IoT / Big data 분석 및 활용

빅데이터 개요

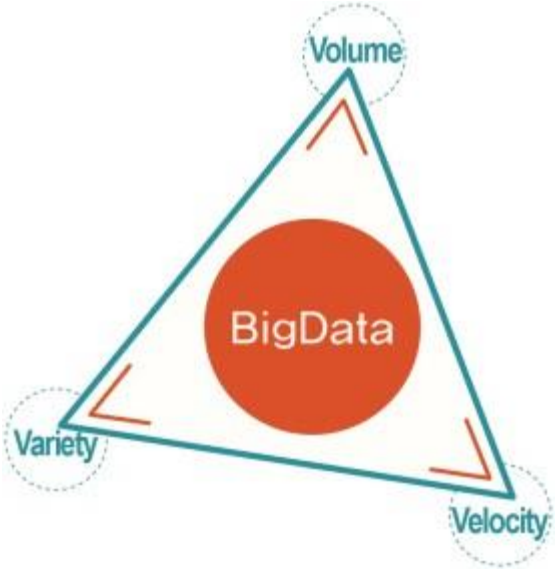
□ 빅데이터란

- 기존 데이터보다 너무 방대하여 기존의 방법이나 도구로 수집/저장/분석 등이 어려운 정형 및 비정형 데이터들을 의미
- 1분 동안 구글에서는 200만 건의 검색, 유튜브에서는 72시간의 비디오, twitter에서는 27만 건의 트윗이 생성됨
- 세계적인 컨설팅 기관인 맥킨지(Mckinsey)는 빅데이터를 기존 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석하는 역량을 넘어서는 규모로서 그 정의는 주관적이며 앞으로도 계속 변화될 것이라고 정의
- 빅데이터를 테라바이트 이상의 데이터라고 정의하기도 하며 대용량 데이터를 처리하는 아키텍처라고 정의하기도 함

빅데이터 개요

빅데이터의 특징

- 크기(Volume), 속도(Velocity), 다양성(Variety)
- 크기는 일반적으로 수십 테라바이트 또는 수십 페타바이트 이상 규모의 데이터 속성
- 속도는 대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성
- 융복합 환경에서 디지털 데이터는 매우 빠른 속도로 생산되므로 이를 실시간으로 저장, 유통, 수집, 분석처리가 가능한 성능을 의미
- 다양성은 다양한 종류의 데이터를 의미하며 정형화의 종류에 따라 정형, 반정형, 비정형 데이터로 분류



빅데이터 개요

□ 빅데이터 플랫폼

- 빅데이터 기술의 집합체이자 기술을 잘 사용할 수 있도록 준비된 환경
- 기업들은 빅데이터 플랫폼을 사용하여 빅데이터를 수집, 저장, 처리 및 관리
- 빅데이터 플랫폼은 빅데이터를 분석하거나 활용하는 데 필요한 필수 인프라(Infrastructure)
- 빅데이터 플랫폼은 빅데이터라는 원석을 발굴하고, 보관, 가공하는 일련의 과정을 이음새 없이(Seamless) 통합적으로 제공함. 이러한 안정적 기반 위에서 처리된 데이터를 분석하고 이를 다시 각종 업무에 맞게 가공하여 활용하여 사용자가 원하는 가치를 정확하게 획득

빅데이터 개요

빅데이터 핵심 기술

- 분할 점령(Divide and Conquer)
- 데이터를 독립된 형태로 나누고 이를 병렬적으로 처리하는 것
- 빅데이터의 데이터 처리란 이렇게 문제를 여러 개의 작은 연산으로 나누고 이를 취합하여 하나의 결과로 만드는 것을 의미
- 대용량의 데이터를 처리하는 기술 중 가장 널리 알려진 것은 아파치 하둡(Apache Hadoop)과 같은 Map-Reduce 방식의 분산 데이터 처리 프레임워크

빅데이터 개요

빅데이터 활용 사례

- 2014년 월드컵과 2016년 올림픽을 개최한 리우데자네이루는 지능형운영센터(IOC)를 통해 도시 관리와 긴급 대응 시스템을 갖춤
- IBM의 분석 솔루션이 적용된 지능형운영센터에는 교통, 전력, 홍수, 산사태 등의 자연재해와 수자원 등을 통합 관리할 수 있는 체계
- IBM이 제공한 고해상도 날씨 예측 시스템은 날씨와 관련한 방대한 데이터를 분석해 폭우를 48시간 이전에 예측
- 싱가포르는 차량의 기하급수적인 증가로 인한 교통체증을 줄이기 위해 교통량 예측 시스템을 도입하였으며, 싱가포르는 이 시스템을 통해 85% 이상의 정확성으로 교통량을 측정

빅데이터의 속성

빅데이터의 공통적 속성 3V

- 데이터의 크기(Volume), 데이터의 속도(Velocity), 데이터의 다양성(variety)을 나타내며 이러한 세 가지 요소의 측면에서 빅데이터는 기존의 데이터베이스와 차별화
- 데이터 크기(Volume)
  - 단순 저장되는 물리적 데이터양을 나타내며 빅데이터의 가장 기본적인 특징
- 데이터 속도(Velocity)
  - 데이터의 고도화된 실시간 처리 의미
  - 데이터가 생성되고, 저장되며, 시각화되는 과정이 얼마나 빠르게 이뤄져야 하는지에 대한 중요성
- 다양성(Variety)
  - 다양한 형태의 데이터를 포함하는 것을 의미. 정형 데이터뿐만 아니라 사진, 오디오, 비디오, 소셜 미디어 데이터, 로그 파일 등과 같은 비정형 데이터도 포함

빅데이터의 새로운 V – 정확성(Veracity)

- 빅데이터 시대에는 방대한 데이터의 양을 분석하여 일정한 패턴을 추출할 수 있다. 빅데이터를 분석하는 데 있어 기업이나 기관에 수집한 데이터가 정확한 것인지, 분석할 만한 가치가 있는지 등을 살펴야 하는 필요성 대두

빅데이터의 속성

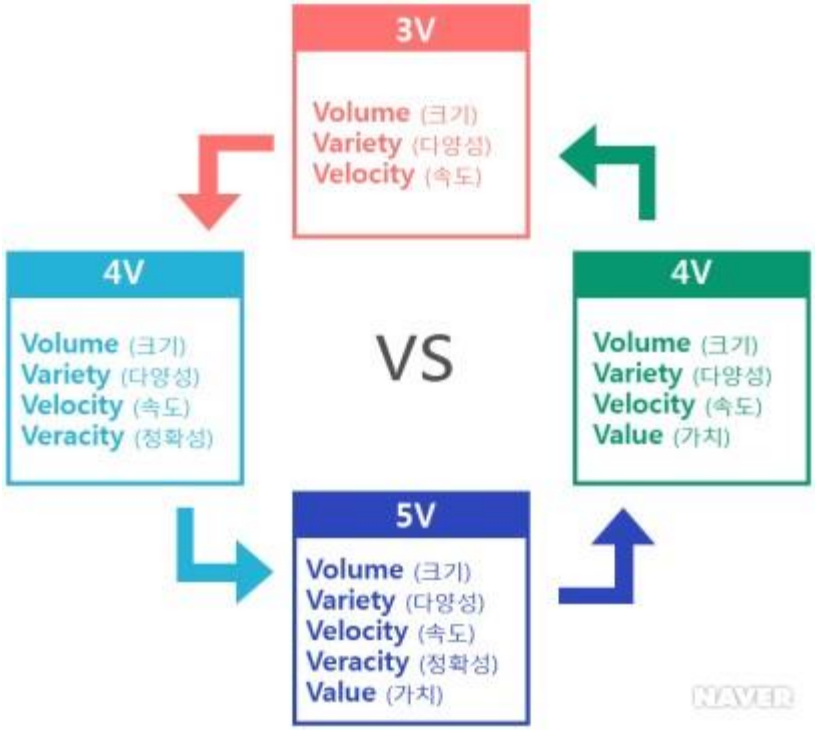
☐ 빅데이터의 새로운 V – 가변성(Variability)

- 소셜미디어의 확산으로 자기 의견을 웹을 통해 자유롭게 게시하는 것이 쉬워졌지만 실제로 자신의 의도와는 달리 자기 생각을 글로 표현하게 되면 맥락에 따라 자신의 의도가 다른 사람에게 오해를 불러일으킬 수 있다. 이처럼 데이터가 맥락에 따라 의미가 달라진다고 하여 빅데이터의 새로운 속성으로 가변성(Variability)이 제시됨

빅데이터의 속성

☐ 빅데이터의 새로운 V – 시각화(Visualization)

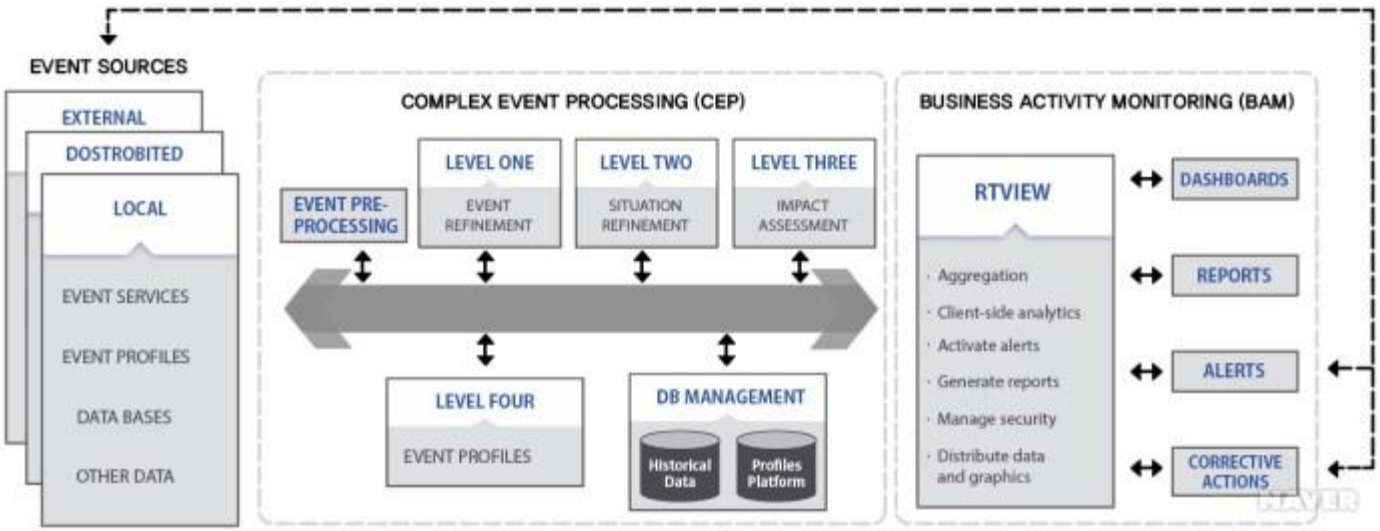
- 빅데이터는 정형 및 비정형 데이터를 수집하여 복잡한 분석을 실행한 후 용도에 맞게 정보를 가공하는 과정을 거치며, 정보의 사용 대상자가 쉽게 이해할 수 있어야 한다. 그렇지 않으면 정보의 가공을 위해 소모된 시간적, 경제적 비용이 무용지물이 될 수 있음



Big data 처리 기술: CEP(Complex Event Processing)

□ CEP의 정의

- RFID 리더, 바코드 스캐너, 기계 장치의 센서와 같이 다양한 IT 환경에서, 최근에는 중요 자원의 위치를 알려주는 GPS(Global Positioning Systems) 정보까지 다양한 데이터가 끊임없이 쏟아지고 있다. 시스템이 지속적으로 데이터를 발생시키고 발생하는 데이터의 양이 점점 늘어가고 있는 상황에서 비즈니스적으로 의미 있는 데이터를 신속하게 추출하고 처리하는 문제는 매우 중요함
- 복잡한 데이터를 처리하는 기술인 CEP(Complex Event Processing)가 등장

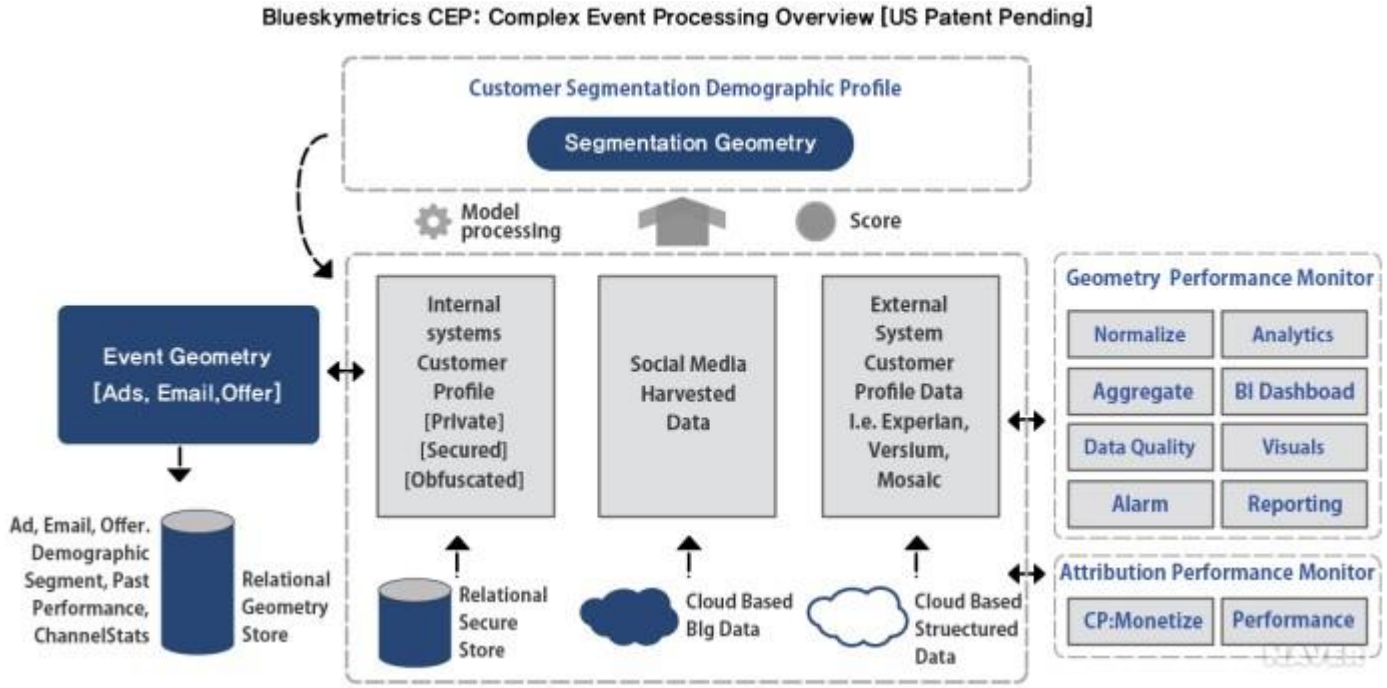


(CEP 과정)

Big data 처리 기술: CEP(Complex Event Processing)

☐ CEP의 정의

- 여러 이벤트 소스로부터 발생한 이벤트를 대상으로 실시간으로 의미 있는 데이터를 추출하여 이에 대응되는 기능을 수행하는 것을 의미. 이때 이벤트 데이터는 스트림 데이터로써 대량으로 지속적으로 입력되는 데이터, 시간 순서가 중요한 데이터, 끝이 없는 데이터
- 스트림 데이터는 전통적인 관계형 데이터베이스에서는 실시간 처리 및 분석을 할 수 없으며, CEP는 바로 이런 스트림 데이터를 실시간으로 분석하는 이벤트 데이터 처리 기술

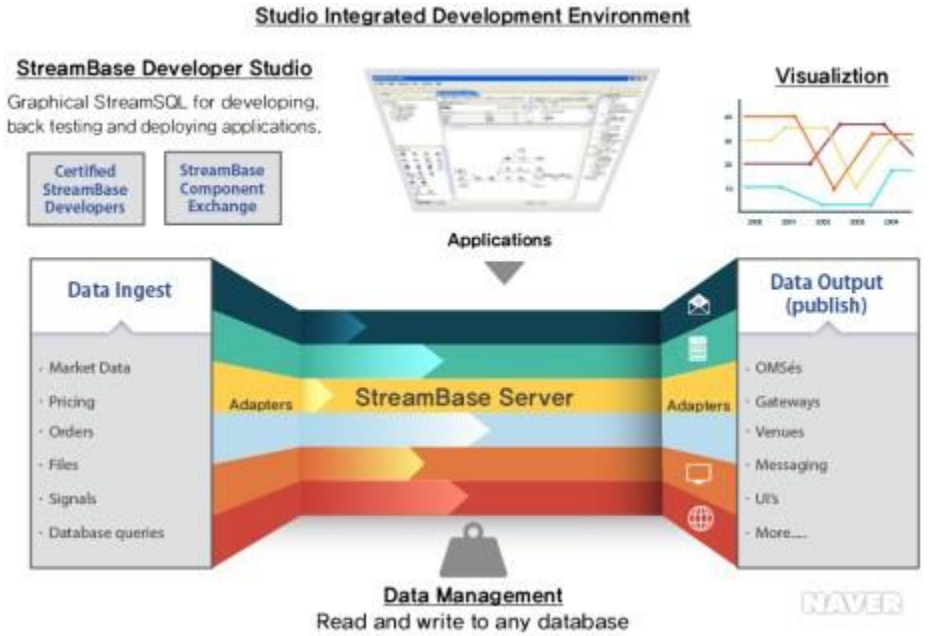


(Blueskymetrics CEP 개요)

Big data 처리 기술: CEP(Complex Event Processing)

☐ CEP의 정의

- 이러한 이벤트들은 세일즈 리드(Sales leads), 주문서 및 소비자 서비스 전화 같은 조직의 여러 계층 상에서 발생
- 새로운 품목, 문자 메시지, 소셜미디어 상의 포스트, 주식시장의 피드, 교통정보, 기상정보 등 여러 종류의 데이터도 이벤트가 됨
- 측정값이 사전의 정의된 시간, 온도 등의 기타 값의 임계치를 넘었을 때의 상태 변화를 이벤트로 정의 가능
- CEP는 실시간으로 패턴을 분석하고 비즈니스 관련 부서에서 IT 및 서비스 부서와 더욱 원활하게 의사소통 통로 제공

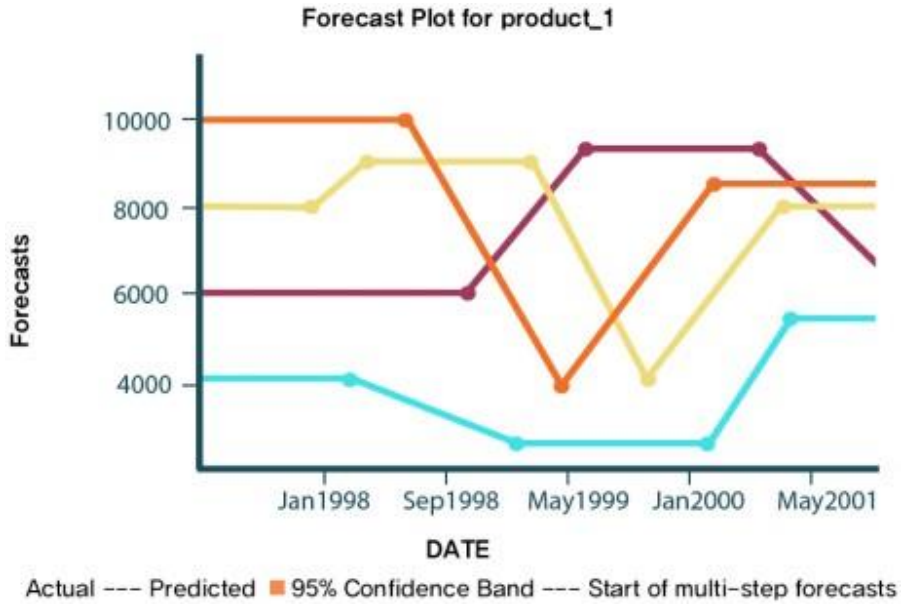


(CEP 아키텍처)

Big data 처리 기술: CEP(Complex Event Processing)

☐ 시계열 데이터베이스와 CEP

- 시계열 데이터는 전형적으로 복합 이벤트 처리와 관련된 분석에 과거 문맥을 제공
- 금융권과 같은 모든 수직적 산업에 적용할 수 있고 협력적으로 BPM(Business Process Management)과 같은 다른 기술들과 함께 적용 가능
- 미래 가격의 움직임의 통계적 임계치를 결정하기 위해서 과거 가격 변동성을 이해할 필요가 있는 금융권의 시나리오를 고려해보면 이는 거래 모델과 매매 비용 분석 모두에 매우 유용
- 어제, 지난주 혹은 지난달에 발생한 일은 오늘 그리고 미래에 일어날 일의 확장이므로 과거 시계열과 실시간 스트리밍 데이터를 단일한 시간 연속체로 간주

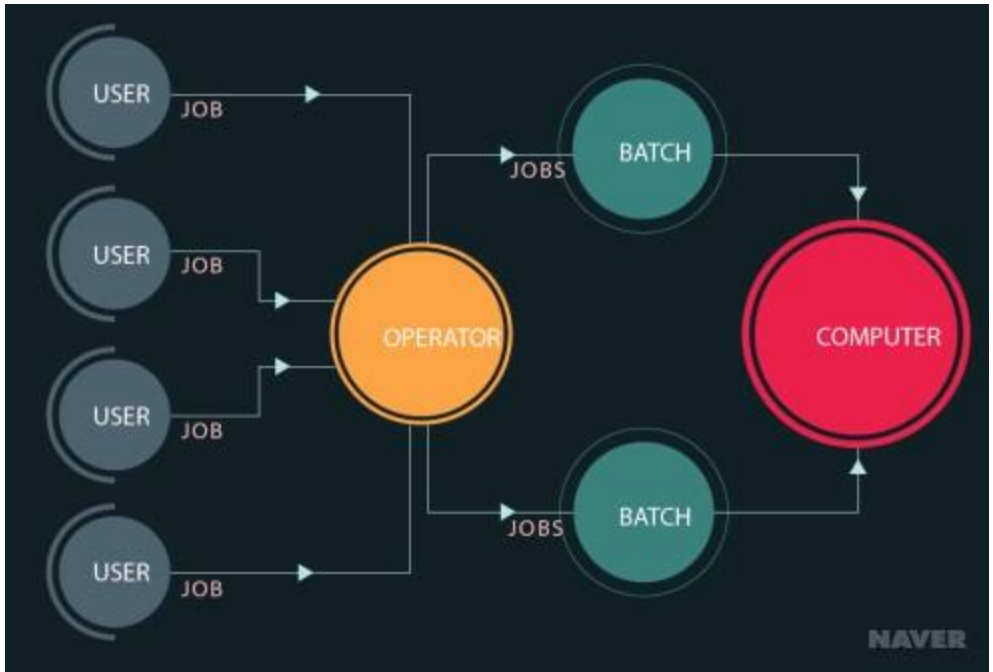


(시계열 데이터베이스)

Big data 처리 기술: Batch 처리

Batch의 정의

- 처리해야 할 데이터를 일정 기간 또는 일정량 정리하여 처리하는 것을 의미
- 컴퓨터 시스템에서는 처리의 대상이 되는 데이터를 일 단위나 월 단위마다 모아두고 그것을 하나로 종합하여 처리
- 처리의 대상이 되는 작업들을 종합하고 일정량을 나눈 다음 처리 작업을 실행한 후에 처리된 데이터들을 통합
- 일괄처리 방식은 컴퓨터 이용 형태로서 오래된 방법이지만, 컴퓨터의 처리 효율을 높일 수 있고, 일정 시점 단위로 처리해야 하는 업무에는 여전히 유용한 방법으로 이용됨

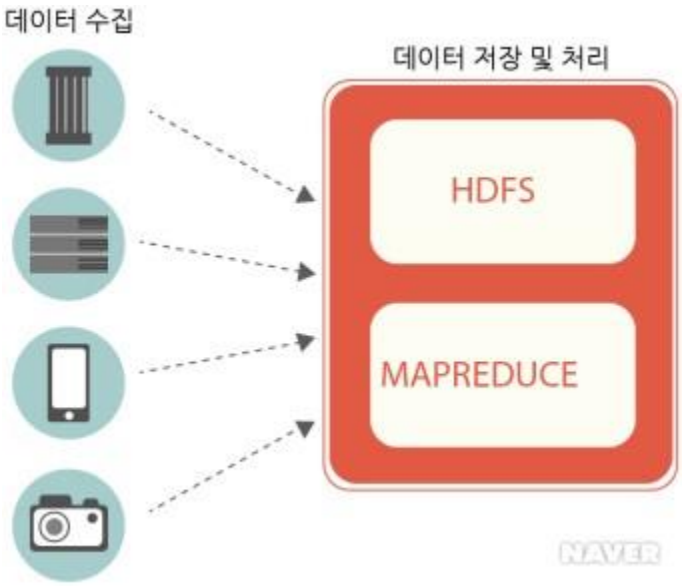


(Batch 처리 시스템)

Big data 처리 기술: Batch 처리

☐ 빅데이터 일괄처리 플랫폼(Hadoop)

- 하둡은 분산 파일시스템인 HDFS(Hadoop Distributed File System)와 분산처리를 위한 맵리듀스(MapReduce)로 구성된 빅데이터 플랫폼
- 그림과 같이 휴대전화, 컴퓨터 등에서 수집된 데이터를 맵리듀스를 통해 처리하고 HDFS를 통해 처리된 데이터를 저장
- 최근 산업계에서는 제조관리, 에너지관리, 네트워크, RFID, 통신, 금융 애플리케이션, 웹 로그 & 클릭 스트림 분석 등은 실시간 빅데이터 처리 기술을 요구
- 하둡은 하둡 에코 시스템 중 하나인 HBase를 이용하여 실시간으로 데이터 분석 가능



Big data 처리 기술: Batch 처리

일괄처리 플랫폼 Cloudera Enterprise

- 클라우데라(Cloudera)는 Cloudera Hadoop용 실시간 쿼리 엔진인 Cloudera Enterprise
- 비정형 · 정형 데이터를 불문하고 대규모 확장 시스템에 존재하는 모든 형태의 데이터를 함께 처리할 수 있으며 실시간 운영 가능
- 중앙 빅데이터 플랫폼을 제공해 기업의 대규모 데이터를 효과적으로 관리 가능

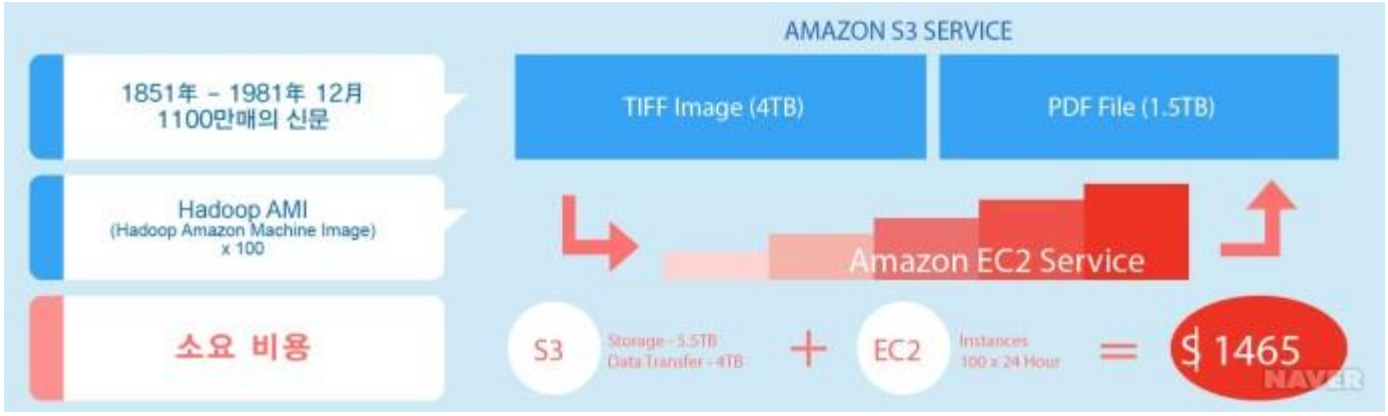
일괄처리 플랫폼 Hive

- Hive는 Hadoop의 분산 처리 프레임 워크 맵리듀스를 프로그래밍 없이 쉽게 개발할 수 있는 도구
- SQL과 유사한 쿼리 언어 HiveQL을 이용할 수 있다는 장점
- HiveQL은 자동으로 맵리듀스 작업으로 변환돼 기존 SQL을 사용했던 엔지니어도 쉽게 맵리듀스의 장점 활용 가능
- Hadoop Hive 등의 주변 기술과 결합해 과거와 같이 정기적으로 실행되는 전형적인 일괄처리뿐만 아니라 방대한 데이터의 일괄처리도 가능하게 됨

Big data 처리 기술: Batch 처리

□ 일괄처리 플랫폼 구현사례

- 뉴욕타임스는 과거 게재했던 기사를 PDF 파일로 서비스하길 원했다. 문제는 130여 년 전 기사부터 서비스를 시작한다는 것이었다. 이 양은 무려 1,100만여 개의 기사에 달했다. 기존 기사를 TIFF 형식으로 스캔하자 약 4테라바이트에 달하는 이미지 데이터가 추출됐다. 뉴욕타임스는 Amazon EC2, S3(일종의 컴퓨터 임대 서비스)를 이용해 100 노드로 구성된 Hadoop 클러스터를 구축했다. 이를 통해 4테라바이트 이미지 데이터를 24시간 만에 약 1.5테라바이트의 PDF 데이터로 변환했다. 작업에 들어간 비용도 불과 240달러였다.



센싱 빅데이터 활용 사례

□ 센싱 빅데이터를 활용한 서비스

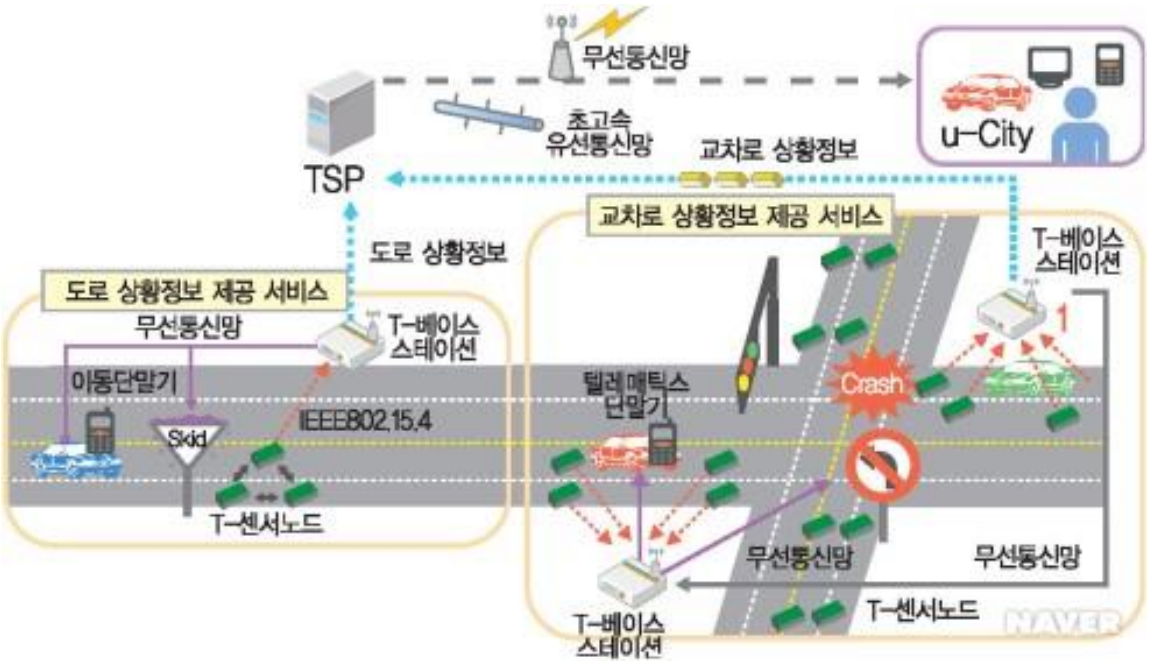
- 지구 환경 연구 분야에서는 관측된 대량의 데이터를 분석하여 기후 변화, 지질 변화 예측 등에 활용
- 지능형 교통서비스, 스마트 그리드 서비스, 범죄예방 서비스, 실시간 환자 모니터링 서비스 등 공공 분야 서비스로 확산

□ 지능형 교통시스템(ITS : Intelligent Transportation System)

- 지능형 교통시스템은 교통시설의 이용 극대화를 토대로 교통서비스의 개선 및 교통사고의 획기적 감소를 목표로 한다. 따라서, 지능형 교통시스템은 첨단 차량 및 도로 시스템(AVHS : Advanced(혹은 Automated) Vehicle and Highway System)으로 도로의 결빙을 방지하거나 차량 간격을 자동으로 조정하여 추돌을 방지하고 차량 간의 운행에서 차선의 급작스런 변경으로 인한 사고를 미연에 방지해주는 서비스 등을 제공한다.
- 다양한 차량 내 센서 정보를 활용한 차량 자체 진단을 통해 사고를 예방하고, 운전자의 졸음상태나 음주상태 등의 정보를 감지하여 경고음을 유발하는 등 안전 운행에 저해되는 요소를 제거하는 것이 가능하다. 이러한 서비스들은 다양한 센서들을 복합적으로 사용하여 판단의 정확도를 높이는 방법들이 시도되고 있다.

센싱 빅데이터 활용 사례

□ 지능형 교통시스템(ITS : Intelligent Transportation System)



(도로, 교차로 상황정보 제공 서비스)

센싱 빅데이터 활용 사례

□ 스마트 그리드(Smart Grid)

- 가정이나 공장에서 전기 사용량과 요금을 실시간으로 파악해 절전 효과를 거두고 발전소를 효율적으로 가동하기 위한 지능형 전력망
- 값싼 심야 전기를 사용해 가전제품을 가동하거나 전기가 남아도는 시간대에 충전했다가 전기가 모자랄 때 꺼내 쓸 수 있다.
- 스마트 그리드망을 구성하는 요소들의 상태 및 미터링 데이터를 지속적으로 수집, 실시간 분석을 통해서 상황 모니터링 및 전력 수요를 예측하는 기술이 적용되어야 함

□ 신생아 심폐 정지 예측 사례

- 캐나다 온타리오 기술대학교(University of Ontario Institute of Technology) 병원의 의사들은 신생아의 몸에 부착된 센서에서 보내는 정보 즉 생리학 데이터 스트림(혈압, 체온, 심전도, 혈중 산소 포화도 등)의 실시간 분석 및 상관관계 분석을 통해 신생아의 생명을 위협할 수 있는 잠재적인 상황을 조기에 감지함으로써, 현재의 의료 장비보다 최대 24시간 전에, 중환자실에 있는 숙련된 간호사보다 조기에 위험 상황을 감지하여, 신생아의 사망률을 낮추고 장기적인 증상 개선이 가능하게 되었다.

센싱 빅데이터 활용 사례

☐ 범죄와 테러 위협에 대처하기 위한 도메인 인식 시스템(DAS : Domain Awareness System)

- 뉴욕 경찰청과 마이크로소프트가 공동으로 개발한 새로운 도메인 인식 시스템은 감시 카메라, 차량 번호판 리더, 방사선 탐지기, 911 전화, 범죄 이력 보고서, 기타 공공 안전 데이터베이스에서 데이터를 취합하고 분석하는 방법을 경찰 관계자들에게 제공한다.
- 위협과 범죄를 야기하는 모든 행위를 종합적으로 보고 판단할 수 있도록 하기 위한 것으로 잠재적인 보안 위협에 대해 실시간으로 경고 메시지를 제공한다. 예를 들어 도시 내 어디서든 방사선 탐지기가 울리면 DAS는 즉시 센터에 알려 원인을 파악하게 한다. 현재는 감시 카메라 영상을 실시간으로 분석해 활용하지는 않지만 취합한 데이터를 통합 관리하면서 신속하게 확인할 수 있도록 하고 있다

비정형 센싱 빅데이터 활용

☐ 외부 현상을 감지하기 위해 활용하는 센서의 형태

- 온도와 습도, 위치 등 단순한 데이터를 센싱하여 정형화된 데이터로 제공
- 차량 번호판 리더, CCTV, 위성 카메라처럼 복합적인 정보를 센싱하여 비정형의 영상 데이터로 제공
- 영상 데이터를 자동으로 분석하는 기술은 아직 정확도가 떨어져 한정된 목적으로 쓰이지만 꾸준히 발전하고 있는 만큼 조만간 센싱 영상 데이터의 실시간 자동 분석을 통한 상황인지가 가능할 것으로 예상
- 영상, 음성 등의 비정형 센싱 데이터는 유용한 정보를 찾아내기 위한 작업이 복잡하고 응용의 목적에 따라 추출하고자 하는 정보가 달라지므로 사용자가 자유롭게 비정형 센싱 데이터를 처리할 수 있는 환경을 제공해야 하며, 기존의 정형 센싱 데이터에 통합해 활용할 수 있어야 더 정확하게 외부 상황을 판단할 수 있음
- 복합이벤트처리 시스템은 최근 비정형 센싱 데이터 활용 및 실시간 발생 센싱 데이터 폭증에 대응하기 위해 분산 처리 기술을 적용한 분산 스트림 컴퓨팅 기술
- 분산 스트림 컴퓨팅은 서비스 구축 목적 및 환경에 따라 비정형 센싱 데이터 처리 로직을 자유롭게 추가하거나, 처리 연산의 부하 증대 문제를 해결하기 위해 분산 처리 기법을 도입한 경우, 센싱 데이터가 폭증함에 따라 센싱 데이터를 분할하여 병렬 처리가 필요한 경우에 주로 활용

빅데이터 수집: 비정형 데이터 마이닝

비정형 데이터

- 비정형 데이터(Unstructured data)란 일정한 규격이나 형태를 지닌 숫자 데이터(numeric data)와 달리 그림이나 영상, 문서처럼 형태와 구조가 다른 구조화되지 않은 데이터를 의미
- 책, 잡지, 문서의료 기록, 음성 정보, 영상 정보와 같은 전통적인 데이터 이외에 이메일, 트위터, 블로그처럼 모바일 기기와 온라인에서 생성되는 데이터
- 대표적인 비정형 데이터인 문서에는 문자가 가장 많은 비중을 차지하고 있지만 숫자와 도표, 그림도 포함하고 있다. 이러한 문서 정보는 정보의 관점에서 보면 유형이 불규칙하고 의미를 파악하기 모호해서 기존의 컴퓨터 처리 방식을 적용하기 어렵다. 기존의 컴퓨터 시스템은 연산과 처리 절차가 숫자 데이터 중심으로 설계되어 있기 때문에 이름이나 성별과 같은 문자 변수는 숫자로 변환해 처리하는 방법을 주로 사용했다. 그러나 이런 방법은 트위터나 블로그처럼 모바일과 온라인에서 생성되는 대규모의 비정형 데이터에 적용하는 것이 불가능하다. 비정형 데이터는 불규칙 정도에 따라 반정형 데이터(semi-structured data)로 구분하기도 한다.

빅데이터 수집: 비정형 데이터 마이닝

텍스트 마이닝(Text mining)

- 대규모의 문서(text)에서 의미 있는 정보를 추출하는 것
- 분석 대상이 비구조적인 문서정보라는 점에서 데이터 마이닝과 차이
- 텍스트 마이닝은 텍스트 분석(text analytics), 텍스트 데이터베이스로부터 지식 발견(KDT, Knowledge Discovery in Textual Database), 문서 마이닝(document Mining) 등으로 불리움
- 정보 검색, 데이터 마이닝, 기계 학습(machine learning), 통계학, 컴퓨터 언어학(computational linguistics) 등이 결합된 분야
- 분석 대상이 형태가 일정하지 않고 다루기 힘든 비정형 데이터이므로 인간의 언어를 컴퓨터가 인식해 처리하는 자연어 처리(NLP, natural language processing) 방법과 관련됨
- 문서 분류(document classification), 문서 군집(document clustering), 메타데이터 추출(metedata extraction), 정보 추출(information extraction) 등으로 구분

빅데이터 수집: 비정형 데이터 마이닝

텍스트 마이닝(Text mining)

- 문서 분류는 도서관에서 주제별로 책을 분류하듯이 문서의 내용에 따라 분류하는 것
- 문서 군집은 성격이 비슷한 문서끼리 같은 군집으로 묶어주는 방법. 통계학의 방법론인 판별분석(discriminant analysis)과 군집분석(clustering)과 유사한 개념으로 분석 대상이 숫자가 아닌 텍스트이다.
- 문서 분류는 사전에 분류 정보를 알고 있는 상태에서 주제에 따라 분류하는 방법이며 문서 군집은 분류 정보를 모르는 상태에서 수행하는 방법이며, 지도 학습(supervised learning), 비지도(자율) 학습(unsupervised learning)이라고 함. 데이터 마이닝에서도 동일한 의미로 사용
- 정보추출은 문서에서 중요한 의미를 지닌 정보를 자동으로 추출하는 방법론

빅데이터 수집: 비정형 데이터 마이닝

☐ 웹 마이닝(Web mining)

- 인터넷을 이용하는 과정에서 생성되는 웹 로그(web log) 정보나 검색어로부터 유용한 정보를 추출하는 웹을 대상으로 한 데이터 마이닝
- 전통적인 데이터 마이닝의 분석 방법론을 사용하기도 하지만 웹 데이터의 속성이 반정형 혹은 비정형이고, 링크 구조를 형성하고 있기 때문에 별도의 분석기법이 필요
- 분석 대상에 따라 웹 구조 마이닝(web structure mining)과 웹 유시지 마이닝(web usage mining), 웹 콘텐츠 마이닝(web contents mining)으로 구분
- 웹 구조 마이닝
  - 웹 사이트의 노드(node)와 연결 구조를 분석하는 기법
  - 웹 페이지가 연결된 구조를 의미하는 하이퍼링크(hyperlink)로부터 패턴을 찾아내거나 웹 페이지 구조를 분석
- 웹 유시지 마이닝
  - 인터넷 이용자의 이용 경로인 웹서버 로그(web server log) 파일 분석을 통해 웹 사이트 개선이나 고객 특성을 반영한 맞춤형 서비스를 지향

빅데이터 수집: 비정형 데이터 마이닝

☐ 웹 마이닝

- 웹 콘텐츠 마이닝은
  - 웹 페이지에 저장된 콘텐츠 로부터 웹 사용자가 원하는 정보를 빠르게 찾는 기법으로 검색엔진에 많이 사용
  - 웹 페이지를 다루고 있는 주제에 따라 자동적으로 분류 가능하며 전통적인 데이터 마이닝 작업과 유사
  - 웹 페이지에서 특정 상품의 설명이나 독자의 상품평과 같은 유용한 정보를 추출하는 작업은 데이터 마이닝과 다름
- 웹 마이닝 분석 과정은 데이터 마이닝 분석 과정과의 차이는 데이터 수집
- 데이터 마이닝에서 데이터는 이미 수집된 상태이거나 데이터웨어하우스같이 구조적으로 잘 정리된 장소에 저장되어 있으나, 웹 마이닝에서 데이터 수집 작업은 실질적으로 수행되어야 함.
- 웹 크롤러(web craw)는 스파이더(spiders), 웜(worm), 로봇(robots) 또는 봇(bots)으로 불리우며, 웹 페이지를 자동으로 내려 받는 프로그램
- 크롤러는 하이퍼링크로 연결된 웹 페이지를 하나하나 찾아가 텍스트와 영상 등 각종 자료를 수집한다. 목적에 따라 일반적인 목적의 검색엔진에서 사용되는 범용 크롤러(universal crawler), 특정 범주에 속하는 페이지만을 탐색하는 포커스 크롤러(focus crawler), 제한된 주제만을 검색하는 토픽 크롤러(topical crawler)

빅데이터 수집: 비정형 데이터 마이닝

□ 웹 마이닝

- 데이터가 수집되고 나면 데이터 마이닝 분석과 동일하게 데이터 전처리 과정(pre-processing)과 웹 데이터 마이닝 분석, 사후 처리(post-processing) 과정 수행
- 트위터나 페이스북과 같은 소셜네트워크서비스(SNS)가 등장하면서 여기에서 발생하는 데이터를 분석하는 소셜 웹 마이닝도 주목받고 있음
- 소셜 웹 마이닝은 SNS에서 사람들의 네트워크 관계와 주고받는 대화 내용을 통해 영향력 있는 사람이 누구인지, 어떤 주제가 관심을 받는지 등을 분석

빅데이터 수집: 비정형 데이터 마이닝

오피니언 마이닝(Opinion mining)

- 어떤 사안이나 인물, 이슈, 이벤트에 대한 사람들의 의견이나 평가, 태도, 감정 등을 분석하는 것
- 특정 주제에 대해 사람들의 주관적인 의견을 모아 문장을 분석하며, 문장 분석에서는 사실과 의견을 구분해 의견을 뽑아내어 긍정과 부정으로 나누고 그 강도를 측정
- 분석 대상은 주로 포털 게시판, 블로그, 쇼핑몰과 같은 대규모의 웹 문서이므로 자동화된 분석방법 사용
- 분석 대상이 텍스트이므로 텍스트 마이닝에서 활용하는 자연어 처리(NLP, natural language processing) 방법, 컴퓨터 언어학(computational linguistics) 등을 활용
- 감정 분석(sentiment analysis)이라고고 함. 이 외에도 브랜드 모니터링(brand monitoring), 버즈 모니터링(buzz monitoring), 온라인 인류학(online anthropology), 시장 영향력 분석(market influence analytics), 대화 모니터링(conversation mining), 온라인 소비자 이해(online consumer intelligence) 등과 같이 다양한 용어로 불리움
- 버즈 모니터링
  - 온라인에서 특정 주제에 대한 여론을 분석하는 것. 기업의 경우 트위터와 인터넷에 올라온 기업 관련 댓글을 실시간으로 분석해 자사 이미지를 파악하고 대응전략을 세우고 있다. 버즈(buzz)는 벌이나 기계가 윙윙대는 소리를 말하는데 소비자가 자발적으로 특정 상품이나 서비스에 대해 긍정적인 입소문을 퍼트려 좋은 이미지를 만드는 것을 의미하는 '버즈 마케팅(buzz marketing)'에서 유래

데이터 시각화(Data Visualization)

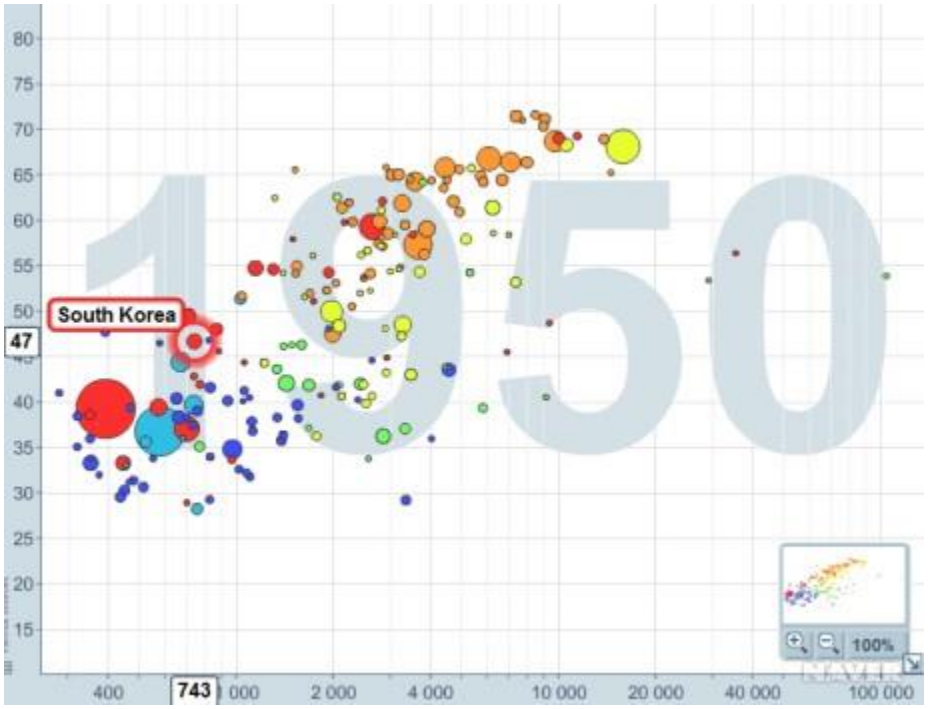
데이터 시각화

- 데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 과정
- 데이터 시각화의 목적은 도표(graph)라는 수단을 통해 정보를 명확하고 효과적으로 전달하는 것
- 의미를 효과적으로 전달하기 위해서는 심미적인 형태와 기능적인 요소가 조화를 이루어야 함
- 이상적인 시각화란 단지 명확하게 의사를 전달하는 데 머물러서는 안 되고 보는 사람을 집중하게 하고 참여하게 만들어야 함.

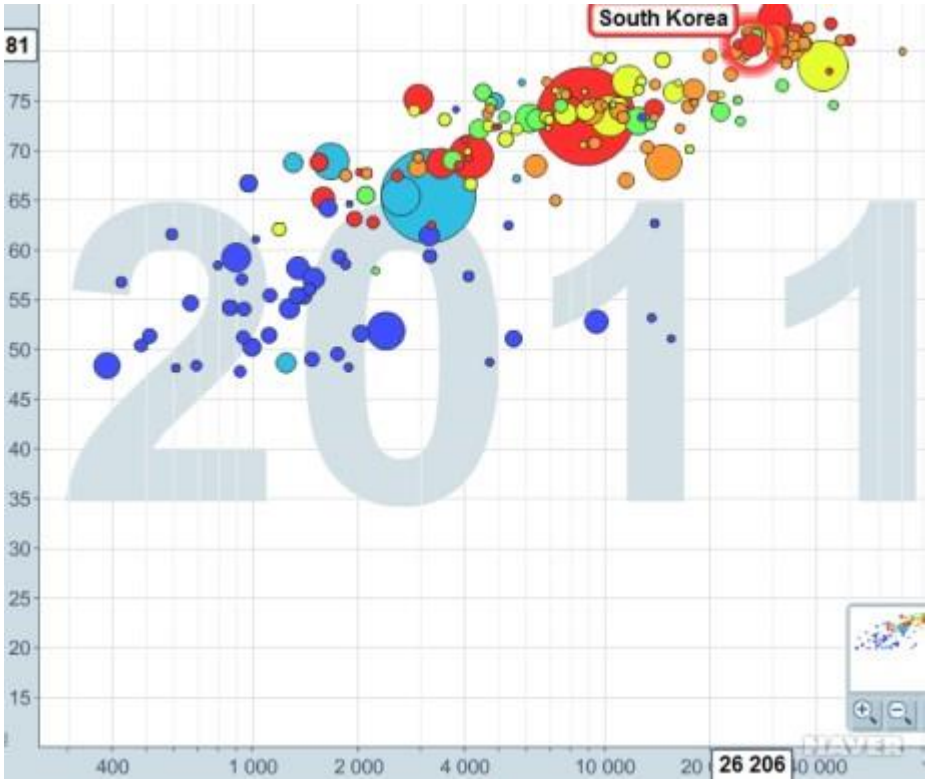
데이터 시각화(Visualization)

□ 데이터 시각화

- 그래프를 통해 공감을 이끌어내는 대표적인 사례로는 한스 로슬링(Hans Rosling) 교수의 그래픽 애니메이션이다. 스웨덴 의대 교수이자 통계학자인 그는 200년간 시간의 변화에 따라 움직이는 국가별 기대수명과 GDP 도표를 통해 국가별 변화를 설득력 있게 표현한다. 그림을 통해 한국은 낙후된 국가에서 선진국으로 진입했음을 확인할 수 있다.



(국가별 기대수명과 GDP, 1950년)



(국가별 기대수명과 GDP, 2011년)

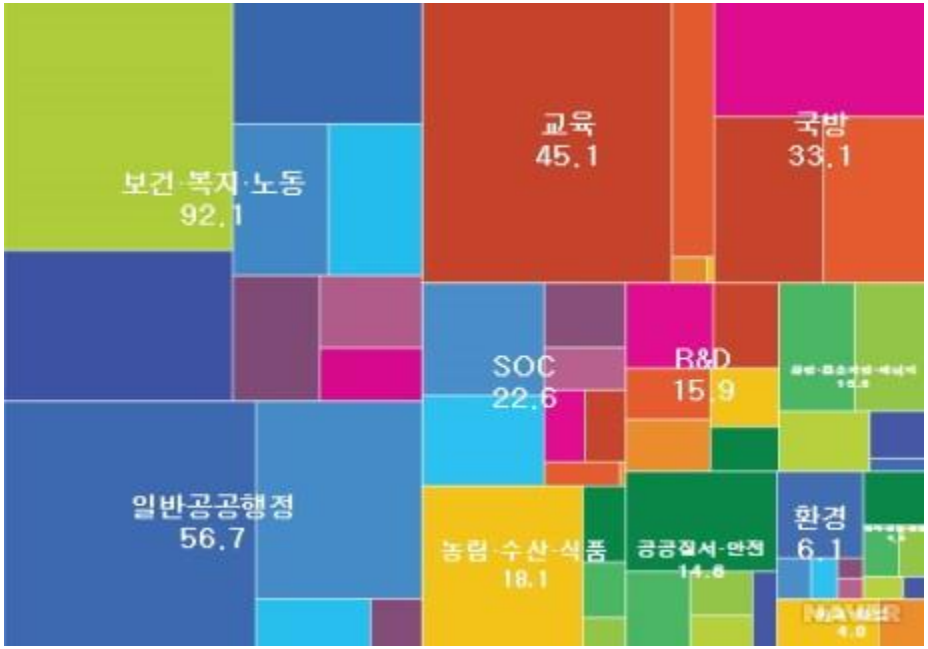
데이터 시각화(Visualization)

데이터 시각화

- 데이터 시각화와 연관된 개념
  - 정보 시각화(information visualization)
    - 일반적으로 소프트웨어 시스템의 파일이나 프로그램 코드, 도서관의 서지 데이터베이스, 인터넷의 관계 네트워크 등과 같은 대규모 비수량 정보를 시각적으로 표현하는 것을 의미
  - 과학적 시각화(scientific visualization)
    - 건축학, 기상학, 의학, 생물학 분야에서 시간의 흐름에 따른 변화를 입체적으로 표현하는 것
    - 과학 분야의 연구결과를 일반인이 쉽게 이해하도록 그림으로 표현하는 사람을 시각화 과학자(visualization scientist)라고 하며, 아름답게 표현된 행성이나 은하계의 사진은 이들이 만든 작품
  - 시각 디자인(visual design)
  - 정보 그래픽(information graphics)
    - 인포그래픽(infographic)이라고도 하며, 정보와 데이터, 지식을 시각적으로 표현하는 것
    - 표지판이나 지도, 언론, 기술보고서, 교육 분야에서 발생하는 복잡한 정보를 빠르고 명확하게 표현하는 것이 핵심
    - 흔히 접하는 교통표지판과 복잡한 지하철 노선도가 대표적인 인포그래픽
    - 한 장에 수많은 데이터를 요약해 표현하는 인포그래픽은 특히 신문이나 방송과 같은 미디어에서 주목
    - 지역별 날씨를 그림으로 나타낸 일기예보 기상도나 기사 내용 중의 통계 수치를 그래프로 나타내는 것
    - 방대한 데이터에서 중요한 정보를 선택해 시각적 효과를 극대화시키는 작업을 데이터 저널리즘(Data Journalism)이라고 지칭
    - 영국의 《가디언(Guardian)》이 시청자 참여를 결합한 데이터 저널리즘을 선도하고 있으며 우리나라의 경우 《연합뉴스》 등에서 다양한 시도를 하고 있음

데이터 시각화(Visualization)

□ 데이터 시각화



인포그래픽으로 표현한 2012년 예산안(단위: 조원)

데이터 시각화(Visualization)

통계그래픽

- 데이터로부터 의미 있는 정보를 추출해 효과적으로 표현하는 통계학의 학문적 특성으로 인해 데이터 시각화와 밀접한 관련이 있다.
- 치 정보를 효과적으로 표현하기 위한 기술통계학(descriptive statistics)의 대표적인 시도로 막대기둥 그림표(histogram)나 누적도수분포표(ogive), 줄기와 잎 그림표(stem-leaf plot), 상자와 수염 그림표(Box-Whisker ' s plot) 등을 사용
- 튜키(John Tukey) 교수는 통계학에서 그래픽의 중요성을 강조했다. 이론통계학 분야에도 크게 기여한 튜키는 저서인 탐색적 데이터 분석(Exploratory Data Analysis)을 통해 통계데이터의 시각적 표현에 관한 많은 업적을 남겼다. 상자와 수염 그림표도 그가 고안했다. 상자와 수염 그림표는 최솟값, 최댓값, 사분위수(Quartile)와 같은 데이터가 가진 특성을 간단한 상자 모양의 그림으로 간결하게 나타냄

데이터 시각화(Visualization)

단어 구름(Word Cloud)

- 문서에 사용된 단어의 빈도를 계산해서 시각적으로 표현하는 것
- 많이 나오는 단어는 크게 표시되기 때문에 한 눈에 문서의 핵심 내용 파악 가능
- 태그 구름(tag cloud)이라고도 하며, 태그는 옷이나 물건에 소재나 취급 방법 등을 설명하기 위해 붙이는 꼬리표다. 웹 페이지나 소셜네트워크서비스(SNS)에서 콘텐츠를 설명하기 위해 붙이는 키워드를 태그라고 부른다. 태그 구름은 웹 사이트에서 태그의 중요도를 글자 크기나 색깔로 표시
- 표현하려는 콘텐츠의 성격에 따라 문서 구름(text cloud)과 데이터 구름(data cloud)으로 구분
- 문서 구름이 문서에 포함된 단어를 시각적으로 표현한 것이라면 데이터 구름은 단어 대신에 숫자 정보를 크기와 색깔로 표현한 것을 의미. 예를 들면, 인구 규모에 따라 국가명의 크기나 색을 달리해서 표현하거나 주식시장에서 주가의 등락과 거래량을 반영해 회사명의 크기와 색을 결정

데이터 시각화(Visualization)

□ 단어 구름(Word Cloud)

태그 구름(Web 2.0과 데이터 구름(국가별 인구 규모)

- 단어들 간 상관관계에 주목하는 분석 방법으로 코워드 분석(co-word analysis)
  - 코워드문장 안에서 함께 사용되는 단어들의 규칙을 조사해서 문서의 주제와 관련된 핵심 개념이 무엇이고 이들의 관계가 어떤지를 식별하는 내용분석 기법이다(He, 1999). 1980년대에 프랑스에서 개발되었다. 단어 간의 관계는 함께 발생하는 빈도수와 단어 간의 친밀도를 지수로 환산하고 이 지수를 기반으로 연계관계를 나타내고 몇 개의 독립적인 그룹으로 구분해 표현한다. 예를 들어 빅데이터와 관련해 학술 잡지에 게재된 논문을 모두 찾은 후 중요한 키워드 간 관계를 파악해 시각화하면 빅데이터 관련 기술, 빅데이터 활용 사례, 데이터 과학자 등과 같은 세부 영역으로 구분될 수 있고 각각의 세부 영역을 대표하는 키워드의 빈도와 키워드 간의 연계를 표현할 수 있다. 이를 통해 연구 분야의 동향을 쉽게 파악할 수 있다

데이터 시각화(Visualization)

데이터 시각화 도구

- 데이터 시각화를 지원하는 도구로는 마이크로소프트의 엑셀(Excel)이나 구글의 스프레드시트(Spreadsheets)처럼 데이터 관리와 그래프 작성을 위해 만들어진 도구가 있다. 구글의 스프레드시트는 데이터를 구글 서버에 저장하기 때문에 인터넷 접속이 가능한 어떤 컴퓨터에서도 작업이 가능하다. 다른 사람과 실시간으로 공동 작업도 가능하고 한스 로슬링의 그래프처럼 시간의 흐름에 따라 움직이는 차트도 만들 수 있다.
- 전문적인 분석을 위한 프로그래밍 언어로는 파이선(python), 피에이치피(PHP) 등이 있고 오픈 소스인 프로세싱(Processing)과 R이 있다.
- R은 통계분석 소프트웨어이면서 통계그래픽 기능도 뛰어난 오픈 소스 프로그램