

чи [Электронный ресурс]. – Режим доступа: http://pitbit.ru/news/programs/yandex_speechkit_novoe_prilozhenie_dlya_raspoznavaniya_russkoy_rechi/.

Сергеева Инна Ивановна

*к.э.н., доцент кафедры информатики и информационных технологий
ФГБОУ ВПО «Орловский государственный институт экономики и торговли»
e-mail: inchiksergeeva@yandex.ru*

Ляхова Ольга Валерьевна

*студентка 3 курса факультета учета и информационных технологий
ФГБОУ ВПО «Орловский государственный институт экономики и торговли»
e-mail: inchiksergeeva@yandex.ru*

Раздолянский Сергей Валерьевич

*студент 3 курса факультета учета и информационных технологий
ФГБОУ ВПО «Орловский государственный институт экономики и торговли»
e-mail: grafcorp01@gmail.com*

УДК 004.81

И.И. Сергеева, О.С. Шапиро

ОСОБЕННОСТИ ИСПОЛЬЗОВАНИЯ ТЕХНОЛОГИИ BIG DATA

В статье дано понятие «больших данных», представлен краткий обзор определенных сложностей, связанных с Big Data, а именно: хранилищами данных объемом в терабайт, петабайт, технологиями и подходами для преодоления сложностей получения значимой информации из Big Data.

Ключевые слова: большие данные, хранилища данных, информационные массивы.

Современное развитие общества характеризуется огромным ростом информационных объемов, циркулирующих в обществе, и проблемами их хранения, транспортировки, защиты. Эти данные могут быть весьма разнообразными: жестко структурированными или неструктурированными вообще, часто обновляемыми и хранящимися в разнообразных источниках. Для определения такого вида информации был введен новый термин «большие данные». Однако его значение гораздо шире, чем просто управление и анализ больших объемов данных. Согласно отчету McKinsey Institute, «Большие данные» выступают как основа для инноваций, ужесточения конкуренции и роста производительности, это не просто стандартные наборы данных, размер которых превосходит возможности СУБД по вводу, хранению, управлению, обработке и анализу инфор-

мации, причем, в основном, они плохо структурированы и не вписываются в форматы типичных БД. Проблемы «Больших данных» ещё и в том, что они могут храниться распределено во множестве репозиториях, чаще всего за пределами организации, где создаются и используются, при этом важно не только обеспечить доступ к огромному объему информационных массивов, но и иметь в наличии весь инструментарий, необходимый для анализа, выявления взаимосвязей, обеспечения возможности получить на этой основе значимые выводы. Консалтинговая компания Forrester дает краткую формулировку: «Большие данные» объединяют техники и технологии, которые извлекают смысл из данных на экстремальном пределе практичности».

Источников «Больших данных» в современном мире великое множество. В их качестве могут выступать непрерывно поступающие данные с измерительных устройств, события от радиочастотных идентификаторов, потоки сообщений из социальных сетей, метеорологические данные, данные дистанционного зондирования земли, потоки данных о местонахождении абонентов сетей сотовой связи, устройств аудио- и видеорегистрации. Собственно, массовое распространение перечисленных выше технологий и принципиально новых моделей использования различно рода устройств и интернет-сервисов послужило отправной точкой для проникновения «Больших данных» едва ли не во все сферы деятельности человека. В первую очередь, научно-исследовательскую деятельность, коммерческий сектор и государственное управление.

Примеров массивов «Больших данных» множество и с каждым часом становится все больше. В 2010 году корпорации мира накопили 7 экзабайтов данных на наших домашних ПК, на ноутбуках хранится 6 экзабайтов информации. В 2010 году в сетях операторов мобильной связи обслуживалось 5 млрд. телефонов. Каждый месяц в сети Facebook выкладывается в открытый доступ 30 млрд. новых источников информации. Ежегодно объемы хранимой информации вырастают на 40%, в то время как глобальные затраты на ИТ растут всего на 5%. По состоянию на апрель 2011 года в библиотеке Конгресса США хранилось 235 терабайт данных. Американские компании в 15 из 17 отраслей экономики располагают большими объемами данных, чем библиотека Конгресса США.

Компании собирают и используют данные самых разных типов как структурированные, так и неструктурированные. Так, по итогам опроса (Cisco Connected World Technology Report), 74% респондентов получают данные из текущей информации, 55% собирают исторические данные, 48% снимают данные с мониторов и датчиков, а 40% пользуются данными в реальном времени, а затем стирают их. Чаще всего данные в реальном времени используются в Индии (62%), США (60%) и Аргентине (58%). В мире сбором неструктурированных данных занимаются 32% опрошенных, а в Китае этот показатель достигает более половины опрошенных (56%).

Основные проблемы, связанные с Big Data, сконцентрированы в 4 областях. Во-первых, это хранение, поскольку объем данных в сотни терабайт или петабайт не позволяет легко хранить и управлять ими с помощью традиционных реляционных баз данных. Во-вторых, большинство из них являются не-

структурированными. Важной характеристикой является скорость передачи в хранилища, возврат для обработки и манипулирования. И наконец, самая важная проблема – анализ Big Data. Все современные средства обработки основаны на репрезентативных выборках или агрегированных подмножествах информации, что приводит к определенному рода ошибкам и неточностям. Однако возможность анализировать всю генеральную совокупность позволит этих ошибок избежать. Таким образом, Big Data требует наличия иного программного обеспечения, аппаратных средств, алгоритмов, методов обработки.

20 марта 2012 года IBM представила три новых решения Smarter Analytics Signature Solutions для выявления фактов мошенничества, оценки рисков и анализа поведения потребителей. По оценкам IDC, компании потратят свыше \$120 млрд. до 2015 года на программно-аппаратные решения и сопутствующие сервисы, направленные на выявление скрытых закономерностей в «больших данных».

– Решение Anti-fraud, Waste & Abuse призвано обнаруживать в режиме реального времени факты мошенничества, связанные с уклонением от уплаты налогов и страховыми выплатами, тем самым предотвращая неправомерные выплаты денежных средств. Кроме этого, соответствующие службы страховых компаний и государственных служб получают оптимальные рекомендации по дальнейшему разрешению возникающих инцидентов. При незначительных нарушениях они могут ограничиваться отправкой письма с требованием вернуть платеж, при более серьезных – поступает предложение провести полноценное расследование.

– Решение Next Best Action должно помочь компаниям лучше узнать своих клиентов и выстроить с ними более содержательные взаимовыгодные отношения. Оно сможет проанализировать не только данные, накопленные в корпоративной CRM-системе, но и информацию из внешних источников, в частности, социальных сетей. Как следует из названия, результатом работы станет свод рекомендаций по дальнейшим действиям в зависимости от предпочтений и модели поведения отдельно взятого клиента.

– Предпосылкой для создания решения CFO Performance Insight стали чрезвычайно высокие темпы роста объемов финансовой информации в компаниях, – по данным IBM, на уровне 70% в год. Новый аналитический инструментарий позволяет учитывать финансовые показатели предыдущих лет, проводить моделирование по принципу «что если» и на основе полученных данных строить прогнозы.

Представленные решения включают разнообразные продукты IBM и услуги подразделения Global Business Services (GBS).

После нескольких лет экспериментов с технологиями Big data и первых внедрений в 2013 году адаптация подобных решений значительно возрастет, прогнозируют в Gartner. Исследователи опросили ИТ-лидеров во всем мире и установили, что 42% опрошенных уже инвестировали в технологии Big data или планируют совершить такие инвестиции в течение ближайшего года (данные на март 2013 года).

Компании вынуждены потратиться на технологии обработки больших

данных, поскольку информационный ландшафт стремительно меняется, требуя новых подходов к обработке информации. Многие компании уже осознали, что большие массивы данных являются критически важными, причем работа с ними позволяет достичь выгод, не доступных при использовании традиционных источников информации и способов ее обработки. Кроме того, постоянное муссирование темы «Больших данных» в СМИ подогревает интерес к соответствующим технологиям.

Когда мы говорим о big data, мы понимаем, что это связано со следующими аспектами: большим объемом информации, ее разнообразием или необходимостью обрабатывать данные очень быстро. С другой стороны, под этим термином часто понимают совершенно конкретный набор подходов и технологий, призванных решить данные задачи. В основе одного из таких подходов лежит система распределенных вычислений, где обработка больших объемов данных требует для себя не одну высокопроизводительную машину, а целую группу таких машин, объединенных в кластер.

Компания DIS Group развивает направление по внедрению технологий класса big data. Специалисты компании прорабатывают различные решения в данной области, включая детальную проработку технических аспектов применения big data-технологий и позиционирования их в инфраструктуре заказчиков:

Увеличивающееся количество гигантских массивов корпоративных данных и все технологии для их создания, хранения, перемещения, анализа, архивирования и извлечения информации – всё это «Большие данные». Огромный объем информации раздвигает границы хранилищ, серверов и периметров безопасности, создавая огромные проблемы для ИТ-отделов, подлежащие решению.

Список литературы:

1. Черняк Л. Большие данные – новая теория и практика // Открытые системы. СУБД. – 2011. – №10. – С. 18-25.
2. Большие данные: как извлечь из них информацию компании / А. Моррисон [и др.] // Технологический прогноз. – 2010. – №3. – С. 42.
3. Preimesberger, Chris Hadoop, Yahoo, 'Big Data' Brighten BI Future. EWeek (15, August, 2011).
4. На сколько большие Большие данные? [Электронный ресурс]. – Режим доступа: <http://www.statsoft.ru/products/Enterprise/big-data.php>.

Сергеева Инна Ивановна

*к.э.н., доцент кафедры информатики и информационных технологий
ФГБОУ ВПО «Орловский государственный институт экономики и торговли»
e-mail: inchiksergeeva@yandex.ru*

Шапиро Ольга Сергеевна.

*студентка 3 курса факультета учета и информационных технологий
ФГБОУ ВПО «Орловский государственный институт экономики и торговли»
e-mail: oshapiro@bk.ru*