DOI 10.37882/2223–2966.2021.10.23

РАЗРАБОТКА ПРОЕКТА СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ДЛЯ ПРОВАЙДЕРА СЕТЕВЫХ УСЛУГ НА БАЗЕ ВЫЧИСЛИТЕЛЬНЫХ КЛАСТЕРОВ

DEVELOPMENT OF A DECISION SUPPORT SYSTEM PROJECT FOR A NETWORK SERVICE PROVIDER BASED ON COMPUTING CLUSTERS

Yu. Sagalaev
A. Sagalaeva
O. Romashkova

Summary. This article discusses the problems of evaluating and analyzing control solutions for expanding and retaining an active audience of network service providers based on the use of computing clusters and decision support systems. The key aspects are identified and the relevance of the use of data mining, information collection and processing for decision support tasks based on the use of scalable and distributed computing cluster architectures is justified. A conceptual modular scheme and diagrams in UML notation for formalizing the system project are developed, further ways of developing the proposed approach to data processing and analysis are described.

Keywords: decision support systems, data mining, computing clusters, targeting.

Сагалаев Юрий Романович
Аспирант, ГАОУ ВО «Московский городской педагогический университет (МГПУ)» г. Москва
yrok472@mail.ru
Сагалаева Анна Игоревна
Аспирант, ГАОУ ВО «Московский городской педагогический университет (МГПУ)» г. Москва
omegaanya@gmail.com
Ромашкова Оксана Николаевна
Д.т.н., профессор, Российская академия народного хозяйства и государственной службы при Президенте РФ (РАНХиГС)», г. Москва
ox-rom@yandex.ru

Аннотация. В данной статье рассмотрена проблематика оценки и анализа управляющих решений для расширения и удержания активной аудитории провайдеров сетевых услуг на основе использования вычислительных кластеров и систем поддержки принятия решений. Обозначены ключевые аспекты и обоснована актуальность применения интеллектуального анализа данных, сбора и обработки информации для задач поддержки принятия решений на основе использования масштабируемых и распределенных архитектур вычислительных кластеров. Разработаны концептуальная модульная схема и диаграммы в нотации UML для формализации проекта системы, описаны дальнейшие пути развития предложенного подхода к обработке и анализу данных.

Ключевые слова: системы поддержки принятия решений, интеллектуальный анализ данных, вычислительные кластера, таргетирование.

Введение

В настоящее время в различных отраслях бизнеса наблюдается устойчивый рост объемов разнородной информации, представляющей определенную ценность с точки зрения принятия управленческих решений по развитию и расширению деятельности организаций [1]. Это особенно актуально для сферы оказания услуг, связанных с обеспечением корпоративного или личного доступа в сеть Интернет, мобильной связи и других телекоммуникационных сервисов для конечных потребителей. Возникает необходимость анализа собранных статистических данных о предпочтениях клиентов, порядке их поведения в сети, цифровом следе, характере оплат, объемах и сезонности потребляемого трафика, что необходимо для сегментации и выделения целевых групп пользователей [2]. Конечной целью обработки и анализа данных в обозначенной прикладной сфере может стать выявление новых скрытых и не очевидных знаний, полезных для разработки специализированных и персонифицированных акционных предложений, новых тарифов, программ лояльности по удержанию целевой аудитории или ее расширению.

Значительные трудности в организации эффективных процессов анализа данных в данной отрасли бизнеса заключаются в необходимости оперативной обработки больших объемов не структурированной или слабоструктурированной информации (Big Data), а также в производительных системах сбора данных и их предварительной предобработке и очистке, преобразованию к единому унифицированному формату [3]. Обозначенные процессы являются причиной воз-

никновения высокой вычислительной нагрузки на аппаратное обеспечение, что часто требует значительных финансовых затрат на приобретение, настройку, конфигурацию и развертывание серверных решений или оплату тарифов использования облачных решений и сервисов. Последние являются удобным средством для виртуализации серверов с целью обеспечения масштабирования и надежной работы систем обработки и анализа данных [4].

Хранение собранных данных из разных источников является особо критичной задачей в контексте решения перечисленных выше задач. По причине существенного увеличения длины SQL команд на агрегированные операции поиска, выборки, вставки и сохранения данных в реляционных системах управления базами данных (СУБД), возрастает сложность формирования транзакционных запросов к базам данных (БД) [5]. В зависимости от используемых аппаратных ресурсов это может существенно снижать быстродействие и эффективность обработки данных в режиме реального времени. Поэтому, в качестве хранилищ Big Data целесообразным на практике является использование не реляционных БД (NoSQL), таких как MongoDB, Redis, HBase, Firebase и др. Данные СУБД позволяют динамически расширять структуру БД без необходимости внесения существенных изменений и имеющиеся представления данных. Преимущества использования NoSQL в рамках обеспечения процесса обработки Big Data обусловлено такими факторами, как: высокий уровень гибкости в обеспечении нужного уровня масштабирования данных; отсутствие ограничений на хранимые типы данных; поддержка удобной для обработки модели представления данных «ключ-значение»; наличие документно-ориентированного подхода [6].

Дополнительной проблемой при интеграции как коммерческих, так и открытых облачных решений является организация производительных вычислительных кластеров (ВК), способных обеспечить достаточный уровень надежности и безотказности процессов обработки информации. В настоящее время для обеспечения процессов обработки и анализа Big Data активно применяются решения на основе ВК с использованием модели распределенных вычислений Map Reduce, что позволяет обеспечить параллельную обработку данных и увеличить скорость проведения [7]. Благодаря использованию распределенной файловой системы (HDFS) становиться возможным гибкая организация процессов обработки и свертки данных на управляющий сервер. Преимуществами данного подхода является автоматизация распределения узлов на ВК в рамках компьютерной сети, что позволяет задействовать вычислительные возможности неограниченного числа хостов, имплементация алгоритмов резервирования данных для обеспечения надежности их хранения, а также поддержка программных реализаций для большинства современных языков программирования высокого уровня. Однако, данных подход может приводить к трудоемким процессам обработки потоковых данных ВК в режиме реального времени, сложностям в развертывании системы в случае наличия не стабильного канала передачи данных с низкой пропускной способностью, а также к ресурсоемкости процедур визуализации данных по итерациям в случае наличия большого числа коротких онлайн транзакций [8].

Одним из возможных путей решения обозначенных выше проблем является проектирование, разработка и использование целевых систем поддержки принятия решений (СППР), основанных на распределенной и масштабируемой архитектуре ВК, имплементирующие приоритетные элементы искусственного интеллекта (в том числе методы и модели интеллектуального анализа данных, машинного или глубокого обучения), позволяющих автоматизировать процессы оценки различных альтернатив и сценариев взаимодействия с целевой аудиторией [9,10]. Это позволит существенно упростить процессы анализа лицам, принимающим решения (ЛПР), в компаниях по стратегическому планированию, развитию бизнеса и решению операционных задач [11,12]. Все это обуславливает актуальность и целесообразность данного исследования.

Цель данной статьи

Автоматизация оценки и анализа управляющих решений по целевому сегменту аудитории провайдеров сетевых услуг на базе проекта системы поддержки принятия решений, построенной на распределенной облачной архитектуре вычислительных кластеров. Практическая необходимость в исследовании заключается в удержании существующих и привлечении новых клиентов для компании на базе формирования целевых акций и специальных предложений для целевой аудитории.

Описание концепции

Построение СППР целесообразно выполнять на базе использования микросервисной архитектуры в рамках частного облачного Hadoop OpenStack IaaS решения, что обеспечит мультиплатформенность и стабильность в управлении отдельными функциями с возможностью дальнейшего быстрого масштабирования и расширения. Предлагаемая СППР имплементирует методы машинного обучения (МО), является распределенной, может быть запущена на разном числе виртуальных серверов, находящихся под управлением ВК, что позволяет обеспечить ее отказоустойчивость и масштабирование на любом числе узлов. Клиент-

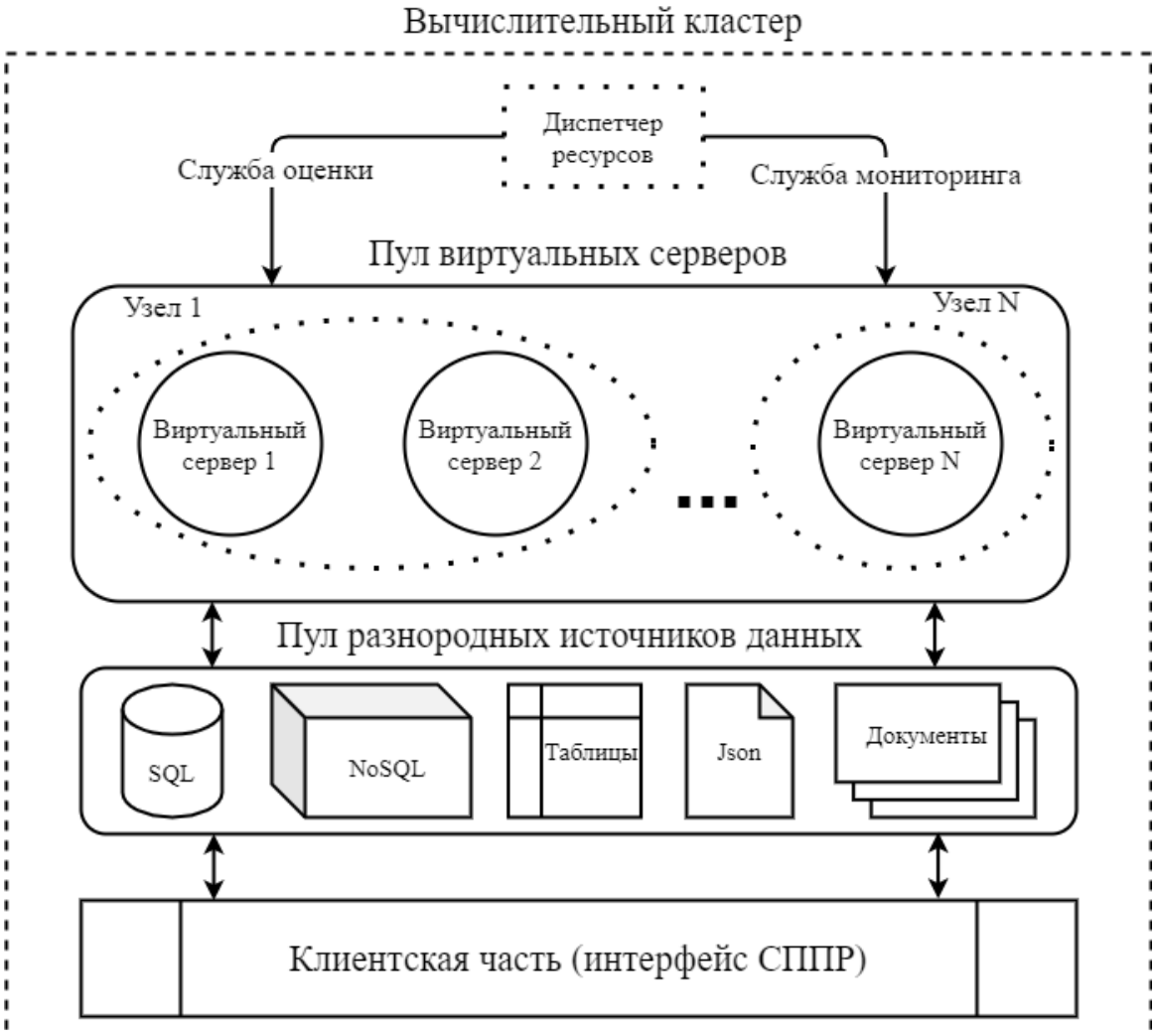

Рис. 1. Схема балансировки нагрузки ВК

ская часть системы может быть представлена в виде веб-приложения, размещенного на удаленном хостинге или на отдельном физическом корпоративном сервере, предоставляющего доступ к данным посредством интерактивного пользовательского интерфейса и визуализацию результатов анализа данных. Программными средствами разработки подобной системы могут быть: язык программирования Python, фреймворк Django и дополнительные библиотеки поддержки процессов обработки и анализа данных Pandas, Sklearn, Pickle, Matplotlib. Гибкость развертывания системы может быть обеспечена благодаря интеграции средств унифицированной контейнеризации, в частности слоев Docker, что позволяет упростить процесс сборки и запуска системы для одновременного использования большим числом пользователей на распределенных узлах ВК. СППР состоит из 5 отдельных модулей, каждый из которых реализует свой функционал по обработке, оценке или анализу данных, в частности:

1. Модуль сбора данных из различных источников. Фактически, реализует операции агрегации из реляционных и нереляционных БД, а также из текстовых файлов. Результатом работы данного модуля является упорядоченный набор json файлов.
2. Модуль предобработки данных. Осуществляет ряд операций по: оценке признаков, статистическому анализу корреляции между ними с целью сокращения размерности, разделению на категориальные, вещественные, целочисленные, устранению пустот и пропущенных значений, удалению выбросов и аномалий. Результатом работы данного модуля является очищенные данные в формате csv.
3. Модуль конфигурации моделей анализа данных. Реализует разбивку данных на обучающие и тестовые множества, создание моделей машинного обучения, их обучение и валидацию результатов

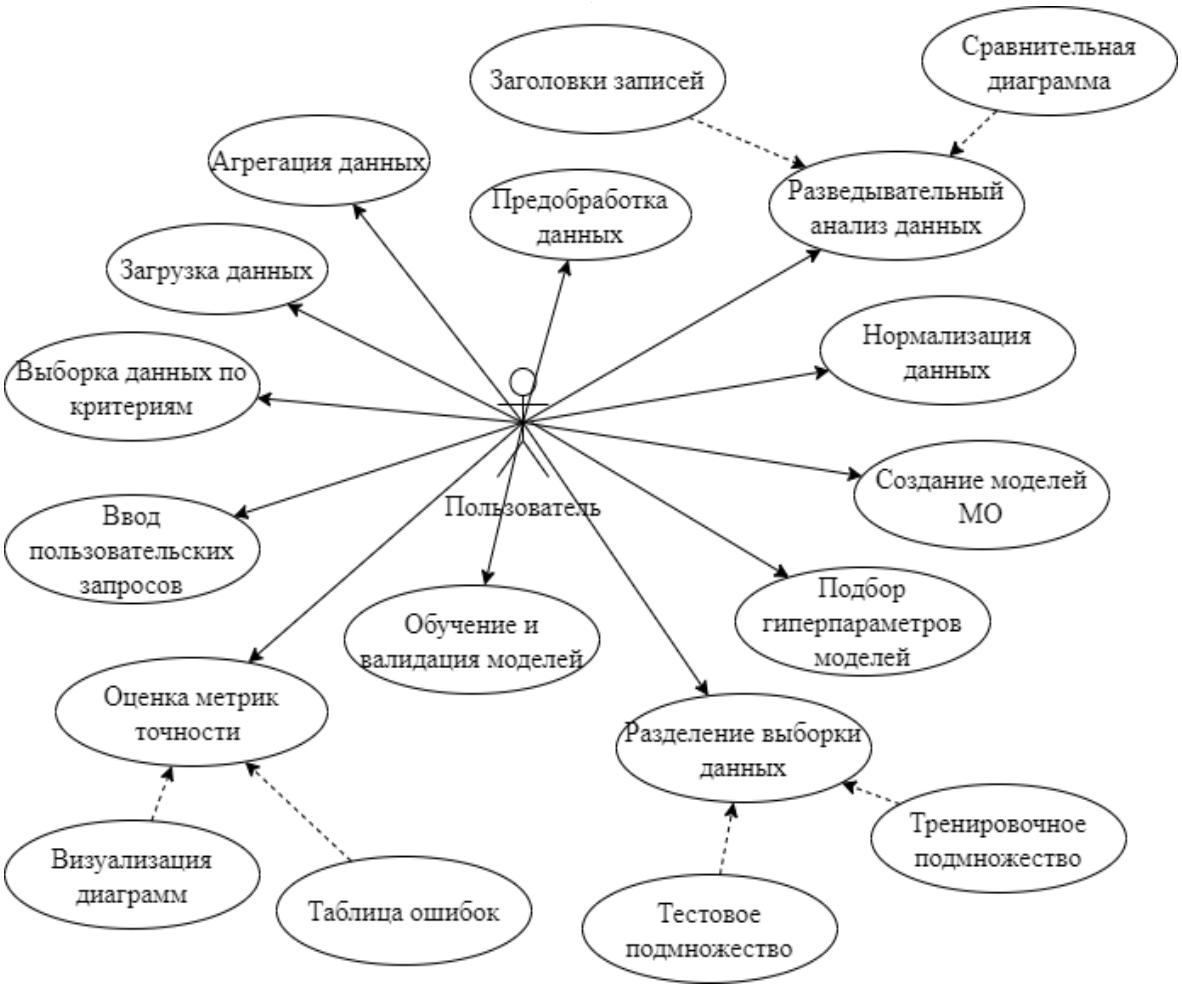

Рис. 2. Диаграмма основных вариантов использования СППР

анализа. Результатом являются объекты обученных моделей, сериализуемые в формате pickle.

4. Модуль оценки точности и адекватности моделей. Обеспечивает расчет метрик оценки качества созданных моделей, в том числе полноты, точности и F1-меры для классификации данных. В результате формируются файлы логов по результатам оценок моделей в виде кросстаблицы для быстрого анализа ЛПР с целью выявления модели, наиболее обобщающей входные данные.
5. Модуль ранжирования и визуализации результатов. Осуществляет отображение результатов работы СППР в клиентском интерфейсе веб-приложения в графическом (диаграммы, графики) и табличном виде. Ключевым результатом является перечень клиентской базы, разделенным по разным сегментам (кластерам), в зависимости от их покупательской способности и возможных сценариях действия (повышение уровня затрат, удержание текущего уровня, снижение до минимального уровня и уход).

В связи с возможным не стабильным (динамическим) ростом уровня вычислительной нагрузки, оказываемой СППР в процессе ее функционирования на ВК, на этапе проектирования системы целесообразно заложить в нее возможности балансировки вычислительных ресурсов путем интеграции механизмов миграции виртуальных серверов между узлами облачной инфраструктуры. Это может быть осуществлено путем разработки специализированного сервиса диспетчеризации вычислительных ресурсов на базе запуска ряда системных служб, отслеживающих степень использования процессоров, оперативной памяти и дискового пространства физических серверов. Концептуальная схема данного процесса приведена на рисунке 1.

ВК при развертывании запускает диспетчер ресурсов, осуществляющий процессы мониторинга и оценки уровня загруженности пула виртуальных серверов, на отдельных узлах которых развернуты модули СППР. Взаимодействие данного пула с разнородными иточниками данных, среди которых предусмотрены как SQL

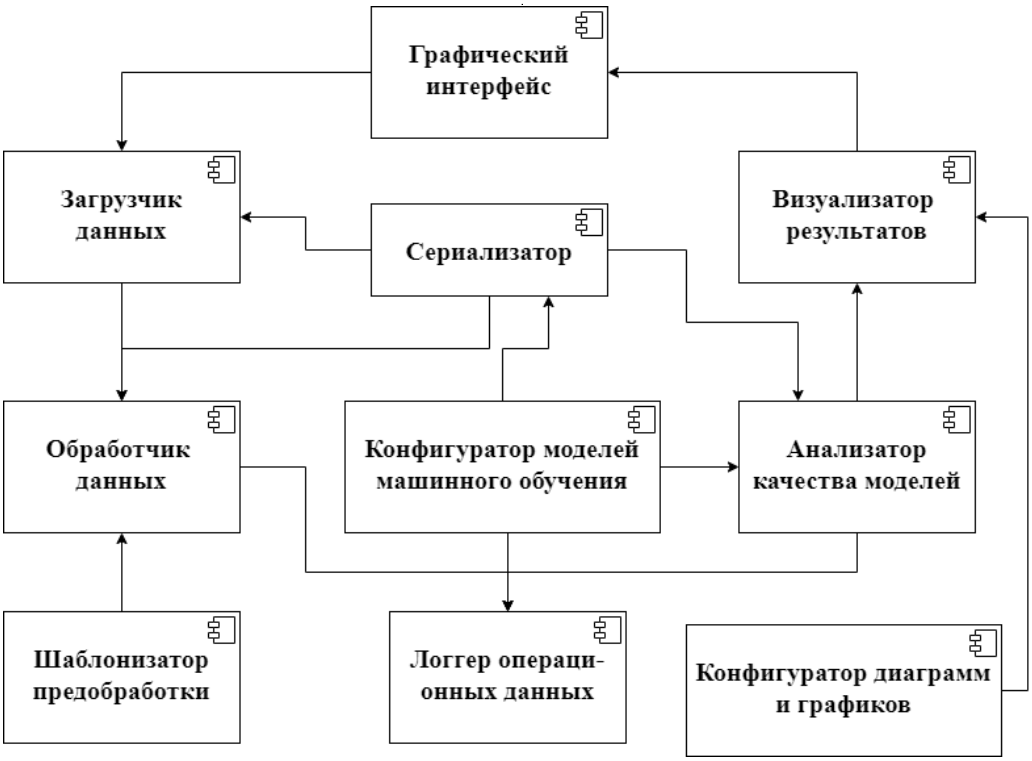

Рис. 3. Диаграмма ключевых компонентов СППР

и NoSQL БД, так и отдельные текстовые или табличные файлы и наборы документов, осуществляется в случае штатного (не превышающего 70% загрузки ресурсов ВК) режима посредством выполнения асинхронных запросов. Результаты обработки данных выводятся на клиенсткую часть СППР на формы графического интерфейса, после чего ЛПР имеет возможность сопоставлять результаты, анализируя различные сценарии по сегментированию и таргетированию клиентов.

Разработка проекта

Для системного и последовательного отображения функционала, порядка работы и процессов обмена данными между модулями проектируемой СППР целесообразным является использование унифицированного языка объектного моделирования UML. Основные возможности взаимодействия с СППР со стороны ЛПР (пользователя системы) отражены на диаграмме вариантов использования (рисунок 2). Пользователь может: загружать данные в систему (выбирать режима автоматической или ручной загрузки данных из файлов или из удаленных источников); агрегировать (объединять) данные в таблицы; выполнять предобработку данных (редактировать записи вручную или путем активации автоматического режима); разведывательный анализ (выводить заголовки записей; группировать их и строить сравнительные диаграммы по статистическим и мета данным); нормализировать данные (выбирая функции нормализации); осуществлять выборку данных по заданным критериям (в качестве которых могут выступать значения полей или временные интервалы); вводить SQL запросы (если источником данных является реляционная БД); разделять выборки данных на тренировочные и тестовые подмножества в различных пропорциях; выбирать, создавать, обучать и выполнять валидацию моделей МО; подбирать вручную или формировать автоматически (на базе метода grid-search) значения гиперпараметров моделей МО для оптимизации их точности; производить оценку метрик точности моделей на базе построения таблиц ошибок и графических диаграмм (ROC и AUC кривых); вывода ранжированного перечня целевых пользователей с предлагаемыми целевыми сценариями действий для повышения уровня их лояльности.

С целью отображения процессов обмена данными СППР между отдельными функциональными элементами системы построена диаграмма компонентов, приведенная на рисунке 3. Набор компонентов графического интерфейса взаимодействуют с загрузчиком данных, результаты работы которого необходимы обработчику данных (функционирующему на основе выбранных шаблонов предобработки) для логирования промежуточных и итоговых операционных результатов. На базе использования логов конфигуратор моделей МО ис-

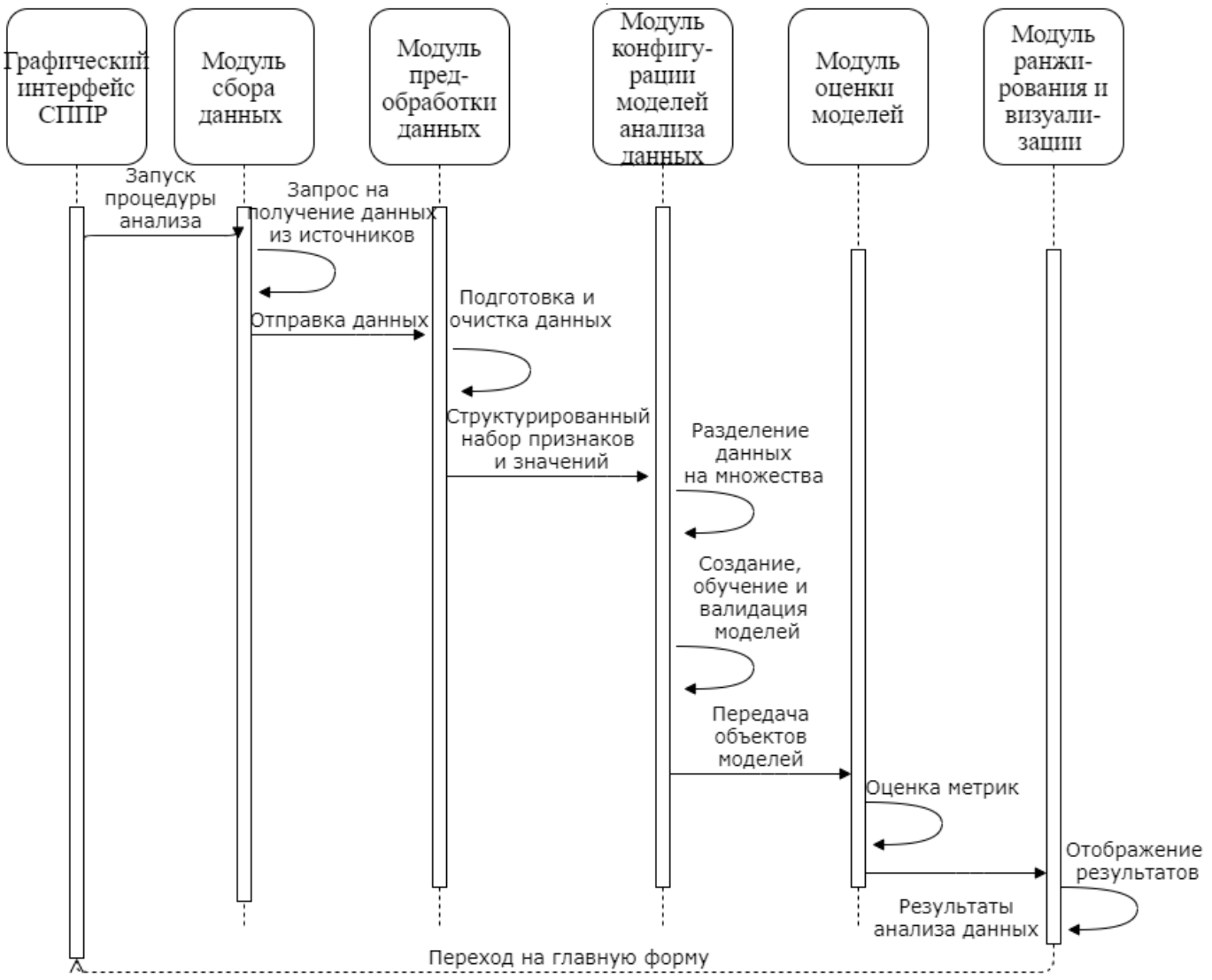

Рис. 4. Диаграмма общей последовательности действий СППР

пользует компонент сериализации (преобразования объектов в файлы) для анализа качества моделей и визуализации результатов в табличном и графическом виде (на основе соответствующих конфигураций), используемых при обновлении графического интерфейса пользователя.

С целью формализации и последовательного отображения процессов обмена данными между отдельными модулями системы, имплементирующими часть из приведенных выше компонентов, целесообразна разработка диаграммы последовательности действий. Рассмотрим данный процесс в рамках одного узла. После выделения ВК стартового объема ресурсов для виртуального сервера, в рамках которого запускаются и разворачиваются модули СППР посредством инициализации процедуры анализа через графический интерфейс пользователя выполняется последовательная активация соответствующих служб и процессов. В частности, осуществляются запросы на получение данных из выбранных источников, их отправка для предобработки, подготовка и очистка, структурирование и передача признаков с соответствующими им значениями, разделение данных на отдельные подмножества с последующим обучением и валидацией моделей. После этого реализуется передача сохранных объектов моделей МО для оценки метрик их качества и выполняется вывод полученных результатов в виде ранжированного перечня альтернативных сценариев поведения руководства компании для привлечения новых и повышения лояльности существующих клиентов в интерактивном режиме отображения.

Важной особенностью предлагаемого проекта СППР является ее динамическая масштабируемость в ряде

случаев (при использовании 80% выделенных ресурсов автоматически задействуются резервные узлы, на которых развертывается Docker образ с нужными модулями, число которых меняется на основе анализа статистики за предварительно заданный период времени).

Заключение

Разработанный проект системы поддержки принятия решений для провайдера сетевых услуг на базе вычислительных кластеров является систематизированным и унифицированным решением задачи распределенного анализа данных, которое может быть использовано в качестве формализованного виденья для дальнейшей программной реализации и развертывания на реальной облачной инфраструктуре с целью коммерческого использования. Возможными путями совершенствование предложенного проектами являются: интеграция большего числа данных для формирования полноценного пула структурированных озер данных в виде распределенных репозиториев для обеспечения целостности информации и ее унификации; внедрение отдельного модуля оперативного анализа данных для формирования быстрых кросс-отчетов; повышение уровня надежности и производительности системы путем оптимизации алгоритмов анализа данных как на программном, так и на аппаратном уровне; снижение общего энергопотребления вычислительного кластера путем более тонкой балансировки ресурсов серверов.

ЛИТЕРАТУРА

1. Шибаев Д.С. Оптимизация методов прогнозирования, обработки и анализа информации в разностуктурных хранилищах даних / Д.С. Шибаев, В.В. Вычужанин, Н.О. Шибаева, Н.Д. Рудниченко // Информатика и математические методы в моделировании. — 2018. — № 1. — С. 78–85.
2. Чехарин Е.Е. Большие данные: большие проблемы / Е.Е. Чехарин // Перспективы науки и образования. — № 3 (21). — 2016. — С. 7–11.
3. Ивутин А.Н., Есиков Д.О., Мельник С.И. Кластерная вычислительная система для решения задач обеспечения устойчивости функционирования распределенных информационных систем// Известия Тульского государственного университета. Технические науки. — 2016. — № 9. — С. 90–95.
4. Сизов В.А. Разработка моделей повышения эффективности сохранности данных в распределенной вычислительной среде на основе динамического резервирования данных// Вестник евразийской науки. — 2018.-Т.10 — № 6. — С. 74.
5. Ледянкин И.А., Легков К.Е. О некоторых концептуальных вопросах разработки параллельных структур вычислительных задач кластерных вычислительных систем // Наукоемкие технологии в космических исследованиях Земли. — 2014. — Т. 6, № 6. — С. 30–38.
6. Борисов В.В., Зернов М.М., Федулов А.С., Якушевский К.А. Исследование характеристик гибридного вычислительного кластера //Системы управления, связи и безопасности. — 2016. — № 4. — С. 129–146.
7. Цебренко К.Н. Анализ вопросов безопасности информационных систем на основе применения высокопроизводительных вычислительных систем // Международный журнал гуманитарных и естественных наук. — 2020. — № 8–1. — С. 103–105.
8. Богатырев В.А., Богатырев А.В., Богатырев С.В. Перераспределение запросов между вычислительными кластерами при их деградации// Известия высших учебных заведений. Приборостроение. — 2014. — Т. 57, № 9. — С. 54–58.
9. Биктимиров М.Р. Тенденции развития технологий обработки больших данных и инструментария хранения разноформатных данных и аналитики / М.Р. Биктимиров, А.М. Елизаров, А.Ю. Щербаков. — № 5. — Т. 19. — 2016. — С. 390–406.
10. Мулюкова К.В. Сравнительный анализ современных инструментов Data Mining // Молодой ученый. — 2019. — № 1. — С. 19–21
11. Boyko V. Concept implementation of decision support software for the risk management of complex technical system / N. Rudnichenko, V. Boyko, S. Kramskoy, Y. Hrechukha, N. Shibaeva // Advances in Intelligent Systems and Computing of the series Advances in Intelligent Systems and Computing. — 2016. — № 512. — P. 255–269.
12. Rudnichenko N. Decision Support System for the Machine Learning Methods Selection in Big Data Mining / N. Rudnichenko, V. Vychuzhanin, I. Petrov, D. Shibaev // Proceedings Of The Third International Workshop on CMIS (CMIS-2020): session 6 "Intelligent Information Technologies" April 27-May 1, 2020. — Zaporizhzhia: NU "Zaporizhzhia Polytechnic" (edited by S. Subbotin), 2020. — P. 872–885.

Сагалаева Анна Игоревна (omegaanya@gmail.com), Ромашкова Оксана Николаевна (ox-rom@yandex.ru).

Журнал «Современная наука: актуальные проблемы теории и практики»