Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.658.2

Гладченко
Александр Валерьевич

Модели управления данными в корпоративных системах

АВТОРЕФЕРАТ
на соискание академической степени
магистра технических наук

по специальности 1-40 80 05 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Научный руководитель
Смолякова О.Г.
к.т.н., доцент

Минск 2014

КРАТКОЕ ВВЕДЕНИЕ

В современном мире огромные корпорации в числе Coca Cola, MTV, и т.п. нуждаются в непостижимых объёмах информации о клиентах, товарах, вспомогательной информации к сопровождающимся элементам, данных о пользователях, правах, действиях, записи действий пользователей, сохранения изменений и т.д. Как правило, никакие данные корпораций в хранилищах данных никогда не удаляются, поэтому последних становится крайне большое количество, и для управления ими нужны соответствующие хранилища и правильно выбранная модель управления данными.

Можно рассмотреть несколько альтернативных стратегий распределения данных, каждая из которых имеет как преимущества, так и недостатки. Основным преимуществом централизованной базы данных является простота. Все операции выполняются под контролем единственного узла. Все запросы на выборку и обновление данных направляются в центральный узел. Недостатком данной стратеги является то, что размер базы данных ограничивается объемом внешней памяти в центральном узле. Кроме того, центральный узел может стать узким местом всей системы с точки зрения надежности, поскольку база данных становится недоступной при появлении ошибки в системе связи и полностью выходит из строя при выходе из строя центрального узла. Это является недопустимым для экологически опасных химических производств.

Стратегия расчленения, при которой единственная копия базы данных в виде непересекающихся подмножеств распределяется по многим узлам сети, не допускает существования копий отдельных частей базы данных. При этом на первый план выдвигается процесс проектирования расчлененных данных с целью получения преимуществ за счет распределения запросов на выборку и обновления по тем узлам, где расположены запрашиваемые данные. В этом случае стоимость связи может быть снижена за счет того, что большая часть запросов к базе данных будет осуществляться в локальных узлах.

При распределении данных с использованием стратегии дублирования в каждом узле сети размещается полная копия базы данных. Основное преимущество этой стратегии относится к области надежности и эффективности выборки, что требует, однако, значительных затрат памяти. Кроме того, с целью согласования множественных копий базы данных необходимо поддерживать их изменения, что является трудно выполнимой задачей, отвлекающей значительные ресурсы системы.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цели и задачи исследования

Целью диссертационной работы является исследование рентабельности использования моделей управления данными в корпоративных системах и создание модифицированной модели управления данными, учитывая критерий быстродействия обработок команд и операций, масштабируемости программных средств при увеличении потребности в дополнительном пространстве для хранения данных.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Определить начальные факторы использования баз данных.
2. Исследовать пути решения коллизионных ситуаций.
3. Исследование и определение скорости обработки команд, запросов, различных вспомогательных действий.
4. Определение количественных затрат на развёртывание, масштабируемости и поддержку программных продуктов.
5. Исследование результатов оптимизации моделей управления данными.
6. Создание собственной модели управления данными.
7. Сопоставление результатов значений данных результатов с представленными хранилищами данных.

Объектом исследования являются модели управления (хранилища, базы) данными такие как Microsoft SQL Server, Allegro Graph, Mongo DB.

Предметом исследования являются быстродействие моделей управления данными, рентабельность, начальные возможности применения в корпоративных системах и после оптимизированные; быстрота реагирование на запросы пользователей добавления, удаления, изменения данных; а также варианты оптимизации операций, серверов и отчёты оптимизации.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность исследования оптимизации и уменьшение затрат на использование различных моделей управления данными. В современном мире без моделей управления данными категорически невозможно обойтись, в силу нужды хранения последних для определения рентабельности многих других ветвей развития корпорации, так и доступа к данным, нужным для работы и сопровождения развития системы в целом. А стоимость оборудования в достаточном количестве велика, потому возникновение вопроса более рентабельной модели управления данными для определённых задач не ново

для современного мира. В узких рамках оборудования серверов различные модели управления данными могут вести себя абсолютно по-разному и выбор верной модели управления может оказаться решающим для заказчика с точки зрения сохранения ресурсов и средств.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии научно-техническими заданиями и планами работ кафедры «Программное обеспечение информационных технологий», и хозяйственными договорами с предприятиями Республики Беларусь.

«Разработать модели, методы, алгоритмы для оценки параметров, повышения надежности и качества функционирования аппаратно-программных средств систем и сетей сложной конфигурации и внедрить в современные обучающие комплексы» (ГБ № 06-2004, № ГР 20111065, научный руководитель НИР – В. В. Бахтизин).

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя О.Г. Смолякова, заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались и обсуждались на XV Республиканской научной конференции студентов и аспирантов «Новые математические методы и компьютерные технологии в проектировании, производстве и научных исследованиях». Материалы работы приняты на использование преподавателями ГГУ им. Ф. Скорины города Гомеля.

Структура и объём диссертации

Диссертация состоит из введения, общей характеристики работы, трёх глав, заключения, списка использованных источников, списка публикаций автора и приложений. В первой главе представлен анализ предметной области, выявлены существенные отличия и проблемы в рамках исследования, показаны направления их решения. Вторая глава посвящена решению проблем

быстродействия, оптимизации, уменьшения времени отклика на действия пользователей, увеличения эффективности использования ресурсов серверов. В третьей главе представлены сравнения характеристик, полученные путём оптимизации моделей управления данных, и результаты экспериментальных исследований метрологических характеристик и практического применения разработанных улучшений моделей управления данных в корпоративных системах.

Общий объём работы составляет 55 страниц, из которых основного текста – 42 страницы, 38 рисунков на 17 страницах, 4 таблицы на 4 страницах, список использованных источников из 59 наименований на 4 страницах и 2 приложения на 4 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** приведён анализ применяемых моделей управления данными, структура хранимых и обрабатываемых данных в файлах, хранимые на дисковом пространстве сервера, где развёрнуты модели управления данными. Проведены анализы каждой модели: MS SQL Server, Mongo DB, Allegro Graph, описаны плюсы и минусы. Описаны обширные возможности каждой из представленной модели управления данными

Современная модель управления данными представляет собой аппаратно-программный комплекс, решающий задачи ввода, обработки и отображения хранимых данных. Базовыми задачами ПО являются инициализация и управление аппаратной частью, прием, первичная обработка и сохранение полученных данных, обеспечение взаимодействия с администратором. К специальным задачам относятся реализация методов и алгоритмов обработки и анализа данных, оптимизация программной и аппаратной части.

Вторая глава посвящена реализации разработке алгоритмов и подходов в оптимизации обработки данных внутри моделей управления данными.

Обеспечение функционирования модели управления данными в режиме реального времени зависит от организации взаимодействия системного ПО и сервера, где расположена модель управления данными.

Существуют хранилища SSD различных типов. Среди них — хранилище SSD на основе DRAM и хранилище SSD на основе технологии флэш-памяти,

такой как одноуровневые ячейки (SLC) и многоуровневые ячейки (MLC). У каждого типа есть свои достоинства и недостатки.

– DRAM. Как обычная оперативная память для компьютера, DRAM отличается очень высоким быстродействием, но ненадежна. Для DRAM требуется постоянный элемент питания, чтобы сохранить данные на время отключения данных. Такие хранилища часто выпускаются в виде плат PCIe, устанавливаемых на системной плате сервера;

– SLC. Быстродействие и жизненный цикл хранилищ на SLC выше, чем у MLC, поэтому SLC используется в хранилищах SSD корпоративного уровня. Однако цена устройств SLC существенно выше, чем у MLC;

– MLC. Обычно флэш-память типа MLC используется в потребительских устройствах и обходится дешевле, чем SLC. Однако у MLC более низкая скорость операций записи и существенно более высокий износ, чем у SLC.

Для оптимизации работы моделей управления данными были использованы следующие стратегии обработки и хранения данных:

1. Дублирование (репликация) позволяет создать полный дубликат базы данных. Так, вместо одного сервера у Вас их будет несколько

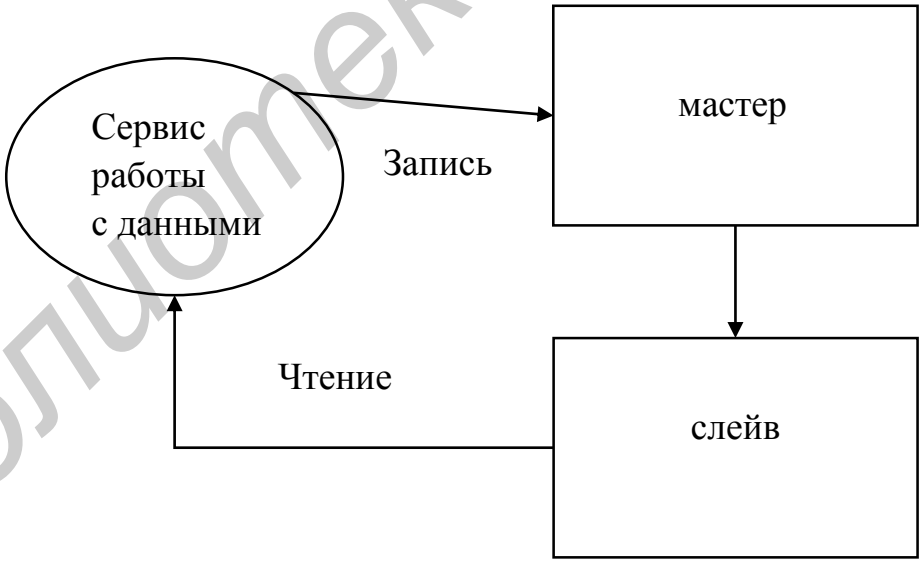

Рисунок 1 – Диаграмма представления стратегии дублирования

– Master — это основной сервер БД, куда поступают все данные. Все изменения в данных (добавление, обновление, удаление) должны происходить на этом сервере;

– Slave — это вспомогательный сервер БД, который копирует все данные с мастера. С этого сервера следует читать данные. Таких серверов может быть несколько.

Репликация позволяет использовать два или больше одинаковых серверов вместо одного. Операций чтения (SELECT) данных часто намного больше, чем операций изменения данных (INSERT/UPDATE). Поэтому, репликация позволяет разгрузить основной сервер за счет переноса операций чтения на слейв.

2. Шардинг (дублирование) — это другая техника масштабирования работы с данными. Суть его в разделении базы данных на отдельные части так, чтобы каждую из них можно было вынести на отдельный сервер. Этот процесс зависит от структуры Вашей базы данных и выполняется прямо в приложении в отличие от репликации

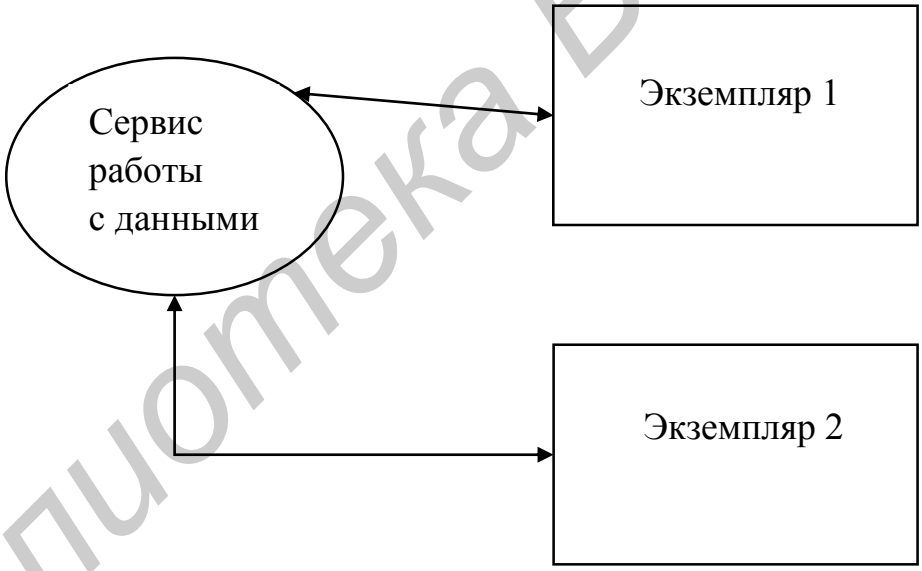

Рисунок 2 – Диаграмма представления стратегии клонирования

Вертикальный шардинг — это выделение таблицы или группы таблиц на отдельный сервер.

Горизонтальный шардинг — это разделение одной таблицы на разные сервера. Это необходимо использовать для огромных таблиц, которые не умещаются на одном сервере. Разделение таблицы на куски делается по такому принципу:

– На нескольких серверах создается одна и та же таблица (только структура, без данных);

– В приложении выбирается условие, по которому будет определяться нужное соединение (например, четные на один сервер, а нечетные — на другой).

Перед каждым обращением к таблице происходит выбор нужного соединения.

Схема, созданной модели управления данными представлена на рисунке 3.

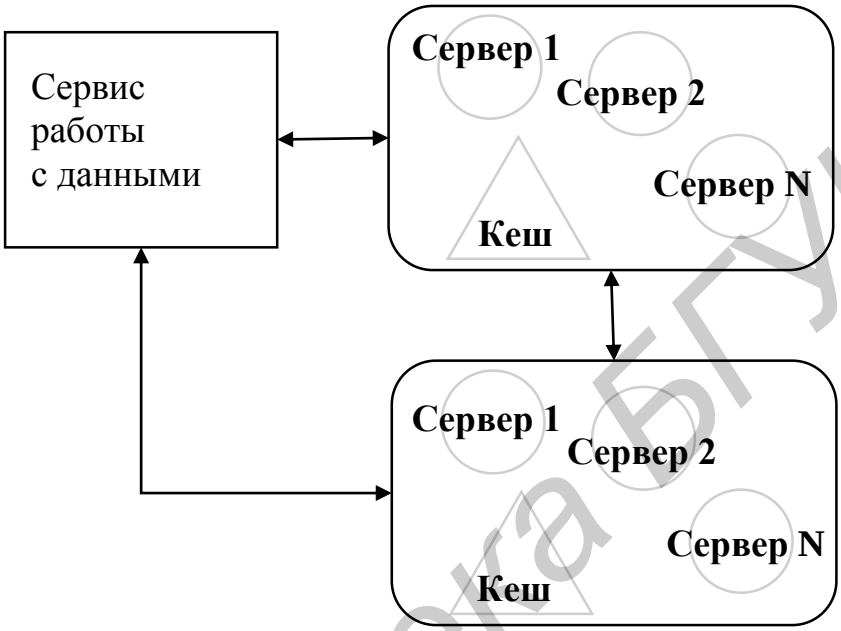

Рисунок 3 – Схема модифицированной модели управления данными.

Для усовершенствования оптимизации серверов и баз данных была применена улучшенная модель управления данными. По результатам исследования, преимущества использования смешения и модификации моделей при помощи кеширования основных данных стали очевидным. Улучшенная модель управления данными состоит из стратегии представления разделением и дублированием данных, а также добавлением промежуточных кэшированных отделов, размер которых соответствует используемому типу базы данных. Результатами применения модифицированной модели управления данными стало уменьшение потребления времени процессора и ресурсов винчестера, а также возросла скорость отклика на транзакции внутри системы в целом.

Увеличение потребления оперативной памяти не усугубляет ситуацию, а только показывает качество проработанной модели, так как скорость поиска по памяти значительно выше, чем по винчестеру.

В **третьей главе** предъявлены результаты модифицированной модели управления данными. Представлены мониторинги и графики, показывающие использование центрального процессора, оперативной памяти и винчестера;

основных компонентов сервера для использования модели управления данными.

На первом этапе настройки использовались для моделей управления данными стандартно предоставляемые разработчиком.

Далее заполнялись хранилища колоссальным количеством данных, для более успешного и результативного тестирования эффективности, рентабельности, исследования поведения, узких мест моделей хранения данных. Записывались результаты.

На последнем этапе оптимизировались сервера для хранения предоставляемых тестированию данных и сами модели управления данными, хранилища. После чего сравнивались полученные результаты с предыдущими полученными, и составлялись выводы в зависимости от результатов поведения моделей хранения данных в старой неоптимизированной и новой оптимизированной средах.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

Учитывая миллионы транзакций, производимые внутри компаний и корпораций таких как Coca Cola, MTV, COUB в течении суток – нужно понимать, что для максимальной эффективности предоставляемой информации пользователям сайтов и внутренних систем компаний нужно использовать не одно хранилище данных, а их совокупность. Но в таком случае, только возрастает ответственность администраторов и соответственно затраты на поддержку, расширение, и обеспечение максимальной эффективности совокупности систем использующие различные хранилища данных.

В работе рассмотрены некоторые задачи проектирования информационных структур, территориально распределенных БД. Показано, что большинство задач сводятся к задачам нелинейного целочисленного программирования с булевыми переменными и ограничениями. Основные трудности поиска точного оптимального решения задач подобного класса связаны с большими размерностями дискретных переменных и неоднозначными требованиями с различными возможностями развёртывания, ограничения средств на разработку и поддержку.

Были сформулированы и обоснованы принципы проектирования территориально структур распределенных баз данных. В качестве глобального критерия синтеза структуры базы данных предложен комплексный критерий

экономического характера, а именно – минимум стоимости хранения, но максимум эффективности путём оптимизации и дефрагментации информации.

Методы и модели управления, описанные в работе, представляют собой методический и математический аппарат для формализации анализа и синтеза информационных структур, территориально распределенных моделей.

Использование материалов данной работы позволяет существенно повысить эффективность использования информационных ресурсов в корпоративных системах за счет выбора оптимальных структур, размещения моделей управления (баз) данными и оптимизации степени дублирования информации.

Можно сделать вывод: данное исследование помогает отечественным разработчикам программного обеспечения и корпоративных систем, а также управляющим проектами предлагать заказчику отечественное исследование по применению различных моделей управления данными (хранилищ), что в свою очередь перенесёт исследование и расширение данной области в Республику Беларусь.

Рекомендации по практическому использованию результатов

1. Для серверов с табличными представлениями данных размер кэша должен быть в два раза большим, чем размер таблицы, хранящая индексы и размер предоставляемый для размещения хранимых процедур вместе взятых.
2. Для серверов с документо-ориентированным хранением должен быть предоставлен размен кэша равный 10%-ам от размера одного осколка по стратегии разделения.
3. Для серверов с графо-ориентированным хранением данных должен предоставляться размер кэша равный размеру осколка с типами структур, используемых в базе данных.
4. Полученные результаты формулируют теоретическую и практическую базу для разработки ПО вместе с моделями управления данными в корпоративных системах в целом. Они могут быть использованы для модернизации и дальнейшего развития существующих систем.
5. Разработанные методы и алгоритмы анализа стрессоустойчивости систем, серверов и надлежащим образом составленных данных могут применяться в автоматизации анализа состояния систем, и исследованием узких мест развития хранилищ данных в целом в рамках корпорации
6. Результаты работы могут использоваться при подготовке администраторов для разработки, обслуживания компьютерных систем, настройки и поддержания серверов.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Гладченко, А.В. Модели управления данными в корпоративных системах // Современные научные исследования и инновации. 2014. № 12 [Электронный ресурс]. URL: http%3A%2F%2Fweb.snauka.ru%2Fissues%2F2014%2F12%2F34565

2. Цурганова, Л.А. Объектно-ориентированное моделирование осадки свайного фундамента на C# / Л.А. Цурганова, А.В. Гладченко // «Новые математические методы и компьютерные технологии в проектировании и научных исследованиях», XIV Республиканская научная конференция студентов и аспирантов (2011, Гомель). XIV Республиканская научная конференция сту-дентов и аспирантов «Новые математические методы и компьютерные техно-логии в проектировании, производстве и научных исследованиях», 21-23 марта 2011 г.: [материалы]: в 2 ч. Ч.1 / редкол.: О.М. Демиденко (гл. ред.) [и др.]. – (гл. ред.) [и др.]. – Гомель: ГГУ им Ф.Скорины, 2011. С. 67-68.

3. Цурганова, Л.А. Компьютерное моделирование системы «Свая – грунтовое основание» на языке программирования C# / Л.А. Цурганова, А.В. Гладченко // Творчество молодых ‘2011. Сб. научных работ т и аспирантов УО «Гомельский госуниверситете имени Ф.Скорины». В 2 ч. Ч.1. Гомель: ГГУ им. Ф. Скорины, 2011. – С. 78–80.