

빅 데이터 소개

순천향대학교 컴퓨터공학과

이 상 정

빅 데이터 소개

학습 내용

- 빅 데이터의 정의
- 빅 데이터 컴퓨팅 주요 개념
- 빅 데이터 파이프라인

1. 빅 데이터의 정의

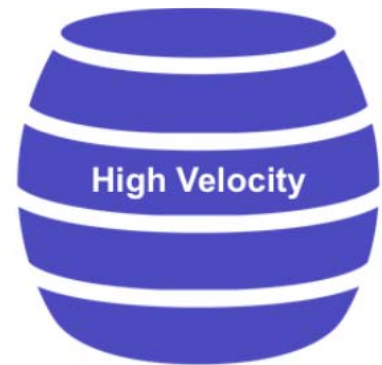
빅 데이터 소개

빅 데이터 (Big Data) 란?

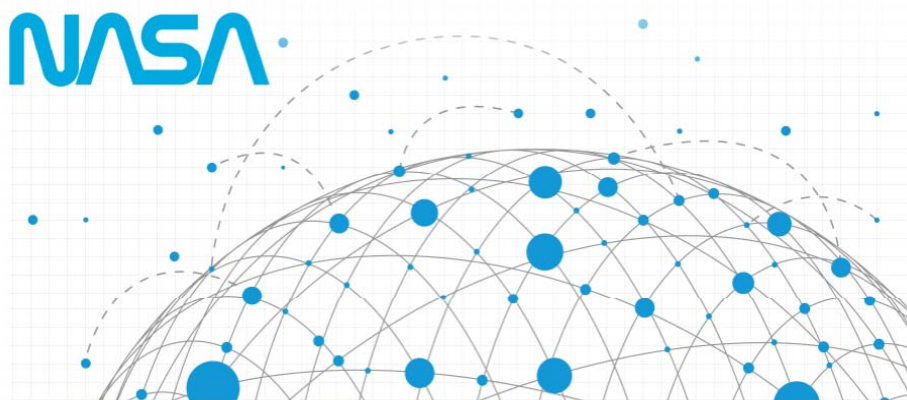
- 빅 데이터 란 기존의 방식으로 표현/저장/처리/분석하기 어려운 다양한 소스의 큰 규모의 자료를 의미
- 빅 데이터 소스
 - 정보 소비자 (consumers)
 - SNS, 인터넷 문서, 웹 로그, 의료 기록, 사진 아카이브, 비디오 아카이브,
 - 과학(science)
 - 기상, 유전, 시뮬레이션, 생태 환경,
 - 산업/정부
 - 센서, 카메라, RFID, 무선 센서 네트워크,

빅 데이터 특성 - 3V

규모 (Volume)
다양성 (Variety)
속도 (Velocity)



규모 (Volume)

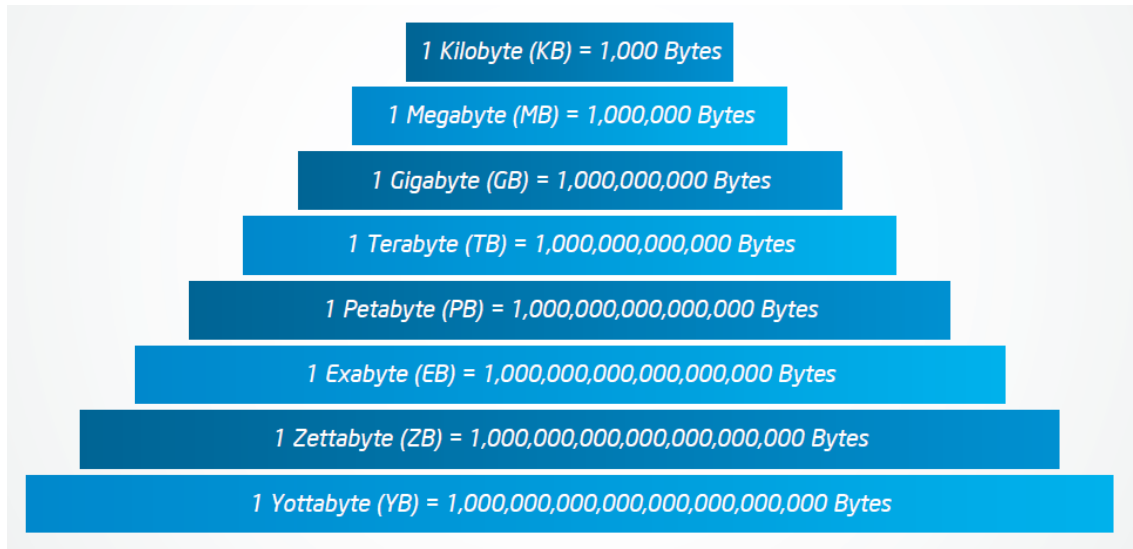


□ NASA 사례

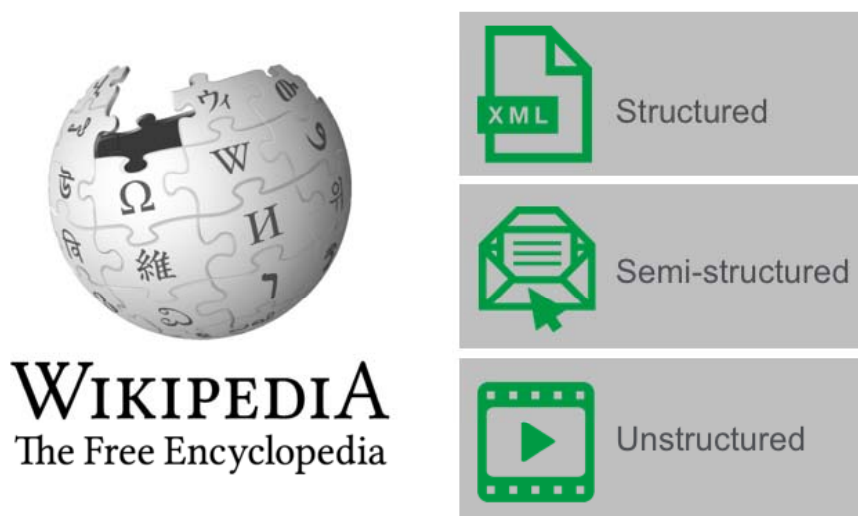
- 2014년 말 기준으로 NASA는 몇 초 단위로 약 1.73 GB 데이터가 생성 및 수집됨
- 이런 대규모 데이터는 **단일의 대규모 데이터로 저장 어려움**

데이터의 규모

- 빅 데이터는 Tera(10^{12}), Peta(10^{15}), Exa(10^{18}) 바이트 규모

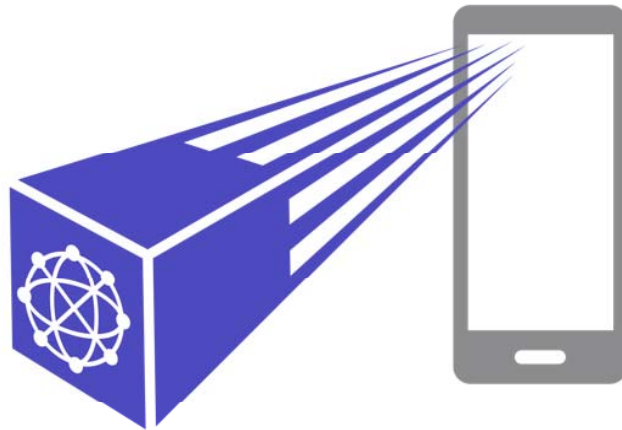


다양성 (Variety)



- 위키피디아(Wikipedia) 사례
 - 데이터 형태가 텍스트 외에 하이퍼링크, 이미지, 오디오, 비디오 파일 등 다양
 - 다양한 데이터 형태를 갖고 비구조화된 데이터의 표현이 어려움

속도 (Velocity)



□ 사물 인터넷에 연결된 센서

- 초 당 수 천개 데이터 포인트(data point) 생성
- NASA와 위키피디아 데이터와는 달리 스마트폰, 센서 데이터는 실시간으로 생성되고, 즉시 유용한 정보로 처리되어야 함
- 빠른 속도로 생성되는 **데이터의 실시간 처리 어려움**

구글(Google)의 빅 데이터 문제



□ 구글은 1996년 이래로 20년째 웹 검색 엔진을 운용

- 효율적인 검색을 위해 인터넷 전체에 대한 목록의 색인을 목표
- 이를 위해서는 **빅 데이터의 3V**를 대처하는 **혁신적인 방법**이 필요

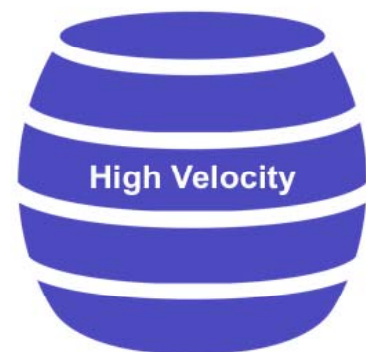
구글의 빅 데이터 문제 – 규모 (Volume)

- 인터넷의 모든 웹 페이지를 단어를 수집하고 색인을 생성
 - 기존의 데이터베이스로는 처리할 수 없는 **방대한 규모의 데이터**



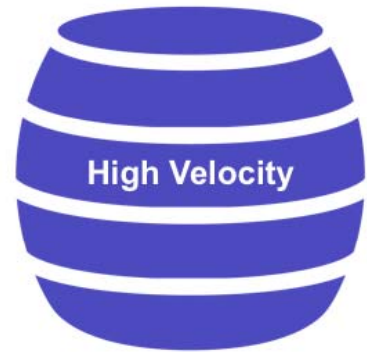
구글의 빅 데이터 문제 – 속도 (Velocity)

- 매일 수백만개의 웹 페이지들이 갱신
 - 이들 갱신된 데이터를 **빠른 속도로** 수집하고 처리



구글의 빅 데이터 문제 – 다양성 (Variety)

- 웹 페이지가 비구조화된 텍스트, 이미지, 오디오 및 비디오 등의 조합으로 구성
 - 모든 데이터의 조각들을 표현하고 저장하는 기존의 데이터베이스는 없음



2. 빅 데이터 컴퓨팅 주요 개념

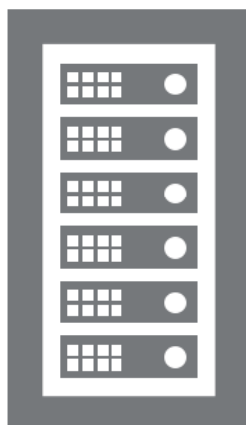
구글의 빅 데이터 문제

- 앞에서 소개한 **구글의 빅 데이터 문제의 해결**을 소개하여 **빅 데이터 컴퓨팅의 핵심 개념**을 소개



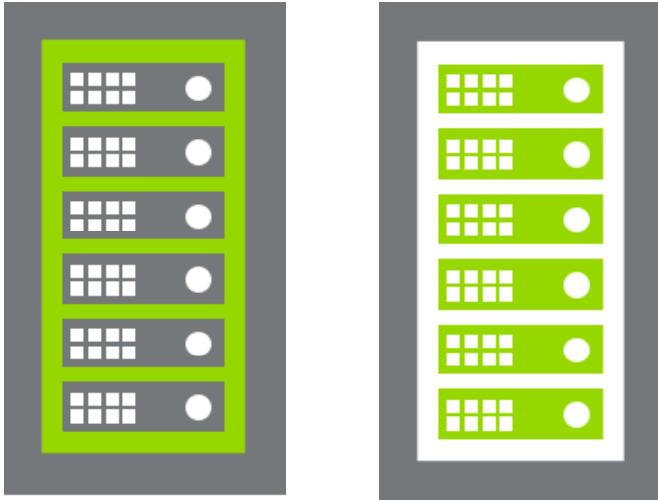
구글의 해결 방안

- **구글의 해결 방안**
 - 상호 연결된 다 수의 저가의 일반 범용 컴퓨터
 - 클러스터 (Cluster)
 - 구글 파일 시스템(GFS, 구글 File System), 빅테이블(Bigtable), 맵리듀스(MapReduce)
 - 분산 파일 시스템, NoSQL, 데이터 병렬처리



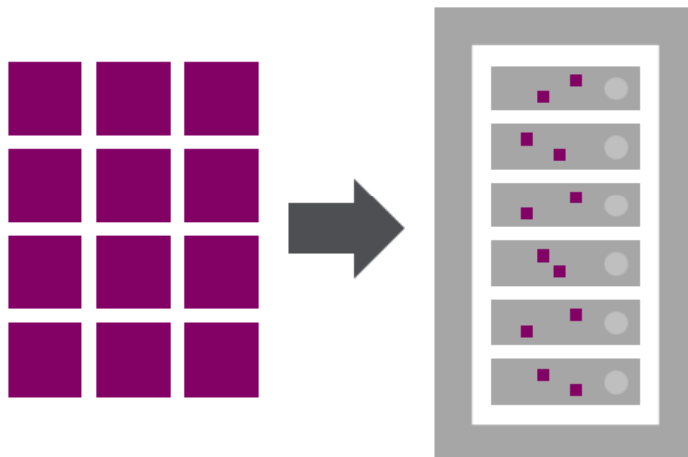
구글의 해결 방안 - 클러스터

- 클러스터는 개별 머신이 노드로 구성된 노드들의 집합
 - 각 노드는 일반 범용 컴퓨터로 고장나면 쉽게 교체 가능



구글의 해결 방안 - GFS

- GFS (Google File System)는 파일들을 여러 조각으로 분할하여 클러스터의 노드에 배분
 - 조각들은 **고장 감내(fault tolerance)**를 위해 서로 다른 노드에 복제



구글의 해결 방안 – 빅테이블 (1)

- 빅테이블(Bigtable)은 GFS 상에서 데이터를 저장하고 조회하는 데이터베이스 시스템

	A	B	C	D
1	Name	Email	Purchase	Time
2	Ahi	tahi@my-co.com	12	2010-03-16 19:06:01
3	Becker	becker@mapr.com	6	2015-05-05 03:21:12
4	Carlisle	carlisle@tcs.net	12	2014-12-12 09:35:03
5	Carlisle	carlisle@tcs.net	12	2015-01-25 11:30:22
6	Cavalero	dancav@abc123.com	6	2016-04-16 07:55:52

구글의 해결 방안 – 빅테이블 (2)

- 빅테이블은 성글고 분산된 지속성의 다차원 정렬 맵 (a sparse, distributed persistent multi-dimensional sorted map)
 - 행 키, 열 키, 타임스탬프로 저장된 데이터를 매핑
(row, column, timestamp) → cell contents

	A	B	C	D
1	Name	Email	Purchase	Time
2	Ahi	tahi@my-co.com	12	2010-03-16 19:06:01
3	Becker	becker@mapr.com	6	2015-05-05 03:21:12
4	Carlisle	carlisle@tcs.net	12	2014-12-12 09:35:03
5	Carlisle	carlisle@tcs.net	12	2015-01-25 11:30:22
6	Cavalero	dancav@abc123.com	6	2016-04-16 07:55:52

구글의 해결 방안 – 빅테이블 (3)

	A	B	C	D
1	Name	Email	Purchase	Time
2	Ahi	tahi@my-co.com	12	2010-03-16 19:06:01
3	Becker	becker@mapr.com	6	2015-05-05 03:21:12
4	Carlisle	carlisle@tcs.net	12	2014-12-12 09:35:03
5	Carlisle	carlisle@tcs.net	12	2015-01-25 11:30:22
6	Cavalero	dancav@abc123.com	6	2016-04-16 07:55:52

	A	B	C	D
1	Name	Email	Purchase	Time
2	Ahi	tahi@my-co.com	12	2010-03-16 19:06:01
3	Becker	becker@mapr.com	6	2015-05-05 03:21:12
4	Carlisle	carlisle@tcs.net	12	2014-12-12 09:35:03
5	Carlisle	carlisle@tcs.net	12	2015-01-25 11:30:22
6	Cavalero	dancav@abc123.com	6	2016-04-16 07:55:52

21

구글의 해결 방안 – 빅테이블 (4)

- ❑ 기존의 데이터를 덮어쓰기(overwriting)하지 않고 시간으로 구분하여 조회

	A	B	C	D
1	Name	Email	Purchase	Time
2	Ahi	tahi@my-co.com	12	2010-03-16 19:06:01
3	Becker	becker@mapr.com	6	2015-05-05 03:21:12
4	Carlisle	carlisle@tcs.net	12	2014-12-12 09:35:03
5	Carlisle	carlisle@tcs.net	12	2015-01-25 11:30:22
6	Cavalero	dancav@abc123.com	6	2016-04-16 07:55:52

구글의 해결 방안 – 빅테이블 (5)

- 행은 **태블릿(tablet)**이라는 부분 테이블로 분할
 - 태블릿은 각 클러스터의 노드에 분산 및 부하균형(load balancing) 단위
- 빅테이블은 **기존의 파일들을 재구성하지 않고** 새로운 노드들을 추가하면서 **대규모 데이터**를 다룰 수 있도록 설계

	A	B	C	D
1	Name	Email	Purchase	Time
2	Ahi	tahi@my-co.com	12	2010-03-16 19:06:01
3	Becker	becker@mapr.com	6	2015-05-05 03:21:12
4	Carlisle	carlisle@tcs.net	12	2014-12-12 09:35:03
5	Carlisle	carlisle@tcs.net	12	2015-01-25 11:30:22
6	Cavalero	dancav@abc123.com	6	2016-04-16 07:55:52

23

구글의 해결 방안 – 맵리듀스

- GFS 에 저장된 데이터 처리를 위해 **맵리듀스(MapReduce)** 패러다임으로 병렬처리
 - 함수형 언어에서 리스트 데이터를 처리할 때 사용하는 함수인 **map 함수**와 **reduce 함수**에 기원
 - **map 함수**는 리스트의 각 원소들에게 어떤 공통된 작업을 처리하고자 할 때 사용
 - **reduce 함수**는 리스트 전체 원소를 모아 하나의 결과를 출력하고자 할 때 사용
 - 함수 사용 예
 - map (+1) [1,2,3,4,5,6,7,8,9,10]**
// 1~10에 각각 1씩 더함 => [2,3,4,5,6,7,8,9,10,11]
 - reduce (+) [1,2,3,4,5,6,7,8,9,10]**
// 1+2+3+4+5+6+7+8+9+10 => 55

맵리듀스 역사

History of MapReduce



아파치 하둡 (Apache Hadoop)

- ❑ 구글 Labs는 자신들의 빅 데이터 해결 방안들은 논문으로 발표
- ❑ 아파치 하둡
 - Yahoo의 Doug Cutting이 구글의 방식을 참조하여 개발
 - 후에 Apache Foundation의 오픈 소스로 공개



MapReduce	MapReduce
Bigtable	HBase
GFS	HDFS

3. 빅 데이터 파이프라인

빅 데이터 소개

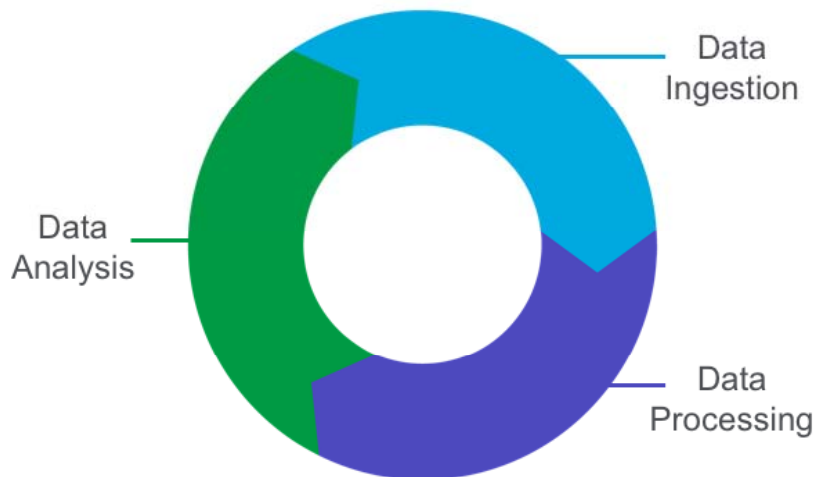
데이터 파이프라인 (1)

- 일반적으로 빅 데이터는 데이터의 소스로 부터 수집 (ingestion) 되어 처리 (processing) 및 분석 (analysis) 의 파이프라인 단계의 과정을 수행



데이터 파이프라인 (2)

- 데이터 파이프라인은 분석 결과가 수집되어 다시 순환 반복되어 처리될 수도 있음

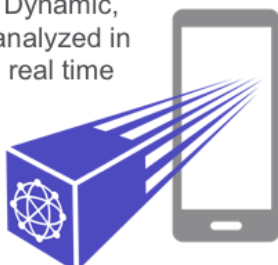


데이터 수집 - 속도 (Velocity)

- 스트리밍 데이터 (streaming data)
 - 사물 인터넷의 센서들과 같이 동적으로 끊임없이 생성되고 실시간에 준하는 (near real time) 분석이 요구
- 배치 데이터 (batch data)
 - 과거의 추이 분석 등을 위해 중앙의 서버로 부터 다운로드되어 대용량의 저장 데이터 로 즉시 분석이 요구되지 않음

Velocity:
Streaming vs. batch

Dynamic,
analyzed in
real time

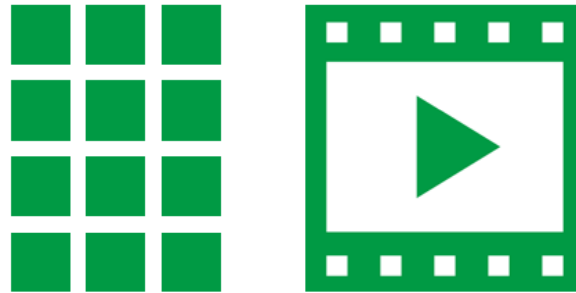


Analyze historical
trends, no
immediate action

데이터 수집 - 다양성 (Variety)

- 빅 데이터의 형태는 구조화, 반구조화, 비구조화된 데이터가 될 수 있음

Variety:
Structured vs. unstructured



데이터 수집 - 규모 (Volume)

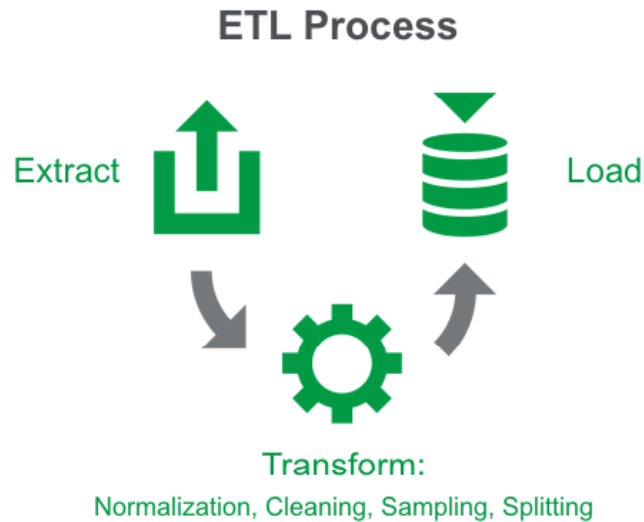
- 빅 데이터 저장 시 실제 데이터의 용량에다가 고장 감내를 위한 복제, 운영체제, 메타 데이터 등의 추가되는 용량을 고려하면 실제 데이터 용량의 약 4배의 공간이 요구

Volume:
Size of data

$$\sim 4 \times \text{Data} = \text{Storage Space}$$

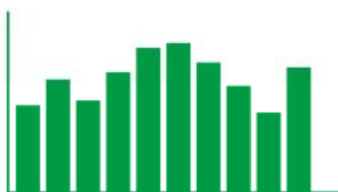
데이터 처리 - ETL

- 수집된 데이터는 ETL 과정으로 처리되어 분석될 준비 완료
 - ETL (Extract, Transform, Load)
 - 변환 과정에서 클리닝, 정규화, 샘플링, 분할이 포함



데이터 분석

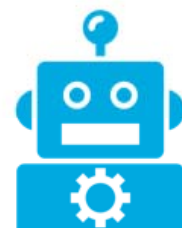
- 데이터 분석
 - 그래프나 차트로 처리된 데이터의 표현
 - 응용 프로그램의 자동화된 의사결정, 알림
 - 기계 학습(machine learning)
 - 딥러닝(deep learning)



Graphs, charts



Decisions, alerts



Machine learning

□ 빅데이터 적용 및 활용 사례 1개 조사

- 사례 예
 - 기업의 고객 정보 수집 및 분석
 - 전자상거래 고객의 선호도 조사
 - 의사결정 지원 시스템
 - 시스템 로그 데이터 수집 및 분석
 - 웹 서버 접근 로그 데이터 수집 및 분석
 - 해킹 및 보안 관련 분석
 - 의료 및 건강 기록 관리 및 분석
 - 농축산 데이터 분석 및 활용
 - IoT 센서 데이터 수집 및 활용
 - 기상 및 환경 데이터 수집 및 분석

.....

□ MapR Academy, <http://learn.mapr.com/>

- Introduction to Big Data
 - <http://learn.mapr.com/ess-100-introduction-to-big-data>

□ 구글 논문

- The Google File System
 - 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October, 2003.
 - <https://research.google.com/archive/gfs.html>
- Bigtable: A Distributed Storage System for Structured Data
 - OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November, 2006.
 - <https://research.google.com/archive/bigtable.html>