УДК 004.91 **DOI** 10.15622/ia.21.5.2

Н.И. Юсупова, Г.Р. Воробьева, Р.Х. Зулкарнеев

ПОДХОД К ИНТЕГРАЦИИ РАЗНОРОДНЫХ ИСТОЧНИКОВ МЕДИЦИНСКИХ ДАННЫХ НА ОСНОВЕ МИКРОСЕРВИСНОЙ АРХИТЕКТУРЫ

Юсупова Н.И., Воробьева Г.Р., Зулкарнеев Р.Х. **Подход к интеграции разнородных источников медицинских данных на основе микросервисной архитектуры.**

Аннотация. Задача обработки медицинской информации в настоящее время в нашей стране и за рубежом решается посредством разнородных медицинских информационных систем, преимущественно локального и регионального уровней. Постоянно возрастающий объем и сложность накапливаемой информации наряду с необходимостью обеспечения прозрачности и преемственности обработки медицинских данных (в частности, к примеру, по бронхолегочным заболеваниям) в различных организациях требует разработки нового подхода к интеграции их разнородных источников. При этом важным требованием к решению поставленной задачи является возможность веб-ориентированной реализации, что позволит сделать соответствующие приложения доступными широкому кругу пользователей без высоких требований к их аппаратно-программным возможностям. В работе рассматривается подход к интеграции разнородных источников медицинской информации, который основан на принципах построения микросервисных веб-архитектур. Каждый модуль обработки данных может быть использован независимо от других программных модулей, предоставляя универсальную точку входа и результирующий набор данных в соответствии с принятой схемой данных. Последовательное выполнение этапов обработки предполагает передачу управления соответствующим программным модулям в фоновом режиме по принципу Cron. В схеме декларируется два вида схем данных – локальная (от медицинских информационных систем) и глобальная (для единой системы хранения), между которыми предусмотрены соответствующие параметры отображения по принципу построения XSLT-таблиц. Важной отличительной особенностью предлагаемого подхода представляется модернизация системы хранения медицинской информации, заключающейся в создании зеркальных копий основного сервера с периодической репликацией соответствующей информации. При этом взаимодействие между клиентами и серверами хранилищ данных осуществляется по типу систем доставки контента с созданием сеанса соединения между конечными точками по принципу ближайшего расстояния между ними, рассчитанного по формуле гаверсинусов. Проведенные вычислительные эксперименты над тестовыми данными по бронхолегочным заболеваниям показали эффективность предложенного подхода как для загрузки данных, так и для их получения отдельными пользователями и программными системами. В целом показатель реактивности соответствующим веб-ориентированных приложений был улучшен на 40% при стабильном соединении.

Ключевые слова: медицинские данные, хранилища данных, интеграция данных, веб-приложения, система доставки контента, микросервисная архитектура.

1. Введение. В настоящее время биомедицина и здравоохранение все чаще классифицируются как области интенсивного использования данных (так называемые «data-intensive field»), где специалисты постоянно сталкиваются с необходимостью обработки и анализа большого объема разнородной информации. При

этом разнородная медицинская информация представляет собой данные различного формата и назначения, получаемые из различных источников по различным протоколам. Для конкретизации решаемой задачи здесь и далее разнородные медицинские данные представляют собой медико-клиническую информацию, а также административно-социальные данные задействованных в соответствующих медицинских процессах объектов, субъектов и ресурсов. Медико-экономическая информация в данной работе не рассматривается и является предметом дальнейших исследований авторов.

В качестве источников медицинских данных выступают клинические данные для поддержки принятия решений различной специализации в виде стандартизированных данных из электронных историй болезни, данные с датчиков мониторинга и различных записывающих лабораторных устройств, данные неотложной помощи, лекарственных препаратах и пр. Здесь же фигурируют и административно-социальные данные пациентов, врачей, медицинских учреждений. При этом одной из основных задач развития технологий обработки таких данных является соблюдение преемственности в лечении пациентов, обеспечение прозрачности клинико-диагностических процессов, повышение эффективности лечения за счет оперативного доступа ко всей необходимой информации.

На сегодняшний день достигнут значительный прогресс в области обработки и хранения медицинской информации [1]. Однако интеграция и управление постоянно растущими объемами разнородных медицинских данных часто сопряжены с рядом проблем, связанных прежде всего с трансформацией непосредственно самой отрасли, обусловленной накоплением массивов цифровых данных, их аналитической обработкой и применением полученных результатов в процессе принятия соответствующих решений. При этом акцент смещается на преобразование крупномасштабных разнородных медицинских данных в информационные продукты и стратегии, что является, в частности, одной из ключевых правительственных инициатив в области персонализованной медицины. Так, с 2011 по 2018 год реализован первый проект по созданию единой государственной информационной системы в сфере здравоохранения, с 2018 года разрабатывается второй проект по созданию единого цифрового контура.

Другой областью повышенного интереса к интеграции и хранению данных в медицине является разработка стратегий для эффективного анализа постоянно растущего массива электронных медицинских карт (ЭМК). При этом одна из основных проблем при

извлечении знаний из электронных медицинских карт заключается в том, что электронные медицинские записи представляют собой крайне разнородные источники данных со сложным массивом количественных, качественных и транзакционных данных. Разрозненные типы данных включают коды МКБ (используемые в основном для ценообразования и оплаты больничных процедур), биохимические и лабораторные тесты, клинические (текстовые) заметки, исторические архивы медицинских вмешательств, методов лечения и даже доставки фармацевтических препаратов. Эти источники данных часто собираются десятками людей (иногда с субъективными критериями) для каждого случая. Следовательно, данные ЭМК довольно сложно анализировать, особенно если рассматривать институциональные и даже многоцентровые уровни с большим количеством пациентов.

Клинические и биомедицинские данные бывают самых разных форматов; они зачастую сложны, неоднородны, слабо аннотированы и, как правило, не структурированы, а объем данных постоянно растет. Каждая из этих проблем: размер, разнообразие, форматирование, сложность, неоднородность, плохие аннотации и отсутствие структуры создают проблемы для эффективного применения в процессе принятия решений.

Проблема усугубляется отсутствием единого подхода к организации медицинских данных, несогласованностью и разнородностью медицинских информационных систем на уровне как отдельных медицинских учреждений, так и территориальных областей. Здесь же важно отметить проблему низкой сопряженности сведений о различных случаях лечения одного и того же человека в конкретной медицинской организации (в частности, при лечении бронхолегочных заболеваний). В подавляющем большинстве случаев в ЭМК дискретно фиксируется только один случай лечения пациента (от поступления до выписки), однако непосредственно интеграция данных предполагает создание единой системы, обеспечивающий доступ ко всем передвижениям пациента в рамках как отдельной организации, так и страны в целом.

Для решения обозначенных проблем требуется подход к интеграции данных, что, безусловно, сопряжено с решением вопросов взаимодействия медицинских информационных систем между собой. Очевидным решением здесь является использование единого информационного хранилища медицинских данных (в частности, по бронхолегочным заболеваниям), обеспечивающего возможность

оперативного импортирования и экспортирования разнородных данных в / из любой информационной системы.

При этом необходимо обеспечить интеграцию источников данных, которые, в свою очередь, существенно различаются по типу (видео, звук, текст, графика, компьютерная анимация и т.д.), назначению, способу записи и возможности удаленной обработки. Также важно учитывать, что зачастую при хранении результатов клинических обследований на различном медицинском оборудовании используются разнообразные форматы данных, нередко специфичные для определенного типа оборудования и не доступные для обработки общедоступными средствами.

2. Состояние вопроса. В настоящее время известен целый ряд российских и зарубежных научно-исследовательских работ, посвященных обозначенному вопросу. Однако наибольшее внимание при этом уделяется созданию такого формата представления медицинских данных, который позволял бы независимо от используемых аппаратно-программных платформ импортировать и экспортировать информацию, а также осуществлять ее обработку, анализ и графическую интерпретацию [2, 3]. Так, к примеру, известны результаты в области разработки единого формата представления медицинской информации, основанного на онтологическом подходе к представлению объектов и субъектов медицинского обслуживания [3-5]. На сегодняшний день формат доступен не только на российском, но и на международном уровне. Однако в настоящий момент вопрос его внедрения в существующие медицинские информационные системы остается открытым.

Другое направление исследований связано с разработкой непосредственно подхода к интеграции без привязки к формату представления данных. Так, согласно некоторым научным школам в данной области, интеграция должна быть проведена поэтапно, начиная от уровня отдельно взятого пациента и данных медицинских приборов и аппаратов до представления сводных данных в единую государственную информационную систему в сфере здравоохранения [6]. В большинстве исследований по интеграции в основу положен онтологический подход, а данные рекомендуется представлять в общепринятом формате HL7 FHIR [7], а также с применением технологий облачных сервисов для реализации хранения больших объемов медицинских данных из гетерогенных источников.

Для обобщения мирового опыта в области интеграции разнородных источников медицинских данных были проанализированы результаты известных научных исследований по

этому вопросу. Результаты сравнительного анализа нескольких распространенных подходов к интеграции медицинской информации приведены в таблице 1.

Обобщая представленные в таблице 1 данные, можно резюмировать следующее. Большинство известных подходов являются узкоспециализированными и представляют собой решение задач обработки и анализа данных по одному или нескольким заболеваниям. При этом реализующая их программная архитектура, как правило, слабо масштабируется – введение новых функций и ресурсов требует существенных трудовых и вычислительных затрат. В большинстве случаев предлагаемые решения локальны и ориентированы на реализацию в пределах одного медицинского учреждения, что сопряжено с соответствующей физической организацией систем долговременного хранения медицинской информации. Немаловажным представляется тот факт, что практически каждое известное решение предполагает применение собственного формата медицинских данных, полученных в результате интеграции гетерогенных источников. При этом ни один из форматов не коррелирует в достаточной мере с известными стандартами представления медицинской информации (например, HL7 FHIR), что существенно затрудняет интеграцию соответствующих решений со сторонними информационными системами. Кроме того, в подавляющем большинстве случаев имеют место десктопные приложения, что, в свою очередь, сопряжено с высокими требованиями к вычислительным ресурсам конечного пользователя. Выполнение запросов к большим объемам медицинской информации так же не оптимизировано, что приводит к низкой реактивности соответствующих приложений.

В этой связи возникает актуальная задача разработки подхода к интеграции данных из разнородных медицинских информационных систем, обеспечивающего возможность веб-ориентированного оперативного обращения с различных интерфейсов взаимодействия для последующей обработки полученных результирующих данных с выгрузкой в установленные структуры, к примеру, в виде электронной медицинской карты пациента.

Таблица 1. Сравнительный анализ подходов к интеграции разнородной медицинской информации

Название, организация, назначение	Преимущества	Недостатки
Масштабируемые информационные системы [8] *Совместные исследования университетов США (Albert Einstein College of Medicine, Beth Israel Deaconess Medical Center, Tufts University School of Medicine и др.)* Цель: создание инфраструктуры оцифрованных медицинских данных по всему миру	– открытый код – веб-интерфейс для доступа к данным – положительные отзывы пользователей с точки зрения эргономики приложения	– в настоящее время система предназначена для медицинской информации только по интенсивной терапии – используется собственный формат данных (нет поддержки стандарта FHIL) – низкая реактивность приложения
Цифровая платформа для анализа и визуализации эпидемиологических данных [9] *ICMR-National Institute of Malaria Research, Индия* Цель: разработка универсального интерфейса, который позволит проводить быстрый и интерактивный анализ эпидемиологических данных по малярии	– простой и понятный интерфейс – единый доступ к большим эпидемиологическим данным, накопленным в рассматриваемом регионе	– только настольные приложения – высокие требования к вычислительным мощностям пользователей – система предназначена для обработки региональных данных по малярии – на данный момент нет возможности масштабирования

Продолжение Таблицы 1

Глобальный хаб для медицинских данных [10] *Clalit Research Institute, Израиль* Цель: разработка единого информационного пространства медицинских данных для реализации программ персонализированной медицины	– наличие единого информационного пространства в рамках региона – использование облачных технологий для хранения данных – веб-ориентированный интерфейс, доступный широкому кругу пользователей без высоких требований к вычислительным ресурсам	– только региональный масштаб – низкая реактивность – сложность масштабирования
Интеграция данных из разнородных источников [11] *HealthCore Inc, США* Цель: описание и оценка процесса интеграции данных из нескольких дополнительных источников для проведения исследований результатов лечения пациентов с раком легкого	– наличие единого информационного пространства в рамках региона	– рассматриваются только 4 источника данных, масштабирование не предусмотрено на настоящий момент – данные представлены в собственном, отличном от стандартного, формате (негативно сказывается на интероперабельности)
Интеграция систем клинических и лабораторных исследований [12] *Институт медицинской информатики, Германия* Цель: создание единой онтологии для	– в основе лежит онтологический подход, позволяющий выявить связи (в том числе неявные) между ресурсами, объектами и субъектами	– в настоящее время существует только десктопная версия, сопряженная с высокими требованиями к вычислительным ресурсам пользователя – не предусмотрено масштабирование системы

Продолжение Таблицы 1

электронных медицинских карт университетской больницы Эрлангена		– требуется стандартизации целевой онтологии и внутренней модели данных, а также интеграции дополнительных отображений в стандартизированные терминологии
Технологии интеграции и визуализации данных (DIVT) [13] *Institute of Biomaterials and Biomedical Engineering, Канада* Цель: анализ эффективности принятия решений врачами на основе интегрированных наборов медицинских данных	– проводится подробный анализ взаимосвязи доступа пользователей с эффективностью использования их времени, точностью принятия решений и когнитивной нагрузкой	– в качестве информационной базы рассматриваются не данные электронных медицинских карт, а рекомендации для врачей интенсивной терапии – не рассматриваются непосредственно подходы к интеграции данных, фокус на оценке эффективности применения единого информационного пространства
Интеграция источников структурированной и неструктурированной медицинской информации [14] *University of Antwerpen, Бельгия* Цель: сравнительная оценка эффективности медицинских прогнозов на основании изолированных и интегрированных источников данных	– рассматриваются возможности применения ранней (одна схема для всех источников) и поздней интеграции (отдельная схема для каждого источника) данных – предложена схема ансамблирования методов ранней и поздней интеграции данных для повышения эффективности медицинского кодирования на их основе	– на текущий момент результаты не масшабируемы и ориентированы на локальное применение в рамках одного учреждения – данные представлены в собственном, отличном от стандартного, формате (негативно сказывается на интероперабельности) – в настоящее время существует только

Продолжение Таблицы 1

		десктопная версия, сопряженная с высокими требованиями к вычислительным ресурсам пользователя – не предусмотрено масштабирование системы
Интеграция данных для прецизионной персонализированной медицины [15] *National Institute of Cardiology 'Ignacio Chávez', Мексика* Цель: развитие персонализированной медицины с использованием методов искусственного интеллекта для интегрированных медицинских данных	– предложен подход к созданию информационной медицинской экосистемы как центра интеграции данных из распределенных источников	– не рассматриваются гетерогенные источники данных – формат данных отличен от стандартизованного или рекомендованного соответствующим организациями
Health Mining [16] *University of Catania, Италия* Цель: создание системы интеграции источников медицинских данных для анализа и поддержки принятия решений на этой основе	– подход ориентирован на интеграцию трех типов данных: клинические данные, социальные данные и данные IoT	– используется только централизованное хранение – приложение предназначено только для локального использования – расширение решения на большее число данных и организаций сопряжено с колоссальными вычислительными затратами

Продолжение Таблицы 1

Облачное хранилище открытых биомедицинских данных из разнородных источников [17] *Stanford University, США* Цель: разработка масштабируемых интеллектуальных инфраструктур биомедицинских данных для повышения качества биомедицинских исследований.	– используются облачные технологии, что при единой точке доступа создает широкие возможности для централизованного хранения больших данных, – основой является онтологический подход, позволяющий выявить связи между задействованными в процессе ресурсами.	– низкая реактивность веб-приложений из-за ограниченных выделенных ресурсов общедоступных конечных точек SPARQL, – сложность сопровождения проекта из-за семантической неоднородности источников данных.

3. Формализация задачи интеграции медицинских данных. В общем виде задача интеграции разнородных источников медицинских данных может быть описана следующим образом. Пусть представлены *N* источников медицинской информации:

$$A = [A_1, \dots, A_N]. \quad (1)$$

При этом множество кортежей данных из источника A_i можно обозначить как:

$$D^{A_i} = \{d_1^{A_i}, \dots, d_m^{A_i}\}. \quad (2)$$

Тогда информационный поток, поступающий от атомарного источника, в момент времени *t* описывается (в контексте обработки медицинских данных) линейной моделью вида:

$$y(t) = y(t_0) + v(t - t_0), \quad (3)$$

где *t* – стартовое время отсчета, $y(t_0)$ – количество сообщений за время *t*, *v* – средняя скорость увеличения информационного потока.

Следовательно, флуктуация каждого информационного потока может быть определена как:

$$\sigma(t_n) = \sqrt{\frac{1}{n}\sum_{i=0}^{n}[y(t_i) - (y(t_0) + v(t_i - t_0))]^2}. \quad (4)$$

Каждый источник характеризуется собственной схемой организации данных, предполагающей описание элементов данных посредством одной из известных моделей (реляционной, сетевой, иерархической, объектной и пр.). Здесь и далее схемы, используемые отдельными источниками данных, предлагается обозначать как «локальные». Пусть в контексте локальной схемы данные, поступающие от одного удаленного источника, представляют собой упорядоченный набор значений, где каждый кортеж может быть описан отношением вида:

$$X = [x_1, x_2 \dots, x_N], \quad (5)$$

где x_1,...,x_n – значения исходного кортежа для n параметров.

При этом предполагается, что данные из разнородных источников аккумулируются централизованно в хранилище, где, в свою очередь, используется глобальная схема данных, к которой должны быть трансформированы все локальные схемы:

$$Y = [y_1, y_2 \dots, y_m], \quad (6)$$

где y_1,...,y_m – значения кортежа в глобальном хранилище для m параметров.

Соответственно, для включения источника в единую систему данных необходим дополнительный компонент трансформации данных, обеспечивающий автоматическое отображение локальной схемы в глобальную с применением соответствующей схемы отображения (по аналогии с известной моделью XML – XSL(T)). Указанный компонент можно представить в виде функции отображения следующим образом:

$$f_p: \{a_1, a_2, \dots, a_n\} \xrightarrow{g} \{b_1, b_2, \dots, b_m\}; g = \{g_1, g, \dots, g_m\}, \quad (7)$$

где f_p – функция преобразования кортежа данных, задающая правила формирования нового кортежа из элементов заданного кортежа; $\{a_1, a_2, \dots, a_n\}$ – исходный кортеж данных; $\{b_1,b_2,\dots,b_m\}$ – формируемый кортеж данных; g=$\{g_1, g_2,\dots,g_m\}$ – эталонный набор параметров, который необходимо реализовать в формируемом кортеже данных.

Одним из значимых аспектов предлагаемого подхода является формирование результирующего набора данных как совокупности непосредственно содержательных данных, а также метаданных, которые характеризуют обстоятельства их получения (к примеру, информация о медицинском учреждении, используемом лабораторном оборудовании и пр.). Если рассматривать информационный поток от удаленного источника в качестве совокупности отдельных (необязательно взаимосвязанных) информационных блоков, то сказанное можно представить следующим образом:

$$D = \{X, M\}, \tag{8}$$

где X – массив кортежей содержательных данных, M – массив кортежей с метаданными для одного источника.

Для описания метаданных предлагается рассматривать источники данных в качестве группы ресурсов в соответствии с принятыми стандартами и спецификациями представления медицинской информации (в данном случае имеется в HL7 FHIR [5]). В общем виде каждый k-й ресурс R представлен совокупностью своих p характеристик – параметров соответствующего провайдера:

$$R^k = \{r_1: \{e_1, f_1\}; \dots; r_p: \{e_p, f_p\}; type\}, \tag{9}$$

где каждая характеристика r_i характеризуется парой ключ e_i и значение f_i; type – флаг типа ресурса.

При этом для повышения информативности ресурсы предлагается разделить на событийные и участники. К первым относятся атрибуты, характеризующие процесс получения данных (параметры оборудования, к примеру), а ко вторым – атрибуты, характеризующие непосредственно участников процесса формирования набора данных (пациенты, медицинский персонал и пр.):

$$\begin{gathered} R = Rs \cup Rp: \forall k\ R^k\mathrm{s} = \{r1: \{e1, f1\}; \dots; rp: \{ep, fp\};' \mathrm{s}'\}; \\ R^k\mathrm{p} = \{r1: \{e1, f1\}; \dots; rp: \{ep, fp\};' \mathrm{p}'\}. \end{gathered} \tag{10}$$

Последовательная обработка кортежей данных из поступающего от провайдера информационного потока (получение данных, их нормализация и приведение к стандартному виду типа HL7) в терминах кортежей данных может быть описано следующим образом. Пусть каждый k-й этап обработки данных приводит к

формированию нового кортежа данных. Этот факт может быть описан таким образом, что зафиксированное значение всех элементов множества X относительно одного значения параметра k представляет вектор состояния:

$$C_i = \langle X_1, k_1; \dots; X_n, k_n \rangle. \quad (11)$$

Тогда множество всех возможных векторов состояний $C = \{C_i, i = 1, \dots, |C|\}$ составляет полное множество состояний получаемых от источника данных, которое можно представить в виде $C = \prod_i k_i$ (в данном случае соотношение приведено для одного кортежа информационного потока, получаемого от отдельно рассматриваемого источника данных).

По аналогии формирование единого информационного пространства можно описать последовательной сменой состояний каждого его кортежа. При этом представляется целесообразным введение так называемого начального этапа $k = 0$, при котором элементы каждого кортежа единого информационного пространства представляют собой неопределенные (NULL) значения. На последнем k_n этапе формируется окончательный набор кортежей в соответствии с описанными выше функциями отображения и трансформации.

4. Подход к территориальному распределению хранилищ данных. Предлагаемый подход к интеграции данных из разнородных источников основан на их последовательном преобразовании с учетом соответствующих метаданных. Ожидаемым результатом данного процесса является формирование единого централизованного хранилища медицинской информации, построенного по принципу CDN (Content Delivery Network [19]).

Выбор CDN в качестве базового для решения выявленной проблемы обусловлен следующим. На протяжении нескольких лет авторами проводились исследования, связанные с повышением реактивности веб-приложений, ориентированных на данные (рассматривались различные прикладные пространственные данные, в частности, геофизические [21, 23]). Как и ожидалось, было установлено, что ключевым фактором, определяющим характеристики реактивности приложений, является объем и сложность запрашиваемых клиентской стороной размещенных на сервере данных. Однако вместе с тем проведенные исследования показали зависимость скорости отклика от расстояния между источником запроса (клиентом) и размещением запрашиваемых данных. В целом, это известная в веб-разработке проблема, заключающаяся в том, что

время ожидания ответа от сервера при направлении к нему запроса со стороны клиента тем больше, чем территориально дальше находится источник запроса от его цели. Известное решение – технология CDN – положена в основу предложенного подхода к модернизации физической организации системы хранения информации.

Известно, что в целом при использовании CDN-подхода распределение соединения «клиент – сервер» осуществляется по-другому. Запрос клиента автоматически переадресуется к географически ближайшему кэширующему серверу в составе CDN, что позволяет доставлять контент существенно быстрее по сравнению с аналогичным запросом, направленным к основному серверу. При этом существенного снижается пропускная нагрузка на основной сервер за счет задействования дополнительных серверов для кэширования данных [20].

Предполагается, что физически единое информационное хранилище должно быть разделено на несколько территориальных кластеров, объединяемых через внешние интерфейсы в общую систему данных. Каждый территориальный кластер включает в себя несколько источников данных, в качестве которых выступают медицинские учреждения и организации, размещенные в пределах некоторого пространственного кластера с заданной геопространственной привязкой. С заданной периодичностью необработанные данные от соответствующих источников поступают в территориально ближайшее хранилище данных выделенного пространственного кластера (рисунок 1).

В результате единое информационное хранилище медицинских данных должно представлять собой географически распределенную сетевую инфраструктуру из множества источников данных по схеме типа «звезда». В центре указанной структуры размещается хранилище данных, для всех остальных связанных с ним хранилищ осуществляется систематическая репликация всех данных, что позволяет хранить копию имеющихся данных непосредственно в каждом территориальном кластере. Поскольку предполагается работа непосредственно с данными, то представляется целесообразной именно процедура репликации, а не кэшировании, как это предусмотрено известным подходом для сети доставки контента.

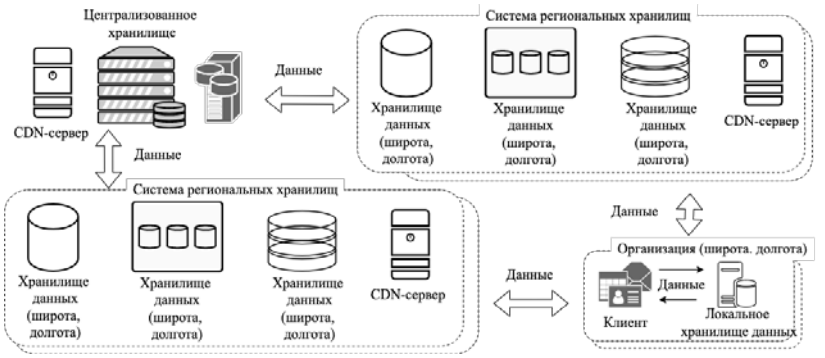

Рис. 1. Фрагмент схемы территориального CDN-распределения источников данных

Для загрузки и получения данных (так же в соответствии с принципами CDN) возможно применение технологии GeoDNS. Данная технология позволяет привязывать к одному доменному имени одновременно несколько IP-адресов. При анализе пришедшего от клиента запроса по соответствующему IP-адресу определяется географическое положение пользователя. В зависимости от полученного значения пользователь для получения / загрузки данных отправляется в территориально ближайшее хранилище данных. При этом расчет расстояния между соответствующими точками, соединенными дугой (расстояние по дуге большого круга), осуществляется по известной формуле гаверсинусов, которая представляет собой расстояние от центральной точки дуги, измеряемой удвоенным данным углом, до центральной точки хорды, стягивающей дугу:

$$d = 2r \arcsin\left(\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)\right), \quad (12)$$

где d – это центральный угол между двумя точками, лежащими на большой дуге; r – радиус сферы; φ_1 и φ_2 – широта первой и второй точек в радианах; λ_1 и λ_2 – долгота первой и второй точек в радианах.

Известно, что в CDN используется принцип кэширования по первому обращению, что означает максимальное время ожидания для пользователя, обратившегося к основному серверу первым. При этом все последующие пользователи будут получать данные, кэшированные на ближайшей к ним точке присутствия. Для преодоления ограничений, накладываемых этой схемой, используются технологии

регионального извлечения: соседние серверы, входящие в состав CDN, забирают контент друг у друга, а не обращаются к оригинальному серверу.

По аналогии с обозначенной схемой представляется целесообразным применить такой же подход к территориальному распределению и репликации данных в рамках предлагаемого единого хранилища медицинских данных. Пользователь, отправивший запрос на получение или загрузку данных, переадресуется к ближайшей точке присутствия и получает доступ к соответствующей копии основного хранилища медицинских данных. При этом если ближайшая точка присутствия не может найти данные (например, между сеансами репликации данных с основным сервером), то начинается поиск по соседним точкам присутствия, откуда и будет направлен ответ пользователю. Иными словами, имеет место подход, также известный как «tiered distribution» (или «многоуровневая раздача»).

Таким образом, за счет сокращения физического расстояния между потребителями и источниками данных можно уменьшить время отклика соответствующего контента. Поскольку одним из основных технических требований является возможность применения данных в веб-ориентированной среде, то время отклика от сервера при получении запроса от клиента на работу с информацией является критичным с точки зрения эргономики веб-приложений. Поэтому сокращение времени отклика за счет использования CDN-подобного подхода к организации хранения данных позволит при этом и повысить реактивность использующих их приложений.

Здесь представляется целесообразным отметить, что в качестве потребителя данных может выступать не только конечный пользователь, но и сторонние программные системы и приложения. При этом способ обращения к данным не зависит от типа обращающегося к ним клиента, поскольку IP-адрес источника запроса может быть выделен из заголовка любого пришедшего на сторону сервера запроса (предполагается, что взаимодействие между клиентом и сервером при этом осуществляется по стандартному интернет-протоколу HTTP(s)).

Еще одним положительным моментом от использования CDN-подхода для работы с медицинскими данными является снижение риска потери доступа к данным из-за потери основного сервера. В случае, если основной сервер или ближайшее для обращающегося к данным клиента хранилище по техническим причинам недоступны, запрос за счет принципа многоуровневой раздачи будет успешно выполнен. Иными словами, данные доступны все время, пока

восстанавливается работоспособность или доступность основного сервера или связанного с ним в сеть хранилища данных.

Уязвимость предлагаемого подхода сопряжена с привязкой к группе IP-адресов для управления распределением клиентских запросов между составляющими единого хранилища медицинских данных. В таком случае возможны блокировки, обусловленные блокировками сервисов, которые являются близкими по группе IP соответствующего CDN-провайдера. В результате таких блокировок возможна и блокировка территориальных компонент всего единого хранилища медицинских данных. Проблема решаема путем запроса изменения IP-адреса у CDN-провайдера.

Предложенная схема может быть (с некоторыми модификациями) адаптирована под различные геополитические регионы. Схема типа «звезда» по хранению данных может быть преобразована в схему типа «снежинка». В этом случае к выделяется основной сервер на федеральном уровне и по одному региональному серверу на уровне каждой области или региона страны. Данные от поставщиков поступают в реплицированное ближайшее хранилище, откуда передаются в региональное хранилище и далее – на основной сервер.

В этом случае процедура репликации должны быть асинхронно рассредоточена между различными уровнями поставщиков данных: от источников к региональным хранилищам, от региональных хранилищ – к основному серверу, и обратно. Техническая сложность реализации такого подхода, а также сопряженные с этим финансовые затраты представляются целесообразными в контексте существенного повышения реактивности использующих данные приложений.

5. Научная новизна предложенных решений. Отличительной особенностью предложенного подхода к интеграции данных является дополнение метаданных геопространственной меткой, характеризующей территориальное размещение источника данных на основании соответствующих параметров геолокации. Если представить сказанное в терминах HL7 FHIR [5] (как говорилось выше), то каждый k-й ресурс-источник данных R с p характеристиками – параметрами, необходимо дополнить следующим образом:

$$R^k = \{r1: \{e_1, f_1\}; \dots; r_p: \{e_p, f_p\}; \text{type}, \text{loc}: \{\text{lat}, \text{lng}\}\}, \quad (13)$$

где каждая характеристика *loc* характеризуется парой атрибутов lat и lng для представления географических широты и долготы источника-ресурса соответственно.

Глобальное информационное хранилище представляет собой совокупность кортежей, прошедших процедуру трансформации в соответствии с установленной схемой данных и дополненных соответствующей ссылкой на метаданные источника данных. Последние, в свою очередь, и содержат необходимую для позиционирования объекта геопространственную информацию. Так, множество кортежей данных из источника A_i можно обозначить как:

$$D^{A_i} = \left\{d_1^{A_i}, \dots, d_m^{A_i}, R^{A_i}\right\}, \tag{14}$$

где R^{A_i}– ссылка на метаданные источника A_i.

В соответствии с принципами CDN предлагается создание комплекса связанных хранилищ данных, репликация которых с централизованным хранилищем позволяет хранить данные в актуальном состоянии, а геопространственная метка обеспечивает возможность обработки запросов с учетом геопространственной локации их источников. Соответствующая информация представлены в метаданных хранилища медицинской информации:

$$\text{loc: \{lat: } \{x, x\_\text{val}\}, \text{lng: } \{y, y\_\text{val}\}\}, \tag{15}$$

где параметр loc метаданных произвольного источника данных представлен двумя парами параметров: соответственно ключ *x* и значение *x_val* для географической широты lat; ключ *y* и значение *y_val* для географической долготы lng местоположения источника данных. При этом при необходимости пространственная характеристика может быть задана в отличной от геодезической системе координат.

При получении данных от удаленного источника выполняется обработка соответствующего запроса, что предполагает непосредственно извлечение и анализ его геопространственной метки. Поскольку взаимодействие между клиентом и сервером предполагает применение протокола HTTP(HTTPS), соответствующая информация может быть получена посредством извлечения IP-адреса клиента непосредственно из заголовка запроса (Request, RemoteAddr):

$$\begin{gathered}\text{Request} = \{\text{Header}, \text{Body}\};\ \text{Header} = \{\text{General}, \text{Additional}\};\\ \text{Additional} = \{\text{Vary}, \text{Accept} - \text{Ranges}, \text{RemoteAddr}\},\end{gathered} \tag{16}$$

где Request – пакет запроса от клиента, Header – параметры заголовка запроса, Body – параметры тела запроса, General и Additional – основная и дополнительная части заголовка запроса соответственно, Vary – параметр, характеризующий возможность использования кэшированных ответов, Accept-Ranges – маркер составного запроса, RemoteAddr – IP-адрес хоста, с которого поступил запрос.

Параметр RemoteAddr трансформируется в пару значений типа «широта – долгота». Полученная метка сравнивается последовательно со всеми n зарегистрированными хранилищами данных R на предмет выявления кратчайшего расстояния между ними (по большой дуге):

$$\begin{gathered}\text{RemoteAddr} = \{\text{lat: } \{\text{x}, \text{x\_val}\}, \text{lng: } \{\text{y}, \text{y\_val}\}\}; \\ \min(\ \ |\text{RemoteAddr}, \text{Ri.loc}|, \text{i} = 1, \dots, \text{n}).\end{gathered} \tag{17}$$

На основании полученного результата определяется то хранилище, которое физически ближе остальных находится к источнику запроса. Данное хранилище в текущем сеансе запроса помечается как целевое:

$$\begin{gathered}\text{Target} = R^k : \forall l \neq k \ , \\ |\text{RemoteAddr}, R^l.loc| > \ |\text{RemoteAddr}, R^k.loc|.\end{gathered} \tag{18}$$

На последующем шаге формируется совокупность кортежей данных, передаваемая непосредственно в целевое хранилище Target. При этом выделенная из запроса по IP-адресу геопространственная метка также дополняет каждый из сформированных кортежей.

В результате формируется массив геопространственных данных, к которым, в свою очередь, применимы не только статистические модели и методы, но также и методы геостатистики, позволяющие в том числе оценить особенности соответствующего пространственного распределения соответствующих объектов, субъектов и процессов, а также оценить специфику их пространственной анизотропии при работе непосредственно с глобальным хранилищем.

Помимо геопространственной метки кортежи данных сопровождаются и временной меткой, что позволяет провести их обработку методами анализа пространственно-временных данных и оценить характер их анизотропии не только в пространстве, но и во времени.

В результате соответствующий кортеж, поступивший в запросе от источника данных A_i, при направлении в целевое хранилище будет иметь следующий вид:

$$D^{A_i} = \left\{d_1^{A_i}, \dots, d_m^{A_i}, R^{A_i}, \text{loc}, \text{timetag}\right\}, \quad (19)$$

где R^{A_i}– ссылка на метаданные источника A_i, loc – пространственная характеристика источника данных, заданная по ключу парой координат типа «широта – долгота», timetag – временная метка получения данных в формате UTC.

Для синхронизации с главным информационным хранилищем выполняется периодическая процедура репликации данных. При этом метки источников данных в соответствующих кортежах дополняются меткой соответствующего им целевого хранилища, принявшего запрос на размещение данных:

$$D^{A_i} = \left\{d_1^{A_i}, \dots, d_m^{A_i}, R^{A_i}, \text{loc}, \text{timetag}, \text{target}\right\}, \quad (20)$$

где target – ссылка на метаданные целевого хранилище данных.

Помимо непосредственно медицинских данных геопространственная метка должна сопровождать и запрос на их извлечение, поступающий от потребителя данных и характеризующий его физическое местоположение по результатам процедуры геолокации. Соответствующая информация по аналогии с вышеизложенным форматом запроса содержится в его заголовке (Request, RemoteAddr) и представляет собой IP-адрес источника запроса. На основании IP-адреса выделяется непосредственно геопространственная метка в виде пары геопространственных координат (географические широта и долгота).

Далее (аналогично обработке запроса на загрузку данных) выделенные геопространственные данные последовательно сопоставляются с метаданными распределенных хранилищ медицинских данных. Из них выявляется то хранилище, которое является ближайшим к исходной точке запроса с точки зрения соответствующего расстояния по большой дуге. Выделенное хранилище данных в рамках текущего сеанса взаимодействия с клиентом помечается как целевое.

В целевое хранилище направляется поступивший от клиента запрос на извлечение данных. При этом в результирующий набор кортежей, полученный при выполнении соответствующего запроса, добавляется пространственно-временная метка, характеризующая метаданные хранилища, из которого были извлечены данные, а также время формирования кортежа в формате UTC. Соответствующая информация помещается в заголовок отклика, формируемого на сервере.

Между основным и распределенным хранилищами выполняется процедура репликации данных. По умолчанию данная процедура должна выполняться с заданной периодичностью, например, один раз в сутки (каждые 24 часа). Соответственно обновление кортежей данных во всех хранилищах может быть представлено следующим образом:

$$D(t+\Delta t) \quad = \begin{Bmatrix} d_1\ (t+\Delta t), \dots, d_m(t+\Delta t), R \quad , \\ \text{loc}, (t++\Delta t)(UTC), \text{target} \end{Bmatrix}, \tag{21}$$

где t – временная характеристика запроса на загрузку данных, Δt – промежуток времени, по истечении которого данные должны быть обновлены и дополнены (при необходимости); R – ссылка на метаданные источника данных, loc – пространственная характеристика источника данных, заданная по ключу парой координат типа «широта – долгота», timetag – временная метка получения данных в формате UTC.

Вместе с тем во избежание трудоемких с точки зрения вычислительных затрат последовательных опросов источников «на местах» со стороны центрального хранилища данная процедура может быть дополнена следующим образом.

Представляется целесообразным введение дополнительного программного триггера, срабатывание которого привязано к изменению состава кортежей в соответствующем локальном хранилище информации. При выполнении заданного в программном сценарии условия на обновление данных выполняется процедуры их передачи (и обновления) сначала в центральное хранилище и далее в распределенные хранилища за исключением того, которое было непосредственно причиной срабатывания соответствующего триггера.

При этом во избежание возможных коллизий, связанных с началом одной операции обновления данных в центральном или распределенных хранилищах информации в архитектуру соответствующей информационной системы, целесообразно ввести программный диспетчер, в фоновом режиме отслеживающий входящий и исходящий информационные потоки в хранилищах.

6. Модульная схема интеграции данных. В работе [21] предложена и успешно апробирована схема создания единого информационного пространства геомагнитных данных на основе комбинирования принципов консолидации и федерализации информации. Представляется целесообразным несколько адаптировать предложенную схему под особенности создания и применения

медицинской информации с учетом описанной выше CDN-подобной организации физического хранения данных (рисунок 2).

Формирование централизованного хранилища медицинской информации осуществляется посредством фонового (по принципу Cron) последовательного выполнения связанных вычислительных процессов по получению, обработке и физическому размещению соответствующей информации, поступающей из территориально распределенных гетерогенных источников (в частности, разнородных медицинских информационных систем). При этом принцип Cron [22] предполагает составление графика рабочих процессов («workers»), которые запускаются автоматически (в фоновом режиме) и выполняются после настройки без непосредственного участия разработчика, сохраняя информацию о ходе своей работы в соответствующих программных логах, которые размещены на сервере в заданном каталоге (обычно корневом для проекта) и доступны для анализа.

Каждый этап обработки поступающих от гетерогенных источников медицинских данных представлен в виде одного или группы связанных программных модулей. При этом порядок выполнения вычислительных операций инкапсулируется таким образом, что каждый программный модуль может быть интерпретирован как сторонними, так и внешними потребителями как «черный ящик» с известными форматами представления входной и выходной информации.

Представляется целесообразным в контексте веб-ориентированной реализации рассматриваемого подхода применить элементы микросервисной архитектуры. В этом случае каждый отдельный функциональный модуль обработки данных (или группа связанных модулей) могут быть использованы независимо друг от друга, в том числе и сторонними программными системами и библиотеками. Для этого у каждого из представленных модулей-сервисов предполагается наличие собственного программного интерфейса (по принципу организации API – прикладных программных интерфейсов).

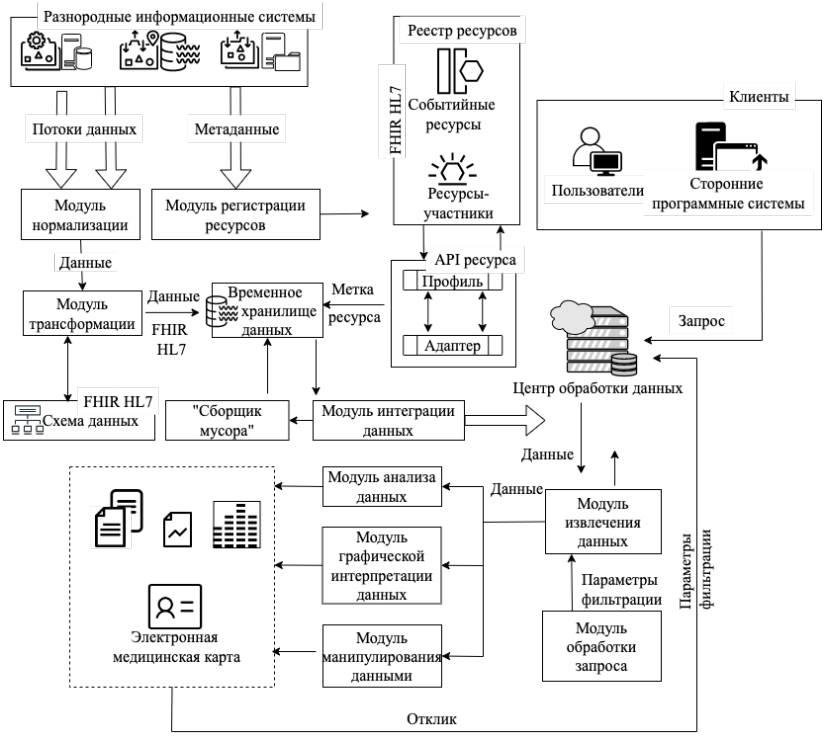

Рис. 2. Общая схема реализации предлагаемого подхода к интеграции данных

В результате обращение к каждому программному модулю осуществляется по его интерфейсу таким образом, что соответствующий вызов выполняется по названию (метке) интерфейса с передачей значений необходимых входных параметров и возвратом искомого результата по установленному протоколу. В качестве основного формата представления выходных данных используется JSON (JavaScript Object Notation – это формат, реализующий неструктурированное текстовое представление структурированных данных, основанное на принципе пар ключ-значение и упорядоченных списках [24]), который в настоящее время является фактически стандартом де-факто передачи данных в программных системах. В предлагаемой схеме выделены два типа данных – непосредственно содержательная информация (сами медицинские данные) и метаданные (информация об источниках данных). При этом содержательная информация, поступающая от внешних источников,

может быть представлена в структурированном, полуструктурированном и неструктурированном виде (в текстовом, CSV-подобных, мультимедиа, графических и иных форматах). В этой связи непосредственно данные непосредственно от своих провайдеров поступают в модуль нормализации, где проходят предварительную обработку для приведения к унифицированному виду. Так, к примеру, графические и мультимедиа данные преобразуются в BLOB-объекты [25], представляющие собой пригодные для обработки и анализа массивы двоичных данных, которые, как правило, применяются для хранения данных MIME-типа.

При этом для повышения эффективности дальнейшей обработки данные представляются в результирующем JSON-формате, что позволяет применять к ним библиотечные модели и методы трансформации. При этом сформированные BLOB-массивы также становятся частью выходных JSON-потоков в качестве секций формата CDATA (известного из спецификации W3C XML).

На последующем этапе данные формата JSON преобразуются в формат, специфичный для медицинских данных. В качестве такового выступает стандарт HL7, который в настоящее время применяется во многих медицинских учреждения в нашей стране и за рубежом для электронного обмена документами (особенно в тех медицинских организациях, где пациент получает непосредственно интенсивную помощь, в частности, в больницах). HL7 включает в себя концептуальные стандарты (HL7 RIM), стандарты приложений (HL7 CCOW), документальные стандарты (HL7 CDA), и стандарты обмена сообщениями (HL7 v2., v3.0 и HL7 FHIR). Согласно известным научным работам (например, [26]), самым актуальным и многообещающим из этих стандартов является HL7 FHIR (Health Level 7 - Fast Healthcare Interoperability Resources) [26].

Преобразование JSON-данных в указанный формат осуществляется на основании соответствующей схемы данных, используемой модулем трансформации. В схеме данных представлены основные параметры трансформации / отображения для последующего формирования результирующего набора данных.

Основой стандарта HL7 FHIR являются ресурсы (FHIR Resources). При этом каждый ресурс представляет собой отдельную независимую структурированную единицу информации, используемая при обмене медицинскими данными. Большинство ресурсов — это отображение реального мира в цифровых данных. Так, к примеру, на верхнем уровне абстракции предлагается выделять такие виды

ресурсов, как пациенты, обращения в медицинские организации и учреждения, а также результаты осмотров и исследований.

При дальнейшей детализации ресурсы декомпозируются на метаданные соответствующих провайдеров данных, которыми выступают конкретные медицинские информационные системы отдельных медицинских организаций и учреждений. Формирование ресурсов осуществляется вычислительным модулем регистрации ресурсов, на вход которого поступают метаданные источников медицинской информации, направляемой в рассматриваемое централизованное хранилище данных.

При этом предлагается выделение двух видов ресурсов. Первая группа обозначается как «событийные ресурсы» и включает в себя результаты всех действий, осуществляемых между пациентом и медицинским учреждением. Здесь могут быть представлены результаты осмотров, лабораторных исследований, программы лечения, обращения пациента в лечебное учреждение и пр.

Вторая группа ресурсов представляет собой характеристику участников взаимодействия в рассматриваемых процессах. Это непосредственно пациенты, медицинские учреждения, медицинский персонал, назначенные лечебные препараты и мероприятия, используемое лабораторное оборудования (фактически, здесь представлены как активные, так и пассивные участники медицинских процессов).

Все сформированные и зарегистрированные ресурсы размещаются в одноименном реестре, данные в котором представлены в соответствии со стандартом HL7 FHIR. При этом для унификации доступа к ресурсам каждому из них ставится в соответствие отдельный API, отличительной особенностью которого является самоидентифицируемость за счет представления метаданных в соответствующей структуре, именуемой профилем ресурса. Для программного доступа к ресурсу через его API используется его профиль, так называемая точка входа, принимающая входные данные и выдающая в результате уникальную метку соответствующего ресурса. Метка представляет собой свертку JSON-данных в структуру по спецификации JSON Web Token (JWT) [27, 28], что в итоге имеет вид объекта, который определен в открытом стандарте RFC 7519 [29].

Фоновое выполнение процесса трансформации содержательных данных предполагает наличие промежуточного (временного) хранилища, в котором (помимо хранения) выполняется дополнительная обработка данных, связанная с объединением

непосредственно данных и метки ресурса, характеризующего обстоятельства и источники их получения.

Данные из временного хранилища с заданной периодичностью поступают в модуль интеграции данных, в которой информационные потоки объединяются в единый массив и направляются в централизованное хранилище (центр обработки данных). Важно отметить, что в схему интеграции дополнительно включен так называемый «сборщик мусора», выполнение которого запускается завершением работы модуля интеграции данных [30, 31]. Указанный программный модуль отвечает за очищение временного хранилища данных после передачи обработанных данных в централизованное хранилище.

Непосредственно обработка пользовательских запросов (включая управление CDN-серверами) осуществляется посредством группы модулей, отвечающих за получение и декомпозицию поступивших с клиентской стороны запросов. Выделенные параметры запросы являются базой для выполнения вертикальной и горизонтальной фильтрации данных, размещенных в централизованном хранилище. Результаты извлечения данных поступают на вход одному или группе программных модулей, отвечающих за их анализ и интерпретацию конечному пользователю в виде наборов данных, графиков, таблиц, либо в формате электронной медицинской карты соответствующего пациента. Результаты интерпретации передаются на клиентскую сторону в качестве отклика и доступны как непосредственно для рендеринга, так и для применения сторонними средствами обработки данных.

7. Оценка эффективности предложенного подхода. Возможности предложенного подхода к интеграции данных из гетерогенных источников медицинской информации были проанализированы на примере тестовых данных, сформированных по аналогии с данными пациентов с бронхолегочными заболеваниями трех городских клинических больниц г. Уфа.

Данные представлены наборами значений параметров обследования 500 гипотетических пациентов и содержат информацию в текстовом и графическом форматах (составлены анонимизированные тестовые данные по каждому пациенту, характерные для программ лечения в рассмотренных медицинских учреждениях: электронные медицинские карты по каждому пациенту в формате XML, данные МКБ в формате CSV, биохимические и лабораторные тесты (в текстовом и графическом форматах), клинические заметки медицинского персонала (в текстовом формате), программы лечения

(в текстовом формате) и назначения фармацевтических препаратов (в текстовом формате)). Для повышения наглядности были развернуты тестовые сервера для хранения данных на виртуальном хостинге Beget со следующими характеристиками: веб-сервер с процессором 72 * Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz, Apache 2.4.5.1, сетевое журналируемое хранилище данных типа «ключ» – «значение» Redis (REmote DIctionary Server) [32].

Для имитации применения нескольких разнородных систем хранения созданы аналогичные копии сервера данных (в общей сложности были развернуты 4 копии), на которых в случайном порядке были размещены медицинские данные. Дополнительно также было выделено серверное хранилище с аналогичными техническими характеристиками для имитации централизованного хранилища (так называемого основного сервера), предназначенного для хранения результатов интеграции данных из соответствующих гетерогенных источников.

Вычислительный эксперимент проводился в несколько этапов для данных небольшого объема (порядка 150-200 Мб). На первом этапе осуществлялась загрузка данных с сервера, принятого за сервер отдельной организации. Далее информационный поток через обработку и унификацию данных перенаправлялся в хранилище следующего уровня (имитирующего региональный уровень). На завершающем этапе данные передаются на основной сервер.

Эксперименты были проведены по двум направлениям. С одной стороны, исследовались характеристики реактивности в случае применения стандартной монолитной веб-ориентированной архитектуры. С другой стороны, вычислительный эксперимент был проведен в рамках архитектуры по предложенному подходу к интеграции данных. Клиентская сторона при этом тестировалась со следующим характеристиками: CPU Intel Core i5 10300H ГГц, оперативная память 4 ГБ, скорость интернет-соединения ~52.4 Мбит/с. Результаты вычислительного эксперимента показали, что подход с монолитной архитектурой обеспечивает получение серверного отклика за 10-12 с при стабильном Интернет-соединении. Предложенный подход за счет применения CDN позволяет сократить значение указанного параметра до 6-7 с.

Вторая часть экспериментов с тестовыми данными была направлена на оценку изменения скорости получения серверного отклика при направлении простого одномерного запроса к центральному хранилищу. Подход, основанный на монолитной архитектуре и отсутствии территориального распределения системы

хранения данных, показал значение указанного параметра в 15-17 с, а предложенный подход с микросервисной архитектурой и CDN – 9-10 с. При этом были установлены пути повышения реактивности соответствующих веб-ориентированных приложений посредством применения технологии Node.js на сервере (предназначенной для работы с высоконагруженными приложениями) для определения ближайшей точки доступа с расчетом соответствующих расстояний по географическим координатам клиента [33].

8. Заключение. В настоящее время одним из магистральных направления развития информационных технологий является разработка систем цифрового здравоохранения, предназначенных в том числе для повышения эффективности процессов принятия решений на основе медицинских данных. Одним из важнейших факторов успешности реализации данного направления является повышение эффективности и прозрачности обработки соответствующих данных, а также реактивности соответствующих инструментально-программных средств. Значимым препятствием на пути решения поставленной задачи является постоянно возрастающие объем и сложность медицинских данных, а также разнообразие их источников. В конечном итоге требуется разработка подхода, который в результате своей программной реализации позволил бы решить поставленные задачи.

В работе предлагается подход к интеграции разнородных источников медицинской информации, отличительной особенностью которого является применение в архитектуре соответствующей программной системы принципов организации микросервисных приложений. Предложена многоуровневая организация системы сбора, обработки и хранения медицинской информации с дополнительным предоставлением метаданных, характеризующих источники информации и обстоятельства их получения (к примеру, используемое лабораторное оборудование и пр.).

Представляется целесообразным отметить еще одну отличительную особенность предложенного подхода, которая заключается в применении системы доставки контента (CDN) для организации системы хранения данных. Предложено создание копий основного сервера хранения данных с различной геопространственной привязкой, которая является определяющим фактором при направлении клиентского запроса на получение / загрузку соответствующей медицинской информации. Также

предусматривается (с заданной периодичностью) репликация данных между основным и зеркальными серверами-хранилищами данных.

Результаты проведенных вычислительных экспериментов (на тестовых данных, автоматически сформированных на основе реальных по бронхолегочным заболеваниям) в заданных технических условиях клиент-серверного веб-ориентированного взаимодействия показали, что применение предложенного подхода к решению перечисленных выше проблем позволит существенно сократить время получения отклика от серверной части (по сравнению с традиционным монолитным подходом) при загрузке данных в среднем на 39 % и при выполнении запросов на выборку – на 41%.

Литература

1. Snyder M., Zhou W. Big data and health // The Lancet. Digital Health. 2019. Vol. 1, iss. 6. P. E252-E-254
2. Комолов А.В. Обзор медицинских стандартов передачи электронной информации // Аллея науки. 2019. Т. 2. № 2(29). С. 909-913
3. Martínez-Costa C., Schulz S. HL7 FHIR: Ontological Reinterpretation of Medication Resources // Studies in Health Technology and Informatics. 2017. No. 235. P. 451–455. doi:10.3233/978-1-61499-753-5-451.
4. Mukhiya S., Rabbi F., Pun V. [et al.]. A GraphQL approach to Healthcare Information Exchange with HL7 FHIR // Procedia Computer Science. 2019. No. 160. P.338-345. doi:10.1016/j.procs.2019.11.082.
5. Hong N., Wang K., Wu S. [et al.] An Interactive Visualization Tool for HL7 FHIR Specification Browsing and Profiling // Journal of Healthcare Informatics Research. 2019. No. 3. doi:10.1007/s41666-018-0043-8.
6. Елоев М.С. Опыт внедрения медицинской информационной системы в многопрофильном амбулаторно-поликлиническом учреждении // Военно-медицинский журнал. 2014. Т. 335. № 9. С. 4-13
7. Alqudah A., Al-Emran M., Shaalan K. Medical data integration using HL7 standards for patient's early identification // PLOS ONE. 2021. No. 16. P. e0262067. doi:10.1371/journal.pone.0262067.
8. Brogan J., del Pilar M., López A. [et al.] Scalable data systems require creating a culture of continuous learning // EBioMedicine Home (Part of Lancet Discovery Science). 2021. Vol. 74, P. 103738, doi: https://doi.org/10.1016/j.ebiom.2021.103738
9. Prakash C., Amit Sh. National Institute of Malaria Research-Malaria Dashboard (NIMR-MDB): A digital platform for analysis and visualization of epidemiological data // The Lancet Regional Health. 2022. P. 100030.
10. Balicer R., Arnon A. Digital health nation: Israel's global big data innovation hub // The Lancet. 2017. Vol. 389, iss. 10088, p. 2451-2453. doi: https://doi.org/10.1016/S0140-6736(17)30876-0
11. Grabner M, Molife C, Wang L, Winfree K, Cui Z, Cuyun Carter G, Hess L. Data Integration to Improve Real-world Health Outcomes Research for Non–Small Cell Lung Cancer in the United States: Descriptive and Qualitative Exploration // JMIR Cancer 2021;7(2):e23161. DOI: 10.2196/23161
12. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H-U, Bürkle T,et al. Ontology-Based Data Integration between Clinical and Research Systems // PLoS ONE. 2015. No. 10(1). P. e0116656. pmid:25588043.

13. Lin YL, Trbovich P, Kolodzey L, Nickel C, Guerguerian A. Association of Data Integration Technologies With Intensive Care Clinician Performance: A Systematic Review and Meta-analysis // JAMA Netw Open. 2019. No. 2(5). P. e194392. doi:10.1001/jamanetworkopen.2019.4392.
14. Scheurwegs E., Luyckx K. [et al.]. Data integration of structured and unstructured sources for assigning clinical codes to patient stays // Journal of the American Medical Informatics Association. 2016. Vol. 23, Iss. e1. P. e11–e19, https://doi.org/10.1093/jamia/ocv115
15. Martínez-García M., Hernández-Lemus E. Data Integration Challenges for Machine Learning in Precision Medicine // Front. Med. 2022. No. 8:784455. doi: 10.3389/fmed.2021.784455.
16. Di Stefano A., La Corte A., Scatá M. Health Mining: a new data fusion and integration paradigm // Proceedings of CIBB. 2014. Vol. 1. P. 98-107.
17. Kamdar M.R., Fernández J.D., Polleres A. [et al.] Enabling Web-scale data integration in biomedicine through Linked Open Data // Digit. Med. 2019. No. 2. P. 90. https://doi.org/10.1038/s41746-019-0162-5
18. Dhayne H., Haque R., Kilany R., Taher Y. In Search of Big Medical Data Integration Solutions. A Comprehensive Survey // IEEE Access. 2019. PP. 1-10. doi:10.1109/ACCESS.2019.2927491.
19. Kük E., Erel-Ozcevik M. Access protocol aware controller design for eMBB traffic in SD-CDN // Computer Networks. 2022. No. 205. P. 08686. doi:10.1016/j.comnet.2021.108686.
20. Zerwas J., Poese I., Schmid S., Blenk A. On the Benefits of Joint Optimization of Reconfigurable CDN-ISP Infrastructure // IEEE Transactions on Network and Service Management. 2021. PP. 105-112. Doi:10.1109/TNSM.2021.3119134.
21. Vorobev, A.; Soloviev, A.; Pilipenko, V.; Vorobeva, G.; Sakharov, Y. An Approach to Diagnostics of Geomagnetically Induced Currents Based on Ground Magnetometers Data // Appl. Sci. 2022, 12, 1522. https://doi.org/10.3390/app12031522
22. Choi, B. Python Network Automation Labs: cron and SNMPv3. In: Introduction to Python Network Automation. Apress, Berkeley, CA, 2021.doi:10.1007/978-1-4842-6806-3_15.
23. Vorobev, A.V., Pilipenko, V.A., Enikeev, T.A., Vorobeva, G.R. Geoinformation system for analyzing the dynamics of extreme geomagnetic disturbances from observations of ground stations // Computer Optics. 2020. No. 44(5). P. 782–790.
24. Barlas K., Stefaneas P. An Algebraic Specification / Schema for JSON // Journal of Engineering Research and Sciences. 2022. No. 1. doi:10.55708/js0105025.
25. Rajendran L., Veilumuthu R. An Efficient Distributed Model for XMLised Blob Data Generation // International Journal of Computer Applications. 2011. No. 22. doi:10.5120/2561-3519.
26. Yang Z., Jiang K., Lou M. [et al.] Defining health data elements under the HL7 development framework for metadata management // Journal of Biomedical Semantics. 2022. No. 13. doi:10.1186/s13326-022-00265-5.
27. Rahmatulloh A., Gunawan R., Nursuwars F. Performance comparison of signed algorithms on JSON Web Token // IOP Conference Series: Materials Science and Engineering. 2019. No. 550. P. 012023. doi:10.1088/1757-899X/550/1/012023.
28. Beltran V. Characterization of web single sign-on protocols // IEEE Communications Magazine. 2016. No. 54. P. 24-30. doi:10.1109/MCOM.2016.7514160.
29. Jones M., Bradley J., Sakimura N., JSON Web Token (JWT)., RFC 7519, doi:10.17487/RFC7519, May 2015, https://www.rfc-editor.org/info/rfc7519.

30. Cai Sh., Chen K., Liu M. [et al.] Garbage collection and data recovery for N2DB // Tsinghua Science and Technology. 2022. No. 27. P. 630-641. doi:10.26599/TST.2021.9010016.
31. Garcia A., May D., Nutting E. Integrated Hardware Garbage Collection // ACM Transactions on Embedded Computing Systems. 2021. No. 20. P. 1-25. doi:10.1145/3450147.
32. Zhang Q., Bernstein P., Berger D., Chandramouli B. Redy: remote dynamic memory cache // Proceedings of the VLDB Endowment. 2021. No. 15. P. 766-779. doi:10.14778/3503585.3503587.
33. Tserpes K., Pateraki M., Varlamis I. Strand: scalable trilateration with Node.js // Journal of Cloud Computing. 2019. No. 8. doi:10.1186/s13677-019-0142-y.

Юсупова Нафиса Исламовна — д-р техн. наук, профессор, кафедра вычислительной математики и кибернетики, ФГБОУ ВО Уфимский государственный авиационный технический университет. Область научных интересов: интеллектуальные методы обработки информации и управления с приложениями в социальных, экономических и технических системах. Число научных публикаций — 560. yussupova@ugatu.ac.ru; улица Карла Маркса, 12, 450008, Уфа, Россия; р.т.: +7(908)350-3285.

Воробьева Гульнара Равилевна — д-р техн. наук, профессор, доцент, кафедра вычислительной математики и кибернетики, ФГБОУ ВО Уфимский государственный авиационный технический университет. Область научных интересов: геоинформационные и веб-технологии, системы хранения и обработки информации. Число научных публикаций — 141. gulnara.vorobeva@gmail.com; улица Карла Маркса, 12, 450008, Уфа, Россия; р.т.: +7(908)350-3285.

Зулкарнеев Рустэм Халитович — д-р мед. наук, профессор, кафедра пропедевтики внутренних болезней, ФГБОУ ВО Башкирский государственный медицинский университет Минздрава России. Область научных интересов: пульмонология, медицинские информационные системы. Число научных публикаций — 30. zurustem@mail.ru; улица Карла Маркса, 9/1, 450076, Уфа, Россия; р.т.: +7(908)350-3285.

Поддержка исследований. Работа выполнена при финансовой поддержке РНФ (проект № 22-19-00471).

DOI 10.15622/ia.21.5.2

N. YUSUPOVA, G. VOROBEVA, R. ZULKARNEEV

APPROACH TO SOFTWARE INTEGRATION OF HETEROGENEOUS SOURCES OF MEDICAL DATA BASED ON MICROSERVICE ARCHITECTURE

Yusupova N., Vorobeva G., Zulkarneev R. **Approach to Software Integration of Heterogeneous Sources of Medical Data Based on Microservice Architecture.**

Abstract. The task of processing medical information is currently being solved in our country and abroad by means of heterogeneous medical information systems, mainly at the local and regional levels. The ever-increasing volume and complexity of the accumulated information, along with the need to ensure transparency and continuity in the processing of medical data (in particular, for bronchopulmonary diseases) in various organizations, requires the development of a new approach to integrating their heterogeneous sources. At the same time, an important requirement for solving the problem is the possibility of web-oriented implementation, which will make the corresponding applications available to a wide range of users without high requirements for their hardware and software capabilities. The paper considers an approach to the integration of heterogeneous sources of medical information, which is based on the principles of building microservice web architectures. Each data processing module can be used independently of other program modules, providing a universal entry point and the resulting data set in accordance with the accepted data schema. Sequential execution of processing steps implies the transfer of control to the corresponding program modules in the background according to the Cron principle. The schema declares two types of data schemas - local (from medical information systems) and global (for a single storage system), between which the corresponding display parameters are provided according to the principle of constructing XSLT tables. An important distinguishing feature of the proposed approach is the modernization of the medical information storage system, which consists in creating mirror copies of the main server with periodic replication of the relevant information. At the same time, the interaction between clients and data storage servers is carried out according to the type of content delivery systems with the creation of a connection session between end points based on the principle of the nearest distance between them, calculated using the haversine formula. The computational experiments carried out on test data on bronchopulmonary diseases showed the effectiveness of the proposed approach both for loading data and for obtaining them by individual users and software systems. Overall, the reactivity score of the corresponding web-based applications was improved by 40% on a stable connection.

Keywords: medical data, data warehouses, data integration, web applications, content delivery system, microservice architecture.

Yusupova Nafisa — Ph.D., Dr.Sci., Professor, Computational mathematics and cybernetics department, Ufa State Aviation Technical University. Research interests: intelligent methods of information processing and management with applications in social, economic and technical systems. The number of publications — 560. yussupova@ugatu.ac.ru; 12, Karl Marx St., 450008, Ufa, Russia; office phone: +7(908)350-3285.

Vorobeva Gulnara — Ph.D., Dr.Sci., Professor, Associate professor, Computational mathematics and cybernetics department, Ufa State Aviation Technical University. Research interests: geoinformation and web technologies, systems of information storing and processing.

The number of publications — 141. gulnara.vorobeva@gmail.com; 12, Karl Marx St., 450008, Ufa, Russia; office phone: +7(908)350-3285.

Zulkarneev Rustem — Ph.D., Dr.Sci., Professor, Department of propaedeutics of internal diseases, Bashkir State Medical University of the Ministry of Health of Russia. Research interests: pulmonology, medical information systems. The number of publications — 30. zurustem@mail.ru; 9/1, Karl Marx St., 450076, Ufa, Russia; office phone: +7(908)350-3285.

Acknowledgements. This research is supported by RSF (grant 22-19-00471).

References

1. Snyder M., Zhou W. Big data and health. The Lancet. Digital Health. 2019. Vol. 1, iss. 6. P. E252-E-254
2. Komolov A.V. [Review of medical standards for the transmission of electronic information]. Alleya nauki – Alley of Science. 2019. vol. 2. no. 2(29). pp. 909-913 (in Russ.)
3. Martínez-Costa C., Schulz S. HL7 FHIR: Ontological Reinterpretation of Medication Resources. Studies in Health Technology and Informatics. 2017. no. 235. pp. 451–455. doi:10.3233/978-1-61499-753-5-451.
4. Mukhiya S., Rabbi F., Pun V. [et al.]. A GraphQL approach to Healthcare Information Exchange with HL7 FHIR. Procedia Computer Science. 2019. no. 160. pp.338-345. doi:10.1016/j.procs.2019.11.082.
5. Hong N., Wang K., Wu S. [et al.] An Interactive Visualization Tool for HL7 FHIR Specification Browsing and Profiling. Journal of Healthcare Informatics Research. 2019. No. 3. doi:10.1007/s41666-018-0043-8.
6. Eloev M.S. [Experience in implementing a medical information system in a multidisciplinary outpatient clinic]. Voenno-medicinsky Journal – Military Medical Journal. 2014. vol. 335. no. 9. pp. 4-13
7. Alqudah A., Al-Emran M., Shaalan K. Medical data integration using HL7 standards for patient's early identification. PLOS ONE. 2021. no. 16. p. e0262067. doi:10.1371/journal.pone.0262067.
8. Brogan J., del Pilar M., López A. [et al.] Scalable data systems require creating a culture of continuous learning. EBioMedicine Home (Part of Lancet Discovery Science). 2021. Vol. 74, P. 103738, doi: https://doi.org/10.1016/j.ebiom.2021.103738
9. Prakash C., Amit Sh. National Institute of Malaria Research-Malaria Dashboard (NIMR-MDB): A digital platform for analysis and visualization of epidemiological data. The Lancet Regional Health. 2022. P. 100030.
10. Balicer R., Arnon A. Digital health nation: Israel's global big data innovation hub. The Lancet. 2017. Vol. 389, iss. 10088, p. 2451-2453. doi: https://doi.org/10.1016/S0140-6736(17)30876-0
11. Grabner M, Molife C, Wang L, Winfree K, Cui Z, Cuyun Carter G, Hess L. Data Integration to Improve Real-world Health Outcomes Research for Non–Small Cell Lung Cancer in the United States: Descriptive and Qualitative Exploration. JMIR Cancer 2021;7(2):e23161. DOI: 10.2196/23161
12. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H-U, Bürkle T,et al. Ontology-Based Data Integration between Clinical and Research Systems. PLoS ONE. 2015. No. 10(1). P. e0116656. pmid:25588043.
13. Lin YL, Trbovich P, Kolodzey L, Nickel C, Guerguerian A. Association of Data Integration Technologies With Intensive Care Clinician Performance: A Systematic Review and Meta-analysis // JAMA Netw Open. 2019. No. 2(5). P. e194392. doi:10.1001/jamanetworkopen.2019.4392.

14. Scheurwegs E., Luyckx K. [et al.]. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. Journal of the American Medical Informatics Association. 2016. Vol. 23, Iss. e1. P. e11–e19, https://doi.org/10.1093/jamia/ocv115
15. Martínez-García M., Hernández-Lemus E. Data Integration Challenges for Machine Learning in Precision Medicine. Front. Med. 2022. No. 8:784455. doi: 10.3389/fmed.2021.784455.
16. Di Stefano A., La Corte A., Scatá M. Health Mining: a new data fusion and integration paradigm. Proceedings of CIBB. 2014. Vol. 1. P. 98-107.
17. Kamdar M.R., Fernández J.D., Polleres A. [et al.] Enabling Web-scale data integration in biomedicine through Linked Open Data. Digit. Med. 2019. No. 2. P. 90. https://doi.org/10.1038/s41746-019-0162-5
18. Dhayne H., Haque R., Kilany R., Taher Y. In Search of Big Medical Data Integration Solutions. A Comprehensive Survey. IEEE Access. 2019. PP. 1-10. doi:10.1109/ACCESS.2019.2927491.
19. Kük E., Erel-Ozcevik M. Access protocol aware controller design for eMBB traffic in SD-CDN. Computer Networks. 2022. No. 205. P. 08686. doi:10.1016/j.comnet.2021.108686.
20. Zerwas J., Poese I., Schmid S., Blenk A. On the Benefits of Joint Optimization of Reconfigurable CDN-ISP Infrastructure. IEEE Transactions on Network and Service Management. 2021. PP. 105-112. Doi:10.1109/TNSM.2021.3119134.
21. Vorobev, A.; Soloviev, A.; Pilipenko, V.; Vorobeva, G.; Sakharov, Y. An Approach to Diagnostics of Geomagnetically Induced Currents Based on Ground Magnetometers Data. Appl. Sci. 2022, 12, 1522. https://doi.org/10.3390/app12031522.
22. Choi, B. Python Network Automation Labs: cron and SNMPv3. In: Introduction to Python Network Automation. Apress, Berkeley, CA, 2021.doi:10.1007/978-1-4842-6806-3_15.
23. Vorobev, A.V., Pilipenko, V.A., Enikeev, T.A., Vorobeva, G.R. Geoinformation system for analyzing the dynamics of extreme geomagnetic disturbances from observations of ground stations. Computer Optics. 2020. No. 44(5). P. 782–790.
24. Barlas K., Stefaneas P. An Algebraic Specification / Schema for JSON. Journal of Engineering Research and Sciences. 2022. No. 1. doi:10.55708/js0105025.
25. Rajendran L., Veilumuthu R. An Efficient Distributed Model for XMLised Blob Data Generation. International Journal of Computer Applications. 2011. No. 22. doi:10.5120/2561-3519.
26. Yang Z., Jiang K., Lou M. [et al.] Defining health data elements under the HL7 development framework for metadata management. Journal of Biomedical Semantics. 2022. No. 13. doi:10.1186/s13326-022-00265-5.
27. Rahmatulloh A., Gunawan R., Nursuwars F. Performance comparison of signed algorithms on JSON Web Token. IOP Conference Series: Materials Science and Engineering. 2019. No. 550. P. 012023. doi:10.1088/1757-899X/550/1/012023.
28. Beltran V. Characterization of web single sign-on protocols. IEEE Communications Magazine. 2016. No. 54. P. 24-30. doi:10.1109/MCOM.2016.7514160.
29. Jones M., Bradley J., Sakimura N., JSON Web Token (JWT)., RFC 7519, doi:10.17487/RFC7519, May 2015, https://www.rfc-editor.org/info/rfc7519.
30. Cai Sh., Chen K., Liu M. [et al.] Garbage collection and data recovery for N2DB. Tsinghua Science and Technology. 2022. No. 27. P. 630-641. doi:10.26599/TST.2021.9010016.
31. Garcia A., May D., Nutting E. Integrated Hardware Garbage Collection. ACM Transactions on Embedded Computing Systems. 2021. No. 20. P. 1-25. doi:10.1145/3450147.

32. Zhang Q., Bernstein P., Berger D., Chandramouli B. Redy: remote dynamic memory cache. Proceedings of the VLDB Endowment. 2021. No. 15. P. 766-779. doi:10.14778/3503585.3503587.
33. Tserpes K., Pateraki M., Varlamis I. Strand: scalable trilateration with Node.js. Journal of Cloud Computing. 2019. No. 8. doi:10.1186/s13677-019-0142-y.