Технологии обработки больших данных

«38.04.05 – Бизнес-информатика направленность интеллектуальное управление цифровым предприятием»

http//vikchas.ru

https://www.famous-scientists.ru/3653/

Лекция 1 «Общие вопросы работы с данными»

Часовских Виктор Петрович
д-р техн. наук, профессор кафедры ШИиКМ

ФГБОУ ВО «Уральский государственный экономический университет»

Екатеринбург 2022

Данные - определения и интерпретация

Согласно «Информационно-коммуникационные технологии в образовании» ГОСТ Р 52653-2006 , **данные** – представление информации в формализованном виде, пригодном для передачи, интерпретации и обработки.

Согласно «Система стандартов по информации, библиотечному и издательскому делу» ГОСТ 7.0-99 , **данные** – информация, обработанная и представленная в формализованном виде для дальнейшей обработки.

Словарь Ушакова

Данные - сведения, обстоятельства, служащие для какого-нибудь вывода, решения.

Современный экономический словарь. 1999

Данные

1) факты и характеризующие их числовые, количественные показатели: имена, даты событий сведения об экономических процессах, местах действия;

2) сведения, обработанные Специальным образом для принятия решений, информация.

Словарь экономических терминов

Данные

1) факты, не связанные друг с другом: имена, даты, числа;

2) сведения, обработанные специальным образом для принятия решений, информация.

Краткий словарь по вычислительной технике, информатике и метрологии

Данные

сведения, полученные путем измерения, наблюдения, логических или арифметических операций, представленные в форме, пригодной для постоянного хранения, обработки и передачи.

Полный словарь терминов и понятий мобильной связи

Данные

информация, представленная в формализованном виде, пригодном для автоматизированной обработки.

Словарь Ожегова

Данные

1. Сведения, необходимые для какого-то вывода, решения.

2. Свойства, способности, качества как условия или основания для чего-то.

Словарь Ефремовой

Данные

Сведения, факты, характеризующие кого-л., что-л., необходимые для каких-л. выводов, решений.

Свойства, способности, качества как условия или основания, необходимые для чего-л.

Энциклопедия Брокгауза и Ефрона

Данные

— В вопросах математики Д. суть величины, значения которых известны или предполагаются известными; зная их, требуется в рассматриваемом вопросе определить искомые неизвестные величины.

Д. (Δεδόμενα) есть заглавие одного из сочинений Эвклида, составляющего продолжение его "Элементов".

Кембриджский словарь

Данные

Это информация, особенно факты и числа, собранные для последующего использования при принятии решений. Данные – это информация в электронной форме, пригодная для хранения и использования компьютером.

Философия рассматривает преобразование сведений в данные, данных в информацию, а информации – в знания.

Кибернетика – Норберт Винера об информации
«… чем более вероятно сообщение, тем меньше оно содержит информации. Клише, например, имеют меньше смысла, чем великолепные стихи».
«Передача информации возможна лишь как передача альтернатив».
«Идеальная информация не содержит в себе ничего поддающегося измерению, следовательно, доступная измерению информация не может быть идеальной».

«…сообщество простирается лишь до того предела, до которого простирается действительная передача информации. Можно дать некоторую меру сообщества, сравнивая число решений, поступающих в группу извне, с числом решений, принимаемых в группе. Мы измеряем тем самым автономию группы. Мера эффективной величины группы – это тот размер, который она должна иметь, чтобы достичь определенной установленной степени автономии».

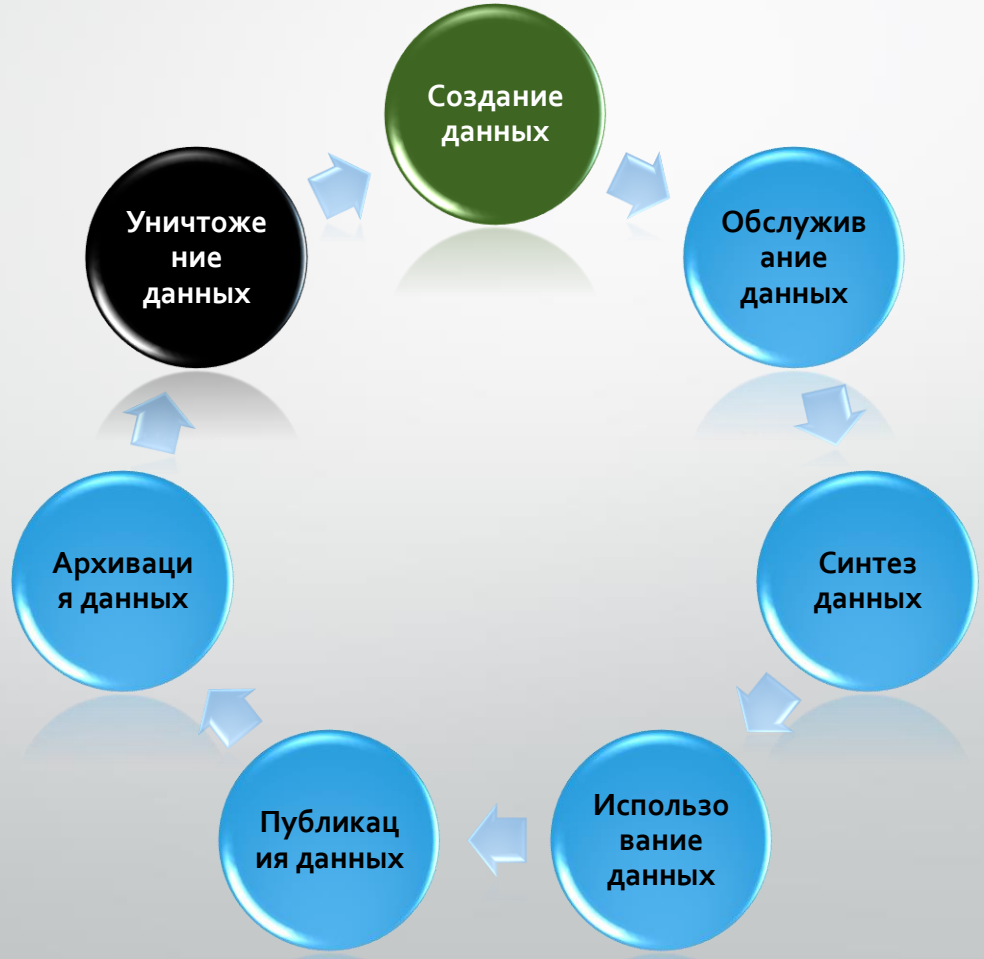
«Информация – это обозначение содержания, полученного из внешнего мира в процессе нашего приспособления к нему и приспосабливания к нему наших чувств».

«… подобно тому как энтропия есть мера дезорганизации, информация есть мера организации».

Информация - это снятая неопределенность (К. Шеннон)

Жизненный цикл данных

Жизненный цикл данных – это последовательность этапов, которую конкретная порция данных проходит от начального этапа создания или получения до момента архивации или удаления. Основные этапы жизненного цикла данных представлены на рисунке.



Создание данных (Data Generation/Data Capture).

На этом этапе данные генерируются или захватываются.
Этот этап обычно еще делят на три типа получения данных:

1. Приобретение данных (Data Acquisition) Получение организацией данных, уже сгенерированных вне предприятия.
2. Запись данных (Data Entry) Создание новых данных оператором или компьютером. Данные имеют ценность для предприятия.
3. Регистрация сигналов (Signal Reception)

Захват данных устройствами. Особенно важно в системах управления, но в последнее время особенно ценно при использовании такого подхода, как Интернет вещей

Обслуживание данных (Data Maintenance)

После того, как данные были созданы, необходимо их хранить и обслуживать. Необходимо осуществлять доставку данных в точку, где их будут использовать или производить над ними манипуляции (например, операции синтеза).

Можно говорить о том, что обслуживание данных – это обработка данных без получения из извлечения из них полезной информации для предприятия.

Зачастую обслуживание данных включает в себя такие действия с данными, как перемещение, интеграция, очистка, обогащение и ETL-процессы (Extract, Transform, Load).

Обслуживание данных обычно подразумевает применение широкого спектра методов из области управления данными (Data Management).

Синтез данных (Data Synthesis)

Это сравнительно недавно появившаяся стадия в жизненном цикле данных.

Используется не во всех моделях жизненных циклах данных.

Синтез данных – это процесс получения дополнительной ценности из данных при помощи использования индуктивной логики и сторонних информационных источников.

Это стадия, на которой с данными работаю аналитики, причем они могут использовать в своей работы методы моделирования рисков, актуарного моделирования, моделирования для принятия инвестиционных решений и др.

На этой стадии используется индуктивная логика, не дедуктивная.

Индуктивная логика требует использования экспертного мнения, т.к. именно компетенции экспертов необходимы для построения моделей скоринга (*оценки кредитоспособности*) и др.

Использование данных (Data Usage)

До сих пор шла речь об использовании данных внутри одного предприятия, которые возможно были подвержены очистке и обогащению на стадии обслуживания данных и использовались совместно с дополнительным третьими источниками данных на стадии синтеза данных.

На стадии использования данных они применяются в качестве полезной информации для задач, которые должны выполняться и управляться на основе данных.

Эти задачи могут быть вне жизненного цикла данных. Тем не менее, данные становятся все более значимой частью бизнес-процессов предприятий. Данные сами могут быть продуктом или услугой (или быть частью продукта или услуги), предлагаемой предприятием.

Использование данных имеет специальные задачи в рамках управления данными (Data Governance).

Одна из задач заключается в законном использовании данных в требуемом виде. Это называется “разрешенное использование данных” (permitted use of data). Могут существовать регулирующие или договорные ограничения на то, как фактически можно использовать данные, а часть роли управления данными (Data Governance) заключается в обеспечении соблюдения этих ограничений.

Публикация данных (Data Publication)

При использовании данных возможна ситуация, когда данные отправляются за пределы предприятия. В этом случае говорят о публикации данных.

Публикация данных – это вынос данных за пределы предприятия.

Примером этого процесса может быть маклер, рассылающий ежемесячные отчеты клиентам. Все данные, которые были разосланы, уже не могут быть отозваны. Если были разосланы данные с неверными значениям, то такие данные не могут быть исправлены, поскольку они уже становятся недоступны для предприятия. Управление данными (Data Governance) может потребоваться,
чтобы помочь принять решение о том, как будут обрабатываться неверные данные, которые были отправлены из предприятия.

Архивация данных (Data Archival)

Данные могут быть использованы как однократно, так и несколько раз.

Но затем рано или поздно жизненный цикл данных начинает подходить к концу.

Первая стадия этого состояния заключается в архивации данных.

Архивация данных – это копирование данных в пассивную среду, в которой они хранится, для тех случаев, когда они понадобятся снова в активной производственной среде, и удаление этих данных из всех активных производственных сред.

Архив данных – это просто место, где хранятся данные, без их обслуживания, использования или публикации. В случае необходимости данные могут быть восстановлены из архива.

Уничтожение данных (Data Purging)

Уничтожение данных – это последовательность операций для выполнения необратимого удаления данных, делающая невозможным как восстановление данных, так и получение остаточной информации (Data Remanence) о них.

Это одна из самых сложно реализуемых процедур управления данными.

Даже с теоретической точки зрения существует команда записи значения в ячейку памяти, но команды стирания значения как таковой нет.

Для уничтожения данных необходимо изготовить высокопроизводительный источник случайных чисел и перезаписать ими весь носитель информации (перезаписи области хранения недостаточно, так как сохраняется информация об исходном количестве данных).

Иначе говоря, при уничтожении данных необходимо не только сделать недоступными от восстановления на физическом уровне сами данные, но и связанную с ними информацию в других наборах данных.

Уничтожение данных регламентируется в ГОСТ Р 50739-95 (первый цикл - запись все "0", второй цикл - запись псевдослучайных чисел), причём классы защищенности данных и протоколы работы устанавливаются на уровне руководящих документов Гостехкомиссии (ФСТЭК, Федеральной службы по техническому и экспортному контролю).

В настоящее время в РФ предусмотрено семь классов защиты информации.

Согласно действующему на настоящему моменту пункту 19.2 приказа ФСТЭК N17 от 11.02.2013 “При выводе из эксплуатации машинных носителей информации, на которых осуществлялись хранение и обработка информации, осуществляется физическое уничтожение этих машинных носителей информации”.

Метаданные

При сборе данных возникают метаданные, содержащие какую-либо информацию о собранных данных.

Например, время создания набора данных, авторство и первоисточник, размер и кодировка данных – все это метаданные. В соответствии с «ГЕОГРАФИЧЕСКИЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ» ГОСТ Р 52438-2005 метаданные – это сведения о данных ((пространственные) метаданные: Данные о пространственных данных.

Примечание - Пространственные метаданные, описывающие набор пространственных данных, в общем случае могут содержать сведения о составе, статусе (актуальности и обновляемости), происхождении, местонахождении, качестве, форматах представления, условиях доступа, приобретения и использования, авторских правах на данные, применяемых системах координат, позиционной точности, масштабах и других характеристиках.).

Структурированные метаданные называют онтологией или схемой метаданных.

Онтология определяет общий словарь для ученых, которым нужно совместно использовать информацию в предметной области. Она включает машинно-интерпретируемые формулировки основных понятий предметной области и отношения между ними”.

Онтологии обычно используются в таких областях, как искусственный интеллект, семантическая паутина, системная инженерия, биомедицинская информатика, библиотечное дело, информационная архитектура и др. Все онтологии нужны для организации информации

Концепция метаданных была распространена на мир систем и включает в себя любые "данные о данных" - имена таблиц, столбцов, программ и тому подобное.

Таким образом, по сути, метаданные-это "данные, которые описывают структуру и порядок использования организацией информации, а также системы, используемые ею для управления этой информацией".

Создать модель метаданных-значит создать "корпоративную модель" самой индустрии информационных технологий.

Использование

Основная цель метаданных - ускорить и обогатить поиск ресурсов. В общем случае поисковые запросы с использованием метаданных могут избавить пользователей от выполнения более сложных операций фильтрации вручную.

Метаданные должны помочь преодолеть семантический разрыв. Сообщая компьютеру, как связаны данные и как эти отношения могут быть оценены автоматически, можно будет обрабатывать еще более сложные операции фильтрации и поиска.

Например, если поисковая система понимает, что "Ван Гог" был "голландским художником", она может ответить на поисковый запрос "Голландские художники" ссылкой на веб-страницу о Винсенте Ван Гоге, хотя точный термин "голландские художники" никогда не встречается на этом сайте; сегодня это невозможно. Этот подход также называется представлением знаний. Особый интерес она представляет для семантической сети и искусственного интеллекта.

Некоторые метаданные предназначены для оптимизации алгоритмов сжатия. Например, если есть метаданные, которые позволяют компьютеру отличать передний план от фона в видео, он может сжимать обе части независимо друг от друга и таким образом достигать более высоких скоростей сжатия.

Некоторые метаданные предназначены для представления переменного содержимого.

Например, если программа просмотра изображений знает наиболее важную область изображения — например, ту, где находится человек, — она может уменьшить изображение до этой области и таким образом показать пользователю наиболее интересные детали на маленьком экране, например на экране мобильного телефона.

Подобный вид метаданных предназначен для того, чтобы слепые люди могли "читать" диаграммы и рисунки, например, преобразуя их для специальных устройств вывода или читая описание с помощью синтеза голоса.

Для автоматизации рабочих процессов можно использовать и другие описательные метаданные.

Например, если инструмент знает содержание и структуру данных, он может автоматически преобразовать их и передать другому инструменту в качестве входных данных. Таким образом, пользователи могут сохранить множество действий копирования и вставки, необходимых при анализе данных с помощью различных инструментов.

Типы метаданных

Метаданные можно отличить по их ...

Содержание. Метаданные могут описывать либо сам ресурс, например имя и размер файла, либо его содержимое, например "Видео показывает мальчика, играющего в футбол".

Изменчивость. Что касается всего ресурса, метаданные могут быть либо неизменяемыми, например, название файла не меняется, независимо от того, какая часть файла рассматривается, либо изменяемыми, например, описания сцен видео меняются.

Логическая функция. Существует три слоя логических функций, лежащих друг на друге: нижний-это субсимволический слой, содержащий сами необработанные данные, на символическом уровне-метаданные, описывающие содержимое необработанных данных, а самый верхний логический слой содержит метаданные, позволяющие логически рассуждать с использованием символического уровня.

Жизненный цикл метаданных

Есть три фазы, которые следует рассматривать независимо друг от друга:

Творение. Даже на ранних этапах планирования и проектирования необходимо отслеживать все созданные метаданные. Неэкономично начинать прикреплять метаданные только после завершения производственного процесса.

Например, если метаданные, созданные цифровой камерой во время записи, сохраняются не сразу, то впоследствии они должны быть восстановлены вручную с большими усилиями. Поэтому необходимо, чтобы различные группы производителей ресурсов сотрудничали, используя совместимые методы и стандарты.

Манипуляция. Метаданные должны адаптироваться при изменении их ресурса. Они должны быть объединены при объединении двух ресурсов. Эти операции редко выполняются современным программным обеспечением, например, программное обеспечение для редактирования изображений обычно не отслеживает метаданные цифровых камер, хранящиеся в формате EXIF.

Разрушение. Это может быть полезно для хранения метаданных даже после того, как их ресурс был уничтожен, например, в истории изменений в текстовом документе или для архивирования удалений файлов из-за управления цифровыми правами. Ни один из современных стандартов метаданных не учитывает эту фазу.

Хранение

Метаданные могут храниться либо внутри, то есть в том же файле, что и данные, либо снаружи, то есть в отдельном файле. Обе возможности имеют свои преимущества и недостатки:

Внутреннее хранилище позволяет передавать метаданные вместе с их данными, поэтому они всегда под рукой и ими легко манипулировать. Этот метод создает высокую избыточность и не позволяет удерживать метаданные вместе.

Внешнее хранилище позволяет объединять метаданные, например, в базу данных, для более эффективного поиска. Избыточность отсутствует, и метаданные могут передаваться одновременно при использовании потоковой передачи. Однако, поскольку большинство форматов используют URL для этой цели, метод того, как метаданные связаны с их данными, должен рассматриваться с осторожностью. Что делать, если метаданные могут быть оценены только при наличии подключения к WWW?

Кроме того, возникает вопрос о формате данных: хранение метаданных в удобочитаемом формате, таком как XML, может быть полезно, поскольку пользователи могут понимать и редактировать их без каких-либо инструментов вообще.

С другой стороны, эти форматы не оптимизированы для емкости памяти, то есть может быть полезно хранить их в двоичном нечеловеческом формате, чтобы ускорить передачу и сэкономить память.

Виды метаданных

В общем случае существует два различных класса метаданных: структурные или управляющие метаданные и направляющие метаданные.

Структурные метаданные используются для описания структуры компьютерных систем, таких как таблицы, столбцы и индексы.

Метаданные руководства используются для того, чтобы помочь людям найти конкретные предметы, и обычно выражаются в виде набора ключевых слов на естественном языке.

Метаданные реляционной базы данных

Каждая система реляционных баз данных имеет свои собственные механизмы хранения метаданных.

Примеры метаданных реляционных баз данных включают:

Таблицы всех таблиц в базе данных, их имена, размеры и количество строк в каждой таблице.

Таблицы столбцов в каждой базе данных, в каких таблицах они используются и тип данных, хранящихся в каждом столбце.

В терминологии базы данных этот набор метаданных называется каталогом.

Стандарт SQL определяет единое средство доступа к каталогу, называемое the INFORMATION_SCHEMA, но не все базы данных реализуют его, даже если они реализуют другие аспекты стандарта SQL.

Большие данные. Системы управления Большими данными

Большие данные – это данные, которые не помещаются в оперативную память компьютера.

По сути это определение обозначает то, что свойство “быть большим” является не самостоятельным свойством данных, а зависит от характеристики системы, применяемой для их обработки.

Например, обычному человеку затруднительно запомнить какая именно температура была в нашем городе каждый день за прошедший месяц. Таким образом, три десятка значений вполне могут быть примером Больших данных. Однако вот человек уверенно сообщает “прошедший месяц был холодным”. Это сообщение несет информацию об обработанных данных: по мнению собеседника, средняя температура за прошедший месяц была ниже, чем обычно в этом месяце за несколько десятков лет.

Другим примером могут быть данные об объектах, которые теоретически несут важную информацию, однако имеющие такой размер, что эти данные практически невозможно не только обработать или сохранить, но даже собрать.

Рассмотрим к примеру набор данных, содержащий координаты и скорости молекул в воздушном столбе над территорией аэропорта. Имеются также метаданные с описанием в какой момент проводилось измерение и что это за молекула. Такой набор данных несет информацию о погодных условиях над аэропортом, включая температуру, давление, влажность, облачность, особые погодные условия – проходящий торнадо или падающий град. С другой стороны, для корректной обработки данные для всех молекул должны быть достаточно полны и репрезентативны для статистической обработки.

В результате такого мысленного эксперимента мы понимаем, что для эффективной работы с большими данными нужна модель данных, позволяющая сформировать методы работы с данными.

Данные могут быть различных типов. Информацию, полученную в результате учёта или измерения каких-либо объектов или параметров, называют мастер-данными (Master Data).

Например, учёт количества, замеры координат и скоростей конкретных молекул – это мастер-данные.

Транзакционные данные – это данные, отображающие результат выполнения каких-либо операций.

Например, данные о взаимодействии молекул между собой, а именно о пересечении границ рассматриваемой области, о траектории конкретной молекулы, об испарении капель дождя – это транзакционные данные. Транзакционные данные описывают взаимодействие объектов друг с другом или с окружающим миром, которые можно получить при помощи обработки мастер-данных

Ретроспективные данные (Historical data) – это данные, снабженные метками времени.

Например, с одной стороны мы можем сохранять данные о координате и векторе скорости каждой молекулы, но если у нас есть набор координат в зависимости от времени, то скорость молекулы становится лишней, она вычисляется исходя из модели, описываемой ньютоновской механикой.

Ссылочные данные (справочники, НСИ, нормативно-ссылочная информация, Reference Data, Lookup Data, Dictionaries) – это базовые неизменяемые данные, заранее известные из внешних источников, такие как нормативы, сокращения, акронимы, словари, стандарты.

Например, удельные веса молекул, зависимость температуры замерзания и кипения от давления, зависимость средней скорости молекул (скорости звука) от температуры

Формат данных.

Структурированные данные имеют заранее определенный формат.

Полуструктурированные или слабоструктурированные данные – это данные, зачастую собранные из различных источников.

Структура данных документирована, но в зависимости от источника данных конкретный формат представления информации может быть разным.

Неструктурированные данные требуют обязательной обработки и последующей валидации перед использованием.

Например, данные о координатах и скоростях молекул, в которых некоторые координаты пропущены или некоторые записи повторяются, являются полуструктурированными.

Нам нужно понять, почему так произошло и перед использованием либо исключить такие данные (что может привести к систематической ошибке), либо, исходя из модели данных, восстановить пропущенные значения.

Данные, в которых координаты измеряются в разных единицах измерения, числа иногда записаны словами, иногда латинскими цифрами, а иногда в виде сканированного изображения почерка лаборанта, являются неструктурированными данными.

Обычно Большие данные описываются при помощи следующих характеристик.

Объем (Volume) – количество сгенерированных и хранящихся данных. Размер данных определяет значимость и потенциал данных, а также то, могут ли они быть рассмотрены как Большие данные.

Разнообразие (Variety) – тип данных. Большие данные могут состоять из текста, изображений, аудио, видео. Большие данные при сопоставлении друг с другом могут дополнять отсутствующие данные.

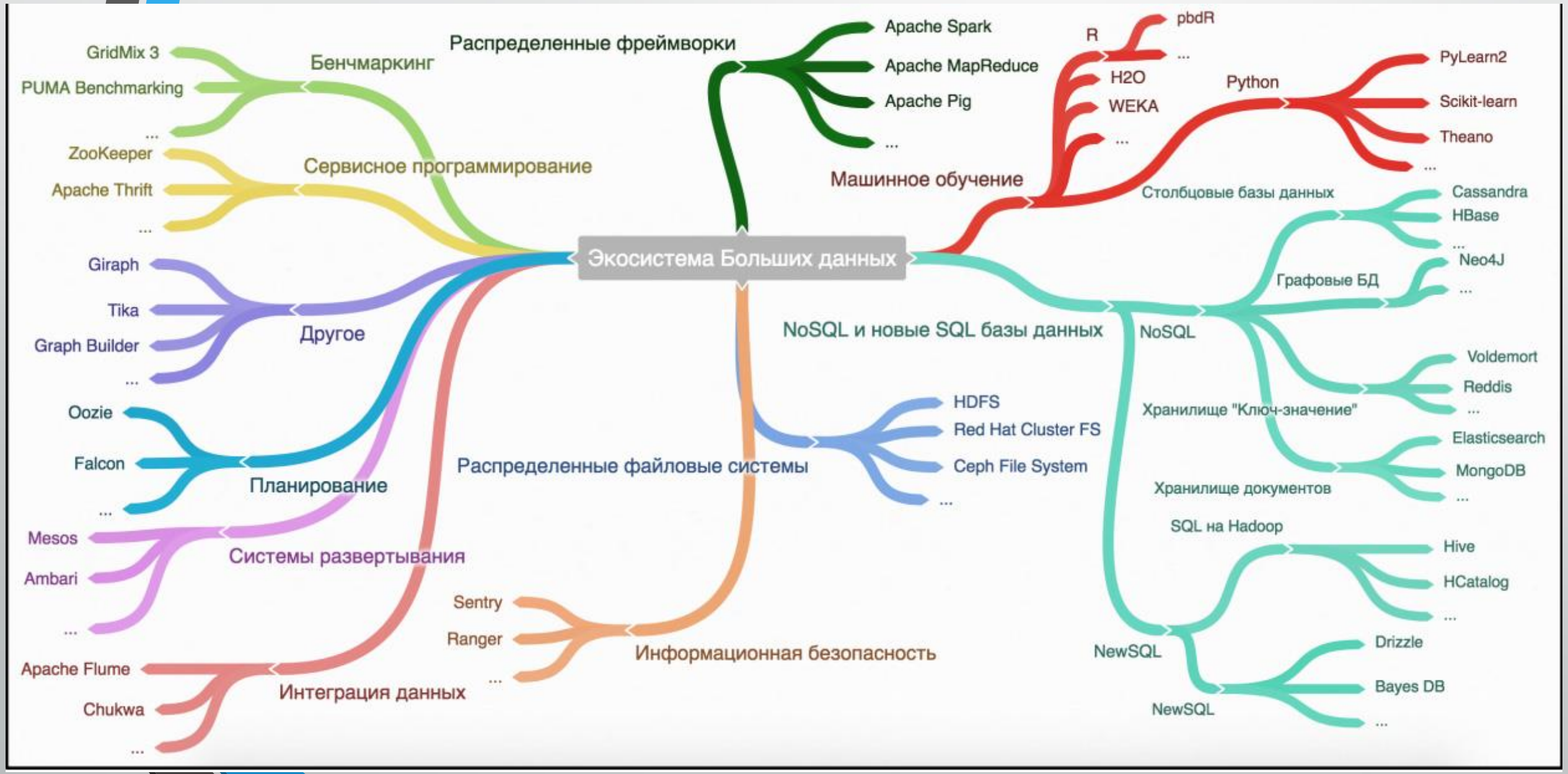
Скорость (Velocity) – скорость. Здесь подразумевается скорость, с которой данные генерируются и обрабатываются. Очень часто Большие данные используются в режиме реального времени.

Изменчивость (Variability) – противоречивость наборов данных может препятствовать их обработке и управлению ими.

Достоверность (Veracity) – качество данных напрямую влияет на точность проведения анализа данных.

Интеллект – карта больших данных

Большие данные могут быть классифицированы в соответствии с несколькими главными компонентами.



Распределенные файловые системы Для хранения и обработки Больших данных созданы распределенные системы хранения данных, в том числе распределенные файловые системы, позволяющие использовать внешнее файловое пространство системы хранения для обработки данных на нодах, входящих в вычислительных кластер. Зачастую удобно использовать распределенные файловые системы, арендуемые как отдельный облачный сервис, например, Azure? Google , Amazon , Yandex.

Распределенные фреймворки Обработка находящихся на распределенных системах хранения данных ведется параллельно на компьютерах, составляющих узлы (nodes) вычислительного кластера. Для организации вычислений разработчики систем обработки используют распределенные фреймворки. Большинство фреймворков доступны по лицензии. Существуют также облачные фреймворки, арендуемые как отдельный облачный сервис.

Бенчмаркинг. Этот класс инструментов был разработан для оптимизации инсталляции Больших данных при помощи использования стандартизированных профилей (Profiling suites).

Бенчмаркинг и оптимизация инфраструктуры Больших данных зачастую не является сферой ответственности дата-ученых, это область ответственности для отдельных профессионалов, специализирующихся на IT инфраструктуре. Использование оптимизированной инфраструктуры может существенно снизить стоимость используемого оборудования.

Серверное программирование

Предположим, что вы сделали приложение для прогнозирования результатов футбольных матчей мирового класса и вы хотите разрешить другим использовать прогнозы, сделанные вашим приложением. Тем не менее, вы не имеете представления об архитектуре или технологии всех, кто стремится использовать ваши прогнозы. Сервисные инструменты позволяют предоставлять приложения на Больших данных другим приложениям в качестве службы. Дата ученым иногда приходится предоставлять свои модели через службы. Наиболее известным примером здесь является REST-сервис; REST означает репрезентативную передачу состояния (REpresentational State Transfer, REST). Она часто используется в качестве обмена данными с веб-сайтами.

Планирование Инструменты планирования позволяют автоматизировать повторяющиеся задачи и запускать задания на основе таких событий, как добавление нового файла в папку. Они похожи на такие инструменты, как CRON в Linux, но специально разработаны для работы в отказоустойчивом кластере. Вы можете использовать их, например, для запуска задачи MapReduce всякий раз, когда в каталоге имеется новый набор данных.

Системы развертывания Настройка инфраструктуры Больших данных – непростая задача, и развертывание новых приложений в кластере Больших данных – это зона ответственности инженеров по Большим данным. Они в значительной степени автоматизируют установку и настройку компонентов Больших данных.

Интеграция данных Допустим, что уже есть распределенная файловая система, и теперь необходимо перенести данные из одного источника в другой. В таких случаях используют фреймворки для интеграции данных, такие как Apache Sqoop и Apache Flume. Этот процесс похож на процесс извлечения, преобразования и загрузки (Extract, Transform and Load, ETL) в традиционном хранилище данных.

Информационная безопасность Средства обеспечения безопасности Больших данных позволяют осуществлять централизованный контроль доступа к данным. Безопасность Больших данных стала самостоятельной дисциплиной, и дата-ученые обычно сталкиваются с ней только как потребители данных. Безопасностью Больших данных занимаются эксперты по информационной безопасности.

Машинное обучение Если у вас есть Большие данные, то было бы неплохо получить из них полезный контент. Это можно сделать при помощи использования методов машинного обучения, статистики и прикладной математики. Появилась возможность писать программы с формулами и алгоритмами, а затем загружать в программы различные данные. На сегодняшний день, когда появилось огромнейшее количество данных, один компьютер уже не в состоянии справиться с задачей их обработки. Некоторые алгоритмы, разработанные в прошлом веке, увы, не смогут справиться с этой задачей, даже если теоретически можно было бы подключить к решению задачи все компьютеры Земли. Это связано с временной сложностью алгоритма.

Одна из самых больших проблем со старыми алгоритмами заключается в том, что они недостаточно масштабируются.

Учитывая объем данных, которые необходимо анализировать сегодня, это становится проблематичным. Для обработки этого объема данных требуются специализированные структуры и библиотеки. Например, в языке Python есть следующие библиотеки: Scikit-learn (библиотека машинного обучения), PyBrain (для работы с нейронными сетями), NLTK (для обработки естественного языка), Pylearn2 (еще одна библиотека машинного обучения), TensorFlow (библиотека глубокого обучения, есть программный интерфейс API для языка Python), Keras (библиотека для работы с нейронными сетями) и другие. Существует также Apache Spark – программный каркас с открытым исходным кодом для реализации распределенной обработки неструктурированных и слабоструктурированных данных.

Базы данных NoSQL и новые SQL базы данных **SQL 2019, Adabas**

Использование реляционных баз данных для обработки Больших данных крайне неэффективно из-за высоких накладных расходов. Традиционно для обработки Больших данных используются базы данных типа “ключ – значение” ADABAS.

База данных вида “ключ – значение”, по сути, представляет собой ассоциативный массив (Hash, Dict), то есть множество, состоящее из пар (Key, Value). В некоторых реализациях на множестве ключей вводится отношение порядка, и мы можем получить значения последовательно по мере возрастания ключа. В других случаях сортировка по ключу неустойчива при одинаковых ключах и при неоднократных выборках можно получить различную последовательность пар.

Большое количество баз данных можно разделить на следующие типы:

- Столбцовые базы данных (Column databases). Данные хранятся в столбцах, что позволяет алгоритмам выполнять гораздо более быстрые запросы.
- Хранилища документов (Document stores) Хранилища документов больше не используют таблицы, но сохраняют каждое наблюдение в документе. Это позволяет использовать гораздо более гибкую схему данных.
- Потоковые данные (Streaming data). Данные собираются, преобразуются и агрегируются не в партиях, а в реальном времени.

- Хранилища для ключей (Key-value stores). Данные не хранятся в таблице; для каждого значения назначается ключ (как рассказано об этом выше).
- SQL на Hadoop – пакетные запросы на Hadoop, использующие фреймворк MapReduce в фоновом режиме.
- Новый SQL (New SQL). Этот тип сочетает масштабируемость баз данных NoSQL с преимуществами реляционных баз данных. Здесь используется интерфейс SQL и реляционная модель данных.
- Графовые базы данных (Graph databases). Это тип баз данных, использующих графовые структуры для семантических запросов с узлами и ребрами и свойствами для представления и хранения данных. Классическим примером этого типа является социальная сеть.