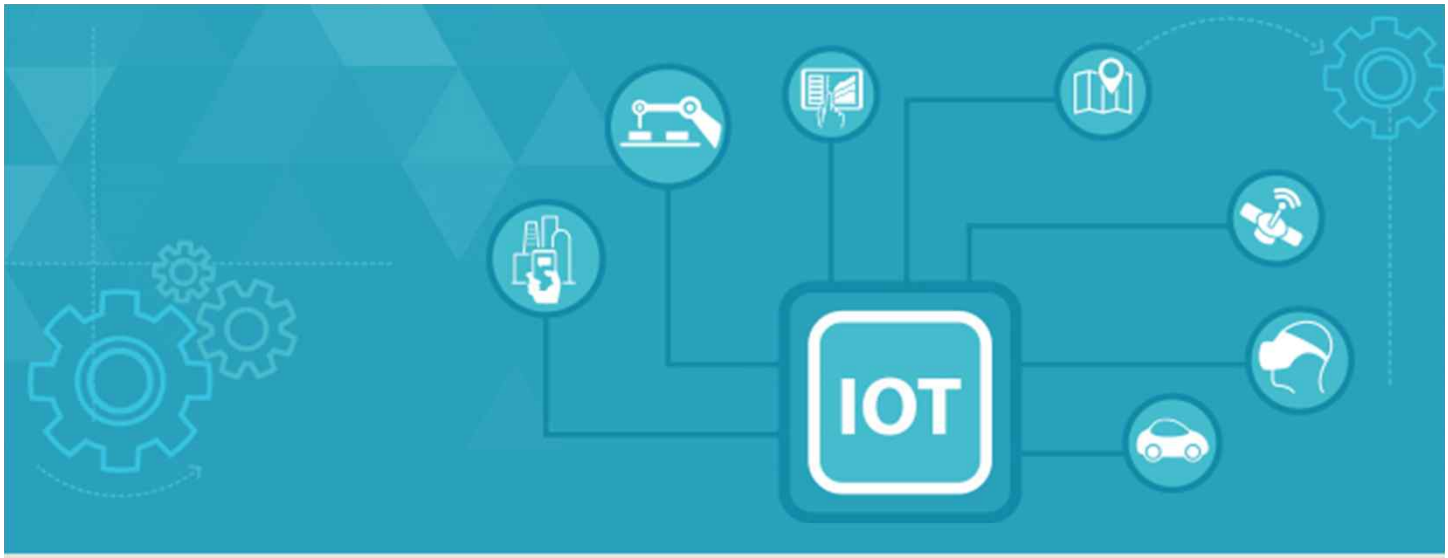


4차 산업혁명 신산업 기술 이해

4차 산업혁명과 빅데이터 준비하기

한국기술교육대학교
온라인평생교육원
OLEI

학습내용

- 빅데이터 니즈(Needs)
- 빅데이터 처리
- 빅데이터 준비

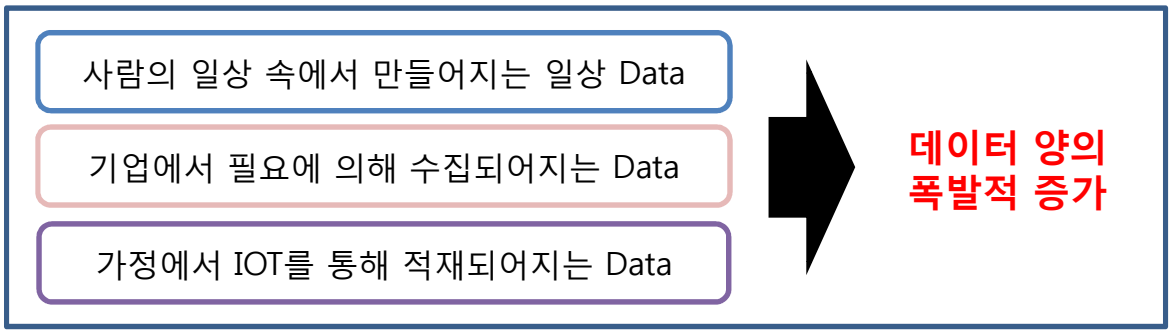
학습목표

- 빅데이터 정의와 오해를 찾아, 빅데이터의 기원 속에서 필요성을 설명할 수 있다.
- 빅데이터의 처리 과정에 대한 순서와 단계별 의미를 설명할 수 있다.
- 빅데이터 수집, 분석, 표현에 있어 필요한 내용과 도구들을 설명할 수 있다.

빅데이터 니즈(Needs)

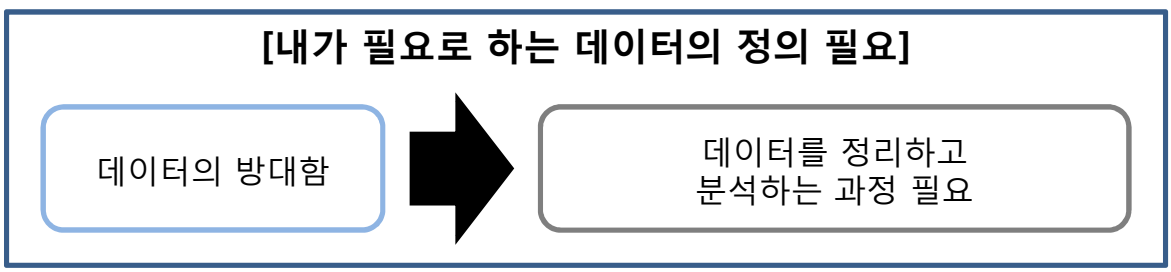
1. 빅데이터(Big Data)의 현재

- 수많은 양의 데이터가 매일 생성되고 있는 세계
- 사람들은 온라인에서 많은 시간을 보내며 자신의 일상을 다양한 형태로 공유
- 스마트폰, 자동차, 스마트홈 등에 탑재되어 있는 수많은 센서들이 현실 세계에 적재되어, 세계를 감지하고, 데이터를 생성 및 공유



2. 빅데이터(Big Data)에 대한 질문

① 내가 필요한 Data는 무엇인가?
② 그 Data는 어떻게 수집하지?
③ 이 많은 Data는 다 필요한가?
④ 이 Data는 어떻게 처리하지?
⑤ 필요한 기술과 처리과정은?
⑥ 이 상황에서 필요한 표현은?



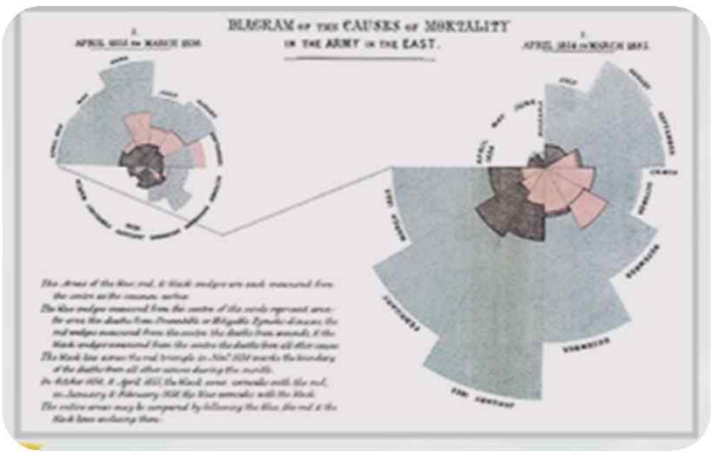
빅데이터 니즈(Needs)

3. 역사 속 빅데이터(Big Data)

1) 나이팅게일의 로즈 다이어그램

- 1849년 이집트 여행 중 알렉산드리아 병원 참관하고, 정규 간호 교육의 중요성을 절실히 느끼게 되어 전쟁의 참상 기사를 읽은 뒤 자극 받음
- 부모의 반대를 무릅쓰고, 독일 카이저벨트의 프로테스탄트 학교에서 간호학을 공부
- 1853년 런던 숙녀병원의 간호부장이 됨
- 1854년 11월 4일 크림전쟁에 참여 전쟁이 끝난 그 해 7월에 복귀
- 크림전쟁의 야전병원에서 환자 대부분 전쟁으로 죽는 것이 아니라 전염병 때문에 죽는다는 것을 발견하고 비위생적인 상황을 개선하려 노력
- 설득을 위해 필요한 Data를 수집하고, 표현하기 위해 방법을 고민하고 그림으로 표현하여 윗사람들 설득
- 무질서한 병원에 규율을 세워 환자 사망률은 42퍼센트에서 2퍼센트로 뚝 떨어짐

[나이팅게일]

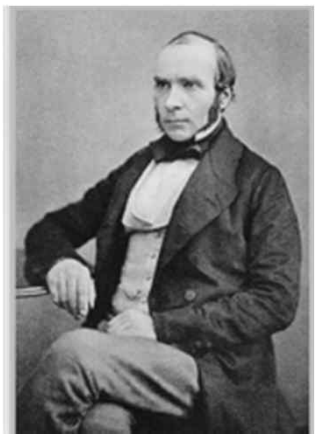


통계적인 자료를 다이어그램으로 보여줌
역사적인 가치가 매우 뛰어난 걸작으로 인정

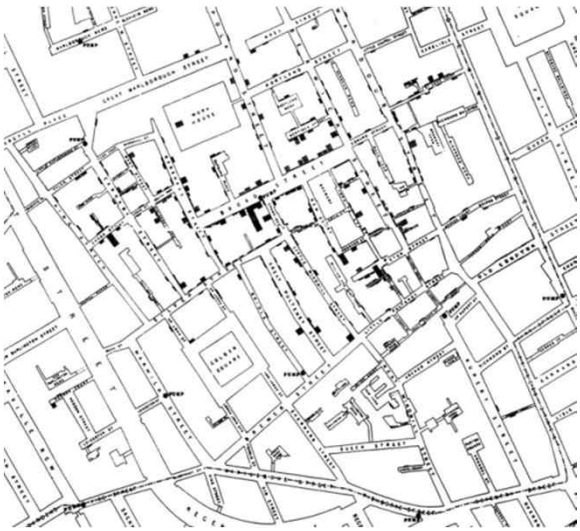
3. 역사 속 빅데이터(Big Data)

2) 존 스노우의 콜레라 맵

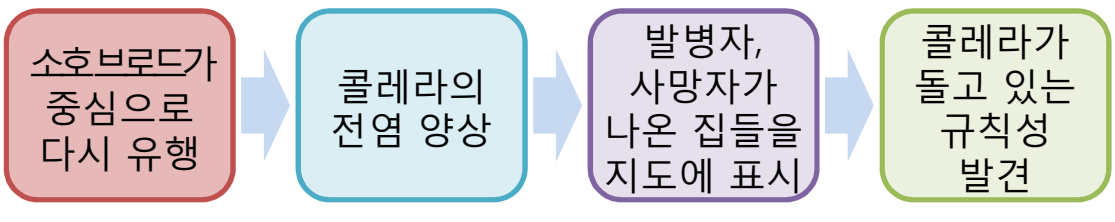
- 빅토리아 여왕 시대의 영국의사
- 마취통증의학과 의사로서 효과적인 마취방법 등을 연구하고 개발
- 역학의 선구자
- 1854년
  - 런던 소호에서 창궐한 콜레라가 오염된 물을 통해서 퍼졌다는 것을 연구를 통해 밝혀 내어 수많은 목숨을 구함
  - 당시 런던에서는 콜레라가 주기적으로 창궐



[존 스노]

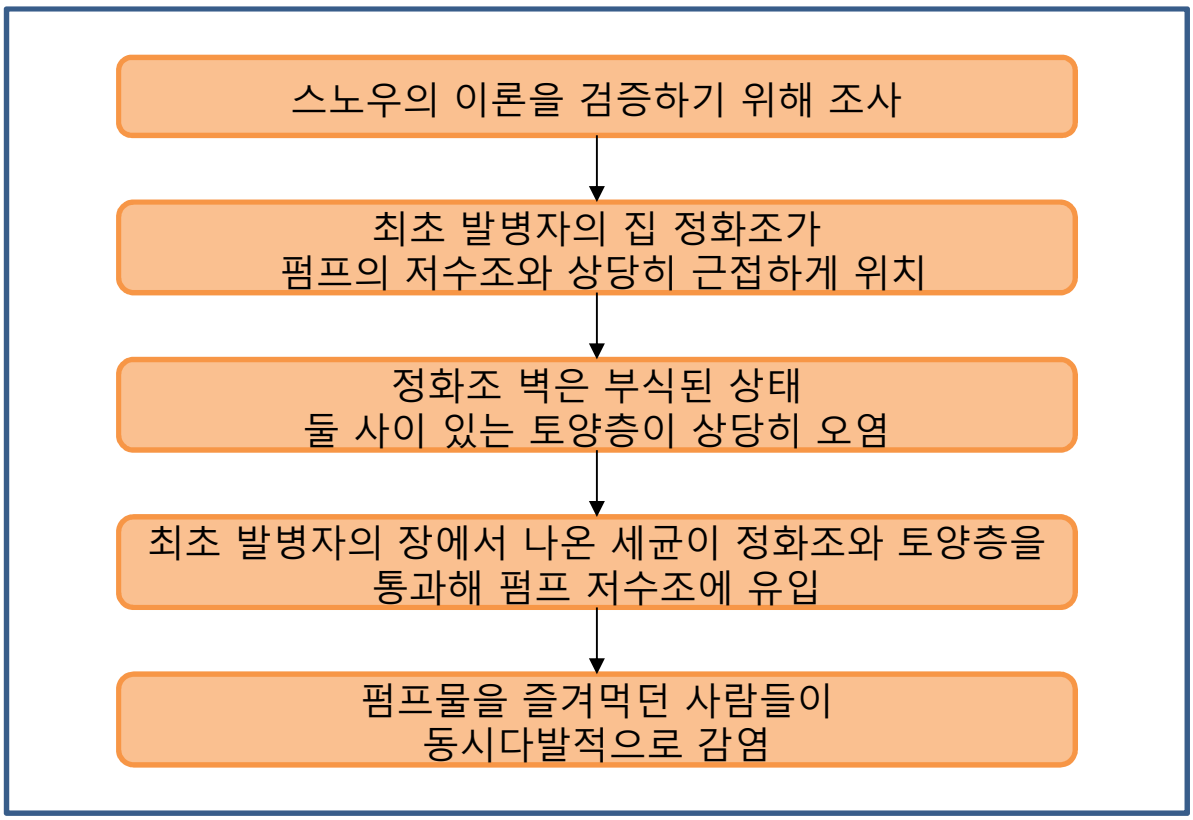


[콜레라 맵]



3. 역사 속 빅데이터(Big Data)

2) 존 스노우의 콜레라 맵



- 역학의 방법론은 존 스노우가 한 방법들을 빅데이터와 첨단 기법을 동원해 발전시킨 것일 정도로 역학의 성립에 많은 기여를 함

빅데이터 니즈(Needs)

4. 빅데이터 니즈와 오해

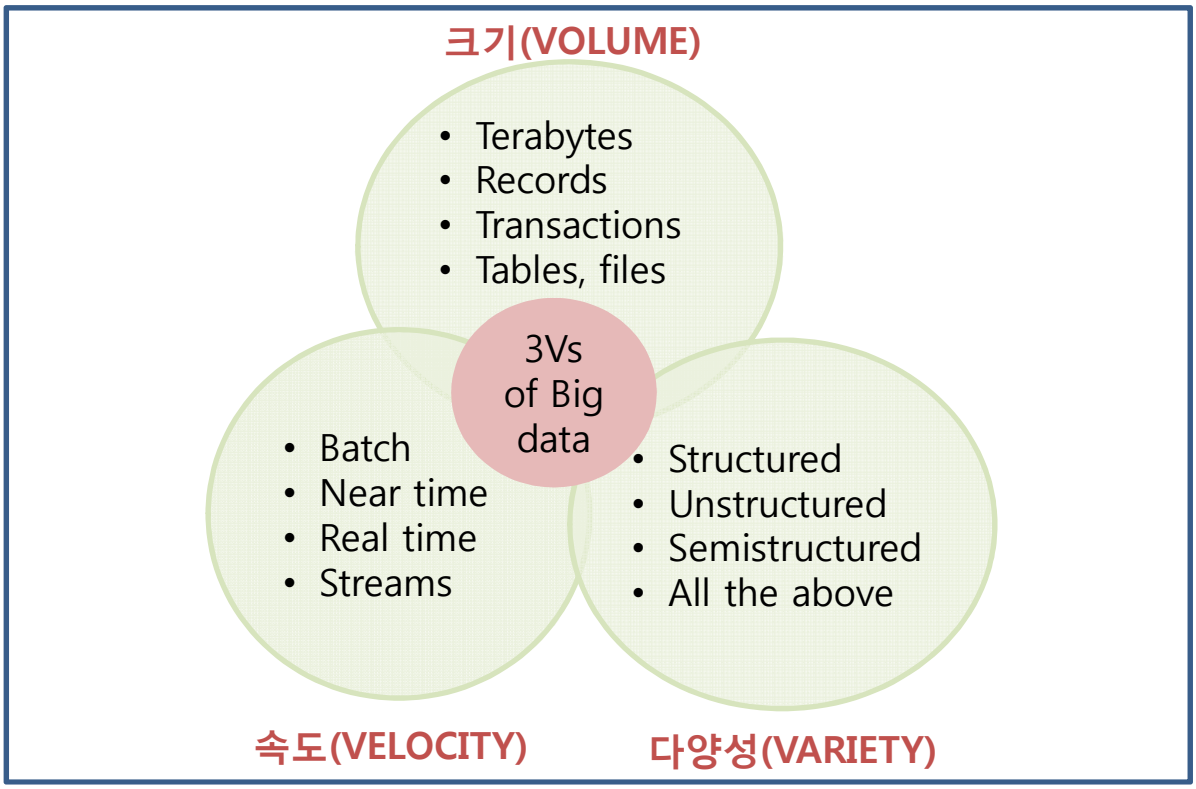
1) 역사 속 빅데이터의 증명

- 빅 데이터는 단순한 Data가 아님
- 필요에 의해 Data가 수집
- Data의 패턴을 통해 인사이트(통찰력) 고찰

2) 빅데이터에 대한 정의와 오해

- 3V : 빅데이터를 설명할 때 항상 먼저 나오는 것
- **V**olume, **V**elocity, **V**ariety

데이터의 속성에 초점이 맞춰져
별개의 데이터로 오해 발생



4. 빅데이터 니즈와 오해

2) 빅데이터에 대한 정의와 오해

- 빅데이터(Big Data) 시작 단계에서는 Data 수집에 초점이 맞춰짐
- Data 축적만이 인사이트(통찰력)을 가져오는 것이 아님
- 필요로 하는 방향으로 Data가 하고자 하는 방향을 맞추어 줘야 한다는 것이 중요함
- 방향을 맞춰 이루어지는 빅데이터(Big Data)는 목적을 이루기 위한 수단이 됨
- 목적이 확실하다면 Data의 양도 형태도 중요한 것은 아님

빅데이터가 없는 것이 아니라

- 무엇이 필요한지, 모르고 있는 것은 아닐까요?
- 목적과 그에 따른 Data에 대한 고민이 부족한 것은 아닐까요?

빅데이터 처리

1. 빅데이터 처리 과정

1) 빅데이터 정의

기관명	빅데이터 정의 및 특징
포레스트	• 가치를 얻기 위한 데이터와 무엇을 할 것인지 아는 사람이 기업에게 필요하다는 것을 의미하는 기술 • 볼륨, 속도, 다양함, 다양성으로 현재의 기술로 감당하기 어려운 규모의 데이터 • 경제적 가치를 창출하는 데이터
SERI	• 거대한 데이터 집합으로 대규모 데이터와 관련된 기술 및 도구 포함
가트너	• 3V로 정의 : 크기(Volume), 다양성(Variety), 복잡성(Complexity) • 크기(Volume):데이터 규모가 크다는 것을 의미 • 다양성(Variety):로그기록, 소셜, 위치정보 등 데이터의 종류 증가로 텍스트와 멀티미디어 등 비정형화된 데이터의 유형이 다양화 되는 것을 의미 • 복잡성(Complexity):구조화되지 않은 데이터, 데이터 저장 방식의 차이, 중복성 문제 등 데이터 종류가 확대되고 외부 데이터의 활용 등으로 관리대상이 증가됨으로써 점차적으로 데이터 관리 및 처리가 복잡화되고 심화되어 새로운 처리 및 관리기법이 요구되는 상황을 의미

빅데이터 처리

1. 빅데이터 처리 과정

1) 빅데이터 정의

기관명	빅데이터 정의 및 특징
SAS	• 4V로 정의:크기(Volume), 다양성(Variety), 속도(Velocity), 가치(Value) • 크기(Volume), 다양성(Variety)는 가트너 정의와 동일 • 속도(Velocity) : 센서나 모니터링 등 사물정보, 스트리밍 정보 등 실시간성 정보가 증가하고 있고 이러한 실시간성으로 인한 데이터 생성, 이동과 유통의 속도가 증가하고 있으며 대규모 데이터 처리 및 가치있는 실시간성 정보활용을 위해 데이터 처리 및 분석 속도가 매우 중요함 • 가치(Value) : 새로운 가치를 창출하는 것
노무라 연구소	• 빅데이터를 처리할 수 있는 인재, 조식, 데이터 처리, 축적, 분석기술, 데이터 자원 등을 빅데이터의 3요소로 정의 • 3요서의 조화로운 발전이 데이터의 특성과 컴퓨터 파워의 발달에 따라 실생활 적용이 빠르게 확산될 것으로 전망

빅데이터 처리

1. 빅데이터 처리 과정

2) 정보 기술 패러다임의 변화

- 1990년 이후 인터넷이 확산되면서 정형 Data와 비정형 Data가 대량으로 발생하면서 정보 홍수개념이 등장하고 이후, 오늘날 빅데이터 개념으로 정의
- 개인화 서비스와 SNS의 확산으로 기본 인터넷 서비스 환경 재구성을 바탕으로 증가하는 전 세계 디지털 Data 양은 ZB 단위로 2년마다 2배씩 증가하여 2020년에는 약 40ZB가 될 것
- 스마트폰의 보급으로 데이터가 매우 빠르게 축적되어 ZB 시대를 스마트 시대라고도 함

PC 시대		인터넷 시대	모바일 시대	스마트 시대
패러다임 변화	디지털, 전산화	온라인화, 정보화	소셜화, 모바일화	지능화, 개인화, 사물 정보화
정보 기술 이슈	PC, PC통신 데이터 베이스	초고속 인터넷, www. 웹 서버	모바일 인터넷, 스마트폰	빅데이터, 차세대 PC, 사물 네트워크
핵심 분야 (서비스)	PC, OS	포털, 검색 엔진, Web 2.0	스마트폰, 웹 서비스, SNS	미래 전망, 상황 인식, 개인화 서비스
대표 기업	MS, IBM	구글, 네이버, 유튜브	애플, 페이스북, 트위터	구글, 삼성, 애플, 페이스북, 트위터
정보 기술 비전	1인 1PC	클릭 e-Korea	손 안의 PC, 소통	IT everywhere, 신가치창출

빅데이터 처리

1. 빅데이터 처리 과정

3) 빅데이터 분류

- 정형과 비정형은 DB 스키마라는 표준방식으로 정의하느냐 일반적인 파일시스템 형태로 유지하느냐의 차이로 구분
- **정형 Data** : 단순하게 보면, 행(가로), 열(세로)의 Table형으로 저장되는 Table, xls 파일, csv 파일 등의 형태
- **비정형 Data** : Data의 저장이 중요하지 않으나 단순한 경향파악이 주요 이유일 경우에 해당 (이 경우 일정시간경과 후 소멸)
- 저장할 필요가 없거나 생성속도가 빨라서 저장기술이 생성속도를 따라갈 수 없을 경우는 **비정형 Data**로 유지

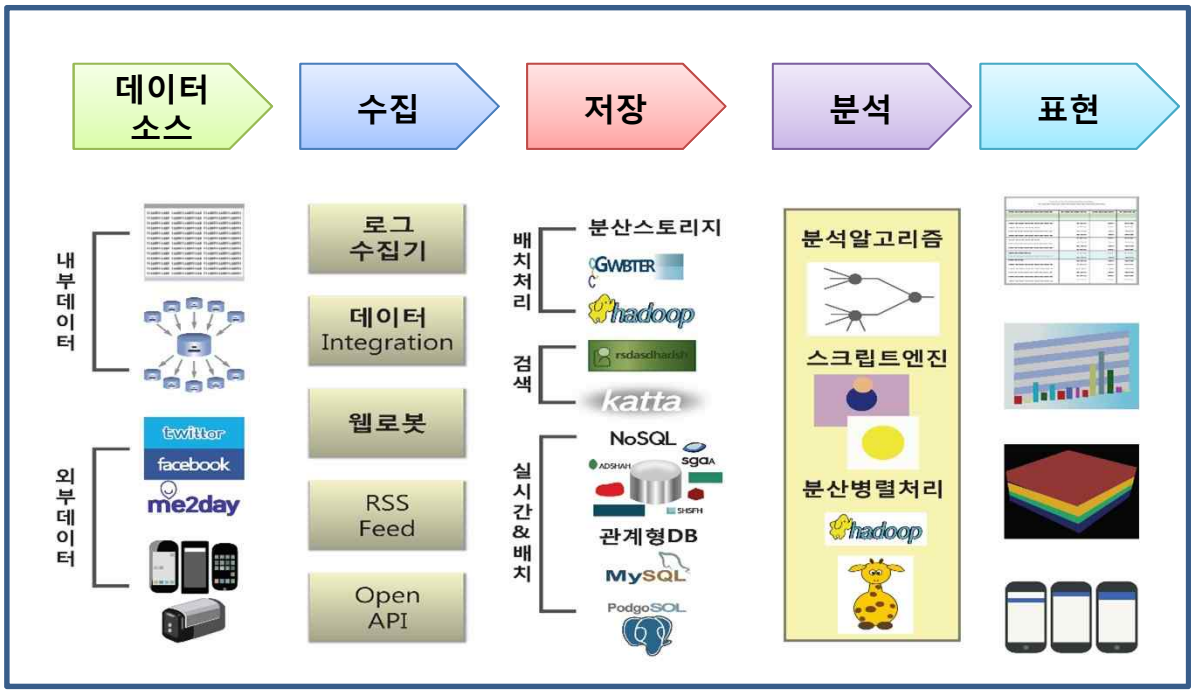
4) 빅데이터 처리 흐름

- 필요한 Data 의 위치를 파악하여, 수집 이후 저장
- 저장된 Data를 정제 가공
- 대상에 따라 보이기 위한 시각화 표현 작업
- 내부 파일 시스템, 데이터베이스 관리 시스템, 센서 등에서 정형 데이터 수집
- 인터넷으로 연결된 외부에서 비정형 데이터를 수집

빅데이터 처리

1. 빅데이터 처리 과정

4) 빅데이터 처리 흐름



2. 빅데이터 요소 기술

1) 빅데이터 처리에 필요한 기술 분류

- 크롤링 : R, Pyrhon, JavaScript
- **저장 기술** : RDBMS의 대용량 확장 System, MongGo DB, Hadoop, Spark와 같은 실시간 처리, 분산 병렬 처리 System
- **분석 기술** : 적절한 통계방법론을 사용
- **시각화 기술** : R Python 등의 언어에서 표현해 내는 방법과 태블로와 같은 시각화 도구 이용

2. 빅데이터 요소 기술

1) 빅데이터 처리에 필요한 기술 분류

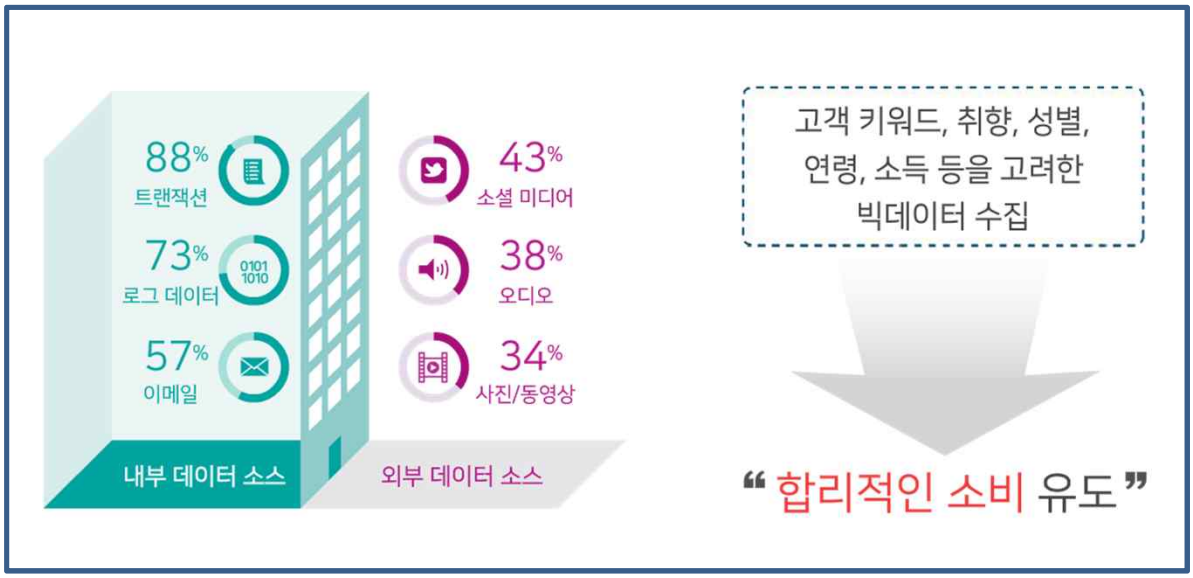
과정	설명	해당기술
생성	조직의 내부와 외부에 존재하는 여러 데이터를 생성하는 기술	• 데이터베이스 • 파일관리시스템 • 인터넷으로 연결된 파일 등
수집	조직의 내부와 외부에서 생성되는 여러 데이터 소스로부터 필요로 하는 데이터를 검색하여 수동 또는 자동으로 수집하는 과정과 관련된 기술로 단순 데이터 확보가 아닌 검색, 수집, 변환을 통해 정제된 데이터를 확보하는 기술	• 로그수집기 • 크롤링 • 센싱 • RSS Reader, Open API • ETL등
저장	작은 데이터라도 모두 저장하고 실시간으로 저렴하게 데이터를 처리하고 처리된 데이터를 더 빠르고 쉽게 분석하도록 효율적으로 저장하는 기술	• 분산 파일 시스템 • NoSQL • 병렬 DBMS등
처리	엄청난 양의 데이터의 저장, 수집, 관리, 유통, 분석을 처리하는 일련의 기술	• 실시간 처리 • 분산병렬처리 • 맴리듀스 등
분석	데이터를 효율적으로 정확하게 분석하여 비즈니스 등의 영역에 적용하기 위한 기술로 이미 여러 영역에서 활용해온 기술	• 통계분석 • 데이터 마이닝 • 텍스트 마이닝 • 평판분석 • 소셜 네트워크 분석 등
시각화	자료를 시각적으로 묘사하는 기술로, 빅데이터는 기존의 단순 선형적 구조의 방식으로 표현하기 힘들기 때문에 필수적	• 정보 편집 기술 • 정보 사각화 기술 • 시각화 도구 등

빅데이터 처리

1. 빅데이터와 통계

1) 빅데이터는 어디서 수집되는가?

- 대부분의 기업들은 빅데이터 활용 시, 필요한 인사이트를 확보하기 위해 내부 Data를 분석하는데 초점이 맞춰져 있지만 소셜 미디어와 같은 인터넷 공간의 Data를 주목해야 시장의 흐름을 파악할 수 있음



2) 통계의 필요성

① 설득의 자료

② 사건 원인 분석

③ 미래 예측

평균	표준편차	분산
회귀 분석	상관 분석	교차 분석

1. 빅데이터와 통계

3) 통계학

- 통계학이란?

① 수량적 비교를 기초로 하여, 많은 사실을 통계적으로 관찰하고 처리하는 방법을 연구하는 학문으로 정의

② 빅데이터 기획과 모아진 Data를 정제하기 위해 필요한 학문

③ 통계 수치를 구할 수 있는 언어로는 R과 Python

2. 빅데이터 도구

1) R 프로그래밍

- R 프로그래밍 언어(줄여서 R)는 통계 계산과 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경임
- 뉴질랜드 오클랜드 대학의 로버트 젠틀맨(Robert Gentleman)과 로스 이하카(Ross Ihaka)에 의해 시작되어 현재는 R 코어팀이 개발
- R은 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 있으며, 패키지 개발이 용이하여 통계학자들 사이에서 통계 소프트웨어 개발에 많이 쓰임
- 1993년 처음 공개됨
- R은 그동안 일부 통계학자들만 애용하는 마이너한 언어였는데, 최근 데이터 분석 붐이 일면서 3~4년 사이 그 인기가 매우 높아짐

2. 빅데이터 도구

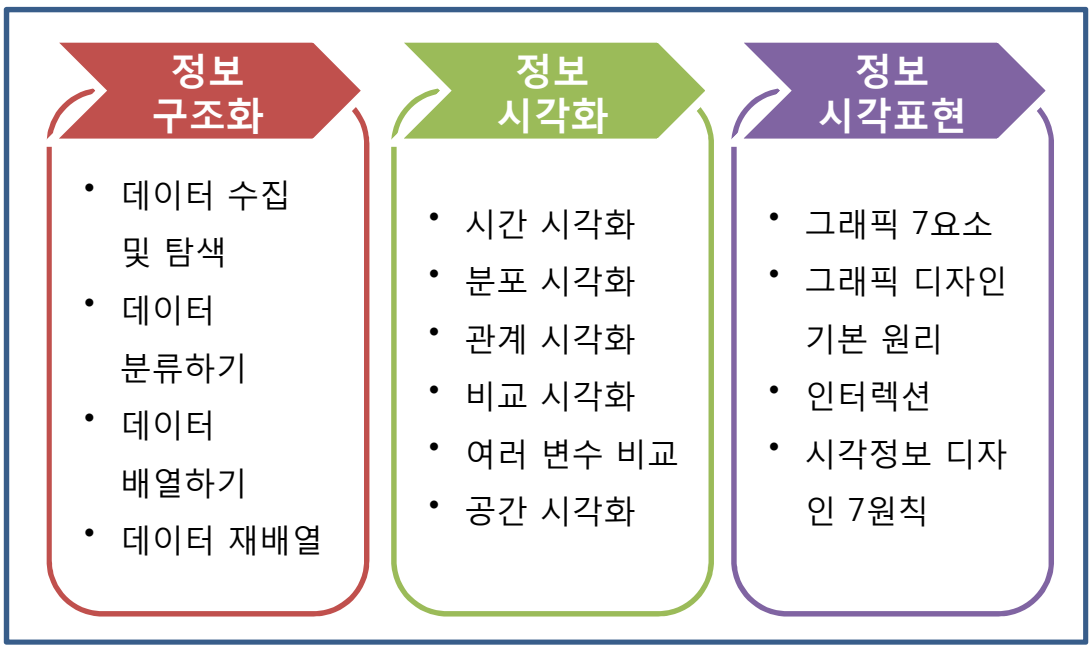
2) Python 프로그래밍

- 파이썬(Python)은 1991년프로그래머인 귀도 반 로섬 (Guido van Rossum)이 발표한 고급 프로그래밍 언어
- 독립적인 플랫폼이며 인터프리터식, 객체지향적, 동적 타이핑 (dynamically typed) 대화형 언어
- 파이썬이라는 이름은 귀도가 좋아하는 코미디 〈Monty Python ' s Flying Circus〉에서 따온 것
- 데이터 분석을 위해 파이썬을 사용하기 위해서 Ipython, Numpy, Scipy, Pandas, Matplotlib, Beautiful Soup 등 다양한 라이브러리들을 사용

3. 빅데이터 시각화

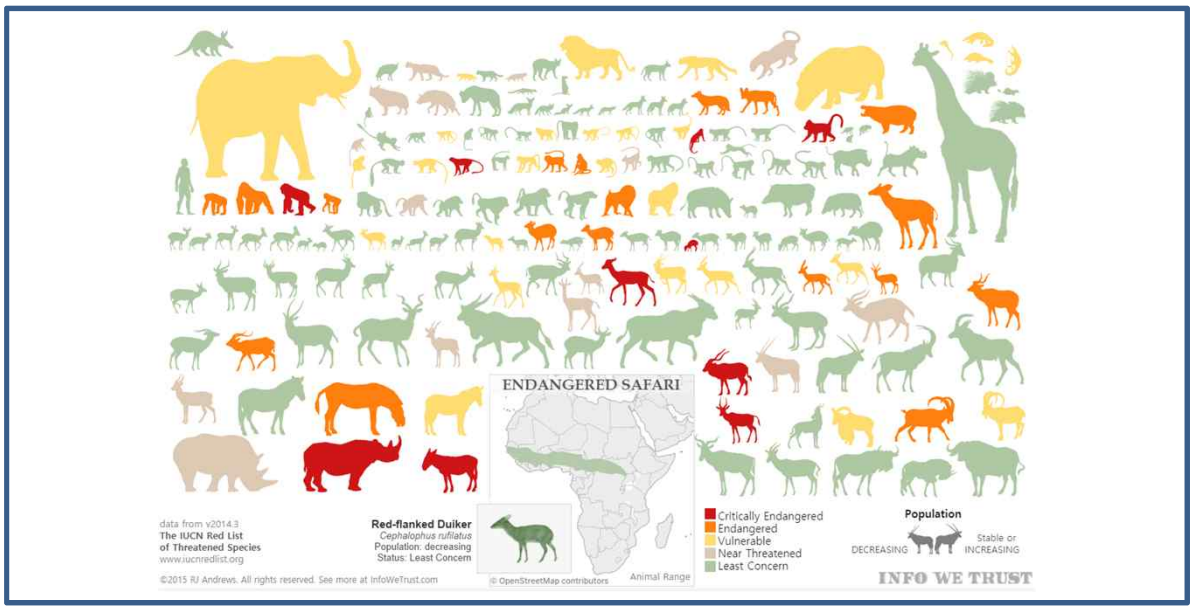
1) 시각화 방법론

- 정보 구조화, 정보 시각화, 정보 시각표현의 3단계



3. 빅데이터 시각화

2) 시각화 사이트 예: 멸종 위기의 사파리



- 몸집이 큰 아프리카 포유류의 개체 수에 대한 안정성과 멸종 위기 상태
- 총 다섯 가지 색으로 멸종위기 단계를 구분하며 포유류의 이미지를 클릭하면 서식지가 표시됨

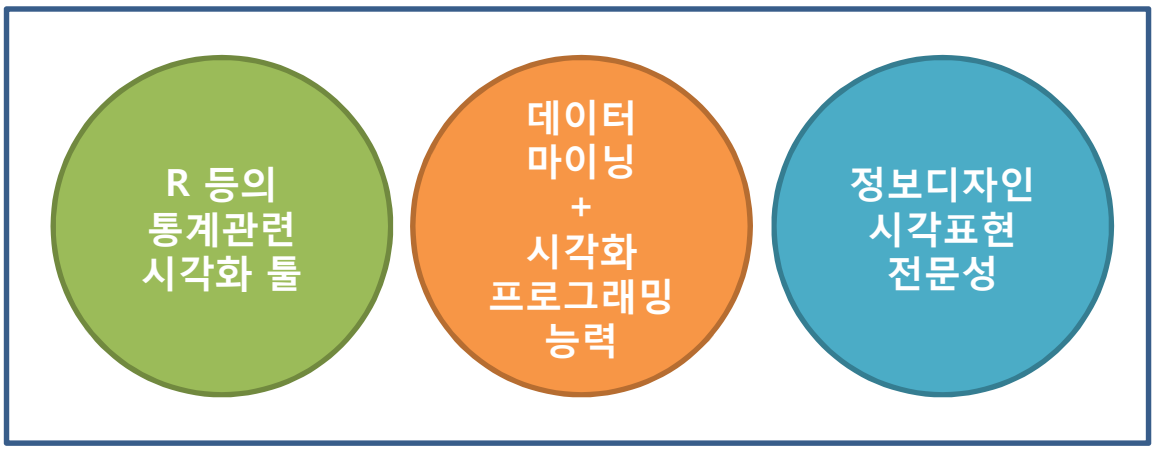
3) 시각화 도구 예: 태블로(Tableau)

- 2003년 1월 미국 캘리포니아의 Tableau Softwar사에서 개발
- 일반 사람들이 사각화를 통해 Data 를 이해할 수 있게 만들자라는 목적으로 만들어 짐
- MySQL,MS SQL, ORACLE등의 OLAP 데이타 백앤드에 연결해서 AD HOC 방식으로 리포트를 뽑아낼 수 있는 BI 리포팅 도구
- AD HOC방식이란, OLAP의 필드를 가지고, X,Y,Z 측으로 지정하여 분석하고 리포트나 그래프 등을 표현할 수 있는 툴

3. 빅데이터 시각화

4) 시각화 요건

- R, Python, 태블로 등의 통계 관련 시각화 도구 및 프로그래밍 능력과 더불어 시각화 표현이 전문성을 갖추는 것이 빅데이터 시각화 질을 향상
- 빅데이터에 대한 준비과정에는 상당히 많은 내용이 포함



핵심정리

1. 빅데이터 니즈(Needs)

1) 빅데이터의 현재

- 사람들의 일상, 기업, 가정에서 많은 데이터들이 생성되고 공유하고 있으며 이런 세계에서 데이터의 양은 폭발적으로 증가함

2) 빅데이터에 대한 질문

- 방대한 데이터 속에서 필요로 하는 데이터를 수집하기 위해서는 내가 필요로 하는 데이터의 정의가 필요함

3) 역사 속 빅데이터

- 나이팅게일의 로즈 다이어그램은 통계적인 자료를 다이어그램으로 보여준 사례로 역사적인 의미를 지님
- 존 스노의 콜레라 맵은 빅데이터와 첨단 기법을 통원해 발전시킨 것일 정도로 역학의 성립에 많은 기여함

2. 빅데이터 처리

1) 빅데이터 처리 과정 및 요소 기술

- 스마트폰의 보급으로 데이터가 매우 빠르게 축적되어 제타바이트 시대를 스마트 시대라고도 함
- 빅데이터는 데이터 소스를 수집, 저장, 분석하여 표현하는 절차를 거쳐 완성함

핵심정리

3. 빅데이터 준비

1) 빅데이터 수집 방법

- 기업에서 빅데이터를 활용 할 때, 내부 데이터 분석 뿐만 아니라 소셜 미디어와 같은 인터넷 공간의 Data 데이터를 주목해야 시장의 흐름을 파악
- 고객들의 키워드, 취향, 성별, 연령, 소득을 고려한 방대한 빅데이터를 수집해 면밀히 분석하고 고객들을 합리적인 소비로 유도함

2) 통계의 필요성

- 주어진 수많은 데이터틀에서 어떤 가치를 찾아낼 것인가에는 데이터 분석가의 경험과 다양한 통계 관련된 지식들이 필요함