제4장 관계 데이터베이스 정규화

1. 정규화 개요

// 이상(anomaly)

이상은 릴레이션에 어떤 **연산**을 수행할 때 발생되는 **곤란한 현상**을 말한다.
이상은 릴레이션에 데이터 **중복**이나 잘못된 **종속** 관계가 존재할 때 발생될 수 있다.

수강

학번	과목	학년	점수	담당교수
1	DB	4	80	홍하은
2	DB	3	90	홍재연
2	OS	3	75	이순신
3	DB	4	95	홍재연
3	OS	4	90	홍재연

기본키 = [학번 + 과목]

(1) 삽입 이상(insertion anomaly)

① 새로운 자료를 릴레이션에 삽입하는 경우에 **불필요한 정보**를 같이 저장해야만 되고, 아니면 자료 자체가 삽입되지 않는 현상이다.
② 예를 들어, **'수강'** 릴레이션에 '학번이 7이고, 학년이 3'이라는 사실만을 삽입하려고 한다. 그러면 삽입되지 않는다. 이유는 {학번과 과목}이 릴레이션의 기본키이기 때문이다. 기본키는 null이 될 수 없다. 만일, 이를 꼭 삽입하려면 '과목'에 원하지 않는 임의 값을 같이 삽입해야 한다.

(2) 삭제 이상(deletion anomaly)

① 어떤 튜플이 삭제될 때 보관되어야 할 정보까지 같이 삭제되는 현상이다. 즉, **정보 손실**이 발생되는 현상을 '삭제 이상'이라 한다.
② 예를 들어, **'수강'** 릴레이션에서 학번 1인 학생이 수강등록을 취소하여, 이 튜플이 삭제되면 이 학생이 4학년이라는 정보까지 같이 삭제된다. 이유는 학생의 학년 정보를 가지는 유일한 튜플이기 때문이다.

(3) 갱신 이상(update anomaly)

① 중복된 튜플 중에서 일부 속성값만을 수정하면 **정보의 모순**이 발생된다. '갱신 이상'이 발생된다.
② 예를 들어, **'수강'** 릴레이션에서 학번 2인 학생의 학년을 3에서 4로 수정해야 하는 경우에 하나의 튜플만을 수정하면 자료의 일관성이 없게 된다.

// 정규화(normalization)

(1) 이상이 발생하는 원인

데이터 중복	• 데이터 중복은 릴레이션을 처리할 때 이상을 발생시키다. • 데이터 중복은 데이터 관리에 여러 가지 치명적인 문재를 발생시키다.
데이터 종속성	• 이상은 '여러 종류의 사실들을 **하나의 릴레이션**으로 취급'하면 발생된다. • 즉, **튜플**을 구성하는 속성들 사이에는 서로 복잡한 종속 관계가 존재한다. • 하나의 릴레이션으로 모든 것을 표현하면 문제가 발생될 수밖에 없다. • 속성들 사이의 복잡한 종속성을 무시한 결과이다.

(2) 이상을 해결하는 방법

릴레이션 분해	• 릴레이션에는 **불필요한 정보가 중복**될 수 있다. • 데이터 **중복**이 많을수록 **이상(문제점)**은 더 많이 발생될 수 있다. • 릴레이션을 작게 **분해**(decomposition)하면 **중복은** 최소화된다. • 하나의 릴레이션은 속성들 사이에 하나의 종속성만 갖도록 작게 분해한다. • 릴레이션을 분해하면 하나의 릴레이션은 하나의 종속성을 갖는 속성들만으로 표현될 수 있다. • 이런 분해 과정을 **정규화**라고 한다.

(3) 스키마 변환(schema transformation)

스키마 변환은 릴레이션을 보다 바람직한 구조로 변환하는 것으로 다음 원칙을 따른다.

정보의 무손실	변환되기 전에 스키마가 표현하고자 하는 정보의 내용이 그대로 포함되어 있어야 한다.
최소의 자료 중복	자료 중복은 경제적인 손실뿐만 아니라 릴레이션 조작시 많은 이상을 발생시키는 요인이 된다.
독립적인 구조 (분리의 원칙)	서로 연관된 자료는 독립된 하나의 릴레이션으로 **분리**하여 표현한다. 이는 자료를 독립적으로 처리할 수 있는 기초가 된다.

// 비정규형 릴레이션

학생				과목	
학번	이름	학년	학과	수강명	담당교수
1	순이	4	전산	디비	홍하은
2	철수	4	전산	소공	홍재연
3	하나	3	토목	역학	이관석

[비정규형 릴레이션]

관계 데이터베이스 정규화 과정에서 비정규형 릴레이션은 정규형으로 변환시키고, 정규형은 다시 분해하여 최적의 릴레이션 구조가 되도록 변환시킨다.

// 정규형(normal form)

먼저, 정규형 사이의 관계는 다음과 같다.

비정규형

↓ **원자값**이 아닌 모든 도메인 분해(도메인이 원자값)

1NF ~ 릴레이션의 **모든 도메인(속성값)**이 **원자값**으로 구성되어 있다.

↓ **부분함수종속** 제거

2NF ~ **1NF이고, 기본키에 속하지 않은 모든 속성**이 **기본키에 완전함수종속**이다.

↓ **전이함수종속**(이행함수종속) 제거

3NF ~ **2NF이고, 기본키에 속하지 않은 모든 속성**이 **기본키**에 **전이함수종속**이 **아니다.**

↓ **결정자**이면서 **후보키**가 **아닌 것** 제거

BCNF ~ **3NF이고, 릴레이션**의 **모든 결정자**가 **후보키**이다.

↓ 함수종속이 아닌 **다치종속**(MVD) 제거

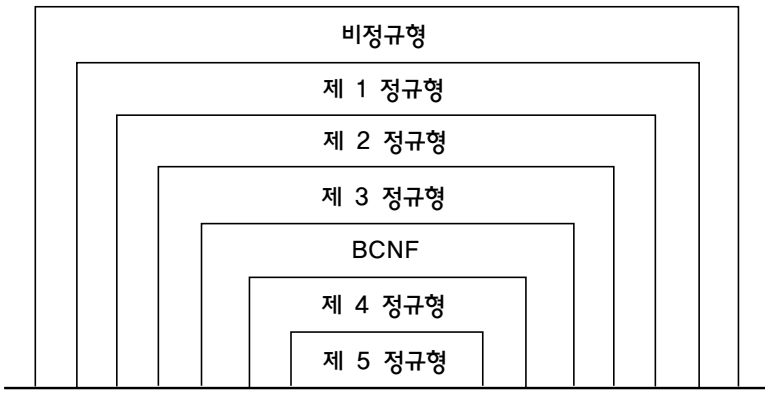
4NF ~ **BCNF이고, 모든 다치종속**이 **함수종속**이면 **4NF**에 속한다.

↓ 후보키를 통하지 않은 **조인종속**(JD) 제거

5NF ~ **4NF이고, 모든 조인종속**이 릴레이션의 **후보키**에 의해서만 **성립**된다.

☞ 3NF에는 전이함수종속이 **없다.**

정규형 사이의 포함 관계는 다음과 같다.



- 정규형은 차수가 높을수록 보다 **강력한 제약조건**이 적용된 것이다.
- 릴레이션의 분해는 자료 중복을 최소화하여 이상이 발생되지 않도록 하는 것이지
- 모든 릴레이션을 제5정규형이 되도록 분해하는 것은 아니다.
- 데이터베이스 특성에 따라 설계자가 결정해야 한다.

기출문제 분석

1. 학생의 학번, 성명, 소속학과, 지도교수에 대한 데이터와 학생이 수강한 교과목의 교과목번호, 학번, 학점에 대한 데이터를 한 곳 에서 관리하기 위하여 릴레이션 R과 함수적 종속 FD를 아래와 같이 구성하였다. 릴레이션 R을 운영·관리하는 과정에서 발생할 수 있는 이상(anomaly)에 해당되지 않는 것은? (단, 밑줄은 릴레이션의 기본키를 의미한다) [2010년 국가 7급]

릴레이션	R (학번, 성명, 교과목번호, 학점, 지도교수, 소속학과)
함수종속	FD : {학번→성명, 학번→지도교수, 학번→소속학과, (학번,교과목번호)→학점}

① 학생의 학번과 교과목번호를 입력하는 동시에 지도교수와 소속학과를 입력할 수 있다.
② 학생의 지도교수를 변경할 경우 학생이 수강하여 학점을 취득한 교과목 수만큼 변경작업을 반복하여야 한다.
③ 학생이 수강한 교과목과 학점을 입력할 때마다 해당 학생의 성명, 지도교수, 소속학과가 반복적으로 저장된다.
④ 학생이 한 교과목만 수강 신청하여 학점을 얻은 후 학점포기로 해당 튜플을 삭제할 경우 학생의 소속학과가 파악되지 않을 수 있다.

♤ 이상(anomaly)

// 다음과 같은 릴레이션 구조이다.

R

학번	성명	교과목번호	학점	지도교수	소속학과
1	이순신	DB_1	90	P1_홍재연	전산
1	이순신	SE_2	80	P1_홍재연	전산
2	강감찬	DB_1	80	P2_홍하은	토목
2	강감찬	SE_2	90	P2_홍하은	토목
3	임꺽정	DB_1	70	P3_이삼오	건축
4		DB_1		P1_홍재연	전산

· 학생의 학번과 교과목번호를 입력하는 동시에 지도교수와 소속학과를 입력할 수 있다.(○)
→ 동시에 입력할 수 있는 것은 이상에 해당되지 않는다.

· 갱신이상 : 학생의 지도교수를 변경하려면, 학생이 수강한 교과목 수만큼 변경해야 한다.
· 삽입이상 : 학생이 수강한 과목과 학점을 입력할 때마다 성명, 지도교수, 소속학과가 반복 저장
· 삭제이상 : 학생이 한 과목만 수강 신청한 후 포기하면 학생의 소속학과가 사라진다.

정답 : ①

2. 관계형 데이터베이스(relational database)에 대한 설명으로 옳은 것을 〈보기〉에서 모두 고르면? [2018년 국회 9급]

----------<보 기>----------

ㄱ. 스키마 변환 시 정보의 무손실, 자료 중복의 감소, 관련된 구조 간의 통합의 원칙을 준수하여야 한다.

ㄴ. 관계대수(relational algebra)의 연산에서 피연산자는 모두 릴레이션이지만 연산결과는 릴레이션이 아니다.

ㄷ. 릴레이션에 연산을 수행 시 삽입이상(insertion anomaly), 삭제이상(deletion anomaly), 갱신이상(update anomaly)이 발생 할 수 있다.

ㄹ. 튜플을 구성하는 속성 사이에 존재하는 종속관계를 고려하지 않고 하나의 릴레이션으로 표현하여 이상(anomaly)을 해결 할 수 있다.

ㅁ. 릴레이션이 여러 속성을 표현할 때 이를 작게 분해(decomposition)하는 과정을 정규화(normalization)라고 한다.

ㅂ. 릴레이션들은 관계대수(relational algebra)로 조작이 가능하다.

----

① ㄱ, ㄴ, ㄷ ② ㄴ, ㄷ, ㄹ ③ ㄴ, ㄹ, ㅁ
④ ㄷ, ㄹ, ㅂ ⑤ ㄷ, ㅁ, ㅂ

♤ **관계형 데이터베이스**

ㄱ. 스키마 변환 시 정보의 무손실, 자료 중복의 감소, 관련된 구조 간의 통합의 원칙을 준수하여야 한다.**(×)**

→ 관련된 구조 간의 분해의 원칙을 준수하여야 한다.

→ 독립적인 구조로 분리

→ 서로 연관된 자료는 독립된 하나의 릴레이션으로 분리하여 표현한다.

→ 이는 자료를 독립적으로 처리할 수 있는 기초가 된다.

ㄴ. 관계대수(relational algebra)의 연산에서 피연산자는 모두 릴레이션이지만 연산결과는 릴레이션이 아니다.**(×)**

→ 관계대수의 연산결과도 모두 릴레이션이다.

ㄹ. 튜플을 구성하는 속성 사이에 존재하는 종속관계를 고려하지 않고 하나의 릴레이션으로 표현하여 이상(anomaly)을 해결 할 수 있다.**(×)**

→ 속성 사이에 존재하는 종속관계를 고려하여 릴레이션을 분해해야 한다.

→ 즉, 튜플을 구성하는 속성들 사이에는 서로 복잡한 종속 관계가 존재한다.

정답 : ⑤

3. 정규화에 대한 설명으로 옳지 않은 것은? [2016년 국가 7급]

① 정규화의 목적은 각 릴레이션에 분산된 종속성을 하나의 릴레이션으로 통합하는 것이다.
② 정규화 과정을 거치지 않으면 여러 다른 종류의 정보를 하나의 릴레이션에 표현하여 그 릴레이션을 조작할 때 이상 현상이 발생할 수 있다.
③ 데이터 간에 존재하는 함수종속은 이상 현상의 원인이 될 수 있다.
④ 정규화가 잘못되면 데이터의 불필요한 중복이 발생하여 릴레이션 조작 시 문제를 유발할 수 있다.

♤ **정규화**

- 하나의 릴레이션을 구성하는 속성들 사이에는 다양한 데이터 종속 관계가 존재할 수 있다.
- 정규화는, 하나의 릴레이션에는 기본적으로 하나의 종속성이 표현되도록 **분해**한다.

정답 : ①

4. <보기>에서 설명하는 생년월일, 주소의 속성 종류는? [2021년 서울 7급]

----<보기>----

- **고객 개체의 생년월일 속성은 연, 월, 일로 의미를 세분화할 수 있다.**
- **고객 개체의 주소 속성은 시(도), 구(군), 동, 우편번호 등으로 의미를 세분화할 수 있다.**

① 단일 값 속성 ② 다중 값 속성
③ 단순 속성 ④ 복합속성

♤ **복합속성(composite attribute)**

- **복합속성**은 하나의 속성이 **여러 개의 속성**으로 분리될 수 있는 속성이다.

학생

학번	이름	주소
1	김유신	서울 동작구 노량진동 1
2	이순신	서울 종로구 광화문동 2

- 주소가 **복합**속성이다.
- 복합속성은 무조건 분리하는 것은 아니다.
- 복합속성은 요구사항에 따라 필요시 분리한다.

정답 : ④

5. 정규화(normalization)에 대한 설명 중 옳은 것만을 모두 고르면? [2021년 국가 7급]

ㄱ. 데이터의 정규화는 중복을 최소화하고 삽입, 삭제, 수정 이상을 최소화하기 위해서 함수적 종속성과 기본키를 기반으로, 주어진 릴레이션 스키마를 분석하는 과정이다.

ㄴ. 릴레이션 스키마 R의 모든 원소들의 도메인(domain)이 나눌 수 있는 단위로 되어있을 때, R이 제1정규형에 속한다.

ㄷ. 제2정규형이 되기 위해서는 릴레이션 R이 제1정규형이고 기본키가 아닌 속성이 기본키에 부분함수종속이어야 한다.

ㄹ. 제3정규형이 되기 위해서는 릴레이션 R이 제2정규형이고, 릴레이션 R의 함수종속 관계에서 이행적함수종속을 제거해야 한다.

ㅁ. 제4정규형이 되기 위해서는 릴레이션 R이 제3정규형이고, 함수종속 관계에서 모든 결정자가 후보키이면 된다.

---

① ㄱ, ㄹ ② ㄴ, ㄷ

③ ㄱ, ㄹ, ㅁ ④ ㄷ, ㄹ, ㅁ

♤ **정규화**

ㄴ. 릴레이션 스키마 R의 모든 원소들의 도메인(domain)이 **나눌 수 있는 단위**로 되어있을 때, R이 제1정규형에 속한다.**(×)**

→ R의 모든 원소들의 도메인이 **나눌 수 없는 단위**로 되어있을 때, 제1정규형에 속한다.

ㄷ. 제2정규형이 되기 위해서는 릴레이션 R이 제1정규형이고 기본키가 아닌 속성이 기본키에 **부분함수종속**이어야 한다.**(×)**

→ 제2정규형이 되기 위해서는 제1정규형이고 기본키가 아닌 속성이 기본키에 **완전함수종속**이어야 한다.

ㅁ. **제4정규형**이 되기 위해서는 릴레이션 R이 제3정규형이고, 함수종속 관계에서 모든 결정자가 후보키이면 된다.**(×)**

→ **보이스/코드 정규형**(BCNF; boyce/codd normal form)에 대한 설명이다.

// **제4정규형**을 간단하게 정의하면 다음과 같다.

함수종속이 아닌 다치종속(MVD)이 제거되면 4NF에 속한다.

- **다치종속**은 하나의 릴레이션에 **다가속성**이 **2개이상** 존재할 때 발생한다.
- 원래, 관계 데이터베이스에서는 속성 값으로 다중값(다가속성)을 허용하지 않는다.(제1NF 제약)
- 다치종속은 머리가 두 개인 이중 화살표 기호 →→로 나타낸다.
- 예 : 과목 →→ 교수, 교수 = {P1, P2}

정답 : ①

6. BCNF(boyce-codd normal form)를 만족하기 위한 조건만을 모두 고른 것은? [2015년 국가 7급]

ㄱ. 모든 결정자가 후보키이여야 한다.
ㄴ. 후보키에 속하지 않는 모든 애트리뷰트가 기본키에 이행함수종속 되어 있지 않다.
ㄷ. 릴레이션의 모든 애트리뷰트가 원자값을 갖는다.
ㄹ. 후보키에 속하지 않는 모든 애트리뷰트가 기본키에 부분함수종속 되어 있지 않다.

---

① ㄱ, ㄷ
② ㄱ, ㄴ, ㄹ
③ ㄴ, ㄷ, ㄹ
④ ㄱ, ㄴ, ㄷ, ㄹ

♤ BCNF

• 주어진 내용 모두가 BCNF를 만족하기 위한 조건들이다.

// 데이터베이스 정규형

비정규형
↓ 원자값이 아닌 모든 도메인을 분해한다.
1NF ~ 릴레이션의 모든 도메인(속성값)이 원자값으로 구성되어 있다.
↓ 부분함수종속을 제거한다.
2NF ~ 1NF이고, 기본키에 속하지 않은 모든 속성이 기본키에 완전함수종속이다.
↓ 전이함수종속을 제거한다.
3NF ~ 2NF이고, 기본키에 속하지 않은 모든 속성이 기본키에 전이함수종속이 아니다.
↓ 결정자가 후보키가 아닌 함수종속을 제거한다.
BCNF ~ 3NF에 속하고, 릴레이션의 모든 결정자가 후보키이다.
↓ 함수종속이 아닌 다치 종속(MVD)을 제거한다.
4NF ~ BCNF에 속하고, 모든 다치 종속이 함수종속이면 4NF에 속한다.
↓ 후보키를 통하지 않은 조인 종속(JD)을 제거한다.
5NF ~ 4NF이고, 모든 조인 종속이 릴레이션의 후보키에 의해서만 성립된다.

• 3NF에는 전이함수종속이 없다.

정답 : ④

7. <보기>의 관계 데이터베이스 설계의 함수적 종속성과 정규형에 대한 설명 중 괄호에 들어갈 용어는? [2020년 서울 7급]

----<보기>----------

릴레이션 R이 BCNF에 속하고 모든 ( )이 함수종속(functionally dependent)이면 릴레이션 R은 4NF에 속한다.

---

① 다치 종속성(multivalued dependency) ② 이행 종속성(transitive dependency)
③ 부분 종속성(partial dependency) ④ 조인 종속성(join dependency)

♤ **정규형**

비정규형
↓ **원자값**이 아닌 모든 도메인 분해(도메인이 원자값)
1NF ~ **릴레이션의 모든 도메인(속성값)이 원자값으로 구성되어 있다.**
↓ **부분함수종속** 제거
2NF ~ **1NF이고, 기본키에 속하지 않은 모든 속성이 기본키에 완전함수종속이다.**
↓ **전이함수종속**(이행함수종속) 제거
3NF ~ **2NF이고, 기본키에 속하지 않은 모든 속성이 기본키에 전이함수종속이 아니다.**
↓ **결정자**이면서 **후보키가 아닌 것** 제거
BCNF ~ **3NF이고, 릴레이션의 모든 결정자가 후보키이다.**
↓ 함수종속이 아닌 **다치종속**(MVD) 제거
4NF ~ **BCNF이고, 모든 다치종속이 함수종속이면 4NF에 속한다.**
↓ 후보키를 통하지 않은 **조인종속**(JD) 제거
5NF ~ **4NF이고, 모든 조인종속이 릴레이션의 후보키에 의해서만 성립된다.**

☞ 3NF에는 전이함수종속이 **없다.**

// 다치종속(MVD, multivalued dependency) - 다가종속

다치종속은 속성 A, B, C를 가지는 릴레이션 R(A, B, C)에서
속성 쌍 (A, C)-값에 대응하는 B값의 집합이 A값에만 종속되고 C값과는 독립적일 때,
릴레이션 R에서 B가 A에 다치종속이라 하며, **A↠B**로 표기한다.

- **다치종속**은 하나의 릴레이션에 **다가속성**이 **2개이상** 존재할 때 발생한다.
- 원래, 관계 데이터베이스에서는 속성 값으로 다중값(다가속성)을 허용하지 않는다.(제1NF 제약)
- 다치종속은 머리가 두 개인 이중 화살표 기호 ↠로 나타낸다.
- 예 : 과목 ↠ 교수, 교수 = {P1, P2}

정답 : ①

8. <보기>는 관계형 데이터베이스의 정규화 작업을 설명한 것이다. 제1정규형, 제2정규형, 제3정규형, BCNF를 생성하는 정규화 작업을 순서대로 나열한 것은? [2016년 계리직]

-----<보기>----------
ㄱ. 결정자가 후보키가 아닌 함수종속성을 제거한다.
ㄴ. 부분함수종속성을 제거한다.
ㄷ. 속성을 원자값만 갖도록 분해한다.
ㄹ. 이행적함수종속성을 제거한다.
---

① ㄱ → ㄴ → ㄷ → ㄹ
② ㄱ → ㄷ → ㄹ → ㄴ
③ ㄷ → ㄱ → ㄴ → ㄹ
④ ㄷ → ㄴ → ㄹ → ㄱ

♤ **정규형 사이의 관계**

ㄷ. 속성을 원자값만 갖도록 분해한다. → 제1정규형
↓
ㄴ. 부분함수종속성을 제거한다. → 제2정규형
↓
ㄹ. 이행적함수종속성을 제거한다. → 제3정규형
↓
ㄱ. 결정자가 후보키가 아닌 함수종속성을 제거한다. → BCNF

정답 : ④

9. 보이스 코드 정규형(BCNF: boyce-codd normal form)을 만족하기 위한 조건에 해당하지 않는 것은? [2019년 국가 9급]

① 조인(join) 종속성이 없어야 한다.
② 모든 속성 값이 원자 값(atomic value)을 가져야 한다.
③ 이행적 함수종속성이 없어야 한다.
④ 기본키가 아닌 속성이 기본키에 완전함수종속적이어야 한다.

♤ **정규형**

• 조인(join) 종속성이 없어야 한다.(×) → 조인 종속이 제거된 것은 5NF이다.

정답 : ①