УДК 004.65:004.7:004.75:004.738.5

В.И. Гриценко, А.А. Урсатьев

Большие Данные и инструментарий для аналитики

Проанализированы материалы известных зарубежных аналитических корпораций, исследовательских компаний в области ИТ и международных научных центров. Сквозь призму свойств и характеристик, природу и потенциальную ценность Больших Данных прослежено влияние на трансформацию ИТ к новому набору ключевых технологий, формирующих платформу обработки для извлечения новых знаний, обнаружения неочевидных связей и углубленного понимания, проникновения в суть явлений и исследуемых процессов.

Ключевые слова: Большие Данные, мультиструктурированные данные, управление данными для аналитики, аналитические хранилища, логические хранилища данных.

Проаналізовано матеріали відомих закордонних аналітичних корпорацій, дослідницьких компаній в області ІТ, міжнародних наукових центрів. Крізь призму властивостей і характеристик, природу й потенційну цінність Великих Даних простежено вплив на трансформацію ІТ до нового набору ключових технологій, що формують платформу обробки для добування нових знань, виявлення неочевидних зв'язків і поглибленого розуміння, проникнення в суть явищ і досліджуваних процесів.

Ключові слова: Великі Дані, мультиструктуровані дані, керування даними для аналітики, аналітичні сховища даних, логічні сховища даних.

Введение. Данные – неотъемлемый атрибут средств вычислительной техники. Исследование данных – научное направление, сформировалось спустя определенное время после создания первых вычислительных машин и накопления опыта автоматизации инженерных и научно-технических расчетов. Эти направления исследований имеют выраженную, достаточно сложную историю развития и прямую зависимость от научно-технического уровня и возможностей средств переработки информации и их использования.

Значительным рывком в этой области стало развитие автоматизированных систем управления, проектирования и обработки данных. Как известно, такие системы базируются на больших объемах данных, имеющих достаточно сложную структуру: от массивов до совокупности данных, организованных в базы данных (БД) под управлением СУБД, и хранилища данных. Бизнес-решения в течение многих десятилетий принимались на основе системы, известной как *Business Intelligence* (*BI*), которая традиционно взаимодействовала со структурированными данными из относительно ограниченного пула корпоративных данных.

Такая форма взаимодействия существенно сужала область поиска информации для принятия решений. Ощущалась необходимость в привлечении достаточно широкого и разнообразного спектра, различных по своей природе цифровых данных для аналитики, объем которых в мире стремительно расширяется [1–10], и будет возрастать более чем в два раза каждые два года. Источником данных могут быть научные исследования, производственная, социальная и другие сферы деятельности мировых сообществ.

Получаемые данные [1–10] относятся непосредственно к словосочетанию Большие Данные (*Big Data*), которое получило широкую известность в эпоху петабайта, когда решение проблемы больших потоков данных приобрело особое значение для развития приоритетных направлений мировой науки [11, 12].

Уже показано [1], что использование Больших Данных позволит создать прирост национальных и мировой экономик, существенно повысит эффективность функционирования и конкурентоспособность организаций частного и государственного секторов, управление социальными проектами. На современном этапе данные превращаются в капитал, который не-

посредственно участвует в формировании и управлении различных сфер деятельности человека.

Цифровой мир и Большие Данные

Последнее десятилетие ознаменовалось значительным числом публикаций [1–10] об исследованиях «цифрового Мира или цифровой вселенной» – данных, которые создаются и ежегодно копируются. Приведем примеры. Хранилище Европейского института биоинформатики (*European Bioinformatics Institute*, *EBI*)[1] в Великобритании на время подготовки публикации содержало 20 Пб биологических данных и резервных копий о генах, белках и малых молекулах; Пекинский институт геномики (*Beijing Genomics Institute*, *BGI*) ежедневно генерирует шесть терабайт геномных данных; Хранилище данных в *CERN* Европейской лаборатория физики элементарных частиц в Швейцарии ежегодно увеличивается примерно на 15 Пб данных, генерируемых при столкновении частиц на Большом адронном коллайдере и др.

Рост объемов данных, получаемых в процессе проведения научных и производственных экспериментов, наблюдается не только в биологии, но и в таких областях как астрономия, физика, химия, геология, медицина и др. Широкое использование мультимедиа в здравоохранении и клиенто-ориентированной индустрии в значительной степени способствует росту больших данных. Видео генерирует огромное количество данных. Каждая минута наиболее часто используемого видео во время хирургической операции генерирует в 25 раз больший объем данных (в минуту), чем содержат цифровые изображения высокого разрешения при компьютерной томографии (*Computed Tomography*, *CT*). При том, что каждое из этих неподвижных изображений по объему уже в тысячи раз занимает больше байт, чем одна страница текста или числовых данных. Заметим, что более 95 процентов клинических данных, полученных в здравоохранении, в настоящее время представлено в видеоформате. На мультимедийные данные уже приходится более половины магистрального интернет-трафика (т.е. трафика, передаваемого по крупнейшим каналам между основными Интернет-сетями). Рост этой доли будет наблюдаться и в дальнейшем [1].

В промышленности проектирование и производство изделий, где, например автомобильные и аэрокосмические компании могут оценить сотни или тысячи виртуальных прототипов космической станции и наилучший дизайн автомобиля, порождает значительное число данных. Еще одним примером служат новые крупномасштабные эксперименты, генерирующие ежедневно петабайт разнородных данных для комплексной имитационной модели [13].

Росту быстро расширяющихся массивов данных будут продолжать способствовать основные установившиеся тенденции в области традиционных транзакционных БД, мультимедийного контента, в том числе видеонаблюдений, и все большую популярность будут приобретать социальные медиа и распространение *IoT* – Интернета вещей.

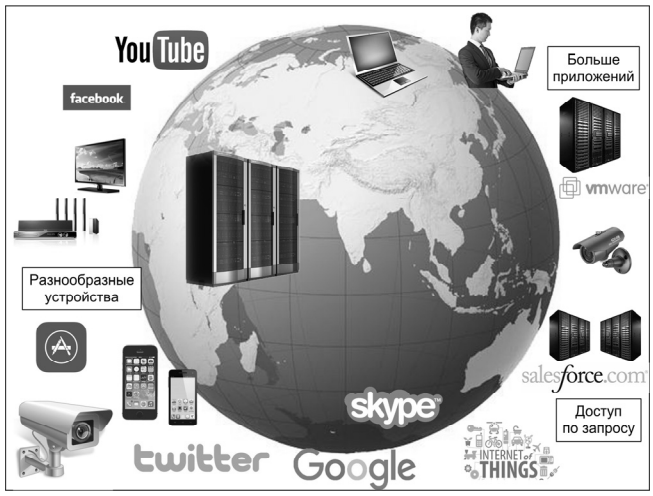
Развитие персонификации в решении повсеместных проблем требуют сбора данных с большей частотой и детализацией, захватывая каждую транзакцию клиента, добавляя персональную информацию, а также собирая подробную информацию о поведении потребителей в различных средах. В потребительском рынке, ежедневно общаясь, выполняя поиск и просматривая всевозможные товары, совершая покупки, обмениваясь мнениями потребители создают собственные огромные «следы» данных.

Для описания цифрового следа, который пользователи оставляют на сайтах, был принят специальный термин – *exhaust data* (выхлоп, выброс или отработанные данные). Под этим подразумевается побочный продукт взаимодействия пользователей в Интернете: где и что они выбирают, как долго смотрят на страницу, где проводят курсором мыши, что печатают и т.д. Эти отработанные данные, привнося свой вклад в растущий объем больших данных, могут оказать существенную пользу. Так, например, цифровые потоки медиа-данных активизируют приток

[1] Входит в Европейскую лабораторию молекулярной биологии.

мнений и отношений, ценных для управления взаимоотношениями с клиентами [1].

Один из примеров источников роста Больших Данных приведен на рисунке.



Социальные данные включают в себя потоки обратной связи с клиентами – микро-блогами сайтов, такими как *Twitter*, медиа-платформы *Facebook*, *Foursquare* и другими – очевидные примеры новых сегментов роста больших данных. Уже созданы системы, в которых потребители генерируют почти непрерывный поток данных о себе, распространяемый быстрыми темпами благодаря *сетевому эффекту*. Смартфоны стали отличной иллюстрацией того, как используемые мобильные устройства создают дополнительные источники данных, учитываемых в настоящее время и включающих в себя текстовые сообщения, историю посещаемых страниц, географическое положение благодаря *GPS* и другую информацию [1, 7, 14]. Эта деятельность одновременно увеличивает потребность в аналитическом потенциале.

Распространение сенсоров приложений в *IoT*, т.е. датчиков и исполнительных механизмов, встроенных в физические объекты и связанных с вычислительными ресурсами преимущественно беспроводными сетями – это еще одна тенденция, стимулирующая рост больших данных [1, 3, 15, 16]. Интернет вещей – следующий виток эволюции Интернета обещает массу сервисов для промышленности, в частности, транспорта, «умных» энергосетей, систем безопасности, здравоохранения, образования и потребительской электроники. *Gartner* прогнозирует, что через 10–15 лет *IoT* будет объединять 26 млрд устройств – рост в сравнении с текущими годами в 30 раз [16]. Возрастут [3] до 10 процентов данные встраиваемых систем, сигналы которых служат основным компонентом *IoT*.

Исследования *McKinsey* предполагают [1], что число взаимодействующих устройств[2], соединенных с *IoT* [15], будет расти со скоростью, превышающей 30 процентов в год в течение следующих пяти лет.

Некоторые из ожидаемых секторов роста – это коммунальные службы, обеспечивающие установку «умных» счетчиков и интеллектуальных приборов; здравоохранение, которое разворачивает удаленный мониторинг здоровья; розничная торговля, все чаще использующая метки *RFID* на товарах потребления и автомобильная промышленность, увеличивающая количество датчиков в транспортных средствах.

Следует упомянуть и о вживляемых медицинских устройствах. В будущем сенсоры всех видов, в том числе импланты, будут собирать виртуальные и реальные биометрические данные. Мониторить медицинскую эффективность, сравнивать физическую активность и здоровье, отслеживать вспышки вирусов – все это будет возможно в режиме реального времени [7].

Определение Больших Данных. До настоящего времени общепринятое определение больших данных отсутствует. Вместе с тем предпочтение отдают двум из них, приводимых далее. Первое [1]: «Большие Данные – это наборы данных, размеры которых выходят за пределы возможностей по сбору, хранению, управлению и анализу, присущих обычному программному обеспечению БД». Другое[3], по мнению авторов, аналогично первому, дополняя

[2] Межмашинное взаимодействие – физическая основа *IoT* – набор технологий, позволяющих устройствам обмениваться информацией, вернее, данными или передавать их в одностороннем порядке. Посредством *M2M* осуществляется доступ к удаленным объектам.

[3] *Merv Adrian*. Big Data. Teradata Magazine Q1/2011, P. 1–5. © 2011 Teradata Corporation – *docshare*04.*docshare.tips*/*files*/20905/209055375.*pdf*

его в части съема данных с объекта и доставки их к устройству обработки, и звучит так: Большие Данные – это данные, сбор, управление и обработку которых невозможно осуществить с помощью наиболее часто используемых аппаратных средств и программных инструментов в течение допустимого для пользователя времени.

Эти оба определения субъективны, так как не ставится вопрос, каковым должен быть набор данных, чтобы считать его Большими Данными. С этими оговорками Большие Данные могут варьироваться. Поэтому исследователи многих стран приходят к заключению, что определение Больших Данных будет изменяться во времени по мере развития технологий.

Ключевые характеристики, свойства Больших Данных. Характеристики предопределяются особенностями среды больших данных – это и значительные объемы данных, большая часть которых представлена нетрадиционными форматами, требование обработки со скоростью их поступления и комбинирование типов информации из множества источников. Различают в основном три ключевые характеристики, определяющие большие данные. Это так называемые «три *V*»: объем (*Volume*), разнообразие (*Variety*) и скорость (*Velocity*). Отдельные исследователи вносят в группу ключевых характеристик и ценность (*Value*) [1, 7, 13, 14].

Объем (*Volume*). Эта характеристика – наиболее заметный параметр Больших Данных. Исходя из приведенного определения термина *Большие Данные* [1], его следует воспринимать как относительный. Некоторые отрасли вырабатывают, скорее всего, гигабайт или терабайт данных, в отличие от петабайт или экзабайт, получаемых при исследованиях физики элементарных частиц на Большом адронном коллайдере, или присущих таким компаниям как автомобильные, аэрокосмические и социальные сети в рассмотренных ранее примерах. Тем не менее, эти, казалось бы, небольшие выборки могут нуждаться в интенсивной и сложной обработке и анализе. Индустрия финансовых услуг демонстрирует эту динамику изменений. Определенные виды работ с большими данными могут потребовать рассмотрения множества записей, в то время как каждая из них может быть объемом всего в несколько байт (например, тикер[4] – код акции компаний в информации о котировках ценных бумаг). Напротив, архивы электронной почты могут накапливать несколько петабайт данных, содержащих ценные предложения, наиболее информативные записи отложенных и текущих дел, учет проектов, юридических записей, контрактов и др.

Разнообразие (*Variety*). Сам по себе цифровой мир, безусловно, включает в себя все типы данных. Однако большинство генерируемых новых данных неструктурированы (см. рисунок). Знания об этих данных возможны, если они каким-то образом промаркированы[5] – практика, которая приводит к метаданным. Последние – один из быстро растущих сегментов цифрового мира, хотя они и составляют небольшую часть этого мира (около 23 процентов). Прогнозируют, что к 2020 г. треть данных цифровой вселенной (более чем 13 Зб) будут отнесены к объему больших данных, в случае, если они будут размечены и проанализированы. Неразмеченная информация ожидает извлечения из нее ценной информации [7], так как при добавлении услуг или выполнения маркетинговых кампаний новые виды данных необходимы для извлечения полезной и необходимой информации.

Традиционные форматы данных, как правило, относительно хорошо определены в схеме данных и изменяются медленно. В отличие от них, нетрадиционные форматы – многочисленные неструктурированные и полуструктурированные данные, имеющие различную природу: файлы текстов, документы, электронные пись-

[4] Тикер (*ticker, ticker symbol, stock symbol*), код акции, код инструмента – краткий набор букв латинского алфавита (цифр и вспомогательных символов), который используется для обозначения акций компаний. Тикеры используются для упрощения представления, например, тикер *Google – GOOG*.

[5] В качестве маркеров конкретных покупателей используются банковские и именные скидочные карты.

ма, посты в социальных медиа-, видео- и аудиозаписи, *GPS*-координаты, *web*-страницы, системные регистрационные журналы и другое – демонстрируют стремительные темпы роста. Это многообразие данных охватывает термином *мультиструктурированные данные*, поскольку большинство из них имеет структуру, не поддающуюся описанию некой единой схемой. Полуструктурированные данные подразумевают логическую схему и формат, который может быть понятным, но не дружественным к пользователю (например, *weblogs*).

Сочетание широкого спектра типов данных из множества источников различной природы является ключевым критерием при определении, может ли приложение рассматриваться в контексте больших данных [1, 7, 13, 15, 17]. Этот существенный аспект больших данных влияет как на технические решения, так и на возможные результаты. Комбинирование типов информации – сложная техническая задача: какова относительная весомость сообщения в твитере относительно записи потребителя? Как объединить огромное количество изменяющихся записей пациентов с опубликованными медицинскими исследованиями и геномными данными в поиске лучшего лечения для конкретного пациента? Примером этого может служить неоднородный контент внутренних оперативных данных из *ERP*-системы с частично структурированными данными из файлов веб-журналов, которые идентифицируют поведение онлайновых клиентов с анализом структуры поведения из неструктурированного текста комментариев. Другой пример – расширенное моделирование погоды/климата, в формировании которых используются метеорологических данные на протяжении 100 лет, и новые физические модели поведения океана, изменения уровней *CO*, что объединяется с поступающими данными со спутниковых каналов для обеспечения моделирования в реальном режиме времени [13].

По мнению авторов, именно здесь уместно упомянуть о некоторых из затронутых *IDC* проблем пользователей больших данных [13]. В опросе, проведенном *IDC* в начале 2012 г., ИТ и бизнес-группы, выразили озабоченность по поводу необходимости переоценки измеряемой совокупности данных для поддержки принятия решений. При этом наиболее часто упоминалась проблема оценки релевантности больших данных. Часто организации пересматривают условия анализа существующих и новых источников данных для улучшения или изменения процессов принятия компетентных решений в рамках своих организаций.

Скорость (*Velocity*), с которой информация принята, проанализирована и передана. Ключевым в оценке требований Больших Данных к скоростным характеристикам операций над ними, является понимание задач конечных пользователей: моделирование сложных явлений, обнаружение неочевидных связей и скрытых закономерностей, обработка сообщений в социальной сети *Twitter* и других бизнес-организационных процессов [13].

В то же время, поисковые машины должны обрабатывать миллиарды запросов для определения совпадений рекламных объявлений, однако нет насущной необходимости выполнять этот анализ в режиме реального времени. Другими словами, необходима достоверная информация в нужное время и с необходимой степенью точности.

Так, система *BI* ранее опиралась, как правило, на хранилища, обеспечивающие доступ к использованным оперативным данным. В требованиях к *BI*-системам все чаще высказывалось желание работать со «свежими данными». Временной лаг в несколько часов или дней для помещения оперативных данных в хранилище с целью их последующего анализа в определенных ситуациях уже считается недопустимым – полная и точная информационная картина сегодня нужна в реальном времени непосредственно в ходе выполнения бизнес-процесса [18].

Ценность (*Value*). В контексте Больших Данных *ценность* относится как к стоимости оборудования и технологий (капитальные или постоянные затраты), так и к стоимости, привнесенной большими данными. Сбор информации имеет решающее, но не исчерпывающее значение,

поскольку существенная часть ценности находится в применении, а не хранении как таковом. Эффект от использования больших данных из разнообразных источников существенно различается. Обычно он скрыт в нетрадиционных данных. Задача состоит в определении того, что ценно, а затем осуществляется трансформация и извлечение этих данных для анализа совместно с существующими историческими данными, так как их комбинированный набор повысит ценность в силу синергетического эффекта (греч. *synergos* – вместе действующий).

Конечная ценность больших данных будет оцениваться на основе одного или более из трех критериев [1, 13, 14, 17] с учетом оценки полезности и точности информации и сокращения времени получения ответа.

Суть стоимости данных заключается в их неограниченном повторном использовании – альтернативной ценности. Абсолютная ценность данных может намного превышать ту, которую удается извлечь при первичном использовании. Инновационные компании могут извлечь скрытую ценность и потенциально получить огромные преимущества, т.е. ценность данных следует рассматривать как возможности их дальнейшего использования, а не только нынешнего [1].

Ярким примером извлечения скрытой ценности из использованных данных служит работа [19] Интернет-компании *Google* в научном журнале *Nature*, в которой опубликован один из способов раннего выявления эпидемии гриппа путем мониторинга обращений за медицинской помощью в форме онлайн-запросов к поисковым системам. Поскольку относительная частота определенных запросов ощутимо коррелирует с процентом посещений врача, когда пациент представлен гриппоподобными симптомами, можно оценить текущий уровень недельной активности гриппа в каждом регионе США с отставанием примерно на день. Прогноз компании *Google* был основан на повторном анализе большого набора данных, использованных ранее и сохраненных.

Другой пример извлечения скрытой ценности – полезная информации из цифрового следа – «выброса данных» – побочного продукта других видов деятельности сообщества пользователей. Многие компании собирают выбросы данных и осуществляют их обработку для улучшения существующих или разработки новых служб, применяя принцип рекурсивного обучения на основе собранных данных. Каждое действие пользователя считается сигналом, который анализируется и возвращается в систему обучения. Выбросы данных – это механизм, положенный в основу многих компьютеризированных служб, таких как распознавание голоса, спам-фильтры, переводчики и др. Когда пользователь указывает в программе распознавания голоса, что она неправильно поняла произнесенное слово, он, по сути, обучает систему, совершенствуя ее.

Методы и технологии анализа Больших Данных

Широкий спектр методов и технологий разрабатывается и адаптируется для слияния, манипулирования, анализа и визуализации структурированных и неструктурированных данных значительных объёмов и многообразия для получения воспринимаемых человеком результатов обработки больших данных. Они предполагают статистические методы анализа данных, информатику, прикладную математику и экономику и свидетельствуют о гибком междисциплинарном подходе. С позиции экспертов [1], Большие Данные – это продвинутые аналитические решения, которые разрабатываются учеными разных профилей и опираются на результаты анализа больших данных. Они должны быть встроены в инструменты анализа так просто и привлекательно, что ими будут стремиться пользоваться ежедневно.

IDC определяет методы обработки Больших Данных как новое поколение технологий и архитектур, предназначенных для извлечения экономичной выгоды из очень больших объемов разнообразных данных, обеспечивающих высокую скорость съема и анализа. Это определение справедливо для аппаратных средств, программного обеспечения и услуг, которые объединяют, организовывают, анализируют и представляют результаты интеллектуальной отчетности [13, 14].

Аналитика, разработанная для принятия решений, в отличие от традиционных транзакционных систем поддержки текущего функционирования предприятия, использует для достижения своих целей хранилища данных – *DW* (*Data Warehouse*). Так, например, продажа товаров или авиабилетов проводится с использованием БД, предназначенных для обработки коротких транзакций, а анализ динамики продаж за определенный период, позволяющий спланировать работу с поставщиками или выявить пассажиропоток по направлениям и категориям пассажиров, – посредством хранилища данных. Фактором, приведшим к разделению аналитических и транзакционных систем, являются разные требования, предъявляемые двумя подсистемами (*OLAP* и *OLTP* [18]) к вычислительным ресурсам. Сложные, непредвиденные запросы могут привести к непредсказуемой нагрузке на оперативные БД и тогда обработка коротких записей, имеющих достаточно высокие требования к времени отклика и пропускной способности системы, станет практически недостижимой. Необходима среда хранения данных, используемых для аналитики [20, 21].

Хранилища данных, в общем случае, в зависимости от принятой архитектуры, включают в себя средства *ETL* извлечения, преобразования и загрузки данных в хранилище; витрины данных, которые хранятся в структурах, оптимальных для решения конкретных задач пользователей; репозитории данных, метаданных и др. Эффективная работа аналитических систем требует единства репозитория, для наполнения которого необходимо предварительно согласовать разнородные данные из различных источников. Структура репозитория *DW* призвана максимально полно и быстро удовлетворять потребности в информации [20–22].

Хранилища данных[6] в настоящее время содержат более широкие решения управления данными для аналитики: способность принимать, преобразовывать внешние данные и управлять ими в сочетании с традиционными внутренними источниками. По мнению *Gartner* [23, 24], *DW* – это совокупность наборов данных двух или более разрозненных источников, используемых при необходимости совместно в рамках интегрированной, допускающей изменение во времени, стратегии управления информацией. Его логическое проектирование подразумевает возможность гибкого подключения дополнительных источников данных без привнесения существенной модификации в работающую систему. *DW* может быть значительно больше, чем объем данных, размещаемых в отдельной СУБД (*DBMS*), особенно в случаях управления распределенными данными.

Ориентация хранилищ на выполнение аналитической обработки предполагает наличие средств управления данными для аналитики (*Data Management Solution for Analytics*, *DMSA*). Вместе они составляют системы, которые выполняют аналитическую обработку, необходимую для поддержки принятия решений, и могут быть расширены к новым структурам и типам данных, таких как *XML*, текст, документы, геопространственная и другая информация, а также доступ и управление внешними файловыми системами.

DMSA определяется как завершенная система ПО, которая поддерживает и управляет данными в одной или ряде различных файловых систем для одной (обычно) или нескольких БД, выполняющих реляционную обработку, даже если данные не хранятся в реляционной структуре, и обеспечивает доступ к данным из независимых аналитических инструментов и интерфейсов. Кроме того, *DMSA* должна поддерживать доступность данных для независимого прикладного программного обеспечения (*front-end*), включать в себя механизмы для изоляции различных типов рабочих нагрузок друг от друга, а также управлять различными параметрами доступа конечного пользователя в рамках управляемых экземпляров данных.

Собственно хранилища могут содержать *DMSA* или быть частью более значительной

[6] См. отчеты *Gartner* «Магический квадрант для хранилищ данных и средств управления данными для аналитики» (*Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics*). – 12 *Febr.* 2015 и 25 *Febr.* 2016.

системы, в том числе с использованием независимых средств управления. *DMSA* не конкретный класс или тип технологии, они могут состоять из комбинации различных технологических решений. Основное требование к ним – управление хранением и доступом к данным в некоторой запоминающей среде (жесткие диски, флэш-память, твердотельные накопители и др.) и возможность предоставления открытого доступа к данным.

На сегодня главным направлением развития технологии создания аналитических *DW* являются гибридные, хранящие структурированные данные и контент, технологические платформы, которые расширят хранилища данных за пределы какой-либо текущей практики [25], и логические хранилища (*Logical Data Warehouse*, *LDW*) – новая архитектура, сочетающая преимущества традиционных хранилищ (репозитория) [22] с альтернативными стратегиями управления и доступа к данным [23–27].

Само понятие *хранилище данных* сегодня уже не означает реляционный интегрированный репозиторий. По *Gartner*, с некоторого времени *DW* из решений, функционирующих исключительно как репозиторий, стали трансформироваться в направлении систем, поддерживающих согласованную обработку и логику предоставления информации в соответствии с новой моделью логического хранилища данных [23–27]. Из этого следует название *логические DW* – поскольку оно фокусируется на логике информации. *LDW* занимают граничное положение между подходом централизованных хранилищ [20–22] и управляемым сервисом данных для аналитики.

По сути, *LDW* – это архитектурный слой в вершине хранилища, состоящий преимущественно из сервисов и метаданных, который оказывает услугу предоставления данных по запросу аналитического приложения, – своеобразная логическая информационная платформа доставки. Требуемые данные в едином представлении доставляются без перемещения в пул ресурсов приложений.

Данные собираются через метаданные, виртуализацию данных, распределенную обработку, где высокопроизводительные инструменты заняты конкретными задачами их подготовки, сложной оптимизацией и возможностями управления. В результате обеспечивается доступ ко всем видам информационных ресурсов, что является значительным шагом в интеграции информации. Логическое хранилище данных предоставляет платформу информационных услуг для приложений.

Инфраструктура централизованного управления данными гибко управляет основными компонентами структуры хранилища: репозиторием данных и метаданных, виртуализацией данных и распределенными вычислениями, таксономическим и онтологическим разрешением (процесс разделения данных на структурированный набор отдельных компонентов по соответствующей схеме), соглашением об уровнях обслуживания *SLA* и др. [23–27]. Она также предусматривает системный аудит и механизм принятия решений, объединенные в общую инфраструктуру, чтобы определить, какие из доступных решений в области данных больше всего соответствуют условиям соглашения об уровне обслуживания и результатам системного аудита [26].

Поскольку *LDW* рассматриваются в контексте Больших Данных с присущими им высокими требованиями к параметрам обработки, то, сместив акцент с концепции репозитория, внимание акцентируется на времени доступа к данным и производительности. Так, основные решения, используемые в *DW*, основаны на архитектуре с массовым параллелизмом обработки (*Massively Parallel Processing*, *MPP*) – классе параллельных вычислительных систем, состоящих из множества узлов, организованных по принципу *shared nothing* или *shared everything*.

В первом случае каждый узел независим и самодостаточен, т.е. системные ресурсы (память, диски) не разделяются с другими узлами, во втором – узлы используют разделяемые ресурсы. Каждый подход имеет свои преимущества и недостатками – *shared nothing* не позволяет утилизировать все ресурсы достаточно эффективно. *Shared everything* выполняет эту задачу более эффективно, но вследствие согла-

сования использования разделяемых ресурсов, нуждается в дополнительном времени.

Преимущества *MPP*-архитектуры очевидны – это линейная масштабируемость, обеспечивающая стабильные и предсказуемые параметры производительности. Тенденция в области *DW* такова, что будущее за архитектурой *MPP* при соблюдении соответствия функциональным требованиям *ACID*, когда коммуникационная сеть между узлами должна обладать высокой пропускной способностью и отказоустойчивостью.

Колоночная (*columnar*) или полиморфная модель хранения данных в задачах аналитики также позволяет существенно сократить время доступа к требуемой информации при оптимизации чтения больших объемов данных с дисков. При этом аналитические выборки в *SQL*-запросе пользователя, как правило, содержат не больше семи-восьми полей. Объемы данных при этом могут быть достаточно большими (по 300–500 Тб и даже более чем с петобайтными размерами). Поколоночное представление данных используется в системах с подавляющим большинством операций типа *чтение*, как правило, в аналитических системах класса *BI – ROLAP –* и хранилищах данных [28, 29].

Производительность и ценность информации как актива, определяемые исходя из простоты доступа к информации и возможности применять ее разными способами, как пишут аналитики *Gartner*, станет новым и наиболее существенным показателем ценности хранилища данных. В противном случае то огромное количество информации в корпоративных хранилищах и в Интернете может остаться невостребованным из-за размера данных, их сложности и несопоставимой природы, короткого периода актуальности ряда данных и различающихся требований обработки и объединения. Концепт *LDW*-хранилища позволяет анализировать и контролировать Большие Данные, чтобы содержащаяся в них информация становилась явной, доступной и актуальной. Это один из новых перспективных подходов к хранилищам и управлению аналитическими данными [23, 25–27].

Наряду с *LDW*-хранилищем продолжают успешно развиваться и традиционные корпоративные хранилища данных – *Enterprise Data Warehouse* (*EDW*), представляющие собой вариант интегрированной, предметно-ориентированной и физически централизованной системы управления данными, построенной на базе аппаратных средств, оптимизированных для выполнения сложных запросов. Получил распространение также класс хранилищ, в первую очередь ориентированных на высококвалифицированных пользователей (исследователей данных), решающих нестандартные аналитические задачи. Это – контекстно независимые хранилища (*Context-Independent Data Warehouse*), в которых имеется возможность изменять схемы чтения данных, что позволяет получать новые информационные срезы и извлекать дополнительные необходимые сведения. Достигается это путем использования средств поиска, механизмов графов и других расширенных возможностей для создания новых информационных моделей [24].

Кроме того, нельзя принебречь гибким подходом (*agile approach*) к созданию аналитических *DW*, вызванным высокими затратами при неспособности быстро адаптироваться к изменению условий внешней экономической среды [25].

Создание аналитических баз данных – сложный, требующий затрат времени и значительных средств процесс. При этом, как показывает зарубежный опыт, 70–80 процентов проектного бюджета тратится прежде, чем будет получена какая-либо ценность для бизнеса. Это означает, что необходимо минимизировать усилия, связанные с интеграцией и согласованием данных. В модели гибкого внедрения используются только данные, необходимые для решения специфических проблем бизнеса, вместо того, чтобы предпринимать огромные усилия по интеграции больших объемов данных.

Существует достаточно примеров, когда сторонники гибкого подхода работают с представителями бизнес-подразделений, чтобы определить, скажем, «сотню» элементов данных, которые увеличат эффективность и ценный результат можно будет достигнуть быстрее – за недели или месяцы, а не кварталы или годы.

Как следствие, гибкие методы снижают риски и создают системы с высоким уровнем принятия их пользователями [25].

Заключение. Сложные структуры, возрастающие объемы Больших Данных и короткий период актуальности ряда из них дали импульс интенсивному развитию новых технологий сред хранения и обработки. Появление широкого класса нереляционных и колоночных, наряду с традиционными, БД – горизонтально масштабируемых, поддерживающих мультиструктурированное хранение данных, минимизация времени доступа к требуемой информации путем оптимизации чтения данных с дисков, применение архитектуры с *MPP*, поддержка систем БД в оперативной памяти и многие другие решения, которые авторы предполагают раскрыть в последующих работах, позволили существенно увеличить производительность вычислений, способствующих скорейшему достижению результатов аналитических исследований Больших Данных в различных сферах деятельности человека.

Значительное место занимает подготовка данных к анализу. Используемый консолидированный анализ даже при небольшом числе внешних источников данных становится слишком затратным, а собственно создание БД требует времени и значительных средств для введения их в эксплуатацию. В связи с этим интерес смещается с традиционных хранилищ в сторону гибридных и логических решений. В ограниченном применении возможно использование гибкого подхода, направленного на уменьшение размерности Больших Данных.

Ориентация хранилищ на широкое применение, в том числе и в задачах интеллектуальной аналитической обработки с присущим им глубоким проникновением в суть исследуемых явлений, требует простого обеспечения доступа ко всем имеющимся видам информационных ресурсов, возможности их использования разными способами, а также доставки требуемых данных к ресурсам приложений без их перемещения.

Согласуя концепт логических хранилищ в целом, *LDW*-хранилище представляется авторам в виде слабо связанной, обеспечивающей необходимую гибкость, многослойной архитектуры, которая содержит новые технологии и прогрессивные средства выявления, извлечения и подготовки широкого спектра данных. В таком случае анализ с учетом логики обработки информации предоставляет определенный информационный ресурс приложению в качестве услуги. Так будет обеспечена простая доступность и актуальность информации, содержащейся в Больших Данных.

Большие Данные – сложная межотраслевая научно-техническая проблема, к которой очевиден научный и коммерческий интерес. Будущие технологии обработки Больших Данных с включением в ее контуры методов и средств интеллектуализации получит широкое применение в экономической, производственно-технологической сфере и в других важных областях деятельности человека.

1. *Big* Data: The next frontier for innovation, competition, and productivity / J. Manyika, M. Chui, B. Brown et al. – May 2011. – http://www.mckinsey.com/businessfunctions/business-technology/our-insights/big-data-the-next-frontier-for-innovation
2. *EMC*. Research and analysis of IDC «Digital universe study» commissioned by EMC Corporation EMC. – 2014. – http://ukraine.emc.com/leadership/digital-universe/index.htm#Archive (In Russian).
3. *The Digital* Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. EMC Digital Universe with Research & Analysis by IDC. – April 2014. – http://www.emc.com/leadership/digital-universe/2014iview/index.htm
4. *The expanding* digital universe March 2007 / J.F. Gantz, D. Reinsel, Chute Chr. et al. – 24 марта 2015. – http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf
5. *Gantz J*., *Reinsel D.* The Digital Universe Decade – Are You Ready? – May 2010. – http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf
6. *EMC* NEWS. Press Rellease. New Digital Universe Study Reveals Big Data Gap: Less Than 1% of World's Data is Analyzed; Less Than 20% is Protected. – http://www.emc.com/about/ news/press/2012/20121211-01.htm
7. *Gantz J*., *Reinsel D.* The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. – Dec. 2012. – http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf
8. *Lesk M*. How Much Information Is There In the World? – 1997. – www.lesk.com/ mlesk/ksg97/ksg.html

9. *Lyman P*., *Hal R. Varian* How much information? 2003 (School of Information Management and Systems, Univ. of California at Berkeley). – http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/
10. *Hilbert M*., *López P.* The World's Technological Capacity to Store, Communicate, and Compute Information // Published Online Feb. 10 2011, Science 1 April 2011. – **332,** N 6025. – P. 60–65. DOI: 10.1126/science. 1200970. – http://www.sciencemag. org/content /332/6025/ 60.full
11. *Big* Data // Nature. – 2008. – **455**, N 7209. – P. 1–136. – http://www.nature.com/nature/journal/v455/n7209/ index.html
12. *Marx V*. Biology: The big challenges of big data // Nature International weekly journal of science. – 2013. – **498**, N 7753. – P. 255–260. – http://www.nature.com/nature/journal/ v498/n7453/full/498255a.html
13. *Olofson Carl W*., *Vesset Dan* Big Data: Trends, Strategies, and SAP Technology. – August 2012. – https://www.sap.com/bin/sapcom/en_ae/downloadasset.2012-09-sep-26-13. idc-report-big-data-trends-strategies-and-saptechnology-pdf.html; – http://www.itexpocenter.nl/iec/sap/BigDataTrendsStrategiesandSAPTechnology.pdf
14. *Gantz J*., *Reinsel D.* Extracting Value from Chaos. – June 2011. – https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf
15. *Chui M*., *Loffler M.*, *Roberts R.* The Internet of Things // McKinsey Quarterly. – March 2010. – http://www.mckinsey.com/industries/high-tech/our-insights/the-internet-of-things
16. *Uont R.*, *Shilit B.* The mechanisms of the Internet of things, Otkrytye sistemy. – 2015. – № 1. – C. 38– 42. (In Russian).
17. *Oracle*: Big Data for the Enterprise. – June 2013. – http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf
18. *Gritsenko V.I*., *Oursatyev A.A.* Information Technologies: the Tendency, the Ways of the Development // Upr. sist. mas. – 2011. – № 5. – C. 3–20. (In Russian).
19. *Detecting* influenza epidemics using search engine query data / J. Ginsburg, M. Mohebbi, R. Patel et al. – // Nature. – 2009. – **457**. – P. 1012–1014. – http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html
20. *Asadullaev S.* Data Warehouse Architectures-1, -2, -3. – 2009. – http://www.ibm.com/developerworks/ru/library/sabir/axd_1/index.html ... axd_3/ index.html (In Russian).
21. *Asadullaev S.* Data, metadata and NSI: triple storage strategy. – 2009. – http://www.ibm.com/developerworks/ru/library/r-nci/index.html (In Russian).
22. *The Practice* of Building Data Warehousing: The SAS System. – Open systems. – 1998. – № 4–5. –http://www.osp.ru/dbms/1998/04-05/13031592/ (In Russian).
23. *Mark A. Beyer*, *Roxane Edjlali.* Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics. – 12 Feb. 2015. – http://www.gartner.com/technology/reprints.do?id=1-2A21OQO&ct=150217&st=sg
24. *Roxane Edjlali*, *Mark A. Beyer.* Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics. – 25 Feb. 2016. – https://www.gartner.com/doc/reprints?id=1-2ZFVZ5B&ct=160225&st=sb
25. *Data* warehouses: the market is being transformed. On the materials of foreign sites // Intersoft Lab. – 2012. – http://www.iso.ru/print/rus/journal/document10179.phtml (In Russian).
26. *TDWI*. The Logical Data Warehouse: What it is and why you need it. – June 24, 2015. – https://tdwi.org/webcasts/2015/06/the-logical-data-warehouse-what-it-is-and-why-you-need-it.aspx
27. *ThoughtWeb*. Logical Data Warehousing for. Благодаря исследованиям Gartner – http://imagesrv. gartner.com/media-products/pdf/samples/ sample3.pdf
28. *Column* DBMS – the principle of operation, advantages and scope. – 28 Jan. 2011. – http://habrahabr.ru/post/95181/ (In Russian).
29. *Whitehorn M*. Big Data Technologies emerge to battle large, complex data sets. – 05 Dec. 2011. – http://www.computerweekly.com/news/2240111952/Big-data-technologies-emerge-to-battle-large-complex-data-sets или https://www.prj-exp.ru/dwh/big_data_technologies_emerge_to_battle.php

Поступила 08.08.2017
Тел. для справок: +38 044 526-4159 (Киев)
E-mail: *aleksei@irtc.org.ua*

UDC 004.65:004.7:004.75:004.738.5

V.I. Gritsenko[1], A.A. Oursatyev[2]

[1] Corresponding Member of the NAS of Ukraine, International Research and Training Centre of Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave., 40, Kyiv, 03187, Ukraine,

[2] PhD in Techn. Sciences, Leading Research Associate, International Research and Training Centre of Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave., 40, Kyiv, 03187, Ukraine, aleksei@irtc.org.ua

Big Data and Tools for Analytics

Keywords: Big Data, multi-structured data, data management solution for analytics, analytic warehouse, logical data warehouse.

Introduction. The materials of the well-known foreign analytical corporations, the research IT-companies, and the international scientific centers are analyzed. The influence on transformation of the IT to a new set of key technologies creating

the platform for their processing is traced through a prism of the properties and characteristics, the nature and potential value of Big Data.

Purpose. It is important to create and research the methods and technologies for processing the Big Data for extracting new knowledge, discovering non-obvious links and in-depth understanding of the phenomena and the investigated processes and the prospects of their development.

Methods. The informational and analytical methods and technologies for data processing, the methods for data assessment and forecasting, taking into account the development of the most important areas of the informatics and information technology.

Results. Consolidated analysis, even with a small number of external data sources, is too expensive. Actually, creating a DB takes time and the significant resources for putting them into operation. In this regard, interest shifts from the traditional stores to hybrid and logical solutions. In a limited application it is possible to use a flexible approach aimed at reducing the dimension of the Big Data.

The storages orientation for the widespread use in the intelligent analytical processing problems, with their inherent deep insight into the investigated phenomena, requires the simple access to all available types of the information resources, the possibility of using them in the different way, and the delivery of the required data to the application resources without their movement.

Agreeing on the concept of the logical storage in general, we imagine the LDW storage as a loosely coupled, providing the necessary flexibility, multi-layered architecture, which includes new technologies and the progressive means of identifying, extracting and preparing a wide range of data. In this case, the analysis, taking into account the logic of the information processing, provides a certain information resource to the application as a service. This will ensure the simple availability and relevance of the information contained in the Big Data.

Conclusion. Big Data is a complex inter-branch scientific and technical problem. In the future, the Big Data processing technologies, with the inclusion into its contours the methods and means of intellectualization, will be widely used in the economic, industrial and technological spheres and other important fields of human activity.

●

Для соответствия научно-метрическим базам при подаче статей к рассмотрению, авторы должны подать метаданные на английском языке:

– ФИО

– место и адрес работы каждого автора

– расширенную аннотацию (до 2000 знаков с пробелами и рубриками: *Introduction*, *Purpose*, *Methods*, *Results*, *Conclusion*)

– список пристатейной литературы в переводе или транслитерации.

При оформлении списков литературы к расширенной аннотации на английском языке можно пользоваться сайтом

***http*://*translit.net* для русских ссылок**

***http*://*ukrlit.org*/*transliteratsiia* для украинских.**