CHAPTER 01

데이터베이스 기본 개념

Database Fundamental Concepts

Lecture Note

박상돈 조교수

대전대학교 컴퓨터공학과

목 차

01. 데이터베이스의 필요성

데이터베이스를 배우기 전에, 가장 기본적인 질문에서 출발한다. "데이터(Data)"와 "정보(Information)"라는 단어는 일상에서 자주 사용되지만, 이 두 개념을 명확하게 구분할 수 있는 사람은 의외로 많지 않다. 바로 이 두 개념의 차이를 이해하는 것이 데이터베이스 학습의 출발점이다.

1.1 데이터와 정보

(1) 데이터 (Data)

데이터(Data)란, 현실 세계에서 단순히 관찰하거나 수집한 사실(Fact)이나 값(Value)을 말한다. 핵심은 아직 가공되지 않은 원시(Raw) 상태의 자료라는 점이며, 데이터 그 자체만으로는 의미를 가지기 어렵다. 숫자, 문자, 기호 등 다양한 형태를 취할 수 있다.

예시: "서울"이라는 단어 하나만 봤을 때, 이것이 지역 이름인지, 카페 이름인지, 아이돌 그룹 이름인지 맥락 없이는 알 수 없다. "35"라는 숫자도 나이인지, 기온인지, 시험 점수인지, 몸무게인지 단독으로는 전혀 알 수 없다. "2025-08-01"이라는 날짜 역시 누구의 생일인지, 어떤 행사 날짜인지, 계약 만료일인지 그 자체로는 판단할 수 없다. 이처럼 맥락이 없는 날것 그대로의 상태가 바로 데이터이다.

(2) 정보 (Information)

정보(Information)란, 데이터를 의사 결정에 유용하게 활용할 수 있도록 처리(가공)하고 조직화한 결과물이다. 쉽게 말하면, 데이터에 맥락(Context)과 의미(Meaning)를 부여한 것이 정보이다. 정보는 의사결정에 직접 활용할 수 있으며, 분석 및 가공 과정을 통해 생성된다.

예시: 앞서 예를 든 데이터들 — "서울", "35도", "2025년 8월 1일", "맑음" — 을 합치고 분석하면, "2025년 8월 1일 서울 기온이 35도이고 맑음. 기상청 기준 35도 이상이므로 폭염 경보 발령이 필요"라는 정보가 도출된다. 기상청 담당자는 이 정보를 바탕으로 폭염 경보 발령 여부를 의사결정할 수 있다.

핵심 개념

데이터는 원재료이고, 정보는 가공된 제품이다.

데이터 → 처리(가공) → 정보

비유: 닭(데이터) → 조리(처리) → 치킨(정보)

편의점 알바 비유: 매일 팔린 상품 이름과 수량(삼각김밥 30개, 컵라면 15개, 아이스크림 50개 등)은 데이터이다. 이 데이터가 수 개월치 쌓이면 분석을 통해 패턴을 발견할 수 있다. 예를 들어, "화요일에 삼각김밥이 유독 잘 팔린다(근처 학교 급식 없는 날)", "금요일 저녁에 맥주 판매량이 급증한다(불금 효과)", "여름에는 아이스크림 재고를 2배로 늘려야 한다" 같은 결론이 정보이다. 이 정보를 바탕으로 점장은 발주 수량을 조절하고, 재고를 관리하며, 알바 스케줄을 편성할 수 있다.

시험 출제 포인트

서술형: "데이터와 정보의 차이를 서술하시오"

객관식: "다음 중 데이터에 해당하는 것은?"

단답형: "데이터에서 정보가 만들어지는 과정을 무엇이라 하는가?" → 정보 처리

1.2 정보 처리 (Information Processing)

정보 처리(Information Processing)의 정의는 다음과 같다. 데이터(Data)에서 정보(Information)를 추출하는 과정 또는 방법으로, 데이터를 상황에 맞게 분석하고 해석하여 데이터 간의 의미 관계를 파악하는 것이다. 핵심 키워드는 "분석", "해석", "의미 관계 파악"이다.

정보 처리의 흐름

데이터(Data) → 처리·가공(분석/해석/변환) → 정보(Information)

정보 처리의 구체적 예시

예시 1 — 기상 데이터: 데이터 "서울", "35도", "2025년 8월 1일", "맑음"을 합치고 분석하면, "35도는 기상청 기준 폭염 경보에 해당하므로 발령이 필요하고, 서울 시민에게 외출 자제를 권고하고, 독거노인 가정 안전 확인을 지시하고, 무더위 쉼터를 추가 개방해야 한다"는 정보 및 후속 의사결정이 가능해진다.

예시 2 — 수강 신청 데이터: 데이터베이스시스템 수강 신청자 45명, 강의실 정원 40명, 신청 시간 9시~9시 3초. 이를 처리하면, "정원 5명 초과이므로 5명에게 대기번호를 부여하고, 수요가 높으니 추가 분반 개설 또는 정원 증원을 검토해야 한다"는 정보가 도출된다.

예시 3 — 카페 매출 데이터: 일주일간 아메리카노 판매 기록(월 120잔, 화 115잔, 수 130잔, 목 125잔, 금 180잔, 토 200잔, 일 190잔)을 분석하면, 평일 평균 134잔, 주말 평균 195잔으로 주말 매출이 평일 대비 약 45% 높다는 정보를 얻는다. 이를 바탕으로 "금요일부터 원두 발주량 1.5배 증가, 주말 아르바이트 1명 추가 배치, 월·화 할인 이벤트 실시" 등의 의사결정이 가능하다.

데이터베이스의 존재 이유는 바로 이 정보 처리를 효율적으로 수행하기 위해서이다. 데이터가 아무리 많아도 가공하지 않으면 쓸모가 없다. 100만 건의 고객 거래 데이터가 하드디스크에 쌓여 있는데 아무도 분석하지 않는다면, 그것은 디스크 용량만 차지하는 쓸모없는 파일에 불과하다. 데이터를 체계적으로 저장하고, 필요할 때 빠르게 꺼내서 가공할 수 있다면 고객 행동 패턴 파악, 매출 예측, 마케팅 전략 수립 등이 가능해진다. 이것을 가능하게 해주는 도구가 바로 데이터베이스이다.

1.3 정보 시스템과 데이터베이스

(1) 정보 시스템 (Information System)

정보 시스템(Information System)이란, 조직 운영에 필요한 데이터를 수집·저장·가공·분배하는 체계적인 시스템이다. 핵심 단계는 "수집 → 저장 → 가공 → 분배"의 네 단계이다.

기업 환경을 예로 들면, 영업부에서는 고객 주문·매출 데이터, 생산부에서는 제품 생산량·불량률 데이터, 인사부에서는 직원 급여·근태 데이터, 재무부에서는 수입·지출·세금 데이터, 마케팅부에서는 광고 효과·고객 반응 데이터가 발생한다. 이런 데이터가 각 부서에서 따로 관리되면 조직 전체의 큰 그림을 볼 수 없다. 모든 부서의 데이터를 한 곳에 모아 저장하고, 필요할 때 가공하여 각 부서에 유용한 정보로 제공하는 전체적인 체계가 정보 시스템이다.

(2) 정보 시스템의 종류

구분	설명	역할

MIS (경영 정보 시스템)	Management Information System	기업의 경영 관리에 필요한 의사결정 정보를 제공하는 시스템. 재무, 인사, 생산, 마케팅 등 다양한 분야에서 활용
DSS (의사결정 지원 시스템)	Decision Support System	MIS에서 한 단계 더 나아가, 데이터를 깊이 분석하여 의사결정의 방향까지 제시하는 시스템

MIS 예시: "이번 분기 매출이 전 분기 대비 20% 증가", "A 제품 재고가 안전 수준 이하로 하락", "이번 달 직원 이직률이 5% 초과"

DSS 예시: "A 지역 20대 여성 고객이 급증 중이므로, 해당 세그먼트 타겟 마케팅을 강화하면 매출을 15% 더 늘릴 수 있음"

(3) 데이터베이스 계층 구조: DB vs. DBMS vs. DBS

MIS든 DSS든, 모든 정보 시스템의 기반에는 데이터베이스가 있다. 데이터가 없으면 정보도 없고, 정보가 없으면 의사결정도 불가능하기 때문이다. 여기서 반드시 구분해야 하는 세 가지 용어가 있다.

용어	정의	비유
DB (Database)	데이터의 저장소. 데이터가 실제로 저장된 곳.	창고 (물건이 쌓여 있는 곳)
DBMS (Database Management System)	DB에 저장된 데이터를 관리하는 소프트웨어. 데이터의 추가, 수정, 삭제, 검색 등의 작업을 수행.	창고 관리인 (물건 정리, 검색, 반출 담당)
DBS (Database System)	DB + DBMS + 사용자 + 응용 프로그램을 합친 전체 체계.	창고 + 관리인 + 이용자 + 물류 시스템

핵심 구분 — MySQL은 무엇인가?

MySQL은 DBMS이다. 데이터베이스 자체가 아니라, 데이터베이스를 관리하는 소프트웨어이다.

DB = MySQL이 관리하는 실제 데이터가 저장된 곳

DBMS = MySQL 소프트웨어 자체

DBS = MySQL + 데이터 + 사용자 + 응용 프로그램 전체

대표적인 DBMS: Oracle, MySQL, PostgreSQL (2장에서 상세 학습)

섹션 1 핵심 요약

데이터에서 정보를 만들어야 하는데, 그러려면 데이터를 체계적으로 저장하고 관리하는 도구가 필요하다. 그것이 데이터베이스이다.

02. 데이터베이스의 정의와 특징

이 섹션은 1장에서 시험 출제 비중이 가장 높은 핵심 파트이다. 객관식, 단답형, 서술형 등 어떤 형태로든 출제될 수 있다.

2.1 데이터베이스(DB)의 정의

> **데이터베이스의 교과서적 정의**
> "특정 조직의 여러 사용자가 공유하여 사용할 수 있도록 통합해서 저장한 운영 데이터의 집합"

이 한 문장 안에 네 개의 핵심 키워드가 포함되어 있다. 하나씩 살펴보자.

키워드 1: 공유 데이터 (Shared Data)

특정 조직의 여러 사용자가 함께 소유하고 이용할 수 있는 공용 데이터이다. 한 사람만 쓰는 데이터가 아니라, 여러 사용자가 같이 쓰는 데이터를 의미한다.

예시: 대학교의 학생 데이터(학번, 이름, 학과, 성적 등)는 교무처(성적 관리, 졸업 심사), 학생처(장학금 심사), 도서관(대출 기록), 총무처(등록금 확인), 취업지원센터(취업 통계) 등 최소 5개 이상의 부서가 공유하여 사용한다. 모든 부서가 하나의 학생 데이터베이스를 공유하기 때문에, 각 부서가 따로 학생 명단을 관리할 필요가 없고, 데이터 불일치 문제도 방지된다.

키워드 2: 통합 데이터 (Integrated Data)

데이터의 중복을 최소화한 통합된 데이터 집합이다. 불필요한 중복은 제거하되, 성능상의 이유

로 통제 가능한 중복은 일부 허용할 수 있다.

핵심: "완전히 제거"가 아니라 "최소화"이다! 이 차이가 매우 중요하다. DBA(데이터베이스 관리자)가 의도적으로 속도 향상을 위해 만든 통제된 중복은 괜찮다. 문제가 되는 것은 같은 데이터가 여기저기에 중복 저장되어 불일치(Inconsistency)가 발생하는 불필요한 중복이다.

예시: 학생 김철수의 전화번호가 교무처 시스템에는 "010-1234-5678"로, 학생처 시스템에는 "010-9999-8888"로 되어 있다면, 어느 것이 맞는지 알 수 없다. 이러한 데이터 불일치 문제를 해결하기 위해 통합 데이터베이스가 등장했다.

함정 문제 주의!

"통합 데이터에서 중복은 완전히 제거된다. (O/X)" → 정답: X

중복을 '최소화'하는 것이지 '완전히 제거'하는 것이 아니다!

키워드 3: 저장 데이터 (Stored Data)

컴퓨터가 접근할 수 있는 저장 매체(디스크, SSD 등)에 저장된 데이터이다. 물리적으로 영구 보존이 가능해야 한다. 사람의 머릿속에만 있는 지식이나 종이 서류장에만 있는 데이터는 데이터베이스라고 부를 수 없다. 데이터베이스는 반드시 컴퓨터가 읽고 쓸 수 있는 디지털 형태로 저장되어야 한다.

키워드 4: 운영 데이터 (Operational Data)

조직의 주요 기능을 수행하기 위해 지속적으로 유지·관리해야 하는 필수 데이터이다. 임시 데이

터(테스트용 더미 데이터, 일시적 계산 중간 결과, 임시 메모 등)는 운영 데이터에 해당하지 않는다.

운영 데이터의 예: 학생 정보, 교수 정보, 강의 시간표, 성적, 등록금 납부 내역

운영 데이터가 아닌 예: 시스템 테스트용 가짜 학생 데이터, 임시 보고서 초안, 연습 테이블

> **암기법 — 공통저운(共統貯運)**
>
> 공유(共) + 통합(統) + 저장(貯) + 운영(運)
>
> "공통적으로 저장하고 운영한다"로 기억
>
> 시험: "DB의 정의에 포함되는 4가지 키워드를 쓰시오" → 공유, 통합, 저장, 운영

2.2 데이터베이스의 4 대 특징

특징 1: 실시간 접근 (Real-Time Accessibility)

사용자의 데이터 요구에 실시간으로 빠르게 응답할 수 있어야 한다. 질의(Query)에 대해 즉각적인 결과를 반환하며, 수 초 이내의 응답 시간 보장이 필수적이다. 온라인 트랜잭션 처리(OLTP: Online Transaction Processing) 환경에서 핵심적인 요구사항이다.

실생활 예시: 은행 ATM에서 잔액 조회 시 30분 뒤에 결과가 나온다면 아무도 사용하지 않을 것이다. 수강 신청에서 버튼을 눌렀는데 결과가 1시간 후에 나온다면 다른 과목도 이미 마감되어 있을 것이다. 인터넷 쇼핑에서 검색 결과가 1분 후에 나온다면 고객은 이탈한다. 아마존의 연구에 따르면, 페이지 로딩 시간이 0.1초만 늘어나도 매출이 1% 감소한다고 한다.

특징 2: 계속 변화 (Continuous Evolution)

데이터베이스는 정적(Static)이 아니라 동적(Dynamic)이다. 새로운 데이터가 들어오고, 기존 데이터가 수정되고, 불필요한 데이터가 삭제되면서, 항상 최신의 정확한 데이터 상태를 유지해야 한다. 이러한 데이터 변화는 CRUD라는 네 가지 기본 연산으로 이루어진다.

CRUD — 소프트웨어의 기본 데이터 처리 연산

C (Create): 새로운 데이터 삽입. 예) 신입생 입학 시 학생 테이블에 새 레코드 추가

R (Read): 데이터 검색·조회. 예) 특정 학생의 성적 조회

U (Update): 기존 데이터 수정. 예) 학생 이사 후 주소 변경

D (Delete): 불필요한 데이터 삭제. 예) 폐강된 강좌 삭제

모든 소프트웨어의 데이터 처리는 결국 이 네 가지의 조합이다. 웹 개발, 앱 개발, 게임 개발 등 어떤 분야든 데이터를 다루는 모든 작업은 Create, Read, Update, Delete 중 하나에 해당한다.

CRUD	SQL	REST API (HTTP)
C (Create)	INSERT	POST
R (Read)	SELECT	GET
U (Update)	UPDATE	PUT
D (Delete)	DELETE	DELETE

특징 3: 동시 공유 (Concurrent Sharing)

여러 사용자가 동시에 같은 데이터 또는 서로 다른 데이터를 사용할 수 있어야 한다. 데이터의

정확성을 보장하면서 동시 접근을 허용하는 기술을 동시성 제어(Concurrency Control)라고 하며, 트랜잭션 격리 수준(Isolation Level)을 통해 충돌을 방지한다. 이 주제는 13주차 10장 "회복과 병행 제어"에서 상세히 학습한다.

실생활 예시: 수강 신청 시 잔여석이 1석인데 두 명이 동시에 신청 버튼을 누른 경우, 반드시 한 명만 성공하고 다른 한 명은 "마감" 처리되어야 한다. 콘서트 티켓 예매에서도 마지막 1장에 두 명이 거의 동시에 결제를 시도하면, 한 명만 성공해야 한다. 두 명 다 결제되면 좌석은 1개인데 티켓이 2장 발행되는 대참사가 벌어진다.

특징 4: 내용으로 참조 (Content Reference)

데이터의 물리적 주소(디스크의 몇 번째 섹터, 몇 번째 블록)가 아니라, 데이터의 내용(Content)으로 검색한다. 사용자가 찾고자 하는 조건을 제시하면 DBMS가 해당 데이터를 자동으로 찾아 반환한다.

비유: 파일 시스템 방식(옛날 방식)은 "3층 A구역, 4번째 선반, 왼쪽에서 7번째 책 주세요"처럼 물리적 위치를 정확히 알아야 한다. 데이터베이스 방식은 "컴퓨터 과학 관련 책 중 2020년 이후 출판된 것 전부 찾아주세요"처럼 원하는 조건만 말하면 된다.

SQL 예시: `SELECT * FROM 학생 WHERE 학과 = '컴퓨터공학과'`

→ "학생 테이블에서 학과가 컴퓨터공학과인 데이터를 전부 찾아라." 사용자는 'What(무엇)'만 명시하고, 'How(어떻게 찾을지)'는 DBMS가 처리한다. 이를 선언적(Declarative) 방식이라 부른다.

암기법 — 실계동내(實繼同內)

실시간 접근 + 계속 변화 + 동시 공유 + 내용 참조

"실시간으로 계속 동시에 내용으로"

시험: "DB의 4가지 특징을 쓰고 각각 설명하시오" → 실계동내를 떠올리고 풀이

섹션 2 핵심 요약

DB의 정의 4대 키워드: 공유, 통합, 저장, 운영 (공통저운)

DB의 4대 특징: 실시간 접근, 계속 변화, 동시 공유, 내용 참조 (실계동내)

CRUD = Create(삽입), Read(조회), Update(수정), Delete(삭제) — 모든 SW의 기본 연산

03. 데이터 과학 시대의 데이터

데이터에도 여러 종류가 있다. 모든 데이터가 같은 형태가 아니며, 분류 기준은 크게 두 가지이다. 형태(구조화 정도)에 따른 분류와 특성(질적/양적)에 따른 분류이다.

3.1 형태에 따른 데이터 분류

데이터를 구조화 정도에 따라 분류하면, 정형 데이터, 반정형 데이터, 비정형 데이터의 세 가지로 나뉜다.

구분	정형 데이터 (Structured)	반정형 데이터 (Semi-Structured)	비정형 데이터 (Unstructured)
정의	미리 정해진 스키마에 따라 행과 열로 구성된 데이터	고정 스키마 없이 데이터 내부에 구조(메타데이터)가 포함된 데이터	정해진 구조가 없는 자유 형태의 데이터
구조화 정도	높음	중간	없음
대표 예시	RDBMS, CSV, Excel	HTML, XML, JSON, 센서 데이터	텍스트, 이미지, 영상, 음성, PDF
대표 시스템	Oracle, MySQL, PostgreSQL	MongoDB 등 NoSQL	Hadoop, Spark
전체 비중	약 20%	-	약 80% 이상

(1) 정형 데이터 (Structured Data)

미리 정해진 구조(Schema)에 따라 저장된 데이터로, 행(Row)과 열(Column)로 구성된다. 스키마(Schema)란 데이터의 구조를 미리 정의한 것으로, 예를 들어 테이블 생성 시 "학번은 정수형 8자리, 이름은 문자열 최대 20자"처럼 지정하는 것이다.

정형 데이터 예시 — 학생 정보 테이블:

학번	이름	학과	학년	평점
20210001	김철수	컴퓨터공학	3	3.8
20210002	이영희	전자공학	2	4.2
20210003	박민준	컴퓨터공학	4	3.5

정형 데이터의 5가지 특징: (1) 고정된 스키마 보유 — 구조가 미리 정해져 일관성 보장 (2) SQL로 검색·조작 가능 — 9주차, 10주차에 집중 학습 (3) 데이터 무결성 보장 — 스키마에 맞지 않는 데이터는 입력 거부 (4) 관계형 데이터베이스에서 관리 (5) 트랜잭션 지원 — 오류 시 원래 상태로 복구 가능

(2) 반정형 데이터 (Semi-Structured Data)

고정된 스키마가 없지만, 데이터 내부에 구조(메타데이터)에 대한 설명이 함께 존재하는 데이터이다. 완전히 표 형태는 아니지만 구조가 아예 없지도 않은 중간 형태로, 태그(Tag)나 키(Key)를 통해 데이터 스스로 "나는 이런 구조야"라고 설명한다.

XML 예시:

```
<학생목록>
  <학생>
    <학번>20210001</학번>
    <이름>김철수</이름>
    <학과>컴퓨터공학</학과>
  </학생>
</학생목록>
```

JSON 예시:

```
{
  "학생": [{
    "학번": "20210001",
    "이름": "김철수",
    "학과": "컴퓨터공학"
  }]
}
```

정형 vs. 반정형의 핵심 차이: 정형 데이터는 스키마가 외부에 미리 정의(CREATE TABLE로 구조 선언)되고, 반정형 데이터는 스키마가 데이터 안에 내포되어 있다. 반정형은 구조를 유연하게 변경할 수 있어(학생 A에는 전화번호 필드가 있고 학생 B에는 없어도 OK) 웹 서비스, 모바일 앱, IoT 센서 등에서 널리 사용된다.

시험 출제 포인트

"반정형 데이터의 예를 3가지 쓰시오" → HTML, XML, JSON

"JSON은 어떤 유형의 데이터인가?" → 반정형 데이터

(3) 비정형 데이터 (Unstructured Data)

정해진 구조가 전혀 없는 자유 형태의 데이터이다. 전체 데이터의 약 80% 이상을 차지하며 가장 빠르게 증가하고 있다.

유형	예시	규모
텍스트	SNS 게시물, 이메일, 뉴스 기사, 리뷰, 웹 문서	-
이미지	사진, 의료 영상(X-ray, CT), 위성 이미지	전 세계 하루 약 14억 장 촬영
영상/음성	유튜브 동영상, 팟캐스트, CCTV 영상, 음악	유튜브 1분마다 500시간 분량 업로드
문서 파일	PDF, Word, PowerPoint	-

비정형 데이터의 급증이 빅데이터 기술(Hadoop, Spark), NoSQL 데이터베이스(MongoDB, Cassandra), AI/딥러닝 기반 분석 기술의 발전을 견인했다. 이 수업에서는 주로 정형 데이터를 다루는 관계형 데이터베이스를 학습하지만, 실제 세계에서는 비정형 데이터가 압도적으로 많다는 큰 그림을 반드시 인지해야 한다.

시험 출제 포인트

"PDF 파일은 어떤 유형인가?" → 비정형 데이터

"엑셀 파일은 어떤 유형인가?" → 정형 데이터

"전체 데이터에서 비정형 데이터가 차지하는 비중은?" → 약 80% 이상

3.2 특성에 따른 데이터 분류

데이터의 특성에 따라 크게 범주형 데이터(Categorical Data)와 수치형 데이터(Numerical Data)로 나뉘며, 각각은 다시 두 가지 하위 유형으로 세분된다. 이 분류는 통계학에서 유래했지만, 데이터베이스 설계에서도 적절한 데이터 타입과 제약 조건 설정을 위해 매우 중요하다.

대분류	소분류	특징	대표 예시

범주형 (Categorical)	명목형 (Nominal)	서열(순서)이 없는 값. 단순히 범주를 구분하는 용도. 크기 비교·산술 연산 불가.	혈액형, MBTI, 성별, 국적, 전공
범주형 (Categorical)	순서형 (Ordinal)	서열(순서)이 있는 값. 순서 비교는 가능하지만 값 간 간격이 불균등. 산술 연산 부적합.	학년, 학점(A~F), 회원등급, 만족도(상/중/하)
수치형 (Numerical)	이산형 (Discrete)	단절된(불연속적) 숫자 값. 주로 개수(Count)를 세는 데이터. 정수 값.	판매량, 합격자 수, 교통사고 건수, 수강 인원
수치형 (Numerical)	연속형 (Continuous)	연속적 숫자 값(실수). 측정(Measurement)을 통해 얻는 데이터. 두 값 사이에 무한한 중간값 존재.	키, 몸무게, 온도, 속도, 주가

(1) 범주형 데이터 (Categorical / Qualitative Data)

종류를 나타내는 값을 가진 데이터로, 질적(Qualitative) 데이터라고도 한다.

명목형 (Nominal): 서열이 없는 값이다. A형이 B형보다 높다거나, INTJ가 ENFP보다 크다는 개념은 없다. 값 간 크기 비교, 순서 비교, 산술 연산 모두 의미가 없다. 할 수 있는 것은 빈도(Frequency)를 세는 것 정도이다(예: 우리 반 A형 12명, B형 8명).

순서형 (Ordinal): 서열이 있는 값이다. 학년(1<2<3<4), 학점(A>B>C>D>F), 회원 등급(골드>실버>브론즈) 등은 순서 비교가 가능하다. 그러나 A 학점과 B 학점의 실력 차이가 B와 C의 차이와 같다고 보장할 수 없으므로, 값 간 간격이 불균등하다. 산술 연산도 엄밀히는 부적합하다(GPA 평균은 편의상 계산할 뿐 수학적으로 정확하지 않다).

범주형 데이터의 핵심

범주형 데이터는 연산의 대상이 아니다.

"1학년 + 2학년 = ?" → 의미 없는 연산 (3학년이 되는 것이 아님)

(2) 수치형 데이터 (Numerical / Quantitative Data)

크기 비교와 산술 연산이 가능한 데이터로, 양적(Quantitative) 데이터라고도 한다. 범주형과 달리 평균, 합계, 표준편차 등의 통계 분석에 직접 활용할 수 있다.

이산형 (Discrete): 단절된 숫자 값이다. "셀 수 있는 데이터"로 기억하면 된다(Count). 아이스크림을 150.5개 팔 수는 없고, 합격자가 42.7명일 수는 없다. 항상 정수 값을 취하며, 값 사이에 중간값이 존재하지 않는다. 판매량, 합격자 수, 교통사고 건수, 가족 수, 수강 인원 등이 해당한다.

연속형 (Continuous): 연속적 숫자 값이다. "잴 수 있는 데이터"로 기억하면 된다(Measure). 키 172cm와 173cm 사이에 172.1, 172.15, 172.153... 등 무한한 중간값이 존재한다. 소수점 이하 값이 의미를 가지며, 측정을 통해 얻는다. 키, 몸무게, 온도, 속도, 주가 등이 해당한다.

이산형 vs. 연속형 판별 기준 (절대 안 헷갈리는 방법)

셀 수 있으면 이산형, 잴 수 있으면 연속형.

"몇 개?" → 이산형 | "얼마나?" → 연속형

판매량(몇 개 팔았어? → 이산형) vs. 키(얼마나 커? → 연속형)

연습 문제

다음 데이터의 유형을 각각 분류하시오.

데이터	유형	이유
MBTI	명목형 (범주형)	INTJ, ENFP 등 사이에 순서가 없음
학점 (A~F)	순서형 (범주형)	A>B>C>D>F로 순서가 있으나 간격 불균등
수강 인원	이산형 (수치형)	42.5명은 불가능, 셀 수 있는 정수 값
체온	연속형 (수치형)	36.5도와 36.6도 사이에 무한한 값, 잴 수 있음
성별	명목형 (범주형)	남/여 사이에 순서 개념 없음
회원 등급	순서형 (범주형)	골드>실버>브론즈로 서열 존재
일일 판매량	이산형 (수치형)	개수를 세는 데이터, 정수 값
몸무게	연속형 (수치형)	측정값, 소수점 이하 의미 있음

1 장 최종 요약

섹션 1: 데이터베이스의 필요성

데이터는 원재료, 정보는 가공된 제품이다. 데이터 → 처리(정보 처리) → 정보의 흐름을 기억하라.

정보 시스템의 기반에 데이터베이스가 있으며, DB(저장소), DBMS(관리 소프트웨어), DBS(전체 시스템)는 서로 다른 개념이다.

섹션 2: 데이터베이스의 정의와 특징

DB의 정의 4대 키워드: 공유, 통합, 저장, 운영 (암기법: **공통저운**)

주의: 통합 데이터의 중복은 '완전 제거'가 아니라 '최소화'이다.

DB의 4대 특징: 실시간 접근, 계속 변화, 동시 공유, 내용 참조 (암기법: **실계동내**)

CRUD: Create(삽입), Read(조회), Update(수정), Delete(삭제) — 모든 소프트웨어의 기본 데이터 처리 연산

섹션 3: 데이터 과학 시대의 데이터

형태별 분류: 정형(표 형태, RDBMS) / 반정형(XML, JSON, 데이터 안에 구조 내포) / 비정형(이미지, 영상, 텍스트 등, 전체의 80% 이상)

특성별 분류: 범주형(명목형: 서열 없음 / 순서형: 서열 있음, 간격 불균등) / 수치형(이산형: 셀 수 있음 / 연속형: 잴 수 있음)

다음 시간 예고 — 2장: 데이터베이스 관리 시스템(DBMS)

파일 시스템의 문제점(데이터 중복, 데이터 종속성 등)

DBMS가 이 문제를 어떻게 해결하는지

DBMS의 장단점과 발전 과정

복습 필수 항목 (다음 시간까지 반드시 암기)

DB의 정의 4대 키워드: 공유, 통합, 저장, 운영 (공통저운)

DB의 4대 특징: 실시간 접근, 계속 변화, 동시 공유, 내용 참조 (실계동내)