Описание и формирование периметра базы данных по систематизации и хранению разноструктурированных данных

А.А. Нечитайло[1], О.И. Васильчук[2], А.А. Гнутова[1]

[1]Самарский национальный исследовательский университет им. академика С.П. Королева, Московское шоссе 34А, Самара, Россия, 443086

[2]Поволжский государственный университет сервиса, Гагарина 4, Тольятти, Россия, 445017

Аннотация. Для хранения больших данных, как правило, используются реляционные базы данных. Для многостороннего исследования и анализа процессов, происходящих в крупных экономических системах финансисты, экономисты и другие технические специалисты используют большие данные с фактическими названиями предприятий, городов, регионов и т.д. Поэтому при использовании физических названий исследуемых регионов техническим специалистам интуитивно более удобно применять нереляционные базы данных.

1. Введение

«Big data» предусматривает процесс управлением и анализ больших объемов данных, который начал в мире активно развиваться с 2011 г. при этом инструменты анализа данных стали получать информацию из большего количества разноструктурированных и разноудаленных источников, что вызвано повсеместным внедрением цифровых технологий в различных сферах (бизнес, медицина, развлечения и т.д.). Так, в частности, согласно Прогноза социально-экономического развития Российской Федерации на период до 2036 года «Система здравоохранения будет функционировать в рамках единого цифрового контура на основе единой государственной информационной системы в здравоохранении (ЕГИСЗ), который даст возможность для сбора, хранения, обработки («big data») и анализа больших массивов информации» [6]. Одна из конечных целей данной работы включает обработку и интеллектуальный анализ больших данных («big data»), параллельные вычисления с целью создания систем принятия решений в реальном масштабе времени. Для решения таких задач нужно определять не только взаимосвязи (алгоритмы, модели и др.) конечной цели со средствами ее достижения и существующими ограничениями, но и формы описания и формирования периметра базы данных.

2. Постановка задачи

Задача синтеза рациональных схем выбора альтернатив и оценивания их качества состоит в том, чтобы из множества конкурирующих стратегий решения некоторой проблемы, на основе анализа условий и последствий ее реализации выбрать лучшую (оптимальную). Значимым дополнением к сказанному является то, что под условиями понимается не некоторая неподвижная картина сегодняшнего дня, но и условия, которые могут сложиться за

время реализации стратегии. Приятие обоснованных оптимальных решений при этом невозможно без устойчивого и оперативного получения надежных больших массивов данных.
Принимая во внимание вышеизложенное и учитывая последние тенденции, в ближайшем будущем основными источниками массивов информации станут: интернет вещей (IoT), социальные СМИ, метеорологические данные, GPS-сигналы из транспортных средств, данные о местонахождении абонентов мобильных сетей, Google Trends, сайты для поиска работы и другие альтернативные источники информации.
Особое внимание по систематизации и хранению разноструктурированных данных уделяют ЦБ РФ и ФНС РФ, в связи с этим, бизнесу предстоит решить ряд системных и технологических вопросов, препятствующих внедрению анализа «big data» в повседневную практику.
Среди них - отсутствие у компаний стратегий использования методов и данных анализа больших данных, недостаток современных технологических решений, отсутствие соответствующих навыков и понимания ключевых потоков формирования массивов данных.
Исследование проблем, связанных с внедрением «big data» в деятельность экономических субъектов, направленных на обеспечение экономической безопасности и развитие бизнеса, показывает, что усиление контроля со стороны ЦБ РФ и ФНС РФ направлено, в первую очередь, на формирование периметра базы данных по систематизации и хранению разноструктурированных данных юридических лиц в едином информационном пространстве.
Центральные банки по всему миру создали или создают департаменты для работы с большими данными («big data»), чтобы лучше понимать экономику, которой они управляют в надежде однажды получить технологии, позволяющие мониторить состояние экономики в режиме реального времени. Существующий мировой тренд представлен в таблице 1.

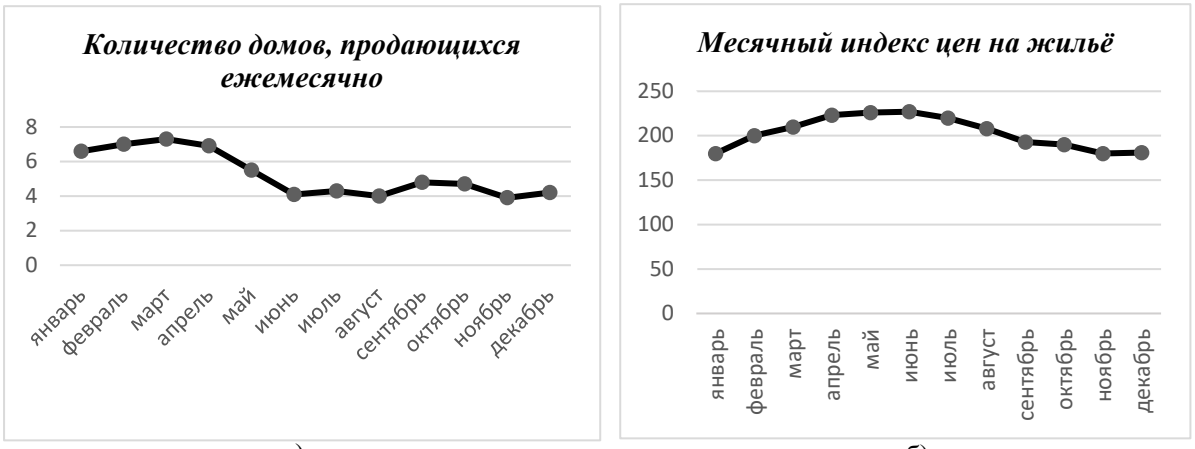
Таблица 1. Деятельность Центрального Банка стран по продвижению «big data».

Регион	Описание деятельности Центрального Банка страны по продвижению «big data»
Россия	Банк России опубликовал первое исследование, посвященное анализу на основе «Больших данных» («big data»). В докладе «Оценка экономической активности на основе текстового анализа» представлена методика расчета опережающего индикатора экономической активности в России, который построен на базе ежедневного контекстного анализа новостных сайтов с применением машинного обучения. Создана система мониторинга новостей, большие данные могут предсказать потребительское поведение на большом отрезке времени
Япония	Банк Японии использует большие данные с 2013 г. для анализа экономической статистики, что помогает регулятору строить более точные прогнозы. Планируется использовать большие данные для прямого сбора экономических данных, вместо того чтобы полагаться на результаты опросов.
Китай	Народный банк Китая будет активнее использовать «big data», искусственный интеллект и облачные вычисления, чтобы повысить свою способность распознавать, предотвращать и сокращать межотраслевые и межрыночные финансовые риски. В Китае большими данными интересуются в контексте слежения за потребителями и, главным образом, для контроля за должниками. Одна из главных проблем Китая — быстрое образование «пузырей» и склонность населения к участию в финансовых пирамидах. В мае

	местный ЦБ заявил, что планирует использовать большие данные вместе с искусственным интеллектом для отслеживания подобных рисков
США	В процессе принятия решений по денежно-кредитной политики регулятор продолжает полагаться на традиционные наборы данных. Экономисты Федеральной резервной системы (ФРС) часто используют «big data» при изучении конкретных вопросов, таких как динамика расходов после ураганов. Тем не менее ФРС видит множество недостатков в больших данных, особенно ограниченные периоды времени, которые охватывают эти сверхнасыщенные наборы данных. Это существенно снижает их ценность для прогнозирования. Кроме того, наборы данных часто производятся частными компаниями, ориентированными на нечто иное, чем экономический анализ. Это может сделать большие данные менее надежным, и ФРС опасается применять их для разработки политики.Тем не менее, в отдельных проектах большие данные уже используются. Например, для анализа потребительских и государственных расходов после ураганов. Проблема больших данных, по мнению экономистов, заключается в слишком малой глубине выборки, что существенно снижает возможности анализа. К тому же, часто данные собираются частными компаниями, которые преследуют собственные интересы (Коммерческие банки: Более 60% банков в Северной Америке считают, что «big data» даёт конкурентное преимущество, более 90% – что тот, кто справится с «big data», выиграет в будущем, только 37% банков имеют работающие проекты)
Еврозона	ЕЦБ изучает большие данные с 2013 г. Информация о примерно 40 тыс. ежедневных транзакциях на денежном рынке станет основой альтернативной ставки, поскольку традиционные бенчмарки становятся ненадежными. Регулятор также приобрел большой набор данных о ценах фактических покупок потребителей и изучает способы измерения инфляции в режиме реального времени. Аналитики ЕЦБ отслеживают Google Trends, чтобы оценить изменение безработицы, и используют алгоритмы для анализа сообщений в СМИ, чтобы оценить, рассматривается ли риторика регулятора как «ястребиная» или «голубиная». Однако ЕЦБ сохраняет осторожность. Так же как есть опасения по поводу фейковых новостей, которые доминируют в социальных медиа, существует риск того, что фейковая или по крайней мере низкого качества статистика вытеснит более качественные данные в публичном дискурсе. Информацию о 40 тыс. ежедневных транзакциях ляжет в основу альтернативной учетной ставки. ЕЦБ также приобрел данные о ценах реальных покупок граждан и ищет способы интернет-скрейпинга для измерения инфляции в реальном времени.
Великобритания	Создан Совет по «big data», который теперь называется Группой по управлению данными, а также лабораторию данных и аналитическое подразделение.

	Аналитики Банка Англии недавно использовали «big data», чтобы оценить последствия изменений обменного курса. Они также создают платформу для данных торговых репозитариев.
Индия, Сингапур, Индонезия	Индия сталкивается с проблемой безопасности и неприкосновенности личных данных, поэтому ЦБ страны больше озабочен кибербезопасностью в контексте больших данных. Сингапур создал Группу по анализу данных, в задачу которой входит сбор больших данных, которые пока будут анализироваться вручную, без применения ИИ-технологий. Основная задача, как в Индии, — борьба с отмыванием денег и терроризмом. Департамент статистики Банка Индонезии исследует социальные сети, новостные сайты и другие источники для анализа потребительских настроений. Недавно он начал получать данные от интернет-магазинов.

Приведенный ниже рисунок 1 является иллюстрацией использования «big data» банками для предсказания продаж домов на рынке США через Google Trends. Методика основана на том, что люди ищут дома гораздо больше непосредственно перед покупками [3].



а) б)

Рисунок 1. Иллюстрация изменения цен и количества продаваемых домов в США. (а) количество домов, продающихся ежемесячно; (б) месячный индекс цен на жилье.

Как видно, управленческие решения формируется на основе полученной информации и способу её передачи по функциональным единицам организации. Качество, достоверность, своевременность будет определенно влиять на эффективность управленческого решения. Век информационных технологий позволяет формировать, укрупнять, модернизировать информацию, в связи с чем возникают проблемы, которые приводят к избытку информации и ухудшению ее качества [5].

По мнению специалистов, объем полезной информации по отношению ко всей полученной информации будет уменьшаться год от года. Считается, что уже на сегодняшний день далеко не все данные ценны — по оценкам IDC, к 2020 году доля полезной информации составит лишь 35% от всей сгенерированной [1].

Для того, чтобы использование информации, которую получает менеджер было эффективным, необходимо правильно определить полезна ли полученная информация, и будет ли она важна для принятия управленческого решения. И только после этого выбрать правильный инструментарий (алгоритмы, модели, системы, компетентность и т.д.). Экспериментальное сравнение реляционных и нереляционных баз данных, проведенное авторами, подтверждает экспертные оценки специалистов о том, что управление тысячами

атрибутов, что требуется для экономических исследований - в реляционных базах данных неэффективно.
В связи с чем становится весьма актуальной для экономики проблема описания и формирования периметра базы данных по систематизации и хранению разноструктурированных документарных данных. Схематичное представление этого приводится на рисунке 2.

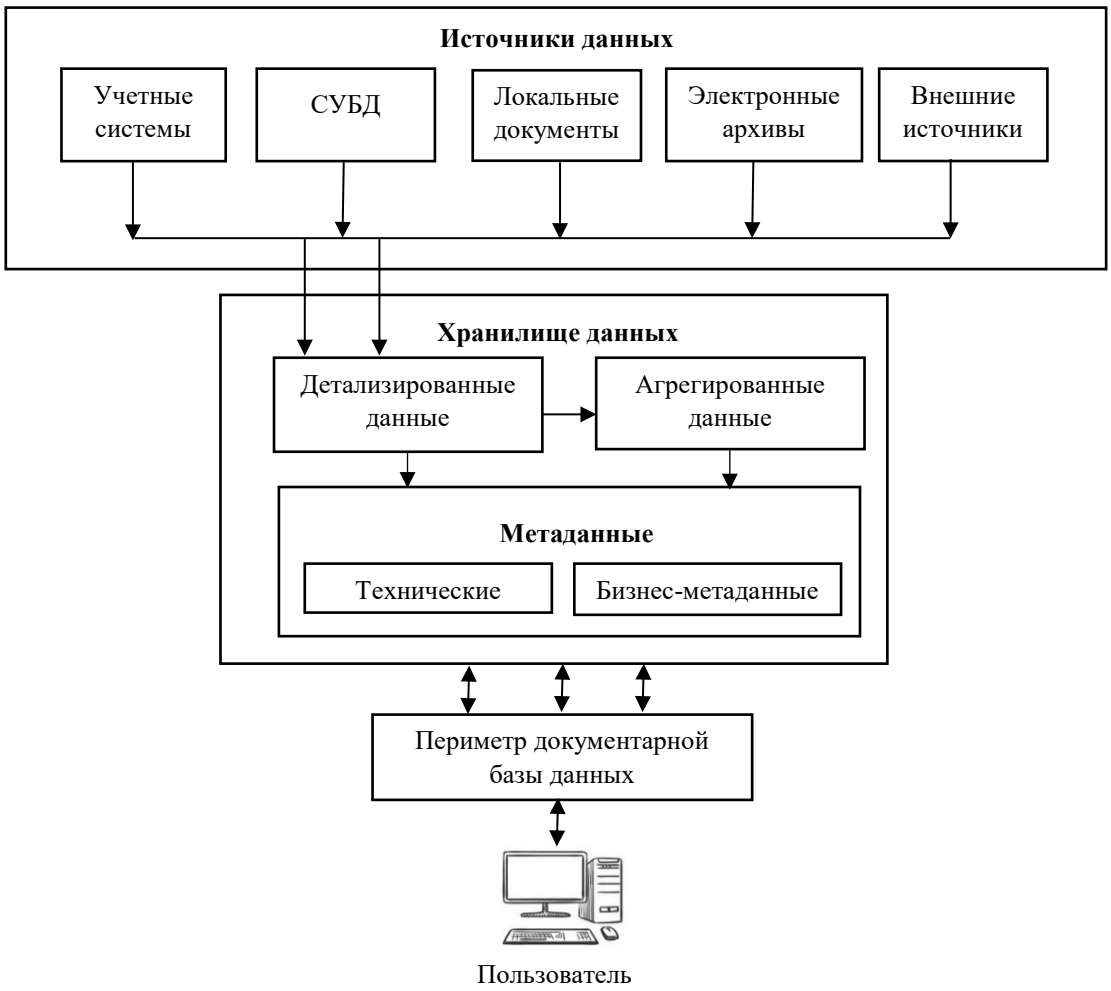


Рисунок 2. Иллюстрация структуры периметра базы данных по систематизации и хранению разноструктурированных данных в экономике.

Документные базы данных интуитивно понятны разработчикам, поскольку данные на уровне приложения обычно представляются как документ JSON. Разработчики могут сохранять данные с помощью той же документной модели, которую они используют в коде приложения. В документной базе данных все документы могут иметь одинаковую или различную структуру данных. Каждый документ является самоописываемым (т.е. содержит схему, которая может быть уникальной) и не обязательно зависит от какого-либо другого документа. Документы группируются в «коллекции», которые по своему назначению схожи с таблицами в реляционных базах данных.
Например, файл JSON для описания элемента книги в простой базе данных книг может выглядеть как следующий код.

```
[
  {
    "year" : 2013,
    "title" : "Turn It Down, Or Else!",
    "info" : {
      "directors" : [ "Alice Smith", "Bob Jones"],
      "release_date" : "2013-01-18T00:00:00Z",
      "rating" : 6.2,
      "genres" : ["Comedy", "Drama"],
      "image_url" : "http://ia.media-
imdb.com/images/N/O9ERWAU7FS797AJ7LU8HN09AMUP908RLlo5JF90EWR7LJKQ7@@._V1
_SX400_.jpg",
      "plot" : "A rock band plays their music at high volumes, annoying the neighbors.",
      "actors" : ["David Matthewman", "Jonathan G. Neff"]
    }
  },
  {
    "year": 2015,
    "title": "The Big New Movie",
    "info": {
      "plot": "Nothing happens at all.",
      "rating": 0
    }
  }
]
```

При использовании документной базы данных каждая сущность, отслеживаемая приложением, может храниться как отдельный документ. Документная база данных позволяет разработчику с удобством обновлять приложение при изменении требований. Кроме того, если необходимо изменить модель данных, то требуется обновление только затронутых этим изменением документов. Для внесения изменений нет необходимости обновлять схему и прерывать работу базы данных. При использовании документной базы данных атрибуты каждой транзакции можно описать в одном документе, что упрощает управление и повышает скорость чтения. Изменение атрибутов одной транзакции не повлияет на другие транзакции.

Анализ популярных документарных баз данных: Amazon DocumentDB (совместима с MongoDB), Amazon DynamoDB, MongoDB и Couchbase, проведенный авторами на основе исследования литературных источников и экспертных мнений показал перспективность применения для решения экономических задач документарной базы MongoDB в том числе с использованием решения AWS MongoDB Quick Start (также доступного в формате PDF) для развертывания кластера MongoDB в облаке AWS.

При решении задач организации современного производства требуется учитывать все большее число факторов различной природы, являющихся предметом исследования различных областей знаний. В этих условиях один человек не может принять решение о выборе факторов, влияющих на достижение цепи, не может определить существенные взаимосвязи между целями и средствами; в формировании и анализе модели принятия решения должны участвовать коллективы разработчиков, состоящие из специалистов различных областей знаний, между которыми нужно организовать взаимодействие и взаимопонимание; а проблема принятия решений становится проблемой коллективного выбора целей, критериев, средств и вариантов достижения цели, т.е. проблемой коллективного принятия решения на основе современных методов обработки больших данных. Это приводит к тому, что постановка задачи становится проблемой, для решения которой нужно разрабатывать специальные подходы, приемы, методы. В таких случаях возникает необходимость определить область проблемы принятия решения (проблемную ситуацию); выявить факторы

влияющие на ее решение; подобрать приемы и методы, которые позволяют сформулировать или поставить задачу таким образом, чтобы решение было принято.
Если удается получить выражение (алгоритм, методику и др.), связывающее цель со средствами, то задача практически всегда решается. Эти выражения могут представлять собой не только простые соотношения, подобные рассмотренному, но и более сложные, составные критерии (показатели), аддитивного или мультипликативного вида. Конечно, в этом случае могут возникнуть вычислительные сложности, при преодолении которых может потребоваться вновь обратиться к постановке задачи. Однако полученное формализованное представление задачи позволяет в дальнейшем применять и формализованные методы анализа проблемной ситуации.
Принятие решений — это научное направление, которое начало складываться в середине прошлого века. Задачей этого направления является синтез рациональных схем выбора альтернатив и оценивания их качества, которая состоит в том, чтобы из множества конкурирующих стратегий решения некоторой проблемы, на основе анализа условий и последствий ее реализации выбрать лучшую (оптимальную). Существенным дополнением к последней фразе является то, что под условиями понимается не некоторая застывшая картина «сегодня», но и те условия, которые могут сложиться за время реализации стратегии.
Это научное направление отличается тем, что к выбору критерия оптимальности нужно подходить творчески. Согласно такому подходу, критерий оптимальности не является неким экстремумом функции одной переменной, а представляет собой область многомерного пространства признаков, в которой некоторые частные параметры могут являться неоптимальными. Подразумевается, что речь идет о том, что все частные функции полезности рассматриваются не как равновесные, а как иерархически упорядоченная система функций полезности, обладающих разными весами (выбор которых, наряду с выбором самих функций, собственно, и составляет содержание процесса принятия решения).

3. Выводы

Таким образом, для принятия решения нужно получить выражение, связывающее цель со средствами се достижения с помощью вводимых критериев оценки достижимости цели и оценки средств. Если такое выражение получено, то задача решена.
В классической теории принятия решений центральный вопрос связывают с аксиоматикой «рационального» выбора. В итоге, при обращении к методам классической теории принятия решений выбор сводится к бинарным отношениям предпочтения. Однако классические рациональные основания выбора не универсальны, а представляют собой лишь ограниченную часть оснований, на которых могут строиться разумные и естественные механизмы выбора решений. С целью упрощения построения и взаимодействия указанных механизмов (алгоритмов, методик и др.) для разных отраслей народного хозяйства целесообразным является построение типовых периметров (возможно интерфейсов) баз сбора и хранения больших данных (big data).
Число и сложность подобных проблем, для которых невозможно сразу получить критерий эффективности в аналитической форме, но мере развития цивилизации возрастает; возрастает также и цена неверно принятого решения. Для проблем принятия решения характерно, как правило, сочетание качественных и количественных методов. Принятие решений в системах управления промышленностью часто связано с дефицитом времени: лучше принять не самое хорошее решение, но в требуемый срок, так как в противном случае лучшее решение может уже и не понадобиться. Поэтому решение часто приходится принимать в условиях неполной информации (ее неопределенности или дефицита), и нужно обеспечить возможность как можно в более сжатые сроки определить наиболее значимые для принятия решений сведения и наиболее объективные предпочтения, лежащие в основе принятия решения.

4. Литература

[1] How Central Banks Are Using Big Data to Help Shape Policy [Электронный ресурс]. – Режим доступа: https://www.bloomberg.com/news/articles/2017-12-18/central-banks-are-turning-to-big-data-to-help-them-craft-policy (15.11.2018).

[2] The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales [Электронный ресурс]. – Режим доступа: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2022293 (05.11.2018).

[3] Вашко, Т.А. Технология дублирования информации как средство повышения качества принятия решений // Проблемы современной экономики. – 2011. – № 4. – С. 137-141.

[4] Насколько велик Интернет? [Электронный ресурс]. – Режим доступа: https://geektimes.ru/company/asus/blog/275032/ (25.10.2018).

[5] Проблемы принятия эффективного управленческого решения [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/problemy-prinyatiya-effektivnogo-upravlencheskogo-resheniya (25.10.2018).

[6] Прогноз социально-экономического развития Российской Федерации на период до 2036 года [Электронный ресурс]. – Режим доступа: http://economy.gov.ru/wps/wcm/connect/9e711dab-fec8-4623-a3b1-33060a39859d/prognoz2036.pdf?MOD=AJPERES&CACHEID=9e711dab-fec8-4623-a3b1-33060a39859d (15.11.2018).

[7] ЦБ используют big data для формирования финансовой политики [Электронный ресурс]. – Режим доступа: http://www.vestifinance.ru/articles/95398 (20.12.2018).

Description and formation of the database perimeter for systematization and storage of multi-structured data

A.A. Nechitaylo[1], O.I. Vasilchuk[2], A.A. Gnutova[1]

[1]Samara National Research University, Moskovskoe Shosse 34A, Samara, Russia, 443086
[2]Volga Region State University of Service, Gagarin st. 4, Togliatti, Russia, 445677

Abstract. For storage of big data, as a rule, relational databases are used. Financial resources, economists, and other technical data are used to provide more extensive research and analysis of the processes occurring in large economic systems. Therefore, when using the physical names of the studied regions, non-relational databases are technically more convenient.