Этапы проектирования баз данных

Невозможно создать БД без подробного ее описания, также как и не возможно сделать какое-либо сложное изделие без чертежа и подробного описания технологий его изготовления. Другими словами, нужен *проект*.

Проектом принято считать эскиз некоторого устройства, который в дальнейшем будет воплощен в реальность.

Процесс проектирования БД представляет собой процесс переходов от неформального словесного описания информационной структуры предметной области к формализованному описанию объектов предметной области в терминах некоторой модели. Конечной целью проектирования является построение конкретной БД. Очевидно, что процесс проектирования сложен и поэтому имеет смысл разделить его на логически завершенные части – *этапы*.

Можно выделить пять основных этапов проектирования БД:

1. Сбор сведений и системный анализ предметной области.
2. Инфологическое проектирование.
3. Выбор СУБД.
4. Даталогическое проектирование.
5. Физическое проектирование.

Сбор сведений и системный анализ предметной области

- это первый и важнейший этап при проектировании БД.

В нем необходимо провести подробное словесное описание объектов предметной области и реальных связей, присутствующих между реальными объектами. Желательно чтобы в описании определялись взаимосвязи между объектами предметной области. В общем случае выделяют два подхода к выбору состава и структуры предметной области:

1. ***Функциональный подход*** – применяется тогда, когда заранее известны функции некоторой группы лиц и комплексы задач, для обслуживания которых создается эта БД, т.е. четко выделяется минимальный необходимый набор объектов предметной области под описание.
2. ***Предметный подход*** – когда информационные потребности заказчиков БД четко не фиксируются и могут быть многоаспектными и динамичными. В данном случае минимальный набор объектов предметной области выделить сложно.

В описание предметной области включаются такие объекты и взаимосвязи, которые наиболее характерны и существенны для нее. При этом БД становится предметной, и подходит для решения множества задач (что кажется наиболее заманчивым). Однако трудность всеобщего охвата предметной области и невозможность конкретизации потребностей пользователей приводит к избыточно сложной схеме БД, которая для некоторых задач будет неэффективной.

Рекомендуется использовать компромиссный вариант, который, с одной стороны, ориентирован на конкретные задачи, а с другой стороны, учитывает возможность расширения приложения.

Системный анализ должен заканчиваться подробным описанием информации об объектах предметной области, которая должна храниться в БД, формулировкой конкретных задач, которые будут решаться с использованием данной БД с кратким описанием алгоритмов их решения, описанием выходных и входных документов при работе с БД.

Инфологическое проектирование

– частично формализованное описание объектов предметной области в терминах некоторой семантической модели.

Зачем нужна инфологическая модель, и какую пользу она дает проектировщикам?

Дело в том, что процесс проектирования длительный, требует обсуждений с заказчиком и специалистами в предметной области. Кроме того, при разработке серьезных корпоративных информационных систем проект базы данных является фундаментом, на котором строится вся система в целом.

Инфологическая модель должна включать такое формализованное описание предметной области, которое легко будет восприниматься не только специалистами в области БД.

Описание должно быть настолько емким, чтобы можно было оценить глубину и корректность проработки проекта БД.

Широкое распространение получила модель Чена «*Сущность-связь*» (Entity Relationship), она стала фактическим стандартом в инфологическом моделировании, и получило название ER – модель.

Выбор СУБД осуществляется на основе различных требований к БД и, соответственно, возможностей СУБД, а также в зависимости от имеющегося опыта разработчиков.

Даталогическое проектирование есть описание БД в терминах принятой даталогической модели данных. В реляционных БД даталогическое или логическое проектирование приводит к разработке схемы БД, т. е. совокупности схем отношений, которые адекватно моделируют объекты предметной области и семантические связи между объектами.

Основой анализа корректности схемы являются функциональные зависимости между атрибутами БД.

В некоторых случаях между атрибутами отношений могут появиться нежелательные зависимости, которые вызывают побочные эффекты и аномалии при модификации БД.

Под модификацией понимают внесение новых данных в БД, удаление данных из БД, а также обновление значений некоторых атрибутов. Для ликвидации возможных аномалий предполагается проведение нормализации отношений БД. Этап логического проектирования не заключается только в проектировании схемы отношений.

В результате выполнения этого этапа, как правило, должны быть получены следующие результирующие документы:

- Описание концептуальной схемы БД в терминах выбранной СУБД.
- Описание внешних моделей в терминах выбранной СУБД.
- Описание декларативных правил поддержки целостности БД.
- Разработка процедур поддержки семантической целостности БД.

Физическое проектирование

заключается в увязке логической структуры БД и физической среды хранения с целью наиболее эффективного размещения данных, т.е. отображение логической структуры БД в структуру хранения.

Решается вопрос размещения хранимых данных в пространстве памяти, выбора эффективных методов доступа к различным компонентам «физической» БД, решаются вопросы обеспечения безопасности и сохранности данных.

Ограничения, имеющиеся в логической модели данных, реализуются различными средствами СУБД, например, при помощи индексов, декларативных ограничений целостности, триггеров, хранимых процедур. При этом опять-таки решения, принятые на уровне логического моделирования определяют некоторые границы, в пределах которых можно развивать физическую модель данных. Точно также, в пределах этих границ можно принимать различные решения. Например, отношения, содержащиеся в логической модели данных, должны быть преобразованы в таблицы, но для каждой таблицы можно дополнительно объявить различные индексы, повышающие скорость обращения к данным.

Кроме того, для повышения производительности могут использоваться возможности параллельной обработки данных. В результате БД может размещаться на нескольких сетевых компьютерах. С другой стороны могут использоваться преимущества многопроцессорных систем.

Для обеспечения безопасности и сохранности данных решаются вопросы способы восстановления после сбоев, резервного копирования информации, настройка систем защиты под выбранную политику безопасности и т.д.

Таким образом, ясно, что решения, принятые на каждом этапе моделирования и разработки базы данных, будут сказываться на дальнейших этапах. Поэтому особую роль играет принятие правильных решений на ранних этапах моделирования.

Контрольные вопросы

1.Что такое проект?

2. Какие этапы проектирования БД принято выделять?

3. В чем назначение системного анализа?

4. Какие подходы могут применяться в системном анализе предметной области?

5. Что представляет собой этап инфологическое проектирование?

6. В чем различие инфологического и даталогического этапов проектирования?

7. Какие документы и модели необходимо получить при завершении этапа даталогического проектирования?

8. Назовите результаты физического проектирования

Теория нормализации

Переход от инфологического проектирования к даталогическому производится с учетом выбора СУБД. В данном курсе мы изучаем РМД и, следовательно, выбираем реляционную СУБД. Прежде всего, необходимо построить корректную схему БД, ориентируясь на РМД. Основой анализа корректности схемы являются так называемые функциональные зависимости между атрибутами БД. Некоторые функциональные зависимости атрибутов являются нежелательными из-за побочных явлений и аномалий, которые они могут вызвать.

Обычно различают следующие проблемы:

- избыточность данных;
- аномалии обновления;
- аномалии удаления;
- аномалии ввода.

Избыточность данных характеризуется наличием в кортежах отношений повторяющейся информации. Многократное дублирование данных приводит к неоправданному увеличению занимаемого объема внешней памяти

Аномалии обновления, прежде всего, связанны с избыточностью данных, что приводит к проблемам при их изменении. При изменении повторяющихся данных придется многократно изменять их значения, однако, если изменения будут внесены не во все кортежи, возникнет несоответствие информации, которое называется аномалией обновления.

Аномалии удаления могут возникать при удалении записей из ненормализованных таблиц и характеризуются вероятностью удаления не всех дублированных кортежей.

Аномалии ввода возникают при добавлении в таблицу новых записей, обычно в поля с ограничениями NOT NULL (не пустые). Кода в отношение на данный момент времени невозможно ввести однозначную информацию.

Для ликвидации нежелательных функциональных зависимостей есть специальный формальный механизм называемый нормализацией.

В процессе нормализации происходит устранение избыточности и противоречивости хранимых данных.

Нормальные формы

Теория нормализации основана на концепции нормальных форм. Каждой нормальной форме соответствует набор ограничений. Отношение находится в нормальной форме, ели оно удовлетворяет свойственному данной форме набору ограничений.

В теории реляционных БД обычно выделяется следующая последовательность нормальных форм:

- *первая нормальная форма (1НФ);*
- *вторая нормальная форма (2НФ);*
- *третья нормальная форма (3НФ);*
- *нормальная форма Бойса-Кодда (БКНФ);*
- *четвертая нормальная форма (4НФ);*
- *пятая нормальная форма, или нормальная форма проекции соединения (5НФ или ПС/НФ).*

Основные свойства нормальных форм:

- каждая следующая нормальная форма, в некотором смысле, улучшает свойства предыдущей;
- при переходе к следующей нормальной форме свойства предыдущих нормальных форм сохраняются.

Определение 1НФ.

Отношение находится в первой нормальной форме тогда и только тогда, когда каждый его атрибут содержит атомарные значения и отношение не содержит повторяющихся групп.

Пусть необходимо автоматизировать процесс отпуска товаров со склада по накладной, примерный вид накладной на рисунке 1.

Накладная № 234

Дата	Покупатель	Адрес
10.01.2002	ООО «Геракл»	г. Москва, ул. Стромынка, 20

Отпущен товар	**Количество**	**Ед. изм.**	**Цена ед. изм**	**Общая стоимость**
Тушенка	1000	банки	25	25000
Сахар	50	КГ	10	500
Макароны	300	кг	10	3000

ИТОГО 28500

По накладной можно сформировать следующее отношение удовлетворяющее 1НФ (рисунок 2):

ОТПУК ТОВАРОВ

Номер накладной
Дата
Покупатель
Город
Адрес
Товар
Количество
Ед.изм.
Цена ед.изм.
Общая стоимость

Определение 2НФ.

Отношение находится во второй нормальной форме тогда и только тогда, когда оно находится в первой нормальной форме и каждый не ключевой атрибут функционально зависим от атрибутов первичного ключа.

Прежде всего, необходимо определить понятие функциональной зависимости.

Функционально зависимым считается атрибут, значение которого однозначно определяется значением другого атрибута, т.е. значение одного атрибута зависит от значения другого.

Функциональная зависимость значения атрибута Y от значения атрибута X обозначается следующим образом: X→Y.

Необходимо отметить, что атрибут, указываемый в левой части называется детерминантом.

Продолжим рассмотрение описанного выше примера. Для приведения отношения к 2НФ необходимо вначале выделить *первичный ключ*.

Первичный ключ возможно определить из следующих рассуждений.

Если бы по одной накладной отпускался бы только один товар, то первичным ключом являлся бы атрибут «Номер накладной», однако по одной накладной отпускается несколько различных товаров, следовательно, первичный ключ должен состоять из двух атрибутов «Номер накладной» и «Товар», только в этом случае будет обеспечено свойство уникальности.

Рассмотрим функциональные зависимости атрибутов от первичного ключа, при этом проще начинать рассмотрение с частей первичного ключа, в данном случае с атрибута «Номер накладной»:

Номер накладной → Покупатель

Номер накладной → Дата

Номер накладной → Город

Номер накладной → Адрес

Определим функциональные зависимости от атрибута «Товар»:

Товар → Ед.изм

Товар → Цена ед.изм

Оставшиеся атрибуты определяются первичным ключом:
Номер накладной, Товар → Количество

Номер накладной, Товар → Общая стоимость

В результате мы получили три различные категории, у каждой из которых свой первичный ключ.

После проведения вышеуказанного анализа отношения производим его декомпозицию и определяем типы связей (рисунок 3)

НАКЛАДНАЯ
Номер
 накладной
Дата
Покупатель
Город
 Адрес

ОТПУСК ТОВАРОВ
Номер накладной
Товар
 Количество
 Общая стоимость

ТОВАР
Товар
Ед. изм.
 Цена ед.из

В результате получаем отношения, все атрибуты которых полностью, функционально зависимы от своих первичных ключей

Определение 3НФ. *Отношение находится в третьей нормальной форме тогда и только тогда, когда оно находится во второй нормальной форме и не содержит транзитивных зависимостей между не ключевыми атрибутами и, следовательно, удовлетворяют условию 2НФ.*

Функциональная зависимость атрибутов X и Y отношения называется транзитивной, если существует такой атрибут Z, что имеются функциональные зависимости $X \rightarrow Z$ и

$Z \rightarrow Y$, но отсутствует функциональная зависимость $Z \rightarrow X$.

В результате анализа отношения «НАКЛАДНАЯ» определяем следующую транзитивную зависимость:

Номер накладной → Покупатель

Номер накладной → Город

Номер накладной → Адрес

Покупатель → Город

Покупатель → Адрес

В результате анализа отношения «ОТПУСК ТОВАРОВ», также определяется транзитивная зависимость следующего вида:

Номер накладной, Товар → Количество

Номер накладной, Товар → Общая стоимость

Количество → Общая стоимость

От транзитивной зависимости в отношении «НАКЛАДНАЯ» можно избавиться простой декомпозицией отношения. С атрибутом «Общая стоимость» отношения «ОТПУСК ТОВАРОВ» можно поступить еще проще, отказаться от использования этого атрибута, т.к. общую стоимость можно всегда получить, зная цену единицы товара и какое его количество продано, следовательно, не имеет смысла использовать внешние носители для хранения этих данных. В результате получим следующую схему отношений приведенную к 3НФ (рисунок 4).

НАКЛАДНАЯ
Номер накладной
Дата
Покупатель

ОТПУСК ТОВАРОВ
Номер накладной
Товар
Количество

ТОВАР
Товар
Ед. изм.
Цена ед.изм

ПОКУПАТЕЛЬ
Покупатель
Город
Адрес

В большинстве случаев достижение третьей нормальной формы, или даже формы Бойса-Кодда считается достаточным для реальных проектов БД. Четвертая и пятая считаются нормальными формами высших порядков

Контрольные вопросы

1. Для чего необходим процесс нормализации?
2. Какие аномалии могут возникать при использовании ненормализованных отношений и почему?
3. Определите процессы синтеза и декомпозиции.
4. Назовите определение 1НФ.
5. Назовите определение 2НФ. 6
6. Назовите определение 3НФ.
7. Что дает приведение БД к 3НФ?
8. Почему приведение к 3НФ считается достаточным для большинства проектов БД?