На правах рукописи

Чан Ван Фу

МЕТОДЫ ОБРАБОТКИ РАЗНОРОДНЫХ ДАННЫХ В ПРОАКТИВНЫХ СИСТЕМАХ УПРАВЛЕНИЯ ТРАНСПОРТНОЙ ИНФРАСТРУКТУРОЙ

05.13.01 - Системный анализ, управление и обработка информации
(информационные технологии и промышленность)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Волгоград – 2019

Работа выполнена на кафедре «Системы автоматизированного проектирования и поискового конструирования» в федеральном государственном бюджетном образовательном учреждении высшего образования «Волгоградский государственный технический университет» Министерства науки и высшего образования Российской Федерации.

Научный руководитель	доктор технических наук, доцент **Щербаков Максим Владимирович**.
Официальные оппоненты:	**Чопоров Олег Николаевич**, доктор технических наук, профессор, ФГБОУ ВО «Воронежский государственный технический университет», кафедра систем информационной безопасности, профессор; **Иващенко Антон Владимирович**, доктор технических наук, профессор, ФГБОУ ВО «Самарский государственный технический университет», кафедра «Вычислительная техника», заведующий кафедрой.
Ведущая организация	ФГБОУ ВО «Астраханский государственный технический университет».

Защита диссертации состоится 26 июня 2019 г. в ___ часов ___ минут на заседании диссертационного совета Д 212.028.08, созданного на базе Волгоградского государственного технического университета по адресу: 400005, г. Волгоград, пр. им. Ленина, 28, ауд. В-1001.
С диссертацией можно ознакомиться в библиотеке Волгоградского государственного технического университета и на сайте www.vstu.ru.

Автореферат разослан «_____»_____ 2019 г.

Ученый секретарь
диссертационного совета Орлова Юлия Александровна

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Проактивные системы поддержки принятия решений являются развитием компьютерных систем поддержки принятия управленческих решений (СППР) за счет (i) применения принципа проактивных вычислений, т.е. переноса действий человека на более высокий уровень управления и (ii) использования алгоритмов интеллектуальной обработки данных и машинного обучения. Принятие проактивных или предупреждающих решений может привести к экономии средств при управлении процессами, в частности, управления транспортной инфраструктурой на основе анализа больших объемов разнородных данных. В концепции проактивных систем лежит схема: обнаружить – спрогнозировать – принять решение –действовать и принципы построения СППР на основе обработки событий.

При реализации проактивного управления транспортной инфраструктурой критической задачей является задача эффективной обработки разнородных данных, получаемых с различных источников. Качество и своевременность данных влияет на оперативность и результативность принятия решений. В современной научной дискуссии присутствует вопрос о разработке подходов эффективного сбора и обработки данных, поиска новых способов хранения и предварительного анализа. Экспоненциальный рост объема данных и увеличение сетевой полосы пропускания, предоставляемой для передачи данных, открывает новые возможности управления транспортной инфраструктурой, но при этом возникают проблемы эффективной обработки данных.

По способу расположения источников данных следует говорить о (i) сосредоточенных данных (один централизованный источник) и (ii) территориально-распределенных данных. Под разнородными данными понимаются данные, имеющие различную структуру. В результате анализа литературы и реальных бизнес-процессов выделена следующая классификация типов данных: машинные данные сенсорного типа (sensors data), событийные данные (log data), визуальные данные, текстовые данные (textual data), данные социальных сетей (social data), георапределенные данные (geospatial data).

Для решения задач проактивного управления транспортной инфраструктурой необходимо решить вопрос сбора, хранения и обеспечения эффективного (универсального доступа) к разнородным данным большого объема. Важнейшим фактором здесь является время доступа к разнородным данным в процедурах принятия решений.

Степень научной разработанности темы. Вопросами синтеза методов и технологий проактивной поддержки принятия решений, в том числе и для интеллектуального управления транспортной инфраструктурой, занимались следующие ученые: Давид Тенненхаус (D Tennenhouse), Шилов Н.Г., Мотиенко А.И., Тарасов А.Г., Охтелев М.Ю., Вонт Р. Разработку подходов предсказательного моделирования вели следующие ученные: Аверин Г.В., Звягинцева А.В., Бетелин В.Б. Изучением вопросов сбора и слияния разнородных данных в системах поддержки принятия решений занимались ученые: Буймистрюк Г., Коваленко В.Н., Лебедев С.В., Слипецкий Д.Я., Бахадор Халедги (Bahador Khaleghi), Алаа Хамис (Alaa Khamis), Фахаредаин О. Карай (Fakhareddine O. Karray), Давис Л. Халл (David L. Hall), Джемс Линас (James Llinas),

Сарвес Рават (Sarvesh Rawat), Сурабги Рават (Surabhi Rawat), Иван Мигуел Пирес (Ivan Miguel Pires), Нуно М. Гарсиа (Nuno M. Garcia), Джоши Р. (R. Joshi).

Следует отметить, что в исследованиях проблема сбора, слияния и предварительной обработки разнородных данных при реализации системы поддержки принятия решений остается актуальной.

Объектом исследования является процесс сбора и обработки разнородных данных в проактивных системах управления транспортной инфраструктурой.

Предметами исследования является совокупность методов сбора и предварительной обработки разнородных данных в проактивных системах управления транспортной инфраструктурой.

Методология и методы исследования. В процессе выполнения работы были применены следующие методы: системного анализа, сбора и обработки разнородных данных, в том числе лог-данных, видеопотоков данных, математической статистики, параллельных высокопроизводительных вычислений.

Цель исследования. Целью работы является повышение эффективности процесса сбора и предварительной обработки данных в проактивных системах управления транспортной инфраструктурой за счет разработки модели и методов сбора и распределенной обработки разнородных данных. Под эффективностью понимается скорость доступа к данным в процессе принятия управленческих решений.

Для достижения поставленной цели были сформулированы следующие задачи.

1. Выполнить системный анализ процессов сбора и слияния разнородных данных в проактивных системах управления транспортной инфраструктурой.

2. Выполнить обзор и анализ современного состояния исследований в области методов сбора и предварительной обработки разнородных данных, а также технологических стеков обработки больших данных.

3. Разработать модель и методы сбора и предварительной обработки разнородных данных в проактивных системах управления транспортной инфраструктурой.

4. Выполнить проектирование и разработать программное обеспечение, реализующее предложенные методы обработки информации;

5. Провести испытания предлагаемых методов, программного продукта и обосновать эффективность предлагаемых положений.

Методология и методы исследования: системный анализ, теория принятия решений, математическая статистика, методы распределенных и параллельных вычислений, методы машинного обучения и интеллектуальной обработки данных.

Научная новизна. заключается в совокупности модели и методов сбора и предварительной обработки разнородных данных в проактивных системах управления транспортной инфраструктурой, включающей в себя:

1. модель хранения разнородных данных в соответствии с концепцией «озеро данных», которая отличается новыми структурами шаблона объекта и шаблона параметров объекта, что позволяет распределено хранить как необработанные, так и структурированные разнородные данные, в соответствии с предопределенной схемой, и приводит к сокращению времени доступа к данным;

2. метод сбора и предварительной обработки данных, поступающих из разнородных источников, основным отличием которого является впервые реализованные

механизмы преобразования данных к требуемому формату непосредственно в процессе их передачи с разделением на микропотоки данных, что позволило снизить время выполнения запросов;

3. новая грамматика унифицированных SQL-подобных запросов к разнородным данным, отличающаяся расширением оператора SELECT в командах DML языка SQL, позволяющая формировать запросы без учета специфики данных;

4. метод анализа разнородных данных в режиме реального времени в системах управления транспортной инфраструктурой с элементами дополненной реальности, отличительными особенностями которого являются: (1) механизм уточнения положения транспортных средств на основе анализа координат наблюдателя, камеры и движущегося объекта; (2) идентификация объекта с использованием подходов распознавания изображений; (3) обработка событийных данных, полученных от транспортных средств.

Теоретическая значимость работы состоит в разработке модели и методов сбора и предварительной распределенной обработки разнородных данных. Содержащиеся в диссертационной работе анализ, выводы и предложения могут быть использованы для разработки систем проактивного управления транспортной инфраструктурой.

Практическая значимость работы состоит в разработанном программном обеспечении, реализующем предложенные методы. Разработаны «Распределенная система слияния и предобработки разнородных данных с разных источников» (свидетельство о государственной регистрации программы для ЭВМ № 2017660307 от 20 сентября 2017 г.) и «Программное обеспечение обнаружения мошенничества в телекоммуникационных предприятиях» (свидетельство о государственной регистрации программы для ЭВМ № 2017611602 от 7 февраля 2017 г.). Программы прошли апробацию в телекоммуникационной компании ОАО «Indochina Telecom Mobile», которая осуществляет сбор и обработку данных городской транспортной инфраструктуры.

Основные положения, выносимые на защиту.

1. модель хранения разнородных данных в соответствии с концепцией «озеро данных», применение которой снижает время на дальнейшую обработку;

2. метод сбора и предварительной обработки данных, поступающих от разнородных источников, позволяет снизить время выполнения запросов к разнородным данным за счет предварительной обработки и структуризации данных в процессе их передачи;

3. новая грамматика унифицированных SQL-подобных запросов к разнородным данным, расширяющая команды DML языка SQL, позволяет формировать запросы без учета специфики данных;

4. метод анализа разнородных георасределенных данных в режиме реального времени позволяет повышать эффективность проактивных систем управления транспортной инфраструктурой с элементами дополненной реальности;

5. программное обеспечение, реализующее предложенные решения в распределенной среде вычислений, позволяет разрабатывать эффективные системы проактивного управления транспортной инфраструктурой.

Достоверность результатов исследований подтверждается проведенными экспериментальными исследованиями на сгенерированных данных, а также внедрением и использованием рекомендаций, содержащихся в диссертационном исследовании.

Апробация работы. Основные положения исследования докладывались и обсуждались на следующих научных конференциях: Distributed Computer and Communication Networks : 21st International Conference, DCCN 2018 (Moscow, Russia, September 17–21, 2018), Proceedings of the IV International research conference «Information technologies in Science, Management, Social sphere and Medicine» (ITSMSSM 2017), 20th Distributed Computer and Communication Networks – DCCN 2017 : 20th International Conference (Moscow, Russia, September 25-29, 2017), 7th International Conference on Information, Intelligence, Systems & Applications (IISA) (Greece, 13-15 July 2016), Наука и современность – 2016 : сб. матер. XLV междунар. науч.-практ. конф. (г. Новосибирск, 26 мая, 14 июня 2016 г.), XXI Региональная конференция молодых исследователей Волгоградской области (г. Волгоград, 8-11 ноября 2016 г.).

Публикации. По теме диссертации издано 12 печатных работ, в том числе 4 статьи в изданиях, рекомендованных ВАК, 4 работы в зарубежных изданиях, индексируемых в базах научного цитирования Scopus и Web of Science. По результатам работы созданы 2 программных продукта, которые получили Свидетельства о государственной регистрации.

Личный вклад автора. В диссертации представлены результаты исследований, выполненных самим автором. Личный вклад автора состоит в постановке задач исследования, разработке теоретических и прикладных методов их решения, в обработке, анализе, обобщении полученных результатов и формулировке выводов. В публикациях с соавторами авторский вклад распределяется пропорционально.

Структура и объём диссертации. Диссертационная работа состоит из введения, четырех глав, заключения, а также библиографического списка из 136 наименований и 2 приложений. Общий объем работы – 141 страница, в том числе 43 рисунка и 7 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность работы, представлена характеристика работы, выделены объект и предмет исследования, определены цель и задачи исследования, формируются научная новизна, методология и методы исследования, практическая значимость работы, приведены структура работы, основные положения, выносимые на защиту.

В первой главе проведен анализ проактивных систем поддержки принятия решений управления транспортной инфраструктурой, их архитектуры и особенности. Критической задачей в проактивных системах является задача сбора и предварительной обработки данных для дальнейшего принятия решений. Представлена классификация данных по различным критериям. Выделены следующие типы данных, используемых в проактивных системах поддержки принятия решений: машинные данные сенсорного типа (sensors data), событийные данные (log data), визуальные данные – изображения или видео, текстовые данные (textual data), данные социальных сетей

(social data), георaспределенные данные (geospatial data). Данные могут быть структурированными, неструктурированными и полуструктурированными. Поэтому в системах управления транспортной инфраструктурой разнородность данных является критичной и требует реализации эффективных подходов слияния разнородных данных. Выполнен анализ особенностей и проблем сбора и слияния разнородных данных в проактивных системах поддержки принятия управленческих решений. Выделены следующие особенности: (1) необходимость предопределения форматов данных, (2) необходимость обеспечения качества данных; (3) необходимость минимизации участия человека в процессе сбора разнородных данных; (4) обеспечение сбора данных в реальном времени. Отметим, что сбор данных и слияние данных являются критическими процессами в проактивных системах управления транспортной инфраструктурой. Важнейшим фактором здесь является время доступа к данным в процедурах принятия решений. Рассмотрены две концептуальные модели обработки данных: (i) λ-модель и (ii) модель Kappa, используемые при разработке проактивных систем поддержки принятия решений. Модель λ-архитектуры для сбора и распределенной обработки данных является предпочтительной для синтеза проактивных систем управления транспортной инфраструктурой. Сформулирована цель и даны постановки задачи диссертационного исследования.

Во второй главе приведен обзор современного состояния исследований в области сбора и обработки разнородных данных. Выполнен обзор методов слияния разнородных данных, применяемых в системах управления и поддержки принятия решений. Методы слияния разнородных данных направлены на (i) улучшение качества разнородных данных, получаемых с различных источников и (ii) снижение времени их обработки. Основная проблема при использовании методов слияния – время обработки данных, представленных в различных форматах.

Рассмотрены различные подходы хранения разнородных данных: реляционные базы данных, базы данных NoSQL, распределенные системы хранения данных. Для решения задачи проактивного управления транспортной инфраструктурой на основе обработки разнородных данных целесообразно применять распределенные хранение и обработку данных. В главе сделан обзор различных подходов (технологических стеков) для реализации системы обработки потоковых (в режиме реального времени) и пакетных данных. Анализ технологических стеков показывает целесообразность использования λ-архитектуры для реализации методов сбора и обработки разнородных данных при проактивном управлении транспортной инфраструктурой.

Рассмотрены современные методы интеллектуального анализа данных и их возможность применения в проактивных системах управления транспортной инфраструктурой. Во-первых, методы интеллектуального анализа данных могут быть применены в процессе слияния данных, для повышения качества изображений, и/или для распознания объектов в видеопотоках. Во-вторых, анализ методов интеллектуального анализа данных позволил сформулировать требования к методам сбора и предварительной обработки данных.

Предложена концепция хранилища данных по принципу «озера данных» (ХДОД) на основе λ-архитектуры, в которой реализуются как пакетная, так и потоко-

вая обработки данных. Данная концепция может быть использована для проектирования архитектуры проактивных систем управления транспортной инфраструктурой на основе обработки разнородных данных.

В третьей главе описываются предложенные методы эффективного сбора и предварительной обработки разнородных данных в проактивных системах управления транспортной инфраструктурой.

Модель распределенного хранения разнородных данных в концепции «озеро данных». В результате системного анализа процесса сбора и структуры данных была предложена модель хранения разнородных данных в соответствии с концепцией озера данных (далее по тексту модель ХДОД):

$$S = \langle \{DT\}_{i=1}^{n_{ST}}, \{SS\}_{j=1}^{m_{SS}}, \{E\}_{k=1}^{p_E}, IS, DS \rangle, \quad (1)$$

где, $\{DT\}_{i=1}^{n_{ST}}$ – множество шаблонов данных, n_{ST} – количество шаблонов данных, $\{SS\}_{j=1}^{m_{SS}}$ – методы разбиения разнородных данных, m_{SS} – количество типов данных, $\{E\}_{k=1}^{p_E}$ - множество исполнителей задач сбора данных, IS – метод индексирования данных в хранилище озера данных, DS – структура озера разнородных данных. Таким образом, предложенная модель ХДОД отличается от имеющихся, наличием шаблонов объектов и шаблонов параметров объектов, которые позволяют распределено хранить как сырые разнородные данные, так и структурированные данные в соответствии с предопределенной схемой, что позволяет снизить временные затраты на доступ к данным.

Рассмотрим компоненты модели ХДОД, схема которой представлена на рисунке 1.

Компонент ***Object Template*** предназначен для шаблонизации некоторого объекта O_i, где $i = 1,...,n$. Каждый объект может иметь множество источников данных M_{ds}, которые созданы в соответствии с шаблонам ***Data Source Template***. Пусть имеется некоторый объект, структура которого будет представлена следующим образом:

$$O =< ds_1, ds_2, .., ds_m >$$

Каждый источник данных может иметь множество параметров с различными типами данных M_p в соответствии с шаблонами ***Parameter Template***. Структура такого произвольного источника данных ds представлена в виде:

$$ds =< p_1, p_2, .., p_k >$$

Метод сбора и предварительной обработки разнородных данных в ХДОД для проактивного управления транспортной инфраструктурой. В рамках предложенной модели ХДОД разработаны метод сбора и хранения разнородных источников данных в соответствии с *лямбда*-архитектурой для потоковой и пакетной обработки данных. Разработаны: (i) алгоритм синхронизации разнородных данных, (ii) алгоритм разделения данных и (iii) метод индексирования разнородных данных.

Предложена модификация метода, в результате которой метод включает следующие шаги:

1. Определение требуемой схемы данных. Наблюдаемые объекты описываются набором гетерогенных данных. Схема данных для хранения такого рода данных представлена в соответствии с форматом.

$$sD = \langle gId, timestamp, (lat, long), attrD \rangle,$$

где *gId* – глобальный идентификатор объекта, который является уникальным для каждого наблюдаемого объекта; *timestamp* – временная метка, определяемая при наблюдении; *(lat, lon)* – координаты местоположения объекта на временной отметке как пара широты и долготы; *attrD* – список (словарь) пар ключ-значение, описывающий функции объекта и его значения, например, «*speed : 68*».

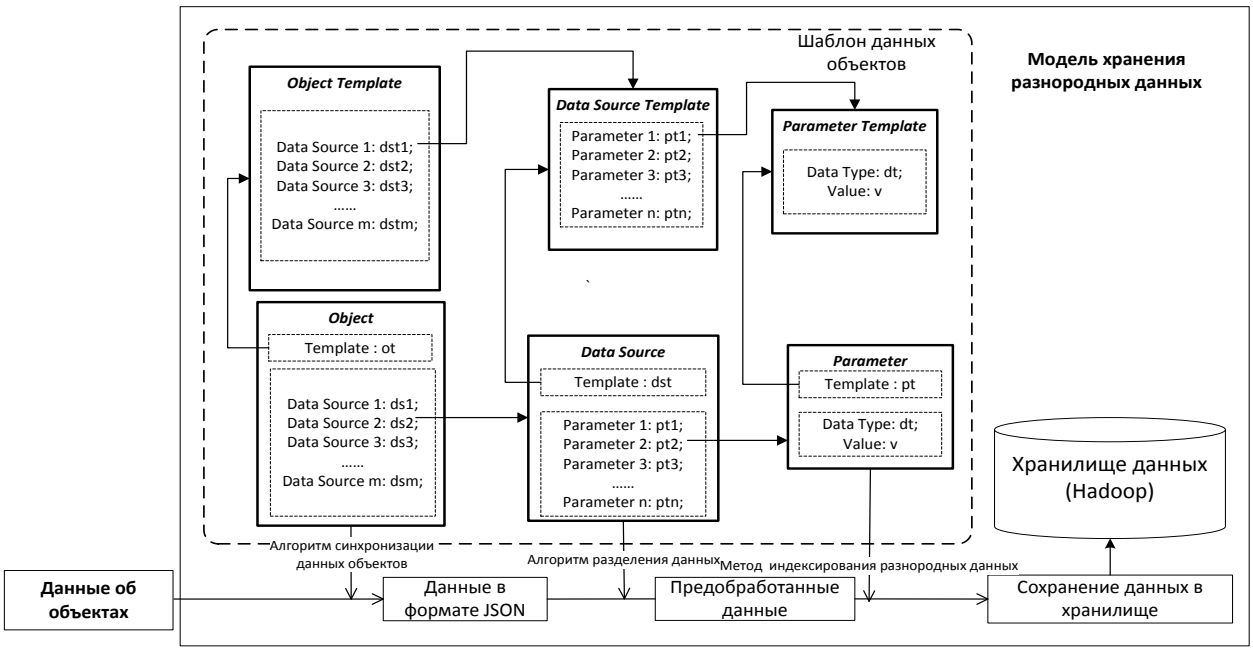

Рисунок 1 – Модель хранения разнородных данных

2. Описание источников данных и настроек сборщиков данных. Мы используем сборщики данных терминов для программного обеспечения, имеющего доступ к источникам данных, и сбора данных о наблюдаемых объектах. Высокоуровневое описание произвольных источников данных определяется форматом.

$$sC = \langle sId, acs, (lat, lon), attrS \rangle,$$

где *sId* - уникальный идентификатор источника данных (например, ссылка на веб-службу, откуда поступают данные), *acs* – это список значений ключа для источника данных, помеченный как *sId*, *attrS* – внутренняя схема данных полученных от источника данных. Считаем, что *attrS* содержит обязательный тип параметра, который принимает значение из списка *<type: JSON, type: VIDEO>*.

3. Построение схем привязки данных. На этом этапе создается связь между исходной схемой источника данных и требуемой схемой. Эта ссылка представлена в виде набора *R*, содержащего пары атрибутов из набора *attrD* схемы *sD* и атрибутов из набора *attrS* в схеме *sC*.

$$R = \{r_{i,j}\}; r_{i,j} = \langle attrD_i, attrS_j \rangle,$$

отметим, что:

$\exists r_{lon}: \langle lon^{(sC)}, lon^{sD} \rangle$ и, $\exists r_{lat}: \langle lat^{(sC)}, lat^{sD} \rangle$.

4. Реализация алгоритмов преобразования данных. В соответствии с настройками привязки R и алгоритмами $\{\alpha_k\}_{k=1}^{||R||}$ реализовано преобразование данных из исходной схемы в желаемую.

$$\forall r_{i,j} \in R; \exists \alpha_{i,j}: v(attrD_i) \rightarrow v^*\big(attrS_j\big),$$

где *v* указывает значение атрибута. Следует отметить, что преобразования могут быть простыми и сложными. В простом случае, например, для типа JSON может быть установлено однозначное соответствие двух схем данных. Если тип VIDEO, видеопоток преобразуется с использованием кадрового преобразования с временной привязкой атрибута времени привязки и географическими координатами *(lat, lon)* для каждого кадра. Кроме того, *attrD* задает параметры с идентификатором видео, идентификатором кадра, именем фрейма и указанием пути к сохраненному видеофайлу или сохраненному изображению (кадру). Если для распознавания образов используются средства компьютерного зрения, *attrD* содержит атрибуты для его использования. Кроме того, необходимо включить атрибуты *attrD*, объявляющие оценку производительности распознавания.

5. Разделение данных. Схема *DS* для разделения потоков данных в микро-потоки

$$DS_{\alpha_k} = \langle df, \alpha_k, \{mdf_l\}_{l=1}^{L_{\alpha_k}} \rangle,$$

где df – исходный поток данных, mdf_l – l поток данных в памяти для определенного алгоритма α_k, L_{α_k} – количество потоков.

На этом этапе данные разбиваются на потоки данных, подлежащие обработке в распределенной архитектуре, в соответствии с предопределенными задачами, например, вычисляя количество объектов с равными атрибутами из набора атрибутов *attrD*.

6. Вставка обработанных данных в базу данных. Когда данные преобразуются в соответствии с определенной схемой, они вставляются в базу данных. Это позволяет извлекать разнородные данные из базы данных без дополнительных манипуляций с данными.

Метод реализации запросов к разнородных данным в ХДОД на основе унифицированной SQL-подобной грамматики. Для решения проблемы унификации доступа к данным в данной работе была спроектирована архитектура системы обработки разнородных данных с использованием унифицированных запросов для ХДОД.

Предлагаемый метод реализации запросов к ХДОД с унифицированной грамматикой SQL-подобных запросов включает в себя следующие шаги. 1. Построение унифицированного запроса на SQL-образном языке. 2. Разбор запроса, сформированного на SQL-образном языке. 3. Формирование запроса к ХДОД. 4. Агрегация и вывод пользователю результатов на запрос к хранилищу разнородных данных. Предложена новая грамматика унифицированных SQL-подобных запросов к разнородным данным, отличающаяся расширением оператора SELECT в командах DML языка SQL, позволяющая формировать запросы без учета специфики данных. Общая грамматика инструкции, реализующего выбор данных (select) имеет следующий вид:

query : select_stmt where_stmt;

где, *select_stmt* – оператор выбора,

where_stmt – оператор условий фильтрации результатов;

На рисунке 2 представлена схема инструкции SELECT с оператором *from_stmt* выбора источников данных.

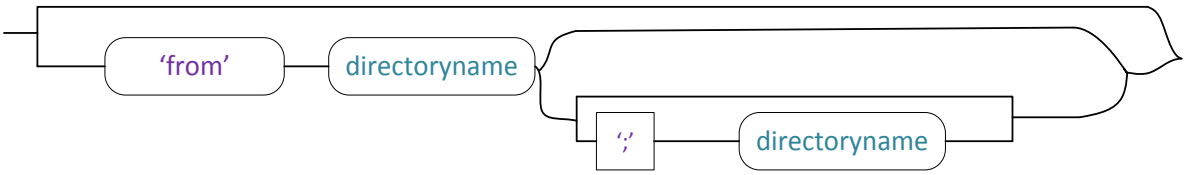

Рисунок 2 – Схема инструкции SELECT с оператором *from_stmt*

Метод анализа разнородных данных в режиме реального времени в системах управления транспортной инфраструктурой с элементами дополненной реальности. Разработка систем управления транспортной инфраструктурой с элементами дополненной реальности (Augmented Reality, AR) требует анализа обоих типов данных: видеопотоков (распознавание или сегментация изображений) и событийных данных, полученных с транспортных средств (далее по тексту ТС). Информация должна быть согласована и предоставлена для конечного пользователя для дальнейшего принятия решений. Время обработки запросов и предоставление результатов пользователю является критическим показателем. В работе предлагается подход для объединения данных о ТС источников анализа и позиционирования в системе управления транспортной инфраструктурой с элементами дополненной реальностью. Отличительными особенностями метода являются: (1) механизм уточнения положения транспортных средств на основе анализа координат наблюдателя, камеры и движущегося объекта; (2) идентификация объекта с использованием подходов распознавания изображений; (3) обработка событийных данных, полученных от транспортных средств. Этапы реализации метода.

Этап 1. Наблюдатель, использующий AR-систему, начинает захватывать изображение участка дороги в определенное время τ (время начала события).

Этап 2. Пакет данных *DP* о запуске события отправляется на сервер. Пакет содержит данные о наблюдателе (его местоположении) и времени начала события τ.

Этап 3. В момент времени поступления пакетов данных на сервер, формируется и отправляется запрос видеопотока VS_c в соответствии с местоположением наблюдателя и местоположением камеры.

Этап 4. Запускается процедура распознавания транспортного средства по изображениям, извлеченным из видеопотоков VS_c в интервале времени $[\tau - \varepsilon, \tau + \varepsilon]$.

Этап 5. Если транспортное средство распознано, выполняются следующие действия.

(a) Сохранить временную метку τ^* распознанного ТС.

(b) Создать и отправить запрос R_1 к *DD,* содержащий время распознавания τ* и пару $(long_c, lat_c)$.

(c) Выбрать из *DD* все ТС, которые имеют одинаковое местоположение в интервале времени $[\tau - \varepsilon, \tau + \varepsilon]$. Одинаковое местоположение означает, что $L < d$, где d - предопределенный порог. Значение L вычисляется таким образом:

$L = R \cdot c$, где, R - радиус Земли (константа), $c = 2 \cdot a\,tan\,2\,(\sqrt{a}, \sqrt{1-a})$

$a = sin^2(\,(lat_v - lat_c)/2) + cos(\,lat_c) \cdot cos(\,lat_v) \cdot sin^2(\,(long_v - long_c)/2)$

(d) Если ТС выбрано на предыдущем шаге с идентификатором *ID**, создать и отправить запрос R_2 на *SD*. Этот запрос содержит уникальный идентификатор *ID** ТС.

(e) Получить и подготовить пакет данных ответа DP_r, содержащий информацию о ТС с идентификатором *ID**.

Этап 6. Отправить пакет данных DP_r конечному наблюдателю.

В четвертой главе описаны фреймворк для генерации событий в разработке проактивных систем, называемый EVGEN, представлена архитектура системы сбора и предварительной обработки данных, проведены испытания и обоснована эффективность предлагаемых моделей и методов.

Предложен фреймворк EVGEN (сокращенно «EVent GENerator») для генерации потока данных событий транспортной инфраструктуры для тестирования предлагаемых методов обработки данных. Фреймворк позволяет создавать поток JSON-данных в соответствии с предварительными настройками: форматом данных (структурой) и интенсивностью потока данных. Архитектура системы базируется на основе модели обмена сообщениями публикации-подписки (Publish-Subscribe), в дальнейшем по тексту PS-модель. Фреймворк EVGEN был использован для генерации данных о транспортной инфраструктуре для оценки проблемы транспортного трафика.

Для апробации и тестирования эффективности предложенных в главе 3 модели и методов, была спроектирована архитектура проактивной системы управления транспортной инфраструктурой на основе лямбда-архитектуры системы распределенной обработки данных. Архитектура решения представлена на рисунке 3.

Программная реализация системы управления транспортной инфраструктурой выполнена с использованием языков программирования Java, Python в IDE Eclipse.

Метод сбора и слияния разнородных данных (ПМ) был апробирован при решении задачи сбора данных о транспортной ситуации по данным с движущихся ТС (в виде логов) и видеопотоков. Предполагается, что пользователю необходимо получать данные о перемещениях ТС в заданный промежуток времени [*t0, tk*] в заданной географической окрестности из всех имеющихся источников данных, т.е. из лог-файлов и из видеопотоков. В качестве базового метода, с которым осуществлялось сравнение производительности предлагаемых подходов, использовалась реализация подхода П1, обрабатывающего «сырые» данные в ХДОД без предварительной обработки. Время обработки данных ПМ складывалось из времени предобработки исходных данных и времени выполнения запроса.

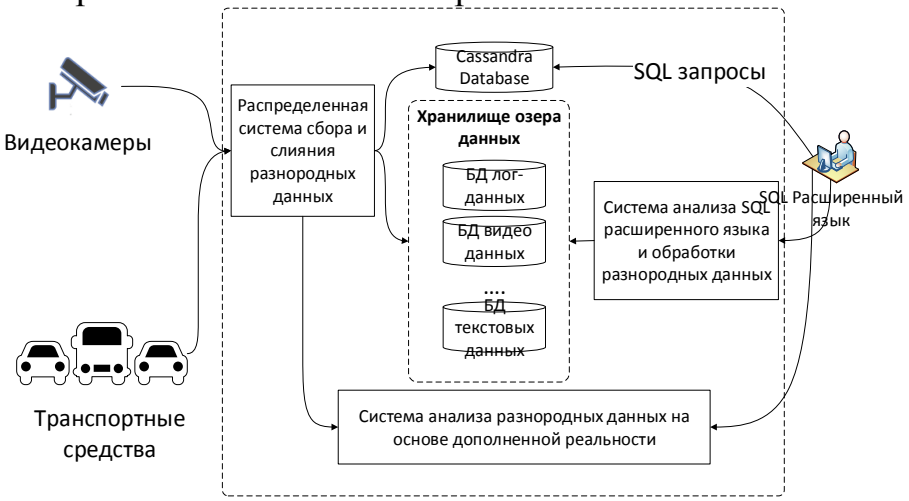

Рисунок 3 – Архитектура проактивной системы управления транспортной инфраструктурой на основе слияния и анализа разнородных данных

Эксперименты проводились для различной степени интенсивности поступаемых данных: для лог-файлов – 1 млн сообщений в секунду и для одного источника с видеопотоком. Среднее время реализации ПМ составило соответственно 0,14 с и 0,75 с, т.е. в сумме

0,89 с. Также менялась интенсивность запросов: 10, 30, 300 и 3000 запросов в секунду. Каждый эксперимент проводился 10 раз. Так, для 10 запросов в секунду время запроса П1 составило 0,703±0,16 с против 1,087±0,32 с для предлагаемого метода, что не является значительной разницей. Однако уже для 30 запросов в секунду П1 показал время 1,925±0,09 с, а предлагаемый метод – 1,368±0,25 с, что в 1,4 раза меньше П1. Для 300 запросов в секунду время выполнения предлагаемого метода составило 2,904±0,33 с, что в 1,7 раз меньше чем для П1 (4,995±0,28 с). Для 3000 в секунду время выполнения предлагаемого метода составило 14,153±1,04 с, что в 3,24 раза меньше, чем для П1 (45,899±2,56 с). Следует отметить, что для подтверждения превосходства предлагаемого метода использовалась непараметрическая процедура тестирования двух независимых выборок, предложенная Сигелом с уровнем значимости $\alpha = 0,01$.

Для тестирования эффективности работы модифицированного метода (ПМ-2) решения было проведено несколько экспериментов продолжительностью 30 мин 38 сек (365 завершенных партий, 1 318 747 записей). Сравнение проводилось с методом ПМ. В таблице 1 показано, что для экспериментов с 3 брокерами среднее время обработки уменьшается. Итак, для части А: сокращение в ~1,77 раза, для части В сокращение составляет около 2,6 раза.

Таблица 2 – Результаты экспериментов для определенного профиля нагрузки: среднее время обработки и средняя задержка для различной конфигурации

Части экспериментов	Брокер	Интенсивность (event/sec)	Среднее время обработки, sec
Part A:video & log	1 broker	100	0.339 ± 0.802
Part B: log	1 broker	1500	3.095 ± 17.81
Part C:video & log	1 broker	1000	2.195 ± 16.290
Part A: video & log	3 brokers	100	0.191 ± 0.118
Part B: log	3 brokers	1500	1.519 ± 8.741
Part C:video & log	3 brokers	1000	1.066 ± 6.12

И, наконец, для части С сокращение в 2.05 раза. Основной вывод заключается в том, что увеличение количества брокеров для наблюдаемой интенсивности является разумным.

Результаты апробации и обоснования эффективности метода выборки разнородных данных из ХДОД на основе предлагаемой SQL-подобной грамматики. В качестве запроса (запрос 1) использовалось выражение «Вычислить количество транспортных объектов, которые отправляют данные на сервер или появляются на участке дороге, охватываемом камерой camera1 во временном интервале *[t1;t2]*. Запрос с использованием предложенной грамматики представлен в следующем виде: *select count from Datalog, camera1 where modified > 1516051435073 and modified < 1516661699999;* где, *Datalog* - путь к JSON-файлу, содержащему данные о ТС; *camera1* - путь к файлу, содержащему изображения с камеры № 1. Эксперименты проводились для различных объемов разнородных данных (лог-файлов и видео файлов). Среднее время разбора SQL-запроса и реализации одного запроса для разнородных данных (2Мб лог-файлов и 288.1Мб видео файлов в виде изображений) составило соответственно 0,0135 с и 148,3 с, т.е. в сумме 148,31 с. Также менялась конфигурация экспериментов: (1) 2Мб

лог-файлов и 230.4Мб видео файлов в виде изображений; (2) 1Мб лог-файлов и 172.8Мб видео файлов в виде изображений; (3) 1Мб лог-файлов и 115.2Мб видео файлов в виде изображений; (4) 1Мб лог-файлов и 57.6Мб видео файлов в виде изображений.

Для апробации метода анализа разнородных данных в режиме реального времени в системах управления транспортной инфраструктурой с элементами дополненной реальности была спроектирована архитектура системы на основе библиотеки OpenCV 3.0 для языка программирования Java. Время обработки одного кадра видео (изображения) составляет 10 - 35 мс, поэтому система подходит для работы в реальном времени. Объем файлов журналов данных в динамической базе данных составляет ~2 Мб, время обработки событийных данных составляет 20 – 80 мс. Время поиска данных транспортного средства в DD зависит от объема событийных данных. В результате тестирования работоспособности предлагаемой системы показана возможность реализации метода на практике для большого числа различных источников.

В заключении сформулированы основные результаты и выводы по работе.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ РАБОТЫ

Основным результатом исследования является разработка теоретических положений, направленных на повышение эффективности функционирования проактивных систем управления транспортной инфраструктурой за счет разработки модели и методов сбора и предварительной распределенной обработки разнородных данных. Получены следующие результаты.

1. В результате системного анализа процесса сбора и слияния разнородных данных в проактивных системах управления транспортной инфраструктурой, выделены основные операции сбора и обработки разнородных данных, для которых имеется проблема совершенствования методов обработки данных.
2. В результате анализа современного состояния исследований в области сбора и обработки разнородных данных, предложена концепция хранилища данных, формируемая по принципу «озера данных» на основе λ-архитектуры, в которой реализуются как пакетная, так и потоковая обработка данных.
3. Предложены модель и методы эффективного сбора и предварительной обработки разнородных данных в соответствии с концепцией «озеро данных» и наличием механизмов преобразования данных к нужному формату в процессе их передачи и разделением на микропотоки данных, что позволило снизить время выполнения запросов к разнородным данным. Для реализации методов разработана новая грамматика унифицированных SQL-подобных запросов к разнородным данным, позволяющая формировать запросы без учета специфики данных.
4. Выполнены испытания и показана эффективность предлагаемых методов. Для обоснования эффективности при различной интенсивности потоков данных разработан фреймворк EVGEN для генерации потока данных событий транспортной инфраструктуры. Так для различных архитектур и настроек методов сбора и обработки разнородных данных среднее время обработки разнородных данных уменьшилось в

~1,77 – 3,2 раз. Предлагаемые методы целесообразно применять в проактивных системах управления транспортной инфраструктурой в которых критично время выборки данных из ХДОД.

5. Разработана архитектура проактивной системы управления транспортной инфраструктурой на основе лямбда-архитектуры распределенной обработки данных.

6. Результаты апробированы в телекоммуникационной компании ОАО «Indochina Telecom Mobile», которая осуществляет сбор и обработку данных о транспортной инфраструктуре, получены 2 свидетельства о государственной регистрации программы для ЭВМ.

На основе полученных результатов можно сделать выводы об эффективности предлагаемых моделей и методов для сбора и предварительной обработки разнородных данных в проактивных системах управления транспортной инфраструктурой.

Перспективы использования работы. Дальнейшее развитие данного направления связно с формированием моделей адаптации схем данных для обработки возрастающего объема неструктурированных данных о транспортной инфраструктуре.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, включённых в Перечень ВАК

1. **Чан Ван Фу** Метод сбора и слияния разнотипных данных в проактивных системах интеллектуальной поддержки принятия решений / Ван Фу Чан, М.В. Щербаков, Туан Ань Нгуен, Д.А. Скоробогатченко // Нейрокомпьютеры: разработка, применение. - 2016. - № 11. - С. 40-44.

2. **Чан, Ван Фу** Обзор архитектур систем поддержки принятия решений, использующих аналитику данных в режиме реального времени / Ван Фу Чан, М.В. Щербаков, Туан Ань Нгуен // Известия ВолгГТУ. Сер. Актуальные проблемы управления, вычислительной техники и информатики в технических системах. - Волгоград, 2016. - № 3 (182). - С. 95-100.

3. Разработка метода проактивного обнаружения мошенничества потребителей услуг телекоммуникационной компании / Туан Ань Нгуен, М.В. Щербаков, **Ван Фу Чан**, А.Г. Кравец // Прикаспийский журнал: управление и высокие технологии. - 2016. - № 4. - С. 43-52.

4. **Чан Ван Фу** Грамматика запросов для хранилища разнородных данных в проактивных системах // Ван Фу Чан, М.В. Щербаков, Ван Квонг Сай // Программные продукты и системы, 2018.- Т. 31.- № 4. С. 659–666.

Публикации в изданиях, индексируемых в международных базах научного цитирования (Scopus, Web of Science)

5. **Чан, Ван Фу** Yet Another Method for Heterogeneous Data Fusion and Preprocessing in Proactive Decision Support Systems: Distributed Architecture Approach [Электронный ресурс] / Ван Фу Чан, М.В. Щербаков, Туан Ань Нгуен // Distributed Computer and Communication Networks – DCCN 2017 : 20th International Conference (Moscow, Russia, September 25-29, 2017) : Proceedings / ed. by V.M. Vishnevskiy [et al.]. – [Springer International Publishing AG], 2017. – P. 319-330. – URL : http://www.springer.com/gp/book/9783319668352?wt_mc=ThirdParty.SpringerLink.3.EPR653.About_eBook. – (Ser. Communications in Computer and Information Science ; Vol. 700).

6. Clustering helps to determine the changes in telecom subscribers' behavior [Электронный ресурс] / А.В. Голубев, Ань Туан Нгуен, М.В. Щербаков, **Ван Фу Чан** // Proceedings of the IV International research conference «Information technologies in Science, Management, Social sphere and Medicine» (ITSMSSM 2017) / ed. by O.G. Berestneva [et al.]. – [Published by Atlantis Press], 2017. – P. 239-243. – (Ser. Advances in Computer Science Research (ACSR) ; Vol. 72). – URL : https://www.atlantis-press.com/proceedings/itsmssm-17.

7. **Чан, Ван Фу** EVGEN: A framework for event generator in proactive system design [Электронный ресурс] / Ван Фу Чан, М.В. Щербаков, Туан Ань Нгуен // 7th International Conference on Information, Intelligence, Systems & Applications (IISA) (Greece, 13-15 July 2016) / Institute of Electrical and Electronics Engineers (IEEE). – [Publisher: IEEE], 2016. – DOI: 10.1109/IISA.2016.7785402. – URL : http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7774711.

8. **Чан, Ван Фу** On-the-Fly Multiple Sources Data Analysis in AR-Based Decision Support Systems [Электронный ресурс] / Ван Фу Чан, М.В. Щербаков, Ван Квонг Сай // Distributed Computer and Communication Networks : 21st International Conference, DCCN 2018 (Moscow, Russia, September 17–21, 2018) : Proceedings / eds.: V. M. Vishnevskiy, D. V. Kozyrev ; V. A. Trapeznikov Institute of Control Sciences of RAS (ICS RAS, Russia, Moscow), Peoples' Friendship University of Russia (RUDN University, Russia, Moscow). – Cham : Springer Nature Switzerland AG, 2018. – P. 481-492. – (Book ser.: Communications in Computer and Information Science (CCIS) ; vol. 919). – URL : https://link.springer.com/chapter/10.1007/978-3-319-99447-5_41.

Прочие публикации

9. Нгуен, Туан Ань Архитектура проактивной системы сбора и обработки геопространственных данных / Туан Ань Нгуен, **Ван Фу Чан** // Наука и современность – 2016 : сб. матер. XLV междунар. науч.-практ. конф. (г. Новосибирск, 26 мая, 14 июня 2016 г.) / под общ. ред. С.С. Чернова ; Центр развития научного сотрудничества (ЦРНС). - Новосибирск, 2016. - С. 122-127.

10. **Чан, Ван Фу** Разработка программного обеспечения для интеграции разнородных данных в проактивных системах поддержки принятия решений / Ван Фу Чан // XXI Региональная конференция молодых исследователей Волгоградской области (г. Волгоград, 8-11 ноября 2016 г.) : тез. докл. / редкол.: А.В. Навроцкий (отв. ред.) [и др.] ; Комитет молодёжной политики Волгогр. обл., Совет ректоров вузов Волгогр. обл., Волгоградский гос. техн. ун-т. - Волгоград, 2016. - С. 142.

Свидетельства о регистрации программ для ЭВМ

11. Свид. о гос. регистрации программы для ЭВМ № 2017660307 от 20 сентября 2017 г. Российская Федерация. Распределённая система слияния и предобработки разнородных данных с разных источников / М.В. Щербаков, **Ван Фу Чан**, Туан Ань Нгуен, Ван Квонг Сай; ВолгГТУ. - 2017.

12. Свид. о гос. регистрации программы для ЭВМ № 2017611602 от 7 февраля 2017 г. Российская Федерация. Программное обеспечение обнаружения мошенничества в телекоммуникационных предприятиях / М.В. Щербаков, Туан Ань Нгуен, **Ван Фу Чан**; ВолгГТУ. - 2017.

Подписано в печать. 04.2019 г. Заказ №_____ . Тираж 100 экз. Усл. печ. л. 1,0
Формат 60 х 84 1/16. Бумага офсетная. Печать офсетная.
Типография ИУНЛ Волгоградского государственного технического университета.
400005, г. Волгоград, просп. им. В.И.Ленина, 28, корп. № 7