СТРАТЕГИИ БИЗНЕСА
анализ | прогноз | управление
электронный научно-экономический журнал
№ 5 (37) 2017

Big data – новый подход формирования бизнес-знаний

Т.С. Скорочкина
старший преподаватель кафедры «Бизнес-информатика»
Финансовый университет при Правительстве Российской Федерации
Г. Москва, Российская Федерация

T.S. Skorochkina
Assistant Professor, Department of Business- informatics
Financial University under the Government of the Russian Federation
Moscow, Russian Federation

Аннотация: Статья посвящена вопросу формирования бизнес-знаний с использованием новых неоднородных источников и специфических технологий, объединённых понятием «большие данные». Были проанализированы различные определения данного термина и характерные особенности новой технологии. Сформулированы основные отличия «больших данных» от традиционных, на основании чего сделаны выводы о значимости этих данных в процессе формирования конкурентного преимущества.
Ключевые слова: Big data, «большие данные», аналитика, источники данных, традиционные данные, интернет-технологии, методы извлечение данных, типы данных, структурированные данные, неструктурированные данные, мультиструктурные данные, обработка данных, анализ данных.

BIG DATA – NEW APPROACH TO BUSINESS KNOWLEDGE CREATION

Abstract: The article is devoted to the topic of creation of business knowledge with usage of new heterogeneous information sources and specific technologies, united under the concept – “big data”. It was analyzed various definitions of the “Big data” term as well as specifics of that particular technology. It was formulated some key differentiations of “Big data” from traditional technologies. Based on that analysis it was made a conclusion regarding the value of this technology for gaining a competitive advantage.
Keywords: Big data, analytics, data sources, traditional data, internet-technologies, methods of data sourcing, data types, structural data, unstructured data, data processing, data analysis.

В современном мире знания являются важнейшим активом компаний. Бизнес-знания в общем смысле – это понимание того, что происходит и как эту ситуацию выгодно использовать с коммерческой точки зрения.

Переход от принятия интуитивных управленческих решений к обоснованным, опирающимся на точные данные, происходит уже достаточно давно. Количество источников данных и разнообразие средств углубленной аналитики, используемых для принятия таких решений, постоянно увеличивается.

По мнению Билла Фрэнкса, директора по аналитике глобальных партнёрских программ компании Teradata, ничто так сильно не повлияет на сферу передовой аналитики в ближайшие годы, как постоянное появление новых и мощных источников данных. Уже сегодня можно с уверенностью заявить о наступлении эпохи революционных подходов в аналитической сфере, связанных с использованием «больших данных»[1].

Сам термин Big data появился относительно давно, но массовый интерес к данному явлению существенно вырос именно в последние несколько лет. В основном это связано с двумя ключевыми факторами: активным развитием и внедрением интернет-технологий как основного источника новых данных и одновременно развитием технологий в части возможности хранения и обработки гигантских массивов информации.

Единого общепринятого определения «больших данных» на сегодняшний день не существует. Такие лидеры ИТ-рынка, как Oracle, Intel, Microsoft, предлагают собственные варианты[2], резюмируя которые можно сформулировать характерные черты данного понятия: Big data – это огромные массивы данных, получаемые как из традиционных реляционных баз данных, так и из новых источников неструктурированных данных: документы, почта, блоги, социальные сети и пр., для обработки которых применяются в том числе машинное самообучение и методы искусственного интеллекта.

В свою очередь известная аналитическая и консалтинговая компания Gartner определила своё понимание «больших данных» через основные свойства, которыми эти данные должны обладать (3 V). И в последнее время всё больше экспертов поддерживают именно данное определение – «Big data – это масштабные (volume), с высокой скоростью

[1] Билл Френкс. Укрощение больших данных. М.: «Манн, Иванов и Фербер», 2014. С.29
[2] The Big Data Conundrum: How to Define It? [Electronic resource] // MIT Technology Review [Official website]. 3.10.2013. URL: http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/ (accessed: 1.03.2017)

передачи (velocity), многообразные (variety) информационные активы, которые требуют рентабельных инновационных технологий обработки для извлечения полезной информации и принятия обоснованных решений»[3].

Ниже представлено несколько значимых цифр, характеризующих масштабы Big Data[4]:

- За месяц 600 миллионов пользователей Facebook добавляют в сеть 30 миллиардов единиц контента;
- Пользователи Twitter выполняют 32 миллиарда поисковых запросов в месяц;
- Компания Zynga, занимающаяся сетевыми виртуальными играми, ежедневно обрабатывает более петабайта игровой информации;
- Пользователи Google в 2011 году выполняли почти 5 миллиардов поисковых запросов в день;
- McKinsey & Company считает, что почти в каждой отрасли американской экономики компании с численностью персонала более 1000 человек накапливают в среднем больший объём информации, чем Библиотека Конгресса США.

Любая активность в сети, хотим мы того или нет, фиксируется и запоминается. Поисковые запросы, просмотры интересующих товаров в интернет-магазине или просто «лайк» интересной ссылки или забавной картинки моментально сохраняются на различных виртуальных серверах. Большие данные появляются везде, и их умелое применение, по мнению экспертов, окажется серьёзным конкурентным преимуществом. Основной вопрос в том, как их умело применить. Для ответа на него необходимо определить ключевые различия между традиционными данными, которые десятилетиями считались важнейшим ресурсом компаний, и новым типом «больших данных».

По сути, принципиальных различий два: метод получения данных и их тип.

Извлечение традиционных данных можно охарактеризовать как классический подход (и до недавнего времени это был единственный способ). Данные извлекались целенаправленно по определённым условиям и «складывались» в жесткие преднастроенные структуры. Этот процесс всегда инициировался и контролировался человеком. Иными словами, мы заранее знали, что мы хотим извлечь, в каком объёме, в каком виде и как мы это будем хранить. В некоторой степени это было обусловлено исторически сложившейся потребностью в эффективном использовании пространства.

С «большими данными» всё в точности наоборот. В подавляющем большинстве случаев такие данные генерируются автоматически без какого-либо участия человека. И здесь, как ни странно, скрыта основная ценность этих данных. Когда мы самостоятельно инициируем сбор определённых данных и знаем, в каком виде и в какой структуре они будут храниться, то мы собственно изначально знаем, для каких целей эти данные будут использоваться и в итоге какую информацию из них можно получить. Big data представляют нам те сведения, которые мы сами специально не стали бы собирать, ввиду их неочевидной ценности. Но, обладая большим массивом, казалось бы, бессмысленных данных и применив специфические алгоритмы обработки и анализа, в результате можно получить очень неожиданные теории и гипотезы. Конечно, не все данные и даже небольшая их часть имеют скрытую ценность и поддаются анализу. «Большие данные» можно сравнить со снежной лавиной, которая несется с огромной скоростью, не переставая увеличиваться. Естественно, что в процессе неконтролируемого роста объёма накапливается огромное количество избыточных фактов и просто информационного мусора. С этим связана ключевая проблема эффективного использования «больших данных» – необходимо отсортировать этот мусор и извлечь ценные и релевантные фрагменты информации. Сложность и масштабность этой проблемы становятся очевидны после определения возможных типов «больших данных».

Традиционные данные хранились (и хранятся) в настроенных структурах, и, естественно, в большинстве своём относятся к структурированному типу. Это существенно упрощало их обработку и анализ.

Тип «больших данных» определяется типом их источника, а, как было сказано выше, источником может выступать как таблица обычной базы данных, так и пользовательские блоги, содержащие информацию на естественном языке. В связи с этим тип у этого рода данных может быть совершенно любой, и для простоты понимания многие специалисты определяют их как неструктурированные, хотя корректнее сказать, что это мультиструктурные данные. Таким образом, возвращаясь к главному вопросу – как умело применить «большие данные», нужно сначала решить проблему их извлечения; не формирования, так как этот процесс проходит по большей части автоматически, а именно осознанного извлечения, очистки от «шелухи», что, по сути, можно отнести уже к предварительному этапу анализа.

Кроме того, из мультиструктурности big data следует очень важный вывод – не стоит рассматривать «большие данные» сами по себе. Изолированно они не несут особой ценности в силу характера источников. Ценность и собственно конкурентное преимущество возникают в случае, когда результат обработки «больших данных» объединяется с информацией и знаниями, полученными из традиционных источников компании. Источники «больших данных» качественно дополняют традиционные источники, но ни в коем случае не заменяют их[5].

В связи с тем что всё больше компаний осознают необходимость работы с новыми типами данных, они естественно оказываются перед выбором подходящего инструмента[6]. В настоящее время на ИТ-рынке представлен отдельный класс продуктов и технологий, обеспечивающих решение проблемы извлечения и обработки «больших данных».

По данным IDC, рынок решений для работы с «большими данными» в этом году превысит планку в $16 млрд. Эксперты Allied Market Research обещают, что рынок решений для Hadoop в долгосрочной перспективе подрастет в 25 раз: с $2 млрд в 2013 г. до $50 млрд к 2020 г.[7]

---

[3] Svetlana Sicular. Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s. [Electronic resource] // Forbes [Official website]. 27.03.2013. URL: https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#3552d26e42f6 (accessed: 1.03.2017)

[4] Т. Дэвенпорт, К.Дж. Хо. О чем говорят цифры. М.: «Манн, Иванов и Фербер», 2014. С.12

[5] Билл Френкс. Укрощение больших данных. М.: «Манн, Иванов и Фербер», 2014. С.47

[6] Екатерина Кочеткова. Платформы для Big Data: сравнение вендоров. [Электронный ресурс] // CNews Аналитика [Офиц. сайт]. 02.04.2014. URL http://www.cnews.ru/articles/platformy_dlya_big_data_sravnenie_vendorov (дата обращения: 1.03.2017)

[7] Там же.

Таким образом, чтобы как минимум остаться на плаву, а в перспективе развиваться и укреплять свои позиции, современным компаниям критически важно понять значимость того информационного мусора, который нас окружает. Всё, что раньше никак не связывали с бизнес-данными и не признавали источниками, имеющими какую-то информационную ценность, сегодня, в силу развития технологий, приобретает совершенно другой смысл, и активный рост рынка решений для работы с «большими данными» яркое тому подтверждение. В современном мире невозможно добиться конкурентного превосходства, опираясь на стандартную информацию, в той или иной мере известную всем участникам рынка. Необходимо искать новые теории, получать неявные и на первый взгляд совершенно неочевидные знания, опираясь на все источники данных, которые компания в силах обработать; кроме того, немалую роль играет скорость получения этих знаний. Только те, кто действительно осознает значимость происходящих сегодня перемен и сумеет правильно их использовать в своей деятельности, имеют реальные шансы добиться существенного успеха.

Список литературы:

1. *Билл Френкс.* Укрощение больших данных. М.: «Манн, Иванов и Фербер», 2014. 352 с.
2. *Т. Дэвенпорт, К.Дж. Хо.* О чем говорят цифры. М.: «Манн, Иванов и Фербер», 2014. ыыыыыыыыыыыы 224 с.
3. The Big Data Conundrum: How to Define It? [Electronic resource] // MIT Technology Review [Official website]. 3.10.2013.
4. Svetlana Sicular. Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s. [Electronic resource] // Forbes [Official website]. 27.03.2013.
5. *Екатерина Кочеткова.* Платформы для Big Data: сравнение вендоров. [Электронный ресурс] // CNews Аналитика [Офиц. сайт]. 02.04.2014.