Эту книгу хорошо дополняют:

Большие данные
Виктор Майер-Шенбергер

Великий переход
Николас Карр

Новый цифровой мир
Эрик Шмидт

Bill Franks

Taming the Big Data Tidal Wave

Finding Opportunities in Huge Data Streams with Advanced Analytics

John Wiley & Sons, Inc.

Билл Фрэнкс

Укрощение больших данных

Как извлекать смысл из гигантских потоков данных с помощью продвинутой аналитики

Перевод с английского Андрея Баранова

Издательство «Манн, Иванов и Фербер»
Москва, 2014

УДК 330.47
ББК 65.051.03
Ф93

Фрэнкс, Б.

Ф93 Укрощение больших данных: как извлекать смысл из гигантских потоков данных с помощью продвинутой аналитики / Билл Фрэнкс ; пер. с англ. Андрея Баранова. — М. : Манн, Иванов и Фербер, 2014. — 352 с.

ISBN 978-5-00057-146-0

По убеждению Билла Фрэнкса, ведущего аналитика всемирно известной компании Teradata, уже сейчас наступила эпоха совершенно новых подходов в аналитической сфере и в использовании больших объемов данных. Что такое большие данные, каково их значение, каковы методы, технологии и принципы новейшей аналитики и как это повлияет на последующее развитие бизнеса — в этой книге вы найдете подробную, четко структурированную, изложенную простым языком и наиболее полную информацию о больших данных.

УДК 330.47
ББК 65.051.03

ISBN 978-5-00057-146-0

Оглавление

Эта книга посвящается Стейси, Джесси и Даниэль.
Они мирились с тем, что многие ночи и выходные
я посвящал этой книге

Предисловие

Хотите вы этого или нет, но в ближайшее время на вас обрушится огромное количество данных. Возможно, уже обрушилось. Возможно, вы уже на протяжении некоторого времени пытаетесь справиться с этим, понять, как хранить данные для последующего доступа, как исправлять ошибки и недостатки или классифицировать их. Теперь вы готовы извлечь смысл из этого огромного набора данных путем их анализа, чтобы узнать что-то о своих клиентах, своем бизнесе или о некоторых аспектах своей организационной среды. А возможно, вы пока далеки от этого, но уже видите свет в конце туннеля управления данными.

В любом случае вы пришли по адресу. Билл Фрэнкс предполагает, что вскоре мир наводнят не только большие данные, но и книги о больших данных. Я предсказываю (без всякой аналитики), что эта книга будет отличаться от прочих. Во-первых, она одна из первых на эту тему. Но, самое главное, она сконцентрирована на ином.

Большинство книг о больших данных будут посвящены управлению большими данными: тому, как собирать их в базу данных или хранилище данных, или тому, как структурировать и классифицировать их. Если вы много читаете о Hadoop, MapReduce или других методах хранения данных, это значит, что вы наткнулись на книгу, посвященную управлению большими данными.

Это, конечно, важная работа. Независимо от их объема и качества данные мало чем полезны, если их не поместить в такую среду и формат, которые позволят получить к ним доступ и проанализировать их.

Сама по себе тема управления большими данными не обеспечивает движения вперед. Для того чтобы извлечь пользу из данных, необходимо проанализировать их и совершить какое-либо действие

на основании результатов анализа. Так же как традиционные инструменты управления базами данных не обеспечивали автоматический анализ данных о транзакциях, полученных из традиционных систем, системы Hadoop и MapReduce не производят автоматическую интерпретацию данных, полученных от сайтов, картирования генов, анализа изображений или других источников больших данных. Даже до наступления эпохи больших данных многие организации многие годы (а иногда и десятилетия) занимались исключительно управлением данными, не извлекая из них никакой пользы в плане улучшения качества анализа и принятия решений.

Думаю, эта книга акцентирует внимание именно на том, на чем нужно. Она в первую очередь посвящена эффективному анализу больших объемов данных, а не управлению ими. Она начинается с данных и переходит к таким темам, как фреймовое представление решения, построение аналитического центра и создание аналитической культуры. Разумеется, здесь упоминается об управлении большими данными, однако основное внимание уделено созданию, организации, подбору персонала и воплощению аналитических инициатив, которые позволяют извлечь из входных данных пользу.

На тот случай, если вы этого не заметили: в настоящее время тема аналитики крайне актуальна в бизнес-среде. Я занимался в основном вопросами конкуренции компаний в области аналитики, и мои книги и статьи по этой теме были самыми популярными из всех, что я когда-либо писал. Конференции на тему аналитики проводятся повсеместно. У таких крупных консалтинговых фирм, как Accenture, Deloitte и IBM, имеется большой практический опыт в этой области. Многие компании, государственные и даже некоммерческие организации сделали аналитику своим стратегическим приоритетом. Сегодня наблюдается повышенный интерес к проблеме больших данных, однако в центре внимания должны по-прежнему оставаться способы приведения этих данных в форму, позволяющую проанализировать их и использовать в процессе принятия решений.

Билл Фрэнкс находится в уникальном положении: он может описать пересечение области больших данных и аналитики. Его компания Teradata, в отличие от других поставщиков систем хранения данных, всегда была максимально сосредоточена именно на анализе данных и извлечении из них пользы для бизнеса. И хотя компания

хорошо известна как поставщик корпоративных инструментов для хранения данных, Teradata в течение многих лет также предоставляла набор аналитических приложений.

За последние несколько лет Teradata наладила тесное партнерство с SAS — ведущим поставщиком аналитического программного обеспечения — для разработки высокомасштабируемых инструментов проведения анализа больших баз данных. Эти инструменты, которые часто подразумевают встроенный анализ в среде хранилища данных, предназначены для таких мощных аналитических приложений, как системы обнаружения мошенничества в режиме реального времени и крупномасштабного скоринга* покупательского поведения потребителей. Билл Фрэнкс — скоринг-директор по аналитике этого партнерства и поэтому имеет доступ к идеям и опыту в области проведения крупномасштабного анализа и «обработки в базе данных». Вероятно, лучшего источника на эту тему просто не существует.

Так что же еще особенно интересного и важного содержится в этой книге?

- Глава 1 включает в себя обзор концепции больших данных и объясняет, что «размер не всегда имеет значение». На протяжении всей книги Фрэнкс отмечает, что бóльшая часть данных вообще бесполезна и очень важно уметь отфильтровывать ненужные данные.
- Обзор источников больших данных в главе 3 — интересный, полезный и необыкновенно подробный каталог. Подход к веб-данным и веб-аналитике в главе 2 может заинтересовать людей и организации, которые стремятся понять поведение потребителей, совершающих покупки через интернет. Этот подход выходит далеко за рамки обычной веб-аналитики, ориентированной на отчетность.
- Глава 4, посвященная «эволюции масштабируемости аналитических систем», представит вам технологические платформы для

* Скоринг (англ. score — подсчет очков) — система оценки кредитоспособности, в основу которой положены численные статистические методы обработки анкет потенциальных заемщиков. Суть ее в том, что за каждую позицию анкеты («стаж работы» или «количество детей») потенциальный заемщик получает некое количество баллов. В зависимости от суммы набранных баллов принимается решение об одобрении или отказе в выдаче кредита. *Прим. ред*.

больших данных и аналитики с такой точки зрения, которую вы больше нигде не найдете. В ней автор также описывает такие современные технологии, как MapReduce, и разумно утверждает, что анализ больших данных потребует использования комбинации сред.

- Эта книга содержит ультрасовременные сведения о том, как создавать аналитические среды и управлять ими, — эти сведения вы также нигде больше не найдете. Если вы хотите познакомиться с новейшими размышлениями на тему «аналитических песочниц» и «аналитических наборов данных предприятия» (это была новая для меня тема, однако теперь я знаю, что́ они собой представляют и какое значение имеют), вы найдете их в главе 5, которая также содержит важные замечания по поводу необходимости в системах и процессах управления моделями и скорингом.
- В главе 6 рассматриваются доступные сегодня типы аналитического программного обеспечения, в том числе программной среды R с открытым исходным кодом. Обычно очень трудно найти здравое рассуждение о сильных и слабых сторонах различных аналитических сред, однако здесь оно представлено. И наконец, описание методов анализа будет понятно даже далеким от техники людям.
- Третья часть книги сосредоточена на том, как управлять человеческим и организационным аспектами аналитики. В этом автор также опирается на здравый смысл. Мне, например, особенно понравился акцент на фреймовом представлении проблем и решений в главе 7. Слишком многие аналитики принимаются за анализ, не задумываясь о более важных вопросах, связанных с постановкой проблемы.
- Недавно меня спросили, описывал ли кто-нибудь, кроме меня, аналитическую культуру. Я сказал, что не знаю, однако это было до того, как я прочитал четвертую часть книги Фрэнкса. Она связывает аналитическую и инновационную культуру так, как никто прежде этого не делал.

Хотя книга содержит технические сведения, она доступна для широкой аудитории, в том числе для людей с ограниченными техническими

познаниями. Совет Фрэнкса по поводу инструментов для визуализации данных касается всей книги: «Чем проще, тем лучше. Прибегайте к усложнению только в случае крайней необходимости».

Если ваша организация собирается заняться аналитикой — а так и должно быть! — вам придется столкнуться со многими аспектами, затронутыми в этой книге. Даже если вы не специалист в технических вопросах, необходимо ознакомиться с некоторыми темами, связанными с обеспечением аналитических возможностей компании. Если же вы как раз являетесь техническим специалистом, то многое узнаете о человеческом аспекте аналитики. Если вы читаете это предисловие в книжном магазине или просматриваете описание книги на сайте, смело покупайте ее. Если вы ее уже купили, немедленно приступайте к чтению!

Томас Дэвенпорт,
заслуженный профессор информатики и управления,
Бэбсон-колледж.
Сооснователь и директор по исследованиям
Международного института аналитики

Введение

Вы получили электронное письмо: вам предлагают приобрести персонализированную компьютерную систему. Кажется, магазин прочитал ваши мысли, поскольку всего несколько часов назад вы просматривали информацию о компьютерах на его сайте...

Вы отправились в магазин за компьютером, и по пути поступает предложение купить со скидкой кофе в кофейне, мимо которой вы проезжаете: можете получить 10%-ную скидку, если заедете в течение ближайших 20 минут...

Пока пьете кофе, приходит извинение от производителя товара, на качество которого вы пожаловались вчера на своей странице в Facebook, а также на сайте компании...

Наконец, возвращаетесь домой, а вас ждет предложение приобрести специальную броню для вашей любимой онлайн-видеоигры, которая поможет пройти некоторые места, на которых вы застряли...

Звучит неправдоподобно? Думаете, это картины далекого будущего? Нет, эти сценарии возможны уже сегодня! Большие данные. Передовая аналитика. Аналитика больших данных. Кажется, что сегодня уже не обойтись без этих понятий. Люди обсуждают, пишут и продвигают идеи больших данных и передовой аналитики. Теперь к их суждениям можно добавить и эту книгу.

Что реально, а что нет? Уж слишком много внимания к этой теме! Может быть, анализ больших данных — не более чем шумиха? Разговоров на эту тему и правда много, однако эпоха преобразований в сфере аналитических возможностей и эффективного использования больших объемов данных действительно наступила. За ажиотажем, поднятым в средствах массовой информации, стоит нечто очень реальное

и мощное. Шумиха вокруг больших данных объясняется тем, что и предприятия, и потребители взволнованы ожиданием тех преимуществ, которые со временем предоставит анализ больших данных.

Большие данные, в свою очередь, становятся источником новых данных, которые стимулируют аналитические инновации в бизнесе, правительстве и академических кругах. Эти нововведения в состоянии радикально изменить взгляд организаций на свой бизнес. Большие данные обеспечат информацию, которая поможет принимать более взвешенные решения, и в некоторых случаях они будут разительно отличаться от тех, что принимаются сегодня. Анализ больших данных даст такое понимание, о котором сегодня можно только мечтать. Вы увидите, что укрощение волны больших данных и укрощение новых источников данных осуществляется аналогичными способами. Тем не менее дополнительные возможности, которые предоставляют большие данные, требуют использования новейших инструментов, технологий, методов и процессов. Старые способы анализа просто не сработают. Пришло время, когда передовые аналитические методы должны перейти на следующий уровень. Именно этому посвящена книга.

«Укрощение больших данных» не просто название книги. Скорее, это попытка определить, какие предприятия выиграют, а какие проиграют в следующем десятилетии. Подготовившись и взяв на себя инициативу, организации сумеют оседлать волну больших данных, чтобы достичь успеха, вместо того чтобы быть ею раздавленными. Что нужно знать и как подготовиться, чтобы подчинить себе большие данные и извлечь из них ценные новые сведения? Сядьте поудобнее и приготовьтесь это выяснить!

Целевая аудитория

В последние годы появилось бесчисленное количество книг, посвященных передовым методам анализа, а также ряд книг о больших данных. Эта книга подходит к вопросу с иной точки зрения. Основное внимание уделено объяснению, что такое большие данные и как с помощью аналитики их можно использовать, а также рассказать о подходах к созданию и развитию передовой аналитической экосистемы мирового класса в современной среде больших данных. Эта книга адресована широкому кругу читателей. Профессиональный ли вы аналитик,

предприниматель, использующий результаты работы аналитиков, или вам просто интересна тема больших данных — в этой книге вы найдете для себя что-нибудь полезное.

В книге нет подробных технических описаний; технические детали используются лишь в той мере, в какой необходимо обеспечить высокий уровень понимания обсуждаемой темы. Цель — помочь читателям понять и начать применять эти концепции, а также определить области для дальнейшего исследования. Эта книга скорее руководство, чем учебник, и она доступна для читателей, далеких от технических вопросов. В то же время те, кто уже глубоко понимает тему, между строк смогут увидеть технический подтекст.

Обзор содержания

Книга состоит из четырех частей, каждая из которых охватывает один аспект укрощения больших данных. В первой части объясняется, что такое большие данные, каково их значение и способы применения. Вторая часть касается инструментов, технологий и методов, необходимых для анализа и успешного использования больших данных. Третья часть посвящена людям, командам и принципам анализа, которые позволяют обеспечить эффективность. Четвертая часть подводит итог и фокусируется на том, как внедрить передовые методы анализа с помощью центра аналитических инноваций и изменения культуры. Приведем более подробное описание тем каждой части и главы.

Часть I. Появление больших данных

В первой части идет речь о том, что такое большие данные, почему они важны, в чем состоят преимущества их анализа. Описаны десять источников больших данных и то, как эти источники могут быть использованы организациями для улучшения своего бизнеса. Если читатели не знают, что такое большие данные или насколько широко их применение, первая часть даст ответы на эти вопросы.

Глава 1. Что такое «большие данные» и каково их значение? Эта глава начинается с обзора темы больших данных. Затем приводится ряд соображений о том, как организации могут их использовать. Для того чтобы помочь своим организациям справиться с волной больших

данных, читателям следует разобраться в содержимом данной главы так же хорошо, как в остальных главах.

Глава 2. Веб-данные: первые большие данные. Вероятно, наиболее широко используемый и самый известный источник больших данных на сегодняшний день — это данные, собранные с помощью сайтов. Журналы, которые содержат историю посещения пользователями веб-страниц, — настоящая сокровищница информации, которая только и ждет, чтобы ее проанализировали. Организации в целом ряде отраслей уже интегрировали подробные данные о клиентах, полученные с помощью сайтов, в собственную аналитическую среду. В этой главе показано, как эти данные расширяют возможности и изменяют процесс принятия различных бизнес-решений.

Глава 3. Источники больших данных и их ценность. Здесь мы подробно рассмотрим еще девять источников больших данных, чтобы объяснить, что представляет собой каждый источник данных, а также перечислим некоторые способы их применения в бизнесе. Одни и те же базовые технологии могут привести к возникновению нескольких источников больших данных в различных отраслях, а различные отрасли могут воспользоваться преимуществами одних и тех же источников данных. Большие данные имеют очень широкую сферу применения.

Часть II. Укрощение больших данных: технологии, процессы и методы

Часть II посвящена технологиям, процессам и методам, необходимым для укрощения больших данных. За последние годы увеличились возможности масштабируемости этих трех факторов. Организации не могут далее полагаться на устаревшие подходы и желают оставаться конкурентоспособными в мире больших данных. Эта часть книги наиболее «техническая», но все же она доступна для понимания. Читатели познакомятся с рядом концепций, с которыми им предстоит столкнуться в мире анализа больших данных.

Глава 4. Эволюция масштабируемости аналитических систем. Темп роста объема данных всегда предъявлял высокие требования к наиболее масштабируемым из доступных методов анализа. Перед появлением больших данных они уже были близки к своим пределам. Теперь традиционные подходы просто не работают. В этой главе

рассматриваются слияние аналитической среды со средой данных, массивно-параллельные архитектуры, облачные и грид-вычисления, а также модель MapReduce. Каждая из этих парадигм обеспечивает большую масштабируемость и будет играть важную роль в процессе анализа больших объемов данных.

Глава 5. Эволюция аналитических процессов. Значительное увеличение уровня масштабируемости требует обновления аналитических процессов. Глава начинается с описания использования так называемых аналитических песочниц для обеспечения профессиональных аналитиков масштабируемой средой в целях создания передовых аналитических процессов. Далее объясняется, как наборы данных предприятия могут обеспечить бóльшую последовательность и уменьшить риск при создании аналитических данных и одновременном увеличении производительности труда аналитика. В конце главы описывается, как встроенные процессы скоринга позволяют пользователям и приложениям использовать результаты применения передовых аналитических процессов.

Глава 6. Эволюция аналитических инструментов и методов. В этой главе рассматриваются пути развития передовых аналитических инструментов, а также объясняется, как подобные прорывы повлияют на работу профессиональных аналитиков с большими объемами данных. Затрагиваются такие темы, как эволюция визуальных интерфейсов, аналитические точечные решения, инструменты с открытым исходным кодом и инструменты визуализации данных. Рассказывается, как профессиональные аналитики изменили свои подходы к построению моделей для более эффективного использования имеющихся возможностей. Среди описываемых тем: групповое моделирование, экспресс-моделирование и анализ текста.

Часть III. Укрощение больших данных: люди и подходы

Третья часть посвящена людям, которые занимаются анализом, их командам и подходам, используемым для обеспечения высокого качества работы. Наиболее важный фактор при проведении любого анализа, в том числе анализа больших данных, — наличие подходящих людей, которые руководствуются правильными принципами анализа.

Ознакомившись с третьей частью, читатели будут лучше понимать, чем хороший анализ, хороший профессиональный аналитик и хорошая команда аналитиков отличаются от остальных.

Глава 7. Что такое хороший анализ? Подсчет статистики, составление отчета и применение алгоритма моделирования — лишь некоторые из необходимых шагов для обеспечения хорошего анализа. В начале данной главы поясняются отдельные определения, а затем речь идет об обеспечении качественного анализа. Большие данные — довольно сложная тема, поэтому особенно важно понять принципы, излагаемые в этой главе.

Глава 8. Что такое хороший профессионал в области аналитики? Навыки в области математики, статистики и программирования — необходимые, но недостаточные характеристики хорошего профессионального аналитика. Хороший аналитик должен иметь такие качества, как обязательность, творчество, деловая смекалка, навыки проведения презентации и интуиция. В этой главе описано, почему каждая из этих черт имеет большое значение для профессионального аналитика и почему ими не стоит пренебрегать.

Глава 9. Что такое хорошая аналитическая команда? Как организации следует создавать и поддерживать команды аналитиков, чтобы обеспечить оптимальный эффект? Каким образом команды вписываются в организацию? Как они должны работать? Кто должен отвечать за создание передовой аналитики? Здесь затронуты часто встречающиеся проблемы и изложены принципы, которые необходимо иметь в виду при создании аналитической команды.

Часть IV. Объединение пройденного: аналитическая культура

В четвертой части изложены хорошо известные базовые принципы, которым должна следовать организация, чтобы успешно внедрять инновации, используя передовые средства анализа и большие данные. Поскольку это фундамент многих дисциплин, внимание сосредоточено на том, какое отношение данные принципы имеют к передовой аналитике в современной корпоративной среде. Описываемые концепции, вероятно, знакомы читателям в отличие от способов их применения к области передовой аналитики и больших данных.

Глава 10. Создание условий для внедрения инноваций в сфере аналитики. Глава начинается с обзора некоторых принципов, лежащих в основе успешного внедрения инноваций. Далее объясняется, как они применяются в мире больших данных и передовой аналитики, с помощью концепции центра аналитических инноваций. Цель состоит в том, чтобы показать читателям, как можно обеспечить внедрение аналитических инноваций и укрощение больших данных в своих организациях.

Глава 11. Создание культуры инноваций и открытий. Глава посвящена созданию культуры инноваций и открытий. Она написана легко и непринужденно и дает пищу для размышлений о том, что требуется для создания культуры, способной к инновационному анализу. Изложенные в главе принципы хорошо известны. Тем не менее их стоит еще раз проанализировать, а затем подумать о том, как их применить к большим данным и передовой аналитике.