1과목 데이터의 이해

작성자 : 윤소영

ADsP 헬프 데스크

구분	링크
종합 정보 안내	https://eduatoz.notion.site/f55a34175dd14f42b14161140a27c1d2?v=6be282fd277c43e2a2bf014db0514fb0
1과목 QnA 정리문서	https://colab.research.google.com/drive/1cCk43pbnapr0mxOSiC8ssN3ya5TJwSqG
2과목 QnA	https://colab.research.google.com/drive/1J7Ul9W-bVobaAob0wQKIsho3Uo8HkeVq
3과목 R – QnA	https://colab.research.google.com/drive/1VmixW_RYpn8_XycGAjgqSCZ00sR-_8Ke
3과목 통계분석 – QnA	https://colab.research.google.com/drive/1QDuCKk86lKTP8oxOTwOo1D2935Tu-5-e
3과목 정형분석 – QnA	https://colab.research.google.com/drive/1_NOLfLHlYrmAxBXcpcqNV1pwHme9C4IO
NoSQL 읽기자료	https://meetup.toast.com/posts/274
과목별 요약 강의 듣기 (R은 시험대비만 가능)	https://youtube.com/playlist?list=PLnp1rUgG4UVZO4ndD_HITLiBb8GlrtUOI
추가 강의 듣기	https://youtube.com/playlist?list=PLnp1rUgG4UVaHL5KKWkJxpT02X7Fh6ggv

링크는 변경이 없으며, 내용은 매 시험 때마다 계속 추가 됩니다.

교재 하단의 페이지는 영상 강의 페이지와 맞춘 것입니다.
일련번호가 아님에 유의하세요!
영상과 일치하지 않는 페이지가 있으면 imbgirl@naver.com 으로 문의 주세요 ^^!

합격을 기원합니다!

01 - 01

데이터 이해 - 데이터의 이해

1-01. 데이터 유형

데이터의 정의

- 데이터는 개별 데이터 자체로는 의미가 중요하지 않은 객관적인 사실(fact)

- 추론, 예측, 전망, 추정을 위한 근거(basis)로 기능하는 특성을 갖음
- 다른 객체와의 상호 관계 속에서 가치를 갖음

데이터 유형

정성적 데이터 (qualitative data)	▪ 자료의 성질, 특징을 자세히 풀어 쓰는 방식 ▪ 언어, 문자로 기술 (예: 설문조사의 주관식 응답, SNS에 올린 글, 기상특보) ▪ 비정형 데이터 형태로 저장, 분석에 시간과 비용이 필요함
정량적 데이터 (quantitative data)	▪ 수치, 기호, 도형으로 표시 (예: 지역별 온도, 풍속, 강우량) ▪ 데이터 양이 증가하더라도 저장, 분석이 용이함

1-02. 암묵지 vs 형식지

- 가장 널리 알려진 지식의 차원은 Polanyi에 의해 구분된 “암묵지와 형식지”이다

암묵지

- 학습과 체험을 통해 개인에게 습득 (현장 작업과 같은 경험을 통해 획득)
- 시행착오와 오랜 경험을 통해 개인에게 습득된 무형 지식
- 예)김장김치 담그기, 자전거 타기
- 공유되기 어려움

형식지

- 교과서, 매뉴얼, 비디오, DB 등으로 형상화 된 지식을 의미
- 예) 회계, 재무 관련 대차대조표에 요구되는 지식의 매뉴얼
- 외부로 표출되어 여러 사람이 공유할 수 있는 지식

1-03. 암묵지와 형식지 상호작용

- "지식경영" 이란?

> 개인의 암묵지와 집단에서의 형식지가 나선형의 형태로 회전하면서 생성, 발전, 전환되는 지식의 발전을 기반으로 한 기업의 경영

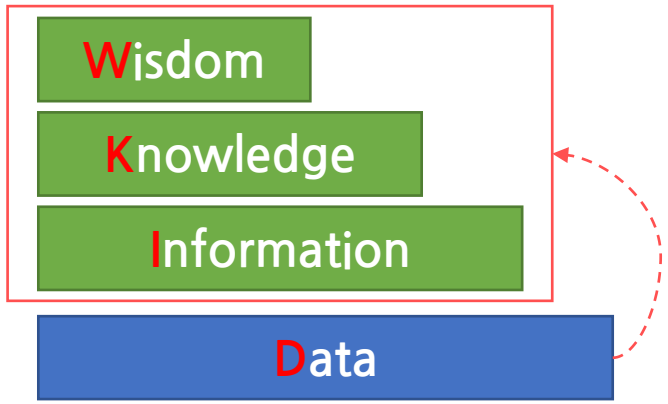
- 암묵지, 형식지의 4단계 지식전환 모드

1단계 : 공통화 (암-암)	▪ 암묵적 지식 노하우를 다른 사람에게 알려주는 것
2단계 : 표출화 (암-형)	▪ 암묵적 지식 노하우를 책이나 교본 등 형식지로 만드는 것
3단계 : 연결화 (형-형)	▪ 책이나 교본(형식지)에 자신이 알고 있는 새로운 지식(형식지)을 추가하는 것
4단계 : 내면화 (형-암)	▪ 만들어진 책이나 교본(형식지)을 보고 다른 직원들이 암묵적 지식(노하우)을 습득

'ㄱ'과 'ㄴ' 사이에 '표연'이 있다!

1-04. 데이터와 정보의 관계

- Data → Information → Knowledge → Wisdom 계층구조
  - 데이터를 가공 처리하여 얻을 수 있는 것 : 정보, 지식, 지혜!



A 마트의 다른 상품들도 B마트보다 쌀 것이라고 판단한다
상대적으로 저렴한 A마트에서 연필을 사야 겠다
A마트의 연필이 더 싸다
A마트는 100원에, B마트는 200원에 연필을 판매한다

데이터(Data)	▪ 타 데이터와의 상관관계가 없는 가공하기 전의 순수한 수치나 기호
정보(Information)	▪ 데이터의 가공 및 상관/연관 관계 속에서 의미가 도출된 것
지식(Knowledge)	▪ 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물
지혜(Wisdom)	▪ 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 아이디어

- 데이터베이스는 "통합, 저장, 공용, 변화되는 데이터"를 특징으로 한다!

통합 데이터 (Integrated)	▪ 데이터베이스에 같은 내용의 데이터가 중복되어 있지 않다는 것을 의미
저장 데이터 (Stored)	▪ 자기디스크나 자기테이프 등과 같이 컴퓨터가 접근할 수 있는 저장매체에 저장되는 것을 의미
공용 데이터 (Shared)	▪ 여러 사용자에게 서로 다른 목적으로 데이터베이스의 데이터를 공동으로 이용되는 것을 의미
변화되는 데이터 (Changed)	▪ 새로운 데이터의 추가, 기존 데이터의 삭제, 갱신으로 항상 변화하면서도 항상 현재의 정확한 데이터를 유지해야 한다는 것을 의미

DBMS, RDBMS, ODBMS

DBMS

- 사용자와 데이터베이스 사이에서 사용자의 요구에 따라 정보를 처리해주고 데이터베이스를 관리해주는 소프트웨어

RDBMS

- 관계형 데이터베이스 관리 시스템
- 정형화된 테이블로 구성된 데이터 항목들의 집합체
- MySQL(오픈소스 RDBMS), Oracle Database(상용 RDBMS)
- SQL : RDBMS의 데이터를 관리하기 위해 설계된 특수 목적의 프로그래밍 언어 챔벌린과 레이먼드 F. 보이스가 처음 개발하였음

ODBMS

- 객체 지향 데이터베이스 관리 시스템
- 객체들을 생성하여 계층에서 체계적으로 정리하고, 다시 계층들을 하위 계층이 상위 계층으로부터 속성과 방법들을 물려받을 수 있는 DBMS
- 복잡한 데이터 구조를 표현 및 관리하는 DBMS

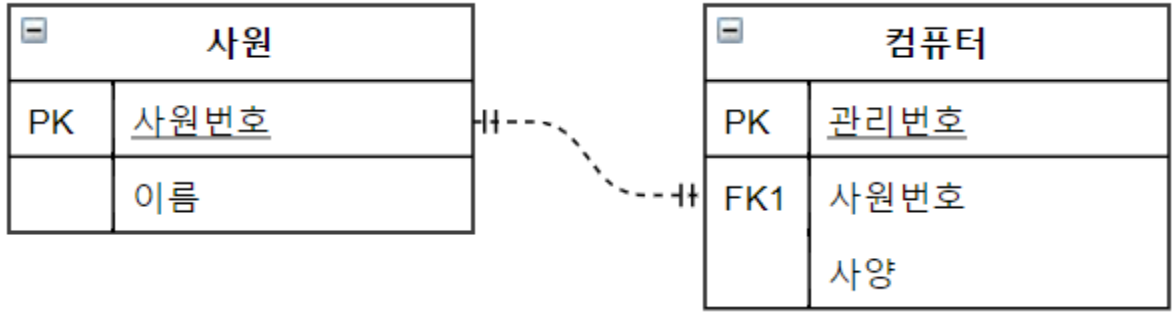
Customer

Column Name	Data Type	Key	Null	Description
Emailid	varchar2(50)	Primary key	Not null	User Email Id as primary key
FName	varchar2(20)		Not null	User first name
LName	varchar2(20)		Not null	User last name
Address	varchar2(20)		Not null	User address for deliver address
Cnumber	number		Not null	User contact number
Password	Varchar2(20)		Not null	User password for login

Validate

Column Name	Data Type	Key	Null	Description
Emailid	varchar2(50)	Foreign key	Not null	User Email Id as Foreign key
Hashcode	varchar2(256)		Not null	Hashcode MD5 generate combining emailid and date
Valid_date	Date		Not null	Link valid only for 48 hours
Vtype	varchar2(10)		Not null	Verification type describe for register/password link generate
Flag	boolean		Not null	True-Link used False-Link not used

```
SELECT Col1, Col2
        , CASE WHEN Col3 BETWEEN 10 AND 19 THEN '10'
                    WHEN Col3 BETWEEN 20 AND 29 THEN '20'
                    WHEN Col3 BETWEEN 30 AND 39 THEN '30'
                    WHEN Col3 BETWEEN 40 AND 49 THEN '40'
                    WHEN Col3 BETWEEN 50 AND 59 THEN '50'
                    WHEN Col3 >= 60 THEN '60+'
          END AGE_RANGES
          , COUNT(CASE WHEN Col4 = '1' THEN Col 5 END) Feature1
          , COUNT(CASE WHEN Col4 = '2' THEN Col 5 END) Feature2
          , COUNT(CASE WHEN Col4 = '3' THEN Col 5 END) Feature3
FROM Table 1
 GROUP BY  Col1, Col2
. CASE WHEN Col3 BETWEEN 10 AND 19 THEN '10'
```



데이터베이스 설계 절차

요구조건분석 -> 개념적 설계 -> 논리적 설계 -> 물리적 설계

요구조건 분석	▪ 데이터베이스 사용자, 사용목적, 사용범위, 제약조건 등을 정리, 명세서 작성
개념적 설계	▪ E-R 모델, 정보를 추상적 개념으로 표현하는 과정, DBMS 독립적 E-R 다이어그램 작성
논리적 설계	▪ 자료를 컴퓨터가 이해할 수 있도록 특정 DBMS의 논리적 자료 구조로 변환
물리적 설계	▪ 논리적 구조로 표현된 데이터를 물리적 구조의 데이터로 변환하는 과정

NoSQL(Non-SQL, Non-Relational, Not Only SQL …)

- 관계형 데이터베이스보다 덜 제한적인 일관성 모델을 이용하는 데이터의 저장 및 검색을 위한 메커니즘 제공, 디자인 단순화, 수평적 확장성, 세세한 통제 등을 포함
- 기존의 RDBMS가 갖고 있는 특성 뿐만 아니라 다른 특성들을 부가적으로 지원함

NoSQL 저장방식 도구 : MongoDB, Apache HBase, Redis

- MongoDB : 데이터 교환 시 비산(BSON: Binary JSON) 문서 형태로 저장하여 여러 서버에 분산 저장 및 확장이 용이하며, 방대한 데이터 처리가 빠르다는 장점이 있다. C++로 작성됨
- Apache HBase : 하둡 플랫폼을 위한 공개 비관계형 분산 데이터 베이스이다. 구글의 빅테이블(BigTable)을 본보기로 삼았으며 자바로 쓰여졌다.
- Redis : Remote Dictionary Server의 약자, "키-값" 구조의 비정형 데이터를 저장하고 관리하기 위한 오픈 소스 기반의 비관계형 데이터베이스 관리 시스템(DBMS)이다.

1-09. 기업 내부 데이터베이스 솔루션 - 1

- 시대별 기업 내부 데이터베이스 솔루션
  - 1980년대 : OLTP, OLAP, 2000년대 : CRM, SCM

OLTP

- On-Line Transaction Processing, 온라인 거래 처리, 예) 상품주문, 회원 정보 수정
- 주 컴퓨터와 통신회선으로 접속되어 있는 복수의 사용자 단말에서 발생한 트랜잭션을 주 컴퓨터에서 처리하여 그 결과를 사용자에 되돌려 보내 주는 처리형태

OLAP

- On-Line Analytical Processing, 온라인 분석 처리, 예) 10년간 A사의 직급별 임금 상승률
- 다차원으로 이루어진 데이터로부터 통계적인 요약 정보를 제공할 수 있는 기술, 다차원의 데이터를 대화식으로 분석하기 위한 SW

1-09. 기업 내부 데이터베이스 솔루션 - 2

- 시대별 기업 내부 데이터베이스 솔루션
  - 1980년대 : OLTP, OLAP, 2000년대 : CRM, SCM

CRM

- Customer Relationship Management
- 고객별 구매 이력 데이터베이스를 분석하여 고객에 대한 이해를 돕고 이를 바탕으로 각종 마케팅 전략을 통해 보다 높은 이익을 창출할 수 있는 솔루션

SCM

- Supply Chain Management
- 제조, 물류, 유통업체 등 유통공급망에 참여하는 모든 업체들이 협력을 바탕으로 정보기술(Information Technology)을 활용, 재고를 최적화하기 위한 솔루션
- 기업이 외부 공급업체 또는 제휴업체와 통합된 정보시스템으로 연계하여 시간과 비용을 최적화 시키기 위한 것
- 자재구매 데이터, 생산, 재고 데이터, 유통/판매 데이터, 고객 데이터로 구성됨

분야별 기업 내부 데이터베이스 솔루션 - 제조부문

Data Warehouse

- 기업 내의 의사결정 지원 애플리케이션을 위한 정보를 제공하는 하나의 통합된 데이터 저장 공간
- ETL : 추출, 변환, 적재(Extract, transform, load)
  주기적으로 내부 및 외부 데이터베이스로부터 정보를 추출하고 정해진 규약에 따라 정보를 변환한 후에 정보를 적재함
- 데이터들은 시간적 흐름에 따라 변화하는 값을 일정기간 유지

데이터 웨어하우스의 4대 특성

- 데이터의 통합 : 데이터들은 전사적 차원에서 일관된 형식으로 정의됨
- 데이터의 시계열성 : 관리되는 데이터들은 시간의 흐름에 따라 변화하는 값을 저장함
- 데이터 주제 지향적 : 특정 주제에 따라 데이터들이 분류, 저장, 관리됨
- 비소멸성(비휘발성) : Batch 작업에 의한 갱신이외에 변하지 않음 (빈번한 삽입, 삭제 아님)

분야별 기업 내부 데이터베이스 솔루션 - 제조부문

Data Mart	▪ 전사적으로 구축된 데이터 웨어하우스로부터 특정 주제, 부서 중심으로 구축된 소규모 단일 주제의 데이터 웨어하우스 ▪ 재무, 생산, 운영과 같이 특정 조직의 특정 업무 분야에 초점을 두고 있음
ERP	▪ Enterprise Resource Planning, 제조업을 포함한 다양한 비즈니스 분야에서 생산, 구매, 재고, 주문, 공급자와의 거래, 고객 서비스 제공 등 주요 프로세스 관리를 돕는 여러 모듈로 구성된 통합 애플리케이션 소프트웨어 패키지

분야별 기업 내부 데이터베이스 솔루션 - 제조부문

BI (Business Intelligence)

- 기업의 Data Wearhouse에 저장된 데이터에 접근해 경영의사결정에 필요한 정보를 획득하고 이를 경영활동에 활용하는 것
- 데이터를 통합/분석하여 기업 활동에 연관된 의사결정을 돕는 프로세스를 말함
- 가트너는 '여러 곳에 산재하여 있는 데이터를 수집하여 체계적이고 일목요연하게 정리함으로써 사용자가 필요로 하는 정보를 정확한 시간에 제공할 수 있는 환경'으로 정의함
- 하나의 특정 비즈니스 질문에 답변하도록 설계

ad hoc report

- BI와 빅데이터 분석의 차이점을 표현한 키워드
- Optimization, forecast, insight : 빅데이터 분석 관련 키워드임

BA (Business Analytics)

- 경영 의사결정을 위한 통계적이고 수학적인 분석에 초점을 둔 기법
- 성과에 대한 이해와 비즈니스 통찰력에 초점을 둔 분석 방법
- 사전에 예측하고 최적화하기 위한 것으로 BI 보다 진보된 형태

1-09. 기업 내부 데이터베이스 솔루션 - 6

분야별 기업 내부 데이터베이스 솔루션 - 금융부문

블록체인 (Block Chain)

- 기존 금융회사의 중앙 집중형 서버에 거래 기록을 보관하는 방식에서 벗어나 거래에 참여하는 모든 사용자에게 거래 내용을 보내주며 거래 때마다 이를 대조하는 데이터 위조 방지 기술

그 외에 EAI, EDW, ERP, e-CRM 등이 있다

분야별 기업 내부 데이터베이스 솔루션 - 유통부문

KMS

- Knowledge Management System
- 지식관리시스템의 약자, 조직 내의 지식을 체계적으로 관리하는 시스템을 의미

RFID

- 무선주파수(RF, Radio Frequency)를 이용하여 대상을 식별할 수 있는 기술
- RF 태그에 사용 목적에 알맞은 정보를 저장하여 적용 대상에 부착한 후 판독기에 해당되는 RFID 리더를 통해 정보를 인식함

기업 내부 데이터베이스 솔루션 인지 아닌지 구분할 수 있어야 함 : SCM, CRM, ERP, KMS은 기억

01 - 02

데이터 이해 - 데이터의 가치와 미래

1-10. 빅데이터란?

빅데이터 정의

- 빅데이터는 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터이다
- 빅데이터는 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처이다
- 데이터의 양(Volume) 데이터 유형과 소스 측면의 다양성(Variety), 데이터 수집과 처리 측면에서 속도(Velocity)가 급격히 증가하면서 나타난 현상이다

빅데이터 - 4V(ROI, Return On Investment, 투자자본수익률 관점에서 보는 빅데이터)

Volume	▪ 데이터의 크기, 생성되는 모든 데이터를 수집
Variety	▪ 데이터의 다양성, 정형화된 데이터를 넘어 텍스트, 오디오, 비디오 등 모든 유형의 데이터를 대상으로 함
Velocity	▪ 데이터의 속도, 사용자가 원하는 시간 내 데이터 분석 결과 제공, 업데이트 속도 빠름
Value	▪ Value는 '비즈니스 효과 요소', Volume, Variety, Velocity는 '투자비용 요소'이다

빅데이터의 출현 배경

- 산업계에서 일어난 변화를 보면 빅데이터의 현상은 양질 전환 법칙으로 설명할 수 있다
  양질 전환 법칙 : 일정한 양이 누적되면 어느 순간 질적인 비약이 이루어짐
  기업들이 보유한 데이터가 '거대한 가치 창출이 가능할 만큼 충분한 규모 ' 에 도달
- 학계의 거대 데이터 활용 과학 확산
  학계에서도 빅데이터를 다루는 현상들이 늘어나고 있다. 대표적 사례는 인간 게놈 프로젝트가 있다
- 디지털화, 저장기술, 인터넷 보급, 모바일 혁명, 클라우드 컴퓨팅 등 관련 기술 발전과 관련이 있다
  - 클라우드 컴퓨팅 : 빅 데이터 분석에 경제적 효과를 제공해준 결정적 기술
    클라우드 분산 병렬처리 컴퓨팅은 대용량 데이터 처리 비용을 획기적으로 줄임
- 소셜 미디어, 영상 등 비정형 데이터의 확산
- 데이터 처리 기술 발전

- 빅데이터는 "석탄/철, 원유, 렌즈, 플랫폼 " 이다!

석탄, 철	▪ 빅데이터는 석탄, 철이 산업혁명에서 했던 역할을 지금의 제조업 뿐 아니라 서비스 분야의 생산성을 획기적으로 끌어올려 혁명적 변화를 가져올 것으로 기대된다
원유	▪ 각종 비즈니스, 공공기관 대국민 서비스, 경제 성장에 필요한 '정보'를 제공하여, 산업 전반의 생산성을 향상시킬 것으로 기대된다
렌즈	▪ 현미경이 생물학 발전에 미쳤던 영향만큼 데이터가 산업 전반에 영향을 미칠 것이다 ▪ 구글 'Ngram Viewer'를 통해 수천만 권의 책을 디지털화
플랫폼	▪ 비즈니스 측면에서는 '공동 활용의 목적으로 구축된 유/무형의 구조물 ' 을 의미함 ▪ 페이스북과 같이 다양한 사업자들이 공동으로 사용하는 플랫폼을 빅데이터 형태로 제공할 것으로 예상 ▪ 각종 사용자 데이터와 센서 데이터를 수집하고 API를 공개하면 서드파티 사용자들이 활용하는 플랫폼 역할을 기대

1-13. 빅데이터의 가치 산정, 본질적 변화

빅데이터의 가치 산정이 어려운 이유

데이터의 활용 방식	▪ 재사용이나 재조합, 다목적용 데이터 개발 등이 일반화되면서 특정 데이터를 언제,어디서,누가 활용할지 알 수 없다
새로운 가치 창출	▪ 데이터가 기존에 없던 가치를 창출함에 따라 그 가치를 측정하기 어렵다
분석기술의 발달	▪ 분석 기술의 발달로 지금은 가치 없는 데이터도 새로운 분석 기법의 등장으로 거대한 가치를 만들어내는 재료가 될 가능성이 있다

빅데이터가 만들어내는 본질적인 변화

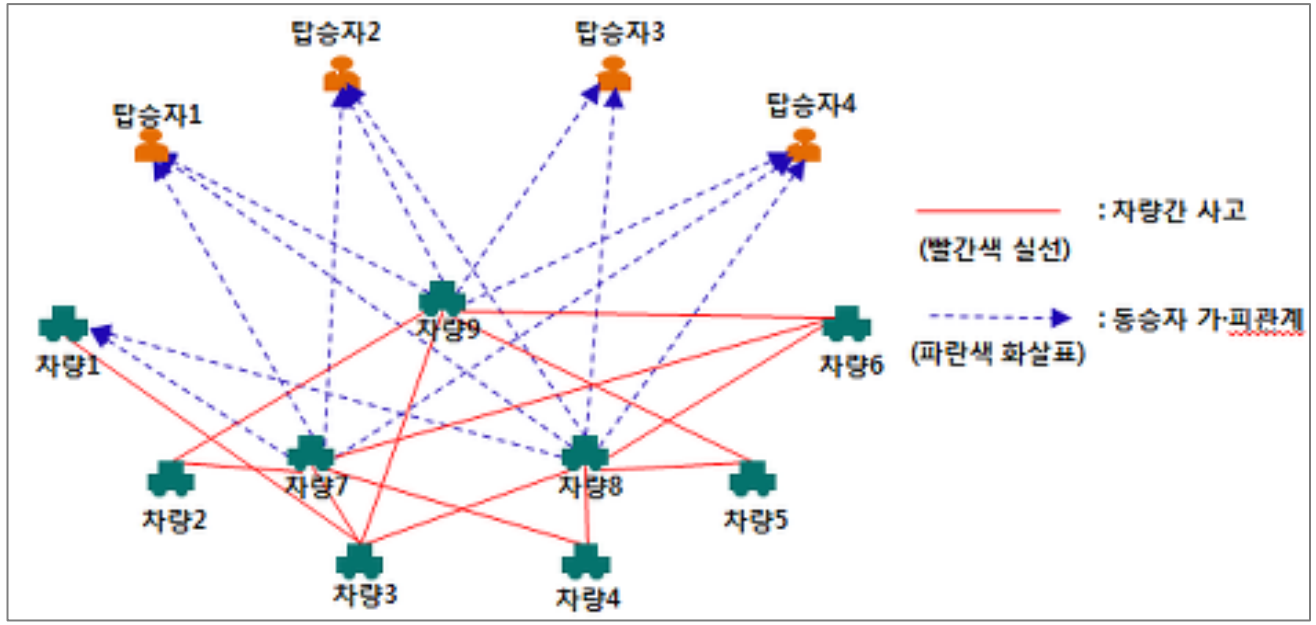
사전처리	사후처리
표본조사	전수조사
질(Quality)	양(Quantity)
인과관계	상관관계

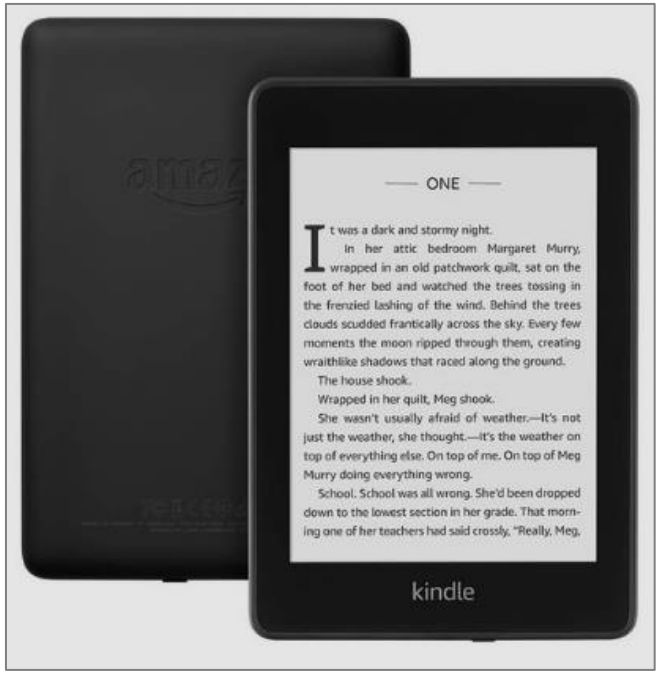
사전처리 => 표준화된 문서 포맷

사후처리 => 데이터를 모은 뒤 그 안에서 숨은 정보를 찾아냄

1-14. 빅데이터 활용 사례

- 구글 검색엔진, 월마트의 구매 패턴 분석, IBM 왓슨 - 의료 분야에 활용
- 정부의 실시간 교통정보 활용, CCTV 국가 안전에 활용
- 사회관계망분석을 통한 현상분석, 가수의 팬 음악청취 기록 분석 활용
- 아마존의 킨들(Kindle, 전자책 전용 단말기)에 쌓이는 전자책 읽기 관련 데이터 분석해 저자들에게 제공





1-15. 빅데이터 활용 기법 -1/2

연관규칙학습 Association rule Learning	▪ 변수간 주목할 만한 상관관계가 있는지 찾아내는 방법 ▪ 우유구매자가 기저귀도 같이 구매하는가? ▪ 커피를 사는 사람들이 탄산음료도 많이 구매하는가?
유형분석 Classification tree Analysis	▪ 사용자는 어떤 특성을 가진 집단에 속하는가? 와 같은 문제 해결에 사용함 ▪ 문서를 분류하거나 조직을 그룹으로 나눌 때, 온라인 수강생들을 특성에 따라 분류할 때 사용함
유전 알고리즘 Generic Algorithms	▪ 최적화가 필요한 문제의 해결책을 자연선택, 돌연변이 등과 같은 메커니즘을 통해 점진적으로 진화 시켜 나가는 방법 ▪ 최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가? ▪ 응급실에서 의사를 어떻게 배치하는 것이 가장 효율적인가?

1-15. 빅데이터 활용 기법 -2/2

기계학습

- 훈련 데이터로부터 패턴을 학습해 '예측'하는 일에 활용되고 있음
- 기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화 중 어떤 것을 가장 보고 싶어할까? (넷플릭스 추천 시스템)

회귀분석

- 선형함수로 나타낼 수 있는 수치데이터 분석
- 사용자의 만족도가 충성도에 어떤 영향을 미치는가?

감정분석

- 특정 주제에 대해 말하거나 글을 쓴 사람의 감정을 분석함
- 소셜 미디어에 나타난 의견을 바탕으로 고객이 원하는 것을 찾아낼 때 활용함
- 호텔에서 고객의 논평을 받아 서비스를 개선하기 위해 활용함

소셜 네트워크 분석

- = 사회관계망분석(SNA)
- 영향력 있는 사람을 찾아낼 수 있으면, 사람들 간 소셜 관계를 파악할 수 있다

"감정분석, 소셜 네트워크 분석을 구별하자!"

1-16. 빅데이터 위기요인과 통제방안 - 1/3

- 빅데이터 위기요인의 종류에는 사생활 침해, 책임 원칙의 훼손, 데이터의 오용이 있다
- 1. 사생활 침해

위기요인

- 우리를 둘러싼 정보 수집 센서들의 수가 점점 늘어나고 있고, 특정 데이터가 본래 목적 외에 가공 처리돼 2차 3차적 목적으로 활용될 가능성이 증가
- 익명화(Anonymization) : 사생활 침해를 방지하기 위해 데이터에 포함된 개인 식별 정보를 삭제하거나 알아볼 수 없는 형태로 변환 하는 것

통제방안

- 동의제에서 책임제로 전환
- 개인정보의 활용에 대한 개인이 매번 동의하는 것은 경제적으로도 매우 비효율적임
- 사생활침해 문제를 개인정보 제공자의 동의를 통해 해결하기 보다는 개인 정보 사용자에게 책임을 지움으로써 개인정보 사용 주체가 보다 적극적인 보호 장치를 강구하게 하는 효과가 발생할 것으로 기대됨

1-16. 빅데이터 위기요인과 통제방안 - 2/3

2. 책임 원칙의 훼손

위기요인	▪ 빅데이터 기반분석과 예측 기술이 발전하면서 정확도가 증가한 만큼, 분석 대상이 되는 사람들은 예측 알고리즘의 희생양이 될 가능성이 증가함 ▪ 그러나 잠재적 위험 사항에 대해서도 책임을 추궁하는 사회로 변질될 가능성이 높아 민주주의 사회 원칙을 크게 훼손할 수 있다 ▪ (예) 범죄 예측 프로그램을 통해 범죄 전 체포
통제방안	▪ 기존의 책임원칙을 강화할 수 밖에 없다

1-16. 빅데이터 위기요인과 통제방안 - 3/3

3. 데이터의 오용

위기요인

- 빅데이터는 일어난 일에 대한 데이터에 의존함
- 그것을 바탕으로 미래를 예측하는 것은 적지않은 정확도를 가질 수 있지만, 항상 맞을 수는 없음
- 주어진 데이터에 잘못된 인사이트를 얻어 비즈니스에 직접 손실을 불러 올 수 있음

통제방안

- 데이터 알고리즘에 대한 접근권 허용 및 객관적 인증방안을 도입 필요성 제기

알고리즈미스트

- 데이터 분석 알고리즘으로 부당한 피해를 보는 사람을 방지하기 위해서 생겨난 직업
- 데이터 분석 알고리즘으로 인해 피해를 입은 사람을 구제하는 전문가

1-17. 개인 정보 비식별화 기법

개인 정보 비식별화 기법

데이터 마스킹 (Masking)	▪ 다양한 유형의 데이터 관리 시스템에 저장된 정보를 보호하는 데 사용되는 프로세스 (카드 뒤 4자리 숨기기, 주민번호 뒤 6자리 숨기기)
데이터 범주화	▪ 변수가 가질 수 있는 가능한 값들을 몇 개의 구간으로 범주화 ▪ 홍길동, 35세 => 홍씨, 30-40세
가명	▪ 개인식별 정보를 삭제, 알아볼 수 없는 형태로 변환 ▪ 홍길동, 국제대 재학 => 임꺽정, 한성대 재학
잡음 첨가	▪ 자료 값에 잡음을 추가하거나 곱해 원래 자료에 약간의 변형을 가하여 공개
총계 처리 / 평균값 대체	▪ 데이터의 총합 값을 보임으로 개별 데이터의 값이 보이지 않도록 함
데이터 값 삭제	▪ 데이터 셋의 값 중 필요 없는 값 또는 개인 식별에 중요한 값 삭제

01 - 03

데이터 이해 - 가치 창조를 위한 데이터 사이언스와 전략 인사이트

1-18. 빅데이터 열풍

- IT 솔루션은 "공포 마케팅"이 잘 통하는 영역

도입만 하면 모든 문제를 한번에 해소할 것처럼 강조하다 나중에는 합류하지 못하면 위험에 처할지도 모른다는 공포 분위기 조성!

거액의 투자를 하지만, 어떻게 활용하고 어떻게 가치를 뽑아내야 할지 첫 번째 물음부터 다시 해야 하는 사태가 벌어짐

빅데이터 열풍 또한 유사한 패턴과 흐름을 갖는다

기존 분석 프로젝트를 포장해 놓은 것이 많음

성공적인 인터넷 기업!

데이터 분석과 함께 시작되고 분석이 내부 의사결정에 결정적 정보를 제공함

성공하지 못한 인터넷 기업!

데이터 분석에 기초해 전략적 통찰을 얻고, 효과적인 의사결정을 내리고 성과를 만들어 내는 체계가 없었음

1-19. 빅데이터 분석

- 빅데이터 분석, 'Big' 이 핵심이 아니다
  - 데이터의 양이 아닌 유형의 다양성과 관련이 있음
  - 음성, 텍스트, 이미지, 비디오 ➔ 다양한 정보 원천의 활용
- 전략적 통찰이 없는 분석의 함정
  - 한국의 경영 문화는 여전히 분석을 국소적인 문제 해결 용도로 사용하는 단계
  - 기업의 핵심 가치와 관련해 전략적 통찰력을 가져다 주는 데이터 분석을 내재화 하는 것이 어려움
- 일차적인 분석 vs 전략 도출을 위한 가치 기반 분석
  - 일차적 분석을 통해서도 부서나 업무 영역에서 상당한 효과를 얻을 수 있음
  - 일차적 분석 경험이 증가하고 분석의 활용 범위를 더 넓고 전략적으로 변화시켜야 함

1-20. 대표적인 일차적 분석 애플리케이션 사례

산업과 분석 애플리케이션의 사례

금융서비스	▪ 신용점수 산정, 사기 탐지, 고객 수익성 분석
소매업	▪ 재고 보충, 수요예측
제조업	▪ 맞춤형 상품 개발, 신상품 개발
에너지	▪ 트레이딩, 공급, 수요예측
온라인	▪ 웹 매트릭스, 사이트 설계, 고객 추천

1-21. 데이터 사이언스의 정의

데이터 사이언스의 정의

- 데이터로부터 의미 있는 정보를 추출해내는 학문
- 정형, 반정형, 비정형의 다양한 유형의 데이터를 대상으로 함
- 분석 뿐 아니라 이를 효과적으로 구현하고 전달하는 과정까지 포함한 포괄적 개념
- 데이터 공학, 수학, 통계학, 컴퓨터 공학, 시각화, 해커의 사고방식, 해당 분야의 전문 지식을 종합한 학문 => 총체적(holistic) 접근법을 사용함
- 과학과 인문학의 교차로에 서 있다고 할 수 있음 => 스토리텔링, 커뮤니케이션, 창의력, 직관력 필요

데이터 사이언스의 핵심 구성요소

- IT(Data Management)
- 분석
- 비즈니스 컨설팅

1-22. 데이터 사이언스의 정의

다른 학문과의 차이점

	데이터 사이언스	통계학	데이터 마이닝
분석 대상	정형, 비정형, 반정형 등 다양한 데이터 유형	정형화된 데이터	
분석 방법	분석 + 시각화 + 전달을 포함한 포괄적 개념		분석에 초점
학문 접근	종합적 학문 또는 총체적 접근법		

1-23. 데이터 사이언티스트의 역량

- 가트너(Gartner)가 본 데이터 사이언티스트의 역량

> 데이터 관리, 분석 모델링, 비즈니스 분석, 소프트 스킬 => 공통점은 호기심에서 시작

- 데이터 해커, 애널리스트, 커뮤니케이션, 신뢰받는 어드바이저 등의 조합이라 할 수 있다
- 하드 스킬과 소프트 스킬 능력을 동시에 갖추고 있어야 한다
- 데이터 처리 기술 이외에 사고방식, 비즈니스 이슈에 대한 감각, 고객들에 대한 공감 능력이 필요하다

1-23. 데이터 사이언티스트의 역량

데이터 사이언티스트가 갖춰야 하는 스킬!

하드 스킬

- Machine Learning, Modeling, Data Technical Skill
- 빅데이터에 대한 이론적 지식 : 관련 기법에 대한 이해와 방법론 습득
- 분석 기술에 대한 숙련 : 최적의 분석 설계 및 노하우 축적

소프트 스킬

- 통찰력 있는 분석 : 창의적 사고, 호기심, 논리적 비판
- 설득력 있는 전달 : Storytelling, Visualization
- 다분야 간 협력 : Communication

- 데이터 사이언디스트들은 주로 데이터 처리나 분석 기술과 관련된 ( )만을 요구 받는 것 처럼 보인다.
- 하지만 이러한 ( )은 훌륭한 데이터 사이언티스트가 갖춰야 하는 능력의 절반에 불과하다. 나머지 절반은 통찰력 있는 분석, 설득력 있는 전달, 협력 등 ( )이다

1-23. 데이터 사이언티스트의 역량

데이터 사이언티스트가 효과적 분석모델 개발을 위해 고려해야 하는 사항

- 분석 모델이 예측할 수 없는 위험을 살피기 위해 현실 세계를 돌아보고 분석을 경험과 세상에 대한 통찰력과 함께 활용한다
- 가정들과 현실의 불일치에 대해 끊임 없이 고찰하고 모델의 능력에 대해 항상 의구심을 갖는다
- 분석의 객관성에 의문을 제기하고 분석 모델에 포함된 가정과 해석의 개입 등의 한계를 고려한다
- 모델 범위 바깥의 요인은 판단하지 않는다

- 데이터 사이언티스트에 요구되는 인문학적 사고의 특성과 역할

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

1-25. 최근의 사회경제적 환경의 변화

최근의 사회경제적 환경의 변화 (인문학 열풍의 이유)

- 단순 세계에서 복잡한 세계로의 변화 : 다양성과 각 사회의 정체성, 연결성, 창조성 키워드 대두
- 비즈니스의 중심이 제품생산에서 서비스로 이동 : 고객에게 얼마나 뛰어난 서비스를 제공 여부가 관건
- 경제와 산업의 논리가 생산에서 시장창조로 바뀜 : 무형자산이 중요

데이터 기반 분석의 상관관계, 통계적 분석의 인과관계

- 신속한 의사결정을 원하는 비즈니스에서는 실시간 '상관관계' 분석에서 도출된 인사이트를 바탕으로 수익을 창출할 수 있는 기회가 점점 늘어나고 있음
- '상관관계' 를 통해 특정 현상의 발생 가능성이 포착되고, 그에 상응하는 행동을 하도록 추천되는 일이 점점 늘어날 것
- 데이터 기반의 '상관관계' 분석이 주는 인사이트가 '인과관계'에 의한 미래 예측을 점점 더 압도해 가는 시대가 도래하고 있음

의사 결정 오류

로직(논리) 오류

- 부정확한 가정을 하고 테스트를 하지 않는 것

프로세스 오류

- 결정에서 분석과 통찰력을 고려하지 않은 것
- 데이터 수집이나 분석이 너무 늦어 사용할 수 없게 되는 것
- 대안을 진지하게 고려하지 않은 것

가치 패러다임의 변화

Digitalization → Connection → Agency

데이터 사이언스의 한계와 인문학

- 모든 분석은 가정에 근거함 => 잘못된 분석은 안 좋은 결과를 가져올 수 있음
- 모델의 능력에 대해 항상 의구심을 갖고
- 가정과 현실의 불일치에 대해 계속 고찰하고
- 분석 모델이 예측할 수 없는 위험을 살펴야 함