МОДУЛЬ 1

Лекция №1. Аналитика данных

Постоянное увеличение скорости обработки данных и пропускной способности, непрерывное изобретение новых инструментов для создания, обмена и потребления данных, а также постоянное появление новых создателей и потребителей данных по всему миру гарантируют, что рост объема данных не прекращается.

Данные порождают еще больше данных в постоянном добродетельном цикле.

Современная экосистема данных включает в себя целую сеть взаимосвязанных, независимых и постоянно развивающихся сущностей.

Она включает в себя данные, которые необходимо интегрировать из разрозненных источников, различные виды анализа и навыки для получения глубоких знаний,

активные заинтересованные стороны, которые должны сотрудничать и действовать в соответствии с полученными знаниями, а также инструменты, приложения и инфраструктуру для хранения, обработки и распространения данных по мере необходимости.

Начнем с источников данных.

Данные доступны в виде различных структурированных и неструктурированных наборов данных, содержащихся в тексте, изображениях, видео, потоках кликов, разговорах пользователей, платформах социальных сетей, Интернете вещей, или IoT-устройствах, событиях реального времени, которые передают данные, старые базы данных и данные, полученные от профессиональных поставщиков данных и агентств.

Никогда прежде источники не были столь разнообразными и динамичными.

Когда вы работаете с таким количеством различных источников данных, первым шагом является получение копию данных из исходных источников в хранилище данных.

На этом этапе вы только приобретаете необходимые вам данные, работаете с форматами данных, источниками и интерфейсами, через которые эти данные могут быть получены.

Надежность, безопасность и целостность получаемых данных - вот некоторые из проблем, которые необходимо решить на этом этапе.

После того как необработанные данные собраны в одном месте, их необходимо организовать, очистить и оптимизировать для доступа конечных пользователей.

Данные также должны соответствовать требованиям и стандартам, принятым в организации. Например, соответствие руководящим принципам, регулирующим хранение и использование персональных данных, таких как медицинские, биометрические или бытовые данные в случае устройств IoT.

Соблюдение таблиц основных данных в организации для обеспечения стандартизации

Еще одним примером является соблюдение таблиц основных данных в организации для обеспечения стандартизации основных данных во всех приложениях и системах организации.

Основные задачи на этом этапе могут включать управление данными и работу с хранилищами данных которые обеспечивают высокую доступность, гибкость, доступность и безопасность.

Наконец, у нас есть заинтересованные стороны, приложения, программисты, аналитики и специалисты по анализу данных. все они получают эти данные из корпоративного хранилища данных.

Основные проблемы на этом этапе могут включать интерфейсы, API и приложения, которые могут донести эти данные до конечных пользователей в соответствии с их конкретными потребностями.

Например, аналитикам данных могут понадобиться исходные данные для работы с бизнесом. Заинтересованным сторонам могут понадобиться отчеты и информационные панели.

Приложениям могут понадобиться пользовательские API для получения этих данных.

Важно отметить влияние некоторых новых и развивающихся технологий, которые формируют сегодняшнюю экосистему данных и ее возможности.

Например, облачные вычисления, машинное обучение и большие данные.

Благодаря облачным технологиям каждое предприятие сегодня имеет доступ к безграничному хранилищу данных, высокопроизводительным вычислениям, технологиям с открытым исходным кодом, большим данным, высокопроизводительным вычислениям, технологиям с открытым исходным кодом, технологиям машинного обучения и новейшим инструментам, и библиотекам.

Специалисты по изучению данных создают прогностические модели путем обучения алгоритмов машинного обучения на прошлых данных, Кроме того, большие данные.

Сегодня мы имеем дело с массивами данных, которые настолько массивны и разнообразны, что традиционные инструменты и методы анализа перестают быть адекватными. традиционные инструменты и

методы анализа уже не подходят, что открывает путь для новых инструментов и методов, а также новых знаний и представлений.

Мы узнаем больше о больших данных и их влиянии на формирование бизнес-решений в дальнейшем в этом курсе.

ЛЕКЦИЯ 2. КЛЮЧЕВЫЕ ИГРОКИ В ЭКОСИСТЕМЕ ДАННЫХ

Сегодня организации, которые используют данные для раскрытия возможностей и применяют эти знания для дифференциации себя, являются лидерами в будущем.

Будь то поиск закономерностей в финансовых операциях для выявления мошенничества, использование рекомендательных для повышения конверсии, поиск информации в социальных сетях для выявления мнения клиентов или персонализация предложений брендов на основе анализа поведения клиентов свои предложения на основе анализа поведения клиентов, лидеры бизнеса осознали, что данные являются ключом к конкурентному преимуществу.

Чтобы получить ценность от данных, необходимо огромное количество наборов навыков и людей, играющих различные роли.

В этом видео мы рассмотрим роли инженеров по данным, аналитиков данных, ученых по данным, бизнес-аналитики и аналитики бизнес-аналитики (BI-аналитики), которые помогают организациям использовать огромные объемы данных и превращать их в практические выводы.

Все начинается с инженера данных.

Инженеры по данным - это люди, которые разрабатывают и поддерживают архитектуры данных и делают данные доступными для бизнес-операций и анализа.

Инженеры по данным работают в экосистеме данных для извлечения, интеграции и организации данных из разрозненных источников.

Чистое преобразование и подготовка данных проектируют, хранят и управляют данными в хранилищах данных. Они обеспечивают доступ к данным в форматах и системах, которые необходимы различным бизнес-приложениям, а также заинтересованные стороны, такие как аналитики данных и специалисты по исследованию данных, могут использовать их.

Инженер по данным должен обладать хорошими знаниями в области программирования, глубокими знаниями систем и технологических архитектур, а также глубокое понимание реляционных баз данных и нереляционных хранилищах данных.

Теперь давайте рассмотрим роль аналитика данных.

Вкратце, аналитик данных переводит данные и цифры на простой язык, чтобы организации аналитики данных проверяют и очищают данные для получения глубокого понимания, выявляют корреляции, поиск закономерностей и применение статистических методов.

Анализировать и добывать данные, визуализировать данные для интерпретации и представления результатов анализа данных.

Аналитики - это люди, которые отвечают на такие вопросы, как, например, является ли опыт поиска пользователей в целом хороший или плохой с функцией поиска на нашем сайте? или каково популярное восприятие людей относительно наших инициатив по ребрендингу?

Или есть ли корреляция между продажами, одного продукта и другого?

Аналитики данных требуют хорошего знания электронных таблиц, написания запросов и использования статистических инструментов для создания графиков и информационных панелей.

Современные аналитики данных также должны обладать некоторыми навыками программирования. Им также необходимы сильные аналитические способности и умение рассказывать истории.

А теперь давайте посмотрим, какую роль в этой экосистеме играют специалисты по анализу данных.

Специалисты по анализу данных анализируют данные для получения действенных выводов и строят модели машинного обучения или глубокого обучения, которые обучаются на прошлых данных для создания прогнозирующих моделей.

Специалисты по изучению данных - это люди, которые отвечают на такие вопросы, как "Сколько новых подписчиков в социальных сетях я могу получить в следующем месяце или какой процент моих клиентов я потеряю из-за конкуренции в следующем квартале, или является ли эта финансовая операция необычной для данного клиента?

Специалистам по работе с данными требуются знания математики, статистики и хорошее понимание языков программирования, баз данных и построения данных, языков, баз данных и построения моделей данных. Они также должны обладать знаниями в своей области.

Также существуют бизнес-аналитики и BI-аналитики.

Бизнес-аналитики используют результаты работы аналитиков данных и ученых по данным для изучения возможных возможные последствия для их бизнеса и действия, которые они должны предпринять или рекомендовать.

BI-аналитики делают то же самое.

Их внимание сосредоточено на рыночных силах и внешних влияниях, которые формируют их бизнес. Они предлагают интеллектуальные решения для бизнеса путем организации и мониторинга данных по различным бизнес-функциям и исследуя эти данные для извлечения идей и действий, которые улучшают эффективность бизнеса. Проще говоря, инженерия данных преобразует необработанные данные в пригодные для использования.

Аналитика данных использует эти данные для получения глубоких выводов.

Специалисты по анализу данных используют аналитику данных и инженерию данных для прогнозирования будущего на основе данных из прошлого.

Бизнес-аналитики и аналитики бизнес-аналитики используют эти идеи и прогнозы для принятия решений, которые приносят пользу и способствуют развитию бизнеса.

Интересно, что нередко специалисты по данным начинают свою карьеру с одной и переходят к другой роли в экосистеме данных, дополняя свои навыки.

3 ЛЕКЦИЯ. ОПРЕДЕЛЕНИЕ АНАЛИЗА ДАННЫХ

Анализ данных - это процесс сбора, очистки, анализа и добычи данных, интерпретации результатов и составления отчета о полученных результатах. С помощью анализа данных мы находим закономерности в данных и корреляции между различными точками данных.

Именно благодаря этим закономерностям и корреляциям формируется понимание и делаются выводы.

Анализ данных помогает компаниям понять свои прошлые результаты и обосновать принятие решений для будущих действий.

Используя анализ данных, компании могут проверить правильность выбранного курса действий, прежде чем принять решение.

Это экономит ценное время и ресурсы, а также обеспечивает больший успех. Мы рассмотрим четыре основных типа анализа данных, каждый из которых имеет свою цель и место в процессе анализа данных.

Описательный анализ помогает ответить на вопросы о том, что произошло за определенный период времени путем обобщения прошлых данных и представления выводы заинтересованным сторонам. Она помогает предоставить важную информацию о прошлых событиях.

Например, отслеживание прошлых результатов деятельности на основе ключевых показателей эффективности организации или анализ денежных потоков.

Диагностическая аналитика помогает ответить на вопрос. Почему это произошло?

На основе данных описательной аналитики можно копнуть глубже, чтобы найти причину результата.

Например, внезапное изменение посещаемости веб-сайта без очевидной причины или увеличение продаж в регионе, где не было никаких изменений в маркетинге.

Предиктивная аналитика помогает ответить на вопрос: "Что будет дальше?

Исторические данные и тенденции используются для прогнозирования будущих результатов.

Некоторые из областей, в которых бизнес применяет предиктивный анализ, - это оценка рисков и прогнозирование продаж.

Важно отметить, что цель предиктивного анализа не в том, чтобы сказать, что произойдет в будущем, ее цель - предсказать, что может произойти в будущем.

Все прогнозы носят вероятностный характер.

Предписывающая аналитика помогает ответить на вопрос,

Что следует предпринять? Анализируя прошлые решения и события, можно определить вероятность различных исходов.

На основе этого оценивается вероятность различных исходов и принимается решение о том, что делать дальше.

Самоуправляемые автомобили - хороший пример предписывающей аналитики. Они анализируют окружающую среду, чтобы принимать решения относительно скорости, смены полосы движения, маршрута.

Или авиакомпании, автоматически корректирующие цены на билеты в зависимости от спроса клиентов. Цены на бензин, погода или пробки на маршрутах.

Теперь давайте рассмотрим некоторые ключевые шаги в любом процессе анализа данных.

Понимание проблемы и желаемого результата.

Анализ данных начинается с понимания проблемы, которую необходимо решить, и желаемого результата результат, которого необходимо достичь. Где вы находитесь и где вы хотите быть, должны быть четко определены до начала процесса анализа.

Установление четкой метрики.

Этот этап процесса включает в себя принятие решения о том, что будет измеряться.

Например, количество проданного в регионе продукта X и как оно будет измеряться, например.

В течение квартала или во время фестивального сезона сбор данных, как только вы узнаете, что вы собираетесь измерять и как вы собираетесь это измерять, вы определяете необходимые вам данные, источники данных, которые вам нужны источники данных, из которых вам нужно получить эти данные, и лучшие инструменты для этой работы.

Очистка данных. После сбора данных следующим шагом является устранение проблем с качеством данных, которые могут повлиять на точность анализа.

Это очень важный шаг, поскольку точность анализа может быть обеспечена только в том случае - данные чистые.

Вы очистите данные на предмет отсутствующих или неполных значений и выбросов. Например, демографические данные клиента, в которых поле возраст имеет значение 150 - это выброс.

Вы также стандартизируете данные, поступающие из нескольких источников.

Анализ и извлечение данных.

После того как данные будут очищены, вы извлечете и проанализируете их с различных точек зрения. Вам может понадобиться манипулировать данными несколькими различными способами, чтобы понять тенденции, определить корреляции и найти закономерности и вариации. Интерпретация результатов.

После анализа данных и, возможно, проведения дальнейшего исследования, которое может быть итерационным цикл, настало время интерпретировать результаты. В процессе интерпретации результатов вам необходимо оценить, можно ли защитить ваш анализ от возражений, и есть ли какие-либо ограничения или обстоятельства, при которых ваш анализ может оказаться неверным.

Представление своих выводов.

В конечном счете, цель любого анализа - повлиять на принятие решений. Способность донести и представить свои выводы в ясной и действенной форме является столь же важной частью процесса анализа данных, как и сам анализ.

Отчеты, информационные панели, диаграммы, графики, карты, тематические исследования - это лишь некоторые из способов представления данных.

4 Лекция. Мнение практиков об аналитике данных

В этом видео мы послушаем, как как несколько профессионалов в области данных рассказывают о том, как они определяют аналитику данных и что этот термин означает для них.

Я определяю аналитику данных как процесс

- сбора информации
- последующего анализа этой информации
- подтверждения различных гипотез исходя из этой информации.

Для меня аналитика данных также означает рассказывать истории с помощью данных. Использование данных для того, чтобы четко и лаконично донести информацию о состоянии мира до окружающих вас людей.

Анализ данных - это использование информации вокруг вас для принятия решений. Точно так же, как вы встаете каждое утро, вы смотрите новости. Прогноз погоды расскажет какая температура будет в этот день, будет ли дождь. От этого может зависеть, что вы собираетесь надеть или чем заняться. Анализ данных - это не абстрактная концепция, это то, что мы делаем естественным образом, но у него есть техническое название и теперь людям платят за то, чтобы они делали это в гораздо большем или более масштабном опыте. Но на самом деле все не так сложно.

Я объясняю это так: у вас есть проблема, и вам нужно использовать факты для проверки гипотезы, вот тут-то и вступает в игру аналитика данных.

Процесс начинается с определения проблемы, понимание желаемого результата и затем вам нужно создать свою собственную гипотезу.

Чтобы проверить ее, необходимо:

- собрать данные,
- очистить их,
- анализировать данные,
- затем представить ее ключевым заинтересованным сторонам.

Аналитика данных - это любые наборы данных, которые вы можно использовать для анализа информации, все, что поможет вам понять, что происходит. В моем случае, как CPA, я всегда смотрю на финансовое состояние. Я всегда анализирую данные. чтобы предсказать, где кто-то был, где они находятся сейчас и куда они направляются. Эти данные

помогают мне видеть дальше и почти предсказать будущее любой компании, с которой я работаю.

Аналитика данных - это сбор, очистка, анализ, представление, и, в конечном итоге, обмен данными и результатами анализа для того, чтобы помочь донести что именно происходит в вашем бизнесе, что происходит в данных, чтобы вы могли помогать принимать более эффективные решения.

Я бы определил аналитику данных как процесс или, лучше сказать, феномен получения информации собранную от релевантной аудитории, возможно, ваших клиентов или вашей социальной аудитории, и разбивать эту информацию на подмножества, и использование этих данных для принятия решений о продуктах или услугах, которые вы хотите предложить, или в случае цифровой среды, в которой мы сейчас находимся, принятия решений об определенных фрагментах контента, который вы хотите опубликовать так, чтобы это понравилось вашей целевой аудитории.

ЛЕКЦИЯ 5. АНАЛИТИКА ДАННЫХ И АНАЛИЗ ДАННЫХ (2 МИН)

Термины "Анализ данных" и "Аналитика данных" часто используются как взаимозаменяемые, в том числе и в этом курсе.

Однако важно отметить, что существует тонкая разница между терминами и значением слов Анализ и Аналитика. На самом деле некоторые люди утверждают, что эти термины означают разные вещи и не должны использоваться как взаимозаменяемые. Да, техническая разница есть...

Словарные значения таковы:

Анализ - детальное изучение элементов или структуры чего-либо.

Аналитика - систематический вычислительный анализ данных или статистики.

Анализ может проводиться и без цифр или данных, например, бизнес-анализ, психоанализ и т.д. В то время как аналитика, даже если она используется без приставки "данные", почти всегда подразумевает использование данных для проведения численных манипуляций и выводов.

Некоторые эксперты даже говорят, что Анализ данных основан на выводах, основанных на исторических данных, в то время как Аналитика данных предназначена для прогнозирования будущих показателей. Команда разработчиков данного курса не придерживается этой точки зрения, и вы увидите, почему, позже в курсе, когда познакомитесь с

такими терминами, как предиктивная аналитика, предписывающая аналитика и т. д.

Поэтому в этом курсе мы придерживаемся более либеральной точки зрения и используем термины "анализ данных" и "аналитика данных" для обозначения одного и того же. Например, предыдущее видео называется "Определение анализа данных", а предыдущее видео с точкой зрения нескольких профессионалов в области данных называется "Что такое анализ данных". Разница в этих названиях не является преднамеренной.

ЛЕКЦИЯ 6. РЕЗЮМЕ И ОСНОВНЫЕ МОМЕНТЫ

Современная экосистема данных включает в себя сеть взаимосвязанных и постоянно развивающихся объектов, которые включают в себя:

Данные, доступные во множестве различных форматов, структур и источников.

Среда данных предприятия, в которой необработанные данные находятся на стадии обработки, чтобы их можно было организовать, очистить и оптимизировать для использования конечными пользователями.

Конечные пользователи, такие как бизнес-заинтересованные лица, аналитики и программисты, которые потребляют данные для различных целей.

Новые технологии, такие как облачные вычисления, машинное обучение и большие данные, постоянно меняют экосистему данных и возможности, которые она предлагает. Инженеры по данным, аналитики данных, ученые по данным, бизнес-аналитики и аналитики бизнес-аналитики - все они играют жизненно важную роль в экосистеме, позволяющей извлекать из данных глубокие знания и результаты для бизнеса.

В зависимости от целей и результатов, которых необходимо достичь, существует четыре основных типа анализа данных:

- Описательная аналитика, которая помогает расшифровать "Что произошло".
- Диагностическая аналитика, которая помогает понять "Почему это произошло".
- Прогнозирующий анализ, который анализирует исторические данные и тенденции, чтобы предположить "Что произойдет дальше".
- Предписывающая аналитика, которая предписывает "Что следует делать дальше".

Процесс анализа данных включает в себя:

- Понимание проблемы и желаемого результата.
- Установление четкой метрики для оценки результатов.
- Сбор, очистка, анализ и добыча данных для интерпретации результатов.
- Сообщение результатов таким образом, чтобы они повлияли на принятие решений.

МОДУЛЬ 2

1 Лекция. Обязанности аналитика данных

Хотя роль аналитика данных варьируется в зависимости от типа организации и степени в какой она внедрила практику работы с данными, существуют некоторые обязанности, которые типичны для роли аналитика данных в современных организациях.

К ним относятся:

- Получение данных из первичных и вторичных источников данных,
- Создание запросов для извлечения необходимых данных из баз данных и других систем сбора данных,
- Фильтрация, очистка, стандартизация и реорганизация данных при подготовке к анализу данных,
- Использование статистических инструментов для интерпретации наборов данных,
- Использование статистических методов для выявления закономерностей и корреляций в данных,
- Анализ закономерностей в сложных наборах данных и интерпретация тенденций,
- Подготовка отчетов и диаграмм, которые эффективно передают тенденции и закономерности,
- Создание соответствующей документации для определения и демонстрации этапов процесса анализа данных.

В соответствии с этими обязанностями, давайте рассмотрим некоторые навыки, которые необходимы для аналитика данных.

Процесс анализа данных требует сочетания технических, функциональных и "мягких" навыков.

Давайте сначала рассмотрим некоторые технические навыки, которые понадобятся вам в роли аналитика данных.

К ним относятся:

- Опыт использования электронных таблиц, таких как Microsoft Excel или Google Sheets, владение инструментами и

программным обеспечением для статистического анализа и визуализации, такими как IBM Cognos, IBM SPSS, Oracle Visual Analyzer, Microsoft Power BI, SAS и Tableau.
- Владение хотя бы одним из языков программирования, таких как R, Python, и в некоторых случаях C++, Java и MATLAB,
- Хорошее знание SQL и умение работать с данными в реляционных и NoSQL базах данных,
- Способность получать доступ и извлекать данные из хранилищ данных, таких как карты данных, хранилища данных, озера данных и другие, конвейеры данных и другие.
- Знакомство с инструментами обработки больших данных, такими как Hadoop, Hive и Spark.

Мы узнаем больше об особенностях и вариантах использования некоторых из этих языков программирования, баз данных и инструментов обработки больших данных.

Теперь мы рассмотрим некоторые функциональные навыки, необходимые для роли аналитика данных.

К ним относятся:

- Знание статистики, которое поможет вам анализировать данные, проверять результаты анализа и выявлять заблуждения и логические ошибки.
- Аналитические навыки, которые помогут вам исследовать и интерпретировать данные, теоретизировать и делать прогнозы.
- Навыки решения проблем, поскольку конечной целью любого анализа данных является решение проблем.
- Навыки зондирования, которые необходимы для процесса обнаружения, то есть для понимания проблемы с точки зрения различных заинтересованных сторон и пользователей - потому что процесс анализа данных начинается с четкой формулировки проблемы и желаемого результата.
- Навыки визуализации данных, которые помогут вам выбрать методы и инструменты для эффективного представления ваших результатов анализа в зависимости от аудитории, типа данных, контекста и конечной цели анализа.
- Навыки управления проектом, позволяющие управлять процессом, людьми, зависимостями и сроками реализации инициативы.

Это подводит нас к мягким навыкам аналитика данных.

Анализ данных - это и наука, и искусство.

Вы можете стать асом в технических и функциональных знаниях, но одним из ключевых факторов, определяющих ваш успех, будут мягкие навыки.

Сюда входит:

- ваша способность работать в сотрудничестве с деловыми и межфункциональными командами;
- эффективно общаться, чтобы докладывать и представлять свои результаты;
- рассказывать убедительную историю; и заручиться поддержкой вашей работы.

Прежде всего, в основе анализа данных лежит любознательность.

В ходе своей работы вы будете натыкаться на закономерности, явления и аномалии, которые могут показать вам другой путь.

Способность позволить новым вопросам всплыть на поверхность и бросить вызов вашим предположениям и гипотезам делает вас отличным аналитиком.

Вы также можете услышать, как специалисты по анализу данных говорят об интуиции как об обязательном качестве.

Важно отметить, что интуиция в данном контексте - это способность предчувствовать будущее на основе распознавания образов и гипотез и прошлого опыта.

2 ЛЕКЦИЯ. ТОЧКИ ЗРЕНИЯ: КАЧЕСТВА И НАВЫКИ АНАЛИТИКА ДАННЫХ

специалисты по данным рассказывают о качествах и навыках, необходимых для того, чтобы стать аналитиком данных.

Аналитик данных, это человек любопытного от природы. Тот, кто внимателен к деталям и любит работать с компьютерами. Любопытный человек будет искать ответы даже иногда, когда нет вопроса, или они не против исследовать и искать в областях, которые, возможно, не были продуманы ранее.

Должно быть внимание к деталям или поиск закономерностей.

Входите ли вы в комнату и просто считаете, естественно, людей, как обставлена комната, обращая внимание на мелкие детали, и затем наслаждаетесь компьютерами, потому что технологии развиваются так быстро.

Что-то или навык, которому вы научились сегодня, через два-три года может оказаться неприменимым.

Поэтому вам нужно быть способным развивать новые навыки и изучать новое программное обеспечение в зависимости от того, как изменился рынок или отрасль.

Требуются технические навыки и мягкие навыки.

Технические навыки включают в себя python, SQL, R, tableau и power BI, а мягкие навыки или навыки межличностного общения означают, знаете ли вы, какие правильные данные использовать и какой правильный инструмент использовать и как представить данные ключевым заинтересованным сторонам.

И эти навыки требуют деловой хватки и навыков презентации.

Вы должны быть очень ориентированы на детали, вы должны любить цифры, вы должны любить информацию и быть готовым изучать эту информацию и не просто смотреть на поверхность, а погружаться глубже. Так, например, в нашей работе я не могу просто принять банковскую выписку за чистую монету, я должен посмотреть на нее и сравнить. Правильно ли выглядит печать? Особенно в современном мире существует множество случаев мошенничества, недопонимания и людей, которые пытаются получить вашу информацию и использовать ее обманным путем.

Поэтому хороший аналитик данных должен уметь сравнивать прошлогоднюю информацию с информацией этого года, чтобы понять, соответствует ли она действительности, чтобы проверить, правильно ли она выглядит.

Должны ли вы обладать их взглядом и мышлением и а не просто принимать все за чистую монету?

Есть много качеств и навыков, необходимых для работы аналитиком данных, и я делю их на две группы: мягкие навыки и технические навыки.

Я думаю, что самые важные "мягкие" навыки для аналитика данных - это быть действительно любопытным, задавать много хороших вопросов, быть действительно вдумчивым и внимательно слушать, и понимать, как точку зрения пользователя, так и ваших коллег и что им больше всего нужно от данных, и всегда быть готовым учиться, потому что аналитика - это быстро развивающаяся область, поэтому вам придется постоянно учиться и читать, чтобы оставаться на вершине.

Существует множество технических навыков, необходимых для работы аналитиком данных. Самым важным техническим навыком для начинающих аналитиков данных является SQL. Это, безусловно, наиболее широко используемый язык, и каждый раз, когда вы извлекаете данные из базы данных, вам понадобится знание SQL.

И нет ничего лучше, чем аналитик данных с действительно хорошим знанием SQL.

Я думаю, что иногда люди забегают вперед и пробуют множество очень сложных технологий, прежде чем освоить основы SQL. Я думаю, что это действительно большая ошибка. Я думаю, что всегда полезно знать Python и R, которые являются двумя основными языками программирования, используемыми для анализа данных.

Я думаю, что как начинающему аналитику данных, вам не обязательно владеть обоими или одним из них. Но начать хорошо разбираться в одном или другой, будет очень полезно для вашей карьеры.

Еще один важный технический навык, которым должен обладать аналитика данных - это владеть хотя бы одним инструментом визуализации данных и понимать общие принципы визуализации данных.

Сегодня набор навыков аналитика данных гораздо более динамичен, динамичнее, чем раньше.

Поэтому аналитику данных необходимо знать, какую проблему он пытается решить с помощью данных. Вытаскивать эти данные по мере необходимости, в нужной им структуре с помощью SQL из озера данных, в котором они находятся.

Вы знаете, что там будет много разных таблиц, и им нужно будет понять, как их объединить, а затем извлечь данные, очистить их, обработать, манипулировать ими, добывать их, чтобы они могли извлечь из них полезные сведения.

Представить эти выводы кратко, четко, используя хорошие визуализации и приборных панелей, и, другими словами, уметь рассказать хорошую историю с помощью этих данных.

3 ЛЕКЦИЯ. ОДИН ДЕНЬ ИЗ ЖИЗНИ АНАЛИТИКА

День из жизни аналитика данных может включать в себя множество возможностей - от получения данных из различных источников до создания запросов для получения данных от получения данных из различных источников данных до создания запросов для извлечения данных из хранилища данных, перебирание рядов данных в поисках нужных идей, создания отчетов и информационных панелей, и взаимодействия с заинтересованными сторонами для сбора информации и представления выводов, это целый спектр работ. И да, самое важное - очистка и подготовка данных, чтобы выводы имели достоверную основу.

Это является значительной частью того, чем может заниматься любой аналитик данных в своей работе.

Привет. Я Сиварам Джалади.

Я работаю аналитиком данных в Fluentgrid, компании по разработке технологических решений для интеллектуальных сетей расположенной в городе Вишакхапатнам в Индии. Fluentgrid является партнером IBM и лауреатом премии IBM Beacon за свои решения в сегментах "умной" энергетики и "умного" города.

Мы предлагаем интегрированные решения операционного центра для энергетических компаний и "умных" городов, используя нашей платформы оперативного анализа, известной как Fluentgrid Actilligence.

Наш клиент, энергоснабжающая компания в Южной Индии, заметила всплеск жалоб

по поводу завышенных счетов. И частота этих жалоб говорит о том, что здесь есть что-то большее чем просто случайность.

Поэтому меня попросили изучить жалобы и данные по выставленным счетам и посмотреть, смогу ли я что-то обнаружить.

Я начинаю с оценки того, что у меня есть.

Некоторые из очевидных мест, которые, как я знаю, мне предстоит изучить, это данные о жалобах, данные об абонентах и данные о выставленных счетах.

Это и будет моей отправной точкой.

Прежде чем я погружусь в специфику данных, я составлю список вопросов, первоначальных гипотез, с которых я собираюсь начать.

Например, модель потребления энергии абонентами, сообщающих об этой проблеме:

Есть ли диапазон потребления, в котором завышение платы по счетам происходит чаще, чем в других?

Есть ли области концентрации этих жалоб в зависимости от района:

Концентрируются ли жалобы в конкретных населенных пунктах города?

Частота и распространенность жалоб по отдельным абонентам:

- Повторяются ли жалобы на завышение счетов от одних и тех же абонентов?
- Если да, то какова частота повторных случаев?

- Если абоненту один раз выставили завышенный счет, происходит ли завышение счета каждый месяц, начиная с первого.
- или повторные случаи происходят спорадически или вообще не происходят?

По мере того, как я проясняю свои первоначальные гипотезы и набор вопросов, с которых я собираюсь начать.

Далее я определяю наборы данных, которые я собираюсь выделить и проанализировать, чтобы подтвердить или опровергнуть свои гипотезы.

Я извлекаю средние годовые, квартальные и месячные суммы счетов заявителей и ищу диапазон, в котором количество жалоб больше, чем в других.

Затем я просматриваю данные о местонахождении заявителей, чтобы выяснить, есть ли связь между завышенными счетами и почтовыми индексами.

Здесь я вижу, что, похоже, жалобы сосредоточены в определенных районах.

Это выглядело так, будто может что-то добавить. Поэтому вместо того, чтобы перейти к третьей гипотезе, я решил немного углубиться в эти данные.

Далее я извлекаю данные о дате подключения.

Более 95% заявителей были нашими абонентами более семи лет, хотя не все абоненты старше семи лет сталкивались с этой жалобой.

Итак, мы видим некоторую концентрацию по районам, и мы видим значительную концентрацию жалоб, основанных на дате подключения.

Далее, я вытаскиваю марку и серийный номер счетчиков. И вот оно - серийные номера принадлежат одной и той же партии счетчиков, поставленных одним и тем же поставщиком. Концентрация этих счетчиков, а значит и жалоб, происходила из районов, в которых были установлены эти счетчики.

На данном этапе я чувствую себя уверенно, представляя эти выводы заинтересованным сторонам. Я также собираюсь поделиться источниками данных и процессом, в ходе которого я пришел к этому анализу. Это всегда способствует повышению достоверности выводов.

На этом проект может закончиться, а может и вернуться. Может быть, те же жалобы с другими общими чертами, или совершенно другой набор жалоб, на которые нам нужно найти ответы.

4 ЛЕКЦИЯ: ТОЧКИ ЗРЕНИЯ: ПРИЛОЖЕНИЯ АНАЛИЗА ДАННЫХ

Применение аналитики данных в современном мире повсеместно.

Каждый рекламный ролик, который вы видите, кто-то должен был проанализировать и определить либо для потребителя, либо для компании, какой информацией они хотят поделиться.

Так, вы знаете, что четыре из 10 стоматологов или вы увидите информацию, связанную с количеством калорий или реакции на определенные вещи, - все это требует анализа.

Это не то, о чем следует думать отдельно и отдельно, это то, что мы делаем каждый день в нашей жизни. Даже люди, следящие за уровнем сахара при диабете, всегда проводится анализ, так что приложения универсальны.

Итак, самое замечательное в аналитике в наше время - это то, что она очень широко распространена и применима. Каждая отрасль, каждая вертикаль, каждая функция в данной организации может извлечь выгоду из данных и аналитики.

Занимаетесь ли вы анализом процесса продаж, занимаетесь ли вы финансовой отчетностью в конце месяца, создавая предопределенные и стандартизированные отчеты. Или если вы занимаетесь чем-то вроде планирования численности персонала или анализа численности персонала, как я уже сказал, все это можно использовать в любой вертикали, будь то авиакомпании, фармацевтика, банковское дело, все эти и функции внутри них могут выиграть от аналитики.

И в этом климате, в котором мы сейчас находимся в связи с пандемией, есть компании, которые уделяют пристальное внимание покупательским привычкам своих клиентов. Очевидно, что они могут отличаться от того, что ожидали эти компании.

Поэтому сейчас аналитика данных становится более важной, потому что им необходимо убедиться в том, что они могут поворачиваться и идти в ногу со спросом.

И действительно быть в состоянии удовлетворить потребности своих клиентов и покупателей.

все больше и больше применений альтернативной аналитики данных в мире финансов. Например, мы можем использовать анализ настроений твитов и новых историй в дополнение к традиционному финансовому анализу и для более эффективного инвестирования, то есть для принятия лучших инвестиционных решений.

Кроме того, данные спутниковых снимков можно использовать для отслеживания развития промышленной деятельности, а данные о

геолокации можно использовать для отслеживания посещаемости магазинов и для прогнозирования объема продаж.

5 ЛЕКЦИЯ. РЕЗЮМЕ И ОСНОВНЫЕ МОМЕНТЫ

Роль аналитика данных охватывает следующее:

- Приобретение данных, которые лучше всего подходят для конкретного случая использования.
- Подготовка и анализ данных для понимания того, что они собой представляют.
- Интерпретация и эффективное донесение информации до заинтересованных сторон, которым необходимо действовать в соответствии с полученными результатами.
- Обеспечение документирования процесса для последующего использования и повторения.

Для успешного выполнения этой роли аналитикам данных необходимо сочетание технических, функциональных и "мягких" навыков:

- Технические навыки включают в себя различные уровни владения навыками использования электронных таблиц, статистических инструментов, инструментов визуализации, языков программирования и запросов, а также умение работать с различными типами хранилищ данных и платформами больших данных.
- Понимание статистики, аналитические методы, решение проблем, способность рассматривать ситуацию с разных точек зрения, визуализация данных и навыки управления проектами - все это относится к функциональным навыкам, которые необходимы аналитику данных для эффективной работы.
- Мягкие навыки включают в себя умение работать в сотрудничестве, эффективно общаться, рассказывать убедительную историю с помощью данных, заручаться поддержкой и поддержкой заинтересованных сторон. Любопытство к изучению различных путей и интуиция, которая помогает предвидеть будущее на основе прошлого опыта, также являются необходимыми навыками для того, чтобы стать хорошим аналитиком данных.

МОДУЛЬ 3 - ЭКОСИСТЕМА ДАННЫХ И ЯЗЫКИ ДЛЯ СПЕЦИАЛИСТОВ ПО ДАННЫМ

1 ЛЕКЦИЯ. РЕПОЗИТОРИИ. ЭКОСИСТЕМА АНАЛИТИКА

Экосистема аналитика данных включает инфраструктуру, программное обеспечение, инструменты, рамки и процессы, используемые для сбора, очистки, анализа, добычи и визуализации данных.

В этом видео мы сделаем краткий обзор экосистемы, а затем подробно рассмотрим каждой из этих тем в последующих видео.

Сначала поговорим о данных.

В зависимости от того, насколько четко определена структура данных, их можно разделить на структурированные, полу-структурированные или неструктурированные.

Данные, которые имеют строгий формат и могут быть аккуратно организованы в строки и столбцы, - **это структурированные дан**ные. Это те данные, которые вы обычно видите, например, в базах данных и электронных таблицах.

Полу-структурированные данные - это смесь данных, которые имеют последовательные характеристики, и данных, которые не соответствуют жесткой структуре. Например, электронная почта. Электронное письмо содержит смесь структурированных данных, таких как имя отправителя и получателя, а также содержимое электронного письма, что является неструктурированными данными.

И еще есть **неструктурированные данные**:

Данные, которые являются сложными, и в основном качественная информация, которую невозможно свести к строкам и столбцам.

Например, фотографии, видео, текстовые файлы, PDF и содержимое социальных сетей. *Тип данных определяет тип хранилищ данных, в которых эти данные могут собираться и храниться,* а также инструменты, которые можно использовать для запроса или обработки данных.

Данные также поступают в самых разнообразных форматах файлов, собираемых из различных источников данных, начиная от реляционных и нереляционных баз данных, API, веб-сервисов, потоков данных, социальных платформ и сенсорных устройств.

Это подводит нас к хранилищам данных:

Этот термин включает в себя базы данных, хранилища данных, карты данных, озера данных и хранилища больших данных. Тип, формат и источники данных влияют на тип хранилищ данных, которые вы можете использовать для сбора, хранения, очистки, анализа и добычи данных для анализа.

Например, если вы работаете с большими данными, вам понадобятся хранилища больших данных, которые позволяют хранить и обрабатывать большие объемы данных с высокой скоростью, а также фреймворки,

которые позволяют вам выполнять сложную аналитику в реальном времени на больших данных.

Экосистема также включает языки, которые можно классифицировать как языки запросов, языки программирования, языки оболочки и сценариев. От запросов и манипулирования данными с помощью SQL до разработки приложений для работы с данными с помощью Python, и написания сценариев оболочки для повторяющихся операционных задач - это важные компоненты аналитика данных.

Автоматизированные инструменты, фреймворки и процессы для всех этапов аналитического процесса являются частью экосистемы аналитиков данных.

От инструментов, используемых для сбора, извлечения, преобразования и загрузки данных в хранилища данных, до инструментов для сбора данных, очистки данных, добычи данных, анализа и визуализации данных - это очень разнообразная и богатая экосистема.

Электронные таблицы, блокноты Jupyter Notebook и IBM Cognos - вот лишь несколько примеров.

2 ЛЕКЦИЯ. ТИПЫ ДАННЫХ

Данные - это неорганизованная информация, которая обрабатывается для придания ей смысла. Как правило, данные состоят из фактов, наблюдений, представлений, чисел, знаков, символов, и изображения, которые могут быть интерпретированы для извлечения смысла.

Одним из способов классификации данных является их структура - данные могут быть:

- Структурированными;
- Полу-структурированными или
- неструктурированными.

Структурированные данные имеют четко определенную структуру или придерживаются определенной модели данных могут храниться в четко определенных схемах, таких как базы данных и *во многих случаях могут быть представлены в табличной форме со строками и столбцами.*

Структурированные данные - это объективные факты и числа, которые можно собирать, экспортировать, хранить и организовывать в типичных базах данных.

Некоторые из источников структурированных данных могут включать:

- базы данных SQL и системы онлайновой обработки транзакций (или OLTP), ориентированные на деловые операции.

- Электронные таблицы, такие как Excel и Google Spreadsheets
- Онлайновые формы
- Датчики, такие как системы глобального позиционирования (или GPS) и радиочастотной идентификации (или RFID) метки; и журналы сети и веб-сервера.

Как правило, структурированные данные можно хранить в реляционных базах данных или базах данных SQL.

Кроме того, структурированные данные можно легко исследовать с помощью стандартных методов и инструментов анализа данных.

Полуструктурированные данные - это данные, которые имеют некоторые организационные свойства, но не имеют фиксированной или жесткой схемы. Полуструктурированные данные не могут храниться в виде строк и столбцов, как в базах данных.

Они содержат: *теги и элементы, или метаданные, которые используются для группировки данных и их организации* в иерархию.

Некоторые из источников полу-структурированных данных могут включать:

- Электронную почту
- XML и другие языки разметки
- двоичные исполняемые файлы
- пакеты TCP/IP
- Заархивированные файлы

Интеграция данных из различных источников XML и JSON позволяют пользователям определять теги и атрибуты для хранения данных в иерархической и широко используются для хранения и обмена полу-структурированными данными.

Неструктурированные данные - это данные, которые не имеют легко идентифицируемой структуры и, следовательно, не могут быть организованы в обычной реляционной базе данных в виде строк и столбцов. Они не имеют определенного формата, последовательности, семантики или правил.

Неструктурированные данные могут работать с неоднородностью источников и имеют разнообразные приложения для бизнес-аналитики и анализа.

Некоторые из источников неструктурированных данных могут включать:

- веб-страницы
- ленты социальных сетей
- Изображения в различных форматах (например, JPEG, GIF и PNG)
- Видео- и аудиофайлы

- Документы и файлы PDF
- презентации PowerPoint
- Журналы СМИ; и
- Опросы

Неструктурированные данные могут храниться в файлах и документах (таких как Word doc) для ручного анализа или в базах данных NoSQL, которые имеют свои собственные инструменты анализа для изучения этого типа данных.

Подводя итог, можно сказать, что:

- **Структурированные данные** - это данные, хорошо организованные в форматах, которые могут храниться в базах данных и поддаются стандартным методам и инструментам анализа данных;
- **Полуструктурированные** данные - это данные, которые в некоторой степени организованы и полагаются на метатеги для группировки и иерархии;
- **Неструктурированные данные** - это данные, которые не организованы в виде строк и столбцов в определенном формате.

3 ЛЕКЦИЯ. ФАЙЛОВЫЕ СТРУКТУРЫ

Как специалист по работе с данными, вы будете работать с различными типами и форматами файлов данных.

Важно понимать структуру, лежащую в основе форматов файлов, а также их преимуществами и ограничениями. Это понимание поможет вам принять правильное решение о выборе форматов, которые лучше всего подходят для ваших данных и потребностей в производительности.

Некоторые из стандартных форматов файлов, которые мы рассмотрим в этом видео, включают:

- Делимитированные форматы текстовых файлов,
- Microsoft Excel Open XML Spreadsheet, или XLSX
- Расширяемый язык разметки, или XML,
- Portable Document Format, или PDF,
- JavaScript Object Notation, или JSON,

Текстовые файлы с разделителями - это текстовые файлы, используемые для хранения данных в виде текста, в котором каждая строка или строка содержит значения, разделенные разделителем **«;»,** где разделитель - это последовательность из одного или нескольких символов для указания границы между независимыми сущностями или значениями.

Для разделения значений может использоваться любой символ, но наиболее распространенными разделителями являются запятая, табуляция, двоеточие, вертикальная полоса и пробел.

Значения, разделенные запятыми (или **CSV**) и значения, разделенные табуляцией (или **TSV**), являются наиболее часто используемыми типы файлов в этой категории.

В CSV разделителем является запятая, а в TSV - табуляция.

Когда в текстовых данных присутствуют буквальные запятые, их нельзя использовать в качестве разделителя, TSV служат альтернативой формату CSV.

В бегущем тексте остановки табуляции встречаются нечасто. Каждая строка, или горизонтальная линия, в текстовом файле имеет набор значений, разделенных разделителем и представляет собой запись.

Первая строка работает как заголовок столбца, где каждый столбец может иметь различный тип данных. Например, один столбец может иметь тип даты, а другой - тип строки или целого числа.

Делимитированные файлы допускают значения полей любой длины и считаются стандартным форматом для предоставления простой схемы информации. Они могут обрабатываться практически всеми существующими приложениями. Разделители также представляют собой одно из различных средств для указания границ в потоке данных.

Microsoft Excel Open XML Spreadsheet, или XLSX, - это формат файлов Microsoft Excel Open XML который относится к формату файлов электронных таблиц. Это формат файлов на основе XML, созданный компанией Microsoft.

В файле .XLSX, также известном как рабочая книга, может быть несколько рабочих листов.

Каждый рабочий лист организован в строки и столбцы, на пересечении которых находится ячейка. Каждая ячейка содержит данные.

XLSX использует открытый формат файла, что означает, что он доступен для большинства других приложений. Он может использовать и сохранять все функции, доступные в Excel, а также известен как один из ***более безопасных форматов файлов, поскольку в нем невозможно сохранить вредоносный код.***

Расширяемый язык разметки, или XML, - это язык разметки с установленными правилами кодирования данных.

Формат файлов XML читается как людьми, так и машинами. Это самоописывающийся язык, предназначенный для передачи информации через Интернет. XML в некоторых отношениях похож на HTML, но имеет и отличия. Например, в .XML не используются предопределенные теги,

как в .HTML. XML не зависит от платформы и языка программирования и поэтому упрощает обмен данными между различными системами.

Portable Document Format, или PDF, - это формат файлов, разработанный компанией Adobe для представления документов независимо от прикладного программного обеспечения, аппаратного обеспечения и операционных систем, что означает, что его можно что означает, что документы можно просматривать одинаково на любом устройстве. Этот формат часто используется в юридических и финансовых документах, а также может применяться для заполнения данных, например, в формах.

JavaScript Object Notation, или JSON, - это текстовый открытый стандарт, разработанный для передачи структурированных данных через Интернет.

Формат файла представляет собой независимый от языка формат данных, который может быть прочитан в любом языке программирования. JSON прост в использовании, совместим с широким спектром браузеров и считается одним из лучших инструментов для обмена данными любого размера и типа, даже аудио и видео. По этой причине многие API и веб-службы возвращают данные в формате JSON.

4 ЛЕКЦИЯ. ИСТОЧНИКИ ДАННЫХ

Как мы уже говорили в одном из наших предыдущих видеороликов, источники данных никогда не были такими динамичными и разнообразными, как сегодня. и разнообразными, как сегодня.

В этом видео мы рассмотрим некоторые распространенные источники, такие как:

- Реляционные базы данных,
- плоские файлы и наборы данных XML,
- API и веб-сервисы,
- веб-скрапинг,
- потоки данных и каналы.

Как правило, организации имеют внутренние приложения для поддержки управления повседневной деятельностью, операциями с клиентами, персоналом и рабочими процессами, бизнес-деятельностью, операциями с клиентами, работой с персоналом и рабочими процессами.

Эти системы используют реляционные базы данных, такие как SQL Server, Oracle, MySQL и IBM DB2, для хранения данных в структурированном виде. Данные, хранящиеся в базах и хранилищах данных, могут быть использованы в качестве источника для анализа.

Например, данные из системы розничных операций могут быть использованы для анализа продаж в различных регионах, а данные из системы управления взаимоотношениями с клиентами можно использовать для составления прогнозов продаж.

Вне организации существуют и другие общедоступные и частные наборы данных. Например, правительственные организации постоянно публикуют демографические и экономические данные.

Существуют также компании, которые продают конкретные данные, например, данные о точках продаж или финансовые данные, или данные о погоде, которые предприятия продают на рынке или данные о погоде, которые предприятия могут использовать для определения стратегии, прогнозирования спроса и принимать решения, связанные с дистрибуцией или маркетинговыми акциями, среди прочего.

Такие наборы данных обычно предоставляются в виде плоских файлов, файлов электронных таблиц или XML-документов.

Плоские файлы хранят данные в обычном текстовом формате, с одной записью или строкой в строке, и каждое значение отделяется разделителями, такими как запятые, точки с запятой или символы табуляции.

Данные в плоском файле отображаются в одной таблице, в отличие от реляционных баз данных, которые содержат несколько таблиц. Одним из наиболее распространенных форматов плоских файлов является CSV, в котором значения разделяются запятыми.

Файлы электронных таблиц - это особый тип плоских файлов, которые также организуют данные в табличном формате - строки и столбцы.

Но электронная таблица может содержать несколько рабочих листов, и каждый рабочий лист может быть сопоставлен с отдельной таблицей. Хотя данные в электронных таблицах представлены в виде обычного текста, файлы могут храниться в пользовательских форматах и включать дополнительную информацию, такую как форматирование, формулы и т.д.

Microsoft Excel, который хранит данные в формате .XLS или .XLSX, вероятно, является самой распространенной электронной таблицей.

Другие форматы включают Google sheets, Apple Numbers и LibreOffice.

Файлы XML содержат значения данных, которые идентифицируются или маркируются с помощью тегов. В то время как данные в плоских файлах являются "плоскими" или отображаются в одной таблице, XML-файлы могут поддерживать более сложные структуры данных, например, иерархические.

Некоторые распространенные виды использования XML включают данные из онлайн-опросов, банковских выписок и других неструктурированные наборы данных.

Многие поставщики данных и веб-сайты предоставляют API, или интерфейсы прикладных программ, и веб-сервисы, с которыми могут взаимодействовать несколько пользователей или приложений и получать данные для обработки или анализа.

API и веб-сервисы обычно ожидают входящих запросов, которые могут быть в форме веб-запросов от пользователей или сетевых запросов от приложений, и возвращают данные в виде обычного текста, XML, HTML, JSON или медиафайлов.

Давайте рассмотрим несколько популярных примеров использования API в качестве источника данных для анализа данных:

Использование API Twitter и Facebook для получения данных из твитов и постов для выполнения таких задач, как поиск мнений или анализ настроений, который заключается в обобщении количества положительных и критических отзывов по определенному вопросу, такому как политика правительства, продукт, услуга или удовлетворенность клиентов в целом.

API фондового рынка, используемые для получения таких данных, как цены на акции и товары, прибыль на акцию и исторические цены, для торговли и анализа.

API поиска и проверки данных, которые могут быть очень полезны для аналитиков данных для очистки и подготовки данных, а также для соотнесения данных - например, для проверки того, к какому городу или, например, проверить, к какому городу или штату относится тот или иной почтовый или почтовый индекс.

API также используются для извлечения данных из источников баз данных, как внутри организации, так и за ее пределами.

Веб-скрейпинг используется для извлечения нужных данных из неструктурированных источников.

Известный также под названиями screen scraping, web harvesting и web data extraction, web scraping позволяет загружать определенные данные из неструктурированных источников. Таких как веб-страницы на основе заданных параметров.

Веб-скрейперы могут, помимо прочего, извлекать текст, контактную информацию, изображения, видео, товары и многое другое с веб-сайта. Некоторые популярные виды использования веб-скрейпинга включают:

- сбор информации о товарах с сайтов розничной торговли, производителей и сайтов электронной коммерции для обеспечения
- сравнение цен,
- получение информации о продажах из открытых источников данных,
- извлечение данных из сообщений и авторов на различных форумах и в сообществах,
- а также сбор обучающих и тестовых наборов данных для моделей машинного обучения.

Некоторые из популярных инструментов для веб-скреппинга включают BeautifulSoup, Scrapy, Pandas и Selenium.

Потоки данных - еще один широко используемый источник для агрегирования постоянных потоков данных, поступающих из таких источников, как приборы, IoT-устройства и приложения, GPS-данные с автомобилей, компьютерные программы, веб-сайты и социальные сообщения.

Эти данные, как правило, имеют временные метки, а также геометки для географической идентификации.

Некоторые потоки данных и способы их использования включают:

- биржевые и рыночные тикеры для финансовой торговли,
- потоки розничных транзакций для прогнозирования спроса и управления цепочками поставок,
- видеонаблюдение и видеоматериалы для обнаружения угроз,
- потоки данных социальных сетей для анализа настроений,
- потоки данных датчиков для мониторинга промышленного или сельскохозяйственного оборудования,
- потоки веб-кликов для мониторинга производительности веб-сайтов и улучшения дизайна,
- и события авиарейсов в реальном времени для перебронирования и изменения расписания.

Некоторые популярные приложения, используемые для обработки потоков данных, включают Apache Kafka, Apache, Spark Streaming и Apache Storm.

RSS (или Really Simple Syndication) каналы - еще один популярный источник данных.

Они обычно используются для сбора обновленных данных с онлайн-форумов и новостных сайтов, где данные обновляются на постоянной основе.

Используя программу для чтения фидов, которая представляет собой интерфейс, преобразующий текстовые файлы RSS в поток обновленных данных, обновления передаются на пользовательские устройства.

5 ЛЕКЦИЯ. ЯЗЫКИ ДЛЯ ПРОФЕССИОНАЛОВ В ОБЛАСТИ ДАТА АНАЛИТИКИ

В этом видео мы познакомимся с некоторыми языками, имеющими отношение к работе специалистов по работе с данными. Их можно разделить на следующие категории: языки запросов, языки программирования и shell-сценарии. Владение хотя бы одним языком из каждой категории необходимо для любого специалиста по работе с данными.

Проще говоря:

Языки запросов предназначены для доступа к данным в базе данных и манипулирования ими; например, SQL.

Языки программирования предназначены для разработки приложений и управления поведением приложений; например, Python, R и Java; и языки оболочки и сценариев, такие как Unix/Linux Shell и PowerShell, идеально подходят для повторяющихся и трудоемких операционных задач.

SQL, или язык структурированных запросов, - это язык запросов, предназначенный для доступа к информации из реляционных баз данных, в основном, хотя и не только, реляционных баз данных и манипулирования ими. Используя SQL, мы можем написать набор инструкций для выполнения таких операций, как вставка, обновление и удаление записей в базе данных; создавать новые базы данных, таблицы и представления; и писать хранимые процедуры - что означает, что вы можете написать набор инструкций и вызвать их для последующего использования.

Вот некоторые преимущества использования SQL:

SQL является переносимым и может использоваться независимо от платформы, его можно использовать для запросов к данным в самых разных базах данных и хранилищах данных, хотя каждый поставщик может иметь некоторые вариации и специальные расширения. Он имеет простой синтаксис, схожий с английским языком. Его синтаксис позволяет разработчикам писать программы с меньшим количеством строк, чем некоторые другие языки программирования.

Используя основные ключевые слова, такие как select, insert, into и update, он может быстро и эффективно извлекать большие объемы данных. Он работает в системе интерпретаторов, что означает, что код может быть выполнен сразу после его написания, что делает создание прототипов быстрым и легким.

SQL - один из самых популярных языков запросов. Благодаря большому сообществу пользователей и огромному объему документации, накопленному за многие годы, SQL продолжает оставаться единым языком запросов.

Благодаря большому сообществу пользователей и огромному объему документации, накопленной за многие годы, он продолжает оставаться единой платформой для всех своих пользователей по всему миру.

Python - это широко используемый язык программирования высокого уровня общего назначения с открытым исходным кодом. Его синтаксис позволяет программистам выражать свои концепции в меньшем количестве строк кода, по сравнению с некоторыми более старыми языками.

Python считается одним из самых простых языков для изучения и имеет большое сообщество разработчиков.

Благодаря своей ориентации на простоту и читабельность, а также низкой кривой обучения, он является идеальным инструментом для начинающих программистов.

Он отлично подходит для выполнения высокопроизводительных вычислительных задач на огромных объемах данных, которые в противном случае могут быть чрезвычайно трудоемкими и громоздкими.

Python предоставляет такие библиотеки, как Numpy и Pandas, которые облегчают эту задачу за счет использования параллельной обработки. В нем есть встроенные функции почти для всех часто используемых концепций.

Python поддерживает множество парадигм программирования, таких как объектно-ориентированное, императивное, функциональное, и процедурные, что делает его подходящим для широкого спектра случаев использования.

Теперь давайте рассмотрим некоторые причины, которые делают Python одним из самых быстрорастущих языков программирования в мире.

Он прост в изучении - Python позволяет использовать меньшее количество строк кода для выполнения задач по сравнению с другими языками. Это язык с открытым исходным кодом - Python является бесплатным и использует модель разработки на основе сообщества. Он работает в средах Windows и Linux и может быть перенесен на различные платформы. Он пользуется широкой поддержкой сообщества и имеет множество полезных аналитических библиотек.

В нем есть несколько библиотек с открытым исходным кодом для работы с данными, визуализации данных, статистики, и математики, и это лишь некоторые из них.

Его обширный набор библиотек и функциональных возможностей также включает:

- Pandas для очистки и анализа данных,
- Numpy и Scipy для статистического анализа,
- Beautifulsoup и Scrapy для веб-скрейпинга,
- Matplotlib и Seaborn для визуального представления данных в виде гистограмм, гистограмм,
- и круговых диаграмм,
- Opencv для обработки изображений.

R - это язык программирования с открытым исходным кодом и среда для анализа данных, визуализации данных, машинного обучения и статистики. Широко используемый для разработки статистического программного обеспечения и проведения анализа данных, он особенно известен своей способностью создавать убедительные визуализации, что дает ему преимущество перед некоторыми другими языками в этой области.

К основным преимуществам R относятся следующие:

- Это независимый от платформы язык программирования с открытым исходным кодом,
- Он может быть сопряжен со многими языками программирования, включая Python,
- Он обладает высокой расширяемостью, что означает, что разработчики могут продолжать добавлять функциональные возможности, определяя новые функции,
- Он облегчает работу со структурированными и неструктурированными данными, что означает, что он
- имеет более широкие возможности работы с данными,
- В нем есть такие библиотеки, как Ggplot2 и Plotly, которые предлагают эстетические графики для своих пользователей.
- Вы можете создавать отчеты с данными и скриптами, встроенными в них; также интерактивные веб-приложения, которые позволяют пользователям играть с результатами и данными,
- Этот язык является доминирующим среди других языков программирования для разработки статистических инструментов.

Java - это объектно-ориентированный, основанный на классах и платформо-независимый язык программирования, первоначально разработанный компанией Sun Microsystems.

Он занимает одно из первых мест среди используемых сегодня языков программирования. Java используется в ряде процессов на протяжении всего процесса анализа данных, включая очистку данных, импорт и экспорт данных, статистический анализ и визуализация данных.

Фактически, большинство популярных фреймворков и инструментов, используемых для работы с большими данными, как правило, написаны на Java, такие как Hadoop, Hive и Spark.

Он идеально подходит для проектов, критичных к скорости.

Оболочка **Unix/Linux Shell** - это компьютерная программа, написанная для системы UNIX. Это серия команд UNIX, записанных в обычном текстовом файле для выполнения конкретной задачи.

Написание сценария оболочки является быстрым и простым.

Он наиболее полезен для выполнения повторяющихся задач, которые могут занять много времени, если набирать по одной строке за раз.

Типичные операции, выполняемые сценариями оболочки, включают:

- работа с файлами,
- выполнение программ,
- задачи системного администрирования, такие как резервное копирование дисков и оценка системных журналов,
- сценарии установки сложных программ,
- выполнение обычного резервного копирования,
- выполнение пакетов

PowerShell - это кроссплатформенный инструмент автоматизации и конфигурационная среда от Microsoft, который оптимизирован для работы со структурированными форматами данных, такими как JSON, CSV, XML, и REST API, веб-сайтами и офисными приложениями.

Он состоит из оболочки командной строки и языка сценариев.

PowerShell основан на объектах, что позволяет фильтровать, сортировать, измерять, группировать, сравнивать и выполнять многие другие действия над объектами по мере их прохождения через конвейер данных.

Это также хороший инструмент для поиска данных, построения графических интерфейсов, создания диаграмм, приборных панелей и интерактивных отчетов.

ИТОГИ МОДУЛЯ 3:

В этом уроке вы узнали следующую информацию:

Экосистема аналитики данных включает инфраструктуру, программное обеспечение, инструменты, рамки и процессы, используемые для сбора, очистки, анализа, добычи и визуализации данных.

В зависимости от того, насколько четко определена структура данных, данные можно разделить на категории:

- **Структурированные данные**, то есть данные, которые хорошо организованы в форматах, которые можно хранить в базах данных.
- **Полу-структурированные данные**, то есть данные, которые частично организованы и частично имеют свободную форму.
- **Неструктурированные данные**, то есть данные, которые не могут быть организованы условно в строки и столбцы.

Данные поступают в самых разных форматах файлов, таких как разделенные текстовые файлы, электронные таблицы, XML, PDF и JSON, каждый из которых имеет свои преимущества и ограничения в использовании.

Данные извлекаются из различных источников данных, начиная от реляционных и нереляционных баз данных и заканчивая API, веб-сервисами, потоками данных, социальными платформами и сенсорными устройствами.

После того как данные определены и собраны из различных источников, их необходимо поместить в хранилище данных, чтобы подготовить к анализу. Тип, формат и источники данных влияют на тип хранилища данных, которое может быть использовано.

Специалистам по работе с данными необходим целый ряд языков, которые помогут им извлекать, подготавливать и анализировать данные. Их можно классифицировать следующим образом:

- Языки запросов, такие как SQL, используемые для доступа к данным из баз данных и манипулирования ими.
- Языки программирования, такие как Python, R и Java, для разработки приложений и управления поведением приложений.
- Языки оболочки и сценариев, такие как Unix/Linux Shell и PowerShell, для автоматизации повторяющихся операционных задач.

МОДУЛЬ 4 UNDERSTANDING DATA REPOSITORIES AND BIG DATA PLATFORMS

1 ЛЕКЦИЯ. ТИПЫ БАЗ ДАННЫХ

Хранилище данных - это общий термин, используемый для обозначения данных, которые были собраны, организованы, и изолированы таким образом, чтобы их можно было использовать для деловых операций или добывать для отчетности и анализа данных. Это может быть небольшая или большая инфраструктура базы данных с одной или несколькими базами данных, которые собирают, управляют и хранят наборы данных.

В этом видеоролике мы представим обзор различных типов хранилищ, в которых могут находиться ваши данные, таких как базы данных, хранилища данных и хранилища больших данных

Начнем с баз данных. База данных - это собрание данных или информации, предназначенное для ввода, хранения, поиска и поиска, извлечения и изменения данных. А **система управления базами данных, или СУБД**, - это набор программ, которые создают и поддерживают базу данных.

Она позволяет хранить, изменять и извлекать информацию из базы данных с помощью функции *называемой запросом*. Например, если вы хотите найти клиентов, которые были неактивны в течение шести месяцев или более, используя функцию запроса, система управления базой данных получит данные обо всех клиентах из базы данных, которые были неактивны в течение шести месяцев и более.

Несмотря на то, что база данных и СУБД означают разные вещи, эти термины часто используются как взаимозаменяемые. Существуют различные типы баз данных. На выбор базы данных влияют несколько факторов, таких как тип и структура данных, механизмы запросов, требования к латентности, транзакции. механизмы, требования к задержкам, скорость транзакций и предполагаемое использование данных.

Здесь важно упомянуть два основных типа баз данных - **реляционные и нереляционные базы данных.**

Реляционные базы данных, также называемые РСУБД, построены на организационных принципах плоских файлов, где данные организованы в табличном формате со строками и столбцами, следующими четко определенной структуре и схеме.

Однако, в отличие от плоских файлов, РСУБД оптимизированы для операций с данными и запросов, включающих в себя множество таблиц и гораздо большие объемы данных.

Язык структурированных запросов, или SQL, является стандартным языком запросов для реляционных баз данных. Затем появились нереляционные базы данных, также известные как NoSQL, или "Не только SQL".

Нереляционные базы данных появились в ответ на объем, разнообразие и скорость, с которой данные генерируются сегодня, в основном под влиянием достижений в области облачных вычислений, Интернета вещей и распространения социальных сетей.

Созданные для скорости, гибкости и масштабирования, нереляционные базы данных сделали возможным хранение данных в виде без схем или в свободной форме.

NoSQL широко используется для обработки больших данных. Хранилище данных работает как центральное хранилище, которое объединяет информацию, поступающую из разрозненных источников и консолидирует ее с помощью процесса **извлечения, преобразования и загрузки, также известного как процесс ETL**, в одну всеобъемлющую базу данных для аналитики и бизнес-аналитики.

На самом высоком уровне процесс **ETL** помогает вам извлекать данные из различных источников данных, преобразовать данные в чистое и пригодное для использования состояние и загрузить их в *хранилище данных предприятия*.

С хранилищами данных связаны понятия **Data Marts и Data Lakes**, которые мы рассмотрим позже. Марки данных и хранилища данных исторически были реляционными, поскольку большая часть традиционных данных предприятия хранится в RDB.

Однако с появлением технологий NoSQL и новых источников данных стали использоваться ***нереляционные хранилища данных***.
Хранилища данных также используются для создания Хранилищ данных.

Другой категорией хранилищ данных являются хранилища больших данных, которые включают в себя распределенную вычислительную инфраструктуру и инфраструктуру хранения для хранения, масштабирования и обработки очень больших данных.

В целом, хранилища данных помогают изолировать данные и сделать отчетность и аналитику более эффективными и достоверными, а также служат в качестве архива данных.

2 ЛЕКЦИЯ. РЕЛЯЦИОННАЯ БАЗА ДАННЫХ ИЛИ РСУБД

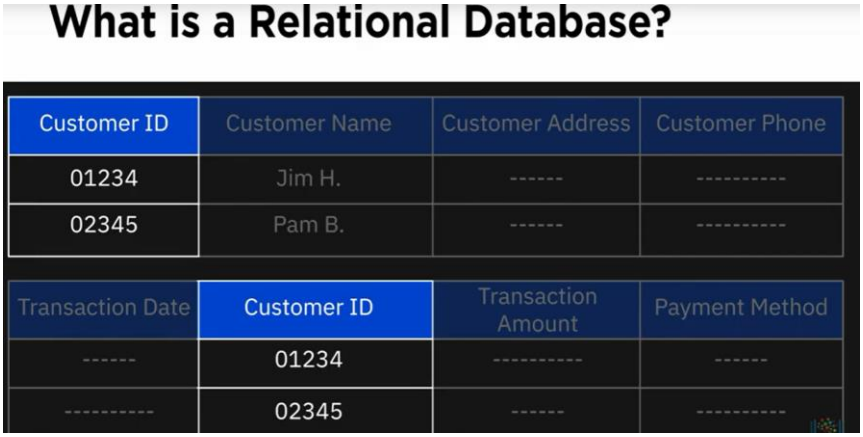
Реляционная база данных - это набор данных, организованных в структуру таблиц, в которой таблицы могут быть связаны, или соотнесены, на основе данных, общих для каждой из них.

Таблицы состоят из строк и столбцов, где строки являются "записями", а столбцы "атрибуты".

Рассмотрим пример таблицы "Клиент", в которой хранятся данные о каждом клиенте в компании. Столбцы, или атрибуты, в таблице клиентов следующие

- «идентификатор» компании,
- название компании,
- адрес компании и
- основной телефон компании;

и каждая строка – это запись о клиенте.



Теперь давайте разберемся, что означает связь между таблицами на основе данных, общих для каждой из них. Наряду с таблицей клиентов, компания также ведет таблицы транзакций, которые содержат данные, описывающие множество отдельных транзакций, относящихся к каждому клиенту.

Столбцы таблицы транзакций могут включать

- дату транзакции,
- идентификатор клиента,
- сумма транзакции
- способ оплаты.

Таблица клиентов и таблицы транзакций могут быть связаны на основе общего поля **Customer ID.**

Вы можете запросить таблицу клиентов для создания отчетов, таких как выписка по клиенту, которая объединяет все транзакции за определенный период. Эта возможность связывания таблиц на основе общих данных позволяет вам получить совершенно новую таблицу из данных, содержащихся в одной или нескольких таблицах с помощью одного запроса.

Это также позволяет понять взаимосвязи между всеми имеющимися данными и получить новые знания для принятия лучших решений.

Реляционные базы данных используют структурированный язык запросов, или SQL, для запроса данных. ***Реляционные базы данных основаны на организационных принципах плоских файлов, таких как электронные таблицы, данные организованы в строки и столбцы в соответствии с четко определенной структурой и схемой.***

Реляционные базы данных по своей конструкции идеально подходят для оптимизированного хранения, поиска и обработки данных для больших объемов данных, в отличие от электронных таблиц, которые имеют ограниченное количество строк и столбцов.

Каждая таблица в реляционной базе данных имеет уникальный набор строк и столбцов, а связи могут быть определены между таблицами, что сводит к минимуму избыточность данных.

Более того, вы можете ограничить поля базы данных определенными типами данных и значениями, что минимизирует несоответствия и приводит к большей согласованности и целостности данных. Реляционные базы данных используют **SQL для запроса дан**ных, что дает вам преимущество в обработке. Вы сможете обрабатывать миллионы записей и извлекать большие объемы данных за считанные секунды.

Более того, архитектура безопасности реляционных баз данных обеспечивает контролируемый доступ к данным, а также обеспечивает соблюдение стандартов и политик управления данными. Реляционные базы данных варьируются от небольших настольных систем до массивных облачных систем.

Реляционные базы данных могут быть:

- с открытым исходным кодом и внутренней поддержкой,
- с открытым исходным кодом и коммерческой поддержкой, или
- коммерческие системы с закрытым исходным кодом.

IBM DB2, Microsoft SQL Server, MySQL, Oracle Database и PostgreSQL - вот некоторые из популярных реляционных баз данных.

Облачные реляционные базы данных, также называемые "база данных как услуга", находят все более широкое применение, поскольку они имеют доступ к безграничным возможностям вычислений и хранения данных, предлагаемым облаком.

Некоторые из **популярных облачных реляционных баз данных** включают Amazon Relational Database Service (RDS), Google Cloud SQL, IBM DB2 on Cloud, Oracle Cloud и SQL Azure.

РСУБД - это зрелая и хорошо документированная технология, что облегчает ее изучение и поиск квалифицированных, талантливых специалистов.

Одним из наиболее значительных преимуществ реляционной базы данных является ее способность создавать значимую информацию путем объединения таблиц.

Некоторые из других преимуществ включают:

Гибкость: Используя SQL, во время работы базы данных и выполнения запросов вы можете добавлять новые столбцы, добавлять новые таблицы, переименовывать отношения и вносить другие изменения.

Сокращение избыточности: Реляционные базы данных сводят к минимуму избыточность данных. Например, информация о клиенте содержится в одной записи в таблице "Клиент", а в таблице транзакций, относящихся к клиенту, хранится ссылка на таблицу "Клиент.

Простота резервного копирования и аварийного восстановления: Реляционные базы данных предлагают простые возможности экспорта и импорта, что делает резервное копирование и восстановление более простым. Экспорт может происходить во время работы базы данных, что упрощает восстановление после сбоя. Облачные реляционные базы данных выполняют непрерывное зеркалирование, что означает, что потеря данных при восстановлении может измеряться секундами или меньше.

Соответствие стандарту ACID: ACID означает атомарность, согласованность, изоляцию и долговечность. Соответствие ACID подразумевает, что данные в базе данных остаются точными и последовательными несмотря на сбои, а транзакции в базе данных обрабатываются надежно.

Теперь мы рассмотрим некоторые случаи использования реляционных баз данных:

Обработка транзакций в режиме онлайн: OLTP-приложения ориентированы на задачи, ориентированные на транзакции, которые выполняются с высокой скоростью. Реляционные базы данных хорошо подходят для OLTP-приложений, потому что они могут обслуживать большое количество пользователей; они поддерживают возможность вставки, обновления или удаления небольших объемов данных; они также поддерживают частые запросы и обновления, а также быстрое время отклика.

Хранилища данных:

В среде хранилищ данных реляционные базы данных могут быть оптимизированы для онлайновой аналитической обработки (или OLAP), где исторические данные анализируются для получения бизнес-аналитики.

IoT-решения:

Решения для Интернета вещей (IoT) требуют скорости, а также способности собирать и обрабатывать данные с пограничных устройств, что требует легковесного решения базы данных.

Это подводит нас к ограничениям РСУБД:

РСУБД плохо работают с полуструктурированными и неструктурированными данными и, следовательно, не подходит для обширной аналитической работы с такими данными.

Для миграции между двумя СУБД схемы и тип данных должны быть идентичными в исходной и конечной таблицах. **Реляционные базы данных имеют ограничение на длину полей данных, что означает, что если вы пытаетесь ввести в поле больше информации, чем оно может вместить, информация не будет будет сохранена.**

Несмотря на ограничения и эволюцию данных в наше время больших данных, облачных вычислений, устройств IoT и социальных сетей, *РСУБД* продолжают оставаться доминирующей технологией для работы со *структурированными данными*.

2 ЛЕКЦИЯ NOSQL

NoSQL, что означает "не только SQL", или иногда "не SQL" - это **нереляционна**я база данных, которая обеспечивает гибкие схемы хранения данных.

Базы данных NoSQL существуют уже много лет, но только недавно стали более популярными в эпоху облаков, больших данных и большого объема веб- и мобильных приложений.

Сегодня их выбирают за такие качества, как масштабируемость, производительность и простота использования. Важно подчеркнуть, что "нет" в слове "NoSQL" - это сокращение от "не только", а не реальное слово "нет".

Базы данных NoSQL создаются для конкретных моделей данных и имеют гибкие схемы, что позволяет программистам создавать и управлять современными приложениями.

Они не используют традиционный дизайн базы данных строка/столбец/таблица с фиксированными схемами и обычно не используют язык структурированных запросов (или SQL) для запроса данных, хотя некоторые из них могут поддерживать SQL или SQL-подобные интерфейсы.

NoSQL позволяет хранить данные без схем или в свободной форме.

Любые данные, будь то структурированные, полу-структурированные или неструктурированные, могут храниться в любой записи. В зависимости от модели, используемой для хранения данных, существует четыре общих типа NoSQL базы данных.

Хранилище ключей-значений, основанные на документах, основанные на столбцах и основанные на графах.

Данные в базе данных типа "ключ-значение" хранятся в виде набора пар "ключ-значение". Ключ представляет собой атрибут данных и является

уникальным идентификатором. И ключи, и значения могут быть любыми - от простых целых чисел или строк до сложных документов JSON.

Хранилища ключей-значений отлично подходят для хранения данных пользовательских сессий и пользовательских предпочтений, для создания рекомендаций и целевой рекламы в реальном времени, а также для кэширования данных в памяти.

Однако если вы хотите иметь возможность запрашивать данные по определенному значению, вам нужны связи между значениями данных или необходимо иметь несколько уникальных ключей, хранилище ключевых значений может оказаться не самым лучшим вариантом.

Redis, Memcached и DynamoDB - вот некоторые известные примеры в этой категории.

Базы данных на основе документов:

Базы данных на основе документов хранят каждую запись и связанные с ней данные в одном документе. Они позволяют гибко индексировать, выполнять мощные специальные запросы и анализировать коллекции документов.

Базы данных на основе документов предпочтительны для платформ электронной коммерции, хранения медицинских записей, CRM-платформ, и аналитических платформ.

Однако если вы хотите выполнять сложные поисковые запросы и многооперационные транзакции, база данных на основе документов может оказаться для вас не лучшим вариантом.

MongoDB, DocumentDB, CouchDB и Cloudant - вот некоторые из популярных баз данных на основе документов.

На основе столбцов:

Модели на основе столбцов хранят данные в ячейках, сгруппированных в виде столбцов данных, а не строк. Логическая группировка столбцов, то есть столбцов, к которым обычно обращаются вместе, называется семейством столбцов.

Например, имя клиента и его анкетные данные, скорее всего, будут доступны вместе, но не истории покупок. Поэтому данные об имени и профиле клиента можно сгруппировать в семейство столбцов. Поскольку колоночные базы данных хранят все ячейки, соответствующие столбцу, в виде непрерывной записи на диске, доступ к данным и их поиск становятся очень быстрыми.

Базы данных столбцов могут отлично подойти для систем, требующих интенсивных запросов на запись, хранения данные временных рядов, погодные данные и данные IoT.

Но если вам необходимо использовать сложные запросы или часто менять шаблоны запросов, возможно, это не лучший вариант для вас.

Самые популярные колоночные базы данных - **Cassandra и HBase**.

Графовые базы данных:

Базы данных на основе графиков используют графическую модель для представления и хранения данных. Они особенно полезны для визуализации, анализа и поиска связей между различными частями данных.

Круги - это узлы, в них содержатся данные. Стрелки отражают взаимосвязи. Графовые базы данных - отличный выбор для работы со связанными данными, то есть данными, которые которые содержат множество взаимосвязанных отношений.

Графовые базы данных отлично подходят для социальных сетей, рекомендаций товаров в режиме реального времени, сетевых диаграмм, выявления мошенничества и управления доступом.

Но если вы хотите обрабатывать большие объемы транзакций, это может оказаться не лучшим выбором потому что графовые базы данных не оптимизированы для аналитических запросов большого объема.

Neo4J и CosmosDB - одни из самых популярных графовых баз данных.

NoSQL была создана в ответ на ограничения традиционной технологии реляционных баз данных. Основным преимуществом NoSQL является способность обрабатывать большие объемы структурированных, полу-структурированных, и неструктурированных данных.

Некоторые из других преимуществ включают:

- способность работать как распределенные системы, масштабируемые на несколько центров обработки данных, что позволяет им использовать преимущества инфраструктуры облачных вычислений;
- Эффективная и экономически выгодная архитектура масштабирования, которая обеспечивает дополнительную емкость
- и производительность при добавлении новых узлов; и
- Более простая конструкция, лучший контроль над доступностью и улучшенная масштабируемость, что позволяет вам быть более гибкими, гибче и быстрее выполнять итерации.

Подведем итоги ключевых различий между реляционными и нереляционными базами данных:

Схемы Реляционные СУБД жестко определяют, как все данные, вводимые в базу данных, должны быть типизированы и, в то время как базы данных NoSQL могут быть схемно-агностическими, что позволяет хранить неструктурированные и полуструктурированные данные и манипулировать ими.

Обслуживание высококлассных коммерческих реляционных систем управления базами данных обходится дорого, в то время как базы данных NoSQL специально разработаны для хранения и обработки данных для недорогого оборудования.

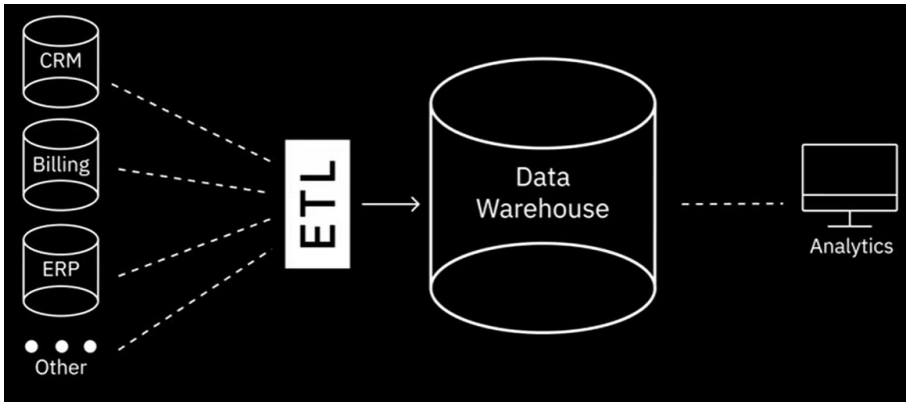
Реляционные базы данных, в отличие от большинства NoSQL, поддерживают ACID-совместимость, что обеспечивает надежность транзакций и восстановление после сбоев.

РСУБД - это зрелая и хорошо документированная технология, а значит, риски более или менее ощутимы по сравнению с NoSQL, которая является относительно новой технологией.

Тем не менее, базы данных NoSQL не стоят на месте и все чаще используются для критически важных приложений.

3 ЛЕКЦИЯ

Ранее в этом курсе мы рассмотрели базы данных, хранилища данных и хранилища больших данных. Теперь мы немного углубимся в изучение хранилищ данных, карт данных и озер данных а также узнаем о процессе ETL и конвейерах данных.



Хранилище данных работает как многоцелевое хранилище для различных случаев использования. К тому времени, когда данные поступают в хранилище, они уже смоделированы и структурированы для конкретной цели, то есть они готовы к анализу.

Как организация, вы можете выбрать хранилище данных, когда у вас есть огромные объемы данных из ваших операционных систем, которые

должны быть легко доступны для отчетности и анализа. Хранилища данных служат единым источником истины, хранящим текущие и исторические данные, которые были очищены, приведены в соответствие и классифицированы.

Хранилище данных - это многоцелевой инструмент оперативной и результативной аналитики. Хранилище данных - это подраздел хранилища данных, созданный специально для определенной бизнес-функции, цели или сообщества пользователей.

Идея заключается в том, чтобы предоставить заинтересованным сторонам данные, которые наиболее актуальны для них, когда они в них нуждаются.

Например, отделы продаж или финансов получают доступ к данным для составления квартальных отчетов и прогнозов. Поскольку Data Mart (витрина данных) данных предлагает аналитические возможности для ограниченной области хранилища данных, он обеспечивает изолированную безопасность и изолированную производительность.

Самая важная роль Data Mart- это отчетность и аналитика для конкретного бизнеса.

Озеро (data lake) данных - это хранилище, в котором могут храниться большие объемы структурированных, полуструктурированных и неструктурированных данных в их родном формате, классифицированных и помеченных метаданными.

Таким образом, если в *хранилище данных* хранятся данные, *обработанные для конкретных нужд*, то *озеро данных* - это *пул необработанных данных*, где каждый элемент данных классифицирован и помечен метаданными то есть уникальными идентификаторами такими как хэш, метатеги для дальнейшего использования.

Вы можете выбрать озеро данных, если вы генерируете или имеете доступ к большим объемам данных на постоянной основе, но не хотите ограничиваться конкретными или заранее определенными сценариями использования.

В отличие от хранилищ данных, в озере данных хранятся все исходные данные без каких-либо исключений. При этом данные могут включать все типы и источники данных. Озера данных иногда также используются в качестве перевалочного пункта хранилища данных.

Наиболее важная роль озера данных – использование их в предиктивной и расширенной аналитике.

Теперь мы переходим к процессу, который лежит в основе получения ценности из данных - извлечению, процессу преобразования и загрузки, или ETL.

ETL - это способ преобразования необработанных данных в данные, готовые к анализу. Это автоматизированный процесс, в ходе которого вы собираете необработанные данные из определенных источников, извлекаете информацию, которая соответствует вашим потребностям в отчетности и анализе, очистка, стандартизация и преобразование этих данных в формат, пригодный для использования в контексте вашей организации; и загрузить их в хранилище данных.

Хотя ETL является общим процессом, фактическая работа может быть очень разной по использованию, полезности, и сложности.

Извлечение - это этап, на котором данные из источников собираются для преобразования. Извлечение данных может осуществляться посредством:

- **пакетной обработки**, то есть исходные данные перемещаются большими кусками из источника в целевую систему через запланированные промежутки времени.

Инструменты для пакетной обработки включают Stitch и Blendo.

- **Потоковая обработка**, то есть исходные данные извлекаются из источника в режиме реального времени и преобразуются во время транспортировки и перед загрузкой в хранилище данных.

Инструменты для потоковой обработки включают Apache Samza, Apache Storm и Apache Kafka.

- **Преобразование** включает в себя выполнение правил и функций, которые преобразуют необработанные данные в данные, которые можно использовать для анализа.

Например:

- приведение форматов дат и единиц измерения к единому виду во всех исходных данных,
- удаление дубликатов данных,
- отфильтровывание ненужных данных,
- обогащение данных, например, разделение полного имени на имя, отчество и фамилию,
- установление ключевых связей между таблицами, применение бизнес-правил и проверки данных.

Загрузка - это этап, на котором обработанные данные переносятся в конечную систему или хранилище данных.

Это может быть:

- Первоначальная загрузка, то есть заполнение всех данных в хранилище,

- инкрементная загрузка, то есть периодическое применение текущих обновлений и модификаций по мере необходимости;

или

- Полное обновление, то есть удаление содержимого одной или нескольких таблиц и повторная загрузка свежими данными.
- Проверка нагрузки, которая включает в себя проверку данных на наличие отсутствующих или нулевых значений, производительность сервера,
- мониторинг сбоев нагрузки, являются важными частями этого этапа процесса.

Очень важно следить за сбоями в нагрузке и убедиться, что созданы правильные механизмы восстановления.

ETL исторически использовался для пакетных рабочих нагрузок в больших масштабах. Однако с появлением инструментов потокового ETL они все чаще используются для обработки потоковых событийных данных в реальном времени.

Часто можно встретить взаимозаменяемые термины ETL и конвейеры данных. И хотя и те, и другие перемещают данные от источника к месту назначения, **конвейер данных** - это более широкий термин, который охватывает весь процесс перемещения данных из одной системы в другую.

ETL является подмножеством конвейеров данных.

Конвейеры данных могут быть спроектированы для пакетной обработки, для потоковых данных, а также для комбинации пакетной и потоковой обработки данных.

В случае потоковых данных обработка или преобразование данных происходит в непрерывном потоке. Это особенно полезно для данных, требующих постоянного обновления, таких как данные от датчика, отслеживающего дорожное движение.

Конвейер данных - это высокопроизводительная система, которая поддерживает как длительные пакетные запросы, так и небольшие интерактивные запросы. Местом назначения конвейера данных обычно является озеро данных, хотя данные также могут загружаться в различные целевые пункты назначения, такие как другое приложение или инструмент визуализации.

Существует ряд решений для конвейерной обработки данных, наиболее популярными **среди которых являются ApacheBeam и DataFlow.**

ТОЧКИ ЗРЕНИЯ

В этом видео мы послушаем, как несколько профессионалов в области данных рассказывают о некоторых факторах, которые они учитывают при выборе наиболее подходящего хранилища данных для своей организации.

При выборе подходящей базы данных необходимо учитывать ряд факторов.

- На сценарий использования, то есть то для чего будет использоваться хранилище данных,
- будет ли оно использоваться для хранения структурированной информации, полуструктурированной или неструктурированной информации, если вы знаете заранее какова схема данных,
- есть ли требования к производительности,
- работаете ли вы с данными в состоянии покоя, потоковыми данными или данными в движении
- нужно ли шифровать данные?
- знаете ли вы, с каким объемом данных вы работаете?
- нужна ли вам система больших данных
- каковы требования к хранению данных
- нужно ли часто обновлять данные и часто обращаться к ним или нужно просто хранить и держать в хранилище в течение длительного времени и необходимы, например, для резервного копирования.
- Затем у вашей организации могут быть определенные стандарты, которые они могут ввести в действие в отношении того, какие базы данных или какие хранилища данных вам разрешено использовать для различных видов задач.

Так что все эти факторы необходимо учитывать, поэтому, когда мы рассматриваем вопрос о том, какое хранилище данных мы хотим выбрать, мы рассматриваем эти факторы, мы смотрим на то, какой объем данных требуется и затем мы также смотрим на тип доступа, который нам нужен, будем ли мы обращаться к ним в короткие промежутки времени или мы выполняем длительные запросы к ним, использую ли я их для обработки транзакций или я использую ее для аналитики, архивации или хранения данных?

- Мы также смотрим на совместимость, насколько это новое хранилище данных совместимо с нашей существующей экосистемой языков программирования, инструментов и хранилищ данных, и любых процессов, которые у нас есть, такие как функции безопасности, которые предоставляет нам это хранилище, и самое главное - масштабируемость.
- Так как мы можем быть довольны его производительностью сегодня, но достаточно ли оно масштабируемо, может ли оно масштабироваться вместе с потребностями организации

Я не часто могу выбирать тип хранилища данных, которое использует моя организация. Очень немногие организации используют одно хранилище данных в наши дни в моей команде, в которой я работаю.

- У нас есть набор предпочтительных решений.
- У нас есть предпочтительная корпоративная реляционная база данных.
- У нас есть предпочтительная реляционная база данных с открытым исходным кодом для некоторых небольших проектов и для микросервисов
- и еще у нас есть предпочтительный источник неструктурированных данных, так что это три наших основных источника.

Главное - подумать о навыках, которые есть в вашей организации и рассмотреть стоимость различных решений.

В нашем случае у нас есть несколько экспертов по db2, так что наша корпоративная база данных - это db2, однако есть и другие проекты, которые используют другие базы данных.

Несколько раз, мы пошли в разных направлениях, чтобы понять, где мы действительно хотим быть. Хостинговая платформа тоже имеет значение, потому что теперь это не просто хочу ли я использовать ibm db2 или хочу ли я использовать microsoft sql server от другого производителя или не между этими двумя вариантами, а в том, когда я буду делать эти данные на aws rds.

Может быть, мне стоит рассмотреть aurora от amazon, может быть, мне стоит рассмотреть реляционные предложения от google. Есть так много различных вариантов, которые вы должны рассмотреть, если есть решение о том как должны храниться данные, как должны извлекаться данные, а также решение о месте хранения - все это очень важные вопросы, когда вы принимаете решение о хранении данных.

Я бы сказал, что структура данных, характер приложения и объем, в котором данные поступают в вашу базу данных, все эти факторы определяют характер источника данных.

В большинстве случаев реляционной базы данных будет достаточно, однако есть граничные случаи, когда реляционные базы данных, такие как ibm db2, oracle или postgres, не обязательно справятся со своей задачей, поэтому в зависимости от конкретного случая использования.

Например, если вы анализируете или получаете гигабайты или терабайты данных в день, то хранилища документов, такие как mongodb или белые хранилища колонок, такие как cassandra, могут быть подходящими для вас, в то же время, если вы пытаетесь построить системы рекомендаций продуктов, пытаетесь показать сеть отношений между различными людьми в социальных сетях, тогда графовые структуры данных, такие как neo4j или apache tinker pop идеально

подойдут вам в то же время, если вы изучаете терабайты или петабайты данных для аналитики, движок hadoop с mapreduce будет хорошим выбором для вас, поэтому все сводится к характеру приложения, объему данных, к структуре данных, прежде чем вы сможете выбрать правильную базу данных или источник данных для вашего сценария использования.

4 ЛЕКЦИЯ

В этом цифровом мире каждый оставляет за собой след. От наших привычек путешествовать до наших тренировок и развлечений, растущее число подключенных к интернету устройств, с которыми мы взаимодействуем ежедневно, записывают огромное количество данных о нас.

Для этого даже существует название "Большие данные". Компания Ernst and Young предлагает следующее определение: большие данные относятся к динамичным, разрозненным объемам данных, создаваемыми людьми, инструментами и машинами. Они требуют новых, инновационных и масштабируемых технологий для сбора, размещения и аналитической обработки огромного количества данных, для получения в реальном времени информации о бизнесе, связанной с потребителями, риском, прибылью, производительностью, управлением производительностью, и повышения акционерной стоимости. Единого определения больших данных не существует, но есть определенные элементы, которые являются общими в различных определениях, такие как скорость (velocity), объем (volume), разнообразие (variety), правдивость (veracity) и ценность (value).

Вот эти "V" больших данных

Скорость - это скорость, с которой накапливаются данные. Данные генерируются чрезвычайно быстро в процессе, который никогда не останавливается. Потоковые, локальные и облачные технологии, работающие в режиме близком к реальному времени, и могут обрабатывать информацию очень быстро.

Объем - это масштаб данных или увеличение объема хранимых данных. Движущими факторами объема являются увеличение количества источников данных, датчики с более высоким разрешением и масштабируемая инфраструктура.

Разнообразие - это разнообразие данных. Структурированные данные аккуратно укладываются в строки и столбцы в реляционных базах данных, в то время как неструктурированные данные не организованы заранее определенным образом, например, твиты, сообщения в блогах, фотографии, числа и видео. Разнообразие также отражает то, что данные поступают из различных источников; машин, людей и процессов, как внутренних, так и внешних по отношению к организации.

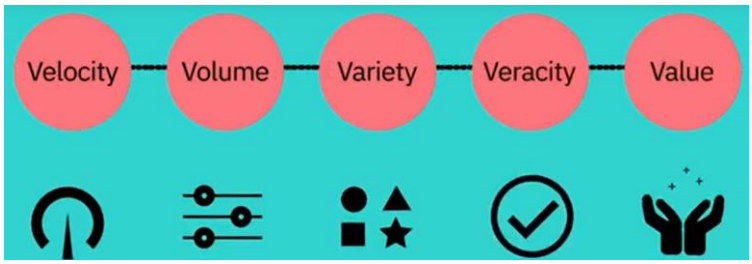
Движущими факторами являются мобильные технологии, социальные медиа, носимые технологии, гео-технологии, видео и многие, многие другие.

Достоверность- это качество и происхождение данных, их соответствие фактам и точность. Атрибуты включают последовательность, полноту, целостность и неоднозначность. К движущим факторам относятся стоимость и необходимость отслеживания.

При большом количестве доступных данных не утихают споры о точности данных в цифровую эпоху. Является ли информация реальной или ложной?

Ценность - это наша способность и необходимость превращать данные в ценность. Ценность - это не только прибыль. Она может иметь медицинские или социальные преимущества, а также удовлетворенность клиентов, сотрудников или личная удовлетворенность.

Основная причина, по которой люди тратят время на понимание больших данных- это извлечение из них ценности.



<u>Давайте рассмотрим несколько примеров "V" в действии.</u>

Скорость. Каждые 60 секунд на YouTube загружаются часы видеоматериалов, которые генерируют данные. Подумайте о том, как быстро накапливаются данные за несколько часов, дней и лет.

Объем. Население мира составляет около 7 миллиардов человек, и подавляющее большинство из них пользуются цифровыми устройствами. Мобильные телефоны, настольные и портативные компьютеры, носимые устройства и так далее. Эти устройства генерируют, получают и хранят данные объемом примерно 2,5 квинтиллиона байт каждый день. Это эквивалентно 10 миллионам DVD-дисков blu-ray.

Разнообразие. Давайте подумаем о различных типах данных.

Текст, фотографии, фильмы, звук, данные о состоянии здоровья с носимых устройств и многие различных типов данных от устройств, подключенных к Интернету вещей.

Достоверность. Восемьдесят процентов данных считаются неструктурированными, и мы должны разработать способы получения надежных и точных данных. Данные должны быть классифицированы, проанализированы и визуализированы. Сегодня специалисты по анализу данных извлекают информацию из больших данных и справляются с проблемами, которые представляют собой эти огромные массивы данных. Масштабы собираемых данных означают, что использование обычных инструментов анализа данных не представляется возможным, однако альтернативные инструменты, использующие распределенные вычислительные мощности, могут решить эту проблему.

Такие инструменты, как *Apache Spark, Hadoop и их экосистема обеспечивают способы извлечения, загрузки, анализа и обработки данных на распределенных вычислительных ресурсах*, что позволяет получить новые знания и идеи и знания. Это дает организациям больше возможностей для связи со своими клиентами и обогатить предлагаемые ими услуги.

Поэтому в следующий раз, когда вы наденете свои смарт-часы, разблокируете смартфон, или будете отслеживать свою тренировку, помните, что ваши данные начинают путешествие, которое может провести их по всему миру, через анализ больших данных и обратно к вам.

ЛЕКЦИЯ 5

Технологии обработки больших данных предоставляют способы работы с большими наборами структурированных, полуструктурированными и неструктурированными данными, чтобы из больших данных можно было извлечь ценность.

В некоторых других видеороликах мы обсуждали такие технологии обработки Больших Данных, как базы данных NoSQL и озера данных.

В этом видео мы поговорим о трех технологиях с открытым исходным кодом и о роли, которую они играют в аналитике больших данных которую они играют в анализе больших данных - Apache Hadoop, Apache Hive и Apache Spark.

Hadoop - это набор инструментов, обеспечивающих распределенное хранение и обработку больших данных.

Hive - это хранилище данных для запросов и анализа данных, построенное поверх Hadoop.

Spark - это распределенная система анализа данных, предназначенная для выполнения сложного анализа данных в режиме реального времени.

Hadoop, основанная на java платформа с открытым исходным кодом, позволяет распределенно хранить и обрабатывать больших массивов данных на кластерах компьютеров. В распределенной системе Hadoop узел - это отдельный компьютер, а совокупность узлов образует кластер.

Hadoop может масштабироваться от одного узла до любого количества узлов, каждый из которых обеспечивает локальное хранение данных и вычисления.

Hadoop представляет собой надежное, масштабируемое и экономически эффективное решение для хранения данных с без требований к формату.

Используя Hadoop, вы можете:

- Включать новые форматы данных, такие как потоковое аудио, видео, данные о настроениях в социальных сетях и потоки кликов, наряду со структурированными, полуструктурированными и неструктурированными данными, которые традиционно не используются в хранилищах данных.
- Обеспечить доступ к данным в режиме реального времени для всех заинтересованных сторон.
- оптимизация затрат на корпоративное хранилище данных за счет консолидации данных по всей организации и перемещения "холодных" данных организации и перемещения "холодных" данных, то есть данных, которые не используются часто, в систему на базе Hadoop.

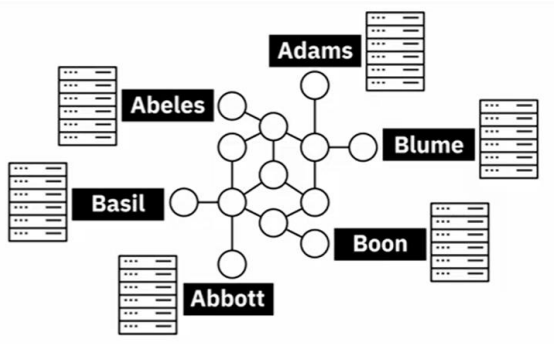
Одним из четырех основных компонентов Hadoop является Hadoop Distributed File System, или HDFS, которая представляет собой систему хранения больших данных, работающую на нескольких аппаратных средствах, подключенных через сеть.

HDFS обеспечивает масштабируемое и надежное хранение больших данных за счет разделения файлов на нескольких узлов.

Она разделяет большие файлы на несколько компьютеров, обеспечивая параллельный доступ к ним. Таким образом, вычисления могут выполняться параллельно на каждом узле, где хранятся данные. Кроме того, система реплицирует блоки файлов на разных узлах для предотвращения потери данных, что делает ее отказоустойчивой.

Давайте поймем это на примере.

Рассмотрим файл, содержащий номера телефонов всех жителей США; номера телефонов для людей с фамилией, начинающейся на A, могут храниться на сервере 1, B - на сервере 2, и так далее.



В Hadoop части этого телефонного справочника будут храниться в кластере. Чтобы восстановить всю телефонную книгу, вашей программе потребуются блоки с каждого сервера в кластере.

HDFS также реплицирует эти небольшие фрагменты на два дополнительных сервера по умолчанию, что обеспечивает доступность при отказе одного из серверов. В дополнение к более высокой доступности, это дает множество преимуществ.

Это позволяет кластеру Hadoop разбивать работу на более мелкие фрагменты и запускать эти задания на всех серверах кластера для лучшей масштабируемости.

Наконец, вы получаете преимущество локальности данных, которая представляет собой процесс перемещения вычислений ближе к узлу, на котором находятся данные.

Это очень важно при работе с большими массивами данных, поскольку минимизирует перегрузку сети и увеличивает пропускную способность.

Некоторые другие преимущества использования HDFS включают:

- Быстрое восстановление после аппаратных сбоев, поскольку HDFS создана для обнаружения сбоев и автоматического восстановления.
- Доступ к потоковым данным, поскольку HDFS поддерживает высокую пропускную способность данных.
- Размещение больших массивов данных, поскольку HDFS может масштабироваться до сотен узлов, или компьютеров, в одном кластере.
- Переносимость, поскольку HDFS переносима на различные аппаратные платформы и совместима с различными базовыми операционными системами.

Hive - это программное обеспечение для хранилища данных с открытым исходным кодом для чтения, записи и управления большими данными файлов, которые хранятся непосредственно в HDFS или других системах хранения данных, таких как Apache HBase.

Hadoop предназначен для длительного последовательного сканирования, и, поскольку Hive основан на Hadoop, запросы имеют очень высокую задержку, что означает, что Hive не подходит для приложений, которым требуется очень быстрое время отклика.

Кроме того, Hive основан на чтении, поэтому он не подходит для обработки транзакций, которые, как правило включает в себя высокий процент операций записи.

Hive лучше подходит для задач хранилища данных, таких как ETL, отчетность и анализ данных и включает инструменты, обеспечивающие простой доступ к данным через SQL.

Это подводит нас к Spark, универсальному механизму обработки данных, предназначенному для извлечения и обработки больших объемов данных и обработки больших объемов данных для широкого спектра приложений, включая интерактивную аналитику, обработку потоков, машинную обработку, аналитику, аналитику и обработку данных.

Аналитика, обработка потоков, машинное обучение, интеграция данных и ETL. Он использует преимущества обработки в памяти для значительного увеличения скорости вычислений и выливает данные на диск только в случае нехватки памяти.

Spark имеет интерфейсы для основных языков программирования, включая Java, Scala, Python, R, и SQL.

Он может работать с использованием своей автономной технологии кластеризации, а также поверх других инфраструктур. Он может получать доступ к данным из большого количества источников данных, включая HDFS и Hive, что делает ее очень универсальной.

Способность быстро обрабатывать потоковые данные и выполнять сложную аналитику в режиме реального времени является ключевой областью применения Apache Spark.

ИТОГИ

В этом уроке вы узнали следующую информацию:

Хранилище данных - это общий термин, обозначающий данные, которые были собраны, организованы и изолированы таким образом, чтобы их можно было использовать для отчетности, аналитики, а также в архивных целях.

К различным типам хранилищ данных относятся:

- Базы данных (**Databases)**, которые могут быть реляционными или нереляционными, каждая из которых соответствует набору организационных принципов, типов данных, которые они могут хранить, и инструментов, которые могут быть использованы для запроса, организации и получения данных.
- Хранилища данных (**Data Warehouses)**, которые консолидируют поступающие данные в одно всеобъемлющее хранилище.
- Data marts, которые по сути являются подразделами хранилища данных, созданными для изоляции данных для конкретной бизнес-функции или случая использования.
- Озера данных (**Data Lakes**), которые служат хранилищами больших объемов структурированных, полуструктурированных и неструктурированных данных в их собственном формате.
- Хранилища больших данных (**Big Data Stores,**), которые обеспечивают распределенную вычислительную инфраструктуру и инфраструктуру хранения для хранения, масштабирования и обработки очень больших массивов данных.

ETL, или Extract Transform and Load, процесс - это автоматизированный процесс, который преобразует необработанные данные в готовые для анализа данные путем:

- Извлечения данных из источников.
- Преобразования необработанных данных путем их очистки, обогащения, стандартизации и проверки.
- Загрузка обработанных данных в конечную систему или хранилище данных.

Конвейер данных, иногда используемый как взаимозаменяемое понятие с ETL, охватывает весь путь перемещения данных из источника в целевое озеро данных или приложение с помощью процесса ETL.

Большие данные относятся к огромным объемам данных, которые производятся каждый момент каждого дня людьми, инструментами и машинами. Огромная скорость, объем и разнообразие данных бросают вызов инструментам и системам, используемым для работы с обычными данными. Эти проблемы привели к появлению инструментов и платформ обработки, разработанных специально для Больших Данных, таких как Apache Hadoop, Apache Hive и Apache Spark.

МОДУЛЬ 5

На этом этапе у вас есть понимание проблемы и желаемого результата – вы знаете "Где вы находитесь" и "Где вы хотите быть".

У вас также есть четко определенная метрика - вы знаете "Что будет измеряться" и "Как это будет измерено".

Следующим шагом для вас будет определение данных, необходимых для вашего сценария использования.Процесс определения данных начинается с определения информации, которую вы хотите собрать.

На этом этапе вы принимаете решения относительно (a) конкретной информации, которая вам нужна; и (b) возможных источников этих данных.

Ваши цели определяют ответы на эти вопросы.

Возьмем пример компании, которая хочет создать целевые маркетинговые кампании на основе возрастной группы, которая больше всего покупает их продукцию.

Их цель - разработать такие рекламные кампании, которые наиболее привлекательны для этого сегмента и побуждают их в дальнейшем влиять на своих друзей и сверстников, чтобы они покупали эти продукты.

Исходя из этого сценария использования, некоторые из очевидных сведений, которые вы будете определять, включают в себя профиль клиента, история покупок, местоположение, возраст, образование, профессия, доход и семейное положение.

Чтобы получить еще более глубокое представление об этом сегменте, вы можете также решить собрать данные о жалобах клиентов этого сегмента, чтобы понять, с какими проблемами они сталкиваются поскольку это может отбить у них охоту рекомендовать вашу продукцию.

Чтобы узнать, насколько они были удовлетворены решением своих проблем, вы можете собрать оценки из опросов, проводимых службой поддержки клиентов.

Развивая эту тему, вы, возможно, захотите понять, как эти клиенты отзываются о вашей продукции в социальных сетях и сколько их знакомых участвуют с ними в этих обсуждениях. Например, сколько лайков, акций и комментариев получают их сообщения.

Следующим шагом в этом процессе является определение плана сбора данных.

Вам необходимо установить временные рамки для сбора определенных вами данных. Некоторые из необходимых вам данных могут потребоваться на постоянной основе, а некоторые - в течение определенного период времени.

Например, для сбора данных о посетителях веб-сайта вам может потребоваться обновлять показатели в режиме реального времени.

Но если вы отслеживаете данные по определенному событию, у вас есть определенная дата начала и окончания сбора данных.

На этом этапе вы также можете определить, какой объем данных будет достаточным для проведения достоверного анализа.

Определяется ли объем сегментом, например, все клиенты в возрастном диапазоне от 21 до 30 лет; или набор данных из ста тысяч клиентов в возрастном диапазоне от 21 до 30 лет.

Вы также можете использовать этот шаг для определения зависимостей, рисков, плана смягчения последствий и некоторых других подобных факторов, имеющих отношение к вашей инициативе.

Цель плана должна заключаться в том, чтобы установить четкость, необходимую для выполнения. На третьем этапе необходимо определить методы сбора данных. На этом этапе вы определите методы сбора необходимых вам данных. Вы определите, как вы будете собирать данные из определенных вами источников данных, таких как внутренние системы, сайты социальных сетей или сторонние поставщики данных.

Ваши методы будут зависеть от типа данных, сроков, в течение которых вам нужны данные, и объема данных.

После завершения разработки плана и методов сбора данных вы можете реализовать стратегию сбора данных и начать сбор данных.

Вы будете вносить изменения в свой план по мере его выполнения, поскольку условия меняются по мере того, как по мере реализации плана на местах.

Данные, которые вы определите, источник этих данных и методы, которые вы используете для сбора данных, влияют на их качество, безопасностm и конфиденциальностm.

Ни одно из этих соображений не является одноразовым, они актуальны на протяжении всего жизненного цикла процесса анализа данных.

Работа с данными из разрозненных источников без учета того, как они соотносятся с метрикой качества, может привести к неудаче.

Для того чтобы быть надежными, данные должны быть свободными от ошибок, точными, полными, актуальными, и доступными.

Вам необходимо определить признаки качества, метрику и контрольные точки для того, чтобы убедиться в том, что ваш анализ будет основан на качественных данных.

Вам также необходимо следить за вопросами, относящимися к управлению данными, такими как безопасность, регулирование, и соблюдение требований.

Политика и процедуры управления данными относятся к удобству использования, целостности и доступности данных.Штрафы за несоблюдение требований могут исчисляться миллионами долларов и могут подорвать доверие к не только вашим результатам, но и вашей организации.

Еще одним важным аспектом является конфиденциальность данных.

Собираемые вами данные должны соответствовать требованиям конфиденциальности, лицензии на использование и соответствие установленным нормам.

Необходимо предусмотреть проверки, валидации и аудиторский след.

Потеря доверия к данным, используемым для анализа, может поставить под угрозу процесс, привести к подозрительным и повлечь за собой штрафные санкции.

Определение правильных данных - очень важный этап процесса анализа данных.

Если все сделано правильно, это гарантирует, что вы сможете взглянуть на проблему с разных точек зрения и ваши выводы будут достоверными и надежными.

Лекция 2 Gathering Data. Video: Data Sources

Источники данных могут быть внутренними или внешними по отношению к организации, и они могут быть первичными, вторичными или сторонними источниками данных. Давайте рассмотрим несколько примеров, чтобы понять, что мы понимаем под первичными, вторичными и сторонними источниками данных.

Термин "первичные данные" относится к информации, полученной непосредственно вами из источника.

Это могут быть внутренние источники, такие как данные из организации, CRM, HR или приложения рабочего процесса. Это может также включать данные, которые вы собираете непосредственно в ходе опросов, интервью, дискуссий, наблюдений и фокус-групп.

Вторичные данные относятся к информации, полученной из существующих источников, таких как внешние базы данных, исследовательские статьи, публикации, учебные материалы и Интернет или финансовые отчеты, доступные в качестве открытых данных. Сюда

также можно включать данные, собранные в ходе внешних опросов, интервью, дискуссий, наблюдений и фокус-групп.

Данные третьих лиц - это данные, приобретенные вами у агрегаторов, которые собирают данные из различных источников и объединяют их в всеобъемлющие наборы данных исключительно с целью продажи данные. Сейчас мы рассмотрим некоторые из различных источников, из которых вы можете собирать данные.

Базы данных могут быть источником первичных, вторичных и сторонних данных. Большинство организаций имеют внутренние приложения для управления своими процессами, рабочими процессами и клиентами.

Внешние базы данных доступны по подписке или для покупки. Значительное количество предприятий уже перешли или в настоящее время переходят на облачные технологии, которые все чаще становятся источником доступа к информации в режиме реального времени и по запросу информации. Интернет является источником общедоступных данных, которые доступных для компаний.

И частным лицам для бесплатного или коммерческого использования. Веб – это богатый источник данных, находящихся в общественном достоянии. Они могутвключают учебники, правительственные документы, документы и статьи, предназначенные для общественного потребления, сайты социальных сетей и интерактивные платформы, такие как Facebook, Twitter, Google, YouTube. Инстаграм все чаще используется для получения данных и мнений пользователей. Предприятия используют эти источники данных источники данных для получения количественных и качественного анализа. Существующие и потенциальные потребители. Сенсорные данные, производимые носимыми устройствами, умными зданиями, умными городами, смартфонами, медицинскими приборами и даже бытовые приборы, являются широко используемым источником данных.

Обмен данными - это источник данных третьей стороны, который предполагает добровольный обмен данными между поставщиками и потребителями данных потребителями, частными лицами, организациями и правительствами могут быть как поставщиками, так и потребителями данных. Данные, которыми обмениваются

Обмен данными может включать данные, поступающие от бизнес-приложений, сенсорных устройств, активности в социальных сетях, местоположения или данные о поведении потребителей.

Опросы собирают информацию с помощью анкет распространяемых среди определенной группы людей. Например, оценка заинтересованность существующих клиентов в расходах на обновленную

Например, оценка заинтересованности существующих клиентов в расходах на обновленную версию продукта. Опросы могут проводиться в Интернете или на бумаге.

Данные переписи населения также являются широко используемым источником для сбора данные о домохозяйствах, такие как данные о благосостоянии и доходах или данные о населении, например. Интервью являются источником для сбора качественных данных, таких как мнения и опыт участников. Например, Например, интервью проводится для того, чтобы понять повседневные проблем, с которыми сталкивается руководитель службы поддержки клиентов. Интервью могут быть телефонными через Интернет или очным наблюдением.

Исследования включают наблюдение за участниками в определенной

среде или во время выполнения определенной задачи. Например, наблюдение за тем, как пользователи перемещаются по сайту электронной торговли, чтобы оценить.

Легкость, с которой они могут найти товары и сделать покупку данные, полученные в ходе опросов, интервью, наблюдения.

Исследования могут быть доступны в виде первичных, вторичных и сторонних данные. Источники данных никогда не были такими динамичными и разнообразными, как они есть сегодня. Они также постоянно развиваются.

Дополнение ваших первичных данных вторичными и сторонними источники данных могут помочь вам исследовать проблемы и решения новыми и значимыми способами.

Лекция 3

В этом видео мы узнаем о различных методах и инструментах, доступных для сбора данных из источников данных, рассмотренных ранее в курсе, таких как базы данных, веб, данные датчиков, биржи данных и некоторые другие источники, используемые для конкретных потребностей в данных.

Мы также узнаем об импорте данных в различные типы хранилищ данных. SQL, или язык структурированных запросов, - это язык запросов, используемый для извлечения информации из реляционных баз данных.

SQL предлагает простые команды для указания того, что нужно извлечь из базы данных, таблицу, из которой нужно извлечь информацию, сгруппировать записи с совпадающими значениями, определить последовательность отображения результатов запроса и ограничение количества результатов, которые могут быть возвращены запросом, а также множество других возможностей и функций.

К нереляционным базам данных можно обращаться с помощью SQL или SQL-подобных инструментов.

Некоторые нереляционные базы данных поставляются с собственными инструментами запросов, такими как CQL для Cassandra и GraphQL для Neo4J.

Интерфейсы прикладного программирования (или API) также широко используются для извлечения данных из различных источников данных.

API вызываются из приложений, которым требуются данные, и получают доступ к конечной точке, содержащей данные.

Конечные точки могут включать базы данных, веб-службы и рынки данных. API также используются для проверки данных. Например, аналитик данных может использовать API для проверки почтовых адресов и почтовых индексов.

Веб-скрейпинг, также известный как скрейпинг экрана или сбор веб-данных, используется для загрузки конкретных данных с веб-страниц на основе заданных параметров.

Среди прочего, веб-скрейпинг используется для извлечения таких данных, как текст, контактная информация, изображения, видео, подкасты и товары с веб-страниц.

RSS-каналы - еще один источник, обычно используемый для сбора обновленных данных с онлайн-форумов и новостных сайтов, где данные обновляются на постоянной основе.

Потоки данных - популярный источник для агрегации постоянных потоков данных, поступающих из источников таких как приборы, IoT-устройства и приложения, а также GPS-данные с автомобилей.

Потоки данных и каналы также используются для извлечения данных из социальных сетей и интерактивных платформ.

Платформы обмена данными позволяют обмениваться данными между поставщиками и потребителями данных. Обмен данными имеет набор четко определенных стандартов обмена, протоколов и форматов, необходимых для обмена данными.

Эти платформы не только облегчают обмен данными, но и обеспечивают безопасность и управление. Они обеспечивают рабочие процессы лицензирования данных, деидентификацию и защиту персональной информации, правовые рамки и карантинную аналитическую среду.

Примерами популярных платформ обмена данными являются AWS Data Exchange, Crunchbase, Lotame, и Snowflake.

Для удовлетворения конкретных потребностей в данных можно использовать множество других источников данных. Например, для изучения маркетинговых тенденций и расходов на рекламу такие исследовательские компании, как Forrester и Business Insider, как известно, предоставляют надежные данные.

Исследовательские и консультационные фирмы, такие как Gartner и Forrester, являются широко известными источниками, которым можно доверять в вопросах стратегических и оперативных рекомендаций.

Аналогичным образом, существует множество надежных имен в области данных о поведении пользователей, использовании мобильных и использования Интернета, рыночных опросов и демографических исследований. Данные, которые были определены и собраны из различных источников данных, теперь необходимо загрузить или импортировать в хранилище данных, прежде чем их можно будет обрабатывать, добывать и анализировать.

Процесс импорта включает в себя объединение данных из различных источников для обеспечения комбинированного представление и единый интерфейс, с помощью которого можно выполнять запросы и манипуляции с данными.

В зависимости от типа данных, их объема и типа целевого хранилища, вам могут понадобиться различные инструменты и методы.

Конкретные хранилища данных оптимизированы для определенных типов данных. Реляционные базы данных хранят структурированные данные с четко определенной схемой. Если вы используете реляционную базу данных в качестве системы назначения, вы сможете только

хранить структурированные данные, такие как данные из OLTP-систем, электронных таблиц, онлайн-форм, датчиков, сетевых и веб-журналов.

Структурированные данные можно хранить и в NoSQL.

Полуструктурированные данные - это данные, которые обладают некоторыми организационными свойствами, но не имеют жесткой структуры.

Например, данные из электронной почты, XML, заархивированные файлы, двоичные исполняемые файлы и протоколы TCP/IP. Полуструктурированные данные можно хранить в кластерах NoSQL.

Для хранения и обмена полуструктурированными данными обычно используются XML и JSON. JSON также является предпочтительным типом данных для веб-сервисов. Неструктурированные данные - это данные, которые не имеют структуры и не могут быть организованы в

схемы, такие как данные с веб-страниц, ленты социальных сетей, изображения, видео, документы, журналы СМИ и опросы.

Базы данных NoSQL и озера данных обеспечивают хороший вариант для хранения и манипулирования большими объемами неструктурированных данных.

В озерах данных можно хранить данные всех типов и схем. Инструменты ETL и конвейеры данных предоставляют автоматизированные функции, которые облегчают процесс импорта данных.

Такие инструменты, как Talend и Informatica, и языки программирования, такие как Python и R, и их библиотеки широко используются для импорта данных.

ИТОГИ

В этом уроке вы узнали:

Процесс выявления данных начинается с определения информации, которую необходимо собрать, что, в свою очередь, определяется целью, которую вы стремитесь достичь.

После определения данных следующим шагом будет определение источников, из которых вы будете извлекать необходимые данные, и составление плана сбора данных. На этом этапе также принимаются решения о сроках, в течение которых вам нужен набор данных, и о том, какой объем данных будет достаточным для проведения достоверного анализа.

Источники данных могут быть внутренними или внешними для организации, они могут быть первичными, вторичными или сторонними, в зависимости от того, получаете ли вы данные непосредственно из первоисточника, извлекаете их из внешних источников данных или приобретаете у агрегаторов данных.

Некоторые из источников данных, из которых вы можете собирать данные, включают базы данных, Интернет, социальные сети, интерактивные платформы, сенсорные устройства, обмен данными, опросы и наблюдательные исследования.

Данные, которые были определены и собраны из различных источников данных, объединяются с помощью различных инструментов и методов, чтобы обеспечить единый интерфейс, с помощью которого можно запрашивать данные и манипулировать ими.

Данные, которые вы определяете, источник этих данных и методы, которые вы используете для сбора данных, влияют на качество, безопасность и конфиденциальность, которые необходимо рассмотреть на этом этапе.

МОДУЛЬ 6 - РАБОТА С ДАННЫМИ

Обработка данных, также известная как мульчирование данных, представляет собой итеративный процесс, включающий в себя исследование, преобразование, проверку и предоставление их для достоверного и значимого анализа.

Он включает ряд задач, связанных с подготовкой необработанных данных для четко определенной цели, где необработанные данные на этом этапе - это данные, которые были собраны из различных источников данных в хранилище данных.

Обработка данных включает в себя ряд задач, связанных с подготовкой данных к анализу.

Как правило, это 4-этапный процесс, включающий обнаружение, преобразование, проверку и публикацию.

Фаза обнаружения, также известная как фаза исследования, заключается в том, чтобы лучше понять ваши данные в отношении вашего сценария использования.

Цель состоит в том, чтобы выяснить, как лучше всего очистить, структурировать, упорядочить, и отобразить имеющиеся у вас данные для вашего сценария использования.

На следующем этапе, этапе преобразования, происходит основная часть процесса обработки данных.

Он включает в себя задачи, которые вы решаете для преобразования данных, такие как структурирование, нормализация, денормализация, очистка и обогащение данных.

Начнем с первой задачи преобразования - структурирования.

Эта задача включает действия, которые изменяют форму и схему ваших данных. Поступающие данные могут иметь различные форматы.

Например, некоторые данные могут поступать из реляционной базы данных, а некоторые – из веб-интерфейсов API.

Для того чтобы объединить их, необходимо изменить форму или схему данных. Это изменение может быть простым - изменить порядок полей в записи или наборе данных или сложным - объединение полей в сложные структуры.

Объединения и союзы - это наиболее распространенные структурные преобразования, используемые для объединения данных из одной или нескольких таблиц.

То, как они объединяют данные, отличается друг от друга.

Объединения объединяют столбцы. При объединении двух таблиц столбцы из первой исходной таблицы объединяются со столбцами из второй исходной таблицы в одной строке.

Таким образом, каждая строка результирующей таблицы содержит столбцы из обеих таблиц.

Объединения объединяют строки.

Строки данных из первой исходной таблицы объединяются со строками данных из второй таблицы в одну таблицу.

Каждая строка в результирующей таблице берется из одной исходной таблицы или другой. Преобразование также может включать нормализацию и денормализацию данных.

Нормализация направлена на очистку базы данных от неиспользуемых данных и уменьшение избыточности и несогласованности.

Данные, поступающие из транзакционных систем, например, где постоянно выполняется ряд операций вставки, обновления, и удалений выполняется на постоянной основе, подвергаются значительной нормализации.

Денормализация используется для объединения данных из нескольких таблиц в одну таблицу, чтобы она может быть запрошена быстрее.

Например, нормализованные данные, поступающие из транзакционных систем, обычно денормализуются перед выполнением запросов для составления отчетов и анализа.

Еще один тип преобразования - очистка. Задачи очистки - это действия, которые устраняют нарушения в данных, чтобы получить достоверный и точный анализ.

Неточные, отсутствующие или неполные данные могут исказить результаты анализа, поэтому их необходимо учитывать.

Также данные могут быть необъективными, иметь нулевые значения в соответствующих полях или есть выбросы.

Например, вы можете захотеть выяснить демографическую информацию о продаже определенного но в полученных вами данных не указан пол.

Вам нужно либо найти источник этих данных и объединить их с существующим набором данных, либо либо вам нужно удалить и не учитывать записи, в которых отсутствует это поле.

Далее в курсе мы рассмотрим еще много примеров очистки данных. Обогащение данных - это четвертый тип преобразования. Когда вы рассматриваете имеющиеся у вас данные, чтобы найти дополнительные точки данных, которые могут сделать ваш анализ более значимым, вы обогащаете данные.

Например, в крупной организации, где информация разрознена по системам, вам может потребоваться обогатить набор данных, предоставляемый одной системой, информацией, доступной в других системах или даже в общедоступных базах данных.

Рассмотрим сценарий, в котором вы продаете ИТ-периферию предприятиям и хотите проанализировать и хотите проанализировать структуру покупок ваших клиентов за последние пять лет.

У вас есть основные таблицы клиентов и таблицы транзакций, в которых вы собрали информацию о клиентах и историю покупок.

Дополните ваш набор данных данными о производительности этих предприятий, которые, возможно, доступны в виде общедоступных данных, может быть ценным для вас, чтобы понять факторы, влияющие на их акторы, влияющие на их решения о покупке.

Вставка метаданных также обогащает данные.

Например, вычисление оценки настроения на основе журнала отзывов клиентов, сбор геопозиционных данные о погоде на курортах для анализа тенденций заполняемости, или фиксация времени публикации и тегов для сообщения в блоге.

После преобразования следующим этапом работы с данными является их проверка. Здесь вы проверяете качество структурирования, нормализации, очистки и обогащения данных, и обогащения.

Правила валидации относятся к повторяющимся шагам программирования, используемым для проверки согласованности, качества, и безопасности имеющихся данных.

Это подводит нас к публикации - четвертой фазе процесса обработки данных. Публикация включает в себя доставку выходных данных для последующих проектов. Публикуется преобразованная и проверенная версия входного набора данных вместе с метаданные об этом наборе данных.

Наконец, важно отметить критическую важность документирования шагов и соображений которые вы предприняли для преобразования необработанных данных в данные, готовые к анализу.

Все этапы работы с данными являются итерационными по своей природе. Для того чтобы повторить шаги и пересмотреть свои соображения по выполнению этих шагов, крайне важно документировать все соображения и действия.

Лекция 3. инструменты для обработки данных

В этом видео мы рассмотрим некоторые из популярных программ и инструментов для работы с данными, такие как: Excel Power Query / Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python и R. Начнем с самого базового программного обеспечения, используемого для ручного сбора данных и ручной обработки данных - электронных таблиц.

Электронные таблицы, такие как Microsoft Excel и Google Sheets, имеют множество функций и встроенных формулы, которые помогут вам выявить проблемы, очистить и преобразовать данные.

Существуют дополнения, которые позволяют импортировать данные из нескольких различных типов источников а также очищать и преобразовывать данные по мере необходимости - например, Microsoft Power Query для Excel и Google Sheets Query для Google Sheets.

OpenRefine - это инструмент с открытым исходным кодом, который позволяет импортировать и экспортировать данные в широком разнообразии форматах, таких как TSV, CSV, XLS, XML и JSON.

Используя OpenRefine, вы можете очищать данные, преобразовывать их из одного формата в другой и расширять данные с помощью веб-служб и внешних данных.

OpenRefine прост в освоении и использовании. Он предлагает операции на основе меню, что означает, что вам не нужно запоминать команды или синтаксис. Google DataPrep - это интеллектуальная облачная служба данных, которая позволяет визуально исследовать, очищать и подготавливать структурированные и неструктурированные данные к анализу.

Это полностью управляемая услуга, что означает, что вам не нужно устанавливать и управлять программным обеспечением или инфраструктурой.

DataPrep чрезвычайно прост в использовании.

При каждом действии вы получаете предложения о том, каким должен быть ваш идеальный следующий шаг. DataPrep может автоматически определять схемы, типы данных и аномалии. Watson Studio Refinery, доступная через IBM Watson Studio, позволяет обнаруживать, очищать, и преобразовывать данные с помощью встроенных операций.

Она преобразует большие объемы необработанных данных в качественную информацию, готовую к употреблению для аналитики.

Data Refinery обеспечивает гибкость при изучении данных, хранящихся в различных источниках данных. Он автоматически определяет типы и классификации данных, а также автоматически применяет применимые политики управления данными.

Trifacta Wrangler - это интерактивный облачный сервис для очистки и преобразования данных. Он берет беспорядочные, реальные данные, очищает и перестраивает их в таблицы данных, которые Затем их можно экспортировать в Excel, Tableau и R. Он известен своими функциями совместной работы, позволяя нескольким членам команды работать одновременно.

Python имеет огромную библиотеку и набор пакетов, которые предлагают мощные возможности для работы с данными.

Давайте рассмотрим некоторые из этих библиотек и пакетов.

Jupyter Notebook - веб-приложение с открытым исходным кодом, широко используемое для очистки и преобразования данных, статистического моделирования, а также визуализации данных.

Numpy, или Numerical Python, - самый базовый пакет, который предлагает Python. Он быстрый, универсальный, совместимый и простой в использовании. Он обеспечивает поддержку больших, многомерных массивов и матриц, а также высокоуровневых математических функции для работы с этими массивами.

Pandas предназначен для быстрого и простого анализа данных.

Он позволяет выполнять сложные операции, такие как слияние, объединение и преобразование огромных массивов данных, выполняемые с помощью простых однострочных команд.

Используя Pandas, вы можете предотвратить распространенные ошибки, возникающие в результате неправильного согласования данных, поступающих из разных источников.

R также предлагает ряд библиотек и пакетов, специально созданных для работы с беспорядочными данными, например Dplyr, Data.table и Jsonlite.

Используя эти библиотеки, вы можете исследовать, манипулировать и анализировать данные. Dplyr - это мощная библиотека для работы с данными. Она имеет точный и понятный синтаксис.

Data.table помогает быстро объединять большие наборы данных. Jsonlite - это надежный инструмент для разбора JSON, который отлично подходит для взаимодействия с веб-интерфейсами API.

Инструменты для работы с данными имеют различные возможности и размеры. Ваше решение о выборе лучшего инструмента для ваших нужд будет зависеть от следующих факторов специфических для вашего случая использования, инфраструктуры и команды - таких, как поддерживаемый размер данных, структуры данных, возможности очистки и преобразования, потребности инфраструктуры, простота использования и обучаемость.

ЛЕКЦИЯ 4

Согласно отчету Gartner о качестве данных, некачественные данные ослабляют конкурентоспособность организации и подрывают важнейшие цели бизнеса.

конкурентоспособность организации и подрывает важнейшие бизнес-цели.

Отсутствующие, непоследовательные или неверные данные могут привести к ложным выводам и, следовательно, неэффективным решениям.

решениям.

А в мире бизнеса это может дорого обойтись.

Наборы данных, собранные из разрозненных источников, могут иметь ряд проблем, включая недостающие

значения, неточности, дубликаты, неправильные или отсутствующие разделители, противоречивые записи,

и недостаточные параметры.

В некоторых случаях данные могут быть исправлены вручную или автоматически с помощью инструментов и скриптов для обработки данных.

но если их невозможно исправить, их необходимо удалить из набора данных.

Хотя термины Data Cleaning и Data Wrangling иногда используются как взаимозаменяемые,

важно помнить, что очистка данных - это лишь часть всего процесса очистки данных.

Wrangling.

Очистка данных является очень важной и неотъемлемой частью фазы преобразования

в рабочем процессе преобразования данных.

Типичный рабочий процесс очистки данных включает в себя: проверку, очистку и верификацию.

Первым шагом в процессе очистки данных является выявление различных типов проблем

и ошибок, которые может содержать ваш набор данных.

Вы можете использовать сценарии и инструменты, которые позволяют вам определить определенные правила и ограничения и

проверить данные на соответствие этим правилам и ограничениям.

Для проверки можно также использовать профилирование данных и инструменты визуализации данных.

Профилирование данных помогает вам исследовать исходные данные, чтобы понять структуру, содержание,

и взаимосвязи в ваших данных.

Оно позволяет выявить аномалии и проблемы с качеством данных.

Например, пустые или нулевые значения, дублирующиеся данные или попадает ли значение поля в

в ожидаемый диапазон.

Визуализация данных с помощью статистических методов может помочь вам обнаружить отклонения.

Например, построение графика среднего дохода в демографическом наборе данных может помочь вам обнаружить выбросы.

Это подводит нас к фактической очистке данных.

Методы, которые вы примените для очистки набора данных, будут зависеть от вашего сценария использования и

типа проблем, с которыми вы сталкиваетесь.

Давайте рассмотрим некоторые из наиболее распространенных проблем с данными.

Начнем с недостающих значений.

С недостающими значениями очень важно бороться, поскольку они могут привести к неожиданным или необъективным

результаты.

Вы можете отфильтровать записи с пропущенными значениями или найти способ получить источник этой

информацию, если она является неотъемлемой частью вашего сценария использования.

Например, недостающие данные о возрасте из демографического исследования.

Третьим вариантом является метод, известный как импутация, который рассчитывает недостающее значение на основе

статистических значений.

Ваше решение о выбранном способе действий должно быть основано на том, что является лучшим

для вашего случая использования.

Вы также можете столкнуться с дублирующими данными - точками данных, которые повторяются в вашем наборе данных.

Их необходимо удалить.

Другой тип проблем, с которыми вы можете столкнуться, - это нерелевантные данные.

Данные, которые не вписываются в контекст вашего сценария использования, можно считать нерелевантными

данные.

Например, если вы анализируете данные об общем состоянии здоровья части населения,

их контактные номера могут быть неактуальными для вас.

Очистка может включать в себя также преобразование типов данных.

Это необходимо для того, чтобы значения в поле хранились в соответствии с типом данных этого поля.

Например, числа хранятся как числовой тип данных или дата хранится как тип данных даты.

Также может потребоваться очистка данных для их стандартизации.

Например, для строк вы можете захотеть, чтобы все значения были в нижнем регистре.

Аналогично, форматы дат и единиц измерения должны быть стандартизированы.

Затем возникают синтаксические ошибки.

Например, белые пробелы или лишние пробелы в начале или конце строки - это синтаксическая ошибка.

ошибка, которую необходимо исправить.

Сюда же можно отнести исправление опечаток или формата, например, название штата вводится

в полной форме, например, Нью-Йорк, а в некоторых записях - в сокращенной форме, например, NY.

Данные также могут иметь отклонения, или значения, которые значительно отличаются от других наблюдений

в наборе данных.

Выбросы могут быть неправильными, а могут и не быть.

Например, если поле "Возраст" в базе данных избирателей имеет значение 5, вы знаете, что это неверные данные и их нужно исправить.

данные и должны быть исправлены.

Теперь рассмотрим группу людей, где годовой доход находится в диапазоне от ста

тысяч до двухсот тысяч долларов - за исключением одного человека, который зарабатывает миллион долларов

в год.

Хотя эта точка данных не является неверной, она представляет собой выброс, и на нее нужно обратить внимание.

В зависимости от вашего сценария использования, вам может понадобиться решить, не исказит ли включение этих данных

результаты не в лучшую сторону.

Это подводит нас к следующему шагу в рабочем процессе очистки данных - проверке.

На этом этапе вы проверяете результаты, чтобы определить эффективность и точность, достигнутые в результате операции по очистке данных.

операции по очистке данных.

Вам необходимо повторно проверить данные, чтобы убедиться, что правила и ограничения, применяемые к данным, остаются в силе после внесения исправлений.

после внесенных вами исправлений.

И, в конце концов, важно отметить, что все изменения, предпринятые в рамках операции по очистке данных

должны быть задокументированы.

Не только сами изменения, но и причины их внесения, а также качество

хранящихся в настоящее время данных.

Отчетность о том, насколько здоровы данные, является очень важным шагом.

ТОЧКИ ЗРЕНИЯ

ца

В этом сегменте профессионалы в области данных делятся

какая часть их работы связана со сбором,

очистку и подготовку данных для анализа.

Я бы сказал, что относительно большая часть

моей работы связана со сбором,

подготовка и очистка данных для анализа.

Я работаю в компании с

очень хорошая команда инженеров по обработке данных.

Поэтому мне не приходится выполнять такую работу

как некоторым другим специалистам по анализу данных.

Но, тем не менее, любой человек, который тесно работает с данными,

будь то специалист по данным,

аналитик данных, инженер машинного обучения,

действительно должен

понимать, откуда берутся данные.

Неизбежно, что ни один набор данных не является идеальным.

Всегда будут компромиссы или небольшие ошибки.

Поэтому очень важно потратить

значительную часть своего времени,

понимая подчеркнутые данные, которые были использованы для

создания набора данных и

какие потенциальные проблемы могут быть связаны с этими данными.

Моя работа в качестве CPA включает в себя много анализа.

Финансовые отчеты, деятельность по счетам,

оценка процессов и средств контроля.

Сбор данных может быть довольно простым, если,

бухгалтерская информация хранится

в главной бухгалтерской системе

или в центральном хранилище, где данные легко собрать.

Вероятно, около 30 процентов работы

это разложить все по полочкам.

И когда вы переходите к аналитике,

вы сможете погрузиться в самую суть.

Итак, вам нужно отслеживать данные,

убедиться в их точности,

убедиться, что все сходится.

Убедитесь, что у вас есть вся информация.

Например, по финансовым отчетам,

я должен убедиться, что люди предоставили

мне 12 месяцев [неслышно] отчетов,

что я не упустил никаких данных, а если упустил,

что у меня достаточно информации, чтобы быть в состоянии спрогнозировать или

прогнозировать или даже оглянуться назад, чтобы

оценить, что было сделано

в [неслышно] на основе того, что у меня есть.

Это, безусловно, полезно.

В этом сегменте профессионалы в области данных рассказывают о

шагах, которые они предпринимают для обеспечения надежности данных.

Один из важнейших шагов для того, чтобы

чтобы убедиться, что ваши данные надежны,

это запуск сводной статистики по отдельным столбцам

данных и убедиться, что

что они соответствуют действительности.

Например, если у вас есть столбец

где-нибудь записываются посещения

в месяц на веб-сайт, и вы

запустите сводную статистику по этому столбцу,

вы получите минимум,

среднее значение, медиану, максимум,

и вы увидите что-то забавное, например,

в одном месяце было отрицательное количество посещений или что-то в этом роде.

Вы знаете, эти данные ненадежны.

Финансовая информация, в частности, должна быть надежной.

Она должна быть не предвзятой.

В ней не должно быть ошибок.

Это лишь некоторые из многих характеристик.

которые необходимы для того, чтобы на данные можно было положиться.

Поэтому, прежде чем вникать в детали, нужно провести так называемую логическую

прежде чем вникать в детали транзакции.

Имеет ли она смысл на высоком уровне?

Если вы ожидали, что доход от продаж увеличится,

но видите, что она резко снизилась,

то сначала разберитесь в этой части.

Верен ли мой источник?

Выполняю ли я запрос в правильном периоде?

Правильно ли я выбрал счет главной книги?

Начните с этого, убедитесь, что

что основные вопросы целостности данных были решены в первую очередь.

Как только мы убедимся, что данные надежны,

тогда мы можем начать углубляться

в анализ и формировать выводы

о финансовых показателях

на основе нашего анализа данных.

ИТОГИ

В этом уроке вы узнали следующую информацию:

После того как определенные вами данные собраны и импортированы, следующий шаг - сделать их пригодными для анализа. Именно здесь и начинается процесс Data Wrangling или Data Munging. Обработка данных - это итеративный процесс, который включает в себя исследование, преобразование и проверку данных.

Преобразование необработанных данных включает в себя задачи, которые вы решаете для:

Структурное манипулирование и объединение данных с помощью джойнов и юнионов.

Нормализация данных, то есть очистка базы данных от неиспользуемых и избыточных данных.

Денормализация данных, то есть объединение данных из нескольких таблиц в одну таблицу, чтобы их можно было быстрее запросить.

Очистка данных, которая включает в себя профилирование данных для выявления проблем с качеством, визуализацию данных для выявления выбросов и устранение таких проблем, как недостающие значения, дубликаты данных, нерелевантные данные, несоответствующие форматы, синтаксические ошибки и выбросы.

Обогащение данных, которое включает в себя рассмотрение дополнительных точек данных, которые могут повысить ценность существующего набора данных и привести к более содержательному анализу.

Для процесса обработки данных существует множество программного обеспечения и инструментов. Некоторые из наиболее часто используемых включают Excel Power Query, Spreadsheets, OpenRefine, Google DataPrep, Watson Studio Refinery, Trifacta Wrangler, Python и R, каждый из которых имеет свой набор характеристик, сильных сторон, ограничений и областей применения.

MODULE 7 - ANALYZING AND MINING DATA

a

Прежде чем мы поймем, что такое статистический анализ, его связь с анализом данных и, в частности, с

добычи данных, давайте сначала разберемся, что такое статистика. Статистика - это отрасль математики

занимающаяся сбором, анализом, интерпретацией и представлением числовых или количественных данных.

данных. Она окружает нас повсюду в нашей повседневной жизни. Говорим ли мы о среднем

доходах, среднем возрасте или самых высокооплачиваемых профессиях - все это статистика. Сегодня статистика

применяется в различных отраслях для принятия решений на основе данных. Например, исследователи используют

статистику для анализа данных, полученных при производстве вакцин, с целью обеспечения безопасности и эффективности,

или компании, использующие статистику для снижения оттока клиентов путем более глубокого изучения их потребностей.

требования. Теперь давайте рассмотрим, что такое статистический анализ. Статистический анализ - это применение

статистических методов к выборке данных с целью выработки понимания того, что представляют собой

что представляют собой эти данные. Он включает в себя сбор и тщательное изучение каждого образца данных в наборе

элементов, из которых могут быть взяты образцы. Выборка в статистике - это репрезентативная

выборка, взятая из общей совокупности, где совокупность - это дискретная группа людей или

вещей, которые могут быть идентифицированы по крайней мере по одной общей характеристике для целей

сбора и анализа данных. Например, в определенном случае использования население может представлять собой всех

люди в штате, имеющие водительские права, и выборка из этой популяции, которая является

частью, или подмножеством, населения могут быть мужчины-водители старше 50 лет. Статистические

Статистические методы в основном полезны для того, чтобы убедиться, что данные интерпретированы правильно, а очевидные взаимосвязи

являются значимыми, а не просто случайными. Всякий раз, когда мы собираем данные из выборки, существует

есть два различных типа статистики, которые мы можем использовать. Описательная статистика для обобщения информации

о выборке; и инференциальная статистика, позволяющая делать выводы или обобщения о

более широкой популяции. Описательная статистика позволяет представить данные в осмысленном виде.

что позволяет упростить интерпретацию данных. Данные описываются с помощью сводных диаграмм,

таблиц и графиков без каких-либо попыток сделать выводы о населении, из которого взята выборка.

из которой взята выборка. Цель состоит в том, чтобы облегчить понимание и визуализацию

необработанные данные, не делая выводов относительно выдвинутых гипотез. Например,

мы хотим описать результаты теста по английскому языку в определенном классе из 25 учеников. Мы записываем

результаты тестов всех учащихся, вычисляем сводную статистику и строим график.

Некоторые из распространенных показателей описательного статистического анализа включают центральную тенденцию,

дисперсия и перекос: Центральная тенденция, или нахождение центра выборки данных. Некоторые

распространенные показатели центральной тенденции включают среднее значение, медиану и моду. Эти показатели

говорят вам, куда попадает большинство значений в вашем наборе данных. Так, в предыдущем примере среднее значение

или среднее математическое значение для класса из 25 учеников будет представлять собой общую сумму

оценок всех 25 учеников, деленная на 25, то есть на количество учеников. Если

вы упорядочите вышеприведенный набор данных от наименьшего значения баллов до наибольшего значения баллов из

25 студентов и выбрать среднее значение - то есть значение с 12 значениями слева

и 12 значений справа от значения балла, то это значение будет медианой для данного набора данных.

набора данных. Если 12 студентов набрали меньше 75%, а 12 студентов набрали больше

75%, то медиана равна 75. Медиана уникальна для каждого набора данных и не подвержена влиянию выбросов.

Режим - это значение, которое встречается наиболее часто в наборе наблюдений. Например, если

наиболее распространенный балл в этой группе из 25 студентов составляет 72%, то это и есть мода для

этого набора данных. Итак, вы видите, как рассмотрение вашего набора данных через эти значения может помочь

получить более четкое представление о наборе данных. Дисперсия - это мера изменчивости в

наборе данных. Общепринятыми мерами статистической дисперсии являются дисперсия, стандартное отклонение,

и диапазон. Дисперсия определяет, насколько далеко точки данных отстоят от центра, т.е. от распределения значений,

распределения значений. Когда распределение имеет меньшую вариативность, значения в наборе данных

более последовательны. Однако, когда изменчивость выше, точки данных более разнородны,

и экстремальные значения становятся более вероятными. Понимание изменчивости может помочь вам понять вероятность

наступления того или иного события. Стандартное отклонение показывает, насколько тесно группируются ваши данные

вокруг среднего значения. А диапазон показывает расстояние между наименьшим и наибольшим значениями в

в вашем наборе данных. Перекос - это мера того, является ли распределение значений симметричным

вокруг центрального значения или перекошено влево или вправо. Перекошенные данные могут влиять на то, какие типы анализа

можно проводить. Это некоторые из основных и наиболее часто используемых инструментов описательной статистики.

инструменты, но есть и другие инструменты, например, использование корреляции и диаграмм рассеяния

для оценки взаимосвязи парных данных. Второй тип статистического анализа - это

инференциальная статистика. Инференциальная статистика использует данные из выборки, чтобы сделать выводы

о более крупной совокупности, из которой была взята выборка. Используя методы инференциальной

статистики вы можете сделать обобщения, которые применяют результаты выборки к популяции

как единое целое. Некоторые распространенные методы инференциальной статистики включают проверку гипотез, доверительные интервалы и регрессионный анализ.

интервалы и регрессионный анализ: Проверка гипотез - например, для изучения эффективности

вакцины путем сравнения результатов в контрольной группе, проверка гипотез может сказать вам, является ли

Эффективность вакцины, наблюдаемая в контрольной группе, скорее всего, существует и в популяции

также и в популяции. Доверительные интервалы включают в себя неопределенность и ошибку выборки, чтобы создать

диапазон значений, в который, скорее всего, попадет фактическое значение в популяции. Регрессионный анализ

включает в себя проверку гипотез, которая помогает определить, существуют ли взаимосвязи, наблюдаемые в

Данные выборки действительно существуют в популяции, а не только в выборке. Существуют различные

пакеты программного обеспечения для проведения статистического анализа данных, такие как Statistical Analysis System

(или SAS), Statistical Package for the Social Sciences (или SPSS) и Stat Soft. Статистика

составляют основу добычи данных, поскольку: Обеспечивает меры и методологии, необходимые для

и выявления закономерностей, которые помогают определить различия между случайным шумом

и значимыми результатами. Как добыча данных, о которой мы узнаем больше в этом курсе,

и статистика, как методы анализа данных, помогают в принятии более эффективных решений.

ЛЕКЦИЯ 2. ЧТО ТАКОЕ ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ?

Добыча данных или процесс извлечения знаний из данных,

это сердце процесса анализа данных. Это

междисциплинарная область, которая включает в себя использование технологий распознавания образов

технологий распознавания образов, статистического анализа и

математических методов. Его цель - выявить корреляции

в данных, поиск закономерностей и

вариации. Понимание тенденций и прогнозирование вероятностей.

Вы часто будете слышать о закономерностях и тенденциях в контексте

анализа данных, поэтому давайте сначала разберемся в этих понятиях.

Распознавание образов - это обнаружение закономерностей или

общности в данных.

Рассмотрим данные журнала регистрации входов в приложение в какой-либо

организации. Они содержат такую информацию, как

имя пользователя, метка времени входа, время, проведенное в каждом сеансе входа, и

выполненные действия. Когда мы анализируем эти данные, чтобы получить

понимание привычек или поведения пользователей, например,

время дня, когда максимальное количество пользователей обычно входит в систему, или пользовательские

роли, которые обычно проводят максимальное количество часов, заходя в приложение

приложение или модули в приложении рабочего процесса, которые

при изучении данных вручную или с помощью инструментов

для выявления закономерностей, скрытых в

данных. Тенденция, с другой стороны, - это общая тенденция

набора данных изменяться с течением времени. Например, глобальное

потепление в краткосрочной перспективе, например, год от года

температура может оставаться неизменной или повышаться или понижаться на несколько

градусов, но в целом глобальные температуры продолжают

повышаться в течение длительного времени, что делает глобальное потепление тенденцией.

Добыча данных находит применение в различных отраслях промышленности и

дисциплинах. Например, профилирование поведения клиентов

потребностей и располагаемого дохода для того, чтобы предложить целевые

кампании, финансовые учреждения, отслеживающие

транзакций на предмет необычного поведения и выявления

мошеннических операций с использованием моделей интеллектуального анализа данных.

Использование статистических моделей для прогнозирования вероятности развития у пациента

конкретных заболеваний и определения приоритетов лечения.

Доступ к данным об успеваемости учащихся для прогнозирования уровня достижений

уровня успеваемости и целенаправленных усилий по оказанию поддержки там, где это необходимо.

где это необходимо. Помощь следственным органам в развертывании полицейских сил

там, где вероятность преступлений выше, и согласовывать поставки и

логистику с прогнозами спроса.

Существует несколько методов, которые можно использовать для обнаружения закономерностей и

построить точные модели для обнаружения, будь то описательные,

диагностическое, прогностическое или предписывающее моделирование. Давайте

разберем некоторые из наиболее часто используемых методов.

Классификация - это техника, которая классифицирует атрибуты в

целевые категории, например, классификация клиентов на низких,

средние или высокие траты на основе того, сколько они зарабатывают.

Кластеризация похожа на классификацию, но предполагает

группировку данных в кластеры, чтобы их можно было рассматривать как группы.

Например, кластеризация клиентов на основе географических

Аномалия или обнаружение выбросов - это метод, который

помогает найти закономерности и данные, которые не являются нормальными или

неожиданными. Например, скачки в использовании кредитной карты

которые могут указывать на возможное злоупотребление.

Поиск ассоциативных правил - это метод, который помогает установить

взаимосвязь между двумя событиями данных. Например.

покупка ноутбука часто сопровождается

покупка охлаждающей подставки. Последовательные модели - это

техника, позволяющая проследить серию событий, происходящих в определенной

последовательности. Например, отслеживание покупательского маршрута от момента входа в интернет-магазин до момента выхода из него.

момента входа в интернет-магазин до момента выхода из него.

Группировка по сродству - это техника, используемая для обнаружения совместного возникновения

в отношениях. Эта техника широко используется в интернет-магазинах

для перекрестных продаж и повышения продаж своей продукции путем рекомендации

продукты людям, основываясь на истории покупок других людей.

которые приобрели тот же товар.

Деревья решений помогают строить модели классификации в

форме древовидной структуры с множеством ветвей, где каждая

каждая ветвь представляет собой вероятное событие. Эта техника помогает

построить четкое понимание взаимосвязи между

входом и выходом.

Регрессия - это метод, который помогает определить характер взаимосвязи между двумя переменными, которая может быть причинно-следственной.

взаимосвязи между двумя переменными, которая может быть причинно-следственной

или корреляционной. Например, на основе таких факторов, как

местоположение и площадь, может быть использована регрессионная модель

для прогнозирования стоимости дома.

Добыча данных, по сути, помогает отделить шум от реальной

информации и помогает предприятиям сосредоточить свои усилия только на том.

на том, что имеет значение.

ЛЕКЦИЯ

а

В этом видео мы узнаем о некоторых широко используемых программах и инструментах для добычи данных

добычи данных, таких как: Электронные таблицы, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio;

и SAS. Электронные таблицы, такие как Microsoft Excel и

Google Sheets, обычно используются для выполнения основных задач по поиску данных. Электронные таблицы могут

могут использоваться для размещения данных, экспортированных из других систем в легкодоступном

и легко читаемом формате. Для демонстрации конкретных аспектов данных можно использовать поворотные таблицы,

что очень важно, когда вам нужно отсортировать и проанализировать огромное количество данных. Они

также облегчают сравнение между различными наборами данных. Дополнения, доступные

для Excel, такие как Data Mining Client for Excel, XLMiner и KnowledgeMiner for

Excel, позволяют выполнять общие задачи по добыче данных, такие как классификация, регрессия, ассоциативные правила, кластеризация и моделирование.

правила, кластеризация и построение моделей. GoogleSheets также имеет множество надстроек, которые можно

использоваться для анализа и добычи данных, например, анализ текста, добыча текста, Google Analytics.

R - один из наиболее широко используемых языков для выполнения статистического моделирования и вычислений

статистиками и специалистами по добыче данных. В состав R входят сотни библиотек, специально

созданных для операций по добыче данных, таких как регрессия, классификация, кластеризация данных, поиск ассоциаций и правил, анализ текстов, анализ выбросов.

извлечение правил, извлечение текста, обнаружение выбросов и анализ социальных сетей. Некоторые из популярных

Пакеты R включают tm и twitteR. tm, фреймворк для приложений интеллектуального анализа текста в R, предоставляет функции для интеллектуального анализа текста.

функции для интеллектуального анализа текста. twitteR - основа для интеллектуального анализа твитов. R Studio - это

популярная интегрированная среда разработки с открытым исходным кодом (IDE) для работы с языком программирования R.

языком программирования R. Библиотеки Python, такие как Pandas и NumPy.

широко используются для добычи данных. Pandas - это модуль с открытым исходным кодом для работы со структурами данных

и анализа. Возможно, это одна из самых популярных библиотек для анализа данных в Python.

Он позволяет загружать данные в любом формате и предоставляет простую платформу для организации,

сортировки и манипулирования этими данными. Используя Pandas, вы можете: выполнять основные численные вычисления

такие как среднее значение, медиана, мода и диапазон; вычислять статистику и отвечать на вопросы, касающиеся

корреляции между данными и распределения данных; исследовать данные визуально и количественно;

визуализировать данные с помощью других библиотек Python. NumPy - это инструмент для математических

вычислений и подготовки данных на языке Python. NumPy предлагает множество встроенных функций

и возможностей для добычи данных. Блокноты Jupyter Notebooks стали предпочтительным инструментом для

Data Scientists и Data Analysts при работе с Python для выполнения поиска данных и статистического

анализа. SPSS расшифровывается как Statistical Process for Social

наук. Хотя название предполагает его первоначальное использование в области социальных наук, он

SPSS широко используется для расширенной аналитики, анализа текстов, анализа тенденций, проверки достоверности

предположений и перевода бизнес-проблем в решения на основе науки о данных.

SPSS имеет закрытый исходный код и требует лицензии для использования. SPSS имеет простой в использовании интерфейс, который

Для выполнения сложных задач требуется минимальное кодирование. Она включает в себя эффективные инструменты управления данными

и пользуется популярностью благодаря своим возможностям глубокого анализа и точным результатам.

IBM Watson Studio, включенная в пакет IBM Cloud Pak for Data, использует коллекцию инструментов с открытым исходным кодом, таких как Jupyter.

инструментов с открытым исходным кодом, таких как блокноты Jupyter, и расширяет их инструментами IBM с закрытым исходным кодом.

что превращает его в мощную среду для анализа данных и науки о данных. Она доступна

через веб-браузер в общедоступном облаке, частном облаке и в виде приложения для настольных компьютеров.

Watson Studio позволяет членам команды совместно работать над проектами, которые могут варьироваться от простого исследовательского анализа до создания моделей машинного обучения и искусственного интеллекта.

анализа до создания моделей машинного обучения и искусственного интеллекта. Она также включает в себя SPSS Modeller

которые позволяют быстро разрабатывать прогностические модели для бизнес-данных.

SAS Enterprise Miner - это комплексная графическая рабочая среда для добычи данных. Он предоставляет мощные

возможности для интерактивного исследования данных, что позволяет пользователям выявлять взаимосвязи

внутри данных. SAS может управлять информацией из различных источников, добывать и преобразовывать данные,

и анализировать статистику. Он предлагает графический интерфейс пользователя для нетехнических пользователей. С помощью

SAS вы можете: выявлять закономерности в данных, используя ряд доступных методов моделирования;

исследовать взаимосвязи и аномалии в данных; анализировать большие данные; проверять достоверность

выводов, полученных в процессе анализа данных. SAS очень прост в использовании благодаря своему синтаксису

и также легко отлаживается. Она способна работать с большими базами данных и обеспечивает высокую безопасность для своих пользователей.

безопасность для своих пользователей. В этом видео мы познакомились лишь с некоторыми из инструментов для анализа данных.

доступных сегодня. . Ваше решение о выборе лучшего инструмента для ваших нужд будет определяться

размер и структура данных, которые поддерживает инструмент, его функции, возможности визуализации данных.

возможностями, потребностями инфраструктуры, простотой использования и обучаемостью. Довольно часто

использовать комбинацию инструментов для добычи данных, чтобы удовлетворить все ваши потребности.

ИТОГИ

В этом уроке вы узнали следующую информацию:

Статистика - это отрасль математики, занимающаяся сбором, анализом, интерпретацией и представлением числовых или количественных данных.

Статистический анализ предполагает использование статистических методов для того, чтобы понять, что представляют собой данные.

Статистический анализ может быть:

Описательным; он дает краткое представление о том, что представляют собой данные. Общие показатели включают центральную тенденцию, дисперсию и перекос.

Инференциальный; который включает в себя выводы или обобщения о данных. Общие показатели включают проверку гипотез, доверительные интервалы и регрессионный анализ.

Data Mining, проще говоря, это процесс извлечения знаний из данных. Он включает в себя использование технологий распознавания образов, статистического анализа и математических методов для выявления корреляций, закономерностей, вариаций и тенденций в данных.

Существует несколько методов, которые могут помочь в добыче данных, например, классификация атрибутов данных, объединение данных в группы, установление взаимосвязей между событиями, переменными, входом и выходом.

Для анализа и добычи данных существует множество программ и инструментов. Некоторые из наиболее часто используемых включают электронные таблицы, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio и SAS, каждый из которых имеет свой набор характеристик, сильных сторон, ограничений и областей применения.

MODULE 8 - COMMUNICATING DATA ANALYSIS FINDINGS

Лекция 1

нца

Процесс анализа данных начинается с понимания проблемы, которую необходимо решить, и

желаемого результата, который должен быть достигнут.

А заканчивается он передачей полученных результатов таким образом, чтобы они повлияли на принятие решений.

Проекты по работе с данными являются результатом совместных усилий, охватывающих все бизнес-функции, в которых участвуют

людей с многопрофильными навыками, а полученные результаты включаются в более крупную

бизнес-инициативу.

Успех вашей коммуникации зависит от того, насколько хорошо другие смогут понять и довериться вашим выводам, чтобы предпринять дальнейшие действия.

вашим выводам, чтобы предпринять дальнейшие действия.

Поэтому, как аналитики данных, вы должны рассказать историю с помощью своих данных, визуализировав их наглядно

и создавая структурированное повествование, явно ориентированное на вашу аудиторию.

Прежде чем приступить к созданию коммуникации, вам необходимо восстановить связь со своей аудиторией.

Начните с того, что задайте себе следующие вопросы: кто моя аудитория?

Что для них важно?

Что поможет им доверять мне?

Ваша аудитория в основном будет представлять собой разнообразную группу - с точки зрения бизнес-функций.

которые они представляют, играют ли они оперативную или стратегическую роль в организации, насколько

насколько на них влияет проблема, и другие подобные факторы.

Ваша презентация должна быть построена на том уровне информации, которым уже обладает ваша аудитория.

имеет.

Основываясь на своем понимании аудитории, вы решите, какая информация и в каком объеме

необходима для лучшего понимания ваших выводов.

Очень заманчиво привести все данные, с которыми вы работали, но вы должны

следует подумать, какие фрагменты более важны для вашей аудитории, чем другие.

Презентация - это не свалка данных.

Факты и цифры сами по себе не влияют на решения и не побуждают людей к действию.

Вы должны рассказать убедительную историю.

Включите в презентацию только ту информацию, которая необходима для решения бизнес-проблемы.

Слишком много информации заставит вашу аудиторию с трудом понять, что вы хотите сказать.

что вы хотите сказать.

Начните презентацию с демонстрации своего понимания бизнес-проблемы.

аудитории.

Легко отступить от предположения, что все мы знаем, для чего мы здесь, но

отражение вашего понимания проблемы, которую необходимо решить, и результата, который

необходимо достичь, - это отличный первый шаг к тому, чтобы завоевать их внимание и зародить доверие.

доверие.

Говорить на языке бизнес-сферы организации - еще один важный фактор

в установлении связи между вами и вашей аудиторией.

Следующим шагом в разработке вашей коммуникации является структурирование и организация вашей презентации

для достижения максимального эффекта.

Ссылайтесь на собранные вами данные.

Помните, что данные, являющиеся основой всего, что вы сообщаете, являются

как черный ящик для аудитории.

Если вы не сможете установить достоверность ваших данных, люди не будут знать, что они могут

могут доверять вашим выводам.

Поделитесь своими источниками данных, гипотезами и проверками.

Работайте над созданием достоверности ваших выводов по ходу дела - не упускайте из виду

ключевые предположения, сделанные в ходе анализа.

Организуйте информацию в логические категории на основе имеющейся у вас информации - есть ли у вас

например, у вас есть качественная и количественная информация?

Обдуманно выбирайте подход "сверху вниз" или "снизу вверх" в своем изложении.

Оба подхода могут быть эффективными - в зависимости от вашей аудитории и конкретного случая использования.

Будьте последовательны в своем подходе.

Важно определить, какие форматы коммуникации будут наиболее полезны для вашей аудитории.

Нужно ли им взять с собой резюме, фактологический бюллетень или отчет?

Как ваша аудитория собирается использовать представленную вами информацию, это должно определять

выбранные вами форматы.

Информация должна быть изложена таким образом, чтобы побуждать к действию.

Если ваша аудитория не понимает значимости вашей идеи или не убеждена в ее

полезности, инсайт не будет иметь никакой ценности.

Эссе в тысячу слов не окажет такого же воздействия, как визуальный образ, создавая четкий мысленный

образ в сознании вашей аудитории.

Мощная визуализация рассказывает историю через графическое изображение фактов и цифр.

Визуализация данных - графики, диаграммы, схемы - это отличный способ оживить данные.

Независимо от того, показываете ли вы сравнение, взаимосвязь, распределение или состав, у вас есть инструменты.

которые помогут вам показать закономерности и выводы гипотез.

Данные имеют ценность благодаря историям, которые они рассказывают.

Ваша аудитория должна быть способна доверять вам, понимать вас и относиться к вашим выводам и идеям.

Устанавливая достоверность ваших выводов, представляя данные в рамках повествования и

подкрепляя их визуальными впечатлениями, вы можете помочь своей аудитории получить ценные

выводы.

ТОЧКИ ЗРЕНИЯ

до конца

В этом видео мы послушаем, как

профессионалов в области данных о том, какую

роли, которую играют рассказывание историй в жизни аналитика данных.

Роль повествования в

жизни аналитика данных невозможно переоценить.

Очень важно

по-настоящему хорошо рассказывать истории с помощью данных.

Я думаю, что люди естественным образом

понимают мир через истории.

Если вы пытаетесь убедить кого-либо

сделать что-либо с помощью данных,
первое, что вы должны сделать, это рассказать четкую,
краткую, убедительную историю.
Я также думаю, что может быть очень полезно для
аналитику данных разработать историю
в любое время, когда они работают с
набором данных, чтобы помочь себе лучше
понять, что лежит в основе набора данных и что он делает.
Всегда будет существовать баланс между
рассказом ясной, связной,
простой историей и уверенностью в том, что
что вы передаете все сложности.
которые вы можете найти в данных.
Я думаю, что найти этот баланс может быть очень сложно,
но это очень важно.
Искусство рассказывать истории
играет важную роль в жизни аналитика данных.
Не имеет значения, как много или
сколько замечательной информации вы получили.
Если вы не можете найти способ
донести ее до вашей аудитории,
будь то потребитель или
директорского или руководящего уровня,
тогда все напрасно.
Вы должны найти способ

донести это, и обычно

лучше всего сделать это в визуальной форме или через рассказ истории,

чтобы они поняли, как

эта информация может быть полезной.

Я должен сказать, что умение рассказывать истории - это важный навык.

Это как последняя миля в доставке информации.

Многие люди могут справиться с

с технической стороной через короткий период обучения.

Однако способность извлекать ценность из

данных и доносить их до слушателей - в дефиците.

Если вы думаете о долгосрочной карьере,

я считаю, что очень важно

знать, как рассказать убедительную историю с использованием данных.

Повествование абсолютно необходимо для данных и аналитики.

Это то, как вы на самом деле передаете свое послание.

Каждый может показать цифры,

но если у вас нет истории,

если у вас нет убедительной причины действовать,

то, в конечном счете, то, что вы представляете, не будет

не найдет отклика у вашей аудитории.

Они провели исследование в

Стэнфорде, где люди представляли

свои презентации, и в этих презентациях были просто KPI,

цифры статистики, но они также рассказывали историю.

После этого членов аудитории опросили,

что они запомнили из каждой из этих презентаций,

и именно эти истории запомнились им.

Да, в них по-прежнему были факты и

цифры, содержащиеся в истории,

но именно таким образом вы доносите их до слушателей.

Эмоциональная связь

с историей, с пониманием,

с данными - это то, как вы заставите людей

предпринять те действия, которые вы хотите и должны предпринять.

Лекция 3. Визуализация

ца

Визуализация данных - это дисциплина передачи информации с помощью визуальных элементов.

таких как графики, диаграммы и карты.

Ее цель - сделать информацию простой для восприятия, интерпретации и запоминания.

Представьте себе, что вам приходится просматривать тысячи строк данных, чтобы сделать выводы и сравнить их с визуальным представлением тех же данных.

с визуальным представлением этих же данных, обобщающим выводы.

Используя визуализацию данных, вы можете представить краткое описание взаимосвязей, тенденций и

закономерности, скрытые в данных, которые если и не невозможно, то очень трудно расшифровать

на основе массива данных.

Чтобы визуализация данных имела ценность, необходимо выбрать визуализацию, которая наиболее

эффективно донести ваши выводы до аудитории.

А для этого вам нужно начать с того, чтобы задать себе несколько вопросов.

Какую взаимосвязь я пытаюсь установить?

Хочу ли я сравнить относительную долю составных частей одного целого, например,

вклад различных продуктовых линий в общий доход компании?

Хочу ли я сравнить несколько величин, например, количество проданных продуктов и выручку

за последние три года?

Или я хочу проанализировать одно значение с течением времени, что в данном примере может означать, как

продажи одного конкретного продукта изменились за последние три года.

Нужно ли мне, чтобы моя аудитория увидела корреляцию между двумя переменными?

Например, корреляцию между погодными условиями и бронированиями на горнолыжном курорте.

Хочу ли я обнаружить аномалии в данных - например, найти в них значения, которые могут

потенциально исказить результаты?

Вопрос, на который я пытаюсь ответить, - это не просто всеобъемлющий вопрос в

дизайн и процесс визуализации данных - вы должны быть в состоянии ответить на этот вопрос для

вы должны быть в состоянии ответить на этот вопрос для вашей аудитории с каждым набором данных и информацией, которые вы визуализируете.

Также необходимо учитывать, должна ли визуализация быть статичной или интерактивной.

Интерактивная визуализация, например, может позволить вам изменять значения и видеть

влияние на связанную переменную в режиме реального времени.

Итак, подумайте о ключевых выводах для вашей аудитории, предвосхитите ее информационные потребности

и вопросы, которые у них могут возникнуть, а затем спланируйте визуализацию, которая четко донесет ваше

сообщение четко и эффективно.

Давайте рассмотрим несколько основных примеров типов графиков, которые вы можете создать для визуализации

ваших данных.

Гистограммы отлично подходят для сравнения связанных наборов данных или частей одного целого.

Например, на этой гистограмме вы можете увидеть численность населения 10 различных стран

и как они соотносятся друг с другом.

Диаграммы столбцов сравнивают значения рядом друг с другом.

Их можно эффективно использовать, чтобы показать изменения во времени.

Например, показать, как меняются просмотры страниц и время сеансов пользователей на вашем сайте

от месяца к месяцу.

Хотя гистограммы и столбцовые диаграммы похожи, за исключением ориентации, их нельзя всегда использовать

не всегда могут быть взаимозаменяемыми.

Например, столбчатая диаграмма может лучше подходить для отображения отрицательных и положительных значений.

Круговые диаграммы показывают разбивку объекта на составляющие и долю этих составляющих по отношению друг к другу.

частей по отношению друг к другу.

Каждая часть круговой диаграммы представляет собой статическое значение или категорию, а сумма всех категорий

равна ста процентам.

В данном примере в маркетинговой кампании с четырьмя маркетинговыми каналами - социальными сайтами, нативной рекламой, платными агентами влияния и платной рекламой.

реклама, платные агенты влияния и живые мероприятия - вы можете увидеть общее количество лидов, сгенерированных

по каждому каналу.

Линейные диаграммы отображают тенденции.

Они отлично подходят для того, чтобы показать, как изменяется значение данных в зависимости от непрерывной переменной.

Например, как изменились продажи вашего продукта или нескольких продуктов с течением времени.

время - непрерывная переменная.

Линейные диаграммы можно использовать для понимания тенденций, закономерностей и вариаций в данных;

а также для сравнения различных, но связанных наборов данных с несколькими сериями.

Визуализация данных также может использоваться для построения информационных панелей.

Приборные панели организуют и отображают отчеты и визуализации, поступающие из нескольких источников данных

в единый графический интерфейс.

Вы можете использовать приборные панели для мониторинга ежедневного прогресса или общего состояния бизнес-функции

или даже конкретного процесса.

Приборные панели могут представлять как оперативные, так и аналитические данные.

Например, у вас может быть маркетинговая панель, с помощью которой вы отслеживаете текущую маркетинговую

кампанию на предмет охвата аудитории, созданных запросов и конверсии продаж в режиме реального времени.

В рамках этой же приборной панели вы также можете видеть, как коэффициент конверсии этой

кампании по сравнению с коэффициентом конверсии некоторых успешно проведенных кампаний в

прошлом.

Приборные панели - это отличный инструмент, позволяющий представить полную картину с высоты птичьего полета и в то же время

позволяя при этом детализировать информацию до следующего уровня по каждому параметру.

Приборные панели: просты для понимания обычным пользователем

облегчают сотрудничество между командами; и позволяют генерировать отчеты на ходу.

Используя приборные панели, вы можете увидеть результат изменения данных и показателей практически мгновенно - и

это может помочь вам оценить ситуацию с разных точек зрения, на ходу, без

без необходимости возвращаться к чертежной доске.

Лекция 4. Софт для визуализации данных

онца

В этом видео мы рассмотрим некоторые из наиболее часто используемых программ для визуализации данных

и инструменты.

К ним относятся: Электронные таблицы, Jupyter Notebook и библиотеки Python, R-Studio и R-Shiny,

IBM Cognos Analytics, Tableau и Microsoft Power BI.

Некоторые из них представляют собой комплексные решения для анализа данных, в то время как другие предназначены специально для

визуализации данных - от бесплатных инструментов с открытым исходным кодом до коммерческих решений.

Электронные таблицы, такие как Microsoft Excel и Google Sheets, возможно, являются наиболее часто используемым программным обеспечением для создания графических представлений данных.

используемое программное обеспечение для создания графических представлений наборов данных.

Электронные таблицы просты в освоении и имеют тонну документации и видеоуроков, доступных

в Интернете для ознакомления.

Excel предоставляет несколько типов диаграмм, начиная от базовых столбчатых, линейных, круговых и разворотных диаграмм,

до более сложных вариантов, таких как диаграммы рассеяния, линии тренда, диаграммы Ганта, водопад

диаграммы и комбинированные диаграммы (с помощью которых можно объединить несколько типов диаграмм).

Excel также предоставляет рекомендации по наилучшему визуальному представлению для вашего набора данных.

Чтобы сделать диаграммы более презентабельными, можно добавить заголовок диаграммы, изменить цвета элементов,

и добавить метки к данным.

Google Sheets также предлагает аналогичные типы диаграмм для визуализации, хотя Excel имеет

больше встроенных опций на основе формул, чем в Google Sheets.

Как и Excel, Google Sheets может помочь вам выбрать подходящую визуализацию.

Все, что вам нужно сделать, - это выделить данные, которые вы хотите визуализировать, и нажать кнопку диаграммы.

вы получите список предложенных диаграмм, наиболее подходящих для ваших данных.

Диаграммы и отчеты автоматически обновляются как в Excel, так и в Google Sheets, по мере изменения

при изменении базовых данных.

Google Sheets предпочтительнее Excel, когда требуется совместная работа нескольких пользователей.

Jupyter Notebook - это веб-приложение с открытым исходным кодом, которое предоставляет отличный способ для изучения данных

и создавать визуализации.

Чтобы использовать Jupyter Notebook, не обязательно быть экспертом по Python.

Python предоставляет множество библиотек, которые используются для визуализации данных.

Давайте рассмотрим несколько таких библиотек.

Matplotlib - это широко используемая библиотека визуализации данных Python.

Она предоставляет различные виды двухмерных и трехмерных графиков и гибкость для создания графиков несколькими

различными способами.

Используя Matplotlib, вы можете создавать высококачественные интерактивные графики и диаграммы с помощью всего нескольких

строк кода.

Он имеет большую поддержку сообщества и кросс-платформенную поддержку, так как является инструментом с открытым исходным кодом.

Bokeh предоставляет интерактивные графики и диаграммы и известен тем, что обеспечивает высокопроизводительную

интерактивности при работе с большими или потоковыми наборами данных.

Bokeh предлагает гибкость в применении взаимодействия, макетов и различных вариантов стилизации для

визуализации.

Он также может преобразовывать визуализации, написанные в некоторых других библиотеках Python, таких как

Matplotlib, Seaborn и Ggplot.

Dash - это Python-фреймворк для создания интерактивных веб-визуализаций.

Используя Dash, вы можете создавать высокоинтерактивные веб-приложения с помощью кода Python.

Знание HTML и javascript полезно, но не является обязательным условием.

Dash легко поддерживается, является кроссплатформенным и мобильным.

Используя R-Studio, вы можете создавать базовые визуализации, такие как гистограммы, столбчатые диаграммы, линейные диаграммы,

квадратные диаграммы и диаграммы рассеяния; и расширенные визуализации, такие как тепловые карты, мозаичные карты,

3D-графики и коррелограммы.

Shiny - это пакет R, который помогает создавать интерактивные веб-приложения, которые можно размещать как отдельные приложения

на веб-странице.

Эти веб-приложения легко отображают объекты R, такие как графики и таблицы, и могут быть сделаны

доступ к ним может получить любой желающий.

С помощью Shiny можно также создавать приборные панели.

Простота работы с Shiny способствовала его популяризации среди специалистов по работе с данными.

IBM Cognos Analytics - это комплексное аналитическое решение.

Некоторые из функций визуализации, предоставляемых Cognos, включают: Импорт пользовательских визуализаций;

Функция прогнозирования, которая обеспечивает моделирование данных временных рядов и прогнозы на основе данных

представленных в соответствующих визуализациях; Рекомендации для визуализаций на основе

ваших данных; Условное форматирование, которое позволяет вам видеть распределение ваших данных и

выделение исключительных точек данных, например, выделение высоких и низких цифр продаж при превышении

Cognos известен своими превосходными визуализациями и наложением данных на физический мир с помощью геопространства.

на физический мир с использованием геопространственных возможностей.

Tableau - это компания-разработчик программного обеспечения, которая производит интерактивные продукты для визуализации данных.

Используя продукты Tableau, вы можете создавать интерактивные графики и диаграммы в виде приборных панелей

и рабочих таблиц с помощью жестов перетаскивания.

Tableau также предлагает возможность публикации результатов в виде историй.

В Tableau можно импортировать сценарии на языках R и Python и использовать преимущества его функций визуализации.

которые намного превосходят возможности других языков.

Возможности визуализации Tableau просты и интуитивно понятны в использовании.

Tableau совместим с файлами excel, текстовыми файлами, реляционными базами данных и облачными базами данных.

источниками, такими как Google Analytics и Amazon Redshift.

Power BI - это облачный сервис бизнес-аналитики от Microsoft, который позволяет вам

создавать отчеты и информационные панели.

Это мощный и гибкий инструмент, известный своей скоростью и эффективностью, а также простым в использовании интерфейсом.

интерфейс с функцией перетаскивания.

Power BI совместим с различными источниками, включая Excel, SQL Server и облачные хранилища данных.

хранилищами данных, что делает его отличным выбором для специалистов по работе с данными.

Power BI обеспечивает возможность совместной работы и обмена настраиваемыми информационными панелями и интерактивными

отчетами в безопасном режиме, даже на мобильных устройствах.

Приборная панель Power BI состоит из множества визуализаций на одной странице, которые помогают вам рассказать свою историю.

Эти визуализации, называемые плитками, прикрепляются к приборной панели.

Приборная панель является интерактивной, что означает, что изменение одной плитки влияет на другие.

Решая, какие инструменты использовать, необходимо учитывать простоту использования и цель

визуализации.

Что касается доступных инструментов и возможностей визуализации, которые они предлагают

-если вы можете визуализировать это, вы можете создать это.

ИТОГИ

В этом уроке вы узнали следующую информацию:

Данные имеют ценность благодаря историям, которые они рассказывают. Для того чтобы эффективно донести свои выводы, вам необходимо:

Убедиться, что ваша аудитория может доверять вам, понимать вас и относиться к вашим выводам и идеям.

Установить достоверность ваших выводов.

Представить данные в структурированном виде.

Поддержать свое сообщение сильными визуализациями, чтобы сообщение было ясным и четким и побуждало аудиторию к действию.

Визуализация данных - это дисциплина передачи информации с помощью визуальных элементов, таких как графики, диаграммы и карты.

Цель визуализации данных - сделать информацию легкой для восприятия, интерпретации и запоминания.

Чтобы визуализация данных принесла пользу, необходимо:

Подумать о ключевых выводах для вашей аудитории.

Предвидеть их информационные потребности и вопросы, а затем спланировать визуализацию, которая четко и эффектно донесет ваше сообщение.

Существует несколько типов графиков и диаграмм, позволяющих представить любые данные, например, столбчатые диаграммы, диаграммы в столбцах, круговые диаграммы и линейные диаграммы.

Вы также можете использовать визуализацию данных для построения информационных панелей. Приборные панели организуют и отображают отчеты и визуализации, полученные из нескольких источников данных, в едином графическом интерфейсе. Они просты для восприятия и позволяют создавать отчеты на ходу.

Решая, какие инструменты использовать для визуализации данных, необходимо учитывать простоту использования и цель визуализации. Некоторые из популярных инструментов включают электронные таблицы, Jupyter Notebook, библиотеки Python, R-Studio и R-Shiny, IBM Cognos Analytics, Tableau и Power BI.

МОДУЛЬ 9

онца

Вакансии аналитика данных существуют в промышленности, правительстве и

академических кругах. В каждой отрасли, будь то банковское дело и финансы, страхование,

здравоохранение, розничная торговля или информационные технологии, есть место

для квалифицированных аналитиков данных. Эти должности востребованы как в

крупном бизнесе, так и в начинающих компаниях.

По данным Forbes, мировой рынок аналитики больших данных

рынок, объем которого в 2018 году составил 37,34 миллиарда долларов США, в дальнейшем

как ожидается, будет расти со среднегодовым темпом роста

12,3% с 2019 по 2027 год и достигнет 105,08 миллиарда долларов США

долларов США к 2027 году. В настоящее время спрос на

квалифицированных аналитиков данных значительно превышает предложение, что

Это означает, что компании готовы платить за наем

квалифицированных аналитиков данных.

Существует большое разнообразие должностных обязанностей для аналитиков данных.

аналитиков данных. Карьерный путь открыт для вас, мы

в целом классифицируем их на аналитиков данных, специалистов

роли и специалиста по доменам

роли. Роли специалистов по анализу данных предназначены для аналитиков данных, которые

хотят оставаться сфокусированными и расти в технических и функциональных

аспектах своей роли. На этом пути. Вы можете начать свою

карьеру в качестве помощника или младшего аналитика данных и пройти путь

через аналитика, старшего аналитика, ведущего аналитика и

должности главного аналитика.

Границы между этими ролями, годы опыта

которые дают право на переход на следующий уровень, и характер опыта, который необходимо приобрести для продвижения вверх.

опыт, который необходимо приобрести для продвижения вверх, может варьироваться в зависимости от

отрасли, размера организации и того, насколько велика ваша команда.

команда. В небольших командах, например, вы можете приобретать

опыт во всех аспектах анализа данных, начиная со сбора данных и заканчивая их визуализацией и представлением.

от сбора данных до визуализации и представления результатов заинтересованным сторонам.

заинтересованных сторон, и это может произойти за короткий промежуток времени.

В больших командах и организациях роли могут

обычно раздваиваются в зависимости от вида деятельности, что означает, что вы

вы можете получить опыт в одной конкретной фазе

процесса, прежде чем перейти к следующему. Это помогает вам оттачивать свои

навыки в одной части процесса, прежде чем вы перейдете к следующей.

следующую. На пути от младшего специалиста по анализу данных до ведущего

или главного аналитика данных, вы будете постоянно совершенствовать

свои технические, статистические и аналитические навыки от

от базового уровня до уровня эксперта. Вы будете демонстрировать

свою способность работать с широким набором инструментов и

платформами. Различные аспекты процесса анализа данных и

разнообразные случаи использования с точки зрения технических навыков, вы

вы можете начать со знания только одного инструмента для составления запросов и языка программирования

язык. Какой-либо один тип хранилища данных или ограниченный набор

инструментов визуализации. По мере накопления опыта от вас

ожидается, что вы научитесь и продемонстрируете свою способность работать

со все большим количеством инструментов, языков, данных, репозиториев

и новейшими технологиями, ваши коммуникативные навыки,

навыки презентации, навыки работы с заинтересованными сторонами и навыки управления проектами

навыки управления проектами - все это необходимо оттачивать и поднимать на более высокий уровень

постепенно. Будучи ведущим или основным аналитиком, вы также можете

нести ответственность за создание процессов в вашей команде, разработку

рекомендации по программному обеспечению и инструментам. Команда должна работать над

повышением квалификации команды и расширением ее состава, чтобы включить

больше специалистов. В некоторых организациях эти

обязанности могут быть возложены на менеджера уровня

человек, который поднялся по карьерной лестнице и руководит командой специалистов по анализу данных.

аналитиков. Доменные специалисты, также известные как функциональные

аналитики, - это аналитики, которым требуется специализация в какой-либо

конкретной области и считаются авторитетами в своей области.

например, в области здравоохранения, продаж, финансов, социальных сетей или цифрового

маркетинг. Они могут быть не самыми технически подкованными людьми.

Эти роли носят такие названия, как наш аналитик, аналитик по маркетингу,

аналитик по продажам, аналитик по здравоохранению или аналитик по социальным сетям.

Далее следуют должности, связанные с аналитикой. К ним относятся

такие роли, как менеджеры проектов, менеджеры по маркетингу и менеджеры по персоналу.

менеджеры. Это рабочие места, где навыки аналитики ведут к повышению

эффективности и результативности. Достаточно большое количество данных

аналитиков - это вакансии, связанные с аналитикой. Поскольку все больше и больше

все больше организаций полагаются на данные для принятия решений.

Как у аналитика данных у вас также есть возможность изучать и

изучать новые навыки, чтобы получить доступ к другим профессиям, связанным с данными

Такие профессии, как инженерия данных или наука о данных. Для

Например, если вы начинаете работать младшим аналитиком данных и вам

вам очень нравится работать с озерами данных и хранилищами больших данных,

вы можете приобрести дальнейший опыт работы с этими технологиями

и развиться в карьере до инженера по большим данным. Если

вас больше привлекает деловая сторона вопроса, вы можете

также изучить навыки, необходимые для создания бизнеса.

Боковой переход в бизнес-аналитику или бизнес-аналитику.

аналитику.

Хотя карьерный ландшафт аналитика данных очень обширен, хорошо то.

то, что у вас есть множество доступных ресурсов.

чтобы помочь вам стать успешным на вашем пути в качестве

аналитика данных, все, что вам нужно сделать, это ухватиться за те возможности, которые вы

или те, которые открываются перед вами, и

и учиться по ходу дела.

Точки зрения

онца

В этом видео мы послушаем, как

специалистов по работе с данными о том.

как они пришли в эту профессию.

Моя нынешняя роль специалиста по данным не существовала

существовала до того, как я занял эту должность.

Я понял, что в нашей компании существует

в нашей компании существует потребность в предоставлении

данных более быстрым и эффективным способом,

чем обращение в отдел ИБ, который

провести встречу для обсуждения,

выработать требования, а затем

у них был бы конечный продукт, который

люди не были удовлетворены.

Но вам приходилось идти в конец очереди

и проходить весь процесс заново,

чтобы получить то, что вы искали.

Удовлетворяя потребность компании

предоставлять отчеты за две недели,

я создал базу данных компании.

которая имеет доступ к большей информации.
У нас есть аналитики, которые теперь могут
удовлетворить эту неудовлетворенную потребность компании.
Я попал на должность специалиста по данным случайно.
На самом деле я работал над своей докторской диссертацией по
экономики в Университете Иллинойса, Урбана-Шампейн,
когда мой коллега предложил
что степень магистра в области
по статистике также будет отличным дополнением.
Так я попал в
программу по статистике в Иллинойсе.
Но как только я начал заниматься этим,
я был очень увлечен и
и, так сказать, пути назад уже не было.
Другими словами, моя первоначальная цель
стать экономистом на самом деле
превратилась в карьеру, наполненную данными, моделированием,
аналитикой, сбором информации, коммуникацией,
визуализацией и, конечно же,
в основе всего этого лежит решение проблем на основе данных.
Я попал на должность аналитика данных в
компании, занимающейся финансовыми данными, фактически случайно.
В то время моя компания начала нанимать
аналитика данных по акциям в [неслышно], Китай,
и мне очень повезло присоединиться к команде,

потому что они искали кого-то, кто
обладает навыками финансового анализа,
которые я могу привнести в работу.
После этого моя команда начала нанимать людей,
с техническими навыками, такими как Python, R и Sickle.
У меня всегда была любовь к цифрам.
Одна из вещей, которая происходит, это то, что когда
ты так много работаешь с цифрами,
они начинают рассказывать историю,
и способность
смотреть на эти цифры и рассказывать
эту историю - вот что меня привлекает.
У меня всегда был такой уровень знаний о цифрах,
или просто меня всегда привлекала
анализу данных и будь то электронные таблицы Excel,
или QuickBooks,
или любые наборы данных, которые
могут помочь получить информацию, которую мы ищем,
особенно в финансовой индустрии, где мы
мы смотрим на прибыль и убытки,
и баланс, и что происходит
когда одна компания покупает другую.
Мы всегда смотрим на эти данные, чтобы поговорить с ними,
и рассказать об истории компании и ее будущем.
Я получил свою нынешнюю должность

data scientist сразу после окончания программы,

это была магистратура по науке о данных.

До поступления в магистратуру

я работала и аналитиком данных, и менеджером по аналитике.

до конца

В этом видео мы послушаем, как профессионалы в области данных рассказывают о том.

о том, что работодатели ищут в аналитике данных.

Работодатели ищут честных аналитиков данных.

В процессе найма я спрошу,

если бы вам пришлось выбирать только одно,

что бы вы предпочли: уложиться в срок или получить правильный ответ?

Я всегда ищу человека, который бы ответил,

Я хочу быть уверенным, что информация верна.

Пропуск сроков не так вреден, как

когда компания принимает

многомиллионное решение на основе неверной информации,

или кто-то потеряет работу из-за того, что

или кто-то потерял работу из-за того, что информация не была получена или была представлена неверно.

Гораздо важнее быть честным.

Я думаю, что главное, что ищут работодатели в

аналитиков данных - это человек, который умеет ясно излагать свои мысли.

Если вы делаете самый блестящий анализ в мире,

но не сможете донести его до внешних заинтересованных сторон,

тогда это ничего не стоит.

Я думаю, что этот навык очень востребован.

Я думаю, что еще одна вещь, на которую, очевидно, обращают внимание компании

когда они ищут аналитика данных.

это свободное владение цифрами,

способность понимать сложный анализ,

способность понимать тесты AB

и о чем говорят результаты AB-тестов,

и последствия этих результатов.

Я также думаю, что все чаще

работодатели ищут

аналитиков данных с действительно сильными навыками SQL.

Еще одна вещь, которую работодатели ищут в

аналитиков данных - это менталитет роста

и готовность к обучению,

потому что отрасль меняется очень быстро.

Я думаю, что они ищут навыки программирования,

включая Python, R, SQL.

В то же время, они ищут и личностные качества.

Ориентированы ли вы на детали,

нравится ли вам работать с данными,

и умеете ли вы решать проблемы, и так далее, и тому подобное.

Как работодатель, я постоянно нанимаю людей.

Что я ищу?

Мы ищем людей, которые

ориентированных на детали и в некоторой степени переборчивых.

Они не просто хотят делать то, что перед ними,

они хотят идти дальше.

Мы ищем людей, которые имеют более высокие устремления,

а также способных мыслить нестандартно.

Если я говорю: "Сделай ABC",

они не просто сделают это,

они сделают это плюс

[неслышно] и предложат мне несколько альтернатив.

Люди, способные устранять неполадки.

Если что-то пойдет не так, они

не просто остановятся и скажут,

Боже мой, мне нужно поговорить с начальником.

Они скажут: вот проблема,

вот мои мысли.

Вот два возможных решения

как вы можете решить это так, чтобы

работа и компания

могли продолжать двигаться вперед. Это то, что вам нужно.

Не просто ориентированный на детали и не просто хорошо разбирающийся в цифрах.

Вы также должны быть человеком, который

может мыслить нестандартно,

и уметь решать проблемы и устранять неполадки.

Это то, что работодатели будут

сейчас больше, чем когда-либо.

Они ищут способность разбираться в данных,
а под знанием данных мы понимаем несколько вещей.
Уметь работать с данными в различных форматах,
уметь думать о них.
Под этим мы подразумеваем,
знать, какие данные вам нужны
для решения поставленных задач.
Умение работать с данными очень важно.
Решение проблем - еще один очень важный навык.
То есть, если есть
проблема, представленная аналитику данных,
он должен быть в состоянии знать, как решить
эту проблему, используя данные
в каком бы формате они ни были представлены,
уметь анализировать их и
представить выводы, которые позволят решить проблему.
Они также должны быть очень динамичными в этом,
если им будет представлен
внезапно совершенно другой набор данных,
который выглядит совсем не так, как раньше,
они должны уметь адаптироваться к этим изменениям.
Вот почему качество быть
динамичность и адаптивность также важны.
Они также должны быть способны
быстро овладевать техническими навыками.

Под этим мы подразумеваем, что если

один SQL DIAdem используется в одной среде,

они должны быть способны

работать в другой парадигме.

Если в одном месте используется

RStudio, но они знают Python,

они должны быть способны быстро освоить

RStudio быстро, и все такое.

Способность к быстрому обучению, динамичность,

и знание данных - вот несколько вещей, которые

работодатели ищут в хорошем аналитике данных.

ИТОГИ

В этом уроке вы узнали следующую информацию:

Роли аналитиков данных востребованы в любой отрасли, будь то банковское дело и финансы, страхование, здравоохранение, розничная торговля или информационные технологии.

В настоящее время спрос на квалифицированных аналитиков данных значительно превышает предложение, что означает, что компании готовы платить высокую цену за наем квалифицированных аналитиков данных.

Должности аналитиков данных можно классифицировать следующим образом:

Роли специалистов по анализу данных - на этом пути вы начинаете с должности младшего аналитика данных и продвигаетесь до уровня главного аналитика, постоянно совершенствуя свои технические, статистические и аналитические навыки от базового до экспертного уровня.

Роли специалистов в конкретной области - эти роли подойдут вам, если вы получили специализацию в конкретной области и хотите работать, чтобы стать авторитетом в своей области.

Рабочие места с поддержкой аналитики - эти роли включают в себя рабочие места, где наличие аналитических навыков может повысить вашу производительность и выделить вас среди коллег.

Другие профессии, связанные с данными - В современной экосистеме данных существует несколько других ролей, таких как инженер по данным, инженер по большим данным, специалист по данным, бизнес-аналитик или аналитик бизнес-аналитики. Если вы повысите свою квалификацию в соответствии с требуемыми навыками, вы сможете перейти на эти роли.

Существует несколько путей, которые вы можете рассмотреть для того, чтобы получить профессию аналитика данных. К ним относятся:

Получение академической степени в области аналитики данных или таких дисциплин, как статистика и информатика.

Онлайновые специализации по нескольким курсам, предлагаемые такими учебными платформами, как Coursera, edX и Udacity.

Переход к анализу данных в середине карьеры путем повышения квалификации. Например, если у вас есть техническое образование, вы можете сосредоточиться на развитии технических навыков, характерных для Анализа данных. Если у вас нет технического образования, вы можете запланировать обучение некоторым базовым технологиям, а затем работать, начиная с должности начального уровня.

Практическое задание

Одной из операционных задач розничной торговли является поддержание оптимального предложения товаров при сокращении неиспользуемых запасов. Сегодня все больше розничных компаний используют аналитические методы для отслеживания данных о продажах и товарных запасах и составления прогнозов на основе исторических данных.

Представьте, что вы - аналитик данных в отделе планирования ведущей розничной компании в США. Ваша задача - спрогнозировать спрос на три ключевые категории товаров на текущий год, чтобы иметь достаточные запасы для удовлетворения круглогодичного спроса.

Прежде чем вы сможете спрогнозировать уровень запасов для этих категорий товаров, вам необходимо изучить историю продаж этих категорий товаров. Вам необходимо проанализировать тенденции и закономерности, скрытые в исторических данных, и понять некоторые факторы, которые определяют эти тенденции.

Вот примерный набор данных, в котором собраны данные о продажах этих трех категорий товаров за 2018 и 2019 годы. Набор данных также фиксирует, проводила ли розничная компания целевые маркетинговые кампании в определенном квартале, а также периоды, в течение которых на данную категорию товаров действовали скидки. В реальном сценарии набор данных отражал бы гораздо больше деталей, но для наших целей мы представляем более упрощенный набор данных.

Month-Year of Purchase	Product Category	Units Sold	Discount	Targeted Campaign	Customer Ratings	Defects Reported
Jan-Mar 2018	Designer Clothes	1275	0%	N	9.6	12
Jan-Mar 2018	Fitness Gadgets	4250	15%	N	9.5	12
Jan-Mar 2018	Travel Accessories	1670	0%	N	9.3	8
Apr-Jun 2018	Designer Clothes	1825	0%	N	8.9	23
Apr-Jun 2018	Fitness Gadgets	3760	0%	N	7.7	32
Apr-Jun 2018	Travel Accessories	1720	0%	N	9.1	7
Jul-Sep 2018	Designer Clothes	3150	10%	Y	9.3	15
Jul-Sep 2018	Fitness Gadgets	1330	0%	N	8.5	8
Jul-Sep 2018	Travel Accessories	3550	0%	Y	9.1	12
Oct-Dec 2018	Designer Clothes	4715	20%	Y	7.1	48
Oct-Dec 2018	Fitness Gadgets	6450	20%	Y	8.7	22
Oct-Dec 2018	Fitness Gadgets	6450	20%	Y	8.7	22
Oct-Dec 2018	Travel Accessories	5430	20%	N	9.2	9
Jan-Mar 2019	Designer Clothes	1375	0%	N	8.6	6
Jan-Mar 2019	Fitness Gadgets	1765	0%	N	8.9	5
Jan-Mar 2019	Travel Accessories	1475	0%	N	7.9	23
Apr-Jun 2019	Designer Clothes	2175	0%	Y	8.8	8

Apr-Jun 2019	Fitness Gadgets	1925	0%	N	8.6	8
Apr-Jun 2019	Travel Accessories	1215	0%	N	8.2	6
Apr-Jun 2018	Flat Screen	4750	0%	Y	7.8	34
Jul-Sep 2019	Designer Clothes	3100	15%	N	9.2	14
Jul-Sep 2019	Fitness Gadgets	2530	0%	N	8.7	12
Jul-Sep 2019	Travel Accessories	3275	0%	Y	8.7	13
Oct-Dec 2019	Designer Clothes	6425	25%	Y	9.3	12
Oct-Dec 2019	Fitness Gadgets	7125	20%	Y	9.6	16
Oct-Dec 2019	Travel Accessories	6510	30%	Y	9.3	8

Описательные методы анализа, то есть методы, которые помогают понять, что произошло, включают выявление тенденций, закономерностей и корреляций в данных. Некоторые общие события, на которые вам, возможно, придется обратить внимание в этом наборе данных, включают:

Квартальные показатели продаж товарной категории за несколько лет.

Колебания спроса в праздничные дни.

Новый год (январь), День святого Валентина (февраль), День матери (май), День отца (июнь), День независимости (июль), День благодарения (ноябрь) и Рождество (декабрь).

Изменение количества продаж продукции в течение нескольких недель после проведения целевой маркетинговой кампании.

Изменение количества продаж продукта, когда продукт доступен по сниженной цене.

Корреляция между увеличением или уменьшением продаж одного продукта, что приводит к соответствующему увеличению или уменьшению продаж другого продукта.

Взаимосвязь между оценками покупателей, накопленными за продукт в одном квартале, и их влиянием на количество продаж в следующем квартале.

Взаимосвязь между количеством жалоб, полученных на бракованную продукцию в одном квартале, и их влиянием на количество продаж в следующем квартале.

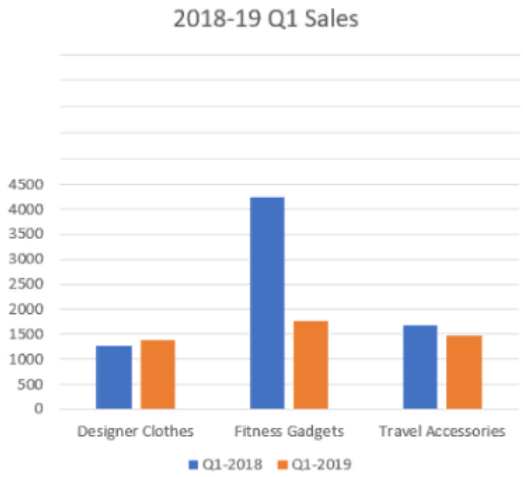
Прежде чем анализировать данные на предмет выявления закономерностей и аномалий, необходимо:

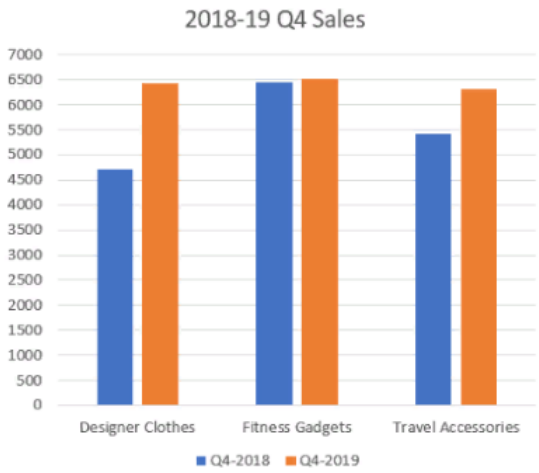
Определить и собрать все точки данных, которые могут иметь отношение к вашему сценарию использования.

Например, категория товара, месяц и год продажи, скидки на товар, целевые маркетинговые кампании, рейтинги товара и сообщения о дефектах.

Очистите данные.

Необходимо выявить и устранить проблемы в данных, которые могут привести к ложным или неполным выводам, например, недостающие данные, избыточные данные и неверные данные.





Наконец, когда вы получите результаты, вы создадите соответствующие визуализации, которые донесут ваши выводы до аудитории. На приведенных ниже графиках показаны визуализации, которые можно использовать для выявления скрытых тенденций в данных.

инструкции

Вы сделали первый шаг к тому, чтобы стать аналитиком данных. Этот курс познакомил вас с компонентами современной экосистемы данных и ролью анализа данных в этой экосистеме. Вы также узнали о знаниях, навыках, инструментах, обязанностях и карьерных перспективах аналитика данных. Прежде чем вы погрузитесь в мир анализа данных, уделите время оценке своего понимания видеоматериалов и материалов, которые вы только что прослушали. Обратитесь к статье "Использование аналитики данных для прогнозирования и планирования запасов" и попробуйте ответить на эти открытые вопросы.

Критерии оценки

Итоговое задание составляет 10% от вашей итоговой оценки.

Вы будете оценены за выполнение следующих 7 заданий:

Задание 1: Представьтесь. (1 pt)

Задание 2: Кратко объясните, почему вы хотите изучать аналитику данных? (1 pt)

Задание 3: Перечислите как минимум 2 точки данных, необходимых для анализа тенденций и закономерностей, которые помогут вам спрогнозировать запасы продукции на текущий год. (2 pts)

Задание 4: Обратитесь к таблице данных в прочитанном тексте и укажите 2 ошибки/проблемы, которые могут повлиять на точность ваших выводов. (2 балла)

Задание 5: Обратитесь к таблице данных из прочитанного и определите 2 корреляции, которые вы заметили в наборе данных. (2 балла)

Задание 6: Кратко объясните 2 наблюдения или идеи, которые вы можете извлечь из представленной визуализации. Вы также можете обратиться к набору данных, представленному в тексте. (2 балла)

Задача 7: "Данные имеют ценность благодаря историям, которые они рассказывают. Ваша аудитория должна быть способна доверять вам, понимать вас и относиться к вашим выводам и идеям". Обратитесь к видеоролику Модуля 8 "Обзор общения и обмена результатами анализа данных" и перечислите два способа, с помощью которых вы можете помочь своей аудитории доверять вам, понимать вас и относиться к вам. (2 pt)