УДК 681.3.07

А. В. Балдин, А. В. Брешенков

АНАЛИЗ ПРОБЛЕМЫ ПРОЕКТИРОВАНИЯ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ СУЩЕСТВУЮЩЕЙ ИНФОРМАЦИИ ТАБЛИЧНОГО ВИДА

Выполнен краткий аналитический обзор современной теории проектирования реляционных баз данных, сформулированы ее достоинства и недостатки. Введено понятие информации табличного вида. Сформулированы мотивы преобразования информации табличного вида в реляционные базы данных. Выполнен анализ задач, возникающих в процессе преобразования информации табличного вида в реляционные базы данных.

Сегодня трудно переоценить значение информации и информационных систем. Особое место в составе информационных систем принадлежит базам данных (БД). В настоящее время практически не осталось таких областей человеческой деятельности, где бы они не использовались. При этом потребность в БД и системах управления базами данных (СУБД) постоянно растет.

К числу наиболее распространенных моделей построения БД относятся реляционные модели данных (РМД). Достоинства РМД побудили к проведению большого числа теоретических и практических разработок в области теории проектирования реляционных БД (РБД), в области разработки инструментальных средств, ориентированных на их создание. В частности, БД проектируются в соответствии с предложенными теоретиками БД этапами проектирования, модели БД строятся в соответствии с требованиями к РМД, реляционные таблицы проектируются с учетом требований нормализации. Существующие теоретические положения проектирования БД позволяют разработчику обоснованно назначить ключевые и индексные поля, формировать связи между таблицами, обеспечивать безопасность данных и выполнять много полезных мероприятий, обеспечивающих разработку высококачественных программных систем.

Однако даже основоположники РМД, в частности К. Дж. Дейт, признают, что традиционная теория проектирования РБД пока далека от совершенства, а “проектирование БД — это скорее искусство, чем наука” [1]. Это связано с тем, что проектные решения принимают, исходя из анализа предполагаемых схем данных, без возможности учета реальных данных.

Кроме того, большая часть информации, в том числе и информация табличного вида (ИТВ), которая нуждается в автоматизированной обработке, находится вне БД и даже вне ЭВМ [2].

Применение методов и автоматизированных средств проектирования РБД на основе использования существующей ИТВ, с одной стороны, позволит свести к минимуму недостатки современной теории, которая вынуждена отталкиваться от гипотетических данных, а с другой стороны, в случае необходимости, позволит выполнить эффективное преобразование ИТВ в РБД. А такая необходимость, как показывают экспертные исследования и собственный опыт разработок, возникает часто.

В современной методологии проектирования РБД выделяется четыре этапа проектирования: формулировка и анализ требований, концептуальное проектирование, датологическое проектирование, физическое проектирование [1–3].

Формулировка и анализ требований связаны с определением сферы применения БД, сбором информации об использовании данных, выделением информационных потоков, определением семантики, экспертными оценками, определением целей разработки. Этот этап заканчивается техническим заданием (ТЗ) на БД, которое при наличии ИТВ в значительной степени уже сформулировано. Действительно, определен состав данных, их семантика, имеются сведения об информационных потоках, об использовании данных. Таким образом, при наличии ИТВ процесс формирования ТЗ несколько упрощается.

Концептуальное проектирование (или инфологическое проектирование) посвящено построению модели предметной области. Она строится на основе применения теории РБД с использованием РМД. Эта модель абстрагирована от инструментальных систем СУБД.

Именно на этом этапе реляционный подход к проектированию БД проявляет себя в полной мере. Он позволяет выявить концептуальные ошибки в проекте БД, заложить эффективные решения на ранних этапах ее разработки. Именно на этом этапе в наибольшей мере проявляется специфика подхода к проектированию РБД на основе использования ИТВ. Это объясняется тем, что данные уже существуют, хотя их форма представления в общем случае не удовлетворяет требованиям к РМД.

Датологическое проектирование связано с построением модели данных на основе инфологической модели. Причем в качестве модели данных используют сами данные. Модель данных строится в терминах инструментальной СУБД. На этом этапе, как и на предыдущем, проверяется адекватность модели, ее непротиворечивость и расширяемость.

При проектировании РБД на основе использования существующей ИТВ происходит слияние инфологического и датологического этапов проектирования. Это связано с тем, что при наличии данных разработчик имеет возможность использовать инструментальные средства СУБД на ранних этапах проектирования БД.

Физическое проектирование позволяет привязать датологическую модель к среде хранения. На этом этапе выбираются носитель данных, внутренние форматы их хранения, методы доступа к данным, методы сжатия данных, реализуются меры по безопасности данных. Этот этап мало связан с концептуальной моделью данных и практически не зависит от того, использовалась ли ИТВ при проектировании РБД. Более того, современные инструментальные СУБД берут на себя значительную часть решения проблем физического проектирования.

Основным понятием РМД является ***отношение***.

В РМД считается классическим следующее определение отношения. Пусть задано множество из n типов, или доменов, T_i ($i = 1, \ldots, n$), причем все они необязательно должны быть различными. Тогда r будет ***отношением*** определенным на этих типах, если оно состоит из двух частей, заголовка и тела (заголовок еще иногда называют схемой, переменной или отношением или ***интенсионалом отношения***, а тело — расширением, значением переменной или отношением или ***экстенсионалом отношения***), где

заголовок — это множество, состоящее из n атрибутов вида $A_i{:}T_i$; здесь A_i — имена атрибутов отношения r, а T_i — соответствующие имена типов;

тело — это множество, состоящее из m кортежей t; здесь t — множество компонентов вида $A_i{:}v_i$, в которых v_i — значение типа T_i, т.е. значение атрибута A_i в кортеже t [1].

Отношение можно представить как таблицу, где каждая строка — это кортеж, а каждый столбец — множество значений одного атрибута. Таблица, соответствующая отношению из k атрибутов, должна удовлетворять следующим свойствам:

- каждая строка представляет собой кортеж из k значений, принадлежащих k столбцам;
- каждый кортеж содержит точно одно значение (соответствующего типа) для каждого атрибута;
- порядок столбцов фиксирован $(1, 2, \ldots, k)$;
- порядок строк произволен;
- любые две строки различаются хотя бы одним элементом;
- строки и столбцы могут обрабатываться в любой последовательности, определяемой применяемыми операциями обработки.

На основе этих требований можно судить о некоторых проблемах представления данных в виде реляционных таблиц в процессе традиционного инфологического проектирования БД, в частности:

• из-за отсутствия реальных данных неочевиден выбор атрибута или атрибутов, которые обеспечили бы реализацию тезиса — любые две строки различаются хотя бы одним элементом. Этот выбор субъективен и далеко не всегда лучший. Использование ИТВ для назначения соответствующих атрибутов позволяет этот процесс формализовать и добиться наилучшего решения;

• назначение соответствующего типа для каждого атрибута также субъективно и впоследствии при заполнении таблиц реальными данными может оказаться неверным. Назначение типа атрибутов в ИТВ формализуется, так как основывается не на опыте и интуиции разработчика, а на анализе реальных данных.

Далее рассмотрим вопросы обеспечения целостности данных. Ограничение целостности — это логическое выражение, связанное с РБД, результатом которого всегда должно быть значение TRUE.

Обеспечение целостности отношений тесно связано с понятием ключей. Допустим, что K — множество атрибутов отношения R. В таком случае K является потенциальным ключом для R тогда и только тогда, когда оно обладает одновременно двумя приведенными ниже свойствами:

уникальностью — ни одно допустимое значение R никогда не содержит два разных кортежа с одним и тем же значением K;

необратимостью — никакое строгое подмножество K не обладает свойством уникальности.

Анализ требований к ключам позволил сделать следующие выводы.

Процедура назначения внешних ключей при традиционном проектировании РБД нетривиальна, не формализована, субъективна и не гарантирует лучшего решения. Она основывается на анализе предлагаемых схем отношений, которые далеко не всегда могут гарантировать принятие во внимание реальных данных.

При проектировании РБД на основе ИТВ процедуру назначения внешних ключей можно формализовать и таким образом добиться оптимального решения, удовлетворяющего требованиям к внешним ключам, так как анализ в этом случае проводится на основе использования существующих данных, а не на основе предлагаемых схем отношений.

В реляционной теории БД фундаментальной является концепция функциональной зависимости (ФЗ) [3]. Пусть R — отношение, а X и Y — произвольные подмножества множества атрибутов отношения R.

Тогда Y функционально зависимо от X, X→Y, если для любого допустимого значения переменной отношения R каждое значение множества X отношения R связано точно с одним значением множества Y отношения R. Иначе говоря, для любого допустимого значения переменной отношения R, если два кортежа переменной отношения R совпадают по значению X, они также совпадают и по значению Y.

Анализ ФЗ и мнений экспертов позволил сделать следующие выводы.

1. Единственный способ определения ФЗ для схемы отношения заключается в том, чтобы внимательно проанализировать семантику атрибутов. В этом смысле зависимости являются фактически высказываниями о реальном мире. Они не могут быть доказаны [3].

2. При традиционном проектировании РБД, несмотря на наличие правил вывода, выявление всех возможных ФЗ — процесс чрезвычайно трудоемкий и субъективно зависимый. Вероятность необнаружения всех ФЗ при анализе "вручную" очень велика. А невыявленные функциональные зависимости могут существенно сказаться на качественных характеристиках проектируемой РБД, на ее целостности, непротиворечивости, неизбыточности.

3. Процесс выявления ФЗ на основе имеющихся данных подлежит формализации, для которой можно использовать вычислительные машины, что, в свою очередь, позволит свести к минимуму невыявленные ФЗ, существенно ускорить процесс проектирования РБД.

К числу основополагающих принципов проектирования РБД относятся принципы ***нормализации отношений*** — аппарата ограничений на формирование отношений. Он позволяет устранить дублирование, обеспечивает непротиворечивость хранимых данных, уменьшает затраты на ведение БД. Анализ принципов нормализации и мнений экспертов позволил сделать следующие выводы:

• принципы нормализации при традиционном проектировании РБД являются не более и не менее, чем соображениями здравого смысла, записанными в формальном виде [3];

• попытка формализации процесса выявления и исключения транзитивных зависимостей на базе анализа только схемы отношения вряд ли может быть успешной. Это связано с тем, что даже профессиональный разработчик далеко не всегда может сформулировать критерии зависимостей даже для конкретного отношения, опираясь только на схему отношения. В случае наличия реальных данных в реляционных таблицах выявление и исключение транзитивных зависимостей можно автоматизировать и выполнять, исходя из содержимого этих данных.

При наличии только гипотетических данных и схемы отношения, которыми располагает разработчик РБД при традиционном проектировании, выявление многозначных зависимостей — задача чрезвычайно

трудоемкая, и ее решение не гарантировано от ошибок. Тем более является нетривиальной задача исключения многозначных зависимостей. При проектировании РБД, основываясь на реальных данных, задачи выявления и исключения многозначных зависимостей можно формализовать, тем самым снизить трудоемкость и улучшить качество проектирования.

При традиционном проектировании РБД активно используются средства семантического моделирования данных, при этом используется понятие сущностей. Сущности отображают объекты реального мира. Между сущностями могут быть связи. Связь определяется как ассоциация, объединяющая несколько сущностей. Сущности, включенные в связь, называются ее участниками, а число участников связи называются ее степенью. Связи в модели "сущность–связь" могут быть следующих типов: "один к одному", "один ко многим", "многие к одному" и "многие ко многим". На основе сущностей и связей строятся диаграммы "сущность–связь" (ER-диаграммы), которые в определенном смысле являются проектами БД. В результате анализа концепции построения ER-диаграммы сделаны следующие выводы:

• одной из существенных проблем, которая возникает при формировании ER-диаграмм, является назначение связей. При отсутствии реальных данных, во-первых, не очевидны участники связей, а во-вторых, — далеко не всегда определены типы связей;

• положение дел существенно меняется при наличии ИТВ. Из контекста данных, содержащихся в таблицах ИТВ, можно не только выявить реальные связи между сущностями, но и определить их тип. Причем при разработке соответствующих методов и алгоритмов этот процесс можно автоматизировать, что позволит исключить дефекты проектирования при назначении связей и существенно снизить трудоемкость решения данной проблемы.

Предложено неформальное определение ИТВ. Информация табличного вида имеет следующие свойства:

это информация, которая интуитивно воспринимается ее потребителями как таблица;

в табличном представлении информации нередко отсутствуют разделители строк и столбцов;

элементы данных нередко размещаются в нескольких строках;

типы элементов данных, соответствующих одному столбцу, могут различаться;

заголовки ИТВ могут включать в себя подзаголовки;

заголовки и/или подзаголовки одноименного столбца нередко размещаются в нескольких строках.

Укрупненную модель ИТВ можно представить следующим образом:

$$\text{ИТВ} = \{\text{NT}_1, \text{NT}_2, \ldots, \text{NT}_i, \ldots, \text{NT}_q\},$$

где q — число таблиц в наборе; NT_i — i-я нереляционная, неформализованная таблица набора,

$$\text{NT}_i = (\Pi_{i1}, \Pi_{i2}, \ldots, \Pi_{ij}, \ldots, \Pi_{in}),$$

где Π_{ij} — j-е поле i-й таблицы, обладающее перечисленными свойствами; n — число полей таблицы.

Как следует из модели, ограничения на исходные структуры не являются строгими, что обеспечивает широкий круг потенциальтных пользователей средств автоматизированного проектирования РБД на основе ИТВ.

В качестве одного из многочисленных примеров ИТВ может служить каталог электрооборудования автомобилей, фрагмент которого приведен на рис. 1.

В этой ИТВ нет разделителей заголовков, нет разделителей строк, присутствуют подзаголовки. Все эти разделители интуитивно воспринимаются человеком и представляют неразрешимую проблему для компьютера. Ни одно из существующих программных средств не переведет эту таблицу хотя бы в формат электронных таблиц (ЭТ) Excel, не говоря уже о том, чтобы получить реляционную таблицу, которую можно сразу использовать в существующей базе данных или вновь проектируемой. На рис. 2 приведен результат импорта данного текстового файла в формат ЭТ.

Результат говорит сам за себя; несколько столбцов объединились, а некоторые остались раздельными. Из-за отсутствия разделителей

Тип изделия	Код ОКП, ТУ или ГОСТ	Uном В	Устанавливаемые приборы					Количество	
			Спидометр тахометр	Указатель тока или напряжения	Приемник указателя топлива	Приемник указат. давлен., манометр	Приемник указателя температуры	Сигнализаторов	Ламп освещения
АР40.3801.000	...	12	АР40.3802	АР40.3812	АР40.3806	-	АР40.3807	21	8
	ТУАР.3801.004-98		АР40.3801 060	000	000		000		
АР41.3801.000	...	12	АР40.3802	АР40.3812	АР40.3806	-	АР41.3807	21	8
	ТУАР.3801.004-98		АР41.3801 060	000	00		00		
АР52.3801.000	...	12	АР52.3802	-	АР51.3806	-	АР51.3808	12	5
	ТУАР.3801.001-95		000		000		000		
			АР52.3801 060						
АР60.3801.000	...	12	АР60.3801	АР60.3812	АР60.3806	АР60.3801	АР60.3801	32	7
	ТУАР.3801.002-96		100	000	000	200	200		
АР60.3801.000-01	...	12	АР60.3801	АР60.3812	АР60.3806	АР60.3801	АР60.3801	15	7
	ТУАР.3801.002-96		100	000	000	200	200		
АР60.3801.000-04	...	12	АР60.3801	АР60.3812	АР60.3806	АР60.3801	АР60.3801	32	7
	ТУАР.3801.007-99		100-04	000	000	200	200		
АР60.3801.000-05	...	12	АР60.3801	АР60.3812	АР60.3806	АР60.3801	АР60.3801	16	7
	ТУАР.3801.007-99		100-04	000	000	200	200		

Рис. 1. Фрагмент каталога электрооборудования автомобилей

	A	B	C	D	E	F	G	H	I	J	K
1	Тип	Код	ОКП,	Uном	Устанавливаемые приборы				Количество	Мас-	Завод-	Облас
2	изделия	ТУ или ГОСТ		В					са,	изготови-	(по основным моделям	
3			Спидометр	Указатель	Приемник	Приемник	Приемник	Сиг-	Ламп	кг	тель	
4			тахометр	тока или	указате-	указат.	указате-	нали-	осве			
5				напряже-	ля топ-	давлен.,	ля тем-	зато-	ще-			
6				ния	лива	манометр	пературы	ров	ния			
7												
8	АР40.3801.000	...		12	АР40.3802	АР40.3812	АР40.3806	-	АР40.3807	21	8	-	АО RAR	Автомо
9		ТУАР.3801.004-98		АР40.3801	000	000		000					Москвич-2141	
10			060											
11	АР41.3801.000	...		12	АР40.3802	АР40.3812	АР40.3806	-	АР41.3807	21	8	-	АО RAR	Автомо
12		ТУАР.3801.004-98		АР41.3801	000	00		00					Москвич-2141	
13			060											
14	АР52.3801.000	...		12	АР52.3802	-	АР51.3806	-	АР51.3808	12	5	1.08	АО RAR	Автомоби.
15		ТУАР.3801.001-95		000		000		000						
16				АР52.3801										

Рис. 2. Результат импорта текстового файла в формат электронной таблицы

строк в результирующей таблице смешаны строки и понять, какие данные относятся к какой строке, практически невозможно. Даже если и удается привести подобную ЭТ к пригодному для работы виду, не будут реализованы следующие преимущества РБД:

- БД позволяют не только вводить данные в таблицы, но и контролировать правильность ввода;
- БД позволяют хранить и обрабатывать значительно бóльшие, по сравнению с ЭТ, объемы данных;
- в БД возможно создание связей между таблицами, что позволяет совместно использовать данные из нескольких таблиц, при этом для пользователя они будут представляться одной таблицей;
- предоставляя связи между отдельными таблицами БД позволяют избежать дублирования данных, сэкономить память компьютера, а также увеличить скорость и точность обработки информации;
- у БД значительно больше возможностей при работе нескольких пользователей с одними и теми же данными, при этом все пользователи гарантированно будут работать с актуальными данными;
- БД в отличие от ЭТ имеют развитую систему защиты от несанкционированного доступа;
- БД позволяет организовать многопользовательский доступ для значительно большего, по сравнению с ЭТ, числа пользователей.

Важно отметить, что перечисленные преимущества БД перед ЭТ еще в большей степени являются преимуществами БД перед ИТВ, представленной на бумаге, в текстовом формате или в формате текстовых процессоров.

Преимущества РБД перед ИТВ, рассмотренные преимущества использования существующих данных для качественного решения

проблем проектирования РБД, а также насущная потребность в БД на предприятиях, занимающихся обработкой информации, — все это определяет актуальность разработки средств автоматизированного проектирования РБД на основе существующей ИТВ.

Даже неформального определения ИТВ и краткого анализа традиционной теории проектирования РБД достаточно, чтобы сформулировать основные задачи, которые необходимо решить для реализации автоматизированного процесса преобразования ИТВ в РБД. К ним относится разработка методик:

- преобразования нереляционных таблиц в реляционные;
- нормализации заполненных реляционных таблиц;.
- назначения ключевых полей в заполненных реляционных таблицах;
- выявления и формирования связей между заполненными реляционными таблицами.

Предлагаются следующие пути решения сформулированных задач:

в рамках методики преобразования нереляционных таблиц в реляционные необходима разработка: алгоритма исключения подзаголовков, алгоритма выявления и исключения подзаголовков внутри таблицы, алгоритма выявления и исключения заголовков (расположенных в первом столбце таблицы), алгоритма исключения дублирования строк в таблице, метода приведения значений данных одноименного столбца к одному типу;

в рамках методики нормализации заполненных реляционных таблиц необходимо разработать методы приведения таблиц к 1, 2, 3 и 4-й нормальным формам;

в рамках методики выявления и формирования связей между заполненными реляционными таблицами необходимо разработать методы выявления связей “один к одному”, “один ко многим”, “многие к одному” и “многие ко многим”.

В работе [4] разработаны названные алгоритмы и методы.

Проведение перечисленных мероприятий, с одной стороны, обеспечит построение реляционной БД с уже внесенными данными в таблицы БД, а с другой стороны, позволит использовать в ходе разработки и эксплуатации БД существующую теорию построения реляционных БД.

В результате анализа проблемы проектирования РБД на основе использования существующей ИТВ сделаны следующие выводы.

1. Традиционная теория проектирования РБД пока далека от совершенства.

2. В случае если имеется ИТВ, возможна формализация проектных процедур и получение проектных решений, имеющих лучшие характеристики.

3. Методология проектирования РБД на основе существующей ИТВ должна органично сочетаться с проверенной годами традиционной методологией проектирования РБД.

4. Информация табличного вида представлет собой информацию, которая интерпретируется затнтересованными в ней людьми двумерными таблицами, а они не удовлетворяют требованиям к реляционным таблицам.

5. Мотивами преобразования ИТВ в РБД являются потребности использования сформулированных преимуществ РБД, а также внедрение в существующие БД данных из ИТВ.

СПИСОК ЛИТЕРАТУРЫ

1. Дейт К. Дж. Введение в системы баз данных. – 7-е изд. / Пер. с англ. – М.: ИД “Вильямс”, 2001. – 1072 с.
2. Григорьев Ю. А., Ревунков Г. И. Банки данных: Учеб. для вузов. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2002. – 320 с.
3. Дейт К. Дж. Введение в системы баз данных. – 8-е изд.: Пер. с англ. – М.: ИД “Вильямс”, 2005. – 1328 с.
4. Брешенков А. В. Методы решения задач проектирования реляционных баз данных на основе использования существующей информации табличного вида. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2007. – 154 с.

Статья поступила в редакцию 23.01.2007

Александр Викторович Балдин родился в 1951 г., окончил МВТУ им. Н.Э. Баумана в 1974 г. Д-р техн. наук, начальник отдела интеграции информационных систем МГТУ им. Н.Э. Баумана. Автор 73 научных работ в области автоматизации и моделирования процессов управления и баз данных.

A.V. Baldin (b. 1951) graduated from the Bauman Moscow Higher Technical School in 1974. D. Sc. (Eng.), head of department for integration of information systems of the Bauman Moscow State Technical University. Author of 73 publications in the field of automation and simulation of management processes and data bases.