Database and Big Data Systems

Lab 4

Today

- Recap
- Schema Design

Recap - Functional Dependencies

sid	name	address
123	Agus	Labrador Park
456	Bron	Bukit Brown
789	Hannah	Sungei Buloh
012	Dewi	Bukit Puaka

More formally,
A functional dependency $X_1, \ldots, X_m \rightarrow Y_1, \ldots, Y_n$
holds in a relation R if:

$$\forall t, t' \in R,\ t[X_1] = t'[X_1] \cap \cdots \cap t[X_m] = t'[X_m] \rightarrow t[Y_1] = t'[Y_1] \cap \cdots \cap t[Y_n] = t'[Y_n]$$

- Why?
  - Let you specify constraints

Functional Dependencies

- How to derive other FDs: Armstrong Axioms

Reflexivity:

$$Y \subseteq X \implies X \longrightarrow Y$$

e.g., {sid} ⊆ {sid,course}, so, {sid,course} → {sid}

Augmentation:

$$X \longrightarrow Y \implies XZ \longrightarrow YZ$$

e.g., sid → name, so, sid,course → name,course

Transitivity:

$$X \longrightarrow Y \wedge Y \longrightarrow Z \implies X \longrightarrow Z$$

e.g., sid,course → sid and sid → name, so sid,course → name

Functional Dependencies

- Canonical cover:
  - Given a set **F** of FDs $\{f_1, \ldots, f_n\}$
  - Canonical cover: smallest $\mathbf{F_c}$ such that $\mathbf{F_c^+ = F^+}$
  - Let you find **super key**

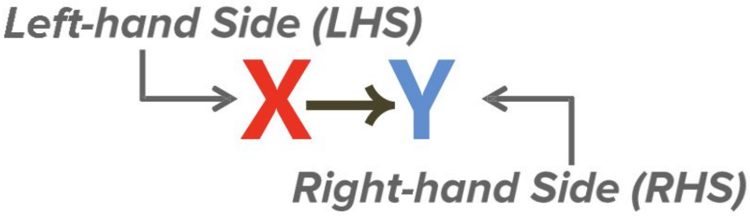
F

sid,course → grade
sid → name, address
sid,name → name, address
sid,course → grade,name

Functional Dependencies

How to compute the canonical cover?

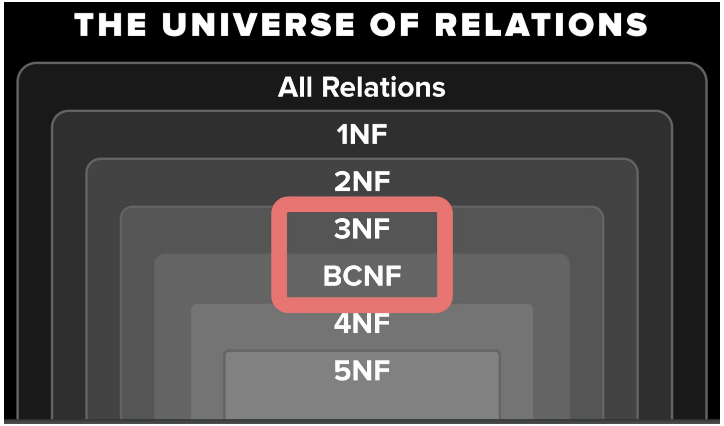
1. Put FDs in into standard forms. i.e. all FDs have single attribute on the RHS
2. Minimize the LHS of each FD. i.e. there are some other FDs among the LHS attributes
3. Delete redundant FDs. i.e. duplicate, or can be derived through transitivity.



Extraneous: removing the attributes doesn't affect the closure

Normalization

- Data normalization = refining relations
  - Given relation R, with FDs
  - Identify bad FDs
    - Bad →redundancy
  - Decompose it to better relations
    - "Better" = some normal forms
    - Desirable ones: 3NF and BCNF
  - Any decomposition?
    - Lossless-join
    - Dependency preserving



Decomposition

loans(bname,bcity,assets,cname,loanId,amt)

bname	bcity	assets	cname	loanId	amt
Downtown	Pittsburgh	$9M	Andy	L-17	$1000
Downtown	Pittsburgh	$9M	Oswin	L-23	$2000
Compton	Los Angeles	$2M	Andy	L-93	$500
Downtown	Pittsburgh	$9M	Damian	L-17	$1000

Lossless Joins

→ Motivation: Avoid information loss.

→ Goal: No noise introduced when reconstituting universal relation via joins.

→ Test: At each decomposition $R=(R_1 \cup R_2)$, check whether $(R_1 \cap R_2) \rightarrow R_1$ or $(R_1 \cap R_2) \rightarrow R2$.

bname	assets	cname	loanId
Downtown	$9M	Andy	L-17
Downtown	$9M	Oswin	L-23
Compton	$2M	Andy	L-93
Downtown	$9M	Damian	L-17

⋈

loanId	bcity	amt
L-17	Pittsburgh	$1000
L-23	Pittsburgh	$2000
L-93	Los Angeles	$500

✓

bname	bcity	assets	cname	loanId	amt
Downtown	Pittsburgh	$9M	Andy	L-17	$1000
Downtown	Pittsburgh	$9M	Oswin	L-23	$2000
Compton	Los Angeles	$2M	Andy	L-93	$500
Downtown	Pittsburgh	$9M	Damian	L-17	$1000

Decomposition

To test whether the decomposition $R=\{R_1,\dots,R_n\}$ preserves the FD set F:

→ Compute F^+

→ Compute G as the union of the set of FDs in F^+ that are covered by $\{R_1,\dots,R_n\}$

→ Compute G^+

→ If $F^+ = G^+$, then $\{R_1,\dots,R_n\}$ is Dependency Preserving

R(A,B,C,D)

$F+ = \{A \rightarrow B,\ AB \rightarrow D,\ A \rightarrow D,\ C \rightarrow D\}$

Dependency Preserving?

R1(A,B,C) R2(C,D)

How about?

R1(A,B,D) R2(C,D)

Normal Forms

- 1NF:
  - Flat schema
- 2NF:
  - All non-key attributes fully dependent of candidate key
- BCNF:
  - For any non-trivial $X \rightarrow Y$ in F+, then X is a super key
- 3NF:
  - For any non-trivial $X \rightarrow Y$ in F+, then
    - X is a super key
    - Or, Y is part of a candidate key

BCNF

Given a relation **R** and a FD set **F**:

Step #1 – Compute **F+**

Step #2 – **Result** $\leftarrow$ {**R**}

Step #3 – While $R_i \in$ **Result** not in BCNF, do:

→ **(a)** Choose **(X→Y) ∈ F+** such that **(X→Y)** is covered by R_i and $X \nrightarrow R_i$

→ **(b)** Decompose R_i on **(X→Y)**:

$\boxed{R_{i,1} \leftarrow X \cup Y}$ ← $R_{i,1}$ includes Y

$\boxed{R_{i,2} \leftarrow R_i - Y}$ ← $R_{i,2}$ does not include Y

$\text{Result} \leftarrow (\text{Result} - \{R_i\}) \cup \{R_{i,1}, R_{i,2}\}$

BCNF Decomposition

- More optimized algorithm:
  - Fewer sub-tables
  - We’ll be using this instead of the simple one

$Normalize(R)$
  $C \leftarrow$ the set of all attributes in R
  find X **s.t.** $X^+ \neq X$ **and** $X^+ \neq C$
  if X is not found
  then “R is in BCNF”
  else
    decompose R into $R_1(X^+)$ and $R_2((C - X^+) \cup X)$
    $Normalize(R_1)$
    $Normalize(R_2)$

Exercise 1

Given the relation R(A,B,C,D,E) with the following FDs:

$C \rightarrow E$

$BD \rightarrow AE$

$A \rightarrow BC$

Find a candidate key.

$BD \rightarrow ABDCE$

Exercise 1

Given the relation R(A,B,C,D,E) with the following FDs:

$C \rightarrow E$

$BD \rightarrow AE$

$A \rightarrow BC$

Find a candidate key.

What is the minimal cover?

Step 1. $C \rightarrow E$, $BD \rightarrow A$, $BD \rightarrow E$, $A \rightarrow B$, $A \rightarrow C$

Step 2. Same as above, can't further simplify LHSs

Step 3. $BD \rightarrow E$ is redundant, $BD \rightarrow A$, $A \rightarrow C$, $C \rightarrow E \Rightarrow_{\text{trans}} BD \rightarrow E$

$C \rightarrow E$, $BD \rightarrow A$, $A \rightarrow B$, $A \rightarrow C$

Exercise 1

Given the relation R(A,B,C,D,E) with the following FDs:

$C \rightarrow E$

$BD \rightarrow AE$

$A \rightarrow BC$

Find a candidate key.

What is the attribute closure ?

$C \rightarrow E$, $BD \rightarrow A$, $A \rightarrow B$, $A \rightarrow C$

$C+ = \{ C, E \}$

$BD+ = \{ B, D, A, C, E \}$

$A+ = \{A , B, C, E\}$

Exercise 1

Given the relation R(A,B,C,D,E) with the following FDs:

$C \rightarrow E$

$BD \rightarrow AE$

$A \rightarrow BC$

Find a candidate key.

$BD \rightarrow ABDEC$

$AD \rightarrow ADBCE$

Exercise 2

Given the relation R(A,B,C,D,E) with the following FDs:

$A \rightarrow C$

$C \rightarrow DE$

$B \rightarrow AE$

Candidate keys

Fc = { $A \rightarrow C$, $C \rightarrow D$, $C \rightarrow E$, $B \rightarrow A$} (Note $B \rightarrow E$ is redundant)

A+ = {A, C, D, E}, C+ = {C , D, E }, B+ = {A, B, C, D, E}, D+ = {D}, E+ = {E}

Thus, B is the candidate key.

Exercise 2

Given the relation R(A,B,C,D,E) with the following FDs:

$A \rightarrow C$

$C \rightarrow DE$

$B \rightarrow AE$

Is R in 2NF?

$B \rightarrow AECDB$

$A \rightarrow C$ and $C \rightarrow DE$ are fine, because neither A or C is part of a candidate key.

Thus R is in 2NF!

Is R in BCNF?

R is not in BCNF, because A is not a super key!

Exercise 2

Given the relation R(A,B,C,D,E) with the following FDs:

$A \rightarrow C$

$C \rightarrow DE$

$B \rightarrow AE$

Is R in 2NF?

Yes, B is candidate key

Is R in BCNF?

No, $A \rightarrow C$, but A is not a super key

Exercise 3

Given the relation R(A,B,C,D,E,F,G) with the following FDs:

$E \rightarrow C$

$G \rightarrow AD$

$B \rightarrow E$

$C \rightarrow BF$

What attributes that E functionally determines?

$E \rightarrow ECBF$

Decompose R into BCNF, using the more optimized algorithm.

Exercise 3

Given the relation R(A,B,C,D,E,F,G) with the following FDs:

$E \rightarrow C$

$G \rightarrow AD$

$B \rightarrow E$

$C \rightarrow BF$

What attributes that E functionally determines?

$E \rightarrow ECBF$

$E^+ = \{B,C,E,F\}$

Decompose R into BCNF, using the more optimized algorithm.

Exercise 3

Given the relation R(A,B,C,D,E,F,G) with the following FDs:

$E \rightarrow C$

$G \rightarrow AD$

$B \rightarrow E$

$C \rightarrow BF$

Decompose R into BCNF, using the less optimized algorithm.

- R(A,B,C,D,E,F,G)
  - **R1(E,C)**
  - R2(A,B,D,E,F,G)
    - **R21(G,A,D)**
    - R22(B,E,F,G)
      - **R221(B,E)**
      - R222(B,F,G) <= this is not in BCNF, because of $B \rightarrow E \rightarrow C \rightarrow F$
        - **R2221(B,F)**
        - **R2222(B,G)**

Exercise 3

Given the relation R(A,B,C,D,E,F,G) with the following FDs:

$E \rightarrow C$

$G \rightarrow AD$

$B \rightarrow E$

$C \rightarrow BF$

Decompose R into BCNF, using the more optimized algorithm.

R(A,B,C,D,E,F,G)
- E+ = {B,C,E,F}
- **R1(B,C,E,F)**
- R2(A,D,E,G)
  - G+ = {G,A,D}
  - **R21(G,A,D)**
  - **R22(G,E)**

Exercise 4: Optional

Given the relation R(A,B,C,D,E,F,G) with the following FDs:

$E \rightarrow C$

$G \rightarrow AD$

$B \rightarrow E$

$C \rightarrow BF$

Decompose R into 3NF with FD preserved.

Exercise 4: Optional

Given the relation R(A,B,C,D,E,F,G) with the following FDs:

$E \rightarrow C$

$G \rightarrow AD$

$B \rightarrow E$

$C \rightarrow BF$

Decompose R into 3NF with FD preserved. From exercise 3, we have

R1(B,C,E,F), R21(G,A,D), R22(G,E)

Compute minimal cover, $E \rightarrow C$, $G \rightarrow A$, $G \rightarrow D$, $B \rightarrow E$, $C \rightarrow B$, $C \rightarrow F$
F1 = $\{E \rightarrow C, B \rightarrow E, C \rightarrow B, C \rightarrow F\}$
F21 = $\{G \rightarrow A, G \rightarrow D\}$

Hence the above is also 3NF and FDs are preserved.
Consider an extra $C \rightarrow A$, which is not preserved by the above decomposition, we need to add R3(A,C)