МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и информационных технологий

МЕТОДЫ И ПОДХОДЫ ОРГАНИЗАЦИИ РАСПРЕДЕЛЕННОГО ХРАНЕНИЯ ДАННЫХ В ПРОГРАММНОМ КОМПЛЕКСЕ «НЕКА»

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 271 группы

направления 09.04.01 – Информатика и вычислительная техника

факультета компьютерных наук и информационных технологий

Дмитриевой Кристины Андреевны

Научный руководитель
доцент кафедры ДМиИТ, к.ф.-м.н. _____ И.Д.Сагаева

Заведующий кафедрой
доцент кафедры ДМиИТ, к.ф.-м.н. _____ Л.Б. Тяпаев

Саратов 2019

ВВЕДЕНИЕ

Одной из востребованных отраслей информационного мира является управление и хранения данных. Нетрудно заметить, что любая современная система не обходится без большого объема информации. Полученную тем или иным образом информацию необходимо где-то располагать для дальнейшей обработки. Для данной цели используются базы данных, которые представляют собой совокупность собранной информации, организованную в соответствии с выбранной схемой данных.

Когда объем информации перестает умещаться на локальной машине, то переходят к рассмотрению способов распределенного хранения данных. Обычно в сферу рассмотрения попадают распределенные базы данных, параллельные базы данных и NoSQL базы данных с использованием различных механизмов, позволяющие распределено управлять информацией. Так как для исследований важно, чтобы рабочая станция не только содержала полученные данные, но и могла быть использована для других целей, а также не была дорогостоящей, необходимо подойти к организации системы, учитывая все основные нюансы. Для дальнейшего исследования остановимся подробнее на способе хранения информации при помощи распределённой базы данных.

Распределенные базы данных – это совокупность множества взаимосвязанных баз данных, распределенных в информационной сети. Она состоит из узлов приема запросов и узлов данных. На разных узлах сети появляется возможность хранить разные осмысленные куски всего массива информации, а также использовать узел системы и для других необходимых целей, например, расчетных.

Одним из составляющих распределенной базы данных являются узлы системы, на которых происходит сбор информации для дальнейшей ее обработки. В качестве узлов системы обычно используют реляционные системы управления базами данных (СУБД), от правильного выбора которых напрямую зависит производительность всего комплекса в целом. Наиболее

востребованными свободно распространяемыми серверами баз данных являются следующие СУБД:

MySQL представляет собой систему клиент-сервер, которая содержит многопоточный SQL-сервер, обеспечивает поддержку различных вычислительных машин БД.

PostgreSQL – свободно распространяемая объектно-реляционная СУБД. Данная СУБД поддерживает большую часть стандарта SQL и предлагает множество современных функций, а также возможность всячески расширять используемый интерфейс.

Другой составляющей распределенной базы данных является словарь. Словарь представляет собой отдельную небольшую базу данных, которая хранит всю служебную информацию о системе. Данные в словаре имеют слабоструктурированную форму, что позволяет использовать в качестве модели данных не только в классическую реляционную модель, но и в NoSQL. Для этих целей отлично подходят документно-ориенторованные системы, наиболее популярной из них является MongoDB.

MongoDB – документно-ориентированная нереляционная СУБД с открытым исходным кодом, которая обеспечивает высокую производительность, доступность и автоматическое масштабирование. В MongoDB документы представляются JSON подобными объектами, с которыми в свою очередь удобно и быстро работать.

Таким образом, стоит серьезно подойти к выбору СУБД, как для узлов системы, так и для словаря. Для этого необходимо представлять возможности и ограничения каждой из выбранных серверов.

Помимо этого, на производительность распределенной базы данных будет влиять организация выполнения запроса, которую необходимо производить согласно современным тенденциям. В данном случае акцент делается на организацию выполнения запроса распределено, так как данные могут находиться на совершенно разных узлах.

А также не стоит забывать про контроль целостности данных. Ограничение целостности в базах данных – это логическое выражение, результат вычисления которого всегда должен быть истиной. Иначе данные не будут иметь никакого значения, если при работе системы разрушается их целостность.

Целью магистерской работы является анализ существующих подходов и разработка методов организации распределенного хранения данных.

Для достижения поставленной цели необходимо решить следующие задачи:

- литературный обзор технологий хранения большого объема информации;
- разработать структуру словаря распределенной базы данных;
- провести сравнительный анализ быстродействия системы с использованием реляционной и нереляционной модели данных для формирования словаря
- рассмотреть особенности обработки запросов в базах данных;
- разработать универсальный алгоритм работы модуля обработки запросов с использованием технологии параллельных вычислений;
- провести сравнительный анализ различных подходов получения результирующей распределенной выборки;
- рассмотреть основные особенности сохранения целостности данных в базах данных;
- разработать алгоритм сохранения ссылочной целостности данных в распределенной базе данных.

Основное содержание работы

1 Особенности работы с большими объемами данных

Одной из важнейших областей применения компьютеров является переработка и хранение больших объемов информации в различных сферах деятельности человека. Основой для информационной системы является база данных.

1.1 Модели данных БД

1.1.1 Структурированные модели данных

К БД структурированного типа относят иерархическую, сетевую и реляционную модели.

Первые две модели в настоящие время не используются из-за своих ограничений [1, 2]. Реляционные же БД напротив получили большое распространение в информационном мире.

1.1.2 Неструктурированные модели данных

Нереляционный способ структуризации данных заключается в избавлении от ограничений при хранении и использовании информации и имеет основные особенности, которые заключаются в исключение излишнего усложнения, высокой пропускной способности и неограниченного горизонтального масштабирования [3].

Нереляционные БД принято разделять на четыре вида: «ключ-значение», документно-ориентированные, колоночные и графовые модели [4]. Каждая из моделей служит для определенного класса задач.

1.2 Основные технологии хранения большого объема информации

При увеличении объемов информации, локальной БД становится недостаточно, и тогда используют распределенные и параллельные БД, которые строятся в большинстве случаев на основе реляционной БД. А также нереляционные БД с применением различных дополнений для распределенного хранения данных.

1.3 Основные этапы моделирования БД

При разработке БД выделяют три основных уровня моделирования, при помощи которых происходит переход от предметной области к конкретной реализации БД: концептуальная модель, логическая модель БД, физическая модель данных [5].

1.4 Обработка запросов в БД

1.4.1 Особенности выполнения распределенного запроса

Для того чтобы получить информацию из любой БД необходимо сформировать запрос на выборку и исполнить его. Для локальной БД весь процесс обработки запроса состоит обычно из двух шагов: декомпозиции запроса и его оптимизации [6, 7].

Обработка распределенных запросов – задача, более сложная и требует дополнительных манипуляций при получение результирующей выборки. Между шагами декомпозиции и оптимизации запроса включаются еще два действия: локализация данных и глобальная оптимизация запроса [8, 9].

1.5 Целостность данных в БД

Для сохранения достоверности данных, их структуры используют ограничение целостности. Ограничение целостности в БД – это логическое выражение, результат вычисления которого всегда должен быть истиной [6].

Ограничение целостности обычно разделяют на четыре вида: сущностная целостность, доменная целостность, ссылочная целостность, пользовательская целостность [10]. В РБД, помимо всего вышеизложенного, необходимо поддерживать целостность на разных узлах системы. Такую согласованность называют глобальной ссылочной целостностью [11], которая сама по себе является одной из главных проблем РБД.

2 Программный комплекс «НЕКА»

Программный комплекс «НЕКА» [12] представляет собой клиент – серверное приложение и создавался для исследования и построения РБД. Разработанный программный комплекс можно представить в виде структурной схемы, представленной на рисунке 2.1, которая включает в себя следующие

компоненты: словарь РБД, серверная часть программы, клиентская часть программы.

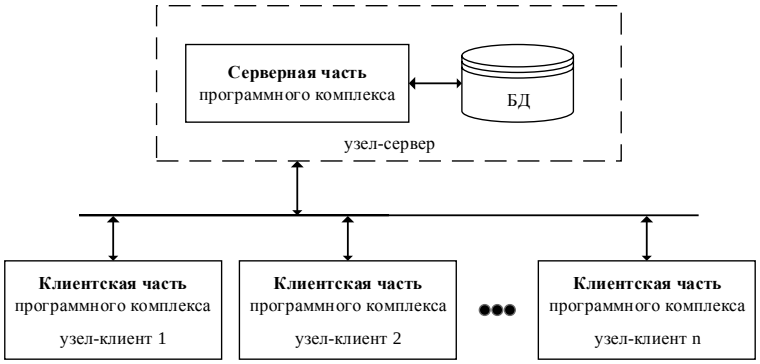

Рисунок 2.1 – Схема программного комплекса «НЕКА»

3 Узлы РБД в программном комплексе «НЕКА»

В работах [13, 14] было проведено исследование СУБД для дальнейшего их применения в качестве узлов РБД. Авторами работы [14] был разработан комплекс тестов, оценивающий производительность, а также был проведен сравнительный анализ использования выбранных СУБД в качестве узла РБД.

4 Словарь РБД в программном комплексе «НЕКА»

4.1 Проектирование словаря РБД

В реляционной модели логическая структура представляет собой набор таблиц. Каждая, из которой является либо самим объектом, либо взаимосвязью между ними [6]. Логическая схема реляционного словаря представлена на рисунке 4.1.1.

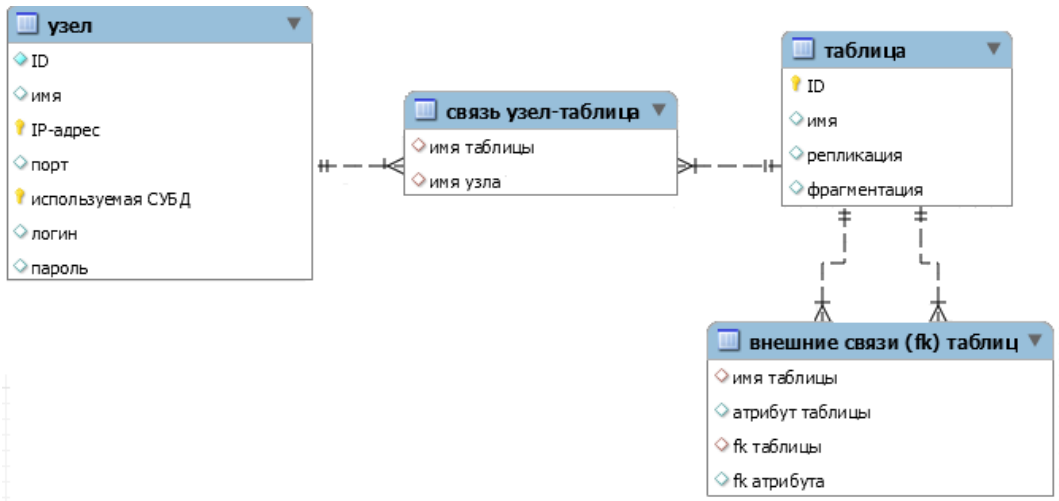

Рисунок 4.1.1 – Структура реляционного словаря

Документно-ориентированные СУБД оперируют абстрактным понятием документ. Они подразумевают инкапсуляцию и кодирование сохраняемой информации в некотором стандартном формате. Логическая схема нереляционного словаря представлена на рисунке 4.1.2.

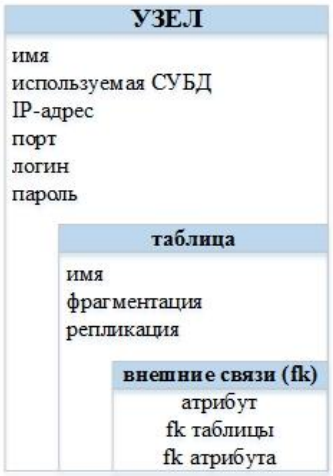

Рисунок 4.1.2 – Структура нереляционного словаря

В рамках исследования будем рассматривать в качестве реализации реляционного словаря СУБД MySQL 5.7 и PostgreSQL 9.3; в качестве нереляционного – MongoDB 3.6.

4.2 Определение исследуемых операций

В процессе работы комплекс «НЕКА» многократно производит операции над словарем. Наиболее используемыми операциями над словарем является чтение, запись и удаление служебной информации. Таким образом, выделим четыре группы операций для проведения исследования: установление соединение со словарем, операции записи служебной информации в словарь, операции удаления служебной информации из словаря, операции чтения служебной информации из словаря.

4.3 Сравнительный анализ быстродействия реляционного и нереляционного словарей

Важнейшей частью научных исследований является построение математических моделей и численного эксперимента, результаты которого требуют дальнейшей обработки. Как правило, для решения этой задачи используют статистические методы планирования эксперимента, повышающие

эффективность исследования, основанного на экспериментальном подходе, а также выявлении свойств исследуемых объектов и проверке справедливости гипотез.

Для проведения сравнительного анализа, согласно выбранным выше критериям, был проведен планируемый эксперимент и построены регрессионные модели. В рамках проводимого эксперимента будем рассматривать в качестве реализации реляционного словаря СУБД MySQL 5.7 и PostgreSQL 9.3; в качестве нереляционного – MongoDB 3.6.

По полученным регрессионным моделям времени выполнения операций словарем, можно сказать, что у каждой СУБД есть проблемные зоны. Стоит отметить, что из исследуемых СУБД не рекомендуется использовать в качестве словаря РБД СУБД PostgreSQL, так как чтение со словаря происходит довольно часто, а как было отмечено выше, у данной СУБД этот параметр являемся слабой зоной.

СУБД MongoDB справляется быстрее со всеми исследуемыми операциями, кроме удаления. А поскольку данная операция используется намного реже, чем остальные, то есть основания полагать, что данную СУБД целесообразнее использовать в качестве словаря РБД.

5 Обработка запросов РБД

5.1 Алгоритм обработки запросов пользователей к РБД

Предлагаемый алгоритм обработки запросов пользователей к РБД заключается в обработки запроса пользователя посредством выделения вида операции, фильтрации слов запроса, параллельной обработки данных словаря и параллельной отправки узлам РБД подзапросов с последующим получением объединенного результата.

5.2 Соединение результирующей распределенной выборки средствами СУБД

Одним из вариантов реализации функционала для РБД является применение средств самой СУБД только в другом использование. Выполнение запроса с использованием данного подхода условно можно разделить на

несколько этапов: формирование подзапросов из исходного запроса для частичных результатов на разных узлах; выделение главное узла по каким-то критериям; выполнение подзапросов и получение частичных результатов; копирование схем необходимых таблиц и перенос их на главный узел; перенос частичных результатов на главный узел; выполнение исходного запроса на главном узле; удаление вспомогательных данных с главного узла.

5.3 Соединение результирующей распределенной выборки с помощью программной реализации операции JOIN

Другой подход реализации функционала для РБД является программная реализация некоторых компонентов СУБД. Выполнение запроса с использованием данного подхода условно можно разделить на три этапа: формирование подзапросов из исходного запроса для частичных результатов на разных узлах; выполнение подзапросов на разных узлах и получение результатов; выполнение операции соединения таблиц JOIN.

5.4 Сравнительный анализ различных подходов получения результирующей распределенной выборки

Рассмотрев различные подходы к получению распределенной выборки необходимо провести сравнение их быстродействия. В результате проведенного сравнительного анализа было выяснено, что даже самая тривиальная программная реализация операции JOIN для соединения распределенных на разных узлах таблицах работает гораздо быстрее, чем та же операция, выполненная с помощью средств СУБД.

6 Поддержание целостности данных в программном комплексе «НЕКА»

6.1 Алгоритм сохранения ссылочной целостности данных в РБД

Ссылочная целостность данных является одной из главных проблем РБД [6]. Алгоритм учитывает особенности данных операций и производит контроль над целостностью данных, не позволяя изменить данные заведомо не допустимым образом.

ЗАКЛЮЧЕНИЕ

Результатом магистерской работы является колоссальное исследование в области организации хранения информации. В рамках работы был проведен литературный обзор технологий хранения большого объема информации. Также разобраны основные теоретические аспекты исследуемого вопроса, а именно организация узлов и словаря РБД. Помимо теоретической базы результатами работы является разработанные подходы и алгоритмы, которые позволяют увеличить производительность распределенной системы хранения данных, а также упрощают ее проектирование.

Одним из составляющих распределенной системы является словарь, в котором располагается вся служебная информация о ней. Данные в словаре имеют слабоструктурированную форму, что позволяет использовать в качестве модели данных не только классическую реляционную модель, но и NoSQL. В работе были разработаны две модели словаря: реляционная и нереляционная (документно-ориентированная) модели. Для аргументированного выбора наилучшего подхода формирования словаря в работе был проведен сравнительный анализ, в результате которого было выявлено, что использование документно-ориенторованной БД в качестве словаря увеличивает производительность системы, нежели традиционные реляционные БД для данной цели.

Другим немаловажным составляющим распределенной системы управления информацией является выполнение запросов. В результате проведенного литературного обзора исследуемого вопроса, было выяснено необходимость разработки обобщенного алгоритма обработки распределенного с упором на параллельны технологии. Данный алгоритм позволяет обрабатывать запрос от пользователя посредством выделения вида операции, фильтрации слов запроса, параллельной обработки данных словаря и параллельной отправки узлам РБД подзапросов с последующим получением объединенного результата. В данной работе было разобрано два подхода: соединение результирующей распределенной выборки средствами СУБД и с

помощью программной реализации операции JOIN. Для установления более производительного подходы был проведен их сравнительный анализ, в результате которого было выявлено преимущество программной реализации операции JOIN перед другим методом. Таким образом, можно говорить о целесообразности использования подхода с программной реализацией операции JOIN.

При управлении информацией важно сохранить ее логический смысл и не утратить ее целостность. Поддержание целостности в распределенных системах является одной из главных проблем их разработки, так как необходиомо учитывать все связи между данными находящимися на разных узлах сети. Для решения данной проблемы в работе был разработан алгоритм сохранения ссылочной целостности данных в РБД, что значительно упрощает реализацию данного функционала в системе.

Все разработанные алгоритмы и подходы были внедрены в программный комплекс «НЕКА», что значительно увеличило его производительность.

Таким образом, поставленные цель и задачи магистерской работы полностью выполнены.

По результатам исследования были опубликованы работы:

- Тимофеева Н.Е., Дмитриева К.А., Сагаева И.Д. «Анализ современных технологий хранения сверхбольших объемов информации» // Программные продукты, системы и алгоритмы. 2018. №1. С. 14 – 19. DOI: 10.15827/2311-6749.18.1.3 ISSN 2311-6749;
- Тимофеева Н.Е., Дмитриева К.А. «Разработка универсального программного комплекса для распределенного хранения данных с использованием параллельной обработки запросов» // Сборник тезисов молодежной научной конференции Девятая международная научно практическая школа «Высокопроизводительные вычисления на GRID-системах», 5-10 февраля 2018, ISSN 978-5-98450-571-0.

– Тимофеева Н.Е., Дмитриева К.А. Универсальный алгоритм обработки запросов с использованием технологии параллельных вычислений // Научно-технический вестник Брянского государственного университета. 2018 №2 С.211-217. ISSN 2413-9920. DOI: https://doi.org/10.22281/2413-9920-2018-04-02-211-217
– Тимофеева Н.Е., Дмитриева К.А. Сравнительный анализ реляционной и нереляционной модели хранения служебной информации централизованной распределенной базы данных // Вестник Российского нового университета Серия: «Сложные системы: модели, анализ, управление», 2019, Выпуск 1, С 66-74, DOI: 10.25586/RNU.V9187.19.01.P.066

Также были приняты в печать материалы конференций:

– Тимофеева, Дмитриева "Сравнительный анализ быстродействия системы с использованием реляционной и нереляционной модели данных для формирования словаря, Всероссийская научно-техническая конференция на тему «Информационно-управляющие, телекоммуникационные системы, средства поражения и их техническое оснащение», г. Пенза, 15 мая 2019 г.
– Тимофеева, Дмитриева: «Сравнительный анализ разных подходов получения выборки в распределенной базе данных", Всероссийская научно-техническая конференция на тему "Информационно-управляющие, телекоммуникационные системы, средства поражения и их техническое оснащение», г. Пенза, 15 мая 2019 г.
– Н.Е. Тимофеева, К.А. Дмитриева «Алгоритм сохранения ссылочной целостности данных в распределенной базе данных», XVIII Международной конференции имени А.Ф. Терпугова «Информационные технологии и математическое моделирование», г. Саратов 26-30 июня 2019г.

Результаты работы публично представлены:

- на конкурсе молодежных проектов «Технологический предприниматель – 2017» в НЧИ КФУ с проектом «Разработка программного комплекса для управления и ведения распределенной базы данных»;
- на IX Международной молодежной научно-практической школе «Высокопроизводительные вычисления на Grid системах» с докладом «Разработка универсального программного комплекса для распределенного хранения данных с использованием параллельной обработки запросов», Северный (Арктический) федеральный университет имени М. В. Ломоносова, г. Архангельск 5– 10 февраля 2018г.
- на Всероссийской научно-технической конференции на тему «Информационно-управляющие, телекоммуникационные системы, средства поражения и их техническое оснащение» с докладом «Сравнительный анализ быстродействия системы с использованием реляционной и нереляционной модели данных для формирования словаря», г. Пенза, 15 мая 2019 г.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1 Нестеров, С.А. Базы данных: учебник и практикум для академического бакалавриата / С. А. Нестеров. М.: Издательство Юрайт, 2016. ,230 с.

2 Types of Database Models | Database Management System [Электронный ресурс] // Database Management System [Электронный ресурс]: [сайт] URL: http://www.databasemanagementsystemsz.com/types-of-database-models/ (дата обращения 01.11.2018). Загл. с экрана. Яз. англ.

3 Клеменков, П. А. Большие данные: современные подходы к хранению и обработке / П. А. Клеменков, С. Д. Кузнецов // Труды ИСП РАН. 2012. С. 143-158

4 Гладкий, М. В. Модель распределенных вычислений MapReduce / М. В. Гладкий // Труды БГТУ. Серия 6: Физико-математические науки и информатика. 2016. №6 (188). С.194-198

5 Змитрович, А.И. Базы данных и знаний: учеб. пособие / А. И. Змитрович, В. В. Апанасович, В. В.Скакун. М.: Изд. центр БГУ, 2007

6 Дейт К. Дж. Введение в системы баз данных. / Дейт К. Дж. 8-е изд., пер. с англ. М.: Вильяме, 2005. 1328 с.

7 Т.С. Карпова. Базы данных: модели реализации СПб.: Питер,2001. 304с.: ил.

8 Date C. J. 1987. What is distributed database? InfoDB, 2:7

9 Г.М. Ладыженский. Технология "клиент-сервер" и мониторы транзакций. //Открытые информационные системы №3, 1994

10 Чуканов К.В., Чичикин Г.Я. Целостность баз данных [Электронный ресурс] // Наука, техника и образование. – 2018. – №11 (52). URL: https://cyberleninka.ru/article/n/tselostnost-baz-dannyh (дата обращения: 13.04.2019).

11 Moldovan, Grigor & Valeanu, Madalina Integrity constraints in distributed databases. – 2006.

12 Тимофеева Н.Е., Полулях К.А. Программный комплекс для управления распределенной базой данных // Программные продукты и системы. 2017. Т. 30. № 4. С. 663–667. DOI: 10.15827/0236-235X.030.4.663-667

13 Полулях К. А., Тимофеева Н. Е., Гераськин А. С. «Исследование баз данных MySQL и PostgreSQL на возможность применения в узлах распределенной вычислительной системы», Материалы Международной научной конференции Компьютерные науки и информационные технологии, 30 июня – 2 июля 2016г., г. Саратов, с 322 – 324, ISBN 978-5-9999-2651-7.

14 Тимофеева Н.Е., Гераськин А.С., Полулях К.А. «Исследование и построение моделей нагрузочного тестирования СУБД для повышения скорости и производительности распределенной вычислительной системы», Вестник Волгоградского государственного университета. Серия 1: Математика. Физика. №1 (38) 2017, с. 75-89, ISSN 2409-1782.