Лекция 2. Жизненный цикл аналитики данных. ETL-процессы. Краудсорсинг

Большие данные мало просто собрать, их нужно как-то использовать, например, чтобы строить прогнозы развития бизнеса или проверять маркетинговые гипотезы. А для использования данные требуется структурировать и анализировать.

Что же такое **аналитика данных** – это процесс преобразования первичных данных в знания и полезную информацию, которую можно использовать для принятия решений, что позволит оптимизировать деятельность: увеличить доход, сократить затраты или достичь других важных результатов.

Технологии для аналитики больших данных:

1. **Предиктивная аналитика** – предсказание будущего на основе собранных данных.
2. **Краудсорсинг** – ручной анализ силами большого количества людей.
3. **Смешение и интеграция данных** – приведение данных из разных источников к одному виду, уточнение и дополнение данных.
4. **Распознавание образов** – отнесение исходных данных к определенному классу с помощью выделения существенных признаков, характеризующих эти данные, из общей массы данных.
5. **Машинное обучение и нейронные сети** – создание программ, которые умеют анализировать и принимать решения, выстраивая логические связи.
6. **Имитационное моделирование** – построение моделей на основе больших данных, которые помогают провести эксперимент в компьютерной реальности, без влияния на реальное положение вещей.
7. **Статистический анализ** – подсчет данных по формулам и выявление в них тенденций, сходств и закономерностей.
8. **Пространственный анализ** – набор алгоритмов (функций), обеспечивающих анализ местоположения (размещения), связей и иных пространственных отношений пространственных объектов, включая анализ зон видимости/невидимости, анализ соседства, анализ сетей, создание и обработку цифровых моделей рельефа и т.д.
9. **Data Mining** – технология добычи новой значимой информации из большого объема данных.

10. **Визуализация** – представление больших данных и результатов их анализа в виде удобных графиков и схем, понятных человеку.

Приведем жизненный пример того, как и при каких условиях необходимо производить анализ данных.

Представим, для аналитика стоит задача оценить продажи ряда магазинов (куда поставляется товар) и каждый магазин ведет свой учет проданных товаров. Реальность такова, что формы учета будут заполнены разными способами, то есть у них будет разная структура и разный формат хранения (некоторая форма таблиц).

Решим эту задачу способом, который первый приходит в голову, то есть предоставим простейший способ решения данной задачи. Пусть у нас есть N таблиц и нам нужно их собрать вместе в одну таблицу. Напишем N скриптов, которые преобразуют эти таблицы, и один сборщик, который собирает их вместе. У такого подхода есть положительные и отрицательные стороны.

К **минусам** можно отнести:

- необходимо поддерживать N скриптов одновременно (где N в порядках тысяч);
- при изменении структуры отчетов магазинов во времени (например, в магазине появился новый сотрудник) необходимо искать и переписывать отдельные скрипты;
- при появлении нового магазина, необходимо писать новый скрипт;
- при изменении формата отчетности необходимо вносить изменения во все скрипты;
- сложная отладка и поддержка, так как магазины не уведомляют об изменении структуры и не следуют никаким спецификациям.

Немного усложним задачу и приблизим ее к реалиям современных компаний. Итак, производитель товаров на самом деле работает не напрямую с магазинами, а через некоторых посредников. Посредники посещают магазины и непосредственно своими действиями пытаются стимулировать продажи. Соответственно, они являются материально заинтересованными лицами и информацию, которую они выдают, приходится перепроверять.

На основании вышесказанного давайте переформулируем задачу. Пусть у нас есть N магазинов и K дистрибьюторов. Можем ли мы агрегировать данные магазинов и сравнить их с результатами дистрибьюторов? При этом у всех данные имеют разную структуру и формат.

Эта задача обладает несколькими основными характеристиками. Входные данные представляют собой огромное количество разнообразных форматов, к которым добавляются отчеты дистрибьюторов. Как правило,

задача характеризуется очень низким качеством данных, в том числе дублированием, несогласованностью и ошибками. На основе полученных результатов и сравнения данных отдел по закупкам принимает решения о том, сколько, что, кому и по какой цене отгружать. То есть решение этой задачи непосредственно влияет на финансовые показатели компании, что безусловно важно.

Существуют также другие варианты решения подобной задачи. Например, самописное решение. Компании производителю будет необходимо нанять специалиста не по профилю компании и критичное ПО будет зависеть от данного специалиста. Если он уйдет, то компания будет вынуждена срочно искать замену, которая сможет поддерживать ПО, и качество будет напрямую зависеть от нанятого специалиста.

Также всегда есть возможность закупить ПО у третьей стороны. Отрицательной стороной такого подхода являются цена, качество и время интеграции. Как правило, цена и время интеграции слишком высоки для среднего производителя и в том числе требует существенных временных затрат сотрудников. Выбор поставщика также не тривиален.

Есть вариант применить SaaA решения, но методология еще нова для рынка и многие компании скептически относятся к подобным сервисам.

Процесс анализа на уровне компании в общем случае выглядит следующим образом: данные консолидируются, определенным образом трансформируются (агрегируются) и загружаются в систему для анализа.

Business Intelligence

Термин Business Intelligence был введен аналитиками Gartner как «процесс, ориентированный на бизнес-пользователя и включающий доступ и исследование информации, ее анализ, выработку интуиции и понимания, которые ведут к улучшенному и неформальному принятию решений».

Иными словами, можно сказать, что BI - это методы и инструменты для перевода необработанной информации в осмысленную, удобную форму. При этом технологии BI обрабатывают большие объемы неструктурированных данных, чтобы найти стратегические возможности для бизнеса.

Business Intelligence имеет отношение к процессу превращения данных в знания, а знаний - в действия бизнеса для получения выгоды; является деятельностью конечного пользователя, которую облегчают различные аналитические и групповые инструменты и приложения, а также инфраструктура хранилища данных.

Главная цель BI состоит в том, чтобы интерпретировать большое

количество данных, заостряя внимание лишь на ключевых факторах эффективности, моделируя исход различных вариантов действий, отслеживая результаты принятия решений.

BI-платформа

Для анализа данных на сегодняшний день существует множество решений. BI-платформа - это такая программа, которая предоставляет пользователю удобные инструменты анализа фактически любых данных (будь то файл Excel либо промышленное хранилище данных). Проще говоря, BI- платформа - это продвинутая система отчетности.

Современные системы класса BI располагают следующими основными функциями: возможность подключения к различным источникам данных (от файла Excel до универсального ODBC подключения); возможность построения как простых отчетов (типа график или таблица), так и сложных параметризированных отчетов с комбинированной структурой и ссылочными связями Drill-Trough, Drill-Up/Drill-Down); возможность прозрачной работы с разными источниками данных (например, Excel и SQL Server) с полноценной обработкой связей между ними; возможность интерактивной работы с данными (формирование отчетов «на лету»); возможность представления реляционных данных как многомерные; возможность распределения прав доступа, используя как внутренние источники аутентификации, так и внешние NTLM, LDAP и т. д.); возможность запуска формирования отчетов как вручную, так и автоматически по расписанию; возможность автоматической рассылки сформированных отчетов; возможность построения отчетов в различных форматах (Excel, HTML, PDF и т. д.).

BI состоит из множества компонентов и обладает различными характеристиками: многомерная агрегация и размещение данных в хранилищах; денормализация баз данных; маркировка и стандартизация данных (ETL); отчетность в режиме реального времени с аналитическими оповещениями; статистические выводы и вероятностное моделирование.

BI должна быть разделена на этапы: информационный поиск, аналитическая обработка в реальном времени (OLAP), предупреждения об отклонениях от ожидаемых показателей, бизнес-аналитика, бизнес-отчётность.

Разница между Business Intelligence и аналитикой заключается в том, что последняя включает в себя комплекс методов для анализа уже чистых данных. В отличии от аналитики, основная задача BI - принятие решений для бизнеса.

Средства BI

В соответствии с подходом аналитиков выделяют три основных типа инструментальных средств BI: средства предоставления информации, средства интеграции, средства анализа.

Средства предоставления информации подразделяются на средства создания отчетов, которые дают возможность создавать форматированные интерактивные отчеты, и на информационные панели показателей, являющиеся одной из составных частей отчетов, представляющие информацию в виде интуитивно понятного графического изображения, включая диаграммы, круговые шкалы, светофоры и т.п. Также существуют такие инструменты, как генераторы нерегламентированных запросов - данная функция, известная также как создание отчетов в режиме самообслуживания, дает пользователям возможность получать ответы на возникающие вопросы.

Наряду со средствами создания отдельных BI-приложений, BI-платформа должна предоставлять средства программной разработки для интеграции приложений в общий бизнес-процесс или обеспечивать их встраивание в другое приложение. BI - платформа должна давать разработчикам возможность создания BI-приложений без кодирования, на основе применения мастеров для визуального редактирования.

OLAP

Одним из средств анализа данных является технология OLAP (Online Analytical Processing — Оперативная аналитическая обработка данных). Это класс приложений и технологий, предназначенных для сбора, хранения и анализа многомерных данных в целях поддержки принятия управленческих решений. Технология OLAP позволяет аналитикам, менеджерам и управляющим сформировать свое собственное видение данных, используя быстрый, единообразный, оперативный доступ к разнообразным формам представления информации.

Одновременный анализ по нескольким измерениям определяется как многомерный анализ. Каждое измерение включает направления консолидации данных, состоящие из последовательных уровней обобщения, где каждый вышестоящий уровень соответствует большей степени агрегации данных по соответствующему измерению.

OLAP представляет собой инструмент для анализа больших объемов данных в режиме реального времени и обеспечивает следующие возможности работы с многомерными данными: гибкий просмотр информации, произвольные срезы данных, детализация, свертка или

консолидация, вращение, сравнение во времени.

Продвинутая визуализация

Инструменты продвинутой визуализации позволяют представлять данные для более эффективного их восприятия посредством использования интерактивных картинок и диаграмм вместо таблиц.

Обычно пользователи в динамическом режиме могут менять графическое представление, использовать масштабирование, объединять данные, изменять цвета.

Предиктивное моделирование и Data Mining

Предиктивное моделирование - это процесс создания (или выбора) модели для предсказания вероятности наступления некоторого события.

Интеллектуальный анализ данных Data Mining - компьютерная техника извлечения знаний, которая использует искусственный интеллект для распознавания образов и выделения значимых закономерностей из данных, находящихся в хранилищах или во входных / выходных потоках.

Эти методы основываются на статистическом моделировании, нейронных сетях, генетических алгоритмах и др.

Частная методология Text Mining решает задачи навигации в больших текстовых массивах, поиск взаимосвязей между ключевыми понятиями текстов, структуризация хранилищ документов, поиск информации, выраженный на естественном языке, распределение по рубрикам. Информация, найденная в процессе использования методов Data Mining, должна описывать новые связи между свойствами, предсказывать значения одних признаков на основе других и т.д. Найденные знания должны быть применимы и по отношению к новым данным с некоторой степенью достоверности. Когда извлеченные знания непрозрачны для пользователя, должны существовать методы постобработки, позволяющие привести их к интерпретируемому виду.

Задачи, решаемые методами Data Mining, включают: классификацию - отнесение объектов (наблюдений, событий) к одному из заранее известных классов; регрессию, в том числе задачи прогнозирования; установление зависимости непрерывных выходных от входных переменных; кластеризацию - группировку объектов (наблюдений, событий) на основе данных (свойств), описывающих сущность этих объектов; ассоциацию - выявление закономерностей между связанными событиями; последовательные шаблоны - установление закономерностей между связанными во времени событиями, то есть обнаружение зависимости, согласно которой если произойдет событие X, то спустя заданное время

произойдет событие У; анализ отклонений - выявление наиболее нехарактерных шаблонов.

Инструменты анализа

Данных становится всё больше и больше, поэтому сейчас как никогда важно иметь необходимый инструментарий для анализа данных и принятия решений. Выделим четыре популярных аналитических системы: MS Excel Power Query, MS Power BI, Pyramid Analytics, Компоненты аналитики MS SQL server (MDS, SSIS, SSAS).

Power Query

Power Query позволяет искать и открывать данные из различных источников доступных онлайн и через корпоративные сети. Он умеет загружать в Excel данные разных типов, форматов и структур, а также из совершенно разных источников: из сети, из файла, из баз данных, из публичных источников данных и корпоративных репозиториев данных (встроена поддержка ETL), из ряда других источников: SharePoint List, OData feed, Active Directory, Facebook и др.

Power Query обладает как положительными характеристиками, так и отрицательными. Из плюсов можно выделить: работает как с табличными моделями, так и с многомерными, умеет подключать дополнительные источники. Минусы данной программы: сложен в освоении, достаточно медлителен, нет возможности разделения доступа, ограничения на размер файлов/записей.

MS Power BI

Power BI - это инструмент создания интерактивных бизнес отчетов с возможностью совместной работы, визуализации и интерактивной работы.

Power BI обладает возможностью быстрой разработки информативных бизнес отчетов и панелей (в сети) - с возможностью взаимодействия и исследования данных. Также поддерживается автоматическое обновление BI- отчетов и визуализации при изменении данных, поддержка языка запросов, в том числе и Power Query, создание каталога данных с индексами для поиска, язык запросов близкий к естественному (для бизнес-аналитика) и возможность интерактивной работы. Поддерживается работа с мобильными устройствами.

Power BI является новым современным продуктом с дружелюбным интерфейсом и легким в освоении. Из недостатков можно выделить то, что

решение «сырое» (некоторые компоненты могут работать нестабильно), оно не работает с OLAP кубами и имеет урезанный функционал в сравнении с конкурентами.

Pyramid Analytics

Pyramid Analytics - облачная платформа бизнес-аналитики, имеющая три ключевых компонента: интеллектуальный анализ данных, интерактивная работа с данными и визуализацией, представление данных аудитории.

Платформа обладает возможностью совместной аналитики и моделирования данных, а также рядом других полезных возможностей: интеграция с R, работа с Big Data, интерактивная визуализация данных.

Продукт легок в освоении, работает с огромным количеством источников и имеет очень широкую функциональность. Главный и единственный недостаток - высокая цена.

Компоненты аналитики MS SQL server (MDS, SSIS, SSAS)

SQL Сервер позволяет проводить анализ внутри своей экосистемы. У него есть обширный набор компонент, сфокусируемся на трех наиболее известных.

Master Data Services — процессы и инструменты управления мастер-данными компании. Мастер-данные - это данные бизнеса: о клиентах, продуктах, услугах, персонале, технологиях, материалах и т.д.

SQL Server Integration Services — миграция и интеграция данных.

SQL Server Analysis Services - OLAP и data mining внутри SQL сервера.

Business Intelligence vs. Data Science

Business Intelligence отвечает на вопросы: что произошло в прошлом квартале? как много товара мы продали? в чем проблема? в какой ситуации? т.е. говорит в основном о том, что было в прошлом и анализирует исторические данные, полученные на текущий момент.

BI оперирует структурированными данными, традиционными источниками и управляемыми наборами данных.

Data Science отвечает на вопросы: а что, если...? какой оптимальный сценарий для нашего бизнеса? что будет дальше? т.е. ставит цель предугадать, сделать прогноз. Области применения - оптимизация, прогнозирование, статистический анализ.

DC работает с большими структурированными и неструктурированными наборами данных.

Понятие жизненного цикла аналитики данных

Жизненный цикл аналитики данных - это последовательность действий, которую нужно выполнить на наборе входных данных для эффективного достижения цели аналитики с помощью выбранных методов анализа. Жизненный цикл аналитики данных может включать в себя: выявление проблем анализа данных, сбор набора данных, проектирование, анализ данных, визуализацию данных.

Для начала необходимо понять, какова реальная конечная цель проекта, какую пользу он может принести, оценить его критерии успешности или провала, выбрать технологии для сбора, трансформации и анализа данных. Нужно постараться ответить на такие дополнительные вопросы: достаточно ли у вас ресурсов на реализацию проекта? с кем вы будете контактировать по ходу проекта? достаточно ли у вас входных данных? были ли уже в компании попытки анализа? достаточно ли у вас времени на реализацию проекта? с какими проблемами вы можете столкнуться?

В этап выявления проблемы (изучения) входят также такие пункты: определение набора необходимых знаний, которые нужны для ориентации в предметной области; наличие доступных ресурсов (люди, инструменты, данные); структурирование данных с точки зрения аналитики; изучение истории бизнеса.

После изучения следует перейти к **подготовке данных**. Определите используемые программные средства, то есть выберите ПО, БД, инструменты анализа и визуализации. Получите, очистите и загрузите данные в систему, после чего оцените количество и качество данных.

На этапе **проектирования модели** нужно определить методы, технологии, рабочие процессы, необходимые для расчета модели и объем данных. Также важно определить корреляцию между переменными: столбцами таблиц и полями данных.

Построение модели заключается в том, чтобы разработать наборы данных для тестирования, обучения и производства. Затем оценить жизнеспособность и надежность данных для использования в модели. И в конечном итоге выбрать рабочее окружение, то есть аппаратные и программные средства, на которых можно настроить процесс.

В результате предыдущего шага получились некие результаты. Теперь необходимо определить, удалось ли вам достичь результата на основе критериев проекта, а также определить ключевые результаты исследования. В зависимости от целевой аудитории можно разработать диаграммы и графики и сформулировать итоги и рекомендации.

Последний этап - **практическая реализация**. В процессе этого этапа происходит доставка финальных рекомендаций, отчетов, кода и технических документов, запуск пилотного проекта, реализация модели в производственной среде и интеграция аналитических оценок в панель управления или в операционной системе.

Начиная работать с большими данными, многие сталкиваются с трудностями, которые можно избежать, выполняя простейшие советы, такие как: всегда проводить глубокое изучение предметной области, разбивать задачу на более мелкие, делать аналитику гибкой и масштабируемой, предусмотреть возможность повторения каждого из этапов с возможностью внесения изменений в предыдущий этап, быть готовым к негативному результату, внимательно оценивать затраченное время на каждый этап.

Смешение и интеграция данных

Так как работа с **Big Data** часто связана со сбором разнородных данных из разных источников, чтобы работать с этими данными, их нужно собрать воедино. Просто загрузить их в одну базу нельзя – разные источники могут выдавать данные в разных форматах и с разными параметрами. Тут и поможет смешение и интеграция данных – процесс приведения разнородной информации к единому виду.

Чтобы использовать данные из разных источников, используют следующие методы:

1. *Приводят данные к единому формату*: распознают текст с фотографий, конвертируют документы, переводят текст в цифры.
2. *Дополняют данные*. Если есть два источника данных об одном объекте, информацию от первого источника дополняют данными от второго, чтобы получить более полную картину.
3. *Отсеивают избыточные данные*: если какой-то источник собирает лишнюю информацию, недоступную для анализа, ее удаляют.

Смешение и интеграция данных нужны, если есть несколько разных источников данных, и нужно анализировать эти данные в комплексе.

Например, ваш магазин торгует офлайн, через маркетплейсы и просто через Интернет. Чтобы получить полную информацию о продажах и спросе, надо собрать множество данных: кассовые чеки, товарные остатки на складе, интернет-заказы, заказы через маркетплейс и так далее. Все эти данные поступают из разных мест и обычно имеют разный формат. Чтобы работать с ними, их нужно привести к единому виду.

Традиционные методы интеграции данных в основном основаны на **процессе ETL** (Extract, Transform, Load) – извлечение, преобразование и загрузка.

Это (ETL) специальный комплекс аппаратно-программных средств, включаемый в систему для реализации процессов в управлении хранилищами данных, который включает в себя:

- извлечение данных из внешних источников;
- их трансформацию и очистку, чтобы они соответствовали потребностям бизнес-модели;
- и загрузку их в хранилище данных.

С точки зрения процесса **ETL**, архитектуру хранилища данных можно представить в виде трёх компонентов:

- *источника данных*: содержит структурированные данные в виде таблиц, совокупности таблиц или просто файла (данные в котором разделены символами-разделителями);
- *промежуточной область*: содержит вспомогательные таблицы, создаваемые временно и исключительно для организации процесса выгрузки.
- *получателя данных*: хранилище данных или база данных, в которую должны быть помещены извлечённые данные.

Перемещение данных от источника к получателю называют *потоком данных*. Требования к организации потока данных описываются аналитиком. ETL следует рассматривать не только, как процесс переноса данных из одного приложения в другое, но и как инструмент подготовки данных к анализу.

Извлечение данных в ETL

Начальным этапом процесса ETL является процедура извлечения записи из источников данных и подготовка их к процессу преобразования. При разработке процедуры извлечения данных в первую очередь необходимо определить частоту выгрузки данных из OLTP-систем или отдельных источников. Выгрузка данных занимает определённое время, которое называется окном выгрузки.

Процедуру извлечения данных можно реализовать двумя способами:

- извлечение данных с помощью специализированных программных средств;
- извлечение данных средствами той системы, в которой они хранятся.

После извлечения данные помещаются в так называемую «промежуточную область», где для каждого источника данных создаётся своя таблица или отдельный файл, или и то, и другое.

О преобразовании данных мы поговорим подробнее чуть позже, а сейчас пара слов о **загрузке данных**.

Процесс загрузки заключается в переносе данных из промежуточных таблиц в структуру хранилища данных. При очередной загрузке в хранилище данных переносится не вся информация из источников, а только та, которая была изменена в течение промежуточного времени, прошедшего с предыдущей загрузки. При этом выделяют два потока:

- *поток добавления* — в хранилище данных передается новая, ранее не существовавшая информация;
- *поток обновления* (дополнения) — в хранилище данных передается информация, которая существовала ранее, но была изменена или дополнена.

Для распределения загружаемых данных на потоке используются средства данных. Они фиксируют состояние данных в некоторые моменты времени и определяют, какие данные были изменены или дополнены.

Вот теперь подробнее остановимся на **преобразовании данных.**

Цель этого этапа — подготовка данных к размещению в хранилище данных и приведение их к виду более удобному для последующего анализа. При этом должны учитываться некоторые выдвигаемые аналитиком требования, в частности, к уровню качества данных. Поэтому в процессе преобразования может быть задействован самый разнообразный инструментарий, начиная с простейших средств ручного редактирования данных и заканчивая системами, реализующими сложные методы обработки и очистки данных. В процессе преобразования данных в рамках ETL чаще всего выполняются следующие операции:

- преобразование структуры данных;
- агрегирование данных;
- перевод значений;
- создание новых данных;
- очистка данных.

Преобразование данных (трансформация) зависит от целей, задач, алгоритмов анализа. Разные задачи анализа потребуют различные методы преобразования.

Преобразование данных – широкое понятие. В контексте аналитики данных, задача преобразования – представить информацию в таком виде, чтобы её можно было максимально эффективно использовать с точки зрения решаемых задач аналитики данных.

Преобразование данных выполняется в компонентах информационных систем в зависимости от целей, преследуемых на конкретном этапе преобразования:

- в процессе переноса, загрузки данных в интегрированный источник

или в области их временного хранения (ETL);
- при подготовке данных к анализу в бизнес-приложении (SRD).

Целями преобразования данных могут быть:
- обеспечение технической, логической совместимости данных;
- подготовка данных к извлечению;
- перенос данных в хранилище;
- и др.

Например, адрес представляет собой набор данных разного типа: улица, город – текстовый, дом, квартира, индекс -числовой. Часто адреса записаны в одну строку, а для анализа необходимы конкретные компоненты адреса. С помощью трансформации можно разбить адрес на поля и преобразовать в нужный формат.

Не все данные перед поступлением в бизнес-приложения, прошли предварительную подготовку в системе. Трансформация данных в системах часто носит технический характер, а значит, слабо связана с возможными алгоритмами, методами, целями анализа.

Базовые операции преобразования в аналитических платформах и ETL-инструментах:

- *Параметры полей*

Изменение имен, типов, меток, назначения полей исходных данных.

Например, для работы требуются данные в числовом виде, а нужное поле в исходном наборе данные имеет строковый тип. Данные поля необходимо преобразовать к числовому типу, чтобы работа с этими данными стала возможной.

- *Квантование*

Разбивает диапазон значений числового признака на интервалы, присваивает значениям, попавшим в интервалы номера интервалов или другие метки.

- *Фильтр строк*

Оставляет записи, удовлетворяющие заданным условиям.

- *Сортировка*

По алгоритму, заданному пользователем, изменяет порядок следования записей исходного набора данных. Часто сортировка позволяет упростить визуальный анализ данных: например, определить наибольшее/наименьшее значения признаков.

- *Обогащение данных*

Операции, позволяющие в случае недостаточности данных для анализа в исходной выборке данных, дополнить её недостающей информацией, взятой из других выборок, методами слияния (например, объединение двух

таблиц по одноименным полям), объединения (к записям одной выборки добавляются записи другой), соединения (к записям одной выборки добавляются все выбранные поля), дополнения (используются одноимённые поля для дополнения одной таблицы полями из других таблиц, отсутствующими в первой) данных.

- *Табличная подстановка значений*

На основе *таблицы подстановки* (содержит пары исходное значение – новое значение) в исходной выборке данных производит замену значений.

Способ автоматической корректировки значений: каждое значение выборки проверяется на соответствие исходного значения таблицы подстановки. Если соответствие найдено, значение выборки меняется на новое значение.

- *Группировка*

Обобщается нужная информация, объединяется в минимально необходимое количество значений и полей (например, в случае когда в таблице данных информация разбросана по полям и записям, «разбавлена» посторонними данными.

- *Вычисляемые значения*

В случае, когда информацию можно получить на основе вычислений над имеющимися данными, в аналитическое приложение включен калькулятор. Например, известны цена и количество проданного товара, необходимо рассчитать сумму прибыли. Механизм работы калькулятора должен поддерживать работу с различными типами данных (не только с числовыми). Например, выполнять логические операции, выделять подстроку и др.

- *Преобразование упорядоченных данных*

Оптимизирует представление упорядоченных данных в целях обеспечения их дальнейшего анализа. Например, решение задачи прогнозирования временного ряда, группировка по временному периоду.

- *Транспонирование*

Делает строки столбцами, столбцы строками.

Итак. Крупные предприятия собирают, хранят и обрабатывают разные типы данных из множества источников, таких как системы начисления заработной платы, записи о продажах, системы инвентаризации и других. Вся эта информация извлекается, преобразуется и переносится в хранилища данных с помощью **ETL**-систем.

Рассмотрим **ETL-системы**, какие платные и общедоступные решения для работы с данными есть на рынке.

ETL-система извлекает данные из разных источников, преобразует их в соответствии с требованиями к формату хранилища данных, а затем загружает в это хранилище.

Схема всегда выглядит так: сначала извлечение данных из одного или нескольких источников, потом их подготовка к интеграции, после этого идет загрузка, и извлеченные данные попадают в общую базу.

Современные инструменты ETL собирают, преобразуют и хранят данные из миллионов транзакций в самых разных источниках данных и потоках. Эта возможность предоставляет множество новых возможностей: анализ исторических записей для оптимизации процесса продаж, корректировка цен и запасов в реальном времени, использование машинного обучения и искусственного интеллекта для создания прогнозных моделей, разработка новых потоков доходов, переход в облако и многое другое.

ETL используется для:

- перемещения данных в облако в процессе *облачной миграции.*
- перемещения данных в хранилище данных (база данных, куда передают данные из различных источников, чтобы их можно было совместно анализировать в коммерческих целях).
- объединения маркетинговых данных в маркетинговой интеграции (перемещение всех маркетинговых данных – о клиентах, продажах, из социальных сетей и веб-аналитики – в одно место, чтобы их можно было проанализировать).
- переноса данных в одно место от разных IoT, бизнес-данных.
- осуществления процесса репликации данных (*Репликация базы данных* – данные из исходных баз данных копируют в облачное хранилище. Это может быть одноразовая операция или постоянный процесс, когда ваши данные обновляются в облаке сразу же после обновления в исходной базе).

Популярные ETL-системы:

Cloud Big Data – PaaS-сервис для анализа больших данных (**Big Data**) на базе Apache Hadoop, Apache Spark, ClickHouse. Легко масштабируется, позволяет заменить дорогую и неэффективную локальную инфраструктуру обработки данных на мощную облачную инфраструктуру. Помогает обрабатывать структурированные и неструктурированные данные из разных источников, в том числе в режиме реального времени. Развернуть кластер интеграции и обработки данных в облаках можно за несколько минут, управление осуществляется через веб-интерфейс, командную строку или API.

IBM InfoSphere – инструмент ETL, часть пакета решений IBM Information Platforms и IBM InfoSphere. Доступен в различных версиях (Server Edition, Enterprise Edition и MVS Edition). Помогает в очистке, мониторинге, преобразовании и доставке данных, среди преимуществ:

масштабируемость, возможность интеграции почти всех типов данных в режиме реального времени.

PowerCenter – набор продуктов **ETL**, включающий клиентские инструменты PowerCenter, сервер и репозиторий. Данные хранятся в хранилище, где к ним получают доступ клиентские инструменты и сервер. Инструмент обеспечивает поддержку всего жизненного цикла интеграции данных: от запуска первого проекта до успешного развертывания критически важных корпоративных приложений.

iWay Software предоставляет возможность интеграции приложений и данных для удобного использования в режиме реального времени. Клиенты используют их для управления структурированной и неструктурированной информацией. В комплект входят: iWay DataMigrator, iWay Service Manager и iWay Universal Adapter Framework.

Microsoft SQL Server – платформа управления реляционными базами данных и создания высокопроизводительных решений интеграции данных, включающая пакеты ETL для хранилищ данных.

OpenText – платформа интеграции, позволяющая извлекать, улучшать, преобразовывать, интегрировать и переносить данные и контент из одного или нескольких хранилищ в любое новое место назначения. Позволяет работать со структурированными и неструктурированными данными, локальными и облачными хранилищами.

Oracle GoldenGate – комплексный программный пакет для интеграции и репликации данных в режиме реального времени в разнородных IT-средах. Обладает упрощенной настройкой и управлением, поддерживает облачные среды.

Pervasive Data Integrator – программное решение для интеграции между корпоративными данными, сторонними приложениями и пользовательским программным обеспечением. Data Integrator поддерживает сценарии интеграции в реальном времени.

Pitney Bowes предлагает большой набор инструментов и решений, нацеленных на интеграцию данных. Например, Sagent Data Flow – гибкий механизм интеграции, который собирает данные из разнородных источников и предоставляет полный набор инструментов преобразования данных для повышения их коммерческой ценности.

SAP Business Objects – централизованная платформа для интеграции данных, качества данных, профилирования данных, обработки данных и отчетности. Предлагает бизнес-аналитику в реальном времени, приложения для визуализации и аналитики, интеграцию с офисными приложениями.

Sybase включает Sybase ETL Development и Sybase ETL Server. Sybase ETL Development – инструмент с графическим интерфейсом для создания и проектирования проектов и заданий по преобразованию данных. Sybase ETL

Server – масштабируемый механизм, который подключается к источникам данных, извлекает и загружает данные в хранилища.

Open source ETL-средства

Большинство инструментов ETL с открытым исходным кодом помогают в управлении пакетной обработкой данных и автоматизации потоковой передачи информации из одной системы данных в другую. Эти рабочие процессы важны при создании хранилища данных для машинного обучения.

Некоторые из бесплатных и открытых инструментов ETL принадлежат поставщикам, которые в итоге хотят продать корпоративный продукт, другие обслуживаются и управляются сообществом разработчиков, стремящихся демократизировать процесс.

Open source ETL-инструменты интеграции данных:

Apache Airflow – платформа с удобным веб-интерфейсом, где можно создавать, планировать и отслеживать рабочие процессы. Позволяет пользователям объединять задачи, которые нужно выполнить в строго определенной последовательности по заданному расписанию. Пользовательский интерфейс поддерживает визуализацию рабочих процессов, что помогает отслеживать прогресс и видеть возникающие проблемы.

Apache Kafka – распределенная потоковая платформа, которая позволяет пользователям публиковать и подписываться на потоки записей, хранить потоки записей и обрабатывать их по мере появления. **Kafka** используют для создания конвейеров данных в реальном времени. Он работает как кластер на одном или нескольких серверах, отказоустойчив и масштабируем.

Apache NiFi – распределенная система для быстрой параллельной загрузки и обработки данных с большим числом плагинов для источников и преобразований, широкими возможностями работы с данными. Пользовательский веб-интерфейс **NiFi** позволяет переключаться между дизайном, управлением, обратной связью и мониторингом.

CloverETL (теперь **CloverDX**) был одним из первых инструментов **ETL** с открытым исходным кодом. Инфраструктура интеграции данных, основанная на **Java**, разработана для преобразования, отображения и манипулирования данными в различных форматах. **CloverETL** может использоваться автономно или встраиваться и подключаться к другим инструментам: RDBMS, JMS, SOAP, LDAP, S3, HTTP, FTP, ZIP и TAR. Хотя продукт больше не предлагается поставщиком, его можно безопасно загрузить с помощью **SourceForge**. **CloverDX** по-прежнему поддерживает CloverETL в соответствии со стандартным соглашением о поддержке.

Jaspersoft ETL – один из продуктов с открытым исходным кодом TIBCO Community Edition, позволяет пользователям извлекать данные из различных

источников, преобразовывать их на основе определенных бизнес-правил и загружать в централизованное хранилище данных для отчетности и аналитики. Механизм интеграции данных инструмента основан на Talend. Community Edition прост в развертывании, позволяет создавать витрины данных для отчетности и аналитики.

Apatar – кроссплатформенный инструмент интеграции данных с открытым исходным кодом, который обеспечивает подключение к различным базам данных, приложениям, протоколам, файлам. Позволяет разработчикам, администраторам баз данных и бизнес-пользователям интегрировать информацию разного формата из различных источников данных. У инструмента интуитивно понятный пользовательский интерфейс, который не требует кодирования для настройки заданий интеграции данных. Инструмент поставляется с предварительно созданным набором инструментов интеграции и позволяет пользователям повторно использовать ранее созданные схемы сопоставления.

*Почему стоит отказаться от локальных **ETL**-решений.*

Традиционные локальные **ETL** чаще всего поставляются в комплекте с головной болью. Например, создаются собственными силами, поэтому могут быстро устареть или не иметь сложных функций и возможностей. Они дороги и требуют времени на обслуживание, а также поддерживают только пакетную обработку данных и плохо масштабируются.

Локальные платформы **ETL** были важнейшим компонентом инфраструктуры предприятий на протяжении десятилетий. С появлением облачных технологий, **SaaS** и больших данных выросло число источников информации, что вызвало рост спроса на более мощную и сложную интеграцию данных, например, **Hadoop.**

Hadoop

Специальные инструменты экосистемы больших данных от **Hadoop** до баз данных **NoSQL** также имеют собственный подход для извлечения, преобразования и загрузки данных.

Hadoop – одно из решений для хранения и анализа больших данных. Его используют Google, Amazon, Facebook, Twitter, eBay и другие гиганты рынка. При этом технология подходит для любого бизнеса, работающего с объемами данных свыше терабайта, оптимизирована для работы на виртуальных машинах, удобно масштабируется. Поэтому облачные провайдеры предлагают ее компаниям как сервис в облаке, который легко внедрить и применять.

Hadoop помогает хранить и обрабатывать массивы информации, готовить ее для выгрузки в другие сервисы, собирать статистику. По сути,

это конструктор, на основе которого строят хранилища данных под потребности бизнеса.

Лучше всего **Hadoop** подходит для работы с неструктурированными данными – неупорядоченной информацией без определенной структуры, которую сложно классифицировать и разбить на группы. Например, с файлами документов, сообщениями, аудио- и видеозаписями, изображениями.

Система может искать нужные сведения в огромном архиве, получать из массива «пустой» информации небольшое количество значимой для компании. Например, подсчитать уникальных пользователей в трафике с миллионов IP-адресов.

Так, крупная сеть универмагов может собирать и обрабатывать информацию о поведении и предпочтениях клиентов из интернета, обрабатывать ее, помещать в хранилище. Там данные объединяют с информацией о продажах, анализируют, в результате становится ясно, какие действия на сайте приводят к покупкам.

Hadoop состоит из нескольких инструментов, в частности файловой базы данных и готовых решений для их обработки, его преимущества:

- **Хранение и быстрая обработка любых данных**. **Hadoop** можно настроить так, чтобы он обрабатывал информацию со всех интернет-ресурсов и социальных сетей компании, систем работы с клиентами, промышленных объектов и датчиков, финансовых отчетов и других источников. Архивы данных в **Hadoop** устроены так, что к ним можно получить доступ, как только они становятся нужны.
- **Высокая мощность вычислений**. Именно поэтому **Hadoop** быстро обрабатывает данные. Мощность зависит от числа вычислительных узлов: чем их больше, тем она выше.
- **Устойчивость к отказам**. В случае аппаратного сбоя, например, если узел вышел из строя, данные пойдут на другой узел, что исключает ошибки.
- **Копии данных сохраняются в системе автоматически**. Не нужно обрабатывать данные перед сохранением. **Hadoop** обрабатывает и неструктурированные данные, например: тексты, изображения, видео.
- **Масштабируемость**. Вы можете добавлять дополнительные узлы, если объем данных увеличится.

Функции Hadoop: для чего можно использовать технологию.

Hadoop подходит для управления безопасностью и рисками, оптимизации маркетинга, финансового анализа, научных и маркетинговых исследований, индексации веб-сайтов, анализа «озер данных» – большого объема неструктурированной информации, собранной компанией.

По данным исследования **Syncsort**, 71% компаний применяют **Hadoop** не только для решения новых проблем с большими данными, но и чтобы улучшить работу с типами информации, которую они используют много лет.

Hadoop для анализа «озер данных»

«Озера данных» – несистематизированная информация, которую компания собирает из разных источников для дальнейшего анализа. Такие данные могут пригодиться в будущем или их обязывает хранить закон.

Когда информация хранится в разных источниках и форматах, она недоступна для анализа, моделирования, прогнозирования, а значит, бесполезна для компании. С помощью **Hadoop** собранные данные можно распределить и структурировать, настроить аналитику для построения моделей и проверки предположений.

Hadoop для обработки данных из соцсетей

В социальных сетях есть массивы данных, анализ которых важен, чтобы понять потребности клиентов. **Hadoop** помогает извлекать информацию для обогащения клиентских профилей: идентификационные данные, семейное положение, интересы, образование, социальный статус и т. д.

Аналитика помогает управлять репутацией компании, таргетировать рекламу на нужную аудиторию, повышать эффективность социальных сетей как канала продаж.

Hadoop для анализа отношения к бренду

Hadoop может собирать и анализировать мнения и эмоции, которые пользователи высказывают в социальных сетях, блогах, онлайн-обзорах, отзывах. Это помогает понять, как люди относятся к продуктам и услугам компании или ее конкурентов, оценить репутацию на рынке, скорректировать продвижение продукта, спрогнозировать продажи.

Hadoop для обработки данных о поведении клиентов

Hadoop может быть полезен для сбора и оценки данных о вовлеченности и поведении клиентов на сайте компании. Платформа собирает данные, откуда пользователи приходят на сайт, на какую веб-страницу, по какому поисковому запросу, куда переходят, сколько времени проводят на сайте, что покупают и с каких страниц уходят.

Анализируя эту информацию, компании могут оптимизировать путь пользователя к покупке, повысить конверсию страниц, сделать удобнее сайт и корзину интернет-магазина, спрогнозировать, какие товары купят пользователи.

Hadoop для обеспечения безопасности и управления рисками

Hadoop анализирует данные серверных журналов и помогает реагировать на нарушения безопасности.

Серверные журналы генерирует компьютер, там собраны данные о работе сети, важные для безопасности и соответствия нормативам. **Hadoop**

подходит для извлечения ошибок, подсчета сбоев системы, получения информации об использовании корпоративных сетей и кибератаках.

С помощью **Hadoop** определяют причины нарушения безопасности, оценивают и моделируют риски, обнаруживают сетевые вторжения. Это помогает разработать способы защиты от злоумышленников.

Hadoop для анализа геоданных

Компании розничной торговли, автомобильной промышленности, производства и магазины могут с согласия клиентов собирать данные об их передвижениях через смартфоны и планшеты, затем хранить и анализировать информацию. Это позволяет прогнозировать визиты покупателей, делать пользователям предложения с учетом их геолокации, строить оптимальные маршруты для транспорта. **Hadoop** поможет сохранить, оптимизировать и обработать огромное количество геоданных.

Hadoop для обработки данных от интернета вещей

Hadoop подходит для обработки данных с различных устройств интернета вещей. Это могут быть персональные **IoT**, например: фитнес-трекеры, которые отдают информацию о местоположении и привычках пользователя, или устройства умного дома.

IoT также применяют в городских экосистемах и промышленности для поддержки процесса производства и управления инфраструктурой, мониторинга транспортных средств и грузов, разработки умных инженерных систем, например, электро-, газо- и водоснабжения.

Обработка данных от систем **IoT** позволяет компаниям сократить расходы, улучшить качество выпускаемой продукции, оптимизировать производство и увеличить продажи.

Hadoop для создания корпоративного центра данных

С помощью **Hadoop** обрабатывают и анализируют массивы внутренних данных компании, получаемых в процессе работы или взаимодействия с клиентами.

Например, ритейлеры анализируют данные о покупках, складских остатках, ассортименте магазинов. Транспортные компании анализируют движение и скорость автомобилей, время грузоперевозок. Банки прогнозируют и оценивают число транзакций, поток клиентов, риск мошеннических действий.

Также на базе **Hadoop** можно создать корпоративный центр данных, из которого пользователи будут брать информацию для работы.

Как компании используют Hadoop

По результатам исследования iDatalabs, технологию чаще используют компании, работающие в сферах программного обеспечения, IT-технологий и услуг, рекрутинга, образования, здравоохранения.

Сфера деятельности	Как применяют Hadoop
Ритейлеры и продавцы услуг	Собирают данные о продажах и транзакциях, поведени покупателей на сайте, информацию из соцсетей и с других ресурсо финансовую информацию, отчетность об ассортименте и складски остатках. Зная, как ведут себя клиенты, можно делать персональны предложения и акции, предлагать востребованные товар разрабатывать программы лояльности, повышать продажи
Предприятия, работающие в ресурсоемких отраслях	Поставщики коммунальных услуг, нефти и газа, промышленны производства, фабрики и заводы используют информацию с датчиков, внутренних сервисных служб, внешних производителе активов. Так можно прогнозировать интервалы техническог обслуживания, цены на продукцию и другие важные фактор уменьшить затраты на производство и оптимизировать рабочи процессы
Финансовые организации, в том числе банки	Анализируют финансовую информацию и риски, выявляю мошеннические действия и разрабатывают защиту от них. Банк работают с большими объемами данных о клиентах и транзакция их анализ помогает предсказывать количество посетителей отделениях, остатки средств в банкоматах, приток и отто корпоративных клиентов. Hadoop успешно справляется с таким задачами: по данным Syncsort, ⅔ организаций финансовой отрасл отмечают, что платформа делает бизнес более гибким и повыша операционную эффективность.
Организации здравоохранения, частные клиники	Около 80% медицинских данных – неструктурированные. Сбор анализ такой информации помогает снизить риск мошенничества с страховками, увеличить прибыльность медцентров, проводит научные исследования, выявлять факторы риска заболеваний оценивать эффективность лечения
Транспортные компании	Используют Hadoop для сбора и анализа данных о транспортировк грузов, перемещении автомобилей, сроках доставки. Это помога уменьшить расходы на топливо, прогнозировать лучшие маршрут определять сроки технического обслуживания транспорта

У **Hadoop** есть обширная экосистема дополнительных проектов с открытым исходным кодом, поэтому большинству компаний сложно внедрять и применять технологию. Например, нужны отдельные специалисты, которые занимаются построением хранилищ данных. Это затрудняло использование **Hadoop** как самостоятельного решения.

Сейчас настроенные инструменты **Hadoop** можно получить в виде облачного сервиса. Такие решения упрощают внедрение **Hadoop**, поскольку не требуют капитальных затрат для пилотных проектов. Кроме того, провайдеры берут на себя экспертное администрирование **Hadoop**, что снимает с пользователей необходимость искать экспертов в штат и делает применение и масштабирование **Hadoop** дешевле и проще.

Наконец, интеграция облачных решений **Hadoop** с недорогими S3-хранилищами снижает затраты на хранение больших данных – обслуживание локальной инфраструктуры обходится дороже.

Hadoop лучше всего подходит:

– Для хранения и обработки неструктурированных данных объемом от одного терабайта – такие массивы сложно и дорого хранить в локальном хранилище.
– Для компонуемых вычислений – когда нужно собрать множество схожих разрозненных данных в одно целое. Также подходит для выделения полезной информации из массива лишней.
– Для пакетной обработки, обогащения данных и **ETL** – извлечения информации из внешних источников, ее переработки и очистки под потребности компании, последующей загрузки в базу данных.

Чтобы не устанавливать и не настраивать компоненты **Hadoop** самостоятельно, можно подключить **Hadoop** в виде облачного сервиса, с бесплатным тестированием.

После интеграции большие данные подвергаются дальнейшим манипуляциям: анализу и так далее.

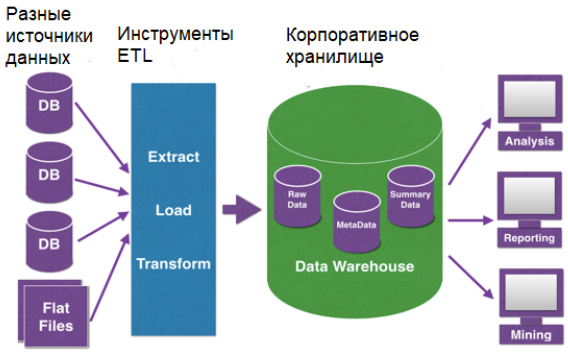


Схема выглядит примерно так: данные извлекают, очищают и обрабатывают, помещают в корпоративное хранилище данных, а потом забирают для анализа.

Краудсорсинг

Обычно анализом **Big Data** занимаются компьютеры, но иногда его поручают и людям. Для этих целей существует **краудсорсинг** – привлечение к решению какой-либо проблемы большой группы людей.

Краудсорсинг (crowdsoursing) – понятие, состоящее из 2-х слов:

croud – толпа,

source – источник.

Получаем: источник чего-то, что мы черпаем в толпе. Слово – аналог аутсорсинга – когда мы передаём какое-то задание внешнему исполнителю.

А краудсорсинг – это когда мы передаём какое-то задание, какую-то функцию внешнему исполнителю, которым является толпа (в фигуральном смысле), т.е. большому количеству людей. Мы размещаем задание, как правило, в интернете. Мы передаём это задание всем, кто может прочитать это задание, увидеть его.

Классическим примером краудсорсинга является Википедия – энциклопедия, в которой можно посмотреть любой термин, любое понятие. Самое интересное, что Википедия в настоящий момент – самая большая энциклопедия в мире, она в 15 раз превышается британскую энциклопедию. Но самое главное не это: количество ошибок на тысячу статей в Википедии меньше, чем в британской энциклопедии. Т.е. мы получили продукт - энциклопедию, абсолютный продукт, который создан всеми людьми, которые участвовали в его создании. Это не коммерческий продукт, никто не получил денег за него.

Рассмотрим пример, когда использовался мировой потенциал для решения задачи. Это касается канадской золотодобывающей компании Голдкорп, которая в своё время столкнулась с необходимостью расширения своей деятельности на том участке, на котором она добывала золото, но она не смогла самостоятельно обработать данные геологоразведки на прилегающих территориях. Она не смогла определить где закладывать новые рудники. Компания разместила данные геологоразведки в интернете и попросила всех желающих обработать эти данные и подсказать где могут быть значимые залежи золота. Большое количество коллективов включились в работу. Компания назначила премию: кто лучше выполнит работу получит приз в пол млн.$. Выиграла Fractal Graphics, состоящая из 4-х математиков. Они никогда не работали с такими данными, но получили результат, им удалось указать те места, где были действительно значимые запасы золота. А компания Голдкорп увеличила величину своих запасов и повысила свою капитализацию примерно на 800 млн $.

Вывод: краудсорсинг обладает очень высокой рентабельностью.

Краудсорсинг в России

Изначально инициаторами появления краудсорсинга в России было государство. Когда государство пришло к мысли, что развивать экономику нужно инновационными методами, госкорпорации стали принуждать к инновациям. При этом принуждение происходило таким образом, что госкорпорациям не объясняли, что такое инновации. Им необходимо было самим понять где и как найти эти открытые инновации, никто их на рынке не предлагал. Поэтому стали искать, смотреть как это сделано на западе и компании пришли к выводу, что самая удобная технология для проведения

открытых инноваций – краудсорсинг. Первым, кто пришел к такой мысли – Герман Греф, на многие годы краудсорсинг стал лидирующей технологией в Сбербанке.

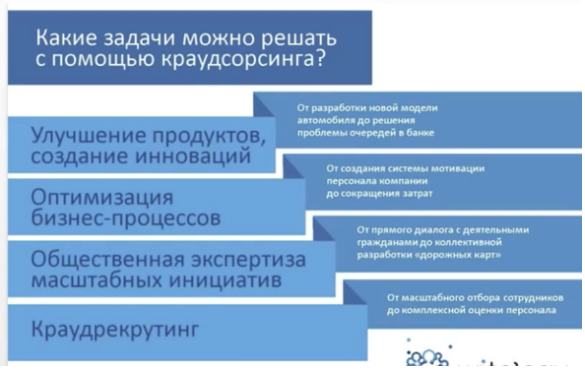
Проекты:

Сбербанк: Очередей. Нет!

Азбука вкуса: Выйди из себя.

Российское законодательство (предложения предпринимателями планы мероприятий по упрощению, удешевлению и ускорению действующих на территории РФ процедур с ведением бизнеса), здравоохранение («Вместе за достойную медицину»), краудрекрутинг для Росатома.

Пример простого краудсорсинга с простыми данными компании: предположим, у вас есть большой объем сырых данных. Например, записи о продажах магазинов, где товары часто записаны с ошибками и сокращениями. К примеру, дрель Dexter с аккумулятором на 10 мАч записана как «Дрель Декстр 10 мАч», «Дрель Dexter 10», «Дрель Dexter акк 10» и еще десятком других способов. Вы находите группу людей, которые готовы за деньги вручную просматривать таблицы и приводить такие наименования к одной форме.



Краудсорсинг хорош, если задача разовая и для ее решения нет смысла разрабатывать сложную систему искусственного интеллекта. Если анализировать большие данные нужно регулярно, система, основанная на **Data Mining** или машинном обучении, скорее всего, обойдется дешевле краудсорсинга. Кроме того, машины лучше справятся со сложным анализом, основанном на математических методах, например, со статистикой или имитационным моделированием.