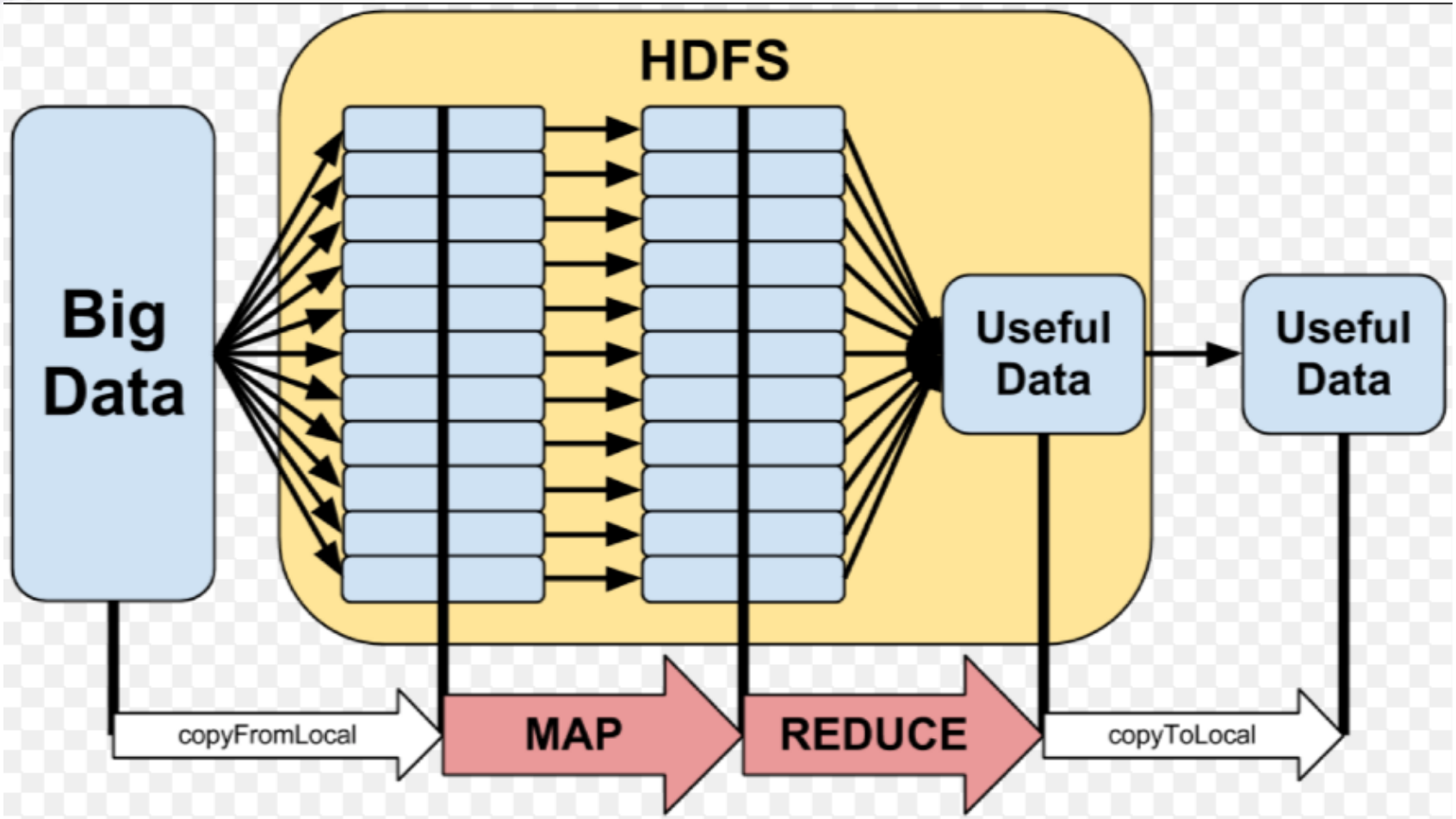
맵리듀스

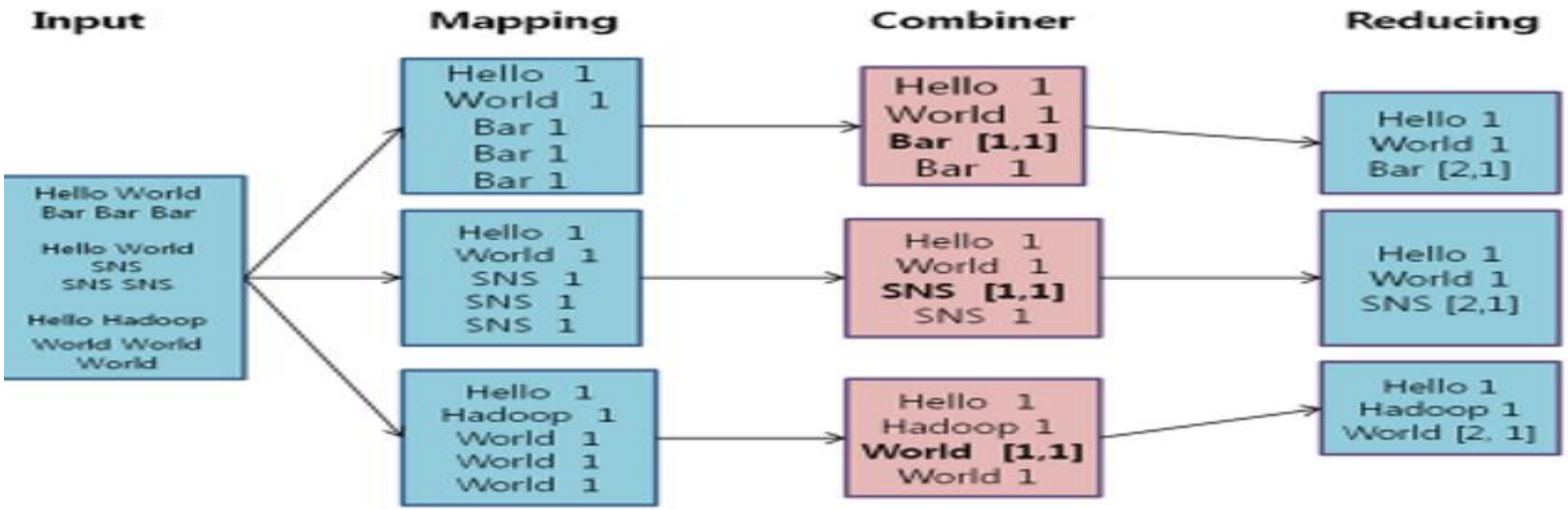
Contents

맵리듀스 처리 과정

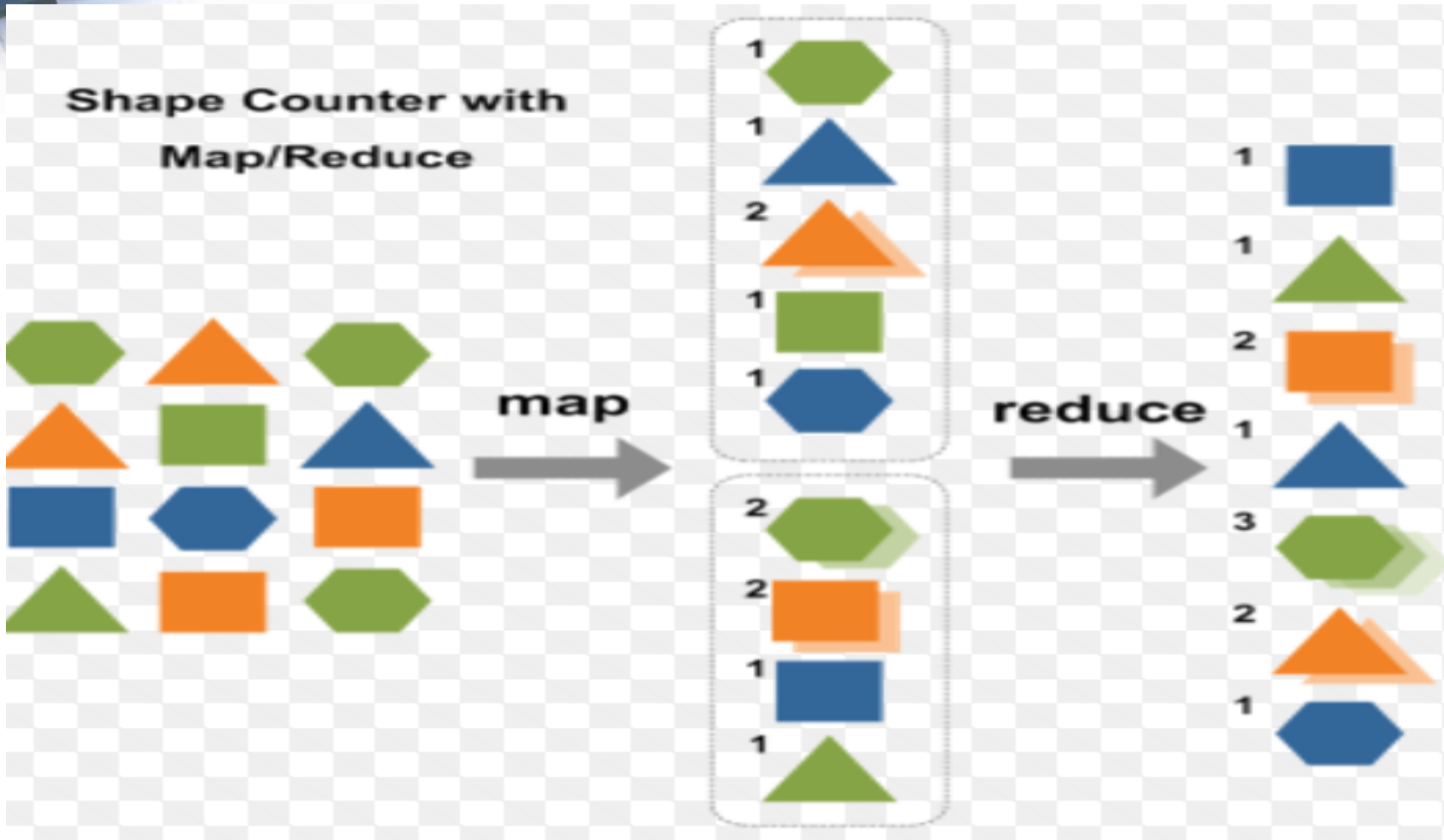

HDFS
Big Data
Useful Data
Useful Data
copyFromLocal
MAP
REDUCE
copyToLocal

맵리듀스란?

- 맵 : 입력 파일 한 줄씩 읽어 데이터 변형
  - **(k1, v1) → list(k2, v2)**
- 리듀스 : 맵의 결과 데이터 집계
  - **(k2, list(v2)) → (k3, list(v3))**



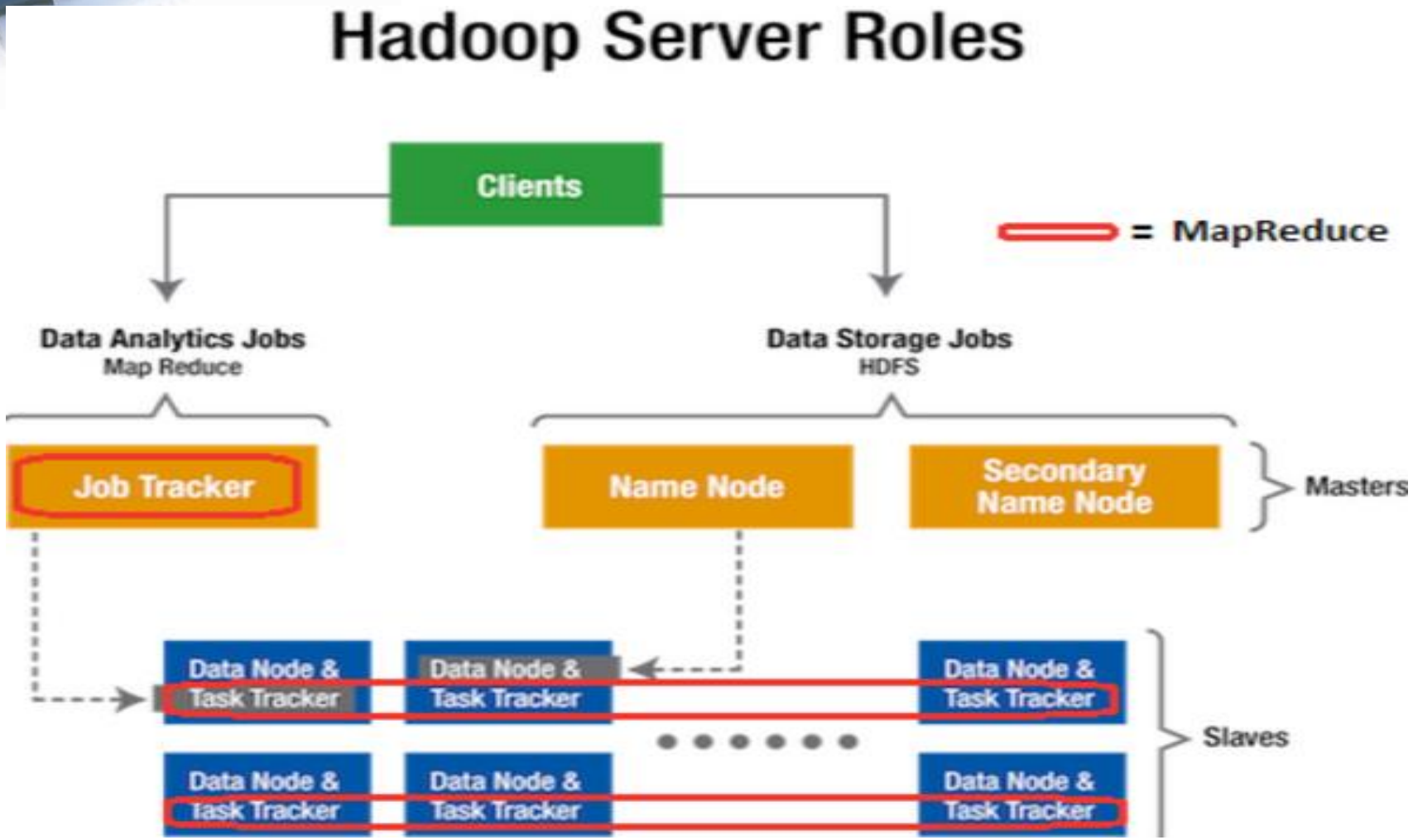
맵리듀스란?



맵리듀스 아키텍쳐

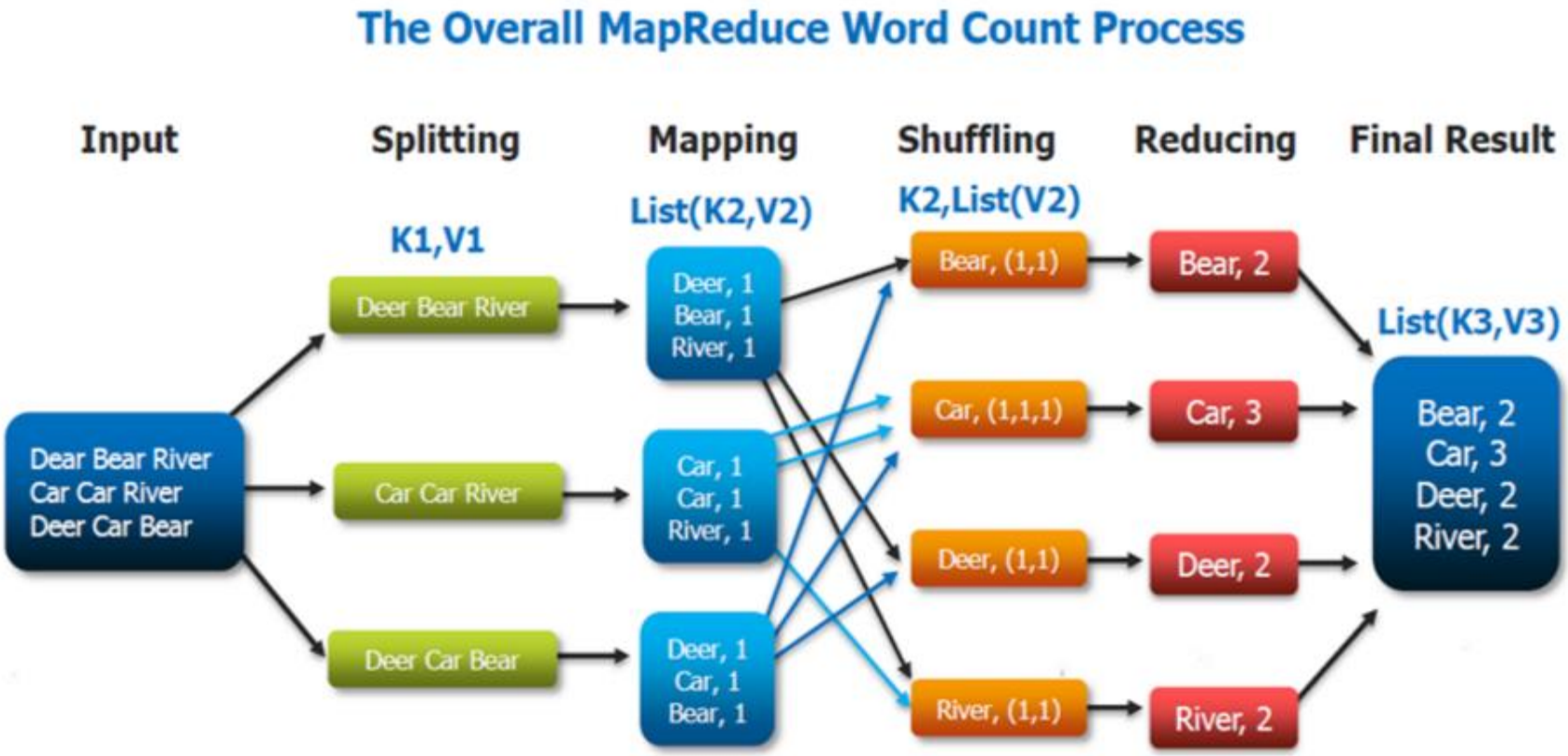
- 핵심 전략
  - 개발자 : 분석 로직 구현
  - 프레임웍 : 데이터분산과 병렬 처리
- 시스템 구성
  - 클라이언트 : 맵리듀스 **API** (주문자)
  - 잡트랙커 : **job** 이라는 하나의 작업단위의 스케쥴링 관리 및 모니터링 (홀서빙자)
  - 태스크트래커 : 맵리듀스 실행, 데이터노드 실행 (주방장-주방보조[**task**])

아키텍쳐


Hadoop Server Roles
Clients
= MapReduce
Data Analytics Jobs
Map Reduce
Data Storage Jobs
HDFS
Job Tracker
Name Node
Secondary Name Node
Masters
Data Node & Task Tracker
Data Node & Task Tracker
Data Node & Task Tracker
Data Node & Task Tracker
Data Node & Task Tracker
Data Node & Task Tracker
Slaves

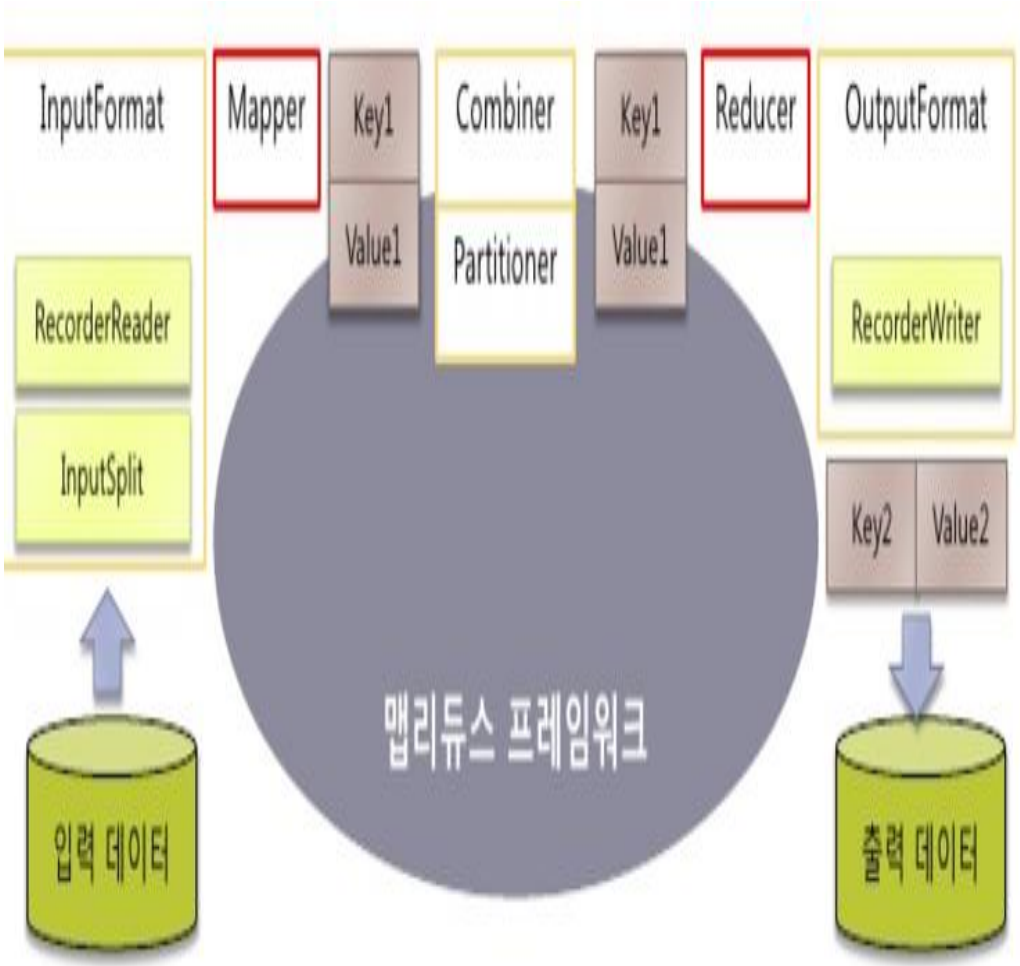
데이터 플로우

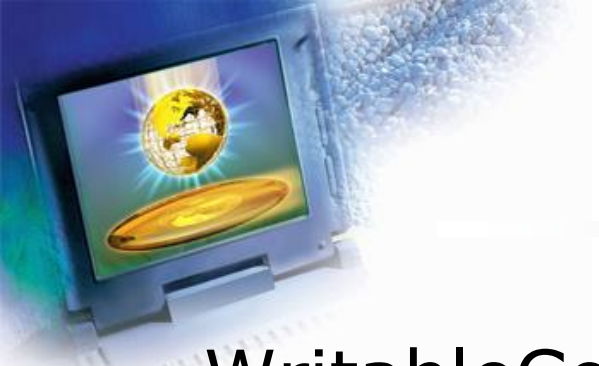
- 맵단계 ➔ 셔블 단계 ➔ 리듀스 단계



맵리듀스 프로그래밍 요소

- 데이터 타입
- InputFormat
- Mapper
- Partitioner
- Reducer
- Combiner class
- OutputFormat





데이터 타입

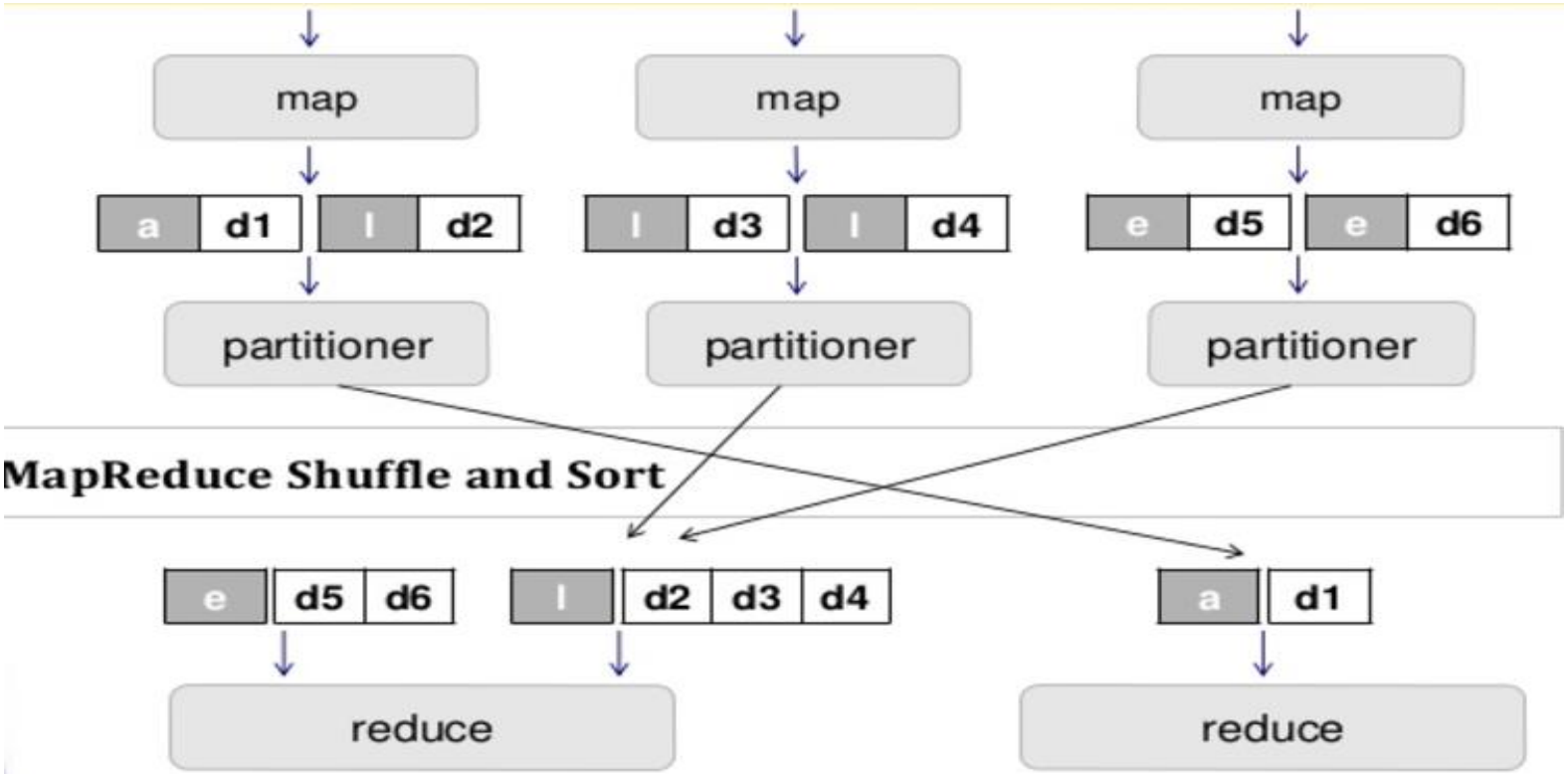
- WritableComparable

Implement 3 methods

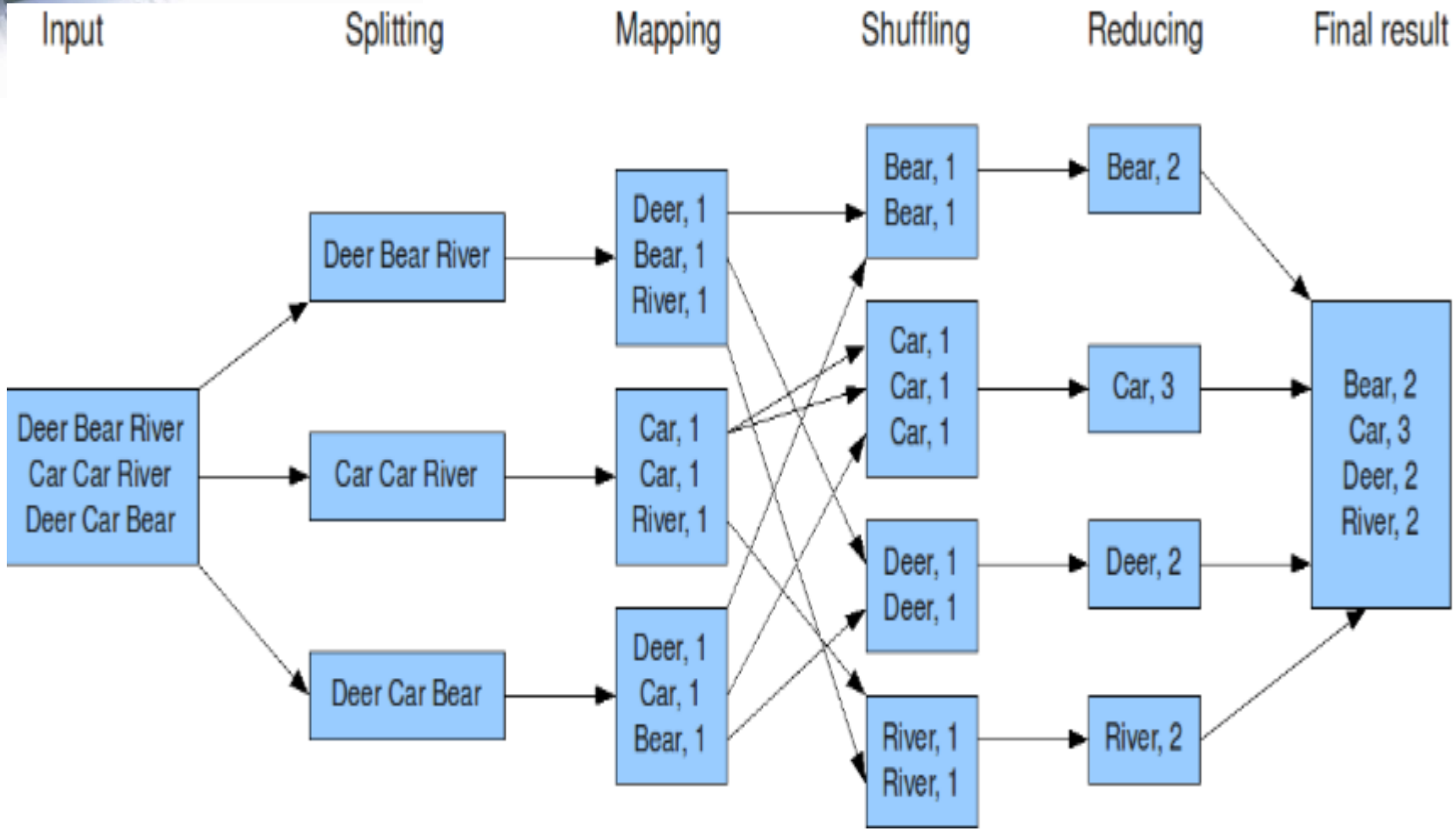
- write(DataOutput)
  - Serialize your attributes
- readFields(DataInput)
  - De-Serialize your attributes
- compareTo(T)
  - Identify how to order your objects
  - If your custom object is used as the key it will be sorted prior to reduce phase

파티셔너

- 출력 데이터가 어떤 리듀스 태스크로 전달될지 결정 (HashPatitioner)

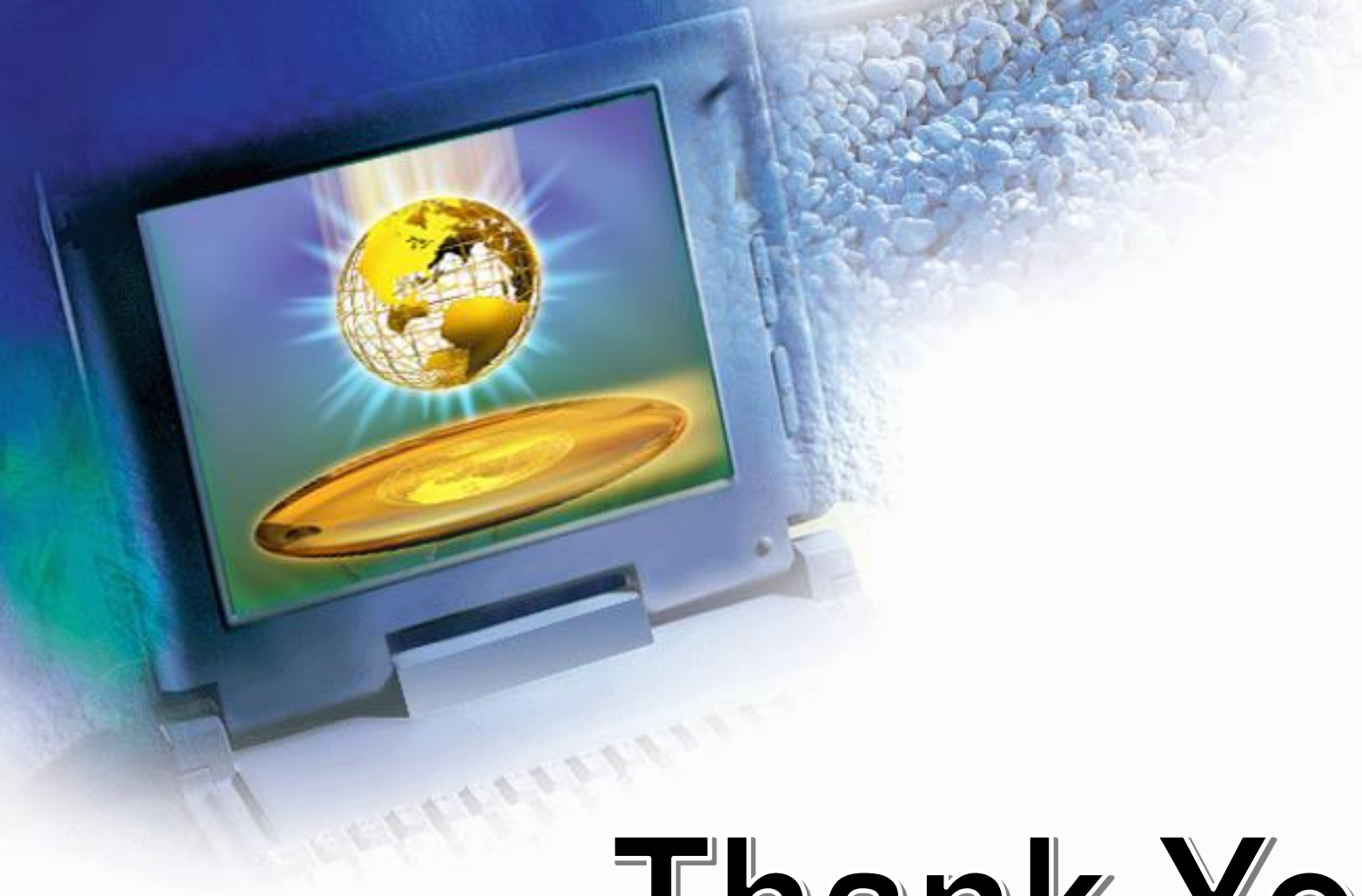


맵리듀스 프로그래밍(word count)



구현 프로세스

- 매퍼 구현
- 리듀서 구현
- 드라이버 클래스 구현
- wordcount 빌드
- wordcount 실행
- 웹에서 실행 결과 확인



Thank You !