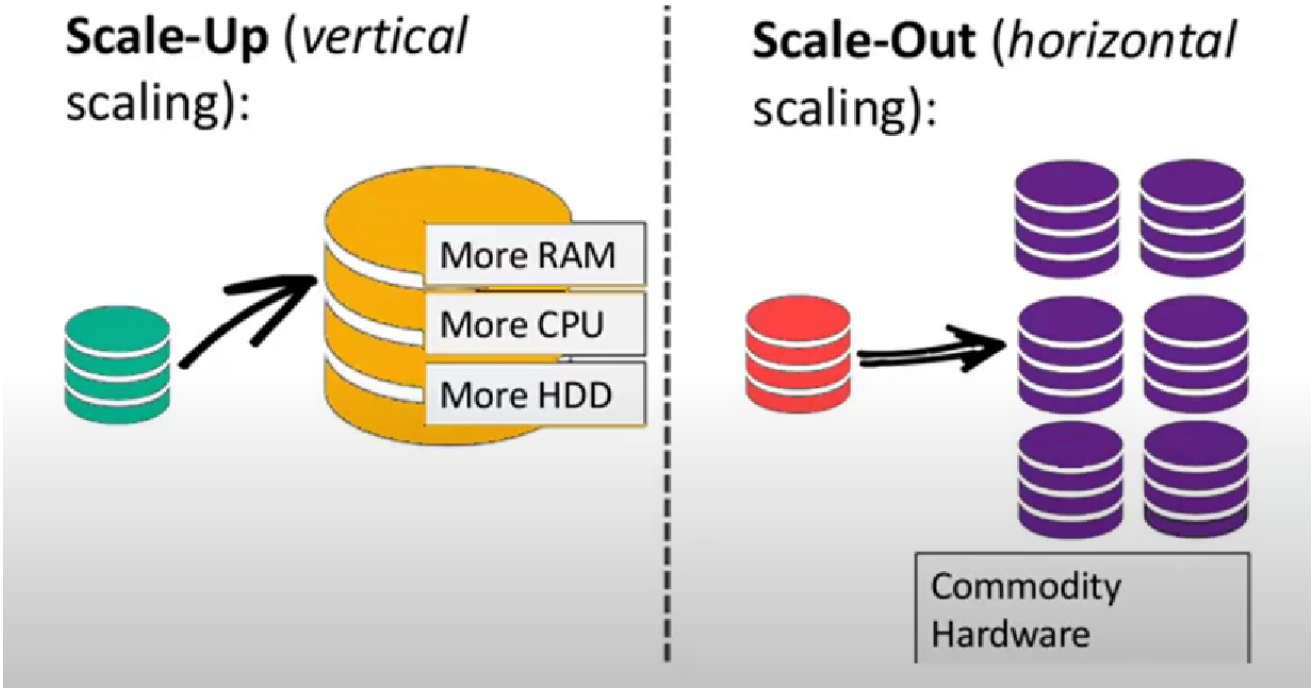
4. 분산 파일 시스템 (Distributed File System)

1. 분산 파일 시스템(DFS, Distributed File System)
   : 분산 된 서버에 파일을 저장하고, 저장된 데이터를 빠르게 처리할 수 있게 만든 시스템

- 분산 파일 시스템
  - 물리적으로 서로 다른 컴퓨터끼리 네트워크로 연결하여, 사용자에게 동일하게 보이는 파일 접근 공간을 제공해 주는 시스템
  - 블레이드 (blade) 서버: 프로세서가 장착된 회로판, 기억장치, 그리고 선반에 장착된 네트워크 접속부로 구성된 아주 얇은 컴퓨터

SCALE-UP or SCALE-OUT



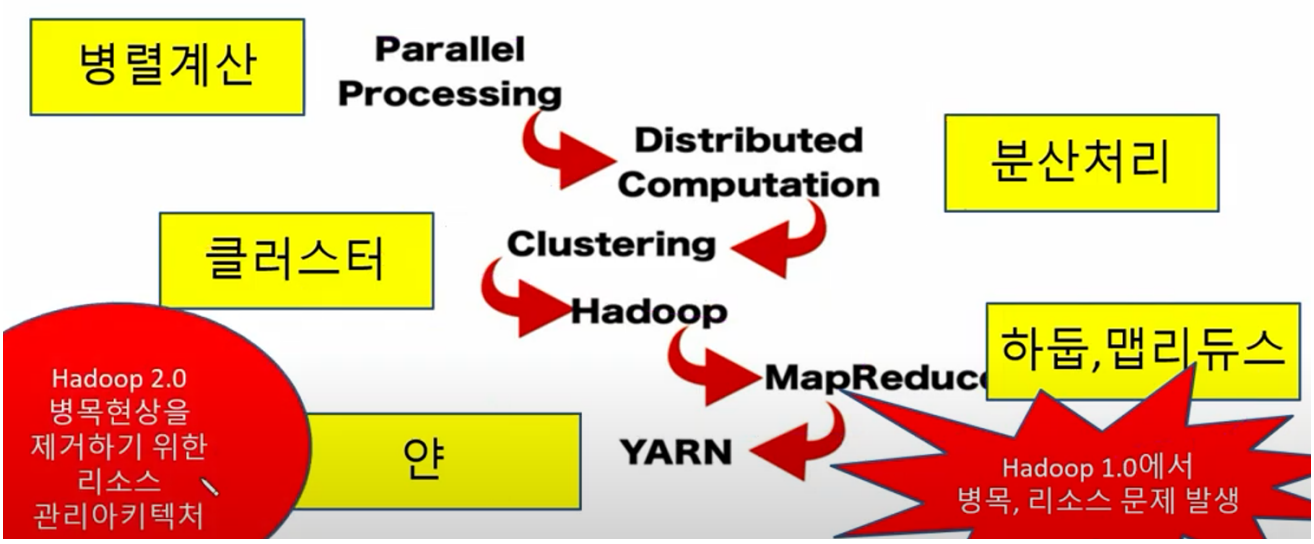
- Scale-Up (비용 부담)

- Scale-Out (분산 처리 시스템, 확장성이 커짐)
  ScaleUp보다 ScaleOut이 더 많은 양의 정보를 처리할 수 있는 개념

    1. 구글 파일 시스템(GFS, Google File System)
    2. 하둡 분산 파일 시스템(HDFS, Hadoop Distributed File System)

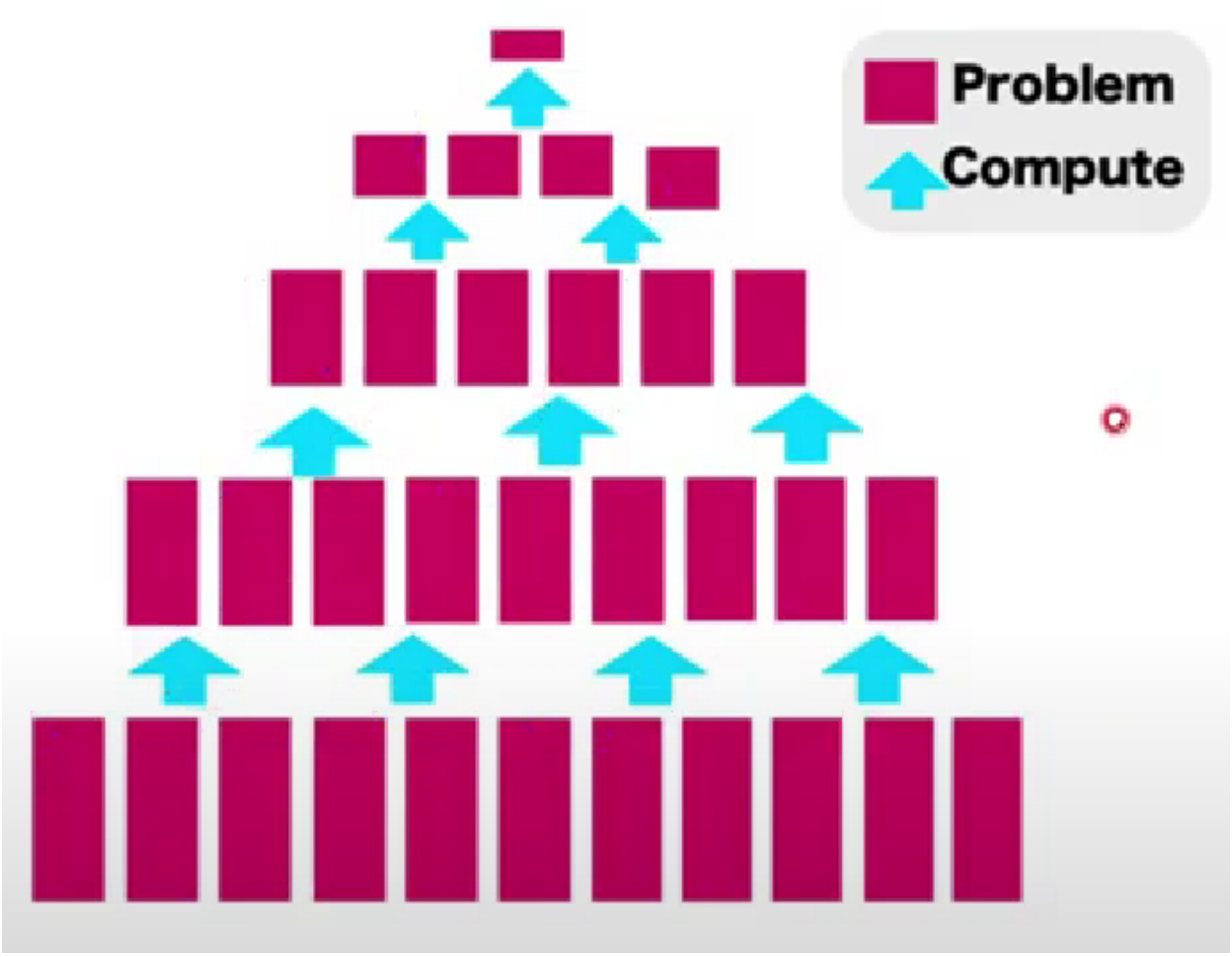
2. 분산 저장 및 처리 (Keyword)

1. 구글의 GFS - 청크, 클라이언트, 마스터
2. 맵리듀스 - Map + Reduce
3. 하둡 HDFS - yarn, spark
4. 하둡 에코 시스템 - hive, pigm impala, zookeeper, oozie



- 병렬계산 : 쪼갠 문제를 동시에 해결 (순차적x)
- 분산처리 : 문제를 쪼개서 여러 컴퓨터로 나누어서 따로 처리 및 저장
- 클러스터 : 분산처리 전 컴퓨터를 여러개로 나눠서 묶어줄 수 있는통신과 관리하는 프로그램이 필요함
- 하둡, 맵리듀스, 얀 : 하둡 1.0 -〉 병목, 리소스 문제 발생 -〉 트래픽 조절-〉 얀(하둡 2.0)

3. (참고) Divide and Conquer 알고리즘
   : 복잡한 문제를 여러개의 단순한 문제로 바꾼 뒤 해결




```
- 하나의 문제를 풀기 위해 독립된 여러 개의 작은 문제로 쪼갠 후 동시계산해서 취합 => 독립된 문제 병렬계산
```

- 가장 큰 장점 -〉 병렬성(Parallesim) -〉 기존에는 슈퍼컴퓨터가...
- 병렬처리 (CPU가 좋은 슈퍼컴퓨터)
  : 여러 개의 프로세스를 통해 독립된 문제를 통시에 처리
- 클러스터
  : 여러 대의 컴퓨터가 네트워크를 통해 연결된 하나의 시스템처럼 동작하는 컴퓨터의 집합 (컴퓨터들을 하나로 묶어줄 수 있는 S.W가 필요하니까 포함)
  -〉 저렴한 컴퓨터 + 네트워크 + S.W(분산처리)

※ 한 컴퓨터로 병렬계산(동시에 문제풀이)를 시킴 =〉 비쌈
-〉 클러스터!! 저렴한 컴퓨터 여러 대를 연결시키고, 한 컴퓨터처럼 병렬계산 시키자 **(여러 컴퓨터, 네트워크, 소프트웨어)**

4. 클러스터(Cluster)