ИНТЕГРАЦИЯ ГЕТЕРОГЕННЫХ ИСТОЧНИКОВ ДАННЫХ НА ОСНОВЕ РЕКУРСИВНОЙ ДЕКОМПОЗИЦИИ

Алексей Кашников, Людмила Лядова

Abstract: Управление данными на современном предприятии характеризуется наличием большого количества разнородных источников данных, не связанных едиными механизмами управления, в том числе и слабоструктурированных или неструктурированных данных и т.п. При этом модель данных, лежащая в основе большинства систем, – реляционная – не является эффективной для решения многих задач. Раздельно существуют системы аналитической обработки и оперативного управления данными, системы управления документами и пр. Различные задачи требуют использования различных моделей представления данных. На этом фоне ставится задача интеграции гетерогенных данных, эффективное решение которой требует создания модели интеграции (или «интеграции» различных моделей данных), которую можно было бы рассматривать как основу для реализации системы, поддерживающей оперативное управление разнородными данными и их аналитическую обработку. Предлагается модель интеграции данных, в которой должны поддерживаться унифицированное представление разнородных источников данных, управление ограничениями целостности, управление выполнением операций манипулирования данными и запросов, согласование данных из разных источников, возможность расширения и настройки на новые источники данных. Существующие подходы к интеграции гетерогенных данных имеют ограничения, которые не позволяют в полной мере говорить об их универсальности. Предлагаемый подход к интеграции основан на рекурсивной декомпозиции источников данных, при которой каждый источник данных последовательно разбивается на атомарные элементы данных, причем на каждом уровне рекурсивной вложенности данные и их описания представляются единообразно. Такая модель позволяет осуществлять интеграцию различных источников данных на любом уровне посредством задания связей между произвольными элементами схемы, ограничений целостности и допустимых операций. Разработанная модель представления источников данных, которая позволяет осуществлять многоуровневую интеграцию гетерогенных источников данных в единое информационное пространство, обеспечивает поддержку ограничений целостности на любом уровне интеграции источников данных, а также поддержку структурных и ассоциативных связей между источниками данных на любом уровне интеграции. Обеспечивается возможность динамического изменения схемы данных, а также расширяемость системы за счет новых моделей данных.

Keywords: модель данных, гетерогенные системы, интеграция.

ACM Classification Keywords: H. Information Systems: H.2. DATABASE MANAGEMENT (E.5): H.2.1. Logical Design – Data models.

Введение

Активное развитие информационной инфраструктуры предприятий привело к возникновению проблем, препятствующих эффективному использованию имеющихся данных и, как следствие, принятию качественных управленческих решений:

Большое количество разнородных источников данных с различными механизмами управления.

Реляционная модель в качестве «универсальной» основы управления данными. Между тем, реляционная модель данных не является адекватной для решения многих современных задач – различные задачи требуют различных моделей представления данных.

«Излишняя» функциональность современных систем управления базами данных (БД) и, как следствие, излишняя дороговизна. Современные коммерческие СУБД предоставляют огромный объем функциональности, большая часть которой остается невостребованной.

Большой объем данных, значительная часть которых представлена слабоструктурированными или неструктурированными данными.

Раздельное существование систем аналитической обработки данных и систем оперативного управления данными.

Единое информационное пространство является ключевым фактором успешности современного бизнеса. В Клермонтском отчете [1] говорится, что одной из важнейших целей сообщества баз данных является переход от управления традиционными БД к задаче управления наборами структурированных, полуструктурированных и неструктурированных данных, распределенных по многим репозиториям предприятий и узлам Internet. Как следствие, возникает задача интеграции гетерогенных данных, эффективное решение которой требует создания *модели интеграции* (или «интеграции» различных моделей данных).

Общей целью данной работы является разработка *системы интеграции гетерогенных источников данных*. В данной статье рассматривается задача разработки *модели интеграции данных*, в которой должны быть отражены следующие аспекты:

- унифицированное представление источников данных;
- управление ограничениями целостности;
- управление выполнением операций манипулирования данными и запросами;
- согласование данных из разнородных источников;
- возможность расширения и настройки на новые источники данных.

Существующие подходы к интеграции гетерогенных данных имеют следующие особенности, которые не позволяют в полной мере говорить об их универсальности:

Вопросы интеграции гетерогенных данных рассматриваются преимущественно с позиций осуществления унифицированных запросов, а не комплексного управления данными.

Манипулирование данными рассматривается преимущественно при интеграции только реляционных СУБД (либо в случае, если нереляционная СУБД имеет реляционную оболочку).

В качестве единой модели данных рассматривается реляционная схема, либо в более общем случае – табличное представление данных.

Для достижения поставленной цели должны быть решены также следующие задачи:

разработка языка запросов и манипулирования интегрированными данными;

разработка программной модели и архитектуры системы интеграции гетерогенных источников данных;

разработка механизмов отображения существующих источников данных на интегрированную модель.

Рассматриваемый в данной работе подход к интеграции основан на *рекурсивной декомпозиции источников данных*, при которой каждый источник данных сводится к неделимым элементам данных, причем на каждом уровне рекурсивной вложенности данные и их описания представляются единообразно. Такая модель позволяет осуществлять интеграцию различных источников данных на любом уровне декомпозиции посредством задания связей между произвольными элементами схемы, ограничений целостности и допустимых операций.

Гетерогенные информационные системы: подходы к интеграции данных

Необходимость в интеграции гетерогенных данных возникает в различных условиях и требует различных подходов, соответствующих этим условиям и требованиям, которые предъявляются к информационным системам. В некоторых случаях достаточно обеспечить работающие информационные системы шлюзами

для обмена данными, в других – обеспечить единое представление информационного пространства для возможности выполнения запросов, охватывающих различные источники данных, в третьих – необходимо предоставить комплексный инструментарий по управлению данными в гетерогенной среде, включая выполнение транзакций и поддержку ограничений целостности. Более глубокая интеграция требует более сложного решения специфических проблем, обусловленных гетерогенной средой. Анализ различных аспектов функционирования гетерогенных систем демонстрирует необходимость применения иных подходов к управлению данными, чем в случае разработки традиционных систем. Разработчики информационных систем (ИС), часто сталкиваясь с набором слабо связанных источников данных, вынуждены каждый раз решать низкоуровневые задачи управления данными в разнородных коллекциях. В число этих задач входят обеспечение возможностей поиска и запрашивания данных; соблюдение правил, ограничений целостности, соглашений об именовании и т.д.; отслеживание происхождения данных; обеспечение доступности, восстановления и контроля доступа; управляемое развитие данных и метаданных и пр. Гетерогенность данных разделяется на физическую и семантическую [16].

Физическая гетерогенность подразумевает различия в представлении данных; может выражаться в различии типов данных, реализаций моделей данных (например, различные реляционные системы). Могут различаться также языки описания процедур, триггеров, языки запросов, манипулирования и определения данных. Кроме этого, могут различаться и сами модели данных. Данные могут быть представлены не только в БД, но и в электронных таблицах или почтовых файлах и т.д. Проблема физической гетерогенности решается посредством *введения стандартов взаимодействия* (ODBC, DAO, OLE DB, ADO, ADO.NET). Достоинством универсальных механизмов является возможность применения одних и тех же средств доступа к разным типам источников, поэтому приложения легко модифицировать, если необходима замена СУБД. Но за универсальность приходится платить невозможностью доступа к уникальной функциональности, специфичной для конкретной СУБД, снижением производительности.

Семантическая гетерогенность проявляется в различиях в наименованиях данных, значениях и логических структурах [17]. В данной области проводится большое количество исследований, но до сих пор не появилось какого-либо стандарта. Прототипы реляционных языков (MSQL, IDEAL) были разработаны для решения этой проблемы. Многие специальные функции этих языков нуждаются в выразительной мощности, выходящей за пределы существующих реляционных систем, диалектов SQL. Требуются мультиреляционные операции и логика исчисления предикатов более высокого порядка, тогда как SQL основан на реляционной алгебре и логике предикатов первого порядка.

Гетерогенная система должна иметь собственные *механизмы управления целостностью*. Каждая локальная БД может иметь свои механизмы поддержания целостности, но контроль целостности на глобальном уровне необходим, так как на нем могут появиться новые связи и новые ограничения. Еще одна задача – поддержка *транзакций*. В гетерогенной среде концепция глобальной транзакции, состоящей из строго определенных субтранзакций, может быть слишком строгой.

В качестве новой абстракции управления данными в таких сценариях вводится понятие пространства данных [12]. *Пространство данных* должно содержать всю информацию, необходимую конкретной организации, несмотря на формат и местоположение этой информации, а также моделировать развитый набор связей между репозиториями данных. Пространство данных моделируется как набор участников и связей. *Участниками пространства данных* являются индивидуальные источники данных: они могут быть реляционными БД, репозиториями XML, текстовыми БД, Web-сервисами и пакетами программного обеспечения. Они могут храниться или быть потоками данных (локально управляемыми системами потоков данных) или даже сенсорными установками. Некоторые участники могут поддерживать выразительные языки запросов, а другие – быть неинтеллектуальными и поддерживающими лишь ограниченные интерфейсы для формулировки запросов (структурированные файлы, Web-сервисы и пр.). Участники могут быть структурированными (например, реляционными БД), полуструктурированными (XML, коллекции кода) или полностью неструктурированными. Некоторые источники будут поддерживать операции обновления, другие – допускать только добавление (в целях архивации), третьи – не допускать

изменений вообще. Пространство данных должно *моделировать любой вид связи между двумя (или несколькими) участниками* (один участник является представлением или репликой другого участника; один участник был «произведен» из других; источники создавались независимо, но отражают одну и ту же физическую систему и т.п.). Пространства данных могут *вкладываться* одно в другое, *перекрываться*, поэтому в пространстве данных должны содержаться *правила разграничения доступа*.

Отличительными *свойствами систем пространств данных* (DSSP) являются следующие:

DSSP должны работать с данными и приложениями в разнообразных форматах, доступных от многих систем через различные интерфейсы: от DSSP требуется поддержка всех данных пространства данных, без каких-либо исключений (как это бывает при использовании СУБД).

Хотя DSSP обеспечивает средства интегрированного поиска, запрашивания, обновления и администрирования пространств данных, те же самые данные часто могут быть доступны для чтения и обновления через собственный интерфейс системы, непосредственно управляющей данными: в отличие от СУБД, DSSP не имеет полного контроля над своими данными.

Могут обеспечиваться разные уровни услуг по обработке запросов к DSSP, в некоторых случаях они могут возвращать приблизительные ответы (если некоторые источники становятся недоступными, DSSP может обеспечить наилучший из возможных результат на основе данных, доступных во время выполнения запроса).

DSSP должны поддерживать средства для обеспечения более тесной интеграции данных пространства, если это необходимо.

Системы объединенных мультибаз данных, как общее решение для проблем межоперационных разнородных систем данных, обеспечивают однородный доступ к данным, сохраненным в множественных БД, которые включает несколько различных моделей данных [7, 5, 14, 16]. *Система мультибаз данных* (MDBS) – это система, которая постоянно находится «невидимо» над существующими БД и файловыми системами, называемыми локальными, и представляет для пользователя иллюзию единой БД. В MDBS поддерживается единая глобальная схема БД, в которой пользователи формируют запросы и осуществляют модификации данных, а локальные системы баз данных фактически поддерживают все данные пользователя. Глобальная схема интегрирует схемы локальных БД.

MDBS транслирует глобальные запросы в запросы к соответствующей локальной БД для фактической обработки, объединяет результаты и генерируют конечный результат для пользователя. Кроме того, MDBS координирует и аварийно прекращает глобальные транзакции для локальных систем БД, которые обрабатывают их, чтобы поддержать непротиворечивость данных внутри локальных баз данных.

Модели интеграции данных

Существующие модели интеграции данных можно разделить на три категории: модель с глобальной схемой, интегрированные базы данных и подход на основе языка мультибаз данных [11].

При реализации *модели с глобальной схемой* обязательным условием интеграции выступает наличие *глобальной схемы данных* [16]. Целью создания такой модели является интеграция данных из различных источников и предоставление пользователю унифицированного представления этих данных. Такое представление является согласованным объединением данных, к которому пользователь может адресовать запросы. В системе, основанной на такой модели, одной из главных задач является установление соответствия (отображения) между множеством источников данных и глобальной схемой. *Схема источника* описывает его структуру, в которой находятся реальные данные, а *глобальная схема* обеспечивает интегрированное виртуальное представление источников. Соответствия в *отображении* устанавливают связь между элементами глобальной схемы и схем источников. Определение виртуальной системы с глобальной схемой является достаточно общим и позволяет охватить все подходы, встречаемые в литературе. Конкретные подходы отличаются характеристиками отображения и выразительной мощностью различных языков схем и запросов.

Интегрированные базы данных основаны на моделях, в которых глобальная схема отсутствует, каждый источник представляется в структурированном виде, а *связи устанавливаются непосредственно между источниками* [11]. Такая схема предоставляет локальным базам больше возможности управления разделяемой информацией, управление получается децентрализованным. Степень интеграции не обязана быть полной, как в случае с глобальной схемой, – она зависит от потребностей пользователей, и, соответственно, такая система может быть либо сильно связанной, либо слабо связанной. Архитектура системы предусматривает наличие *общей модели данных* и *внутреннего командного языка*.

Подход *языков мультибаз данных* (МБД) предназначен для систем, которые не используют предопределенных схем интеграции. Вместо глобальной схемы на пространстве БД определяется *общее пространство имен*. Язык МБД предназначен для предоставления языковых конструкций, позволяющих выполнять запросы, которые охватывают одновременно несколько БД. Такой язык имеет возможности, которых нет в обычном языке. С его помощью пользователи определяют источники данных, способ интеграции, передачи и представления данных (пример – MRDSM с языком MSQL) [11].

Недостатками подхода с глобальной схемой являются необходимость предопределения схемы данных, статическая структура, отсутствие унифицированного представления источников данных. Недостатками подхода интегрированных баз данных являются большая автономность локальных систем, приводящая к децентрализации управления, более сложная архитектура систем управления. Недостатком подхода языков мультибаз данных является то, что задача интеграции фактически перекладывается на плечи пользователей системы, которые обеспечиваются соответствующими инструментами. Большинство систем интеграции ориентированы на *выборку информации* из разнородных источников, они не предоставляют средств манипулирования данными.

Для реализации системы управления данными, обеспечивающей интеграцию гетерогенных источников данных, был выбран подход интегрированных баз данных, поскольку он:

обеспечивает высокую гибкость системы, что позволяет динамически менять ее конфигурацию;

сохраняет достаточную степень независимости источников данных;

позволяет сформировать единое целостное представление пространства данных, абстрагируя от физических и концептуальных характеристик гетерогенной среды.

Система интеграции гетерогенных источников данных

Система управления данными должна поддерживать такие возможности, как:

многоуровневая интеграция разнородных источников данных в единое информационное пространство (т.е. интеграция не только на уровне самих источников данных, но и на уровне их элементов);

выполнение операций определения, выборки данных и манипулирования данными;

поддержка ограничений целостности на любом уровне интеграции источников данных;

поддержка произвольных связей между источниками данных на любом уровне интеграции;

представление физической схемы данных в виде логической схемы предметной области.

Далее приводится формальное описание системы управления данным, модели интеграции данных, рассматриваются свойства модели, а также архитектура системы.

Модель управления данными системы интеграции

Виртуальная система управления данными V – это тройка $< S,Q,P >$, где S – схема представления источников данных, Q – множество схем сущностей для работы с данными, P – множество отображений схемы представления источников данных на схемы сущностей:

$$F = \{p : S \to M \mid M \in Q\}.$$

Система интегрирует разнородные источники данных в единую систему, с общей схемой для реализации единого механизма управления данными, независимого от моделей данных отдельных источников.

В основе *модели данных – рекурсивное представление источников данных*. Данный способ представления позволяет моделировать произвольно сложные источники данных, сводя их к все более простым элементам. Каждый *источник данных* (на самом верхнем уровне вложенности существует один источник – сама схема) представляется в виде набора пар <*атрибут*, *источник данных*>, который разделяет данный источник на составные источники. Данные текущего уровня помещаются в атрибуты пар. Самый нижний уровень вложенности соответствует *неделимому* (*атомарному*) элементу данных. Промежуточные уровни источников данных представляют собой *метаданные*. Очевидно, что при таком подходе к моделированию обработка данных и метаданных осуществляется единообразно. Каждому источнику данных ставится в соответствие *набор связей с другими источниками* (любого уровня), *набор ограничений целостности* и *набор допустимых операций*.

В рекурсивном определении *источники данных* (ИД) обозначим $s_{l,i}^k$:

$$S = s_{0,1}^0 \text{ – источник данных самого верхнего уровня;}$$

$$s_{l,i}^k = \begin{Bmatrix} < \{ < a_j, s_{i,q}^{k+1} > | j \in [1,J], J \in N \}, L_i^k, C_i^k(s_{l,i}^k), O_i^k >, \; k \in [0, K(i)) \\ < \varnothing, L_i^k, C_i^k(s_{l,i}^k), O_i^k >, \; k = K(i) \end{Bmatrix},$$

$$i, l, q \in [1, I], I \in N, K : [1, I] \to N \cup 0$$

Источник помечается тремя индексами: *l* – номер родительского источника данных; *i, q* – номера источников данных в рекурсивном определении; *k* – абсолютный уровень рекурсивной вложенности; *K*(*i*) – функция, возвращающая максимальный абсолютный уровень рекурсивной вложенности для данного ИД; a_j – атрибуты элементов данных (предназначены для хранения данных ИД).

Множество связей данного источника данных обозначим L_i^k, $L_i^k = \{ < \{ s_{l,i}^k \}, t | t \in T, n > \}$, где *T* – множество атрибутов связей, *n* – название связи; индекс *i* играет роль уникального ключа (у всех ИД в схеме данный индекс уникален).

Множество предикатов, задающих *ограничения целостности* на множестве источников данных, обозначается $C_i^k(S)$. Множество *S* может быть множеством источников данных следующего уровня вложенности, либо является вектором $\langle S_1, ..., S_n \rangle$, $n \in N$, где S_i – множество источников данных уровня вложенности *k*+1 для *i*-го источника данных, располагающегося на *k*-м уровне. Ограничения целостности для источника данных задаются в виде предикатной формулы, определенной на множестве источников данных следующего уровня вложенности либо на пространстве наборов источников данных, являющихся источниками данных следующего уровня вложенности для произвольно взятых источников данных схемы.

Источник данных $s_{l,i}^k$ находится в *согласованном состоянии*, если все предикаты из множества предикатов, задающих ограничения целостности для данного источника данных и всех источников данных, находящихся на более высоких уровнях вложенности, принимают истинное значение. Таким образом, $s_{l,i}^k$ – в согласованном состоянии, если $\forall s : s \in S_i^k \bigcup \{ s_{l,i}^k \}$ выполняется условие $\forall p : p \in C(s), p(s) = true$, где *s* – произвольный источник данных, *p(s)* – некоторый предикат, *C(s)* – множество предикатов, задающих ограничения целостности для источника данных *s*. Источник данных находится в согласованном состоянии, если в согласованном состоянии находятся все ИД на следующем уровне рекурсивной вложенности, и он сам удовлетворяет собственным ограничениям ценности.

Рассмотрим задание различных видов ограничений целостности для источника данных *s*. Пусть *x* – переменная, обозначающая источник данных следующего уровня вложенности. Тогда:

Ограничения на диапазон допустимых значений представляются как $P(x) = a < A(x) < b, a,b \in M$, где A(x) – атрибут источника данных x, M – множество, в котором принимают значения атрибуты ИД.

Ссылочные ограничения: $P(x,y) = \forall x \exists y (x \to y)$.

Агрегированные ограничения: $P(x) = max(A(x)) > a, a \in M$, где A(x) – атрибут ИД x, M – некоторое множество, в котором принимают значения атрибуты источника данных.

Множество операций для данного источника данных обозначим O_i^k:

$$O_i^k = \begin{Bmatrix} \{p : S_i^{k+1} \to S_i^{k+1'}\}, k < K(i) \\ \{p : s_i^k \to s_i^{k'}\}, k = K(i) \end{Bmatrix},$$

где S_i^{k+1} – множество источников данных уровня вложенности *k+1* для *i*-го источника данных, располагающегося на *k*-м уровне.

Все *операции источников данных* могут быть разделены на четыре типа: выборка; обновление; удаление; добавление. Операции добавления и удаления некоторого источника данных определяются на множестве источников данных следующего уровня вложенности. Каждая операция может включать в себя произвольное количество операций, выполняющихся как транзакция.

Пусть *x* – добавляемый источник данных. Операция *добавления* для источника данных $s_{l,i}^k$ есть

$$f : S_i^{k+1} \to S'^{k+1}_i, k < K(i), S'^{k+1}_i = S_i^{k+1} \bigcup \{x\}.$$

Пусть *x* – удаляемый источник данных. Операция *удаления* для источника данных $s_{l,i}^k$ есть

$$f : S_i^{k+1} \to S'^{k+1}_i, k < K(i), S'^{k+1}_i = S_i^{k+1} \setminus \{x\}.$$

Операция *обновления* для источника данных $s_{l,i}^k$ есть

$$f : s_{l,i}^k \to s_{l,i}^{k\prime}, S_i^{k+1} = S'^{k+1}_i.$$

В процессе обновления источника данных могут изменяться значения атрибутов, ограничения целостности, набор связей, набор операций. При этом изменение состава нижележащих источников данных выполняется посредством операций добавления и удаления.

Операции добавления и удаления источников данных, находящихся в согласованном состоянии, сохраняют согласованным состояние источника данных, для которого эти операции выполнялись.

В системе поддерживаются *связи* следующих типов: структурные и ассоциативные.

Структурные связи являются неявными и обусловлены рекурсивной структурой схемы данных. Декомпозиция источника данных на ряд более мелких обусловливает появление структурной связи между родительским и дочерними источниками (родительские отношения здесь рассматриваются с точки зрения вложенности источников). Структурные связи образуют *связи между разными уровнями вложенности в рамках одного фрагмента иерархии источников*, соединяют источники разных уровней рекурсивной вложенности.

Ассоциативные связи позволяют связать источники данных, которые не связаны непосредственными структурными связями. Ассоциативные связи соединяют источники одинаковых или разных уровней вложенности разных фрагментов иерархии. Если рассматривать структурные связи как вертикальные, то ассоциативные можно рассматривать как горизонтальные.

Каждая связь имеет атрибут, который характеризует действия, направленные на поддержание ссылочной

целостности во время выполнения операций удаления над связанными источниками данных. Атрибут может принимать одно из следующих значений:

c – каскадное удаление: при удалении источника данных s, на который ссылается источник данных t, источник данных t также будет удален;

r – запрет удаления: если при удалении источника данных s существует источник данных t, ссылающийся на источник данных s, то операция удаления не будет разрешена; чтобы ее осуществить, необходимо либо сначала удалить источник t, либо перенаправить связь от него на другой источник;

d – удаление связи: при удалении источника данных s удаляется связь, ссылающаяся на него от некоторого источника t; при этом источник данных t продолжает существовать независимо.

Схемой сущностей называется логическое представление предметной области в виде сущностей и связей между ними, позволяющее работать с ИД схемы *S* в виде, независимом от конкретного источника данных. Каждая схема ориентирована на работу с некоторым подмножеством ИД. Определим схему сущности как *M* = <*E, L*>. Здесь *E* – множество сущностей, *L* – множество связей между этими сущностями (L = <e1, e2, t>, e1 ∈ E, e2 ∈ E, t ∈ T, где *T* – множество типов связей сущностей; ei – сущность, представляемая парой ei = <A, O>, где *A* – множество атрибутов, а *O* – множество операций сущности). Ключевым для сущностной схемы является отображение *p,* которое устанавливает соответствие между подмножеством ИД и схемой сущности $p : S \to E$, где *S* – некоторое подмножество схемы ИД. Для одной схемы источников данных может существовать несколько схем сущностей для решения различных задач предметной области.

Декомпозиция источника данных – это процесс последовательного структурного разбиения ИД, в ходе которого формируется рекурсивная структура представления схемы данных; при этом процесс разбиения осуществляется до тех пор, пока данные не станут далее неделимыми. Декомпозиция некоторого ИД, представленного в модели данных, в которой существуют только структурные и ассоциативные связи, может быть не единственной. Способ декомпозиции некоторого источника данных, представленного в модели данных, в которой существуют только структурные и ассоциативные связи, не влияет на максимальную глубину рекурсивной вложенности и на общее количество ИД в схеме.

Восстановление схемы данных – процесс получения схемы некоторой модели данных из схемы, полученной в результате декомпозиции схемы источников данных.

Две схемы ИД являются *эквивалентными*, если по ним восстанавливаются идентичные схемы в некоторой модели данных. Различные способы декомпозиции приводят к эквивалентным схемам ИД.

Рассмотрим возможность декомпозиции схем реляционных, иерархических и сетевых моделей данных. Как было сказано выше, декомпозиция той или иной схемы может быть не единственной.

Утверждение 1. Любая реляционная схема может быть отображена в схему источников данных.

Утверждение 2. Любая иерархическая схема может быть отображена в схему источников данных.

Утверждение 3. Любая сетевая схема может быть отображена в схему источников данных.

Данные утверждения доказываются существованием алгоритмов декомпозиции, построенных авторами.

Будем считать эквивалентными схемы реляционной, иерархической и сетевой моделей, если совпадает количество типов записей и отношений, наборы их атрибутов, за исключением первичного и внешнего ключей, а внешние ключи реляционной схемы соответствуют групповым отношениям иерархической и сетевой модели.

Утверждение 4. Любой способ декомпозиции эквивалентных схем данных иерархической, реляционной и сетевой моделей приводит к эквивалентным схемам источников данных.

Архитектура системы

Архитектура системы интеграции гетерогенных источников данных является многослойной, в ней можно выделить три уровня (рис. 1):

- уровень данных (I);
- уровень унифицированного представления данных (II);
- уровень логического представления данных (III).

На нижнем уровне системы *располагаются источники данных* и *драйверы* доступа к ним. Источники данных могут представлять собой реляционные, сетевые, иерархические, объектно-ориентированные, объектно-реляционные базы данных, электронные таблицы, XML-файлы, потоки данных, неструктурированные источники. Для доступа к источникам данных могут использоваться драйверы ODBC, провайдеры OLE DB, ADO.NET, либо иные специфические драйверы, которые поддерживают интерфейс выполнения манипуляций над источником данных, которые от него требуются в данной системе.

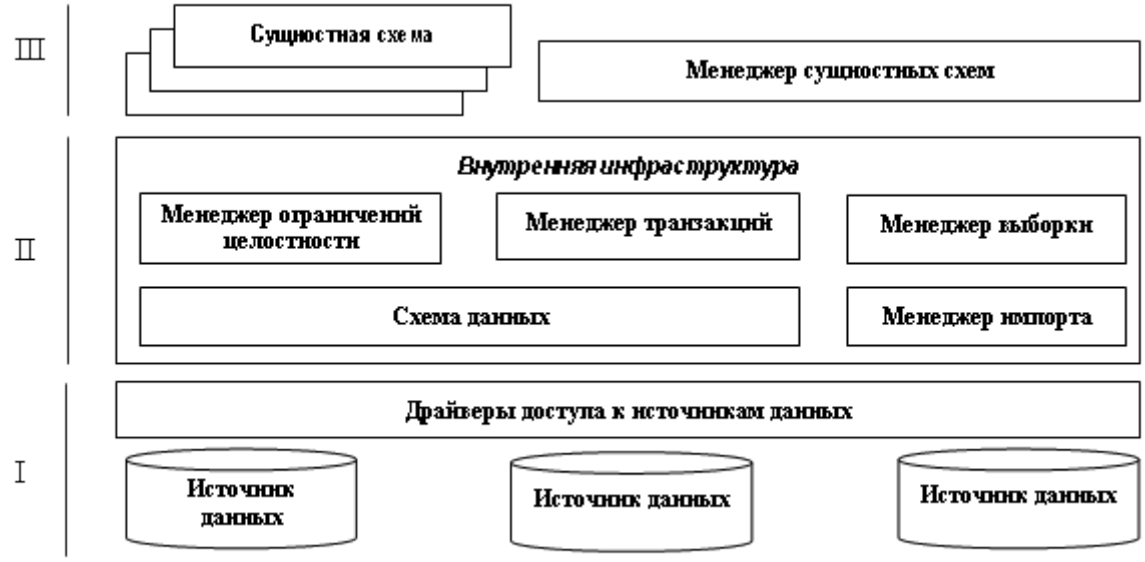


Рис. 1. Архитектура системы интеграции данных

На уровне *внутренней инфраструктуры* реализуется физическое представление схемы интеграции ИД, средства по поддержанию схемы в согласованном состоянии и выполнения базовых операций. Кроме этого, на данном уровне находится компонент, отвечающий за добавление в схему новых ИД.

Схема данных представляет собой внутреннюю рекурсивную структуру моделей источников данных в соответствии с описанной выше математической моделью.

Менеджер транзакций осуществляет выполнение глобальных транзакций в системе. Перед выполнением транзакции в объекте сохраняется информация о состоянии источников данных, которые будет охватывать транзакция, до ее начала. Это необходимо для осуществления процедуры отката в случае, если будет выполнено условие отката транзакции. На глобальном уровне транзакции выполняются последовательно и не могут пересекаться.

Менеджер ограничений целостности контролирует выполнение ограничений целостности во время выполнения операций глобальных транзакций, а также периодически с некоторым интервалом времени. Выполнение *глобальных транзакций* может состоять из произвольных операций манипулирования источниками данных, что может привести к рассогласованию данных в силу нарушения установленных ограничений целостности. Данный компонент осуществляет проверку после выполнения каждой базовой операции над источником данных на наличие нарушений ограничений целостности. Если нарушение было зафиксировано, то осуществляется откат операции и, по возможности, всей транзакции. Периодическая проверка выполнения ограничений целостности нужна на случай, если в источниках данных произошли локальные транзакции под управлением самих источников, что привело к нарушению ограничений целостности, установленных в данной схеме. В случае обнаружения такого нарушения пользователю

(администратору) сообщается об этом и работа с этим ИД становится невозможной до тех пор, как согласованность данных на глобальном уровне не будет восстановлена.

Менеджер выборки данных отвечает за соединение источников данных и представление данных в виде, отвечающем потребностям пользователя. На вход компонент получает запросы, сформулированные в терминах сущностной схемы, а на выходе выдает результат выборки данных из ИД.

Менеджер импорта осуществляет импорт в систему схем данных определенного типа. В основе работы компонента лежат алгоритмы декомпозиции источников данных. Расширение спектра поддерживаемых системой моделей данных требует добавления средств интерпретации данных моделей в этот компонент.

Внешнее представление – это логическое представление интегрированной схемы, отражающее сущности и взаимосвязи конкретных предметных областей.

Схема сущностей обеспечивает работу с системой в терминах предметной области. Компонент обеспечивает высокоуровневое представление источников данных, позволяющее решать задачи предметной области, а не тратить ресурсы на интерпретацию схемы интеграции. Построение подобной схемы сущностей, по сути, означает определение отображения схемы источников данных на предметно-ориентированную схему сущностей.

Менеджер схем сущностей позволяет создавать и редактировать новые схемы сущностей и устанавливать отображение между источниками данных и сущностями.

Заключение

В ходе проведенных исследований разработана модель представления источников данных, которая позволяет осуществлять многоуровневую интеграцию гетерогенных источников данных в единое информационное пространство, обеспечивает поддержку ограничений целостности на любом уровне интеграции источников данных; поддержку структурных и ассоциативных связей между источниками данных на любом уровне интеграции.

Проведенный анализ существующих систем интеграции данных позволяет оценить преимущества предложенного подхода: возможность интеграции данных на любом уровне декомпозиции; возможность динамического изменения схемы данных; расширяемость системы за счет новых моделей данных; универсальное представление связей, ограничений целостности и операций над данными.

Продолжение работы планируется осуществлять в следующих направлениях: разработка средств разрешения проблемы семантической гетерогенности данных, механизмов поддержки распределенных транзакций в гетерогенной среде, механизмов отображения физической схемы источников данных в предметно-ориентированные схемы, алгоритмов декомпозиции и интеграции других моделей данных (в дополнение к построенным для реляционной, иерархической и сетевой моделей) и др.

Благодарности

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

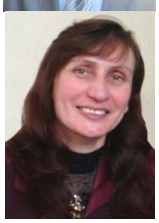
Библиографический список

1. Агравал Р. и др. Клермонтский отчет об исследованиях в области баз данных [Электронный источник]. [Режим доступа: свободный: http://citforum.ru/database/articles/claremont_report/].
2. Кашников А.В. Средства поддержки переносимости информационных систем, основанных на метаданных. Тезисы докладов XV Международной студенческой школы-семинара – М.: МИЭМ, 2007. С.412-414.
3. Кашников А.В. Средства поддержки переносимости информационных систем, основанных на использовании метаданных // Технологии Microsoft в теории практике программирования. Новосибирск, 2007. С. 59-61.

4. Кашников А.В., Лядова Л.Н. Реализация механизма поддержания ссылочной целостности в CASE-системе METAS // Актуальные проблемы математики, механики, информатики: Материалы международной научно-методической конференции, посвященной 90-летию высшего математического образования на Урале. Перм. гос. ун-т, Пермь, 2006. С.175-177.
5. Разнородные (гетерогенные) информационные системы и системы мультибаз данных [Электронный источник]. [Режим доступа: свободный: http://www.interin.ru/page.php?id=170&pg=&print=1].
6. Blunschi L., Dittrich J.-P., Girard O.R., Kirakos S.K., Salles M.A.V. A Dataspace Odyssey: The iMeMex Personal Dataspace Management System // Conference on Innovative Data Systems Research, 2007.
7. Breitbart Y., Garcia-Molina H., Silberschatz A. Overview of multidatabase transaction management // VLDBJ, 1(2):181, 1992.
8. Carey M.J., Haas L.M., Schwarz P.M., Arya M., Cody W.F., Fagin R., Flickner M., Luniewski A.W., Niblack W., Petkovic D., Thomas J., Williams J.H., Wimmers E.L. Towards Heterogeneous Multimedia Information Systems: The Garlic Approach // Proceedings of the Fifth International Workshop on Research Issues in Data Engineering (RIDE): Distributed Object Management, 1995.
9. Chawathe S., Garcia-Molina H., Hammer J., Ireland K., Papakonstantinou Y., Unman J., Widom J. The TSIMMIS project: Integration of heterogeneous information sources // Proceedings of the 100th Anniversary Meeting. Information Processing Society of Japan, Tokyo, Japan, October 1994. P. 7-18.
10. Dittrich J.-P., Salles M.A.V. iDM: A Unified and Versatile Data Model for Personal Dataspace Management // VLDB, 2006.
11. Elmagarmid A., Rusinkiewicz M., Sheth A. Management of Heterogeneous and Autonomous Database Systems // Morgan Kaufmann Publishers, San Francisco, California, USA, 1998.
12. Franklin M., Halevy A., Maier D. From Databases to Dataspaces: A New Abstraction for Information Management // SIGMOD Record, 34(4): 2005. P. 27-33.
13. Hull R. Managing semantic heterogeneity in databases: a theoretical prospective // Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, May 11-15, 1997, Tucson, Arizona, United States. P. 51-61.
14. Kent W. The Breakdown of the Information Model in Multi-Database Systems // SIGMOD, Record 20(4), 1991.
15. Lenzerini M. Data Integration: A Theoretical Perspective // PODS, 2002.
16. Litwin W. From database systems to multidatabase systems: Why and how // British National Conference on Databases, Cambridge Press, 1988.
17. Reddy M.P., Prasad B.E., Reddy P.G. A Methodology for Integration of Heterogeneous Databases // IEEE Transactions on Knowledge and Data Engineering, Vol. 6, No. 6, 1994. P. 920-933.
18. Sheth A.P., Rusinkiewicz M., Karabatis G. Using Polytransactions to Manage Interdependent Data // Database Transaction Models for Advanced Applications. 1992. P. 555-581,
19. Stonebraker M., Madden S., Abadi D., Harizopoulos S., Hachem N., Helland P. The End of an Architectural Era (It's Time for a Complete Rewrite). Proceedings of VLDB, Vienna, Austria, 2007.
20. Thomas G., et al. Heterogeneous distributed database systems for production use // ACM Computing Surveys, 22, 1990. P. 237-266.

Сведения об авторах

Алексей Кашников *– НИУ «Пермский государственный университет», ассистент кафедры математического обеспечения вычислительных систем; Россия, г. Пермь, 614990, ул. Букирева, д. 15; e-mail: Kashnikov@psu.ru.*
Major Fields of Scientific Research: управление данными, анализ данных, моделирование.



Людмила Лядова *– Пермский филиал Национального исследовательского университета «Высшая школа экономики», доцент кафедры информационных технологий в бизнесе; Россия, г. Пермь, 614070, ул. Студенческая, д. 38; e-mail: LyadovaLN@hse.perm.ru.*
Major Fields of Scientific Research: метамоделирование; технология DSM; CASE-средства; языковые инструментарии; предметно-ориентированные языки, DSL.