# I. 데이터의 이해

1. 데이터 [ 정성적 데이터 : 언어, 문자
정량적 데이터 : 수치, 도형, 기호

2. 지식경영

1) 암묵지 : 학습과 경험을 통해 개인에게 축적된 내면화된 지식 (무형)

공식의 지식으로 공통화

외부에 표출돼 다른 사람에게 공유되기 어려움

2) 형식지 : 문서나 매뉴얼처럼 형상화된 지식

언어, 기호, 숫자로 표출화된 지식

개인의 지식으로 연결화

전달과 공유 용이

## 3. DIKW

1) 데이터 (Data) : 가공하기 전의 순수한 데이터, 객관적인 사실

2) 정보 (Information) : 데이터들 가공, 상관관계간 이해를 통해 패턴 인식 후 의미부여한 데이터, 데이터 간 관계 분석

3) 지식 (Knowledge) : 상호 연결된 패턴 이해해 이를 토대로 예측한 결과물
유의미한 정보를 분류하고 개인적인 경험을 적용해 고유의 지식으로 내재화, 적용

4) 지혜 (Wisdom) : 원리에 대한 깊은 이해를 바탕으로 도출되는 창의적인 아이디어

# 데이터베이스

: 데이터의 가치, 체계적으로 정리된 데이터의 집합

## 1. 데이터베이스의 특징

- ✓ 통합된 데이터 : 중복X
- ✓ 저장된 데이터 : 저장매체에 저장
- ✓ 공용데이터 : 서로 다른 목적, 공동 데이터 이용
- ✓ 변화되는 데이터 : 계속 변화하면서도 항상 현재의 정확한 데이터 유지

## 2. 데이터베이스의 특성

- ✓ 정보의 축적 및 전달 : 기계가독성, 검색가능성, 원격조작성 = 원거리에서도 즉시 온라인으로 이용
- ✓ 정보이용 : 이용자의 정보 요구에 따라 다양한 정보를 정확하고 신속하게 획득 가능
- ✓ 정보관리 : 정보를 체계적으로 축적하고 새로운 내용추가나 갱신이 용이
- ✓ 정보기술 발전 : 하드웨어, 정보 전송을 위한 네트워크 기술등의 발전을 견인할 수 있음
- ✓ 경제/산업

## 3. 데이터베이스의 활용

1) OLTP (Online Transaction Processing) : 단순한 정보의 '수집', 단순자동화, 데이터 갱신 위주

2) OLAP (Online Analytical Processing) : 정보위주의 분석 처리, OLTP에서 처리된 트랜잭션 데이터를 분석해
제품의 판매추이, 구매동향 파악 등을 프로세싱, 데이터 조회 위주
쉽고 빠르게 다차원적인 데이터에 접근해 의사결정에 활용할 수 있는 정보 얻음

3) CRM (Consumer Relationship Management) : 고객관계관리, 고객별 DB 분석해 고객에 대한 이해를 돕고 이를 바탕으로
마케팅 전략 수립

4) SCM (Supply Chain Management) : 공급망관리, 기업이 외부 공급업체와 통합된 정보시스템을 연계해
시간과 비용을 최적화시키기 위한 것

5) ERP (Enterprise Resource Planning) : 전사적 자원관리, 경영자원을 하나의 통합 시스템으로 재구축

6) RTE (Real Time Enterprise) : 회사의 주요 경영정보를 통합관리하는 실시간 기업의 새로운 기업경영시스템
회사 전 부문의 정보를 하나로 통합

7) BI (Business Intelligence) : 기업이 보유하고 있는 데이터를 정리하고 분석해 기업의 의사결정에 활용하는 프로세스

8) EAI (Enterprise Application Integration) : 기업 내 상호연관된 모든 app을 연동해 필요한 정보를 중앙집중적으로 통합, 관리

9) KMS (Knowledge Management System) : 기업경영을 지식이라는 관점에서 새롭게 정의하는 접근 방식

+) 객체지향 DBMS : 멀티미디어 등 복잡한 데이터 구조를 관리하는 DBMS

데이터 웨어하우스 (DW) : 방대한 조직 내 분산된 DBMS를 통합, 운영시간성을 가지는 비휘발성 데이터의 집합

SQL : DB와 통신을 위해 고안된 언어

## II. 데이터의 가치와 미래

**# 빅데이터** : 일반적인 DB SW로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터

→ 3V : Volume (양), Variety (다양성), Velocity (속도)

### 1. 기능

- 산업혁명의 석탄과 철: 생산성 확기적으로 끌어올림
- 리테기의 원유: 기존에 없던 새로운 범주의 산업 만들어냄
- 렌즈: 데이터가 산업 전반에 영향을 미침
- 플랫폼: 그 자체로 플랫폼 역할

### 2. 빅데이터가 만들어낸 변화

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

## 3. 위기요인과 통제방안

- 사생활 침해 : 동의에서 책임으로
- 책임원칙 훼손: 결과기반 책임원칙 고수
- 데이터 오용: 알고리즘 접근허용

## 4. 데이터 활용의 3요소

1) 데이터 : 모든것의 데이터화, 제조업에서 서비스업으로
2) 기술: 알고리즘, 인공지능
3) 인력: 데이터사이언티스트, 알고리즈미스트
   - 데이터사이언티스트: 빅데이터의 가치를 창출하기위해 필요
   - 알고리즈미스트: 데이터사이언티스트가 한 일로 인해 부당하게 피해가 발생하는 것을 막기위해 필요

## 5. 빅데이터 활용 기본 테크닉

- 연관규칙 학습 · 상관관계
- 유형분석: 분류
- 유전 알고리즘: 최적화, 점진적으로 진화
- 기계 학습: 훈련데이터로 부터 학습한 알려진 특성을 활용해 '예측'
- 회귀분석: 두 변인 (독립변수, 종속변수)의 관계 파악
- 소셜 네트워크 (사회관계망) 분석 : 오피니언리더 (영향력 있는 사람)을 찾아낼 수 있음
  고객들 간 관계파악
- 감정분석

# Ⅲ. 가치 창조를 위한 데이터 사이언스와 전략 인사이트

## # 데이터사이언스

### 1. 데이터사이언스

- 과학과 인문의 교차로
- 데이터로부터 의미있는 정보를 추출(발견)하고 효과적으로 구현하고 전달
- 정형 / 비정형의 다양한 데이터를 대상
- 총체적 접근법
- 전략적 통찰 ~ Soft Skill

### 2. 데이터사이언티스트의 역량

- 강력한 호기심
- 인문학적 통찰에 근거한 합리적 추론
- Analytics (분석) & IT 전문성 & 비즈니스 캔버싱 (커뮤니케이션, 스토리텔링, 시각화)
- Hard Skill + Soft Skill

1) Hard Skill : 빅데이터에 대한 이론적 지식, 분석기술에 대한 숙련

2) Soft Skill : 통찰력있는 분석, 선구력있는 전달, 다분야간 협력

창의적사고, 호기심 스토리텔링, 시각화 커뮤니케이션

3. 인문학 열풍의 이유

1) Convergence → Divergence : 표준화 / 이성화 → 다양성 / 감성성 / 창조성

2) 제품생산 → 서비스 : 효율경제 → 체험경제

3) 생산 → 시장창조 : 공급자중심의 가격경쟁 → 암묵적이고 함축적 지식인 무형자산, 산출은 → 창조 과정

4. 가치 패러다임의 변화

1) 1단계 : 디지털화 (Digitalization) - 가치를 형상화, 표준화 / 아날로그 세상을 어떻게 효과적으로 디지털화하는가

2) 2단계 : 연결 (Connection) - 다양한 디지털 정보를 필요한 사람에게 연결해서 효과적이고 효율적으로 정보 연결 + 제공

3) 3단계. 에이전시 (Agency) - 개인과 기기, 사물에 이르는 방대한 정보를 하이퍼 연결을 통해 필요한 정보를 효과적으로 제공하고 관리할 수 있는 시대로 발전

5 한계

- 인간의 해석이 개입 (사람에 따라 다른 해석과 결론)
- 모든 분석은 가정에 근거

## # Data 관련 기술

### 1. 개인정보 비식별 기술

1) **데이터 마스킹** : 데이터 속성 유자한 채, 새롭고 읽기 쉬운 데이터로 익명으로 생성(데이터 변조)
개인의 사생활 침해 방지, 응답자의 비밀사항 보호하며 통계자료의 유용성은 충분한 확보

2) **가명처리**

3) **총계처리**

4) **데이터 값 삭제**

5) **데이터 범주화**

### 2. 무결성과 레이크

1) **데이터 무결성** : DB내의 데이터에 대한 정확성, 일관성, 유효성, 신뢰성을 보장하기 위해서 데이터 변경/수정 시 여러가지 제한을 둠
데이터의 정확성 보증

- 개체 무결성, 참조 무결성, 범위 무결성

2) **데이터 레이크** : 수 많은 정보 속에서 의미 있는 내용을 찾기 위해 방식에 상관없이 데이터 저장

# I. 데이터 처리 프로세스

## # ETL : Extraction, Transformation, Load · 데이터 이동과 변환

- Extraction (추출): 데이터 획득
- Transformation (변형): 데이터 클렌징, 품질화, 통합 또는 비즈니스 룰 적용
- Loading (적재): 변형처리가 완료된 데이터를 목표 시스템에 적재

; 데이터 웨어하우스 (DW), 운영 데이터 스토어 (ODS), 데이터 마트 (DM) 에 대한 데이터 적재작업의 핵심 구성요소
데이터 통합, 데이터 이동, 마스터 데이터 관리에 걸쳐 활용
데이터 비정규화: 수능 향상을 위해 테이블을 대신 합치는 것

## - ETL 작업단계

1) Interface : 데이터 획득 위한 인터페이스 메커니즘 구현
2) Staging ETL: 획득된 데이터를 스테이징 테이블에 저장
3) Profiling ETL: 스테이징 테이블에서 데이터 특성 식별 데이터 품질 측정
4) Cleansing ETL: 데이터 보정
5) Integration ETL: 데이터 충돌을 해소하고 클렌징된 데이터 통합
6) Demoralizing ETL: 운영 보고서 생성, 데이터 웨어하우스 (DW) / 데이터마트 (DM) 에 데이터를 적재하기 위해 데이터 비정규화 수행

# ODS : Operational Data Store : 데이터에 추가작업을 위해 다양한 데이터 원천으로부터의 데이터를 추출 / 통합한 DB

- ODS 를 위한 데이터 통합은 데이터 클렌징 / 중복제거 / 비즈니스 룰 데비 무결성 점검 등의 작업 포함
- 일반적으로 원자성 (개별성) 지닌 하위 수준 데이터 ( 실시간 근접 트랜잭션, 가격 등) 를 저장하기 위해 설계

## - ODS 구성단계

1) Interface : 데이터 획득
2) Staging
   - 획득한 데이터를 스테이징 테이블에 저장
   - 통제 정보 추가 : 적재 타임스탬프, 데이터 값에 대한 Check Sum
   - 일괄 (Batch) 작업 형태의 정기적인 ETL과 실시간 데이터 획득 방식을 혼용하여 구성할 수 O
3) Profiling
   - 범위, 도메인, 유일성 확보 등의 규칙을 기준으로 데이터 특성 식별 & 데이터 품질점검
   - 선행 조건에 따라 데이터 프로파일링 요건 설정 → 데이터 프로파일링 수행 → 결과 통계 처리 → 품질보고서 생성 및 유지
4) Cleansing
5) Integration : 정제 완료된 데이터를 ODS 내의 단일 통합 테이블에 저장
6) Export : 통합된 데이터에 익스포트 규칙과 보안규칙을 반영해 테이블을 생성한 후, DBMS, DW, DM에 적재

# # 데이터 웨어하우스 (DW)

## 1. 데이터 웨어하우스의 특징

- 주제중심: 업무항목 기준으로 구조화
- 영속성: 읽기 전용, 삭제 x
- 통합성: 대부분 운영 시스템에 의해 생성된 데이터들의 통합본
- 시계열성: 시간순

## 2. 스타스키마 & 스노우 플레이크 스키마

### 1) 스타스키마 = 조인스키마

제3정규형으로 모델링 / 비정규화된 제2정규형으로 모델링

- 단일 사실 테이블을 중심으로 다수의 차원 테이블로 구성
- 전통적인 관계형 DB를 통해 다차원 DB 기능구현

### 2)

| | 스타스키마 | vs | 스노우 플레이크 스키마 |
|---|---|---|---|
| 장점 | DW 스키마중 가장 단순<br>복잡도 낮아 이해하기 쉬움<br>쿼리 작성 용이, 조인테이블 개수↓ | | 데이터 중복제거<br>→ 시간 단축 |
| 단점 | 차원테이블의 비정규화에 따른<br>데이터 중복으로 해당 테이블에<br>데이터 적재시 많은 시간 소요 | | 스키마 구조의 복잡성↑<br>→ 조인테이블 개수↑, 쿼리 난이도↑ |
| 차원테이블 | 비정규화된 제2정규형 | | 제3정규형 정규화 |

# CDC : Change Data Capture : DB내 데이터에 대한 변경을 식별해 필요한 후속처리를 자동화

- Push: Source Data에서 변경을 식별 → Target에 변경 데이터를 적재
- Pull: Target에서 Source Data를 정기적으로 살펴봐 필요시 데이터 다운로드

1. Time Stamp on Rows : 마지막 변경 타임스탬프 값보다 더 최근의 타임스탬프 값을 갖는 레코드를 변경된 것으로 식별
2. Version Numbers on Rows: 버전을 기록하는 컬럼을 두고 식별된 레코드 버전보다 더 높은 버전을 보유한 레코드를 변경된 것으로 식별
   레코드들의 최신버전은 기록/관리하는 참조테이블을 항상 운용
3. Status on Rows : 타임스탬프나 버전기법에 대한 보완 목적, 사람이 직접 결정 (True / False), 업무규칙 적용 가능
4. Time / Version / Status on Rows: 세개의 모두 활용, 정교한 쿼리 생성에 활용해 개발유연성 제공
5. Triggers on Rows: DB 트리거 활용해 등록된 다수 대상 시스템에 변경 데이터 배포
   → 트리거 영향: 시스템 관리 복잡도 증가, 변경관리 어려워짐, 확장성↓, 전반적 유지보수성↓
6. Event programming: CDC를 어플리케이션에 구현해 다양한 조건에 의한 CDC 메커니즘 구현, 어플리케이션 개발 부담과 복잡도↑
7. Log Scanner on Database: DB의 트랜잭션 로그 스캐닝 및 변경 내역 해석 활용
   영향도 최소화: DB & DB 사용 app & 트랜잭션 무결성
   변경 식별 지연시간 최소화
   DB 스키마 변경 불필요
   DBMS 에 따라 트랜잭션 로그 관리 메커니즘이 상이해 다수의 DB를 사용하는 환경에서 적용시 작업규모 증가