Databases

TDA357/DIT621– LP3 2023

Lecture 7

Ana Bove

(much of the material is based on material from
both Thomas Hallgren and Jonas Duregård)

February 2nd 2023

Recall Last Lecture

- Functional dependencies:
  - $X \to A$ if any two rows that agree on the values in X also agree on the values in A;
  - Properties: reflexivity, transitivity and augmentation;
  - Closures: $X^+ = \{a \mid X \to a\}$;
  - Superkeys ($S \subseteq X^+$) and keys (minimal superkey);
  - Minimal basis/cover;
- Given $R(S)$, the FD $X \to A$ is a BCNF-violation if X is not a superkey ($X \subseteq S \subseteq X^+$);
- Given a (minimal basis) set of FD, $R(S)$ in BCNF if for all non-trivial $X \to A$, X is a superkey ($X \subseteq S \subseteq X^+$) (alt. if there is no BCNF-violation);
- Normalisation algorithm: given $R(S)$ and a BCNF-violation $X \to A$, we split $R(S)$ into $R_1(X^+)$ and $R_2(X \cup (S - X^+))$ and recursively normalise them (R and S become now irrelevant).

Overview of Today's Lecture

- FD examples;
- BCNF decomposition and dependency preservation;
- Multivalued dependencies (MVD);
- 4NF and its normalisation algorithm;
- MVD examples.

Example: FD and BCNF-violation

Consider a relation schema $R(a, b, c, d, e, f)$ with the following FD. Identify 3 different BCNF-violations and compute the closure of the left-hand side of the dependency.

$$
\begin{aligned}
a &\to b \\
a\,c &\to d \\
a\,c &\to e \\
a\,c &\to f \\
d &\to a \\
d &\to c \\
a\,f &\to e \\
e &\to a
\end{aligned}
$$

$a \to b$: $\{a\}^+ = \{a, b\}$.

$a\,f \to e$: $\{a, f\}^+ = \{a, b, e, f\}$.

$e \to a$: $\{e\}^+ = \{a, b, e\}$.

For the rest we have that $\{a, c\}^+ = \{d\}^+ = \{a, b, c, d, e, f\}$.

Example: FD and Data

Consider the schema $R(a, b, c, d)$ with the following FD:

$$a \to b$$
$$b \to a$$
$$c \to d$$

Construct a table with these FD and no other FD with only one attribute on the left side.
Give 2 possible sets of keys.

a	b	c	d
1	7	2	2
1	7	3	3
4	4	2	2
5	6	5	2

Keys: $\{a, c\}$ or $\{b, c\}$

The solution could look like this:

a	b	c	d
1	1	2	2
1	1	2	3

This solution is however incorrect, since it for example violates $c \to d$. Also, $a \to c$ is a FD but shouldn't.

The first two lines make sure that there are no dependencies from a and/or b to c or d.
The third line does the same in the other direction.
The last makes sure that $d \to c$ does not hold.
Observe that a and b do not need to be equal!

BCNF Decomposition and Dependency Preservation

Given the FD: $x \to z$, $y\,z \to w$ normalise $R(x, y, z, w)$

- We obtain the derived FD $x\,y \to w$:
  - from $x \to z$ we get $x\,y \to z$ by augmentation;
  - we have $y \to y$ by reflexivity and then $x\,y \to y$ by augmentation;
  - $x\,y \to z$ and $x\,y \to y$ give us $x\,y \to y\,z$;
  - by transitivity of $x\,y \to y\,z$ and $y\,z \to w$ we get $x\,y \to w$.

- Using $x \to z$ to decompose and the FD (including derived ones) to determine the keys, we get:
  - $R_1(\underline{x}, z)$ in BCNF
  - $R_2(\underline{x}, \underline{y}, w)$ also in BCNF

- Note that $y\,z \to w$ is not guarantee!

Multivalued Dependencies (MVD)

Consider the following example with courses, books, authors and teachers such that:

- Databases has a book and two teachers;
- Automata has a teacher and two books.

course	**book**	**author**	**teacher**
Databases	DTCB	Ullman	Jonas
Databases	DTCB	Ullman	Ana
Databases	DTCB	Widom	Jonas
Databases	DTCB	Widom	Ana
Automata	Finite Automata	Carroll	Nisse
Automata	Automata Theory	Hopcroft	Nisse

This relation has no non-trivial FD...

... however, there is risk for update and deletion anomalies!

Multivalued Dependencies (Cont.)

course	book	author	teacher
Databases	DTCB	Ullman	Jonas
Databases	DTCB	Ullman	Ana
Databases	DTCB	Widom	Jonas
Databases	DTCB	Widom	Ana
Automata	Finite Automata	Carroll	Nisse
Automata	Automata Theory	Hopcroft	Nisse

- Course determines the set of teachers, which are independent of the other values (books and authors);
- Course determines the set of books and authors, which are independent of the other values (teachers);
- There might be *multiple values* of teachers/books and authors related to a course.

$$\text{course} \twoheadrightarrow \text{teacher}$$
$$\text{course} \twoheadrightarrow \text{book author}$$

Multivalued Dependencies (MVD)

Informally, a MVD $X \twoheadrightarrow Y$ holds in a relation for $R(a_1, ..., a_n)$ if whenever
there are tuples $\langle \overline{x}, \overline{y_1}, \overline{z_1} \rangle$ *and* $\langle \overline{x}, \overline{y_2}, \overline{z_2} \rangle$ then
there are also tuples $\langle \overline{x}, \overline{y_1}, \overline{z_2} \rangle$ *and* $\langle \overline{x}, \overline{y_2}, \overline{z_1} \rangle$ with

- $\overline{x}$ are the values for the attributes in X;
- $\overline{y_1}$ and $\overline{y_2}$ are the values for the attributes in Y;
- and $\overline{z_1}$ and $\overline{z_2}$ are the values for the rest of the attributes (those in $Z = \{a_1, ..., a_n\} - X - Y$).

That is, we have all possible combination of values for Y ($\overline{y_1}$ and $\overline{y_2}$) and Z ($\overline{z_1}$ and $\overline{z_2}$) for those rows with the same value in X!

MVD are not easy to figure out nor to check!

Multivalued Dependencies (MVD)

Definition: A *MVD $X \twoheadrightarrow Y$ holds* in a relation for $R(a_1, ..., a_n)$ if for any tuples t_1 and t_2 such that $t_1.X = t_2.X$, there exists tuples t_3 and t_4 such that

- $t_1.X = t_2.X = t_3.X = t_4.X$;
- $t_1.Y = t_3.Y$ and $t_2.Y = t_4.Y$;
- and for $Z = \{a_1, ..., a_n\} - X - Y$ (the rest of the attributes), then $t_1.Z = t_4.Z$ and $t_2.Z = t_3.Z$.

t_3 and t_4 could actually be the same as t_1 and t_2 for some of the choices of t_1 and t_2!

Multivalued Dependencies (MVD): Back to the Example

row	course	book	author	teacher
r_1	Databases	DTCB	Ullman	Jonas
r_2	Databases	DTCB	Ullman	Ana
r_3	Databases	DTCB	Widom	Jonas
r_4	Databases	DTCB	Widom	Ana
r_5	Automata	Finite Automata	Carroll	Nisse
r_6	Automata	Automata Theory	Hopcroft	Nisse

course $\twoheadrightarrow$ teacher: $X = \{\text{course}\}$, $Y = \{\text{teacher}\}$, $Z = \{\text{book, author}\}$ (the rest)

$r_1.X = r_2.X = r_3.X = r_4.X =$ Databases

$r_1.Y = r_3.Y =$ Jonas and $r_2.Y = r_4.Y =$ Ana

$r_1.Z = r_2.Z =$ DTCB+Ullman and $r_3.Z = r_4.Z =$ DTCB+Widmon

course $\twoheadrightarrow$ book author: $X = \{\text{course}\}$, $Y = \{\text{book, author}\}$, $Z = \{\text{teacher}\}$ (the rest)

$r_1.X = r_2.X = r_3.X = r_4.X =$ Databases

$r_1.Y = r_2.Y =$ DTCB+Ullman and $r_3.Y = r_4.Y =$ DTCB+Widmon

$r_1.Z = r_3.Z =$ Jonas and $r_2.Z = r_4.Z =$ Ana

For the rows with same value in X, the cartesian product with values of Y and Z should be in the table!

Properties of Multivalued Dependencies

Let $R(a_1, \ldots, a_n)$ be a relational schema, $S = \{a_1, \ldots, a_n\}$ the set of attributes of R, and $X, Y, V, W \subseteq S$.

Replication: if $X \to Y$ then $X \twoheadrightarrow Y$;

Complementation: If $X \twoheadrightarrow Y$ then $X \twoheadrightarrow S - Y$;

Transitivity: If $X \twoheadrightarrow Y$ and $Y \twoheadrightarrow W$ then $X \twoheadrightarrow W - Y$;

Augmentation: If $X \twoheadrightarrow Y$ and $V \subseteq W$ then $X\,W \twoheadrightarrow Y\,V$;

Trivial: $X \twoheadrightarrow Y$ is trivial if $Y \subseteq X$ or if $S \subseteq X \cup Y$.

Observe the difference in the definition of a trivial dependency!

4th Normal Form (4NF)

Definition: A relational schema $R(S)$ with set of attributes $S = \{a_1, ..., a_n\}$ is in *4th normal form (4NF)* if

- The relational schema is in BCNF;
- For all non-trivial MVD $X \twoheadrightarrow Y$, X is a superkey
  (that is, it determines –following the FD– all the attributes in S).

  If X is not a superkey, then $X \twoheadrightarrow Y$ is a *4NF-violation*.

4NF Normalisation Algorithm

To *normalise* a relational schema $R(S)$ with $S = \{a_1, ..., a_n\}$:

- Find a 4NF-violation: that is, a non-trivial MVD $X \twoheadrightarrow Y$ ($Y \not\subseteq X$ or $S \not\subseteq X \cup Y$) such that X is not a superkey ($S \not\subseteq X^+ = \{a \mid X \to a\}$);
- If there is no such MVD then R is already in 4NF;
- Otherwise decompose $R(S)$ into $R_1(X \cup Y)$ and $R_2(S - Y)$ and normalise them both.

  Note: Again, R is not of interest anymore since it has been replaced by R_1 and R_2 and the set S of attributes is replaced by $X \cup Y$ and $S - Y$ respectively.

Decomposition of Data

R: Courses

course	book	author	teacher
Databases	DTCB	Ullman	Jonas
Databases	DTCB	Ullman	Ana
Databases	DTCB	Widom	Jonas
Databases	DTCB	Widom	Ana
Automata	Finite Automata	Carroll	Nisse
Automata	Automata Theory	Hopcroft	Nisse

course $\twoheadrightarrow$ teacher

R_1: CourseTeachers

course	teacher
Databases	Jonas
Databases	Ana
Automata	Nisse

R_2: CourseBooks

course	book	author
Databases	DTCB	Ullman
Databases	DTCB	Widom
Automata	Finite Automata	Carroll
Automata	Automata Theory	Hopcroft

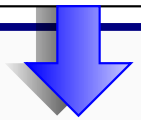
Lossless Join: All Data Back!

R_1: CourseTeachers

course	teacher
Databases	Jonas
Databases	Ana
Automata	Nisse

R_2: CourseBooks

course	book	author
Databases	DTCB	Ullman
Databases	DTCB	Widom
Automata	Finite Automata	Carroll
Automata	Automata Theory	Hopcroft



Query: R_1 NATURAL JOIN R_2

R: Courses

course	book	author	teacher
Databases	DTCB	Ullman	Jonas
Databases	DTCB	Ullman	Ana
Databases	DTCB	Widom	Jonas
Databases	DTCB	Widom	Ana
Automata	Finite Automata	Carroll	Nisse
Automata	Automata Theory	Hopcroft	Nisse

What about the Primary Keys?

R_1: CourseTeachers

course	teacher
Databases	Jonas
Databases	Ana
Automata	Nisse

R_2: CourseBooks

course	book	author
Databases	DTCB	Ullman
Databases	DTCB	Widom
Automata	Finite Automata	Carroll
Automata	Automata Theory	Hopcroft

$R_1(\underline{\text{course}}, \underline{\text{teacher}})$

$R_2(\underline{\text{course}}, \underline{\text{book}}, \underline{\text{author}})$

MVD Example

Identify a non-trivial MVD that holds in the data and is a 4NF-violation, and provide a schema in 4NF based on that MVD.

code	examDate	student	classification
TDA357	20200115	Emil	CS
TDA357	20200115	Emilia	CS
TDA357	20200320	Emil	CS
TDA357	20200115	Emil	DB
TDA357	20200115	Emilia	DB
TDA357	20200320	Emil	DB
XYZ123	20200115	Emil	CS

code $\twoheadrightarrow$ classification (or equivalently: code $\twoheadrightarrow$ examDate, student)

R_1 (<u>code</u>, <u>classification</u>)
R_2 (<u>code</u>, <u>examDate</u>, <u>student</u>)

MVD Example (from exam)

Consider the relation schema `R(g,h,i,j,k)` and the MVD `g h` $\twoheadrightarrow$ `i`.

1. Complete the following table so that the dependency is valid

g	h	i	j	k
0	0	1	2	2
0	0	1	3	3
0	0	4	2	2
5	5	1	2	2
5	5	1	3	3

2. Convert the schema into a relation schema in 4NF.

MVD Example (from exam, cont.)

1.

g	h	i	j	k
0	0	1	2	2
0	0	1	3	3
0	0	4	2	2
0	0	4	3	3
5	5	1	2	2
5	5	1	3	3

2. Using this MVD we obtain the schemas R1(g,h,i) and R2(g,h,j,k).

Database Design Workflow

- ER-modelling and FD are two complementary (but compatible) methods to find a good relational schema;
- FD can find certain things that ER-diagram cannot, and the other way around;

  FD are particularly useful to find secondary keys (UNIQUE constraints);
- A good database design workflow is:
  - Start by creating the ER-model from the domain description;
  - Obtain a relational schema from the ER-model;
  - Go back to the domain description and identify additional FD;
  - Check if they hold in the ER-schema and add secondary keys when appropriate;
  - Apply BCNF decomposition when needed;
  - Identify MVD and apply 4NF decomposition when needed.

Other Normal Forms

1NF (1970): Only atomic values in the tables;

2NF (1971): 1NF + no partial dependencies,
(a FD $X \to Y$ is a partial dependency if $Z \to Y$ for some $Z \subseteq X$);

3NF (1971): 2NF + no transitive dependencies from the primary key to non-key attributes,
Schemas that result from ER-modelling are in 3NF;

BCNF a.k.a. 3½NF (1974): for each non-trivial functional dependency $X \to Y$, X is a superkey;
(BCNF implies 2NF and also 3NF!);

4NF (1977): BNCF + no MVD violation;

5NF, 6NF, ...

Overview of Next Lecture

- Functions;
- Triggers:
  - On tables;
  - On views;
- Example.

Reading:

Book: chapter 7.5

Notes: chapter 7.4.4–7.4.6