

Лекция 3: нормализация

1. Реляционное представление

Э. Кодд: любое представление данных можно свести к совокупности **отношений**.

- **Отношение (relation)** - двумерные таблицы особого вида.
- У отношения есть **атрибуты** (столбцы) и **кортежи** (строки).
- У каждого атрибута есть имя.

Домены

- Каждый атрибут определяется некоторым допустимым набором значений
- **Домен** — множество значений, допустимых в определенном контексте.
- **Смысл домена:** если значения берутся из одного и того же домена, то они относятся к одному типу — эти значения можно сопоставить (сравнить)

Основные правила (1)

- Заголовок отношения — состоит из фиксированного множества атрибутов.
- Тело отношения — состоит из **меняющегося** во времени множества кортежей.

Основные правила (2)

- Каждый **кортеж** состоит из множества пар атрибут-значение, по **одной паре** для каждого атрибута из заголовка.
- Для любой заданной пары атрибут(A)-значение(v), v является значением из единственного домена D , который **связан** с атрибутом A .

Базовые понятия

- **Степень отношения** — это число его атрибутов (отношение степени один - унарное, степени два — бинарное, степени n — n -арное).
- **Кардинальное число** (мощность отношения) — это число его кортежей.

Пример

ID	Surname	Name	Birthday	Location
1	Иванов	Василий	1980-12-01	г. Москва
2	Георгиев	Сергей	1992-03-12	г. Санкт-Петербург
3	Васильев	Андрей	1987-10-14	г. Оренбург
7	Романов	Кирилл	1991-12-01	NULL

Терминология

Переменная отношения/
Имя таблицы

Атрибут/
Колонка

STUDENT

id	name	surname	gr_id
1	Григорий	Иванов	34
2	Григорий	Иванов	34
3	Иван	Сидоров	37

Заголовок

Тело

Кортеж/Строка

Операции реляц. алгебры

Реляционная алгебра — язык для определения новых отношений на основе существующих.

В реляционной алгебре определен ряд **операций** над отношениями.

Результат операции — **новое** отношение.

В операциях будут использоваться обозначения:

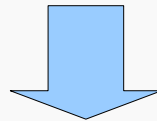
- R, S — отношения (таблицы)
- φ — предикат (условие), $\varphi_1 \wedge \varphi_2$ — составное условие

Операция выборки

- $\sigma_{\varphi}(R)$ — **операция выборки** — в результате операции формируется отношение на основе R , которое содержит только те строки (кортежи), которые удовлетворяют заданному предикату.

Операция выборки

```
SELECT * FROM STUDENTS WHERE  
    STUDENTS.GROUP = '3100' AND  
    STUDENTS.ID >= 150000;
```

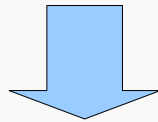

$$\sigma_{(\text{STUDENTS.GROUP}='3100') \wedge (\text{STUDENTS.ID} \geq 150000)}(\text{STUDENTS})$$

Проекция

- $\pi_{attr}(R)$ — проекция — в результате операции формируется новое отношение, содержащее только те атрибуты из R , которые были указаны в проекции:

Проекция

SELECT name, group FROM STUDENTS;



$\pi_{\text{name, group}}(\text{STUDENTS})$

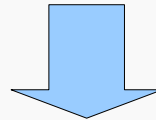
Соединение

$R \bowtie_{\theta} S$ — **соединение** (тета-соединение)

$$R \bowtie_{\theta} S = \sigma_{\theta}(R \times S)$$

Соединение

```
SELECT * FROM STUDENTS  
JOIN EXAMS ON STUDENTS.ID = EXAMS.STUD_ID;
```



STUDENTS ⋈_{STUDENTS.ID=EXAMS.STUD_ID} EXAMS

$$R \bowtie_{\theta} S \equiv S \bowtie_{\theta} R \text{ (коммутативность)}$$

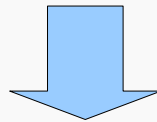
$$R \bowtie_{\theta} (S \bowtie_{\varphi} T) \equiv (R \bowtie_{\theta} S) \bowtie_{\varphi} T \text{ (ассоциативность)}$$

$$\sigma_{\theta \wedge \varphi} (R) \equiv \sigma_{\theta} (\sigma_{\varphi} (R))$$

...

Пример

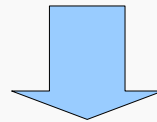
```
SELECT * FROM STUDENTS  
JOIN EXAMS ON STUDENTS.ID = EXAMS.STUD_ID  
WHERE  
    STUDENTS.GROUP = '3100' AND  
    STUDENTS.ID >= 150000;
```



?

Пример

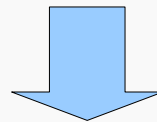
```
SELECT * FROM STUDENTS  
JOIN EXAMS ON STUDENTS.ID = EXAMS.STUD_ID  
WHERE  
  
    STUDENTS.GROUP = '3100' AND  
    STUDENTS.ID >= 150000;
```



$\sigma_{\text{STUDENTS.GROUP} = '3100' \wedge \text{STUDENTS.ID} \geq 150000} (\text{STUDENTS} \bowtie_{\text{STUDENTS.ID}=\text{EXAMS.STUD_ID}} \text{EXAMS})$

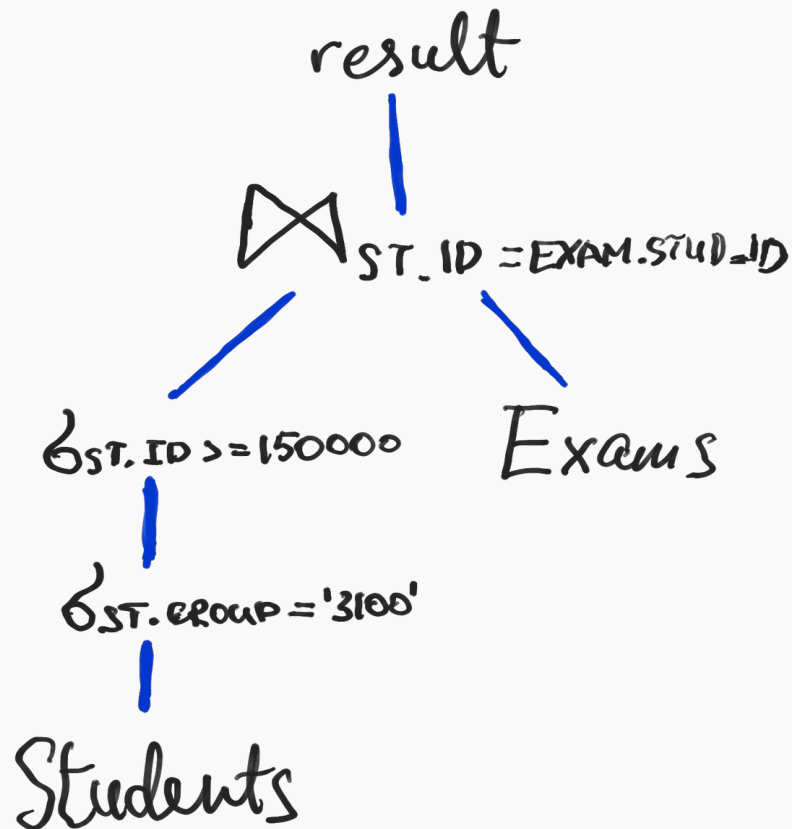
Сокращенная запись

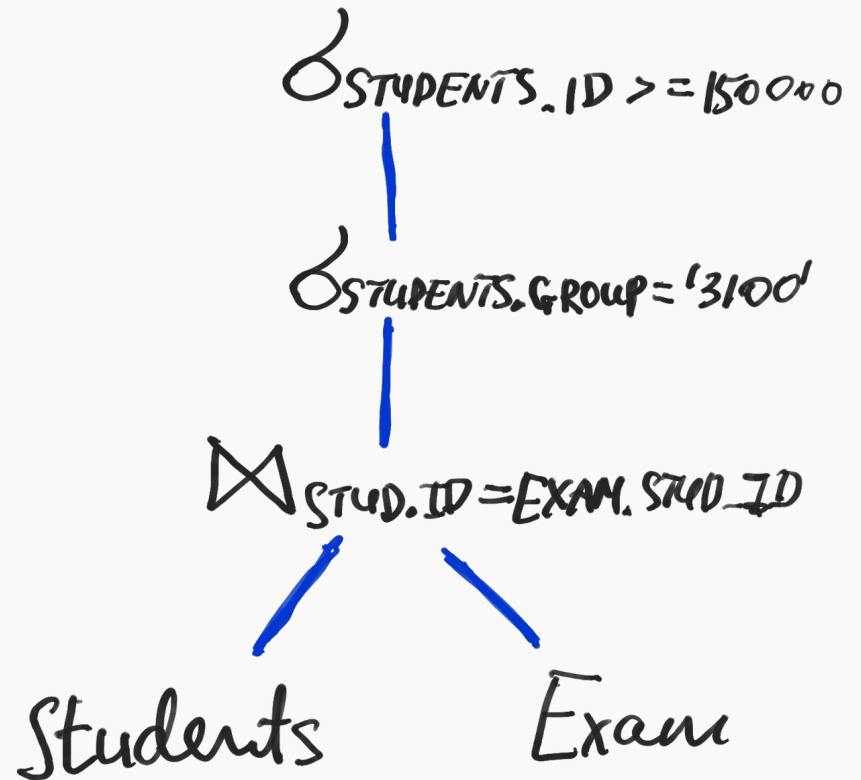
```
SELECT * FROM STUDENTS  
JOIN EXAMS ON STUDENTS.ID = EXAMS.STUD_ID  
WHERE  
    STUDENTS.GROUP = '3100' AND  
    STUDENTS.ID >= 150000;
```



$\sigma_{\text{STUDENTS.GROUP} \wedge \text{STUDENTS.ID}} (\text{STUDENTS} \bowtie_{\text{STUDENTS.ID}=\text{EXAMS.STUD_ID}} \text{EXAMS})$

План выполнения запроса





2. Нормализация

Как проверить полученные отношения?

Вопросы к полученной модели:

- корректны ли полученные отношения?
- правильно ли выявлено распределение атрибутов по отношениям?

Нормализация - формальный метод для проверки/доработки модели на основе ключей и функциональных зависимостей в отношениях.

Возможные проблемы

- Несоответствие смысловых связей реальной предметной области.
- Избыточность данных:

STUDENTS

StudID	StudName	Group	GrMentor
1	Ivan Petrov	P3100	Egor Kirov
3	Vasily Ivanov	P3101	Roman Ivov
34	Gleb Anisimov	P3100	Egor Kirov

АНОМАЛИИ ВСТАВКИ

```
INSERT INTO STUDENTS
```

```
VALUES(57, 'Nina Simonova', 'P3100', 'E. Kirov');
```

```
INSERT INTO STUDENTS
```

```
VALUES(58, 'Petr Uvarov', 'P3100', 'Egor Lomov');
```

STUDENTS

StudID	StudName	Group	GrMentor
1	Ivan Petrov	P3100	Egor Kirov
3	Vasily Ivanov	P3101	Roman Ivov
34	Gleb Anisimov	P3100	Egor Kirov
57	Nina Simonova	P3100	E.Kirov
58	Petr Uvarov	P3100	Egor Lomov

Аномалии модификации

UPDATE STUDENTS

SET GrMentor = 'Eugene Lomov'

WHERE StudName = 'Ivan Petrov';

STUDENTS

StudID	StudName	Group	GrMentor
1	Ivan Petrov	P3100	Eugene Lomov
3	Vasily Ivanov	P3101	Roman Ivov
34	Gleb Anisimov	P3100	Egor Kirov

Аномалии удаления

DELETE FROM STUDENTS

WHERE StudName = 'Vasily Ivanov';

STUDENTS

StudID	StudName	Group	GrMentor
1	Ivan Petrov	P3100	Eugene Lomov
3	Vasily Ivanov	P3101	Egor Kirov
34	Gleb Anisimov	P3100	Egor Kirov

- Данных о группе P3101 больше нет.

Функциональная зависимость

Функциональная зависимость — средство для описания связей между атрибутами отношения.

R — отношение

A_1, A_2 — атрибуты R

Функциональная зависимость

Если в R атрибут A_2 **функционально зависит** от атрибута A_1 , то каждое значение A_1 связано с одним значением A_2 и определяет его.

$$A_1 \rightarrow A_2$$

A_1 — **детерминант** функциональной зависимости.

A_1 и A_2 могут представлять несколько атрибутов.

Пример

- По StudID можно однозначно определить группу:

StudID → *Group*

- Group* не зависит от *StudID* — возможен один и тот же *Group* для разных *StudID*:

Group ↯ *StudID*

STUDENTS

StudID	StudName	Group	GrMentor
1	Ivan Petrov	P3100	Egor Kirov
3	Vasily Ivanov	P3101	Roman Ivov
34	Gleb Anisimov	P3100	Egor Kirov

Пример

- Функциональная зависимость определяется смысловыми связями, на основе которых строится отношение.
- Текущие данные в отношении не влияют на функциональные зависимости:

STUDENTS

StudID	StudName	Group	GrMentor
12	Ivan Petrov	P3101	Egor Kirov
33	Vasily Ivanov	P3102	Roman Ivov
34	Gleb Anisimov	P3103	Egor Ivanov

StudID → *Group*

Функциональные зависимости

STUDENTS

StudID	StudName	Group	GrMentor
12	Ivan Petrov	P3101	Egor Kirov
33	Vasily Ivanov	P3102	Roman Ivov
34	Gleb Anisimov	P3103	Egor Ivanov

StudID \rightarrow Group

StudID \rightarrow GrMentor

Group \rightarrow GrMentor

StudID \rightarrow StudName

StudID, StudName \rightarrow StudName

Функциональные зависимости

STUDENTS

StudID	StudName	Group	GrMentor
12	Ivan Petrov	P3101	Egor Kirov
33	Vasily Ivanov	P3102	Roman Ivov
34	Gleb Anisimov	P3103	Egor Ivanov

- Тривиальная функциональная зависимость:

$\text{StudID, StudName} \rightarrow \text{StudName}$

- Обычно рассматриваются нетривиальные функциональные зависимости.

Минимальное множество функц. зависимостей

Множество функциональных зависимостей
минимально, если:

- у всех зависимостей — один атрибут в правой части;
- $A_1 \rightarrow A_2$ нельзя заменить на $A_3 \rightarrow A_2$ (A_3 — подмножество атрибутов A_1);
- при удалении любой функц. зависимости из изначального множества не получается эквивалентное множество функц. зависимостей;

Аксиомы Армстронга

1) Рефлексивность:

если A_2 — подмножество A_1 , то $A_1 \rightarrow A_2$

2) Дополнение:

если $A_1 \rightarrow A_2$, то $A_1, A_3 \rightarrow A_2, A_3$

3) Транзитивность:

если $(A_1 \rightarrow A_2) \wedge (A_2 \rightarrow A_3)$, то $A_1 \rightarrow A_3$

Нормализация

Нормализация - формальный метод для проверки/доработки модели на основе функциональных зависимостей.

- Выполняется в несколько этапов.
- Приводит отношения в состояние, соответствующее определенному набору правил, которые зависят от выбранной **нормальной формы**.

Ненормализованная форма

Если на пересечении строки и столбца встречается несколько значений:

STUDENTS

StudID	StudName	ExamID	ExamName	ExDate	ProfID	ProfName
123	Ivan Ivanov	34 78	OPD DBMS	14.01.19 29.12.20	55 789	Rebrov A. Uvarov S.
345	Egor Kirov	34 87	OPD History	14.01.19 25.01.19	55 342	Rebrov A. Serov G.

Процесс нормализации

STUDENTS

StudID	StudName	ExamID	ExamName	ExDate	ProfID	ProfName
123	Ivan Ivanov	34 78	OPD DBMS	14.01.19 29.12.20	55 789	Rebrov A. Uvarov S.
345	Egor Kirov	34 87	OPD History	14.01.19 25.01.19	55 342	Rebrov A. Serov G.

В дальнейшем при описании нормальных форм предполагается, что в каждом отношении один потенциальный ключ, который является первичным, определения НФ — не строгие.

Первая нормальная форма (1НФ)

Отношение, на пересечении каждой строки и столбца — **одно** значение.

Вариант 1: сделать из групп значений отдельные строки.

STUDENTS

StudID	StudName	ExamID	ExamName	ExDate	ProfID	ProfName
123	Ivan Ivanov	34	OPD	14.01.19	55	Rebrov A.
123	Ivan Ivanov	78	DBMS	29.12.20	789	Uvarov S.
345	Egor Kirov	34	OPD	14.01.19	55	Rebrov A.
345	Egor Kirov	87	History	25.01.19	342	Serov G.

Первая нормальная форма (1НФ)

Отношение, на пересечении каждой строки и столбца — **одно** значение.

Вариант 2: разбить на таблицы, чтобы исключить группы EXAMS

StudID	ExamID	ExamName	ExDate	ProfID	ProfName
123	34	OPD	14.01.19	55	Rebrov A.
123	78	DBMS	29.12.20	789	Uvarov S.
345	34	OPD	14.01.19	55	Rebrov A.
345	87	History	25.01.19	342	Serov G.

STUDENTS

StudID	StudName
123	Ivan Ivanov
345	Egor Kirov

Вторая нормальная форма (2НФ)

2НФ — 1) отношение в 1НФ и 2) атрибуты, не входящие в первичный ключ, в полной функциональной зависимости от первичного ключа отношения.

A_1, A_2 — атрибуты R

Полная функциональная зависимость: A_2 в полной функциональной зависимости от A_1 , если $A_1 \rightarrow A_2$, но нет зависимостей вида $A_3 \rightarrow A_2$, где A_3 — подмножество A_1 .

Полная функциональная зависимость

Полная функциональная зависимость: A_2 в полной функциональной зависимости от A_1 , если $A_1 \rightarrow A_2$, но нет зависимостей вида $A_3 \rightarrow A_2$, где A_3 — подмножество A_1 .

Из A_1 нельзя удалить атрибут, иначе - потеря функц. зависимости $A_1 \rightarrow A_2$

$\text{StudID}, \text{ExamID} \rightarrow \text{ExDate}$ — полная ф.з.

Вторая нормальная форма (2НФ)

STUDENTS

StudID	StudName	ExamID	ExamName	ExDate	ProfID	ProfName
123	Ivan Ivanov	34	OPD	14.01.19	55	Rebrov A.
123	Ivan Ivanov	78	DBMS	29.12.20	789	Uvarov S.
345	Egor Kirov	34	OPD	14.01.19	55	Rebrov A.
345	Egor Kirov	87	History	25.01.19	342	Serov G.

Чтобы привести к 2НФ — убрать частичные зависимости от ключа:

- 1)удалить атрибуты, зависящие от составляющих ключа из R_1 ;
- 2)новое отношение R_2 : удаленные атрибуты из R_1 + соответствующий детерминант;

Преобразование в 2НФ

STUDENTS

StudID	StudName	ExamID	ExamName	ExDate	ProfID	ProfName
123	Ivan Ivanov	34	OPD	14.01.19	55	Rebrov A.
123	Ivan Ivanov	78	DBMS	29.12.20	789	Uvarov S.
345	Egor Kirov	34	OPD	14.01.19	55	Rebrov A.
345	Egor Kirov	87	History	25.01.19	342	Serov G.

StudID, ExamID → StudName

StudID, ExamID → ExamName

StudID, ExamID → ExDate

StudID, ExamID → ProfID

ProfID → ProfName

Частичная функц. зависимость

Преобразование в 2НФ

STUDENTS

StudID	StudName	ExamID	ExamName	ExDate	ProfID	ProfName
123	Ivan Ivanov	34	OPD	14.01.19	55	Rebrov A.
123	Ivan Ivanov	78	DBMS	29.12.20	789	Uvarov S.
345	Egor Kirov	34	OPD	14.01.19	55	Rebrov A.
345	Egor Kirov	87	History	25.01.19	342	Serov G.

~~StudID, ExamID~~ → StudName

~~StudID, ExamID~~ → ExamName

StudID, ExamID → ExDate

StudID, ExamID → ProfID

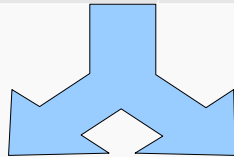
ProfID → ProfName

Частичная функц. зависимость

StudID → StudName

STUDENTS

StudID	StudName	ExamID	ExamName	ExDate	ProfID	ProfName
123	Ivan Ivanov	34	OPD	14.01.19	55	Rebrov A.
123	Ivan Ivanov	78	DBMS	29.12.20	789	Uvarov S.
345	Egor Kirov	34	OPD	14.01.19	55	Rebrov A.
345	Egor Kirov	87	History	25.01.19	342	Serov G.



EXAMS

STUDENTS

StudID	StudName
123	Ivan Ivanov
345	Egor Kirov

StudID	ExamID	ExamName	ExDate	ProfID	ProfName
123	34	OPD	14.01.19	55	Rebrov A.
123	78	DBMS	29.12.20	789	Uvarov S.
345	34	OPD	14.01.19	55	Rebrov A.
345	87	History	25.01.19	342	Serov G.

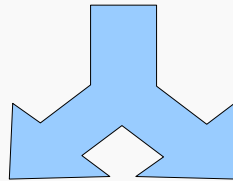
ExamID → ExamName

STUDENTS

StudID	StudName
123	Ivan Ivanov
345	Egor Kirov

EXAMS

StudID	ExamID	ExamName	ExDate	ProfID	ProfName
123	34	OPD	14.01.19	55	Rebrov A.
123	78	DBMS	29.12.20	789	Uvarov S.
345	34	OPD	14.01.19	55	Rebrov A.
345	87	History	25.01.19	342	Serov G.



EXAMS

ExamID	ExamName
34	OPD
78	DBMS
87	History

EXAMS_PARTICIPATION

StudID	ExamID	ExDate	ProfID	ProfName
123	34	14.01.19	55	Rebrov A.
123	78	29.12.20	789	Uvarov S.
345	34	14.01.19	55	Rebrov A.
345	87	25.01.19	342	Serov G.

Вторая нормальная форма

STUDENTS

StudID	StudName
123	Ivan Ivanov
345	Egor Kirov

EXAMS

ExamID	ExamName
34	OPD
78	DBMS
87	History

EXAMS_PARTICIPATION

StudID	ExamID	ExDate	ProfID	ProfName
123	34	14.01.19	55	Rebrov A.
123	78	29.12.20	789	Uvarov S.
345	34	14.01.19	55	Rebrov A.
345	87	25.01.19	342	Serov G.

Вторая нормальная форма (2НФ)

Нет **частичных** зависимостей от потенциальных ключей

StudID → StudName

ExamID → ExamName

StudID, ExamID → ExDate

StudID, ExamID → ProfID

ProfID → ProfName

Частичная функц. зависимость

Третья нормальная форма (3НФ)

3НФ — отношение в 1) 1НФ и 2НФ и 2) все атрибуты, которые не входят в первичный ключ, не находятся в транзитивной функциональной зависимости от первичного ключа.

A_1, A_2, A_3 — атрибуты R

Транзитивная функциональная зависимость — если для A_1, A_2, A_3 из R :

$$A_1 \rightarrow A_2 \wedge A_2 \rightarrow A_3$$

то A_3 транзитивно зависит от A_1 через A_2 (A_1 функционально независим от A_2, A_3).

Преобразование в 3НФ

STUDENTS

StudID	StudName
123	Ivan Ivanov
345	Egor Kirov

StudID → StudName

EXAMS

ExamID	ExamName
34	OPD
78	DBMS
87	History

ExamID → ExamName

EXAMS_PARTICIPATION

StudID	ExamID	ExDate	ProfID	ProfName
123	34	14.01.19	55	Rebrov A.
123	78	29.12.20	789	Uvarov S.
345	34	14.01.19	55	Rebrov A.
345	87	25.01.19	342	Serov G.

StudID, ExamID → ExDate

StudID, ExamID → ProfID

ProfID → ProfName

Транзитивная
зависимость от
первичного ключа

Третья нормальная форма (3НФ)

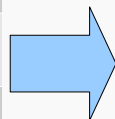
Чтобы привести к 3НФ — убрать транзитивные зависимости:

1) удалить из R_1 атрибуты, транзитивно-зависимые от первичного ключа;

2) новое отношение R_2 : атрибуты (удаленные в 1.) + соответствующий детерминант;

EXAMS_PARTICIPATION

StudID	ExamID	ExDate	ProfID	ProfName
123	34	14.01.19	55	Rebrov A.
123	78	29.12.20	789	Uvarov S.
345	34	14.01.19	55	Rebrov A.
345	87	25.01.19	342	Serov G.



EXAMS_PARTICIPATION

StudID	ExamID	ExDate	ProfID
123	34	14.01.19	55
123	78	29.12.20	789
345	34	14.01.19	55
345	87	25.01.19	342

PROFS

ProfID	ProfName
789	Uvarov S.
55	Rebrov A.
342	Serov G.

Третья нормальная форма

STUDENTS

StudID	StudName
123	Ivan Ivanov
345	Egor Kirov

EXAMS

ExamID	ExamName
34	OPD
78	DBMS
87	History

EXAMS_PARTICIPATION

StudID	ExamID	ExDate	ProfID
123	34	14.01.19	55
123	78	29.12.20	789
345	34	14.01.19	55
345	87	25.01.19	342

PROFS

ProfID	ProfName
789	Uvarov S.
55	Rebrov A.
342	Serov G.

Третья нормальная форма (3НФ)

Нет **транзитивных** зависимостей от потенциальных ключей:

$\text{StudID, ExamID} \rightarrow \text{ExDate}$

$\text{StudID, ExamID} \rightarrow \text{ProfID}$

$\text{ProfID} \rightarrow \text{ProfName}$

Транзитивная
зависимость от
первичного ключа

Нормальная форма Бойса-Кодда (НФБК)

НФБК — отношение в НФБК, когда для всех функциональных зависимостей отношения выполняется условие: детерминант — потенциальный ключ.

$$A_1 \rightarrow A_2$$

A_1 — **детерминант** функциональной зависимости.

Нормализация

- Обычно процесс останавливается на 3НФ или НФБК (в зависимости от предметной области и требований к БД).
- Существуют 4НФ и 5НФ, но используются редко.

Денормализация

- Бывает, что для повышения производительности запросов производится денормализация:
 - несколько отношений объединяют в одно;
- В результате:
 - можно повысить эффективность выполнения некоторых запросов (уменьшается число соединений таблиц);
 - увеличивается избыточность данных;
 - требуется больше усилий на поддержание целостности БД;

При подготовке презентации использовались материалы из:

- Введение в реляционные базы данных / В. В. Кириллов, Г. Ю. Громов, Издательство: BHV, 2009 г.
- Документация PostgreSQL.

<https://www.postgresql.org/about/licence/>

PostgreSQL is released under the PostgreSQL License, a liberal Open Source license, similar to the BSD or MIT licenses.

PostgreSQL Database Management System
(formerly known as Postgres, then as Postgres95)

Portions Copyright © 1996-2020, The PostgreSQL Global Development Group

Portions Copyright © 1994, The Regents of the University of California

Permission to use, copy, modify, and distribute this software and its documentation for any purpose, without fee, and without a written agreement is hereby granted, provided that the above copyright notice and this paragraph and the following two paragraphs appear in all copies.

IN NO EVENT SHALL THE UNIVERSITY OF CALIFORNIA BE LIABLE TO ANY PARTY FOR DIRECT, INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES, INCLUDING LOST PROFITS, ARISING OUT OF THE USE OF THIS SOFTWARE AND ITS DOCUMENTATION, EVEN IF THE UNIVERSITY OF CALIFORNIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

THE UNIVERSITY OF CALIFORNIA SPECIFICALLY DISCLAIMS ANY WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE SOFTWARE PROVIDED HEREUNDER IS ON AN "AS IS" BASIS, AND THE UNIVERSITY OF CALIFORNIA HAS NO OBLIGATIONS TO PROVIDE MAINTENANCE, SUPPORT, UPDATES, ENHANCEMENTS, OR MODIFICATIONS.