데이터의 이해

데이터와 정보

- 데이터: 있는 그대로의 객관적 사실
- 정보: 데이터를 가공

데이터의 유형

- 정량적: 자료를 수치화. 온도, 풍속
- 정성적: 자료의 특징을 풀어 설명. 언어, 문자
- 정형: 정보 형태가 정해짐(관계DB, excel, csv)
- 반정형: 메타데이터(HTML, XML, JSON, RDF)
- 비정형: 형태 정해지지 않음(SNS, 음원, 유튜브)

암묵지, 형식지간 상호작용

- 암묵지: 개인에게 습득되고 겉으로 드러나지 않음
- 형식지: 문서, 메뉴얼 등 형상화된 지식

1. 공통화: 암묵지 지식을 다른 사람에게 알려줌
2. 표출화: 암묵지 지식을 메뉴얼이나 문서로 전환
3. 연결화: 교재, 매뉴얼에 새로운 지식 추가
4. 내면화: 교재, 메뉴얼에서 다른 사람의 암묵지를 터득

DIKW 피라미드

1. data: 있는 그대로의 사실(A 대리점 폰이 100만원, B 대리점 폰이 200만원)
2. information: 데이터를 통해 패턴 인식(A 대리점이 더 싸다)
3. knowladge: 패턴을 통해 예측(A에서 사면 이득이다)
4. wisdom: 창의적인 산물(A의 다른 기종도 B보다 쌀 것이다.)

데이터 단위

KB MB GB TB PB EB ZB YB 테라 페타 엑사 제타 요타 2^10승부터 ~~

데이터베이스 정의와 특징

데이터베이스

- 스키마: 전반적 명세 (외개내)
- 인스턴스: 데이터 개체를 구성하는 속성에 대한 데이터 타입과 값
- 메타데이터
- 인덱스: 정렬, 탐색을 위한 데이터의 이름

DBMS

- 관계DBMS: 테이블로 정리. MySQL, maridDB, Oracle

- NoSQL DBMS: 비정형 데이터.
- DDL, DML, DCL

데이터베이스의 특징

1. 공용 데이터: 여러 사용자가 공동이용
2. 통합된 데이터: 동일 데이터 중복이 없음
3. 저장된 데이터: 저장매체에 저장
4. 변화되는 데이터: CRUD해도 무결성 유지 "공통저변"

디비 설계 절차

- 개념설계: 개념스키마
- 논리설계: 개념ERD를 활용한 논리모델링
- 물리설계: 저장구조설계
- 개논물

데이터베이스 활용

기업 활용 데이터베이스

- OLTP: online transaction processing
  - 데이터를 수시로 갱신 (거래단위마다)
- OLAP: online analytical processing
  - 다차원 데이터를 대화식으로 분석
- CRM: customer relationship management
  - 고객과 관련 자료 분석, 마케팅 활용
- SCM: supply chain management
  - 공급망 연결 최적화
- erp: enterprise resource planning
  - 기업 경영 자원을 효율화
- RTE: real-time enterprise
  - 최신 정보로 빠른 의사결정 지원
- BI: business intelligence
  - 기업 보유 데이터 정리, 분석하는 리포트 중심 도구
- BA: business analytics
  - 통계 기반 비즈니스 통찰력
- Block Chain: 네트워크에 참여하는 모든 사용자가 정보 분산저장
- KMS: knowledge management system
  - 기업의 모든 지식을 포함

데이터 웨어하우스(DW)

- 주제지향성: 분석목적 설정이 중요
- 데이터 통합: 일관된 형식으로 저장
- 시계열성
- 비휘발성: read-only

- ETL(extraction, transform, load)
- ODS(operational Data store)
    - 추출하고, 변환하고, 저장함. ODS에 임시 저장하고 그 후 DW에 저장
- DW에서 이걸 다시 DM으로 변환 가능(데이터마트)

데이터레이크

- 비정형 데이터를 저장함. 하둡과 연계하여 처리함.
- 하둡: 여러 컴퓨터를 하나로 묶어 대용량 데이터를 처리하는 오픈소스 빅데이터 솔루션
    - HDFS: 분산 파일 저장 시스템
    - MapReduce: 분산 데이터를 병렬 처리

데이터의 가치와 미래

빅데이터 출현 배경

- 인터넷 확산
- 스마트폰 보급
- 클라우드 컴퓨팅으로 인한 경제성 확보
- 저장매체 가격하락
- 하둡을 활용한 분산컴퓨팅
- 비정형 데이터 확산

빅데이터의 3V(가트너 정의)

1. Volume(규모): 데이터 양 증가 (ex: 구글 번역 서비스)
2. Variety(다양성): 데이터 유형 증가
3. Velocity(속도): 데이터 생성 / 처리 속도 증가

빅데이터가 만들어낸 변화

"전후양상"

1. 표본조사 -> 전수조사
2. 사전처리 -> 사후처리
3. 질 -> 양
4. 인과관계 -> 상관관계

빅데이터 활용 3대 요소

"인자기"

- 인력, 자원(데이터), 기술

빅데이터의 주요 분석 기법

- 회귀분석: 독립변수, 종속변수, 부동산 가격
- 분류분석: 범주, 고양이, 강아지 이미지 구분

- 연관규칙: 여러 요소들 간 상관관계 존재. 치킨 구매 시 맥주 구매율 증가
- 유전자 알고리즘: 최적화 필요한 문제 해결책, 택배차량 배치 문제, 최대 시청률 위한 프로그램 시간대 배치
- 기계학습: 넷플릭스, RecSys
- 감정분석: 텍스트에서 긍/부 분석
- SNS분석: 사람간 관계, bonding
- 텍스트마이닝: NLP

위기 요인과 통제방안

1. 사생활 침해
    - 제공자에서 사용자로 책임 전환
2. 책임 원칙 훼손
    - ex) 범죄 예측 분석으로 일어나지도 않은 사람 미리 체포
    - 결과에 대해서만 책임
3. 데이터 오용
    - 분석 결과가 항상 옳지 않음
    - 알고리즘 해석 가능한 '알고리즈미스트' 필요
        - 알고리즈미스트: 부당하게 피해 발생한 사람 구제하는 인력

데이터 3법

- 가명정보 개념 사용
    - 이름만 가명으로 하면 공익, 연구 목적 하에 동의 없이 개인정보 활용 가능
    - 개 정 신
    - 개인정보보호법
    - 정보통신망 ~ 법률
    - 신용정보 이용~ 법률
- 개인정보: (홍길동, 33세)
- 가명정보: (홍00, 30대 초반)
- 익명정보: (000, 30대)

개인정보 비식별화

1. 가명처리(황경락, 20세 -> 고경태, 30세)
2. 총계처리(황경락, 20세, 고경태, 30세 -> 평균 25세)
3. 데이터 삭제(주민번호 삭제)
4. 데이터 범주화(황경락, 20~30세)
5. 데이터 마스킹(황경락 -> 황00)

데이터 산업의 발전

1. 처리
    - 프로그래밍 언어를 활용한 데이터 처리
2. 통합
    - DBMS

3. 분석
    - 빅데이터 분석 기술
4. 연결
    - API 활용한 모듈 연결
5. 권리
    - 마이데이터를 활용한 데이터 주권 행사

데이터 사이언스 핵심 구성요소

AI비

1. Analytics: 이론적 지식
2. IT: 프로그래밍 지식
3. 비즈니스 분석: 비즈니스적 능력

- 하드스킬: 이론적. 수학
- 소프트스킬: 스토리텔링, 리더십, 창의력, 분석

빅데이터 가치 패러다임 변화

1. Digitalization: 아날로그 -> 디지털화
2. Connection: 디지털화 정보를 연결
3. Agency: 연결을 효과적으로 관리 "DigitalCA메라"