7. (구글) 맵리듀스(MapReduce)

구글 맵리듀스

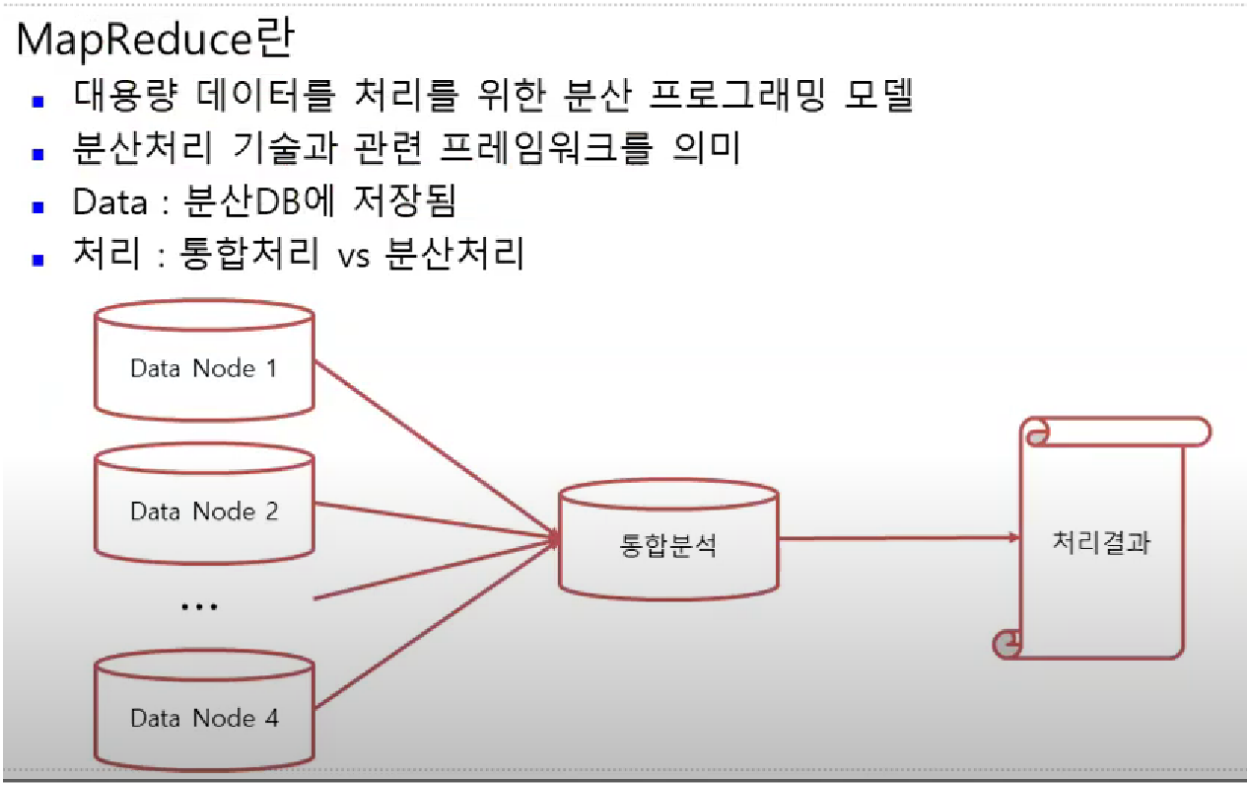
: GFS의 과정에서 맵핑하는 과정이 추가

0. 개발배경
   : 구글 검색(검색언어 추출, 정렬, 인덱스 생성 등)을 위해 개발된 분산환경 병렬 데이터 처리 기법

- 비공유 구조의 여러 노드 PC로 대량의 병렬처리 가능
- 데이터는 키-밸류(key, value)의 쌍으로 존재

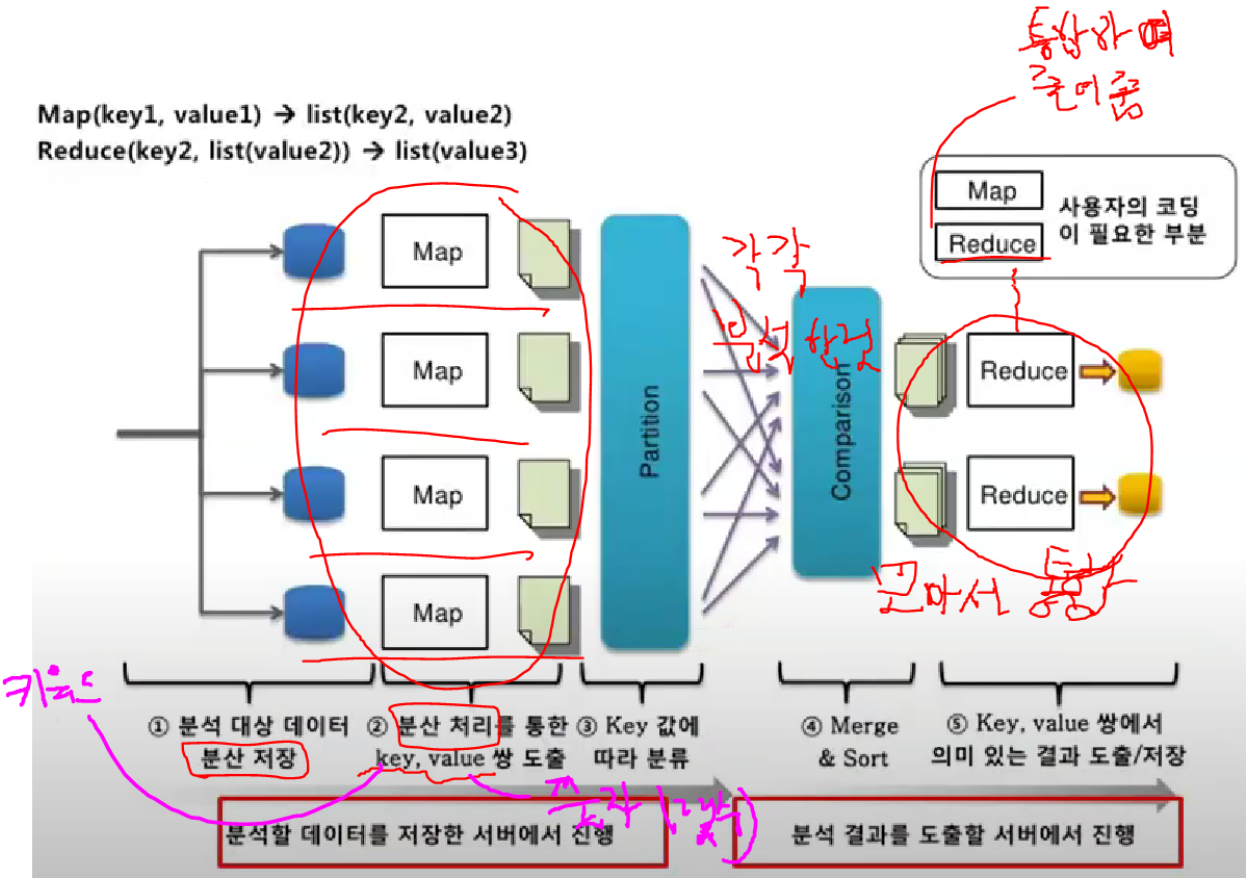
---

0. 맵리듀스(MapReduce)란 =〉 일도 나눠서 처리하자

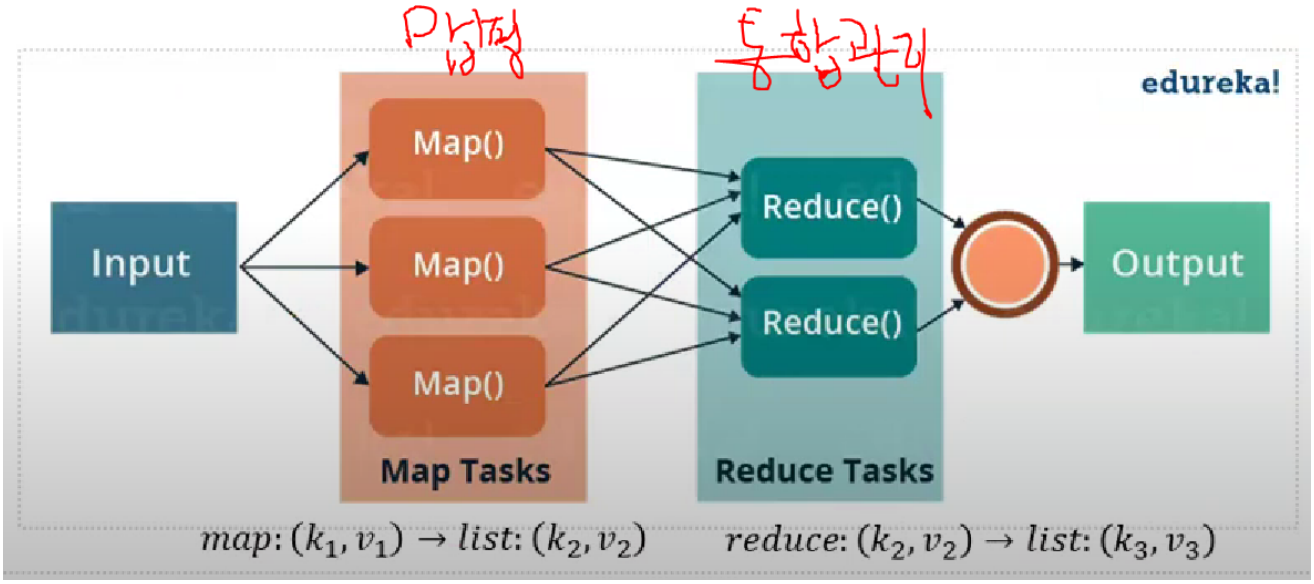


- 데이터를 분산해서 저장한 상태 -〉 키워드가 입력됨 -〉 각 데이터 노드에 있는 데이터들을 갖고와서, 통채로 묶어서 '통합분석'에서 한꺼번에 분석 및 처리
  =〉 많이 저장할 수 있는 분산 저장의 의미는 있지만, 처리할 때마다 다시 갖고 올 것 같으면 뭣하러 나눴음? (분산 저장을 했으면 분산 처리를 해야 유의미하지)
- 즉, '분산 저장'되어 갖고 있는 키워드에 대한 것의 / '분산 처리(분석)'한 결과를 취합하는 것

1. 맵리듀스 = 맵 + 리듀스



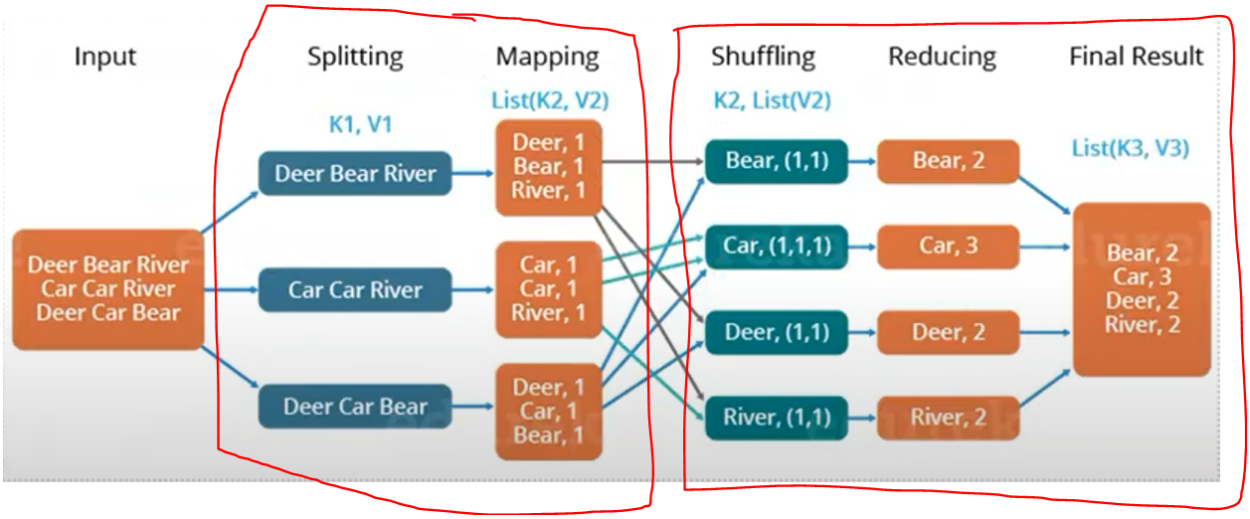
- Map : 흩어져 있는 데이터를 Key, Value의 형태로 연관성 있는 데이터 분류로 묶는 작업 (전체적으로 Key와 Value 값으로 묶어주는 것)
- Reduce : Filtering과 Sorting을 거쳐 데이터를 추출, Map화한 작업 중 중복 데이터를 제거하고, 원하는 데이터를 추출하는 작업 (묶어 준 것을 통합해주는 것)



2. 사례

- 입력과 출력으로써 Key-Value 쌍을 가짐

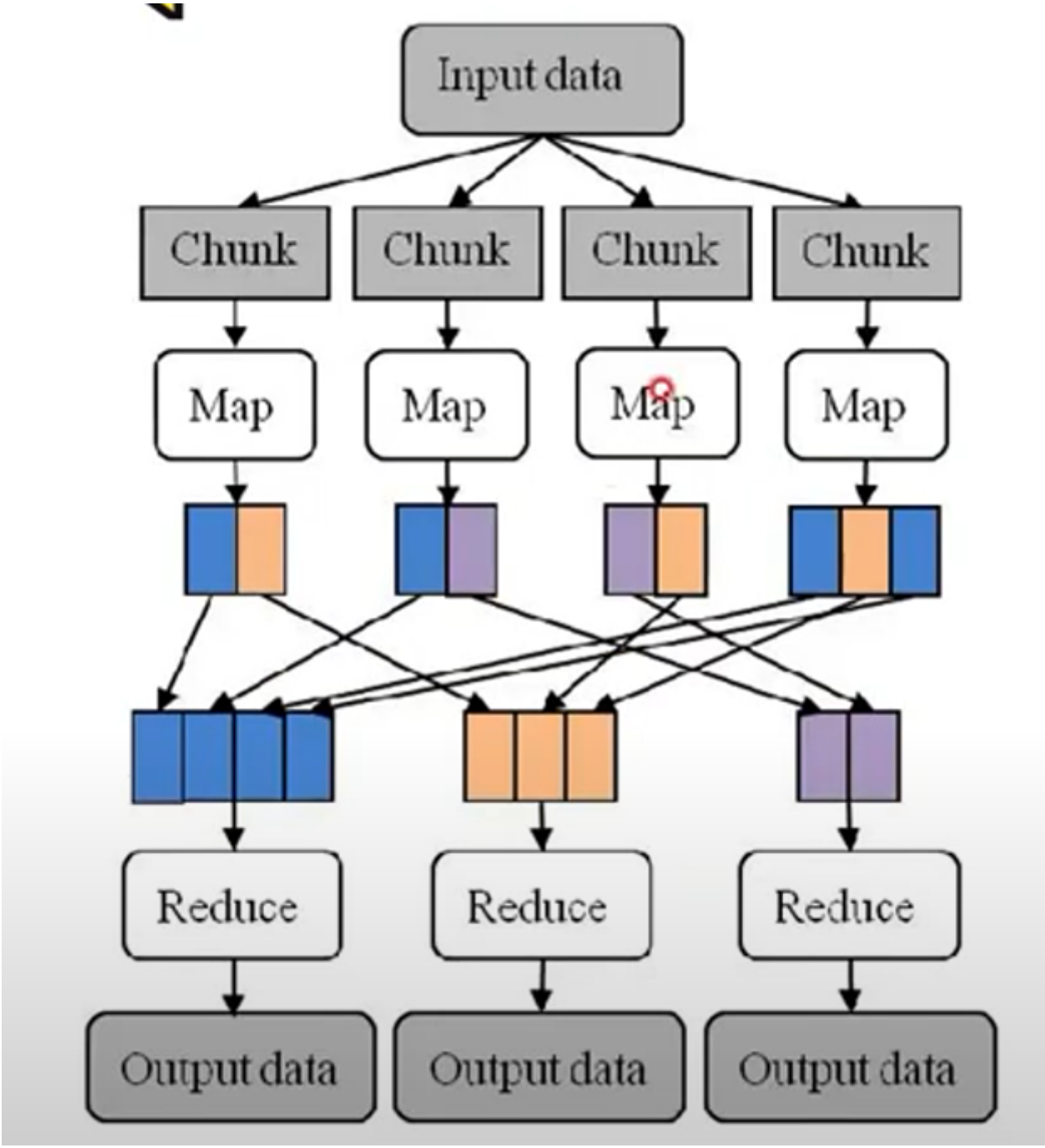
- JAVA로 직접 코딩 -〉 Pig, Hive



---

3. 개념 반복학습

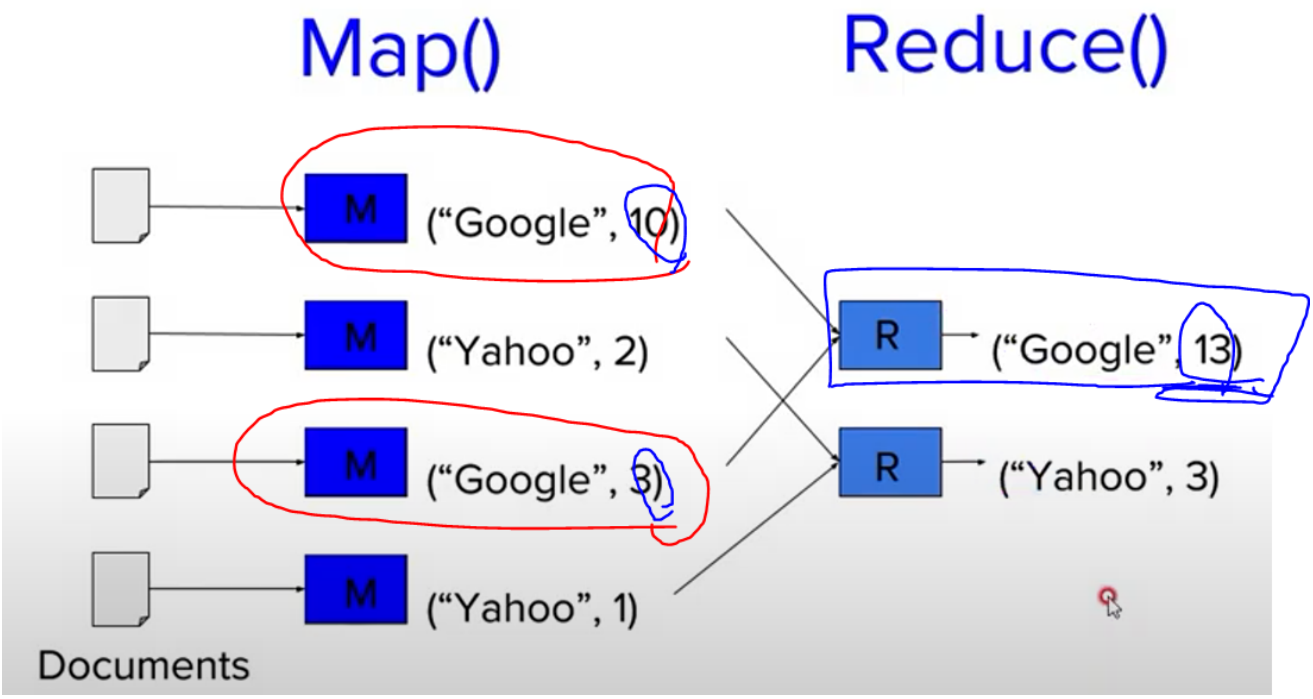
- Map : GFS에서 전달된 청크 단위(64MB)의 데이터를 (Key, Value) 형태의 파일들로 데이터 기록
- Reduce : Map과정에서 분할 및 정리된 (Key, Value) 데이터를 그룹화, 집계 후, GFS에 새로운 (Key, Value)로 저장

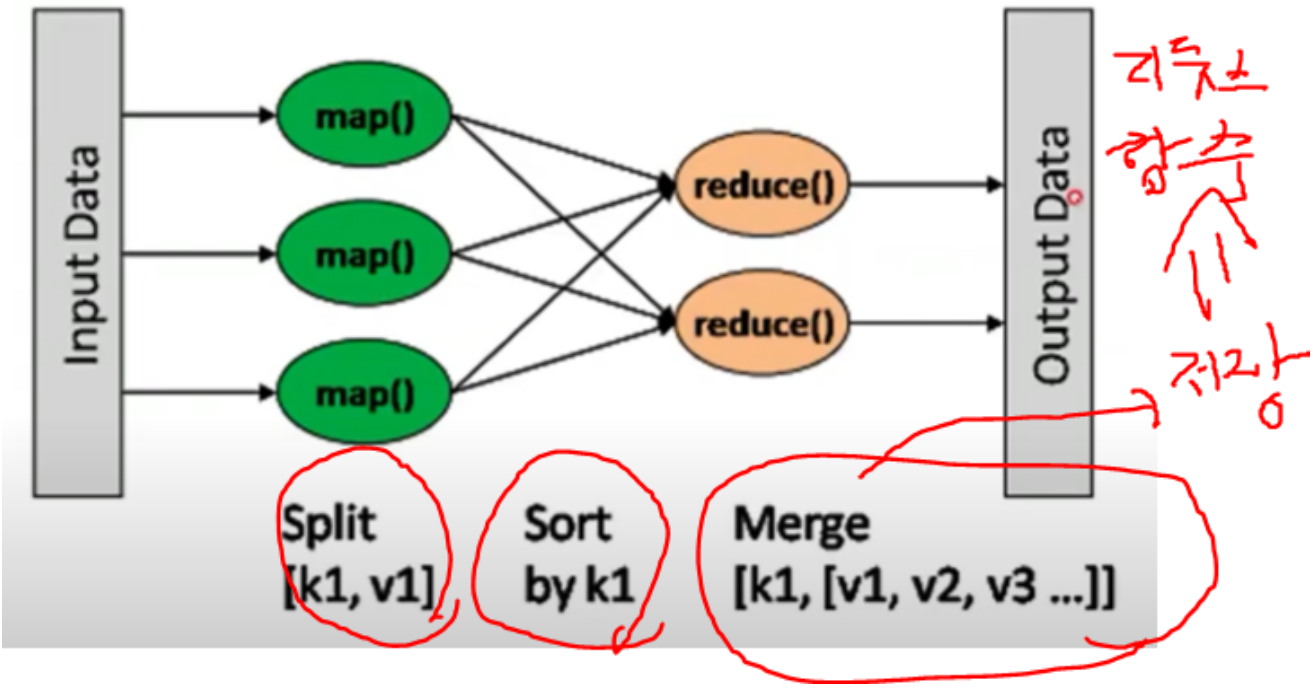


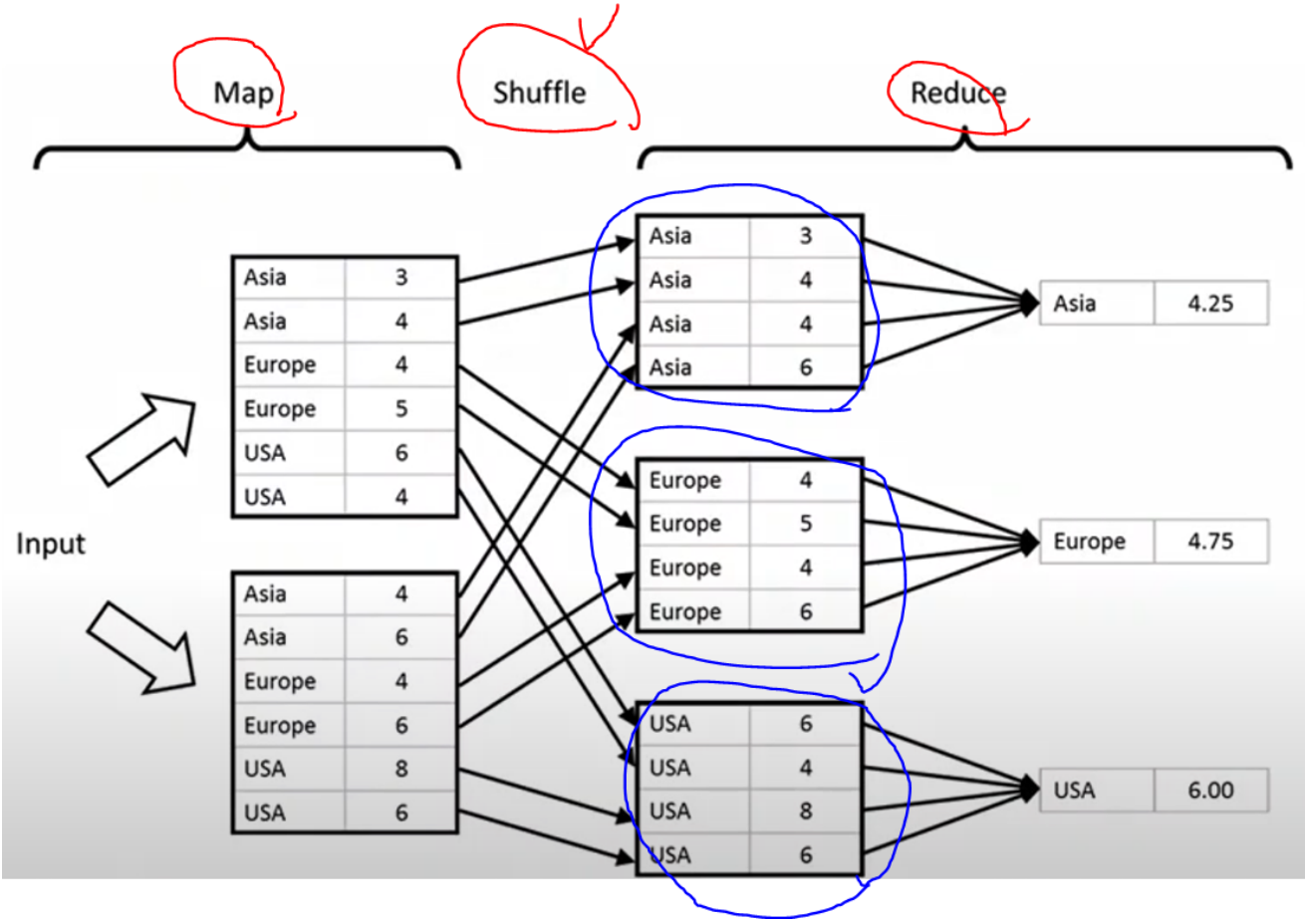
- 맵핑(Mapping)
  : 64MB 단위의 청크 파일을 다시 분류를 하여, Key와 Value로 나누는 과정

- 맵핑 이후 과정 : map -〉 (key, value) -〉 저장 -〉 같은/비슷한 Key값을 모아서 섞음 -〉 하나의 새로운 파일로 만듦(release)

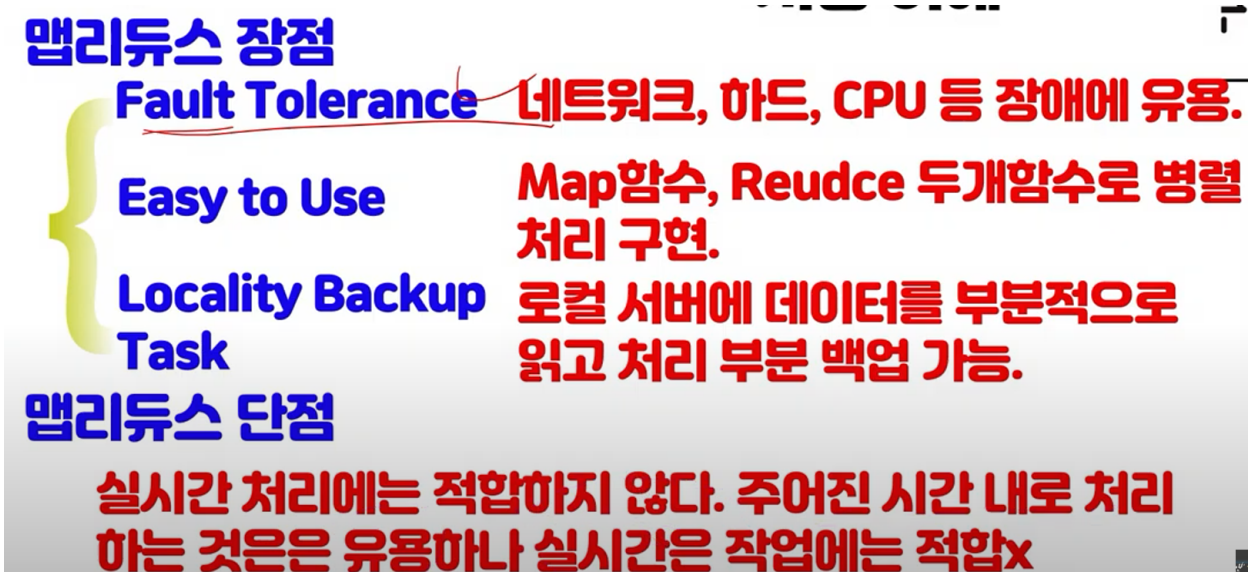
- 리듀스(Reduce)
  : 맵핑한 다음, 비슷한 Key와 Value끼리 모아서 데이터를 줄여나감
  (key, value)로 쪼갠다음 다시 합쳐나간다


Map()
Reduce()
M
("Google", 10)
M
("Yahoo", 2)
M
("Google", 3)
M
("Yahoo", 1)
R
("Google", 13)
R
("Yahoo", 3)
Documents


Input Data
map()
map()
map()
reduce()
reduce()
Output Data
Split
[k1, v1]
Sort
by k1
Merge
[k1, [v1, v2, v3 ...]]

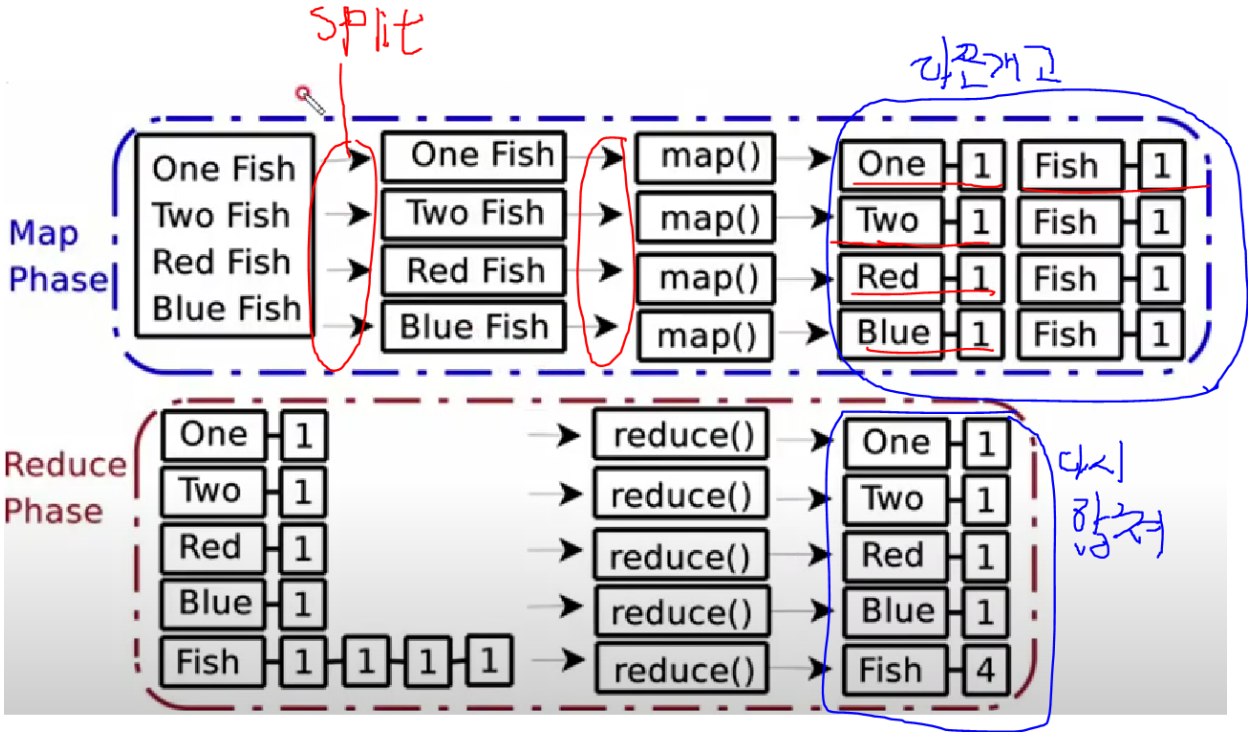


2. 맵리듀스의 장/단점



- 맵핑이 처리 된 후여야 Reduce가능 하므로
  --〉 맵핑이 잘못된 것이 생긴다면 수정도 해줘야하고, 순차적인 걸로만 가능

3. 과정



- Reducing된 데이터가 (Key와 Value)로 저장되는 것임