing stages, Map and Reduce.
- Intermediate processes will take place in between the Map and Reduce phases.
- Sort and shuffle are the tasks taken up by Map and Reduce, which are done intermediate.

MapReduce Architecture

- Architecture of MapReduce basically has two main processing stages, Map and Reduce.
- Intermediate processes will take place in between the Map and Reduce phases.
- Sort and shuffle are the tasks taken up by Map and Reduce, which are done intermediate.

MapReduce Architecture

- Architecture of MapReduce basically has two main processing stages, Map and Reduce.
- Intermediate processes will take place in between the Map and Reduce phases.
- Sort and shuffle are the tasks taken up by Map and Reduce, which are done intermediate.

MapReduce Architecture

- Map() Function
- Reduce() Function

How MapReduce in Hadoop Works?

MapReduce program executes mainly in 4 steps:

- ○ Input splits
- ○ Map
- ○ Shuffle
- ○ Reduce

MapReduce- Map Step

Map Step- Combination of the input splits step and the Map step.

- In the Map step, the source file is passed as line by line. Before input pass to the Map function job, the input is divided into the small fixed-size called Input splits.

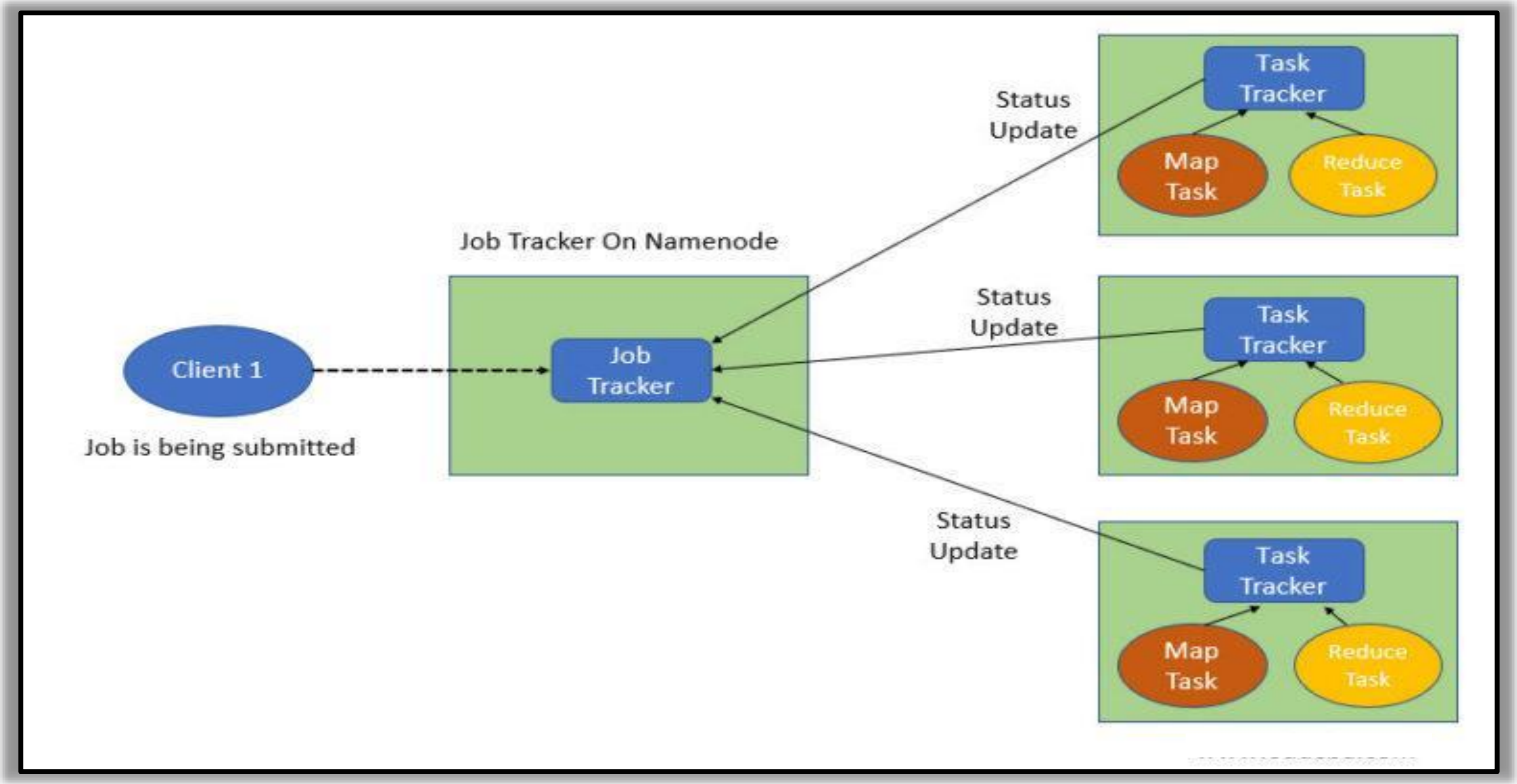
MapReduce- Reduce Step

Reduce Step

- This step is the combination of the Shuffle step and the Reduce.

MapReduce Architecture Components

Working of MapReduce Organizers-

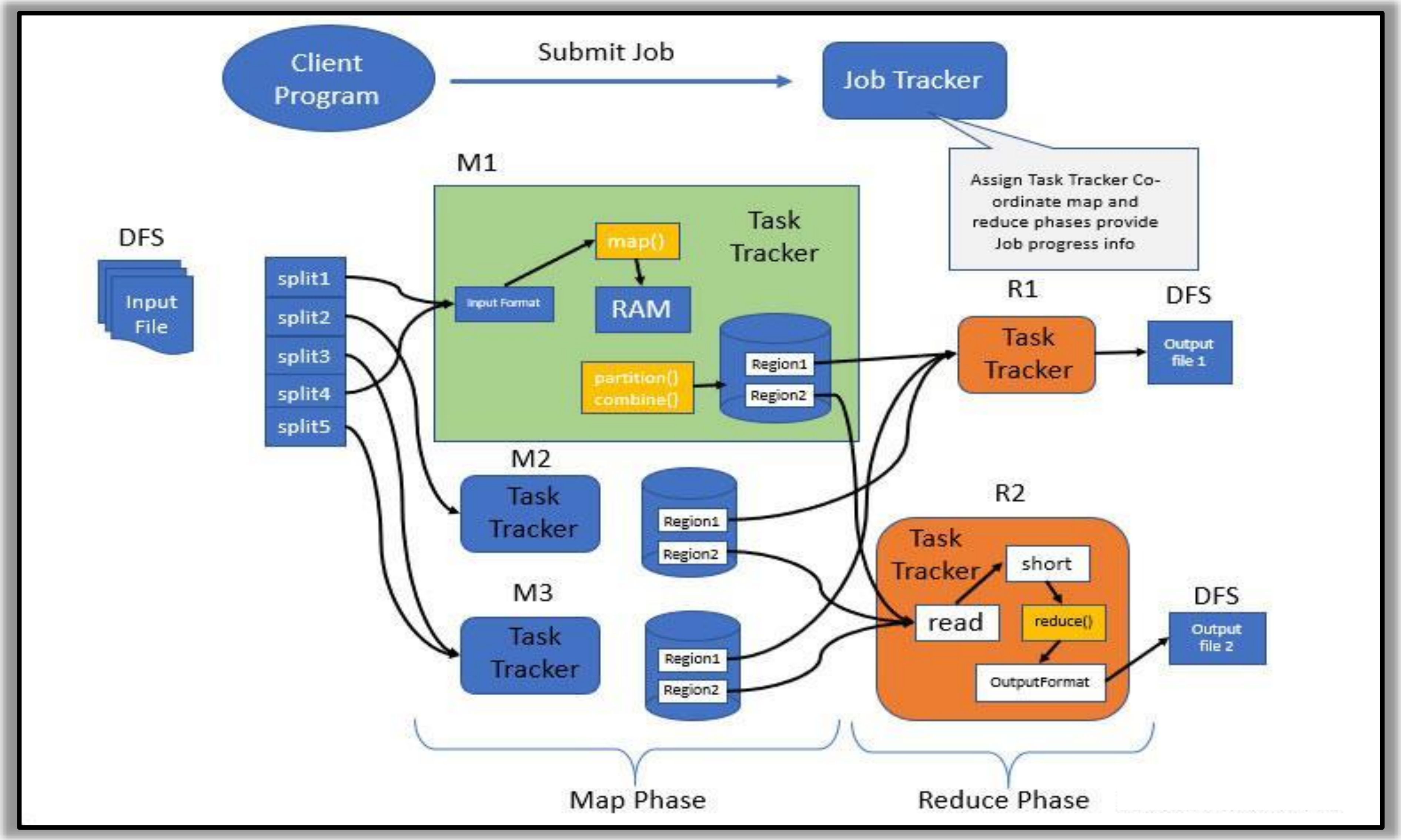


MapReduce Architecture Components

This is how MapReduce organizers work-

- The job is divided into two components: Map tasks (Splits and mapping) and Reduce tasks (Reducing and shuffling).

MapReduce Architecture Components



MapReduce Architecture Components

Map Phase-

- Map phase splits input data into two parts- Keys & Values. Mini reducer which is commonly called a combiner, reducer code places input as combiner.

MapReduce Architecture Components

Processing in Intermediate-

- In the intermediate phase, the map input gets into the sort and shuffle phase.

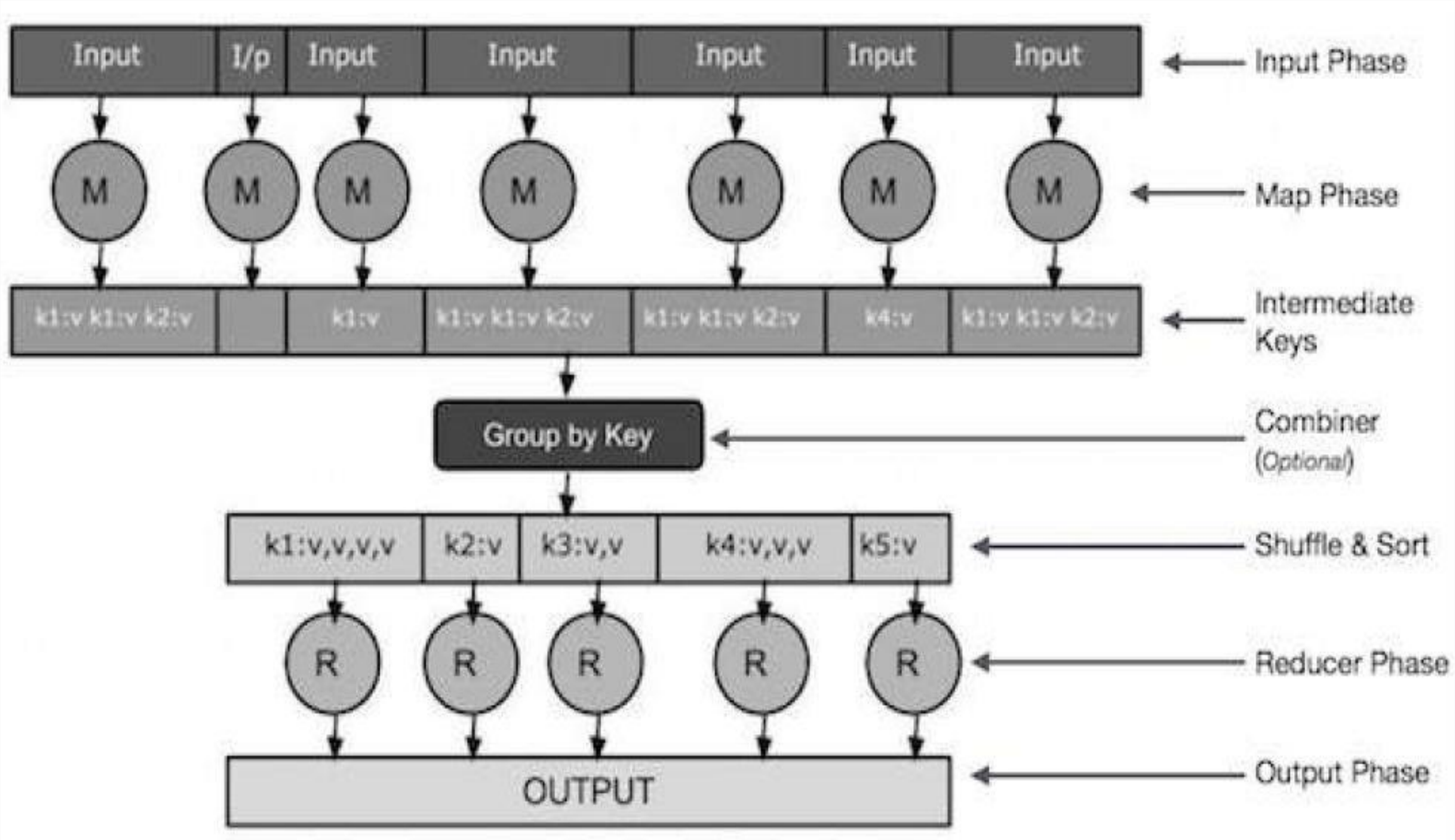
MapReduce Architecture Components

Reducer Phase-

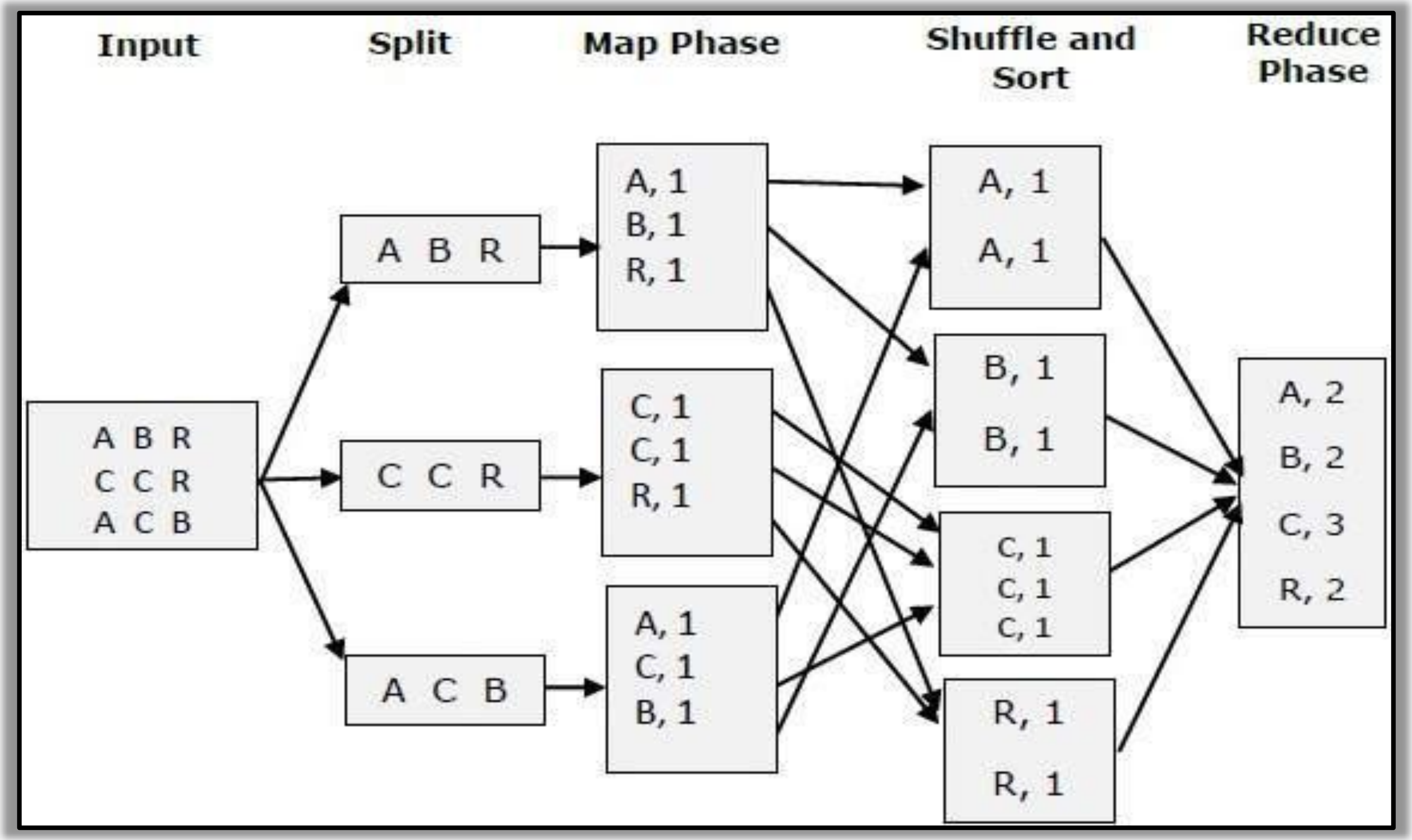
- The reducer takes in data input that is sorted & shuffled.

Output Phase-

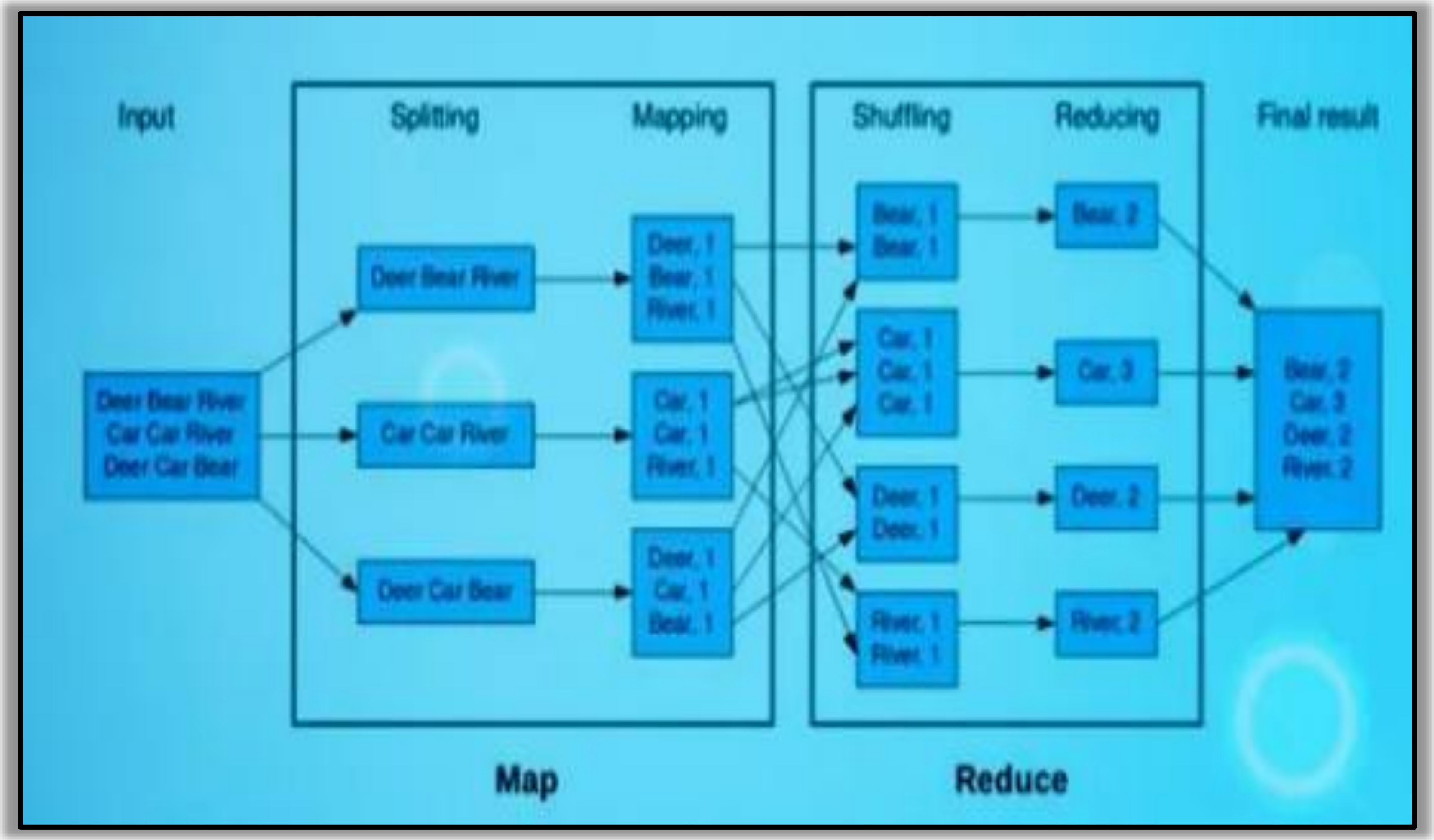
MapReduce Architecture Components



MapReduce Example



MapReduce Example



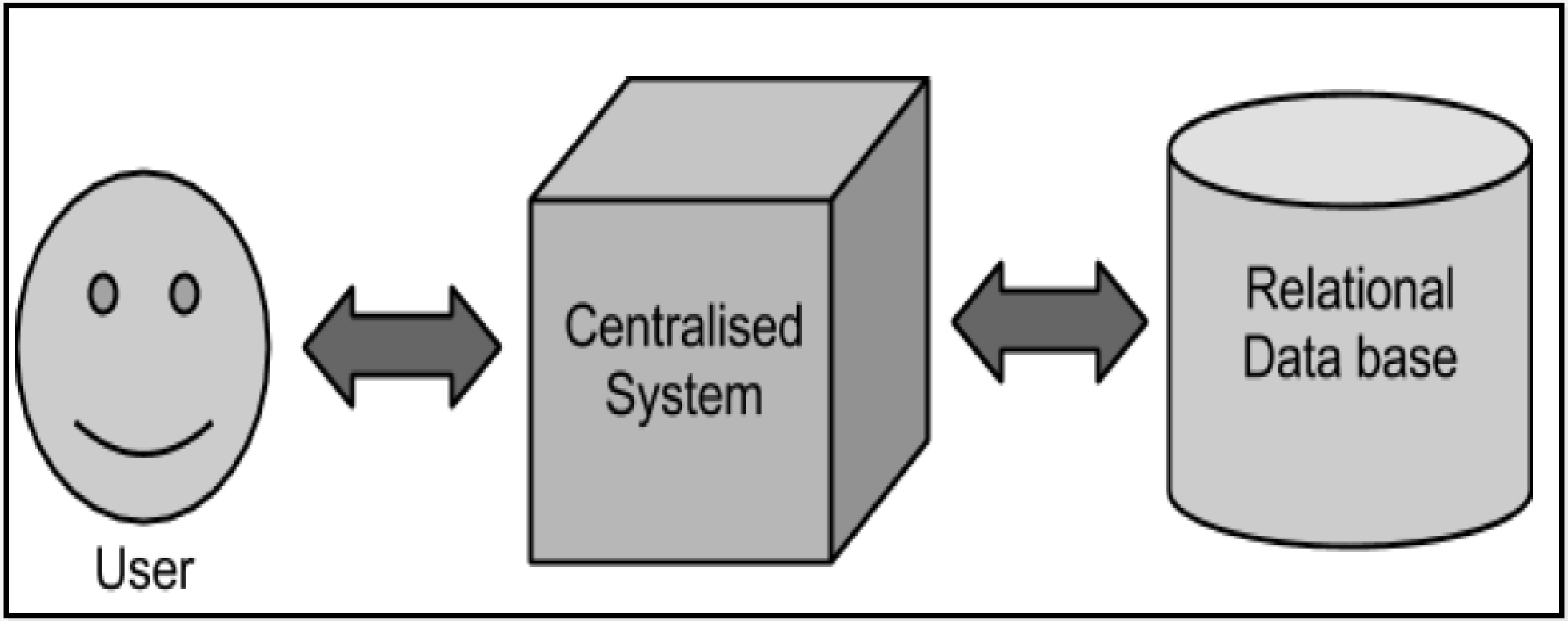
Traditional Approach

Traditional Enterprise Approach

In this approach, an enterprise will have a computer to store and process big data. For storage purpose, the programmers will take the help of their choice of database vendors such as Oracle, IBM, etc.

Traditional Approach

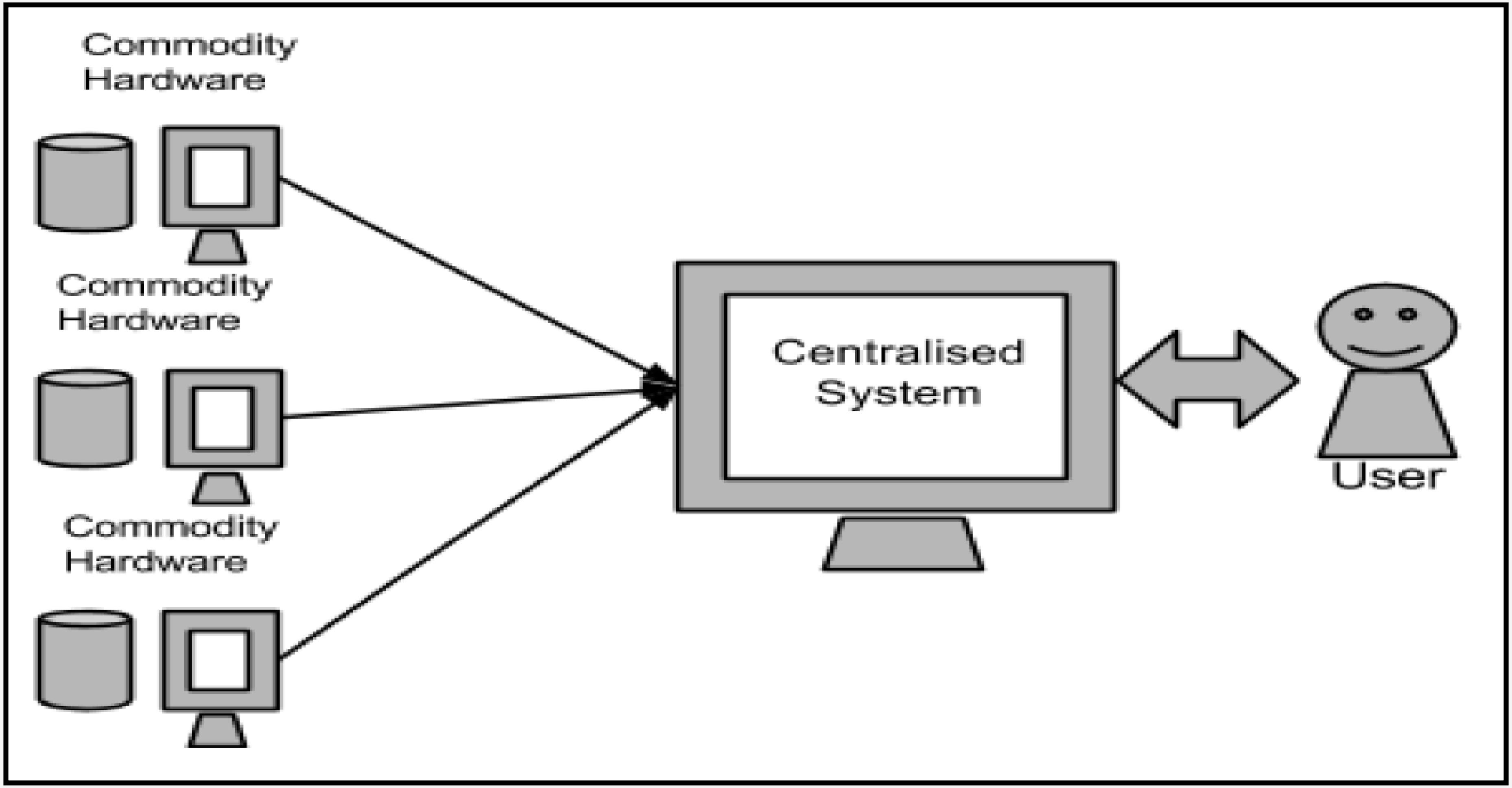
Traditional Enterprise Approach



Google's Solution- MapReduce Algorithm

- Traditional approach works fine with those applications that process less voluminous data that can be accommodated by standard database servers, or up to the limit of the processor that is processing the data.
- Google solved this problem using an algorithm called MapReduce.

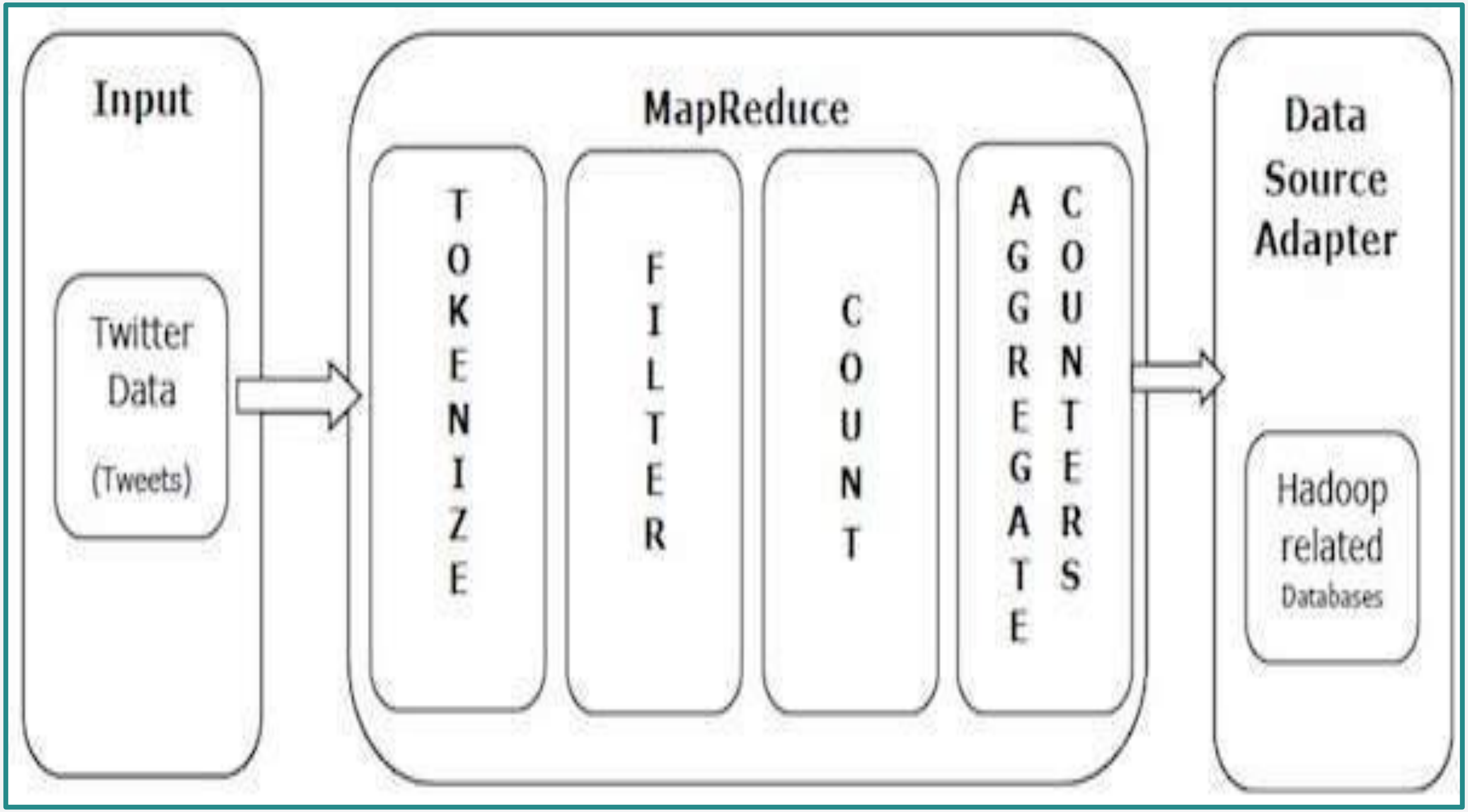
Google's Solution- MapReduce Algorithm



MapReduce Example

- Let us take a real-world example to comprehend the power of MapReduce.
- Twitter receives around 500 million tweets per day, which is nearly 3000 tweets per second.
- The following illustration shows how Twitter manages its tweets with the help of MapReduce.

MapReduce Example



MapReduce Example

As shown in the illustration, MapReduce algorithm performs the following actions –

1. Tokenize
2. Filter
3. Count
4. Aggregate Counters

That’s all for now…