Лекция 1. История BigData. Типы хранения данных, инструменты обработки, компании, использующие BigData

Появление больших данных

Не было бы просто данных — не появились бы большие данные. Данные — это основа понимания. Иногда цепочка «данные — информация — знание» представляют в виде пирамиды, где данные составляют основание, а знание — вершину. Информация строится на основании данных. Мы собираем группы каким-то образом связанных данных и так получаем представление о мире или важную информацию об окружающем нас пространстве. Слова в этой лекции — данные. Информация — это слова, соединенные в предложения, предложения, разделенные на абзацы, а абзацы на текст. И из информации получились знания. Знания — это интерпретация информации для ее использования: вы читаете книгу, обрабатываете информацию, и у вас формируется мнение, появляются собственные идеи, вы предпринимаете какие-то действия.

Данные могут быть и набором цифр, они в свою очередь могут быть представлены различным образом, например, в таблице.

Если вы охотник, то вы, например, знаете или ищете информацию о том, когда в ближайшем к вам лесу бывает больше всего уток. Существуют специальные издания и сайты, где публикуется количество рыбы в той или иной местности по месяцам.

Используя эту информацию, вы принимаете решение о том, когда пойти охотиться на уток или порыбачить.

Хотя может показаться, что большие числа встречаются только в современном мире, а в текстах и хрониках, которые оставила нам история, можно увидеть только маленькие числа, это не так. В Оксфордском университете хранится артефакт, возраст которого составляет около 5 000 лет. В нем рассказывается о победе фараона Нармера над ливанцами к западу от дельты Нила. Описывается, как Египет взял в плен 120 000 человек, захватил 400 000 быков и 1 422 000 козлов. Сотни тысяч и миллионы также упоминаются в египетской Книге мертвых. Для того периода это очень большие данные.

Сложности с большими данными возникли в связи с проведением переписи населения. Первая перепись населения США была проведена в 1790 году. Тогда население Соединенных Штатов составило чуть менее 4 миллионов человек — 3 929 326 человек, включая рабов. Во время последней переписи, которая проводилась в 2010 году, население страны составляло уже 308 745 538 человек. В соответствии со статьей 1 Конституции США перепись населения должна проводиться не реже чем раз в десять лет. Она проводится в годы, заканчивающиеся на «0» . С 1790-го по 1840 год она проводилась шерифами, а в 1840 году появился первый центральный офис Бюро по переписи населения.

И каждый раз людям, занимавшимся переписью населения, казалось, что поставленная перед ними задача обречена на провал. И все — из-за количества данных. Их количество постоянно росло, нужно было обрабатывать и хранить все больше и больше данных, а доступных и удобных инструментов не хватало.

В первые годы все, конечно, делалось вручную. Люди сами чертили таблицы, вносили туда данные, считали без помощи вычислительных машин, пересчитывали по несколько раз, чтобы избежать ошибок. Иногда данные одной переписи не успевали полностью проанализировать до начала следующей! А ведь период между ними составлял десять лет! И следующая перепись приводила ответственных за нее чиновников в еще больший ужас, потому что население росло с каждым годом, и данных во время каждой следующей переписи населения получалось больше, чем во время предыдущей.

Проблема была решена с помощью механизации. В 1890 году при переписи впервые использовали электрическую табулирующую машину Германа Холлерита (1860 - 1929) для обработки данных.

Табулятор — это электромеханическая машина, предназначенная для автоматической обработки (суммирования и категоризации) числовой и буквенной информации, записанной на перфокартах. Результаты выдаются на бумажную ленту или специальные карты. И до появления электронно-вычислительных машин табуляторы использовались по всему миру. Табуляторы достаточно эффективно складывали и вычитали. С умножением и делением было сложнее: требовалось многократное последовательное повторение сложения или вычитания. Многие изобретения появились для облегчения ручного труда. Табулятор — один из них. Идея использования перфокарт принадлежала Джону Биллингсу, будущему тестю Германа Холлерита, который был высокопоставленным чиновником в Бюро по переписи населения США. Ну а зять изобрел машину и оставил свое имя в истории. И не только как создатель табулятора, но и прадедушка компании *IBM*. Он создал компанию для производства своих табулирующих машин, потом продал, и она вошла в состав *International Business Machines,* или *IBM,* в настоящее время — одного из крупнейших в мире производителей и поставщиков аппаратного и программного обеспечения, IT -сервисов и консалтинговых услуг.

Электронные вычислительные машины появились в конце 1940-х годов. В настоящее время обработка данных переписи населения (больших данных) полностью автоматизирована, хотя в сборе данных до сих пор участвуют интервьюеры. Обработка ведется по единым правилам и плану. Первые переписчики о таком не могли и мечтать!

Мы живем в эпоху компьютерной, вычислительной, технологической революции. Современные технологии позволяют и обрабатывать, и хранить информацию.

Современные технологии позволили манипулировать информацией так, как никогда раньше. Современные машины способны не только обрабатывать данные, но и превращать их в информацию. Нам доступен Интернет, новейшие системы связи, мы можем в любой момент связаться с человеком, находящимся в другой части земного шара. А системы, работающие с огромными объемами данных, которые поступают быстро и неструктурированны, не могли бы работать без новейших технологий, позволяющих их обрабатывать и анализировать. До появления этих технологий работать с большими объемами данных было просто непрактично. Обычно использовалась выборка. Возьмем, к примеру, перепись населения. Для выборки тщательно выбиралась небольшая группа, которая, по мнению организаторов, наилучшим образом представляет все население, и анализировалась. Считалось, что она наглядно показывает состояние всего населения. Конечно, в результате была погрешность, но на тот момент это был наилучший инструмент справиться с большим потоком данных.

Термин и характеристики

Автором термина «большие данные» считается Клиффорд Линч, редактор журнала *Nature*. Дата рождения термина — 3 сентября 2008 года, когда вышел специальный номер, тема которого «Как могут повлиять на будущее науки технологии, открывающие возможности для работы с большими объемами данных?». В этом специальном выпуске редакционная коллегия собрала материалы, посвященные взрывному росту объемов обрабатываемых данных и их многообразия, а также технологическим перспективам этого феномена. Высказывались предположения о переходе от количества к качеству.

Английский термин *Big Data* был сформулирован аналогично уже имеющимся: *big business* (большой бизнес) — крупнейшие корпорации; *big oil* (большая нефть) — крупные нефтяные компании США; *big name* (дословно «большое имя») — знаменитость, известная личность.

Термин появился в академической среде и в первый год после появления использовался, когда говорили о росте объемов и многообразия исключительно научных данных. Но уже в 2009 году он просочился в деловую прессу и очень быстро получил широкое распространение. В 2010 году появились первые продукты и решения, относящиеся исключительно к проблеме обработки больших данных. В 2011 году большинство крупнейших поставщиков информационных технологий уже использовали понятие больших данных (например, *IBM, Microsoft).*

Также появились отдельные исследования на эту тему. В том же 2011 году большие данные были названы трендом №2 (после виртуализации) в информационно-технологической инфраструктуре. В 2013 году большие данные были включены в программы американских высших учебных заведений, где изучается наука о данных. 2015 год считается годом перехода к массовому практическому применению больших данных.

Понятие больших данных применяется практически в любой сфере современной информационной деятельности. В первую очередь это, конечно, IT -сфера, а также реклама, торговля и маркетинг, мобильные технологии. Они используются в банковской сфере, телекоммуникациях, энергетике, логистике, промышленности, государственном управлении. Первыми, напомним, их стали использовать метеорологи.

Количество данных постоянно растет, Интернет есть везде, поэтому любой бизнес вынужден думать об этой технологии.

Это ключевой элемент современного информационного пространства. Практически все, что делает отдельный человек, группы людей, человечество в целом, компании из разных сфер бизнеса, правительства, происходит в рамках глобального информационного поля.

Ваша работа, ваш досуг, шопинг, путешествия — все тем или иным способом связано с большими данными. Вы получаете и отправляете письма по электронной почте, вы звоните по телефону и звонят вам, вы серфите в Интернете, и таким образом вы получаете и отправляете биты информации и находитесь внутри системы больших данных. Финансовые операции проходят через Интернет. Все, что вы когда-либо публиковали в социальных сетях, остается во Всемирной паутине. Эти данные не исчезают. Современному человеку не уйти от больших данных. Отдельный человек физически не способен и никак не может успеть осмыслить процессы, которые происходят в информационном поле, в котором он находится.

В настоящее время к большим данным относятся потоки данных объемом свыше 100 Гб в день. В 2003 году в мире было накоплено 5 эксабайтов данных (1 эксабайт равен 1 миллиарду гигабайтов). В 2015 году их было уже более 6,5 зеттабайта (1 зеттабайт = 1024 эксабайтов). К 2020 году прогнозируется 40-4 4 зеттабайта данных. А к 2025 году этот объем вырастет в 10 раз. Мировой доход на рынке больших данных в 2017 году — *US$* 150,8 миллиарда. Их объем настолько велик, что обработка такого количества данных стандартными программными и аппаратными средствами представляется крайне сложной, а иногда просто невозможной.

Кроме основных характеристик *Big Data — volume, velocity, variety* (объем, скорость и разнообразие), есть еще четыре характеристики, которые появились позже. Это *value* (ценность), *veracity* (достоверность), *viability* (жизнеспособность) и *variability* (переменчивость). Несмотря на то, сколько *V* используется для характеристики больших данных, всегда подчеркивается, что физический объем не является основной

или определяющей характеристикой *Big Data.* Другие не менее важны и необходимы для понимания сложности задачи — обработки и анализа больших данных. И ведь любое дело должно быть экономически целесообразным, поэтому «ценность» часто

оказывается вторым *V* при характеристике больших данных после того, как традиционно отдается должное физическому объему.

Big Data — это социально-экономический феномен, который связан с появлением новых технологических возможностей для анализа огромного количества данных. Зачем они простому обывателю? Представьте, что в супермаркете, куда вы обычно ходите за покупками, по какой-то таинственной причине все продукты и товары перемешались. Кекс оказался рядом с молоком, а хлеб с шампунем, яблоки с мясом, рыба с соком. Именно большие данные помогут расставить все по местам, найти нужный товар, узнать срок годности и стоимость. Оперируя большими данными, вы также сможете узнать, чем полезен и чем вреден тот или иной продукт, при каких заболеваниях его нельзя употреблять, а при каких, наоборот, нужно. Полная информация о том, что есть в супермаркете и для чего это нужно, и есть большие данные. Огромные объемы данных обрабатываются, чтобы конкретный человек мог получить нужную ему конкретную информацию для ее дальнейшего применения. Это управление данными является решением проблем отдельного человека, компании, города, страны, мира.

Используя традиционные инструменты, невозможно обработать огромные объемы неоднородной и быстро поступающей информации. Новые современные инструменты для обработки и анализа больших данных позволяют увидеть закономерности, которые не может увидеть человек и даже старые инструменты.

Это помогает оптимизировать все сферы нашей жизни — производство, сбыт, телекоммуникации, даже государственное управление. Большие данные дают конкурентное преимущество.

Допустим, вы хотите узнать баланс вашей банковской карты. Обработка запроса занимает какие-то доли секунды. Это и есть скорости современного информационного рынка. Их требуют большие данные. Современные технологии позволяют обычному пользователю хранить гигабайты информации в своем кармане и у себя дома, бизнес и государственные структуры собирают, обрабатывают и анализируют данные в немыслимых раньше масштабах. Это позволяют современные технологии — технологии *Big Data.*

Большую часть данных генерируют предприятия, с каждым годом они становятся все более важным активом, все больше возрастает роль безопасности. Большие данные поступают из трех источников. Первый — это Интернет, то есть социальные сети, средства массовой информации, разнообразные сайты, форумы и блоги. Второй — это корпоративные архивы. Третий — показания различных приборов и устройств. Главное — научиться обрабатывать и анализировать эти огромные объемы информации. Ведь большие данные — это постоянно меняющаяся картинка. Бели вы правильно получили большие данные, это не только помогает преодолеть неточность, которая всегда сопровождает выборку, но и дает невероятные возможности: к данным из прошлого прибавляются данные из настоящего, и это помогает наилучшим образом справляться с ближайшим будущим. Это возможно, потому что, в отличие от традиционного статистического анализа, большие данные можно постоянно обновлять и учитывать все направления и тенденции.

Как уже говорилось выше, люди, занимающиеся прогнозированием, знают о сезонности и учитывают этот фактор, но большие данные позволяют учитывать множество факторов и вариаций. Можно добавлять новые пакеты или наборы, или ряды

данных и смотреть, помогают ли они делать краткосрочные прогнозы. Например, при прогнозировании продаж уже давно учитываются четыре сезона и праздничные дни.

Но теперь можно посмотреть, как погода в определенные дни влияет на продажи. И это относится не только к товарам, непосредственно связанным с погодой (например, зонтикам или резиновым сапогам), но и колбасе, и открыткам. Это все можно проверить, имея большие данные физически, а также технологии для их обработки и анализа. Если мы увидим, что какой-то фактор оказывает существенное (или какое-то) влияние на продажи, мы учтем его в прогнозировании продаж на соответствующие дни и сделаем все возможное, чтобы удовлетворить спрос.

Большие данные помогают составлять прогнозы, ориентируясь не только по сезонам и дням, но также и в зависимости от мест, где ведется торговля. Они позволяют изучить, какие товары в какой местности пользуются наибольшим спросом.

Например, хаггис (смесь овечьих потрохов с овсянкой, луком и приправами в оболочке для колбасы) — это национальное шотландское блюдо. Вы можете купить его в магазинах Шотландии, в США хаггис продается там, где живет много выходцев из Шотландии, которые традиционно его едят. В ряде штатов про него никогда не слышали. Угорь в желе — это лондонское блюдо, в других частях Англии и других странах оно непопулярно.

Черный пудинг популярен в Великобритании и Ирландии. И продажи можно «точно настроить» в зависимости от спроса на местах.

Есть три основных требования к работе с большими данными — мощные компьютеры, подключение к Интернету и правильный алгоритм. У вас может быть невероятно много данных физически, прекрасная связь с огромным количеством точек, откуда эти данные поступают, но наличия данных и связи недостаточно. Даже бесполезно. С ними же надо работать, а человек способен обрабатывать только небольшое количество данных за один раз, даже самый гениальный математик. С ними просто не справиться. Нужна помощь компьютерных программ, в частности, нужны алгоритмы. Алгоритм — это последовательность действий для выполнения какой-то задачи. В Оксфордском словаре английского языка говорится, что слово происходит от древнегреческого, обозначающего «число», как и «арифметика». Хотя есть и другая версия, и она представляется правильной, поскольку много слов (а то и все), начинающихся с «ал», происходят от арабских слов. В данном случае считается, что алгоритм происходит от имени узбекского математика, астронома, философа и историка Мухаммеда аль-Хорезми, который жил в *IX* веке. В первую очередь он был математиком, и благодаря

ему алгебра стала самостоятельной наукой. Его работы были основными учебниками по математике в европейских университетах на протяжении нескольких веков. В латинизированном варианте его имя звучит как *Algorizmi* или *Algorizmus.* Имя стало нарицательным, таким образом европейские математики стали называть любое вычисление по строго определенным правилам. Позднее понятие расширилось до набора инструкций, описывающих порядок действий для достижения результата в любой сфере деятельности.

Но независимо от происхождения слова, оно относится к набору процедур и правил, которые позволяют нам работать с данными. Одни и те же процедуры и правила могут применяться к различным наборам данных. Многие компьютерные программы включают алгоритмы, но для алгоритма не нужен компьютер, и есть компьютерные программы, которые не включают алгоритм. Пример простого алгоритма — это числа Фибоначчи. Эта числовая последовательность невероятно длинная, а алгоритм для нее очень короткий и простой: каждое последующее число равно сумме двух предыдущих

чисел. То есть инструкция будет звучать: возьмите две единицы и повторно прибавляйте последнее число в ряду к предыдущему, чтобы получить следующее значение.

Если мы говорим о больших данных, алгоритмы могут оказаться очень сложными. Но при этом они все равно будут состоять из процедур и правил, которые позволяют системе анализировать или генерировать данные. Система больших данных хороша настолько, насколько хороши алгоритмы, используемые для доступа к данным, обработки и управления ими. Алгоритм нейтрален. Ему все равно, что подразумевают данные, он просто делает то, что мы запрашиваем. Но нам, как пользователям больших данных, нужно быть очень осторожными с нашими предположениями и точно знать, что делает алгоритм. Самое важное — это правильно интерпретировать результаты. Алгоритм зависит от его разработчиков — они должны сделать правильные предположения о пользователях систем и правильные предположения о выводах, которые можно сделать из данных. Неправильные предположения часто являются причинами неудач при принятии решений.

Например, вы предполагаете, что успеете проскочить, пока на светофоре не загорится красный свет. Вы предполагаете, что успеете пересесть на другой самолет, если между рейсами у вас будет час времени. Но красный свет загорается раньше, чем вы предполагали, и вы попадаете в аварию, или ваш самолет опаздывает, и вам приходится проводить в аэропорту сутки, ожидая следующего стыковочного рейса. Вы встречаете человека в первый раз в жизни и только на основании его внешности и одежды делаете какие-то предположения, которые потом не оправдываются.

Люди постоянно делают предположения о том, что можно и что нельзя, такова человеческая природа, но эти предположения мешают творчеству и новым идеям. И точно так же создатели алгоритмов делают предположения об ограниченности данных и о том, как они будут использоваться. И если нет возможностей для внесения корректировок в алгоритмы, эти предположения будут мешать правильно интерпретировать и использовать данные.

Для примера также можно привести так называемую «проблему 2000 года». Можно считать, что компьютеры получили достаточно широкое распространение в 1960-е годы, когда до 2000 года было еще очень далеко. Разработчики программного обеспечения в XX веке часто использовали для обозначения года только две последние цифры. Соответственно, многие системы предполагали, что год начинается с «19» . То есть при наступлении следующего века такие системы могли предположить, что 2015 год — это 1915 год. Подобное могло привести к серьезным сбоям в работе финансовых программ и систем управления технологическими процессами. Программы вообще могли прекратить работать в 2000 году.

Проблема возникла из-за того, что разработчики программ не подумали о том, что может произойти при смене столетий. Было приложено немало усилий (теперь говорят, что даже больше, чем нужно) и, по некоторым данным, потрачено свыше 300 миллиардов долларов. Но проблема была своевременно обнаружена, проведена соответствующая подготовка, тестирование и профилактика. Хотя теперь говорят, что она была «раздута» с целью получения прибыли. Конечно, следовало проверять системы, управляющие самолетами, и банковские системы, но не базовое офисное программное обеспечение. В любом случае ни больших, ни малых сбоев не было, а это — самое главное. Но мы сейчас о другом — о неправильном предположении или, скорее, о халатности: не была учтена смена цифр с «19» на «20» . Это просто не пришло в голову разработчикам.

И что-то аналогичное может случиться с системами больших данных. Разработчиков алгоритмов для больших данных ждет очень большая работа. Хочется, чтобы они не забывали о «проблеме 2000 года», проводили как можно больше тестов,

проверяли как можно больше предположений и обеспечили возможности легкого внесения исправлений. Ведь обязательно что-то ускользнет, придется делать корректировку, так пусть эта корректировка обойдется не в 300 миллиардов долларов, а дешевле. Так что давайте думать о последствиях [1].

Технология Big data

Огромные объёмы данных обрабатываются для того, чтобы человек мог получить конкретные и нужные ему результаты для их дальнейшего эффективного применения.

Фактически, Big data — это решение проблем и альтернатива традиционным системам управления данными.

Техники и методы анализа, применимые к Big data по McKinsey:

- Data Mining;
- Краудсорсинг;
- Смешение и интеграция данных;
- Машинное обучение;
- Искусственные нейронные сети;
- Распознавание образов;
- Прогнозная аналитика;
- Имитационное моделирование;
- Пространственный анализ;
- Статистический анализ;
- Визуализация аналитических данных.

Горизонтальная масштабируемость, которая обеспечивает обработку данных — базовый принцип обработки больших данных. Данные распределены на вычислительные узлы, а обработка происходит без деградации производительности. McKinsey включил в контекст применимости также реляционные системы управления и Business Intelligence.

Технологии:

- NoSQL;
- MapReduce;
- Hadoop;
- R;
- Аппаратные решения.

Big data: применение и возможности

Объёмы неоднородной и быстро поступающей цифровой информации обработать традиционными инструментами невозможно. Сам анализ данных позволяет увидеть определённые и незаметные закономерности, которые не может увидеть человек. Это позволяет оптимизировать все сферы нашей жизни — от государственного управления до производства и телекоммуникаций.

Решения на основе Big data: «Сбербанк», «Билайн» и другие компании

У «Билайна» есть огромное количество данных об абонентах, которые они используют не только для работы с ними, но и для создания аналитических продуктов, вроде внешнего консалтинга или IPTV-аналитики. «Билайн» сегментировали базу и защитили клиентов от денежных махинаций и вирусов, использовав для хранения HDFS и Apache Spark, а для обработки данных — Rapidminer и Python.

Или вспомним «Сбербанк» с их старым кейсом под названием АС САФИ. Это система, которая анализирует фотографии для идентификации клиентов банка и предотвращает мошенничество. Система была внедрена ещё в 2014 году, в основе системы — сравнение фотографий из базы, которые попадают туда с веб-камер на стойках благодаря компьютерному зрению. Основа системы — биометрическая платформа. Благодаря этому, случаи мошенничества уменьшились в 10 раз.

Big data в Казахстане

4 марта 2020 года Глава государства Касым-Жомарт Токаев провел совещание по цифровизации, в ходе которого ему продемонстрировали информационно-аналитическую систему Smart Data Ukimet. На сегодняшний день уже реализовано 10 отдельных кейсов и подключена 41 информационная система госорганов. До конца года планируется увеличить их количество до 100 и автоматизировать получение и отображение всех ключевых социально-экономических показателей.

Совместно с государственными органами планируется продолжить работу по построению аналитических решений и охватить 15 направлений с реализацией около 50 отдельных кейсов.

«SmartDataUkimet — это проект, который позволит собирать информацию со всех ГО и из различных источников, тем самым даст уникальную возможность, применяя алгоритмы искусственного интеллекта, предоставлять аналитику и делать социально-экономические прогнозы для государства. А самое главное — позволит получать информацию без посредников. То есть, самую точную, напрямую и значительно быстрее», — отметил Асет Турысов, председатель правления АО «Национальные информационные технологии».

Отмечается, что аналитика больших данных является необходимым условием для развития и оптимизации экономики страны. Уже на данном этапе управленческие решения, принятые на основании аналитических данных, позволили достичь значительных результатов, как организационных, так и экономических. На текущий момент реализовано более 10 кейсов в сферах здравоохранения, финансов, образования, социальной защиты. Совокупный экономический эффект уже составил более 187 млрд тенге. [2]

Какие перспективы на рынке Казахстана?

По прогнозу IBM, к 2020 году для специалистов этого профиля откроются более 700 тыс. вакансий. На Западе трансформация уже началась, если сейчас термин Big data на территории СНГ является более популярным среди IT-специалистов, лидеры рынка начали адаптацию обучающими программами, тренингами для классических аналитиков. Кроме того, функционал традиционных должностей будет расширяться, сотрудники будут учиться грамотно работать с большими данными и, следовательно, пользоваться новыми преимуществами, которые дает эта технология. В итоге мы придём к тому, что абсолютно все сотрудники, не только представители IT-отделов, овладеют методами работы с Big Data.

К примеру, в следующем году 62% компаний планируют внедрять machine learning (машинное обучение) и основные методы анализа Big Data, следовательно, организациям нужно будет искать наиболее эффективные способы адаптации сотрудников к этим изменениям.

Технологии по обработке Big Data только заходят на отечественный рынок, на самом деле сейчас никто не может похвастаться крупными внедрениями

и результатами. Со стороны государства ведется большая работа по цифровизации, доказательства тому Цифровой Казахстан, smart city и прочие государственные программы. Барьером для зарубежных компаний является маленькое количество численности населения в стране, то есть качество, количество и сроки окупаемости внедрения сильно возрастают.

В Казахстане крупным заказчиком по обработке данных выступает государство, с периода независимости в стране собралось большое количество данных, их нужно обрабатывать и использовать, чтобы контролировать и конкурировать вне страны.

Одно из крупных внедрений будет происходить в министерстве финансов. Так как это направление стратегически важно для статистики и прогнозирования экономики и ВВП. В нашей стране размер неформальной экономики составляет 26%, в России — 39%, на Украине — 46%, а в Азербайджане и вовсе более 67% экономики находится в тени. Одной из стратегических задач является цивилизация малого и среднего бизнеса, для создания белого рынка. Для более точного анализа и быстрых результатов здесь не обойтись без анализа больших данных.

В Министерстве здравоохранения уже начали базовый уровень цивилизации. ЭПЗ (Электронный паспорт здоровья) позволит создать единую базу данных с историей медицинской карты. Если обратиться в больницу, то данные уже вводят в компьютер. Когда картотека данных будет полностью оцифрована, можно прогнозировать и улучшать работу врачей.

Сейчас база министерства образования и науки интегрирована с базами других госорганов на платформе eGov. Всего в общей сложности в министерстве имеется 73 госуслуги. 25 из них автоматизированы. Идет процесс внедрения НОБД (Национальная образовательная база данных), это подсистема СЭО (Система электронного обучения), предназначенная для автоматизации бизнес-процессов по сбору и обработке первичных статистических данных в сфере образования. В НОБД автоматизирован сбор данных для административных отчетов, заполнявшихся вручную и собиравшихся по цепочке: «организация образования — отдел образования — управление образования — МОН РК». Задачи: сбор ведомственной статистики от первоисточников (организаций образования) в автоматическом режиме; хранение и обработка данных; формирование административной отчетности; обеспечение структурных подразделений МОН РК необходимыми для работы статистическими данными. НОБД обеспечивает полный учет обучающихся; выявляет недостоверную информацию респондентов путем исключения дублирования; упрощает процедуру заполнения Паспортов организаций образования; формирует исторический ряд статистических данных; позволяет формировать нерегламентированные отчеты. Будет полностью внедрена до 2020 года.

Бизнесу также интересны большие данные, огромный интерес проявляют банки, да и в принципе любой крупный или средний бизнес, который хранит данные. Пример тому — «Казпочта» и «Казахтелеком», которые уже начали проявлять интерес.

Рынок по обработке больших данных очень перспективен, если смотреть в будущее. В стране отсутствуют специалисты, спрос на которых с каждым годом увеличивается. Средняя заработная плата на рынке Junior Data Scientist составляет от 200 000 тенге и доходит до 500 000 тенге в месяц [3].

Список использованных источников:

1. Просто Big Data. — СПб.: Страта, 2019. — 148 с.

2. В РК реализован Единый централизованный инструмент сбора, анализа и прогнозирования Big Data. URL: https://profit.kz/news/57188/V-RK-realizovan-Edinij-centralizovannij-instrument-sbora-analiza-i-prognozirovaniya-Big-Data/ (Дата обращения: 12.09.2020).

3. Ахметов С. Big Data в Казахстане: О крупном заказчике, кадрах и перспективах / Капитал, 09.08.2018. URL: https://kapital.kz/tehnology/71257/big-data-v-kazakhstane-o-krupnom-zakazchike-kadrakh-i-perspektivakh.html (Дата обращения: 12.09.2020).