Стандарты в области больших данных

Д.Е. Намиот, В.П. Куприяновский, Д.Е. Николаев, Зубарева Е.В.

***Аннотация*—В этой статье мы хотели бы остановиться на вопросах стандартизации технологии больших данных. Термин (понятие) большие данные используется все больше и больше в связке со многими областями человеческой деятельности, которые часто стандартизованы (или стандартизуются) как на национальном, так и на международном уровнях. Естественно, что в этой связи возникает вопрос и о стандартизации этих самых больших данных. И такого рода работы должны, естественно, начинаться с фиксации того, а что же это такое, и что именно является предметом стандартизации. В работе приводится обзор работ по стандартизации в этой области в мире, оцениваются усилия по стандартизации больших данных на национальном уровне, а также предлагаются направления для развития.**

***Ключевые слова*—большие данные, международные стандарты, национальные стандарты.**

I. Введение

Большие Данные – это и вызов и возможности для бизнеса. К этому направлению приковано большое внимание, об использовании больших данных говорят уже практически во всех сферах бизнеса. К вызовам относятся сложность такого рода систем, постоянный рост объемов данных и появление новых источников, связанных с объектами физического мира. В последнее время к этому добавляются требования к обработке в реальном времени и соответствующей оптимизации архитектур, схем хранения и т.п.

Применение технологии больших данных в самых разных областях, многие из которых стандартизованы или активно стандартизуются, порождает, естественно, вопросы о введении стандартов в области больших данных.

Важно отметить, что с точки зрения IT технологий, стандарты – вещь достаточно обычная. Стандартами ISO (Международной организации по стандартизации) являются, например, такие практические вещи как язык SQL или C++. Стандарты, вообще говоря, не прямого имеют отношения к каким-либо правовым ограничениям (регуляциям), что часто происходит именно в отечественной практике, а должны рассматриваться как то, что в англоязычной литературе называется “best practice”. Вот упомянутые выше языки программирования и есть, между прочим, хороший пример такой интерпретации. Лучшая практика по организации доступа к данным – язык SQL.

Итак, стандарты – это лучшая практика, которая имеет какое-то экономическое выражение (экономическую оценку). Например, открытые стандарты в области Умного Города и Интернета Вещей ускоряют рост на 27% и сокращают стоимость разработок на 30% [1]. Согласно данным Британского института Стандартов (BSI), направление “большие данные” должно принести британской экономике £216 миллиардов к 2017 году и привести к созданию 58 000 новых рабочих мест. И в достижении этого, а также в обеспечении предсказуемости дальнейшего развития, ключевую роль должны сыграть именно стандарты [2].

Здесь необходимо сделать важное замечание относительно того, что стандартизовать в данной области. Формально, раз в названии присутствует слово “данные”, мы должны вести речь о хранении и организации доступа к данным. Которые, в соответствии с классическим определением больших данных (3V-Volume, Velocity and Variety), имеют такую специфику в виде размера, скорости поступления и неоднородности представления. Проблема, однако, в том, что термин big data все больше и больше используется для описания обработки данных (которые часто вовсе и не такие уж большие). Но методы обработки (анализа) имеют гораздо большую вариабельность. Здесь трудно даже представить универсальные “лучшие практики”, кроме совсем уже простых случаев. Поэтому стандарты касаются, в первую очередь, именно собственно данных. В классификации профессий, стандарты ориентированы на data engineering, а не на data science. Хотя, конечно, любые модели данных по определению вводят и класс операций над данными. Но это будут именно операции, а не методы анализа.

В настоящий момент, несколько основных институтов стандартизации вовлечены в работу по стандартам для больших данных. Основные игроки здесь – Международная организация по стандартизации и Международная Электротехническая комиссия (ISO/IEC), Международный Союз Электросвязи (ITU), Британский Институт Стандартов (BSI), Национальный Институт Стандартов и Технологии США (NIST).

В этой статье мы представляем обзор текущего состояния работ в области стандартизации больших данных. Разработка (локализация) стандартов является одной из задач Национальной Технологической Инициативы в России. Пилотные области здесь – большие данные, Умные Города и Интернет Вещей. Авторы статьи являются членами соответствующих

Статья получена 15 сентября 2016.

Намиот Д.Е., МГУ имени М.В. Ломоносова, (email: dnamiot@gmail.com)

Куприяновский В.П., МГУ имени М.В. Ломоносова, (email: vpkupriyanovsky@gmail.com)

Николаев Д.Е., МГТУ имени Н.Э. Баумана (email: d.nikolaev@bmstu.ru)

Зубарева Е.В. — МГУ имени М.В. Ломоносова; ЕГУ имени И.А. Бунина (e-mail: e.zubareva@cs.msu.ru)

рабочих групп по стандартизации (ТК098/РГ1 «Интернет вещей», Т098/РГ2 «Разумный город» и ТК098/РГ3 «Большие данные»), и мы опишем также текущее состояние разработки национальных стандартов в области технологии больших данных.

II. Стандарты больших данных в ISO/IEC

Международная организация по стандартизации и Международная электротехническая комиссия (ИСО / МЭК) создали 3 рабочие группы, ориентированные на стандартизацию следующих технологий: большие данных (ISO/IEC JTC1/WG 9 «Big data»), интернет вещей (ISO/IEC JTC1/WG 10 «Internet of things») и умные города (ISO/IEC JTC1/WG 11 «Smart Cities»). На самом деле, это довольно общий тренд, увязывать большие данные и Интернет Вещей, так как измерения в IoT есть большие данные де-факто.

В соответствии со стандартом ИСО, Рабочая группа по большим данным будет служить в качестве определяющей для главной темы большой программы стандартизации данных и выявления пробелов в области стандартизации. Она будет разрабатывать основополагающие стандарты - в том числе эталонной архитектуры и словарный запас [3]. В настоящее время международная рабочая группа по стандартизации ISO/IEC JTC1/WG 9 «Big data» разрабатывает следующие проекты международных стандартов: комплекс стандартов на эталонную архитектуру больших данных (ISO/IEC 20547) и стандарт на термины и определения (ISO/IEC 20546). Руководитель этой рабочей группы также является цифровым консультантом данных для Национального института стандартов и технологий (NIST), поэтому мы можем ожидать, что в итоговых документах мы увидим многие идеи из NIST здесь.

На момент написания статьи (август 2016 г.) рабочая группа выпустила обзор больших данных и словарь [4]. Он был представлен как RFC – документ (запрос на комментарии). Это относительно небольшой документ. Он содержит основные определения модели данных (например, модели 4V для больших объемов данных - объем, скорость, разнообразие и изменчивость), а также небольшой набор терминов (например, кластерные вычисления, параллельные вычисления и т.д.). На самом деле, этот документ был переведен на русский язык и разослан на согласование по членам ТК098/РГ3 “Большие данные”. В то же время, необходимо прямо еще раз отметить, что это очень небольшой документ, и его явно недостаточно для практической работы (понимаемой именно как представление "наилучшей практики"). Точной информации о дорожной карте от ISO/IEC в этом вопросе у нас нет и, по-видимому, мы должны смотреть на то, что делает NIST в этом направлении. По нашему мнению, именно результаты NIST окажутся в итоге в ISO/IEC. В разделе, который посвящен работам NIST, приводится его (NIST) собственная дорожная карта, составленная с помощью ISO.

III. Стандарты больших данных в ITU

В ITU можно отметить несколько областей активности, касающихся больших данных [5]. В документах ITU указываются следующие области активности:

- высоконадежная, гибкая и масштабируемая сетевая инфраструктура с высокой пропускной способностью и с низкой задержкой. Эта область (с точки зрения МСЭ) имеет много пересечений с разделом облачных вычислений. И именно в этой области, МСЭ представил свой документ как "первый стандарт Big Data".

- Агрегирование и анонимизация наборов данных. Это направление очень важно для таких областей, как Умные Города (концепция города, управляемого данными или город – как сенсор), Интернет Вещей и электронные медицинские приложения [6]. Именно эти работы могли бы послужить базов для отечественных работ в области стандартизации, тяготеющих к правовому регулированию.

- Совместимые платформы. МСЭ планирует целевые вертикальные рынки (домашней автоматизации, электронное здравоохранение) со стандартами совместимости данных. Это должно представлять интерес для отечественных работ в области телемедицины.

- Мультимедиа-аналитика. Мультимедийный контентный анализ IP-трафика, который позволит автоматически и быстро расшифровывать (выделять) события и другие мета-данные для видео-файлов и видео-потоков [7].

- Стандарты для открытых данных. В соответствии с МСЭ, работа по стандартизации должны разработать требования к представлению данных и механизмов для публикации, распространения и открытия наборов данных. Это направление имеет большое пересечение с тематикой Умных Городов. Здесь сразу можно указать на стандарт BSI PAS-212:2016 для обнаружения (раскрытия) данных в умных городах [8]. Опять таки, продвигаемый BSI как "первый стандарт Smart City". Теперь он передается ISO для будущего развития. PAS-212 переведен на русский язык и его локализация в России обсуждается. Стандарты открытых данных, разрабатываемые ITU также должны включать элементы раскрытия данных. Это один из основных моментов – иначе как находить данные? Без кооперации с ISO мы рискуем увидеть несколько несовместимых стандартов.

Как видно из приведенного списка, проводимые работы, на первый взгляд, далеки от того, что принято в России назвать большими данными. Поэтому первая задача для локальных стандартизаторов – это определиться с тем, что они будут рассматривать.

В конце 2015 года, члены МСЭ договорились о международном стандарте для больших данных. Новый стандарт, рекомендация МСЭ-Т Y.3600 "Большие данные - требования и возможности на основе облачных вычислений" [9], был разработан группой экспертов МСЭ-Т, ответственных за будущее сетей, облачных

вычислений и сетевых аспектов мобильной связи. Стандарт описывает, как облачные вычислительные системы могут быть использованы для предоставления услуг Big Data. Главным образом, он описывает требования к облачным вычислениям на основе больших объемов данных (требования по сбору данных, предварительной обработки данных и требования к хранению данных, анализу, визуализации и управлению, безопасности данных и требования по защите, сбору и хранению данных). Кроме того, он содержит определения сервиса больших данных как услуги (BDaaS) - категории облачных сервисов, в которой возможности, предоставляемые клиенту облачных услуг, позволяют собирать, хранить, анализировать, визуализировать и управлять данными с использованием технологий больших объемов данных.

На наш взгляд эта рекомендация заслуживает самого пристального внимания в России, поскольку предоставление “хостинга” больших данных – очень важная задача, хотя бы для учебных заведений, где построение учебных стендов собственными силами будет весьма затруднено.

IV. Стандарты больших данных в NIST

На наш взгляд, NIST предлагает наиболее проработанный стек стандартов по большим данным [10]. Этот так называемый NIST Big Data Interoperability Framework V1.0 включает в себя следующие документы:

- Big Data Definitions
- Big Data Taxonomies
- Big Data Use Cases and Requirements
- Big Data Security and Privacy
- Big Data Architecture White Paper Survey
- Big Data Reference Architecture
- Big Data Standards Roadmap

Определения более ориентированы на практику, чем у ISO и, фактически, посвящены моделям, а не словарным статьям. Документ по таксономии предлагает базовую архитектурную модель (рисунок 1).

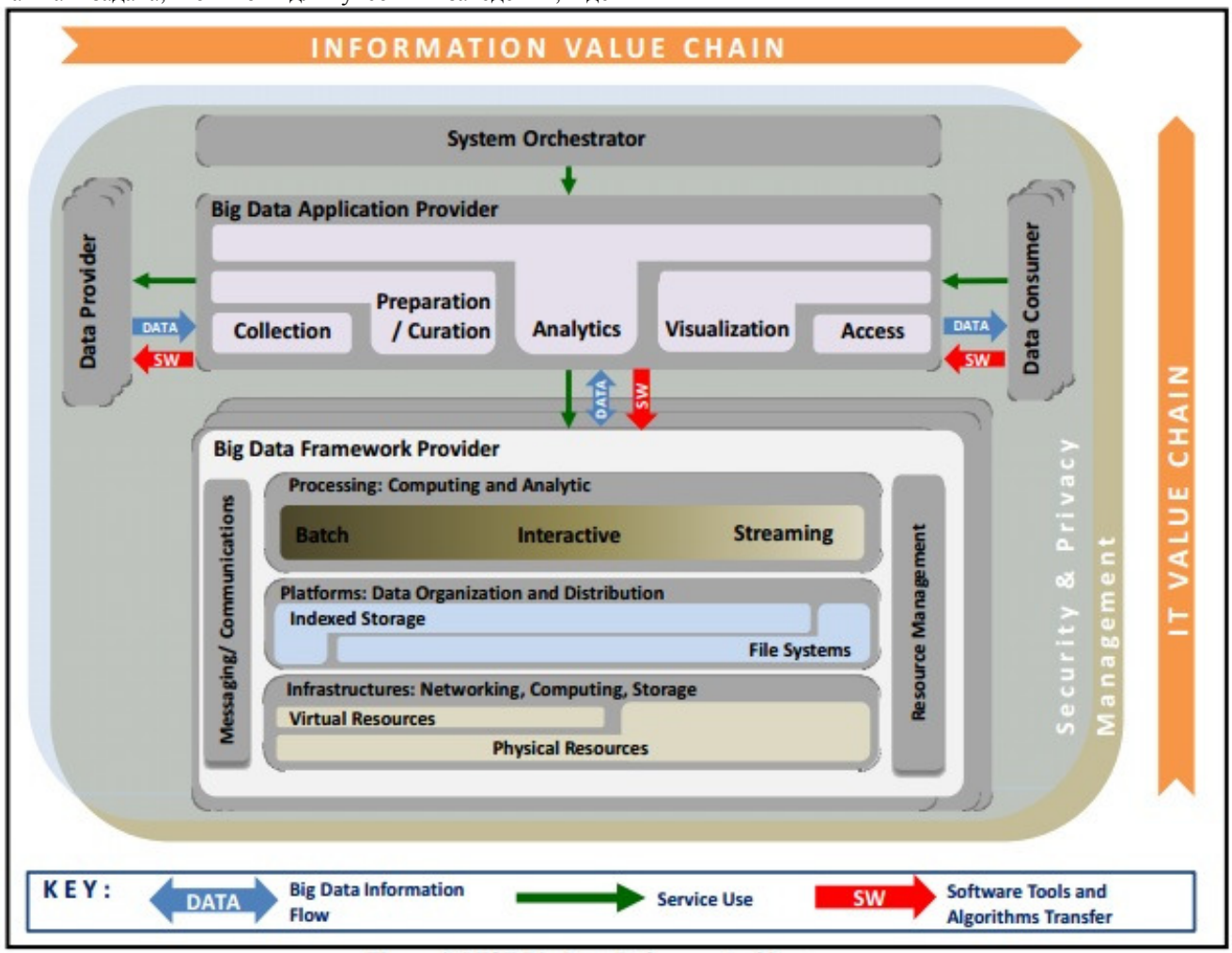

Рис.1 Ссылочная модель Big Data от NIST [11].

Документ с примерами содержит описание реальных применений (описание архитектур данных) из 9 областей: правительственные системы, коммерческие системы, военные применения, здравоохранение, социальные медиа, экология, астрономия и др. Этот раздел может вполне использоваться как учебник (или часть учебного материала).

Документ, касающийся безопасности и приватности также содержит практические примеры из таких областей как здравоохранение, телеметрия, маркетинг и др. Из технологий безопасности рассматриваются вопросы идентичности, авторизация, аудит, безопасность в сети, безопасность устройств.

White Papers Survey рассматривает архитектуры данных и является одним из наиболее интересных элементов в списке NIST. В этом разделе содержатся

предложения и справочные модели (для архитектур данных) от таких компаний, как IBM, Oracle и исследовательских групп в университетах.

Big Data Reference Architecture представляет собой концептуальную модель высокого уровня, разработанную для того, чтобы служить в качестве инструмента для содействия открытому обсуждению требований, проектных структур и операций, присущих большим данных. Она не представляет системную архитектуру конкретной системы больших данных, а скорее является инструментом для описания, обсуждения и разработки архитектуры конкретной системы, используя общую точку (базу) отсчета. Эта модель не привязана к какой-либо конкретной продукции поставщиков [12].

Последний документ - Дорожная карта описывает пересечения между NIST Framework и существующими стандартами (например, SQL). Но самая интересная часть этой дорожной карты - это список направлений развития, который был построен с помощью Объединенного технического комитета 1 (ОТК1) ИСО/МЭК. Исследовательская группа по большим данным определила направления работ, которые могли бы служить в качестве потенциального руководства для ISO в их создании стандартов деятельности больших данных. Исследовательской группой были определены потенциальные пробелы в стандартизации темы больших данных. Эти результаты описывают достаточно широкие области, которые могут представлять интерес для разработчиков и исследователей. Фактически – это “горячие” темы для исследований в области больших данных:

- Модели применения и ссылочные архитектуры. Платформы для больших данных.
- Технические требования и стандартизация для метаданных.
- Модели для приложений (например, пакетная обработка и анализ потоков)
- Языки запросов, включая нереляционные запросы для поддержки различных типов данных (например, XML, RDF, JSON, мультимедиа) и больших объемов данных операций (например, матричных операций).
- Проблемно-ориентированные языки для задач больших данных.
- Семантика для слабой согласованности и согласованности в конечном счете (eventual consistency).
- Расширенные сетевые протоколы для эффективной передачи данных.
- Общие и предметно-ориентированные онтологии и таксономии для описания семантики данных, включая интероперабельность между онтологиями.
- Безопасность и контроль доступа.
- Удаленные, распределенные и федеративные аналитики данных, включая методы обнаружения и обработки ресурсов, а также методы интеллектуального анализа данных.
- Обмен и разделение данных.
- Системы хранения данных (например, in-memory базы данных, распределенные файловые системы, хранилища данных).
- Представление результатов анализа данных (сюда, очевидно, должны входить визуализация и объяснения).
- Энергетическая эффективность работы с большими данными.
- Интерфейс между реляционными (т.е. SQL) и нереляционными (т.е. NoSQL) хранилищами данных
- Оценка качества и достоверности данных.

V Стандарты больших данных в BSI

BSI и его работы в области стандартизации интересны, прежде всего, тесной связью с экономикой.

Таким образом, решение BSI работать в стандартах больших данных также базируется на экономике и экономических данных. В соответствии с BSI, Big Data представляет собой весьма существенную и быстрорастущий возможность на рынке сегодня. Это принесет пользу экономике Великобритании на £ 216 млрд. к 2017 году и привести к созданию 58000 новых рабочих мест [12]. На рисунке 2 показаны проблемы адаптации больших данных согласно BSI.

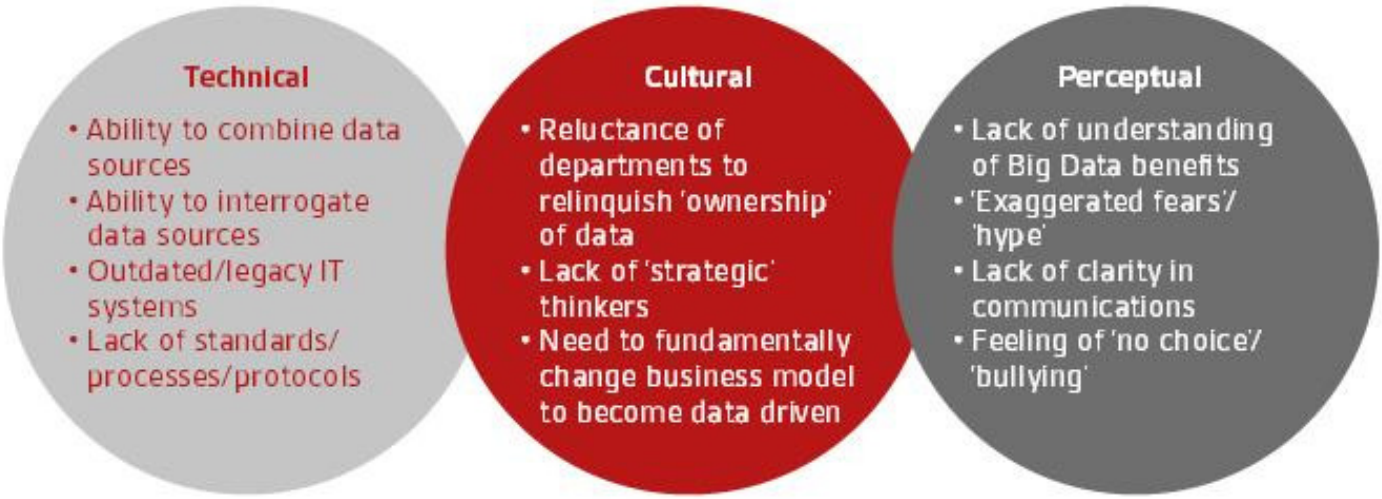

Рис. 2 Проблемы адаптации больших данных [13]

Вышеперечисленные районы определяют направления стандартов. В соответствии с BSI,

следующие темы подлежат стандартизации в больших данных:

- Стандарт на метаданные. В мире больших данных, метаданные являются основой для аналитических отчетов. Множество аналитических отчетов базируются именно на метаданных. BSI предполагает определить наилучшую практику для сбора и хранения метаданных, в том числе, обеспечить качество и полноту собранной информации, а также регламентные процедуры (принципы) для долговременного хранения метаданных (например, как долго они должны храниться). В связи с этим, следует отметить еще раз стандарт для раскрытия (описания) данных в умных городах [8]. Это типичный пример стандарта, связанного с метаданными.
- Стандарты на условия работы с данными. Это также полностью укладывается в идеологию построения «лучших практик». Построение общественного доверия имеет жизненно важное значение. Стандартизуя описание условий работы с данными, BSI собирается обеспечить наилучшую практику для обеспечения того, чтобы условия были просты для понимания и обеспечивали информированное общественное согласие.
- Стандарты сбора данных. Этот стандарт должен быть нацелен именно на процесс сбора данных. Стандартизоваться должно представление того, как (с помощью чего) потребитель принимает решение о том, кто может использовать его данные.
- Стандарты объяснения для проектов Big Data. Многие из инициатив в области больших данных не удается провести (реализовать) из-за общественного сопротивления. Этот стандарт должен определить наилучшую практику для того, как инициативы больших объемов данных должны быть объяснены. Он должен помочь предоставлять инициативы потребителям в ясной и однозначной форме.
- Руководство "Как сделать" для Big Data. На самом деле, это ближайший аналог Примеров использования из NIST. Согласно BSI, этот стандарт должен помочь предприятиям стартовать с проектами Big Data.

Отметим, что аналогичные проекты на русском языке авторам неизвестны.

VI Другие стандарты для больших данных

Из других институтов, которые имеют инициативы, относящиеся к большим данным, отметим:

- Институт Инженеров Электротехники и Электроники (IEEE)
- Международную Электротехническую Комиссию (IEC)
- Инженерный Совет Интернета (The Internet Engineering Task Force - IETF)
- Консорциум Всемирной Паутины (World Wide Web Consortium -W3C)
- Открытый гео-консорциум (Open Geospatial Consortium - OGC)
- Глобальный консорциум промышленных стандартов (Organization for the Advancement of Structured Information Standards - OASIS)
- Пользовательская группа по облачным стандартам (The Cloud Standards Customer Council)

Направления IEEE Будущие Инициативы Big Data стремится объединить информацию о различных начинаниях, происходящих во всем мире, с тем, чтобы поддержать сообщество профессионалов в промышленности, научных кругах и правительствах, работающих над решением проблем, связанных с большими данными [15].

Международная электротехническая комиссия (МЭК) работает с ISO по стандартам больших данных.

IETF разрабатывает и продвигает добровольные стандарты Интернет. Есть несколько проектов, связанных со стандартами больших данных. Например, это сеть телеметрии и анализа данных [16]. Их развитие фокусируется на измерении сети (трафика) и анализа в сетевой среде. Он определяет сетевую телеметрию, описывает архитектуру телеметрической сети, а затем исследует характеристики данных сети.

Работы W3C относятся к семантическим данным [17].

Проект Big Data Europe [18] осуществляется в рамках программы Horizon 2020. Он будет предоставлять решения в следующих областях: гетерогенные связи и интеграция данных, биомедицинская семантическая индексация, крупномасштабная распределенная интеграция данных, мониторинг в режиме реального времени, потоковая обработка и анализ данных, поддержка систем принятия решений, потоковые сети сенсорных данных, гео-пространственная интеграция данных, мониторинг в режиме реального времени, анализ изображений.

Открытый Гео-консорциум создал рабочие группы [19] с особым акцентом на пространственно-временные данные.

OASIS также создал несколько комиссий, связанных с большими данными. Самая интересная, на наш взгляд, работа – это база данных OASIS Key-Value Application Interface (KVDB) TC [20]. В ней определяется открытый программный интерфейс для управления и доступа к данным из систем баз данных на основе модели ключ-значение. Это одна из широко используемых моделей данных в NoSQL.

Пользовательская группа по облачным стандартам публикует описания лучших технических практик для больших данных в облачной среде [21].

Фактически, на момент написания статьи, среди всех перечисленных “лучших практик” национальная (российская) группа по стандартизации больших данных оперирует только со словарем ISO/IEC и некоторыми проектами BSI.

Библиография

[1] Намиот Д.Е., Шнепс-Шнеппе М.А. Об отечественных стандартах для Умного Города // International Journal of Open Information Technologies. 2016. -Т. 4. - №7. С.32-37.
[2] BSI Big Data and standards market research, January 2016

[3] ISO/IEC JTC 1 Forms Two Working Groups on Big Data and Internet of Things https://www.ansi.org/news_publications/news_story.aspx?menuid=7&articleid=5b101d27-47b5-4540-bca3-657314402591 Retrieved: Aug, 2016
[4] ITU-T LIAISON STATEMENT ISO/IEC JTC1/WG9 - ISO/IEC JTC 1/WG 9 N 201 http://www.itu.int/net/itu-t/ls/ls.aspx?isn=12493 Retrieved: Sep, 2016
[5] ITU Big Data http://www.itu.int/en/ITU-T/techwatch/Pages/big-data-standards.aspx Retrieved: Sep, 2016.
[6] Tene, Omer, and Jules Polonetsky. "Big data for all: Privacy and user control in the age of analytics." Nw. J. Tech. & Intell. Prop. 11 (2012): xxvii
[7] Smith, John R. "Riding the multimedia big data wave." Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013.
[8] Kupriyanovsky, Vasily, et al. "On Localization of British Standards for Smart Cities." International Journal of Open Information Technologies 4.7 (2016): 13-21.
[9] ITU Y.3600 http://www.itu.int/itu-t/recommendations/rec.aspx?rec=12584 Retrieved: Aug, 2016
[10] NIST Big Data http://www.nist.gov/itl/bigdata/bigdatainfo.cfm Retrieved: Aug, 2016
[11] Big Data Taxonomies http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-2.pdf Retrieved: Sep, 2016
[12] NBDRA http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6.pdf Retrieved: Aug, 2016
[13] Big Data Standards Workshop http://www.bsigroup.com/en-GB/our-services/events/2016/Big-Data-Standards-Workshop/ Retrieved: Aug, 2016
[14] Big Data Standards and Market Research http://shop.bsigroup.com/upload/275237/The-Big-Data-And-Standards-Market-Research-Report-By-BSI-And-Circle-Research.pdf Retrieved: Aug, 2016
[15] IEEE Big Data http://bigdata.ieee.org/ Retrieved: Aug, 2016
[16] Network Telemetry and Big Data Analysis https://datatracker.ietf.org/doc/draft-wu-t2trg-network-telemetry/ Retrieved: Aug, 2016
[17] Koivunen, Marja-Riitta, and Eric Miller. "W3c semantic web activity." Semantic Web Kick-Off in Finland (2001): 27-44.
[18] Big Data Europe https://www.big-data-europe.eu Retrieved: Aug, 2016
[19] Big Data Working Group http://www.opengeospatial.org/projects/groups/bigdatadwg Retrieved: Aug, 2016
[20] OASIS Key-Value Database Application Interface (KVDB) TC https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=kvdb Retrieved: Aug, 2016
[21] The Cloud Standards Customer Council http://www.cloud-council.org/resource-hub.htm Retrieved: Aug, 2016.

On standards in Big Data area

Dmitry Namiot, Vasily Kupriyanovsky, Danila Nikolaev, Elena Zubareva

Abstract— **In this article, we would like to discuss the issues of standardization of so-called big data. This term (concept) is used more and more in conjunction with many areas of human activity, which are often standardized (or being in the process of standardization) at both the national and international levels. Naturally, this raises the question of standardization of big data too. And this kind of work should, of course, begin with the fixation of, and what it is, and that it is subject to standardization. The paper provides an overview of standardization work in this field across the world, estimated efforts to standardize big data at national level and suggests areas for the development.**

Keywords— **big data, international standards, domestic standards.**