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

УДК 004.383

Коржовник
Дмитрий Александрович

Оптимизация инфраструктуры хранения данных в региональных информационно-аналитических 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
Лапицкая Н.В.
к.т.н., доцент

_____

Минск 2015

КРАТКОЕ ВВЕДЕНИЕ

Фундаментом любой информационной системы являются данные, в случае региональной информационно аналитической системы эти данные представляются в виде развитой инфраструктуры их хранения. Эта инфраструктура есть не что иное, как сеть, объединяющая сервера данных и конфигурируемые вычислительные ресурсы, например другие устройствам хранения данных, пользовательские приложения и сервисы в единую систему. Таким образом, переходя на уровень виртуальной инфраструктуры, так как для организованного функционирования любой системе требуется набор собственного и вспомогательного ПО, на базе соответствующих операционных систем. Для большинство систем по прежнему работают со своими локальными базами данных созданных на основе СУБД. Большинство современных СУБД— реляционные, т.е. представляют данные в виде двумерной таблицы, в которой есть строки (записи) и столбцы (поля записей). Но на практике информационные системы имеют немного другую структуру информационного взаимодействия, когда данные поступают из базы с одного сервера в базу на другом сервере, где както обрабатываться и агрегируются.

Для региональных информационно – аналитических систем, деятельность которых в основном связана с анализом накопленных данных о деятельности организации, которые могут быть четко структурированы, сетевая модель представляет собой более частным случаем иерархической модели. В этой модели запрос, направленный вниз по иерархии, прост (например, какие заказы принадлежат этому покупателю); однако запрос, направленный вверх по иерархии, более сложен (например, какой покупатель поместил этот заказ). Также, трудно представить не иерархические данные при использовании этой модели. Поэтому наиболее характерными операциями над данными для таких систем являются чтение и поиск, так можно выделить отдельно задачу по визуализации этих данных, в то время как новые данные поступают уже в собранном виде из других баз. Именно в этот этап сопряжен с риском столкнуть с проблемами производительности, поскольку внутренняя логика базы источника может быть скрыта и плохо известна разработчику самой аналитической системы. Более того он и не должен об этом беспокоиться, для него главное формат и объем входных данных или выданных, но на практике часто приходится сталкиваться с необходимостью проводить анализа узких мест производительности это может быть сложной задачей.

Именно в выявлении таких узких мест в режиме реального времени хотелось бы использовать компьютеры общего назначения, не требующие высокой квалификации обслуживающего персонала, при это не нарушая SOA принципа развития системы. А для последующего решения проблемы выработать подходящую методологию для оптимизации инфраструктуры хранения данных системы.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является разработка подхода и программного обеспечения для решения задач оптимизации инфраструктуры хранения данных региональной информационно-аналитической системы на базе персональных компьютеров общего назначения.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Определить и проанализировать показатели производительности СУБД на базе ОС семейства Windows подход для их получения в режиме реального времени.
2. Разработать архитектуру программного средства с функциями мониторинга деятельности подсистемы хранения данных информационной системы.
3. Разработать рекомендации для оптимизации инфраструктуры хранения данных информационной системы на основе результатов деятельности ПС.
4. Реализовать ПС и провести экспериментальные исследование на его основе.

Объектом исследования является инфраструктура хранения данных региональной информационно-аналитической системы.

Предметом исследования является программное обеспечение компьютерных систем для решения задач оптимизации инфраструктуры хранения данных.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования компьютеров общего назначения с ОС Windows для решения задачи оптимизации инфраструктуры хранения данных, на основе добавления в инфраструктуру средств ее мониторинга по заданным показателям. И таким образом выявлять проблемы производительности на уровне баз данных входящих в инфраструктуру хранения данных системы, этим самым повысить производительность всей системы.

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:

1. «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 № 11-2004, № ГР 20111065, научный руководитель НИР – В. В. Бахтизин).

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Лапицкой Н.В., заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались и обсуждались на международной конференции посвященной информационным технологиям.

Опубликованность результатов диссертации

По теме диссертации опубликовано 1 статья в сборнике трудов и материалов 51 научно-практической конференции БГУИР.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, четырех глав, заключения, списка использованных источников, списка публикаций автора и приложений.

В первой главе представлен анализ предметной области, выявлены основные существующие проблемы в рамках тематики исследования, показаны направления их решения.

Вторая глава посвящена теоретическому обзору предметной области и выбору информационных технологий, на основе данных экспериментальных исследований. В соответствии с этим были выработаны функциональные требования к программному средству.

В третьей главе предложены методы повышения эффективности предоставления информационных услуг в центрах обработки данных: метод управления нагрузкой, метод распределения памяти в ЦОД. И предложены практические подходы по оптимизации инфраструктуру хранения данных через оптимизацию БД на серверах-узлах информационно аналитической системы.

В четвертой главе предложена практическая реализация ПС и методика его использования в инфраструктуре хранения данных для выявления проблем производительности и последующей оптимизации.

Общий объем работы составляет 64 страниц, из которых основного текста – 40 страниц, 18 рисунков на 12 страницах, 7 таблиц на 6 страницах, список использованных источников из 28 наименований на 2 страницах и 2 приложения на 7 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** представлен анализ предметной области, выявлены основные существующие проблемы в рамках тематики исследования, показаны направления их решения.

В ходе анализа предметной области было установлено, что для большинства информационно-анаитических систем характерна многоуровневая иерархическая инфраструктура хранения данных узлами в которой являются компьютеры общего назначения с развернутыми на них базами данных. Для такого рода инфраструктур целесообразнее всего проводить оптимизацию на основе мониторинга активности на серверах инфраструктуры.

Поскольку функция мониторинга является стандартной для большинства ПО используемого как инструмент администрирования базы данных было рассмотреть несколько аналогов такого ПО предназначенного для работы с MS SQL Server.

Проведенный анализ показал следующее:

1) Проекты DMVStats и SQLSTAT2005 облегчают сбор и анализ сведений, представляемый SQL Server через DMV и DMF. Если их сравнивать, то с одной стороны DMVStats – легче ставится, настраивается и содержит больше готовых отчетов, но не развивается в настоящее время. С другой стороны, SQLSTATS2005 – сделан проще и находится в развитии. Если разработка и поддержка SQLSTATS2005 будет продолжена, в будущем это будет более удобный инструмент для работы.

2) Кроме этих инструментов, можно самим использовать подход, применяемый в них. Это создание своих таблиц и задание, которое по расписанию собирает данные из DMV и DMF для хранения в ваших таблицах для последующего анализа. Для создания SQL запросов вы можете воспользоваться диаграммами, которые показывают отношения системных таблицы и административных представлений: SQL Server System Views Map и SQL Server System Views Map.

3) Рассмотренные аналоги реализуют внешний мониторинг. Они не очень сложны в конфигурировании, но имеют избыточную функциональность, что усложняет наглядность их использования администраторами. Кроме того их сложно использовать для внедрения в инфраструктуру информационной системы согласно принципам SOA – поскольку они являются доступными приложениями что затрудняет реализации мониторинга сетевой (иерархической) инфраструктуры хранения данных в реальном времени.

4) Среди характерных черт можно выделить то, что для получения статистической информации о производительности сервера и его активности они ис-

пользуют DVM, которые являются частью DMO SQL Server, при этом вся пол0443енная информация может быть сохранена в форме отчётов.

На основе проведенного анализа был сделан вывод о том, что для оптимизации инфраструктуры хранения данных информационно-аналитической системы. Необходимо выбрать показатели производительности системы, разработать ПО для их мониторинга, провести экспериментальную апробацию и определить подход для проведения оптимизации с целью решения выявленной проблемы.

Вторая глава посвящена теоретическому обзору предметной области и выбору информационных технологий, на основе данных экспериментальных исследований. В соответствии с этим были выработаны функциональные требования к программному средству.

В ходе моделирования предметной области было установлено, что для частного случая иерархической структуры время получения данных на корневом сервере системе будет равно суммарного времени доступа к данным на каждом из задействованных дочерних узлов. Соответственно для уменьшения это времени нужно решать проблему локально на каждом узле и если можно перераспределять нагрузку на дочерних серверах в соответствии с приоритетом и частотой обращения к ним.

Математическим аппаратом для оценки эффективности каждой отдельной БД ИС можно использовать показатели Временной эффективности, но их применение к современным СУБД не очень объективно, поскольку на СУБД сильно влияет текущая конфигурация.

Поэтому нужно было установить насколько продходят СУБД для задач импорта и поиска данных в иерархической инфраструктуре информационной системы. Для этого был поведен экспериментальный анализ в результат чего был сделан общий вывод:

Производительность сервера данных в подобной инфраструктуре, при полном цикле (прием + передача + очистка) составила 18 000 000 (18 млн.) записей продаж в сутки (при пропорциональной передаче SYSLOG). Производительность сервера данных на MSSQL существенно выше, чем на Oracle (почти в 7 раз при транзитной передаче). Отметим, что сравнение касается не СУБД как таковых, а сервера данных, работающего на этих СУБД. Кроме того, СУБД, как отмечалось ранее, используются с настройками по умолчанию и файлами данных, находящихся на системном диске. MSSQL более производительна.

Как было отмечено ранее расширение информационной сети должно идти согласно парадигме SOA. Главное, что отличает SOA - это использование независимых сервисов с чётко определёнными интерфейсами, которые для выполнения своих задач могут быть вызваны неким стандартным способом, при условии, что сервисы заранее ничего не знают о приложении, которое их вызовет, а приложение не знает, каким образом сервисы выполняют свою задачу.

То самым подходящим выбором с учетом выше описанных условий будет использование технологии .Net и языка C# которые интегрированы в структуру и API взаимодействия со службой SQL Server, которая была выбрана в качестве

базовой СУБД, язык T-SQL. Веб сервисы разрабатываемого ПС смогут быть развернуты на IIS 7.0 на базе ОС семейства Windows.

В **третьей главе** предложены методы повышения эффективности предоставления информационных услуг в центрах обработки данных: метод управления нагрузкой, метод распределения памяти. И предложены практические подходы по оптимизации инфраструктуру хранения данных через оптимизацию БД на серверах-узлах информационно аналитической системы.

Необходимо заметить, что оптимизация настроек программных средств, как самих приложений, так и операционной системы, дает существенно больший прирост производительности системы, чем использование более мощной аппаратуры. Обусловлено это в первую очередь тем, что оптимизация настроек устраняет "узкие места" (bottleneck) на путях следования потоков данных, тогда как новая аппаратура делает "горлышко бутылки" чуть шире и только (хотя иногда и этого достаточно для решения проблем быстродействия).

Оптимизации базовой инфраструктуры— это структурированный, систематический процесс оценки всех возможностей ИТ-инфраструктуры и разработка плана по ее усовершенствованию с целью создания экономически эффективных ИТ-сервисов.

В рамках любой модели оптимизации существует четыре уровня оптимизации: базовый, стандартизированный, рационализированный/расширенный и динамический.

Н а базовом уровне есть этап работы с инфраструктурой хранения данных. Этот этап охватывает способы управления и поддержки серверов, запоминающих устройств, сетевых компонентов и другого оборудования и программного обеспечения центра обработки данных (ЦОД). Сюда входит развертывание операционных систем, установка обновлений и приложений в масштабах всего учебного заведения, а также применение виртуализации для оптимизации ИТ-инфраструктуры и повышения ее управляемости. К этой области относятся такие аспекты безопасности, как брандмауэр, система предотвращения вторжений и антивирус, а также настройка сетевых ресурсов и пропускной способности сети.

Кроме того, этап охватывает эффективные способы работы с хранилищами данных, повышающие безопасность и доступность данных без избыточной нагрузки на ИТ-ресурсы. Именно в этой сфере рассматриваются задачи данной работы. Т.е. на виртуальном уровне инфраструктур через добавление в инфраструктуру средств мониторинга

Метод управления нагрузкой. Нагрузка на ЦОД с течением времени меняется и если выделять необходимые для решения заданий пользователей ресурсы только на основании значения пиковой интенсивности, то часть ресурса не будет востребована в процессе выполнения запроса. Следовательно, используя статистические свойства нагрузки, администрация ЦОД может без потери качества осуществить статистическое мультиплексирование нагрузки, что позволяет предоставить суммарный общий кредит всем клиентам, превышающий общую пропускную способность ЦОД. Для обеспечения выполнения SLA-соглашения предлагается выдавать клиентам ЦОД так называемый «кредит»,

т.е. разрешение на передачу данных со скоростью, не превышающей заданную таким «кредитом». Для уменьшения числа потерянных заявок введен буфер ограниченного объема, куда будут помещаться заявки в случае превышения порогового значения счетчика. Эти ячейки будут повторно поставлены на передачу, когда нагрузка на вход ЦОД уменьшится. Тогда потеря заявок будет происходить только в случае переполнения буфера. Для осуществления такого управления потоками заявок был применен игровой метод, когда управляющее устройство состоит из коллектива вероятностных автоматов, закрепленных за виртуальными каналами, по которым передаются заявки от соответствующего пользователя. Работа автомата задается действиями 1,…,n. Действие означает выбор i-го приращения к заданному кредиту с вероятностью mi. из n возможных при поступлении соответствующей заявки. Выбор действия заявки задается вектором-строкой

$$M=(m_1,\dots,m_n), \quad \text{где } m_i \geq 0, \quad \sum_{i=1}^{n} m_i = 1.$$

Предложен метод распределения памяти в ЦОД, который позволяет сократить число запросов к часто используемым ресурсам и тем самым уменьшить время ответа. Результатом является план распределения памяти с ресурсами по серверам кластера

Приведены рекомендации по локальной оптимизации. Описаны области в которых производительность может быть улучшена посредством использования информации, уже собираемой SQL Server, и могут быть использованы разработанным программным средством.

Приведенные в работе SQL коды могут быть использованы как исходные файлы скриптов передаваемые ПС для выполнения на сервере БД инфраструктуры хранения данных информационной системы., для выявления проблем производительности и последующего их устранения. Приведенный в работе перечень не является полным т.к. ПС позволят передать на выполнение узлу системы на выполнений любой авторский SQL скрипт.

В **четвертой главе** предложена практическая реализация ПС и методика его использования в инфраструктуре хранения данных для выявления проблем производительности и последующей оптимизации.

При разработке архитектуры программного средства можно выделить основные программные модули представленные на Рисунке 1.

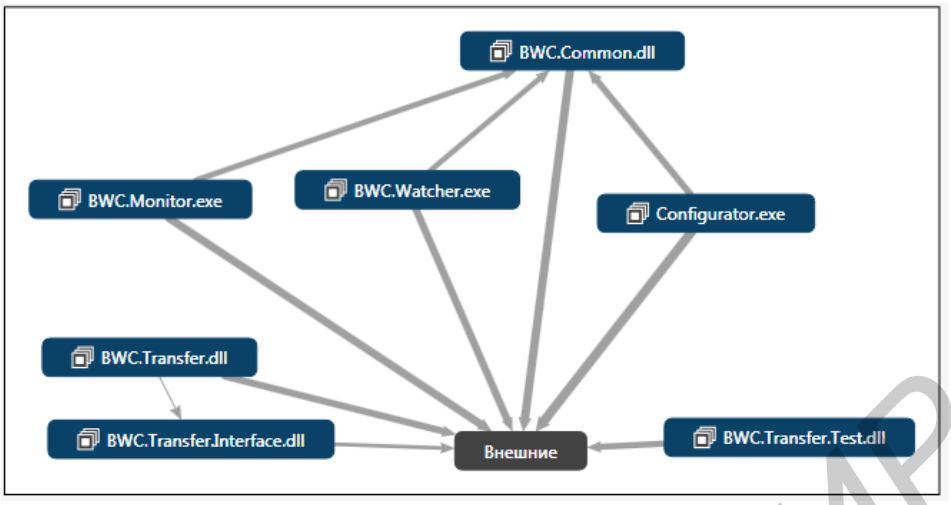

Рисунок 1 – Диаграмма зависимостей программных модулей программного средства

Применение модели итеративной разработки для создания ПС на платформы .NET. и использование шаблона проектирования «модель – представление – контроллер», для реализации модуля BWC.Configurator, позволяет уменьшить связность между объектами, отделить модель от ее представлений, тестировать логику контроллера независимо от представления и упростить логику представления, уменьшить количество обращений к модели.

Экспериментальные исследования ПС показали его эффективность для решения задач мониторинга и выявления узких мест в инфраструктуре хранения данных.

Созданное ПС организовано по модульно-функциональному принципу, позволяет использовать его функциональную модель и структуру, в качестве шаблона для реализации ПС компьютерных систем схожих типов.

Внедрение разработанного ПС в ИТ-инфраструктуру информационно аналитической системы на уровне инфраструктуры хранения данных будет легко выполнимой задачей поскольку соблюдались принципы SOA.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1) Проведен сравнительный анализ некоторых существующих аналогов и проведено экспериментальное нагрузочное СУБД MS SQL, чтобы понять насколько они подходят для выполнения задач хранения данных в инфраструктуре иерархических информационно-аналитических систем.

2) Предложены методы повышения для оценки эффективности работы центров обработки данных, являющихся центральными элементами инфраструктуры хранения данных в информационных системах: метод управления нагрузкой и метод распределения памяти. Разработан гибкий инструментарий и предложена методика по оптимизации БД и Серверов СУБД на базе MS SQL

Server, с целью устранения узких мест на узлах инфраструктуры и повышения производительности информационно аналитической системы при выполнении задач импорта, чтения и поиска данных.

3) Предложена архитектура программного средства для решения задачи внешнего мониторинга информационного обмена внутри инфраструктуры хранения данных построенной по иерархической модели на компьютерах общего назначения. Работа программного средства в режиме реально времени и гибкость в задании анализируемых показателей обеспечивает актуальность собираемых данных о состоянии инфраструктуры.

4) Экспериментально проверено, что соответствие архитектуры ПС парадигме SOA, позволяет его легко включить в структуры информационно аналитической системы и выявлять проблемы производительности

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для разработки ПС. Они могут быть использованы для модернизации и дальнейшего развития существующих систем.
2. Предложен подход по оптимизации инфраструктуры хранения данных и методики для оценки текущей производительности информационно аналитической системы
3. Результаты работы могут использоваться при подготовке персонала для разработки и обслуживания компьютерных систем, решающих задачи администрирования и разработки баз данных.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Коржовник, Д.А Автоматизация создания olap-кубов на базе mssql для оптимизации инфраструктуры региональной информационно аналитической системы. Доклады БГУИР 51-я научная конференция– 2015.