Компьютеризированные пакеты для синтеза и анализа

Красикова Т.Ю., к.э.н., доцент

Big data (Большие данные) в области анализа данных

Стремительное развитие информационных технологий и социальных сервисов служит причиной поиска и разработки информационных решений, которые позволят обрабатывать гигантские объемы входящей информации. В соответствии с исследованием IDC Digital Universe прогнозируется увеличение объема данных на планете до 40 зеттабайтов к 2020 г., это означает, что объем информации на человека будет составлять 5200 Гб.

Для создания необходимых условий по развитию цифровой экономики в Российской Федерации в настоящее время реализуется программа «Цифровая экономика Российской Федерации». Направление, необходимое для повышения конкурентоспособности страны, обеспечения экономического роста национального суверенитета, а также качества жизни самих граждан. Указом Президента РФ от 9 мая 2017 г. № 203 утверждена государственная программа «О Стратегии развития информационного общества в Российской Федерации на 2017 - 2030 г.», в рамках которой планируется развивать технологии Big Data.

Big Data - это серия подходов, технологий и методов, предназначенных для решения проблемы обработки больших объемов структурированных и неструктурированных данных для получения результатов, которые человек способен воспринять. Big Data следует отличать от обычного анализа объемов информации. Большая часть объемов данных представлена в нетрадиционном, неструктурированном для БД формате, таком как: веб-журналы, видеозаписи, текстовые документ, машинный код. Эти данные разбросаны по различным хранилищам, иногда находящимся за пределами фирмы.

К категории **Большие данные (Big Data)** относится информация, которую уже невозможно обрабатывать традиционными способами, в том числе структурированные данные, медиа и случайные объекты. Некоторые эксперты считают, что для работы с ними на смену традиционным монолитным системам пришли новые массивно-параллельные решения.

Big data — это различные инструменты, подходы и методы обработки как структурированных, так и неструктурированных данных для того, чтобы их использовать для конкретных задач и целей.

Фактически, **Big data** — это решение проблем и альтернатива традиционным системам управления данными.

Техники и методы анализа, применимые к **Big data** по McKinsey:

- Data Mining;
- Краудсорсинг;
- Смешение и интеграция данных;
- Машинное обучение;
- Искусственные нейронные сети;
- Распознавание образов;
- Прогнозная аналитика;
- Имитационное моделирование;
- Пространственный анализ;
- Статистический анализ;
- Визуализация аналитических данных.

Горизонтальная масштабируемость, которая обеспечивает обработку данных — базовый принцип обработки больших данных. Данные распределены на вычислительные узлы, а обработка происходит без деградации производительности. McKinsey включил в контекст применимости также реляционные системы управления и Business Intelligence.

Технологии:

- NoSQL;
- MapReduce;
- Hadoop;
- R;
- Аппаратные решения.

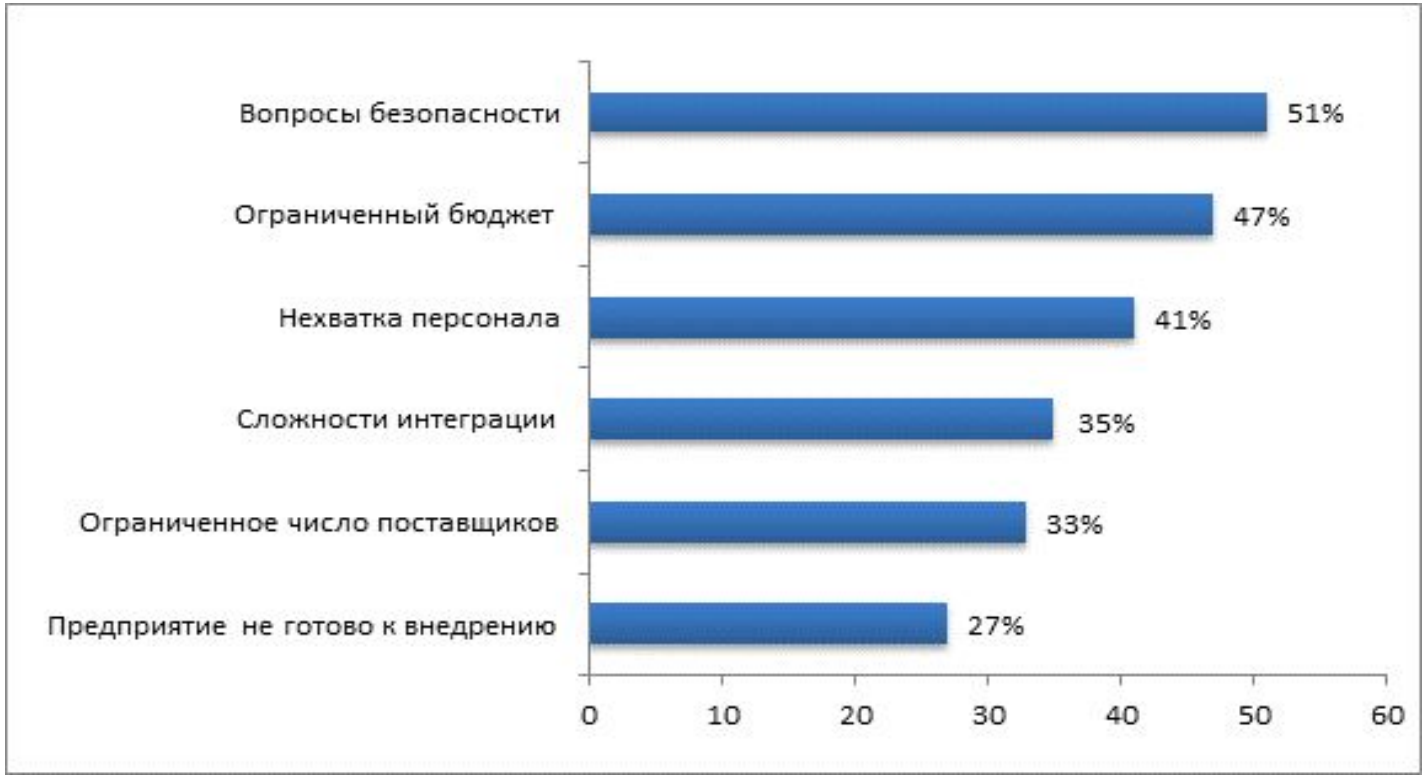
В 2017 году мировой доход на рынке big data должен достигнуть $150,8 млрд, что на 12,4% больше, чем в прошлом году. В мировом масштабе российский рынок услуг и технологий big data ещё очень мал. В 2014 году американская компания IDC оценивала его в $340 млн. В России технологию используют в банковской сфере, энергетике, логистике, государственном секторе, телекоме и промышленности.

Внедрение Big Data подразумевает все способы работы с большой совокупностью данных, постоянно обновляемых и находящихся в разных источниках. Таким образом, организациям требуются инструменты для установления связи между этими данными, в соответствии с анализом которых можно сделать необходимые выводы. Учитывая обстоятельство, что данные постоянно обновляются, обработать и структурировать эту информацию становится еще сложнее, что объясняет возникновение технологий Big Data.

Изучая теоретический и практический опыт, наиболее распространенными программными инструментами для создания информационной инфраструктуры являются Big Data-Map Reduce, Hadoop и NoSQL. Одной из основополагающих технологий считается Hadoop. Это проект фонда Apache Software Foundation, используемый для реализации поисковых и контекстных механизмов высоконагруженных веб-сайтов. Hadoop свободно распространяет набор утилит, библиотек и программный каркас для разработки и выполнения распределенных программ, работающих на кластерах из множества узлов. При импортозамещении Hadoop является основной платформой развития российского рынка, способствующей повышению отечественной конкуренции.

Согласно исследованию Accenture (Аксенчер), проведенному осенью 2017 г. (рис.) практически 60% компаний работают с большими данными и завершают минимум несколько проектов. Довольны результатом 92% этих компаний, а 89% считают, что Big Data крайне важная часть преобразования в бизнесе. В исследовании компании Accenture приняли участие 1000 руководителей из 19 стран мира. Опрос был проведен Economist Intelligence Unit среди 1135 респондентов по всему миру. Многие компании при внедрении новой технологии столкнулись с некоторыми проблемами. Безопасность является важнейшей составляющей для 51%, бюджет для 47%, нехватка необходимых кадров для 41%, затруднения в интегрировании для 35%. В целом оценка довольно оптимистична. 89% компании считают, что подобно интернету, они изменят бизнес также сильно. А свое конкурентное преимущество, по оценке 79% респондентов, потеряют компании, которые не внедрили большие данные.

Основные проблемы при внедрении проектов Big Data



По мнению экспертов-разработчиков Big Data, на расширение сфер применения обработки больших объемов данных сильнее влияет интернет. В результате увеличение количества устройств, которые подключены к Интернету, возникает все больше информации, которую можно применить при управлении бизнесом. Это происходит от того, что производитель изучает потребительский спрос посредством анализа полученных из Интернета данных, что позволяет ему создать рекламу, которая привлечет внимание потребителя.

Эксперты Meta Group в 2001 г. выделили три свойства, которые позволяют отнести данные к Big Data, так как не всю информацию возможно анализировать. Эти характеристики были названы «Три V». Во-первых, это величина физического объема (от англ. volume). Во-вторых, скорость (от англ. velocity), так как постоянно увеличивающийся объем данных требует быстрой обработки. В-третьих, многообразие (от анг. variety), то есть способность обрабатывать различные данные одновременно. В процессе анализа целесообразно выделить четвертую V (veracity — достоверность/правдоподобность данных) и даже пятую V (viability - жизнеспособность).

Специалисты TmaxSoft, выдвинули прогноз, согласно которому следующая «волна» Big Data потребует модернизации СУБД.В соответствии с отчетом IDC из-за увеличения объемов данных на подключенных к Интернету устройствах, доходы, связанные с большими данными, увеличатся с $130 млрд в 2017 г. до более, чем $203 млрд к 2020 г. По мнению экспертов TmaxSoft компании, не имеющие ИТ-инфраструктуру, необходимую для обработки больших объемов данных, не получат выгоду от предполагаемого роста. В накопленных объемах данных содержится важная информация о бизнесе и клиентах, это значит, что создать конкурентоспособный продукт, эффективно функционировать на рынке и получить преимущество по сравнению с другими могут только те компании, которые успешно используют эту информацию.

Однако многие предприятия не располагают соответствующей IT-инфраструктурой для обеспечения необходимой емкости систем хранения, возникает необходимость в приобретении дополнительных ресурсов, способных обрабатывать, анализировать и извлекать информацию из неструктурированных данных, вследствие чего увеличиваются инвестиции в IT-инфраструктуру. Представитель TmaxSoft отмечает, что предприятиям требуется план, включающий в себя источники данных для извлечения, продолжительность жизненного цикла данных, совместимость разных реляционных СУБД и масштабируемость хранения.

Big Data и Business Intelligence являются совершенными дополнениями друг друга. Для анализа текущей ситуации больше подходит бизнес-аналитика. В режиме реального времени пользователи могут получать нужную им информацию. Если компании требуется анализ не только внутренних, но внешних источников, при использовании различных аналитических методов и подходов, требуется применение больших данных. Схожие цели этих технологий делают их союзниками, которые отлично взаимодействуют друг с другом и приносят в разы большую пользу бизнесу.

Big Data также является ключевым компонентом Интернет вещей (Internet of Things, IoT). По оценкам специалистов, в мире в ближайшие 5-10 лет появится около 50 млрд. взаимосвязанных подключаемых устройств, которые будут генерировать зеттабайты данных. Суть Интернет вещей заключается в создании более умной продукции. Встраивание «интернет-чипов» в те устройства, которые ранее не имели вычислительной мощности. Эти чипы используются для обработки данных устройства и распознания привычек потребителей. При увеличении подключаемых устройств соответственно следует и рост объема данных. Именно в этом заключается тесная взаимосвязь IoT с Big Data.