УДК 681.3.06

С. С. Таянский

ТЕХНОЛОГИЯ ПОСТРОЕНИЯ КРУПНОМАСШТАБНОЙ БАЗЫ ДАННЫХ

1. Введение

Концептуальная схема является центральным звеном в архитектуре систем баз данных (БД), обеспечивающим интерпретацию внешнего представления данных в их внутреннее представление в информационных системах. В проблематике моделей данных и систем БД основное место занимают вопросы проектирования и поддержки концептуальной схемы [1, 2]. В настоящей работе рассматривается класс реляционных БД, получивших название несогласованных БД. Подразумевается, что БД состоит из множества отношений, естественное соединение которых необязательно сохраняет эти отношения, то есть проекция естественного соединения на множество атрибутов отношения необязательно равна этому отношению. Одновременно предполагается, что «несогласованность» отношений не является произвольной, а определяется условиями семантической целостности данных, хранимых в БД.

Ключевая роль в поддержке семантической целостности отводится ограничениям целостности специального вида — функциональным зависимостям между атрибутами отношения БД (*F*-зависимости). Понятие функциональной зависимости хорошо известно и детально описано в литературе, например в [2, 3]. Однако свойства этих зависимостей и их влияние на процессы ведения крупномасштабной БД (и, соответственно, на вид схемы БД) изучены недостаточно. Под крупномасштабной БД будем понимать систему с распределенной организацией обработки данных при наличии семантической независимости локальных БД.

Одной из важных задач поддержки целостности крупномасштабной БД является способность глобально поддерживать заданные локальные *F*-зависимости. Сложность поддержки ограничений состоит в том, что проверка *F*-зависимостей внутри локальных отношений суммарно не всегда обеспечивает выполнение ограничений «в масштабе» всего универсального отношения (отношение, полученное путем соединения всех локальных отношений).

Таким образом, целью работы является определение основных свойств операций естественного соединения и проекции при допущении несогласованности значений в отношениях БД, а также определение методики поддержки функциональных зависимостей при интеграции БД на этапе построения крупномасштабной системы.

2. Прикладные средства ведения крупномасштабной базы данных

Запросы, которые в общем случае представляют собой произвольные функции над отношениями, часто используют развитые языки высокого уровня. Наиболее распространенными языками для реляционных систем управления базами данных являются алгебраические языки, позволяющие выражать запросы средствами специализированных операторов, применяемых к отношениям. Такие языки должны обеспечивать не только абстрактные функции, но и дополнительные потребности пользователей.

Рассмотренные в [4] модифицированные операции реляционной алгебры позволяют решить ряд нестандартных задач при ведении крупномасштабной БД. Для расширения возможностей проверки выполнимости *F*-зависимостей введем дополнительные операции над несогласованными отношениями БД.

Пусть U — непустое множество атрибутов. Реляционную алгебру над U будем обозначать $\Re^{RDB} = < U, \Omega >$, где Ω — множество операций реляционной алгебры. При допущении неопределенностей в отношениях, операции алгебры $\Re^{RDB}$ теряют ряд важных свойств, в частности ассоциативность соединения. Расширенную алгебру, дополненную модифицированными операциями проекции и соединения, позволяющими выполнять операции с неопределенностями, будем обозначать $\Re^{\setminus DB} = < U, \Omega' >$, где Ω' — расширенное множество операций реляционной алгебры. Дополнение представляет собой две операции: операцию соединения, в обозначении $><^{U}$, при которой каждая строка попадает в соединение, и операцию проекции в обозначении π_{X}^{U}, исключающую строки с неопределенными значениями. Множество полных отношений (недопускающих неопределенные значения) обозначим *Rel*, а множество частичных отношений (допускающих неопределенности) — $Rel\uparrow$.

Для обобщенных операций проекции и естественного соединения можно выделить область определения из отношений, входящих в множество $Rel\uparrow$. Результат применения этих операций, то есть область значений, при этом также остается в множестве $Rel\uparrow$. С другой стороны, областью определения обобщенных операций также является множество *Rel*, но область значений этих операций может принадлежать множеству $Rel\uparrow$. В случае, когда для каждого кортежа каждого отношения найдется хотя бы один соединимый кортеж, область значений алгебры $\Re^{\setminus DB}$ не выходит из множества *Rel*.

Рассмотрим базу d со схемой $\boldsymbol{R} = \{AB, BD, AC, CD\}$, представленную на рис. 1, и множество функциональных зависимостей $\boldsymbol{F} = \{A \to B, B \to D, A \to C, C \to D\}$.

A	B
a	b
a_1	b_1

B	D
b	d
b_1	d

A	C
a	c
a_1	c_1

C	D
c	d
c_1	d

Рис. 1. *Пример полностью соединимых отношений*

Предположим, необходимо получить универсальное отношение, то есть отношение, включающее все атрибуты рассматриваемой БД, для проверки некоторой F-зависимости из множества $\boldsymbol{F}$. Применим к БД обобщенные операции из алгебры $\Re^{NDB}$ (рис. 2).

A	B	C	D
a	b	c	d
$a1$	$b1$	$c1$	d

Рис. 2. Результат соединения отношений операцией $><^{e}$

Как видно из приведенного примера, результатом операции $><^{U}$ является отношение из Rel, то есть в данном случае эта операция аналогична операции $><$ алгебры $\Re^{RDB}$.

Напомним, что рассматривается класс крупномасштабных БД, где допустимы несогласованные отношения. Предположим, что отношения, изображенные на рис. 1, дополнены кортежами $< a, b_2 >$ и $< a, c_2 >$ (рис. 3).

A	B
a	b
a_1	b_1
a	b_2

B	D
b	d
b_1	d

A	C
a	c
a_1	c_1
a	c_2

C	D
c	d
c_1	d

Рис. 3. Несогласованные отношения

Результат соединения таких отношений операцией $><^{\ell}$ изображен на рис. 4, откуда видно, что универсальное отношение является отношением, входящим в множество $Rel\uparrow$.

A	B	C	D
a	b	c	d
$a1$	$b1$	$c1$	d
a	b_2	$\perp$	$\perp$
a	$\perp$	c_2	$\perp$

Рис. 4. Результат соединения несогласованных отношений (частичное отношение)

При необходимости возврата к исходным отношениям можно воспользоваться модифицированной операцией $\pi^{(\ell}_{X}$, в результате применения которой получим отношения, аналогичные рис. 3.

Таким образом, в алгебре $\Re^{NDB}$ область значений расширяется до множества $Rel \cup Rel\uparrow$. Такое допущение позволяет осуществлять глобальную поддержку F-зависимостей традиционным методом, то есть через универсальное отношение.

Определим отображение φ как пару вида $<\{R_i\}, \boldsymbol{R}>$, тогда области определения алгебр $\Re^{ADB}$ и $\Re^{NDB}$ можно записать формально. Для алгебры $\Re^{ADB}$ функции φ и φ^{-1} будут иметь вид:

$$\varphi_{ADB}: \{R_i\} \to \boldsymbol{R} \in Rel.$$
$$\varphi^{-1}_{ADB}: \{\boldsymbol{R}\} \to R_i \in Rel.$$

Для алгебры $\Re^{NDB}$ отображение φ примет вид:

$$\varphi_{NDB}: \{R_i\} \to \boldsymbol{R} \in Rel \cup Rel\uparrow.$$
$$\varphi^{-1}_{NDB}: \{\boldsymbol{R}\} \to R_i \in Rel.$$

Базовая реляционная алгебра является основным инструментом управления реляционными данными и проектирования структур БД. Простота и лаконичность этой алгебры описана в работе [5]. Развитие распределенных систем привело к соответствующим исследованиям в данной области. Одним из результатов таких исследований является предложенная расширенная алгебра $\Re^{NDB}$.

И, хотя рассмотренная алгебра является инструментом для решения ограниченного класса задач, она играет роль связующего звена теоретических исследований с практическими вопросами глобальной поддержки целостности крупномасштабных БД.

3. Поддержка функциональных зависимостей крупномасштабной базы данных

Положения предыдущего раздела дали основание для исследования вопроса поддержки целостности крупномасштабной БД, основанного на проверке выполнимости локальных F-зависимостей на универсальном отношении. Содержательно этот процесс можно разбить на четыре этапа.

При интеграции информационных систем поддержка F-зависимостей, определенных в локальных отношениях, не гарантирует семантической однозначности данных в ключевых атрибутах. Один из способов однозначной идентификации ключа можно проверить при соединении всех отношений, то есть построением универсального отношения. При этом сократить количество соединяемых отношений можно с использованием модифицированного ограничения на внешний ключ (МОВК) [4], смысл которого сводится к тому, что БД удовлетворяет МОВК относительно $\boldsymbol{F}$, если $\forall R_i \in \boldsymbol{R}$ и $t \in r(R_i)\, \exists t' \in R_i^+$, таких что $\forall R_j \subseteq R_i^+ \;\; \exists t'(R) \in r(R_j)$, и, в частности, $t'(R_i) = t$. Напомним, что R_i^+ означает замыкание R_i относительно $\boldsymbol{F}$. $r(R_i)$ — экземпляр отношения R_i, а t — строку отношения.

Таким образом, первый этап алгоритма поддержки F-зависимости заключается в определении множества отношений, отвечающих МОВК, для каждого ключа во всех локальных отношениях. Как отмечалось выше, такие множества могут содержать как одно отношение (непосредственно содержащее рассматриваемую F-зависимость), так и все отношения БД. Множество схем соединяемых отношений в дальнейшем будем называть *планом соединения*, или *планом*. Схему отношения при соединении по плану соединения будем называть *схемой плана соединения*, или *схемой плана*, и обозначим $plan(\{R_i\})$. При этом возможны три следующие ситуации. Пусть БД имеет схему $\boldsymbol{R} = \{R_1, \ldots, R_n\}$, тогда

1. $plan(\{R_n\}) = \{R_1, R_2, \ldots, R_i, \ldots, R_n\}$;
2. $plan(\{R_p\}) = \{R_j, R_{j+1}, \ldots R_p\}$;
3. $plan(\{R_k\}) = \{R_k\}$.

Таким образом, при обновлении определенных отношений для некоторых планов соединения возможно значительное сокращение временных затрат путем исключения «дорогостоящей» операции естественного соединения.

В тех случаях, когда все же полное соединение неизбежно, необходимо перейти ко второму этапу алгоритма, то есть воспользоваться редуктивным алгоритмом Грэхема, описанным в [4], и определить, к какому классу схем относится схема плана. Принадлежность схемы плана к классу ациклических схем дает возможность осуществлять поддержку F-зависимостей попарным соединением отношений плана.

Сложность связей между локальными отношениями распределенных информационных систем практически исключает успешное завершение редуктивного алгоритма, то есть схема плана соединения в основном относится к классу циклических схем.

На следующем этапе алгоритма необходимо преобразовать циклическую схему к ациклическому виду, точнее, к блочно-ациклическому, то есть необходимо некоторое эквивалентное преобразование схемы плана. Пусть для R_i схема плана имеет вид $plan(\{R_p\}) = \{R_1, R_2, ..., R_{i-1}, ..., R_p\}$ и пусть циклический блок состоит из множества $R'_i = \{R_{i-1}, R_i, R_{i+1}\}$. Тогда в результате перехода к ациклической схеме схема плана примет вид: $plan'(\{R_p\}) = \{R_1, R_2, ..., R'_i, ..., R_p\}$, где отношение со схемой R'_i является результатом естественного соединения отношений циклического блока.

Как отмечалось в [4], преобразование схемы связано с увеличением временных затрат при попарном соединении. С другой стороны, блочно-ациклическое представление схемы исключает полное соединение отношений, что снижает время проверки глобальной выполнимости F-зависимостей, а также исключает сложность работы с неопределенными значениями.

Все перечисленные шаги алгоритма выполняются на этапе интеграции баз данных. Выделение плана соединения, определение его вида, а также построение блочно-ациклической схемы выполняются один раз, на начальном этапе разработки информационной системы.

Следующий шаг связан непосредственно с поддержкой F-зависимостей, которая осуществляется на этапе ведения данных. Поддержка F-зависимостей ведется по ключу, то есть проверяется однозначность значений ключевого атрибута. Распределенное ведение крупномасштабной БД требует, как отмечалось ранее, проверки уникальности ключа во всех локальных отношениях, где существует ключевой атрибут, а именно проверки однозначности ключа в схеме плана соединения. При необходимости предварительная модификация схемы плана обеспечивает выполнение условия, когда попарное соединение эквивалентно соединению в целом, что снижает время проверки однозначности ключа.

Общую схему алгоритма можно представить в виде двух основных этапов. Первый выполняется на стадии проектирования БД и состоит, в свою очередь, из нескольких шагов, описанных выше. Второй этап, непосредственно поддержка F-зависимостей, выполняется при ведении БД и основан на результатах, полученных на предыдущих шагах алгоритма. Содержательная схема алгоритма представлена на рис. 5.

При поддержке целостности БД используется традиционный подход, основанный на F-зависимостях, но предлагаемые средства для решения этой задачи значительно расширяют методы ведения крупномасштабных БД. При этом одним из важных результатов является допущение несогласованности данных в локальных отношениях, за исключением значений в ключевых атрибутах. Это необходимо для сохранения связи между отношениями, а также для сохранения крупномасштабной БД как единого семантического целого.

Обобщенная последовательность шагов проектирования крупномасштабной БД представлена на рис. 6.

Предлагаемая последовательность позволяет определить структуру БД, а также представить локальные БД и разместить их по узлам сети (или интегрировать существующие локальные представления

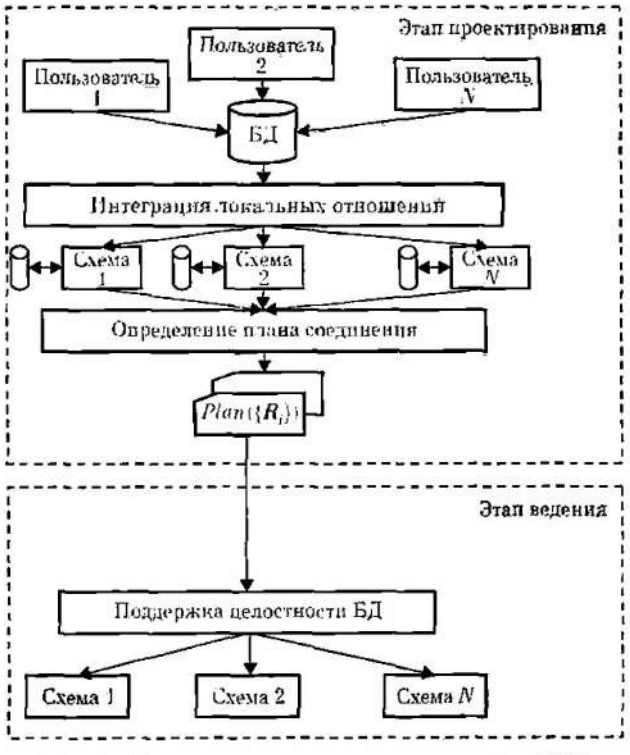

Рис. 5. Схема поддержки крупномасштабной БД

Рис. 6. Этапы построения крупномасштабной БД

в единую глобальную логическую структуру). Кроме того, проектируется технология ведения, обеспечивающая эффективную поддержку локальных БД.

Системы управления с такими возможностями обычно используют одинаковые модели данных, описывающие каждую локальную БД, входящую в систему, и поддерживают справочник о размещении БД. С учетом этого факта, после выполнения проектировщиком этапов 1–4, приложения получают возможность использовать преимущества однородности и согласованности БД, что эквивалентно использованию общесистемного стандарта, принятого в организации централизованных систем. Этап 6 может быть выполнен в каждом узле системы автономно с тем, чтобы учесть особенности локальных БД. Отметим, что выполнение этапов 1, 2, 3 и 6 подобно выполнению этапов при проектировании централизованных БД, поэтому в дальнейшем необходимо обратить внимание на модификацию и дополнения этапов проектирования локальных структур и глобальной логической схемы БД.

4. Заключение

В настоящей статье рассмотрено возможное расширение методологии проектирования БД на крупномасштабные архитектуры. При этом в крупномасштабных системах логически целостная БД может быть фрагментирована и распределена по узлам сети. Как отмечалось в работе, фрагментация и распределенное ведение БД без внимательного централизованного планирования часто приводят к беспорядку и несогласованности данных. Предлагаемая процедура поэтапного проектирования крупномасштабной БД учитывает это важное обстоятельство.

Практическая значимость предложенных результатов состоит в расширении возможностей использования традиционных методов поддержки целостности при интеграции локальных БД в единую крупномасштабную систему. Дальнейшие исследования в этой области следует направить на изучение свойств запросов, влияющих на изменение данных, и построение модели транзакции для обеспечения согласованного взаимного доступа между локальными данными.

Список литературы: 1. *Бойко В. В., Савинков В. М.* Проектирование информационной базы автоматизированной системы на основе СУБД. — М.: Финансы и статистика, 1982. — 174 с. 2. *Богодист В. П., Буслик Н. И., Дедиков Э. А., Жадан А. Н.* Методы проектирования схемы реляционной базы данных // Техника средств связи: Сер. ТЭУ. — М., 1985. — Вып. 2. — С. 49–51. 3. *Гарсиа-Молина Г., Ульман Дж., Уидом Дж.* Системы баз данных. Полный курс: Пер. с англ. — М.: ИД «Вильямс», 2003. — 1088 с. 4. *Мейер Д.* Теория реляционных баз данных. — М.: Мир, 1987. — 608 с. 5. *Дейт К.* Введение в системы баз данных: Пер. с англ. — М.: ИД «Вильямс», 2001. — 1072 с.

Поступила в редакцию 14.09.2006