DOI: 10.17117/na.2016.08.01.278

<http://ucom.ru/doc/na.2016.08.01.278.pdf>

Поступила (Received): 06.08.2016

Лутаев Д.А.
BIG DATA и Machine Learning: возможности
для бизнеса и практические достижения

Lutaev D.A.
BIG DATA and Machine Learning: business
opportunities and practical achievements

В данной статье рассматриваются технологии в области Больших Данных и Машинного обучения, которые признаны прорывными в сфере IT. В работе представлены аналитические данные уровня развития и прогнозы, проиллюстрированы примеры в отдельных отраслях и перспективы применения. Дальнейшее развитие методов машинного обучения будет значительным шагом в сфере вычислительной техники

Ключевые слова: большие данные, машинное обучение, технологии

This article presents us technology in the sphere of Big Data and Machine learning which recognized as breakthrough in the field of IT. Analytical data of the level of development and projections which illustrated using examples in specific industries and application prospects are presented in this work. Further development of machine learning methods will be a significant step in the computing field

Key words: big data, machine learning, technology

Лутаев Даниил Антонович

Аспирант

Уральский федеральный университет им. первого
Президента России Б.Н. Ельцина
г. Екатеринбург, ул. Мира, 19

Lutaev Daniil Antonovich

Graduate

Ural federal university named the first President of
Russia B.N. Yeltsin
Ekaterinburg, Mira st., 19

Большие данные сегодня являются одним из ключевых драйверов развития информационных технологий. В эпоху информационных технологий, бума социальных сетей, по каждому пользователю интернета стало накапливаться значительное количество информации, что дало развитие направлению Big Data.

Термин «Большие Данные» вызывает множество споров. Так, в августе 2015 г., аналитическое агентство Gartner представило исследование (Gartner 2015 Hype Cycle) исключив Big Data из прорывных технологий, добавив машинное обучение (Machine Learning) [2].

Технологии больших данных находятся в центре активного обсуждения основных игроков индустрии больших данных и аналитиков на саммитах, конференциях, мастер-классах и панельных дискуссиях в разных странах мира.

Несколько выводов и опасений касательно перспектив и возможных сценариев развития технологий с большими данными лежат в основе дискуссий:

- «Данные ради данных», их будет сложно применить в разрезе отдельного бизнес-процесса и бизнес-контекста;
- ETL-системы и подходы, используемые сейчас в компаниях, не справятся со сложностью и комплексностью работы с огромными массивами данных;
- Отдельные проекты «отдельными кусками, превращающими процесс в лоскутное одеяло» могут стать проблемой комплексного решения бизнес-задач.

Рис. 1. Цикл зрелости технологий Gartner

Тем не менее, в разных отраслях экономики России, согласно данным исследования CNews Analytics в 2015 году, внедрение технологий больших данных имеет устойчивый раст. Более 1/3 опрошенных компаний (37%) работают по направлению внедрения технологий Big Data, в том числе 20% фактически применяют технологичные решения, а 17% – работают над разработкой и тестированием; 36% респондентов рассматривают такую возможность [3]. Часть опрошенные респондентов подтверждают, что испытывали проблемы работы с большими данными, 21% из которых определяют сферу работы при анализе неструктурированных данных, а 14% из них необходима большая аналитика для решения совершенно новых задач и направлений в бизнесе. Технологии больших в первую очередь востребованы на высококонкурентных рынках, где идет борьба за потребителя, поскольку данных позволяют анализировать поведение потребителей как по группам, так и персонально по клиенту.

Отраслевая структура мирового рынка распространения технологий больших данных представлена на рисунке.

В целом Big Data – это массивы цифровых структурированных и неструктурированных данных (БД в виде таблиц, текстов, изображений, видео и даже голоса). Развитие интернета, социальных сетей, мобильных устройств и M2M стимулировало огромный рост объемов и разнообразия типов. Ниже представлена динамика роста собираемой цифровой информации на примере США.

Рис. 2. Отраслевая структура мирового рынка распространений больших данных

Рис. 3. Динамика роста собираемой цифровой информации в США

Хранить и обрабатывать сегодня можно практически неограниченные объемы. По данным EMC к 2020 году объем Big Data России (количество информации) вырастет до 980 эксабайт.

При этом, существуют мнения, что термин Big Data означает лишь объем накопленной информации, однако понятие намного шире и включает техническую сторону: технологии хранения, вычисления, а также сервисные услуги. Кроме того, большим данным, кроме объема и многообразия технологий, присуща «скорость», то есть данные генерируются с огромной скоростью, вычисления и анализ должны происходить в режиме online.

Одними из самых известных и обсуждаемых проектов в области распределенных вычислений является проекты Hadoop и Apache Spark.

– Hadoop – проект фонда Apache Software Foundation, свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Вокруг Hadoop образовалась целая экосистема из связанных проектов и технологий, многие из которых развивались изначально в рамках проекта, а впоследствии стали самостоятельными [6].

– Spark – проект Apache, предназначенный для кластерных вычислений, представляет собой быструю и универсальную среду для обработки данных, в том числе и для машинного обучения. Spark также имеет API и для R (пакет SparkR), который входит в сам дистрибутив Spark. Но, помимо работы с данным API, имеется еще два альтернативных способа работы со Spark в R. Итого, мы имеем три различных способа взаимодействия с кластером Spark. В данном посте приводится обзор основных возможностей каждого из способов, а также, используя один из вариантов, построим простейшую модель машинного обучения на небольшом объеме текстовых файлов (3,5 ГБ, 14 млн. строк) на кластере Spark развернутого в Azure HDInsight [7].

Большие данные дают новые возможности компаний привлечением клиентов, наращивании объемов бизнеса, оптимизации действующих бизнес-процессов. Кроме того, активное развитие и распространение технологий Big Data является значительным стимулом в развитии инновационной деятельности компании, поскольку предоставляет значительно большие технические возможности.

По прогнозу Bain & Company's Insights Analysis, отраслевая структура использования технологии больших данных к 2020 году будет представлена:

- затраты финансового сектора на Big Data составят 6,4 млрд. USD с динамикой 22% в год;
- затраты интернет-компаний составят 2,8 млрд USD с приростом в 26% за следующие 5 лет;
- затраты госсектора будут сопоставимы с затратами интернет-компаний, имея темп роста чуть ниже 22%;
- затраты сектор телекоммуникаций прогнозируются с приростом 40% в объеме 1,2 млрд USD в 2020 году.

Широкое отраслевое и географическое распространение проиллюстрируем на примере предлагаемых технологий Корпорация EMC – мирового лидера на рынке продуктов, услуг и решений для хранения и управления информацией:

Энергетика, коммунальные предприятия и проектирование. Компания 3TIER использует большие данные для анализа огромных массивов метеорологических и климатических данных, на основе чего предоставляется информация заказчикам о рисках, при использовании возобновляемых источников энергии. Silver Spring Networks использует данные для сокращения расходов энергии и предоставления надежных и экономичных услуг.

Здравоохранение. Национальный фонд медицинского страхования Франции (CNAMTS, National Health Insurance Fund) использует xCP при обработке заявок. Порядка 200 миллионов документов в год обрабатываются со

скоростью 1,25 миллиона страниц в день, в перспективе планируется до 2,75 миллионов страниц.

Детская санитарная часть округа Лос-Анджелес реализовала систему безопасного хранения. Новая система документооборота реализует в автоматическом режиме регистрацию, извлечение и архивацию более 48 000 карт пациентов.

Финансовые услуги и страхование. Generali Hellas благодаря решению EMC Documentum xCP значительно ускорило рассмотрения заявок на страхование «Раньше для принятия решения требовалась неделя. Сегодня это может занять не более двух дней».

Группа Univé-VGZ-IZA-Trias (UVIT) считает, что решение Documentum xCP упростило процесс принятия решений. Заявки могут обрабатываться одновременно несколькими специалистами, так как информация о заявке доступна в цифровой форме.

Viteos Capital Market Services считает, что новые возможности Documentum позволили добиться 15%-ного роста в течение года, что привело к увеличению прибыли и расширению бизнеса за счет предложения заказчикам новых услуг.

Почтовый банк Хорватии (HPB) использует технологии в управлении операционными рисками, за счет упрощения процесса выполнения персоналом функций отмечено сокращение операционных рисков.

Дальнейшее развитие технологий в области Big Data неразрывно связано с широким внедрением **методов машинного обучения** в сочетании с работой с данными. Технологии машинного обучения стали чрезвычайно популярными, благодаря реализации успешных проектов, таких как алгоритм ранжирования Pagerank от корпорации Google, алгоритм рекомендаций Cinematch от крупнейшего дистрибутора медиа-контента Netflix, библиотека алгоритмов обработки изображений («компьютерного зрения») с открытым программным кодом OpenCV и др.

Достижения этой научной дисциплины буквально прорвали технологические и психологические барьеры, которые стояли на пути реализации идеи о возможности более глубокой «переработки» уже существующих информационных массивов с целью извлечения из них нового ценного знания.

Машинное обучение: выявление скрытых зависимостей на основе анализа всего объема данных дает новое качество результатов.

В заключении хотелось бы отметить, согласно прогнозам развития в области сбора и обработки больших данных наибольшую популярность приобретут технологии визуализации данных, анализа медиа файлов и интернета вещей. По мнению ведущих игроков и экспертов в области IT технологий, дальнейшее развитие методов машинного обучения будет значительным шагом в сфере вычислительной техники.

Список используемых источников:

1. Золотых Н.Ю. Машинное обучение и анализ данных (machine learning and data mining). URL: http://www.uic.unn.ru/~zny/ml/Lectures/ml_pres.pdf
2. Исследование аналитического агентства Gartner. URL: <http://www.gartner.com/newsroom/id/3114217>

3. Аналитический обзор рынка Big Data.

URL: <https://habrahabr.ru/company/moex/blog/256747/>

4. Аналитический обзор рынка Больших Данных.

URL: <http://ipoboard.ru/files/cms/5e3af134b9942559eb802ea93a1c9050>

5. Работа с Big Data: основные области и возможности.

URL: http://www.marketing.spb.ru/lib-research/methods/Big_Data.htm

6. Hadoop. URL: <https://ru.wikipedia.org/wiki/Hadoop>

7. R и Spark. URL: <https://habrahabr.ru/post/308534/>

© 2016, Лутаев Д.А.

*BIG DATA и Machine Learning: возможности для
бизнеса и практические достижения*

© 2016, Lutaev D.A.

*BIG DATA and Machine Learning: business
opportunities and practical achievements*