Data Management Systems Introduction to Design Theory

Matteo Devigili

June 07^{th}, 2022

Agenda

- Functional Dependencies
- Data Anomalies
- Normal Forms:
  1. BCNF
  2. Others: 1NF, 3NF

Functional dependencies

"If two tuples of R agree on all of the attributes $A_1, A_2, \ldots, A_n$ then they must also agree on all of another list of attributes $B_1, B_2, \ldots, B_m$. We write this FD formally as $A_1, A_2, ..., A_n \rightarrow B_1, B_2, ..., B_m$ and say that:

- *$A_1, A_2, \ldots, A_n$ functionally determine $B_1, B_2, ..., B_m$"*

Garcia-Molina, Ullman, Widom 2008

Example

Name	Year	Weeks	Degree
NLP	2021/2022	7	Business Analytics
DMS	2021/2022	6	Business Analytics
DMS	2021/2022	6	Actuarial Science
DMS	2021/2022	6	Actuarial Management
D-Viz	2021/2022	6	Business Analytics
D-Viz	2021/2022	6	Actuarial Management
DMS	2020/2021	2	Business Analytics
D-Viz	2020/2021	4	Business Analytics

What is the **FD**?

Example

Name	Year	Weeks	Degree
NLP	2021/2022	7	Business Analytics
DMS	2021/2022	6	Business Analytics
DMS	2021/2022	6	Actuarial Science
DMS	2021/2022	6	Actuarial Management
D-Viz	2021/2022	6	Business Analytics
D-Viz	2021/2022	6	Actuarial Management
DMS	2020/2021	2	Business Analytics
D-Viz	2020/2021	4	Business Analytics

$$\textit{name year} \rightarrow \textit{weeks}$$

Example

Name	Year	Weeks	Degree
NLP	2021/2022	7	Business Analytics
DMS	2021/2022	6	Business Analytics
DMS	2021/2022	6	Actuarial Science
DMS	2021/2022	6	Actuarial Management
D-Viz	2021/2022	6	Business Analytics
D-Viz	2021/2022	6	Actuarial Management
DMS	2020/2021	2	Business Analytics
D-Viz	2020/2021	4	Business Analytics

What about:

$$name\ year \rightarrow degree$$

Key

"We say a set of one or more attributes $\{A_1, A_2, \ldots, A_n\}$ *is a key for a relation* R *if:*

1. *Those attributes functionally determine* **all other attributes** *of the relation. That is, it is impossible for two distinct tuples of* R *to agree on all of* $A_1, A_2, \ldots, A_n$.
2. *No proper subset of* $\{A_1, A_2, \ldots, A_n\}$ *functionally determines all other attributes of* R*; i.e., a key must be* **minimal**."

Garcia-Molina, Ullman, Widom 2008

Superkey

A Superkey satisfies the first condition:

1. *Those attributes functionally determine all other attributes of the relation. That is, it is impossible for two distinct tuples of R to agree on all of $A_1, A_2, \ldots, A_n$."*

Garcia-Molina, Ullman, Widom 2008

Example

Name	Year	Weeks	Degree	Count
NLP	2021/2022	7	Business Analytics	57
DMS	2021/2022	6	Business Analytics	45
DMS	2021/2022	6	Actuarial Science	15
DMS	2021/2022	6	Actuarial Management	9
D-Viz	2021/2022	6	Business Analytics	58
D-Viz	2021/2022	6	Actuarial Management	19
DMS	2020/2021	2	Business Analytics	10
D-Viz	2020/2021	4	Business Analytics	80

the key: {name, year, degree}

possible superkey: {name, year, weeks, degree}

Functional Dependencies

So what?

- Look for FDs;
- Use FDs to design better relation schemas;
- Pay attention to local FDs!

Data Anomalies

- *Redundancy*: unnecessary repetition of information;
- *Update Anomalies*: we may replace information of a tuple, but forget about others;
- *Deletion Anomalies*: after deleting, we may accidentally lose some other information.

Example

Redundancy

Name	Year	Term	Weeks	Degree
NLP	2021/2022	T3	7	Business Analytics
DMS	2021/2022	T3	6	Business Analytics
DMS	2021/2022	T3	6	Actuarial Science
DMS	2021/2022	T3	6	Actuarial Management
D-Viz	2021/2022	T1	6	Business Analytics
D-Viz	2021/2022	T1	6	Actuarial Management
DMS	2020/2021	T2	2	Business Analytics
D-Viz	2020/2021	T2	4	Business Analytics

Example

Update Anomalies

Name	Year	Term	Weeks	Degree
NLP	2021/2022	T3	7	Business Analytics
DMS	2021/2022	T3	5	Business Analytics
DMS	2021/2022	T3	6	Actuarial Science
DMS	2021/2022	T3	6	Actuarial Management
D-Viz	2021/2022	T1	6	Business Analytics
D-Viz	2021/2022	T1	6	Actuarial Management
DMS	2020/2021	T2	2	Business Analytics
D-Viz	2020/2021	T2	4	Business Analytics

Example

Deletion Anomalies

Name	Year	Term	Weeks	Degree
NLP	2021/2022	T3	7	Business Analytics
DMS	2021/2022	T3	6	Business Analytics
DMS	2021/2022	T3	6	Actuarial Science
DMS	2021/2022	T3	6	Actuarial Management
D-Viz	2021/2022	T1	6	Business Analytics
D-Viz	2021/2022	T1	6	Actuarial Management
DMS	2020/2021	T2	2	Business Analytics
D-Viz	2020/2021	T2	4	Business Analytics

Boyce-Codd Normal Form

"A relation R is in BCNF if and only if: whenever there is a nontrivial FD $A_1, A_2, ..., A_n \rightarrow B_1, B_2, ..., B_m$ for R, it is the case that $A_1, A_2, \ldots, A_n$ is a superkey for R

That is, the left side of every nontrivial FD must be a superkey. Recall that a superkey need not be minimal. Thus, an equivalent statement of the BCNF condition is that the left side of every nontrivial FD must contain a key."

Garcia-Molina, Ullman, Widom 2008

Example

BCNF

Name	Year	Term	Weeks	Degree	Count
NLP	2021/22	T3	7	Business Analytics	57
DMS	2021/22	T3	6	Business Analytics	45
DMS	2021/22	T3	6	Actuarial Science	15
DMS	2021/22	T3	6	Actuarial Management	9
D-Viz	2021/22	T1	6	Business Analytics	58
D-Viz	2021/22	T1	6	Actuarial Management	19
DMS	2020/21	T2	2	Business Analytics	10
D-Viz	2020/21	T2	4	Business Analytics	80

Any superkey in our R must contain the key $\{\text{name year degree}\}$.

- Still, $\{\text{name year}\} \rightarrow \{\text{term weeks}\}$ holds

Example

BCNF

Name	Year	Term	Weeks	Degree	Count
NLP	2021/22	T3	7	Business Analytics	57
DMS	2021/22	T3	6	Business Analytics	45
DMS	2021/22	T3	6	Actuarial Science	15
DMS	2021/22	T3	6	Actuarial Management	9
D-Viz	2021/22	T1	6	Business Analytics	58
D-Viz	2021/22	T1	6	Actuarial Management	19
DMS	2020/21	T2	2	Business Analytics	10
D-Viz	2020/21	T2	4	Business Analytics	80

In $\{\text{name year}\} \rightarrow \{\text{term weeks}\}$:

- $\{\text{name year}\}$ is not a superkey for R
- Hence, the existence of $\{\text{name year}\} \rightarrow \{\text{term weeks}\}$ violates BCNF

BCNF: Decomposition

A possible decomposition:

Name	Year	Term	Weeks
NLP	2021/2022	T3	7
DMS	2021/2022	T3	6
D-Viz	2021/2022	T1	6
DMS	2020/2021	T2	2
D-Viz	2020/2021	T2	4

Name	Year	Degree	Count
NLP	2021/2022	Business Analytics	57
DMS	2021/2022	Business Analytics	45
DMS	2021/2022	Actuarial Science	15
DMS	2021/2022	Actuarial Management	9
D-Viz	2021/2022	Business Analytics	58
D-Viz	2021/2022	Actuarial Management	19
DMS	2020/2021	Business Analytics	10
D-Viz	2020/2021	Business Analytics	80

Example

BCNF

Is this in BCNF?

title	*year*	*length*	*genre*	*studioName*	*starName*
Star Wars	1977	124	SciFi	Fox	Carrie Fisher
Star Wars	1977	124	SciFi	Fox	Mark Hamill
Star Wars	1977	124	SciFi	Fox	Harrison Ford
Gone With the Wind	1939	231	drama	MGM	Vivien Leigh
Wayne's World	1992	95	comedy	Paramount	Dana Carvey
Wayne's World	1992	95	comedy	Paramount	Mike Meyers

Garcia-Molina, Ullman, Widom 2008

Start and stop

BCNF

Start:

- Look for any FD that violates BCNF

Stop – trade-off:

- Elimination of Anomalies
- Recoverability of Information
- Preservation of Dependencies

Atomic Values

1NF

"the domain of an attribute must include only atomic (simple, indivisible) values and that the value of any attribute in a tuple must be a single value from the domain of that attribute"

Elmasri and Navathe 2016

Example

1NF

Name	Year	Term	Weeks	Degree
NLP	2021/2022	T3	7	{Business Analytics, Actuarial Science}
DMS	2021/2022	T3	6	{Business Analytics, Actuarial Science, Actuarial Management}
D-Viz	2021/2022	T1	6	{Business Analytics}
DMS	2020/2021	T2	2	{Actuarial Science, Actuarial Management}
D-Viz	2020/2021	T2	4	{Business Analytics}

This violates **1NF**

PostgreSQL

Arrays

PostgreSQL allows columns of a table to be defined as variable-length multidimensional arrays:

```
CREATE TABLE sal_emp (
name              text,
pay_by_quarter    integer[],
schedule          text[][]
);
```

Please see Tutorial 2 on GitHub.

3NF

"Whenever there is a nontrivial FD $A_1, A_2, ..., A_n \rightarrow B_1, B_2, ..., B_m$ *for R, either* $A_1, A_2, \ldots, A_n$ *is a superkey, or those of* $B_1, B_2, ..., B_m$ *that are not among the A's, are each a member of some key (not necessarily the same key)."*

Garcia-Molina, Ullman, Widom 2008

References

- Hector Garcia-Molina, Jeff Ullman, and Jennifer Widom. Database Systems: The Complete Book, Pearson, 2008.
- Ramez Elmasri and Shamkant Navathe. Fundamentals of Database Systems, Global Edition, Pearson Education Limited, 2016.