Big Data: несколько простых вопросов о сложном явлении

Текст Михаил Цымблер

Отличительной чертой современных информационных технологий является феномен Big Data (Больших Данных). Соответствующий термин в настоящее время используется чрезвычайно часто как в научных, так и в популярных публикациях. В этой статье мы попытаемся раскрыть содержание феномена Больших Данных, показав, какие технологии и программные решения объединяются в данном понятии и чем они могут быть полезны рядовому пользователю.

Откуда произошел термин «Большие Данные»?

Проблемой организации эффективной обработки данных больших объемов научное сообщество начало заниматься фактически одновременно с изобретением реляционной модели данных Коддом в 1970 году: его соответствующая статья имела название «A Relational Model of Data for Large Shared Data Banks». В 1975 году возник термин Very Large Database (сверхбольшая база данных). В одной из первых работ сверхбольшой предлагалось называть такую базу данных, вывод содержимого которой занимает более 32 часов. На сегодня эпитетом «сверхбольшая» награждают базу данных, имеющую размер более одного терабайта или количество строк более одного миллиона, хотя эти лимиты имеют тенденцию к увеличению.

Словосочетание «big data», судя по каталогу научных публикаций DBLP.org, по-видимому, впервые было вынесено в заголовок научной статьи в 1999 году: «Automation or Interaction: What's Best for Big Data?». Однако в данном случае это был скорее один из многочисленных синонимов большого объема (big data, very large database, massive dataset и др.), чем самостоятельный термин для обозначения соответствующего явления в информационном обществе.

Исследование количества научных статей в каталоге DBLP.org и web-страниц, проиндексированных поисковым сервисом Google, посвященных тематике Больших Данных (см. рис. 1) показывает взрывной рост интереса научного и пользовательского сообществ в последние три года. Кроме того, неожиданно оказалось, что эта тематика имеет большую популярность по сравнению с суперкомпьютерами. Поражает тот факт, что на начало 2014 года Google проиндексировал примерно столько же web-страниц, релевантных теме Big Data, что и за весь прошлый год.

Интересно, что русский и английский варианты материалов в Википедии дают разную информацию относительно авторства термина «Большие Данные». Русскоязычная статья отдает приоритет К. Линчу, редактору журнала Nature, подготовившему в 2008 году специальный номер этого журнала, который был посвящен данной проблеме. В английском варианте указывается, что термин ввел в 2001 году Д. Лэйни, аналитик консалтинговой компании META Group (сейчас Gartner). Однако практически все источники

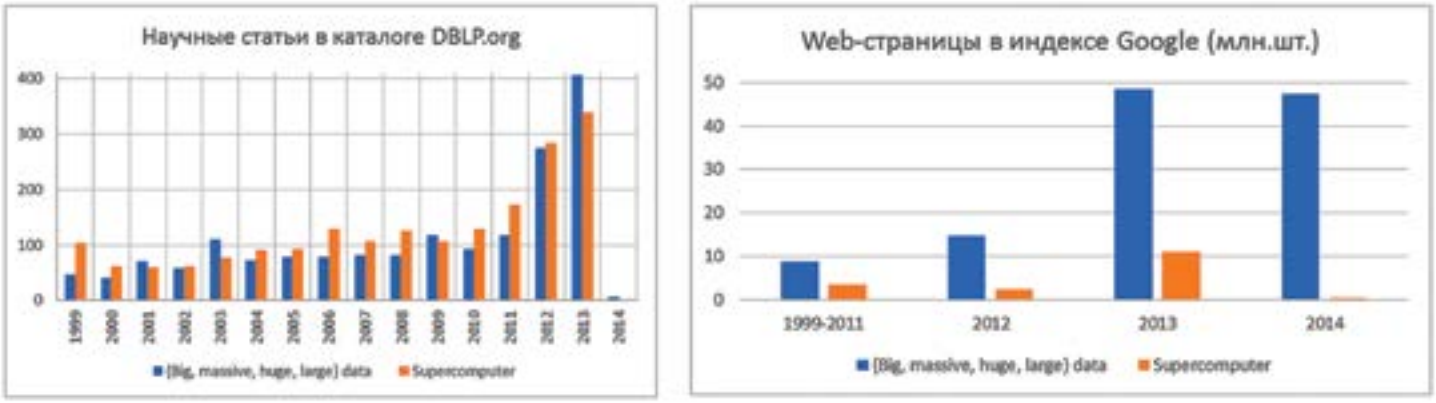

Рис. 1. Количество научных публикаций и web-страниц с ключевыми словами big data и supercomputer

в качестве важнейших характеристик Big Data указывают три «V»: Volume (размер), Velocity (скорость) и Variety (разнообразие). Таким образом, феномен Больших Данных можно определить как наличие в современном информационном обществе беспрецедентно больших объемов данных, с высокой скоростью поступающих из разнообразных предметных областей в различных форматах и требующих новых моделей, методов и алгоритмов, а также аппаратных и программных технологий для эффективного преобразования этих данных в ценную для общества информацию.
Сейчас к вышеуказанным трем «V» Больших Данных часто добавляют (видимо, больше из любви к красивым аббревиатурам) еще два: Value (ценность) — для того чтобы подчеркнуть высокую значимость данного ресурса, и Veracity (достоверность) — для индикации внутренней целостности и непротиворечивости ресурса, которые важны при принятии стратегических решений на основе аналитической обработки этих данных.

Насколько велики Большие Данные?

Согласно исследованию аналитиков компании IDC «The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East», с 2010 по 2020 год объем мировых данных будет удваиваться каждые два года (привет от Г. Мура) и в итоге увеличится в 50 раз, достигнув отметки 40 Зеттабайт, т. е. примерно 5 Терабайт на каждого жителя планеты. Масштаб единиц измерения объемов данных можно понять по рис. 2.
Согласно исследованию аналитиков той же компании, в 2011 году объем мировых данных составил 1.8 Зеттабайт. Чтобы оценить, насколько велики эти объемы, приведем два примера из упомянутого отчета. 1.8 Зеттабайт достаточно, чтобы заполнить информацией 57.5 млрд планшетов Apple iPad емкостью 32 Гигабайта, из которых можно построить «Великую китайскую iPad-стену». 1.8 Зеттабайт «весят» столько же, сколько 200 млн двухчасовых фильмов в формате высокой четкости, которые можно просматривать ежедневно без перерыва в течение 47 млн лет.
Источником таких пугающих объемов данных являются электронные информационные сервисы, бизнес и научные исследования. Объем хранилища социальной сети Facebook ежедневно увеличивается

Рис. 2. Масштаб единиц измерения данных

на 500 Терабайт. Архив Интернета Archive.org к 2012 году достиг размера 10 Петабайт и ежемесячно увеличивается на 20 Терабайт. Сделки Нью-Йоркской фондовой биржи «обходятся» ей ежедневно в 1 Терабайт данных. Эксперименты на Большом Адронном Коллайдере могут генерировать данные со скоростью 1 Петабайт в секунду.

Как используются Большие Данные?

Большие Данные подвергаются интеллектуальному анализу для выявления скрытых трендов и аномалий, которые необходимы при принятии стратегически важных решений в различных областях жизнедеятельности человека. Приведем несколько примеров проектов, связанных с аналитикой Больших Данных.

Проект Hedonometer.org, выполняемый в Университете Вермонта (США), направлен на разработку алгоритмов, с помощью которых можно вычислить (по определению разработчиков) уровень счастья жителей США посредством интеллектуального анализа их микроблогов Twitter. Уровень счастья зависит от частоты определенных слов в твитах пользователей (например, ключевые слова lol, haha, fun повышают уровень счастья, а sick, bad, sad — понижают). Объем ежедневно анализируемых данных составляет около 100 Гигабайт. Имеется визуализация результатов анализа с возможностью отбора периода и региона (см. рис. 3).

В планах разработчиков — добавление учета фраз и предложений, а не только ключевых слов, анализ постов на других языках и в других социальных сетях, а также выявление других эмоций помимо счастья или горя.

Компания Яндекс проанализировала музыкальные пристрастия пользователей сервиса Яндекс. Музыка (см. рис. 4). Карта показывает размер и сходство аудиторий разных исполнителей. Размер круга соответствует количеству пользователей, которые хотя бы раз в год слушали музыканта. Чем ближе круги друг к другу, тем больше у исполнителей общих слушателей.

Корпорация Google разработала сервис Flu Trends, который позволяет определить скорость распространения вируса гриппа в различных странах. Сервис анализирует объем поисковых запросов по темам, связанным с гриппом, позволяя с достаточно высокой точностью оценивать распространение заболевания в мире практически в реальном времени (см. рис. 5).

В рамках подготовки спортсменов сборной к Олимпиаде 2012 года Министерство культуры, СМИ и спорта Великобритании инициировало исследовательскую программу ESPRIT (Elite Sport Performance

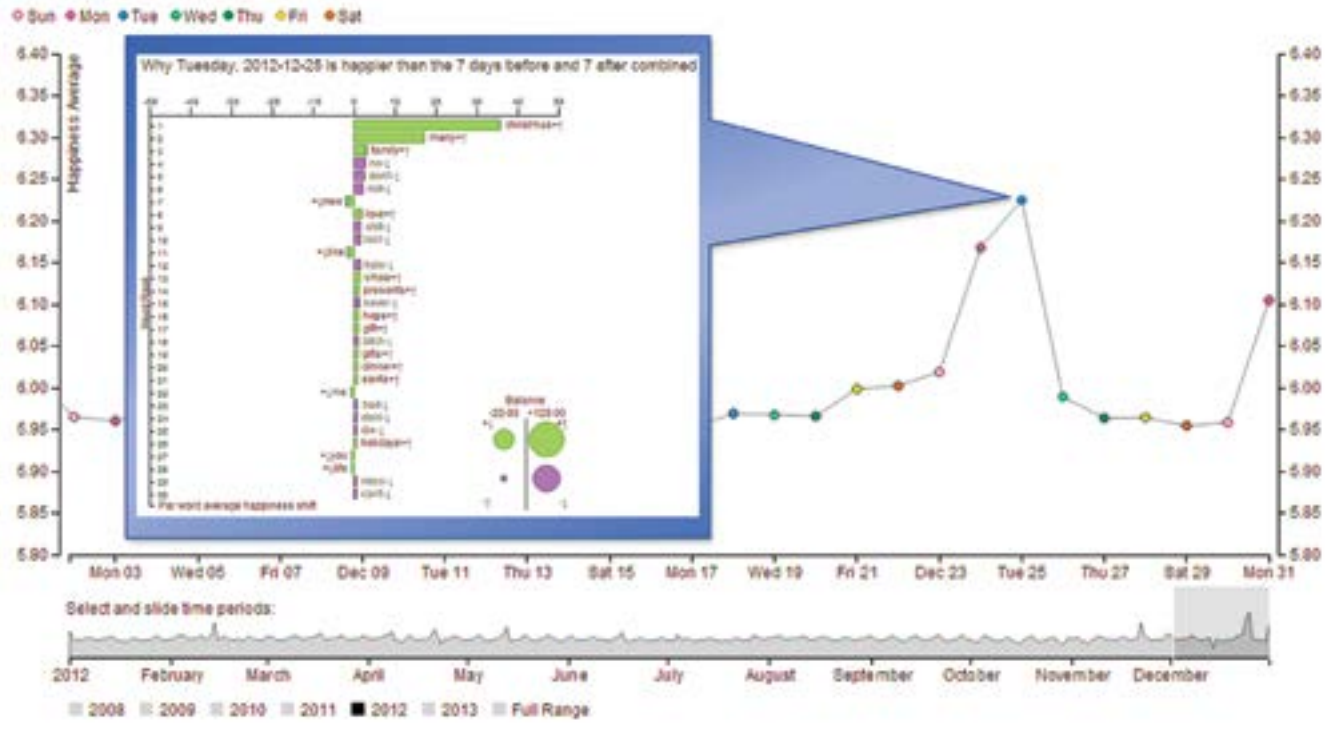

Рис. 3. Визуализация уровня счастья в США на 25.12.2012 в проекте Hedonometer.org

Рис. 4. Карта сервиса Яндекс.Музыка

работки данных для предсказания и определения лечебных мероприятий, а также оценки стоимости и качества мероприятий для пациентов.

В апреле-июне 2014 года в Далласе (Техас, США) пройдет чемпионат мира по обработке Больших Данных — Big Data World Championships (TEXATA 2014). Соревнование включает в себя два заочных тура и финальный очный тур. Во время тура участники должны будут ответить на теоретические вопросы и решить конкретные задачи, касающиеся применения различных технологий для хранения, обработки и интеллектуального анализа данных сверхбольших объемов в следующих областях: здравоохранение, энергетика, страхование, науки о Земле и др. Участники могут использовать любые технологии и программные продукты.

Research in Training). Программа направлена на разработку сенсорного оборудования для получения сверхбольших баз данных показателей физиологической активности атлетов в реальном времени в условиях тренировок и соревнований, а также программного обеспечения, которое позволяет анализировать полученные данные. Напомним, что по сравнению с предыдущей Олимпиадой в Пекине сборная Великобритании поднялась на одну строчку выше в неофициальном медальном зачете, войдя в тройку самых сильных олимпийских команд планеты.

За рубежом проводятся различные конкурсы и чемпионаты по обработке Больших Данных, инициируемые государственными структурами. Например, в США объявлен национальный конкурс «Revolutionizing the Delivery of Health Care through Big Data» по решению проблем обработки Больших Данных в области здравоохранения. Конкурс стал ответом

Рис. 5. Сервис Google Flu Trends

федерального правительства и штатов на проблемы роста расходов на здравоохранение при низком качестве услуг и попыткой внедрения новых моделей поставки и оплаты услуг, которые могут улучшить качество и экономическую эффективность здравоохранения. Целью конкурса является помощь большим и малым государственным и частным организациям здравоохранения по внедрению технологий эффективной аналитической об-

Оргкомитет Международной выставки в области информационных технологий CeBIT'2014 объявил о проведении международного конкурса CODE_n14 среди стартапов в области обработки сверхбольших данных, включая финансовые услуги, здравоохранение, телекоммуникации, автомобилестроение и др. Компании-победители получат возможность представить свои разработки на выставке в Ганновере (Германия) в марте 2014 года.