УДК 004.65

ПРИМЕНЕНИЕ ОРИЕНТИРОВАННЫХ ГРАФОВ ДЛЯ ПОСТРОЕНИЯ СХЕМЫ ГЕТЕРОГЕННОЙ РАСПРЕДЕЛЕННОЙ БАЗЫ ДАННЫХ УРОВНЯ ПРЕДПРИЯТИЯ

[1]Лаврентьев К.А., [2]Пономарчук Ю.В., [1]Фалеева Е.В.

[1]Хабаровская государственная академия экономики и права, Хабаровск, e-mail: klavrentev@mail.ru;
[2]Дальневосточный государственный университет путей сообщения, Хабаровск, e-mail: yulia.ponomarchuk@gmail.com

В работе сформулированы и проанализированы некоторые проблемы разработки и дальнейшей эксплуатации распределенных баз данных, влияющие на возможность интеграции новых узлов в распределенную систему. Описаны методы решения проблем, предлагаемые в современной литературе, приведены их достоинства и недостатки. В работе предлагается информационная модель мультибазы данных с глобальной структурой на основе ориентированного графа для решения проблемы интеграции новых узлов в распределенную систему, а также средства ее разработки и визуализации с целью своевременной автоматизированной модификации и определения мер улучшения производительности. Сравнительный анализ библиотек обработки информации и ее представления в виде графа показал, что наиболее перспективной является свободно распространяемая библиотека Cytoscape Web, поддерживающая основные форматы обмена данными.

Ключевые слова: мультибаза данных, согласованность, оптимизация, интеграция, гетерогенность

APPLICATION OF DIGRAPHS IN DESIGN OF A SCEME FOR HETEROGENEOUS DISTRIBUTED DATABASE AT ENTERPRISE LEVEL

[1]Lavrentev K.A., [2]Ponomarchuk Y.V., [1]Faleeva E.V.

[1]Khabarovsk State Academy of Economics and Law, Khabarovsk, e-mail: klavrentev@mail.ru;
[2]Far Eastern State Transport University, Khabarovsk, e-mail: yulia.ponomarchuk@gmail.com

The paper is devoted to the summary and analyze of some problems of distributed databases architecture development, which affect their ability to integrate new nodes in the information system. Common methods for solution of these problems, their advantages and disadvantages are described according to the contemporary published results. The paper presents an information model of a multidatabase with global structure, which is based on a directed graph, in order to solve the problem of new nodes' integration in the distributed system.

Keywords: multidatabase, consistency, optimization, integration, heterogeneity

В настоящее время сложно представить вид человеческой деятельности, который возможен без использования информационных технологий и автоматизации каких-либо операций. Особенно информационные технологии внедряются в область автоматизации различных бизнес-процессов в корпоративном секторе экономики, где для информационного обеспечения деятельности необходимо ежесекундно передавать десятки гигабайт информации по компьютерным сетям, объединяющим различные каналы связи, коммуникационные устройства и построенным на различных протоколах. Для поддержки интенсивного информационного обмена используются базы данных, распределенные по различным точкам земного шара. Бизнес, в том числе бизнес в сфере высоких технологий, наиболее заинтересован в быстром обмене информацией и низких издержках на её хранение и обработку.

При проектировании и реализации распределенной базы данных возникает множество проблем, связанных с согласованностью и контролем целостности. Целью данной работы является исследование некоторых из них, а также способов их решения. В статье предложена модель организации распределенной базы данных, основанная на ориентированном графе.

Проблемы проектирования распределенных гетерогенных баз данных

Распределенная система уровня предприятия состоит из множества локальных баз данных (БД) подразделений компании. Если все базы данных в ней работают под управлением одной системы управления базами данных (СУБД), то она называется гомогенной, но встречаются такие системы крайне редко. Связано это с тем, что в современных информационных системах чаще приходится строить распределённые БД на основе уже имеющихся унаследованных, т.е. «снизу вверх». При этом необходимо учитывать низкую степень интеграции хранящейся в них информации. Если распределенная система состоит из множества локальных баз под управлением различных

СУБД, то называется гетерогенной и является самым распространённым типом распределённой системы уровня предприятия.

При проектировании и администрировании гетерогенных распределенных систем возникают проблемы, связанные с контролем согласованности данных, проверкой ограничений целостности, выполнения запросов к различным узлам и интеграции новых узлов в систему. Профессор К. Дейт сформулировал 12 основных признаков, которым должна удовлетворять распределенная БД [1]. Среди них для данной работы особенно интересны следующие:

1. Локальная независимость: каждый узел должен полностью контролировать выполняемые им операции.

2. Отсутствие опоры на центральный узел: в распределённой системе не должно быть какого-либо центрального узла, который управлял бы, например, выполнением запросов, управлением транзакциями и т.д.

3. Независимость от физического расположения: пользователи не обязаны знать, где хранятся их данные. Информация может быть переписана с одного узла на другой без изменения работающих с ней приложений.

4. Независимость от типа СУБД: распределенная система должна поддерживать работу со всеми СУБД, которые есть в её составе.

Представленные ниже проблемы проектирования и реализации распределенной БД возникают именно в результате требований ее соответствия вышеупомянутым признакам.

Первая проблема, встречающаяся в гетерогенных распределенных системах – проблема контроля согласованности данных и осуществление проверки ограничений целостности. Распределенная система состоит из множества узлов, каждый из которых управляется различными СУБД. При выполнении операций вставки, обновления или удаления данных необходимо проверить ограничения целостности для изменяемой сущности БД и удостовериться, что данные после изменения сущности будут согласованы. Для иллюстрации приведем следующую ситуацию – пусть в распределённой системе имеются два узла: отдел приема товара и отдел гарантийного ремонта. Также имеются две сущности: документы «Заявка на ремонт» и «Паспорт ремонта», связанные между собой – на основании заявки на ремонт создается «Паспорт ремонта». Среди прочих атрибутов сущность «Заявка на ремонт» имеет атрибут «ФИО клиента» для идентификации владельца техники, находящейся в ремонте. Предположим, что в процессе работы на узле «Отдел приема товара» был удален определенный экземпляр сущности «Заявка на ремонт» и возникла ситуация, в которой на узле «Паспорт ремонта» невозможно определить, кому принадлежит техника, находящаяся в ремонте. В результате возникает рассогласование данных на различных узлах распределенной системы.

Для решения данной проблемы обычно рекомендуют следующие методы:

1. Постоянная репликация данных в распределенной системе.

2. Использование распределенных транзакций и протокола двухфазной фиксации.

Преимущества и недостатки каждого метода будут рассмотрены ниже.

Во-вторых, в распределенных системах может встретиться проблема, связанная с выполнением запросов к различным узлам. Она заключается в том, что при построении запроса, по которому требуется получить информацию с нескольких узлов, необходимо знать схему фрагментации данных распределенной системы и отдельные СУБД, используемые узлами. Вторым аспектом проблемы является необходимость оптимизации распределённых запросов, поскольку их выполнение включает в себя обращение к удаленным узлам и пересылку данных между ними. Минимизация как числа таких обращений, так и объема пересылаемых данных может многократно уменьшить время выполнения запроса, включающее этап глобальной оптимизации, на котором разработчик определяет, какие данные и с какого узла на какой будут пересылаться, а также этап локальной оптимизации на каждом узле.

Важность оптимизации распределенного запроса может быть обоснована следующим примером. Предположим, что на узле А хранится таблица Т1, содержащая 200 строк, на узле В – таблица Т2, содержащая 300 строк, и запрос требует соединения этих таблиц. Очевидно, выгоднее переслать таблицу Т1 на узел В и выполнить соединение на нем, а не наоборот.

Еще одной важной стороной данной проблемы является поддержка и выполнение распределенных транзакций, сложность реализации которых связана с тем, что каждый узел может иметь свою СУБД. Поэтому при построении транзакции узел, который ее создает, может не «знать» структуру узла-получателя, т.е. в этом случае разработчики вынуждены решать проблему контроля согласованности данных в распределенных системах.

Для решения проблемы выполнения запросов и поддержки распределенных транзакций применяют следующие методы:

1) создание мультибаз данных;

2) протокол двухфазной фиксации транзакций.

Прежде чем рассмотреть основные черты каждого метода, стоит описать третью проблему, которая состоит в интеграции новых узлов в систему. Изучение методов ее решения и является целью данной работы. Проблема состоит в том, что при добавлении нового узла необходимо проверить схему данных и сами данные на согласованность с уже имеющейся в системе информацией. Кроме того, интеграция узла требует участия технического специалиста (администратора БД) и, чем больше фрагментирована система, тем больше времени придется затратить на интеграцию нового узла. При этом временные затраты ведут за собой довольно ощутимые финансовые издержки для компании. Для решения проблемы интеграции новых узлов, а вернее, с целью сокращения временных затрат на интеграцию, распределенную систему создают по принципу мультибаз данных.

Существует несколько типов мультибаз данных.

● Мультибазы данных с глобальной схемой: система мультибаз является распределённой и служит внешним интерфейсом для доступа ко множеству локальных СУБД либо структурируется как глобальный уровень над локальными СУБД.

● Федеративные базы данных: в отличие от мультибаз не располагают глобальной схемой, к которой обращаются все приложения. Вместо этого поддерживается локальная схема импорта-экспорта данных. На каждом узле поддерживается частичная глобальная схема, описывающая информацию от тех удалённых источников, данные с которых необходимы для функционирования.

● Мультибазы с общим языком доступа – распределённые среды управления с технологией «клиент – сервер».

Таким образом, мультибаза данных – это распределенная система, в которой тем или иным образом хранится схема данных для работы распределенных запросов и транзакций. С одной стороны, их построение решает проблему интеграции новых узлов. Все, что нужно сделать при интеграции, – это обновить схему данных (глобальную или локальные, в зависимости от используемого типа мультибазы). С другой стороны, специалисту (администратору БД) необходимо потратить много времени на обновление глобальной схемы данных. В дальнейшем необходимо следить за каждым узлом БД и при изменении на нем схемы данных оперативно менять глобальную схему.

Модель гетерогенной распределенной базы данных, построенная на основе ориентированного графа

Для решения сформулированных выше проблем предлагается использовать модель мультибазы данных с глобальной схемой, при этом глобальная схема представлена ориентированным графом. В предлагаемой модели организации распределенной системы ориентированный граф будет состоять из следующих элементов:

● акторы – начальные вершины ориентированного графа. С точки зрения информационной модели – сущности;

● узлы-реципиенты – конечные вершины ориентированного графа. С точки зрения информационной модели – узлы распределенной базы данных;

● ребра – дуги графа, соединяющие акторов с узлами-реципиентами. Весом ребра будем считать количество экземпляров сущности на узле распределенной системы.

Узлы-реципиенты связаны между собой направленными дугами. При этом направление определяет характер взаимодействия узлов (начальная вершина владеет сущностями, на основании которых формируют информацию сущности на конечной вершине), вес дуги показывает количество взаимодействующих сущностей.

Схема данных абстрактной распределенной системы, представленная ориентированным графом, показана на рис. 1. Согласно описанию представленной модели, на нем можно выделить акторов – сущности предметной области, на рисунке они представлены кругами без заливки. Также можно выделить узлы-реципиенты – узлы распределенной системы, на рисунке они представлены кругами с заливкой.

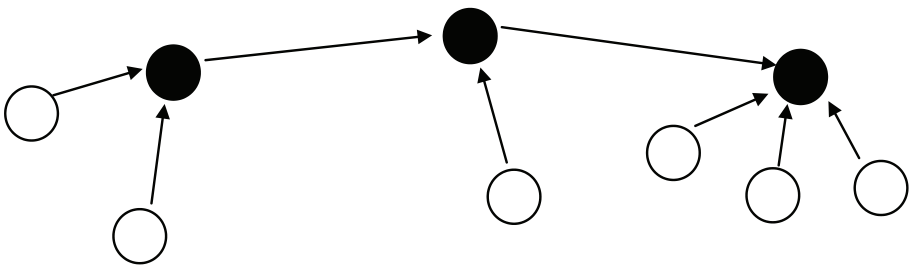

Рис. 1. Схема данных абстрактной распределенной системы

Предлагаемая модель организации схемы распределенной БД обладает рядом преимуществ по сравнению с обычными мультибазами данных с глобальной схемой. У последних есть один серьезный недостаток – схема данных для своего построения требует участия технического специалиста и при ее изменении на каком-либо узле необходимо вручную перестраивать глобальную схему. При построении распределенной СУБД с использованием предлагаемого в данной работе метода система управления сама в автоматизированном режиме будет дополнять глобальную схему данных и перестраивать её в случае изменений на каком-либо узле.

Кроме того, использование предлагаемой организации распределенной системы позволит решить проблему оптимизации распределенных запросов в части определения узла, на котором следует проводить объединение результатов. Такая возможность достигается за счет того, что дуги графа имеют вес, отражающий число экземпляров сущности на данном узле. Эта информация позволяет еще до начала выполнения запроса, при построении его плана сразу определить, где лучше проводить объединение результатов.

Еще одним немаловажным достоинством предлагаемой модели является возможность визуализации глобальной схемы данных, тогда как мультибазы с глобальной схемой не позволяют представить схему в графическом виде. Такая возможность является удобным средством анализа структуры распределенной системы потому, что человеку гораздо проще воспринимать графическую информацию. При использовании визуализации администратор БД может сократить время анализа состояния системы, провести первичный анализ состояния распределенной системы и выявить недостатки или возможности оптимизации.

В предлагаемой модели можно выделить также ряд недостатков: сложность построения распределенных СУБД, работающих с графами для глобальной схемы данных, а также необходимость построения еще одного программного уровня в архитектуре распределенной системы для реализации графовой модели организации.

Анализ практического применения

Представив описание предлагаемого метода, рассмотрим пример разработки модели распределенной системы на основе орграфа. Модель будет построена для распределенной БД вуза, который, допустим, состоит из одного основного учебного подразделения и нескольких филиалов.

На первом этапе построения графовой модели распределенной системы необходимо описать существующую распределённую БД. Для этого сначала выделим узлы распределённой системы и затем распределим сущности БД по узлам.

На основании описания организационной структуры вуза можно выделить узлы распределенной системы: основное учебное подразделение и филиал.

Далее необходимо распределить сущности информационной системы по узлам. Для этого сведем сущности в таблицу. Результат анализа представлен в табл. 1, из данных которой видно, что некоторые сущности повторяются в узлах, следовательно, имеет место фрагментация данных. При добавлении в систему нового филиала администратору необходимо знать, какие сущности фрагментируются. Для того, чтобы ответить на поставленный вопрос, у администратора должна быть полная схема БД, причем актуальная в любой момент времени, или же модель в виде графа, спроектированная по предложенному авторами методу.

Таблица 1

Сущности, распределенные по узлам

Узел	Сущность
Основное учебное подразделение	Студенты
	Преподаватели
	Экзаменационные ведомости
	Дисциплины
	Студенческие группы
	Данные о сотрудниках
	Бухгалтерские отчеты
Филиал	Студенты
	Преподаватели
	Экзаменационные ведомости
	Дисциплины
	Студенческие группы

Для визуализации схемы БД в виде графа авторы предлагают использовать JavaScript-библиотеку (JS-библиотеку). Выбор JavaScript для визуализации графа связан с тем, что этот язык сценариев выполняется в браузере пользователя, а значит, на любой платформе, которая обладает виртуальной java-машиной. Существует множество JavaScript-библиотек для визуализации графов. В результате выполнения сравнительного анализа выявлены 3 распространенные: D3.js, Cytoscape Web и Arbor.js.

Следующим этапом построения модели распределенной БД на основе графа является сравнительный анализ JSt-библиотек

и выбор наиболее подходящей для решаемой задачи, с целью проведения которого авторами были взяты следующие критерии:

1. Поддержка ориентированных графов.

2. Визуализация мощности связей – возможность библиотеки каким-либо образом выделять мощность связи (чаще используется изменение толщины связи).

3. Поддержка распространенных форматов обмена данными, среди которых выбраны JSON и XML. Необходимость их поддержки связана с тем, что информация для построения графа будет приходить с сервера БД.

Для проведения сравнительного анализа сведем выделенные библиотеки и критерии в таблицу. Для оценки пригодности библиотеки по каждому из критериев будем использовать количественную шкалу. Если библиотека поддерживает реализацию критерия, то ставим 1 балл, если нет – 0 баллов. В результате проведен подсчет баллов и выявлена наиболее подходящая библиотека. Результат сравнительного анализа представлен в табл. 2.

Таким образом, в результате сравнительного анализа можно заключить, что наиболее подходящей для визуализации модели распределенной БД на основе графа является js-библиотека Cytoscape Web, разработанная с использованием стандарта HTML5, что позволяет ей обеспечить кроссплатформенность. Она предоставляет разработчику API-функции для управления расположением вершин в графе, задания направления дуг, установки весов и т.д., обеспечивает работу со следующими форматами обмена данными: JSON, GraphML, XGMML. Поддержка специального формата для графовых структур (GraphML) позволяет экспортировать данные в профессиональные пакеты анализа графов.

Заключительным этапом построения графовой модели является построение графа БД (с использованием данных, приведенных в табл. 1) и его визуализация с применением библиотеки Cytoscape Web. При этом для построения и визуализации графа необходимо создать матрицу смежности вершин и массив весов связей.

Матрица смежности вершин графа G с конечным числом вершин n (пронумерованных числами от 1 до n) – это квадратная матрица A размера n, в которой значение элемента a_{ij} равно числу рёбер из i-й вершины графа в j-ю. В матрице смежности каждый элемент может быть равен либо 0, либо 1. Массив весов связей – массив n-элементов, который задает вес каждого ребра в графе. Согласно описанию представленной модели, вес связи – это количество экземпляров сущности в узле распределенной системы.

Результат визуализации модели схемы распределенной БД на основе графа представлен на рис. 2.

Таблица 2

Результаты сравнительного анализа JS-библиотек

Критерий / Библиотека	D3.js	Cytoscape Web	Arbor.js
Поддержка ориентированных графов	0	1	1
Визуализация мощности связей	1	1	0
Поддержка распространенных форматов обмена данными	0	1	1
Итого	1	3	2

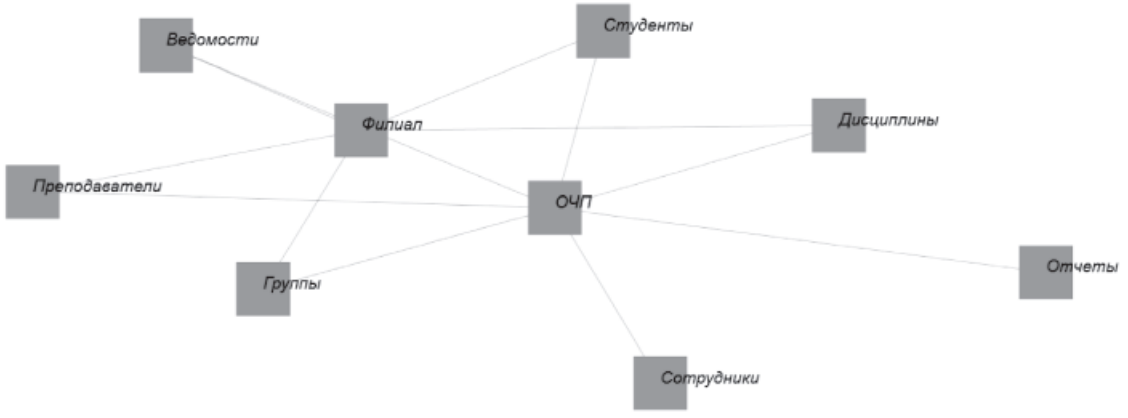

Рис. 2. Визуализация схемы распределённой базы данных

На рис. 2 можно выделить следующие объекты:

● акторы – сущности предметной области, перечисленные в табл. 1, – объекты, которые хранят в себе информацию в распределенной БД. На рисунке представлены узлами: сотрудники, дисциплины, ведомости, отчеты, преподаватели, студенты, группы;

● реципиенты – узлы распределенной БД. Представлены узлами: ОЧП, филиал;

● связи между узлами – дуги графа. Согласно описанию модели отражают связь между сущностями и узлами-реципиентами.

Заключение

Проблемы контроля согласованности и целостности данных в распределенных системах, а также проблема оптимизации распределенных запросов имеют важное значение для улучшения производительности работы распределенной системы и сокращения временных затрат технического специалиста на ее обслуживание. В данной работе описаны проблемы, которые существуют в гетерогенных распределённых системах и предложена модель организации глобальной схемы мультибазы данных в виде ориентированного графа. Также представлены достоинства и недостатки предложенного метода.

В дальнейшем планируется разработка программного обеспечения для работы с предлагаемой моделью и его тестирование на различных распределенных системах.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 13-07-00615.

Список литературы

1. Дейт К.Дж. Введение в системы баз данных: пер. с англ. – 8-е изд. – М.: Издательский дом «Вильямс», 2003. – 1072 с.: ил.

2. Мокрозуб В.Г. Графовые структуры и реляционные базы данных в автоматизированных интеллектуальных информационных системах. – М.: ООО Издательский дом «Спектр», 2011. – 108 с.

3. Седжвик Р. Фундаментальные алгоритмы на C++. Алгоритмы на графах: пер. с англ. – СПб.: ООО «ДиаСофтЮП», 2002. – 496 с.

4. Лаврентьев К., Рябухин С. Применение ориентированных графов на примере исследования инновационной активности правительства Хабаровского края. Хабаровск // Студенческая весна – 2010: материалы XL научной конференции – конкурса научных докладов. – 2010. – С. 46–52.

5. Лаврентьев К.А., Смагин С.И., Пономарчук Ю.В. Проблема интеграции новых баз данных в распределенную информационную систему предприятия // Высокие технологии, исследования, финансы, промышленность: сборник статей Восемнадцатой международной научно-практической конференции «Фундаментальные и прикладные исследования, разработка и применение высоких технологий в промышленности и экономике». 4–5 декабря 2014 года, Санкт-Петербург, Россия / науч. ред. А.П. Кудинов, И.А. Кудинов. – СПб.: Изд-во Политехн. ун-та, 2014. – 200 с.

References

1. Dejt K.Dzh. Vvedenie v sistemy baz dannyh: per. s angl. 8-e izd. M.: Izdatelskij dom «Viljams», 2003. 1072 p.: il.

2. Mokrozub V.G. Grafovye struktury i reljacionnye bazy dannyh v avtomatizirovannyh intellektualnyh informacionnyh sistemah. M.: OOO Izdatel-skij dom «Spektr», 2011. 108 p.

3. Sedzhvik R. Fundamentalnye algoritmy na C++. Algoritmy na grafah: per. s angl. SPb.: OOO «DiaSoftJuP», 2002. 496 p.

4. Lavrentev K., Rjabuhin S. Primenenie orientirovannyh grafov na primere issledovanija innovacionnoj aktivnosti pravitelstva Habarovskogo kraja. Habarovsk // Studencheskaja vesna 2010: materialy XL nauchnoj konferencii konkursa nauchnyh dokladov. 2010. pp. 46–52.

5. Lavrentev K.A., Smagin S.I., Ponomarchuk Ju.V. Problema integracii novyh baz dannyh v raspredelennuju informacionnuju sistemu predprijatija // Vysokie tehnologii, issledovanija, finansy, promyshlennost: sbornik statej Vosemnadcatoj mezhdunarodnoj nauchno-prakticheskoj konferencii «Fundamentalnye i prikladnye issledovanija, razra-botka i primenenie vysokih tehnologij v promyshlennosti i jekonomike». 4–5 dekabrja 2014 goda, Sankt-Peterburg, Rossija / nauch. red. A.P. Kudinov, I.A. Kudinov. SPb.: Izd-vo Politehn. un-ta, 2014. 200 p.

Рецензенты:

Графский О.А., д.т.н., доцент, профессор кафедры «Вычислительная техника и компьютерная графика», Дальневосточный государственный университет путей сообщения, г. Хабаровск;

Ахтямов М.Х., д.б.н., профессор, директор Естественно научного института, Дальневосточный государственный университет путей сообщения, г. Хабаровск.