Лекция
по учебной дисциплине
"Технологии обработки больших данных"

Тема № 1 " Предпосылки формирования больших данных "

I. Учебные цели

1. Дать обучающимся сведения об основных архитектурных принципах построения и применения программного обеспечения средств вычислительной техники при обработке больших данных.

II. Воспитательные цели

1. Воспитывать интерес к дисциплине, стремление совершенствовать свои знания и уверенность в приобретенных знаниях.
2. Стимулировать у обучающихся активную познавательную деятельность, способствовать формированию у них творческого мышления.

III. Расчет учебного времени

Содержание и порядок проведения занятия:

Вступительная часть	– 10 мин.
Основная часть	– 75 мин.
Учебные вопросы:	
1. Термины и определения	– 25 мин.
2. Характеристики и формы представления больших данных	– 25 мин.
3. Цели и прикладные задачи обработки больших данных	– 25 мин.
Заключительная часть	– 5 мин.

IV. Литература

1. Вольфсон, Михаил Борисович Анализ данных : [Электронный ресурс] : учеб. пособие / М. Б. Вольфсон ; рец.: Ю. П. Левчук, А. Л. Алимов ; Федер. агентство связи, Федер. гос. образовательное бюдж. учреждение высш. проф. образования "С.-Петерб. гос. ун-т телекоммуникаций им. проф. М. А. Бонч-Бруевича". - СПб. : СПбГУТ, 2015. - 81 с. : ил. - Библиогр.: с. 81. - (в обл.) : 451.36 р.
2. Ын, А. Теоретический минимум по Big Data : [Электронный ресурс] : всё что нужно знать о больших данных / А. Ын. - СПб. : Питер, 2019. - 208 с. - URL: http://ibooks.ru/reading.php?productid=359225. - ISBN 978-5-4461-1040-7 : Б. ц.

V. Учебно-материальное обеспечение

1. Проектор, материал презентации.
2. Доска, стираемый маркер.

Введение

Мы начинаем изучение дисциплины "Технологии обработки больших данных" и в рамках лекции рассмотрим термины и определения предметной области больших данных, их характеристики и формы представления, а также цели и прикладные задачи обработки больших данных.

Учебные вопросы

1. Термины и определения больших данных

Рассмотрим термины и определения больших данных в соответствии с ГОСТ Р ИСО/МЭК 20546-2019 "Информационные технологии. Большие данные. Обзор и словарь.":

информация: сведения о лицах, предметах, фактах, событиях, явлениях и процессах независимо от формы их представления;

информационный процесс: процесс создания, обработки, хранения, защиты от внутренних и внешних угроз, передачи, получения использования уничтожения информации;

данные (data): реинтерпретируемое представление информации в формализованном виде, пригодном для коммуникации, интерпретации или обработки. Данные могут быть обработаны людьми или автоматическими средствами.

большие данные (big data): большие массивы данных – главным образом, по таким характеристикам данных, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа. Большие данные повсеместно используются множеством различных способов, например, в качестве названия технологии масштабирования, используемой для обработки обширных массивов данных.

облачные вычисления (cloud computing): парадигма для обеспечения сетевого доступа к масштабируемому и гибкому пулу совместно используемых физических или виртуальных ресурсов с системой самообслуживания и администрированием по требованию. Примерами таких ресурсов являются серверы, операционные системы, сети, программное обеспечение, приложения и оборудование для хранения.

аналитика данных (data analytics): составное понятие, состоящее из получения, сбора, проверки и обработки данных, включая их количественную оценку, визуализацию и интерпретацию. Аналитика данных используется для понимания объектов, представленных данными, для прогнозирования конкретных ситуаций и для рекомендаций по шагам для достижения целей. Выводы, полученные из аналитики, используются для различных задач, таких как принятие решений, исследования, устойчивое развитие, проектирование, планирование и т. д.

вариативность данных (data variability): изменения в скорости передачи, формате или структуре, семантике или качестве массива данных

разнообразие данных (data variety): диапазон форматов, логических моделей, временных шкал и семантики массива данных. Разнообразие данных относится к нерегулярным или неоднородным структурам данных, их навигации, запросам и типизации данных.

скорость обработки данных (data velocity): скорость потока, с которой данные создаются, передаются, хранятся, анализируются или визуализируются.

достоверность данных (data veracity): полнота и/или точность данных. Под достоверностью данных понимаются пояснительные данные и самоанализ объектов для поддержки принятия решений в режиме реального времени.

изменчивость данных (data volatility): характеристика данных, относящаяся к скорости изменения этих данных с течением времени.

объем данных (data volume): степень количества данных, оказывающая влияние на ресурсы для вычислений и хранения, а также на управление ими в процессе обработки данных. Объем данных становится важным при работе с большими массивами данных, включая их.

распределенная обработка данных (distributed data processing): обработка данных, в которой выполнение операций распределено между узлами компьютерной сети.

распределенная файловая система (distributed file system): система, управляющая файлами и папками в нескольких сетевых системах.

файл (file): именованный набор записей, рассматриваемый как единое целое.

горизонтальное масштабирование (horizontal scaling): формирование единого логического блока путем соединения нескольких аппаратных и программных средств. Примером горизонтального масштабирования является повышение производительности распределенной обработки данных путем добавления узлов в кластере для дополнительных ресурсов.

Горизонтальное масштабирование для увеличения производительности также называется масштабированием вширь (scale-out).

метаданные (metadata): данные о данных или элементах данных, которые могут включать в себя их описания, а также данные о владении данными, путях и правах доступа и об изменчивости данных.

нереляционная база данных (non-relational database): база данных, не следующая реляционной модели. «NoSQL», что обычно переводится как «не SQL» или «не только SQL», является общеупотребительным термином для обозначения баз данных, не соответствующих реляционной модели.

нереляционная модель данных (non-relational model): логическая модель данных, не следующая реляционной модели хранения и обработки данных.

параллельная работа (parallel): относится к процессу, в котором все события происходят в одном и том же интервале времени, и при этом каждое из них обрабатывается отдельной, но схожей функциональной единицей.

частично структурированные данные (partially structured data): данные, имеющие некую структуру. Частично структурированные данные в индустрии часто называют полуструктурированными. Примерами частично структурированных данных являются записи со свободными текстовыми полями в дополнение к более структурированным полям. Такие данные часто представлены в компьютерно-интерпретируемых/разбираемых форматах, таких как XML или JSON.

реляционная алгебра (relational algebra): алгебра для выражения и манипулирования отношениями.

реляционная база данных (relational database): база данных, данные в которой организованы по реляционной модели.

реляционная модель данных (relational model): модель данных, структура которой основана на реляционных отношениях.

распределение-сборка (scatter-gather): вид обработки больших массивов данных, где необходимые вычисления разделяются и распределяются по нескольким узлам в кластере, а общий результат формируется путем объединения результатов от каждого узла. Обработка методом распределения-сборки обычно требует алгоритмического изменения обрабатывающего программного обеспечения. Примером обработки данных методом распределения-сборки является MapReduce.

потоковые данные (streaming data): данные, передаваемые через интерфейс от непрерывно работающего источника.

структурированные данные (structured data): данные, организованные на основе предопределенного (применимого) набора. Предопределенный набор правил, регулирующих основу для структурирования данных, должен быть четко изложен и опубликован. Предопределенная модель данных часто используется для управления структурированием данных.

SQL: язык баз данных, описанный в Международном стандарте ISO/IEC (ИСО/МЭК) 9075. SQL иногда интерпретируется как язык структурированных запросов (Structured Query Language), но это название не используется в серии ISO/IEC (ИСО/МЭК) 9075.

неструктурированные данные (unstructured data): данные (3.1.5), характеризуемые отсутствием какой-либо структуры, кроме структуры на уровне записи или файла. В целом, неструктурированные данные не состоят из элементов данных. Примером неструктурированных данных является свободный текст.

вертикальное масштабирование (vertical scaling): повышение производительности обработки данных за счет улучшения процессоров, памяти, хранилища или связи. Вертикальное масштабирование для увеличения производительности также называется масштабированием ввысь (scale-up).

Под **протоколом** предметной области взаимосвязи открытых систем, в соответствии с ГОСТ Р 52292-2004 "Информационная технология. Электронный обмен информацией. Термины и

определения", понимается набор семантических и синтаксических правил, определяющий поведение объекта на данном уровне при выполнении коммуникационных функций.

Автоматизированная система (АС), это такая система, в которой часть выполняемых ею функций осуществляется без необходимости выполнения каких-либо действий со стороны персонала её эксплуатирующего. Под автоматизированной системой, в соответствии с ГОСТ 34.003-90 "Информационная технология. Комплекс стандартов на автоматизированные системы. Автоматизированные системы. Термины и определения", понимается система, **состоящая из персонала и комплекса средств автоматизации его деятельности**, реализующая информационную технологию выполнения установленных функций. Под комплексом средств автоматизации понимается совокупность всех компонентов автоматизированной системы за исключением персонала.

Под **информационной технологией** понимаются приемы, способы и методы применения средств вычислительной техники (СВТ) при выполнении функций сбора, хранения, обработки, передачи и использования данных.

2. Характеристики и формы представления больших данных

Данные генерируются с беспрецедентной скоростью машинами, людьми и вещами. Индекс Cisco Visual Networking Index (VNI) прогнозирует рост глобального интернет-трафика и тенденции в области широкополосной связи для мобильных и фиксированных сетей. Согласно Cisco VNI, IP-трафик утроится в течение следующих 3 лет. К 2021 году в мире будет более 26 миллиардов сетевых IP-устройств / подключений (по сравнению с 16,3 миллиардами в 2015 году). В глобальном масштабе IP-трафик достигнет 194,4 эксабайта в месяц в 2021 году. Интернет-видео будет составлять 79 процентов глобального Интернет-трафика к 2021 году. Это больше по сравнению с 63 процентами в 2015 году. К 2021 году мир достигнет трех триллионов минут Интернет-видео в месяц? что составляет пять миллионов лет видео в месяц или около миллиона видео минут каждую секунду. В домах, школах, на работе достижения в области технологий Интернета вещей генерируют большие объемы данных. Куда бы вы ни пошли, и все, что вы делаете в этом цифровом мире, становится новым источником данных. Данные генерируются датчиками, устройствами, видео, аудио, сетями, файлами журналов, транзакционными приложениями, Интернетом и социальными сетями. Большой объем, высокая скорость и большое разнообразие этих наборов данных - ключевая особенность, которая отличает данные от больших данных. Появление этих больших наборов данных требует более совершенных методов, технологий и инфраструктуры для обработки данных и преобразования их в полезную информацию. Данные больше нельзя хранить на нескольких машинах или обрабатывать одним инструментом.

Отличие данных от больших данных может быть установлено следующими **четырьмя характеристиками больших данных**:

объем - описывает объем данных, которые транспортируются и хранятся и задача состоит в том, чтобы найти способы наиболее эффективной обработки увеличивающихся объемов данных, которые, по прогнозам, вырастут в 50 раз к 2020 году до 35 зеттабайт;

скорость - описывает скорость, с которой эти данные генерируются, например, данные, полученные от миллиарда акций, проданных на фондовой бирже, нельзя просто сохранить для последующего анализа, т.к. инфраструктура данных должна быть способна немедленно реагировать на запросы приложений, осуществляющих доступ к данным и их потоковую передачу;

разнообразие - описывает характеристику данных, которые редко находятся в состоянии, полностью готовом для обработки и анализа, такие как неструктурированные данные, которые, по оценкам, составляют от 70 до 90% данных в мире;

достоверность (правдивость) - описывает процесс предотвращения искажения наборов данных из-за неточных данных, например, когда выполняется регистрация в онлайн-аккаунте, то часто используется ложная контактная информация и, т.о., повышение достоверности сбора данных снижает объем необходимой очистки данных.

Хотя здесь перечислены четыре характеристики больших данных, большинство обсуждений, инструментов и документов будут касаться только трех главных характеристик – объем, скорость и разнообразие.

Наличие правильных данных, которые можно преобразовать в информацию, а затем в бизнес-аналитику, имеет решающее значение для достижения целей обработки больших данных. **Источники больших** данных растут в геометрической прогрессии и, в качестве примеров можно привести следующие:

сенсоры и датчики в телеметрии для мониторинга транспорта, умный учета, управления запасами и отслеживание активов, управления автопарком и логистики;

транзакционная информация – фиксируется и сохраняется по мере возникновения событий, а также используется для анализа ежедневных отчетов о продажах и производственных графиков, чтобы определить, сколько запасов нужно запланировать;.

аналитическая информация – используется для решения задач управленческого анализа, для определения того, следует ли организации строить новое производственное предприятие или нанимать дополнительный торговый персонал.

Следует отметить, что большие данные должны быть отсортированы и проанализированы, чтобы иметь ценность.

С ростом важности данных для предприятий и общества возникает много вопросов относительно конфиденциальности и доступности крупных публичных и частных репозиториев больших данных. Важно понимать континуум между открытыми и частными данными для принятия решений о том, какие данные и как именно будут использоваться в организации, а так же какое решение будет использовано для реализации распределенного хранения и обработки больших данных.

Фонд Open Knowledge определяет открытое знание как «любое содержание, информация или данные, которые люди могут свободно использовать, повторное использование и перераспределить без каких - либо правовых, технологических или социальных ограничений.» Открытые данные составляют строительные блоки открытых знаний, а открытые знания - это то, чем становятся открытые данные, когда они полезны и пригодны для использования (см. https://okfn.org/).

Одними из самых наиболее часто используемых форм представления, а, по факту, способами визуализации больших данных являются:

1. Облако тегов. Каждому элементу в облаке тегов присваивается свой весовой коэффициент. Чем выше этот коэффициент, тем больше размер шрифта. Весовой коэффициент зависит от важности элемента, частоты изменения его состояния и других факторов, определяемых экспертами. Это позволяет человеку выявить из всей информации ключевые моменты.
2. Графики и диаграммы. Они помогают быстро представить информацию в наглядном виде.
3. Исторический поток. Он позволяет просматривать всю историю редактирования документа: кто редактировал, что добавил, сколько времени на это потратил.
4. Пространственный поток. Эта технология предоставляет возможность пользователю следить за распределением и перемещением информации по всему миру. С помощью такого отображения данных можно выделить регионы, где данная информация наиболее востребована.
5. Семантическая сеть . Она представляет собой ориентированный граф, отображающий смысловые связи между объектами. Это средство семантического анализа, т. е. учитывающее смысловое сходство между объектами.

3. Цели и прикладные задачи обработки больших данных

В соответствии с Распоряжением Правительства Российской Федерации от 28 июля 2017 г, №1632-р об утверждении программы "Цифровая экономика Российской Федерации", **основными целями** являются:

создание экосистемы цифровой экономики Российской Федерации, в которой данные в цифровой форме являются ключевым фактором производства во всех сферах социально-экономической деятельности и в которой обеспечено эффективное взаимодействие, включая трансграничное, бизнеса, научно-образовательного сообщества, государства и граждан;

создание необходимых и достаточных условий институционального и инфраструктурного характера, устранение имеющихся препятствий и ограничений для создания и (или) развития высокотехнологических бизнесов и недопущение появления новых препятствий и ограничений как в традиционных отраслях экономики, так и в новых отраслях и высокотехнологичных рынках;

повышение конкурентоспособности на глобальном рынке как отдельных отраслей экономики Российской Федерации, так и экономики в целом.

Основными задачами нормативного регулирования в рамки программы "Цифровая экономика Российской Федерации" в части больших данных являются:

обеспечение благоприятных правовых условий для сбора, хранения и обработки данных в том числе с использованием новых технологий, при условии защиты прав и законных интересов субъектов данных и владельцев;

адаптация антимонопольного законодательства к потребностям цифровой экономики;

построение федеральной сети узкополосной связи по технологии LPWAN для сбора и обработки телематической информации;

обеспечение правового режима и технических инструментов функционирования сервисов и использования данных;

обеспечение защиты прав, свобод и законных интересов личности в условиях цифровой экономики;

обеспечение защиту прав и законных интересов бизнеса в условиях цифровой экономики;

обеспечение организационную и правовую защиту государственных интересов в условиях цифровой экономики.

Основными сквозными цифровыми технологиями, которые входят в рамки программы "Цифровая экономика Российской Федерации", являются:

большие данные;
нейротехнологии и искусственный интеллект;
системы распределенного реестра;
квантовые технологии;
новые производственные технологии;
промышленный интернет;
компоненты робототехники и сенсорика;
технологии беспроводной связи;
технологии виртуальной и дополненной реальностей.

Портер М. из Гарварда описывает, как информационных технологии в третий раз за последние 50 лет изменили бизнес:

«Первая волна ИТ в 1960-х и 1970-х годах автоматизировала отдельные действия, такие как выплата стипендий сотрудникам или поддержка разработки и производства продуктов.

Второй волной трансформации бизнеса стало появление Интернета в 1980-х и 1990-х годах, которое позволило координировать и интегрировать внешних поставщиков, каналы сбыта и клиентов в разных регионах.

С IoT мы сейчас находимся в третьей волне, ИТ становятся неотъемлемой частью самого продукта. Встроенные датчики, процессоры, программное обеспечение и возможности

подключения к продуктам (по сути, компьютеры помещаются внутрь продуктов) в сочетании с облаком, в котором хранятся и анализируются данные о продуктах, а также запускаются некоторые приложения, существенно улучшают функциональность и производительность продукта. Огромные объемы данных об использовании новых продуктов позволяют многие из этих улучшений ».

Жизненный цикл анализа данных

Жизненный цикл анализа данных начинается с вопроса, а каждый шаг в жизненном цикле анализа данных включает в себя множество задач, которые необходимо выполнить, прежде чем перейти к следующему шагу.

Жизненный цикл анализа данных состоит из следующих этапов:

Сбор данных - процесс поиска данных и последующего определения того, достаточно ли данных для завершения анализа.

Подготовка данных - этот шаг может включать в себя множество задач по преобразованию данных в формат, подходящий для инструмента, который будет использоваться. Как правило, необходимо внести некоторые коррективы, чтобы помочь ответить на вопрос.

Выбор модели - этот шаг включает в себя выбор метода анализа, который наилучшим образом ответит на вопрос с имеющимися данными. После выбора модели выбирается инструмент (или инструменты) для анализа данных.

Анализ данных - процесс тестирования модели на основе данных и определения надежности модели и проанализированных данных. Удалось ли вам ответить на вопрос с помощью выбранного инструмента?

Представление результатов - обычно это последний шаг для аналитиков данных. Это процесс сообщения результатов лицам, принимающим решения. Иногда аналитика данных просят рекомендовать действия. Можно использовать гистограмму, круговую диаграмму или другое представление, чтобы указать, какое преступление было наиболее распространенным.

Принятие решений - последний этап жизненного цикла анализа данных. Лидеры организаций включают новые знания как часть общей стратегии.

При необходимости, процесс начинается заново со сбора данных.

Сбор и подготовка данных – извлечение, преобразование и загрузка данных

Большая часть данных, которые будут помещены в базу данных для последующего запроса, поступает из множества источников и в широком диапазоне форматов. Извлечение, преобразование и загрузка (ETL) - это процесс сбора данных из этого множества источников, преобразования данных и последующей загрузки данных в базу данных. Данные организации можно найти в текстовых документах Word, электронных таблицах, обычном тексте, презентаций, электронных письмах и файлах PDF. Эти данные могут храниться на различных серверах, которые используют разные форматы.

Процесс ETL состоит из трех этапов:

1. **Извлечение** – получение данных из нескольких источников;
2. **Преобразование** - агрегирование, сортировка, очистка и объединение данных;
3. **Загрузка** – загрузка данных в базу данных для хранения и обработки.

Извлечение данных

На этапе извлечения собираются нужные данные из источника и становятся доступными для обработки. Извлечение преобразует данные в единый формат, готовый к преобразованию. Например, объединение данных с сервера NOSQL и базы данных Oracle предоставит вам данные в разных форматах. Эти данные необходимо преобразовать в единый формат. Кроме того, данные должны быть проверены, чтобы убедиться, что они содержат желаемый тип информации (значение). Это делается с помощью правил проверки. Если данные не соответствуют правилам

проверки, они могут быть отклонены. Иногда эти отклоненные данные исправляются, а затем проверяются. В идеале во время извлечения все необходимые данные из источника (источников) извлекаются с использованием минимальных вычислительных ресурсов, чтобы не влиять на производительность сети или компьютера.

Преобразование данных

На этапе преобразования используются правила для преобразования исходных данных в тип данных, необходимых для целевой базы данных. Сюда входит преобразование любых измеренных данных в один и тот же размер (например, из британских в метрические). На этапе преобразования также требуется несколько других задач. Некоторые из этих задач - объединение данных из нескольких источников, агрегирование данных, сортировка, определение новых значений, вычисляемых на основе агрегированных данных, а затем применение правил проверки. Хотя может показаться, что эти данные полностью готовы к загрузке, обычно еще предстоит работа по их подготовке. Данные (возможно, включая некоторые отклоненные данные) могут пройти другую часть этапа преобразования, известную как «очистка» или «очистка» данных. Часть очистки на этапе преобразования дополнительно обеспечивает согласованность исходных данных.

Загрузка данных

Преобразованные данные загружаются в целевую базу данных. Это может быть простой плоский файл или реляционная база данных. Фактический процесс загрузки сильно различается. Это зависит от типа исходных данных, типа целевой базы данных и типа выполняемого запроса. Некоторые организации могут перезаписать существующие данные совокупными данными. Загрузка новых преобразованных данных может производиться ежечасно, ежедневно, еженедельно или ежемесячно. Это может произойти только тогда, когда в преобразованные данные были внесены определенные изменения. На этапе загрузки применяются правила, определенные в схеме базы данных. Некоторые из этих правил проверяют уникальность и непротиворечивость данных, обязательные поля имеют требуемые значения и т. Д. Эти правила помогают гарантировать, что загрузка и любой последующий запрос данных будет успешным.

Безопасность данных

Безопасность – способность противодействовать определенному множеству угроз, преднамеренных или непреднамеренных дестабилизирующих воздействий на используемые при обработке больших данных средства вычислительной техники, линии связи и технологические процессы (протоколы), которые могут привести к ухудшению качества услуг, характеризующаяся обеспечением конфиденциальности, целостности и доступности принимаемой, обрабатываемой и передаваемой информации (сведений, данных, сообщений):

конфиденциальность – способность не давать права на доступ к информации (сведениям, данным, сообщениям) или не раскрывать её не уполномоченным лицам, логическим объектам или процессам;

целостность – способность не подвергать изменению или аннулированию информации (сведений, данных, сообщений) в результате несанкционированного доступа со стороны не уполномоченных лиц, логических объектов или процессов;

доступность – способность обеспечивать использование информации (сведений, данных, сообщений) по запросу со стороны уполномоченных лиц, логических объектов или процессов.

Одной из наиболее перспективных цифровых технологий является интеллектуальный анализ больших данных, под которым понимается набор процедур, предназначенных для создания описательных и графических сводок данных, предполагающих, что результаты могут выявить интересные закономерности, т.е. процесс, который позволяет нам сформировать гипотезу о данных и найти ответ на новый вопрос. Иногда цель анализа - ответить на конкретные вопросы. В

других случаях у кого-то может быть «предчувствие» или интуиция относительно какого-либо явления в отношении набора данных. Интеллектуальный анализ данных может исследовать причину или следствие этого явления и предоставляет полезный способ изучить данные, чтобы определить, существуют ли какие-либо отношения между наблюдаемыми или собранными данными или есть ли проблемы в данных.

В целом аналитика охватывает более обширную область инструментов, чем статистика. Аналитика использует инструменты математического моделирования в статистике в дополнение к другим формам анализа, таким как машинное обучение, что может включать работу с очень большими наборами данных, которые, в свою очередь, содержат неструктурированные данные.

Заключение

На лекции были рассмотрены термины и определения предметной области больших данных, их характеристики и формы представления, а также цели и прикладные задачи обработки больших данных. Повторение материала лекции с использованием рекомендованной литературы позволит хорошо закрепить теоретический материал. Принципы, способы, методы, и инструменты аналитической обработки больших данных будут рассмотрены на следующей лекции.