데이터 종속성과 정규화

이 장의 주요 내용

- 데이터의 잘못된 논리적 표현으로 인해 발생하는 이상 현상들
- 함수 종속성
- 정규화
  - 제 1 정규형, 제 2 정규형, 제 3 정규형, BCNF
  - 제 4 정규형, 제 5 정규형
- 참고 문헌
  - 데이타베이스 시스템, 이석호 저, 정익사(chapter 11장), 2005년

데이타의 논리적 표현

- 조직체가 가지고 있는 대량의 운용 데이터를 어떻게 조직해야 효율적으로 관리할 수 있는가?
  - 논리적 데이타베이스 설계 문제
    - 관계 모델을 이용하여 어떻게 실세계를 정확히 표현할 것인가?

- 관계 스키마 설계
  - 필요한 애트리뷰트, 엔티티, 관계성을 식별하고 수집
  - 위의 요소들을 릴레이션으로 만듦
    - 고려해야 할 사항
      - 데이타 종속성(data dependancy) : 애트리뷰트들간의 관계성
      - 효율적인 데이타 처리
      - 데이터의 일관성 유지 등

- 설계가 잘못되면 실제 데이터를 처리할 때 부작용이 발생하기 쉬움

이상 현상의 예 – 1/3

- 이상 현상
  - 릴레이션을 처리 할 때 곤란한 현상
- 예) 수강 릴레이션

수강

학번	과목번호	성적	학년
100	C413	A	4
100	E412	A	4
200	C123	B	3
300	C312	A	1
300	C324	C	1
300	C413	A	1
400	C312	A	4
400	C324	A	4
400	C413	B	4
400	C412	C	4
500	C312	B	2
600			2

삭제 이상 발생

갱신 이상 발생
(학년 4 → 3)

삽입 이상 발생

기본키 : {학번, 과목 번호}

이상 현상의 예 – 2/3

- 삭제이상 (deletion anomaly)
  - 예) 200번 학생이 'C123'의 등록을 취소할 경우
    - 과목번호가 기본키 값의 일부분이므로, 투플 전체를 삭제해야 함
    - 3학년이라는 정보도 함께 삭제됨
  - 정의
    - 한 튜플을 삭제함으로써 유지해야 될 정보까지도 삭제되는 연쇄 삭제(triggered deletion)에 의한 정보의 손실이 발생하게 되는 현상

- 삽입이상 (insertion anomaly)
  - 예) 600번 학생이 2학년이라는 사실만을 삽입하려고 할 경우
    - {학번과 과목번호}가 기본 키이므로, 어떤 과목을 등록하지 않는 한 삽입이 불가능
    - 만일 삽입하고자 할 경우, 가상의 임시과목 번호를 함께 삽입해야 함
  - 정의
    - 원하지 않는 데이터도 함께 삽입해야만 되고 그렇지 않으면 삽입이 되지 않는 현상

이상 현상의 예 – 3/3

- 갱신이상 (update anomaly)
  - 예) 학번이 400번 학생의 학년을 4에서 3으로 변경시키고자 할 경우
    - 학번이 400인 4개의 투플에 대해 학년의 값을 모두를 갱신시켜야 함
    - 만일 일부 투플만 변경시킨다면, 학번 400인 학생의 학년은 3과 4, 즉 두 가지 값을 가지게 됨
  - 정의
    - 중복된 투플들 중에서 일부 투플의 애트리뷰트 값만을 갱신시킴으로써 정보의 비일관성(inconsistency)이 생기는 현상

이상의 원인과 해결책

- 이상의 원인
  - 여러 가지 상이한 종류의 정보를 하나의 릴레이션으로 표현하려 하기 때문
  - 즉, 애트리뷰트들 간에 존재하는 여러 가지 데이터 종속 관계를 무리하게 하나의 릴레이션에 표현하려는 데서 발생

- 해결 방안
  - 애트리뷰트들 간의 종속성(dependency)를 분석하여 하나의 릴레이션에는 기본적으로 하나의 종속성이 표현되도록 분해함(decomposition)
  - 이러한 분해 과정을 정규화(normalization)라고 함

스키마 변환

- 일단 만들어진 릴레이션들을 바람직한 형태의 릴레이션들로 다시 변환하는 것

- 스키마 변환의 원리
  - 정보의 무손실
  - 최소의 데이터 중복
  - 분리의 원칙
    - 하나의 독립된 관계성은 하나의 릴레이션으로 분리시켜 표현

함수 종속

- FD : Functional Dependency
  - 애트리뷰트들의 그룹핑을 릴레이션 스키마로 바꾼 것의 적절성을 정형적으로 측정하기 위한 도구
  - 애트리뷰트들의 두 개의 집합 사이의 제약 조건
  - 데이터의 의미를 표현

- 정의
  - 어떤 릴레이션 R에서 X와 Y를 각각 R의 애트리뷰트라고 하자. 애트리뷰트 X의 값 각각에 대해 항상 애트리뷰트 Y의 값이 오직 하나만 연관되어 있을 때
  - 애트리뷰트 Y는 애트리뷰트 X에 함수 종속 $X \rightarrow Y$이라고 함
  - X : 결정자, Y : 종속자

함수 종속 예

- 예) 학생 릴레이션

학번	이름	학년	학과

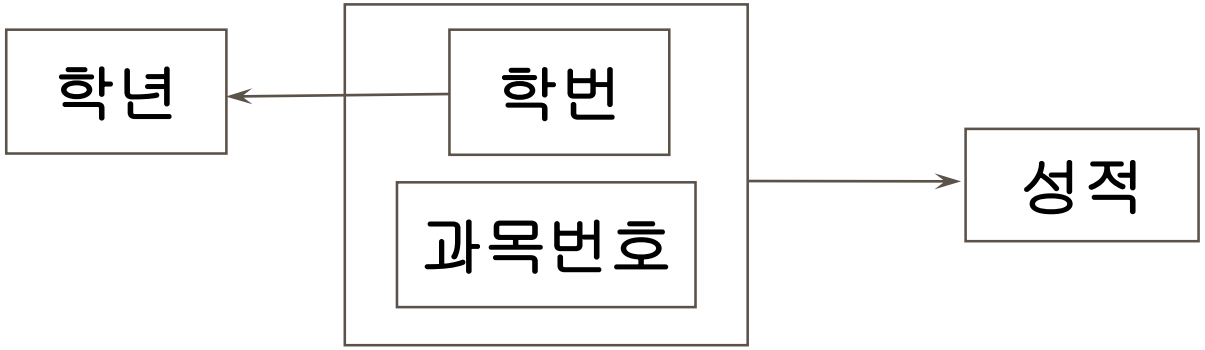
  - 학번 → 이름, 학번 → 학년, 학번 → 학과
  - 이유 : 어떤 학생의 학번이 정해지면, 그 학번에 대응하는 이름, 학년, 학과의 값은 오직 하나만 있기 때문

함수 종속 다이어그램

- FD Diagram
  - 한 릴레이션에서 애트리뷰트들 간의 복잡한 함수 종속 관계를 쉽게 이해하기 위해 도식으로 표현
- 예) 수강 릴레이션

학번	과목번호	성적	학년

  - 함수 종속
    - {학번, 과목 번호} → 성적 : 완전 함수 종속
    - {학번, 과목 번호} → 학년 : 부분 함수 종속
    - 학번 → 학년 : 완전 함수 종속



완전 함수 종속과 부분 함수 종속

- 복합 애트리뷰트 X에 대하여 $X \rightarrow Y$가 성립할 때
- 완전 함수 종속 (full functional dependency)
  - $X' \subset X$ 이고 $X' \rightarrow Y$ 를 만족하는 애트리뷰트 X'이 존재하지 않음
- 부분 함수 종속 (partial functional dependency)
  - $X' \subset X$ 이고 $X' \rightarrow Y$ 를 만족하는 애트리뷰트 X'이 존재함

함수 종속에 대한 추론 규칙

- 어떤 릴레이션 R에 존재하는 함수 종속에 대해 다음과 같은 추론 규칙이 성립됨
  - R1: (반사, reflexive) $A \supseteq B$이면 $A \rightarrow B$이다.
  - R2: (첨가, augmentation) $A \rightarrow B$이면 $AC \rightarrow BC$이고 $AC \rightarrow B$이다
  - R3: (이행, transitive) $A \rightarrow B$ 이고 $B \rightarrow C$이면 $A \rightarrow C$이다.
  - R4: (분해, decomposition) $A \rightarrow BC$이면 $A \rightarrow B$이다.
  - R5: (결합, union) $A \rightarrow B$이고 $A \rightarrow C$이면 $A \rightarrow BC$이다.

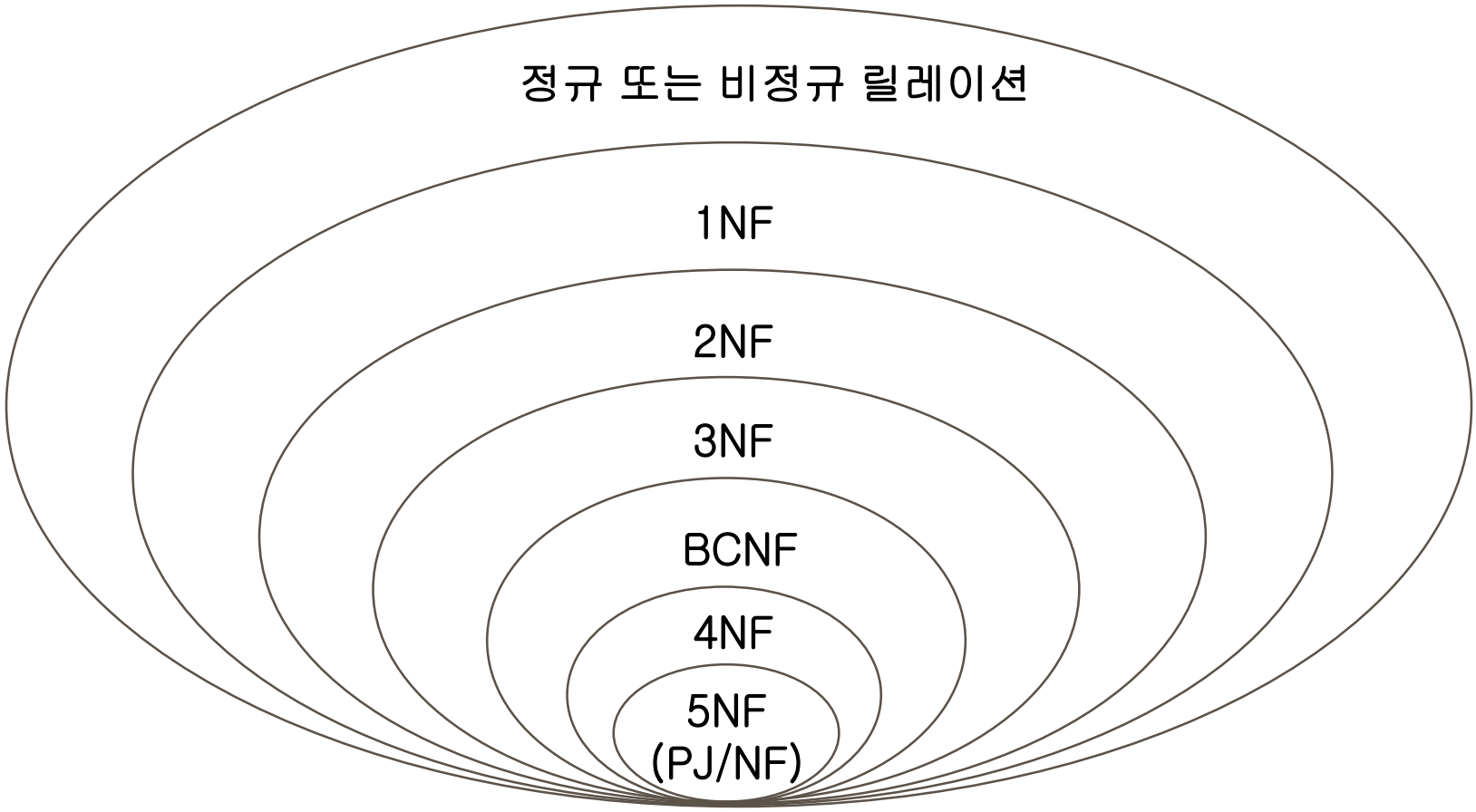
기본 정규형(Normal Form)

- 정규형(Normal Form)
  - 어떤 일련의 제약 조건을 만족하는 릴레이션

- 정규화란?
  - 중복을 최소화하고, 삽입, 삭제, 수정 이상을 최소화하기 위해서 함수적 종속성과 기본 키를 기반으로 릴레이션 스키마를 분석하는 과정
  - 기본적인 아이디어
    - 서로 독립적인 관계는 별개의 릴레이션으로 표현해야 함

- 이렇게 표현된 릴레이션이 어떤 특정의 제약 조건을 만족할 때, 그 제약 조건을 요건으로 하는 정규형에 속한다고 말함
  - 제 1 정규형 ~ 제 5 정규형

정규형들 간의 포함 관계



정규형(Normalization)

- 정규화(Normalization)의 원칙

  정규화 = 스키마 변환 (S -> S')

  ① 무손실 표현
  - 같은 의미의 정보 유지
  - 그러나 더 바람직한 구조

  ② 데이타의 중복성 감소

  ③ 분리의 원칙
  - 독립적인 관계는 별개의 릴레이션으로 표현
  - 릴레이션 각각에 대해 독립적 조작이 가능

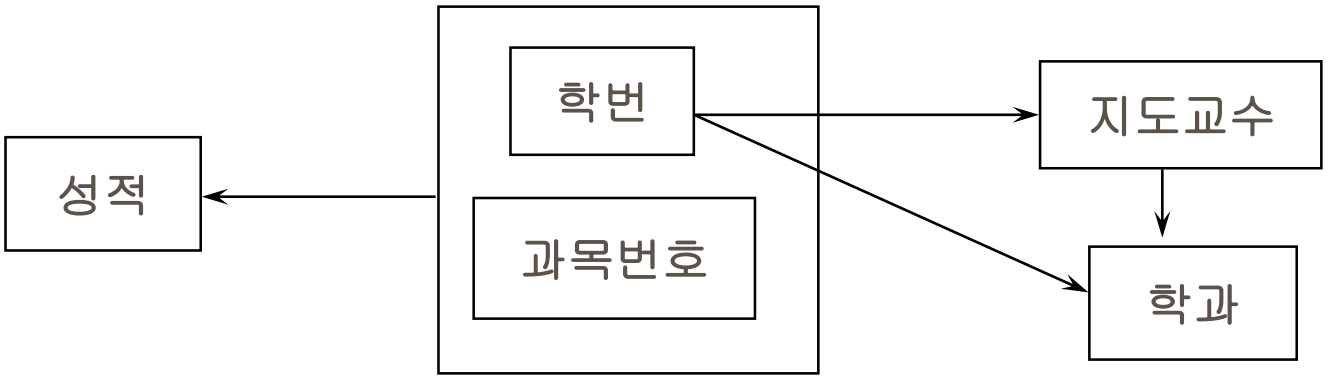
제1정규형 (1NF : First Normal Form)

- 정의
  - 어떤 릴레이션 R에 속한 모든 도메인이 원자값(atomic value)만으로 되어 있다면 제 1 정규형(1NF)에 속한다

- 예) 수강지도 릴레이션

학번	지도교수	학과	과목번호	성적

  - 함수 종속
    - {학번, 과목 번호} → 성적, 학번 →지도교수, 학번 → 학과, 지도교수 → 학과



제1정규형 : 수강 지도 릴레이션의 예

- **수강 지도 릴레이션에서 발생하는 이상 현상들**

수강 지도

학번	지도교수	학과	과목번호	성적
100	P1	컴퓨터	C413	A
100	P1	컴퓨터	E412	A
200	P2	전기	C123	B
300	P3	컴퓨터	C312	A
300	P3	컴퓨터	C324	C
300	P3	컴퓨터	C413	A
400	P1	컴퓨터	C312	A
400	P1	컴퓨터	C324	A
400	P1	컴퓨터	C413	B
400	P1	컴퓨터	C412	C
500	P4			

삭제 이상 발생

갱신 이상 발생
(P1 → P3)

삽입 이상 발생

제1정규형 : 1NF에서의 이상 현상들

- 삽입 이상
  - 예) 500번 학생의 지도교수가 P4라는 사실을 삽입할 경우
  - 500번인 학생이 어떤 교과목을 등록하지 않으면 기본 키{학번, 과목번호}가 null이 되므로, 위의 사실을 삽입할 수 없음

- 삭제 이상
  - 예) 200번 학생이 교과목 C123의 등록을 취소할 경우
  - 학번이 200번에 관련된 투플이 하나만 있는 상황이므로, 이 학생의 지도교수가 P2라는 정보까지 손실됨

- 갱신 이상
  - 예) 400번 학생의 지도교수를 P1에서 P3로 변경할 경우
  - 만일 학번이 400인 4개 투플 중 일부만 지도교수 값을 P3로 변경할 경우 데이터의 일관성을 잃음

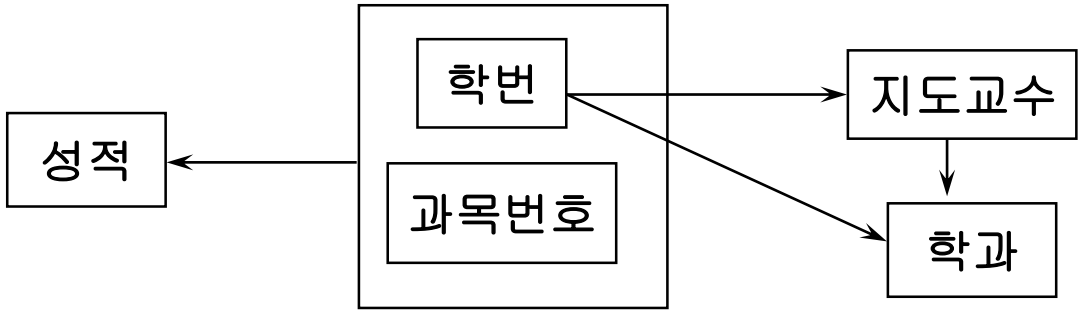
제1정규형 : 이상 현상의 원인 및 해결 방안

- 1NF 이상의 원인
  - 기본 키에 부분 함수 종속된 애트리뷰트가 존재
  - 두 가지 상이한 정보가 포함됨

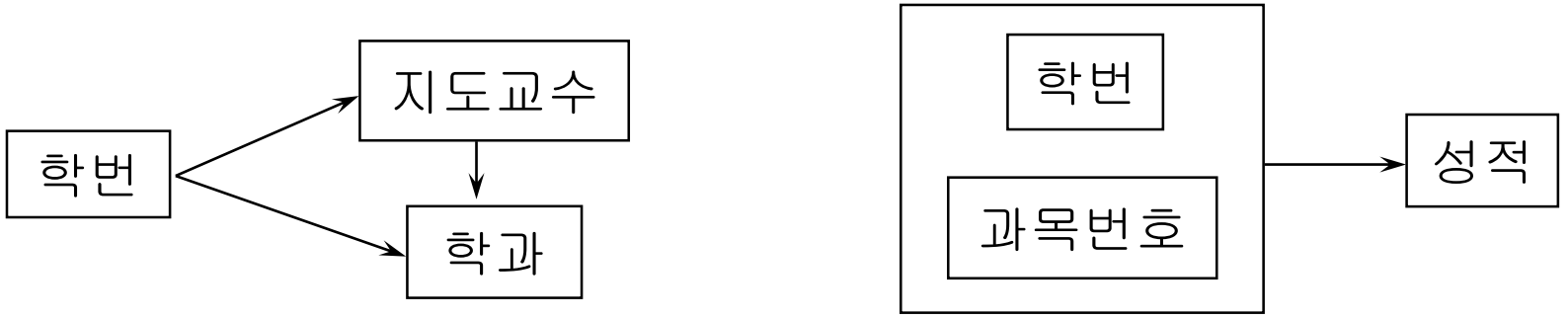
- 해결 방안
  - 프로젝션으로 릴레이션을 분해
  - 부분 함수 종속을 제거 ⇒ 2NF

제1정규형 : 해결 방안 예

- 1NF : 지도교수와 학과가 기본 키인 {학번, 과목번호}에 부분 종속됨



- 해결 방안
  - 수강지도 ⇒ 지도 릴레이션과 수강 릴레이션으로 분해 (2NF)



제2정규형 (2NF)

- **정의**
  - 어떤 릴레이션 R이 1NF이고, 키에 속하지 않는 애트리뷰트 모두가 기본키에 완전 함수 종속이면, 제 2 정규형에 속한다.

- **예) 지도, 수강 릴레이션**

지도

갱신 이상 발생(P1 소속이 컴퓨터→전자과)

학번	지도교수	학과
100	P1	컴퓨터
200	P2	전기
300	P3	컴퓨터
400	P1	컴퓨터

삭제 이상 발생

삽입 이상 발생

어떤 교수가 특정학과에 소속된다는 정보 입력 불가

수강

학번	과목번호	성적
100	C413	A
100	E412	A
200	C123	B
300	C312	A
300	C324	C
300	C413	A
400	C312	A
400	C324	A
400	C413	B
400	C412	C

제 2 정규형 : 지도 릴레이션의 이상

- 삽입 이상
  - 예) 어떤 지도교수가 특정 학과에 속한다는 사실을 삽입하려고 할 때 삽입 불가능
  - 삽입 할 수 있기 위해서는 이 지도 교수의 지도를 받는 학생이 있어야 함
- 삭제 이상
  - 예) 300번 학생의 투플을 삭제하면 지도교수 P3가 컴퓨터공학과에 속한다는 정보가 손실됨
- 갱신 이상
  - 예) 지도교수 P1의 소속이 컴퓨터공학과에서 전자과로 변경된다면 학번이 100과 400번인 두 개의 투플을 모두 변경해야 함

2NF 이상의 원인

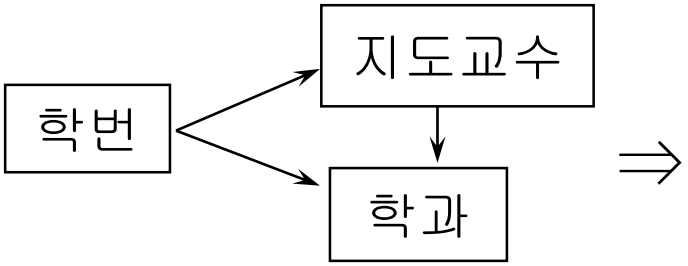
- 2NF 이상의 원인 : 이행적 함수 종속이 존재

- 이행적 함수 종속이란?
  - TD, Transitive Dependency
  - A → B 와 B → C ⇒ A → C
    - 즉, 애트리뷰트 C는 애트리뷰트 A에 이행적 함수 종속

- 2NF 이상의 해결 방안
  - 프로젝션으로 릴레이션 분해
  - 이행적 함수 종속을 제거 ⇒ 3NF

제 3 정규형 (3NF)

- 예 : 지도 ⇒ 학생지도, 지도교수학과 릴레이션



2NF

⇒

학번 → 지도교수

학생지도

학번	지도교수
100	P1
200	P2
300	P3
400	P1

지도교수 → 학과

지도교수학과

지도교수	학과
P1	컴퓨터
P2	전기
P3	컴퓨터

3NF

- 3NF 정의
  - 어떤 릴레이션 R이 2NF이고, 키에 속하지 않은 모든 애트리뷰트들이 기본 키에 이행적 함수 종속이 아닐 때 제 3 정규형에 속한다

제 3 정규형의 문제

- 3NF의 약점
  - i . 복수의 후보키를 가지고 있고,
  - ii . 후보키들이 복합 애트리뷰트들로 구성되며,
  - iii . 후보키들이 서로 중첩되는 경우

  ⇒ 적용 불가능

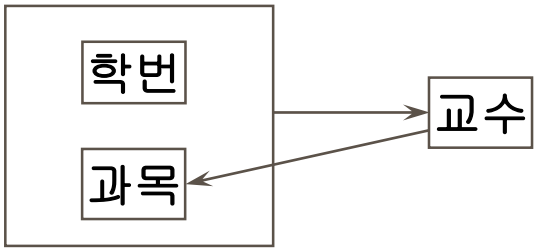
  ⇒ 보다 일반적인 Boyce/Codd Normal Form(BCNF)을 제안

보이스/코드 정규형

- BCNF : Boyce/Codd Normal Form
- 정의
  - 릴레이션 R의 모든 결정자가 후보키이면 릴레이션 R은 BCNF에 속한다.
- 릴레이션 R이 BCNF에 속하면 R은 제1, 제2, 제3 정규형에 속함
- 강한 제3정규형(strong 3NF)이라고도 함

수강과목 릴레이션 예

- **제약 조건: 후보키 : {학번,과목}, {학번,교수}**
  - **각 과목에 대해 한 학생은 오직 한 교수의 강의만 수강한다**
  - **각 교수는 한 과목만 담당한다**
  - **한 과목은 여러 교수가 담당한다.**



3NF

수강과목

갱신 이상 발생
P1이 프로그래밍
→ 자료구조

학번	과목	교수
100	프로그래밍	P1
100	자료구조	P2
200	프로그래밍	P1
200	자료구조	P3
300	자료구조	P3
300	프로그래밍	P4
	자료 구조	**P5**

삭제 이상
발생

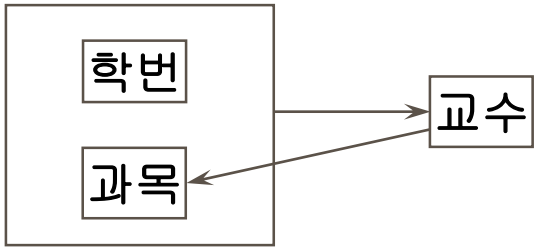
삽입 이상 발생

3NF에서의 이상 현상들

- 삽입 이상
  - 예) 교수 P5가 자료구조를 담당한다는 사실의 삽입은 적어도 수강 학생이 한 사람 있어야 가능
- 삭제 이상
  - 예) 100번 학생이 자료구조를 취소하여 투플을 삭제하면 교수 P2가 자료구조 과목을 담당하고 있다는 정보도 삭제됨
- 갱신 이상
  - 예) 교수 P1이 프로그래밍 과목 대신 자료구조를 담당하게 되면 P1이 나타난 모든 투플들을 변경하여야 함
- 이상 현상의 원인
  - 교수가 결정자이지만 후보키가 아님

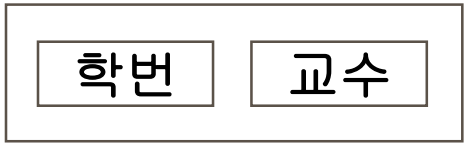
수강과목 릴레이션에 대한 해결 방안

- 수강과목 ⇒ 수강교수, 과목교수 릴레이션



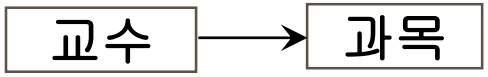
3NF

⇒



수강교수

학번	교수
100	P1
100	P2
200	P1
200	P3
300	P3
300	P4



과목교수

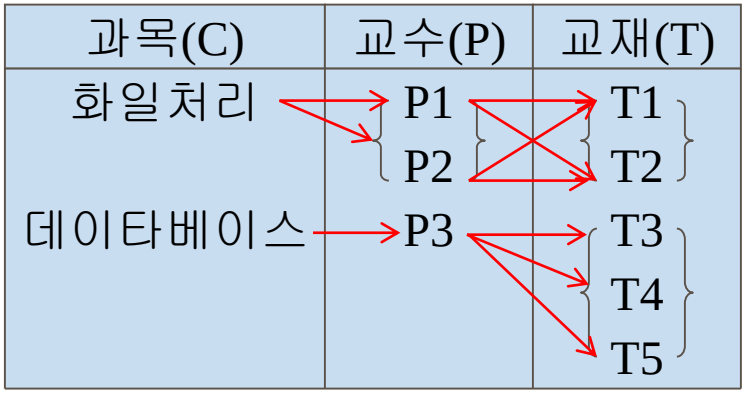
교수	과목
P1	프로그래밍
P2	자료구조
P3	자료구조
P4	프로그래밍

BCNF

제 4 정규형

- 예) 개설 교과목 릴레이션

교과목 목록



⇐ 비정규형 (Repeating Group)

개설 교과목

과목(C)	교수(P)	교재(T)
화일처리	P1	T1
화일처리	P1	T2
화일처리	P2	T1
화일처리	P2	T2
데이타베이스	P3	T3
데이타베이스	P3	T4
데이타베이스	P3	T5

⇐ BCNF
∵ (키에 속하지 않는 결정자 애트리뷰트가 없음)

개설 교과목 릴레이션의 문제점과 해결 방안

- 갱신 이상
  - 새로운 교수 P4가 데이타베이스를 담당한다는 정보 삽입 시, 3개의 교재{T3, T4, T5}에 대한 새로운 3개의 투플을 삽입해야 함

- 이상 현상의 원인
  - 교수와 교재가 서로 무관한 것을 한 릴레이션으로 묶어서 표현한 것이 원인임

- 해결 방안
  - 과목 교수(과목, 교수)릴레이션과 과목 교재(과목, 교재) 릴레이션으로 분해
  - 분해 원리
    - 개설 교과목 릴리에션은 모두 키로 구성되어 있기 때문에 함수 종속이 없음
    - 그러므로, 다치 종속에 의해 분리

다치 종속

- MVD, Multi-Valued Dependency

- 정의
  - A, B, C를 릴레이션 R의 애트리뷰트의 부분 집합이라 할 때 애트리뷰트 쌍(A, C)-값에 대응되는 B-값의 집합이 A-값에만 종속되고 C-값에는 독립이면 B는 A에 다치 종속이라 함
  - 표기 : $A \twoheadrightarrow B$
  - 예) 개설 교과목 릴레이션
    - 과목 $\twoheadrightarrow$ 교수, 과목 $\twoheadrightarrow$ 교재

- MVD를 가진 릴레이션의 분해(Fagin의 정리)
  - R(A,B,C)에서 MVD $A \twoheadrightarrow B|C$이면 R1(A,B)와 R2(A,C)로 무손실 분해 가능

제4정규형 정의

- 정의
  - 릴레이션 R에서 MVD A $\twoheadrightarrow$ B가 존재할 때 R의 모든 애트리뷰트들이 A에 함수 종속(FD) (즉 R의 모든 애트리뷰트 X에 대해 A $\rightarrow$ X 이고 A가 후보키)이면 릴레이션 R은 제 4 정규형에 속한다

- BCNF를 이용한 정의
  - 릴레이션 R이 BCNF에 속하고 모든 MVD가 FD이면 R은 4NF

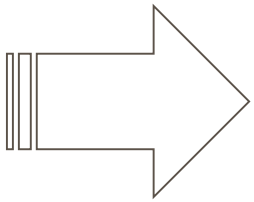
- 의미
  - 어떤 릴레이션 R이 4NF이라면 MVD가 없거나, MVD A $\twoheadrightarrow$ B|C 가 있을 경우 A에 대응되는 B와 C의 값은 하나씩 이어야 하며 이때 A는 후보키라는것을 의미한다.

제 4 정규형 예

- **개설 교과목 릴레이션에 대한 해결 방안**

개설 교과목

과목(C)	교수(P)	교재(T)
화일처리	P1	T1
화일처리	P1	T2
화일처리	P2	T1
화일처리	P2	T2
데이타베이스	P3	T3
데이타베이스	P3	T4
데이타베이스	P3	T5



교과목 교수

과목(C)	교수(P)
화일처리	P1
화일처리	P2
데이타베이스	P3

교과목교재

과목(C)	교재(T)
화일처리	T1
화일처리	T2
데이타베이스	T3
데이타베이스	T4
데이타베이스	T5

4NF

조인 종속 – 1/2

- 가정
  - 릴레이션 R에는 BCNF까지의 어떤 정규형도 위배하는 함수적 종속성이 존재하지 않음
  - 또, 제 4 정규형을 위배하는 다치종속성도 존재하지 않음
  - R을 두 개의 스키마로 무손실 분해를 할 수 없으나, 두 개 이상의 릴레이션 스키마로 무손실 분해 할 수 있는 경우
- 조인 종속성이라는 새로운 종속성을 사용
- 조인 종속성이 존재하는 경우 다중 분해를 통하여 제 5 정규형을 만듦

조인 종속 – 2/2

- 릴레이션이 그의 어떤 프로젝션들을 조인한 결과와 똑같아야 한다는 제약 조건

- 조인 종속 정의
  - 어떤 릴레이션 R의 애트리뷰트들에 대한 부분 집합 $A_1, A_2 \cdots, A_n$이 있다고 하자. 이때 만일 릴레이션 R이 그의 프로젝션 $A_1, A_2, \cdots, A_n$을 모두 조인한 결과와 똑같이 된다면 R은 조인 종속 *($A_1, A_2 \cdots, A_n$)을 만족시킨다고 한다.
  - 즉, 릴레이션R이 조인 종속 *($A_1, A_2 \cdots, A_n$)만족하면 n-분해 릴레이션이다.

조인 종속으로 인한 이상 현상들 – 1/2

- 예) SUPPLY 릴레이션
  - 제약 조건
    - 공급자 s가 부품 p를 공급하고, 프로젝트 c에서 부품 p가 사용되고, 공급자 s가 적어도 하나 이상의 부품을 프로젝트 c에 사용할 때마다
    - 공급자 s는 부품 p를 프로젝트 c에 공급하는 것을 하나의 릴레이션에 표현

- 삽입 이상 현상
  - SUPPLY 릴레이션에서 (S2,P1,C1)의 삽입 시 (S1,P1,C1)의 삽입 필요

SK	PK	CK
S1	P1	C2
S1	P2	C1

조인 종속으로 인한 이상 현상들 – 2/2

- 삭제 이상 현상
  - SUPPLY 릴레이션에서 (S1,P1,C1)의 삭제 시 다른 투플들 중 어느 하나를 함께 삭제하여야 함

SK	PK	CK
S1	P1	C2
S1	P2	C1
S2	P1	C1
S1	P1	C1

- 이상 현상의 원인
  - 3-분해 릴레이션이기 때문

- 해결 방안 : SUPPLY 릴레이션을 3-분해함

해결 방안과 제 5 정규형 정의

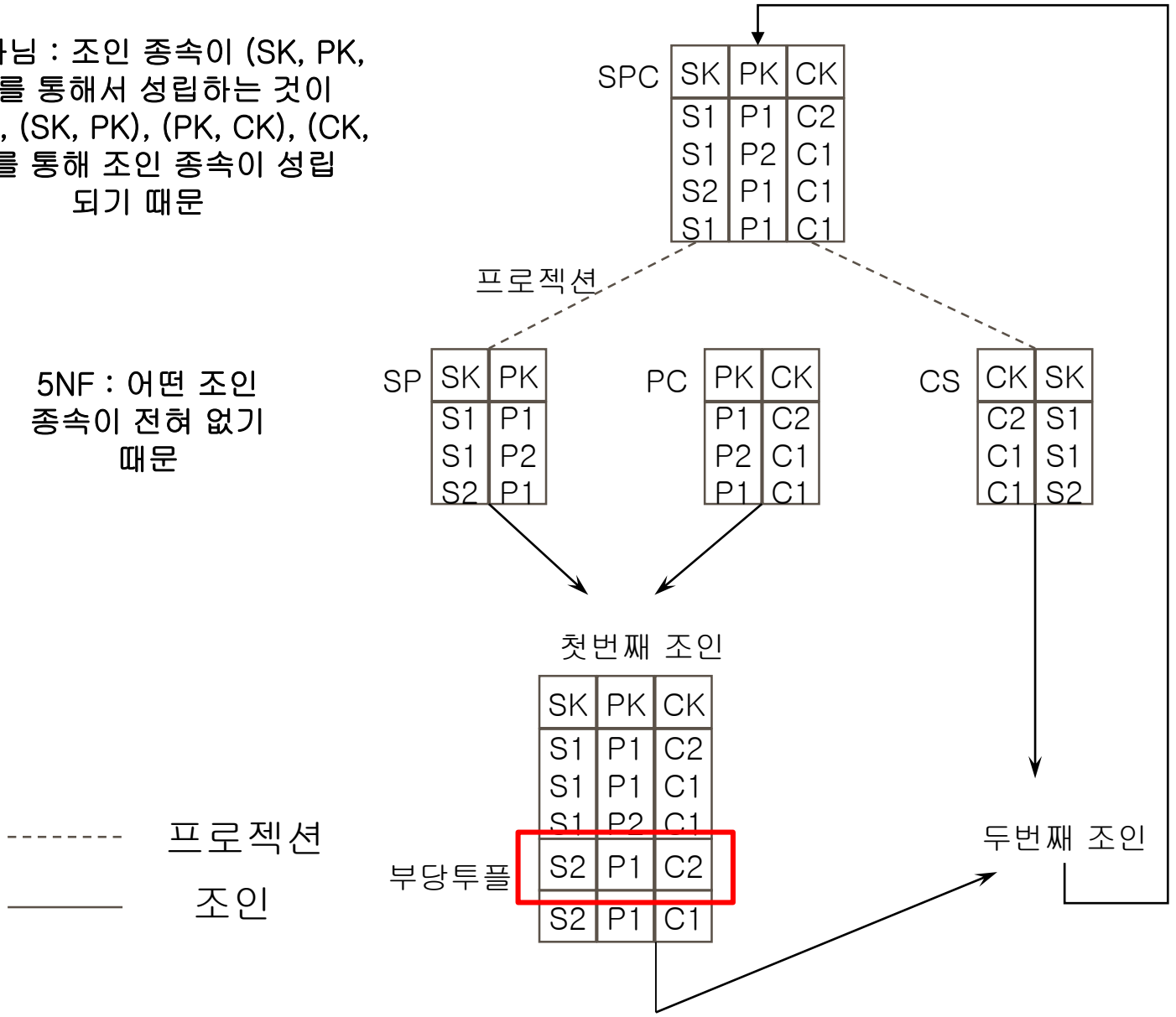
- 해결 방안 : SUPPLY 릴레이션을 3-분해함

- 제 5 정규형 정의
  - 릴레이션 R에 존재하는 모든 조인 종속 (JD)이 릴레이션 R의 후보키를 통해서만 성립된다면 릴레이션 R은 제 5 정규형 또는 PJ/NF(Projection-Join Normal Form)에 속한다.

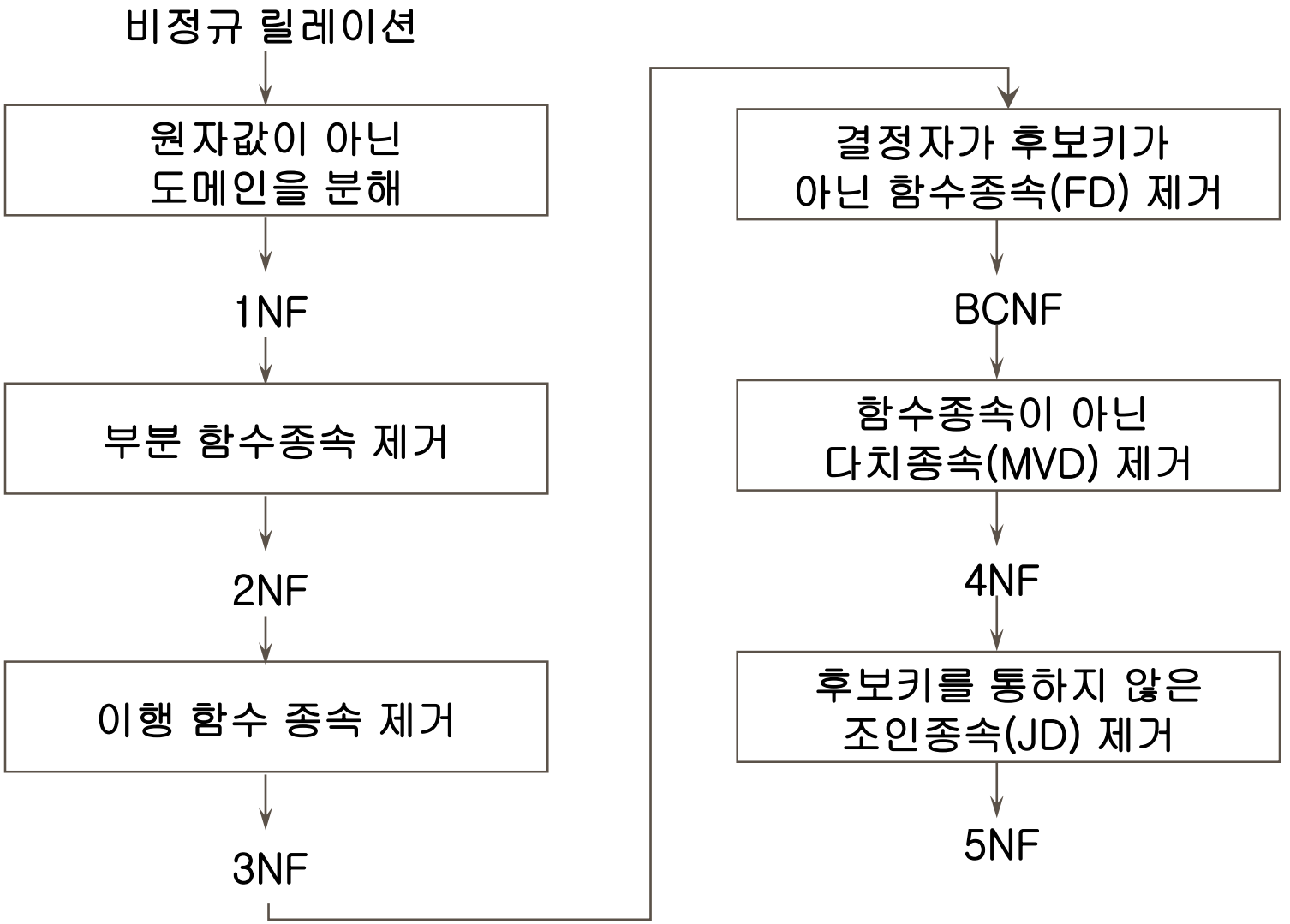
조인 종속성을 가진 릴레이션의 예

5NF 아님 : 조인 종속이 (SK, PK, CK)를 통해서 성립하는 것이 아니고, (SK, PK), (PK, CK), (CK, SK)를 통해 조인 종속이 성립되기 때문

5NF : 어떤 조인 종속이 전혀 없기 때문



결론 : 정규화 과정


비정규 릴레이션
원자값이 아닌
도메인을 분해
1NF
부분 함수종속 제거
2NF
이행 함수 종속 제거
3NF
결정자가 후보키가
아닌 함수종속(FD) 제거
BCNF
함수종속이 아닌
다치종속(MVD) 제거
4NF
후보키를 통하지 않은
조인종속(JD) 제거
5NF

❖ 반(역)정규화(De-Normalization)

☞ Note

- 현실적으로 모든 릴레이션을 반드시 5NF에 속하도록 분해할 필요는 없음
- 학생주소(학번,이름,주소,전화번호) : BCNF가 아님

  기본키 : 학번

  FD : 전화번호→주소

학생전화(학번,이름,전화번호) : 5NF

전화주소(전화번호,주소) : 5NF

이름, 전화번호, 주소는 분리하지 않고 사용하는 것이 검색 성능은 더 좋으므로 위의 5NF으로의 분해는 무의미함