3. 하둡

0. 하둡 : 빅데이터를 다루기 위한 분산 환경 오픈소스 플랫폼을 통칭

1. 하둡의 역사

- 2003 : 구글 GFS
- 2004 : NDFS (Nutch Distributed File System)
- 2005 : 너치 프로젝트 -> NDFS + 맵리듀스
- 2006 : NDFS와 맵리듀스를 너치 프로젝트에서 분리 -> 하둡
- 2006 : 더그 커팅 Yahoo에 합류 -> 하둡 발전

2. 하둡의 성능

1. 당시 데이터 처리

```
- 1테라바이트의 저장공간
- 약 100MB/s 수준의 전송속도
- 전체 드라이브를 읽어오는 데 걸리는 시간 : 약 두 시간 반 이상
```

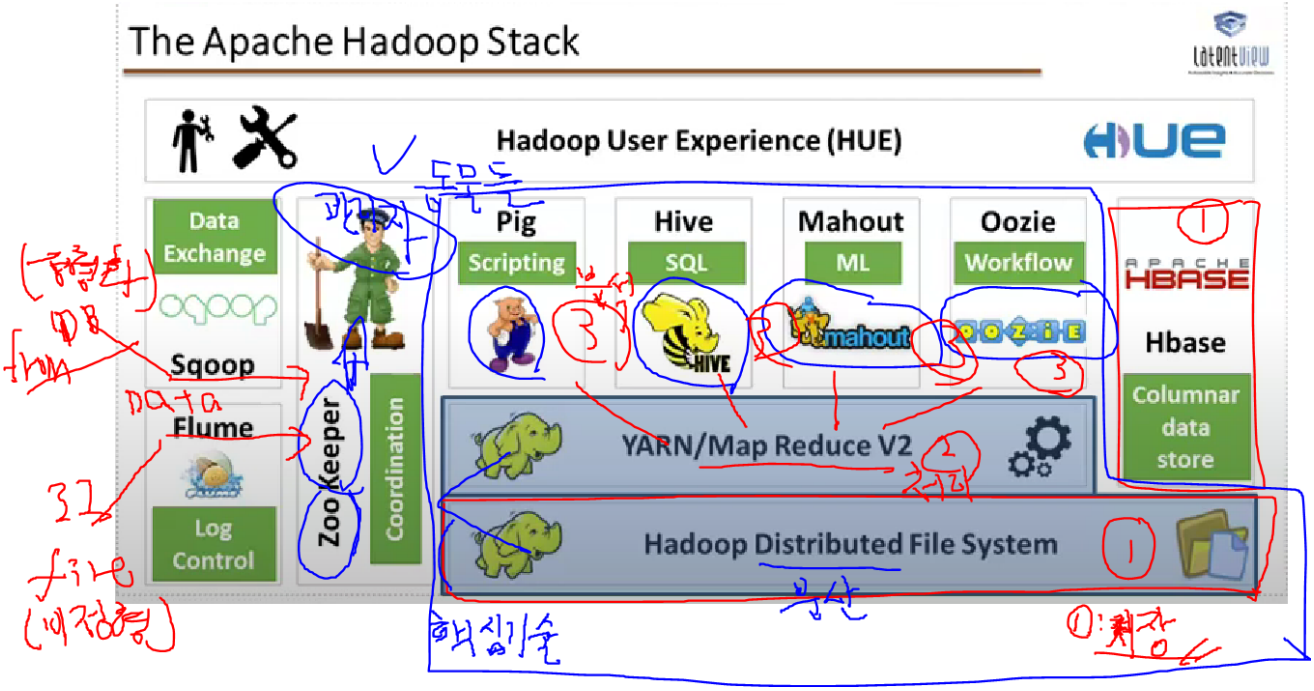
2. 1테라바이트 처리시간 단축

```
- 2008년 4월 : 910개의 노드 클러스터를 이용해 209초 달성
- 2008년 11월 : 구글 (68초)
- 2009년 5월 : 야후 (62초)
```

3. 하둡 : 대용량 데이터를 분산 철기할 수 있는 자바 기반의 오픈소스 프레임워크

- DBMS가 아님. 프레임워크!!

1. HDFS (분산저장) : 빅데이터 파일을 여러 대의 서버에 분산 저장하기 위한 파일시스템
2. 맵리듀스 (분산처리) : 각 서버에서 데이터를 분산처리하는 분산병렬처리를 위한 분석시스템
   =〉 분산되어 저장된 컴퓨터 각각에서 처리



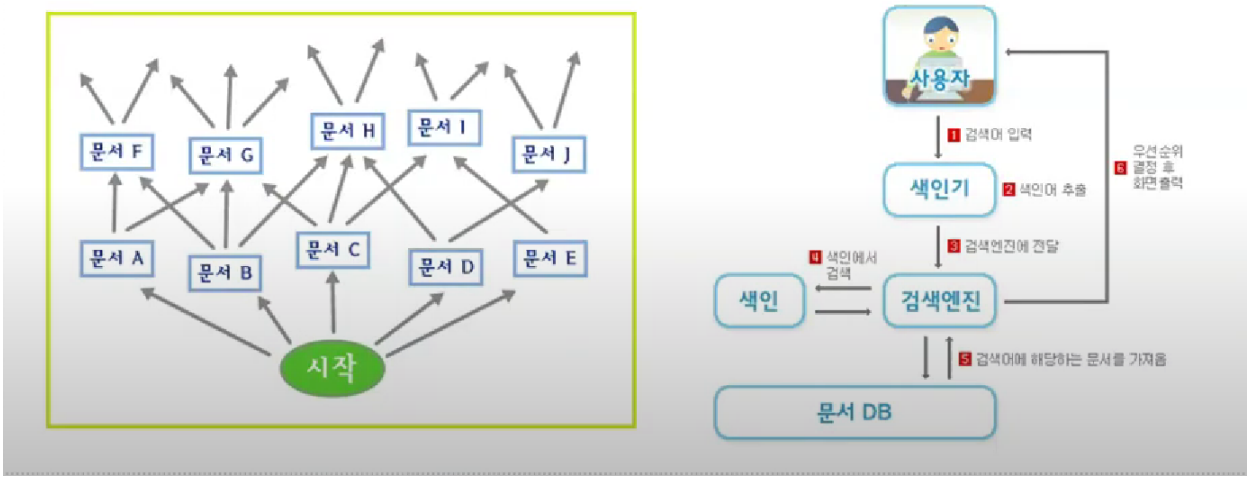
- DBMS의 한계
  : 웹 크롤러 색인처리 과정에서 생성되는 매우 큰 파일 처리 한계

--〉이를 극복하기 위해 나온 것이 '하둡'

- 텍스트 검색 라이브러리로 폭넓게 사용되고 있는 '아파치 루씬(웹 프로그램)'의 창시자인 '더그 커팅'에 의해 시작

- 크롤러와 검색 엔진 시스템 성능 향상

4. 검색엔진의 작동원리
   - 스파이더(spider), 크롤러(crawler)라고 불리는 로봇이 웹에 있는 웹 페이지를 방문해서 모든 내용을 읽어옴
   - 검색에 적합하도록 일정하게 가공하여 저장



- 검색에 적합하도록 -〉 주제. 즉, 색인