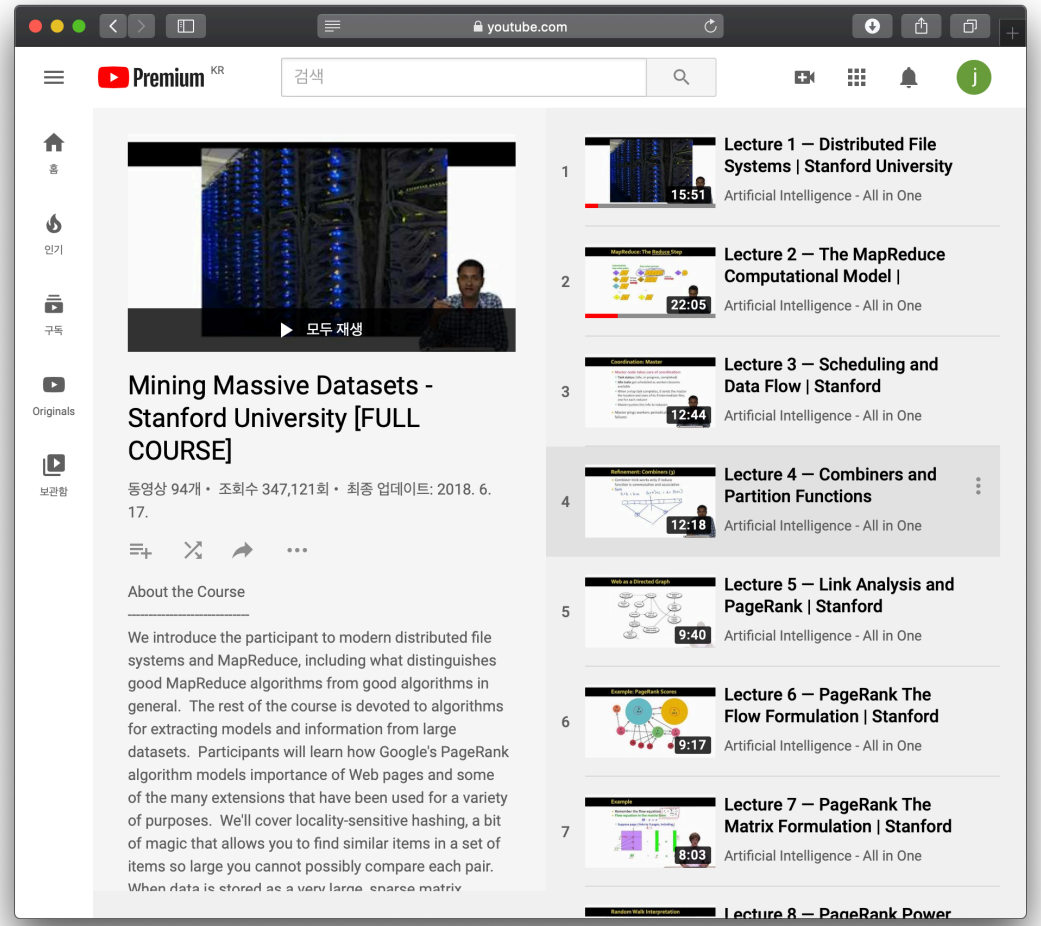
Mining Massive Datasets

Jisung Jeong / jisung0920.github.io

Reference

- Open Course :
  Mining Massive Datasets (Stanford)

- Textbook :
  Mining of Massive Datasets
  (Jure Leskovec, Anand Rajaraman, Jeff Ullman)

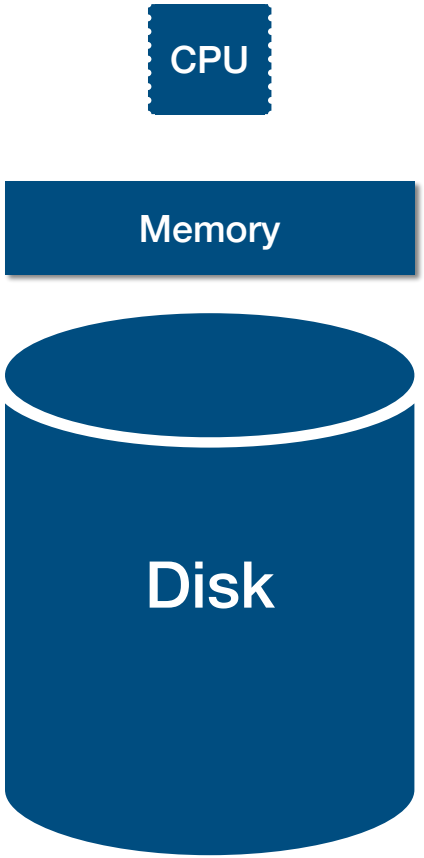


1. MapReduce
2. PageRank
3. Locality-Sensitive Hashing
4. Distance Measures, Nearest Neighbors
5. Frequent Itemsets
6. Communities in Social Networks
7. Stream Algorithms
8. Recommender Systems
9. Dimensionality Reduction
10. Clustering
11. Computational Advertising
12. Machine Learning
13. More About MapReduce
14. More About Locality-Sensitive Hashing
15. More About Link Analysis

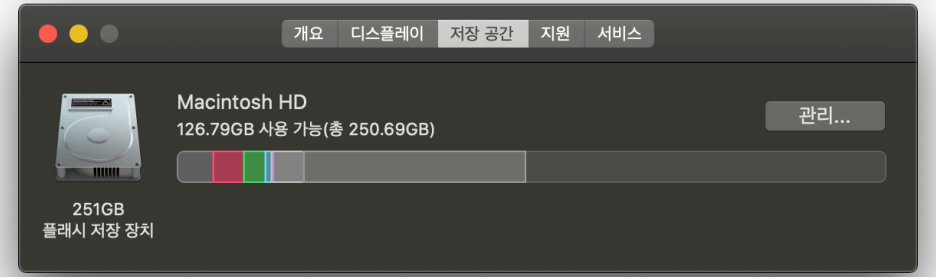
MapReduce

Physical Organization of Compute Nodes

Jisung Jeong / jisung0920.github.io


CPU
Memory
Disk

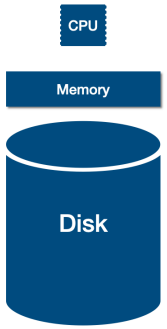

개요
디스플레이
저장 공간
지원
서비스
macOS Catalina
버전 10.15.5
MacBook Air (Retina, 13-inch, 2020)
프로세서 1.1 GHz 쿼드 코어 Intel Core i5
메모리 8GB 3733 MHz LPDDR4
그래픽 Intel Iris Plus Graphics 1536 MB
일련 번호 FVFCM6VSMNHX
시스템 리포트...
소프트웨어 업데이트...
™ 및 © 1983-2020 Apple Inc. 모든 권리 보유. 사용권 계약

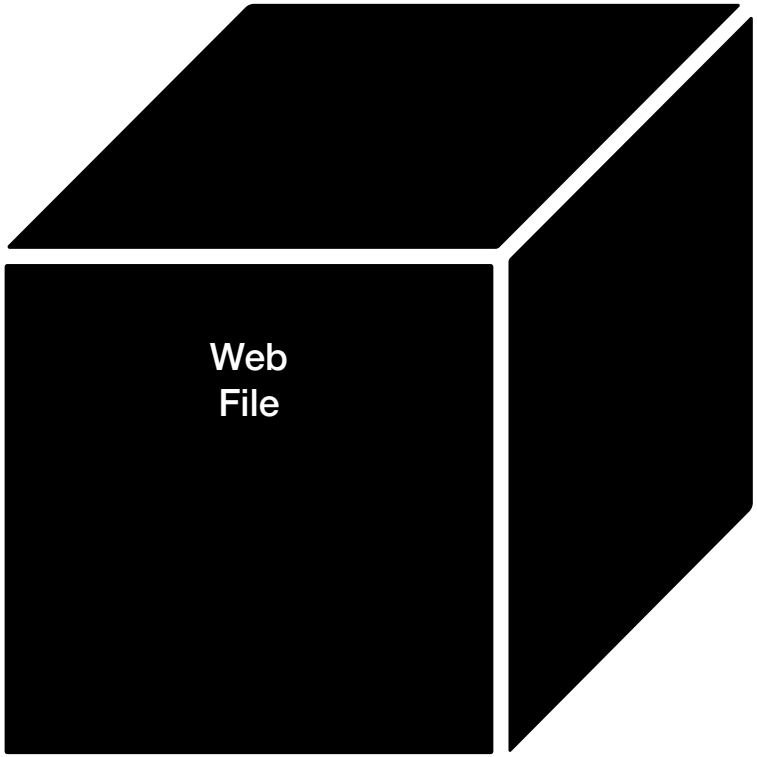

개요
디스플레이
저장 공간
지원
서비스
Macintosh HD
126.79GB 사용 가능(총 250.69GB)
관리...
251GB
플래시 저장 장치

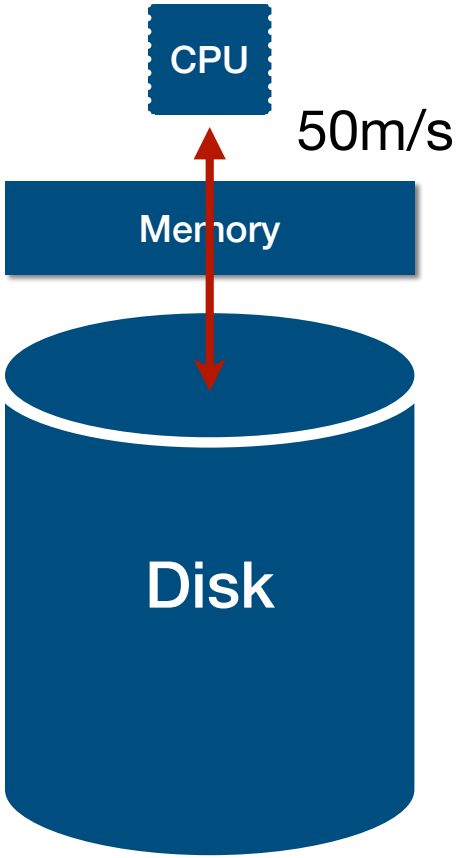
- 웹페이지 수 : 10Billion

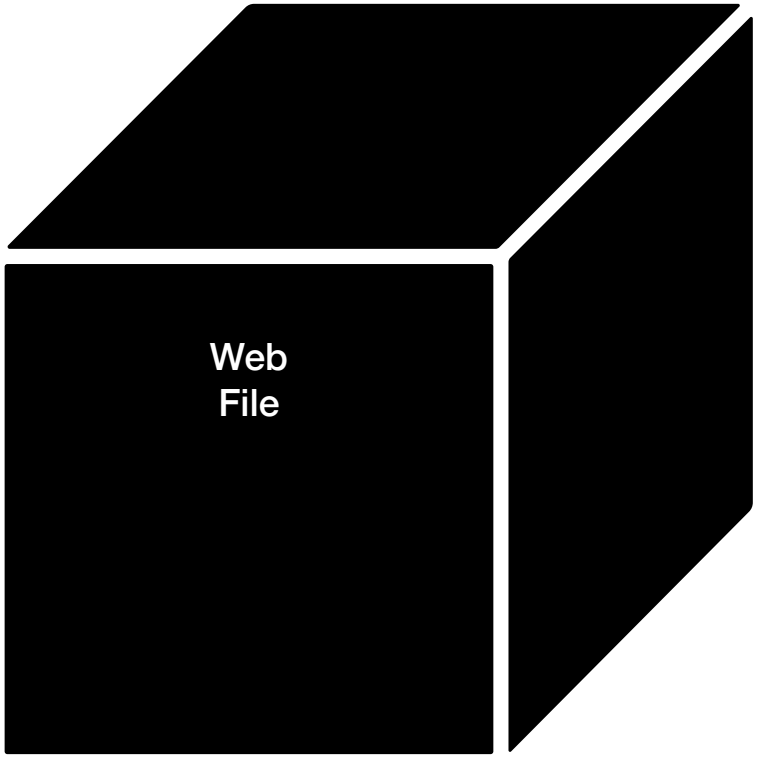
- 웹페이지 크기 : 20KB

= 200TB






CPU
50m/s
Memory
Disk

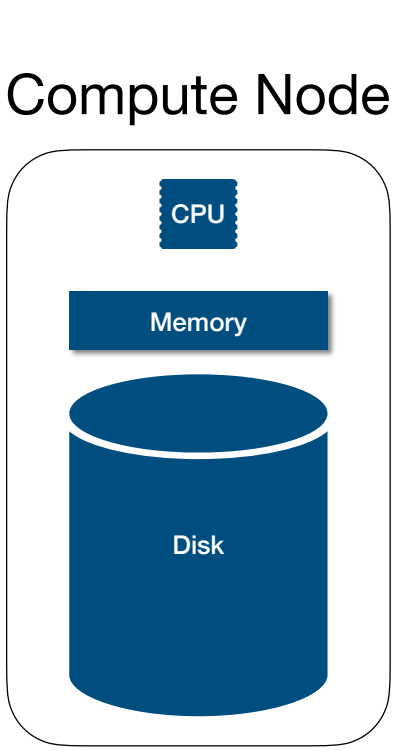

Web
File

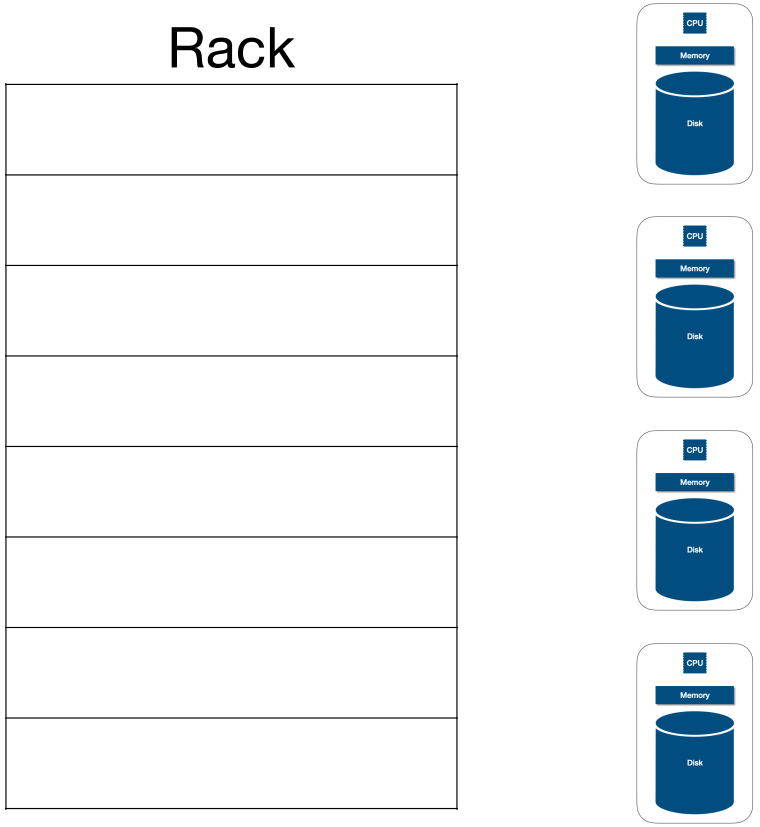
1. Physical Organization of Compute Nodes

2. Distributed File Systems

3. MapReduce

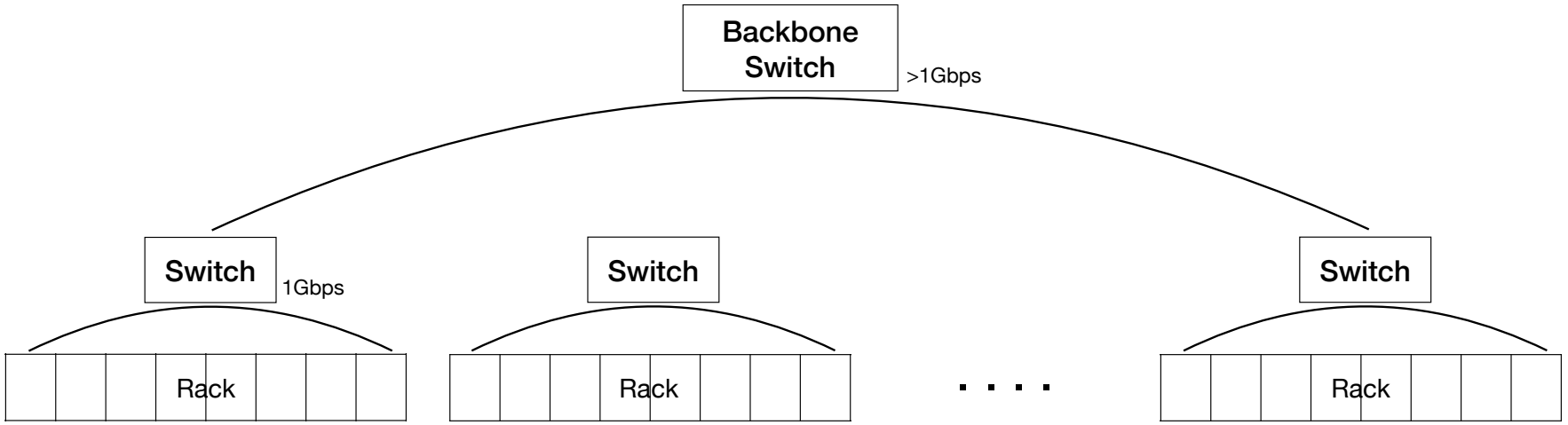
1. Physical Organization of Compute Nodes





1. Physical Organization of Compute Nodes

1. Physical Organization of Compute Nodes



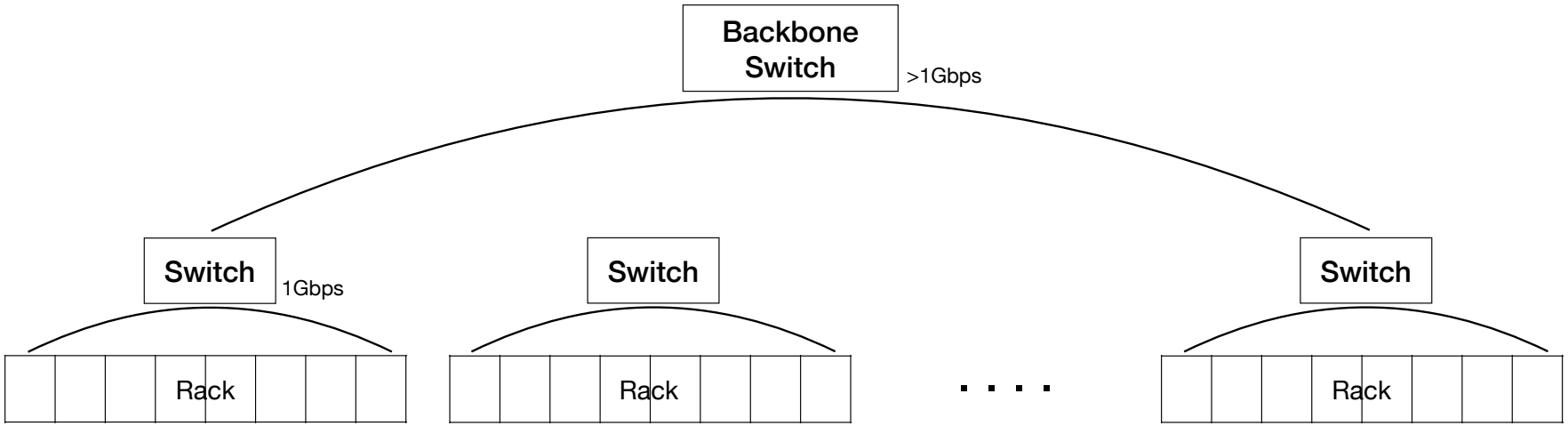
1. Physical Organization of Compute Nodes

J Lab

MapReduce

Distributed File Systems

Jisung Jeong / jisung0920@gmail.com

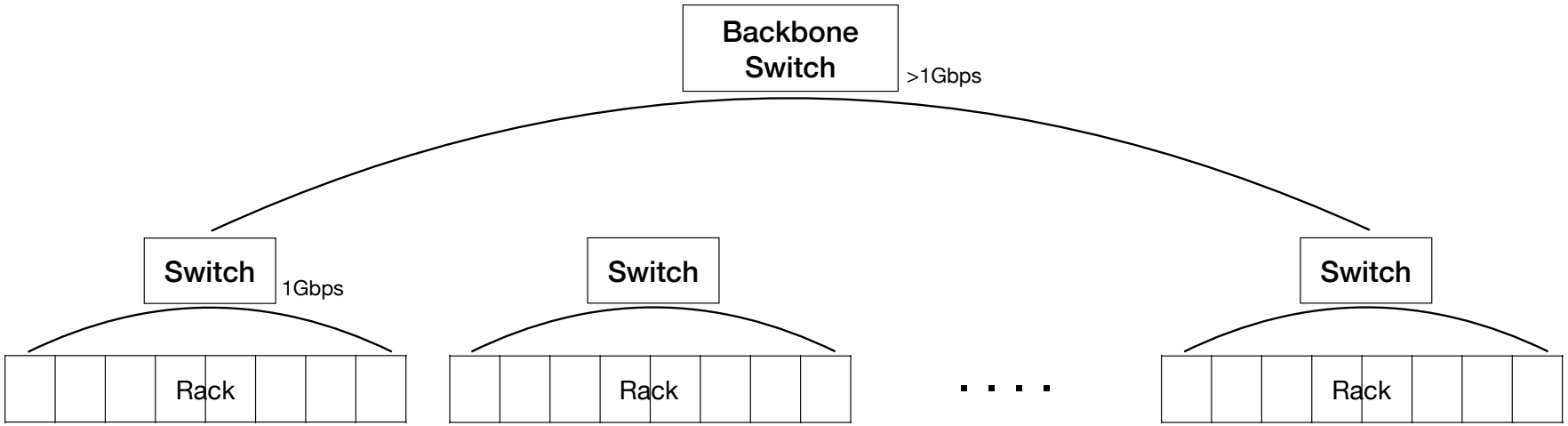

Backbone
Switch
>1Gbps
Switch
1Gbps
Switch
Switch
Rack
Rack
Rack

1. Node Failure

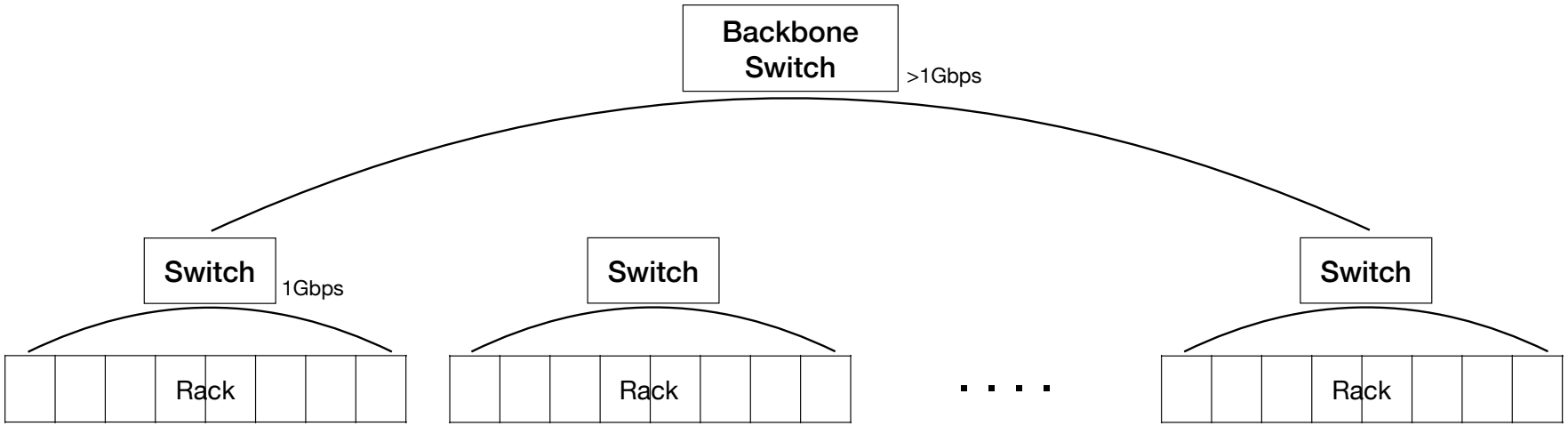
2. Network Bottleneck

3. Programming

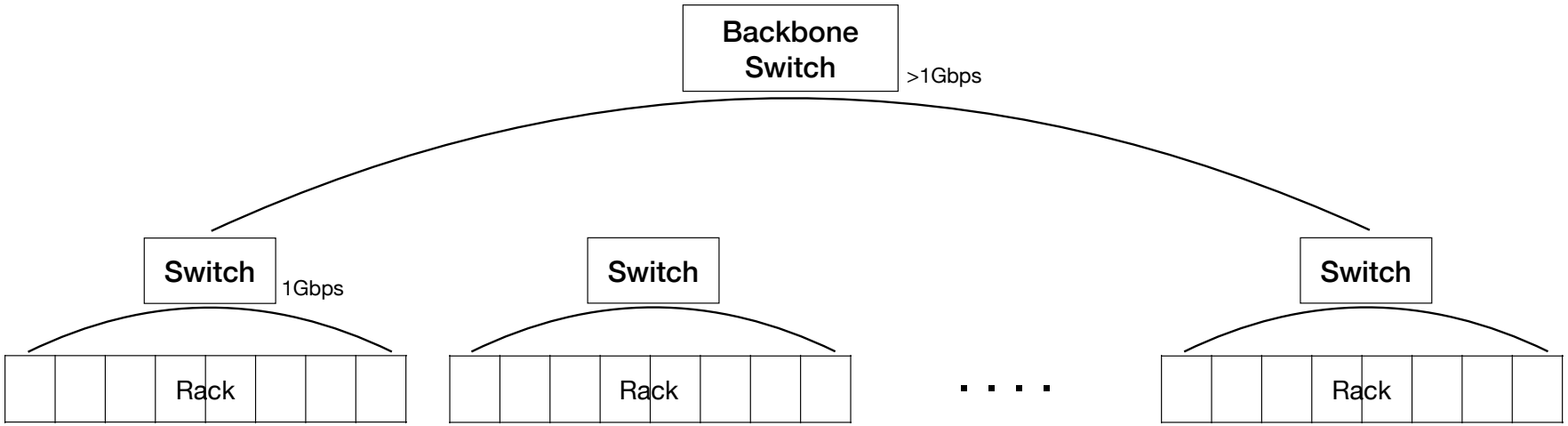
Node Failure



Network Bottleneck

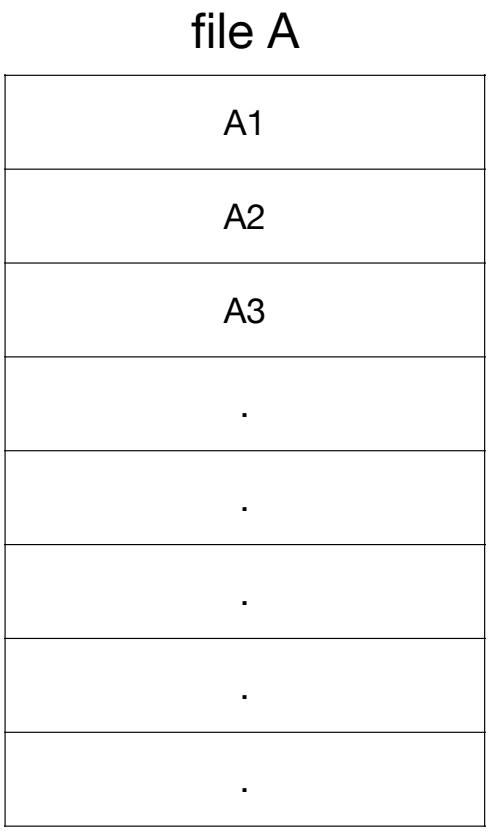


Programming

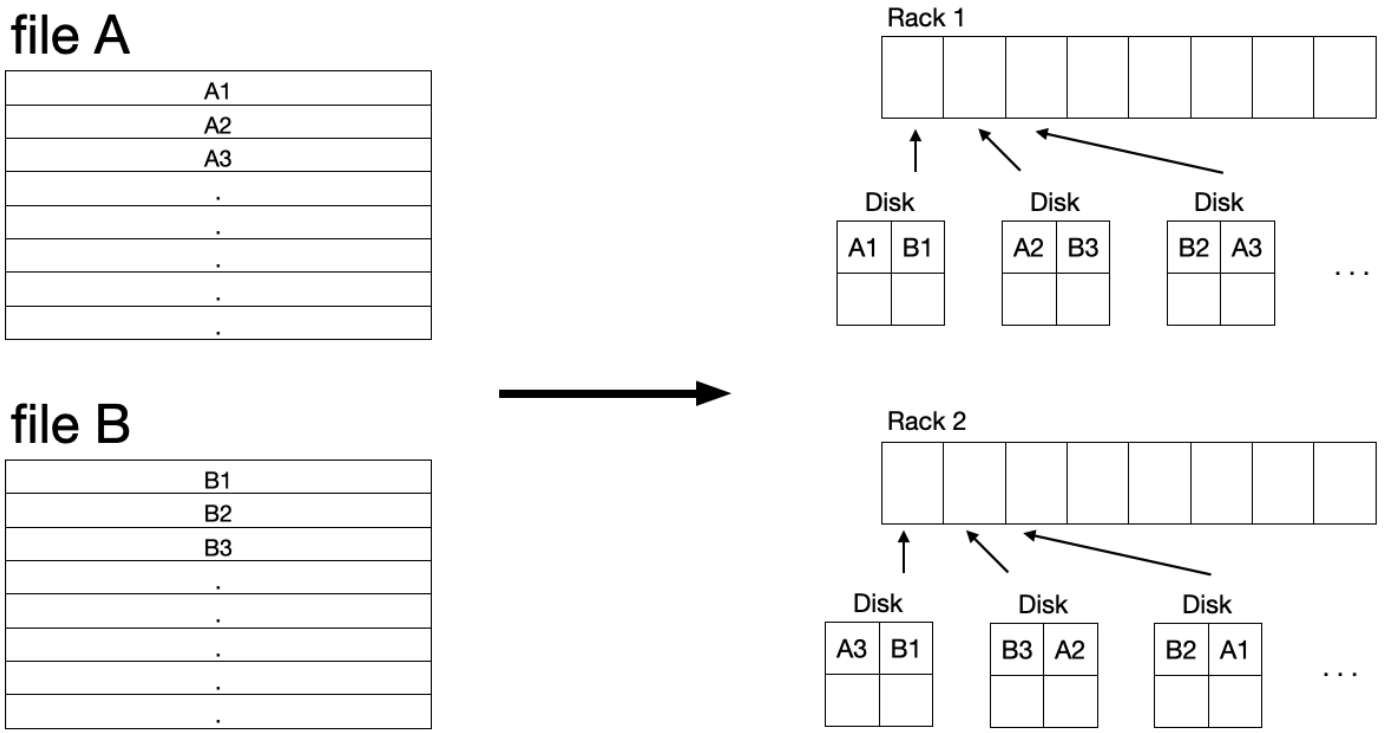


Distributed File System

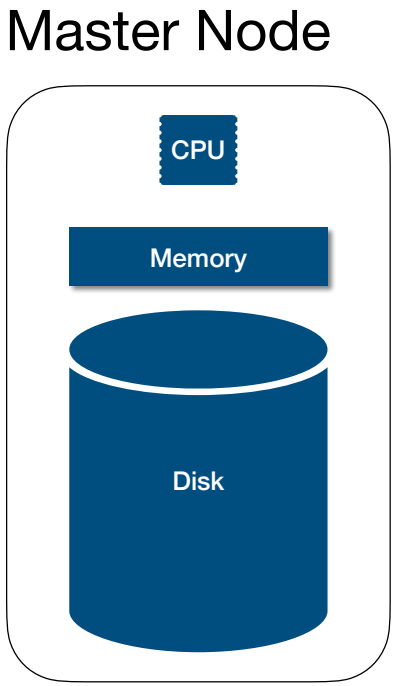
Distributed File System



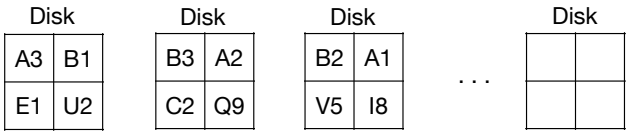
Distributed File System



Distributed File System



Distributed File System

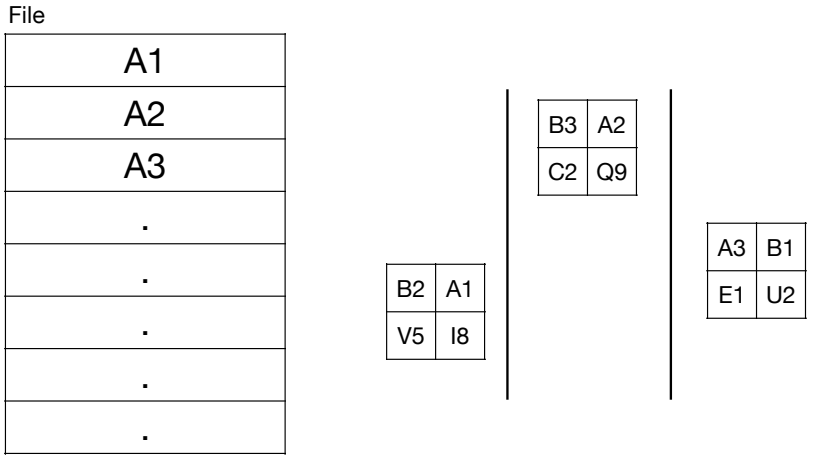


MapReduce

MapReduce

Jisung Jeong / jisung0920@gmail.com

Cluster



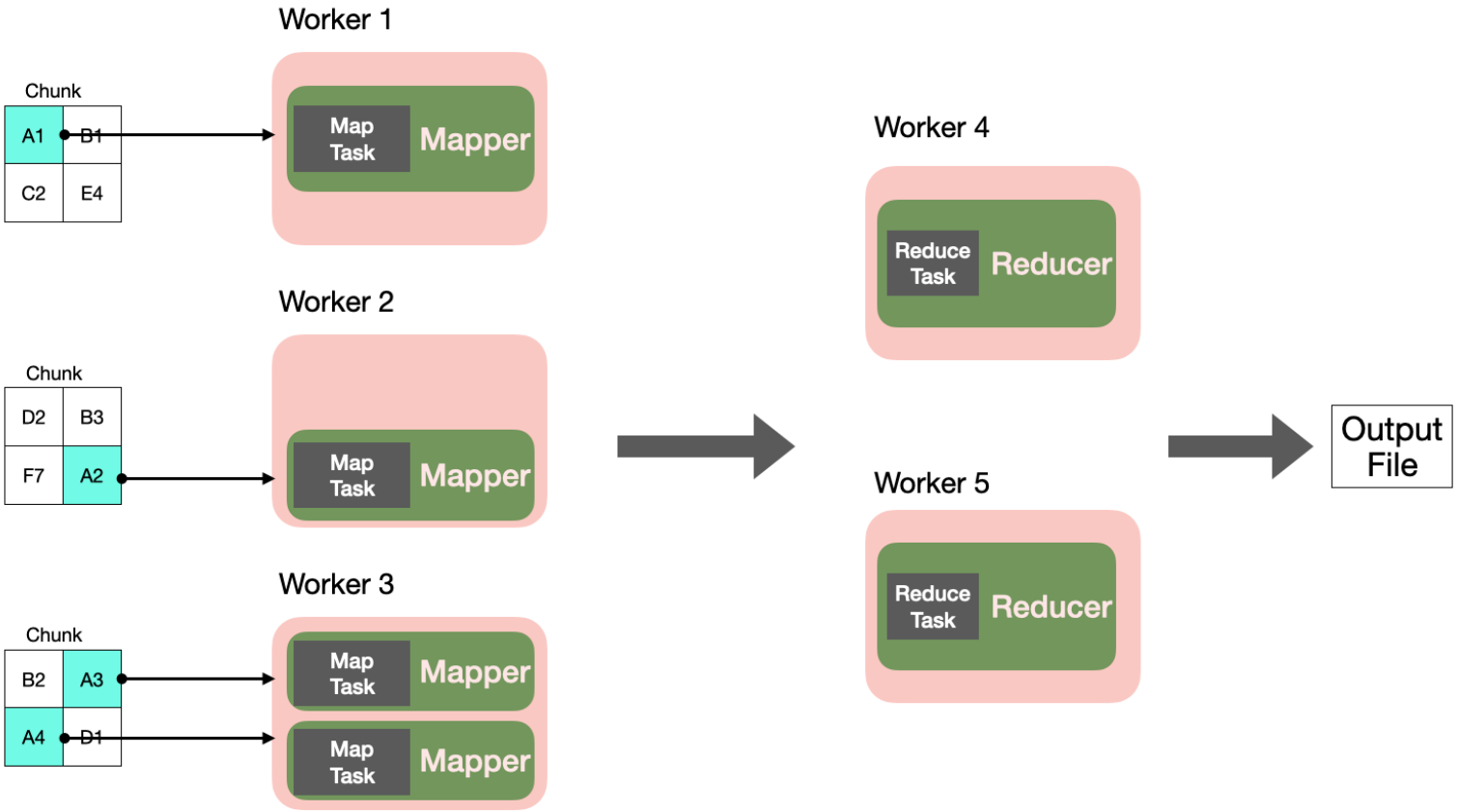
DFS

1. Map

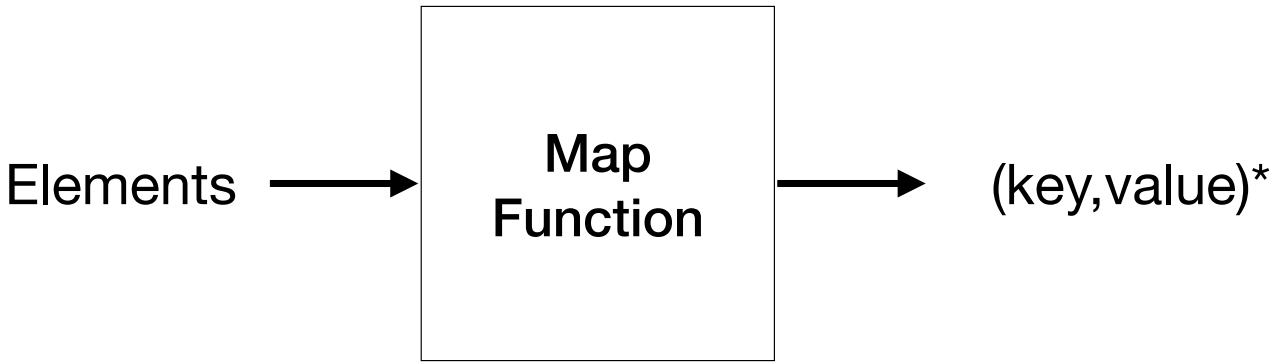
2. Group by Key

3. Reduce

- Worker
- Task
- Map Task
- Reduce Task
- Mapper
- Reducer


Worker 1
Chunk
A1
B1
C2
E4
Map Task
Mapper
Worker 2
Chunk
D2
B3
F7
A2
Map Task
Mapper
Worker 3
Chunk
B2
A3
A4
D1
Map Task
Mapper
Map Task
Mapper
Worker 4
Reduce Task
Reducer
Worker 5
Reduce Task
Reducer
Output File

Map

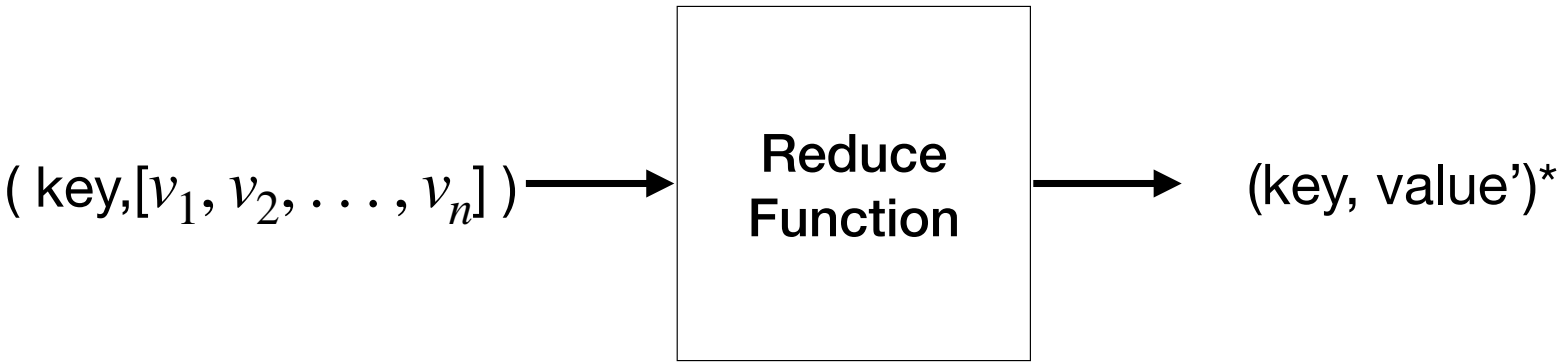


Group by key

$$(\ k_1,\ v_a),\ (\ k_2,\ v_a),\ (\ k_3,\ v_b),\ (\ k_1,\ v_b),\ (\ k_1,\ v_c)\ \ldots$$

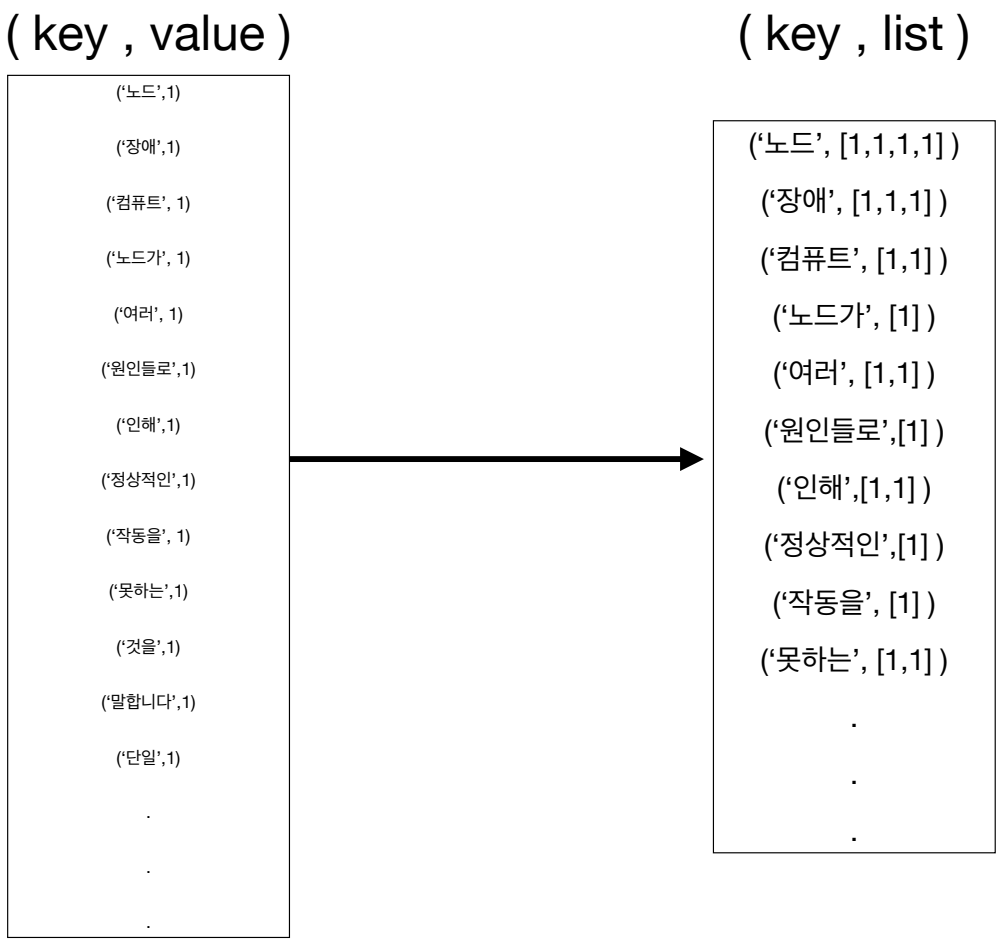
$$(\ k_1,\ [v_a, v_b, v_c]\ )\ ,\ (\ k_2,\ [v_a]\ ), \ldots$$

Reduce

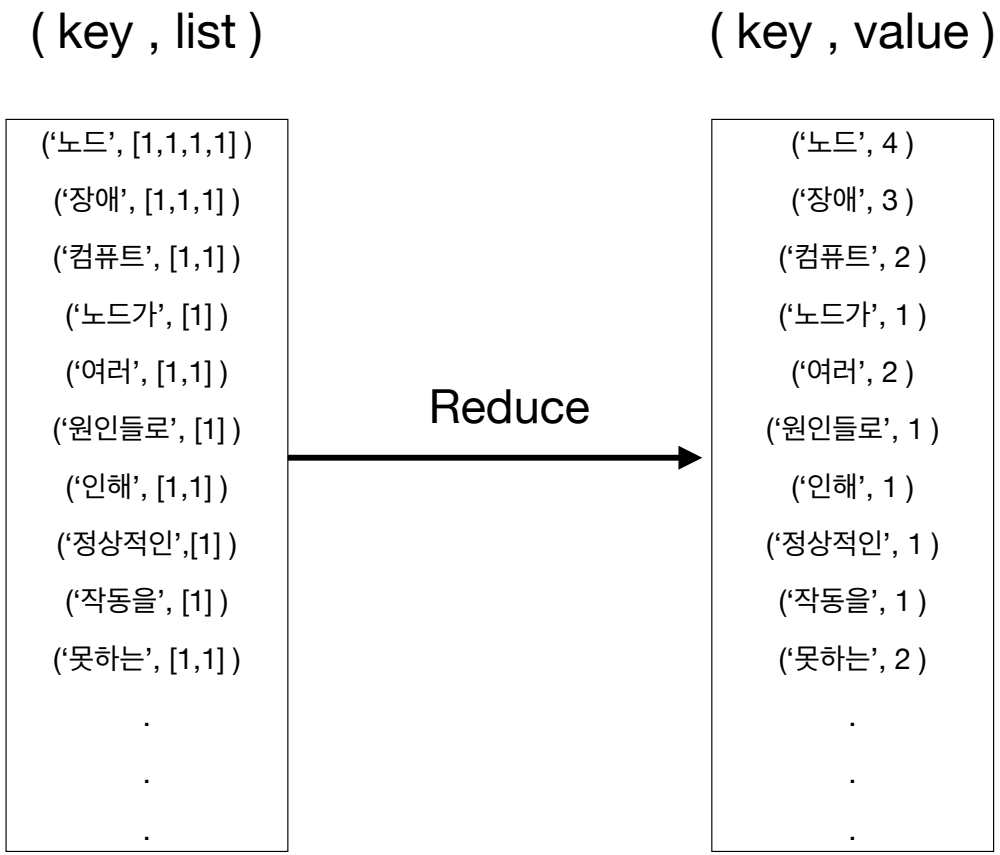


Example - Word Counter

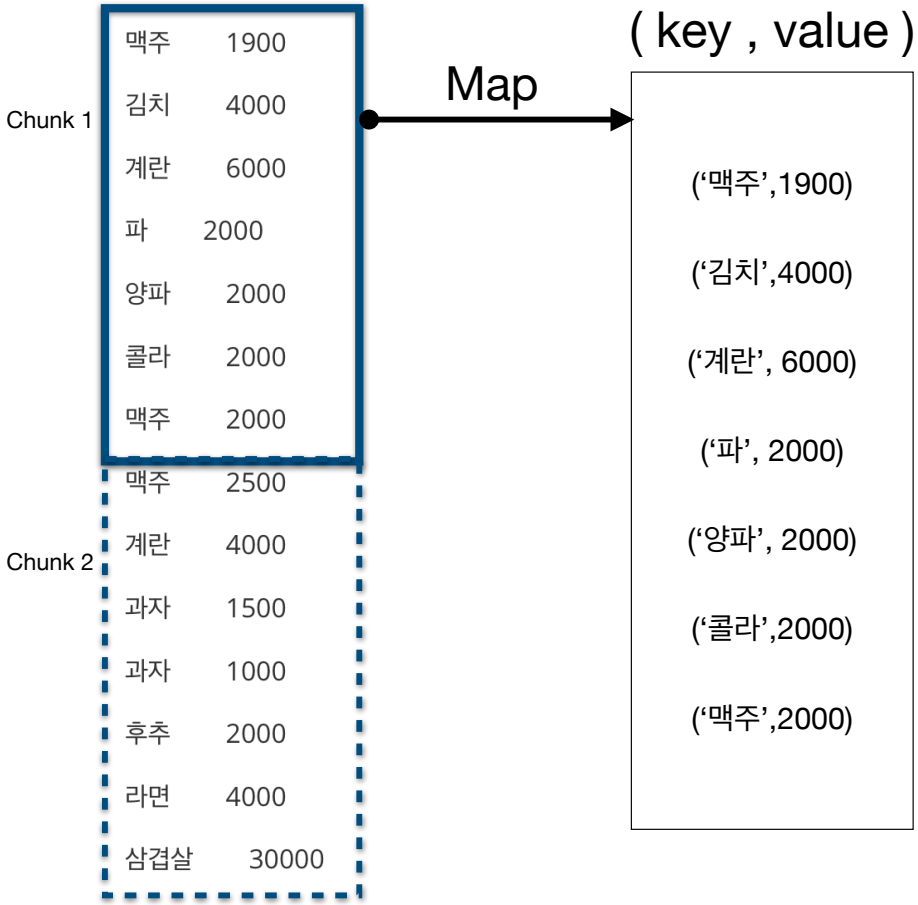
Example - Word Counter



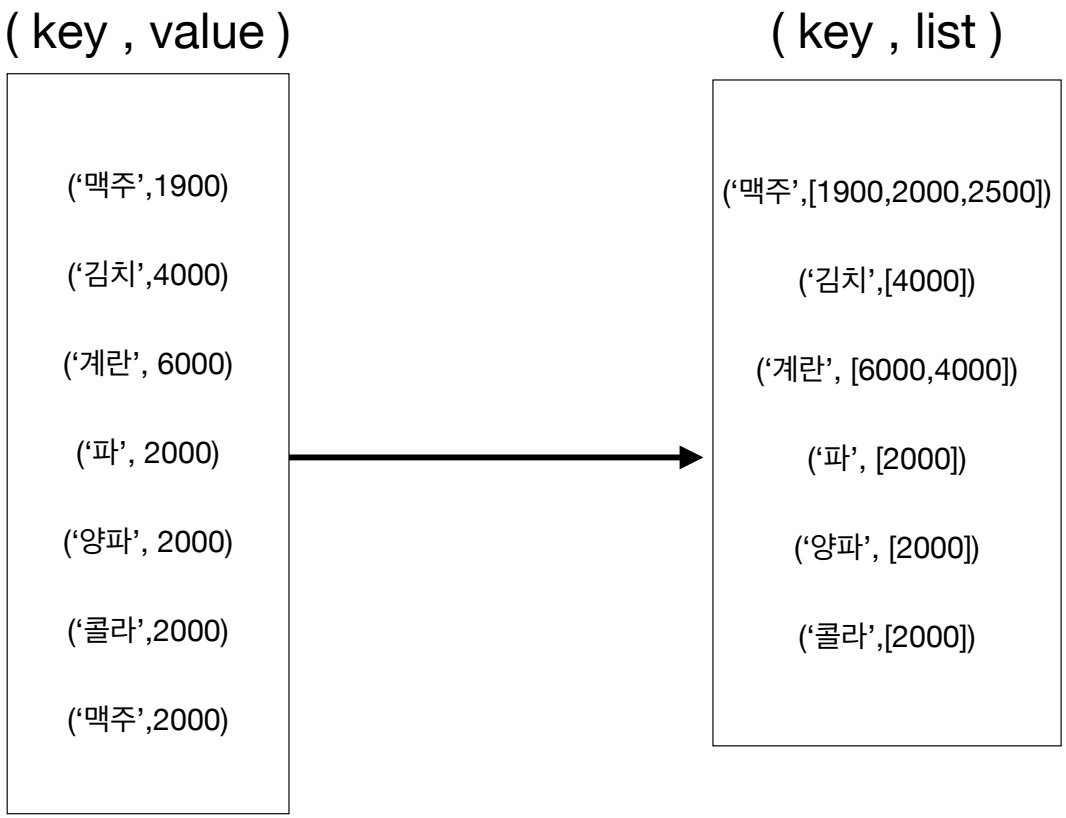
Example - Word Counter



Example - Average Price



Example - Average Price



Example - Average Price

MapReduce

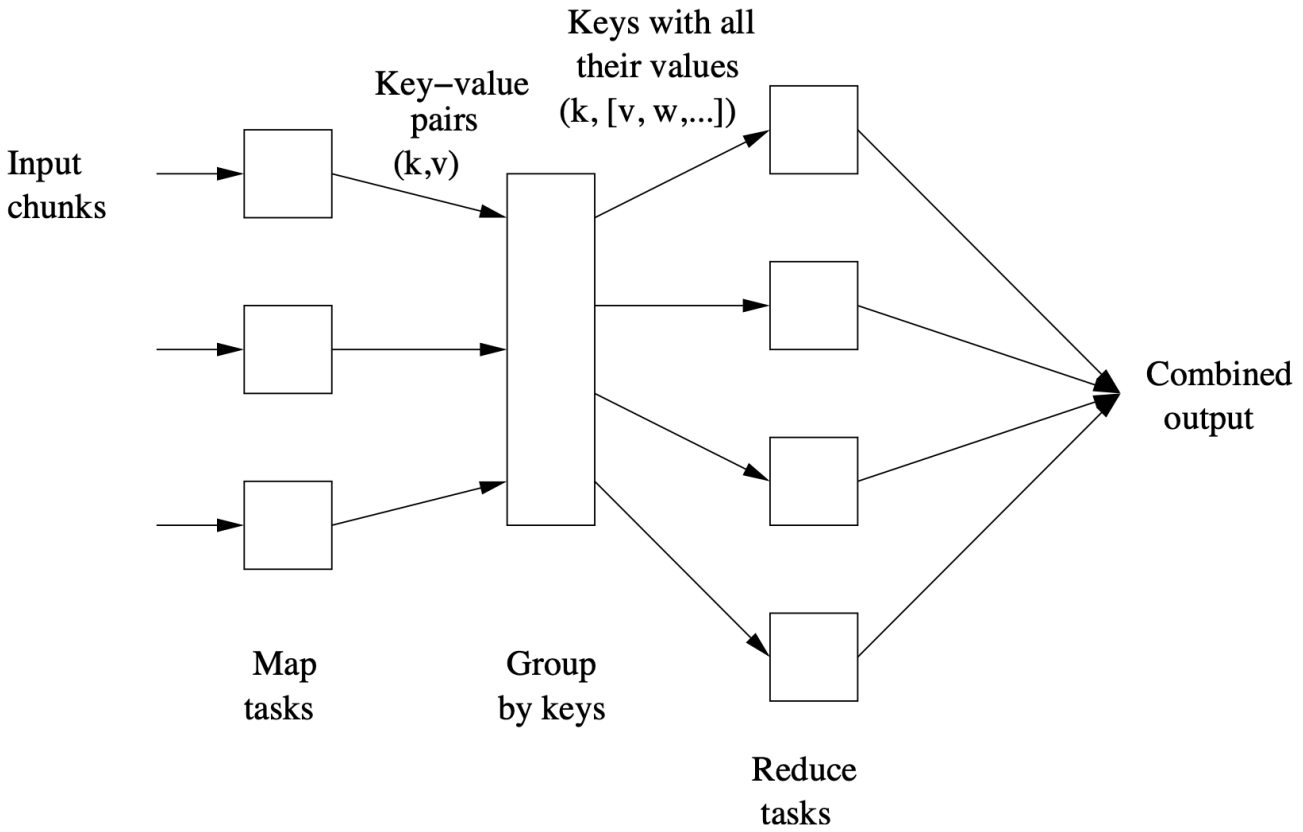


Figure 2.2: Schematic of a MapReduce computation

Mining of Massive Datasets - Figure 2.2

MapReduce

Advanced

Jisung Jeong / jisung0920@gmail.com

MapReduce

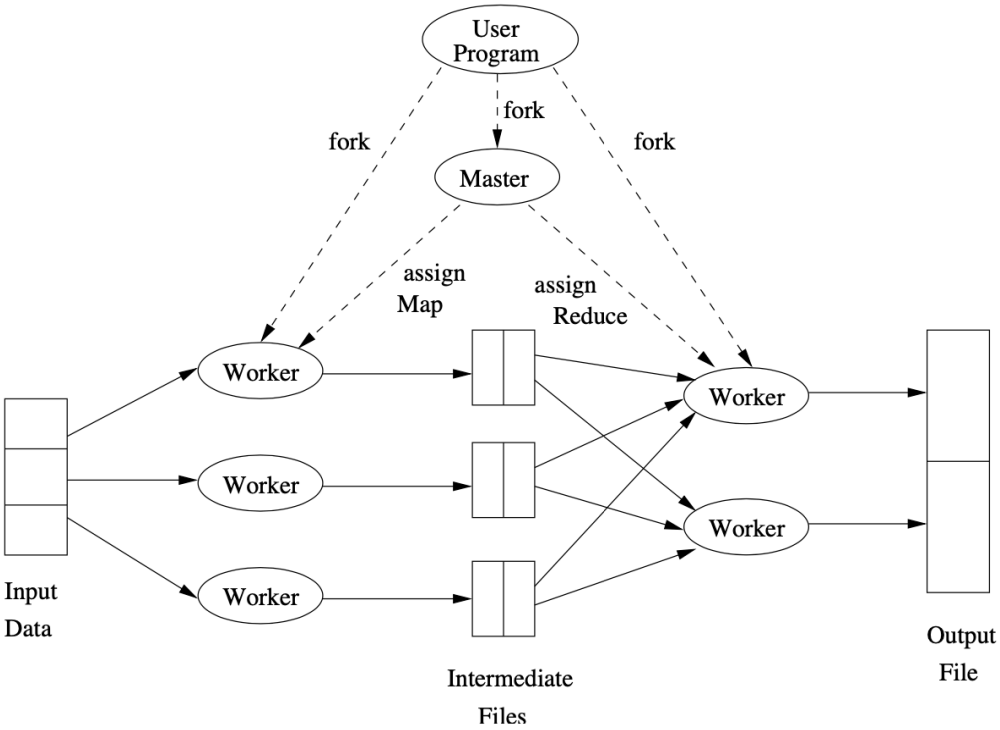


Figure 2.3: Overview of the execution of a MapReduce program

Mining of Massive Datasets - Figure 2.3

Node Failure

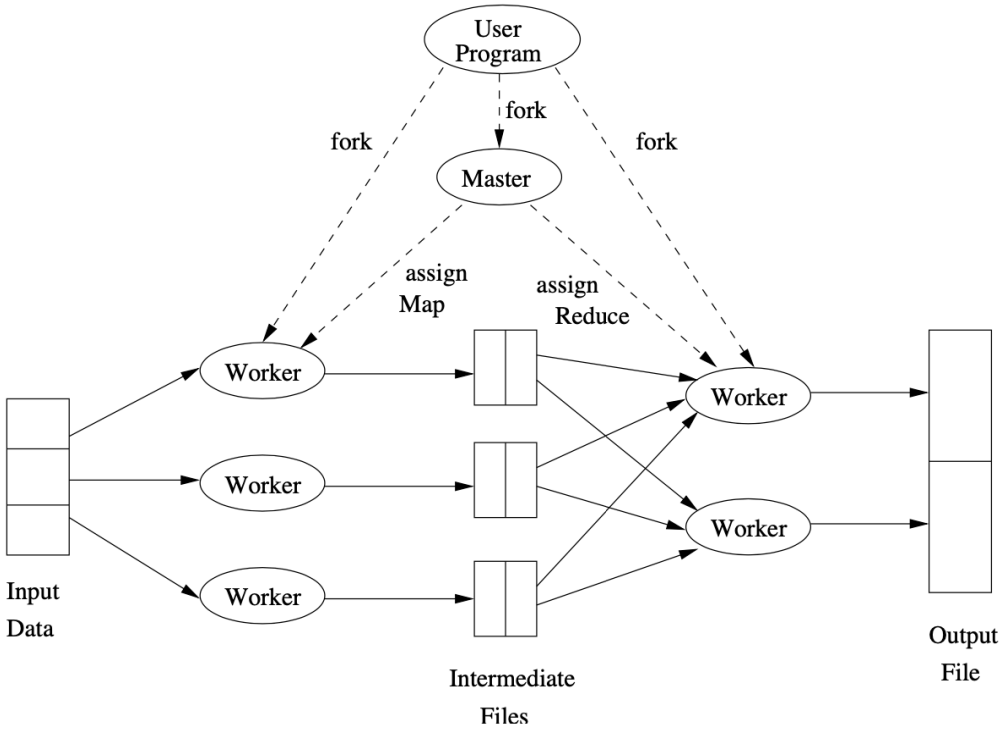
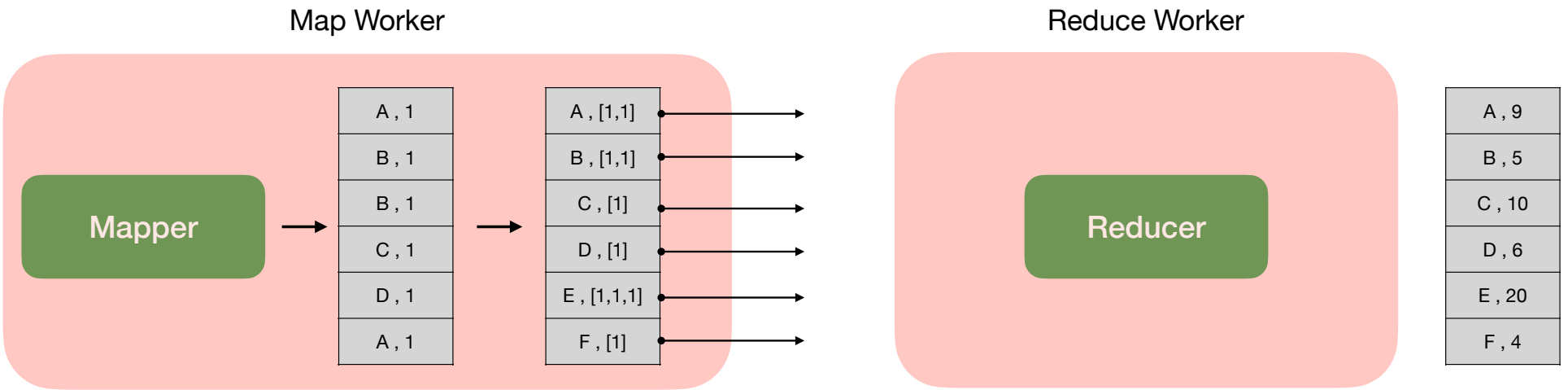


Figure 2.3: Overview of the execution of a MapReduce program

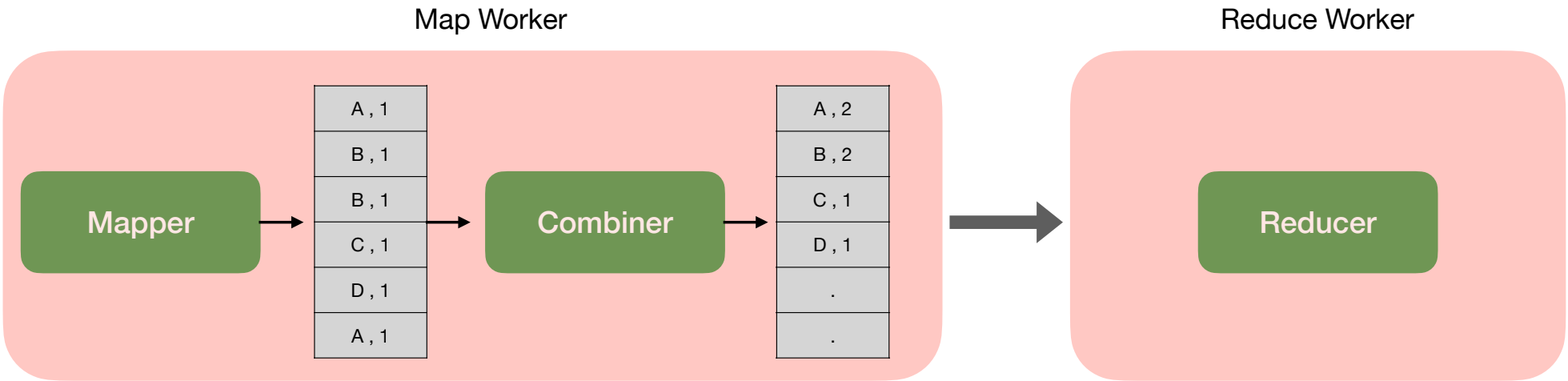
Mining of Massive Datasets - Figure 2.3

- Master
- Map Worker
- Reduce Worker

Combiner


Map Worker
Mapper
A , 1
B , 1
B , 1
C , 1
D , 1
A , 1
A , [1,1]
B , [1,1]
C , [1]
D , [1]
E , [1,1,1]
F , [1]
Reduce Worker
Reducer
A , 9
B , 5
C , 10
D , 6
E , 20
F , 4

Combiner


Map Worker
Reduce Worker
Mapper
A , 1
B , 1
B , 1
C , 1
D , 1
A , 1
Combiner
A , 2
B , 2
C , 1
D , 1
.
.
Reducer

Partition Function

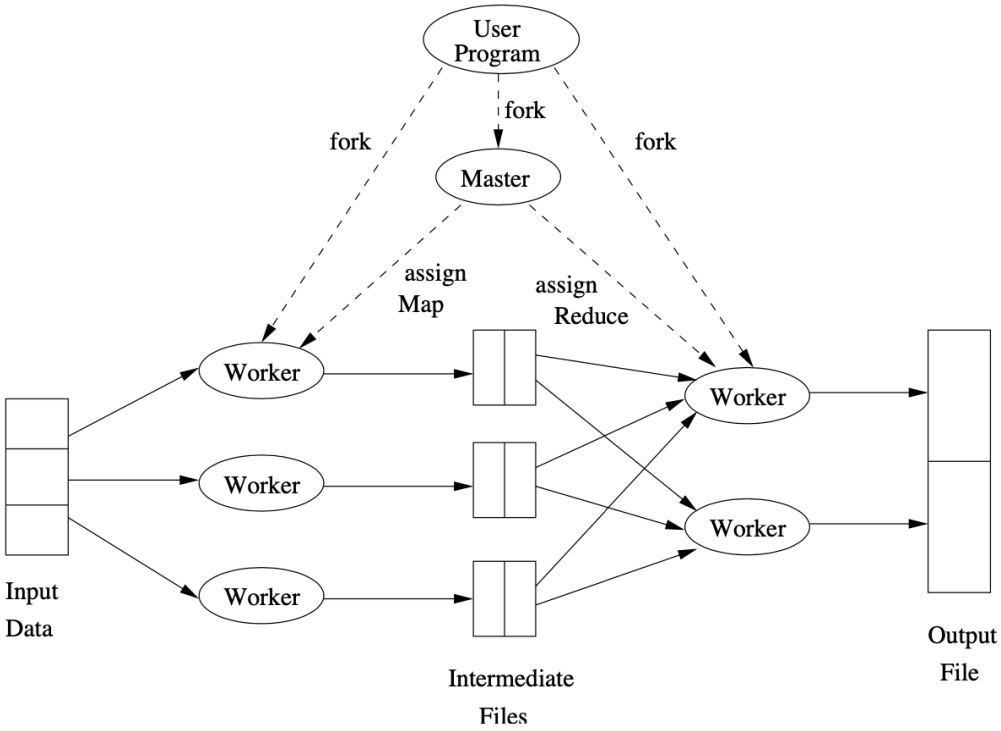


Figure 2.3: Overview of the execution of a MapReduce program

Mining of Massive Datasets - Figure 2.3

Partition Function