ADsP · ADP

(**ADsP** : Advanced Data Analytics Semi-Professional **ADP** : Advanced Data Analytics)

1.과목(데이터의 이해)총정리

1) 데이터의 정의

데이터는 객관적 사실이라는 존재적 특성 동시에

추론, 예측, 전망, 추정을 위한 근거로

기능하는 당위적 특성(~해야만 하는) 의미함.

이는 다른 객체와의 상호관계 속에서일 때 데이터는 가치를 가짐

결국, 데이터는 개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실을 의미함

2) 데이터의 유형

① **정성적데이터**(Qualitative Data) : 언어, 문자 등
형태와 형식이 정해져 있지 않음 (비정형 데이터 형태로 저장, 분석에 시간과 비용이 필요)

· 숫자나 금액으로 환산할 수 없는 것 설문 조사 주관식 응답, 트위터, 페이스북 → 정성적 데이터

② **정량적 데이터(**Quantitative Date) : 수치, 기호, 도형으로 표시
Numerical : 데이터양이 증가하더라도 저장, 분석용이

· 숫자나 금액으로 환산 가능한 것 온도, 풍속, 강수량 → *정량적 데이터*

정량적 vs 정량적 데이터 구분

3) 암묵지 vs 형식지

데이터는 지식경영의 핵심 이슈인 암묵지와 형식지의 상호 작용에 있어 중요한 역할을 함.

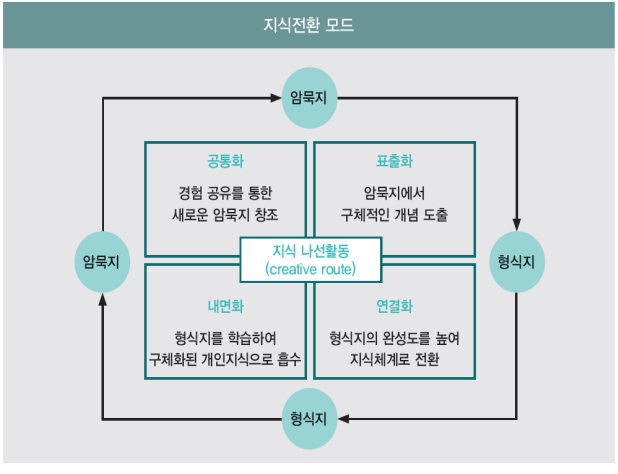
① **암묵지** : 학습과 체험을 통해 개인에게 습득되어 있지만, 겉으로 드러나지 않은 지식, 시행착오와 오랜 경험을 통해 개인에게 습득된 무형의 지식 → 개인에게 체화되어 있으므로 외부에 표출되어 공유가 어려움.

② **형식지** : 교과서, 매뉴얼, 비디오, DB와 같이 형상화된 지식을 의미, 유형의 대상이 있어서 지식의 전달과 공유가 쉬움.

형식지와 암묵지 개념과 사례 구분

4) 암묵지와 형식지의 상호 작용

· 공통화(Socialization) : 암묵지 지식 노하우를 다른 사람에게 알려줌
· 표출화(Externalization) : 암묵지 지식 노하우를 책, 교본 형식으로 전환함.
· 연결화(Combination): 책, 교본에 자신이 알고 있는 새로운 지식을 추가함.
· 내면화(Internalization) : 만들어진 책, 교본을 보고 다른 직원의 암묵적 지식을 습득함.



→ 일련의 지식 순환 과정 중 암묵지가 형식지로 전환 되는 표출화 단계는
개인에게 내재한 경험이 객관적인 데이터로 문서나 매체에 저장·가공·분석하는 과정

5) 데이터와 정보의 관계

DIKW 피라미드(Data → Information → Knowledge → Wisdom) : 데이터, 정보, 지식을 통해 최종적으로 지혜를 얻어내는 과정

① Data: 존재 형식을 불문하고, 타 데이터와의 상관관계가 없는 가공하기 전의 순수한 수치나 기호
연필 가격 : A 마트 100원, B 마트는 200원

② Information: 데이터의 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 의미 부여
A 마트의 연필 가격이 더 싸다.

③ Knowledge: 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물
상대적으로 저렴한 A 마트에서 연필을 사야겠다.

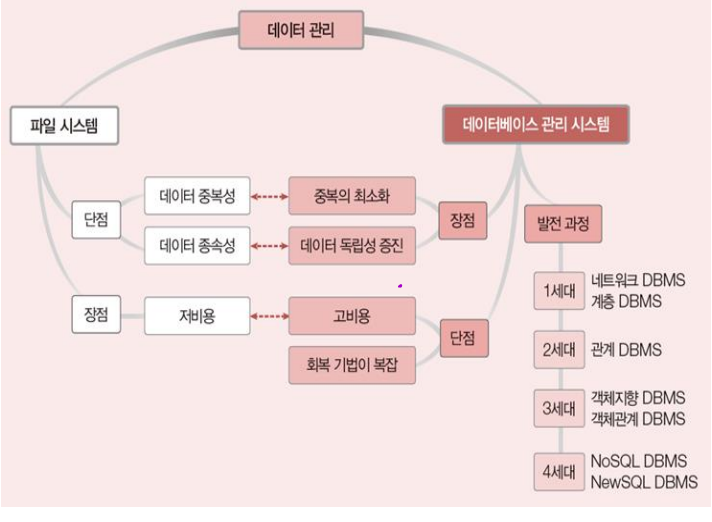
④ Wisdom: 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 아이디어
A 마트의 다른 상품들도 B 마트보다 쌀 것이라고 판단

DIKW 개념 구분과 사례를 이해하자

6) 데이터베이스 정의

동시에 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에
대응해서 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합
소프트웨어로는 **데이터베이스관리시스템** (DBMS:Database Management System)을 의미함.

7) 데이터베이스관리시스템(DBMS) 등장 배경



① 1세대: 네트워크 DBMS, 계층 DBMS
- 복잡하고 변경이 어렵다.

② 2세대: 관계 DBMS(데이터베이스를 데이블 형
구성) 오라클(유료), 액세스, MySQL(무료)

③ 3세대: 객체 지향(Objected) DBMS
- 멀티미디어 데이터의 확산으로 관계형 데이터 모델 표현하기 어려움
- 같은 행위를 갖는 객체는 한 클래스에 속하며, 클래스 연산을 나타내기 위해 메소드 함수 정의함.

3세대: 객체 관계형 모델(ORDBMS)
- 기존의 관계형 모델에 객체 지향형 모델의 장점 선별하여 관계형 모델에 통합한 새로운 개념이 데이터 모델

④ 4세대: NoSQL DBMS
- 소셜 네트워크 서비스 증가 (비정형 데이터)
- 데이터 구조를 미리 정해두지 않기 때문에 비정형 데이터를 저장하고 처리함.

8) 데이터베이스 기본개념

· 스키마(schema) : 구조를 만드는 것
  - 구체적으로 데이터베이스의 구조와 제약조건을 기술

상품, 아이디	카테고리	상품명	상품가격	상품재고
문자형	숫자형	문자형	숫자형	숫자형

·인스턴스(Instance) : 특정 시점의 데이터베이스 내용

·트랜잭션(Transaction) : 데이터베이스의상태를 변화시키기 위해 수행하는 작업의 단위

9) 데이터베이스 설계 순서

① 요구조건 분석

② 개념적 설계 (E-R모델)

③ 논리적 설계 (테이블 설계)

④ 물리적 설계 (데이터 구조화)

10) 데이터베이스의 특징

① 통합된 데이터(Integrated Data): 데이터베이스에서 같은 내용의 데이터가 중복되어 있지 않다는 것 을 의미
② 저장된 데이터(Stored Data): 자기 디스크나 자기 테이프 등과 같이 컴퓨터가 접근할 수 있는 저장매체에 저장되는 것을 의미
③ 공용 데이터(Shared Data): 여러 사용자가 서로 다른 목적으로 데이터베이스의 데이터를 공동 이용
④ 변화되는 데이터(Changed data): 새로운 데이터의 추가, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도
항상 현재의 정확한 데이터를 유지해야 한다는 의미)

4가지 특징 암기하세요!

11) 데이터베이스 특성

① 정보의 축적 및 전달 측면 : 검색 가능성, 원격 조작성을 갖는다
② 정보이용 측면 : 다양한 정보 획득, 원하는 정보 검색(검색)
③ 정보 관리 측면 : 일정한 질서와 구조에 따라 정리, 저장, 검색, 관리할 수 있도록 체계적으로 축적 하고 새로운 내용 추가 용이
④ 정보기술 발전 측면 : 검색관리 소프트웨어, 하드웨어, 네트워크 기술 발전기여
⑤ 경제. 산업적 측면 : 인프라 특성, 효율성 제고, 국민의 편의 증대

4가지 특징 암기하세요!

12) OLTP vs OLAP

① **OLTP**(On-Line Transaction Processing) :
네트워크상의 여러 이용자가 실시간으로 데이터베이스의 데이터를 갱신하거나 조회하는 등의 단위 작업을 처리하는 방식
→ 은행에서 수많은 입출금 등이 일어날 때

② **OLAP**(On-Line Analytic Processing) :
정보 위주의 처리 분석을 의미한다.
의사결정에 활용할 수 있는 정보를 얻을 수 있게 해주는 기술
→ 판매 추이, 구매성향파악, 재무회계 분석 등을 프로세싱하는 것.

13) 데이터웨어하우스 vs 데이터 마트

데이터 마트(Data Mart, DM)는 데이터웨어하우스(Data Warehouse, DW) 환경에서
정의된 접근 계층으로, 데이터웨어하우스에서 데이터를 꺼내 사용자에게 제공하는 역할을 함.
데이터 마트는 데이터웨어 하우스의 부분이며, 대개 특정한 조직, 혹은 팀에서 사용하는 것을 목적으로 함.

14) **CRM**(Consumer Relationship Management) **SCM**(Supply Chain Management)

2000년대 들어서면서 기업 DB 구축의 화두 CRM/SCM

① CRM:선별된 고객으로부터 수익을 창출하고 장기적인 고객 관계를
가능케 함으로써 보다 높은 이익을 창출할 수 있는 솔루션

② SCM:제조, 물류, 유통업체 등 유통 공급망에 참여하는 모든 업체가
협력을 바탕으로 정보기술을 활용, 재고를 최적화하기 위한 솔루션
→ SCM과 CRM은 연동되기 때문에 상호 밀접한 관계

제조/금융/유통 기업 내부 데이터베이스

2000년대 중반 이후에는 중소기업에 대한 DB 구축 투자 증가를 가능하게 됨

실시간 기업(RTE:Real-Time Enterprise)

가트너는 RTE를 '최신 정보를 사용해 자사의 핵심 비즈니스 프로세스들의 관리와 실행 과정에서
생기는 지연 사태를 지속해서 제거함으로써 경쟁하는 기업' 으로 정의
결국, 이를 통해 대기업-중소기업 간의 협업적 IT화 비중이 점차 확대

15) 제조부문

· **ERP** : 제조업을 포함한 다양한 비즈니스 분야에서 생산, 구매, 재고, 주문, 공급자와의 거래, 고객서비스 제공 등 주요 프로세스 관리를 돕는 여러 모듈로 구성된 통합 솔루션

· **BI** (Business Intelligence) : 데이터 기반 의사결정을 지원하기 위한 리포트 중심의 도구

[BI와 BA의 차이점]
BI는 과거나 현재 상태를 설명하기 때문에 '기술적인 분석(Descriptive Analytics)'이라고도 부른다.
BA(Business Analytics)는 소프트웨어로 데이터를 분석해 미래를 예측하거나(예측 분석), 특정 접근법을 적용했을 때 발생할 수 있는 일을 내다보는 (처방적 분석) 기술의 도움을 받는 과정이다.
그래서 BA는 '고급 분석(advanced analytics)'이라고도 불린다.
→ 의사결정을 위한 통계적이고 수학적인 분석에 초점

16) 금융부문

① EAI는 Enterprise Architecture Integration의 약자로 기업 애플리케이션 통합을 의미.
기업 내의 ERP(전사적자원관리), CRM(고객 관계관리), SCM(공급망계획) 시스템이나 인트라넷 등의
시스템 간에 상호 연동이 가능하도록 통합하는 솔루션

② EDW (Enterprise Data Warehouse)는
기존 DW(Data Warehouse)를 전사적으로 확장한 모델인 동시에 BPR과 CRM, BSC 같은
다양한 분석 애플리케이션들을 위한 원천이 됨. 따라서 EDW를 구축하는 것은 단순히 정보를 빠르게 전달하는
대형 시스템을 도입한다는 의미가 아니라 기업 리소스의 유기적 통합, 다원화된 관리체계 정비,
데이터의 중복 방지 등을 위해 시스템을 재설계하는 것.

③ 블록체인 (Blockchain):
데이터 분산 처리 기술. 네 트워크 참여하는 모든 사용자가 모든 거래내용 등의 데이터를 분산,
저장하는 기술을 말함. 블록들을 체인 형태로 묶는 형태이기 때문에 블록체인이라는 명칭이 생겨남.
기존 거래 방식에서 데이터를 위. 변조하기 위해에서는 은행의 중앙 서버를 공격하면 가능했으나
최근 은행 전산망 해킹 사건이 발생했으나 블록체인이면 사 실상 해킹이 불가능함.

17) 유통부문

KMS(Knowledge Management System) 지식관리시스템의 약자.
조직 내의 지식을 체계적으로 관리하는 시스템을 의미함. 이전에는 기업 대부분이 물품을
생산하던 환경이었지만 요즘에는 지적 재산이 매우 중요해짐에 따라 기업을 관리하는 시스템이 등장함

18) 사회기반구조로서 데이터베이스

분야	주요 솔루션
물류	·종합물류정보망 ·부가가치통신망
지리	NGIS,RS
교통	ITS
의료	EDI
교육	NEIS

19) 빅 데이터 (Big Data)

· 데이터의 변화

① Volume(데이터의 크기): 생성되는 모든 데이터를 수집 (트위터 매일 12테라 생성)

② Variety(데이터의 다양성): 정형화된 데이터를 넘어 텍스트/오디오/비디오 등
모든 유형의 데이터를 분석 대상으로 한다. (구글은 2,300억 단어 분석 음성인식 엔진 개발)

③ Velocity(데이터의 속도): 두 가지 관점의 속도를 의미함
즉 사용자 원하는 시간 내 데이터 분석 결과 제공과 데이터의 업데이트되는 속도가 매우 빨라짐.
(브리티시 텔레콤: 1초 60기가의 데이터 전송)

3V와 사례를 암기하세요

20) 기술변화 인재·조직 변화

· 기술변화

① 클라우드 컴퓨팅 활용

② 새로운 데이터 처리, 저장, 분석 기술 및 아키텍쳐

· 인재·조직 변화

데이터 중심 조직/데이터 사이언티스트 요구

21) 빅 데이터 출현 배경

· 빅 데이터 현상은 없었던 것이 새로 등장한 것이 아니라 기존의 데이터, 처리방식, 다루는 사람과 조직 차원에서 일어나는 '변화' 의미

① 산업계-양질 전환 법칙
정보가 지속해서 축적되면서 기업들이 보유한 데이터가
'거대한 가치 창출이 가능한 만큼 충분한 규모에 도달'

② 학계- 빅 데이터를 다루는 현상이 증가
인간게놈프로젝트 통한 유전자 정보 해석

③ 관련 기술발전- 디지털화, 저장기술발전, 인터넷과 모바일 시대 진전에 따른 클라우드 컴퓨팅

22) 빅 데이터 기능

① 빅 데이터는 산업혁명의 석탄, 철에 비유된다.

② 빅 데이터는 원유에 비유된다.

③ 빅 데이터는 렌즈에 비유된다
사례) ngram viewer, 현미경이 있다.

④빅 데이터는 플랫폼에 비유된다.
사례) 페이스북,OS 여기에 해당한다.

23) 빅 데이터가 만들어 내는 본질적 변화

① 정보의 사전처리에서 사후처리 시대로 표준화된 문서 포맷(형식)
② 표본조사에서 전수조사로
③ 질보다 양으로-구글의 자동번역, 결정계수
④ 인과관계에서 상관관계로

→ 데이터 기반의 상관관계 분석이 주는 인사이트가
인과관계 때문에 미래 예측을 점점 더 압도해 가는 시대가 도래하고 있다.

빅 데이터가 만든 본질적인 변화 이해하세요

24) 빅 데이터의 가치 산정이 어려운 이유

① 데이터 활용 방식

데이터의 재사용, 재조합(mashup), 다목적용 데이터 개발 등이 일반화되면서

특정 데이터를 언제, 어디서, 누가 활용할지 알 수 없다.

재사용 사례 – 구글 검색결과를 저장 후 재사용한다.

다목적용 사례 - 전기자동차의 배터리 충전시간

- CCTV(절도범 & 구매정보)

재조합 사례 - 휴대전화 전자파가 뇌종양 관계

② 데이터가 기존에 없던 가치 창출을 한다. 아마존 킨들 전자책 읽기 관련 데이터 분석을 하면 독서 패턴을 알 수 있다.(페이스북 소셜커머스 그래프)

③ 분석 기술의 발달이 데이터에 가치에 영향을 준다. 기존에는 가치가 없는 데이터도 새로운 분석기법으로 가치를 만든다. (SNS 비정형 데이터 이용한 텍스트마이닝 활용)

빅데이터 가치 산정이 어려운 이유와 사례 암기

25) 빅 데이터의 영향

매켄지는 빅 데이터 보고서를 통해 빅데이타가 가치를 만들어 내는 방식으로 5가지 언급

① 투명성 제고로 연구개발 및 맞춤 서비스 제공
② 시뮬레이션을 통한 수요 포착 및 주요 변수 탐색
③ 고객 세분화 및 맞춤 서비스 제공
④ 알고리즘을 활용한 의사결정 보조 혹은 대체
⑤ 비즈니스 모델과 제품, 서비스의 혁신
→ 기업 : 혁신, 경쟁력 제고, 생산성 향상
정부 : 환경탐색, 상황분석, 미래대응
개인 : 목적에 따라 활용

결국, 빅 데이터 활용 확산이 사람들의 생활이 스마트화

26) 빅 데이터 활용 대표 사례

① 기업 활용
- 구글 검색(로그 데이터 활용 기존 페이지랭크 개선)
- 월마트 구매 패턴 분석 (연관규칙)
- IBM 왓슨 인공지능 병원 진료에 활용

② 정부 활용
- 환경탐색(실시간 교통정보수집, 기후정보)
- 상황분석(소셜미디어, CCTV, 통화기록)

③ 개인 활용
- 정치인의 SNA 활용
- 가수 팬들의 청취 분석

빅데이터 활용 사례 암기

27) 빅 데이터 활용 기법

① 연관규칙학습(Association rule learning) : 어떤 수 간에 주목할 만한 상관관계가 있는지를 찾아내는 방법
(마트에서 상관관계가 높은 상품을 함께 진열 → 우유 & 기저귀)

② 유형 분석(Classification tree analysis) : '사용자가 어떤 특성을 가진 집단에 속하는가?'와 같은 문제를
해결하고자 할 때 사용 (온라인 수강생들의 특성에 따라 분류)

③ 유전 알고리즘(Genetic algorithms) : '최대의 시청률을 얻으려면 어떤 프로그램을
어떤 시간대에 방송해야 하는가?'와 같은 문제를 해결할 때 사용. 최적화의 메커니즘을 찾아가는 방법
(연료 효율적인 차를 개발하기 위해 어떻게 원자재와 엔지니어링을 결합해야 하는가?,
응급실에서 의사를 어떻게 배치하는 것이 가장 효율적인가?)

④ 기계 학습(Machine learning): 기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화 중에서
어떤 것을 가장 보고 싶어 할까? 와 같은 문제를 해결할 때 사용. 기계학습은 훈련 데이터로부터 학습한다고
알려진 특성을 활용해 '예측'하는 일에 초점을 맞춘다. (넷플릭스 영화추천 시스템)

⑤ 회귀분석(Regression Analysis): '구매자의 나이가 구매 차량의 타입에 어떤 영향을 미치는가?' 와 같은 질문에 답할 때 사용

⑥ 감정분석(Sentiment Analysis) : '새로운 환불 정책에 대한 고객의 평가는 어떤가?'를 알고 싶을 때 활용
(소셜미디어에 나타난 의견을 바탕으로 고객이 원하는 것을 찾아낼 때 사용된다.)

⑦ 소셜 네트워크 분석(Social network analysis) = 사회 관계망 분석(SNA) : 영향력 있는 사람을 찾아낼 수 있으면,
고객들 간 소셜커머스 관계를 파악할 수 있음

분석기법 정의와 사례를 암기하세요.

28) 빅 데이터 시대의 위기 요인과 통제방안

① 사생활 침해
→ (위기 요인) 빅 데이터 시대가 본격화되면서 우리를 둘러싼 정보 수집 센서(M2M)들의 수가 점점 늘어나고 있고, 특정 데이터가 본래 목적 외에 가공돼 2차·3차적 목적으로 활용될 가능성이 증가하면서 사생활 침해를 넘어 사회·경제적 위협으로 변형될 수 있음.

② 익명화(Anonymization): 사생활 침해를 방지하기 위해 데이터에 포함된 개인 식별 정보를 삭제하거나 알아 볼 수 없는 형태로 변환하는 것을 말한다.
→ (통제방안) 동의에서 책임으로–개인정보의 활용에 대한 개인이 매번 동의하는 것은 경제적으로도 매우 비효율적이다. 따라서 사생활 침해 문제를 개인정보 제공자의 동의를 통해 해결하기보다는 개인정보 사용자에게 책임을 지움으로써 개인정보 사용 주체가 더욱 적극적인 보호 장치를 마련하게 하는 효과가 발생할 것 으로 기대된다. (개인정보 사용자가 책임)

③ 책임 원칙의 훼손
→ (위기 요인) 빅 데이터 기반분석과 예측 기술이 발전하면서 정확도가 증가한 만큼, 분석 대상이 되는 사람들은 예측 알고리즘의 희생양이 될 가능성이 증가한다. 그러나 잠재적 위험 사항에 대해서도 책임을 추궁하는 사회로 변질할 가능성이 커 민주주의 사회 원칙을 크게 훼손할 수 있다.(범죄예측 프로그램)
→ (통제방안) 기존의 책임 원칙을 강화할 수밖에 없다.

④ 데이터의 오용
→ (위기 요인) 빅 데이터는 일어난 일에 대한 데이터에 의존한다. 그것을 바탕으로 미래를 예측하는 것은 적지 않은 정확도를 가질 수 있지만, 항상 맞을 수는 없다. 주어진 데이터에 잘못된 인사이트를 얻어 비즈니스에 직접 손실을 불러올 수 있다.
→ (통제방안) 데이터 알고리즘에 대한 접근권 허용 및 객관적 인증방안을 도입 필요성 제기 이로 인해 알고리즈미스트 역할 요구

위기 요인과 통제방안을 연계하세요.

29) 빅 데이터의 활용에 필요한 3요소

① **데이터** : 모든 것을 데이터화 하는 추세를 빅 데이터 시대에는 피할 수 없다.
특정한 목적 없이 생산된 데이터라도 창의적으로 재활용되면서 가치를 만들어 낼 수 있으므로

② **기술** : 빅 데이터 분석 알고리즘의 진화가 가속화될 것이다. 알고리즘은 데이터양의 증가에 따라
정확도가 증가하는 일반적인 경향이 있다. 그것은 알고리즘을 학습시킬 수 있는 데이터의 양이 증가하면서
알고리즘도 스마트해지는 경향이 있다.

③ **인력** : 데이터 사이언티스트와 알고리즈미스트의 역할이 중요해질 것으로 전망된다.
데이터사이언티스트는 빅 데이터의 다각적 분석을 통해 인사이트를 도출하고 이를 조직 전략 방향 제시에
활용할 줄 아는 기획자로서 전문가 역할을 할 것으로 기대된다.
알고리즈미스트는 데이터 사이언티스트가 한 일로 부당하게 피해가 발생하는 것을 막는 데 필요

→미래의 빅 데이터 현상 3가기 요소가 변화될 것

3가지 활용 요소 암기하세요

30) 빅 데이터 열풍과 회의론

① 시대의 분위기에 합류하기 위해 거액을 투자해 솔루션을 도입한 후 어떻게 활용하고 어떻게 가치를 뽑아내야 할지 첫 번째 물음부터 다시 시작

② 현재 소개되는 많은 빅 데이터 성공사례가 기존의 분석 프로젝트를 포장

③ 빅 데이터 분석도 데이터에서 가치, 즉 통찰을 끌어내 성과를 창출하는 것이 관건

31) 빅 데이터 분석, 'Big'이 핵심 아니다

① 데이터는 크기의 이슈가 아니라, 거기에서 어떤 시각과 통찰을 얻을 수 있느냐의 문제

② 비즈니스의 핵심가치에 집중하고 이와 관련된 분석 평가지표를 개발하고 이를 통해 효과적으로 시장과 고객 변화에 대응할 수 있을 때 빅 데이터 분석은 가치가 있음

③ 빅 데이터와 관련된 걸림돌은 '비용이 아니라 분석적 방법과 성과에 대한 이해 부족'

32) 전략적 통찰이 없는 분석의 함정

① 대부분 성과가 높은 기업일수록 데이터 기반에 의 한 의사결정을 하지만
성과가 우수한 기업들도 가치 분석력 통찰력을 갖췄다고 대답한 비율이 낮다.
→ 기업의 핵심가치와 관련한 전략적 통찰력을 가져다 주는 데이터 분석을 내재화하는 것이 쉬운 일이 아님.

② 아메리카 항공(실패) vs 사우스웨스트 항공(성공)
-(실패 원인) 아메리카 항공은 다른 항공사와 같은 분석전략 사용하여 차별화하지 못함
-(실패 원인) 쓸모없는 수익관리 분석기법 도입
저가 항공사들이 낮은 가격 제시하였기 때문에 분석 자체가 의미가 없다.

→ 단순히 분석이 많이 사용하는 것이 비교우위 의미하지 않음(중요한 것은 전략적 인사이트, 차별성)

33) 일차적인 분석 vs 전략 도출을 위한 가치 기반분석

일차적인 분석 애플리케이션 사례

산업	분석 애플리케이션
정부	사기탐지, 사례관리, 범죄방지, 수익 최적화
금융 서비스	사기탐지, 신용점수 산정, 가격책정
에너지	트레이딩, 공급, 수요 예측
온라인	웹 매트릭스, 사이트 설계
소매업	판촉, 수요 예측, 재고 보충
모든 산업	성과 관리

→일차적인 분석을 통해서도 해당 부서, 업무영역 효과를 얻을 수 있지만, 일차적인 분석은 태생적으로 업계 내부의 문제에만 초점을 두고 있음

→전략적 인사이트 가치 기반을 위해서 인구통계학적 변화, 경제사회 트랜드, 고객 니즈의 변화 고려해야 함

34) 데이터 사이언스 vs 데이터 마이닝 vs 통계학 차이

① 데이터 사이언스란 : 데이터로부터 의미 있는 정보를 추출하는 학문

② 통계학이 정형화된 실험 데이터를 분석 대상으로 하는 것에 비해, 데이터 사이언스는 정형 또는 비정형을 막론하고 다양한 유형의 데이터를 대상으로 총체적 접근법을 사용 (통계학과 차이)

③ 데이터 마이닝은 주로 분석에 초점 두나, 데이터사이언스는 분석뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지 모두 포괄하는 개념(데이터마이닝차이)

④ 결국, 데이터 사이언스란 데이터 공학, 수학, 통계학, 컴퓨터공학, 시각화, 해커의 사고방식, 해당 분야의 전문 지식을 종합한 학문으로 정의

→ **데이터 사이언스의 역할** : 전략적 통찰을 추구하고 비즈니스 핵심 이슈에 답을 하고, 사업의 성과를 견인

35) 데이터 사이언스의 핵심 구성 요소

① IT(Data Management)

② Analytics(분석적 영역)

③ 비즈니스 분석

3가지 핵심 구성 요소 암기하세요.

36) 데이터 사이언티스트가 갖춰야 할 역량(가트너)

① 데이터 관리: 데이터에 대해 이해
② 분석 모델링: 분석론에 대한 지식
③ 비즈니스 분석: 비즈니스 요소에 초점
④ 소프트 기능 : 커뮤니케이션, 협력, 리더십, 창의력

※ 데이터 사이언티스트의 갖춰야 할 역량의 공통점은 호기심이다.
호기심이란 문제의 이면을 파고들고, 질문들을 찾고, 검증 가능한 가설을 세우는 능력 또한,
스토리텔링, 커뮤니케이션, 직관력, 소통능력 필요

37) 데이터 사이언티스트 요구역량 (하드스킬 & 소프트 스킬)

· **Hard Skill**

① 빅 데이터에 대한 이론적 지식: 관련 기법에 대한 이해와 방법론 습득

② 분석 기술에 대한 숙련: 최적의 분석 설계 및 노하 우 축적

· **Soft Skill**

① 통찰력 있는 분석: 창의적 사고, 호기심, 논리적 비판

② 설득력 있는 전달: 스토리텔링, Visualization

③ 다분야 간 협력: Communication

38) 인문학의 부활 이유

① 단순 세계화에서 복잡한 세계로의 변화
- 다양성과 각 사회의 정체성, 연결성, 창조성 키워드 대두

② 비즈니스의 중심이 제품생산에서 서비스로 이동
- 고객에게 얼마나 뛰어난 서비스를 제공하는가 여부가 관건

③ 경제와 산업의 논리가 생산에서 시장 창조
- 무형자산이 중요

39) 데이터 사이언티스트 6가지 핵심 질문

	과거	현재	미래
정보	무슨 일이 일어났는가? -리포팅(보고서)	무슨 일이 일어나고 있는가? -경고	무슨 일이 일어날 것인가? -추출
통찰	어떻게, 왜 일어났는가? -모델링,실험설계	차선 행동은 무엇인가? -권고	최악, 최선의 상황은? -예측, 최적화, 시뮬레이션

정보와 통찰 예를 기억하세요.

40) 데이터 분석 모델링에서 인문학적 통찰력의 적용 사례

신용 리스크 모델을 예로 들자면 인문학적 관점은 3가지로 관점으로 요약

① 성향의 관점
② 행동적 관점
③ 상황적 관점

41) 데이터 사이언스 한계와 인문학

분석과정에서 가정 등 인간의 해석이 개입되는 단계를 반드시 거치게 된다.
아무리 정량적인 분석이라도 명심해야 할 것은 모든 분석에 가정에 근거한다는 사실이다.

→ 항상 인문학자들처럼 모델의 능력에 대해 항상 의구심 가지고, 가정들과 현실의 불일치에 대해 고찰하고,
분석 모델이 예측할 수 없는 위험을 살펴봐야 한다.

42) 빅 데이터 회의론을 넘어 가치 패러다임 변화

· 가치 패러다임 : 경제와 산업 근저에는 다양한 가치 원천이 존재하며, 무작위로 작용하는 것이 아니라 특정 기간 지배적으로 작용함. 이러한 가치 원천은 일정 기간 근본적인 존재로 강력한 힘을 행사하다가 효력이 다하면 다음의 가치 패러다임에 지배적인 지위를 넘겨줌

· 가치 패러디임의 변화

① 디지털화(Digitalization):아날로그의 세상을 어떻게 효과적으로 디지털화하는가가 이 시대의 가치를 창출해 내는 원천

사례) 도스 운영프로그램, 워드/파워포인트와 같은 오피스프로그램 등

② 연결(Connection):디지털화된 정보와 대상들이 서로 연결되어, 이 연결이 얼마나 효과적이고 효율적으로 제공해 주느냐가 이 시대의 성패를 결정함

사례) 구글의 검색 알고리즘, 네이버의 콘텐츠

③ 에이전시(Agency):사물인터넷(IoT)의 성숙과 함께 연결이 증가하고 복잡한 연결을 얼마나 효과적이고 믿을 만하게 관리 하는가가 이슈 데이터 사이언스의 역량에 따라 좌우

43) 개인정보 비식별화

· 개인정보: 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보

· 비식별화 : 정보의 일부 또는 전부를 삭제 또는 대체하거나 다른 정보와 쉽게 결합하지 못하도록 하여
특정 개인을 알아볼 수 없도록 하는 일련의 조치

최근 개인정보 비식별화 출제빈도가 매우 높음.

44) 개인정보 식별요소 제거방법 및 예시

비식별 기술	제거방법	예시
가명처리	식별요소를 다른 값으로 대체	홍길동, 35세, 서울 거주, 한국대 재학 ->임꺽정, 30대 서울 거주, 국제대 재학
총계처리 또는 평균값 대체	데이터를 총합으로 표시하여 개별 데이터 값을 보이지 않도록 함	임꺽정 180cm, 홍길동 170cm ->1-5반 학생 키 합 350cm, 평균 키 175cm
데이터값 삭제	개인 식별을 인식할 수 있는 값 삭제	홍길동, 35세, 서울 거주, 한국대 졸업 ->35세, 서울 거주
범주화	범주의 값으로 변환	홍길동, 35세 ->홍 씨, 30~40세
데이터 마스킹	개인 식별자가 보이지 않도록 처리	홍길동, 35세 -.홍 * *,35세

비식별기술과 예시를 같이 알고 있어야 함.

45) 관계형 데이터베이스 관리시스템(RDBMS) vs 객체 지향 데이터베이스 관리시스템(ODBMS)

구분	RDBMS	ODBMS
주된 장점	· 검증된 시스템 · 대규모 정보처리 가능	복잡한 정보 구조의 모델링 가능
주된 단점	· 제한된 형태의 정보만 처리 가능 · 복잡한 정보 구조의 모델링이 어려움	사용자 정의 타입 및 비정형 복합 정보 타입 지원 가능

46) 데이터 유형 분류

데이터 유형	특징	데이터 종류
정형 데이터	· RDBMS의 고정된 필드에 저장 · 데이터 스키마 지원	RDB 스프레드시트
반정형 데이터	· 데이터 속성인 메타데이터를 가지며, 일반적으로 스토리지에 저장되는 데이터 파일	HTML JSON 웹 문서 센서 데이터
비정형 데이터	· 형태가 구조가 복잡한 이미지, 동영상 같은 멀티미디어 데이터	소셜 데이터 문서 이미지 오디오, 비디오

데이터 유형과 종류를 구분하세요.