Normalization

이상현상(Anomly)

- 삭제 이상 : 튜플 삭제 시 같이 저장된 다른 정보까지 연쇄적으로 삭제되는 현상
- 삽입 이상 : 튜플 삽입 시 특정 속성에 해당하는 값이 없어 NULL을 입력해야 하는 현상
- 수정 이상 : 튜플 수정 시 중복된 데이터의 일부만 수정되어 일어나는 데이터 불일치 현상

학생번호	학생이름	학과	주소	강좌이름	강의실
501	박지성	컴퓨터과	영국 맨체스타	데이터베이스	공학관 110
401	김연아	체육학과	대한민국 서울	데이터베이스	공학관 110
402	장미란	체육학과	대한민국 강원도	스포츠경영학	체육관 103
502	추신수	컴퓨터과	미국 클리블랜드	자료구조	공학관 111
501	박지성	컴퓨터과	영국 맨체스타	자료구조	공학관 111

- 삭제이상은 만약 장미란 학생의 정보를 지우면 강의실 103도 같이 사라져서 다른 튜플들이 강의실 103을 사용하지 못하는 경우에 발생
- 삽입 이상은 튜플을 삽입하는 경우에 해당하는 정보가 없어 NULL을 넣는 현상
- 수정이상은 만약 박지성과 김연아가 데이터베이스 수업을 강의실 110에서 수강할때 박지성의 강의실을 201로 바꾸어도 김연아는 110으로 그대로 유지되고, 같은 수업임에도 강의실이 달라지는 현상이 발생

이는 서로 공유하는 데이터임에도 각자의 튜플에 독립적으로 존재하기 때문에 발생함

따라서 테이블을 분리하여 그 테이블을 통해 강의 제목이나 강의실을 참고하게 하면 이상 현상을 해결할 수 있음

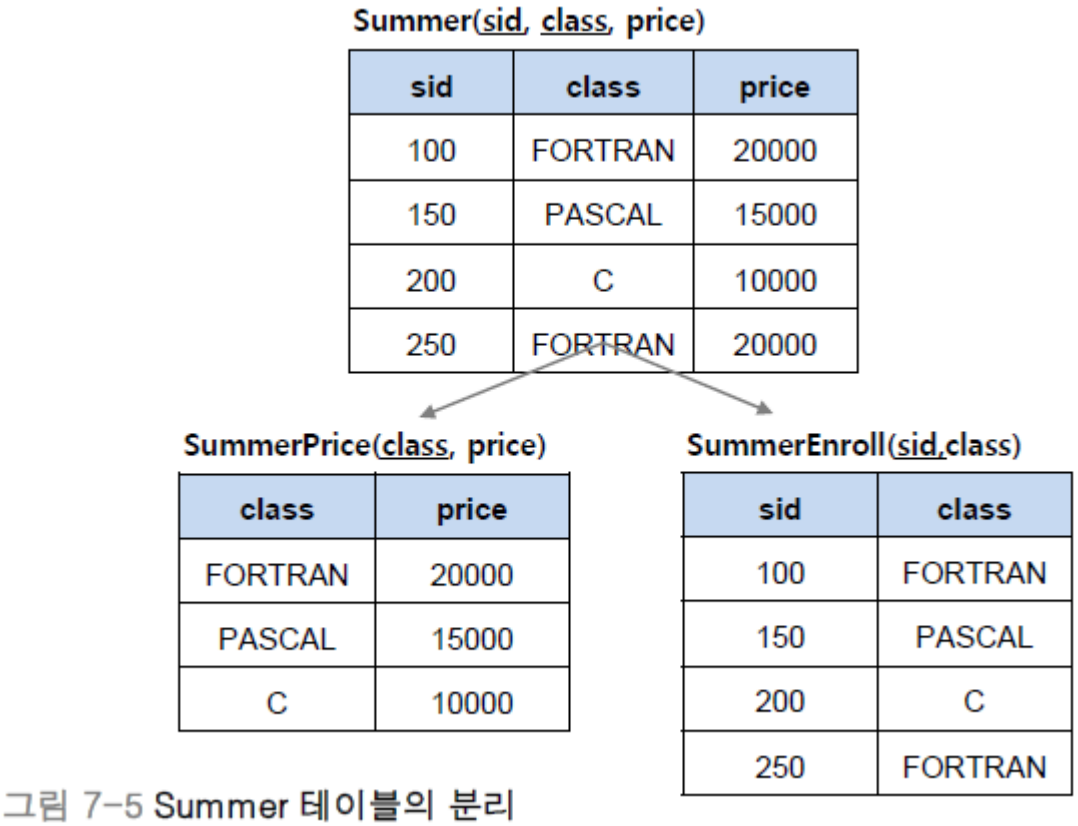


그림 7-5 Summer 테이블의 분리

함수 종속성

- 어떤 속성 A의 값을 알면 다른 속성 B의 값이 유일하게 정해지는 관계를 종속성이라 함
- A→ B로 표기하며 A를 B의 결정자라고 함

학생수강성적

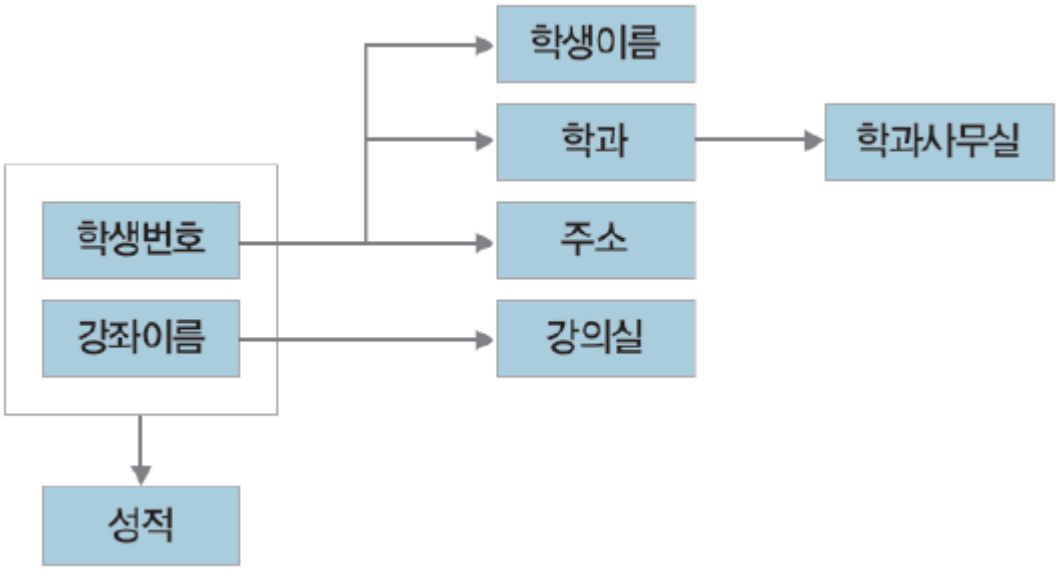
학생번호	학생이름	주소	학과	학과사무실	강좌이름	강의실	성적
501	박지성	영국 맨체스타	컴퓨터과	공학관101	데이터베이스	공학관 110	3.5
401	김연아	대한민국 서울	체육학과	체육관101	데이터베이스	공학관 110	4.0
402	장미란	대한민국 강원도	체육학과	체육관101	스포츠경영학	체육관 103	3.5
502	추신수	미국 클리블랜드	컴퓨터과	공학관101	자료구조	공학관 111	4.0
501	박지성	영국 맨체스타	컴퓨터과	공학관101	자료구조	공학관 111	3.5

- 학생과 수강, 성적의 속성에는 의존성이 존재한다.

- 여기서 의존성은 학생 번호를 알면 학생 이름이 정해지는 관계이다.
- 속성 A의 값을 알면 다른 속성 B의 값이 유일하게 정해지는 의존관계를 속성 B는 속성 A에 종속한다. 또는 속성 A는 속성 B를 결정한다라고 말한다.

함수 종속성 다이어그램

- 릴레이션의 속성: 직사각형
- 속성 간의 함수 종속성: 화살표
- 복합 속성: 직사각형으로 묶어서 그림



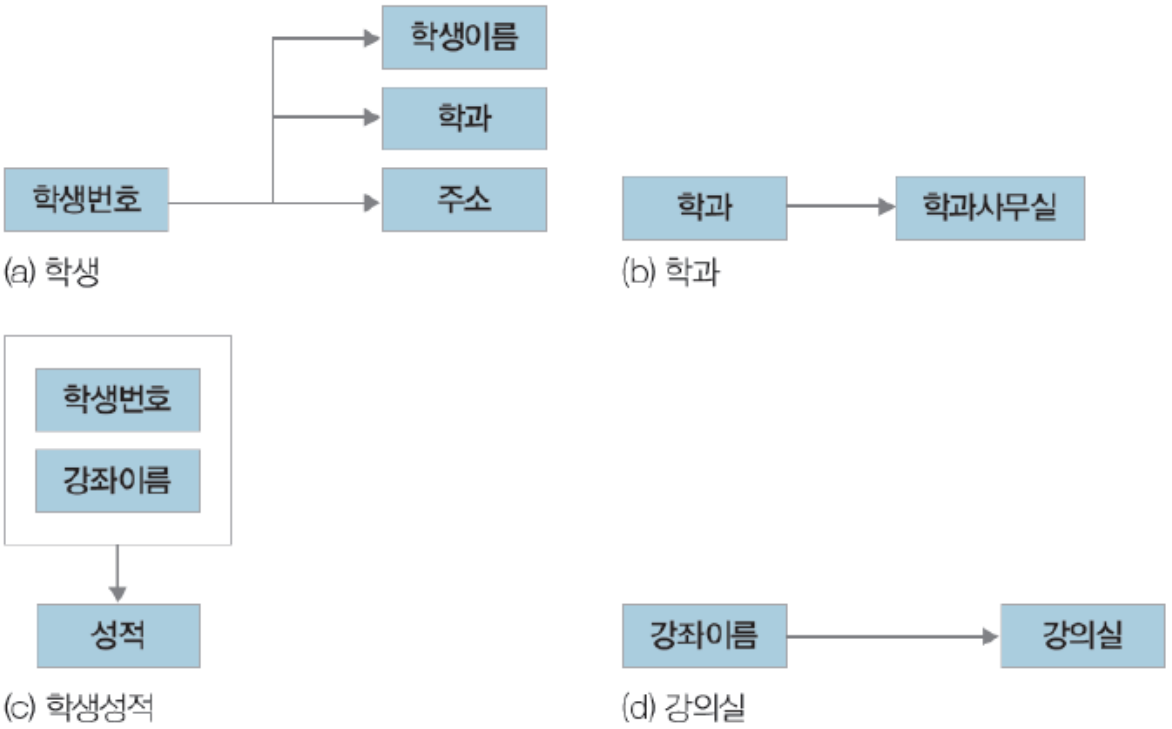
- 학생 번호가 학생이름, 학과, 주소를 결정함
- 복합속성의 경우 그 속성들을 묶어 하나의 직사각형으로 표시한다.

함수 종속성 규칙

적용 규칙	사례	설명
부분집합 규칙 if Y ⊆ X, then X → Y	(학과, 주소) → 학과	학과는 (학과, 주소)의 부분집합 속성이므로, '(학과, 주소) → 학과' 성립
증가 규칙 If X → Y, then XZ → YZ	(학생번호, 강좌이름) → (학생이름, 강좌이름)	'학생번호 → 학생이름'이므로 강좌이름을 추가하여, '(학생번호, 강좌이름) → (학생이름, 강좌이름)' 성립
이행 규칙 : If X → Y and Y → Z, then X → Z	학생번호 → 학과사무실	'학생번호 → 학과', '학과 → 학과사무실'이므로 이행 규칙을 적용하여, '학생번호 → 학과사무실' 성립
결합 규칙 If X → Y and X → Z, then X → YZ	학생번호 → (학생이름, 주소)	'학생번호 → 학생이름', '학생번호 → 주소'이므로 결합 규칙을 적용하여, '학생번호 → (학생이름, 주소)' 성립
분해 규칙 If X → YZ, then X → Y and X → Z	학생번호 → 학생이름, 학생번호 → 주소	'학생번호 → (학생이름, 주소)'이므로 분해하여, '학생번호 → 학생이름', '학생번호 → 주소' 성립
유사이행 규칙 If X → Y and WY → Z, then WX → Z	(강좌이름, 학생이름) → 성적	'학생이름 → 학생번호'(학생이름이 같은 경우가 없다고 가정한다), '(강좌이름, 학생번호) → 성적'이므로 유사이행 규칙을 적용하여, '(강좌이름, 학생이름) → 성적' 성립

이상현상과 결정자

이상 현상은 한 개의 릴레이션에 두 개 이상의 속성이 포함되어 있고 기본키가 아닌 속성이 결정자일 때 발생한다. 이상현상을 해결하기 위해 릴레이션을 분해하면 된다.



정규화

정규화의 기본 목표는 테이블 간에 중복된 데이터를 허용하지 않는다는 것이다.

이를 통해 무결성을 유지할 수 있으며, DB의 저장 용량 역시 줄일 수 있다.

[제1 정규화]

제 1 정규화는 테이블의 컬럼이 원자값을 갖도록 테이블을 분해하는 것이다.

고객취미들(이름, 취미들)

이름	취미들
김연아	인터넷
추신수	영화, 음악
박세리	음악, 쇼핑
장미란	음악
박지성	게임

고객취미(이름, 취미)

이름	취미
김연아	인터넷
추신수	영화
추신수	음악
박세리	음악
박세리	쇼핑
장미란	음악
박지성	게임

개발자	자격증
홍길동	정보처리기사
홍길동	빅데이터 분석기사
장길산	정보보안기사

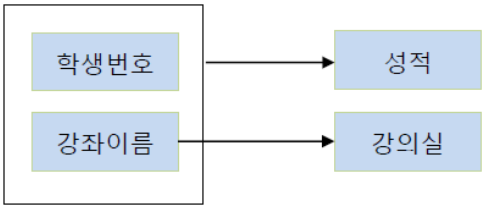
개발자	언어
홍길동	C
홍길동	C++
장길산	JAVA

[제2 정규화]

제 2 정규화는 제 1 정규화를 진행한 테이블에 대해 완전 함수 종속을 만족하도록 테이블을 분해하는 것이다. (기본키의 부분 집합이 결정자가 되어선 안된다는 것)

수강강좌

학생번호	강좌이름	강의실	성적
501	데이터베이스	공학관 110	3.5
401	데이터베이스	공학관 110	4.0
402	스포츠경영학	체육관 103	3.5
502	자료구조	공학관 111	4.0
501	자료구조	공학관 111	3.5



현재 기본키는 (학생번호, 강좌이름)으로 복합 키이다.

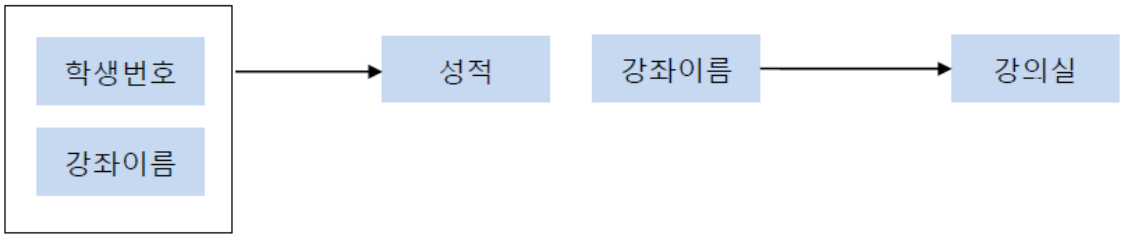
여기서 (강좌이름) —→ (강의실) 와 같이 기본키의 부분키인 강좌이름이 결정자 역할을 할 수 있다.

수강

학생번호	강좌이름	성적
501	데이터베이스	3.5
401	데이터베이스	4.0
402	스포츠경영학	3.5
502	자료구조	4.0
501	자료구조	3.5

강의실

강좌이름	강의실
데이터베이스	공학관 110
스포츠경영학	체육관 103
자료구조	공학관 111



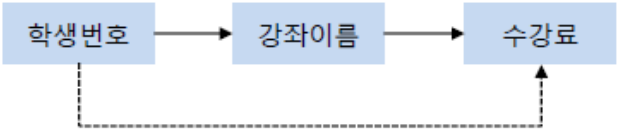
[제3 정규화]

제 3 정규형은 릴레이션 R이 제 2 정규형을 만족하고 기본키에 비이행적으로 종속할때를 의미한다.

이행적 종속은 A → B, B → C 이면 A → C가 성립되는 함수 종속성이다.

계절학기

학생번호	강좌이름	수강료
501	데이터베이스	20000
401	데이터베이스	20000
402	스포츠경영학	15000
502	자료구조	25000



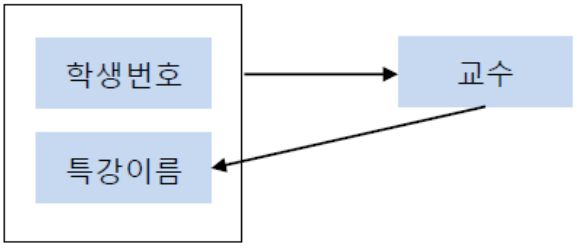
계절학기 릴레이션에서 501 학생의 강좌이름을 스포츠경영학으로 변경하면 수강료가 15000으로 되어야한다. 따라서 속성들을 독립적으로 만드는 것이 아니라 학생 번호로 수강료를 참조할 수 있게 만들면 된다.

[BCNF 정규화]

제 3 정규화를 진행한 테이블에 대해 모든 결정자가 후보키가 되도록 테이블을 분해하는 것이다.

특강수강

학생번호	특강이름	교수
501	소셜네트워크	김교수
401	소셜네트워크	김교수
402	인간과 동물	승교수
502	창업전략	박교수
501	창업전략	홍교수



특강수강 테이블에서 기본키는 (학생번호, 특강이름)이다. 그리고 기본키는 교수를 결정하고 있다.

또한 교수는 특강이름을 결정하고 있다.

문제점은 교수가 특강이름을 결정하는 결정자이지만, 후보키가 아니라는 점이다.

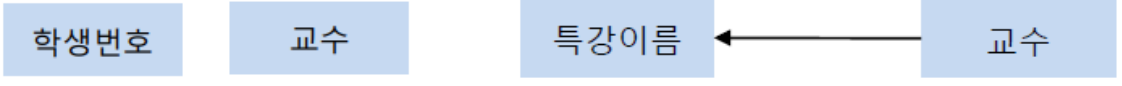
따라서 특강 신청 테이블과 특강교수 테이블로 분해할 수 있다.

특강신청

학생번호	교수
501	김교수
401	김교수
402	승교수
502	박교수
501	홍교수

특강교수

특강이름	교수
소셜네트워크	김교수
인간과 동물	승교수
창업전략	박교수
창업전략	홍교수



[제 4 정규화]

제 4 정규화는 BCNF를 만족해야한다. 또한 다중값 종속이 없어야한다.

다치 종속은 같은 테이블 내의 독립적인 두 개 이상의 컬럼이 또 다른 컬럼에 종속되는 것을 말한다.

A → B인 의존성에서 단일 값 A와 다중 값 B가 존재한다면 다치 종속이라고 한다. A ↠ B 로 표기

개발자	자격증	언어
홍길동	정보처리기사	C
홍길동	빅데이터 분석기사	C++
장길산	정보보안기사	JAVA

제 4 정규화를 통해 분리할 수 있다.

개발자	자격증
홍길동	정보처리기사
홍길동	빅데이터 분석기사
장길산	정보보안기사

개발자	언어
홍길동	C
홍길동	C++
장길산	JAVA

[제 5 정규화]

제 5 정규화는 4 정규화를 만족해야한다. 더 이상 비손실 분해를 할 수 없어야한다.

조인 종속: 하나의 릴레이션을 여러개의 릴레이션으로 분해하였다가, 다시 조인했을 때 데이터 손실이 없고 필요없는 데이터가 생기는 것을 말한다. 조인 종속성은 다치 종속의 개념을 더 일반화한 것 이다.

개발자	자격증	언어
홍길동	정보처리기사	C
홍길동	빅데이터 분석기사	C
홍길동	정보처리기사	C++
홍길동	빅데이터 분석기사	C++
장길산	정보보안기사	JAVA

제 4 정규화 테이블에 대해 조인 연산을 수행한 결과이다.

데이터 손실은 없지만 필요없는 데이터가 추가적으로 생겼기에 제 5 정규화를 만족하지 않는다.

개발자	자격증
홍길동	정보처리기사
홍길동	빅데이터 분석기사
장길산	정보보안기사

자격증	언어
정보처리기사	C
빅데이터 분석기사	C++
정보보안기사	JAVA

개발자	언어
홍길동	C
홍길동	C++
장길산	JAVA

제 5 정규화를 적용한 테이블이다.

반정규화(Denormalization)

> 하나 이상의 테이블에 데이터를 중복해 배치하는 최적화 기법이다.

시스템의 성능 향상, 개발 및 운영의 편의성 등을 위해 정규화된 데이터 모델을 통합, 중복, 분리하는 과정으로, **의도적으로 정규화 원칙을 위배하는 행위**이다.

비정규화는 다른 타협안을 내놓음으로써 그런 단점을 해소하고자 한다. 어느 정도의 데이터 중복이나 그로 인해 발생하는 데이터 갱신 비용은 감수하는 대신 조인 횟수를 줄여 한층 효율적인 쿼리를 날릴 수 있도록 하겠다는 것이다.

장점

- 빠른 데이터 조회

  → 조인 비용이 줄어들기 때문

- 살펴볼 테이블이 줄어들기 때문에 데이터 조회 쿼리가 간단해짐

  → 따라서 버그 발생 가능성도 줄어든다

단점

- 데이터 갱신이나 삽입 비용이 높음
- 데이터 갱신 또는 삽입 코드를 작성하기 어려워짐
- *데이터 간의 일관성*이 깨어질 수 있다. 어느 쪽이 올바른 값인가?
- 데이터를 중복하여 저장하므로 더 많은 저장 공간이 필요

비정규화 대상

1. 자주 사용되는 테이블에 액세스하는 프로세스의 수가 가장 많고, 항상 일정한 범위만을 조회하는 경우
2. 테이블에 대량 데이터가 있고 대량의 범위를 자주 처리하는 경우, 성능 상 이슈가 있을 경우
3. 테이블에 지나치게 조인을 많이 사용하게 되어 데이터를 조회하는 것이 기술적으로 어려울 경우

대부분의 대규모 IT 업체의 경우처럼, 규모 확장성(scalability)을 요구하는 시스템의 경우 거의 항상 정규화된 데이터베이스와 비정규화된 데이터베이스를 섞어 사용한다.

주의점

- 반정규화를 과도하게 적용하다 보면 데이터의 무결성이 깨질 수 있다.
- 입력, 수정, 삭제의 질의문에 대한 응답 시간이 늦어질 수 있다.