14장. 관계형 데이터베이스 설계 알고리즘(Relational Database Design Algorithms)

관계형 데이타베이스 스키마 설계 알고리즘

- 좋은 데이타베이스 스키마 설계를 위해서는 정규형만으로는 불충분함
- 예제: 2개의 애트리뷰트를 갖는 릴레이션은 BCNF이다. 그러면, 모든 릴레이션을 2개 애트리뷰트로 설계하면 좋은 설계인가?
- 좋은 데이타베이스 설계를 보장하기 위해서는 다음의 추가적인 조건들이 필요함
  - 종속성 보존 특성 (dependency preservation property)
  - 무손실 조인 특성 (lossless join property)

릴레이션 분해와 정규형의 부족한 점

- 전체 릴레이션 스키마(universal relation schema) $R = A_1, A_2, \ldots, A_n$
  - 데이타베이스의 모든 애트리뷰트들을 포함하는 릴레이션이다.
  - 분해는 전체 릴레이션 R로부터 시작한다.
- 설계목표 I: R을 m 개의 릴레이션 스키마의 집합인 분해집합 $D = R_1, R_2, \ldots, R_m$로 분해한다.
  - 각 릴레이션 스키마 R_i 는 R의 애트리뷰트들의 부분집합이다.
  - 애트리뷰트 보존 조건: R의 모든 애트리뷰트들은 적어도 하나의 릴레이션 R_i 에 나타나야 한다.
- 설계목표 II: D의 각 릴레이션 R_i 는 BCNF 혹은 3NF에 속하도록 한다.
  - 좋은 데이타베이스 설계는 이들 정규형만으로 부족하다.
  - 예제: 애트리뷰트가 둘 뿐인 릴레이션은 자동적으로 BCNF에 속하게 된다. 이러한 애트리뷰트가 둘인 릴레이션을 조인하게 되면 가짜 투플이 만들어질 수 있다.

분해와 종속성 보존

- 데이타베이스 설계자는 R의 애트리뷰트들에 대해 성립하는 함수적 종속성의 집합 F를 정의한다.
- 분해집합 D는 종속성을 보존해야 한다
  - 즉, 각 릴레이션 R_i 에서 성립하는 모든 함수적 종속성들의 합집합이 F와 동등 (equivalent)해야 한다.
- 종속성 보존은 정형적으로 다음과 같이 정의한다.
  - R_i 상으로 F의 프로젝션($\Pi_{R_i}(F)$)은 F^+ 에 속하는 함수적 종속성 $X \rightarrow Y$들의 집합이며 $(X \cup Y) \subseteq R_i$ 를 만족한다.
  - 만약 $(\Pi_{R_1}(F) \cup \Pi_{R_2}(F) \cup \ldots \cup \Pi_{R_m}(F))^+ = F^+$ 를 만족하면, 분해 집합 $D = R_1, R_2, \ldots, R_m$은 종속성을 보존한다고 정의한다.
  - 이런 특성으로 인해 각 릴레이션 R_i 상의 함수적 종속성들이 개별적으로 성립함을 보임으로써, F의 함수적 종속성들이 성립함을 보장할 수 있다

Claim 1:

- 종속성들의 집합 F에 대해서 종속성을 보존하면서 각 릴레이션 R_i 가 제 3 정규형인 한 분해집합 D를 항상 구할 수 있다.
- 관계 합성(relational synthesis) 알고리즘

---

1. Find a minimal set of FDs G equivalent to F
2. For each X of an FD $X \rightarrow A$ in G,
   create a relation schema R_i in D with the attributes $\{X \cup \{A_1\} \cup \{A_2\} \cup \ldots \cup \{A_k\}\}$, where $X \rightarrow A_1, X \rightarrow A_2, \ldots, X \rightarrow A_k$ are the only dependencies in G with X as left-hand-side;
3. Place any remaining attributes in a single relation schema to ensure the attribute preservation property.

Claim 1A:

- 관계 합성 알고리즘에 의해 생성되는 모든 릴레이션 스키마는 3NF에 속한다.
- 문제점:
  - F에 대한 최소 폐쇄(minimal cover)를 찾아야 한다.
  - 최소 폐쇄를 찾는 효율적인 알고리즘이 존재하지 않는다.
  - F에 대한 여러 개의 최소 폐쇄가 존재할 수 있다. 따라서, 알고리즘의 결과는 어떤 함수적 종속성이 선택되었는가에 따라 다르다.

분해와 무손실 (비부가적) 조인

- 무손실 조인의 비정형적 정의:
  - 분해집합의 릴레이션들을 조인했을 때 어떠한 가짜 투플도 생성되지 않음을 보장하는 특징이다. = 원래데이터 필요 되돌아간다
- 무손실 조인의 정형적 정의:
  - 함수적 종속성의 집합 F를 만족하는 모든 릴레이션 상태 r에 대해 다음의 성질이 만족되면, R의 분해 집합 $D = \{R_1, R_2, \ldots, R_m\}$는 F에 대해서 무손실 조인 특성을 갖는다. ($*$는 D에 포함된 모든 릴레이션의 자연조인이다.)

  ㅠㅠㅋㅋ 자연조인시켰더니 원래대로 돌아왔

$$*(\Pi_{R_1}(r), \ldots, \Pi_{R_m}(r)) = r$$

- 분해집합이 위의 성질을 만족하면, 프로젝션과 조인 연산 후에 가짜 투플이 추가되지 않음이 보장된다.
- 분해된 릴레이션들은 실제로는 기본 릴레이션으로 저장하기 때문에, 조인을 수반한 질의가 의미있는 결과를 생성하기 위해서는 위와 같은 조건이 필요하다.
- 주어진 분해집합 D가 함수적 종속성 집합 F에 대해서 무손실 조인 특성을 가지는 것을 검사하는 알고리즘이 존재한다.

무손실 (비부가적) 조인 특성의 검사

- 검사 알고리즘 (다음의 그림을 이용하여 설명함)

  (a) EMP_PROJ를 EMP_PROJ1과 EMP_LOCS로 분해한 결과가 무손실조인 특성을 갖는지 검사하기 위하여 알고리을 적용함
  (b) EMP_PROJ의 또 다른 분해집합
  (c) 그림 (b)의 분해집합이 무손실 조인 특성을 갖는지 검사하기 위하여 알고리즘을 적용함

(a)

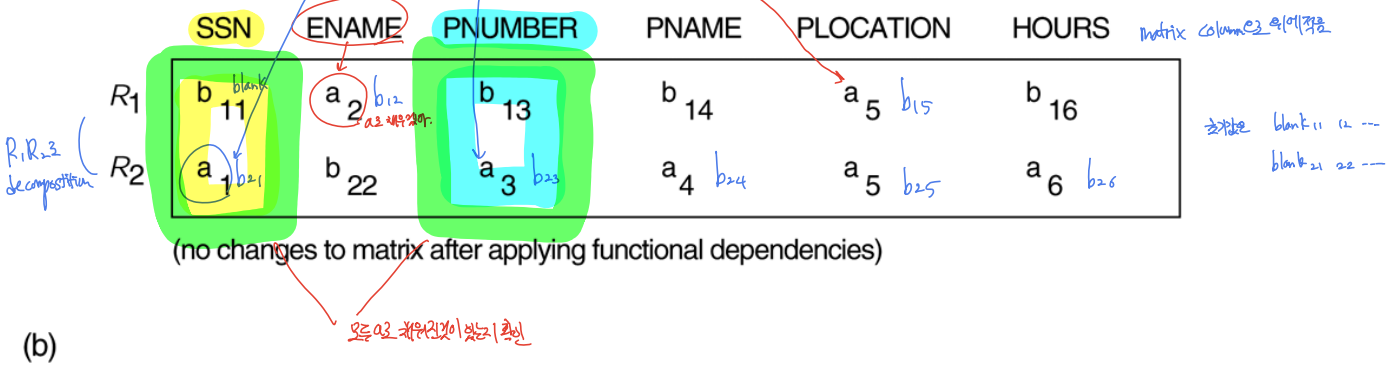
R={SSN, ENAME, PNUMBER, PNAME, PLOCATION, HOURS}

D={R_1, R_2}

R_1=EMP_LOCS={ENAME, PLOCATION}

R_2=EMP_PROJ1={SSN, PNUMBER, HOURS, PNAME, PLOCATION}

F={SSN→ENAME;PNUMBER→{PNAME, PLOCATION};{SSN,PNUMBER}→HOURS}



(b)

EMP

SSN	ENAME

PROJECT

PNUMBER	PNAME	PLOCATION

WORKS_ON

SSN	PNUMBER	HOURS

(c)

R={SSN, ENAME, PNUMBER, PNAME, PLOCATION, HOURS}

D={R_1, R_2, R_3}

R_1=EMP={SSN, ENAME}

R_2=PROJ={PNUMBER, PNAME, PLOCATION}

R_3=WORKS_ON={SSN, PNUMBER, HOURS}

F={SSN→{ENAME;PNUMBER→{PNAME, PLOCATION} ;{SSN,PNUMBER}→HOURS}

	SSN	ENAME	PNUMBER	PNAME	PLOCATION	HOURS
R_1	a_1	a_2	b_{13}	b_{14}	b_{15}	b_{16}
R_2	b_{21}	b_{22}	a_3	a_4	a_5	b_{26}
R_3	a_1	b_{32}	a_3	b_{34}	b_{35}	a_6

(original matrix S at start of algorithm)

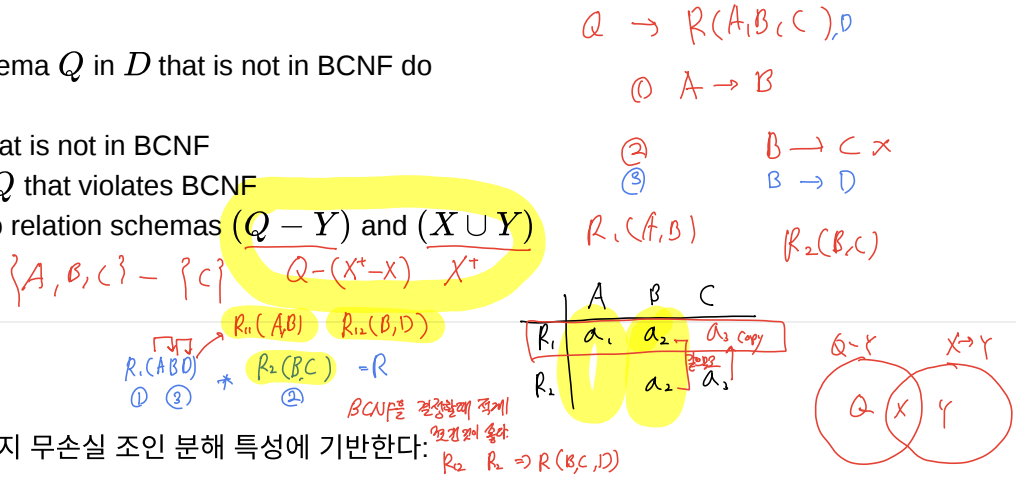
	SSN	ENAME	PNUMBER	PNAME	PLOCATION	HOURS
R_1	a_1	a_2	b_{13}	b_{14}	b_{15}	b_{16}
R_2	b_{21}	b_{22}	a_3	a_4	a_5	b_{26}
R_3	a_1	~~b_{32}~~ a_2	a_3	~~b_{34}~~ a_4	~~b_{35}~~ a_5	a_6

(matrix S after applying the first two functional dependencies - last row is all "a" symbols, so we stop)

분해와 무손실 (비부가적) 조인

- 릴레이션 R을 분해한 릴레이션들이 BCNF 정규형에 속하며, 분해집합이 함수적 종속성의 집합 F에 대해 무손실 조인 특성을 갖게 하는 알고리즘이 존재한다.

Set $D \leftarrow R$
While there is a relation schema Q in D that is not in BCNF do
{
    choose one Q in D that is not in BCNF
    find a FD $X \rightarrow Y$ in Q that violates BCNF
    replace Q in D by two relation schemas $(Q - Y)$ and $(X \cup Y)$
}



- 위 알고리즘은 다음 두 가지 무손실 조인 분해 특성에 기반한다:

- (1) R의 분해집합 $D = \{R_1, R_2\}$가 함수적 종속성의 집합 F에 대해 무손실 조인 특성을 가지기 위한 필요충분조건은 다음의 두 가지 중 반드시 하나를 만족하는 것이다.
  - 함수적 종속성 $((R_1 \cap R_2) \rightarrow (R_1 - R_2))$가 F_+ 에 속한다.
  - 함수적 종속성 $((R_1 \cap R_2) \rightarrow (R_2 - R_1))$가 F_+ 에 속한다.
- (2) R의 분해집합 $D = \{R_1, R_2, \ldots, R_m\}$이 함수적 종속성 집합 F에 대해 무손실 조인 특성을 가지고, R_i의 분해집합 $D_1 = \{Q_1, Q_2, \ldots, Q_m\}$가 F의 R_i 상에의 프로젝션에 대해 무손실 조인 특성을 가진다면, R의 분해집합 $D_2 = \{R_1, R_2, \ldots, R_{i-1}, Q_1, Q_2, \ldots, Q_k, R_{i+1}, \ldots, R_m\}$ 또한 F에 대해 무손실 조인 특성을 가진다.

- 분해집합이 종속성을 보존하며 BCNF 정규형에 속하게 하는 알고리즘은 존재하지 않는다.
- 다음의 수정된 합성 알고리즘은 1) 무손실 조인 특성과 2) 종속성 보존 특성을 만족하며, 3) 제 3 정규형 릴레이션으로 분해하는 것을 보장한다. (주의: BCNF 정규형으로의 분해는 보장 못함)
- 많은 제 3 정규형 릴레이션들은 BCNF 정규형에도 속한다.

무손실 조인과 종속성 보존을 보장하는 제 **3** 정규형 릴레이션으로 분해 **Algorithm**

1. Find a minimal cover for F.
2. For each X of an FD $X \rightarrow Y$ in G
   create a relation schema R_i in D with the attributes $\{X \cup \{A_1\} \cup \{A_2\} \cup \ldots \cup \{A_k\}\}$,
   where $X \rightarrow A_1, X \rightarrow A_2, \ldots, X \rightarrow A_k$ are the only dependencies in G with X as left-hand-side
3. If none of the relations schemas in D contains a key of R, 키를포함하고있지 않다면
   then create one more relation schema that contains
   attributes that form a key for R. key를포함시켜라

Relational synthesis Algo

널값과 허상 투플이 야기하는 문제점

- 널값(null values):
  - 널값이 조인 애트리뷰트에 존재할 때 문제가 발생한다.
  - 널 값이 존재하는 경우 질의를 명기할 때 정규 조인(regular join)과 OUTER 조인의 결과 사이의 차이가 중요해진다.
  - 어떤 질의는 정규 조인을 필요로 하고 어떤 질의는 OUTER 조인을 필요로 한다.
  - 널값 조인의 문제점: (a) 조인 애트리뷰트에 널값이 존재하는 데이타베이스

(a)

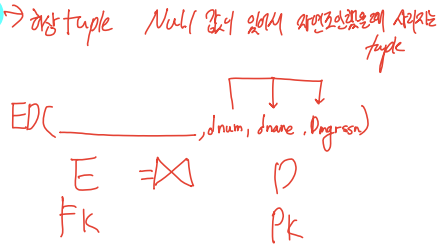
EMPLOYEE

Ename	Ssn	Bdate	Address	Dnum
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4
Wallace, Jennifer S.	987654321	1941-06-20	291 Berry, Bellaire, TX	4
Narayan, Ramesh K.	666884444	1962-09-15	975 Fire Oak, Humble, TX	5
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1
Berger, Anders C.	999775555	1965-04-26	6530 Braes, Bellaire, TX	NULL
Benitez, Carlos M.	888664444	1963-01-09	7654 Beech, Houston, TX	NULL

DEPARTMENT

Dname	Dnum	Dmgr_ssn
Research	5	333445555
Administration	4	987654321
Headquarters	1	888665555

Figure 16.2
Issues with NULL-value joins. (a) Some EMPLOYEE tuples have NULL for the join attribute Dnum. (b) Result of applying NATURAL JOIN to the EMPLOYEE and DEPARTMENT relations. (c) Result of applying LEFT OUTER JOIN to EMPLOYEE and DEPARTMENT.



- 허상 투플(dangling tuples):
  - 정규화 과정에서 릴레이션을 너무 분해하면 허상 투플의 문제가 발생함
  - 분해된 릴레이션을 조인하면 사라지는 투플을 허상투플이라고 함
  - 널값 조인의 문제점:
    - (b) EMPLOYEE와 DEPARTMENT 릴레이션에 자연조인 연산을 적용한 결과
    - (c) EMPLOYEE 릴레이션을 DEPARTMENT 릴레이션과 외부조인한 결과

(b)

Ename	Ssn	Bdate	Address	Dnum	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321
Wallace, Jennifer S.	987654321	1941-06-20	291 Berry, Bellaire, TX	4	Administration	987654321
Narayan, Ramesh K.	666884444	1962-09-15	975 Fire Oak, Humble, TX	5	Research	333445555
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5	Research	333445555
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4	Administration	987654321
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1	Headquarters	888665555

(c)

Ename	Ssn	Bdate	Address	Dnum	Dname	Dmgr_ssn
Smith, John B.	123456789	1965-01-09	731 Fondren, Houston, TX	5	Research	333445555
Wong, Franklin T.	333445555	1955-12-08	638 Voss, Houston, TX	5	Research	333445555
Zelaya, Alicia J.	999887777	1968-07-19	3321 Castle, Spring, TX	4	Administration	987654321
Wallace, Jennifer S.	987654321	1941-06-20	291 Berry, Bellaire, TX	4	Administration	987654321
Narayan, Ramesh K.	666884444	1962-09-15	975 Fire Oak, Humble, TX	5	Research	333445555
English, Joyce A.	453453453	1972-07-31	5631 Rice, Houston, TX	5	Research	333445555
Jabbar, Ahmad V.	987987987	1969-03-29	980 Dallas, Houston, TX	4	Administration	987654321
Borg, James E.	888665555	1937-11-10	450 Stone, Houston, TX	1	Headquarters	888665555
Berger, Anders C.	999775555	1965-04-26	6530 Braes, Bellaire, TX	NULL	NULL	NULL
Benitez, Carlos M.	888665555	1963-01-09	7654 Beech, Houston, TX	NULL	NULL	NULL