Основы технологий баз данных

Б. А. Новиков
Санкт-Петербургский государственный университет, JetBrains Research

Модели данных

Модель данных — это

Система понятий для описания моделей систем

- Структуры данных
- Зависимости
- Операции
  - Ограничения

Описание конкретной системы приложений

- Описание типов и объектов данных
- Возможно, организация хранения
- Описание взаимосвязей
- Ограничения целостности

Модель данных (как инструментарий) включает

- Набор скалярных типов данных
- Методы конструирования и типы сложных объектов
- Средства описания взаимосвязей между объектами данных
- Набор операций над типами данных
- Средства описания ограничений целостности

Назовем их по имени

- (Теоретическая) реляционная модель данных
- Модели данных промышленных реляционных СУБД (SQL)
- Модели плоских файлов
- Модели «сущность-связь»
- Объектно-реляционные модели
- Модели даных языков программирования
- Модели слабоструктурированных данных
- Тернарная модель данных
- Модели для представления графов
- . . . . .

Postgres PROFESSIONAL

Чем отличаются модели данных?

- Методы идентификации объектов внутри и вне модели
- Поиск и связывание объектов
- Массовая или штучная обработка объектов данных

Естественная идентификация

- Объект идентифицируется по значениям его атрибутов, описывающих свойства реального моделируемого объекта
- Относительно легко сопоставить объект модели с реальным
- Не гарантируется уникальность
- Примеры:
  - Цвет, номерной знак и модель автомобиля
  - Номер пенсионного свидетельства и номера других документов
  - Биометрическая информация:

«А ростом он мал, грудь широкая, одна рука короче другой, глаза голубые, волоса рыжие, на щеке бородавка, на лбу другая»

А. С. Пушкин

Идентификация по суррогатам

- Значения ID генерируются системой и не связаны с естественными признаками моделируемого реального объекта
- (Почти) гарантируется уникальность внутри системы
- (Почти) невозможно отождествление объекта в модели с реальным
- Суррогатный идентификатор часто используется как естественный вне системы, которая его создала

Идентификация и изменяемость

- Константы идентифицируют сами себя
- Значения атрибутов, идентифицирующих объект данных, не могут быть изменены
- Если для идентификации используются не все атрибуты, то объект имеет состояние
- Возможность изменения значений атрибутов не зависит от типа идентификации (по естественным или по суррогатным идентификаторам)

Идентификация по взаимосвязям с другими объектами

- Необходимо применять, если объекты неразличимы по значениям атрибутов и не имеют суррогатных ключей
- По положению в пространстве или относительно других объектов
- По адресу в памяти вычислительной системы (обычно это — выход за пределы модели данных)

Навигация или связывание по значениям?

Навигация

- Понятие текущего объекта
- Переходы по заранее установленным связям (по указателям)
- Обычно используется в моделях языков программирования

Связывание по значениям

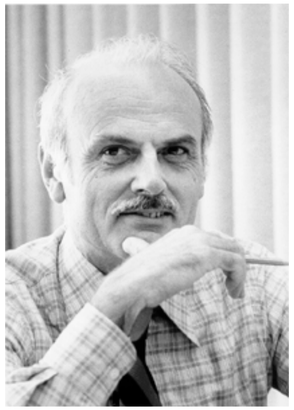
- Динамическое связывание объектов по значениям атрибутов
- Множественность связей
- Обычно применяется в базах данных

Массовая или поштучная обработка

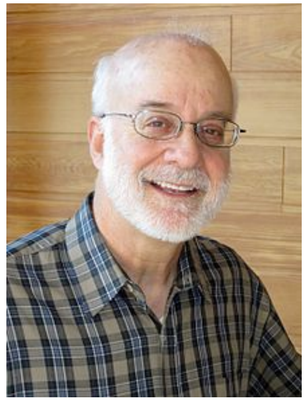
- Зависит от операций, определенных в модели данных
- Массовая обработка естественна при связывании по значениям
- Поштучная обработка естественна при использовании навигации
- Определяет характер нагрузки на аппаратуру (сервер базы данных или вычислительная сеть)
- Необходимо учитывать при выборе критериев эффективности

Теоретическая реляционная модель данных

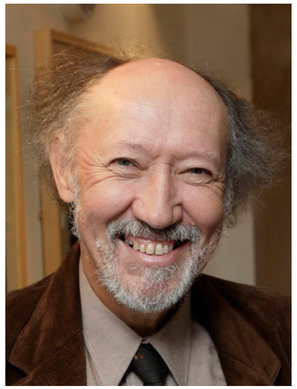
Теоретики



Edgar F. Codd



Ronald Fagin



Chris Date

Домены

- Домен — непустое множество (в математическом смысле), элементы которого рассматриваются как скалярные (в рамках реляционной модели данных)
- На домене определено бинарное отношение равенства $= : D \times D \to \{T, F\}$
- Могут быть определены другие отношения (>, < и т. п), операции (+, —, …) и функции со значениями в том же или другом домине, с одним или несколькими аргументами (вообще говоря, из разных доменов)
- Абстрактные домены: числа, целые числа, текстовые строки, даты, . . .
- Более конкретные: длина, вес, цена, наименование продукта, фамилия, . . .

Отношения

- Отношением (или предикатом) называется функция (нескольких аргументов), принимающая значения True или False
- Аргументы отношения называются атрибутами
- Каждый атрибут принимает значения в некотором домене
- В реляционной модели атрибуты принято обозначать именами (а не номерами)
- Все, что написано выше, можно записать в формальных математических обозначениях

Интерпретация предикатов (отношений)

- истинность предиката интерпретируется как истинность факта, представленного значениями аргументов
- На основе имеющихся отношений можно строить новые, используя логические операции
- Булева логика: многократное повторение утверждений не влияет на истинность этих утверждений

```
exams (name := 'Анна', course := 'Базы данных', grade := 5)
exams (name := 'Анна', course := 'Анализ данных', grade := 5)
exams (course := 'Анализ данных', grade := 4, name := 'Виктор')
exams (name := 'Нина', grade := 5, course := 'Базы данных')
```

Реляционное исчисление

- Любой запрос к базе данных представляет собой предикат, выраженный формулой
- Правила построения формул (из уже имеющихся предикатов):
  - Логические операци: конъюнкция (И), дизъюнкция (ИЛИ), отрицание (НЕ)
  - Скобки
  - Кванторы
- Не допускается рекурсия при определении формул
- Исчисление — декларативный язык запросов

Теоретико-множественная интерпретация

- Предикат можно записать как множество наборов значений атрибутов, для которых значение предиката истинно
- Множества истинности отношений можно записать в виде таблицы
- Все элементы множества различны
- Кортежи (строки) могут идентифицироваться своими значениями
- Теоретическая реляционная модель данных не знает о том, что мир изменчив

exams

name	course	grade
Анна	Базы данных	5
Анна	Анализ данных	5
Виктор	Анализ данных	4
Нина	Базы данных	5

Что такое алгебра?

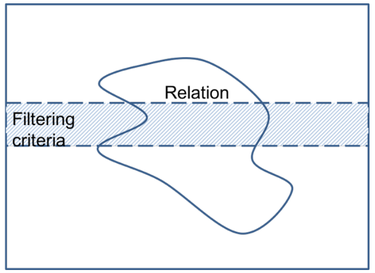
- Нет, не раздел школьной математики
- Множество, снабжение операциями, обычно результат операции в том же множестве
  - Операции могут быть частичными
  - Можно строить алгебраические выражения, применяя операции к результатам
- Примеров много:
  - Целые числа, вещественные числа, …
  - Булева алгебра, …

Реляционная алгебра

- Определена на множестве отношений
- Теоретико-множественные операции (объединение, пересечение, разность) для отношений с одинаковым набором атрибутов
- Селекция (фильтрация)
- Проекция
- Произведение
- Соединение

Реляционная операция селекции

- Выбирает подмножество, удовлетворяющее дополнительным условиям (предикату)
- Предикат селекции:
  - Простые условия на значениях атрибутов, определенные на доменах
  - Логические операции (И, ИЛИ, НЕ)
  - Скобки



exams

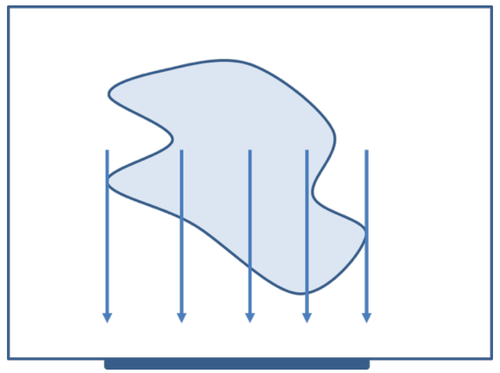
name	course	grade
Анна	Базы данных	5
Анна	Анализ данных	5
Виктор	Анализ данных	4
Нина	Базы данных	5

FILTER [course='Анализ данных' AND grade=5] exams

name	course	grade
Анна	Анализ данных	5

Реляционная проекция

- проекция: P(P(x)) =P(x)
- Реляционная проекция строит отношение с меньшим количеством атрибутов
- Количество кортежей может уменьшиться



exams

name	course	grade
Анна	Базы данных	5
Анна	Анализ данных	5
Виктор	Анализ данных	4
Нина	Базы данных	5

PROJ [name, course] exams

name	course
Анна	Базы данных
Анна	Анализ данных
Виктор	Анализ данных
Нина	Базы данных

PROJ [course, grade] exams

course	grade
Базы данных	5
Анализ данных	5
Анализ данных	4

Произведение

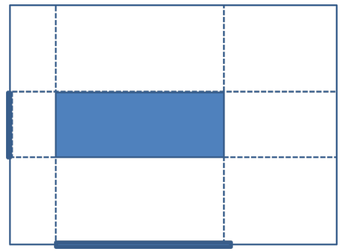
- Прямое (декартово) произведение отношений как множеств

exams

name	course	grade
Анна	Базы данных	5
Анна	Анализ данных	5
Виктор	Анализ данных	4
Нина	Базы данных	5

courses

title	credits
Базы данных	5
Анализ данных	10

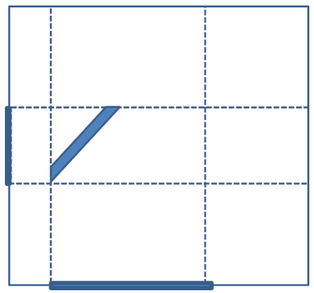


exams PROD courses

name	course	grade	title	credits
Анна	Базы данных	5	Базы данных	5
Анна	Анализ данных	5	Базы данных	5
Виктор	Анализ данных	4	Базы данных	5
Нина	Базы данных	5	Базы данных	5
Анна	Базы данных	5	Анализ данных	10
Анна	Анализ данных	5	Анализ данных	10
Виктор	Анализ данных	4	Анализ данных	10
Нина	Базы данных	5	Анализ данных	10

Соединение

- Прямое произведение с селекцией
- Эквисоединение: соединение по условию равенства значений атрибутов из соединяемых отношений



exams

name	course	grade
Анна	Базы данных	5
Анна	Анализ данных	5
Виктор	Анализ данных	4
Нина	Базы данных	5

courses

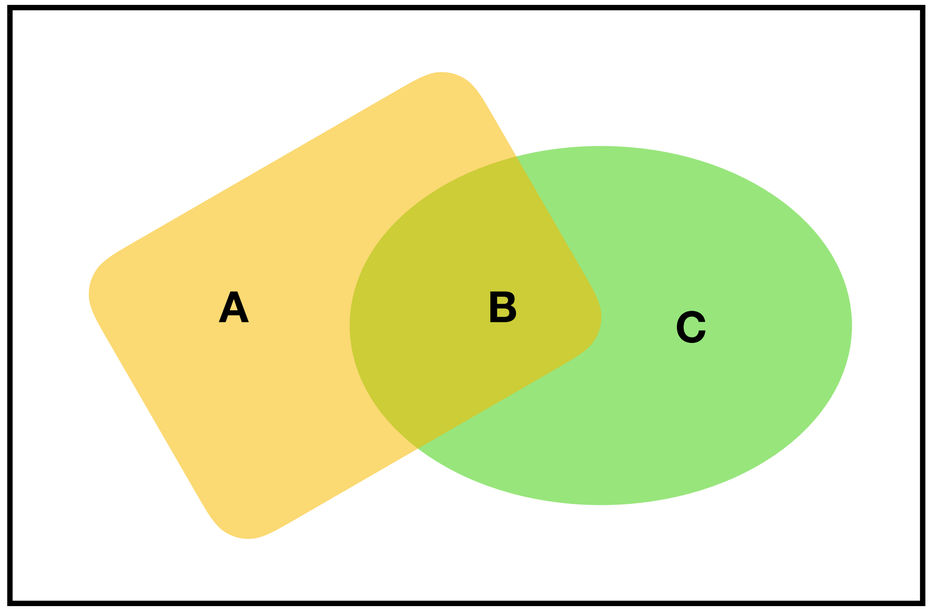
title	credits
Базы данных	5
Анализ данных	10

exams JOIN [course=title] courses

name	course	grade	title	duration
Анна	Базы данных	5	Базы данных	5
Нина	Базы данных	5	Базы данных	5
Анна	Анализ данных	5	Анализ данных	10
Виктор	Анализ данных	4	Анализ данных	10

Теоретико-множественные операции

- Отношения R=AB, S=BC
- Объединение (не путать с операцией соединения): ABC
- Пересечение : B
- Разность A
- Симметричная разность: AC



Эквивалентность реляционных языков запросов

- Любой запрос, выразимый в реляционном исчислении, можно выразить в реляционной алгебре, и обратно.

Алгебраические тождества

$$\sigma\,[p]\,\sigma\,[q]\,R = \sigma\,[p \wedge q]\,R$$

$$\pi\,[A]\,R = \pi\,[A]\,\pi\;\,[AB]\,R$$

$$\sigma[p(A)]\,\pi\,[B]\,R = \pi\,[B]\,\sigma\,[p(A)]\,R,\;\text{ if } A \subseteq B$$

$$\sigma\,[p \vee q]\,R = (\sigma\,[p]\,R)\;\cup\;(\sigma\,[q]\,R)$$

$$\sigma\,[p \vee \nmid q]\,R = \sigma\,[p]\,R\;\setminus\;\sigma\,[q]\,R$$

$$\sigma\,[p(R)]\,(R \rhd\lhd S) = (\sigma\,[p(R)]\,R)\;\rhd\lhd S$$

$$\sigma\,[p]\,(R \cup S) = (\sigma\,[p]\,R) \cup (\sigma\,[p]\,S)$$

$$R \rhd\lhd S = S \rhd\lhd R$$

$$R \cup S = S \cup R$$

$$R \cap S = S \cap R$$

$$(R \rhd\lhd S) \rhd\lhd T = R \rhd\lhd (S \rhd\lhd T)$$

$$(R \cup S) \cup T = R \cup (S \cup T)$$

$$(R \cap S) \cap T = R \cap (S \cap T)$$

$$R \rhd\lhd (S \cup T) = (R \rhd\lhd S) \cup (R \rhd\lhd T)$$

$$R \rhd\lhd (S \cap T) = (R \rhd\lhd S) \cap (R \rhd\lhd T)$$

Вложенные коллекции и 1NF

- Реляционная модель допускает только скалярные атрибуты
- Как хранить более сложные структуры данных?
- 1NF: Первая нормальная форма

GRP	ID	Name
G1	11	N11
	12	N12
	13	N13
G2	21	N21
	22	N22
	23	N23
	24	N24
G3	31	N31
	32	N32

GRP	ID	Name
G1	11	N11
G1	12	N12
G1	13	N13
G2	21	N21
G2	22	N22
G2	23	N23
G2	24	N24
G3	31	N31
G3	32	N32

Функциональные зависимости

- Набор атрибутов Y функционально зависит от набора атрибутов X, если для любого значения X существует ровно одно значение Y.

$$(\textbf{passenger\_id, Flight}) \rightarrow \textbf{seat}$$

- Источник зависимостей: законы реального мира, но не конкретное содержание БД
- Тривиальные зависимости: любой атрибут зависит от полного набора атрибутов отношения.
- Представляют ограничения целостности в реляционной модели

Ключи

Возможный ключ	Набор атрибутов отношения, от которого зависят все атрибуты отношения
Минимальный ключ	Возможный ключ, любое подмножество которого не является возможным ключом
Первичный ключ	Произвольно выбранный минимальный ключ отношения, используемый для идентификации
Внешний ключ	Набор атрибутов, соответствующих первичному ключу (другого) отношения

Аномалии реляционных схем

- Номер студента, фамилия, имя, год поступления, программа, факультет, группа, дисциплина, оценка
- Аномалия вставки:
  - Невозможно добавить группу, в которой нет студентов
  - Невозможно хранить информацию о дисциплине, пока никто не получил оценку
- Аномалия удаления:
  - Вместе с последним студентом удаляется информация о группе
- Аномалия обновления

Нормализация методом разложения

- 2NF: Зависимости от неполного (первичного) ключа

  passenger_id, flight, seat, scheduled_departure **flight → scheduled_departure**

  passenger_id, flight, seat flight, scheduled_departure

- 3NF: Транзитивные зависимости

  flight, departure_airport, city **flight → departure_airport → city**

  flight, departure_airport departure_airport, city

Другие зависимости и нормальные формы

- 4НФ: Многозначные зависимости
- 5НФ: Максимально возможное разложение без потерь (информации)
- Зависимости включения
- И еще много . . .

Синтез нормализованных схем

- Исходные данные: функциональные зависимости между атрибутами (отношения несущественны)
- Существует полиномиальный (по сложности) алгоритм, который строит схему в 3NF

Теоретическая реляционная модель данных: итоги

- Естественная идентификация по значениям атрибутов
- Идентификация определяется функциональными зависимостями
- Связи вычисляются динамически при выполнении операций соединения
- Ассоциативный доступ к данным по значениям (операция селекции)