동국대학교
dongguk university

데이터 플랫폼 이론-01

엄진영

빅데이터 시대에 필요한 것은 무엇일까

- **첫째, 데이터의 90%이상이 비정형 데이터라는 점에 주목**
  - 데이터 분석을 통해 답을 얻으려면 이미지와 MP3, 동영상 파일을 비롯해 소셜 네트워크 서비스에서 끊임없이 올라오는 소식 등의 데이터에 적극 주목

- **둘째, 빅데이터의 특성인 속도와 크기, 다양성 때문에 빅데이터 솔루션의 실질적인 활용이 쉽지 않음**
  - 정형은 물론 비정형의 다양한 데이터를 그것도 엄청난 양으로 존재하는 데이터를 빅데이터 솔루션을 활용해 분석하려면 데이터를 빠르게 처리할 수 있어야 함
  - 현재 시중에 많은 빅데이터 활용 솔루션은 모든 문제를 단번에 해결할 수 있는 특효약이라고 스스로 주장 → 빅데이터 분석의 어려움을 해결할 만큼의 규모나 정교함은 갖추지 못함

➔ 빅데이터 분석을 위해서는 다양한 데이터 소스에서 수집한 데이터를 처리하고 분석하여 지식을 추출하고 이를 기반으로 지능화된 서비스를 제공하는데 필요한 IT 환경

빅데이터 플랫폼을 준비

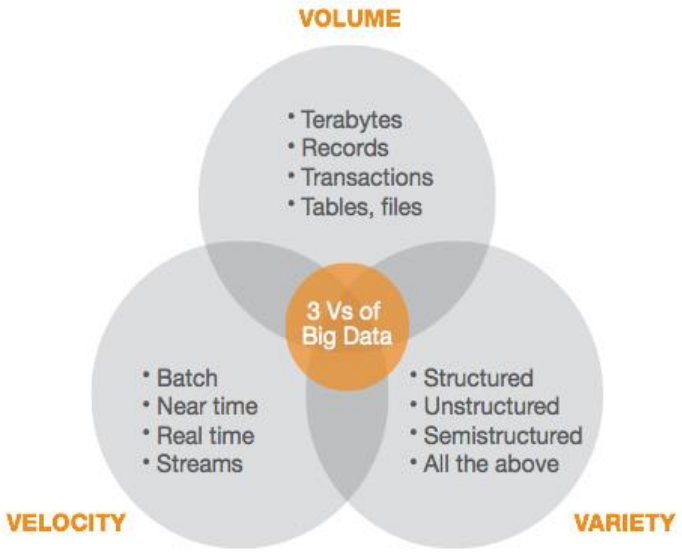
빅데이터 플랫폼 기술

빅데이터 플랫폼
= 빅데이터 처리 플랫폼 기술 + 빅데이터 인프라 기술

Ex) Google : Page Ranking, Google Analytics,
Google Earth, Personalized Search, 맞춤형 배너 광고

빅데이터 ?

- **큰 데이터**
  - 부피가 크고, 변화의 속도가 빠르며, 속성이 매우 다양한 데이터
- **기존 데이터베이스 관리 도구의 데이터 수집 · 저장 · 관리 · 분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술 [복잡성 포함]**
- **3V 로 설명**
  - **Volume(규모), Velocity(속도), Variety(다양성)**



빅데이터를 이해하는 3가지 질문

- **빅데이터로 무엇을 할 수 있을까?**
  - 빅데이터 자체로는 아무 것도 할 수 없다.
  - 분석과 가공의 과정을 거쳐서 새로운 가치를 찾았을 때 빅데이터는 의미가 있다.

- **빅데이터는 새로운 것인가?**
  - 빅데이터는 새로운 것이 아니다.
  - 과학기술의 발전으로 데이터베이스 기술과 데이터 처리기술의 비용 감소로 재조명 된것이다.

- **빅데이터의 예측은 정확한가?**
  - 빅데이터 예측은 과거의 자료를 바탕으로 유사하거나 통계적으로 발생할 확률이 높은 것을 예측
  - 빅데이터의 예측은 관련 데이터가 존재할 때 가능

사람의 눈 vs. 기계의 눈

- **기계가 사람대신 기사를 쓴다? 기계와 사람의 대결?**
  - 빅데이터는 지식이라는 키워드를 널리 사용하게 하는 역할
  - 지식을 생산하는 시스템으로 보기는 어려움

- **왜 사람들이 생각하는 기계에 관심을 보일까?**
  - 기계는 같은 일을 하면 항상 똑같이…
  - 단순하고 반복적인 일을 잘함!!
  - 아무리 반복해도 좋아지지 않음!!
  - 기계가 사람처럼 계속 변화할 수 있도록 방식을 바꾸고 내부적으로 과거의 경험을 지속적으로 반영
  - 사람 ➔ 작은 데이터로 자의적으로 해석
  - 기계 ➔ 많은 데이터로 정확한 기계적 통계 바탕의 결론을 제공

집단지성과 크라우드소싱

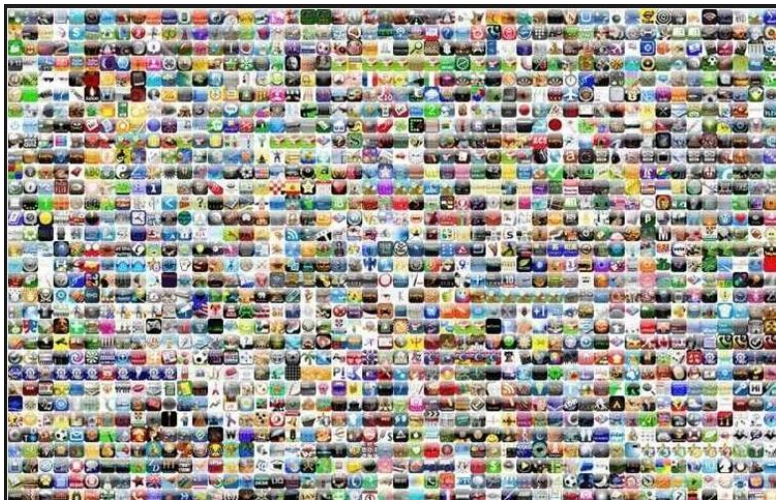
- **왜 데이터가 모이면 더 정확하고 심지어 미래를 예측하는 것도 가능할까?**
  - 빅데이터에 내재되어 있는 가치는 3V중에서 규모(Volume)의 특징을 반영한 집단지성과 다양성(Variety)의 특징을 반영한 크라우드소싱에 근거하여 가치를 창출

최고의 데이터 생산자 : IOT(사물인터넷)

- 사물인터넷(Internet of Things)은 사람, 사물, 공간 등 모든 것들(Things)이 ID를 가지고 다양한 센서와 통신방법을 통하여 인터넷(IP)으로 연결되어, 모든 것들에 대한 정보가 수집되고 공유.교환되어 위치와 환경의 모니터링이 가능한 것

- 사물인터넷이라는 용어의 사용 범위는 M2M(Machine to Machine), IoT를 거쳐 IoE (Internet of Everything)로까지 확장

최고의 데이터 생산자 : 모바일 기기와 SNS



플랫폼?

- Plat + form = [구획 정리된 땅 + 형태] = 용도에 따라 다양한 형태로 사용될 수 있는 공간
- **종합운동장의 장, 놀이터의 터 , 기차역, 연단, 도약 발판, 시장**
- **유튜브**
- **컴퓨터 시스템, 소프트웨어 개발환경, 실행환경**
- **시스템이란 하나의 공통적인 목적을 수행하기 위해 조직화된 요소들의 집합체, 플랫폼은 여러 개의 공통적인 목적을 수행**

하드웨어 플랫폼 종류

- 실제 생활의 다양한 부분에서 '물리적 구조물 ' 이란 의미로 사용

다이빙 플랫폼

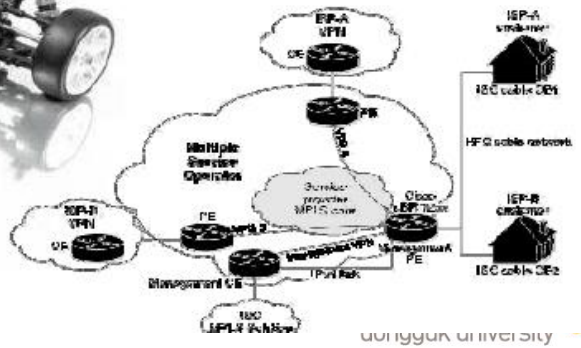
Platform 9 ¾ (해리포터)

IBM Mainframe 플랫폼

iPhone, 세계 최고의 개발자 플랫폼 (apple.com)

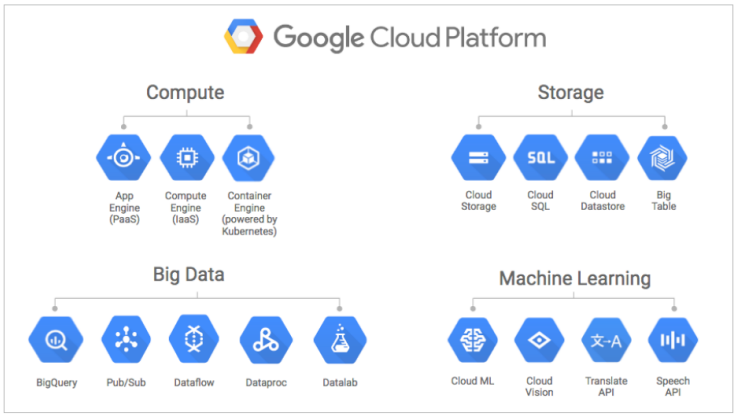
옵티마, 소나타와 플랫폼 공유

네트워크는 Contents Delivery Platform



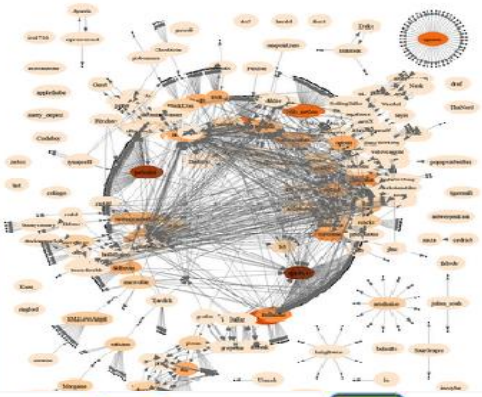
소프트웨어 플랫폼 종류

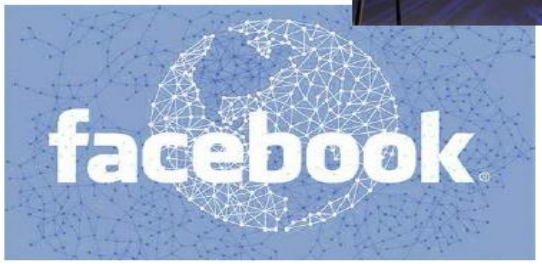
- 다양한 application이 작동하는 기반이 되는 OS 소프트웨어
- 여러가지 기능을 제공해주는 공통 실행환경

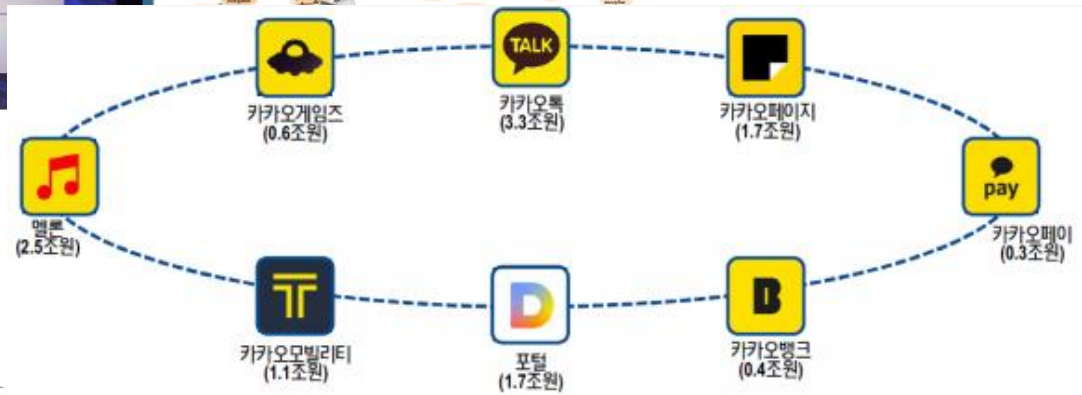


서비스 플랫폼 종류

- **Application 중 특정 서비스는 플랫폼으로 진화**
  - Facebook과 Twitter가 대표적 사례
- **플랫폼화 되면서 자체 생태계 구축**







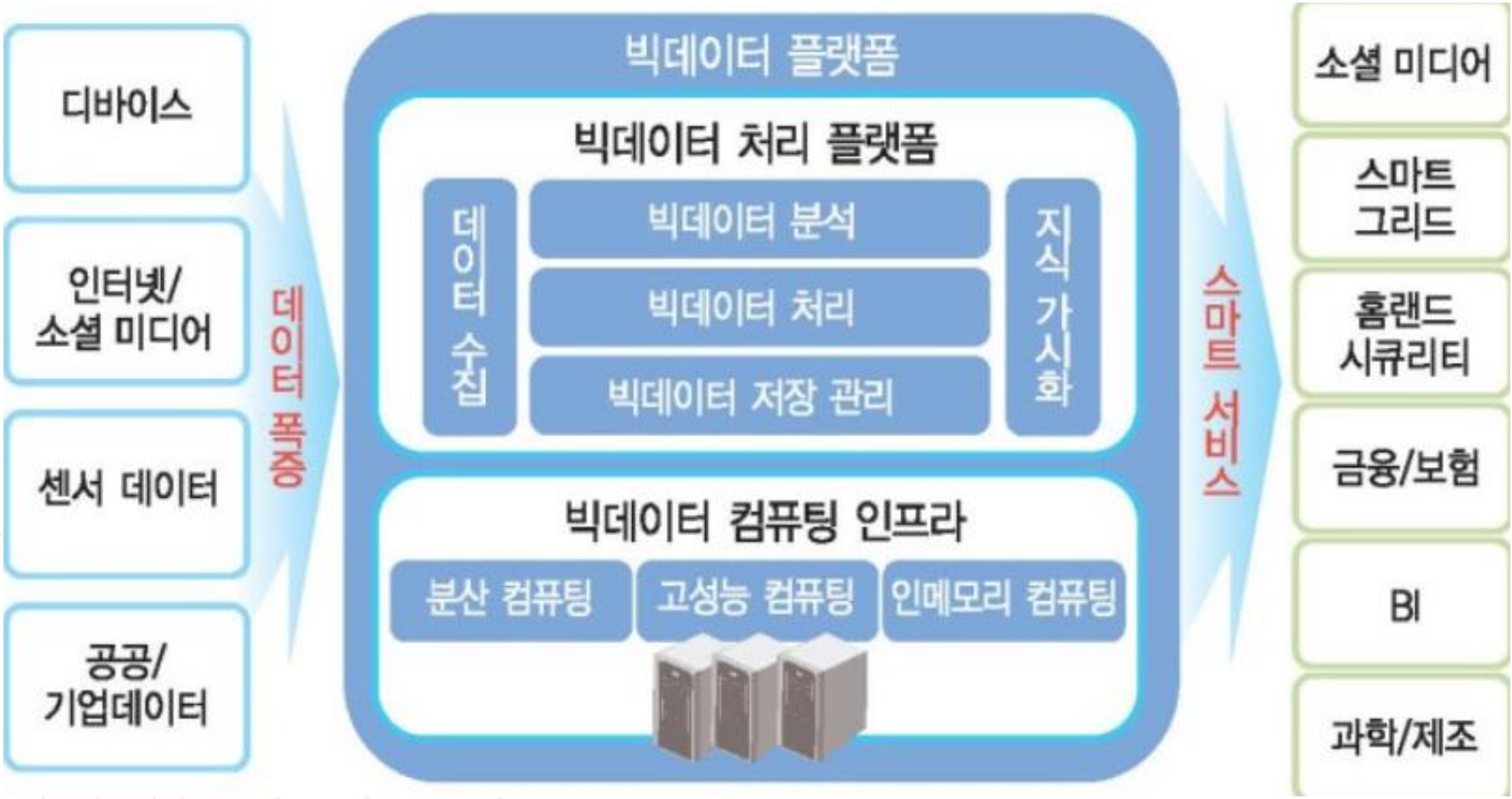
빅데이터 플랫폼

- **빅데이터 기술을 잘 사용할 수 있도록 준비된 환경**
  - 빅데이터 세부 기술의 집합체
- **빅데이터 플랫폼 구축**
  - 하드웨어 인프라 구축
    - ✓ 빅데이터 수용 용량 및 처리, 분석 작업에 대한 부하 등을 감안
  - 소프트웨어 구축
    - ✓ 분석에 필요한 수집, 관리, 분석, 이용자 환경 등에 관련한 소프트웨어를 구축

빅데이터 처리 단계


데이터 소스
수집
저장
처리
분석
표현
내부 데이터
외부 데이터
미디어
수동
자동
로그 수집기
크롤링
센싱
정형·비정형
데이터 등급
장치
일괄 처리
실시간 & 배치
NoSQL
관계형DB
분산 병렬 처리
전처리
분석 방법
Machine Learning

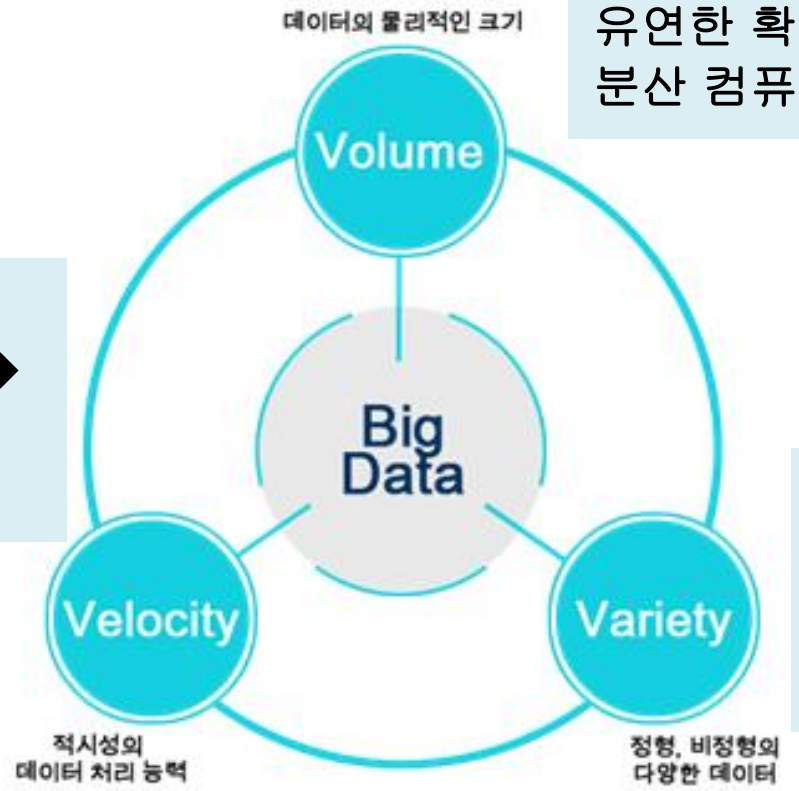
빅데이터 플랫폼


디바이스
인터넷/
소셜 미디어
센서 데이터
공공/
기업데이터
데이터 폭증
빅데이터 플랫폼
빅데이터 처리 플랫폼
데이터 수집
빅데이터 분석
빅데이터 처리
빅데이터 저장 관리
지식 가시화
빅데이터 컴퓨팅 인프라
분산 컴퓨팅
고성능 컴퓨팅
인메모리 컴퓨팅
스마트 서비스
소셜 미디어
스마트 그리드
홈랜드 시큐리티
금융/보험
BI
과학/제조

더 빠르게 더 많은 데이터 활용을 위해

- Twitter, Google ➔ 대규모 클러스터 시스템 구축 운영
- Paypal ➔ 슈퍼컴퓨터 사용

대용량 데이터 ➔
유연한 확장성 ➔
분산 컴퓨팅



비정형 데이터 ➔
고속 데이터 처리 ➔
분산 컴퓨팅,
슈퍼 컴퓨팅

데이터 폭증 ➔
지연 최소화 ➔
인메모리 컴퓨팅,
슈퍼 컴퓨팅

빅데이터 컴퓨팅 인프라

- **빅데이터의 가치는 대량의 다양한 데이터를 고속으로 처리할 수 있는 고 확장성, 고 성능의 컴퓨팅 인프라를 확보하지 않고는 얻을 수 없음**

- **분산 컴퓨팅**
  - 업무를 쪼개어 여러 노드에 분배하여 처리하는 기술, 노드에 장애가 발생했을 때 장애 노드의 작업 내용을 복구하는 기술 등을 통해 시스템의 확장성과 가용성을 제공하는 기술
  - **컴퓨터 클러스터**: 같은 공간내의 서버들을 네트워크 장비로 연결하여 구성한 컴퓨터 시스템으로 용량 확장이 필요할 때마다 쉽게 노드 추가가 가능
  - **컴퓨터 클러스터 기반 분산 컴퓨팅** : 네트워크에 연결된 여러 노드의 처리능력을 이용하여 대규모 문제를 해결하려는 분산처리 모델

빅데이터 컴퓨팅 인프라

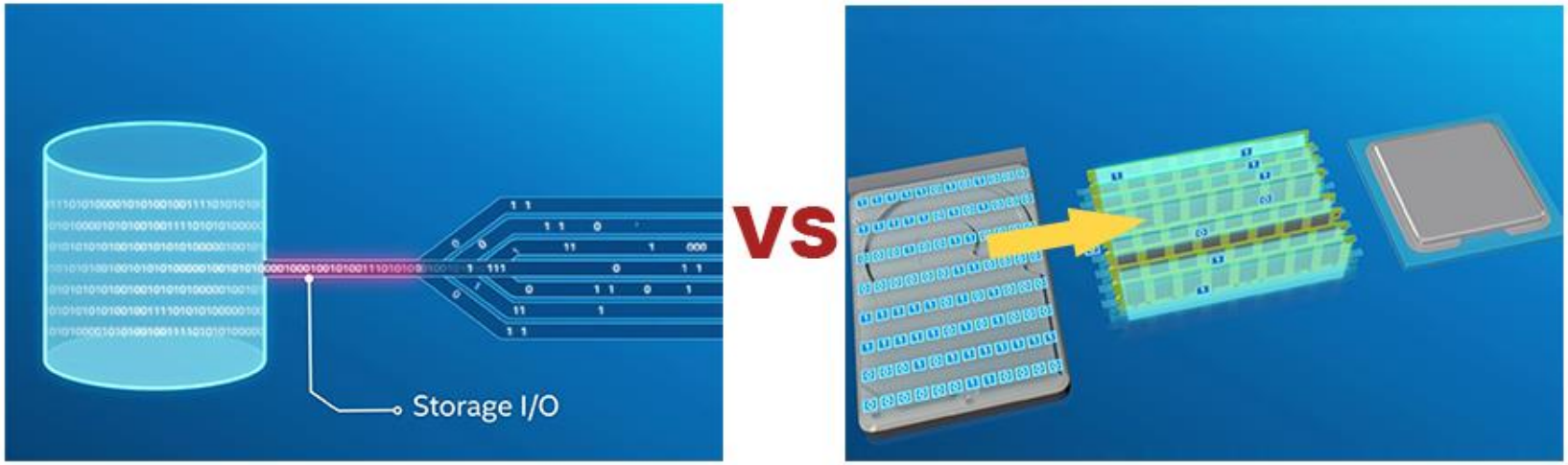
- **고성능 컴퓨팅**
  - 다수의 프로세서 혹은 노드들을 연결하여 고도의 연산 능력을 갖춘 컴퓨터 시스템을 구성하고 이를 최대한 활용하여 고속 처리를 제공하기 위한 분산·병렬 컴퓨팅이 중요
  - 고성능 컴퓨팅을 위한 컴퓨터 시스템 구성
    - ✓ 클러스터 컴퓨터처럼 각 노드의 역할이 사전에 정의되지 않은 범용 노드들을 연결하여 구성
    - ✓ 컴퓨트 노드, 스토리지 노드, 관리 노드 등 노드의 역할을 사전에 정하고 이에 따라 하드웨어를 최적으로 구성
    - ✓ 여러 노드들을 통합하여 하나의 시스템인것처럼 제공하기 위해 다양한 기술을 활용하여 구성

2019년 한국과학기술정보연구원(KISTI)이 보유한 슈퍼컴퓨터 5호기 '누리온'이 세계 15위 슈퍼컴퓨터에 뽑혔다. '누리온'은 연산 속도가 25.7페타플롭스(PF)에 달하고 계산노드는 8437개다. 1PF는 1초에 1000조번 연산이 가능하다. 70억 명이 420년 걸려 마칠 계산을 1시간 만에 끝낼 수 있는 수준이라고 KISTI는 설명했다. 기상청이 보유한 슈퍼컴퓨터 '누리'와 '미리'가 각각 99위와 100위에 랭크됐다.

'톱 500'중 성능은 미국이 38.5%로 1위, 중국이 29.9%로 2위를 차지했다. 두 나라를 합치면 전체의 약 70%에 달한다.

빅데이터 컴퓨팅 인프라

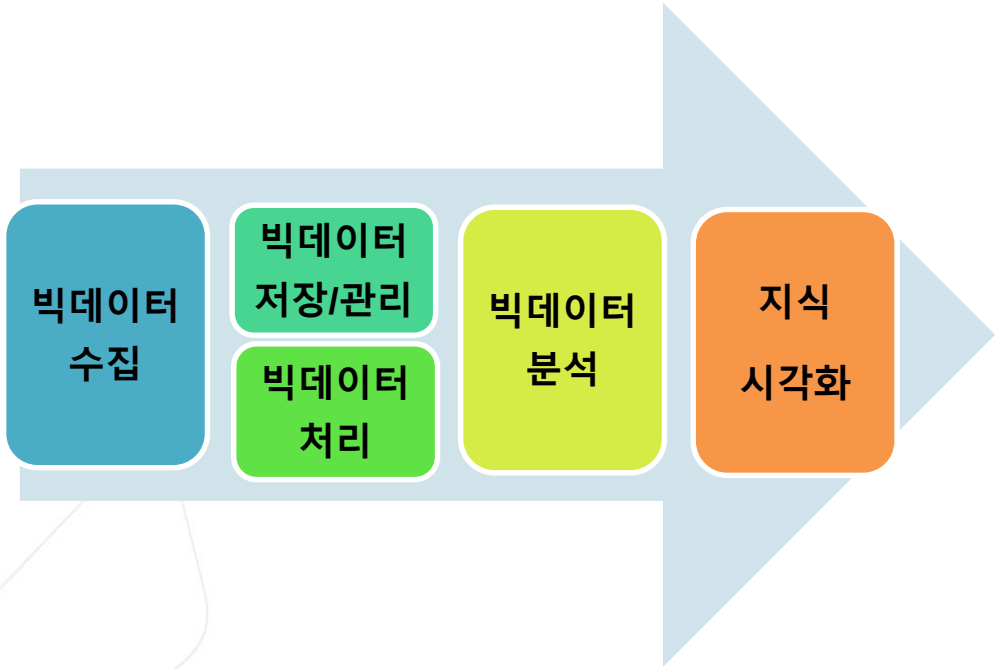
- **인메모리 컴퓨팅**
  - 컴퓨팅 속도저하의 근본원인인 입출력(I/O)을 개선하는 데 중점을 두어, 디스크(보조저장장치)가 아닌 메모리 상에 데이터를 저장해 두고 처리하는 것
    - ✓ 하드 디스크와 일반 메모리를 비교했을때 입출력 속도 차이는 실제 10만배 이상 차이



Storage Computing

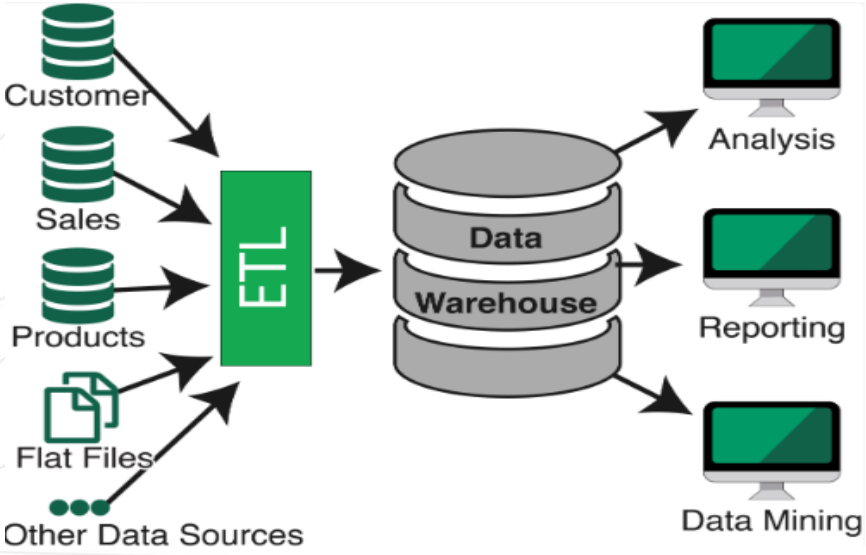
In-Memory Computing

빅데이터 활용을 위한 요소 기술들



빅데이터 플랫폼 요소기술1 – 빅데이터 수집 기술(1)

- **데이터 소스의 위치에 따라 내부 데이터 수집과 외부 데이터 수집으로 구분**
  - 내부 데이터 수집
    - ✓ 자체적으로 보유한 내부 파일 시스템이나 데이터 베이스 관리시스템, 센서 등에 접근하여 데이터를 수집하는 것을 의미
    - ✓ ETL(Extraction, Transformation, Loading)
      - 다양한 소스 시스템으로부터 필요한 데이터를 추출(extract)하여 변환(transformation) 작업을 거쳐 저장하거나 분석을 담당하는 시스템으로 전송 및 적재(loading)하는 모든 과정을 포함
      - 일관성 확보를 위한 정제의 예 : '남', 'M', 'man', 'male', '1' ➔ 'M'으로 표현



빅데이터 플랫폼 요소기술1 – 빅데이터 수집 기술(2)

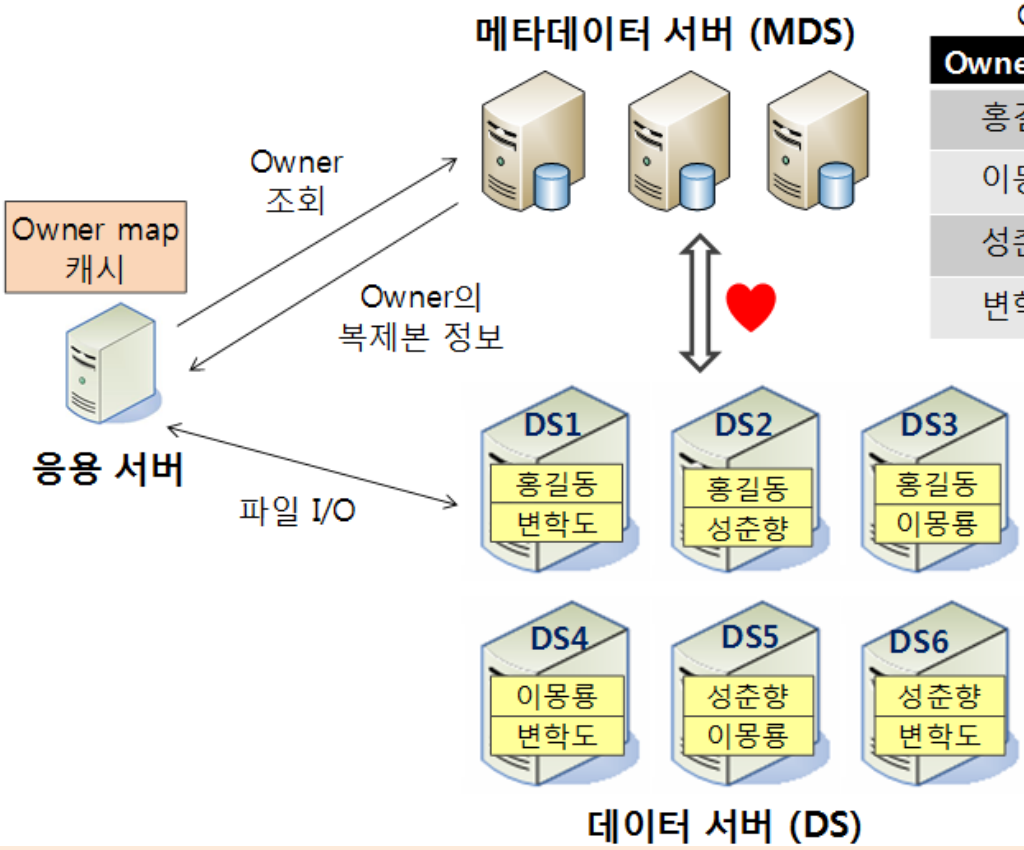
- **데이터 소스의 위치에 따라 내부 데이터 수집과 외부 데이터 수집으로 구분**
  - 외부 데이터 수집
    - ✓ 인터넷으로 연결된 외부에서 데이터를 수집하는 것을 의미
    - ✓ 크롤링 엔진(Crawling Engine)
      - 로봇이 거미줄처럼 얽혀있는 인터넷 링크를 따라다니며 방문한 사이트의 모든 페이지의 복사본을 생성함으로써 문서를 수집

빅데이터 플랫폼 요소기술2 – 빅데이터 저장 관리 기술(1)

- **기존 데이터 저장관리 기술 + 3V 특성 고려**
  - Volume(규모), Velocity(속도), Variety(다양성)
  - 대용량의 다양한 형식을 가진 데이터를 고성능으로 저장하고 검색할 수 있도록 하는 기술
  - 빅데이터가 가지는 대용량, 비정형, 실시간성이라는 특징을 수용할 수 있어야 함
- **데이터 양이 방대**
  - 하나의 노드 능력을 최대화시키는 스케일-업(Scale-up) 방법만으로는 부족
- **데이터 유형이 다양**
  - 기존의 행과 열로 구성된 정형데이터를 위한 관계형 모델에 기반한 방법만으로는 다양한 데이터 유형을 저장하기에는 부적합
- **데이터 생성속도가 빨라짐**
  - 스케일-업 기술과 하드디스크에 기반한 기술이 한계

빅데이터 플랫폼 요소기술2 – 빅데이터 저장 관리 기술(2)

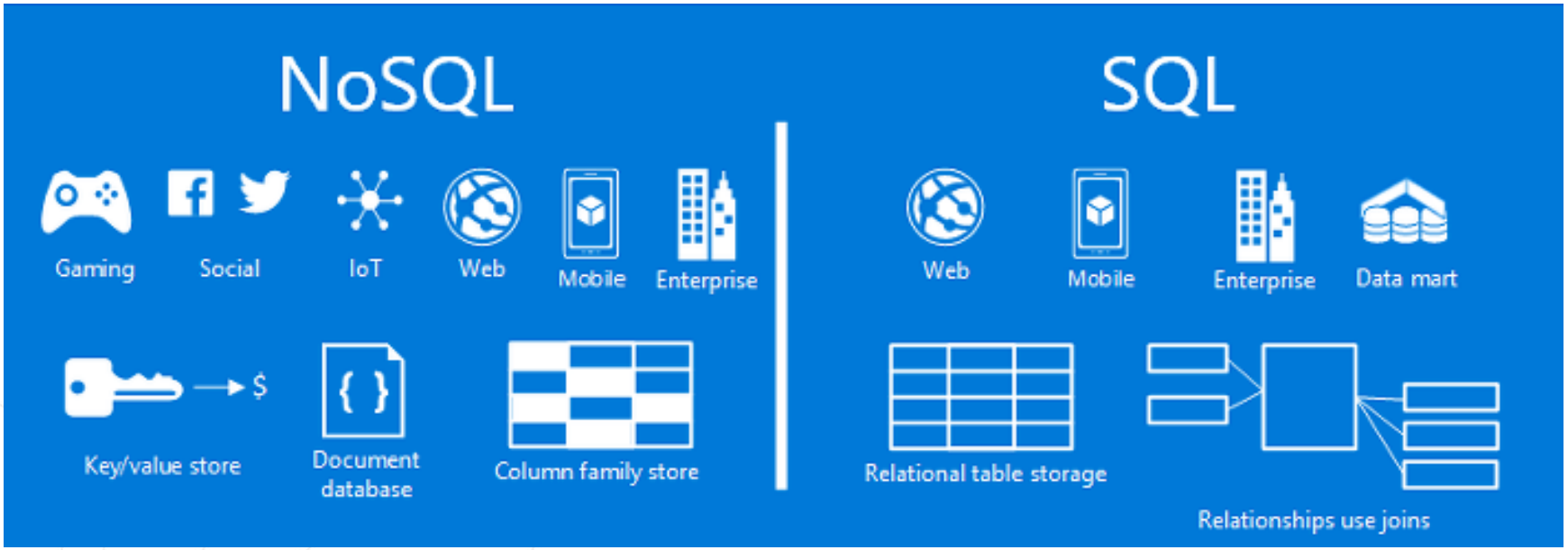
- **분산파일 시스템**
  - 분산 파일 시스템은 막대한 양의 데이터를 저장하고 관리하기 위해 수 많은 서버들에 데이터를 나누고 저장하고 관리하는 파일 시스템



OwFS(Owner-based File System) : NHN이 자체적으로 개발한 분산 파일 시스템

빅데이터 플랫폼 요소기술2 – 빅데이터 저장 관리 기술(3)

- **빅데이터베이스 관리 기술**
  - NoSQL (Not only SQL)
    - ✓ SQL만을 사용하지 않는 데이터베이스 관리 시스템(DBMS)



빅데이터 플랫폼 요소기술2 – 빅데이터 저장 관리 기술(4)

- **NoSQL의 특징**
  - 스키마가 없는 Schema-less 데이터베이스

RDBMS

이름	성별	나이	직업
IML	남	22	Hacker
KIM	남	21	None

NoSQL

이름: "IML"	성별: "남"	나이: 22	직업:"Hacker"
이름:" KIM"	나이: 21	거주지역:"서울 "	

  - 대부분의 NoSQL은 오픈소스

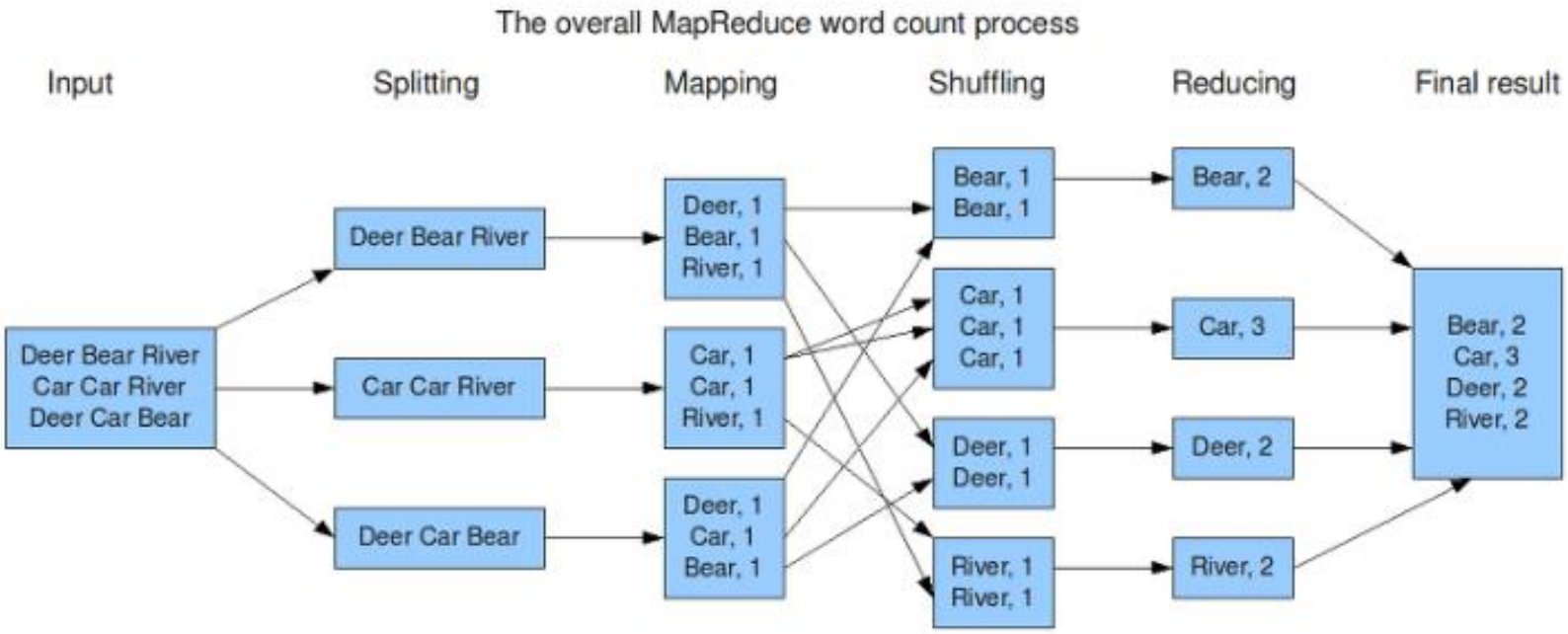
  - 분산 환경으로 서버 구동 중 발생하는 부하를 분산시킬 수 있음

빅데이터 플랫폼 요소기술3 – 빅데이터 처리 기술(1)

- **빅데이터 일괄처리 기술**
  - 빅데이터는 그 규모가 엄청나서 아무리 고가 서버라도 단일 시스템에서 처리하면 처리하는 양보다 오히려 새로 생성되어 쌓이는 데이터가 더 많다?!?!
  - **데이터를 적시에 처리하기 위해서는 쌓인 빅데이터를 여러 서버로 분산해 각 서버 에서 나눠서 처리하고 이를 다시 모아서 결과를 처리하는 분산/병렬 기술 방식을 취한다**

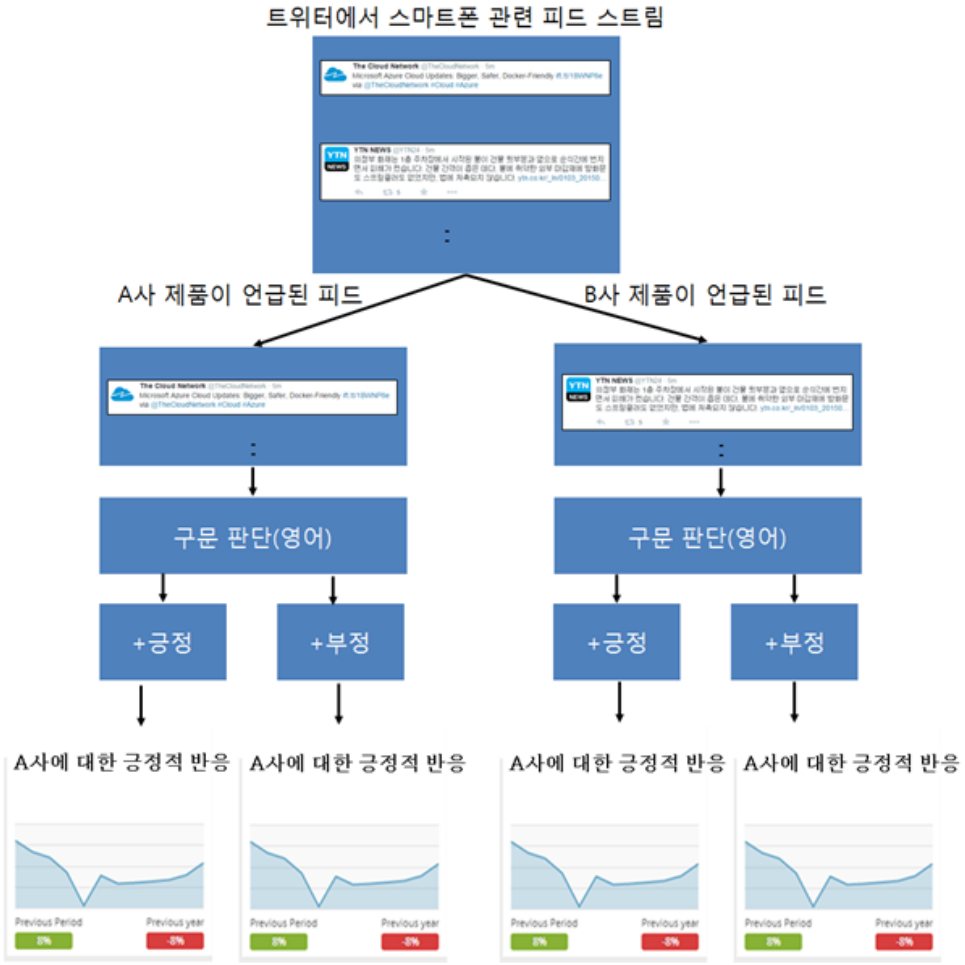
빅데이터 플랫폼 요소기술3 – 빅데이터 처리 기술(2)

- **MapReduce**
  - 빅데이터를 분산 처리하는 맵(map)이라는 기능과 분산 처리된 중간 결과를 모아서 정리하는 리듀스(Reduce)라는 기능을 이용해서 데이터를 병렬로 고속 처리하는 기술



빅데이터 플랫폼 요소기술3 – 빅데이터 처리 기술(3)

- **빅데이터 실시간 처리 기술**
  - 이벤트 기반 실시간 처리 기술, 스트림 처리 기술
  - 끊임없이 입력되는 스트림 데이터를 적정 구간으로 나누어 처리하며, 스트림 데이터가 들어오는 대로 일련의 처리 업무들을 수행하여 그 결과를 연속적으로 제공
    - ✓ 최신 데이터를 기반으로 바로 결과를 얻을 수 있음
    - ✓ 전체 데이터가 처리되기 전이라도 중간 처리 결과를 먼저 제공 ➔ 조기 처리 결과를 얻을 수 있음
  - IBM infosphere streams, Apache Storm, S4



동국대학교 dongguk university

빅데이터 플랫폼 요소기술4 – 빅데이터 분석 기술(1)

- **텍스트마이닝(Text Mining)**
  - 자연어 처리 기술(NLP : Natural Language Processing)로 인간의 언어로 쓰인 비정형 텍스트에서 유용한 정보를 추출하거나 다른 데이터와의 연계성을 파악하면, 분류나 군집화 등 빅데이터에 숨겨진 의미있는 정보를 발견하는 분석 방법

- **웹 마이닝(Web Mining)**
  - 인터넷에서 수집한 정보를 데이터 마이닝 기법으로 분석

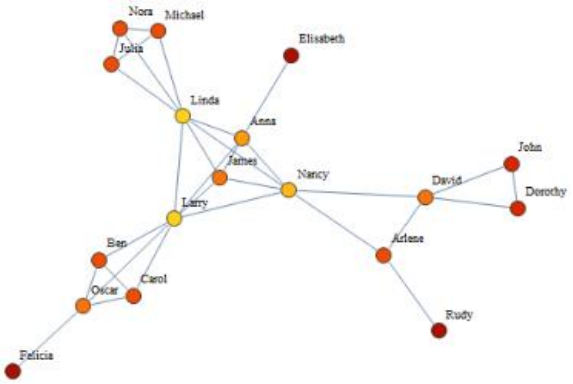
빅데이터 플랫폼 요소기술4 – 빅데이터 분석 기술(2)

- **오피니언 마이닝(Opinion Mining)**
  - 평판 분석, 마케팅에서는 버즈(Buzz, 입소문) 분석이라고도 불림
  - 다양한 온라인 뉴스, SNS, 사용자가 만든 콘텐츠에서 표현된 의견을 추출, 분류, 이해와 자산화하는 컴퓨팅 기술
  - 텍스트 속에서 여러 가지 감정 상태를 식별

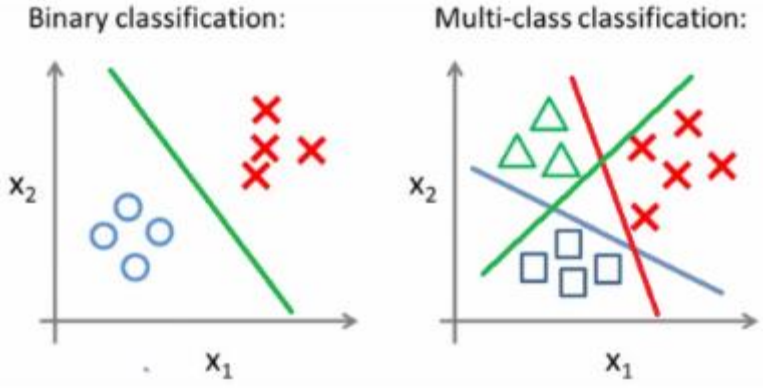
- **리얼리티 마이닝(Reality Mining)**
  - 휴대폰, 모바일 기기 등을 사용하여 인간관계와 행동 양태 등을 추론
  - 통화량, 통화 위치, 통화 상태, 대상, 내용 등을 분석하여 사용자의 인간관계, 행동 특성 등의 정보를 찾아냄

빅데이터 플랫폼 요소기술4 – 빅데이터 분석 기술(3)

- **소셜 네트워크 분석(Social network Analytics)**
  - 수학의 그래프 이론(Graph Theory)을 바탕으로 소셜 네트워크 서비스에서 소셜 네트워크 연결 구조와 연결 강도를 분석하여 사용자의 명성 및 영향력을 측정하는 기법

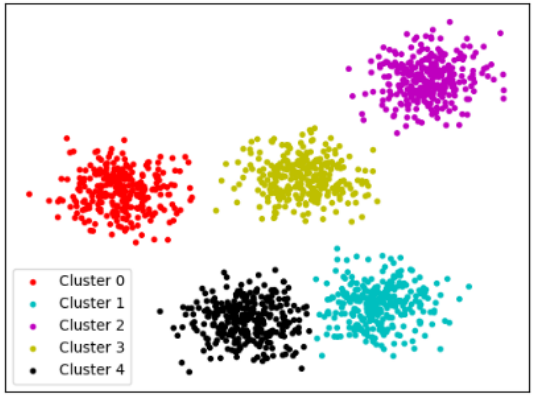


- **분류(Classification)**
  - 이미 알려진 클래스들로 구분되는 훈련 데이터군을 학습시켜 새로 추가되는 데이터가 속할 만한 데이터 군을 찾는 지도학습 방법
  - KNN(K-Nearest Neighbor), 인공신경망을 적용하는 방법 등

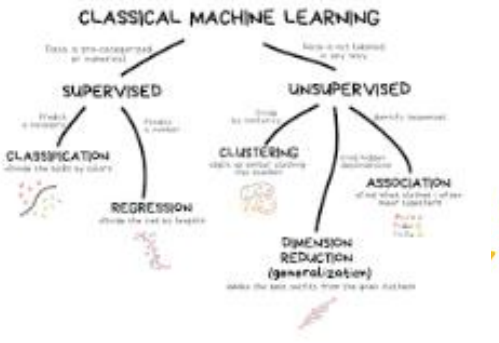


빅데이터 플랫폼 요소기술4 – 빅데이터 분석 기술

- **군집화(Clustering)**
  - 특성이 비슷한 데이터를 합쳐 군으로 분류
  - 훈련 데이터 군을 이용하지 않는 비지도 학습 방법을 사용
  - 관심사나 취미에 따라 군집으로 분류 가능



- **기계학습(Machine Learning)**
  - 인공지능 분야에서 인간의 학습을 모델링한 알고리즘
    - ✓ 흔히 아는 딥러닝 기법도 포함
  - 컴퓨터가 학습할 수 있도록 알고리즘과 기술을 개발하여 수신한 이메일의 스팸 여부와 같은 추상적이거나 특징을 추출하기 어려운 경우 신경망으로 판단할 수 있도록 학습(훈련)

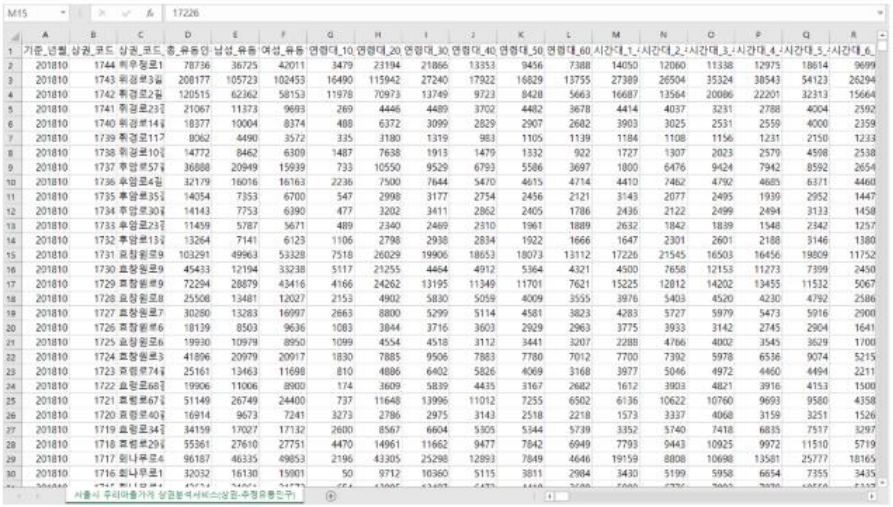


빅데이터 플랫폼 요소기술4 – 빅데이터 분석 기술

- **감정 분석(Sentiment Anaysis)**
  - 문장의 의미를 파악하여 글의 내용에 긍정/부정, 좋음/나쁨의 이진 분류를 하거나, 만족/중간/불만족 처럼 강도 지수화 가능
  - 고객의 감성 트랜드를 시계열로 분석하고 고객 감성 변화에 기업의 신속한 대응과 부정적인 의견의 확산을 방지하는 비즈니스 요구에 활용 가능

빅데이터 플랫폼 요소기술5 – 지식 시각화 기술(1)

- **왜 시각화를 해야할까?**
  - 많은 양의 데이터를 한눈에 볼 수 있다.



기준_년월	상권_코드	상권_코드_명	총_유동인구_수	남성_유동인구_수	여성_유동인구_수	연령대_10_유동인구_수	연령대_20_유동인구_수	연령대_30_유동인구_수	연령대_40_유동인구_수	연령대_50_유동인구_수	연령대_60_이상_유동인구_수	시간대_1_유동인구_수	시간대_2_유동인구_수	시간대_3_유동인구_수	시간대_4_유동인구_수	시간대_5_유동인구_수	시간대_6_
201810	1744	회우정로1	78736	36725	42011	3479	23194	21866	13353	9456	7388	14050	12060	11338	12975	18614	9699
201810	1743	회경로3길	208177	105723	102453	16490	115942	27240	17922	16829	13755	27389	26504	35324	38543	54123	26294
201810	1742	회경로2길	120515	62362	58153	11978	70973	13749	9723	8428	5663	16687	13564	20086	22201	32313	15664
201810	1741	회경로23길	21067	11373	9693	269	4446	4489	3702	4482	3678	4414	4037	3231	2788	4004	2592
201810	1740	회경로14길	18377	10004	8374	488	6372	3099	2829	2907	2682	3903	3025	2531	2559	4000	2359
201810	1739	회경로117	8062	4490	3572	335	3180	1319	983	1105	1139	1184	1108	1156	1231	2150	1233
201810	1738	회경로10길	14772	8462	6309	1487	7638	1913	1479	1332	922	1727	1307	2023	2579	4598	2538
201810	1737	후암로57길	36888	20949	15939	733	10550	9529	6793	5586	3697	1800	6476	9424	7942	8592	2654
201810	1736	후암로4길	32179	16016	16163	2236	7500	7644	5470	4615	4714	4410	7462	4792	4685	6371	4460
201810	1735	후암로35길	14054	7353	6700	547	2998	3177	2754	2456	2121	3143	2077	2495	1939	2952	1447
201810	1734	후암로30길	14143	7753	6390	477	3202	3411	2862	2405	1786	2436	2122	2499	2494	3133	1458
201810	1733	후암로23길	11459	5787	5671	489	2340	2489	2310	1961	1889	2632	1842	1839	1548	2342	1257
201810	1732	후암로13길	13264	7141	6123	1106	2798	2938	2834	1922	1666	1647	2301	2601	2188	3146	1380
201810	1731	효창원로9	103291	49963	53328	7518	26029	19906	18653	18073	13112	17226	21545	16503	16456	19809	11752
201810	1730	효창원로9	45433	12194	33238	5117	21255	4464	4912	5364	4321	4500	7658	12153	11273	7399	2450
201810	1729	효창원로9	72294	28879	43416	4166	24262	13195	11349	11701	7621	15225	12812	14202	13455	11532	5067
201810	1728	효창원로8	25508	13481	12027	2153	4902	5830	5059	4009	3555	3976	5403	4520	4230	4792	2586
201810	1727	효창원로7	30280	13283	16997	2663	8800	5299	5114	4581	3823	4283	5727	5979	5473	5916	2900
201810	1726	효창원로6	18139	8503	9636	1083	3844	3716	3603	2929	2963	3775	3933	3142	2745	2904	1641
201810	1725	효창원로6	19930	10979	8950	1099	4554	4518	3112	3441	3207	2288	4766	4002	3545	3629	1700
201810	1724	효창원로3	41806	20979	20917	1830	7885	9506	7883	7780	7012	7700	7392	5978	6536	9074	5215
201810	1723	효창루74길	25161	13463	11698	810	4886	6402	5826	4069	3168	3977	5046	4972	4460	4494	2211
201810	1722	효창로68길	19906	11006	8900	174	3609	5839	4435	3167	2682	1612	3903	4821	3916	4153	1500
201810	1721	효창로67길	51149	26749	24400	737	11648	13996	11012	7255	6502	6136	10622	10760	9693	9580	4358
201810	1720	효창로40길	16914	9673	7241	3273	2786	2975	3143	2518	2218	1573	3337	4068	3159	3251	1526
201810	1719	효창로34길	34159	17027	17132	2600	8567	6604	5305	5344	5739	3352	5740	7418	6835	7517	3297
201810	1718	효창로29길	55361	27610	27751	4470	14961	11662	9477	7842	6949	7793	9443	10925	9972	11510	5719
201810	1717	회나무로4	96187	46335	49853	2196	43305	25298	12893	7849	4646	19159	8808	10698	13581	25777	18165
201810	1716	회나무로1	32032	16130	15901	50	9712	10360	5115	3811	2984	3430	5199	5958	6654	7355	3435

엑셀 화면을 가득 채우고도 넘치는 양의 데이터를 간단한 선 차트로 요약하여 표현할 수 있습니다. (데이터 출처 : 서울 열린데이터 광장)

빅데이터 플랫폼 요소기술5 – 지식 시각화 기술(2)

- https://www.npr.org/sections/money/2015/05/18/404991483/how-machines-destroy-and-create-jobs-in-4-graphs

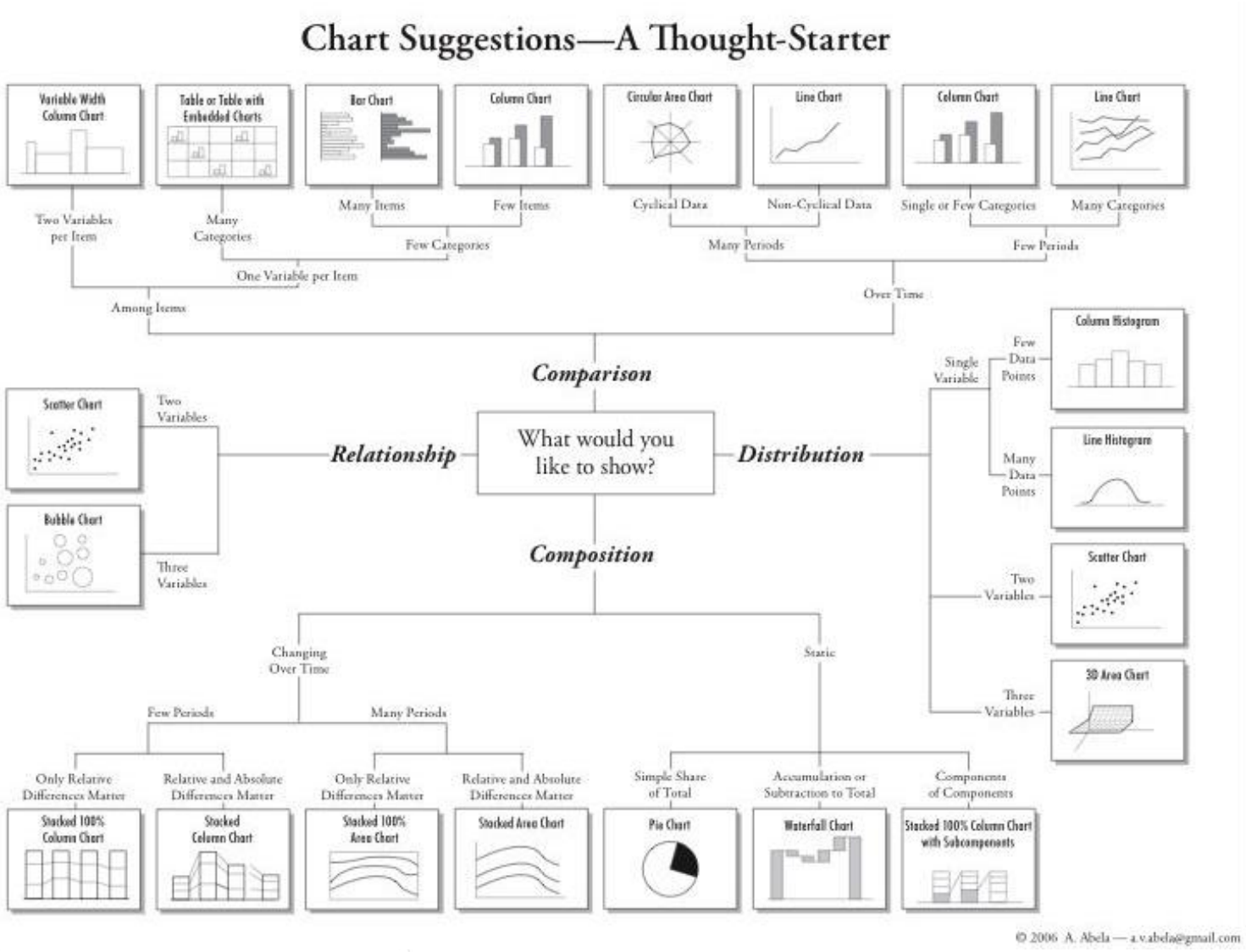
다양한 데이터 시각화 유형의 예 (출처 : datavizcatalogue.com)

빅데이터 플랫폼 요소기술5 – 지식 시각화 기술(3)

- **왜 시각화를 해야할까?**
  - 데이터 분석에 대한 전문지식이 없어도 누구나 쉽게 데이터 인사이트를 찾을 수 있다.
    - ✓ Ex) 데이터 시각화를 하면 우리나라 전국민 약 5천만 명의 우울증 진료현황도 빠르게 알 수 있다.(http://project.newsjel.ly/depressed/)

빅데이터 플랫폼 요소기술5 – 지식 시각화 기술(4)

- Dr. Andrew Abela가 고안한 비교, 관계, 분포, 구성기준에 따른 차트 선택 방법 (https://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html)



빅데이터 플랫폼 요소기술5 – 지식 시각화 기술(5)

- 주의 : 시각에 의해 오류를 범할 수 있다.
- 데이터 시각화의 일반적인 실수 7가지(https://thenextweb.com/dd/2015/05/15/7-most-common-data-visualization-mistakes/)글을 읽어보세요.

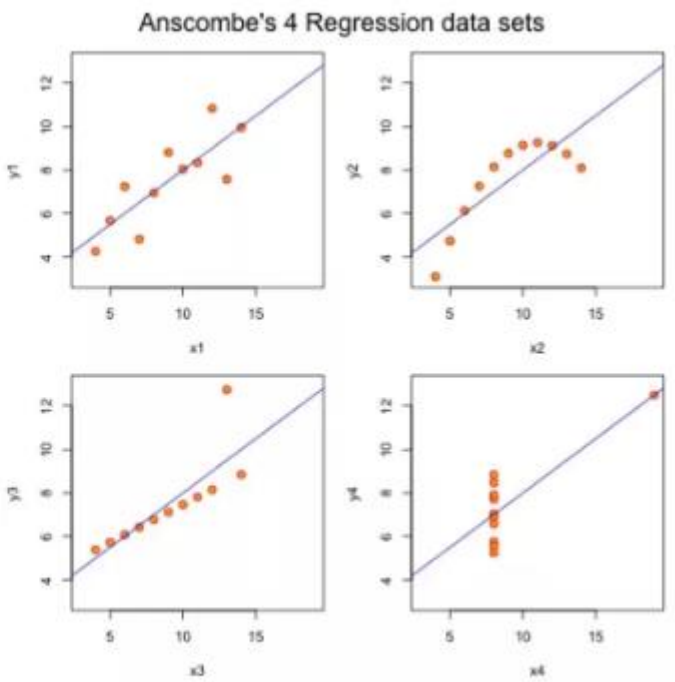
(좌) 오바마 케어 등록 현황을 보도한 Fox News의 시각화 차트 활용 오류, (우) 좌측 차트를 올바르게 수정한 형태

빅데이터 플랫폼 요소기술5 – 지식 시각화 기술(6)

- **왜 시각화를 해야할까?**
  - 요약통계보다 정확한 분석 결과를 도출할 수 있다.
    - ✓ 단순히 데이터 분석결과를 전달할 목적뿐만 아니라 정확한 분석을 위한 데이터 탐색 방법으로 활용

I		II		III		IV	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Mean of X	11.0	**Correlation between X and Y**	0.875
Variance of X	10.0	**Linear regression**	y=3.0+0.5x
Mean of Y	7.5		
Variance of Y	3.75		



1973년 F.J.Anscombe가 개발한 Anscomes' Quarter는 동일한 요약통계(평균, 표준편차, 상관관계_를 가진 4개의 데이터 셋을 산점도로 시각화 했을 때 시각적 패턴이 명확히 다르다는 것을 입증한다. 이는 요약 통계만으로 데이터를 정확하게 볼 수 없다는 것을 의미한다고 해석할 수 있다.

빅데이터 플랫폼 요소기술5 – 지식 시각화 기술(7)

- **왜 시각화를 해야할까?**
  - 효과적인 데이터 인사이트 공유로 데이터 기반의 의사결정을 할 수 있다.

- **빅데이터는 단순한 문자와 숫자의 나열이 아닌 의미와 진실이 숨어있음**
- **직설적인 표현+애매모호함 ➔ 이야기를 어떻게 끌어가서 다양한 시각화 도구로 표현하느냐에 따라 얻을 수 있는 직관이 달라짐**
- **시각화를 통해 데이터를 바라보는 관점이 시간, 분포, 관계, 비교, 공간에 일치할 수 있도록 표현할 수 있는 시각화 방법론+ 사용자의 상상과 창의력➔독창적인 결과**