하둡은 HDFS(Hadoop Distributed File System)라는 데이터 저장소와 맵리듀스(MapReduce)라는 분석 시스템을 통해 분산 프로그래밍을 수행하는 프레임 워크  
데이터를 3군대에 저장함 (reliability)

HDFS : Hadoop의 파일 시스템

Reliability, high bandwidth

Name Node : 블락이 어디에 저장되있는지 정보를 갖고있는 거

Secondary Name Node

Data Node : 블락이 저장되는 곳

어떤 블락들을 가지고 있는지 주기적으로 block reporting

Heartbeat : data node가 살아있는지 알려주는 (10번마다 block report 보냄)

Data Pipelining : 클라이언트의 부담을 줄이기 위해 데이터 저장할 때 다른 놈한테 저장하라고 알려주는

Rack : data node들을 저장하는 곳

Rack awareness : data node failure 때문에

한 데이터 노드에 중복된 블락 x

한 rack에 같은 2개의 블락만 (rack 끼리 switch가 코스트 커서)

MapReduce

Map – Shuffle – Reduce

<http://blog.acronym.co.kr/312>

reverse web link graph : source, target의 순서를 바꿔서 써주는거 이러면 나중에 target을 찾을 때 source를 알 수 있음. 그니까 어떤 웹 페이지를 타겟으로 삼는 웹페이지가 누구 누구이고 그 수도 알 수 있음(웹페이지간 랭킹 파악가능)