УДК 004.67

Технологии Big Data и их применение на современном промышленном предприятии

© П.Д. Иванов, В.Ж. Вампилова

МГТУ им. Н.Э. Баумана, Москва, 105005, Россия

Проведен обзор технологий Big Data на современном этапе развития, проанализированы перспективы их дальнейшего развития. Обоснованы необходимость использования и перспективность применения технологий Big Data. Осуществлен сравнительный анализ платформы Hadoop с ее аналогами. Приведены результаты исследований применения технологий Big Data. Исследованы современное состояние и тенденции развития технологий Big Data в России и за рубежом.

Ключевые слова: *Big Data, большие данные, прогнозирование, программное обеспечение, платформа Hadoop, конкурентное преимущество.*

Технологии Big Data — серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объемов и значительного многообразия. Данные технологии применяются для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения информации по многочисленным узлам вычислительной сети. Они сформировались в конце 2000-х годов в качестве альтернативы традиционным системам управления базами данных и решениям класса Business Intelligence. В настоящее время большинство крупнейших поставщиков информационных технологий для организаций в своих деловых стратегиях используют понятие «большие данные», а основные аналитики рынка информационных технологий посвящают концепции выделенные исследования.

Термин Big Data относится к наборам данных, размер которых превосходит возможности типичных баз данных по хранению, управлению и анализу информации. В настоящее время множество компаний следят за развитием технологий Big Data. Аналитическая компания IDC представила в декабре 2012 г. отчет «Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East», в котором предсказывалось, что объемы информации будут удваиваться каждые 2 года в течение следующих 8 лет. За ближайшие 7 лет количество данных в мире достигнет 40 ЗБ (1 ЗБ = 10^{21} байт), а это значит, что на каждого жителя Земли будет приходиться по 5200 ГБ данных (рис. 1).

В современных условиях организации создают большое количество неструктурированных данных, таких как текстовые документы, изображения, видеозаписи, машинные коды, таблицы и т. д. Вся эта информация хранится во множестве репозиториев, порой даже за пределами организации. Компании могут иметь доступ к огромному

массиву собственных данных и не иметь необходимых инструментов, которые могли бы установить взаимосвязи между этими данными и сделать на их основе значимые выводы. Традиционные методы анализа информации не могут угнаться за огромными объемами постоянно растущих и обновляемых данных, что в итоге и открывает дорогу технологиям Big Data.

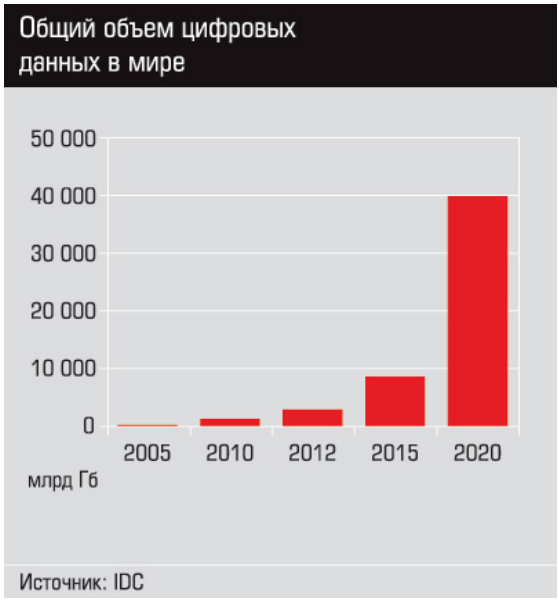

Рис. 1. Общий объем цифровых данных в мире

Можно выделить следующие особенности технологий Big Data [1, 2]:

- работа с информацией огромного объема и разнообразного состава;
- информация весьма часто обновляется и находится в разных источниках;
- качественно отличающийся метод открывающей аналитики для выявления практических знаний, которые непосредственно монетизируются в прибыль;
- наглядное отображение отчетов и возможности сценарного анализа («что, если…»);
- цель применения технологий Big Data — увеличение эффективности работы, создание новых продуктов и повышение конкурентоспособности.

Согласно отчету компании McKinsey «Global Institute, Big data: The next frontier for innovation, competition, and productivity», данные стали важным фактором производства наряду с трудовыми и капитальными ресурсами. Использование больших данных станет основой конкурентного преимущества и роста компаний.

Объем информации на предприятии неуклонно растет за счет данных, полученных с датчиков, измерительных и «умных» устройств. Самыми перспективными устройствами считаются датчики, которые могут передавать данные в режиме реального времени. Все устройства на предприятии с помощью таких датчиков могут быть объединены в сеть, а технологии Big Data позволят обрабатывать информацию, поступающую с них, и проводить необходимые мероприятия в автоматическом режиме. Например, предприятия могут с помощью датчиков получать ежеминутные данные о состоянии своего оборудования и на основе этих данных предсказывать оптимальное время для замены и обслуживания. Слишком ранняя замена приведет к дополнительным расходам, а поздняя — к потере прибыли вследствие простоя оборудования. По оценке компании Cisco, к 2017 г. будет существовать более 1,7 млрд межмашинных соединений [3, 4].

Сферы деятельности, в которых прогнозируется наибольший эффект от применения Big Data, представлены на рис. 2.

Рис. 2. Сферы деятельности с наиболее ощутимым прогнозируемым эффектом от применения больших данных

Технологии Big Data могут быть полезны для решения следующих задач [5]:

- прогнозирование рыночной ситуации;
- маркетинг и оптимизация продаж;
- эффективное сегментирование клиентов;
- совершенствование товаров и услуг;

- принятие более обоснованных управленческих решений на основе анализа Big Data;
- оптимизация портфеля инвестиций;
- повышение производительности труда;
- эффективная логистика;
- мониторинг состояния основных фондов.

Методика и инструменты работы со структурированными данными уже давно созданы. Это реляционная модель данных и системы управления базами данных. Но в современных условиях предприятиям нужно обрабатывать большие объемы неструктурированных данных различных типов (рис. 3), а для этой работы прежние методы не совсем подходят. Нужны новые методики обращения с данными. В настоящее время все более популярной становится модель работы с Big Data, реализованная в проекте Apache Hadoop.

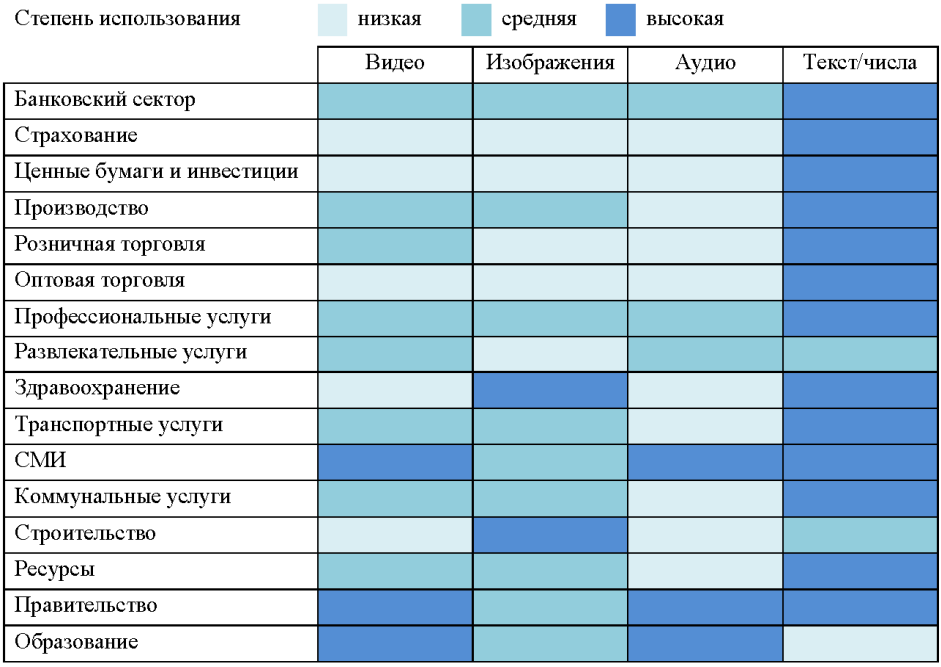

Рис. 3. Превалирующие типы информации для разных сфер деятельности

Большинство продуктов для работы с Big Data обладают высокоэффективной системой обработки огромных объемов информации и ее аналитики в реальном времени.

Ожидаемый эффект от внедрения Big Data может варьироваться в зависимости от типа деятельности и реализуемой политики конкретного предприятия (рис. 4). При работе с большими данными применяют методы манипуляции знаниями: различные методы теории распознавания и классификации, методы разведывательного анализа и обобщения

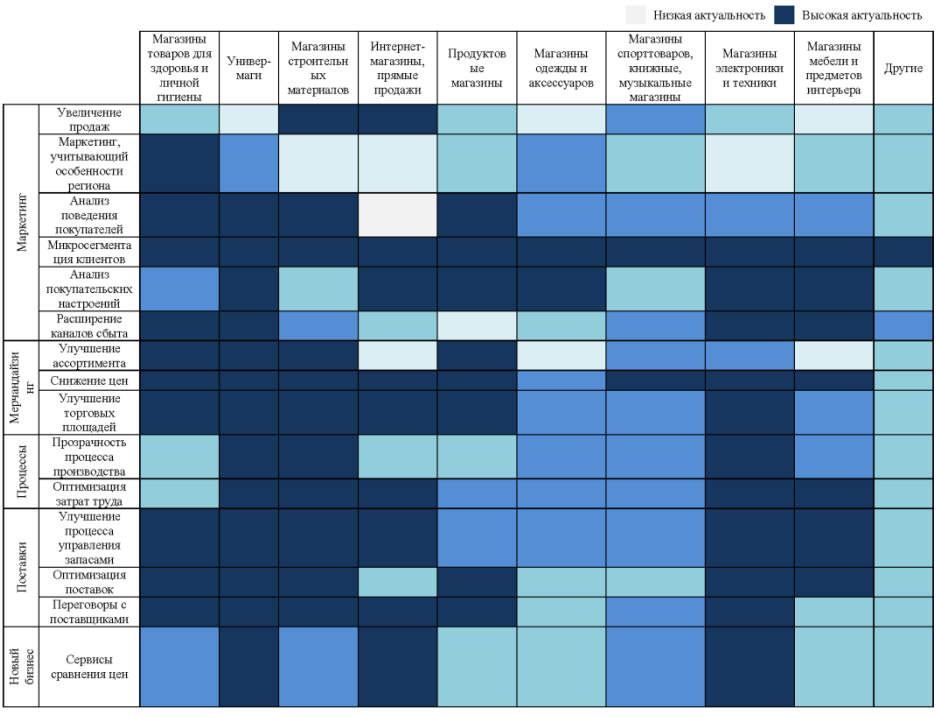

Рис. 4. Зависимость ожидаемого эффекта внедрения Big Data от сферы деятельности и направления политики компании

данных, интеллектуальные подходы в виде генетических алгоритмов, нейросетей и других ответвлений искусственного интеллекта.

Главные задачи платформы Hadoop — хранение, обработка и управление данными.

Основными составляющими платформы Hadoop являются [6, 7]:

- отказоустойчивая распределенная файловая система Hadoop Distributed File System (HDFS), при помощи которой осуществляется хранение;
- программный интерфейс Map Reduce, который является основой для написания приложений, обрабатывающих большие объемы структурированных и неструктурированных данных параллельно на кластере, состоящем из тысяч машин;
- Apache Hadoop YARN, выполняющий функцию управления данными.

Впервые о технологии Hadoop заговорили в 2007 г., и с каждым годом интерес к ней все больше возрастает. Это отражает индекс цитируемости Google (рис. 5).

Платформа Hadoop позволяет сократить время на обработку и подготовку данных, расширяет возможности по анализу, позволяет оперировать новой информацией и неструктурированными данными.

Рис. 5. Индекс цитируемости Google

Результаты проекта по внедрению технологии Hadoop подтверждают целесообразность ее использования (табл. 1).

Таблица 1

Результаты проекта [8]

Платформа	Описание оборудования	Примерная стоимость оборудования, руб.	Среднее время работы одного отчета, мин
БД Oracle	Сервер класса Hiend	12 млн	59
Кластер Hadoop	10 рабочих станций	250 тыс.	66
Hadoop с учетом оптимизации кластера	10 рабочих станций	300 тыс.	40

Решения, построенные на базе технологии Hadoop, обладают рядом существенных преимуществ. Основные из них приведены в табл. 2.

Таблица 2

Преимущества решения на базе Hadoop [8]

Преимущество	Краткое описание
Снижение времени на обработку данных	При обработке данных на кластере можно существенно сократить время на обработку данных
Снижение стоимости оборудования	Применение технологии Hadoop позволяет сократить затраты на оборудование, требуемое для хранения и обработки данных, в десятки раз

Окончание табл. 2

Преимущество	Краткое описание
Повышение отказоустойчивости. Технология позволяет построить отказоустойчивое решение	Выход из строя одного или нескольких узлов кластера влияет только на производительность системы, при этом система продолжает корректно работать и предоставлять сервис конечным пользователям
Линейная масштабируемость	Решение позволяет наращивать производительность просто за счет добавления новых узлов кластера. При этом производительность кластера возрастает линейно
Работа с неструктурированными данными	Технология позволяет осуществлять сложную обработку любых файлов, в том числе неструктурированных, благодаря чему такие данные могут быть эффективно обработаны и использованы

В России доступны решения, использующие технологию Big Data от ведущих производителей (Cisco, HP, IBM, Microsoft, Oracle, Apache), но проектов по реализации очень мало. Отечественный рынок находится в зачаточной стадии развития, но все без исключения аналитики прогнозируют взрывной рост технологий Big Data.

В октябре 2013 г. корпорация EMC провела исследование среди российских компаний, в ходе которого было выявлено, что использование Big Data ведет к существенному улучшению процессов принятия решений, повышает конкурентоспособность компаний и упрощает управление рисками:

- 70 % респондентов в России считают, что анализ данных их компании поможет принимать более обоснованные решения, а 35 % подтверждают, что высшее руководство их компаний полагается на результаты анализа Big Data при принятии стратегических бизнес-решений;
- 31 % респондентов сообщили, что их компании получили конкурентное преимущество в результате внедрения технологий Big Data, а 51 % считают, что отрасли, в которых используются такие технологии, покажут наиболее быстрый рост;
- 51 % респондентов согласны, что технологии анализа Big Data помогут в выявлении и предотвращении кибератак. Это может оказаться ключевым фактором, так как только 67 % респондентов в России уверены, что они смогут полностью восстановить все свои данные [3].

В современном мире, где информация часто обновляется и поступает из разных источников, предприятиям приходится работать с огромными массивами данных. Технологии Big Data позволяют предприяти-

ям хранить, структурировать и анализировать большие объемы информации. Это помогает руководству предприятия находить связь между различными факторами и использовать эту привилегию для получения благоприятного эффекта.

Одним из наиболее перспективных программных обеспечений для работы с Big Data, оптимизированных для промышленных предприятий, является платформа «Hadoop», разработанная компанией «Apache Software Foundation».

Большие данные — это не очередной ажиотаж на ИТ-рынке, это системный, качественный переход к составлению цепочек ценностей, основанных на знаниях. По эффекту его можно сравнить с появлением доступной компьютерной техники в конце прошлого века. Сейчас эта технология находится в фазе ожидания инвесторов: они следят, схлынут ли спекуляции вокруг новой технологии, или же это значимая инновация в стадии проникновения на рынок. В ближайшие 5 лет произойдет исправление недостатков технологии, и к 2018 г. начнется ее широкое распространение.

В то время как недальновидные консерваторы будут применять глубоко устаревшие подходы, предприятия, уже сейчас использующие технологии Big Data, в будущем окажутся на лидирующих позициях.

ЛИТЕРАТУРА

[1] Тиндал Сьюзен. Большие данные: все, что вам необходимо знать. *PC Week/RE,* 2012, № 25 (810). URL: http://www.pcweek.ru/idea/article/detail.php?ID=141962 (дата обращения 10.07.2014)

[2] Gantz John, Reinsel David. *The digital universe in 2020: Big Data, Bigger Digital Shadow s, and Biggest Grow th in the Far East.* URL: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf (дата обращения 10.07.2014)

[3] *Большие данные (Big Data).* URL: http://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_(Big_Data) (дата обращения 10.07.2014)

[4] *Acquia, Examples of Big Data Projects.* URL: http://www.acquia.com/examples-big-data-projects (дата обращения 10.07.2014)

[5] Manyika James, Chui Michael, Brad Brown, Bughin Jacques, Dobbs Richard, Roxburgh Charles, Hung Byers Angela. *Report of McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity.* URL: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (дата обращения 10.07.2014)

[6] Артемов Сергей. *Big Data: новые возможности для растущего бизнеса.* URL: http://www.pcweek.ru/upload/iblock/d05/jet-big-data.pdf (дата обращения 10.07.2014)

[7] *What is Apache Hadoop?* URL: http://hortonworks.com/hadoop/ (дата обращения 10.07.2014)

[8] DIS-group — технология Hadoop. *Практический опыт и экспертиза DIS-group,* 2012. URL: http://www.dis-group.ru) (дата обращения 10.07.2014)

Статья поступила в редакцию 28.08.2014

Ссылку на эту статью просим оформлять следующим образом:

Иванов П.Д., Вампиловв В.Ж. Технологии Big Data и их применение на современном промышленном предприятии. *Инженерный журнал: наука и инновации*, 2014, вып. 8. URL: http://engjournal.ru/catalog/it/asu/1228.html

Иванов Павел Дмитриевич — аспирант, ассистент кафедры предпринимательства и внешнеэкономической деятельности МГТУ им. Н.Э. Баумана.
e-mail: ivanovpd@bmstu.ru

Вампилова Валентина Жаргаловна — студентка кафедры предпринимательства и внешнеэкономической деятельности МГТУ им. Н.Э. Баумана.
e-mail: valentina.vampilova@gmail.com

Big Data technologies and their application in modern industrial enterprise

© P.D. Ivanov, V.Z. Vampilova

Bauman Moscow State Technical University, Moscow, 105005, Russia

Big Data Technology is a series of approaches, tools and methods for processing of structured and unstructured data and a huge amount of significant diversity perceived by man for results that are effective in conditions of continuous growth, the distribution of the numerous sites of computer network, formed in the late 2000s, alternative to traditional systems database management and Business Intelligence decisions. Currently, most of the largest providers of information technology for organizations in their business strategies are using the concept of big data, and basic market analysts of information technologies are devoting dedicated research to this concept. This paper presents an overview of Big Data technologies as of today, and analyzes the prospects for their further development. The necessity and prospects of use of Big Data Technologies are substantiated. Comparative analysis of Hadoop platform with its peers is made. The results of research of technologies Big Data are shown. The current state and development trends of Big Data technologies in Russia and abroad is analyzed.

Keywords: *Big Data, forecasting, software, Hadoop platform, competitive advantage.*

Ivanov P.D., postgraduate, assistant lecturer of the Department of Entrepreneurship and Foreign Economic Activities of the Bauman Moscow State Technical University.
e-mail: ivanovpd@bmstu.ru

Vampilova V.Z., a student of the Department of Entrepreneurship and Foreign Economic Activities of the Bauman Moscow State Technical University.
e-mail: valentina.vampilova@gmail.com