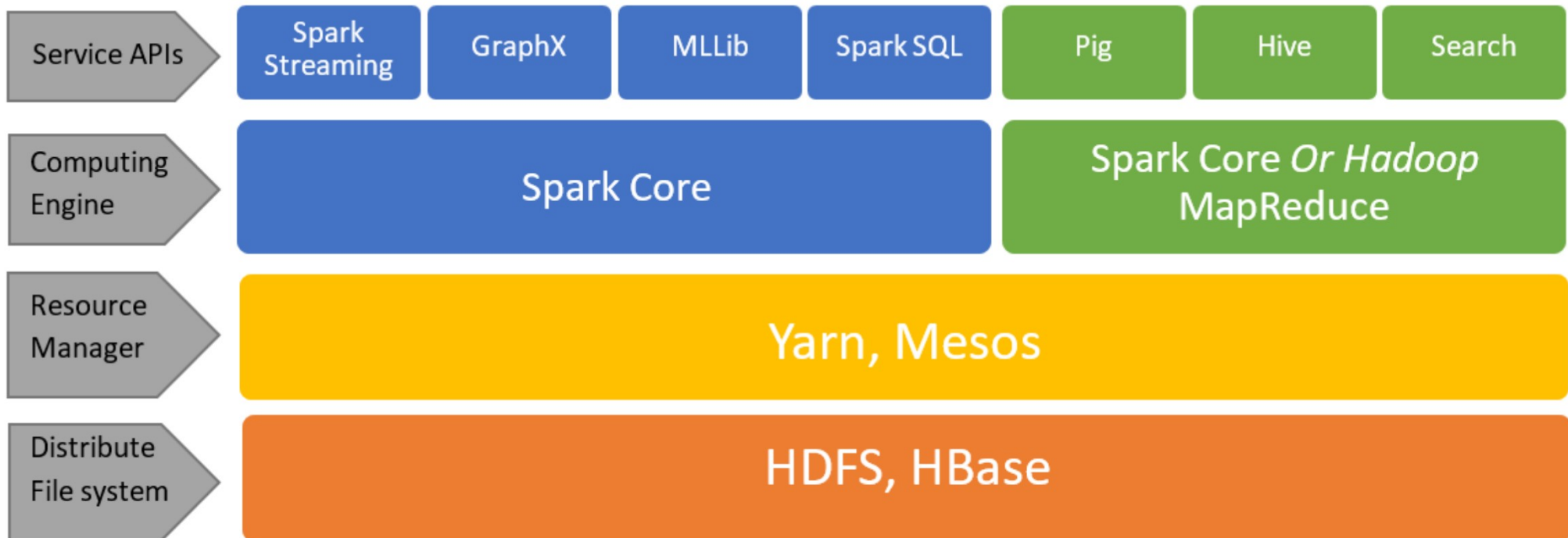
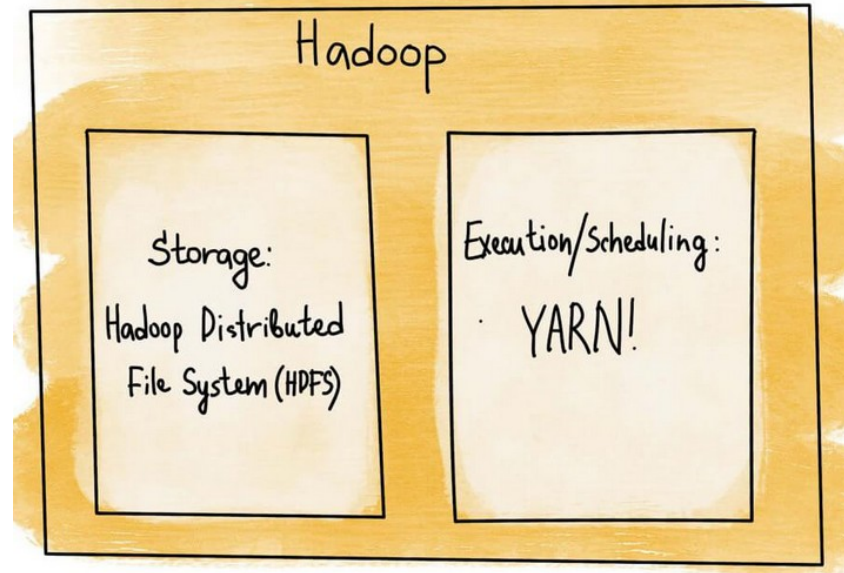
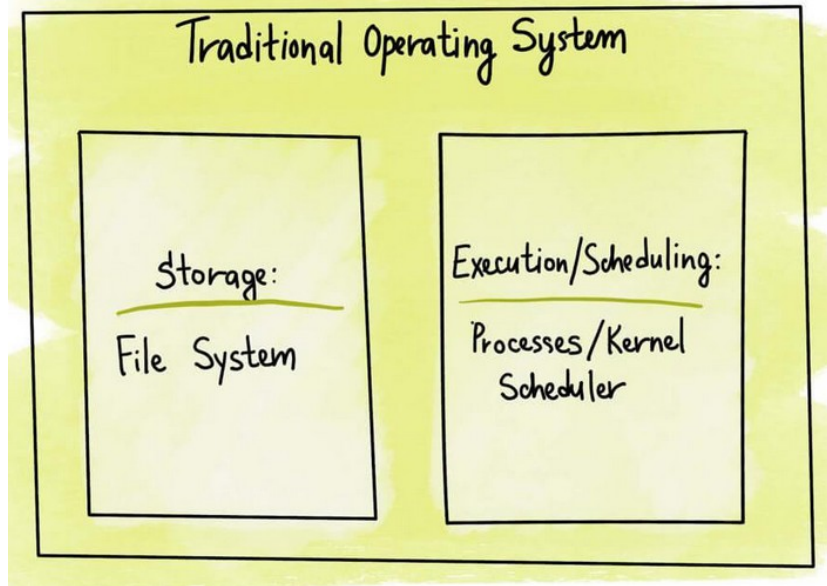


Hadoop | MapReduce



OS Analogy | Recall



Spark Ecosystem

PROGRAMMING LANGUAGES

SCALA

R

JAVA

PYTHON

LIBRARIES

SPARK SQL

MLlib

GRAPHX

STREAMING

ENGINE

SPARK CORE

CLUSTER MANAGEMENT

HADOOP YARN

APACHE MESOS

SPARK SCHEDULER

STORAGE

HDFS

STANDALONE
NODE

CLOUD

RDBMS/NOSQL

Apache Hadoop Framework

Hadoop is a framework that allows us to **store** and **process** large data sets in **parallel** and **distributed** fashion

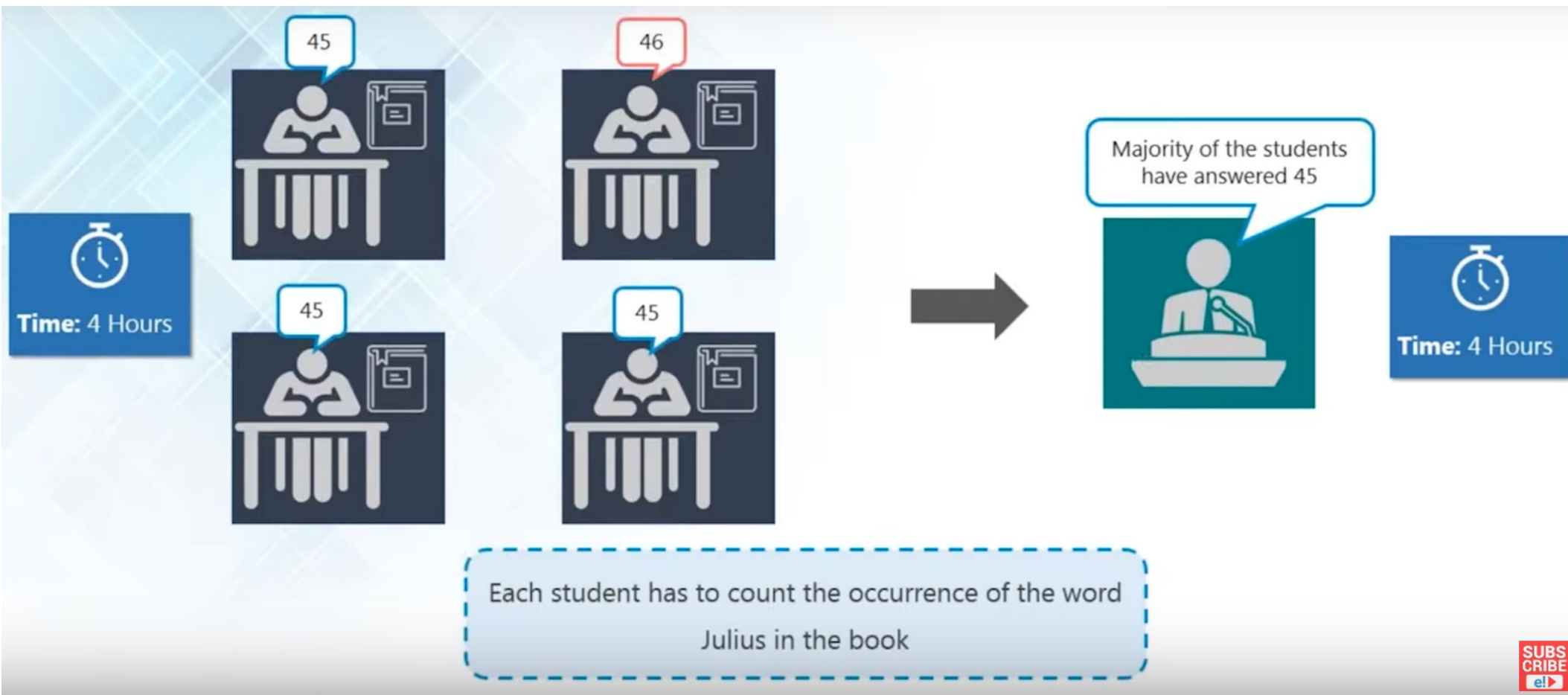


Storage:
Distributed File
System

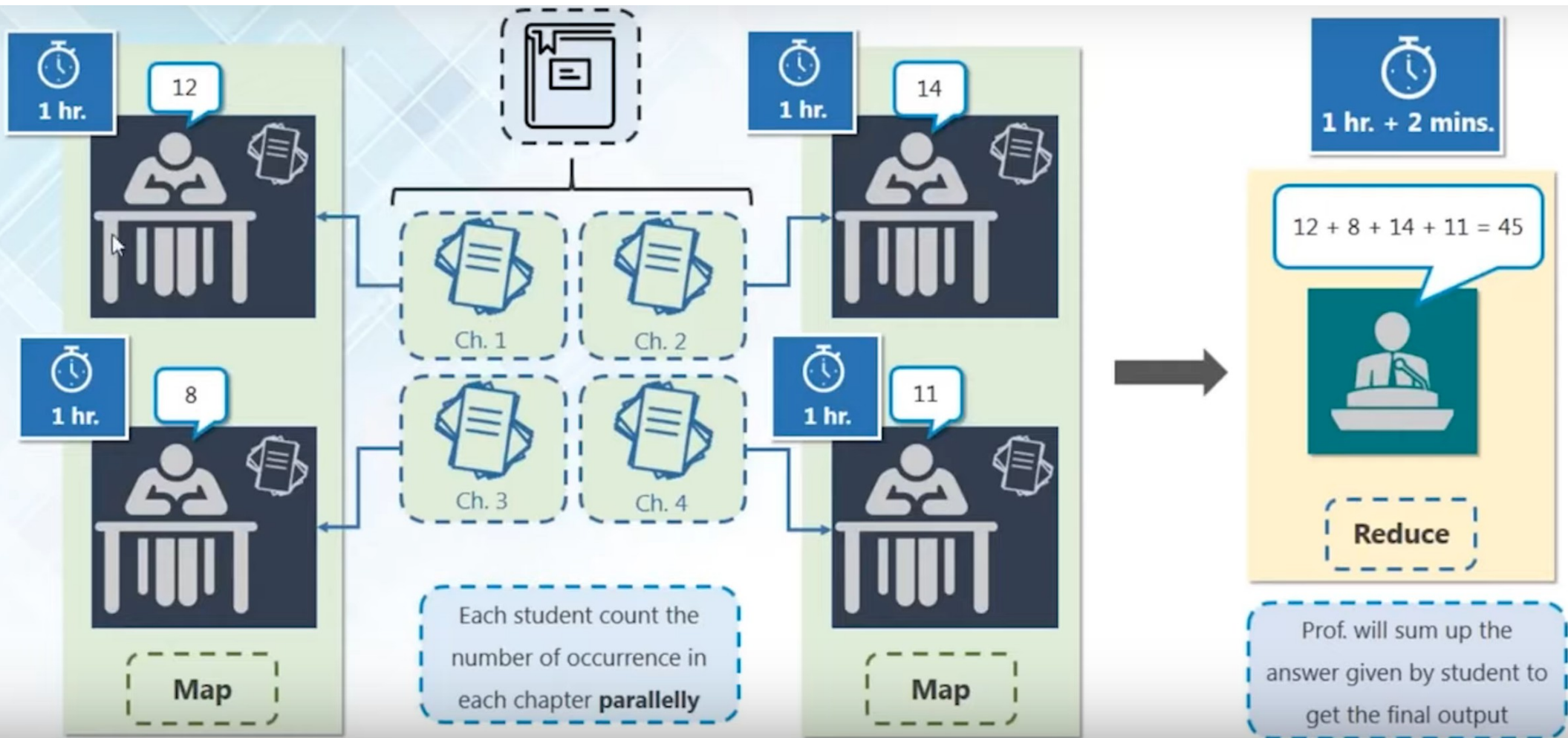


Processing:
Allows parallel &
distributed
processing

Story of MapReduce

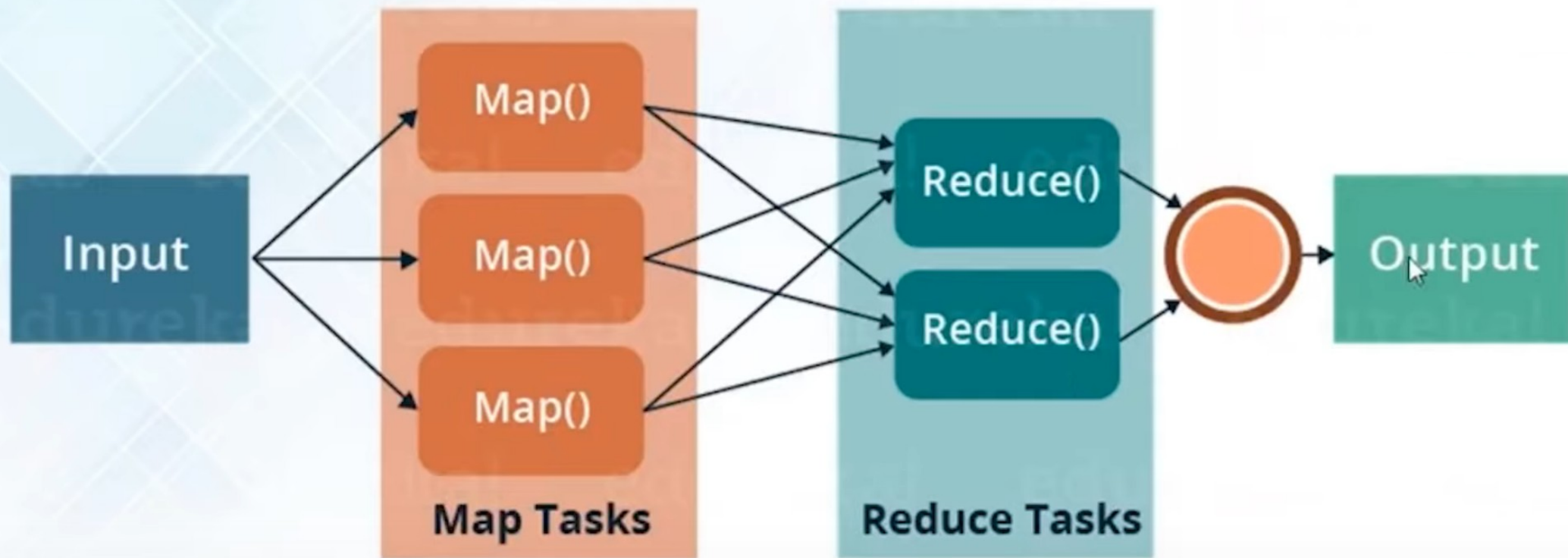


Story of MapReduce

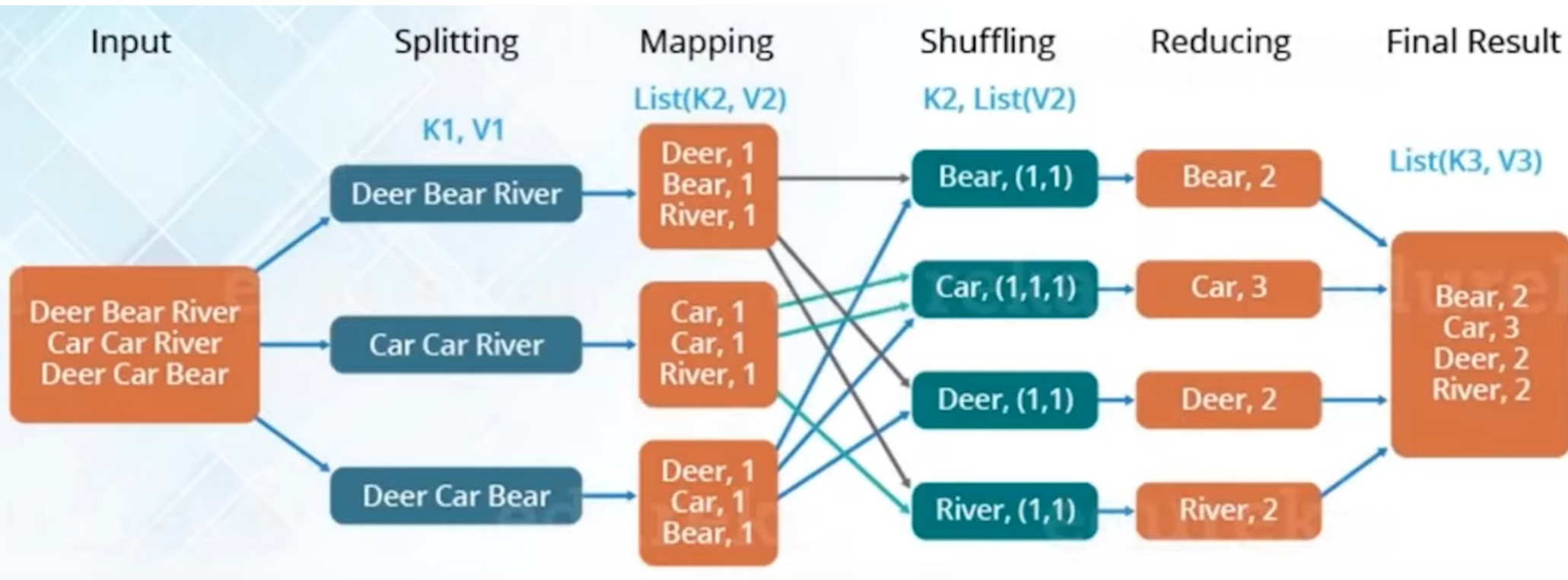


What is MapReduce

MapReduce is a **programming framework** that allows us to perform **distributed** and **parallel** processing on large data sets in a distributed environment



MapReduce Word Count Program



MapReduce Word Count Program

1

Mapper Code:

You write the mapper logic over here i.e. how map task will process the data to produce the key-value pair to be aggregated

2

Reducer Code:

You write reducer logic here which combines the intermediate key-value pair generated by Mapper to give the final aggregated output

3

Driver Code

You specify all the job configurations over here like job name, Input path, output path, etc.

Spark vs Hadoop MapReduce

Factors

Speed

100x times than MapReduce

Faster than traditional system

Written In

Scala

Java

Data Processing

Batch / real-time / iterative /
interactive /graph

Batch processing

Ease of Use

Compact & easier than Hadoop

Complex & lengthy

Caching

Caches the data in-memory &
enhances the system performance

Doesn't support caching of data

Spark

Hadoop MapReduce