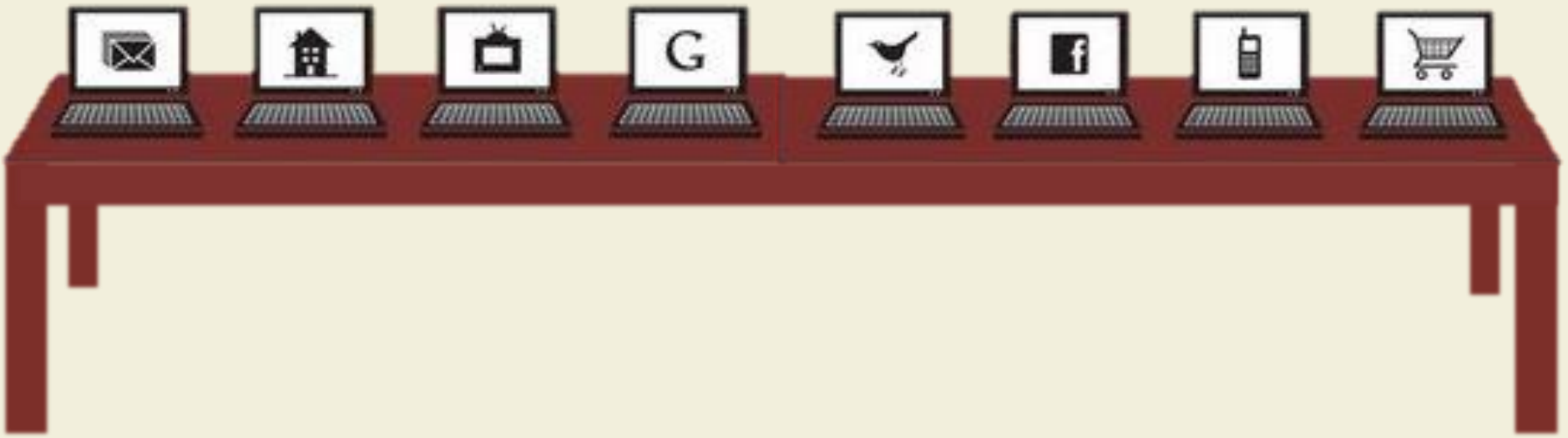
BIG DATA

CONTENTS



PART 1.

인기만점 빅데이터

Big Data, Big Impact :
국제 개발의 새로운 가능성을 여는 중요 기술_ `12

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

세계적 현안을 점검하는 포럼으로 전세계의 이목이 집중된다

Big Data :

향후 3년간 기업들에게 영향을 미칠 10대 전략기술 트렌드_`12,`13

Gartner®

IT나 비즈니스 창조적 파괴를 가할 잠재력을 갖고 있는 기술로,
대규모로 투자할 가치가 있으며 채택이 늦을 수록 기업들에게는 위험 요소로 작용

빅데이터
R&D
이니셔티브

2억달러

빅데이터
전문인력
양성소

5년간
140억원

빅데이터
분석전문가
5000명 양성

교육
커리큘럼

“동네 슈퍼 위기,
빅데이터로
활로 찾는다”


의료
기상
과학
행정
유통
제조업
통신
교통
BIG
DATA

PART 2.

빅 데 이 터
정 의 하 기

보통 수십에서 수천 테라 바이트 정도의 거대한 크기를 갖고

여러 가지 다양한 비정형 데이터를 포함하며

생성-유통-소비가 몇 초에서 몇 시간 단위로 일어나

기존의 방식으로는 관리와 분석이 매우 어려운 데이터

보통 수십에서 수천 테라 바이트 정도의 거대한 크기를 갖고

여러 가지 다양한 비정형 데이터를 포함하며

생성-유통-소비 몇 시간 단위로 일어나

기존의 방식으로는 분석이 매우 어려운 데이터

3V

Variety

Volume

Velocity

Volume

킬로바이트(1000)

메가바이트(1,000,000)_10만

기가바이트(1,000,000,000)_10억

테라바이트(1,000,000,000,000)_1조

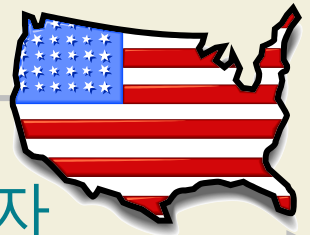
페타바이트(1,000,000,000,000,000)_1000조

엑사바이트(1,000,000,000,000,000,000)_100경

제타바이트(1,000,000,000,000,000,000,000)_10해

요타바이트(1,000,000,000,000,000,000,000,000)_1자

10^{15} 배



Quiz

2012년에는 2,700,000,000,000 기가바이트 (2조 7000억 기가바이트)의 데이터가 쏟아져 나왔다.

2012년에는 (2.7)제타바이트의 데이터가 쏟아져 나왔다.

Quiz

하루 평균 7,500,000,000기가바이트(75억 기가바이트)의 데이터가 쏟아져 나왔다.

하루 평균 (**7.5**)엑사바이트의 데이터가 쏟아져 나왔다.

750만개의 하드디스크

Variety

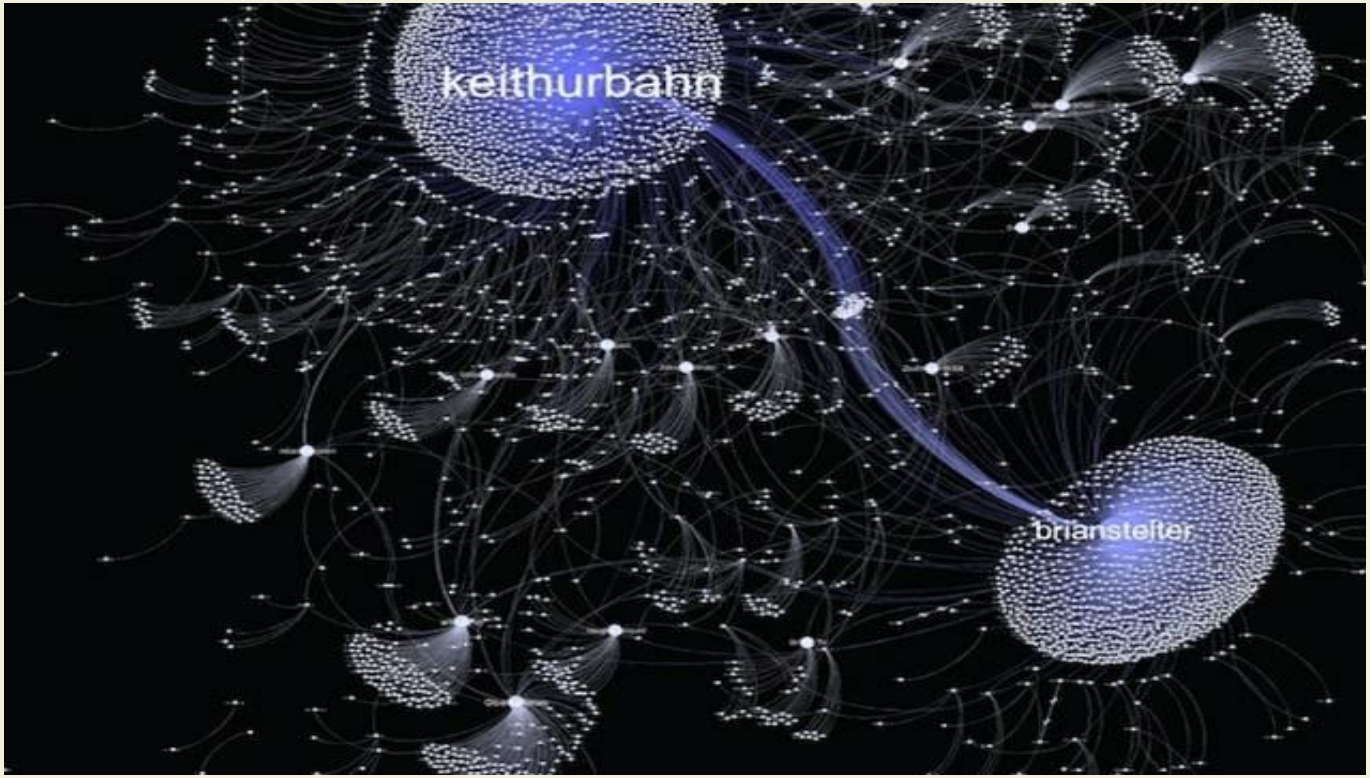
정형 데이터 : 고정된 필드에 저장되는 데이터 (데이터베이스)

온라인 쇼핑몰에서 제품을 주문할 때
이름, 주소, 연락처, 배송주소, 결제정보
등

비정형 데이터: 고정된 필드에 저장되어 있지 않은 데이터

동영상 데이터, 사진과 오디오 데이터,
메신저로 주고받은 대화 내용,
위치정보, 통화 내용 등

Velocity



Ex1) 빈라덴 사망소식 트위터 전파 속도 : 초당 5000회

Ex2) 서울 우면산 산사태의 생생한 모습 : 실시간 화제

기존의 방식으로는 관리와 분석이 매우 어려운 데이터

그리고 이를 관리, 분석하기 위해 필요한

인력과 조직, 관련 기술까지 포괄하는 용어

넓은 의미

PART 3.

어디쓰지 빅데이터

기업마케팅

고객들의 모든 구매 관련 데이터

고객 정보의 일부를 추출

고객들 간의
유사성을 파악

취향에 맞는 책을 추천

고객 구매 정보들 추출

책들 간의
연관성을 파악

다양한 분야의 책을 추천

범죄예방

지난 8년간의 범죄자료 데이터

범죄 시간, 범죄 발생 횟수

범죄 장소, 위험 지역 선정

10%

의료분야

5년간의 상위 5천만 개의 검색어

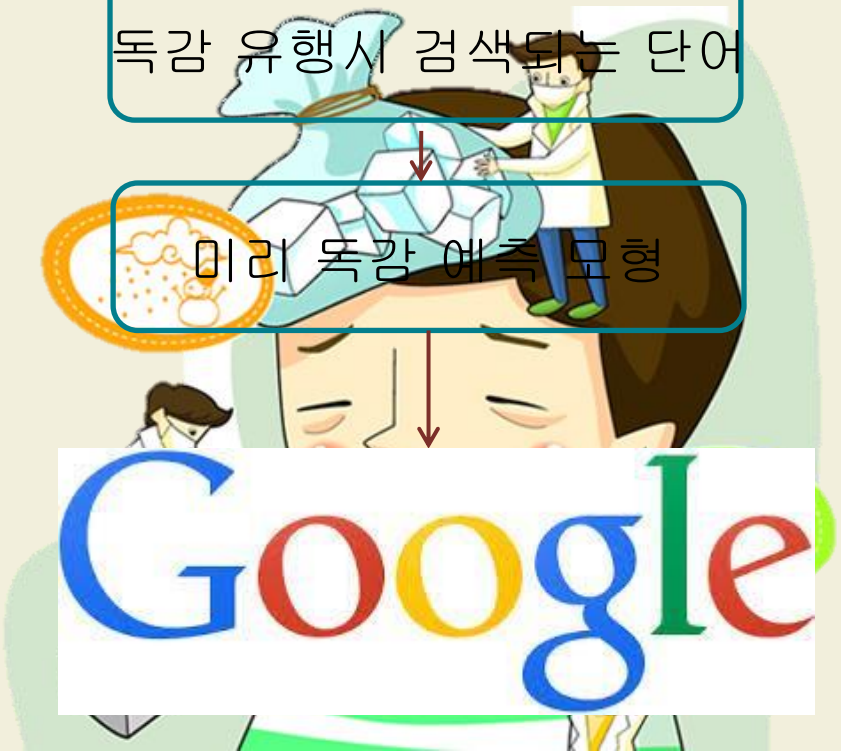
전파속도 ↑ 모니터링 ↓

독감 지역에서의 검색어

2009년 신종 인플루엔자 환자 8만명

독감 유행시 검색되는 단어

미리 독감 예측 모형



교통분야

휴대폰 사용 기록

어느 시각, 어느 지역에서
가장 많이 사용하는지

이용객의 구입품
주요 사용 콘텐츠

노선 결정, 확대

새로운 비즈니스로
해석과 융합

PART 4.

빅데이터 처리기술

ETL 기술...

비정형 데이터 관리,
분산데이터베이스(NoSQL)...

인메모리 기술,
복합 이벤트처리(CEP)...

클라우드 컴퓨팅,
하둡 분산파일시스템...

데이터패턴발견, 데이터
순서화, 자연어 처리...

데이터 시각화
(Visualization)

DataBase

쉽게 데이터 베이스(DB)는 도서관이라고 생각하면 된다. 수많은 데이터들을 일정한 기준에 맞춰서 저장해두고, 사서를 통해서 쉽게 검색(SQL)하고 찾을 수 있다. 이때 사서는 데이터베이스 매니지먼트 시스템(DBMS)이라고 할 수 있다.

Hadoop

대용량 데이터를 처리할 수 있는
대표적인 빅데이터 소프트웨어

Hadoop은 개발자, 더그커팅의 아들이
노란 코끼리 장난감 인형을 하둡이라고
부르는 것을 보고 지은 이름

Hadoop

대용량 데이터를 **분산 처리** 할 수 있는
분산파일시스템

분산파일시스템에 데이터를 저장
→ 찾아가는 정보(index정보)를 이용해서 데이터를 처리

맵리듀스(MapReduce)

수 집

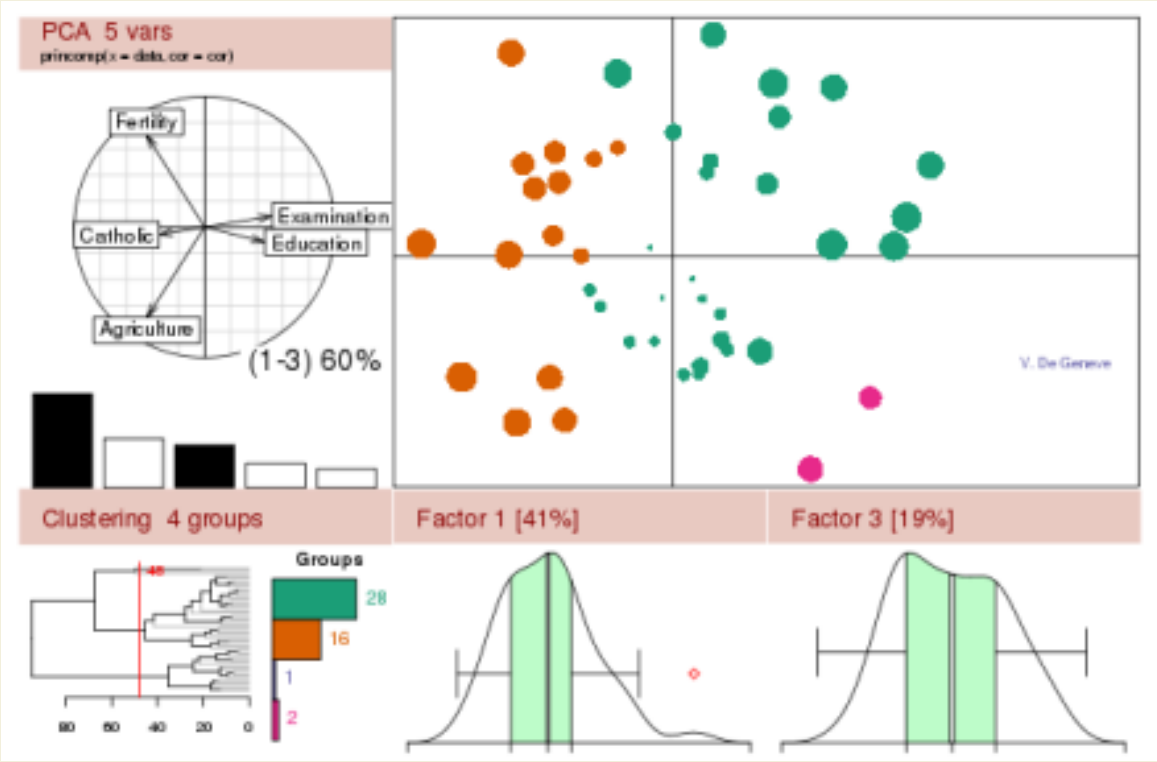
저 장

실시간처리

분산처리

분 석

표 현



R

통계 분석 및 시각화를 위한 프로그래밍 언어

통계학에서 사용하는 Tool

수 집

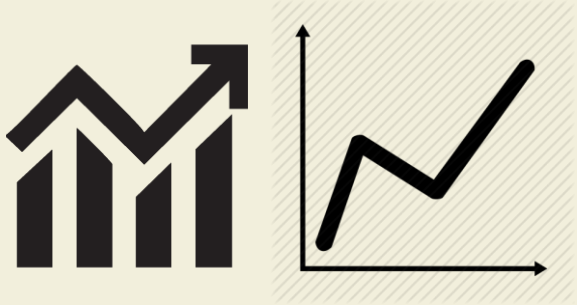
저 장

실시간처리

분산처리

분 석

표 현



방대한 양의 데이터 분석을 직관적이고 신속한

이해와 응용이 가능하도록 하는 것

사물을 바라보는 방식을 바꿈으로써 의사결정권자들이 예전에는 질문하지 못했던 것을 하게 만드는 것

Visualization

심층적인 질문은
뛰어난 전략으로 이어진다