2023-08-02-ADsP-1과목. 데이터의 이해

ADsP-데이터분석가 준전문가-1과목. 데이터의 이해

1장. 데이터의 이해

1절. 데이터와 정보

데이터의 정의

- 1646년 영국 문헌에 처음 등장, '주어진 것'이란 의미로 라틴어 Dare(주다)의 과거분사형
- 추론과 추정의 근거를 이루는 사실(Oxford Dictionary)
- 단순한 객체로서의 가치뿐만 아니라 다른 객체와의 상호관계에서 가치를 갖는 것

데이터의 특성

- 존재적 특성: 객관적 사실(Fact, Raw Material)
- 당위적 특성: 추론, 예측, 전망, 추청을 위한 근거(Basis)

구분	형태	예	특징
정성적 데이터 (Qualitative Data)	언어, 문자 등	회사 매출이 증가함 등	저장, 검색, 분석에 많은 비용이 소모됨
정량적 데이터 (Quantitative Data)	수치, 도형 기호 등	나이, 몸무게, 주가 등	정형화된 데이터로 비용 소모가 적음

> 정성적 데이터:
>
> - **질적 자료**
> - 비정형 데이터, 주관적 내용 통계분석이 어려움
>
> 정량적 데이터:
>
> - **양적 자료**
> - 정형 데이터 , 객관적 내용 통계분석이 용이함
> - 데이터분석에 선호됨

#adsp중요

암묵지: 학습과 경험을 통해 개인에게 체화되어 있지만 겉으로 드러나지 않는 지식

1. 내면화(Internalization)
    - 학습과 체험을 통해 개인이 습득, 겉으로 드러나지 않는 지식, 무형의 지식
    - ex) 김장 김치 담그기, 자전거 타기 등

2. 공통화(Socialization)
    - 내면화된 지식을 조직의 지식으로 만드는 과정

형식지: 문서나 매뉴얼처럼 형상화된 지식 3. 표출화(Externalization) - 개인의 암묵지를 언어나 기호, 숫자 등의 형태로 표출화 4. 연결화(Combination) - 표출화 된 것을 다시 다른 개인이 본인의 지식에 연결

> 내공->표연

암묵지와 형식지의 상호작용 관계

1. 공통화(Socialization): 암묵지를 타인에게 알려준다
2. 표출화(Externalization): 암묵지를 책 등 형식지로 만든다
3. 연결화(Combination): 책 등에 자신이 아는 새로운 지식 추가
4. 내면화(Internalization): 책 등을 보고 타인들이 암묵적 지식 습득

#adsp중요

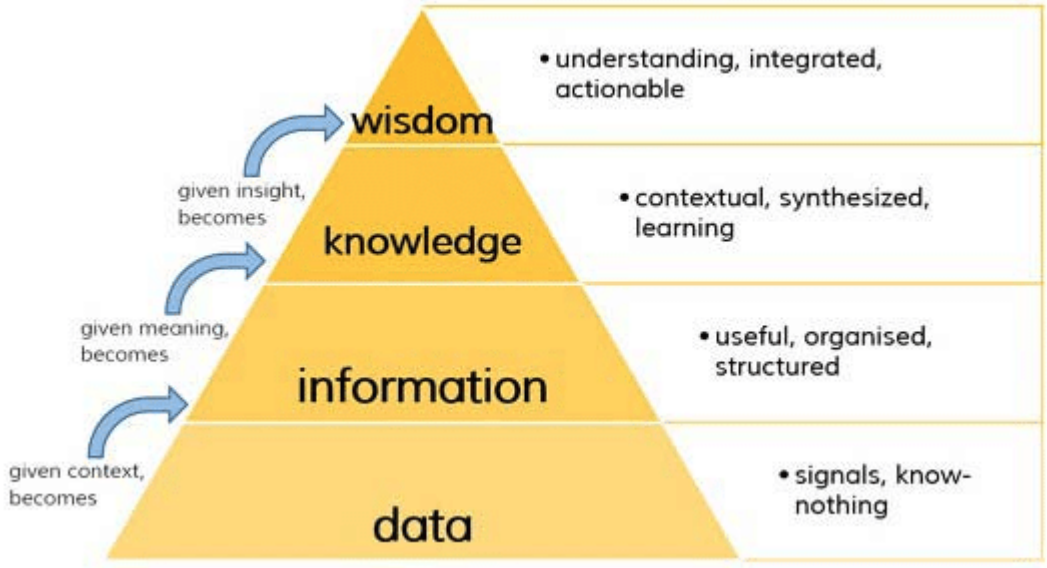
DIKW의 정의

- 데이터(Data): 개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실
- 정보(Information): 데이터의 가공, 처리와 데이터간 연관관계 속에서 의미가 도출된 것
- 지식(Knowledge): 데이터에서 유의미한 정보를 분류하고 개인적인 경험을 결합시켜 고유의 지식으로 내재화된 것
- 지혜(Wisdom): 지식의 축적과 아이디어가 결합된 창의적인 산물

단계	설명	관련 시스템
지혜(Wisdom)	- 지식에 유연성을 더하고, 상황이나 맥락에 맞게 규칙을 적용하는 것 - 근본원리에 대한 깊은 이해를 바탕으로 도출되는 창의적 아이디어 ex) A마트의 다른 상품들도 B마트보다 쌀 것이라고 예측	비즈니스 인텔리전스(BI)
지식(Knowledge)	- 정보를 일반화하고 체계화하여 즉시 적용 및 활용 가능한 형태 - 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물 ex) 연필을 살 때는 A마트로 가는 것이 유리	지식 관리 시스템(KMS), 전사적 지식 포탈(EKP)
정보(Information)	- 데이터 중 사용자가 필요로 하는 데이터 - 사용자의 필요에 의해 정제 및 가공된 데이터 ex) A마트의 연필가격이 더 저렴함	데이터 웨어하우스, OLAP
데이터(Data)	- 관찰, 측정을 통해서 수집된 사실이나 값, 수치, 문자 등 가공되지 않은 원본 데이터 ex) A마트 펜 500원, 연필 200원, 라면 3000원.. B마트 연필 300원	데이터베이스, OLTP, CDC, ETL, 데이터 레이크

#adsp중요

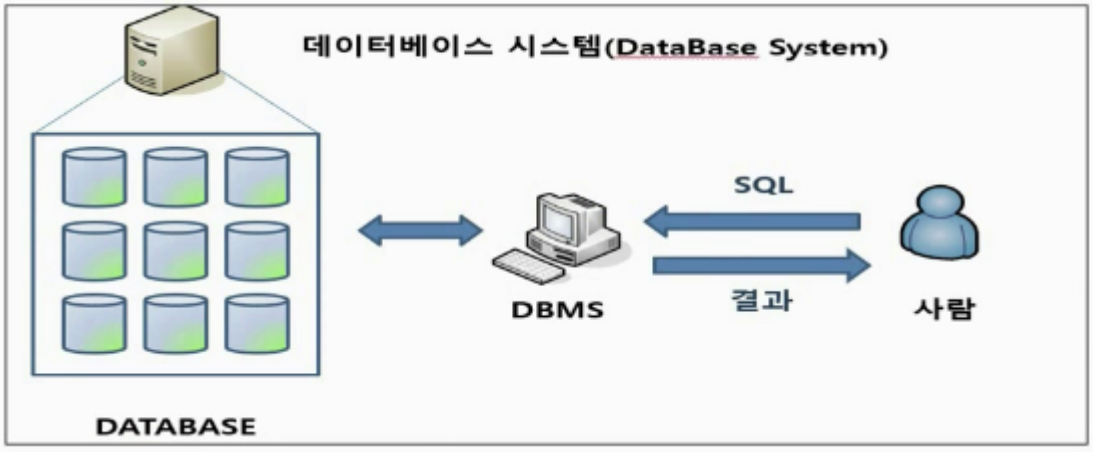
DIWK 피라미드



출처: https://itwiki.kr/w/DIKW_%ED%94%BC%EB%9D%BC%EB%AF%B8%EB%93%9C

2절. 데이터베이스 정의와 특징

- 1950년대: 미국에서 군대의 군비상황을 집중 관리하기 위하여 컴퓨터 도서관을 설립하며 DataBase 탄생



#adsp중요

데이터베이스의 일반적인 특징

데이터베이스 특징	설명
통합된 데이터 Integrated Data	- 동일한 내용의 데이터가 중복되어 있지 않다는 것을 의미 - 데이터 중복은 관리상의 복잡한 부작용을 초래
저장된 데이터 Stored Data	- 자기 디스크나 자기 테이프 등과 같이 컴퓨터가 접근할 수 있는 저장 매체에 저장되는 것을 의미 - 데이터베이스는 기본적으로 컴퓨터 기술을 바탕으로 한 것
공용 데이터	- 여러 사용자가 서로 다른 목적으로 뎅터를 공동으로 이용한다는 것을 의미

데이터베이스 특징	설명
Shared Data	- 대용량화되고 구조가 복잡한 것이 보통
변화되는 데이터 Changable Data	- 데이터베이스에 저장된 내용은 곧 데이터베이스의 현 시점에서의 상태를 나타냄 - 다만 이 상태는 새로운 데이터의 삽입, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 저오학한 데이터를 유지해야 함

데이터베이스의 다양한 측면에서의 특징

측면	특성
정보 이용 측면	- 이용자의 정보 요구에 따라 다양한 정보를 신속하게 획득 - 원하는 정보를 정확하고 경제적으로 찾아낼 수 있다는 특성
정보 관리 측면	- 정보를 일정한 질서와 구조게 따라 정리, 저장, 검색, 관리 할 수 있도록 하여 방대한 양의 정보를 체계적으로 축적하고 새로운 내용의 추가나 갱신이 용이
정보기술 발전 측면	- 데이터 베이스는 정보처리, 검색, 관리 소프트웨어, 관련 하드웨어, 정보 전송을 위한 네트워크 기술의 발전을 견인할 수 있음
경제, 산업 측면	- 다양한 정보를 필요에 따라 신속하게 제공, 이용할 수 있는 인프라라는 특성을 가지고 있어 경제, 산업, 사회 활동의 효율성을 제고하고 국민의 편의를 증진하는 수단으로서 의미를 가짐

3절. 데이터베이스의 활용

1980년대 기업내부 데이터베이스

- OLTP(On-Line Transaction Processing):
  - 호스트 컴퓨터와 온라인으로 접속된 여러 단말간의 처리형태
  - **호스트 컴퓨터가 데이터베이스를 엑세스, 처리하고 결과를 바로 돌려보내는 형태**
  - 데이터베이스의 데이터를 수시로 갱신하는 프로세싱
  - 데이터 갱신 위주
  - ex) 주문입력시스템, 재고관리시스템 등
- OLAP(On-Line Analytical Processing):
  - 다양한 비즈니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근하여 의사 결정에 활용할 수 있는 정보를 얻을 수 있게 해주는 기술
  - OLTP에서 처리된 트랜잭션 데이터를 분석해 제품의 판매 추이, 구매 성향 파악, 재무 회계 분석 등을 프로세싱
  - 데이터 조회 위주

구분	OLTP	OLAP
데이터의 구조	복잡(운영 시스템 계산에 적합)	단순(사업 분석에 적합)
데이터의 갱신	순간적/동적	주기적/정적
응답시간	2, 3 초 ~ 몇 초 이내	수 초 ~ 몇 분까지도 가능
데이터의 범위	과거 30 일 ~ 90 일	과거 5 년 ~ 10 년
데이터 성격	정규/핵심 업무 데이터, mission critical 데이터	비정규/read-only 데이터, index 에 의존
데이터의 크기	수 Giga Byte	수 Tera Byte
데이터의 내용	현재 데이터	기록 보관된, 요약/계산 데이터
데이터 특성	거래 중심	주제 중심
데이터 액세스 빈도	높음	보통 혹은 낮음
데이터의 사용법	고도로 구조화된 연속 처리	고도로 비구조화된 분석처리
질의어의 성격	예언 가능, 주기적	예측하기 어렵고, 특수하다

출처: https://slidesplayer.org/slide/14698809/

2000년대 기업 내부 데이터베이스

고객관계관리 CRM(Customer Relationship Management)

- 기업이 고객 관련 자료를 분석•통합해 고객 중심 자원을 극대화 -> 자원 토대로 고객특성에 맞게 마케팅 활동을 계획•지원•평가 하는 과정
- 일대일 마케팅(One-to-One Marketing), 관계 마케팅(Relationship marketing)에서 진화한 요소들을 기반으로 등장

공급망 관리 SCM(Suppply Chain Management)

- 기업에서 원재료의 생성•유통 등 모든 공급망 단계를 최적화해 수요자가 원하는 제품을 원하는 시간과 장소에 제공하는 것
- 부품 공급업체-생산업체-고객간 거래관계에 있는 기업들 간 IT를 이용한 실시간 정보공유를 통해 시장이나 수요자들이 요구에 기민하게 대응하도록 지원하는 것

분야별 내부 데이터베이스

제조

- ERP(Enterprise Resource Planning)
  - **경영자원을 하나의 통합 시스템으로 재구축**
  - 생산성을 극대화하려는 경영혁신기법을 의미
- BI(Business Intelligence)
  - 보유 데이터를 정리, 분석해 **기업의 의사결정에 활용하는 일련의 프로세스**
- CRM(Customer Relationship Management)
  - 고객관계관리

    - **고객 중심자원을 극대화** -> 고객특성에 맞게 마케팅
- RTE(Real-Time Enterprise)
    - 회사의 주요 경영정보를 통합관리하는 실시간 기업의 새로운 기업경영시스템
    - ERP, SCM, CRM에서 한발 더 나아가 **회사 전 부문의 정보를 하나로 통합**
        - 경영자의 빠른 의사결정을 이끌어냄

금융

- EAI(Enterprise Application Integration)
    - 기업 내 상호 연관된 모든 앱을 유기적으로 연동하여 필요한 **정보를 중앙 집중적으로 통합, 관리 사용**
- EDW(Enterprise Data Warehouse)
    - DW(Data Warehouse)를 전사적으로 확장한 모델로 **BPR과 CRM, BSC 같은 다양한 분석 앱을 위한 원천**이 된다
    - 단순히 정보를 빠르게 전달 보다는 기업 리소스의 유기적 통합, 다원화된 관리 체계 정리, 데이터의 중복 방지 등을 위해 시스템 재설계를 의미

유통

- KMS(Knowledge Management System)
    - **지식관리시스템**
    - 기업 환경이 물품생산사업에서 지적 재산을 중요시하는 지식사회로 발전함에 따라 기업 경영을 지식이라는 관점에서 새롭게 조명하는 접근 방식
- RFID(Radio Frequency, RF)
    - **주파수를 이용해 ID를 식별하는 시스템**
    - **전자태그**
    - 전파를 이용해 먼 거리에서 정보를 인식하는 기술

사회기반구조로서의 데이터베이스

- EDI(Electronic Data Interchange)
    - 주문서, 납품서, 청구서 등 무역에 필요한 각종 서류를 표준화된 양식을 통해 전자적 신호로 바꿔 컴퓨터통신망에 이용하여, 거래처에 전송하는 시스템
- VAN(Value Added Network)
    - 부가가치통신망, 공중 전기통신사업자로부터 통신회선을 차용하여 독자적인 네트워크를 형상하는 것
    - 독자적인 네트워크로 각종 정보를 부포, 영상, 음성 등으로 교환하거나 정보를 축적하거나 또는 복수로 해서 전송하는 등 단순한 통신이 아니라 부가가치가 높은 서비스를 하는 것
- CALS(Commerce At Light Speed)
    - 전자상거래 구축을 위해 기업 내에서 비용 절감과 생산형 향상을 추구할 목적으로 시작
    - 제품의 라이프 사이클(Life Cycle) 전반에 관련된 데이터를 통합한 경영통합정보시스템

2장. 데이터의 가치와 미래

1절. 빅데이터의 이해

빅데이터의 정의

관점에 따른 정의

1. 3V로 요약되는 데이터 자체의 특성 변화에 초점을 맞춘 좁은 **범위의 정의**
2. 데이터 자체뿐 아니라 처리, 분석 기술적 변화까지 포함되는 **중간 범위의 정의**
3. 인재, 조직 변화까지 포함한 **넓은 관점에서의 빅테이터에 대한 정의**

3V

- 양(Volume)
- 다양성(Variety)
- 속도(Velocity)

4V

- 가치(Value)
- 진실성(Veracity)
- 정확성(Validity)
- 휘발성(Volatility)

빅데이터에 거는 기대를 표현한 비유

- 산업혁병의 석탄, 철: 제조업 뿐만 아니라 서비스 분야의 생산성을 획기적으로 끌어올려 사회•경제•문화•생활 전반에 혁명적 변화를 가져올 것으로 기대됨
- 21세기의 원유: 경제 성장에 필요한 정보를 제공함으로써 산업 전반의 생산성을 한 단계 향상시키고, 기존에 없던 새로운 범주의 산업을 만들어낼것으로 전망됨
- 렌즈: 렌즈를 통해 현미경이 생물학 발전에 미쳤던 영향만큼이나 데이터가 산업 발전에 영향을 미칠 것으로 기대됨
- 플랫폼: '공동 활용의 목적으로 구축된 유무형의 구조물'로써의 다양한 서드파티 비즈니스에 활용되면서 플랫폼 역할을 할 것으로 전망됨

과거에서 현재로의 변화

- 사전처리 -> 사후처리
- 표본조사 -> 전수조사
- 질 -> 양
- 인과관계 -> 상관관계

2절. 빅데이터의 가치와 영향

빅데이터 가치 산정이 어려운 이유

- 데이터 활용방식
- 새로운 가치 창출
- 분석 기술 발전

빅데이터가 미치는 영향

분야	영향
기업	혁신, 경쟁력제고, 생산성향상
청부	환경 탐색, 상황분석, 미래대응
개인	목적에 따른 활용

3절. 비즈니스 모델

빅데이터를 활용한 기본 테크닉

테크틱	내용	예시
연관규칙 학습	변인들 간에 주목할 만한 상관관계가 있는지를 찾아내는 방법	커피를 구매하는 사람이 탄산음료를 더 많이 사는가?
유형분석	문서를 분류하거나 조직을 그룹으로 나눌 때, 또는 온라인 수강생들을 특성에 따라 분류할 떄 사용	이 사용자는 어떤 특성을 가진 집단에 속하는가?
유전자 알고리즘	최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 점진적으로 진화시켜 나가는 방법	최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?
기계학습	훈련 데이터로부터 학습한 알려진 특성을 활용해 예측하는 방법	기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화 중에서 어떤 것을 가장 보고 싶어할까?
회귀분석	독립변수를 조작함에 따라, 종속변수가 어떻게 변하는지를 보면서 두 변인의 관계를 파악할 때 사용	구매자의 나이가 구매 차량의 타입에 어떤 영향을 미치는가?
감정분석	특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석	새로운 환불 정책에 대한 고객의 평가는 어떤가?
소셜네트워크분석	특정인과 다른 사람이 몇 촌 정도의 관계인가를 파악할 때 사용하고, 영향력있는 사람을 찾아낼 떄 사용	고객들 간 관계망은 어떻게 구성되어 있나?

4절. 위기 요인과 통제 방안

- 사생활 침해
- 책임 원칙 훼손
- 데이터 오용
- 동의에서 책임으로
- 결과 기반 책임 원칙 고수
- 알고리즘 접근 허용

5절. 미래의 빅데이터

기본 3요소

- 데이터: 모든 것의 데이터화
- 기술: 진화하는 알고리즘, 인공지능
- 인력: 데이터 사이언티스트, 알고리즈미스트

> - 데이터 사이언티스트: 빅데이터에 대한 이론적 지식과 숙련돈 분석 기술을 바탕으로 통찰력, 전달력, 협업 능력을 두루 갖춘 전문이력으로써 빅데이터의 다각적 분석을 통해 인사이트를 도출하고 이를 조직의 전략 방향제시에 활용할 줄 아는 기획자
> - 알고리즈미스트: 데이터 사이언티스트가 한 일로 인해 부당하게 피해가 발생하는 것을 막는 역할을 하며 알고리즘 코딩 해석을 통해 빅데이터 알고리즘에 의해 부당하게 피해를 입는 사람을 구제하는 전문인력

3장. 가치 창조를 위한 데이터 사이언스와 전략 인사이트

1절. 빅데이터 분석과 전략 인사이트

빅데이터 열풍과 회의론

- '빨리 끓어 오른 냄비가 빨리 식는다', 빅데이터 열풍에 대한 거품 우려
- 회의론의 문제점은 빅데이터 분석에서 찾을 수 있는 수많은 가치들을 제대로 발굴해 보기 전에 활용 자체를 사전에 차단해 버릴 수 있다

빅데이터 회의론의 원인 및 진단

투가효과를 거두지 못했던 부정적 학습효과 -> 과거의 고객관계관리(CRM)

- 과거 CRM의 부정적 학습 효과
- 빅데이터 성공사례가 기존 분석 프로젝트를 포함해 놓은 것이 많다
- 단순히 빅데이터에 포커스를 두지 말고 **분석을 통해 가치를 만드는 것에 집중**해야 한다

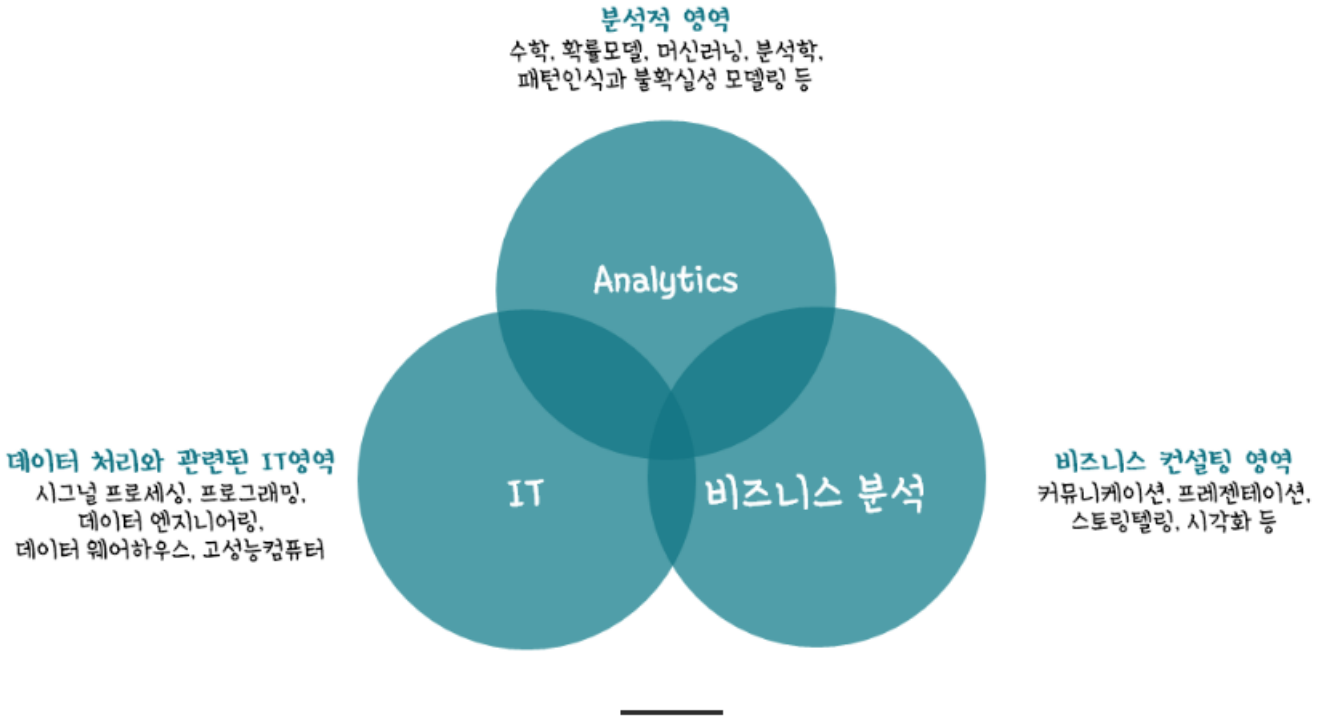
산업별 분석 애플리케이션

산업	일차원적 분석 애플리케이션
금융 서비스	신용점수 산정, 사기 탐지, 가격 책정, 프로그램트에이딩, 클레임분석, 고객 수익성 분석
소매업	판촉, 매대 관리, 수요 예측, 재고 보충, 가격 및 제조 최적화
제조업	공급사슬 최적화, 수요예측, 재고 보충, 보증서 분석, 맞춤형 상품 개발, 신상품 개발
운송업	일정 관리, 노선 배정, 수익 관리
헬스케어	약품 거래, 예비 진단, 질병 관리
병원	가격 책정, 고객 로열티, 수익 관리
에너지	트레이딩, 공급/수요 예측
커뮤니케이션	가격 계획 최적화, 고객 보유, 수요 예측, 생산능력 계획, 네트워크 최적화, 고객 수익성 관리

산업	일차원적 분석 애플리케이션
서비스	콜센터 직원관리, 서비스-수익 사슬 관리
정부	사기 탐지, 사례 관리, 범죄 방지, 수익 최적화
온아린	웹 매트릭, 사이트 설계, 고객 추천
모든 사업	성과 관리

2절. 전략 인사이트 도출을 위한 필요 역량

데이터 사이언스의 영역



출처: https://yourforest17.tistory.com/50

데이터 사이언티스트의 역할

- **스토리텔링, 커뮤니케이션, 창의력, 열정, 직관력, 비판적 시각, 글쓰기 능력, 대회능력** 등을 갖춰야 한다

Hard Skill: - **빅데이터에 대한 이론적 지식**: 관련 기법에 대한 이해와 방법록 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 - **다분야 간 협력**: 커뮤니케이션

통찰력 있는 분석

- 직관과 전략, 경영 프레임워크 경험의 혼합을 통해 통찰력있는 분석을 수행할 수 있어야 함
- 좁은 시각보다 넓은 시각으로 나무보다 숲을 볼 수 있어야 함

3절. 빅데이터 길고 데이터 사이언스의 미래

- 디지털 환경의 진전으로 '빅'데이터 생성되고 있다
- 빅데이터 분석은 여러 분야에서 상당한 가치를 발휘하고 있다

데이터 사이언스의 한계

- 분석과정에서 가정 등 인간의 해석이 개입되는 단계를 반드시 거치게 된다
- 사람에 따라 다른 해석과 결론을 내릴 수 있다
- 아무리 정량적인 분석이라도 모든 분석은 가정에 근거함

데이터 사이언스와 인문학

- 인문학을 이용해 빅데이터와 데이터 사이언스가 데이터에 묻혀 있는 잠재력을 풀어내고, 새로운 기회를 찾고, 누구도 보지 못한 창조의 밑그림을 그릴 수 있는 힘을 발휘하게 될 것

최신 빅데이터 상식

DBMS와 SQL

DBMS(Data Base Management System)

- 데이터베이스를 공유하며 사용할 수 있는 환경을 제공하는 소프트웨어
- 데이터베이스를 구축하는 틀, 효율적인 데이터 검색, 저장 기능 등을 제공
- ex) 오라클, 인포믹스, 엑세스

데이터베이스 관리시스템 종류

- 관계형 DBMS
  - 컬럼과 로우로 이루어진 테이블 형태를 가짐
- 객체지향 DBMS
  - 정보를 '객체'형태로 표현
- 네트워크 DBMS
  - 레코드들이 노드로, 레코드들 사이의 관계가 간선으로 표현되는 그래프를 기반으로 하는 데이터베이스
- 계층형 DBMS
  - 트리 구조를 기반으로 하는 계층 데이터베이스

다양한 측면에서의 데이터베이스 특성

- 데이터베이스의 정보의 축적 및 전달 측면
  - 기계 가독성, 검색 가능성, 원격 조작성
- 정보 이용 측면
  - 이용자의 정보요구에 따라 다양한 정보를 신속 획득, 원하는 정보를 정확•경제적으로 찾아낼 수 있음
- 정보 관리 측면

    - 정보를 일정한 질서와 구조에 따라 정리•저장하고 검색•관리할 수 있도록 하여 방대한 양의 정보를 체계적으로 축적하고 새로운 내용 추가나 갱신이 용이
- 정보기술발전의 측면
    - 정보처리, 검색•관리 소프트웨어, 관련 하드웨어, 정보전송을 위한 네트워크 기술등의 발전은 견인
- 경제•산업적 측면
    - 인프라로서 특성을 가지고 있어 경제, 산업, 사회 활동의 효율성을 재고하고 국민의 편의를 증진하는 수단

데이터베이스의 설계 절차

1. 요구사항 분석
2. 개념적 설계
3. 논리적 설계
4. 물리적 설계
5. 구현

Relationship

- 관리하고자 하는 업무 영역 내의 특정한 두 개의 엔티티 사이에 존재하는 많은 관계 중 특별히 관리하고자 하는 직접적인 관계를 의미
- 1:1 관계, 1:M 관계, M:N관계

데이터 웨어하우스와 ETL

- 데이터 웨어하우스(Data Warehouse): 방대한 조직 내에서 분산 운영되는 DBMS를 효율적으로 통합하여 조정•관리하는 역할을 함
- 특징

특징	설명
주제지향성(Subject Oriented	업무 중심이 아닌 주제 중심
통합성(Integrated)	혼재한 DB로부터의 데이터 통합
시계열성(Time Variant)	시간에 따른 변경 정보를 나타냄
비휘발성(Non-Volatile)	데이터 변경 없이 리포팅을 위한 read only 사용

- ETL(Extract, Transform, Load): 데이터 웨어하우스 구축 시 데이터를 운영 시스템에서 추출하여 가공(변환, 정제)ㅎㄴ 수 데이터 웨어하우스에 적재하는 과정

NoSQL

- 데이터의 폭발적인 증가에 대응하기 위해 발달된 분삭 데이터베이스 기술
- 확장성, 가용성 높은 성능 제공
- 비관계형 데이터베이스 관리 시스템이지만 SQL 계열 쿼리언어를 사용
- Key-Value형태로 자료 저장

- ex) MongoDB, Hbase, Redis, Cassandra

SQL(Structured Query Language)

- **데이터베이스의 하부 언어**
- 단순한 질의 기능부터 완전한 데이터의 정의와 조작 기능을 갖춤
- **영어 문장과 비슷한 구문**으로 초보자들도 비교적 쉽게 사용가능

SQL의 분류

- DDL(Data Definition Language)
- DML(Data Manipulation Language)
- DCL(Data Control Language)

SQL 집계함수

- AVG
- COUNT
- SUM
- STDDEV
- MIN
- MAX

SQL 주요 구문

- WHERE
- ORDER BY
- GROUP BY
- HAVING

Data에 관련된 기술

개인정보 비식별 기술

〈 예시 〉 비식별 조치 방법

처리기법	예시	세부기술
가명처리 (Pseudonymization)	• 홍길동, 35세, 서울 거주, 한국대 재학 → 임꺽정, 30대, 서울 거주, 국제대 재학	① 휴리스틱 가명화 ② 암호화 ③ 교환 방법
총계처리 (Aggregation)	• 임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팥쥐 150cm → 물리학과 학생 키 합 : 660cm, 평균키 165cm	④ 총계처리 ⑤ 부분총계 ⑥ 라운딩 ⑦ 재배열
데이터 삭제 (Data Reduction)	• 주민등록번호 901206-1234567 → 90년대 생, 남자 • 개인과 관련된 날짜정보(합격일 등)는 연단위로 처리	⑧ 식별자 삭제 ⑨ 식별자 부분삭제 ⑩ 레코드 삭제 ⑪ 식별요소 전부삭제
데이터 범주화 (Data Suppression)	• 홍길동, 35세 → 홍씨, 30~40세	⑫ 감추기 ⑬ 랜덤 라운딩 ⑭ 범위 방법 ⑮ 제어 라운딩
데이터 마스킹 (Data Masking)	• 홍길동, 35세, 서울 거주, 한국대 재학 → 홍○○, 35세, 서울 거주, ○○대학 재학	⑯ 임의 잡음 추가 ⑰ 공백과 대체

무결성과 레이크

데이터 무결성(Data Integrity)

- 데이터에 대한 정확한 일관성, 유효성, 신뢰성, 정확성을 보장
- 데이터 변경/수정 시 여러 가지 제한을 둠
- 유형:
  - 무결성(Entity Integrity)
  - 참조 무결성(Referential Integrity)
  - 범위 무결성(Domain Integrity)

데이터 레이크(Data Lake)

- 수 많은 정보 속에서 의미있는 내용을 찾기 위해 방식에 상관없이 데이터를 저장하는 시스템
- 대용량의 형 및 비정형 데이터를 저장 할 뿐만 아니라 접근도 쉽게 할 수 있는 대규모 저장소
- ex) Apache Hadoop, Teredata Integrated Big Data Platform 1700

빅데이터 분석 기술

하둡(Hadoop)

- 여러 개의 컴퓨터를 하나인 것처럼 묶어 대용량 데이터를 처리하는 기술
- 분산파일 시스템(HDFS)을 통해 수 천대의 장비에 대용량 파일을 저장할 수 있는 기능 제공
- 맵리듀스(Map Reduce)로 HDFS에 저장된 대용량의 데이터들을 대상으로 SQL을 이용해 사용자의 질의를 실시간으로 처리

Apache Spark

- 실시간 분산형 컴퓨팅 플랫폼
- 스칼라로 작성이 되어 있지만 스칼라, 자바, R, 파이썬, API 지원
- In-Memory 방식으로 처리하기 때문에 하둡에 비해 처리속도가 빠름

Smart Factory

- 공장 내 설비와 기계에 사물인터넷(IoT)가 설치됨
- 공정 데이터가 실시간으로 수집되고 데이터에 기반한 의사결정이 이뤄짐으로써 생산성을 극대화할 수 있다

Machine Learning & Deep Learning

머신러닝

- 인공지능의 연구 분야 중 하나
- 인간의 학습 능력과 같은 기능을 컴퓨터에서 실현하고자 하는 기술

딥러닝

- 컴퓨터가 많은 데이터를 이용해 사람처럼 스스로 학습할 수 있게 하기 위하여 인공 신경망(Artificial Neural Network, ANN) 등의 기술을 기반하여 구축한 기계 학습 기술 중 하나
- ex) DNN, CNN, RNN, LSTM, Autoencoder, RBM 등
- 소프트웨어 라이브러리: ex) Tensorflow, Caffe, Torch, Theano, Gensim 등

기타

데이터양의 단위

접두어 (SI)	이름	계산법	접두어 (IEC)	이름	계산법
	1바이트(byte)	1000바이트 =1kB			1024바이트 =1KiB
킬로 (103)	1킬로바이트 (kilobyte)/kB	1000KB=1MB	키비 (210)	1키비바이트 (kibibyte)/KiB	1024KiB=1MiB
메가 (106)	1메가바이트 (megabyte)/MB	1000MB=1GB	메비 (220)	1메비바이트 (mebibyte)/MiB	1024MiB=1GiB
기가 (109)	1기가바이트 (gigabyte)/GB	1000GB=1TB	기비 (230)	1기비바이트 (gibibyte)/GiB	1024GiB=1TiB
테라 (1012)	1테라바이트 (terabyte)/TB	1000TB=1PB	테비 (240)	1테비바이트 (tebibyte)/TiB	1024TiB=1PiB
페타 (1015)	1페타바이트 (petabyte)/PB	1000PB=1EB	페비 (250)	1페비바이트 (pebibyte)/PiB	1024PiB=1EiB
엑사 (1018)	1엑사바이트 (exabyte)/EB	1000EB=1ZB	엑스비 (260)	1엑스비바이트 (exbibyte)/EiB	1024EiB=1ZiB

제타 (1021)	1제타바이트 (zettabyte)/ZB	1000ZB=1YB	제비 (270)	1제비바이트 (zebibyte)/ZiB	1024ZiB=1YiB
요타 (1024)	1요타바이트 (yottabyte)/YB	—	요비 (280)	1요비바이트 (yobibyte)/YiB	—

단위	데이터 양
바이트 (byte, B)	1 byte = 2^0 B
킬로바이트 (Kilobyte, KB)	1024 B = 2^{10} B
메가바이트 (Megabyte, MB)	1024 KB = 2^{20} B
기가바이트 (Gigabyte, GB)	1024 MB = 2^{30} B
테라바이트 (Terabyte, TB)	1024 GB = 2^{40} B
페타바이트 (Petabyte, PB)	1024 TB = 2^{50} B
엑사바이트 (Exabyte, EB)	1024 PB = 2^{60} B
제타바이트 (Zettabyte, ZB)	1024 EB = 2^{70} B
요타바이트 (Yottabyte, YB)	1024 ZB = 2^{80} B

B2B와 B2C

B2B

- 기업과 기업 사이의 거래를 기반으로 한 비즈니스 모델
- 기업이 필요로 하는 장비, 재료나 공사입찰 등

B2C

- 기업과 고객 사이의 거래를 기반으로 한 비즈니스 모델
- 이동 통신사, 여행 회사, 신용카드회사, 옥션, 지마켓 등

블록체인(Block Chain)

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

데이터 유형

유형	내용	예시
정형 데이터	- 형태가 있음 - 연산 가능	- 관계형 데이터베이스 - 스프레드시트

유형	내용	예시
	- 주로 관계형 데이터베이스에 저장됨 - 데이터 수집 난이도가 낮고 형식이 정해져 있어 처리가 쉬운 편	- CSV 등
반정형 데이터	- 형태가 있음 - 연산 불가능 - 주로 파일로 저장됨 - 데이터 수집 난이도가 중간 - 보통 API 형태로 제공되기 때문에 데이터처리 기술이 요구됨	- XML - HTML - JSON - 로그형태 등
비정형 데이터	- 형태가 없음 - 연산이 불가능 - 주로 NoSQL에 저장됨 - 데이터 수집 난이도가 높음 - 텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어려움	- 소셜데이터(트위터, 페이스북) - 영상 - 이미지 - 음성 - 텍스트(word, PDF 등)등