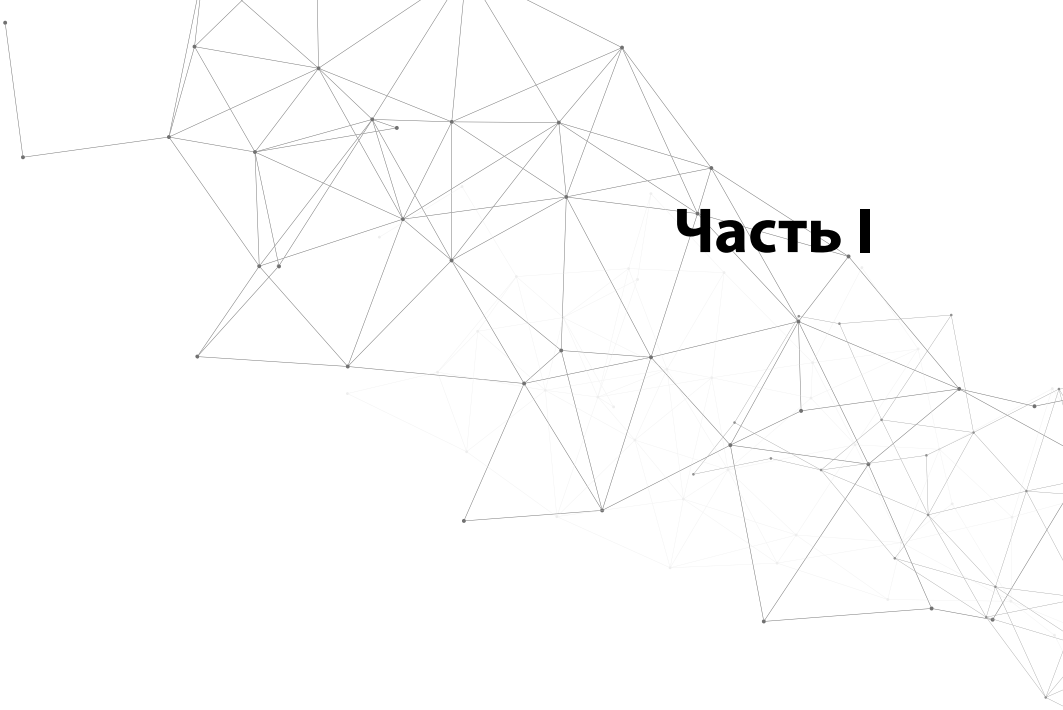
Краткое содержание

Часть I
Основы больших данных

Часть I I
Хранение и анализ больших данных



Часть I

Основы больших данных

Большие данные могут изменять характер бизнеса. Фактически существует множество компаний, деятельность которых основывается на их способности генерировать идеи, которые могут предоставить только большие данные. В первых главах этой книги рассматриваются основы больших данных, в первую очередь с точки зрения бизнеса. Предприятиям необходимо понять, что большие данные — это не только технология, но и способ продвижения организации с помощью технологий.

Часть I имеет следующую структуру:

- Глава 1 дает представление о ключевых понятиях и терминологии, которые определяют саму суть больших данных, а также о потенциале, который они содержат, чтобы оправдать ожидания при решении сложных бизнес-задач. Объясняются разнообразные характеристики, которые позволяют различать наборы больших данных, а также определяются разные типы данных и технологии их анализа.
- Глава 2 ориентирована на то, чтобы ответить на вопрос: почему компании должны быть мотивированы к внедрению систем обработки больших данных вследствие фундаментальных изменений на рынке и в мире бизнеса? Большие данные не являются технологией, связанной с трансформацией бизнеса; однако они позволяют вне-

дрить инновации внутри предприятия при условии, что последнее действует на основе полученных знаний.

- В главе 3 говорится о том, что большие данные — это не просто «традиционный бизнес», и что решение об использовании больших данных должно учитывать рассмотрение множества факторов бизнеса и технологий. Это подчеркивает тот факт, что большие данные открывают для предприятия факторы воздействия внешних данных, которые должны регулироваться и управляться. Так же, как и жизненный цикл аналитики больших данных устанавливает особые требования к обработке данных.

- В главе 4 рассматриваются современные подходы к хранилищу корпоративных данных и Business Intelligence. Затем эта тема детализируется, чтобы показать, что хранение больших данных и ресурсы для их анализа могут использоваться в связке с инструментами корпоративного мониторинга производительности для увеличения аналитических возможностей предприятия и углубления выводов, полученных при обработке данных, предоставляемых Business Intelligence.

Правильно используемые большие данные являются частью стратегической инициативы, основанной на предположении о том, что внутренние данные (полученные внутри бизнеса) не дают всех ответов. Другими словами, большие данные — это не просто проблемы управления данными, которые можно решить с помощью технологии. Речь идет о бизнес-проблемах, решения которых обеспечиваются технологиями, которые могут поддерживать анализ наборов больших данных. По этой причине бизнес-ориентированные рассуждения в части I создают основу для тематики, ориентированной на технологии и описанной в части II.

Глава 1

Понимание больших данных

Концепты и терминология

Характеристики больших данных

Различные типы данных

Введение в исследование конкретных случаев из практики

Большие данные представляют собой сферу деятельности, направленную на анализ, обработку и хранение больших коллекций данных, которые часто генерируются разрозненными источниками. Решения и методы для больших данных становятся востребованными тогда, когда традиционные методы анализа данных, технологии их обработки и хранения оказываются недостаточными. Если более конкретно, то большим данным присущи специфические особенности, такие как объединение разнообразных несвязанных наборов данных, обработка больших объемов неструктурированных данных и выявление скрытой информации в оперативном режиме, без задержек.

Несмотря на то что большие данные появились как новая дисциплина, изначально они развивались в течение многих лет. Управление большими массивами данных и их анализ были многолетней проблемой — начиная от трудоемких подходов к ранним переписям населения до актуарной науки по вычислению страховых взносов. Эти два направления стали причиной развития науки о больших данных.

В дополнение к традиционным аналитическим подходам, основанным на статистике, большие данные добавляют новые методы, усиленные вычислительными ресурсами и подходами к выполнению аналитических алгоритмов. Такие изменения очень важны, поскольку наборы данных продолжают увеличиваться в размерах, становясь все более разнообразными, сложными и ориентированными на не-

прерывный поток. В то время как статистические подходы использовались для приблизительных показателей численности населения с помощью выборки, как в библейские времена, достижения в области вычислительной науки позволили обрабатывать наборы данных целиком, делая ненужными такие выборки.

Анализ больших наборов данных является междисциплинарной задачей, которая сочетает в себе математику, статистику, компьютерные науки и специальные знания предметной области. Такая смесь навыков и точек зрения привела к некоторой путанице относительно того, что включает в себя область больших данных и их анализ. Следовательно, ответ будет зависеть от точки зрения того, кто отвечает на этот вопрос. Границы того, что представляет собой проблема больших данных, также варьируются из-за изменяющихся и совершенствующихся перспектив программного обеспечения и технологий аппаратного обеспечения. Это связано с тем, что определение больших данных учитывает влияние их характеристик на архитектуру самой среды обработки. Тридцать лет назад один гигабайт данных мог бы быть проблемой больших данных и требовать специальных целенаправленных вычислительных ресурсов. Теперь гигабайты данных — обычное явление, они могут быть легко переданы, обработаны и сохранены на клиенториентированных устройствах.

Данные внутри сред больших данных, как правило, накапливаются благодаря сбору данных внутри предприятия с помощью приложений, датчиков и внешних источников. Данные, обработанные с помощью методов решений для больших данных, могут использоваться непосредственно корпоративными приложениями или переправляться в хранилище для улучшения существующих данных.

Результаты, полученные при обработке больших данных, могут привести к широкому спектру выгод и преимуществ, таких как:

- операционная оптимизация

- новые знания для выполнения поставленных заданий
- выявление новых рынков
- точные прогнозы
- обнаружение неисправностей и мошенничества
- более детализированные факты
- совершенствование процесса принятия решений
- научные открытия

Несомненно, применимость и потенциальные выгоды от больших данных очень широкие. Тем не менее, здесь существует множество проблем, которые необходимо учитывать при выборе подходов аналитики больших данных. Эти проблемы должны быть понятны и взвешены с учетом ожидаемых преимуществ, чтобы получить обоснованные решения и планы. Более подробно эти темы рассматриваются во второй части книги.

Концепты и терминология

В качестве отправной точки мы должны определить и понять некоторые фундаментальные концепты и термины.

Наборы данных

Коллекции или группы связанных данных обычно называются наборами данных. Каждую группу или элемент группы (единицу данных) разделяет один и тот же набор атрибутов или свойств, как и все остальные в этом же наборе данных. Некоторыми примерами наборов данных служат:

- твиты, сохраненные в файле
- коллекция графических файлов в каталоге
- строки, извлечение из таблицы базы данных, сохраненные в файле формата CSV

- исторические погодные наблюдения, сохраненные как XML-файлы

На рисунке 1.1 показаны три набора данных, основанные на трех различных форматах данных.

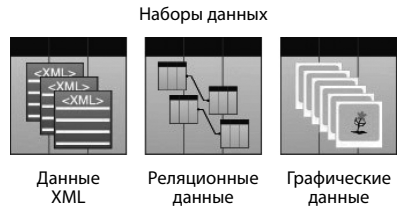


Рисунок 1.1
Наборы данных могут находиться в различных форматах.

Анализ данных

Анализ данных представляет собой процесс исследования данных для поиска фактов, отношений, шаблонов, идей и/или тенденций. Общая цель анализа данных заключается в том, чтобы поддерживать принятие более обоснованных решений. Простым примером анализа данных является анализ данных по продаже мороженого с целью определить, каким образом количество проданных рожков связано со среднесуточной температурой воздуха. Результаты такого анализа могли бы стать обоснованием для решения о количестве мороженого, которое магазин должен заказать в зависимости от прогноза погоды. Выполнение анализа данных помогает установить закономерности и взаимосвязи между анализируемыми данными. На рисунке 1.2 показано изображение символа, обозначающего процесс анализа данных.

Рисунок 1.2
Символ, который используется для обозначения анализа данных.

Аналитика данных

Аналитика данных — это более широкий термин, который охватывает и сам анализ данных. Аналитика дан-

ных является дисциплиной, которая охватывает управление полным жизненным циклом данных, а именно сбор, очистку, организацию, хранение, анализ и регулирование данных. Термин включает в себя разработку методов анализа, научных методик, а также автоматизированных инструментов. В средах больших данных для аналитики были разработаны методы, позволяющие проводить анализ с помощью хорошо масштабируемых распределенных технологий и механизмов, способных к анализу больших объемов данных, поступающих из различных источников. На рисунке 1.3 показано изображение символа, используемого для обозначения аналитики данных.

Рисунок 1.3
Символ, который используется для обозначения аналитики данных.

Обычно жизненный цикл аналитики больших данных включает идентификацию, доставку, подготовку и анализ больших объемов «сырых», неструктурированных данных, с целью извлечения полезной и значимой информации, которая может служить в качестве входных данных для определения моделей, улучшения существующих корпоративных данных и выполнения крупномасштабных поисковых запросов.

Различные виды организаций по-разному используют инструменты и методы аналитики данных. Например, возьмем такие три сектора:

- В средах, ориентированных на бизнес, результаты аналитики данных могут снизить эксплуатационные расходы и облегчить принятие стратегических решений.
- В научной сфере аналитика данных может помочь определить причину явления для улучшения точности прогнозов.
- В средах, базирующихся на сервисах, таких как организации общественного сектора, аналитика данных может помочь усилить акцент на предоставлении высококачественных услуг при низких затратах.

Аналитика данных дает возможность принимать решения на основе данных с привлечением научной поддержки таким образом, чтобы эти решения могли основываться на фактических данных, а не на прошлом опыте или одной только интуиции.

Существуют четыре основные категории аналитики, которые различаются производимыми результатами:

- дескриптивная аналитика
- диагностическая аналитика
- прогностическая аналитика
- прескриптивная аналитика

Различные типы аналитики усиливают разные методы и алгоритмы анализа. Под этим подразумевается, что могут существовать различные требования к данным, их хранению и обработке в целях облегчения предоставления множества разных типов аналитических результатов. Рисунок 1.4 иллюстрирует реальную картину, когда генерирование высокозначимых аналитических результатов увеличивает сложность и стоимость аналитической среды.

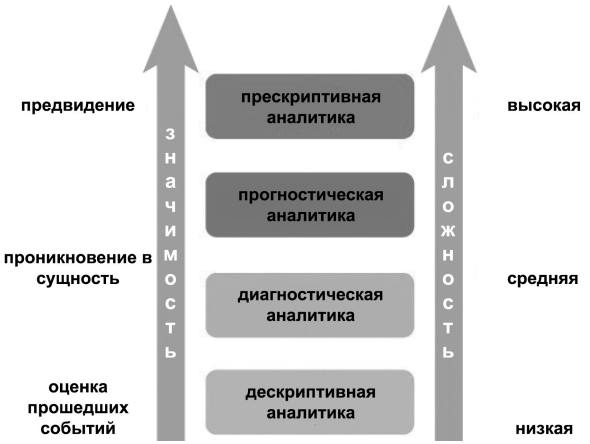


Рисунок 1.4
Важность и сложность увеличиваются от дескриптивной до прескриптивной аналитики.

Дескриптивная аналитика

Дескриптивная аналитика используется для поиска ответов на вопросы о событиях, которые уже произошли. Эта форма аналитики согласовывает данные с контекстом для генерирования информации.

Приведем примеры вопросов:

- Каким был объем продаж за последние 12 месяцев?
- Сколько звонков поступило в службу поддержки, упорядоченных по категориям серьезности и географическому расположению?
- Какова ежемесячная комиссия, заработанная каждым агентом по продажам?

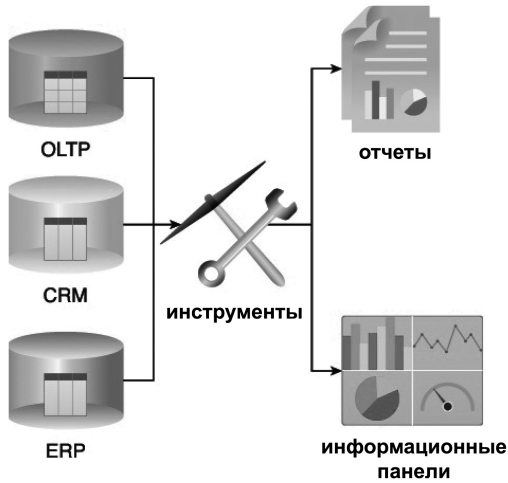
По оценкам, 80 % сгенерированных результатов аналитики являются дескриптивными по своей природе. Ориентированная на смысловую полезность, дескриптивная аналитика обеспечивается с наименьшими затратами и требует относительного базового набора навыков.

Дескриптивная аналитика часто выполняется с помощью специальных отчетов или информационных панелей, как показано на рисунке 1.5. Отчеты обычно статичны по своей природе и отображают исторические данные, которые представлены в форме таблиц или диаграмм. Запросы выполняются в рабочих хранилищах данных внутри предприятия, например, в системе управления взаимоотношениями с клиентами (Customer Relationship Management — CRM) или планирования ресурсов предприятия (Enterprise Resource Planning — ERP).

Диагностическая аналитика

Диагностическая аналитика направлена на то, чтобы определить причину произошедшего события, используя вопросы, которые фокусируются на причинах этого события. Цель

Рисунок 1.5
Рабочие системы, изображенные слева, задействуются с помощью инструментов дескриптивной аналитики для создания отчетов или информационных панелей, изображенных справа.



этого типа аналитики — определить, какая информация относится к данному явлению, чтобы дать возможность ответить на вопросы о том, почему это произошло.

К таким вопросам относятся:

- Почему продажи Q2 были меньше, чем продажи Q1?
- Почему в службу поддержки поступило больше звонков из восточного региона, чем из западного?
- Почему увеличилось число повторных госпитализаций пациентов за последние три месяца?

Диагностическая аналитика дает большую значимость, чем дескриптивная, но требует более продвинутого набора навыков. Обычно диагностическая аналитика требует сбора данных из различных источников и хранения их в структуре, которая подвергается детальному анализу и свертыванию, как показано на рисунке 1.6.

Результаты диагностической аналитики могут быть просмотрены с помощью инструментов интерактивной визуализации, которые позволяют пользователям определять тенден-

ции и шаблоны. Выполнение запросов здесь значительно сложнее, чем в дескриптивной аналитике, и выполняются они на многомерных данных, содержащихся в системах аналитической обработки.

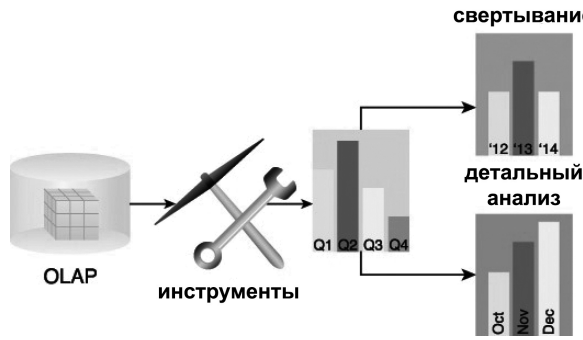


Рисунок 1.6
Диагностическая аналитика может обеспечить данные, которые подходят для их более детального изучения и свертывания.

Прогностическая аналитика

Прогностическая аналитика проводится с целью определить результат события, которое может произойти в будущем. С помощью прогностической аналитики информация усиливается смысловым содержанием. Интенсивность и значимость ассоциативных связей формируют основу моделей, которые используются для создания будущих прогнозов на основе прошлых событий.

Важно учитывать, что у моделей, которые используются для прогностической аналитики, существуют неявные зависимости от условий, в которых происходили прошлые события. Если лежащие в основе причины изменяются, то и модели прогнозирования должны быть откорректированы.

Вопросы обычно формулируются с использованием обоснования «Что, если…», например:

- Какова вероятность невозвращения клиентом кредита, если пропущен ежемесячный платеж?

- Каков процент эффективности лечения пациента, если вместо препарата A будет использоваться препарат B?
- Если клиент приобрел продукты A и B, каковы шансы, что он также купит продукт C?

Прогностическая аналитика призвана предсказать исход события, при этом прогнозы делаются на основе шаблонов, тенденций и исключений, найденных в исторических и текущих данных. Это может привести к выявлению как рисков, так и возможностей.

Такой вид аналитики предполагает использование больших наборов данных, внутренних и внешних, и различных методов анализа этих данных. Эта аналитика обеспечивает высокую значимость результатов и требует еще более совершенного набора навыков, чем дескриптивная и прогностическая. Как правило, инструменты прогностической аналитики используют лежащие в их основе абстрактные способы решения статистических сложных и запутанных тонкостей, предоставляя удобные для пользователя внешние интерфейсы, как показано на рисунке 1.7.

Рисунок 1.7
Инструменты прогностической аналитики могут предоставить удобные для пользователя внешние интерфейсы.

Прескриптивная аналитика

Прескриптивная аналитика основывается на результатах прогностической аналитики, предписывая меры, которые должны быть предприняты. Акцент делается не только на том, какому предписанному варианту лучше всего следовать, но и почему. Другими словами, прескриптивная аналитика дает результаты, на основе которых можно делать выводы, поскольку они встраивают элементы ситуативного понимания.

Таким образом, этот вид аналитики может быть использован для получения преимуществ или снижения риска.

Вопросы могут иметь следующий вид:

- Какой из трех препаратов обеспечивает наилучшие результаты?
- Когда наилучше время для проведения определенной акции?

Прескриптивная аналитика представляет большую значимость, чем любые другие виды аналитики, и, соответственно, требует самого продвинутого набора навыков, а также специализированного программного обеспечения и инструментов. Благодаря ей рассчитываются различные результаты и предлагается оптимальный курс действий для каждого из них. Такая тактика переходит от пояснений к консультациям и может включать в себя моделирование различных сценариев.

Этот вид аналитики охватывает внутренние данные одновременно с внешними. Внутренние данные могут включать в себя текущие и исторические данные о продажах, информацию о клиентах, данные о продуктах и бизнес-правила. Внешние же могут содержать данные из социальных сетей, прогнозы погоды и демографические данные, подготовленные правительством. Прескриптивная аналитика предполагает использование бизнес-правил и больших объемов внутренних и внешних данных для моделирования результатов и прописывает оптимальный план действий, как показано на рисунке 1.8.

Рисунок 1.8
Прескриптивная аналитика включает в себя использование бизнес-правил и внутренних и/или внешних данных для проведения глубокого анализа.

Business Intelligence (BI)

BI позволяет организации получить представление о производительности предприятия путем анализа данных, созданных бизнес-процессами и информационными системами. Результаты анализа могут использоваться руководством для управления бизнесом с целью устранения выявленных проблем или повышения эффективности работы организации. BI применяет аналитику к большим объемам данных в масштабах всего предприятия, которые обычно консолидируются в хранилище корпоративных данных для выполнения аналитических запросов.

Как показано на рисунке 1.9, результат BI может отображаться на информационной панели, которая позволяет менеджерам получать и анализировать результаты и, возможно, уточнять аналитические запросы для дальнейшего исследования данных.

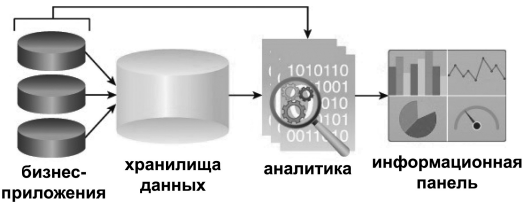


Рисунок 1.9
BI можно использовать для усовершенствования бизнес-приложений, консолидации данных в хранилищах данных и анализа запросов с помощью информационной панели.

Ключевые показатели эффективности (KPI)

KPI — это показатель, который может использоваться для оценивания успеха в конкретном бизнес-контексте. KPI связаны с общими стратегическими целями и задачами предприятия. Они часто используются для выявления проблем эффективности бизнеса и демонстрации соответствия нормативным требованиям. Таким образом, KPI выступают в качестве количественных ориентиров для измерения конкретного аспекта общей эффективности бизнеса. KPI часто отображаются посредством информационной панели KPI, как показано на рисунке 1.10. Информационная панель объединяет отображение нескольких KPI и сравнивает фактические измерения с пороговыми значениями, по которым определяется допустимый диапазон значений KPI.

Рисунок 1.10
Информационная панель KPI выступает в качестве центрального ориентира для оценивания эффективности бизнеса.

Характеристики больших данных

Для того чтобы набор данных можно было считать большими данными, он должен обладать одной или несколькими характеристиками, которые обеспечивают адаптацию его к проектному решению и архитектуре аналитической среды. Большинство этих характеристик данных были первоначально определены Дугом Лейни в начале 2001 года, когда он опубликовал свою статью, описывающую влияние объема, скорости и многообразия данных электронной коммерции на хранилища данных предприятия. К этому списку была добавлена достоверность для расчета более низкого отношения сигнал/шум неструктурированных данных по сравнению со структурированными источниками данных.

В конечном счете, цель заключается в проведении анализа данных таким образом, чтобы высококачественные результаты предоставлялись своевременно, что обеспечит оптимальную значимость для предприятия.

В этом разделе рассматриваются пять характеристик больших данных, которые можно использовать для дифференциации данных, классифицированных как «большие», из других форм. Пять свойств больших данных, показанных на рисунке 1.11, которые обычно называют как «Пять V»:

- объем
- скорость
- многообразие
- достоверность
- ценность

Рисунок 1.11
«Пять V» больших данных.

Объем

Предполагаемый объем данных, который обрабатывается решениями для больших данных, является существенным и постоянно растущим. Большие объемы данных накладывают различные требования по хранению и обработке данных, а также к дополнительной подготовке данных, к процессам сопровождения и управления. На рисунке 1.12 визуально представлен большой объем данных, которые ежедневно создаются организациями и пользователями по всему миру.

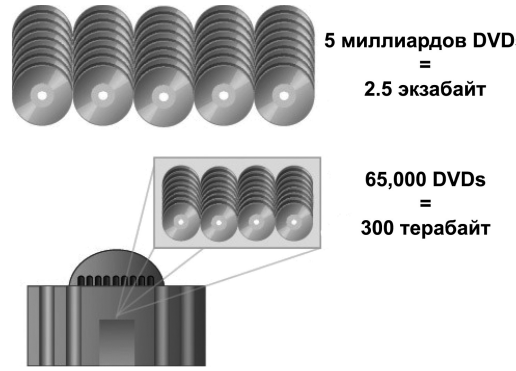


Рисунок 1.12
Организации и пользователи по всему миру создают более 2,5 EB (экзабайтов) данных в день. В качестве сравнения: в настоящее время библиотека Конгресса США содержит более 300 TB (терабайтов) данных.

Типичные источники данных, которые отвечают за генерацию больших объемов данных, могут включать в себя:

- онлайн-транзакции, такие как розничные точки продаж и банкинг
- научные и исследовательские эксперименты, такие как большой адронный коллайдер и атакамский большой антенный телескоп миллиметрового/субмиллиметрового диапазона
- сенсоры, такие как GPS-сенсоры, RFID, смарт-счетчики и телематика
- социальные сети, такие как Facebook и Twitter

Скорость

В средах с большими данными последние могут поступать с высокими скоростями, и при этом огромные массивы данных могут накапливаться за очень короткие промежутки времени. С точки зрения предприятия скорость передачи данных преобразуется в количество времени, которое требуется для обработки данных после их поступления в пределы предприятия. Решение проблемы быстрого притока данных требует от предприятия разработки гибких и доступных решений обработки данных и соответствующих условий для их хранения.

В зависимости от источника данных скорость может быть не всегда высокой. Например, изображения сканирования МРТ генерируются не так часто, как записи в журнале веб-сервера с высоким трафиком. Как показано на рисунке 1.13, скорость передачи данных рассматривается в перспективе, если учитывать, что за минуту могут быть легко созданы следующие объемы данных: 350 тысяч твитов, 300 часов видеоматериалов, загруженных на YouTube, 171 миллион электронных писем и 330 гигабайт данных от сенсоров реактивного двигателя.

Рисунок 1.13
Примеры высокоскоростных наборов больших данных, производимых каждую минуту, включают в себя твиты, видеоматериалы, электронные письма и гигабайты данных, генерируемых реактивным двигателем.