MapRedue: Simplified Data Processing on Large Clusters

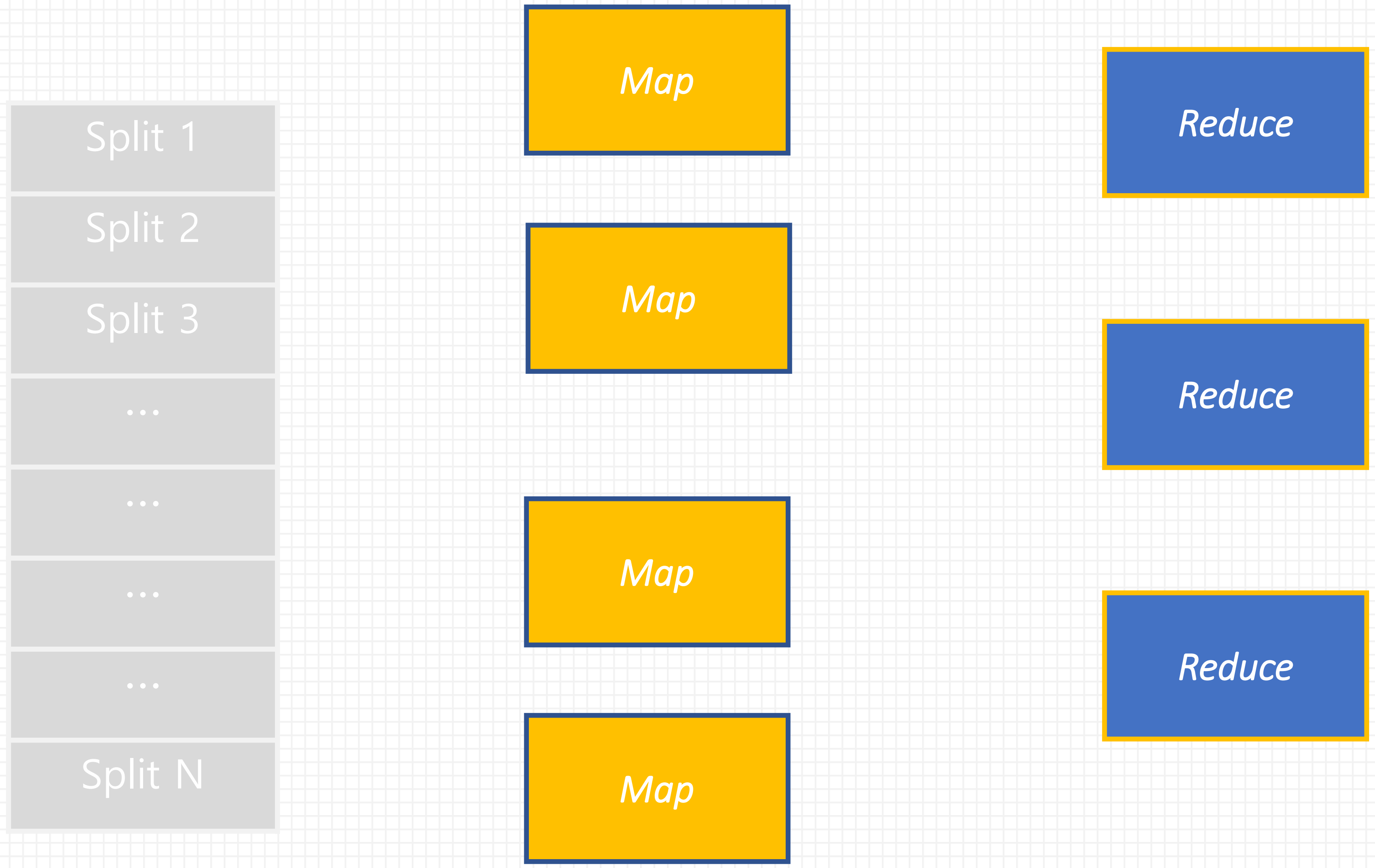
Jeffry Dean
Sanjay Ghemawat

Presenter
KyungHee University
Donghoon Han

About MapReduce

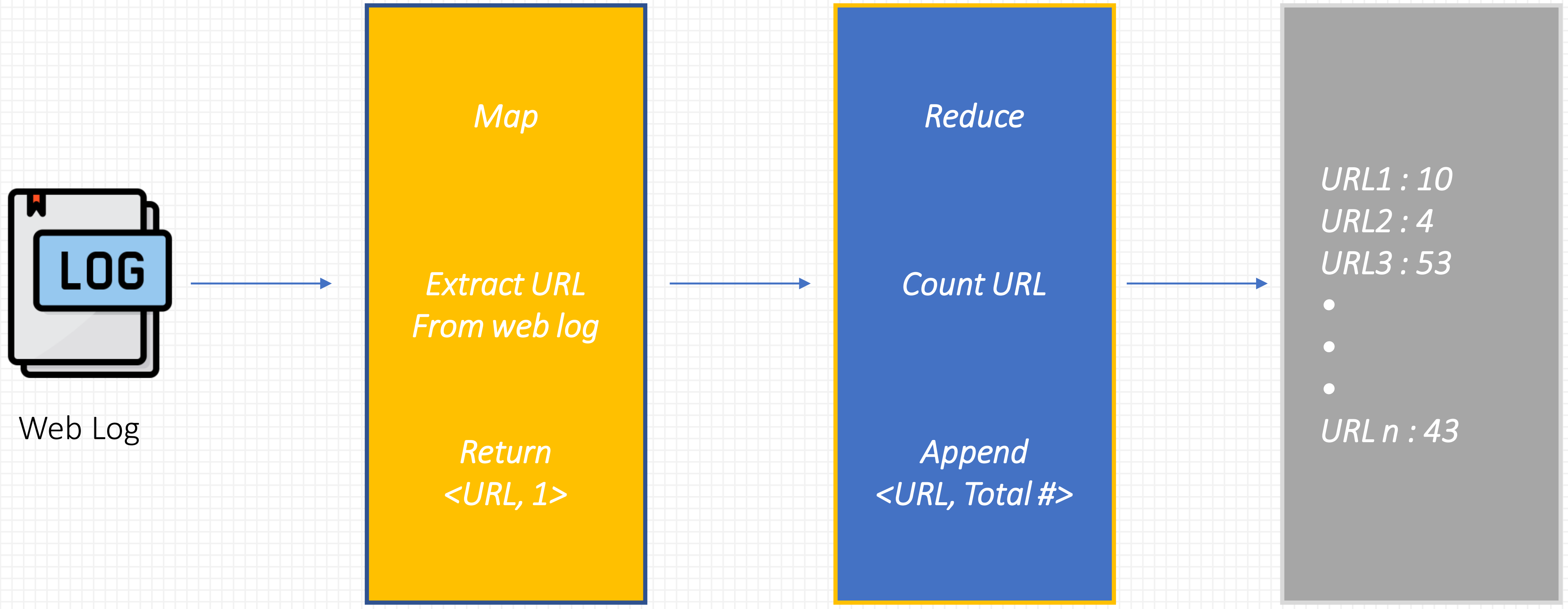
- Software framework developed for processing large cluster of data
- Parallel computation using large cluster of commodity machines
- MapReduce classify data with key value. Then summarizes data with user defined command
- MapReduce consists of "Map function" and "Reduce function"
- Map function processes input and returns intermediate key/value pair
- Reduce function merges all the intermediate pairs and summarize them

Model Explanation



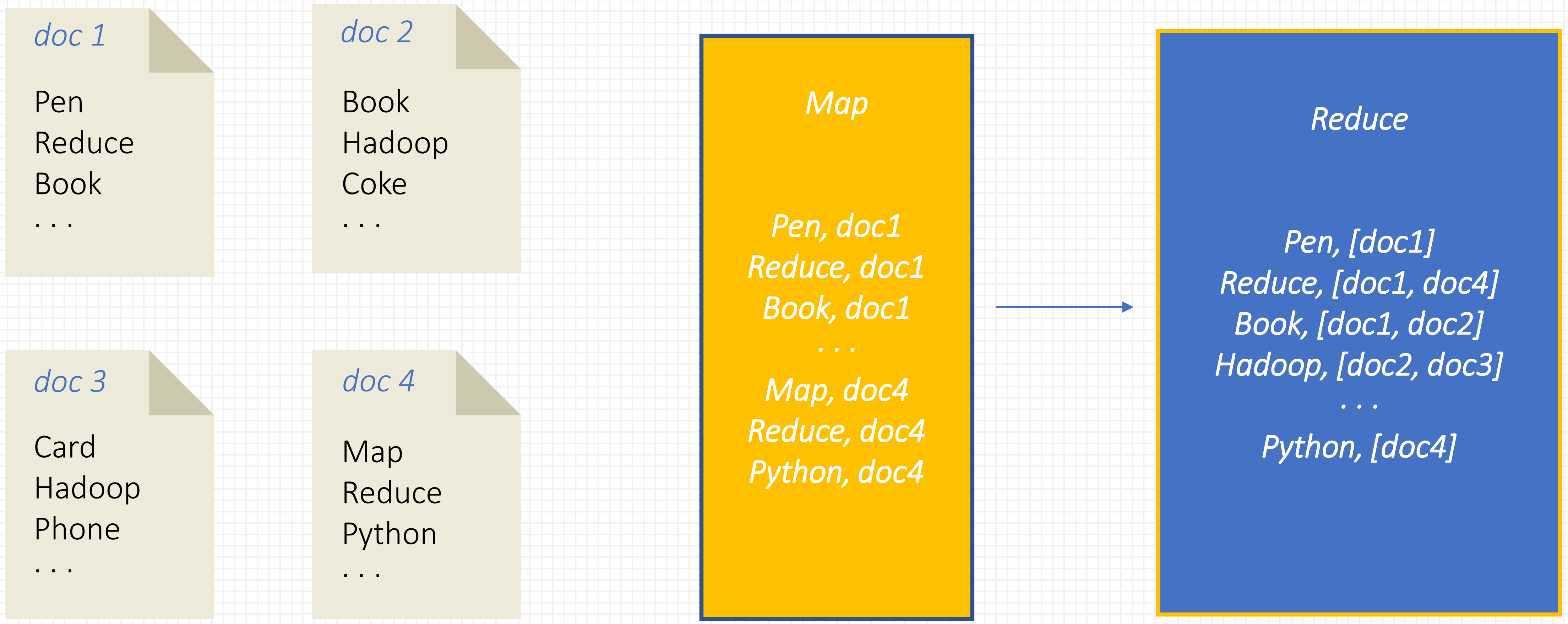
For your Intuition

Count of URL Access Frequency



For your Intuition

Inverted Index



How It Actually Operates

Further Information

Fault Tolerance

Worker Failure
- When worker machine fails,
- Failed workers are marked as idle
- New Map machine's saving location will be

Master Failure
- Master occasionally leaves checkpoint
- If master dies, another candidate will continue from that checkpoint
- If there is only one master machine, execute operation again

Locality

- There will be 'M' mappers and 'R' reducers
- It's good for 'M' and 'R' to outnumber a # of machines
- Spited input pieces should not be to big
- Effective ***"Dynamic load balancing"***

Further Information

Task Granularity

Backup Task

Experiment / Implementation

Experiment Environment

1TB of Input Data
Experiment done in two ways; ‘Grep’ and ‘Sorting’

Last But Not Least

Thank you