Data Management Systems
Introduction to Design Theory

Matteo Devigili

June 07^{th}, 2021

Agenda

- Functional Dependencies
- Data Anomalies
- Normal Forms:
    i. 1NF
    ii. BCNF
    iii. 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 $A_1, A_2, \ldots, A_n \rightarrow B_1, B_2, \ldots, B_m$ and say that:

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

Garcia-Molina, Ullman, Widom 2008

Example

Courses

Table: Courses

Name	Year	Weeks	Degree
NLP	2020/2021	7	Business Analytics
DMS	2020/2021	6	Business Analytics
DMS	2020/2021	6	Actuarial Science
DMS	2020/2021	6	Actuarial Management
D-Viz	2020/2021	6	Business Analytics
D-Viz	2020/2021	6	Actuarial Management
DMS	2019/2020	2	Business Analytics
D-Viz	2019/2020	4	Business Analytics

What is the **FD**?

Example

Courses

Table: Courses

Name	Year	Weeks	Degree
NLP	2020/2021	7	Business Analytics
DMS	2020/2021	6	Business Analytics
DMS	2020/2021	6	Actuarial Science
DMS	2020/2021	6	Actuarial Management
D-Viz	2020/2021	6	Business Analytics
D-Viz	2020/2021	6	Actuarial Management
DMS	2019/2020	2	Business Analytics
D-Viz	2019/2020	4	Business Analytics

$$name\ year \rightarrow weeks$$

Example

Courses

Table: Courses

Name	Year	Weeks	Degree
NLP	2020/2021	7	Business Analytics
DMS	2020/2021	6	Business Analytics
DMS	2020/2021	6	Actuarial Science
DMS	2020/2021	6	Actuarial Management
D-Viz	2020/2021	6	Business Analytics
D-Viz	2020/2021	6	Actuarial Management
DMS	2019/2020	2	Business Analytics
D-Viz	2019/2020	4	Business Analytics

What about:

$name\ year \rightarrow degree$

Keys & Superkeys

- "A *superkey* of a relation schema R = $\{ A_1, A_2, \ldots, A_n \}$ is a set of attributes $S \subseteq R$ with the property that no two tuples t_1 *and* t_2 in any legal relation state r of R will have $t_1[S] = t_2[S]$"
- "A **key** K is a superkey with the additional property that removal of any attribute from K will cause K not to be a superkey anymore" (a key has to be minimal)

Elmasri, Ramez, and Shamkant B. Navathe 2016

Example

Keys & Superkeys

Table: Courses

Name	Year	Weeks	Degree	Count
NLP	2020/2021	7	Business Analytics	57
DMS	2020/2021	6	Business Analytics	45
DMS	2020/2021	6	Actuarial Science	15
DMS	2020/2021	6	Actuarial Management	9
D-Viz	2020/2021	6	Business Analytics	58
D-Viz	2020/2021	6	Actuarial Management	19
DMS	2019/2020	2	Business Analytics	10
D-Viz	2019/2020	4	Business Analytics	80

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

the key: {name, year, degree}

Functional Dependencies

So what?

i. Look for FDs;
ii. Use FDs to design better relation schemas;
iii. 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

Table: Courses

Name	Year	Term	Weeks	Degree
NLP	2020/2021	T3	7	Business Analytics
DMS	2020/2021	T3	6	Business Analytics
DMS	2020/2021	T3	6	Actuarial Science
DMS	2020/2021	T3	6	Actuarial Management
D-Viz	2020/2021	T1	6	Business Analytics
D-Viz	2020/2021	T1	6	Actuarial Management
DMS	2019/2020	T2	2	Business Analytics
D-Viz	2019/2020	T2	4	Business Analytics

Example

Update Anomalies

Table: Courses

Name	Year	Term	Weeks	Degree
NLP	2020/2021	T3	7	Business Analytics
DMS	2020/2021	T3	5	Business Analytics
DMS	2020/2021	T3	6	Actuarial Science
DMS	2020/2021	T3	6	Actuarial Management
D-Viz	2020/2021	T1	6	Business Analytics
D-Viz	2020/2021	T1	6	Actuarial Management
DMS	2019/2020	T2	2	Business Analytics
D-Viz	2019/2020	T2	4	Business Analytics

Example

Deletion Anomalies

Table: Courses

Name	Year	Term	Weeks	Degree
~~NLP~~	~~2020/2021~~	~~T3~~	~~7~~	~~Business Analytics~~
~~DMS~~	~~2020/2021~~	~~T3~~	~~6~~	~~Business Analytics~~
DMS	2020/2021	T3	6	Actuarial Science
DMS	2020/2021	T3	6	Actuarial Management
~~D-Viz~~	~~2020/2021~~	~~T1~~	~~6~~	~~Business Analytics~~
D-Viz	2020/2021	T1	6	Actuarial Management
~~DMS~~	~~2019/2020~~	~~T2~~	~~2~~	~~Business Analytics~~
~~D-Viz~~	~~2019/2020~~	~~T2~~	~~4~~	~~Business Analytics~~

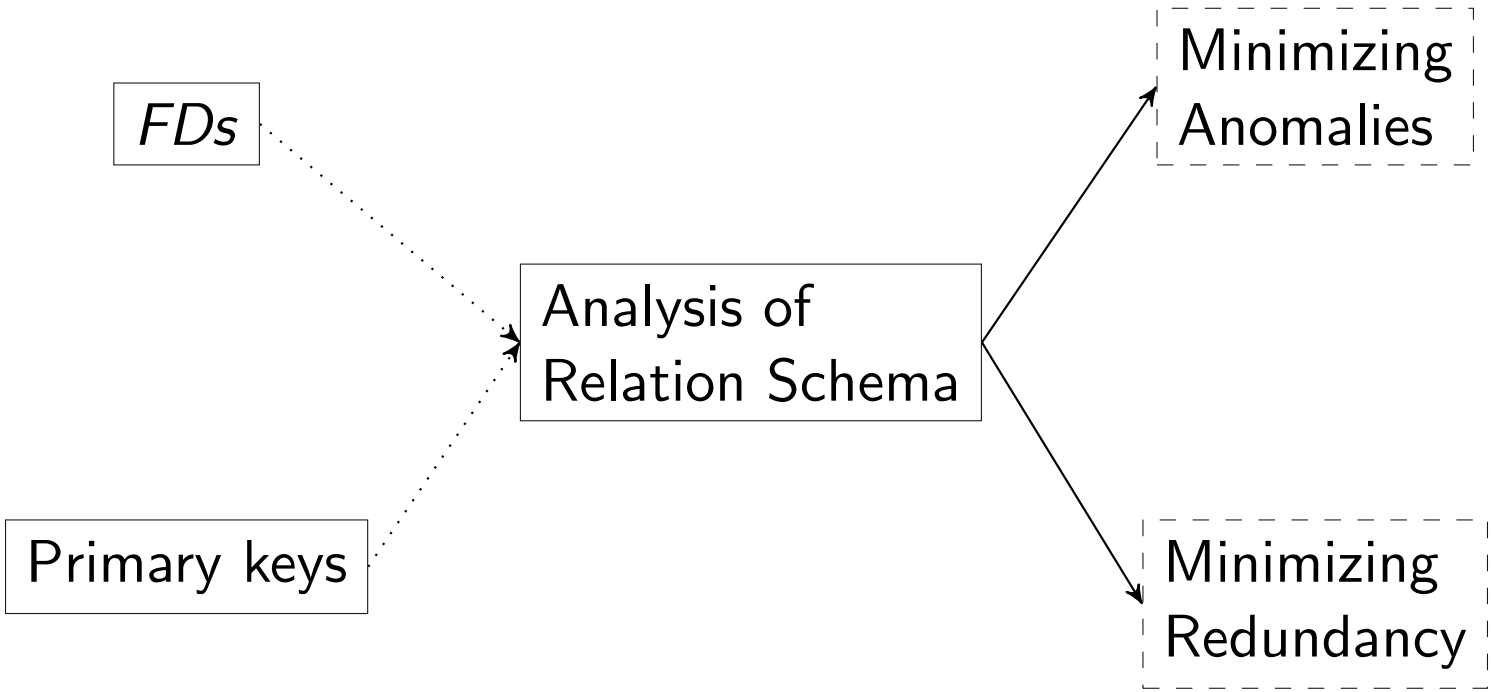
Decomposition

A possible decomposition:

Name	Year	Term	Weeks
NLP	2020/2021	T3	7
DMS	2020/2021	T3	6
D-Viz	2020/2021	T1	6
DMS	2019/2020	T2	2
D-Viz	2019/2020	T2	4

Name	Year	Degree
NLP	2020/2021	Business Analytics
DMS	2020/2021	Business Analytics
DMS	2020/2021	Actuarial Science
DMS	2020/2021	Actuarial Management
D-Viz	2020/2021	Business Analytics
D-Viz	2020/2021	Actuarial Management
DMS	2019/2020	Business Analytics
D-Viz	2019/2020	Business Analytics

Normalization of data



1NF

Atomic values

For the 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, Ramez, and Shamkant B. Navathe 2016

1NF

Example

Name	Year	Weeks	Degree
DMS	2019/2020	6	{Business Analytics, Actuarial Science, Actuarial Management}
D-Viz	2019/2020	10	Business Analytics
DMS	2018/2019	2	Business Analytics
D-Viz	2018/2019	4	{Business Analytics, Actuarial Management}

This violates **1NF**

1NF

PostgreSQL - Arrays

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

PostgreSQL 8.15

Boyce-Codd Normal Form

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

Garcia-Molina, Ullman, Widom 2008

BCNF

Example

Name	Year	Term	Weeks	Degree
DMS	2019/2020	T3	6	Business Analytics
DMS	2019/2020	T3	6	Actuarial Science
DMS	2019/2020	T3	6	Actuarial Management
D-Viz	2019/2020	T1	10	Business Analytics
DMS	2018/2019	T3	2	Business Analytics
D-Viz	2018/2019	T2	4	Business Analytics
D-Viz	2018/2019	T2	4	Actuarial Management

The (super) key is {name year degree}
So, the existence of {name year} $\rightarrow$ {term weeks} violates BCNF.

BCNF

Example

When do we stop decomposing?

Name	Year	Term	Weeks
DMS	2019/2020	T3	6
D-Viz	2019/2020	T1	10
DMS	2018/2019	T3	2
D-viz	2018/2019	T2	4

Name	Year	Degree
DMS	2019/2020	Business Analytics
DMS	2019/2020	Actuarial Science
DMS	2019/2020	Actuarial Management
D-Viz	2019/2020	Business Analytics
DMS	2018/2019	Business Analytics
D-Viz	2018/2019	Business Analytics
D-Viz	2018/2019	Actuarial Management

Decomposition

Trade-off

1. Elimination of Anomalies
2. Recoverability of Information
3. Preservation of Dependencies

Garcia-Molina, Ullman, Widom 2008

3NF

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

Garcia-Molina, Ullman, Widom 2008

Normal Forms

1^{st} Normal Form
2^{nd} Normal Form
3^{rd} Normal Form
Boyce-Codd Normal Form
4^{th} Normal Form
5^{th} Normal Form
...

References

1 Hector Garcia-Molina, Jeff Ullman, and Jennifer Widom. Database Systems: The Complete Book, Pearson, 2008.

2 Elmasri, Ramez, and Shamkant B. Navathe. Fundamentals of Database Systems, Global Edition, Pearson Education Limited, 2016.