Слайд 1. В последнее время технологии больших данных развиваются высокими темпами. Возникает множество высокотехнологичных компаний, предоставляющих новые инструменты по анализу данных. В то же время растёт мировой спрос на специалистов, способных работать с большими объёмами данных, ставить задачи в этой области, понимать сложность и потенциальную стоимость таких работ. Предполагая нарастающий интерес к этой сфере в России, сотрудники образовательных и бизнес-организаций Новосибирского научного центра решили систематизировать опыт работы с технологиями и задачами анализа больших данных. Анализ образовательного рынка в области больших данных показывает нам, что онлайн-образование в сфере аналитики становится всё более популярным и массовым. Существует несколько курсов на английском языке, в т.ч. курсы Data Science and Big Data Analytics от EMC, ряд курсов на платформах Coursera и Udacity. Однако на русском языке подобные курсы пока отсутствуют. Настоящий курс призван ликвидировать данный пробел. Он представляет собой образовательный инструмент для введения в эту, без сомнения вдохновляющую и перспективную, область. Авторы курса предлагают слушателям окунуться в мир методов и технологий анализа больших данных на простых примерах.

Слайд 3. Давайте попробуем установить, что для нас Большие данные на данный момент. Подумайте, вспомните основные ассоциации, попробуйте зарисовать ментальную карту всего, что Вы знаете о больших данных. Выделите затем отдельно то, что на ваш взгляд относится к аналитике больших данных. Возьмите на это 5 минут. Рекомендую воспользоваться каким-нибудь mindmap-инструментом. Итак, нажмите на паузу, а на следующем слайде мы сверимся с вами и вы определите процент попадания.

Слайд 4. Оказывается, всё не так уж и многообразно. Даже тех ассоциаций которые мы имеем на текущий момент достаточно, чтобы вполне чётко себе представить область больших данных. Это конечно же сами данные, их объём.

Слайд 6. Давайте посмотрим на объём с исторической точки зрения. На этом графике мы видим рост объёмов информации, представленной в аналоговом и

цифровом виде. Объём экспоненциально увеличивается. При этом мы видим, что где-то с 2000х годов происходит переломный момент – цифровые носители получают широкое распространение тем самым давая всё большему количеству информации сохраняться и быть доступной уже в цифровом виде. Конечно же здесь мы не учитываем тот факт, что в 1986 году считалась информация, специально отобранная (библиотеки, фильмотеки и т.п.), а в 2002 году уже имеется просто вся информация, в том числе бесчисленные копии фильмов, фотографий и текстов. Отметим, что особый вклад в развитие цифровой эпохи внесли жёсткие диски. Удешевление их производства – основной фактор формирования тренда больших данных.

Слайд 7. Есть и другие причины и факторы возникновения Big Data. Можно ли было говорить об этом десять лет назад? Вряд ли. Собираемые данные были не так заметны. Не во всех машинах стояли датчики для контроля состояния двигателей, не так были распространены средства для мониторинга здоровья, шагомеры и прочие сенсоры. Не так были распространены мобильные устройства. Все эти вещи генерируют данные, которые нужно где-то хранить и обрабатывать.

Кто и когда будет анализировать эти данные и синтезировать новые решения?

Слайд 9. Во-вторых, даже когда происходит развитие технологий, накопление данных осознание новых возможностей не происходит одномоментно. Появление термина большие данные связывают с публикацией в журнале Nature, но также известно и то, что ранее этот термин уже употреблялся. Тем не менее, правильнее говорить именно о 2008 году, т.к. статья констатировала некий общественный статус технологий и запустила массовый процесс осознания её возможностей. Однако только в 2011 году появляется отчёт компании МакКинзи, который очень сильно повлиял на популяризацию тренда. Таким образом, с публикацией данного отчёта общество перешло в фазу создания новых технологий для обработки Big Data (большие инвестиции в рынок технологий BigData) и

одновременной подготовки кадров для этой новой отрасли: Data Engineer, Data Scientists, BD-analyst (нехватка 140 тыс. специалистов).

Что такое аналитика больших данных? Вы уже наверняка знаете основные характеристики больших данных: объём, скорость и разнообразие. Кто-то приводит ещё: ценность, виртуализацию, верификацию и пр. Всё это также относится к большим данным. К ним в ближайшее время будет относиться практически всё. И это не шутка. Технологии позволяют хранить огромные объёмы (Петабайты) на всё меньшем кусочке пространства. Это и развитие сенсоров приводит к тому, что падают затраты на сбор данных.

Слайд 10. Однако возникают затраты на хранение. При условии высоких скоростей генерирования данных, например, сообщений вКонтакте или других соц.сетях, проблема сохранения и обработки на лету тоже становится заметной. С годами компании понимают, что данных уже столько много, что осмысленно поставить вопрос: зачем нам испытывать издержки на их хранение? Появляется мысль о рациональном использовании сохранённых данных: давайте не будем их выбрасывать, а постараемся извлечь из них пользу. Так мы наблюдаем бум разных технологий хранения и обработки данных. Всё это положительно сказывается на рынке систем хранения данных. И чем они надёжнее, быстрее и эффективнее, тем больше их покупают, тем больше данных собирают, тем больше растёт потребность в сборе ещё каких-нибудь данных.

Однако, большой объём данных отнюдь не означать и большую их ценность. Эта зависимость не линейная. Даже если обратиться к собственному опыту – давно ли Вы просматривали свои тысячи фотографий, отснятых за последний год? С ростом объёмов видимая ценность на байт данных падает. Возникает очень много дублей, записей и перезаписей. Всё это приводит к падению качества исходного материала. И это при росте технических качественных характеристик записывающих устройств. Здесь мы сталкиваемся не только с возможностями, которые нам открывают большие данные, но и с проблемой качества исходных данных.

Вот мы и приходим к основным драйверам этого рынка:

- увеличение потоков информации;
- удешевление систем хранения на единицу информации;
- усовершенствование технологий обработки информации.

Слайд 11. Очевидно, что для целей анализа нам особенно важно понимать какие перед нами данные, что они описывают и с какой точностью.

Мы можем заметить, что с началом цифровой эпохи повысилась доступность генерирования и хранения цифровой информации. Это означает, что если раньше задумывались о том, что записать на оставшиеся 30Кб дискеты, то сейчас о гигабайтах не особо задумываются. Если раньше на фотоаппараты Зенит отбирался каждый кадр, тщательно выставлялась экпозиция, то сейчас люди делают до тысячи снимков за день… на один раз пересмотреть, или ни разу. Это обывательский подход к качеству информации и к стоимости одного байта.

Однако та же тенденция сохранятся и в области бизнес-информации.

Один вид информации часто собирается очень тщательно (финансовые прогнозы, планы производства), настолько тщательно, что на рабочих местах пользователей образуется большое число дублей информации без возможности установить, какая информация более актуальная. С этой проблемой призваны бороться ERP, CRM и другие BI-системы структурированного хранения информации.

Если приглядеться уже в архитектуру BI-систем, то можно заметить два типа архитектур хранения бизнес-информации: хранящая текущие состояния объектов и хранящие всю историю изменения объектов. Вторые годятся для дальнейшего анализа и составляют основу бизнес-аналитики больших данных. Главным образом из-за высокого качества исходной информации.

Однако при анализе могут быть поставлены такие вопросы, которые не ставились при сборе. Таким образом, возникает проблема несоответствия цели сбора и использования.

Слайд 12. Компании, разрабатывающие системы хранения как раз и заинтересованы в популяризации больших данных, чтобы обеспечить себе рынок сбыта. Вторыми игроками на этом рынке оказываются компании, которые предлагают решения и услуги по извлечению из собранных данных новых знаний, сегментаций клиентов, новых решений старых и подступающих проблем. Возникают угрозы безопасности и вопросы законности сбора данных – это сдерживает развитие рынка, но с другой стороны придаёт ему устойчивость; но об этом попозже.

Слайд 13. Третьей характеристикой выделяют разнообразие. Действительно, мы наблюдаем большое количество оцифрованной информации в виде каких-то документов, таблиц, баз данных, сайтов и т.п. Если базы данных достаточно понятны в машинной обработке, то XML (полуструктурированные данные) и текстовые документы (неструктурированная информация) представляют определённую проблему, т.к. для них таких нет универсальных методов таких, как SQL для СУБД. Если они и разрабатываются где-то, то ещё так не распространены. Основной проблемой в обработке неструктурированной информации представляется извлечение смыслов текстов, решению которой посвящено целое направление научных исследований (Semantic Web), в т.ч. в корпорациях Google и Яндекс. Есть и другие проблемы с обработкой неструктурированной информации, об этом позже.

Слайд 14. Машинные данные можно условно разделить на несколько типов:

1. структурированная информация
    1. Таблицы СУБД с известными типами данных
2. Полуструктурированная информация

1. XML-документы со своими XSD-схемами

3. Неструктурированная информация.

1. Текстовые документы

2. Видеоконтент

3. Аудиоконтент.

Следует, однако, отметить, некоторую относительность этой классификации. Если вы спросите автора книги, является ли текст его рукописи структурированным, то он, конечно, ответит «да». Ведь он всё структурировал и даже заголовки жирным шрифтом выделил. Режиссер, конечно же, ответит «да» про свой фильм. Но с точки зрения машинной обработки, ничего не знающей о том, что знает автор книги или фильма – эта информация неструктурированная. Документы в виде XML могут дать определённую информацию о структуре, но способ обработки этой информации, т.е. метаинформация может быть неизвестна машине. В идеале было бы хорошо, если к каждому файлу прилагалась машинная инструкция (программа), которая бы описывала, как работать с этой информацией. То есть информация с т.зр. машины становится структурированной, если машина «знает», как её обрабатывать.

Слайд 15. Отдельно следует упомянуть виртуализацию – это технологии которые позволили оторваться от конкретных физических устройств и моделировать их уже в виртуальной среде, что тоже придало некоторое ускорение в развитии облачных технологий, а значит и общей доступности ИТ для более широкого круга компаний и населения.

Слайд 16. Необходимо ещё раз остановиться на понятии Большие данные. Если понимать это как проблему, которая не решается на существующем уровне технологий, что такое понятие очень расплывчато по определению. Если завтра изобретут технологию, решающую проблему, что это уже не большие данные – так

выходит? Выходит, что так. Но это и характерно для проблемы. Сделали технологию – проблема ушла.

Мы видим, что приведённые определения существенно разнятся. И это нормально для молодого направления деятельности. Мы будем считать «большими данными» только первую часть определения русской википедии, т.е. подходы, инструменты и методы обработки данных больших объёмов и многообразия для получения результатов, в условиях непрерывного прироста информации и её значительной распределённости. Также будем понимать и сами данные, обрабатываемые этими методами.

Слайд 17. Проблема в том, что имея видеозапись мы не можем сказать, о каком она объекте, пока не просмотрим.

А можем ли мы автоматически это определить? Распознать фрагменты и присвоить метки – структурировать эти данные.

Слайд 19. Итак, мы утвердились в понимании основных характеристик больших данных.

Рассмотрим более подробно источники (генераторы) данных. На заре цифровой эпохи это были научные установки, эксперименты, заказы крупных корпораций. Теперь же мы имеем устройства в кармане, генерирующие непрерывный поток каких-то данных (интернет-трафик, акселерометр, GPS, и др.). Все устройства будто бы просят подключить их к глобальной сети. Есть уже умные утюги, розетки, сообщающие о своём состоянии в головной центр управления домом [картинку]. Некоторые даже утверждают, что китайские чайники следят за нами и прослушивают ☺ [картинку].

Большие данные генерируются в коммуникациях устройство-устройство и устройство-человек. Например, сервера накапливают в себе информацию о своей работе, логируют всевозможные действия. Данные из этих логов позволяют прогнозировать отказ системы или атаку извне. Люди ежедневно загружают на YouTube терабайты видеопотока. Эти видео-ролики могут многое сказать,

например, о факте падения метеорита или проведении массовых митингов, в том числе в качестве дезинформации. Ещё одним источником больших данных является установка БАК в CERN. Она генерирует порядка 300Тб в секунду. Конечно, не все эти данные записываются на носители. Записывается от силы 1%, а анализируется и того меньше. Есть и другие источники больших данных. Можно провести небольшую их классификацию:

Слайд 20. При таких объёмах информации возникает не только проблема их хранения, но и проблема их последующего извлечения. При настоящей мощности каналов связи, например, мы в Новосибирске не сможем получить всю информацию сгенерированную БАК в Швейцарии. Мы можем передать лишь незначительную часть. Тем не менее, люди каким-то образом проводят анализ этих данных. Итак, мы имеем проблему извлечения и анализа информации и систем хранения. Рассмотрим её подробнее.

Вот установка, которая сохраняет в себе 80Тб информации в день. Есть канал связи, который обеспечивает пропускную способность: 1Гб/сек. К концу дня информация сохранена. Допустим мы хотим её всю извлечь, и не за один день, а за год. Для этого нам нужен новый канал, т.к. этот уже занят записью информации следующего дня. Но что более невероятно, ширина этого канала должна быть в 365 раз больше существующего, чтобы извлечь всю информацию. Из-за этих ограничений очевидно, что нужно обрабатывать информацию прямо там и мощностью тех серверов, где она была сохранена. Такая идея лежит в основе технологий **Hadoop** и модели **MapReduce**.

Слайд 21. Вычислительная модель MapReduce была впервые предложена инженерами из Google Джеффри Дином и Сенджейем Гемаватом в 2004 году. Основная идея заключается в следующем:…

В реализации для GFS, как и в Hadoop, данное решение обладает свойством **локальности**. Т.е. используются те вычислительные ресурсы, на которых сохранена информация.

Слайд 22. Hadoop это ещё одно решение реализующее модель MapReduce.

Файловая система HDFS. Hadoop – это целостное решение для хранения и обработки данных. Сама библиотека – открытая, но есть ряд компаний, которые зарабатывают на обслуживании решений на базе этой библиотеки. Cloudera, например, предоставляет хостинг с Hadoop, а также поставляет шкафы – уже укомплектованные кластеры.

Слайд 23. Теперь, когда мы познакомились с самыми распространёнными технологиями больших данных, можно перейти к вопросам собственно аналитики. В чём состоит анализ больших данных? Следует отметить, что для больших данных пока нет стандартного процесса аналитики. В то время как для «небольших» он есть. Рассмотрим процесс аналитики на примере стандарта CRISP-DM. Здесь есть как процессы загрузки, анализа так и представления результатов. …

Чем же он принципиально отличается от больших данных? Вы уже знаете, что, во-первых: мы не можем просто так взять и извлечь нужные нам данные, перекачать из одного места в другое. Т.е. у нас наложены ограничения на процесс ETL. Нам надо рассмотреть, где именно хранятся эти данные, реализован ли там интерфейс MapReduce? Если нет, какие средства отбора и обработки информации есть в этой системе хранения? Итак, в больших данных нам требуются (1) средства предварительной обработки информации на местах её хранения; (2) возможность запуска алгоритмов анализа прямо на этих данных. Причём, заметим, что для этих алгоритмов должно выполняться свойство локальности: мы не можем внутри них пользоваться данными из разных кластеров, иначе это повлечёт массовую загрузку пропускной способности.

После предварительной обработки данные могут попасть на аналитический сервер, у которого и вычислительные возможности помощнее и есть средства визуализации.

Мы уже поняли, что для извлечения данных необходимо использовать массово-параллельные вычисления, например MapReduce. Т.е. в любом проекте с большими данными нам следует помнить о времени исполнения процесса. Чтобы он не затягивался на месяцы ☺. Когда мы говорим об анализе данных, то мы имеем дело не только с процессом извлечения данных, но и их многократной переработке. Такую переработку можно осуществлять правильно написав функции MapReduce или использовав надстройки над системами кластерных вычислений. Существует ряд решений СУБД, также реализующих парадигму MapReduce, такие как Cassandra, MongoDB. Они позволяют оперировать данными не слишком задумываясь, как они расположены в кластере. Также есть библиотека Mahout от Apache для интеллектуального анализа данных на кластере Hadoop.

Давайте рассмотрим цикл работы аналитика больших данных. Мы уже знаем, что на таких объёмах не все запросы дают мгновенный результат. Поэтому, чтобы не тратить время впустую, каждое ресурсоёмкое действие должно быть или тщательно спланировано или апробировано на данных меньшего объёма.

Слайд 25. Большие данные позволяют нам отойти от некоторых традиционных схем принятия решений и больше положиться на статистические методы. Если Вы видите, что предложенная Вами схема не работает в нескольких случаях, это ещё не повод её отклонять. Может быть количество случает неработоспособности схемы составляет менее 2%. Тогда эта схема очень даже работоспособна.

Это значит, что нам не надо обращать внимание на детали. И конечно же необходима культура, чтобы все гипотезы проверять на объёме данных. Ни одна гипотеза не может быть принята или отклонена на основе лишь её кажущейся правдоподобности. Что это значит на практике? Значит, что в процессе предварительного и глубинного анализа данных необходимо все гипотезы фиксировать и проверять. Выводы могут быть сделаны только на основе проверенных гипотез. Как это выглядит в работе аналитика: вот я обозреваю таблицу с данными. Я строю диаграмму и вижу, что часть данных обладает

закономерностью. Но это я только вижу. Чтобы сделать заключение об этой закономерности, я сначала явно её формулирую, записываю. Затем перевожу в математический критерий, и запускаю алгоритм, выполняющий этот критерий, подчитываю количество объектов, которые подтверждают мою гипотезу и тех, кто опровергает. Вот это соотношение и является доказательством верности гипотезы.

Слайд 26. То о чём мы говорили вначале лекции – это были концепции. Существуют технические решения, использующие Hadoop и MapReduce, которые включают в себя полный комплекс для ведения процесса аналитики: и систему сопряжения с хранилищем, и подсистему извлечения данных, подсистему машинного обучения и систему визуализации, представления и распространения результатов аналитики. Таковы, например, решения от IBM: BigInsights, BigSheets, Stream, решения от SAP: SAP HANA, от EMC Greenplum: Chorus.