Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.75

Пацовский
Алексей Александрович

Модели и алгоритмы удалённого управления данными в распределённой системе

АВТОРЕФЕРАТ

на соискание степени магистра технических наук
по специальности 1-40 80 05 Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных систем

Научный руководитель
Глухова Лилия Александровна
к.т.н., доцент кафедры ПОИТ

Минск 2016

КРАТКОЕ ВВЕДЕНИЕ

В наши дни автоматизация различных технологических и управленческих процессов, без которых немыслимо эффективное решение задач управления промышленным или торговым предприятием, банком, учебным заведением, государственной структурой, основывается на переработке больших объёмов информации. Эффективность автоматизированных информационных управляющих систем в значительной мере зависит от того, насколько обеспечиваются скорость доступа к данным, а также полнота, достоверность, непротиворечивость данных. Обычно информационная система представляет собой интегрированную систему, ядро которой составляет база данных.

В настоящее время всё более популярными становятся алгоритмы и механизмы, позволяющие работать с данными, хранящимися в физической базе данных на сервере, со своего локального компьютера. Ведь в этом случае пользователю нет необходимости устанавливать и настраивать какое-либо специфическое программное обеспечение, пользователь может выполнить все операции удалённо с браузера.

Если в 1980-1990-ых годах один компьютер был мощным устройством, предназначенным для решения широкого круга задач, то сейчас производительности одного компьютера уже иногда не хватает. Также во многих случаях задачи выполняются слишком долго, что негативно сказывается на отношении пользователей к конкретным программным средствам. Поэтому становятся всё более популярными распределённые системы, позволяющие заметно повысить производительность различных операций и в целом программных средств. Если рассматривать распределённую систему в привязке к компьютерным сетям, то распределённая система представляет собой высокоскоростную компьютерную сеть, состоящую из множества компьютеров, которые работают совместно, представая в виде единой связной системы. Их важное преимущество состоит в том, что они упрощают интеграцию различных приложений, работающих на разных компьютерах, в единую систему. Также в случае грамотного использования распределённые системы позволяют получить довольно большой прирост производительности.

Поскольку операции с данными зачастую являются продолжительными, то для их ускорения можно использовать распределённую систему. Таким образом, исследование возможности использования распределённой системы для ускорения операций, связанных с удалённым управлением данными, а также разработка соответствующих моделей и алгоритмов, является актуальной.

Диссертационная работа посвящена разработке моделей и алгоритмов, предназначенных для удалённого управления данными в распределённой системе. Использование разработанных моделей и алгоритмов позволит сократить время выполнения различных операций, предназначенных для работы с данными.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является разработка моделей и алгоритмов удалённого управления данными в распределённой системе.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Произвести обзор предметной области удалённого управления данными в распределённой системе.
2. Исследовать существующие способы организации удалённого управления данными в распределённой системе.
3. Разработать модели удалённого управления данными в распределённой системе.
4. Разработать алгоритмы удалённого управления данными в распределённой системе.
5. Разработать функциональную модель предметной области удалённого управления данными.
6. Разработать программное средство, реализующее разработанные модели и алгоритмы.
7. Провести экспериментальную оценку разработанных моделей и алгоритмов.

Объектом исследования являются распределённые системы.

Предметом исследования является организация удалённого управления данными в распределённой системе.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность использования распределённой системы для ускорения и повышения производительности различных операций и сложных алгоритмов, предназначенных для удалённого управления данными, хранящимися в базе данных.

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

Работа выполнялась в соответствии с научно-техническим заданием и планом работ кафедры «Программное обеспечение информационных технологий» по теме «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 Л. А. Глуховой заключается в формулировке целей и задач исследования.

Апробация результатов диссертации

Основные положения диссертационной работы докладывались и обсуждались на VII Международной научно-методической конференции «Высшее техническое образование: проблемы и пути развития» (Минск, Беларусь, 2014); 51-ой научной конференции аспирантов, магистрантов и студентов БГУИР (Минск, Беларусь, 2015); XVIII Республиканской научной конференции студентов и аспирантов «Новые математические методы и компьютерные технологии в проектировании, производстве и научных исследованиях» (Гомель, Беларусь, 2015); IX Международной научно-методической конференции «Дистанционное обучение – образовательная среда XXI века» (Миснк, Беларусь, 2015).

Опубликованность результатов диссертации

По теме диссертации опубликовано 4 печатных работы, из них 2 работы в сборниках трудов и материалов международных конференций, 1 работа в сборниках трудов и материалах республиканских конференций, 1 работа в сборниках трудов и материалах конференций в БГУИР.

Структура и объем диссертации

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

В первой главе представлен анализ предметной области, проведено исследование существующих способов организации удалённого управления данными в распределённых системах, определены их достоинства и недостатки.

Вторая глава посвящена теоретической разработке математической модели времени обработки данных в распределённой системе. В ней рассмотрены основные архитектурные модели распределённых систем, предназначенных для удалённого управления данными, проведено их исследование на основе разработанного математического аппарата, разработаны алгоритмы, предназначенные для выполнения удалённых операций с огромными массивами данных.

В третьей главе разработана функциональная модель предметной области и спроектировано программное средство, реализующее разработанные

модели и алгоритмы для последующего их экспериментального исследования. В данной главе произведена экспериментальная оценка разработанных моделей и алгоритмов и сделаны соответствующие выводы.

Общий объем работы составляет 97 страниц, из которых основного текста – 89 страниц, 23 рисунка на 23 страницах, 8 таблиц на 7 страницах, список использованных источников из 36 наименований на 3 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** проведено исследование организации удалённого управления данными в распределённых системах, в процессе которого выполнен обзор предметной области удалённого управления данными в распределённой системе, а также исследованы существующие способы организации удалённого управления данными в распределённых системах. На основание произведённого анализа выполнена постановка задачи.

Показано, что модель клиент-сервер является базовой моделью и используется практически в любой распределённой системе. Рассматривая множество приложений типа клиент-сервер, предназначенных для организации доступа пользователей к базам данных, в том числе и для удалённого управления данными, следует выделять следующие три уровня:

- уровень пользовательского интерфейса;
- уровень обработки;
- уровень данных.

Уровень пользовательского интерфейса содержит всё необходимое для непосредственного общения с пользователем. Уровень пользовательского интерфейса обычно реализуется на клиентах. Этот уровень содержит программы, посредством которых пользователь может взаимодействовать с приложением.

Многие приложения модели клиент-сервер построены из трех различных частей: части, которая занимается взаимодействием с пользователем, части, которая отвечает за работу с базой данных или файловой системой, и средней части, реализующей основную функциональность приложения. Эта средняя часть логически располагается на уровне обработки. В противоположность пользовательским интерфейсам или базам данных на уровне обработки трудно выделить общие закономерности.

Уровень данных в модели клиент-сервер содержит программы, которые предоставляют данные обрабатывающим их приложениям. Специфическим свойством этого уровня является требование сохранности. Это означает, что когда приложение не работает, данные должны сохраняться в опреде-

ленном месте в расчете на дальнейшее использование. В модели клиент-сервер уровень данных обычно находится на стороне сервера.

На рисунке 1 представлена обобщённая модель архитектуры клиент-сервер. Все существующие распределённые архитектуры так или иначе опираются на классическую архитектуру. На рисунке можно отчётливо видеть клиентскую и серверную часть.

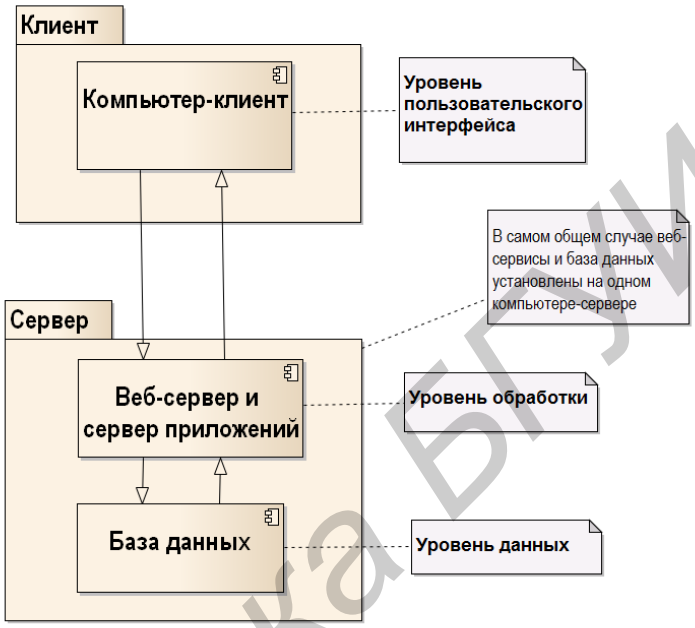

Рисунок 1 – Обобщённая модель архитектуры клиент-сервер

Потенциальным узким местом классической архитектуры является сервер баз данных. Если сервер баз данных становится перегруженным, единственным выходом из положения является покупка более мощной машины. Машины, используемые для поддержки серверов баз данных, обычно стоят недешево, поскольку они должны уметь справляться с пиковой рабочей нагрузкой. Поэтому классической архитектуре свойственны ограничения по отношению и к масштабируемости, и к стоимости – двум важным целям «облачных» вычислений.

Архитектура клиент-сервер с распределённой базой данных отличается от классической архитектуры клиент-сервер наличием распределённой базы данных, которая, как правило, устанавливается на нескольких компьютерах-серверах и позволяет добиться большей масштабируемости.

Однако одним из недостатков рассматриваемых архитектур является их универсальный характер и невозможность их адаптирования под конкретные пользовательские задачи с целью получения максимальной производительности.

Результаты исследований, проведенных в этих направлениях, отражены в работах С. Д. Кузнецова, А. А. Цымбала, К. Дж. Дейта, Т. Канноли, А.

Бондаря, В. Васвани, Э. Таненбаума, М. К. Буза, В. Г. Хорошевского, Л. Е. Карпова, Л. В. Кулагина, Э. Ньюкомера, М. А. Посыпкина, Х. Курди (H. Kurdi), Т. Гуэсми (T. Guesmi) и др.

Вторая глава посвящена разработке моделей и алгоритмов удалённого управления данными в распределённых системах.

Разработана математическая модель времени выполнения операции в распределённой системе, позволяющая оценить ту или иную архитектурную модель. Главным критерием в математической модели является время выполнения операции удалённого управления данными. Разработанную математическую модель можно использовать для определения целесообразности использования распределённой системы для какого-либо алгоритма, а также для определения количества серверов, при котором время выполнения будет минимальным.

Для задачи удалённого управления данными при использовании архитектуры распределённой системы, основанной на классической архитектуре клиент-сервер, общее затраченное время можно вычислить по формуле (1):

$$T = 2T_{\text{вз}} + T_{\text{обр}} + T_{\text{бд}} + 2T_{\text{вз.бд}} \quad (1)$$

где T – общее время выполнения операции удалённого управления данными;

$T_{\text{вз}}$ – время соединения и передачи запроса с компьютера пользователя на сервер и обратно;

$T_{\text{обр}}$ – время обработки данных, выполнения основных вычислений на уровне сервера;

$T_{\text{бд}}$ – время обработки данных на уровне базы данных;

$T_{\text{вз.бд}}$– время соединения и передачи запроса с сервера на базу данных и обратно.

На рисунке 2 изображена модель последовательности выполнения любой операции, связанной с удалённым управлением данными, в распределённой системе.

Использование распределённой системы может значительно сократить время обработки $T_{\text{обр}}$, если алгоритм обработки данных является хорошо распараллеливаемым.

С учётом различных задержек, возникающих в результате использования распределённых систем, время обработки можно определить выражением (2):

$$T_{\text{обр}}(N) = K_1 + \frac{K_2}{N} + N * \left(\delta + t_{\text{соед}} + t_{\text{пер}}\right), \quad (2)$$

где K_1 – время, в процессе которого задача выполняется на центральном сервере, данное время нет возможности распараллелить;

K_2 – время, эквивалентное времени выполнения части задачи, которая может быть распараллелена, без использования распределённых серверов;

N – количество серверов, используемых для обработки данных;

$t_{\text{соед}}$ – среднее время, необходимое для установления соединения между двумя серверами;

$t_{\text{пер}}$ – среднее время, необходимое для передачи данных между двумя серверами;

δ – служебное время, необходимое для инициализации и запуска процесса на одном компьютере.

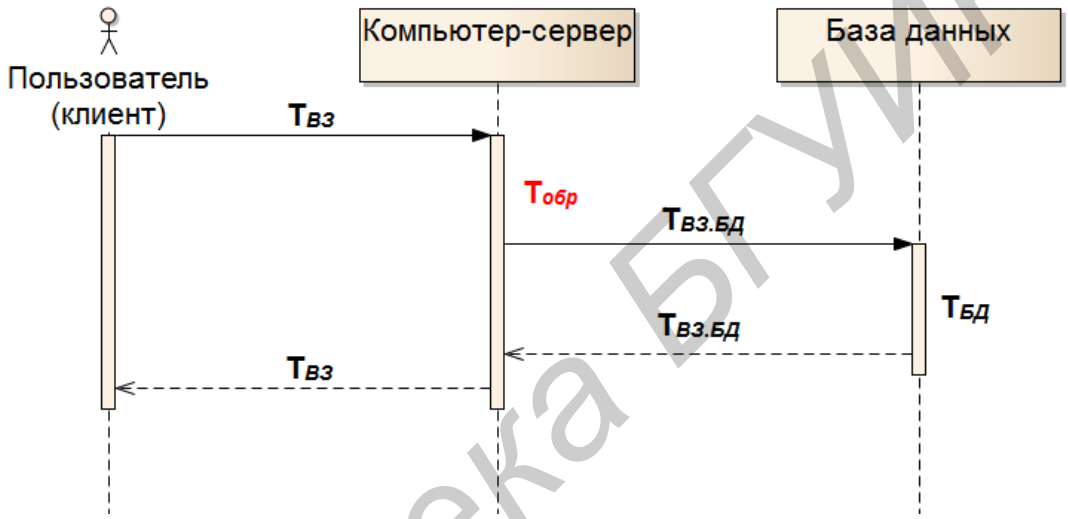

Рисунок 2 – Модель выполнения операции удалённого управления данными

Минимально возможное время обработки можно вычислить по формуле (3):

$$T_{\text{обр.мин}} = K_1 + 2 * \sqrt{K_2 * \left(\delta + t_{\text{соед}} + t_{\text{пер}}\right)}\,. \tag{3}$$

Максимальное количество серверов, при которых время обработки будет наименьшим можно вычислить по формуле (4):

$$N_{\text{макс}} = \left\lfloor \sqrt{\frac{K_2}{\delta + t_{\text{соед}} + t_{\text{пер}}}} \right\rfloor. \tag{4}$$

Минимальное количество серверов, для которого целесообразно использовать распределённую систему, можно определить по формуле (5):

$$N_{\text{мин}} = \left\lceil \frac{K_2 - \sqrt{{K_2}^2 - 4 * K_2 * \left(\delta + t_{\text{соед}} + t_{\text{пер}}\right)}}{2 * \left(\delta + t_{\text{соед}} + t_{\text{пер}}\right)} \right\rceil. \quad (5)$$

На основе разработанной математической модели разработаны и проанализированы следующие архитектурные модели:

– классическая модель клиент-сервер;

– модель клиент-сервер с отдельным сервером для базы данных;

– модель клиент-сервер с отдельным сервером для базы данных и двумя серверами обработки данных;

– модель клиент-сервер с отдельным сервером для базы данных и N серверами обработки данных.

Показано, что использование классической модели клиент-сервер позволяет уменьшить количество серверов, но время выполнения операции над данными довольно большое. По мере увеличения количества серверов время выполнения операции уменьшается, однако затраты на покупку и поддержку серверов возрастают.

Область удалённого управления данными является обширной областью. Сюда входят как задачи выборки, изменения данных, так и более сложные операции с данными. Однако использование распределённой системы имеет смысл лишь для довольно сложных алгоритмов и операций, предназначенных для работы с большими объёмами данных, поскольку простейшие операции с небольшим количеством данных дадут небольшую разницу времени между распределённой и нераспределённой обработкой данных и, зачастую, не будут иметь экономической целесообразности. Поэтому разработаны алгоритм массовой вставки и обновления и алгоритм массового удаления, которые используются для вставки, обновления и удаления огромных объёмов данных.

Разработанные алгоритмы массовой вставки и обновления и массового удаления отличаются от уже существующих тем, что они ориентированы на огромные объёмы данных и использование распределённой системы, а также тем, что они легко реализуются на разработанных архитектурных моделях.

Алгоритм массовой вставки и обновления данных подразумевает выполнение соответствующих операций с данными на основе данных в Excel файлах. Входными данными для алгоритма является массив Excel файлов, выходными данными – результат выполнения операций вставки и обновления записей в базе данных, установленной на сервере.

Алгоритм массового удаления данных подразумевает удаление всех данных, записанных в Excel файлах. Входными данными для алгоритма является массив Excel файлов, выходными данными – результат выполнения операции удаления записей в базе данных, установленной на сервере.

В **третьей главе** произведено экспериментальное исследование разработанных моделей и алгоритмов организации удалённого управления данными в распределённых системах.

Разработана обобщённая модель предметной области удалённого управления данными, общий вид которой представлен на рисунке 3.

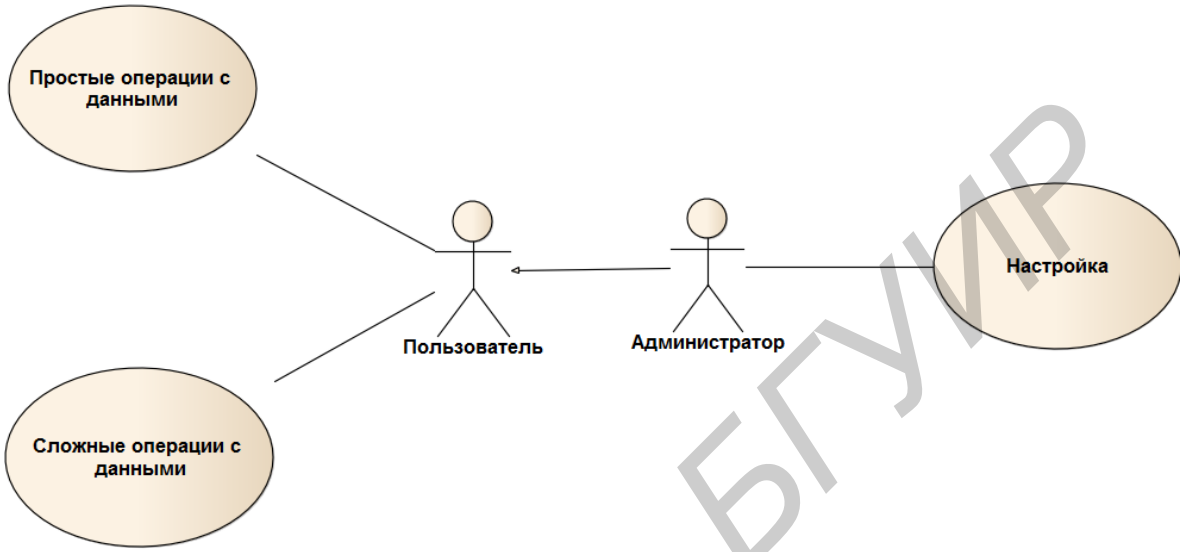

Рисунок 3 – Обобщённая модель предметной области удалённого управления данными

Как видно из рисунка, в данной предметной области используются 2 актёра: пользователь и администратор. У пользователя все операции разделены на простые и сложные операции с данными. К сложным операциям относятся операции массовой вставки и обновления и массового удаления, алгоритмы для которых разработаны во второй главе.

Разработана укрупнённая архитектурная модель программного средства, реализующего разработанные во второй главе модели и алгоритмы. Показано, что в модели архитектуры, использующей распределённую обработку данных, можно выделить такие составные части, как сервер базы данных, центральный сервер, сервера-обработчики и клиентскую часть.

Наиболее подходящим алгоритмом для экспериментального исследования различных архитектурных моделей распределённых систем является разработанный алгоритм массовой вставки и обновления данных. В качестве исходных файлов с данными использовано 6 Excel файлов, каждый из которых имеет 50000 рядов и 10 колонок, из которых одна является первичным ключом, 6 колонок содержат строковые данные, остальные 3 колонки – числовые данные.

В таблице 1 представлены результаты исследования времени обработки данных в архитектурной модели клиент-сервер с отдельным сервером для базы данных и различным количеством серверов обработки данных

Таблица 1 – Результаты исследования времени обработки данных

Количество серверов обработки данных	Количество импортируемых файлов	Общее время выполнения операции, вычисленное эмпирическим путём, с.	Общее время выполнения операции, вычисленное на основе математической модели
1 (нераспределённая обработка)	6	65.5	-
2	6	47.2	46.5
3	6	39.6	38.9
4	6	35.0	33.8
6	6	32.1	30.5
2	8	60.4	60.5
4	8	45.9	45.1

Экспериментальное исследование архитектурных моделей показало:

– разработанная математическая модель соответствует результатам, полученным практическим путём, погрешность составляет не более 6-7%;

– модель, использующая распределённую обработку, позволяет значительно ускорить выполнение операции по сравнению с моделями, не предусматривающими распределённую обработку данных;

– использование или неиспользование распределённой обработки данных зависит в первую очередь от алгоритма: если алгоритм распараллеливается очень хорошо, то имеет смысл использовать распределённую обработку данных, если же алгоритм невозможно распараллелить – то нет никакого смысла использовать распределённую обработку данных.

Экспериментальное исследование разработанных алгоритмов показало:

– разработанные алгоритмы массовой вставки и обновления и массового удаления позволяют значительно быстрее выполнить операции над данными, чем простое циклическое выполнение соответствующих операций;

– в алгоритме массовой вставки и обновления за один вызов хранимой процедуры следует передавать около 2200 рядов с данными, передача меньшего или большего количества может привести к значительным временным задержкам;

– в алгоритме массового удаления за один SQL запрос следует удалять около 700 рядов данных, передача меньшего или большего количества может привести к значительным временным задержкам.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Разработана математическая модель времени выполнения операции в распределённой системе, позволяющая для конкретного алгоритма определить необходимость использования распределённой системы, рассчитать минимальное и максимальное количество серверов, при которых целесообразно использовать распределённую систему, рассчитать минимальное время, которое можно получить при использовании распределённой обработки данных.

2. Разработаны и проанализированы различные архитектурные модели распределённой системы, отличающиеся тем, что они изначально проектировались для предметной области удалённого управления данными, что позволяет увеличить производительность некоторых и трудоёмких операций над данными.

3. Разработаны алгоритмы удалённого управления данными в распределённой системе: алгоритм массовой вставки и обновления данных и алгоритм массового удаления данных, отличающиеся тем, что они предназначены для обработки больших массивов данных в распределённой системе, а также тем, что они предоставляют дополнительные функциональные возможности.

4. Предложены модели архитектуры программного средства удалённого управления данными, реализующего разработанные модели и алгоритмы.

5. Экспериментально проверено, что разработанные алгоритмы массовой вставки и обновления и массового удаления работают быстрее, чем простое циклическое выполнение соответствующих операций. Использование распределённой системы для данных алгоритмов позволяет выполнить соответствующие операции ещё быстрее. Также результаты экспериментальные исследования подтвердили, что разработанная математическая модель позволяет с небольшой погрешностью получить объективное представление о целесообразности использования распределённой системы для того или иного алгоритма.

Рекомендации по практическому использованию результатов

1. Полученные результаты формируют теоретическую и практическую базу для разработки программных средств, оперирующих с данными. Они могут быть использованы для модернизации и дальнейшего развития существующих систем, ориентированных на работу с большими объёмами данных.

2. Разработанная математическая модель может применяться для принятия решения о целесообразности использования распределённой системы в том или ином алгоритме.

3. Результаты работы могут использоваться при подготовке персонала для разработки и обслуживания компьютерных систем, решающих задачи администрирования распределённых систем.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Пацовский, А. А Организация удалённого управления данными в распределённой обучающей системе / А. А. Пацовский, Л. А. Глухова // Высшее техническое образование: проблемы и пути развития: материалы VII Международной научно-методической конференции (Минск, 20-21 ноября 2014). – Минск: БГУИР, 2014. – С. 203-204.
2. Пацовский, А. А Способы организации удалённого управления данными в распределённой системе / А. А. Пацовский // 51-я научная конференция аспирантов, магистрантов и студентов по направлению 4: Компьютерные системы и сети: материалы конф. (Минск, 13-17 апреля 2015). – Минск: БГУИР, 2015. – С. 53.
3. Пацовский, А. А. Организация удалённого управления данными в распределённой системе / А. А. Пацовский, Л. А. Глухова // Новые математические методы и компьютерные технологии в проектировании, производстве и научных исследованиях: материалы XVIII Республиканской научной конференции студентов и аспирантов, Гомель, 23--25 марта 2015 г.: в 2 ч. - Гомель: ГГУ им. Ф.Скорины, 2015. – Ч. 2. – С. 167 – 168.
4. Пацовский, А. А. Архитектурные модели удалённого управления данными в распределённой обучающей системе / А. А. Пацовский, Л. А. Глухова // Дистанционное обучение – образовательная среда XXI века: материалы IX Международной научно-методической конференции (Минск, 3-4 декабря 2015). – Минск: БГУИР, 2014. – С. 277.