ADsP

제 1과목 : 데이터 이해

01장. 데이터의 이해

데이터 : 라틴어로 Dare(주다)의 과거 분사형--〉 주어진 것

- 과거의 관념적이고 추상적인 개념 --〉 기술적이고 사실적인 의미
- 데이터는 추론과 추정의 근거를 이루는 사실
- 데이터는 단순한 객체로서의 가치뿐만 아니라 다른 객체와의 상호관계 속에서 가치를 갖는 것으로 설명

데이터의 특성 :

존재적 특성	객관적 사실 (Fact, Raw Material)
당위적 특성	추론·예측·전망·추정을 위한 근거 (Basis)

데이터의 유형 :

구분	형태	예	특징
정성적 데이터 (Qualitative Data)	언어, 문자	회사 매출이 증가함 등	저장·검색·분석에 많은 비용이 소모 됨
정량적 데이터 (Quantitative Data)	수치, 도형, 기호 등	나이,몸무게,주가 등	**정형화**된 데이터로 비용소모가 적음

- 정성적 데이터 : 비정형 데이터, 주관적 내용, 통계분석이 어려움
- 정량적 데이터 : 정형 데이터 , 객관적 내용, 통계분석이 용이함

지식경영의 핵심 이슈

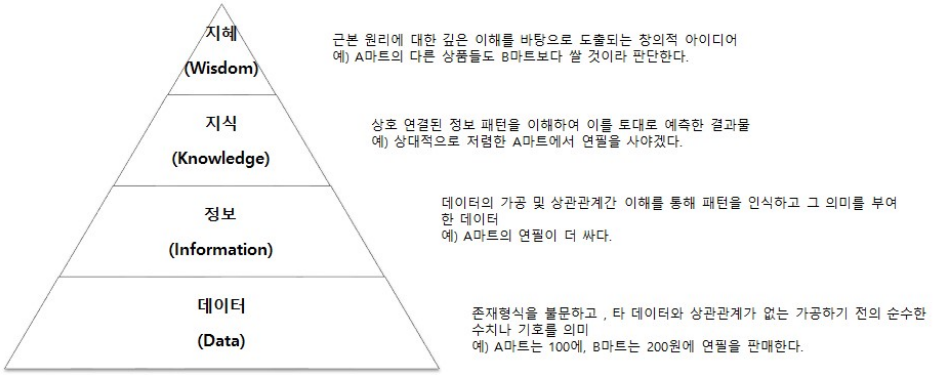
구분	의미	예	특징	상호작용
암묵지	학습과 경험을 통해 개인에게 체화되어 있지만 겉으로 드러나지 않는 지식	김장김치 담그기, 자전거 타기	사회적으로 중요하지만 다른 사람에게 공유되기 어려움	공통화, 내면화
형식지	문서나 매뉴얼처럼 형상화된 지식	교과서, 비디오, DB	전달과 공유가 용이	표출화, 연결화

※암묵지와 형식지의 상호작용 관계 : 공통화 --〉 표출화 --〉 연결화 --〉 내면화

데이터와 정보의 관계

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)	데이터를 통해 도출된 다양한 정보를 구조화하여 유의미한 정보를 분류하고 개인적인 경험을 결합시켜 고유의 지식으로 내재화된 것
지혜(Wisdom)	지식의 축적과 아이디어가 결합된 창의적인 산물



DB의 정의 : 정형 데이터 관리 ---〉 빅데이터 출현으로 비정형데이터 포함

DB의 일반적인 특징 : 통합 / 저장 / 공용 / 운영

- 이 책에서는 운영 대신 변화하는 데이터라고 나옴

1980년대 기업내부 데이터베이스 : OLTP / OLAP

- OLTP(On-Line Transaction Processing) : 호스트 컴퓨터와 온라인으로 접속된 여러 단말 간의 처리 형태. 호스트 컴퓨터가 데이터베이스를 액세서하고, 바로 처리 결과를 반환 --〉 데이터 갱신 위주
- OLAP(On-Line Analytical Processing) : 다양한 비즈니스 관점에서 쉽고 빠르게 다차원적인 데이터에 접근하여 의사 결정에 활용할 수 있는 정보를 얻게 해주는 기술 ---〉 데이터 조회 위주

2000년대 기업내부데이터 베이스 : CRM / SCM

- CRM(Customer Relationship Management) : 고객관계관리, 고객과 관련된 자료를 종합하여 고객 중심으로 극대화
- SCM(Supply Chain Management) : 공급망 관리, 기업에서 원재료 생산·유통 등 모든 공급망 단계를 최적화에 수요자가 원하는 제품을원하는 시간과 장소에 제공. 고객과 거래관계가 있는 기업들 간 IT를 이용한 신시간 정보공유로 수요자 요구에 기민히 대응

각 분야별 내부 데이터베이스

- 제조부문 : ERP -〉CRM으로 발전, RTE를 통한 협업적 IT와 비중 확대
- 금융부문 : EAI, ERP, e-CRM을 통한 정보 공유 및 고객을 위한 정보 활용, DW 도입과 최적화를 위한 BI 기반 시스템 구축. EDW 활성화
- 유통 부문 : CRM과 SCM 구축이 활발히 진행. 상거래를 위한 인프라와 KMS를 위한 백업시스템 구축. RFID 등장

위의 용어들

- ERP(Enterprise Resource Planing) : 인사·재무·생산등 기업에 독립적으로 운영되던 시스템의 경영자원을 하나의 통합 시스템으로 재구축해서 생산성을 극대화하려는 경영혁신기법
- BI(Business Intelligence): 기업이 보유한 수많은 데이터를 정리하고 분석해 기업의 의사결정에 활용하는 프로세스
- CRM
- RTE (Real-Time Enterprise) : 회사의 주요 경영정보를 통합관리하는 실시간 기업의 경영시스템
- EDW(Enterprise Data Warehouse) : 기존 DW를 전사적으로 확장한 모델로 다양한 분석 애플리케이션들을 위한 원천이 됨. 즉 단순한 대형 시스템이 아닌 기업 리소스의 유기적 통합, 다원화된 관리 체계 정비, 데이터 중복 방지들ㅇ을 위해 시스템을 재설계하는 것
- KMS(Knowledge Management System) : 지식관리 시스템

사회기반 구조로서의 데이터베이스

: 사회 각 부문의 정보화가 본격화 되면서 정부를 중심으로 무역, 통관, 물류, 조세, 국세, 조달등 사회간접자본(SOC) 차원에서 EDI를 활용해 부가가치통신망(VAN)을 통해 정보망 형성 – 이제 일반인들도 교통 지리등 손쉽게 정보 획득

위의 용어

- EDI(Electronic Data Interchange) : 주문서, 납품서, 청수더 등 무역에 필요한 각종 서류를 표준화된 양식을 통해 전자적으로 신호로 바꿔 컴퓨터로 거래처에 전송
- VAN(Value Added Network) : 한국전기통신공사로부터 통신회선을 차용하여 독자적인 네트워크를 형성하는 것
- CALS(Commerce At Light Speed) : 전자상거래 구축을 위해 제품의 생명주기 전반에 관련된 데이터들을 통합하고 공유·교환할 수 있도록 한 경영통합정보시스템

02장. 데이터의 가치와 미래

빅데이터의 정의 (관점에 따라)

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 빅데이터에 대한 정의

3V + 4V

3V		
양(Volume) "데이터의 규모“	다양성(Variety) "데이터의 유형과 소스"	속도(Velocity) "데이터의 수집과 처리"
센싱데이터, 비정형 데이터	정형,비정형데이터 (영상,사진)	원하는 데이터의 추출 및 분석 속도

+

4V	
가치(Value)	진실성(Veracity)
정확성(Validity)	휘발성(Volatility)

빅데이터 정의의 범주 및 효과
1단계 : 데이터 변화 (3V)
2단계 : 기술 변화 – 데이터 처리, 저장, 분석 및 클라우드 활용
3단계 : 인재, 조직 변화 : data Scientist와 같은 새로운 인재 필요
※이로써 기존방식으로 얻을 수 없던 통찰 및 가치를 창출함

출현 배경
: 빅데이터 현상은 없었던 것이 새로 등장한 것이 아니라 기존의 데이터, 처리 방식, 다루는 사람과 조직 차원에서 일어나는 '변화'를 말한다

산업계 : 고객 데이터 축적 --〉 데이터의 숨은 가치 발굴
학계 : 거대 데이터 활용, 과학 확산 -〉 아키텍처 및 통계 도구 발전
기술발전 : 관련기술의 발달 -〉 디지털화, 모바일혁명, 클라우드 컴퓨팅 등

빅데이터에 거는 기대를 표현한 비유
1. 산업혁명의 석탄, 철 : 서브 분야의 생산성을 획가적으로 상승시킬 것으로 기대
2. 21세기의 원유 : 새로운 범주의 산업을 만들것으로 기대
3. 렌즈 : 현미경이 생물학 발전에 미친 영향 만큼 산업 발전에 영향을 미칠 것으로 기대
4. 플랫폼 : 다양한 서드파티 비즈니스에 활용되면서 플랫폼 역할을 수행할 것으로 기대

과거에서 현재로의 변화
1. 사전 처리 -〉 사후 처리
2. 표본 조사 -〉 전수 조사
3. 질 -〉 양
4. 인과관계 -〉 상관 관계

빅데이터 가치 산정이 어려운 이유
1. 데이터 활용 방식 : 재사용이나 재조합 등이 일반화 되어, 특정 데이터를 언제·어디서·누가 활용할지 알 수 없게되어 가치를 산정하는것도 어려워짐
2. 새로운 가치 창출 : 빅데이터 시대에 데이터가 '기존에 없던 가치'를 창출하여 측정하기 어려움
3. 분석 기술 발전 : 현재는 가치가 없더라도, 추후 새로운 분석 기법이 등장하면 거대한 가치를 지닌 데이터가 될 수 있음

빅데이터가 미치는 영향
1. 기업 – 혁신, 경쟁력 제고, 생산성 향상
2. 정부 – 환경 탐색, 상황분석, 미래대응
3. 개인 – 목적에 따른 활용
※이를 통해 생활 전반의 스마트화를 이룰 수 있다

빅데이터를 활용한 기본 테크닉 ★★

테크닉	내용	예시
연관 규칙 학습	변인들 간에 주목할만한 상관관계가 있는지 찾아내는 방법	커피를 구매하는 사람이 탄산음료를 더 많이 사는가?
유형분석	문서를 분류하거나, 조직을 그룹으로 나누는, 특성에 따라 분류할때	이 사용자는 어떤 특성을 가진 집단에 속하는가?
유전자 알고리즘	최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 점진적으로 진화	최대의 시청률을 얻으려면 어떤 프로그램이 어느 시간대에 적절한가?
기계학습	훈련 데이터로부터 학습한 알려진 특성을 활용해 예측하는 방법	시청 기록을 바탕으로 사용자가 어떤 것을 보고싶나?
회귀분석	독립변수를 조작함에 따라, 종속변수가 어떻게 변하는지를 보고 두 변인의 관계를 파악할 때 사용	구매자의 나이가 구매차량 타입에 어떤 영향을 미치나?
감정분석	특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석	새로운 환불 정책에 대한 고객 평가는 어떤가?
사회네트워크 분석	특정인과 다른 사람이 몇 촌 정도의 관계인가를 파악할때 사용, 영향력 있는 사람을 찾아 낼 때 사용	고객들 간 관계망은 어떻게 구성되어 있나?

위기 요인과 통제 방안★★★

위기 요인

1. 사생활 침해 --〉 익명화 기술 발전 필요
2. 책임 원칙 훼손 : 분석 대상이 되는 사람들이 예측 알고리즘의 희생양이 될 수 있음. 민주주의 국가에서는 잠재적 위협이 아닌 명확한 결과에 대해 책임을 물어 이에 대한 원리를 훼손할 가능성이 있음 ex) 애니 싸이코패스와 비슷
3. 데이터 오용 : 빅데이터는 일어난 일에 대한 데이터에 의존으로 미래를 예측하더라도 100%의 정확도를 보장할 수 없음. 또한 잘못된 데이터 사용하는 것도 빅데이터의 폐해가 될 수 있음

통제 방안

1. 동의에서 책임으로 : 개인정보 제공 동의를 책임으로 바꿈으로써 사용주체의 적극적인 보호장치를 강구하게 될 수 있음
2. 결과 기반 책임 원칙 고수 : 책임원칙 훼손 위기요인에 대한 통제 방안으로, 기존의 원칙을 좀 더 강화·보완할 필요가 있고, 예측 자료에 의한 불이익을 당할 가능성을 최소화하는 장치를 마련하는 것이 필요 -〉 잘못된 예측 알고리즘을 통한 판단을 근거로 누군가에게 불이익을 줄 수 없고, 이에 따라 피해 최소화 장치 마련
3. 알고리즘 접근 허용 : 데이터 오용에 대한 대응책으로 '알고리즘에 대한 접근권'을 제공하여 예측 알고리즘의 부당함을 반증할 수 있는 방법을 명시해 공개할 것을 주문 -〉 불이익을 당한 사람들은 대변할 전문가가 필요하게 됨

빅데이터 활용의 삼요소

1. 데이터 -〉 모든 것의 데이터화 : 창의적인 분석 가능
2. 기술 -〉 진화하는 알고리즘, 인공지능
3. 인력 -〉 데이터 사이언티스트, 알고리즈미스트 : 다각적 분석을 통한 인사이트 도출이 중요해지고 있음

03장. 가치 창조를 위한 데이터 사이언스와 전략 인사이트

빅데이터 열풍과 회의론이 흘러 나오고 있음.

빅데이터 회의론의 원인 및 진단

1. 투자효과를 거두지 못했던 학습효과 --〉 과거의 고객관계관리(CRM)
   - 도입만 하면 뭐든 해결할 것처럼 마케팅함
   - 막상 거액을 투자해 하드웨어와 솔루션을 도입해도 어떻게 활용하는지 모름
2. 빅데이터 성공사례가 기존 분석 프로젝트를 포함해 놓은 것이 많음
   - (우수고객,이탈예측,구매패턴 분석 등)은 빅데이터가 필요없음
   - 국내 빅데이터 업체들이 CRM분석 성과를 빅데이터 성과처럼 포장

결론 : 빅데이터 분석도 기존의 분석과 마찬가지로, 데이터에서 가치, 즉 통찰을 끌어내 성과를 창출하는 것이 관건이며, 단순히 빅데이터에 포커스를 두지말고 분석을 통해 가치를 만드는 것에 집중해야 함
빅데이터의 크기가 아니라 거기에서 무엇을 얻을 지가 중요.

전략전 통찰이 없는 분석의 함정

- 단순히 분석을 많이 사용하는 것이 곧바로 경쟁우위를 가져다 주지 않는다
- 분석이 경쟁의 본질을 제대로 바라보지 못하면 쓸데없는 분석 결과를 잔뜩 쏟아낸다. 이를 위해 전략적인 통찰력을 가지고 핵심적인 비즈니스 이슈에 집중하여 데이터를 분석하고 차별적인 전략으로 기업을 운영해야함

※예시

- 아메리칸 항공은 수익관리, 가격 최적화에 분석함 --〉 비행경로 및 승무원들 일정 최적화 --〉 초반에는 수익을 많이 냈지만 타 경쟁사들이 비슷한 분석 역량과 수익관리 능력으로 수익 절감됨
- 사우스웨스트 항공은 단순 최적화 모델을 통해 가격 책정 및 운영 --〉 한가지 기종의 비행기로 단순화 -〉 단순 최적화 모델로 좌석 가격 책정 및 운영 결과 경쟁우위 상승 --〉 36년 흑자

산업별 분석 애플리케이션★ (시험에 에너지 나왔음)

산업	일차원적 분석 애플리케이션
금융 서비스	신용점수 산정, 사기 탐지, 가격 책정, 프로그램트레이딩, 클레임분석, 고객 수익성분석
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
서비스	콜센터 직원관리, 서비스-수익 사슬 관리
정부	사기 탐지, 사례 관리, 범죄 방지, 수익 최적화
온라인	웹 매트릭스, 사이트 설계, 고객 추천
모든사업	성과관리

일차적인 분석의 문제점 :

- 일차적인 분석을 통해서도 해당 부서나 업무 영역에서 상당한 효과를 얻을 수 있으나 환경변화와 같은 큰 변화에 제대로 대응하기 어려움.
- 고객 환경 변화를 파악하고 새로운 기회를 포착하기 어려움

전략도출 가치기반 분석

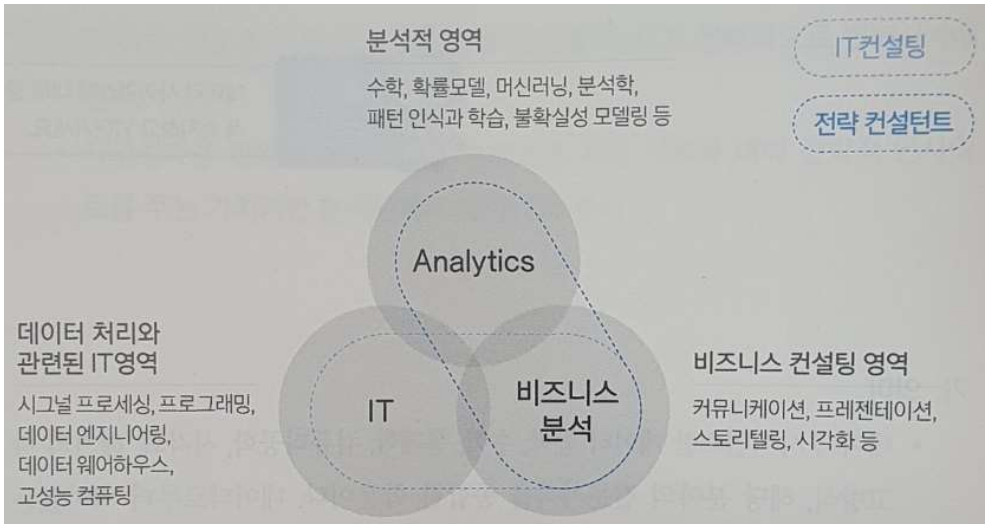
- 분석의 활용 범위를 더 넓고 전략적으로 변화시켜야 함
- 사업성과를 견인하는 요소들과 차별화를 꾀할 기회에 대해 전략적 인사이트를 주는 가치기반 분석단계로 나아가야 한다.

데이터 사이언스의 의미와 역할

: 많은 학문의 전문지식을 종합한 학문.

: 데이터 사이언티스트는 비즈니스의 성과를 좌우하는 핵심이슈에 답을 하고, 사업의 성과를 견인해 나갈 수 있어야함. 소통력 중요

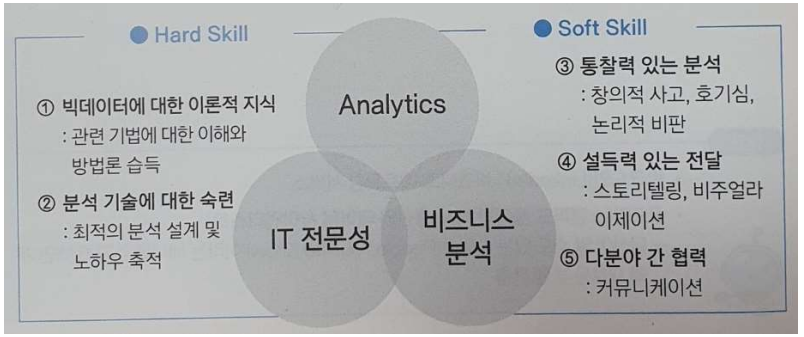
데이터사이언스의 영역



분석 영역 / IT 영역 / 비즈니스 컨설팅 영역

- 데이터 사이언티스트는 스토리텔링, 커뮤니케이션, 창의력, 열정, 직관력, 비판적 시각, 글쓰기 능력, 대화 능력이 필요

요구 역량 : 분석 뿐만 아닌 외적인 인문학적 요소 필요



인문학이 필요한 이유

: 기존 사고의 틀을 벗어나 문제를 바라보고 해결하는 능력, 비즈니스의 핵심가치를 이해하고 고객과 지원의 내면적 요구를 이해하는 능력 등 인문학에서 배울 수 있는 역량이 더 요구됨

외부 환경적 측면에서 본 인문학 열풍 이유

1. 컨버전스 -> 디버전스 : 단순세계화에서 복잡한 세계화로의 변화
2. 생산 -> 서비스 : 비즈니스 중심이 제품생산에서 서비스로 이동
3. 생산 -> 시장창조 : 공급자 중심의 기술경쟁에서 무형자산의 경쟁으로 변화

시대에 따른 가치 패러다임의 변화

- 과거 : 아날로그를 어떻게 디지털로 효과적으로 하는지
- 현재 : 디지털화된 정보와 대상들의 연결을 어떻게 효율적으로 할지
- 미래 : 복잡한 연결을 얼마나 효과적이고 신뢰성높게 관리할 것인지

데이터 사이언스의 한계

- 분석 과정에 반드시 인간의 해석이 개입되는 단계를 반드시 거침
- 인간에 따른 다른 해석과 결론을 내릴 수 있음
- 아무리 정량적인 분석이라도 모든 분석은 가정에 근거한다는 사실

※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※※

최신 빅데이터 상식

DW의 특징 : 주제지향성, 통합성, 시계열성, 비휘발성

개인정보 비식별 기술 :

데이터 마스킹, 가명처리, 총계처리, 데이터값 삭제, 데이터 범주화

Data Lake(데이터 레이크) : 수많은 정보 속에서 의미있는 내용을 찾기 위해 방식에 상관없이 데이터를 저장하는 시스템

빅데이터 분석 기술들

- 하둡 : 여러 개의 컴퓨터를 하나처럼. 맵리듀스로 HDFS(분산파일시스템)에 접근
- Apache Spark : 실시간 분산형 컴퓨팅 플랫폼. IN-memory 방식으로 속도 빠름
- Smart factory : 공장 내 설비 및 기계에 IoT 설치로 생산성 향상
- Machine Learning & Deep Learning : 인공지능 연구 분야

데이터의 유형

- 정형 데이터 : 흔한 관계 테이블
- 반정형 데이터 : XML, JSON 등 스키마나 메타데이터가 있는 것들. 반드시 파싱이 요구
- 비정형 데이터