Adsp Lv1 요약정리

1. 데이터와 정보

- 데이터의 정의
  - 존재적특성: 객관적 사실
  - 당위적 특성: 추론, 예측, 추정의 근거

- 데이터 유형

정성적 데이터	정량적 데이터
- 형태: 언어, 문자	- 형태: 수치, 도형, 기호
- 특징: 비정형 분석에 많은 비용 소모	- 특징: 정형 데이터
- 주관적 내용	- 객관적 내용
- 통계 분석이 어려움	- 통계 분석이 용이

- 지식 경영의 핵심 이슈

* 암묵지
- 형태: 학습, 경험으로 겉으로 드러나지 않음
- 특징: 사회적으로 중요하지만 공유되기 어려움
- 내면화된 지식 -> 조직의 지식으로 공통화

내공표연

*형식지
- 형태: 문서나 메뉴얼, 교과서, 비디오 등
- 특징: 전달과 공유 용이
- 표출화된 지식 -> 개인의 지식으로 연결화

* DIKW 피라미드

1. Data(데이터): 개별 데이터 자체로 중요하지 않은 객관적 사실
(A마트 연필은 100원, B마트 연필은 200원)
2. Information(정보): 데이터 연관관계속 의미 도출
(A마트 연필이 더 싸다)
3. Knowledge(지식): 데이터 가공 간 상관관계속 패턴 인식 및 의미 부여
(상대적으로 저렴한 A마트에서 연필을 사야겠다)
4. Wisdom(지혜)
(A마트가 다른 물품도 B마트보다 저렴할 것이다)

* 데이터베이스 정의와 특징

* 데이터베이스 특징
1. 통합된 데이터: 동일내용 중복X
2. 저장된 데이터: 컴퓨터가 접근할수 있는 저장매체에 저장됨
3. 공용 데이터: 서로다른 사용자가 데이터베이스의 데이터를 공동 이용
4. 변화되는 데이터: 데이터가 변하면서도 항상 현재의 정확한 데이터 유지

* 데이터 유형
1) 정성적 데이터
- 형태: 언어, 문자
- 특징: 저장,검색, 분석에 많은 비용 소모
- 비정형 데이터
2) 정량적 데이터
- 형태: 수치, 도형, 기호
- 정형데이터로 통계 분석에 용이

* 데이터 레이크: 다양한 언어의 Raw데이터를 한 곳에 모아서 저장하는 저장소

* 지식경영 핵심이슈: 암묵지와 형식지의 상호작용
1) 암묵지: 학습과 경험을 통해 개인에게 체화되는 지식
- 상호작용: 내면화 -> 공통화
2) 형식지: 문서나 매뉴얼처럼 표출된 지식
- 상호작용: 표출화 -> 연결화
** 내공표연 (내면화 -> 공통화 -> 표출화 -> 연결화)
* 표출화: 개인에게 축적된 경험을 객관적인 데이터(언어/기호 등)으로 문서나 매체에 저장/가공/분석하는 과정

DIKW피라미드 정보 vs 지식 비교 헷갈리지 말기 (정보는 패턴 인식, 지식은 예측)
- Data(데이터): 타 데이터와의 상관관계가 없는 가공 전 순수한 수치나 기호
Ex) 연필가격은 A마트가 100원, B마트가 200원이다.
- Information(정보): 데이터 가공 및 상관관계 이해를 통한 패턴 인식
Ex) A마트의 연필이 더 싸다.
- Knowledge(지식): 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물
Ex) 더 저렴한 A마트의 연필을 사야겠다.
- Wisdom(지혜): 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 아이디어
Ex) A마트의 다른 상품들도 B마트보다 더 저렴할 것이다.

* 데이터베이스의 활용
- 2000년대 들어서면서 기업 DB구축의 화두: CRM/SCM
1) CRM(Customer Relationship Management)
- 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 정보기술 활용하여 재고를 최적화하는 솔루션
-> CRM과 SCM은 연동되므로 상호 밀접한 관계
오늘날 CRM은 기존 목적은 변화되지 않고, 방법론에서만 변화 중
** 금융/의료/교육부문에서도 다양한 데이터베이스가 활용되고 있음

* 기업내부 데이터베이스
1) OLTP(On-Line Transaction Processing)
- 네트워크상의 여러 이용자가 실시간으로 데이터베이스의 데이터를 갱신하거나 조회하는 처리 방식
Ex) 은행에서 수많은 입출금이 실시간으로 동시에 일어날 때
2)OLAP(On-Line Analytic Processing)
- 정보위주의 처리 분석, 의사 결정 활용에 도움을 주는 기술
- 다차원의 데이터를 대화식으로 분석(대화식이란, 쿼리문으로 db에 원하는 정보를 요청)
Ex) 판매 추이, 구매성향 파악 등

* Data Mining(데이터 마이닝): 대용량 데이터로부터 의미있는 패턴을 찾는 과정

*제조분야
- RTE(실시간 기업, Real Time Enterprise)
: 가트너의 정의; 최신 정보를 사용해 자사의 핵심 비즈니스 프로세스들의 관리와 실행 과정에서 생기는 지연을 제거하여 경쟁하는 기업
- ERP: 경영 자원을 하나의 통합시스템으로 재구축
- BI(Business Intelligence): 기업 의사결정에 활용하는 리포트 중심의 도구
- BA(Business Analytic): 경영 의사결정을 위한 통계적이고 수학적인 분석에 초점을 둔 기법
- EAI: 정보를 중앙 집중적으로 통합 관리 및 사용
- EDW: 다양한 분석 어플리케이션 원천
*유통분야
KMS: 지식관리시스템
RFID: 주파수 이용 ID 식별 시스템

* SCM과 ERP 혼동 주의!!
- SCM은 모든 업체 협력을 바탕으로~~ 재고 솔루션
- ERP: 경영자원 통합 시스템

* 사회기반 구조로서의 데이터베이스
- NEIS, EDI, VAN, CALS, CVO, GIS, LBS, SIM, GPS, ITS등

★ 기업내부와 사회기반 구조의 DB 구분 문제 기출!!

* DB 종류
1세대: 네트워크 DBMS / 계층 DBMS
- 복잡하고 변경이 어려움
2세대: 관계 DBMS(RDBMS)
- DB를 테이블형태로 구성
Ex) 오라클, 액세스, MySQL
★ 3세대: 객체지향 DBMS(ODBMS)
- 멀티미디어 데이터의 확산으로 관계형 데이터 모델 표현의 어려움으로 탄생
- 객체들을 생성하여 계층에서 체계적으로 정리하고, 하위 계층이 상위 계층으로부터 속성과 방법을 물려받을 수 있는 DBMS
3세대: 객체 관계형 DBMS(ORDBMS)
- RDBMS에 ODBMS 장점 선별해서 만든 DBMS
4세대: NOSQL DBMS
- SNS서비스 증가로 비정형데이터가 증가함에 따라 탄생
- 비정형 데이터 전용 DBMS

** [신용평가]는 핀테크 분야에서 빅데이터 활용이 가장 핵심적인 분야
** [클라우드 컴퓨팅]은 빅데이터 분석에 경제적 효과를 제공해준 결정적 기술

★ 빅데이터가 만들어내는 변화
1) 사전처리 -> 사후처리
2) 표본조사 -> 전수조사
3) 질 -> 양
4) 인과관계 -> 상관관계

* 빅데이터에 거는 기대를 표현한 비유
- 산업혁명의 석탄
- 21세기의 원유
- 렌즈
- 플랫폼

* 빅데이터는 규모가 중요하지 않음
* 성과가 높은 기업이라도 분석적 통찰력이 모두 높지는 않음
* 빅데이터의 궁극적 목표: 데이터 분석에 기초한 전략적 분석과 가치 창출

** 빅데이터 활용기법
1) 연관규칙학습
- 어떤 변수간 상관관계가 있는지
EX) 마트에서 상관관계가 높은 상품을 함께 진열(빵과 우유)
2) 유형분석
- 집단을 특성에 따라 분류하고자 할 때
EX) 온라인 수강생들의 특성에 따라 분류
3) 유전 알고리즘
- 최적화 메카니즘
4) 기계학습(ML)
- 훈련데이터로부터 알려진 특성을 활용해 예측
EX) 넷플릭스 영화추천 시스템
5) 회귀분석
- 독립변수를 사용하여 종속변수가 어떻게 변하는지를 보며 관계 파악
EX) 구매자의 나이가 구매 차량 타입에 어떤 영향을 미치는지?
6) 감정분석
- 고객 의견 및 분석
7) 소셜 네트워크 분석

** 빅데이터 위기 요인
1) 사생활 침해
- 통제방안: 개인정보 동의보다는 개인정보 사용자에게 책임을 지움
2) 책임 원칙의 훼손
- 위기요인: 분석 대상이 예측 알고리즘의 희생이 될 수 있고, 잠재적 위험에도 책임을 추궁하여 민주주의 원칙을 훼손할 수 있음
- 통제방안: 기존 책임원칙의 강화
3) 데이터 오용
- 위기요인: 빅데이터는 과거 데이터에 의존하므로, 미래를 항상 옳게 예측할 수 없음, 이를 너무 신뢰하여 손실을 입을 수 있음
- 통제방안: 데이터 알고리즘에 대한 접근권 허용 및 객관적 인증방안 필요 -> 알고리즈미스트 역할 요구

** 개인정보 비식별 기술
1) 데이터 마스킹: 데이터를 익명으로 생성하는 기술 EX) 홍길동; 홍길*
2) 가명처리: 다른 값으로 변경, 난수화 EX) 홍길동; 허균정
3) 총계처리: 데이터 총계 합을 보냄 EX) 전체 등장인물 나이 합: 78세. 평균: 14세
4) 데이터 값 삭제: 개인 식별에 중요하거나 필요 없는 값 삭제
EX) 홍길동, 20세, 울도대 재학 -> 20세, 울도대 제학
5) 데이터 범주화: 범주로 변환 EX) 홍길동, 20세; 홍씨, 20세
-> 이 전체를 포괄적으로 데이터 익명화(Anonymity)라고 함

* 빅데이터 활용 3요소 (인데기)
1) 데이터 2) 기술 3) 인력

* 빅데이터 활용 사례
-구글: Ngram Viewer(책을 디지털화하여 DB검색 서비스), 기존 페이지랭크 알고리즘을 혁신, 실시간 번역 시스템
-넷플릭스: 영화추천 Cinematch 시스템
-월마트: 월마트랩, 고객 구매패턴을 분석해 상품 진열에 활용
-자라: 일일 판매량 실시간 데이터 분석
-아마존: 전자책 데이터 분석하여 저자에게 제공, 추가 구매예상 도서 추천
-라쿠텐: 슈퍼 데이터베이스를 구축해 마케팅 활동 진행

* 데이터 사이언스 영역 (각 예시 구분 문제 기출)
1) 분석적 영역: 수학, 확률 모델, 머신러닝, 분석학 등
2) IT영역: 시그널 프로세싱, 프로그래밍, 데이터 엔지니어링, 데이터 웨어하우스 등
3) 비즈니스 분석 영역: 커뮤니케이션, 시각화, 프레젠테이션, 스토리 텔링 등

* 데이터 사이언티스트 역량
Hard Skill
1) 빅데이터에 대한 이론적 지식 2) 분석 기술에 대한 숙련
Soft Skill
1) 통찰력 있는 분석 2) 설득력 있는 전달 3) 다분야간 협력

* Deep Learning 관련 기법
- RNN, CNN, LSTM, GRUs, AutoEncoder (KNN 아님 주의!!)

* IOT(사물인터넷 기술): 데이터화에 영향을 미치는 기술

* 데이터 웨어하우스: 의사결정에 도움을 주기 위해 기간 시스템의 DB에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 DB
- 시계열성의 특성으로 관리하는 데이터들은 수시적인 갱신이나 변경이 발생가능
- ETL은 주기적으로 내부 및 외부 데이터베이스로부터 정보를 추출하고 정해진 규약에 따라 정보를 변환한 후에 데이터 웨어하우스에 정보를 적재한다.
- 재무, 생산, 운영 등과 같이 특정 조직의 특정 업무 분야에 초점을 맞추어 구축

* 메타데이터: 데이터에 대한 데이터로, 다른 데이터를 설명해주는 데이터

**SQL: DB에 접근할수 있는 DB하부언어, 테이블 단위로 연산 수행
[SQL]은 RDBMS의 데이터 관리를 위해 설계된 특수 목적 프로그래밍 언어로, 챔벌린과 보이스가 개발됨

DML 함수 (기출)

1) AVG: 평균값 반환, 수치형
2) SUM: 총합 반환, 수치형
3) MIN: 최소값 반환, 수치형
4) COUNT: 조건에 맞는 데이터 개수 반환, 수치형/문자형 둘 다 가능
5) STDDEV: 표준편차 반환, 수치형
6) MAX: 최대값 반환, 수치형

활용 예시: SELECT * FROM My_Table WHERE AGE BETWEEN 20 AND 24
Between에 괄호 쳐놓고 단답식 나옴

*SQL 종류

1) DDL: 데이터 정의어, 생성(CREATE)/수정(ALTER)/삭제(DROP) 등
2) DML: 데이터 조작어, 조회(SELECT)/수정(UPDATE)/삭제(INSERT, DELETE) 등
3) DCL: 데이터 제어어, 사용자별 접근 권한 부여 등 보안을 위한 언어

DML 언어가 아닌 것은? 기출 나옴

** [블록체인]은 기존 금융회사의 중앙 집중형 서버에 거래기록을 보관하는 방식에서 벗어나, 거래에 참여하는 모든 사용자에게 거래내역을 보내주며 거래 때마다 이를 대조하는 데이터 위조방지 기술
** [플랫폼 비즈니스 모델]은 다른 이해 관계자들이 보완적인 상품, 서비스를 제공하는 생태계를 구축하고자 하는 상호 의존적인 그룹
# [플랫폼]은 빅데이터 비즈니스 측면에서 "공동 활용의 목적으로 구축된 유무형의 구조물"을 의미하는 빅데이터 기능

* 딥러닝 활용 오픈소스: Theano, Tensor flow, Keras, Torch, caffe

* 데이터베이스 설계절차(요개논물)
- 요구조건 분석 -> 개념적 설계 -> 논리적 설계 -> 물리적 설계

[맵리듀스] 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 발표한 소프트웨어 프레임워크
- 하둡의 계산을 담당, 하둡의 에코시스템이라고도 함
* HDFS: 네트워크에 연결된 기기에 데이터를 저장하는 분산형 파일 시스템
- 하둡의 스토리지를 담당
하둡은 결국, 맵리듀스와 HDFS로 구성된 [플랫폼] 기술

[정보]는 데이터 가공 및 상관관계 간 이해를 통해 패턴을 인식하고 그 의미를 부여한 것

[데이터 웨어하우스] 기업 의사결정 과정 지원을 위한 주제 중심적으로 통합적이며 시간성을 가지는 비휘발성 데이터 집합 (말을 꼬아 어렵게 해놓았지만, 결국 데이터 웨어하우스는 Raw데이터를 "목적"에 맞게 저장한 저장소)
(이와 반대로 데이터 레이크는 Raw 데이터를 그대로 저장한 저장소)

데이터 사이언스는 분석의 "정확성"이 아닌 "통찰력"에 초점을 두고 진행

* 회귀분석 VS 연관성분석
회귀분석: 고객 만족도가 충성도에 어떤 영향을 미치는가?
연관성분석: 맥주를 사는 사람은 치킨도 함께 구매하더라~ 그래서 같이 진열했다~

*Data mashup: 외/내부 여러 데이터 소스들을 이용하여 통합해서 새로운 인사이트를 도출해내는 방법

* 산업별 일차원적인 어플리케이션 분석
1. 금융서비스: 신용점수 산정, 사기 탐지, 프로그램 트레이딩, 고객 수익성분석, 클레임 분석
2. 병원: 가격 책정, 고객 로열티, 수익관리 (병가고수)
3. 에너지: 트레이딩, 공급, 수요 예측 (에트공수)
4. 정부: 사기 탐지, 사례관리, 범죄 방지, 수익 최적화

* 빅데이터 가치 산정이 어려운 이유
1) 데이터 활용 방식: 재사용, 재조합, 다목적용 개발
2) 새로운 가치 창출
3) 분석 기술 발전

** 빅데이터 시대에 가치 패러다임 변화 순서 D → C → A
디지털화(Digitalization) -> 연결(Connection) -> 에이전시(Agency)

** 알고리즈미스트: 통계학이나 비즈니스에 넓고 깊은 지식을 갖추어 알고리즘에 부당함으로 피해 받는 사람들을 구제할 수 있는 능력을 갖춘 직업