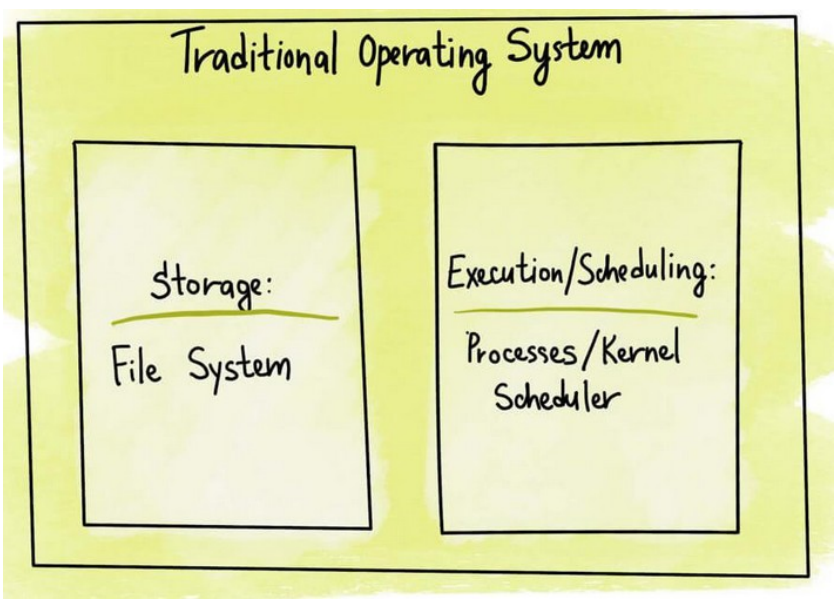
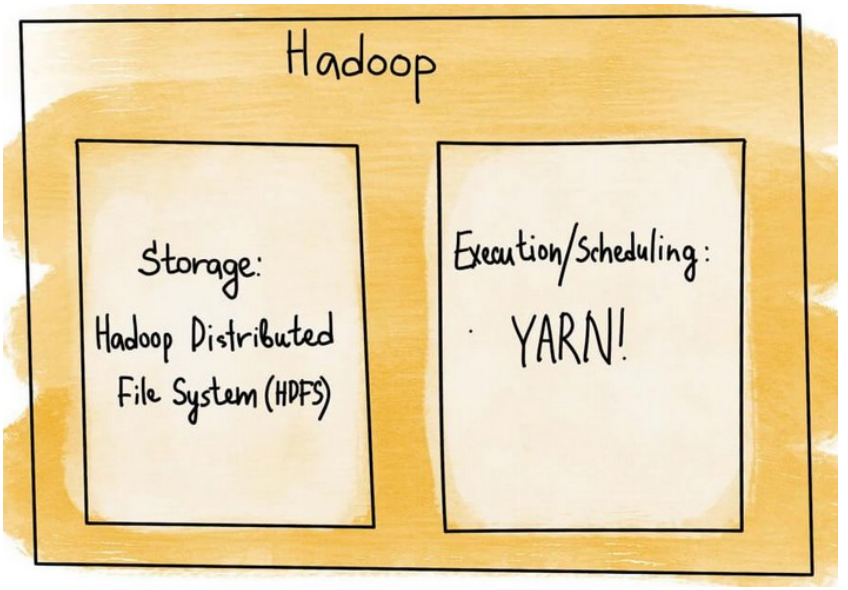
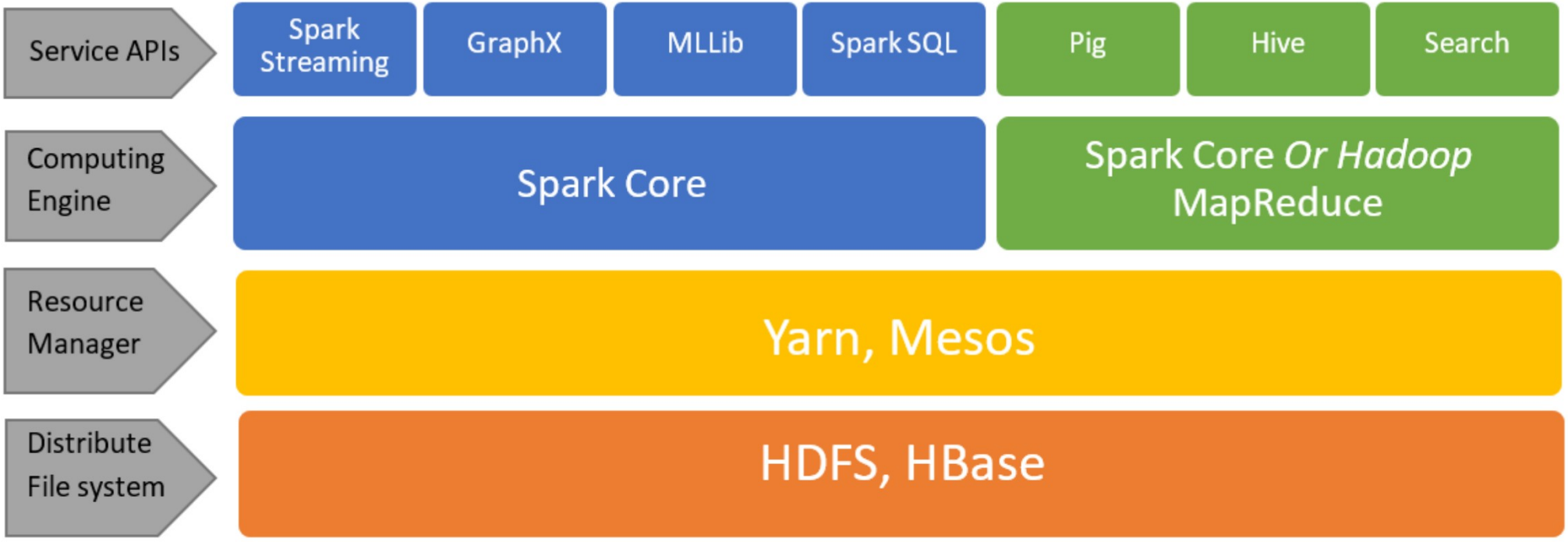
Hadoop | MapReduce

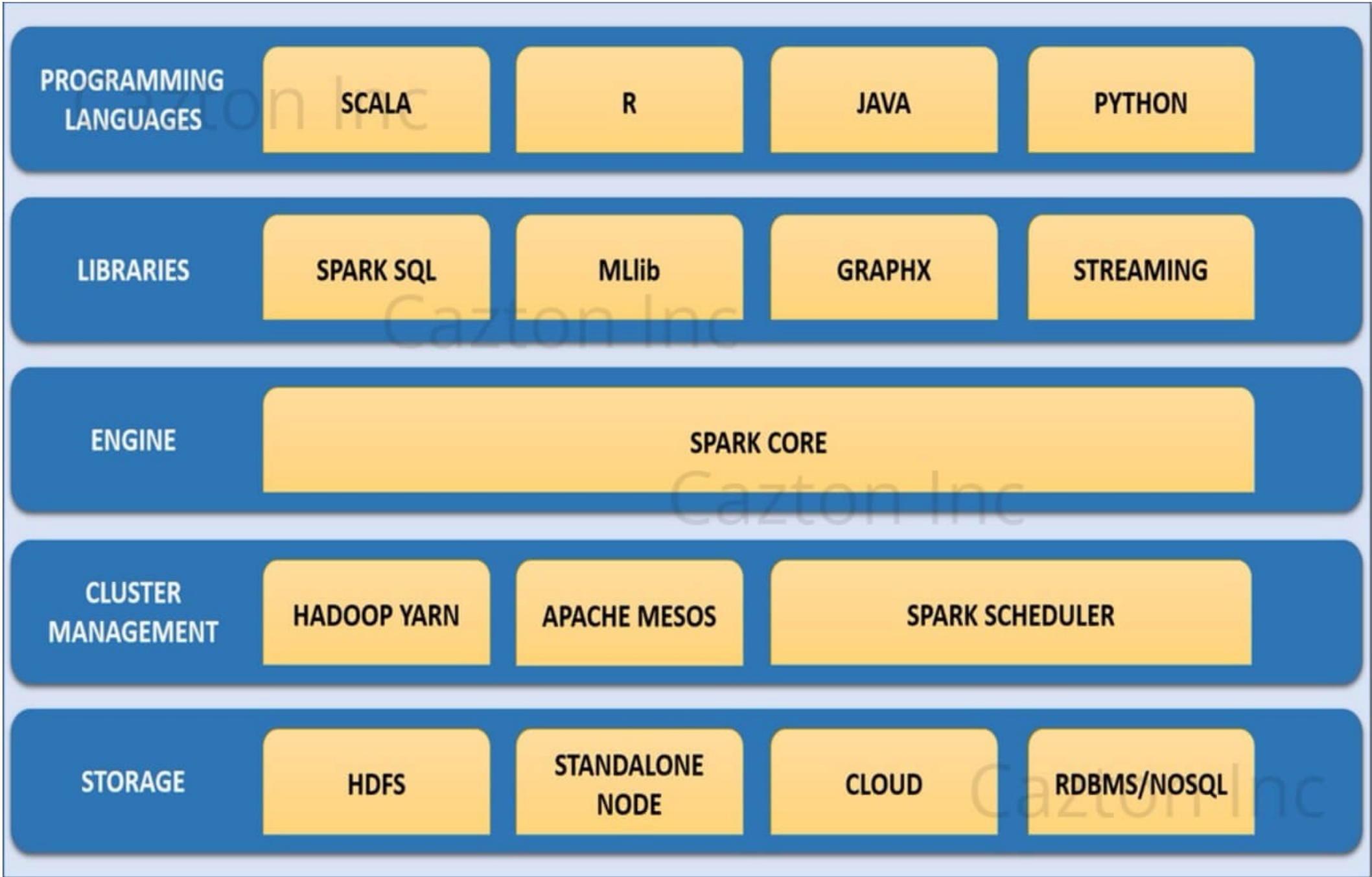
OS Analogy | Recall



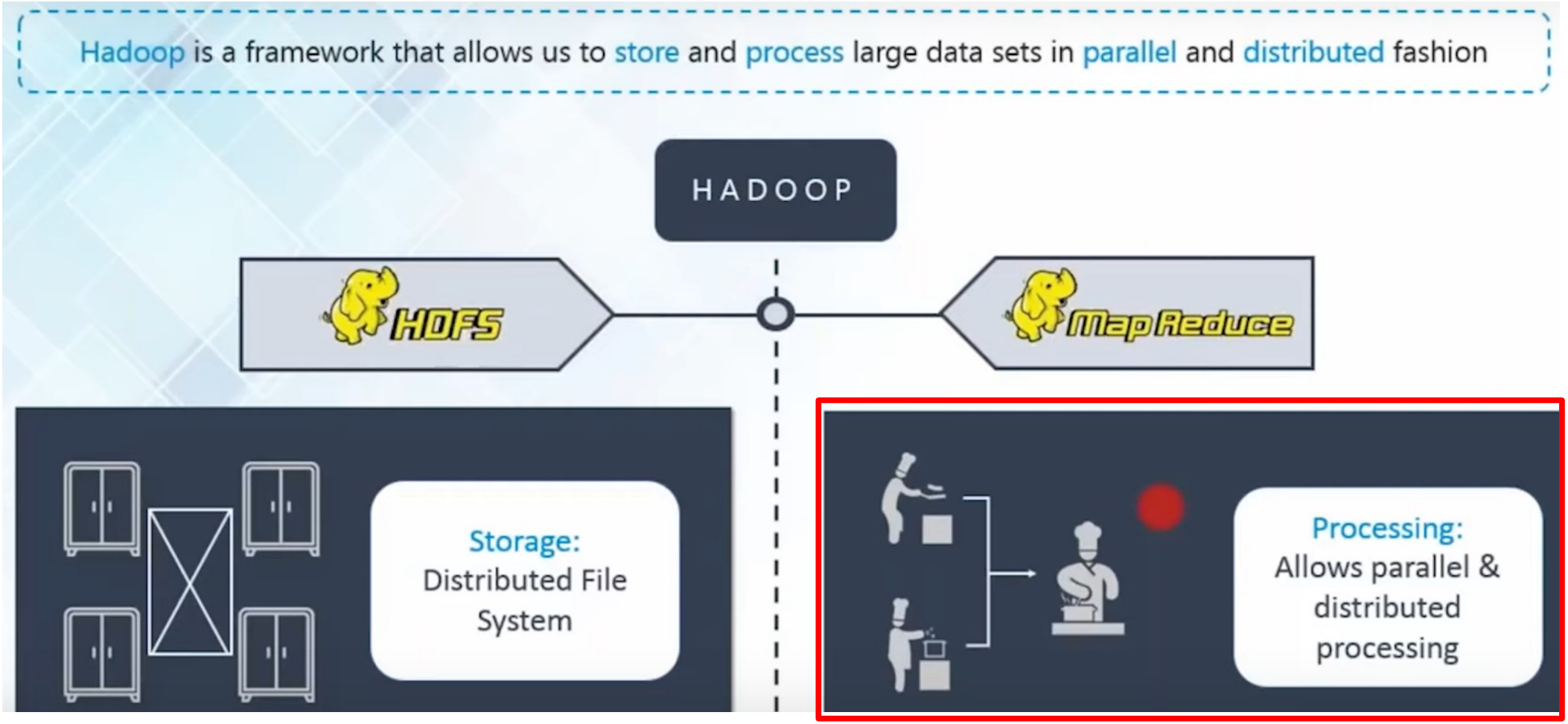




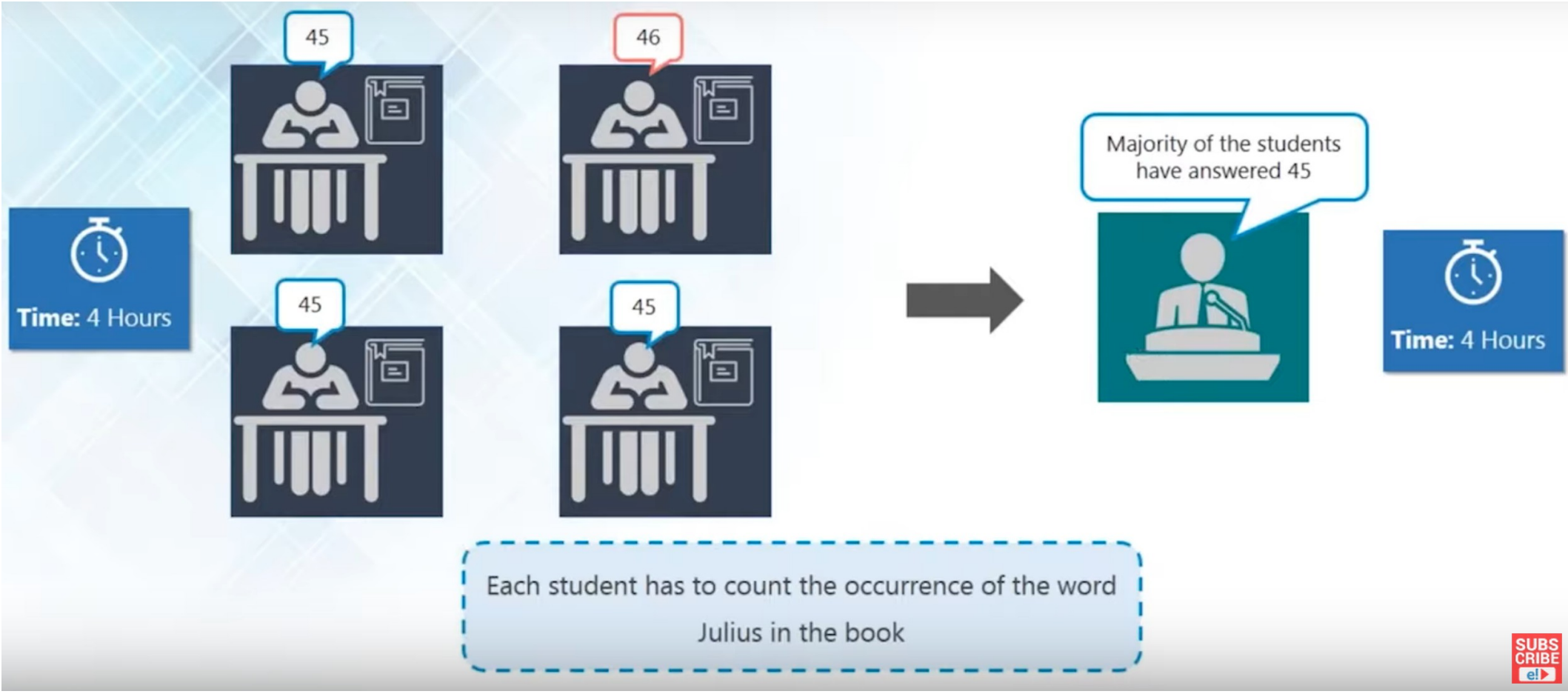
Spark Ecosystem



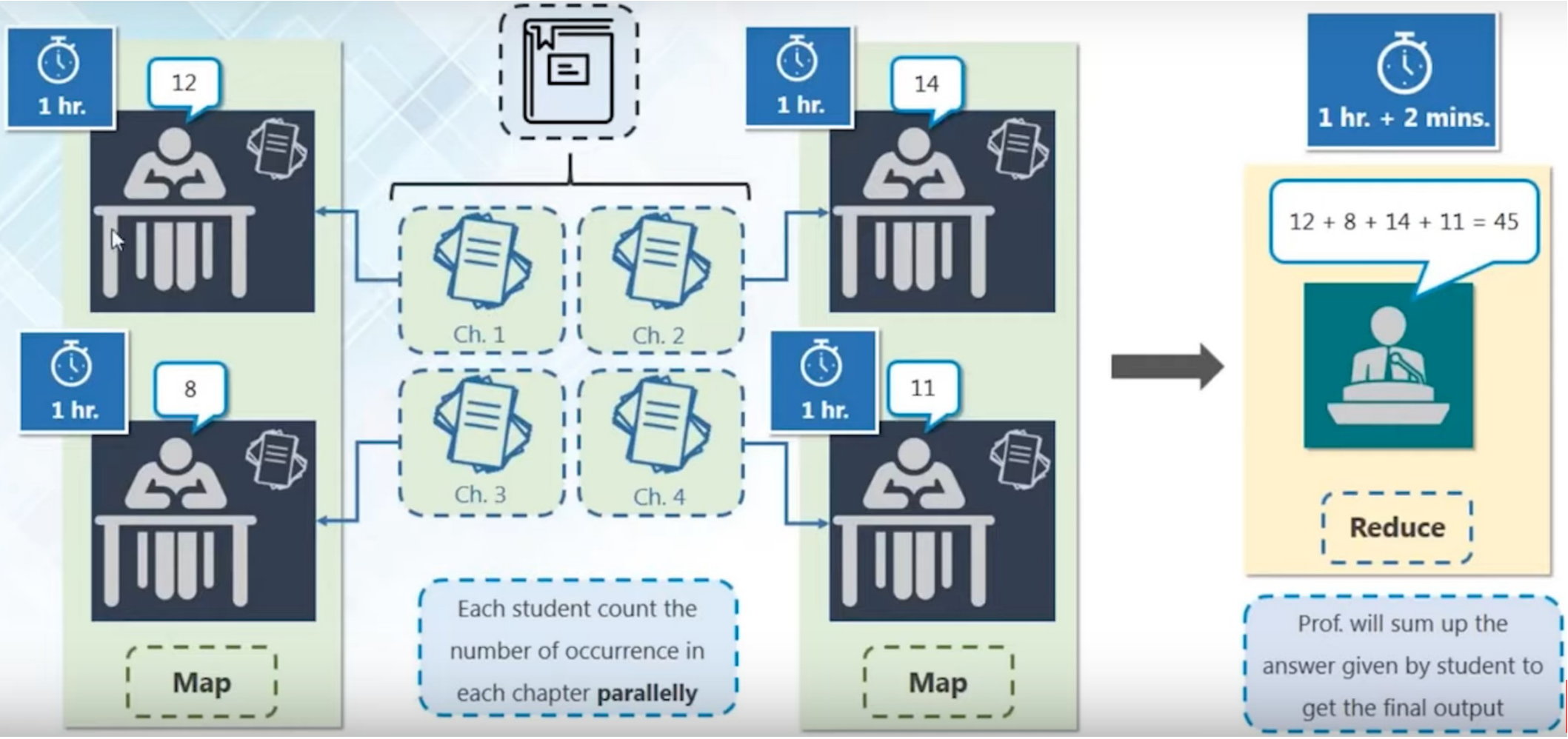
Apache Hadoop Framework



Story of MapReduce

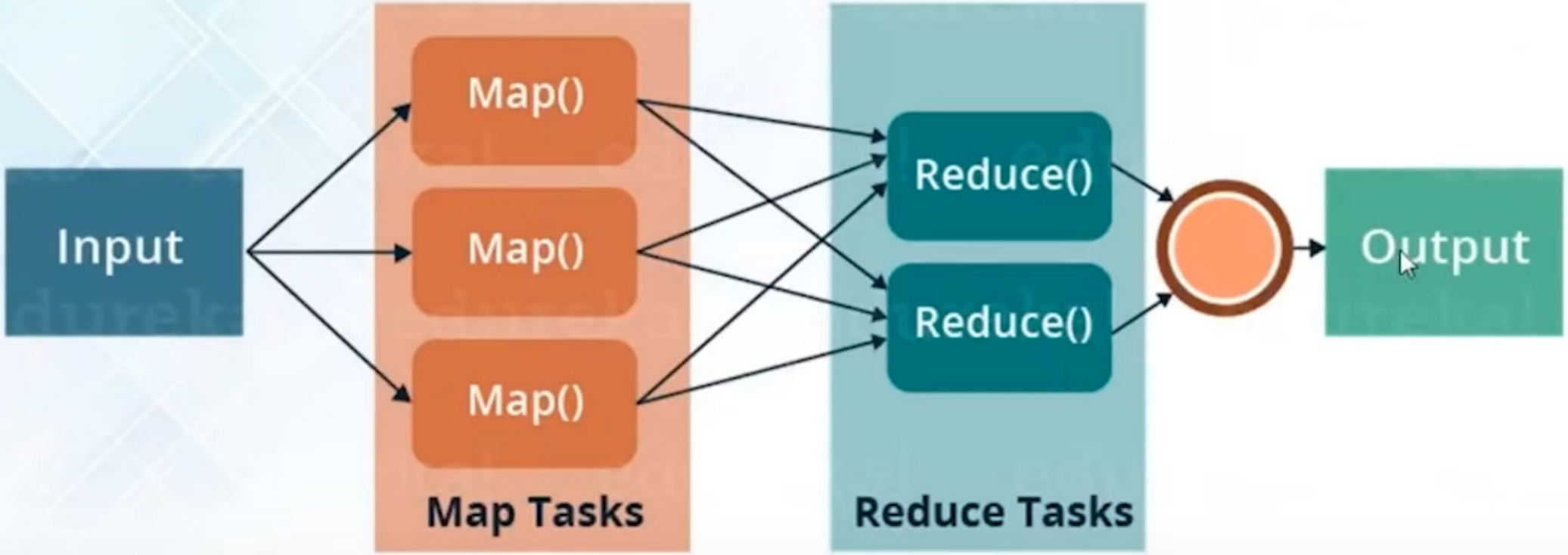


Story of MapReduce

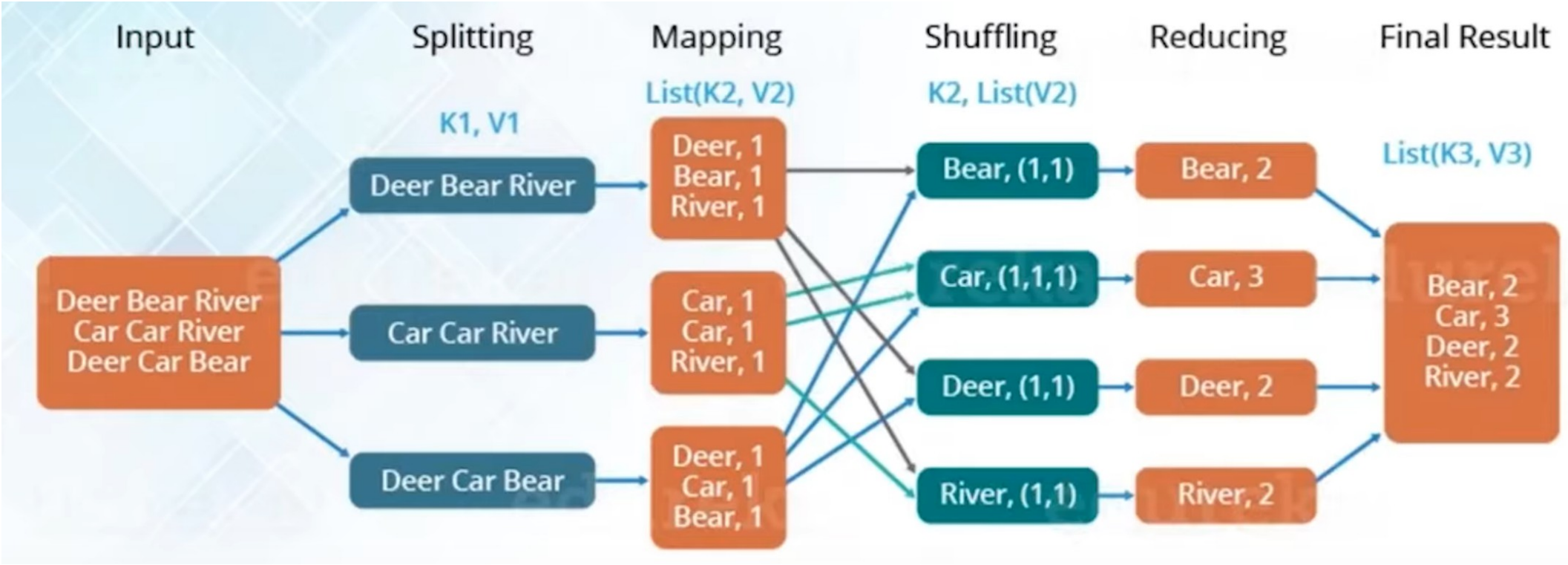


What is MapReduce

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	Spark	Hadoop MapReduce
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