과목1. 데이터 이해

■ 데이터 정의

- 관념적 & 추상적 ⇨ 기술적 & 사실적
- 추론과 추정의 근거를 이루는 사실
- 객관적 사실 (존재적 특성)
- 추론·예측·전망·추정을 위한 근거 (당위적 특성)
- 가치 창출과정에서 가장 기초를 이루는 것

■ 데이터 유형

◆ 정성적 데이터 (qualitative) : 언어·문자 등 비정형 데이터

◆ 정량적 데이터 (quantitative) : 수치·도형·기호 등

■ 지식경영의 핵심 이슈

◆ 암묵지 : 무형의 지식. 개인에게 축적된 내면화 지식 ⇨ 조직의 지식으로 공통화

◆ 형식지 : 형상화된 지식. 언어·기호·숫자 등 표출화 지식 ⇨ 개인의 지식으로 연결화

■ DIKW

◆ Data 데이터 : 객관적인 사실. 가공 전의 순수한 수치나 기호.

◆ Information 정보 : 데이터에서 의미 도출. 패턴 인식. 의미 부여.

◆ Knowledge 지식 : 정보를 구조화하여 분류하고 개인 경험을 결합시켜 고유의 지식으로 내재화

◆ Wisdom 지혜 : 근본원리에 대한 이해를 바탕으로 지식의 축적과 아이디어가 결합된 창의적 산물.

■ 데이터베이스 연혁

50년 : 데이터의 기지라는 뜻으로 데이터베이스 탄생

63년 : '대량의 데이터를 축적하는 기지' 의미로 심포지엄에서 공식 사용. 데이터베이스 관리 시스템 IDS 개발.

65년 : 시스템을 통한 체계적 관리와 저장 등의 의미를 담은 '데이터베이스 시스템' 용어 등장

70년 : 유럽에서 데이터베이스 단일어 일반화

75년 : 우리나라에서 데이터베이스 이용

80년 : 데이터베이스 서비스 시대

85년 : 국내의 데이터베이스 관련 기술의 연구, 개발

■ 데이터베이스 정의

◆ 정형 데이터 관리

- EU : 체계적·조직적으로 정리되고 전자식 또는 기타 수단으로 개별적으로 접근할 수 있는 독립된 데이터.
- 저작권법 : 소재를 체계적으로 배열·구성한 편집물로서 개별적으로 그 소재에 접근하거나 검색할 수 있도록 한 것.

◆ 비정형 데이터 포함

- 컴퓨터용어사전 : 동시에 복수의 적용 업무를 지원할 수 있도록 데이터 수집, 저장, 공급을 위해 일정한 구조에 따라 편성된 데이터 집합.
- 데이터분석전문가가이드 : 상호 관련된 콘텐츠들을 체계적으로 수집·축적하여 다양한 용도와 방법으로 이용할 수 있도록 정리한 정보 집합체.
- 위키피디아 : 관련된 레코드의 집합. 소프트웨어로는 데이터베이스관리시스템(DBMS)을 의미함.

  * DBMS는 이용자가 쉽게 데이터베이스를 구축·유지할 수 있게 하는 소프트웨어. (데이터베이스 시스템 = 데이터베이스 + DBMS)

■ 데이터베이스 특징

◆ 일반적 특징

- 통합된 데이터 : 동일한 내용의 데이터가 중복되어 있지 않음.
- 저장된 데이터 : 컴퓨터가 접근할 수 있는 저장매체에 저장됨.
- 공용 데이터 : 여러 사용자가 서로 다른 목적으로 데이터베이스의 데이터를 공동으로 이용. 대용량화되고 구조가 복잡함.
- 변화되는 데이터 : 새로운 데이터의 삽입, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 정확한 데이터 유지

◆ 정보 축적 및 전달 측면 특징

- 기계가독성 : 대량의 정보를 일정한 형식에 따라 컴퓨터 등의 정보처리기기가 읽고 쓸 수 있음.
- 검색가독성 : 다양한 방법으로 필요한 정보 검색 가능.
- 원격조작성 : 정보통신망을 통하여 원거리에서도 즉시 온라인으로 이용 가능.

◆ 정보 이용 측면 특징

- 이용자의 정보 요구에 따라 다양한 정보를 신속하게 획득 가능.
- 원하는 정보를 정확하고 경제적으로 찾아낼 수 있음.

◆ 정보 관리 측면 특징

- 정보를 일정한 질서와 구조에 따라 정리, 저장, 검색, 관리할 수 있음
- 방대한 양의 정보를 체계적으로 축적하고 새로운 내용의 추가나 갱신이 용이함.

◆ 정보기술 발전 측면 특징

- 데이터베이스는 정보처리, 검색·관리 소프트웨어, 관련 하드웨어, 정보 전송을 위한 네트워크 기술의 발전을 견인할 수 있음.

◆ 경제·산업 측면 특징

- 다양한 정보를 필요에 따라 신속하게 제공·이용할 수 있는 인프라.
- 경제, 산업, 사회 활동의 효율성을 제고하고 국민의 편의를 증진하는 수단.

■ 기업내부 데이터베이스 활용

❖ 시대별

◆ 1980년대

- OLTP (On-Line Transaction Processing)

  호스트 컴퓨터가 데이터베이스를 액세스하고 바로 처리 결과를 돌려보내는 형태.

  데이터베이스의 데이터를 수시로 갱신하는 프로세싱.

  주문입력시스템, 재고관리시스템 등.

- OLAP (On-Line Analytical Processing)

  다양한 비즈니스 관점에서 데이터에 접근하여 의사 결정에 활용할 수 있는 정보를 얻을 수 있게 해주는 기술.

  데이터베이스의 데이터를 조회하는 프로세싱. 다차원의 데이터를 대화식으로 분석하기 위한 기술.

  OLTP에서 처리된 트랜잭션 데이터를 분석해 제품의 판매 추이, 구매 성향 파악, 재무회계 분석 등 프로세싱.

◆ 2000년대

- CRM (고객관계관리; Customer Relationship Management)

  고객 중심 자원 극대화.

  고객 특성에 맞게 마케팅 활동을 계획·지원·평가하는 과정.

- SCM (공급망관리; Supply Chain Management)

  모든 공급망 단계를 최적화해 수요자가 원하는 제품을 원하는 시간과 장소에 제공하는 것.

  거래관계에 있는 기업들간 IT를 이용한 실시간 정보공유.

❖ 분야별

◆ 제조 부문

- 클라이언트/서버 기반의 내부 정보 시스템 ⇨ 웹 기반의 데이터베이스
- ERP ⇨ CRM

- RTE를 통해 협업적 IT화의 비중 확대

* ERP (Enterprise Resource Planning) : 경영자원을 하나의 통합 시스템으로 재구축

* BI (Business Intelligence) : 의사결정에 활용하는 일련의 프로세스. 의사결정 지원을 위한 리포트 중심의 도구.

* CRM (Customer Relationship Management) : 고객 중심 자원 극대화

* RTE (Real-Time Enterprise) : 부문별 전산화에서 회사 전 부문의 정보를 하나로 통합

◆ 금융 부문

- 2000년대 초 : EAI, ERP, e-CRM을 통한 정보 공유 및 통합, 고객 정보의 전략적 활용 시작

- 2000년대 중 : Data Warehouse 도입을 통한 DB 활용 마케팅 강화. DW를 위한 최적화와 BI 기반의 시스템 구축

* EAI (Enterprise Application Integration) : 필요한 정보를 중앙 집중적으로 통합, 관리, 사용할 수 있는 환경

* EDW (Enterprise Data Warehouse) : 기존 DW를 확장한 모델로 BPR, CRM, BSC 같은 다양한 분석 애플리케이션을 위한 원천.
기업 리소스의 유기적 통합, 다원화된 관리 체계 정비, 데이터의 중복 방지를 위해 시스템 재설계

◆ 유통 부문

- CRM과 SCM 구축

- 상거래를 위한 인프라와 KMS를 위한 백업시스템 구축

- RFID의 등장으로 유비쿼터스 시대 준비

* KMS (Knowledge Management System) : 지식관리시스템. 기업 경영을 지식이라는 관점에서 새롭게 조명하는 접근방식

* RFID (Radio Frequency ID) : 주파수를 이용해 아이디를 식별하는 시스템

■ 빅데이터 정의

◆ 좁은 범위의 정의

- 3V로 요약되는 데이터 자체의 특성 변화에 초점
- 규모 volume
- 형태 다양성 variety
- 속도 velocity

◆ 중간 범위의 정의

- 데이터 자체뿐 아니라 처리, 분석 기술적 변화까지 포함
- 새로운 데이터 처리, 저장, 분석 기술 및 아키텍처
- 클라우드 컴퓨팅 활용

◆ 넓은 범위의 정의

- 인재, 조직 변화까지 포함
- 데이터 사이언티스 같은 새로운 인재 필요
- 데이터 중심 조직

* 기존 방식으로는 얻을 수 없었던 통찰 및 가치 창출
* 사업방식, 시장, 사회, 정부 등에서 변화와 혁신 주도

■ 빅데이터 출현 배경

- 빅데이터 현상은 없었던 것이 새로 등장한 것이 아니라 기존의 데이터, 처리방식, 다루는 사람과 조직 차원에서 일어나는 변화 (패러다임 전환)
- 산업계 : 고객 데이터 축적
- 학계 : 거대 데이터 활용, 과학 확산
- 기술 발전 : 관련기술의 발달 (디지털화, 저장기술 발달, 인터넷 보급, 모바일 혁명, 클라우드 컴퓨팅)

■ 빅데이터 기능 (비유)

- 산업혁명의 석탄, 철 : 제조업 뿐 아니라 서비스 분야의 생산성을 획기적으로 끌어올려 사회·경제·문화·생활 전반에 혁명적 변화를 가져옴
- 21세기의 원유 : 경제 성장에 필요한 정보를 제공함으로써 산업 전반의 생산성을 한 단계 향상시키고 기존에 없던 새로운 범주의 산업을 만듬
- 렌즈 : 현미경이 생물학 발전에 미쳤던 영향만큼이나 데이터가 산업 발전에 영향을 줌
- 플랫폼 : 공동 활용의 목적으로 구축된 유,무형의 구조물로써 비즈니스에 활용되면서 플랫폼 역할을 함

■ 빅데이터가 만들어내는 변화

◆ 사전처리 ⇨ 사후처리

- 필요한 정보만 수집하고 필요하지 않은 정보를 버리는 시스템 ⇨ 가능한 한 많은 데이터를 모으고 그 데이터의 숨은 정보 찾기

◆ 표본조사 ⇨ 전수조사

- 데이터 수집 비용 감소와 클라우드 컴퓨팅 기술 발전으로 데이터 처리 비용 감소
- 전수조사를 통해 샘플링이 주지 못하는 패턴이나 정보 발견

◆ 질 ⇨ 양

- 데이터가 지속적으로 추가될 경우 양질의 정보가 오류 정보보다 많아 전체적으로 좋은 결과 산출에 긍정적인 영향을 미친다는 추론

◆ 인과관계 ⇨ 상관관계

- 상관관계를 통해 특정 현상의 발생 가능성 포착되고 그에 상응하는 행동을 하도록 추천
- 데이터 기반의 상관관계 분석이 주는 인사이트

■ 빅데이터 가치 측정 어려움

- 데이터 활용 방식 : 재사용, 재조합, 다목적용 데이터 개발 등이 일반화되면서 특정데이터를 언제·어디서·누가 활용할지 알 수 없음
- 새로운 가치 창출 : 데이터가 기존에 없던 가치를 창출함
- 분석 기술 발전 : 현재는 가치가 없는 데이터일지라도, 추후에 새로운 분석 기법이 등장하면 거대한 가치를 지닌 데이터가 될 수 있음

■ 빅데이터 영향

- 생활 전반의 스마트화

◆ 기업

- 혁신, 경쟁력 제고, 생산성 향상
- 빅데이터를 활용해 소비자의 행동을 분석하고 시장 변동을 예측해 비즈니스 모델을 혁신하거나 신사업 발굴함

◆ 정부

- 환경 탐색, 상황 분석, 미래 대응
- 기상, 인구이동, 각종 통계, 법제 데이터 등을 수집해 사회 변화를 추정하여 관련 정보 추출함

◆ 개인

- 목적에 따른 활용

■ 빅데이터 활용 사례

◆ 기업

- 구글은 사용자의 로그 데이터를 활용한 검색 엔진 개발, 기존 페이지랭크 알고리즘을 혁신하여 검색 서비스를 개선함
- 월마트는 고객의 구매패턴을 분석해 상품진열에 활용함

◆ 정부

- 실시간 교통정보 수집, 기후 정보, 각종 지질 활동, 소방 서비스 등 다양한 국가 안전 확보 활동을 위해 실시간 모니터링을 활용함.
- 의료와 교육 개선을 위해 빅데이터를 활용해 해결책을 모색함.

◆ 개인

- 정치인은 사회관계망 분석을 통해 유세 지역을 선정하고 유권자에게 영향을 줄 수 있는 내용을 선정해 효과적인 선거활동을 함
- 가수는 팬들의 음악 청취 기록 분석을 통해 공연에서 부를 노래 순서를 짜는데 활용함

■ 빅데이터 활용 기본 테크닉

◆ 연관규칙 학습 (Association rule learning)

- 변인들 간에 주목할 만한 상관관계가 있는지 찾아내는 방법
- 커피를 구매하는 사람이 탄산음료를 더 많이 사는가?

◆ 유형 분석 (Classification tree analysis)

- 문서를 분류하거나 조직을 그룹으로 분류할 때 사용하는 방법
- 이 사용자는 어떤 특성을 가진 집단에 속하는가?

◆ 유전자 알고리즘 (Genetic algorithms)

- 최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 점진적으로 진화시켜 나가는 방법
- 최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?

◆ 기계 학습 (Machine learning)

- 훈련 데이터로부터 학습한 알려진 특성을 활용해 예측하는 방법
- 기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화 중에서 어떤 것을 가장 보고 싶어할까?

◆ 회귀 분석 (Regression analysis)

- 독립변수를 조작하며 종속변수가 어떻게 변하는지를 보면서 두 변인의 관계를 파악할 때 사용하는 방법
- 구매자의 나이가 구매 차량의 타입에 어떤 영향을 미치는가?

◆ 감정 분석 (Sentiment analysis)

- 특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석하는 방법
- 새로운 환불 정책에 대한 고객의 평가는 어떤가?

◆ 소셜 네트워크 분석 (Social network analysis)

- 특정인과 다른 사람이 몇 촌 정도의 관계인가를 파악할 때 사용하고 영향력있는 사람을 찾아낼 때 사용하는 방법

■ 빅데이터 시대의 위기요인

◆ 사생활 침해

- 개인정보가 포함된 데이터를 목적 외에 활용할 경우 사생활 침해를 넘어 사회·경제적 위협으로 변형
- 여행 사실을 트위트 한 사람의 집을 강도가 노리는 고전적 사례 발생
- 익명화 기술 발전이 필요함

◆ 책임 원칙 훼손

- 빅데이터 기반 분석과 예측 기술이 발전하면서 정확도가 증가한 만큼, 분석 대상이 되는 사람들은 예측 알고리즘의 희생양이 될 가능성도 증가
- 범죄 예측 프로그램에 의해 범행을 저지르기 전에 체포, 자신의 신용도와 무관하게 대출 거정
- 민주주의 국가의 형사 처벌은 잠재적 위협이 아닌 명확하게 행동한 결과에 대해 책임을 묻고 있음

◆ 데이터 오용

- 데이터 과신, 잘못된 지표의 사용으로 인한 잘못된 인사이트를 얻어 비즈니스에 적용할 경우 직접 손실 발생

■ 위기요인에 따른 통제방안

◆ 동의에서 책임으로

- 개인정보 제공자의 ‘동의’를 통해 해결하기보다 ‘개인정보 사용자의 책임’으로 해결

◆ 결과 기반 책임 원칙 고수

- 특정인의 ‘성향’에 따라 처벌하는 것이 아닌 ‘행동 결과’를 보고 처벌
- 잘못된 예측 알고리즘을 근거로 불이익을 줄 수 없으며, 이에 따른 피해 최소화 장치 마련

◆ 알고리즘 접근 허용

- 예측 알고리즘의 부당함을 반증할 수 있는 방법을 명시해 공개할 것을 주문함
- 불이익을 당한 사람들을 대변할 전문가가 필요하게 됨

■ 빅데이터 활용 3요소

◆ 데이터

- 모든 것의 데이터화
- 수많은 센서들이 인터넷에 연결되는 사물인터넷 시대

◆ 기술

- 진화하는 알고리즘, 인공지능

◆ 인력

- 데이터사이언티스트, 알고리즈미스트 역할 증대

■ 빅데이터 회의론의 원인 및 진단

◆ 투자효과를 거두지 못했던 부정적 학습 효과 (과거의 CRM)

- 도입만 하면 모든 문제를 한번에 해소할 것처럼 강조
- 막상 도입하면 어떻게 활용하고 어떻게 가치를 뽑아내야 할지 난감

◆ 빅데이터 성공사례가 기존 분석 프로젝트를 포함해 놓은 것이 많음

- 굳이 빅데이터가 필요없는 영역이 있음 (우수고객, 이탈예측, 구매패턴 분석 등)
- 국내 빅데이터 업체들이 CRM 분석 성과를 빅데이터 분석으로 과대포장

■ 일차원적인 분석 vs 전략도출 위한 가치기반 분석

◆ 산업별 분석 애플리케이션

- 금융 서비스 : 신용점수 산정, 사기 탐지, 가격 책정, 프로그램 트레이딩, 클레임 분석, 고객 수익성 분석
- 병원 : 가격 책정, 고객 로열티, 수익 관리
- 에너지 : 트레이딩, 공급·수요 예측
- 정부 : 사기 탐지, 사례 관리, 범죄 방지, 수익 최적화
- 소매업 : 판촉, 매대 관리, 수요 예측, 재고 보충, 가격 및 제조 최적화
- 제조업 : 공급사슬 최적화, 수요 예측, 재고 보충, 보증서 분석, 맞춤형 상품 개발, 신상품 개발
- 운송업 : 일정 관리, 노선 배정, 수익 관리
- 헬스케어 : 약품 거래, 예비 진단, 질병 관리
- 커뮤니케이션 : 가격 계획 최적화, 고객 보유, 수요 예측, 생산능력 계획, 네트워크 최적화, 고객 수익성 관리
- 서비스 : 콜센터 직원관리, 서비스-수익 사슬 관리
- 온라인 : 웹 매트릭스, 사이트 설계, 고객 추천

◆ 일차적인 분석의 문제점

- 일차적인 분석을 통해서 해당 부서나 업무 영역에서는 효과를 얻을 수 있음
- 일차적인 분석만으로는 환경변화와 같은 큰 변화에 제대로 대응하거나 고객 환경의 변화를 파악하고 새로운 기회를 포착하기 어려움.
- 급변하는 환경에서는 분석을 일차적 차원에서 점증적, 전술적으로 사용하면 성과는 미미할 수 있음

◆ 전략도출 가치기반 분석

- 해당 사업에 중요한 기회를 발굴하고 주요 경영진의 지원을 얻어낼 수 있음
- 일차원적인 분석을 통해 점점 분석 경험을 쌓아야하고 작은 성공을 거두면 분석의 활용범위를 더 넓고 전략적으로 변화시켜야함
- 사업성과를 견인하는 요소들과 차별화를 꾀할 기회에 대해 전략적 인사이트를 주는 가치기반 분석단계로 나아가야함

■ 데이터 사이언스의 의미

- 공학, 수학, 통계학, 컴퓨터공학, 시각화, 해커의 사고방식, 해당 분야의 전문지식을 종합하여 데이터로 의미있는 정보를 추출하는 학문
- 정형 또는 비정형을 막론하고 숫자, 문자, 영상 등 다양한 유형의 데이터를 대상으로 분석하고 효과적으로 구현하고 전달하는 과정

■ 데이터 사이언스의 구성요소

◆ 데이터 사이언스의 영역

- 데이터 처리와 관련된 IT 영역 : 시그널 프로세싱, 프로그래밍, 데이터 엔지니어링, 데이터 웨어하우스, 고성능 컴퓨팅
- 분석 Analytics 영역 : 수학, 확률모델, 머신러닝, 분석학, 패턴인식과 학습, 불확실성 모델링
- 비즈니스 컨설팅 영역 : 커뮤니케이션, 프레젠테이션, 스토리텔링, 시각화

◆ 데이터 사이언티스트 역할

- 데이터 홍수 속에서 데이터 소스를 찾고 복잡한 대용량 데이터를 구조화, 불완전한 데이터를 서로 연결해야함
- 호기심을 갖고 문제의 이면을 파고들고 질문들을 찾고 검증 가능한 가설을 세우는 능력이 있어야함
- 스토리텔링, 커뮤니케이션, 창의력, 열정, 직관력, 비판적 시각, 글쓰기 능력, 대화능력 등을 갖춰야함

■ 데이터 사이언티스트 요구 역량

◆ Hard Skill

- 빅데이터에 대한 이론적 지식 : 관련 기법에 대한 이해와 방법론 습득
- 분석 기술에 대한 숙련 : 최적의 분석 설계 및 노하우 축적

◆ Soft Skill

- 통찰력 있는 분석 : 창의적 사고, 호기심, 논리적 비판
- 설득력 있는 전달 : 스토리텔링, 비주얼라이제이션
- 다분야간 협력 : 커뮤니케이션

■ 전략적 통찰력과 인문학의 부활

◆ 통찰력있는 분석

- 직관과 전략, 경영 프레임워크 경험의 혼합을 통해 통찰력있는 분석을 수행할 수 있어야함
- 본인 회사 뿐 아니라 전체 업계의 방향과 고객이 무엇을 중시하는지에 대한 이해가 필요함
- 좁은 시각으로 나무만 보는 것이 아니라 넓은 시각으로 숲을 볼 수 있어야함

◆ 외부 환경의 변화로 인문학의 열풍

- 컨버전스 ⇨ 디버전스 : 단순세계화에서 복잡한 세계화로의 변화
- 생산 ⇨ 서비스 : 비즈니스 중심이 제품생산에서 서비스로 이동
- 생산 ⇨ 시장창조 : 공급자 중심의 기술 경쟁에서 무형자산의 경쟁으로 변화

■ 데이터 사이언티스트에 요구되는 인문학적 사고의 특성과 역할

◆ 과거

- 정보 : 무슨 일이 일어났는가 ⇨ 보고서 작성
- 통찰력 : 어떻게, 왜 일어났는가 ⇨ 모델링, 실험설계

◆ 현재

- 정보 : 무슨 일이 일어나고 있는가 ⇨ 경고
- 통찰력 : 차선 행동은 무엇인가 ⇨ 권고

◆ 미래

- 정보 : 무슨 일이 일어날 것인가 ⇨ 추출
- 통찰력 : 최악 또는 최선의 상황은 무엇인가 ⇨ 예측, 최적화, 시뮬레이션

■ 빅데이터 회의론을 넘어 가치 패러다임의 변화

- (과거) 디지털화 : 아날로그 세상을 어떻게 효과적으로 디지털화하는지가 과거의 가치 창출 원천
- (현재) 연결 : 디지털화된 정보와 대상들 서로 연결, 이 연결을 더 효과적이고 효율적으로 제공하는가가 성공 요인
- (미래) 에이전시 : 사물인터넷과 함께 연결이 증가하고 복잡해진 연결을 얼마나 효과적이고 믿을 수 있게 관리하는가가 이슈

■ 데이터 사이언스의 한계

- 분석 과정에서는 가정 등 인간의 해석이 개입되는 단계를 반드시 거침
- 분석 결과가 의미하는 바는 사람에 따라 전혀 다른 해석과 결론을 내릴 수 있음
- 아무리 정량적인 분석이라도 모든 분석은 가정에 근거한다는 사실

■ DBMS 정의

- Date Base Management System의 약자
- 데이터베이스를 관리하여 응용프로그램들이 데이터베이스를 공유하며 사용할 수 있는 환경을 제공하는 소프트웨어
- 데이터베이스를 구축하는 틀을 제공하며 효율적인 데이터 검색, 저장 기능 등을 제공

■ DBMS 종류

◆ 관계형 DBMS

- 데이터를 column과 row를 이루는 하나 이상의 테이블로 정리하며 고유키로 각 row를 식별함
- row는 레코드나 튜플로 부르며, 일반적으로 각 테이블은 하나의 개체 타입을 대표함
- row는 그 개체 종류의 인스턴스를 대표하며 column은 그 인스턴스의 속성이 되는 값들을 대표함

◆ 객체지향 DBMS

- 일반적으로 사용되는 테이블 기반의 관계형 DB와 다르게 정보를 '객체' 형태로 표현하는 데이터베이스 모델

◆ 네트워크 DBMS

- 레코드들이 노드로, 레코드들 사이의 관계가 간선으로 표현되는 그래프를 기반으로 하는 데이터베이스 모델

◆ 계층형 DBMS

- 트리 구조를 기반으로 하는 계층 데이터베이스 모델

■ SQL 정의

- Structured Query Language의 약자
- 데이터베이스를 사용할 때 데이터베이스에 접근할 수 있는 데이터 베이스의 하부 언어
- 단순한 질의 기능 뿐만 아니라 완전한 데이터의 정의와 조작 기능을 갖추고 있음
- 테이블을 단위로 연산을 수행함

■ SQL 집계함수

- AVG : (수치형) 지정한 열의 평균 값 반환
- COUNT : (수치형, 문자형) 테이블의 특정 조건이 맞는 것의 개수 반환
- SUM : (수치형) 지정한 열의 총합을 반환
- STDDEV : (수치형) 지정한 열의 분산 반환
- MIN : (수치형) 지정한 열의 최소값 반환
- MAX : (수치형) 지정한 열의 최대값 반환

◆ 간단한 SQL 문장 해석

```
SELECT NAME, GENDER, SALARY        ⇨ NAME,GENDER, SALARY 라는 이름의 데이터 추출
FROM CUSTOMERS                     ⇨ CUSTOMERS 라는 이름의 테이블 지정
WHERE AGE BETWEEN 20 AND 39        ⇨ AGE 가 20과 39 사이에 있는 데이터 추출
```

■ 개인정보 비식별 기술

- 데이터에서 개인을 식별할 수 있는 요소를 삭제하거나 다른 값으로 대체하는 등의 방법으로 개인을 알아볼 수 없도록 하는 기술

◆ 데이터 마스킹

- 데이터 길이, 유형, 형식과 같은 속성을 유지한 채, 새롭고 읽기 쉬운 데이터를 익명으로 생성
- 예) 홍길동, 35세, 서울 거주, 한국대 재학 ⇨ 홍**, 35세, 서울 거주, **대학 재학

◆ 가명처리

- 개인정보 주체의 이름을 다른 이름으로 변경
- 다른 값으로 대체할 시 일정한 규칙이 노출되지 않도록 주의
- 예) 홍길동, 35세, 서울 거주, 한국대 재학 ⇨ 임꺽정, 30대, 서울 거주, 국내대학 재학

◆ 총계처리

- 데이터의 총합 값을 보임으로서 개별 데이터의 값을 보이지 않도록 함
- 예) 임꺽정 180, 홍길동 170, 이콩쥐 160, 김팥쥐 150 ⇨ 물리학과 학생 키 합 660, 평균 키 165

◆ 데이터값 삭제

- 데이터 공유, 개방 목적에 따라 데이터에 구성된 값 중에 필요 없는 값 또는 개인 식별에 중요한 값 삭제
- 개인과 관련된 날짜 정보는 연단위로 처리
- 예) 홍길동, 35세, 서울 거주, 한국대 졸업 ⇨ 35세, 서울 거주
- 예) 주민등록번호 901206 - 1234567 ⇨ 90년대 생, 남자

◆ 데이터 범주화

- 데이터 값을 범주의 값으로 변환하여 값을 숨김
- 예) 홍길동, 35세 ⇨ 홍씨, 30~40세

■ 무결성과 레이크

◆ 데이터 무결성 (integrity)

- 데이터베이스 내의 데이터에 대한 정확한 일관성, 유효성, 신뢰성을 보장하기 위해 데이터 변경/수정 시 여러 제한을 두어 데이터 정확성 보증
- 무결성 제한의 유형 : 개체 무결성, 참조 무결성, 범위 무결성

◆ 데이터 레이크 (lake)

- 수 많은 정보 속에서 의미있는 내용을 찾기 위해 방식에 상관없이 데이터를 저장하는 시스템
- 대용량의 정형 및 비정형 데이터를 저장할 뿐만 아니라 접근도 쉽게 할 수 있는 대규모의 저장소
- Apache Hadoop, Teradata Integrated Big Data Platform 1700 등과 같은 플랫폼으로 구성된 솔루션 제공

■ 빅데이터 분석 기술

◆ 하둡 (Hadoop)

- 여러 개의 컴퓨터를 하나인 것처럼 묶어 대용량 데이터를 처리하는 기술
- 분산파일시스템(HDFS)을 통해 수 천대의 장비에 대용량 파일을 저장할 수 있는 기능 제공
- 맵리듀스로 HDFS에 저장된 대용량의 데이터들을 대상으로 SQL을 이용해 사용자의 질의를 실시간으로 처리하는 기술 제공

◆ Apache Spark

- 실시간 분산형 컴퓨팅 플랫폼으로써 스칼라로 작성이 되어 있지만 스칼라, 자바, R, 파이썬, API를 지원함
- In-Memory 방식으로 처리하기 때문에 하둡에 비해 처리속도가 빠름

◆ Smart Factory

- 공장 내 설비와 기계에 사물인터넷이 설치되어 공정 데이터가 실시간으로 수집되고 데이터에 기반한 의사결정이 이뤄짐으로써 생산성 극대화

◆ Machine Learning & Deep Learning

- 머신러닝은 인공지능의 연구분야 중 하나로, 인간의 학습 능력과 같은 기능을 컴퓨터에서 실현하고자하는 기술
- 딥러닝은 컴퓨터가 사람처럼 스스로 학습할 수 있게 하기 위하여 인공 신경망 등의 기술을 기반하여 구축한 기계 학습 기술 중 하나

■ B2B와 B2C

- B2B : 기업과 기업 사이의 거래를 기반으로 한 비즈니스 모델 (기업이 필요로 하는 장비, 재료, 공사입찰 등)
- B2C : 기업과 고객 사이의 거래를 기반으로 한 비즈니스 모델 (이동통신사, 여행회사, 신용카드회사, 옥션, 지마켓 등)

■ 블록체인

- 거래정보를 하나의 덩어리로 보고 이를 차례로 연결한 거래장부
- 기존 금융회사의 경우 중앙 집중형 서버에 거래 기록을 보관
- 블록체인은 거래에 참여하는 모든 사용자에게 거래 내역을 보내주며 거래 때마다 이를 대조해 데이터 위조를 막는 방식 사용

■ 데이터 유형

◆ 정형 데이터

- 형태가 있으며 연산이 가능함. 주로 관계형 데이터베이스에 저장됨.
- 데이터 수집 난이도가 낮고 형식이 정해져 있어 처리가 쉬운 편
- 관계형 데이터베이스, 스프레드시트, CSV 등

◆ 반정형 데이터

- 형태가 있으며 연산이 불가능함. 주로 파일로 저장됨. 데이터 내부에 메타데이터를 갖고 있음.
- 데이터 수집 난이도가 중간. 보통 API 형태로 제공되기 때문에 데이터처리 기술(파싱)이 요구됨
- XML, HTML, JSON, 로그형태 등

◆ 비정형 데이터

- 형태가 없으며 연산이 불가능함. 주로 NoSQL에 저장됨.
- 데이터 수집 난이도가 높음. 텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어려움.
- 소셜데이터, 영상, 이미지, 음성, 텍스트 등

❖ 데이터베이스 구성요소

- 메타데이터 : 데이터에 관한 구조화된 데이터로, 다른 데이터를 설명해주는 데이터
- 인덱스 : 데이터베이스 내의 데이터를 신속하게 정렬하고 탐색하게 해주는 구조
- 트리거 : 임의의 테이블에 연관된 데이터베이스 객체로서 테이블에 특정 이벤트 발생 시 활성화됨
- 스키마 : 데이터베이스 구조와 제약 조건에 관한 전반저긴 명세를 기술한 메타데이터 집합
- 데이터마트 : 데이터 웨어하우스에서 정의된 접근계층. 데이터 웨어하우스에서 데이터를 꺼내 사용자에게 제공하는 역할.
- 데이터모델 : 현실 세계의 정보들을 컴퓨터에 표현하기 위해 단순한, 추상화하여 체계적으로 표현한 개념적 모형

❖ 데이터마이닝 : 대용량 데이터에서 의미있는 정보를 추출하여 의사결정에 활용하는 기술

❖ 난수화 : 사생활 침해를 막기 위해 개인정보를 무작위 처리하는 등 데이터가 본래 목적 외에 가공되고 처리되는 것을 방지하는 기술

❖ 데이터 웨어하우스 : 기업의 의사결정 과정을 지원하기 위한 주제 중심적으로 통합적이며 시간성을 가지는 비휘발성 데이터 집합