УДК 004

О. Л. Жижимов [1], **А. А. Лобыкин** [1], **И. Ю. Турчановский** [1]
А. А. Паньшин [2], **С. А. Чудинов** [2]

[1] Институт вычислительных технологий СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

[2] Институт сильноточной электроники СО РАН
пр. Академический, 2/3, Томск, 634055, Россия

E-mail: zhizhim@mail.ru

АВТОМАТИЗИРОВАННАЯ СИСТЕМА СБОРА СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ О СОБЫТИЯХ В РАСПРЕДЕЛЕННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЕ

Описана архитектура автоматизированной системы сбора информации о событиях, происходящих в распределенной информационной системе, на примере системы массовой интеграции ZooSPACE. Архитектура системы разработана на модульном принципе. Разработаны алгоритмы анализа входного информационного потока, структуры и способы хранения данных, способы статистического анализа и представления статистической информации о событиях, происходящих в распределенной информационной системе ZooSPACE.

Ключевые слова: распределенные информационные системы, интеграция гетерогенных данных, Z39.50, LDAP, SRW/SRU, ZooSPACE, сбор статистической информации.

Введение

Современной научной задачей является разработка подходов решения проблем по глобальной интеграции гетерогенных баз данных. Накопленные ресурсы хранятся в разнородных базах данных, форматах и взаимодействуют по разным протоколам передачи информации. Для конечных пользователей существуют проблемы доступа к этой информации. Эта же проблема может осложнять работу как с каталогами, так и с другими научно-информационными ресурсами.

Одним из способов, позволяющим создавать глобальные инфраструктуры, а также решать задачи по интеграции распределенных гетерогенных баз данных, является создание платформы массовой интеграции независимых узлов, функционирующих в соответствии с единой политикой.

Настоящая работа является продолжением цикла статей, посвященных платформе массовой интеграции ZooSPACE [1], и содержит описание организации одной из ее подсистем – подсистемы сбора, формирования и отображения статистической информации о событиях, происходящих в распределенной информационной системе. Важность статистической информации определяется основными требованиями, предъявляемыми к надежной распределенной информационной системе с точки зрения ее управляемости и обеспечения информационной безопасности [1].

Следует заметить, что задача сбора статистической информации в распределенной гетерогенной системе не является тривиальной. Решение этой задачи тесно связано как с архитек-

Жижимов О. Л., Лобыкин А. А., Турчановский И. Ю., Паньшин А. А., Чудинов С. А. Автоматизированная система сбора статистической информации о событиях в распределенной информационной системе // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 1. С. 42–52.

турой системы в целом, так и с ее инфраструктурой в конкретной реализации. Поэтому мы будем иллюстрировать модель сбора статистической информации ее реализацией для платформы ZooSPACE.

Основным протоколом доступа к распределенным информационным ресурсам в ZooSPACE является Z39.50 [1]. В основе Z39.50 лежит идея построения абстрактной модели работы с абстрактной базой данных. Работа с каждой конкретной СУБД согласно Z39.50 должна быть организована только через эту абстрактную модель путем обмена пакетами данных (APDU), содержащими последовательности идентифицируемых по меткам объектов. Все события, которые интересны для анализа эффективности функционирования распределенной информационной системы, требуют протоколирования в формате, адекватно отображающем содержание соответствующих APDU.

При разработке модулей ZooSTAT-S для сбора и статистической обработки информации [2] используется модульная структура с открытым кодом.

В работе рассмотрены способы сбора и обработки информации о состоянии серверов ZooSPACE с целью разработки универсального инструмента сбора статистики, мониторинга запросов / сессий и формирования отчетов за интересуемый период времени. Данный инструмент позволяет администраторам группировать собранную информацию о состоянии распределенных серверов, осуществлять мониторинг и проводить статистический анализ географически распределенных систем.

Обрабатываемые события отображаются в виде различных графиков и таблиц в зависимости от количества сессий, частоты вызова, того или иного запроса во временных интервалах. Кроме перечисленных особенностей, система позволит создавать и формировать зависимости по желанию и усмотрению пользователя.

Основные требования

Система ZooSPACE-S должна использовать программное обеспечение OpenSource и функционировать под управлением операционных системы Linux, FreeBSD. ZooSPACE-S должна выполнять сбор информации с серверов ZooSPACE в формате APDU, обрабатывать ее и выводить результирующие статистики для администраторов системы ZooSPACE-S. Кроме того, для быстрого восстановления системы в случае сбоев и / или отказа сервера выполнять резервное копирование всей чувствительной информации.

Архитектура системы ZooSPACE-S

Система ZooSPACE-S разработана на основе модульной структуры и содержит следующие модули (рис. 1):

- модуль сбора информации о событиях (исходных данных) с серверов ZooSPACE;
- модуль генератора статистик;
- модуль подготовки шаблонов для генератора статистик;
- модуль вывода обработанной статистической информации через web-интерфейс;
- модуль резервного копирования.

Общие характеристики системы ZooSPACE-S

Система ZooSPACE-S разрабатывалась и в текущий момент функционирует на серверной платформе (с процессором Intel Core i5, оперативной памятью 8 Гб, жесткий диск объемом 500 Гб) под управлением FreeBSD 9.0. Ведутся работы по портированию и адаптации для Linux. Система ZooSPACE-S использует программное обеспечение OpenSource:

- Web-сервер Apache 2.2 – для доступа к административной и пользовательской частям системы ZooSPACE-S через Интернет;

[1] ANSI/NISO Z39.50-2003. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Bethesda, Maryland: NISO Press, 2002.

- OpenLDAP 2.4 – для хранения конфигурационных параметров и учетных записей пользователей;
- PHP 5.4 with suhosin-patch и установленными библиотеками bz2, zlib, dom, xml, filter, gd, iconv, json, mbstring, mcrypt, pdo, session, sockets, ldap – для административного управления модулями системы ZooSPACE-S и вывода статистики пользователю через Интернет;
- MySQL 5.5 или PostgreSQL 9.1 – для хранения обработанных данных, промежуточного представления статистической информации, конфигурации системы;
- RRDTool 1.4.7 – Round Robin база данных для хранения обработанных статистических данных;
- Perl 5.14 с установленными классами dbi, ldap, z3950 – для загрузки исходных данных функционирования с удаленных серверов ZooSPACE, их обработки и сохранении в реляционной базе данных;
- jQuery >1.5.1 – для функциональной обработки информации, передаваемой от Web-сервера ZooSPACE-S (AJAX) и ее наглядного представления на системе конечного пользователя;
- Flot 0.7 – библиотека JavaScrip (AJAX), предназначенная для вывода графической информации.

Сбор информации

Модуль сбора информации о событиях с серверов ZooSPACE последовательно опрашивает сервера, указанные в конфигурационном списке, по протоколу Z39.50/SRW,

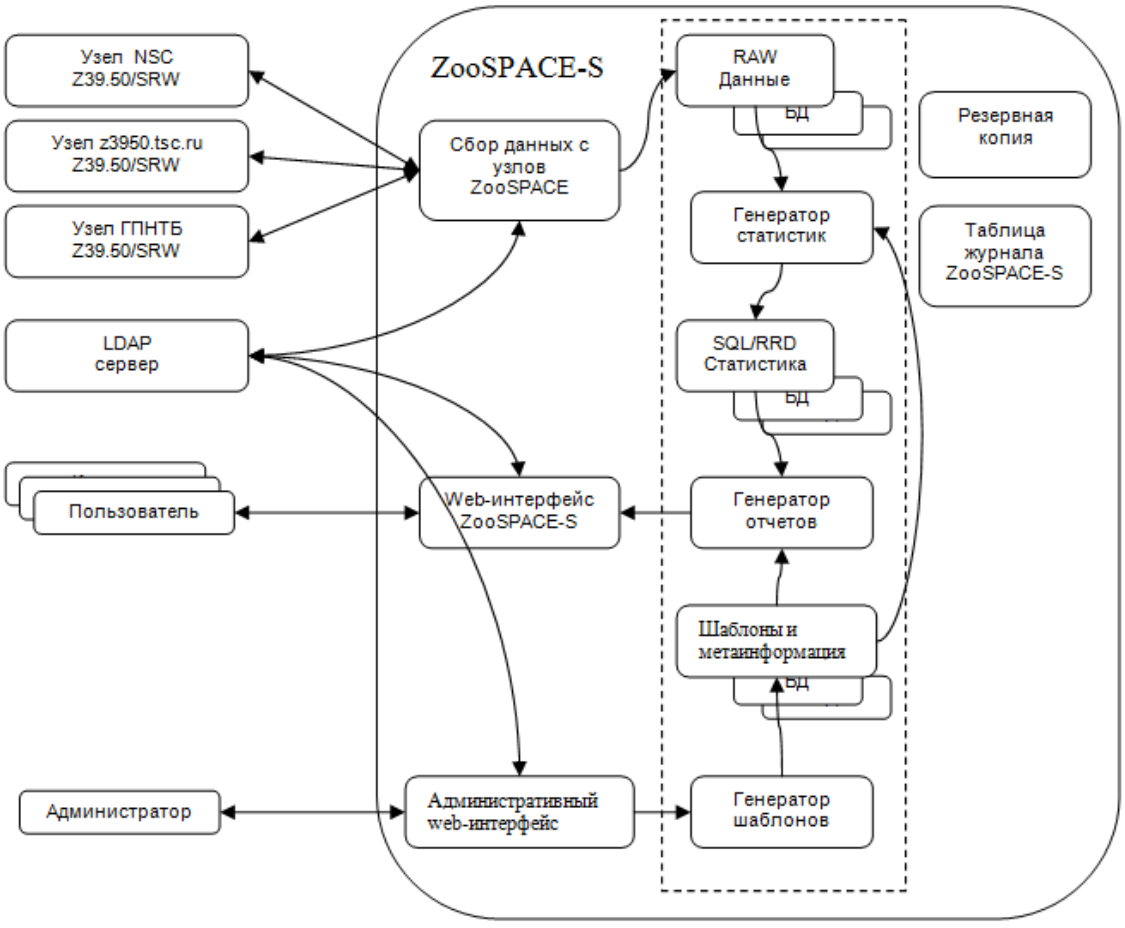

Рис. 1. Архитектура системы сбора статистики ZooSPACE-S

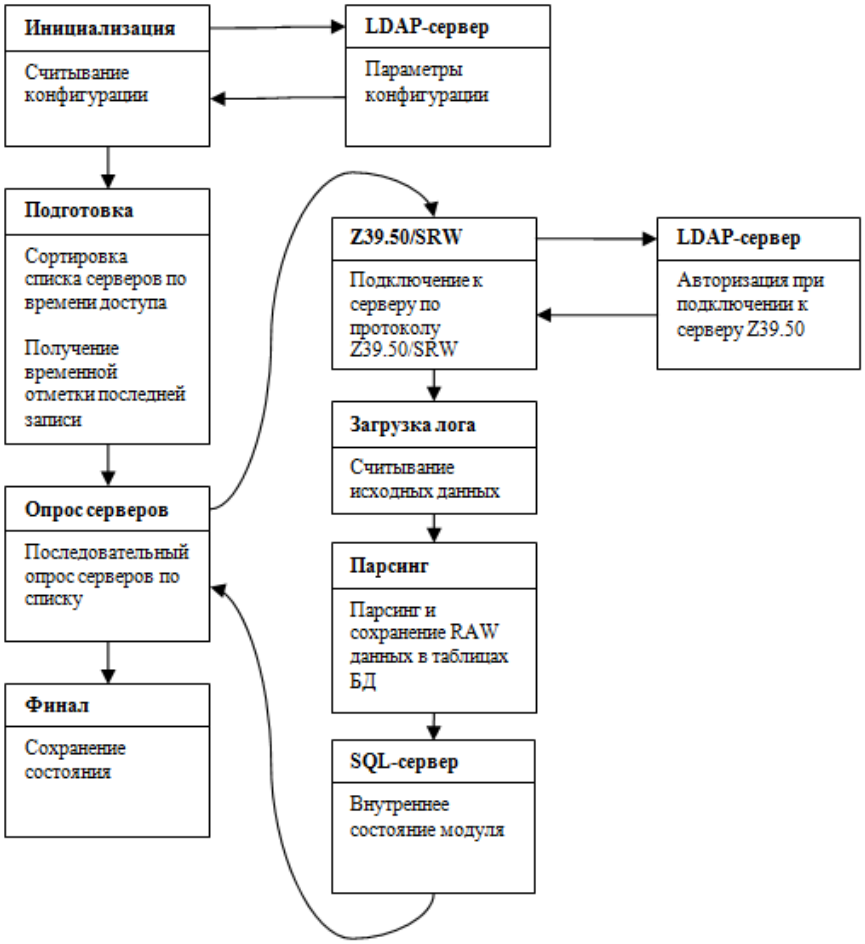

Рис. 2. Функционирование модуля сбора информации

выгружает с них исходные данные в виде пакетов APDU и выполняет их синтаксический анализ. Результат разбора сохраняется в таблицах реляционной базы данных (MySQL или PostgeSQL) как промежуточное представление информации, структура которых позволяет выполнять быстрое построение отчетов.

Модуль сбора информации о событиях на серверах ZooSPACE (рис. 2) состоит из следующих блоков:

- блок считывания конфигурационной информации из LDAP-каталога или файла конфигурации в случае, если LDAP-сервер не доступен;
- блок авторизации и аутентификации при доступе к серверам ZooSPACE;
- блок загрузки исходных данных с серверов ZooSPACE по протоколу Z39.50/SRW;
- блок обработки данных и сохранения обработанных данных в таблицах реляционной БД.

Конфигурационные параметры модуля хранятся в LDAP-каталоге, где указываются список IP-адресов опрашиваемых серверов ZooSPACE, параметры аутентификации для доступа к серверам, периодичность запуска модуля опроса. При отсутствии доступа к LDAP-каталогу используются параметры, заданные в текстовом файле конфигурации [3].

Модуль хранит переменные внутреннего состояния в таблице реляционной БД. При инициализации модуля из этой таблицы считываются для каждого из серверов: время последней обработанной записи, время доступа к серверу, количество ошибок при передаче и т. п. Список сортируется по времени доступа к серверу.

Далее, для каждого сервера ZooSPACE из списка последовательно выполняются следующие процедуры:

1) считывание из LDAP-каталога параметров аутентификации для доступа к серверу Z39.50;
2) подключение к серверу по протоколу Z39.50/SRW;
3) считывание исходных данных с узла Z39.50 с определенного момента времени;
4) предварительная обработка и загрузка исходных данных в таблицы реляционной БД;
5) сохранение внутреннего состояния для текущего сервера в таблице БД.

Указанная последовательность процедур выполняется для каждого сервера.

Генерация статистик событий серверов ZooSPACE

Модуль генератора статистики предназначен для преобразования исходных данных, сохраненных в таблицах сервера БД, в соответствии с правилами, заданными администратором системы, позволяющими выделить существенную информацию о событиях в системе ZooSPACE из потока исходных данных.

Мы опробовали различные структуры для хранения результатов статистической обработки (RRD/SQL/XML). Однако в текущей версии склоняемся к хранению результатов в таблицах реляционной БД. Это связано в первую очередь с более универсальным подходом при формировании отчетов в сочетании с высокой скоростью выполнения сложных запросов к серверу баз данных. Следует отметить, что RRD-базы данных также заслуживают внимания благодаря чрезвычайно высокой скорости работы, но на текущий момент при концептуальной разработке системы ZooSPACE-S желательно иметь универсальный инструмент формирования отчета, поэтому мы использовали SQL-сервер [4].

Модуль генерации статистической информации (рис. 3) загружает из таблиц реляционной БД подготовленные шаблоны с соответствующей им метаинформацией, описывающей

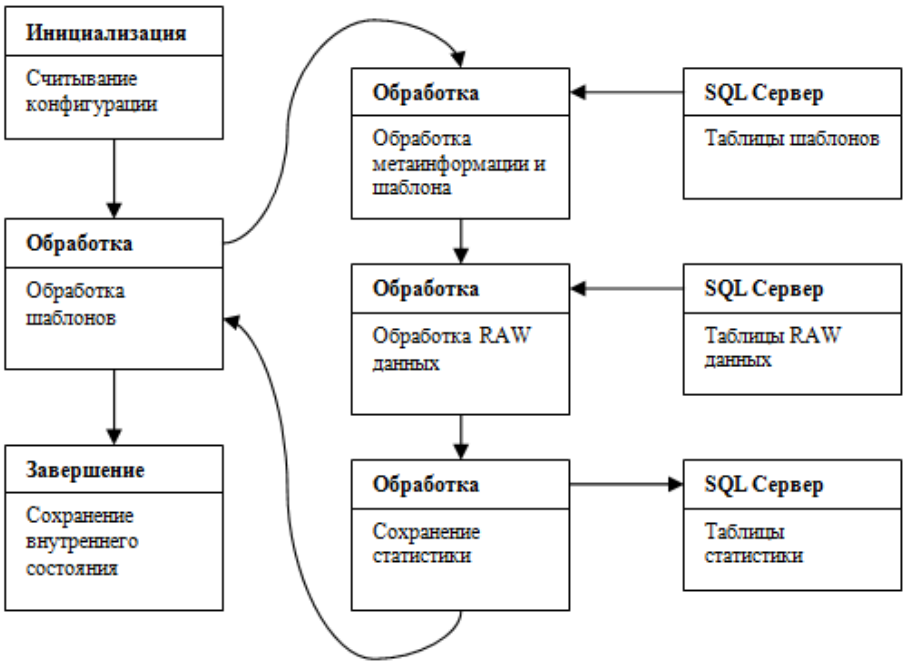

Рис. 3. Функционирование модуля генерации статистической информации

условия применения шаблона, и формируемые таблицы статистик, выполняет обработку исходных данных и генерирует необходимые для визуализации информационные структуры Алгоритм формирования результата статистической обработки универсален и может быть использован для построения широкого класса статистик, например для следующих видов отчетов: количество событий в единицу времени; отсортированный список – частота события от указанного параметра, и т. п.

Подготовка шаблонов генератора статистики событий

Модуль предназначен для подготовки шаблонов и соответствующей им метаинформации описания, проверки их синтаксической и семантической корректности.

Модуль (рис. 4) представляет собой Web-интерфейс администратора, разработанный на языке программирования PHP и состоящий из следующих блоков:

- блок считывания конфигурационной информации из LDAP-каталога или файла конфигурации [3] в случае, если LDAP-сервер не доступен;
- Web-интерфейс для взаимодействия с администратором;
- блок аутентификации;
- блок формирования страницы подготовки шаблона;
- блок проверки корректности шаблона.

Модуль подготовки шаблонов позволяет генерировать статистики 2 основных типов: количество событий в единицу времени, отображаемых в виде графиков; и список событий, зарегистрированных для фиксированного параметра, отсортированный по количеству событий, отображаемых в виде таблицы или круговых диаграмм.

Статистики первого типа могут быть наглядно представлены в виде графиков с временной координатой. Например, в качестве такой статистики можно предложить «число запросов к конкретному серверу в течение последнего месяца», «число ошибочных запросов к серверу в течение дня» и т. п. При формировании таблицы для статистики первого типа указывается логическое выражение, например, (APDU:common: typeAPDU = 1), где число 1 – числовой идентификатор типа APDU – initRequest. Агрегирование по времени выполняется с фиксированной величиной 1 минута. Величина агрегации – 1 минута – может быть изменена в конфигурации. Вероятнее всего, в рабочем проекте для уменьшения размера таблиц эту величину можно установить 10 или 30 минут. Для указанного выше выражения генератор сформирует таблицу реляционной БД, в первом поле которой – отметка времени с шагом в 1 минуту, во втором – число событий, удовлетворяющих указанному выражению за этот промежуток времени [5]. При просмотре страницы статистики пользователем возможна любая агрегация по времени, кратная времени агрегации. Пользователем могут указываться дополнительные условия для генерации статистики (например, указать сервер, базу данных).

Статистики второго типа могут представляться в виде списка. Примером может служить таблица «Количество обращений к серверу с определенного IP за последний месяц», отсортированная по количеству обращений. Вывод информации для конечного пользователя может быть либо в виде списка / таблицы, например «Топ 50», либо в виде круговых диаграмм.

Сгенерированный шаблон вместе с метаинформацией, описывающей условия применения и использования статистики, сохраняется в таблице БД [6].

Конфигурационные параметры модуля хранятся в LDAP-каталоге. При отсутствии доступа к LDAP-каталогу используются параметры, заданные в текстовом файле конфигурации [3]. Для получения доступа к Web-странице сервиса требуется авторизация пользователя. Проверка пароля и прав доступа выполняется через запрос к LDAP-каталогу. После успешного входа в систему администратору предоставляется возможность просмотреть уже имеющийся список шаблонов, создать новый или модифицировать уже имеющийся шаблон.

Пример страницы подготовки нового шаблона «Число запросов ко всем серверам в единицу времени» показан на рис. 5.

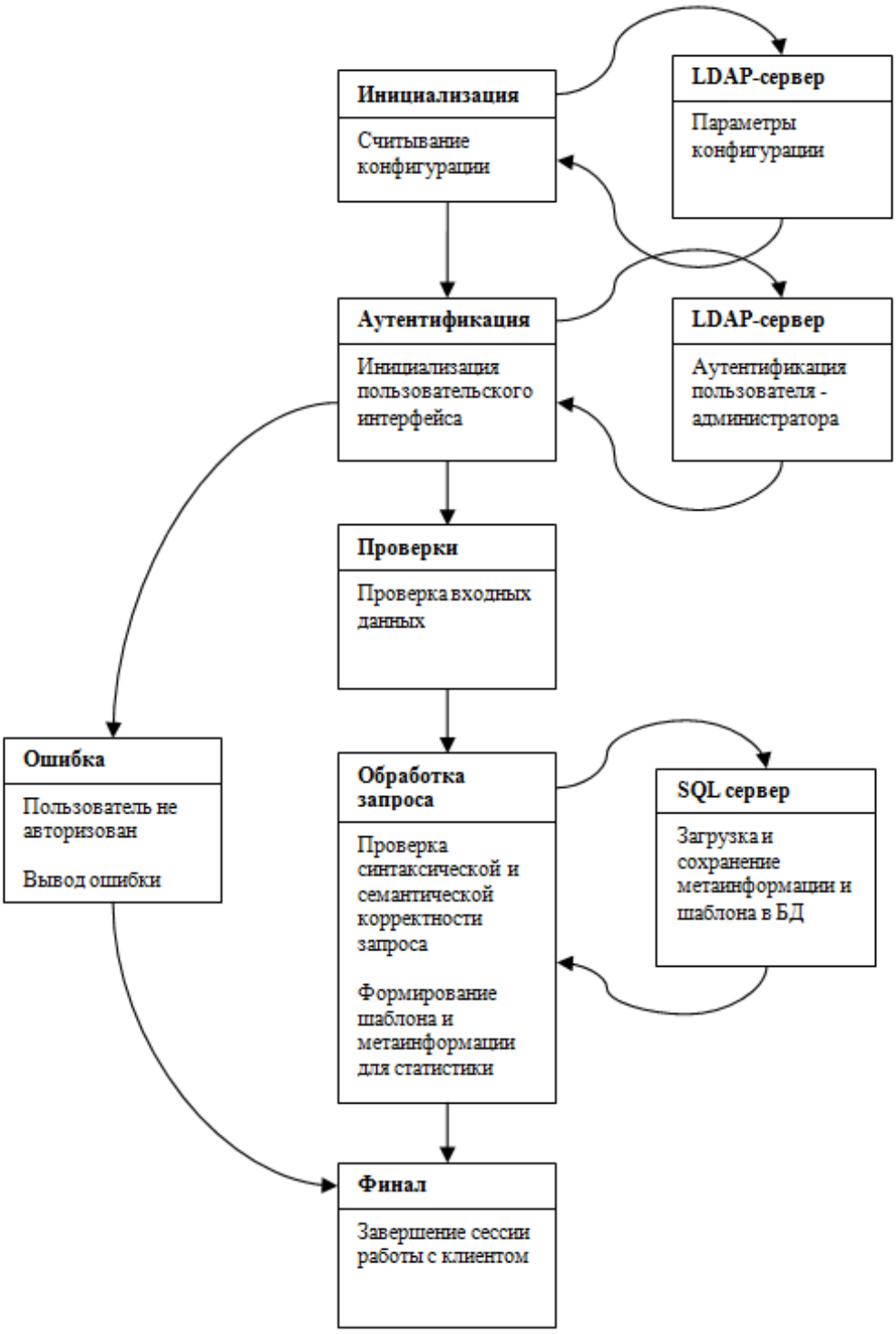

Рис. 4. Функционирование модуля подготовки шаблонов для генератора статистик событий на серверах ZooSPACE

Вывод статистической информации

Модуль вывода предназначен для предоставления Web-доступа пользователям системы ZooSPACE-S к статистической информации. Позволяет в удобной, доступной форме получить информацию о событиях в системе ZooSPACE.

Модуль состоит из следующих блоков (рис. 6):

- блок считывания конфигурационной информации из LDAP-каталога или файла конфигурации в случае, если LDAP-сервер не доступен;
- Web-сервер для взаимодействия с пользователем;
- блок аутентификации;
- блок формирования страницы статистики.

Конфигурационные параметры модуля хранятся в LDAP-каталоге. При отсутствии доступа к LDAP- серверу используются параметры, заданные в текстовом файле конфигурации [3]. Для получения доступа к страницам сервиса требуется авторизация пользователя. Проверка пароля и прав доступа выполняется через запрос к LDAP-каталогу.

После успешного входа в систему пользователю предоставляется возможность посмотреть накопленные статистические данные по любому из серверов ZooSPACE, например, количество запросов к серверу в единицу времени, количество удачных / неудачных запросов в единицу времени, список наиболее активных клиентов, минимальное, среднее, максимальное время сессии и др. Однако список возможных статистик определяется администратором системы, который сформировал соответствующие шаблоны для модуля генератора отчетов и может расширяться.

Генерация страницы, с запрошенной пользователем статистикой, выполняется на основе метаинформации, описывающей шаблон, условия применимости и имени таблицы, содержащей подготовленные для вывода данные.

При формировании Web-страницы из БД считывается метаинформация и шаблон запроса к реляционной БД, выполняются необходимые подстановки и проверки, после чего формируется запрос к сгенерированной заранее таблице, содержащей статистические данные [6]. Результат выводится пользователю в виде графика, круговой диаграммы или списка. На рис. 7 приведен пример генерируемой модулем вывода статистики «Число запросов к серверу в минуту».

Резервное копирование

Модуль резервного копирования (рис. 8) предназначен для сохранения дампов таблиц баз данных, конфигураций модулей ZooSPACE-S, шаблонов статистики и исходных данных.

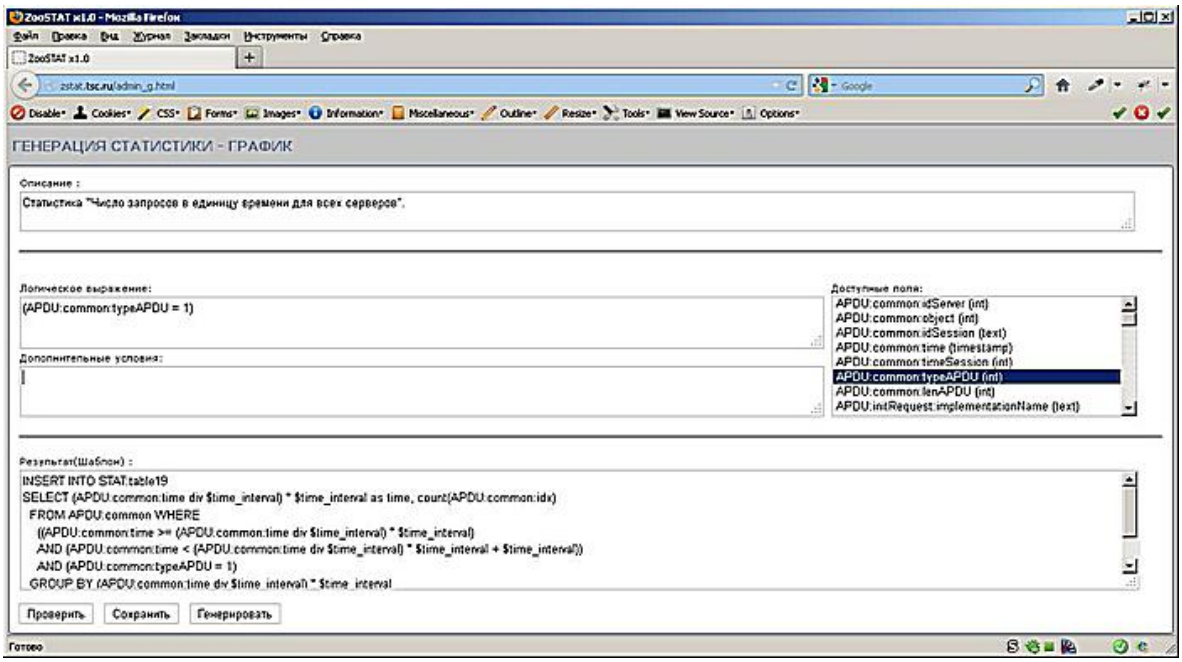

Рис. 5. Пример формирования шаблона для статистики «Число запросов ко всем серверам в единицу времени»

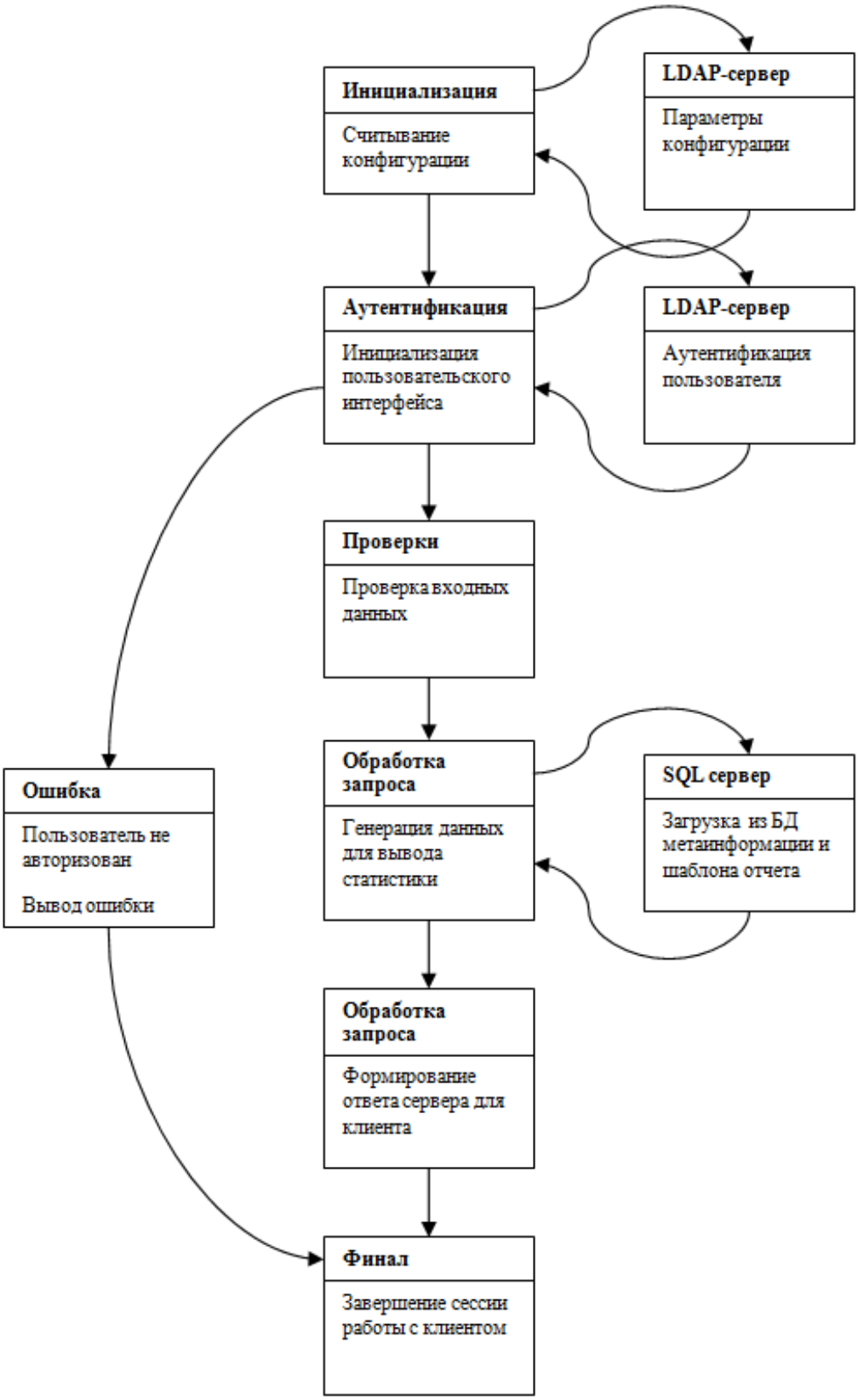

Рис. 6. Функционирование модуля вывода статистической информации

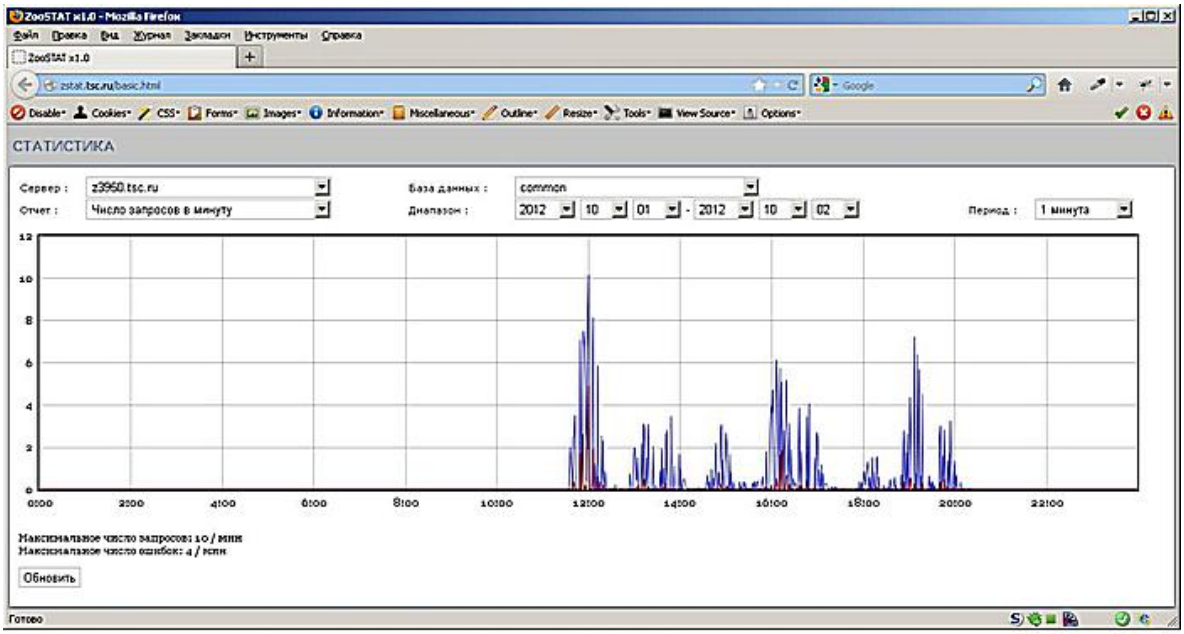

Рис. 7. Пример вывода статистической информации «Число запросов в минуту» работы тестового сервера z3950.tsc.ru в течение суток. Синим цветом обозначено число запросов к серверу в единицу времени, красным – число запросов с ошибками в единицу времени. Период агрегации – 1 минута

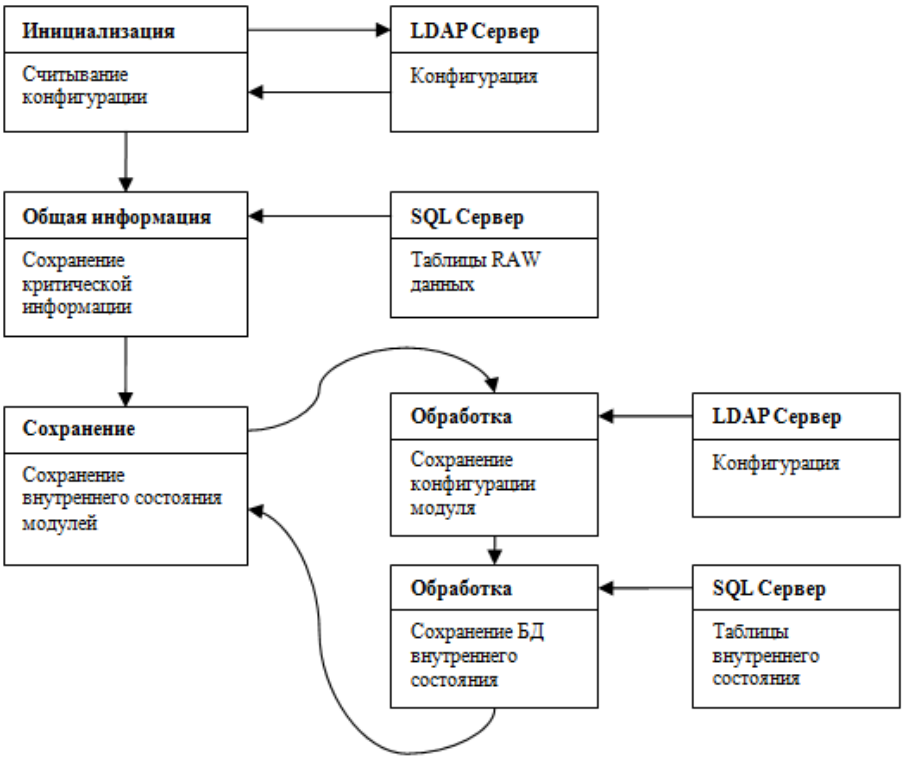

Рис. 8. Алгоритм функционирования модуля резервного копирования

Информация, сохраняемая в резервной копии, позволяет полностью восстановить работоспособность системы при возможных сбоях оборудования.

Модуль резервного копирования запускается с заданной периодичностью и выполняет сохранение дампов таблиц базы данных с конфигурацией модулей, шаблонов статистики и таблиц исходных данных. Таблицы базы данных, содержащие результаты статистической обработки, не сохраняются в резервной копии. После восстановления конфигураций модулей и шаблонов из резервной копии модуль генератора статистики самостоятельно формирует все необходимые таблицы статистик. Периодичность сохранения и список сохраняемых таблиц указываются в LDAP-каталоге и конфигурационном файле модуля. Для уменьшения объема занимаемого резервной копией дискового пространства используется системный вызов утилиты bzip2.

Заключение

Описана система сбора и статистической обработки состояний серверов распределенной информационной системы ZooSPACE, работа которых ориентирована на использование протокола Z39.50. Предлагаемая архитектура и реализация могут быть использованы и для других информационных систем и платформ при минимальной модификации для привязки к конкретным условиям. Мы надеемся, что описанная система ZooSPACE-S послужит эффективным инструментом для анализа и выявления проблемных элементов распределенных информационных систем как в ZooSPACE, так и в других случаях.

Список литературы

1. *Ревнивых А. В., Федотов А. М.* Обзор политик информационной безопасности // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2012. Т. 10, № 3. С. 66–79.

2. *Цимбал А. А., Аншина М. Л.* Технологии создания распределенных систем. Для профессионалов. СПб.: Питер, 2003. 576 с.

3. *Федотов А. М., Шокин Ю. И., Жижимов О. Л., Молородов Ю. И.* Служба директорий LDAP как единая информационная среда // Открытое и дистанционное образование. 2007. № 4 (28). С. 31–41.

4. *Когаловский М. Р.* Энциклопедия технологий баз данных. М.: Финансы и статистика, 2002. 800 с.

5. *Жижимов О. Л., Федотов А. М., Юданов Ф. Н.* Модель управления информационными ресурсами организации // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2010. Т. 8, № 4. С. 81–95.

6. *Мазов Н. А., Жижимов О. Л.* Метаданные и их роль в распределенных информационных системах на основе использования протокола Z39.50: Лекция // Библиосфера. 2006. № 2. С. 51–60.

Материал поступил в редколлегию 27.02.2013

O. L. Zhizhimov, A. A. Lobykin, I. Yu. Turchanovsky, A. A. Panshin, S. A. Chudinov

AN AUTOMATED SYSTEM FOR THE COLLECTION OF STATISTICAL INFORMATION ABOUT EVENTS IN A DISTRIBUTED INFORMATION SYSTEM

The article describes the architecture of an automated system for collecting information about the events in a distributed information system for the example of media integration ZooSPACE. The system architecture is designed on a modular principle. The algorithms analyze the input data flow, structure and storage technologies, methods of statistical analysis and presentation of statistical information about events in a distributed information system ZooSPACE.

Keywords: statistics gathering, state information gathering, distributed information systems, Z39.50, LDAP, SRW/SRU, ZooSPACE.