**Министерство образования и науки РФ**
**Автономная некоммерческая организация высшего образования**
**Самарский университет государственного управления**
**«Международный институт рынка»**

**«Юридический факультет»**
**Кафедра «Государственного и муниципального управления**
**и правового обеспечения государственной службы»**
**Программа высшего образования**
**Направление подготовки «Государственное и муниципальное управление»**

**ДОПУСКАЕТСЯ К ЗАЩИТЕ**

Заведующий кафедрой:
*к.ю.н., доцент Березовский Д.В.*
_____
*(подпись)*

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА**

# «АНАЛИЗ И ИССЛЕДОВАНИЕ ПРОЦЕССА ПРИМЕНЕНИЯ ТЕХНОЛОГИЙ «БОЛЬШИЕ ДАННЫЕ» В ГОСУДАРСТВЕННОЙ ГРАЖДАНСКОЙ СЛУЖБЕ»

Выполнил:
*Самерханова Ж.Н., гр. ГМУ-43*
_____
*(подпись)*

Научный руководитель:
*к.т.н., доцент Никитина Т.В.*
_____
*(подпись)*

Оценка отлично

Самара

2017

# Оглавление

## Введение

Информационное общество, истоки которого лежат в начале 1950-х гг прошлого столетия, с началом нового тысячелетия вступает в новую, активную фазу своего развития. Проникновение информационных технологий во все сферы жизни общества является яркой характеристикой расцвета постиндустриального общества. Наиболее волнующей тенденцией последних лет является повсеместное и многократное увеличение объемов данных. Необходимость накопления и обработки больших объемов данных диктует цифровой индустрии вектор развития технологий хранения и построения эффективных алгоритмов обработки данных с использованием технологий интеллектуального анализа данных.

В секторе государственной гражданской службы объем входящих и обрабатываемых данных также увеличивается. Кроме того, с появлением новых возможностей data analytics, управление и работа с данными может выполняться эффективнее, быстрее и результативнее. На основе вышеизложенных фактов, целью данной выпускной квалификационной работы является – изучить состояние вопроса применения технологии Большие Данные в государственной гражданской службе с последующей разработкой рекомендаций по ее применению. В соответствии с заданной целью мной были поставлены следующие задачи:

1. Изучить теоретические основы технологий Большие Данные;
2. Проанализировать использование технологий Больших Данных в государственной гражданской службе Российской Федерации, зарубежный опыт в данной области;
3. Изучить инструментарий для использования технологий Большие Данные в государственной гражданской службе;
4. Выявить проблемные поля в области государственного регулирования использования технологий Большие Данные;
5. Разработать рекомендации по оптимизации обработки данных в государственной гражданской службе посредством применения технологий Большие Данные.

Объектом исследования являются технологии Большие Данные, предметом исследования: применение технологий Большие Данные в государственной гражданской службе.

Исходя из поставленных задач, в настоящей работе рассмотрены подходы к определению Больших данных, посредством обработки аутентичных зарубежных источников по научной дисциплине Data Science на языке оригинала (английский); приведен набор инструментов и основные параметры использования Больших Данных и их интеллектуального анализа; рассмотрены последние новации законодательства Российской Федерации в области цифровой экономики и защиты персональных данных, выявлены их слабые стороны и зоны роста для успешного прохождения технологий БД в сектор государственной гражданской службы; сформирован набор инициатив по внедрению технологии Большие Данные в сектор государственной гражданской службы, оптимизирующий работу ведомств, а также в качестве инструмента противодействия коррупции, детекции и контроля коррупционных правонарушений.

В работе использованы следующие методы научного исследования: анализ, синтез и моделирование.

# 1. Теоретические основы технологии Большие Данные: состояние, проблема, перспективы

## 1.1. Основные определения технологии Большие Данные

На сегодняшний день организации государственной гражданской службы собирают и обрабатывают огромное количество данных. В большинстве случаев эти данные структурированы. Для того, чтобы ввести понятие Большие Данные и продолжить изложение исследования, необходимо познакомиться с классификацией данных.

Первым классифицирующим признаком служит источник генерации данных ― человек или машина. Данные, созданные человеком, возникают при взаимодействии людей и цифровых или веб систем, таких как онлайн-сервисы, электронные устройства. Примером таких данных могут служить посты в социальных сетях, блогах, сообщения по электронной почте, обмен фото и видео посредством социальных сетей. Данные, созданные машиной, возникают при работе программного обеспечения и аппаратных устройств, и взаимодействия внешней, реальной среды. Например, обработка программой запроса о наличии товара, выбранного онлайн-покупателем; подтверждение транзакции. По идентичному признаку классифицируются данные, получаемые со счетчиков, GPS-систем, сетевых дневников.

Из вышесказанного следует, что данные могут быть получены через множество различных каналов и широкого спектра ресурсов. Кроме того, они могут быть представлены в различных форматах и типах. Ключевое значение в классификации данных в рамках концепции Большие Данные (далее – БД-концепция) имеет тип их внутренней организации (далее – тип). Итак, по типу данные разделяются на структурированные, неструктурированные, полу-структурированные, мета-данные. Рассмотрим подробнее каждый из них.

Структурированные данные чаще всего представлены в текстовом формате в таблицах, неотъемлемым признаком здесь является модель данных**,** соответствующая определенной схеме. Подобная модель используется для ввода

отношений между различными элементами предметной области базы данных, потому применение данной модели возможно при построении реляционной базы данных (см.параграф главу 2, параграф 2). Структурированные данные также возникают в работе корпоративных информационных систем таких как ERP (Планирование ресурсов предприятия) или CRM (Система взаимоотношения с клиентами), в основе которых лежит принцип создания единого хранилища данных, содержащего всю информацию о бизнес-структурах и процессах. Такой формат обработки данных изначально подразумевает соподчиненность, иерархичность с обязательным указанием отношений элементов данных, так как в дальнейшем они подлежат анализу, контролю и обработке. По выражению специалиста по теории и методам анализа данных компании O`Reilly Radar, всемирного лидера в области информационных технологий и Больших Данных, Ди Джея Патила, по совместительству заместителя руководителя Управления науки и техники США, главного специалиста по работе с Данными: «Если Вы не можете измерить процесс, Вам никогда не понять, как им можно управлять» [41]. Примером структурированных данных могут служить записи транзакций банка, оплаты по безналичному счету и другие табуляграммы.

Неструктурированные данные могут быть как текстовыми, так и аудио, видео формата. Отличительным признаком является здесь то, что обработка таких данных напрямую при помощи SQL невозможна. В реляционной базе данных такие аудио сведения или изображения размещаются в качестве BLOB (тип) – массива двоичных данных, предназначенного в том числе для хранения компилированного программного кода.

Полу-структурированные данные обладают определенной формой упорядоченности и систематичности, но связь между ними не установлена. Они выстроены в форме иерархии или расположены в структуре графа. Часто такие данные содержат текст. Примерами служат XML-файлы, JSON-файлы, данные, полученные от датчиков.

Наконец, мета-данные представляют собой информацию о структуре хранилища данных и его отдельных характеристиках. Этот тип данных в основном

создается машиной. В БД-концепции указанный выше тип обладает важным значением: данные явствуют о происхождении, размере, виде, источнике получаемых данных. Примеры мета-данных включают в себя информацию об авторе, дате создания документа; размеры файла и разрешение цифрового снимка.

Существует множество определений понятия Большие Данные–наиболее ёмкое из них описывается по формуле 5V–Volume, Velocity, Variety, Veracity, Value под которой подразумевается новый уровень обработки информации, предполагающий огромный объем данных, с высокой скоростью обновления, большим разнообразием, высокой степенью достоверности и несущее ценность. Но Большие Данные –не просто количественная характеристика гигантского стека информации. Для того, чтобы стать информацией, данные еще должны «вырасти», не только экстенсивно. [35] Прежде всего, применение БД-концепции — это возможность отвечать на различные вопросы, используя серии подходов, методов и инструментов обработки данных, базирующихся на математическом аппарате. Краткий список методов анализа структурированных и не структурированных данных в технологии Большие Данные включает:

1. Методы Data Mining, включающие в себя: обучение ассоциативным правилам, классификацию, кластерный анализ и регрессионный анализ;
2. Прогнозная аналитика;
3. Статистический анализ;
4. Визуализация аналитических данных.

**Концепция эволюции «работающего» знания**

Томас Давенпорт и Лоуренс Прусак, специалисты в области управления знаниями, в книге «Working Knowledge» [37], вышедшей в свет в 1998 году, определили понятия данные, информация, знание. Согласно определению, данные — это совокупность определённых действительных фактов о событиях. Однако, как подчеркивают авторы, данные сами по себе не могут быть основой для принятия управленческих решений по двум причинам. Во-первых, чрезмерно большое количество данных может сбить с толку, отягощая процесс определения, понимания ситуации и целеполагания. Во-вторых, данные сами по себе не содержат

**специфического** значения. Данные априори не несут собой вердиктов, интерпретаций и других заключений, могущих быть сколь-нибудь полезными в процессе принятия решений. Они также не сообщают о собственной значимости или релевантности. Однако они имеют ключевое значение, так как являются источником информации.

Информация, в свою очередь, по определению Давенпорта, ― это данные, способные существенно помочь делу (ориг. Information it is data that makes a difference) [37]. Информация ― это послание, способное оказать влияние на получателя, его суждения и поведение, а также способ восприятия. Данные «вырастают» в информацию под осуществлением над ними следующих операций:

1. **Определение контекста**. Подразумевает ответ на вопросы: «Для чего нам нужны эти данные? С какими намерениями мы добываем эти данные?»

2. **Категоризация**. Подразумевает определение элементов анализа и ключевых компонентов данных.

3. **Вычисления**. Имеется ввиду анализ данных методами математики и статистики.

4. **Доработка/корректировка**. Чистка шумов, удаление ошибок из имеющегося массива данных.

5. **Преобразование**. Преставление данных в доступной, выразительной форме, **доступной для пользователей.** [37]

Таким образом, при осуществлении работы над Большими Данными, задействованы все 5 вспомогательных способов, превращающих данные в информацию.

Наконец, на пути своей эволюции информация становится знанием (knowledge) под воздействием следующих **практик(манипуляций):**

1. **Сравнительный анализ.** Подразумевает ответ на вопрос: «Как полученная информация о ситуации соотносится с обыкновением/реалиями других ситуаций/исходов?»

2. **Определение зоны и степени влияния.** Отвечает на вопрос: «Как данная информация может повлиять на решения и действия получателя?»

3. **Выявление связей.** Определяет, как данный «кусок» информации связан с другими имеющимися сведениями.

4. **Конверсационный анализ.** Получение обратной связи: что думают люди о полученной информации? [37]

Знание развивается с течением времени, через полученный опыт, позволяющий оценить текущее положение с оглядкой на события и исходы, произошедшие в прошлом, помогая понять новые ситуации и явления. Достаточно позитивистским [23] выглядит заявление, что знание рождается из опыта, запоминает повторяющиеся схемы и проводит связи между тем, что происходит в настоящем и тем, что происходило в прошлом. Опыт трансформирует идеи от «что должно было бы случиться» до «что на самом деле случится». Однако именно поэтому знание обладает повышенной ценностью. Располагая большим количеством знаний, мы можем принимать более качественные решения, несмотря на то, что меньший объем знаний кажется более определенным и прозрачным. Определенность и ясность могут показаться более привлекательными, но эти преимущества стоят слишком дорого: очень часто прозрачность достигается в результате пренебрежения неотъемлемыми факторами.

В заключение хочется привести пирамиду эволюции данных [35], разработанной Томасом Эрлом, специалистом в области работы с Большими Данными:

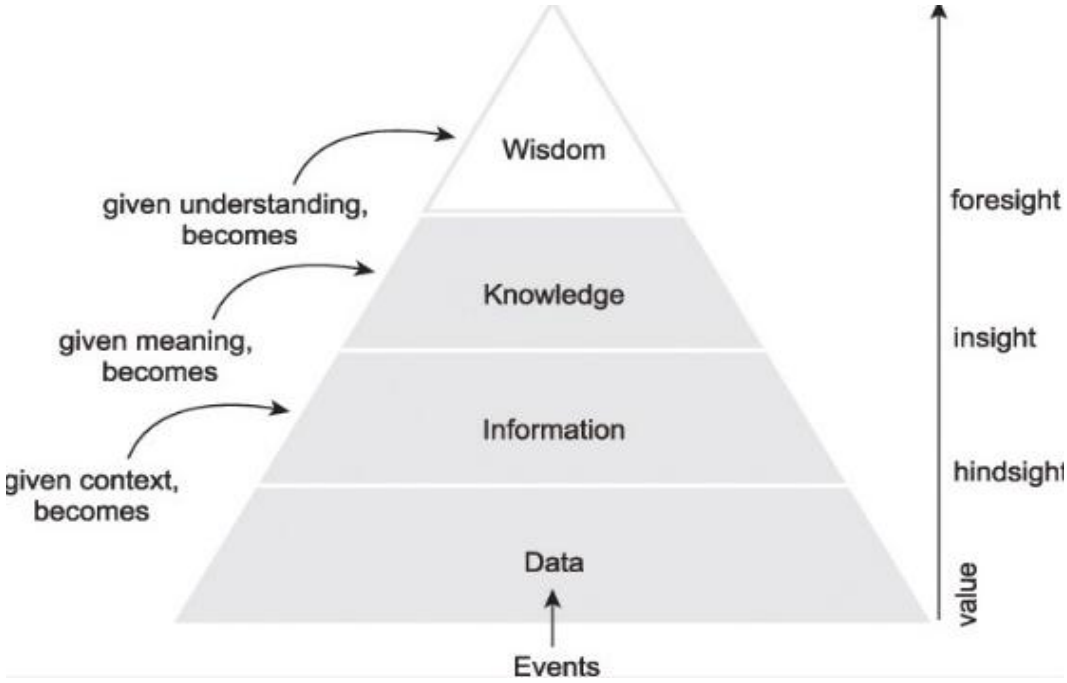


**Рисунок 1 - Эволюция Данных. Автор - Томас Эрл, 2015**

Данная иерархия служит подтверждением и продолжением концепции Давенпорта о получении «работающего» знания. Интересным является приведенная в книге [35] Т.Эрла иерархия с указанием связей и уровня действий:

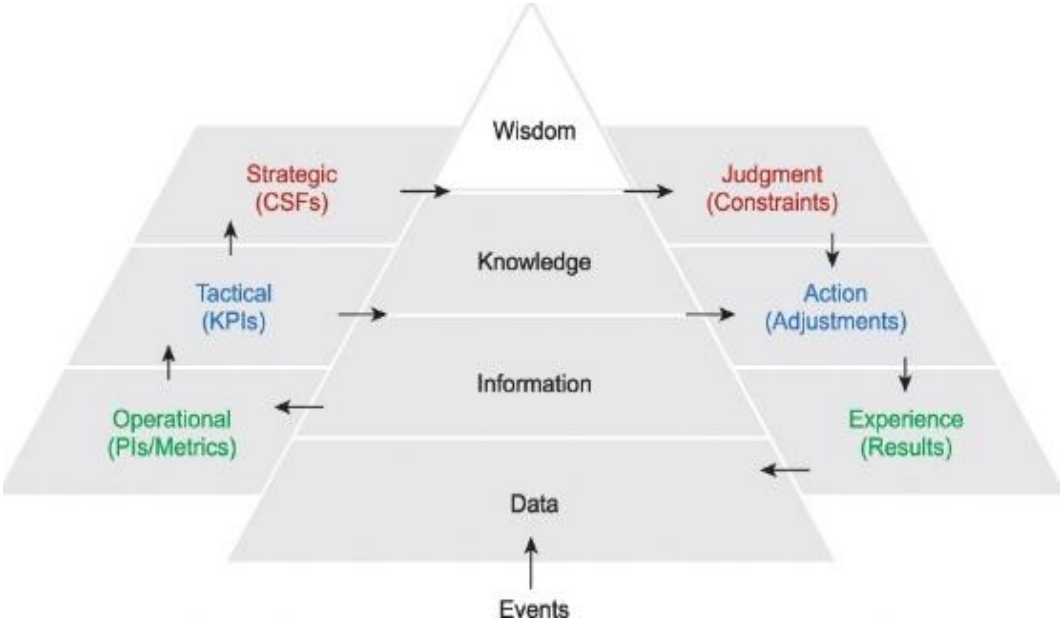


**Рисунок 2 - Эволюция Данных Т.Эрла**

Таким образом, сбор данных – а именно получение достоверных сведений (результатов) о происходящих событиях ― осуществляется на операционном уровне; далее, на тактическом уровне менеджмента производится действие ― вычисления и обработка данных, превращающих их в информацию. Информация на стратегическом уровне подвергается суждению и на своей высшей ступени развития соответствует статусу глубокого познания ― мудрости, способного стать надежным подспорьем при формировании и корректировке стратегии.

## 1.2. Зарубежный опыт использования технологии Большие Данные

На сайте британского Правительства 17 февраля 2017 года была опубликована статья «Использование технологии Большие Данные в Правительстве: вызовы и новые возможности» (ориг. – Big Data in Government: The challenges and opportunities»). Согласно данным этой статьи, Британская Академия и Королевское общество провели независимое расследование, посвященное теме Больших Данных, а именно: какие данные могут быть использованы Правительством и какие методы управления нужны для осуществления смелой идеи.

Джон Мандзони, глава Британского Правительства, в своем послании от 21 февраля 2017 года в свою очередь заявил, что Правительству Великобритании необходимо понять, что сбор, размещение и аналитика Больших данных является центральным звеном национальной инфраструктуры. [42] В качестве примера работы технологии он привел покупку книг, рекомендованных на сайтах, основанием рекомендации которых служила аналитика ранее приобретенных товаров. «Гораздо менее технология ныне была использована в секторе государственной гражданской службы. Настало время изменить это.» [42]. В основном, в докладе Главы Правительства Великобритании речь идет об использовании БД-концепции применительно к выявлению нужд граждан, а также обеспечение релевантных услуг. В интересующей автора тематике выдвигается идея создания «Частного сыщика», информирующего общественность о незаконной

собственности в оффшорных зонах. Это видится возможным благодаря использованию данных Кадастровой службы (Land Registry).

Данный факт напрямую говорит о том, что работа в кооперации с определенными органами государственной гражданской службы возможна и идея автора о применении интеллектуального анализа данных и использовании данных Россреестра, кадастровых служб не является фантасмогорией или очередной научной модой, не имеющей под собой реальных оснований.

Кроме того, утверждает мистер Мандзони, сервисы, основанные на открытых данных, позволят формировать рекомендации служб здравоохранения и других государственных учреждений. Он также сообщил, что в государственном секторе наблюдается нехватка специалистов, работающих в этой области, однако упомянул, что в США этот вопрос стоит еще острее: к 2018 году дефицит кадров составит 190,000 аналитиков Больших Данных (data scientists). Для устранения подобного дефицита в секторе государственной гражданской службы, власти Великобритании осуществляли финансирование тренингов для работников данной технической области и прикладной математики, занимались активным наймом и расширяли возможности карьерного роста для аналитиков данных. Программа, в рамках которой осуществлялась данная деятельность, носит название «Data Science Accelerator Programme». В рамках настоящей программы в 2016 году в городе Ньюпорт, административном центре графства Айл-оф-Уайт, был открыт кампус Науки о Данных. С октября прошлого (2016) года стартовала программа переподготовки по профилю Анализ Данных (Data Analytics), продолжительность курса — 2 года. По мнению Мандзони, специалист любого уровня государственной гражданской службы должен понимать всю полноту власти данных, именно поэтому была разработана программа направления Науки о Данных для непрофильных специалистов. Планируется, что ежегодно переподготовку в Академии будут проходить более 3000 человек.

Кроме того, Британский Парламент издал акт — Digital Economy Act, предусматривающий меры по созданию сервисов и инфраструктуры на основе электронно-технической связи. [40]

Другим важным документом является Стратегия преобразования Правительства — Government Transformation Strategy, опубликована 9 февраля 2017 года, далее — СПП [42] Автор подчеркивает, что Великобритания получила всемирное признание за создание сервиса GOV.UK и является одной из наиболее продвинутых стран в области электронного обеспечения государственных услуг. Сделав свой код открытым, Великобритания позволила десяткам других государств воспользоваться технологией и тем самым сделать работу Правительства прозрачной и более эффективной. СПП предусматривает 3 больших блока преобразования [42]:

**Таблица 1 –**
**Основные направления деятельности в рамках Стратегии Преобразования Правительства (Объединенного Королевства Великобритании и Северной Ирландии)**

| Блок | Пояснение/намерение |
|---|---|
| Изменение в работе сервисов для населения | Улучшать и облегчать процесс работы граждан, бизнес-отрасли и пользователей с государственным сектором; |
| Межведомственная трансформация | Довести до сведения органов государственной гражданской службы ориентиры и методы СПП, в этом фарватере улучшить электронное обслуживание граждан и повысить его эффективность; |
| Внутриправительственная трансформация | Не связана напрямую с улучшением работы сервисов для населения и результатами выполнения политики, но совершенно необходима для внутриправительственной кооперации и более эффективного процесса внедрения электронно-технического преобразования |

Автор считает необходимым подчеркнуть пункт «Извлекать максимум пользы из Данных» (ориг.— Make better use of Data), в котором указано, что Правительству необходимо разрушить границы между ведомствами при работе с

данными в той мере, которая позволит сделать отношения граждан и власти удобными, прозрачными. Иными словами, речь идет об объединении данных и применении интеллектуального анализа.

В рамках СПП предусмотрены следующие меры по достижению поставленных целей:

**Таблица 2 –**
**Меры по достижению целей Стратегии Преобразования Правительства (Transforming Government Strategy)**

| Блок | Пояснение/намерение | |
|---|---|---|
| Изменение в работе сервисов для населения | Улучшать и облегчать процесс работы граждан, бизнес-отрасли и пользователей с государственным сектором; | Изменение способа организации хранения данных Правительства и Правительственных организаций и методов управления настоящих данных; |
| Межведомственная трансформация | Довести до сведения органов государственной гражданской службы ориентиры и методы СПП, в этом фарватере улучшить электронное обслуживание граждан и повысить его эффективность; | Преодоление межведомственных барьеров для объединения данных и последующей эффективной работы с ними; |

| | | |
|---|---|---|
| Внутриправительственная трансформация | Не связана напрямую с улучшением работы сервисов для населения и результатами выполнения политики, но совершенно необходима для внутриправительственной кооперации и более эффективного процесса внедрения электронно-технического преобразования | Сделать Правительство открытым через дальнейшее использование сервисов и механизма загрузки и выгрузки данных с внутренних (Правительство) и внешних источников; Назначение Главного Должностного лица по работе с данными; создание национальной инфраструктуры регистров (базы данных) и обеспечение их безопасности |

Таким образом, с юридической точки зрения процессу внедрения БД-концепции сопутствует 2 профильных документа: Акт о Цифровой Экономике и Стратегия Преобразования Правительства, которые нивелируют границы между правительственными организациями с тем, чтобы создать единую для всех базу данных, либо устранить барьеры на пути объединения отдельных данных для последующей их обработки. Этическая сторона вопроса также рассматривается как в послании мистера Мандзони, так и в СПП. В частности, в мерах по достижению целей есть пункт, который гласит: использовать данные в соответствии с принципами безопасности и корректности, в целях обеспечения этики совместного использования данных, включая ответы на вопросы: что можно и что не следует использовать. [42]

## 2. Анализ использования технологий Большие Данные на примере Департамента по вопросам правопорядка и противодействия коррупции Самарской области

### 2.1. Разработка и использование технологий Большие Данные на примере Департамента по вопросам правопорядка и противодействия коррупции Самарской области

Департамент по вопросам правопорядка и противодействия коррупции Самарской области учрежден Правительством Самарской области Постановлением №506 от 8 октября 2012 года и является непосредственно подчиненным органом Губернатору Самарской области (Департамент управления Делами Губернатора Правительства Самарской области). Департамент является уполномоченным органом государственной власти Самарской области по реализации государственной политики в сфере противодействия коррупции, по профилактике коррупционных нарушений. [13] Департамент при осуществлении своей деятельности в пределах своей компетенции взаимодействует с Управлением Президента Российской Федерации по вопросам противодействия коррупции, территориальными органами федеральных органов исполнительной власти, органами государственной власти Самарской области, иными государственными органами, депутатами законодательных и представительных органов власти, органами местного самоуправления муниципальных образований Самарской области. Автор считает необходимым подчеркнуть, что материально-техническое и финансовое обеспечение деятельности Департамента осуществляется департаментом управления делами Губернатора Самарской области и Правительства Самарской области.[13] В рамках своих полномочий Департамент осуществляет контроль за соблюдением лицами, замещающими государственные должности Самарской области в Администрации Губернатора Самарской области, секретариате Правительства Самарской области, органах исполнительной власти Самарской области, государственными гражданскими служащими Самарской области, замещающими должности государственной гражданской службы Самарской области в

Администрации Губернатора Самарской области, секретариате Правительства Самарской области, органах исполнительной власти Самарской области и лицами, замещающими отдельные должности на основании трудового договора в организациях, созданных для выполнения задач, поставленных перед органами исполнительной власти Самарской области, запретов, ограничений и требований, установленных в целях противодействия коррупции, а также обеспечивает соблюдение государственными гражданскими служащими Самарской области требований законодательства Российской Федерации о контроле за расходами.

В рамках осуществления полномочий, указанных в статье 3 Постановления Правительства Самарской области №506 «О департаменте по вопросам правопорядка и противодействия коррупции», департамент в праве запрашивать в установленном порядке у территориальных органов федеральных органов исполнительной власти, иных органов государственной власти и государственных органов, органов местного самоуправления, организаций независимо от их организационно-правовых форм и форм собственности, а также их должностных лиц информацию, необходимую для выполнения возложенных на Департамент задач. [13] Одной из таких задач является обеспечение достоверности и полноты сведений о доходах, расходах, об имуществе и обязательствах имущественного характера, представленных лицами, замещающими государственные должности, и гражданскими служащими. Ниже представлена таблица основных полномочий Департамента в сфере противодействия коррупции:

**Таблица 3 –**

**Полномочия Департамента по вопросам правопорядка и противодействия коррупции Самарской области согласно Постановлению Правительства Самарской области №506**

| Постановление Правительства Самарской области №506 «О департаменте по вопросам правопорядка и противодействия коррупции» | Полномочия в сфере противодействия коррупции |
|---|---|
| Статья 1, п.1.4 | Взаимодействует с Управлением Президента Российской Федерации по вопросам противодействия коррупции, территориальными органами федеральных органов исполнительной власти, органами государственной власти Самарской области, иными государственными органами, депутатами законодательных и представительных органов власти, органами местного самоуправления муниципальных образований в Самарской области (далее - органы местного самоуправления) и их должностными лицами, организациями, общественными объединениями и гражданами. |
| Статья 2 пункт 2.1 | координация деятельности органов исполнительной власти Самарской области по противодействию |

| | |
|---|---|
| | коррупции; |
| Статья 2, пункт 2.1 | разработка и осуществление в пределах своей компетенции комплекса мероприятий, обеспечивающих реализацию государственной политики в сфере противодействия коррупции на территории Самарской области |
| Статья 2, пункт 2.1 | формирование у лиц, замещающих государственные должности Самарской области, государственных гражданских служащих Самарской области, муниципальных служащих и граждан нетерпимости к коррупционному поведению |
| Статья 2, пункт 2.1 | профилактика коррупционных правонарушений в Правительстве Самарской области, Администрации Губернатора Самарской области, секретариате Правительства Самарской области, органах исполнительной власти Самарской области, организациях, созданных для выполнения задач, поставленных перед органами исполнительной власти Самарской области; |
| Статья 2, пункт 2.1 | **осуществление контроля за соблюдением** лицами, замещающими государственные должности Самарской области в Администрации Губернатора |

| | |
|---|---|
| | Самарской области, секретариате Правительства Самарской области, органах исполнительной власти Самарской области, государственными гражданскими служащими Самарской области, замещающими должности государственной гражданской службы Самарской области в Администрации Губернатора Самарской области, секретариате Правительства Самарской области, органах исполнительной власти Самарской области (далее - лица, замещающие государственные должности, гражданские служащие), и лицами, замещающими отдельные должности на основании трудового договора в организациях, созданных для выполнения задач, поставленных перед органами исполнительной власти Самарской области, запретов, ограничений и требований, установленных в целях противодействия коррупции |
| Статья 2, пункт 2.1 | **обеспечение соблюдения государственными гражданскими служащими Самарской области требований законодательства** Российской Федерации о контроле за расходами, а также иных |

| | |
|---|---|
| | антикоррупционных норм |
| Статья 2, пункт 2.1 | разработка и реализация государственных программ Самарской области в сфере противодействия коррупции на территории Самарской области |
| Статья 2, пункт 2.1 | **осуществление контроля в пределах полномочий Губернатора и Правительства Самарской** области за деятельностью органов местного самоуправления в Самарской области, а также иных организаций и их должностных лиц в рамках предоставленных полномочий; |
| Статья 2, пункт 2.1 | **организация работы по совершенствованию системы контроля** в органах исполнительной власти Самарской области, Администрации Губернатора Самарской области, секретариате Правительства Самарской области, органах местного самоуправления |

Таким образом, использование Больших данных может стать великолепным подспорьем в работе департамента, так как позволяет осуществлять контроль соблюдения государственными гражданскими служащими Самарской области требований законодательства Российской Федерации о контроле за расходами [13], а также иных антикоррупционных норм путем прямой проверки интегрированных данных Росреестра и других Правительственных служб.

Как происходит проверка профилактики антикоррупционных нарушений сегодня?

Порядок установлен федеральным законом №273-ФЗ «О противодействии коррупции» [3]. Помимо этого, региональные исполнительные органы власти (в данной выпускной квалификационной работе процедура профилактики рассмотрена на примере работы Департамента по вопросам правопорядка и противодействия коррупции, далее – Департамента по вопросам правопорядка и СО) также используют памятку госслужащим. Отдел по обработке персональных данных руководствуется Конституцией Российской Федерацией, Постановлением Правительства Самарской области №506, Указом Президента Российской Федерации «О проверке достоверности и полноты сведений, представляемых гражданами, претендующими на замещение должностей федеральной государственной службы, и федеральными государственными служащими, и соблюдения федеральными государственными служащими требований к служебному поведению» №1065.

Для получения сведений, необходимых в рамках осуществления полномочий по обеспечению достоверности и полноты сведений о доходах, расходах, об имуществе и обязательствах имущественного характера, представленных лицами, замещающими государственные должности, и гражданскими служащими и последующего составления материалов проверки соответствия нормам служебного поведения, Отдел по обработке персональных данных Департамента по вопросам правопорядка и СО составляет и направляет запросы в органы Федеральной службы государственной регистрации, кадастра и картографии (далее – Росреестр). Для составления такого запроса требуются паспортные данные лица, а также соглашение на обработку персональных данных. Отдельно необходимо отметить подготовку писем о предоставлении этих данных. Таким образом, вся процедура занимает (рабочих дней): подготовка писем о предоставлении персональных данных, рассылка (5) + получение ответных писем (15), коррекция (5) + подготовка запросов в Росреестр (10) + получение ответов от Росреестра (20) + Обработка полученных результатов (10) = 40 (с учетом параллельности процессов получения ответных

писем, коррекции и подготовки запросов в Росреестр). Ниже приведена диаграмма о распределении рабочего времени на практике в Департаменте по вопросам Правопорядка и Противодействия коррупции:

**Таблица 4 –**

**Распределение рабочего времени и динамика работы в отделе обработки персональных данных Департамента по вопросам правопорядка и противодействия коррупции Самарской области**


За время прохождения практики 20 (25%) часов рабочего времени было потрачено на то, чтобы заполнить форму запроса в Росреестр персональными данными, процесс ожидания ответов занял 58% рабочего времени, в то время как при совместной работе с данными (data sharing) [20] этот процесс может занимать меньше 1 минуты.

По мнению автора, отдельно стоит рассмотреть процедуру проверки наличия конфликта интересов муниципальных служащих, замещающих должность руководителя на основании предоставленных документов о текущей кадровой структуре, включающей персональные данные. Проверка на предмет наличия конфликта интересов происходила методом визуального сравнения и состояла в

поиске наличия совпадений фамилий. Такая «рутинная» работа часто именуется «технической», но ни аппаратное устройство, ни программное обеспечение почему-то в данном случае не используются. Используется человек, который с большой вероятностью может «проглядеть» одинаковые фамилии. Хотя функция сравнения рядов является элементарной функцией MS Excel. И выполнять эту работу можно в данной программе, сводя к минимуму риск ошибки— человеческого фактора.

Очевидно, что наличие единой базы данных лиц, замещающих должности государственной и муниципальной службы, значительно облегчило бы работу, во всяком случае, при осуществлении проверки на предмет конфликта интересов, так как пресловутое изучение печатных данных также требует направления писем о предоставлении последних, что также требует определенных затрат человеко-часов и часов, которые, по мнению автора, необходимо наблюдать.

Однако, на пути создания единой БД существенным препятствием является положения 5 статьи Федерального закона от 27.07.2006 №152-ФЗ «О персональных данных»:

«… П.3: Не допускается объединение баз данных, содержащих персональные данные, обработка которых осуществляется в целях, несовместимых между собой.

П.4: Обработке подлежат только персональные данные, которые отвечают целям их обработки.

П.5: Содержание и объем обрабатываемых персональных данных должны соответствовать заявленным целям обработки. Обрабатываемые персональные данные не должны быть избыточными по отношению к заявленным целям их обработки.» [5]

Если в случае создания единой БД в целях обычного поиска и сверки подозрительных однофамильцев, п.3-5 не станут препятствием, то при развитии и внедрении технологии Большие Данные, а именно интеграции систем персональных данных, данных Росреестра, может возникнуть проблема уже на юридическом уровне. (Здесь мы не будем рассматривать проблемы технического толка, связанные с созданием такой системы).

Юридические коллизии возникают и при самом толковании Базы Данных. Так, в законе от 23.09.92 «О правовой охране программ для электронных вычислительных машин и баз данных» №3523-1 под базой данных понимается любая совокупность данных, которая может быть обработана с помощью электронно-вычислительного устройства [6], однако нельзя согласиться с тем, что подобное толкование полностью отражает широкие возможности информационной технологии, в особенности БД-концепции. Хотя данный закон был отменен введением в действие IV части Гражданского Кодекса РФ, определения базы данных не было дано в новом документе.

На сегодняшний день на территории Российской Федерации в сфере использования Больших Данных на федеральном уровне ведется работа по следующим направлениям:

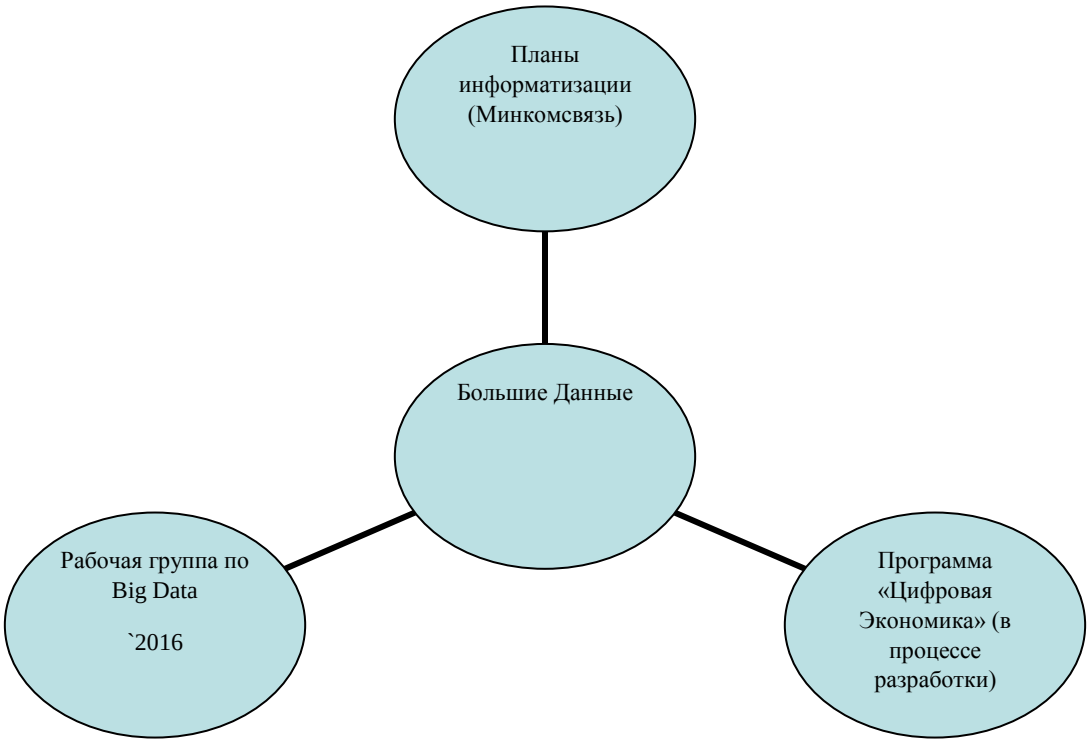


**Рисунок 3 - Законодательные инициативы и регулирование использования Больших Данных в России на современном этапе**

**Рабочая группа по Big Data** в Российской Федерации была создана по инициативе советника Президента Российской Федерации, руководителя компании LiveInternet, Германа Клименко, осенью 2016 года. О формировании рабочей группы для изучения вопроса государственного регулирования Больших Данных сообщил Институт Развития Интернета (далее – ИРИ) ― организация, целью создания и деятельности которой является развитие отрасли современных технологий. Деятельность группы осуществляется на базе Координационного центра Национального Домена сети Интернет. Руководителем группы назначен главный юридический советник ИРИ Сергей Копылов. Планируется осветить ряд вопросов, связанных с регулированием сбора и обработки Больших Данных, в том числе персональных данных. После первого заседания рабочей группы стало известно, что Минкомсвязь допустила возможность корректировки персональных данных, как следует из доклада заместителя Минкомсвязи Алексея Соколова. Категории данных, рассматриваемые на заседании:

**Таблица 5 –**
**Классификация данных по материалам заседания рабочей группы по Большим Данным**

| Категория | Описание проблемы |
|---|---|
| Персональные данные | Регулирование случаев, связанных с получением согласия субъекта персональных данных на их обезличивание, обработку или уничтожение |
| Обезличенные данные | Обезличивание без уничтожения: представляют наибольший интерес для бизнес структур и осуществления процессов таргетирования |
| Данные интернета вещей | Генерируются в результате |

| | |
|---|---|
| | распространения интернета вещей (информация с датчиков индивидуальных девайсов и средств связи) |

Из указанного списка совершенно очевидным является то, что классифицирующим признаком здесь является «тренд», так как согласно проведенному анализу профильной литературы по научной дисциплине Большие Данные, существующей в западной науке, перечисленные Данные классифицированы с различных фундаментальных позиций и единственным объединяющим фактором здесь является «необходимость» (регулирование персональных данных обязательно к пересмотру хотя бы потому, что одобрен и вступает в законную силу Пакет Яровой с 2018 года) и тренд (данные интернета вещей и обезличенные данные обладают потенциалом в BI, большинство представителей рабочей группы – гиганты информационной индустрии МТС, Яндекс, представители IBM). На первой встрече рабочей группы было высказано предложение о формировании единого понимания термина Большие Данные. По данным нам, открытым источникам, найти документ с единодушно одобренной формулировкой не удалось.

Все это позволяет нам заключить, что работа в данной области ведется вполне себе заинтересованными лицами, является много общей и разнузданной. В данной Выпускной квалификационной работе предлагается конкретное применение технологии и рассматриваются проблемы осуществления данной инициативы.

**Планы информатизации России —** профильная программа Российский Федерации по подготовке планов информатизации федеральных органов исполнительной власти и органов управления государственных внебюджетный фондов. Программа осуществляется в соответствии с Постановлением Правительства Российской Федерации №365 «О координации мероприятий по использованию информационно-коммуникационных технологий в деятельности государственных органов» от 24.05.2010, в котором утверждены правила подготовки

планов информатизации государственных органов и отчетов об их выполнении. Так, согласно Правилам, реализация программы осуществляется в два этапа. Согласно сведениям, размещенным на портале Министерства связи и массовых коммуникаций (далее ― Минкомсвязь России), на этапе, соответствующем 2017 году и текущему состоянию реализации, Программа исполняется на II этапе [12]. Это означает, что откорректированные проекты планов информатизации подлежат подготовке и согласованию в части финансирования и(или) перечня мероприятий по информатизации в соответствии с параметрами доведённых лимитов бюджетных обязательств. Ниже приведена таблица действия Программы планов информатизации [11]:

**Таблица 6 - Планы информатизации. Этапы реализации, цели, задачи**

| Этапы | Цели | Задачи |
|---|---|---|
| I этап<br>Подготовка, оценка предварительных проектов планов информатизации; | Обеспечение эффективного расходования средств федерального бюджета и государственных внебюджетных фондов (далее - бюджеты), направляемых на реализацию мероприятий по информатизациию; | Проекты планов информатизации государственных органов; |
| II этап<br>Подготовка, согласование откорректированных проектов планов информатизации; | Обеспечение единства и комплексности при планировании и реализации мероприятий по информатизации, осуществляемых государственными | Проекты государственных программ Российской Федерации, федеральных целевых программ, ведомственных целевых программ, стратегий, концепций и (или) иных |

| | | |
|---|---|---|
| | органами; повышение эффективности реализации мероприятий по информатизации за счет внедрения принципов проектного управления, а также за счет внедрения инструментов общественного контроля за реализацией мероприятий по информатизации; | документов, предусматривающих долгосрочные приоритеты и (или) мероприятия по информатизации, относящиеся к установленной сфере ведения государственных органов; |
| III этап<br>Подготовка, оценка, утверждение итоговых планов информатизации, уточненных (по необходимости) в части планового периода 2015 и 2016 годов и детализированных в отношении очередного 2014 года. | Многократное использование информационных систем, в том числе информационно-коммуникационных технологий, информационно-телекоммуникационной инфраструктуры, создаваемых за счет средств бюджетов; совместимость информационных систем, в том числе информационно-коммуникационных | Проекты федеральных законов, актов Президента Российской Федерации, актов Правительства Российской Федерации, актов государственных органов, в которых содержатся положения, регулирующие отношения в том числе по вопросам использования информационно-коммуникационных технологий, создания, развития, модернизации, эксплуатации информационных систем и |

| | | |
|---|---|---|
| | технологий, информационно-телекоммуникационной инфраструктуры, используемых в различных государственных органах. | информационно-телекоммуникационной инфраструктуры. |

Можем заключить, что идея объединения баз данных Росреестра, индивидуальных сведений лиц, замещающих должности государственной и муниципальной службы, а также данных транзакций вышеуказанных лиц, является не только смелой инициативой в области способов и мер противодействия коррупции, но также соответствует государственной политике в области информатизации органов государственной службы.

Ознакомившись с программой, автор предвосхищает упрек читателя: идея не свежа. Но не следует забывать о том, что программа носит общий характер, не имеет ссылки на конкретные ведомства и технологии. Кроме того, проверка показала [40], что органы государственной власти, включая Минкомсвязь России, не соблюдают сроки, установленные графиками подготовки и утверждения планов информатизации. Так, планы информатизации на 2014 г. утверждены с нарушением установленного срока: Минкомсвязи России – на 84 дня, Федерального казначейства – на 301 день. План Минфина России утвержден после окончания финансового года. Также планы информатизации на 2015 г.: Минкомсвязи России – на 14 дней, Минфина России – на 168 дней, Федерального казначейства – на 230 дней, о чем сказано на сайте пресс-центра Правительства Российской Федерации. Возможно, следует обратить пыл в нечто конкретное и обтекаемое, чем пытаться объять океан пламенеющих перспектив?

Кроме того, в ходе проверки выявлены случаи нарушения Минкомсвязью России сроков подготовки заключений на планы информатизации. Отчеты о

выполнении плана информатизации за 2014 год представлены только 31 органом государственной власти (37%).

**Программа цифровая экономика**. Согласно данным, размещенным на сайте Минкомсвязи, программа должна быть утверждена до 1 июня 2017 года. Однако, Савва Шипов, замминистр экономического развития, заявил в интервью ТАСС 3 июня [5], что работа над законом начнется «уже» этим летом. Налицо очередное несоблюдение сроков реализации. Планируется создание единого облачного хранилища всех баз данных госструктур к 2020 году, однако, на основе вышеизложенных фактов проверки несоблюдения сроков, мы можем предположить некоторые отклонения от заданного временного вектора. В очередной раз мы подчеркиваем, что крупные масштабы программ преобразования не только обладают повышенным риском неуспеха, но и являются коррупциогенными: величина выделяемых средств на крупный проект реформы соответствует ее масштабу, что создает дополнительные источники для махинаций. Мы же предлагаем применение конкретной технологии, в конкретной области и ведомстве, с определенными целями и результатами, прогнозным видением, методами и способами в купе с программой повышения квалификации и развития кадрового потенциала должностей государственной и муниципальной службы.

## 2.2. Техника и методология применения технологии Большие Данные

Первоначально предлагается создать реляционную базу данных чиновников ― лиц, замещающих должности государственной и муниципальной службы (далее РБД) для автоматизированной обработки персональных данных и, в перспективе, выставления индекса коррупциогенности, т.к. реляционная база данных может быть интегрирована с платформой Hadoop (речь о ней пойдет ниже) в сумме с другими источниками данных (неструктурированными). Для того, чтобы продолжить рассмотрение данного вопроса введем понятие Базы Данных и рассмотрим особенности формирования реляционной базы данных.

База данных – это поименованная совокупность взаимосвязанных данных, находящихся под управлением СУБД (Система управления баз данных). Ранее,

определение понятия «База данных» было декларировано в Законе РФ от 23.09.1992 N 3523-1 (ред. от 02.02.2006) «О правовой охране программ для электронных вычислительных машин и баз данных» и являло собой следующее: «База данных – это объективная форма представления и организации совокупности данных (например, статей, расчетов), систематизированных таким образом, чтобы эти данные могли быть найдены и обработаны с помощью ЭВМ» (ст. 1). [6] Однако после введения IV Части Гражданского Кодекса, закон утратил силу. Получается, что на сегодняшний день в законе не определено, что такое база данных, так как в IV части ГК РФ содержатся только положения об изготовителе Базы Данных:

1. Изготовителем базы данных признается лицо, организовавшее создание базы данных и работу по сбору, обработке и расположению составляющих ее материалов. При отсутствии доказательств иного изготовителем базы данных признается гражданин или юридическое лицо, имя или наименование которых указано обычным образом на экземпляре базы данных и (или) его упаковке;

2. Изготовителю базы данных принадлежат:

исключительное право изготовителя базы данных;

право на указание на экземплярах базы данных и (или) их упаковках своего имени или наименования;

право на обнародование базы данных;

право на указание на экземплярах базы данных и (или) их упаковке своего имени или наименования действует и охраняется в течение срока действия исключительного права изготовителя базы данных. [2]

В технической документации некоторых СУБД, а также в некоторых литературных источниках в состав БД включаются не только собственно хранимые данные о предметной области, но и описания БД. Более правильно описания баз данных считать самостоятельными компонентами Банков Данных (система специальным образом организованных данных (БД), программных, технических, языковых, организационно-методических средств, предназначенных для обеспечения централизованного накопления и коллективного многоцелевого использования данных- далее БНД), даже если они и хранятся вместе с самими

данными. Кроме того, в БНД могут присутствовать описания отдельных частей базы данных с точки зрения конкретных пользователей. Такое описание называется подсхемой. Кроме описания баз данных в состав метаинформации, хранимой в БнД, может включаться информация о предметной области, необходимая для проектирования автоматизированной информационной системы, о пользователях БНД, о проектных решениях и т.д. Централизованное хранилище метаинформации называется словарем данных или репозиторием. Роль словарной системы особенно возрастает при использовании средств автоматизированного проектирования информационных систем. Для большинства из них репозиторий является ядром всей системы. Также велика роль репозитория в распределенных системах. В некоторых системах, например, Access, под БД понимают совокупность разных объектов: таблиц, запросов, форм, отчетов, макросов и модулей, то есть понятие базы данных расширено и включает в себя практически все информационные компоненты, созданные для конкретного приложения. В других системах, в частности в Paradox, для обозначения подобной совокупности взаимосвязанных объектов используется понятие «семейство», что, очевидно, терминологически более правильно. При работе с конкретной системой надо, прежде всего, уточнить терминологию, используемую в ней. [19]

**Таблица 7 - Классификация организации данных**

| БД | СУБД | К БнД в целом |
|---|---|---|
| *По форме представления информации*: визуальные; аудио; мультимедиа | *По языкам общения*: открытые; замкнутые; смешанные | *По условиям предоставления услуг*: бесплатные; платные (бесприбыльные коммерческие) |
| *По характеру организации данных*: Неструктурированные; частично структурированные; структурированные (по типу используемой модели: иерархические, сетевые, реляционные, смешанные, мультимодельные) | *По числу уровней в архитектуре*: одноуровневые; двухуровневые; трехуровневые | *По характеру преобладающей обработки информации*: OLTP; OLAP |
| *По типу хранимой информации*: документальные (библиографические, реферативные, полнотекстовые); фактографические; лексикографические | *По выполняемым функциям*: информационные; операционные | *По степени доступности*: общедоступные; с ограниченным кругом пользователей |
| *По характеру организации хранения данных и обращения к ним*: локальные; общие; распределённые | *По сфере возможного применения*: универсальные; специализированные | *По охвату*: территориальные; временные; ведомственные; проблемные |
| *По способу задания метаинформации*: экстенсиональные; интенсиональные | *По «мощности»*: настольные; корпоративные | *По характеру взаимодействия с пользователями*: активные; пассивные |
| | *По ориентации на преобладающую категорию пользователей*: для разработчиков; для конечных пользователей | *По форме собственности*: государственные; негосударственные (частные, групповые, личные) |

Подробнее остановимся на характеристике реляционной базы данных и введем термин реляционной базы данных.

Реляционная база данных — совокупность взаимосвязанных таблиц, каждая из которых содержит информацию определённого типа. База «отношений» была разработана сотрудником компании IBM, математиком Эдгаром Коддом в 1970 году. Впоследствии он разработал 13 правил (в научных кругах известны как 12 правид Кодда – Codd`s 12 rules) для успешного создания и работы реляционной базы данных:

1. Основное правило (Foundation Rule)

Реляционная СУБД должна быть способна полностью управлять базой данных, используя связи между данными. Чтобы быть реляционной системой управления базами данных (СУБД), система должна использовать исключительно свои реляционные возможности для управления базой данных.

2. Явное представление данных (The Information Rule)

Информация должна быть представлена в виде данных, хранящихся в ячейках. Данные, хранящиеся в ячейках, должны быть атомарны. Порядок строк в реляционной таблице не должен влиять на смысл данных.

3. Гарантированный доступ к данным (Guaranteed Access Rule)

Доступ к данным должен быть свободен от двусмысленности. К каждому элементу данных должен быть гарантирован доступ с помощью комбинации имени таблицы, первичного ключа строки и имени столбца.

4. Полная обработка неизвестных значений (Systematic Treatment of Null Values): уеизвестные значения NULL, отличные от любого известного значения, должны поддерживаться для всех типов данных при выполнении любых операций. Например, для числовых данных неизвестные значения не должны рассматриваться как нули, а для символьных данных — как пустые строки.

5. Доступ к словарю данных в терминах реляционной модели (Active On-Line Catalog Based on the Relational Model)

Словарь данных должен сохраняться в форме реляционных таблиц, и СУБД должна поддерживать доступ к нему при помощи стандартных языковых средств, тех же самых, которые используются для работы с реляционными таблицами, содержащими пользовательские данные.

6. Полнота подмножества языка (Comprehensive Data Sublanguage Rule)

Система управления реляционными базами данных должна поддерживать хотя бы один реляционный язык, который

(а) имеет линейный синтаксис,

(б) может использоваться как интерактивно, так и в прикладных программах,

(в) поддерживает операции определения данных, определения представлений, манипулирования данными (интерактивные и программные), ограничители целостности, управления доступом и операции управления транзакциями (begin, commit и rollback).

7. Возможность модификации представлений (View Updating Rule)

Каждое представление должно поддерживать все операции манипулирования данными, которые поддерживают реляционные таблицы: операции выборки, вставки, модификации и удаления данных.

8. Наличие высокоуровневых операций управления данными (High-Level Insert, Update, and Delete)

Операции вставки, модификации и удаления данных должны поддерживаться не только по отношению к одной строке реляционной таблицы, но по отношению к любому множеству строк.

9. Физическая независимость данных (Physical Data Independence)

Приложения не должны зависеть от используемых способов хранения данных на носителях, от аппаратного обеспечения компьютеров, на которых находится реляционная база данных.

10. Логическая независимость данных (Logical Data Independence)

Представление данных в приложении не должно зависеть от структуры реляционных таблиц. Если в процессе нормализации одна реляционная таблица разделяется на две, представление должно обеспечить объединение этих данных, чтобы изменение структуры реляционных таблиц не сказывалось на работе приложений.

11. Независимость контроля целостности (Integrity Independence)

Вся информация, необходимая для поддержания целостности, должна находиться в словаре данных. Язык для работы с данными должен выполнять проверку входных данных и автоматически поддерживать целостность данных.

12. Дистрибутивная независимость (Distribution Independence)

База данных может быть распределённой, может находиться на нескольких компьютерах, и это не должно оказывать влияние на приложения. Перенос базы данных на другой компьютер не должен оказывать влияния на приложения.

13. Согласование языковых уровней (The Nonsubversion Rule)

Если используется низкоуровневый язык доступа к данным, он не должен игнорировать правила безопасности и правила целостности, которые поддерживаются языком более высокого уровня.

Для реляционной базы данных проектирование логической структуры заключается в том, чтобы разбить всю информацию по файлам — отношениям, таблицам, а также определить состав полей (атрибутов) для каждого из этих файлов. Определение ключа каждого из отношений также является задачей логического проектирования реляционной БД. Рассмотрим метод проектирования, основанный на анализе ER-модели [19, 152] и переходе от нее к реляционным отношениям. В основу этого метода положен эмпирический подход. Ниже описан алгоритм перехода от базовой ER-модели к схеме реляционной базы данных. Каждый элемент ER-модели находит свое отражение в схеме базы данных. Для некоторых ситуаций в ER-модели возможно использование нескольких альтернативных решений при их отображении в модель базы данных. Выбор наиболее подходящего решения будет зависеть от разных факторов, часть из которых отображена в ER-модели, а другая часть – нет, т.е. для выбора проектного решения кроме информации непосредственно из ER-модели необходимо дополнительно использовать информацию и из других компонент концептуальной модели предметной области. Любой из уникальных идентификаторов объекта является вероятным ключом полученного отношения. Если объект имеет несколько уникальных идентификаторов, необходимо один из них выбрать в качестве первичного ключа. Часто (но не обязательно) в качестве первичного ключа выбирается самый короткий из вероятных ключей.

Автор предполагает, что при организации реляционной базы данных (программа-максимум) будут задействованы базы данных Росреестра с тем, чтобы нивелировать бюрократические проволочки и получать информацию в режиме реального времени. Это возможно благодаря технологиям облачных вычислений. Под вопросом остается организация виртуального хранилища, взаимодействие ведомственных структур, и, безусловно, безопасность данных, находящемся к тому же под защитой Федерального закона №152-ФЗ «О персональных данных» [5].

При взаимодействии РБД возможно осуществление следующих операций:

1. Объединение таблиц с одинаковой структурой. Результат— общая таблица: сначала первая, затем вторая (конкатенация).

2. Пересечение таблиц с одинаковой структурой. Результат — выбираются те записи, которые находятся в обеих таблицах.

3. Вычитание таблиц с одинаковой структурой. Результат — выбираются те записи, которых нет в вычитаемом.

4. Выборка (горизонтальное подмножество). Результат — выбираются записи, отвечающие определенным условиям.

5. Проекция (вертикальное подмножество). Результат — отношение, содержащее часть полей из исходных таблиц.

6. Декартово произведение двух таблиц Записи результирующей таблицы получаются путем объединения каждой записи первой таблицы с каждой записью другой таблицы.

Преимущества реляционной базы данных очевидны: быстрые ответы на короткие запросы, независимость данных при изменении структуры, но существуют и недостатки, например, в области масштабирования данных, работы с поточными данными.

Размещение Данных. Облачная Инфраструктура.

Поскольку речь идет об облачной инфраструктуре, то задачи по контролю физического доступа являются обязанностью облачного провайдера. Необходимо убедиться воочию, как обеспечивается защита инфраструктуры. В качестве гарантии достоверности предлагается организация экскурсии с целью ознакомления со схемой и процедурами защиты инфраструктуры от неавторизованного доступа.

Положительный показатель достигнут, если провайдер демонстрирует машинные залы, инженерные и другие сервисные помещения, позволяя посмотреть на «кухню» изнутри.

Стоит отметить, что нужно доверять свои данные только дата-центрам, прошедшим сертификацию. Оценить состояние ЦОД возможно и самостоятельно, но на это может уйти очень много времени. Дополнительно следует изучить юридический статус дата-центра и узнать, есть ли у провайдера все необходимые государственные лицензии и контракты на поддержку систем в случае наступления аварийной ситуации. «Организации страдают от широкого и масштабного спектра

угроз, что, безусловно, является поводом для беспокойства. Успешно реализованные атаки оказывают непосредственное влияние на бизнес клиентов и носят деструктивный характер», – говорит представитель компании Arbor Networks Даррен Ансти (Darren Anstee).

Поэтому данные, передаваемые в облако провайдера, должны быть защищены. Важно знать, кто получил доступ к информации, какие операции были с ней проведены, с какого адреса пришел запрос. Все эти вопросы могут быть решены с помощью сервисов управления правами доступа к критическим данным. Также стоит задуматься о создании политики «самоуничтожения» для важной информации, которой не нужно существовать бесконечно долго за пределами корпоративного дата-центра.

Одним из способов защиты передаваемых данных является шифрование. Чтобы защитить информацию должным образом, стоит внедрить шифрование на каждом этапе жизненного цикла данных. Если приложения на мобильных корпоративных устройствах кэшируют данные, такой подход позволит предотвратить утечку в случае потери гаджета.

На рынке представлено множество решений для обеспечения шифрования данных в облаке. Например, одним из них является SecureCloud от Trend Micro. Система шифрует диски виртуальных машин с использованием ключей шифрования, хранимых в SecureCloud, который инициирует процессы шифрования/расшифровывания защищаемых модулей хранения. Архитектурно решение состоит из системы управления, предоставляемой как сервис с доступом через консоль и агентов, установленных на защищаемых виртуальных машинах.

Несколько других популярных сервисов для шифрования данных в облаке предложили пользователи Reddit.

Виртуальный выделенный сервер. Хостинг – это услуга, которая представляет владельцам сайтов вычислительные мощности, для физического хранения на них сайтов, с целью обеспечения постоянного доступа к ним. Хостинг представляет собой дата-центр, в котором установлены стойки или как их ещё называют шкафы. В стойках находится компьютерное оборудование – серверы.

Данные стойки напоминают шкаф с ящиками, где в каждом таком ящике находится отдельный сервер. VPS – это виртуальный выделенный сервер, который в отличие от виртуального хостинга предоставляет возможность самостоятельно администрировать ОС, устанавливать программное обеспечение исходя из потребностей ваших приложений, и арендовать для их работы необходимый объем серверных ресурсов (CPU, RAM, HDD) с возможностью последующего увеличения.

Несмотря на то, что безопасность облачных хранилищ часто принято считать слабой, недостатки физического сервера выравнивают счет. Физический сервер намного дороже, чем виртуальный сервер из-за ресурсов, необходимых для запуска и поддержания сервера. Физическими серверами в целом гораздо сложнее управлять. Это особенно актуально с восстановлением в случае сбоев. Так же, как и на любой другой машине, там будет день, когда из – за целый ряд причин сервер потерпит неудачу. В этих случаях, восстановление из резервных копий является настоящим кошмаром, поскольку сервер должен быть восстановлен с нуля на другой (новый) сервер, а затем данные должны быть восстановлены из резервных копий. Для критически важных производственных систем, это означает, по крайней мере, 8 или более часов простоя. Для предотвращения этого, компании создают кластеры из двух или более серверов, но, конечно, это только увеличит расходы.

Практически невозможно выполнить обновление сервера без дополнительных простоев. Кроме того, стоит отметить, что будущие обновления для выделенного сервера должны быть приняты во внимание при заказе сервера. В противном случае обновление может привести к созданию совершенно нового сервера. Вместо того, чтобы привести к незапланированной миграции услуг и, таким образом, незапланированных простоев сервиса.

Виртуальный частный сервер — легко масштабируемая облачная инфраструктура, построенная требованиям пользователя. VPC позволяет воспроизвести инфраструктуру произвольной сложности, функционируя как виртуальный дата-центр без дополнительных затрат на обслуживание и модернизацию оборудования и программного обеспечения.

Таким образом, существует 2 пути на организации обширной базы данных: реализация через виртуальный выделенный или виртуальный частный сервер с необходимой процедурой проверки безопасности провайдера.

Теперь рассмотрим о реляционной базе данных, ее размещении в хранилище (виртуальном) и непосредственно имеющим к этому отношение облачным вычислениям.

Виртуализация является ключевым компонентом технологий облачных вычислений; она позволяет абстрагироваться от физических деталей аппаратных средств и предоставляет виртуализированные ресурсы для высокоуровневых приложений.

Виртуализированный сервер обычно называется виртуальной машиной. Виртуальные машины позволяют изолировать приложения от обеспечивающих аппаратных средств и от других виртуальных машин. В идеальном случае любая виртуальная машина не воспринимается и не затрагивается другими виртуальными машинами, функционирующими на той же физической машине.

В принципе технологии виртуализации ресурсов добавляют гибкий перестраиваемый уровень программного обеспечения между приложениями и ресурсами, которые используются этими приложениями. Концепция, описывающая виртуализированный сервер баз данных, использует эти преимущества, особенно когда уровень базы данных существующего приложения, спроектированного для применения в обычном центре обработки данных, можно непосредственно портировать на виртуальные машины в публичном облаке.

Как правило, подобный процесс миграции требует лишь минимальных изменений в архитектуре и программном коде развернутого приложения. В подходе, основанном на виртуализированной базе данных, серверы баз данных (как и любые другие программные компоненты) для исполнения перемещаются на виртуальные машины. Хотя инициализация виртуальной машины для каждой копии базы данных порождает определенные издержки с точки зрения производительности, эти издержки оцениваются величиной менее 10%. На практике одно из основных преимуществ подхода на основе виртуализированной базы данных заключается в

том, что при необходимости приложение может иметь полный контроль над динамическим выделением и конфигурированием физических ресурсов уровня баз данных (серверов баз данных).

В результате приложения способны полностью использовать свойство эластичности облачной среды для достижения заранее заданных или настраиваемых целевых показателей масштабируемости или снижения затрат; однако для достижения этих целей требуется компонент для управления доступом, который отвечал бы за мониторинг состояния системы и за выполнение соответствующих действий (например, за выделение увеличенного/уменьшенного объема вычислительных ресурсов) согласно заданным требованиям приложений и стратегиям. Основные обязанности этого компонента состоят в принятии решений о том, когда активировать увеличение или уменьшение количества виртуализированных серверов баз данных, выделяемых программному приложению.

Во многих случаях центры обработки данных используются недостаточно вследствие таких причин, как избыточное выделение ресурсов и изменяющиеся по времени потребности в ресурсах со стороны типичных корпоративных приложений. Мультиаренда (Multi-tenancy) - это механизм оптимизации хостинговых сервисов, при котором несколько клиентов консолидируется в рамках одной операционной системы (на сервере исполняется единственный экземпляр программного обеспечения, обслуживающий несколько клиентов), благодаря чему экономия на масштабе помогает эффективно снижать стоимость вычислительной инфраструктуры.

В частности, мультиаренда позволяет объединять ресурсы в пул, что повышает коэффициент использования ресурсов благодаря избавлению от необходимости выделения ресурсов каждому арендатору в соответствии с его максимальной нагрузкой. Это делает мультиаренду привлекательным механизмом для следующих сторон:

1. Поставщики облачных сервисов (получают возможность обслужить больше клиентов уменьшенным количеством машин)

2. Потребители облачных сервисов (избавляются от обязанности оплачивать аренду всех ресурсов сервера).

База данных как сервис — это концепция, согласно которой сторонний поставщик осуществляет хостинг реляционной базы данных и предоставляет ее как сервис. Такие сервисы в значительной степени избавляют пользователей от необходимости приобретать дорогие аппаратные и программные средства, заниматься обновлениями программного обеспечения, а также привлекать специалистов для выполнения административных задач и технического обслуживания.

Как ожидается, реальная миграция любого приложения на уровне базы данных в сервис реляционной базы данных потребует минимальных усилий, если обеспечивающая реляционная СУБД для существующего приложения будет совместима с предложенным сервисом. Однако поставщик услуг по различным причинам может наложить определенные ограничения или создать те или иные препятствия (например, ограничения по максимальному размеру хостинговой базы данных, максимальному количеству возможных параллельных соединений и т. д.). Кроме того, программные приложения не обладают достаточной гибкостью для управления ресурсами, выделенными другим приложениям (например, для динамического выделения дополнительных ресурсов с целью обслуживания увеличивающейся рабочей нагрузки или для динамического уменьшения выделенных ресурсов с целью уменьшения операционных расходов). Весь процесс управления ресурсами и их распределения контролируется на стороне поставщика сервиса, что требует точного планирования распределяемых вычислительных ресурсов для уровня баз данных и ограничивает способность приложений заказчика в максимальной степени использовать потенциальные преимущества посредством задействования таких особенностей облачной среды, как эластичность и масштабируемость.

Системы управления реляционными базами данных на протяжении нескольких десятилетий рассматривались как универсальное решение для обеспечения сохранения и извлечения данных. Они достигли высокого уровня

зрелости в результате обширных научно-исследовательских усилий и породили большой успешный рынок и множество решений для различных областей бизнеса.

Постоянно растущая потребность в масштабируемости и в новых приложениях породила новые проблемы для традиционных реляционных СУБД, в т. ч. определенную неудовлетворенность результатами применения такого универсального решения в некоторых приложениях веб-масштаба. Ответом на эту проблему было новое поколение недорогого высокопроизводительного программного обеспечения управления базами данных, призванного преодолеть доминирование систем управления реляционными базами данных. Одна из причин перехода к NoSQL-решениям состоит в том, что различные реализации веб-приложений, корпоративных приложений и облачных приложений предъявляют различные требования к своим базам данных, - например, не каждому приложению требуется жесткая согласованность данных.

Еще один пример. Для веб-сайтов с большим объемом трафика, таких как eBay, Amazon, Twitter и Facebook, масштабируемость и высокая доступность ─ это важнейшие требования, которые должны соблюдаться в обязательном порядке. Для этих приложений даже малейшая остановка может иметь существенные финансовые последствия и отрицательно повлиять на доверие потребителей.

Теперь рассмотрим базовые принципы конструирования базы данных типа NoSQL.

По мнению Эрика Брюэра (Eric Brewer), одного из авторов теоремы CAP, тезис о возможности обеспечения не более чем двух из трех свойств может оказаться неверным по следующим причинам.

Разделы встречаются достаточно редко; нет необходимости поступаться согласованностью или доступностью, если система не разбита на разделы.

Выбор между согласованностью и доступностью может делаться в одной и той же системе множество раз (если рассматривать высокую степень гранулярности); подсистемы могут принимать разные решения, которые могут меняться в зависимости от затрагиваемых операций, данных или пользователей.

Эти три свойства существуют скорее в непрерывном состоянии, чем в бинарном. В действительности для результата важнее наличие необходимых уровней каждого свойства, чем полное отсутствие (0%) либо полное присутствие (100%) у системы какого-либо свойства.

Другими словами, теорема CAP лучше всего работает в том случае, когда вы принимаете во внимание нюансы каждого свойства и ориентируетесь на достижение для каждого свойства такого уровня, который необходим для получения заданного результата (исходя из существующих параметров). Э. Брюэр предлагает трехэтапную стратегию для обнаружения разделов и последующего учета их наличия:

Введение в явном виде режима partition mode (наличие разделов), который ограничивает некоторые операции;

Инициирование процесса восстановления, призванного восстанавливать согласованность и компенсировать ошибки, сделанные в процессе разбиения на разделы.

Перейдем к вопросу принципов конструирования баз данных типа NoSQL.

Известная теорема CAP (Consistency, Availability, Tolerance to Partitions ─ Согласованность, Доступность, Устойчивость к разделению) показывает, что распределенная система управления базами данных на практике способна обеспечить выполнение не более двух из трех указанных свойств. Большинство подобных систем жертвует требованием строгой согласованности (дополнительная информация по эволюции теоремы CAP приведена на врезке). В частности, в этих системах применяется политика ослабленной согласованности под названием согласованность в конечном счете (eventual consistency), которая гарантирует, что если к реплицированному объекту не будут применяться никакие новые обновления, то в конечном счете каждое обращение к этому объекту будет возвращать последнее обновленное значение. В отсутствие ошибок максимальный размер окна несогласованности может быть определен на основе таких факторов, как коммуникационные задержки, нагрузка на систему и количество копий, затрагиваемых схемой репликации.

Новые NoSQL-системы обладают несколькими общими конструктивными особенностями:

1. Возможность горизонтального масштабирования пропускной способности с охватом множества серверов.

2. Простой интерфейс/протокол уровня вызова (в отличие от SQL-связывания).

3. Поддержка более слабых моделей согласованности, чем ACID-транзакции в большинстве традиционных реляционных СУБД.

4. Эффективное использование распределенных индексов и оперативной памяти для хранения данных.

5. Возможность динамического описания новых атрибутов или схемы данных.

Перечисленные конструктивные особенности этих систем ориентированы в первую очередь на достижение следующих системных характеристик:

1. Доступность: Система должна быть доступной даже в ситуации отказа сети или отключения всего центра обработки данных.

2. Масштабируемость: Система должна быть в состоянии поддерживать очень большие базы данных с очень высокой частотой запросов при очень низкой задержке.

3. Эластичность: Система должна быть в состоянии удовлетворять меняющиеся требования к приложениям в обоих направлениях (увеличение масштаба или уменьшение масштаба). Кроме того, система должна быть способна корректно реагировать на эти меняющиеся требования и быстро восстанавливать свое устойчивое состояние.

4. Выравнивание нагрузки: Система должна быть в состоянии автоматически перемещать нагрузки между серверами с целью эффективного использования большей части аппаратных ресурсов и избежания любых ситуаций с перегрузкой ресурсов.

5. Отказоустойчивость: Система должна быть в состоянии учитывать тот факт, что даже самые потенциально редкие аппаратные проблемы могут однажды

воплотиться в реальность. Хотя аппаратный отказ остается серьезной проблемой, эта проблема должна решаться на архитектурном уровне базы данных, а не требовать привлечения разработчиков, администраторов и техников с целью создания собственных резервированных решений.

Таким образом, СУБД и облачные вычисления позволяют создавать и размещать в виртуальном хранилище данные больших объемов и осуществлять их эффективную обработку с учетом безопасности хранения данных.

# 3. Разработка рекомендаций по применению технологии Большие Данные на государственной гражданской службе

## 3.1. Программа—минимум по внедрению технологии Большие Данные в Департаменте по вопросам правопорядка и противодействия коррупции

Теперь, когда введено понятие Данных и рассмотрено современное понимание их эволюции, рассмотрены способы систематизации данных и их обработки, проведена исследовательская работа по месту прохождения практики, автор считает необходимым предложить возможные способы улучшения эффективности работы структур исполнительной власти по противодействию коррупции Российской Федерации, на примере Департамента по вопросам правопорядка и противодействия коррупции Самарской области в части полномочий по реализации в государственных учреждениях Самарской области и организациях, созданных для выполнения задач, поставленных органами исполнительной власти Самарской области, мер по профилактике коррупционных правонарушений и последующего осуществления анализа сведений о доходах, об имуществе и обязательствах имущественного характера, представленных гражданами, претендующими на замещение должностей государственной гражданской службы Самарской области, о доходах, расходах, об имуществе и обязательствах имущественного характера, представленных гражданскими служащими, а также о соблюдении гражданскими служащими запретов, ограничений и требований, установленных в целях противодействия коррупции [13]

Как было сказано ранее, предполагается создание реляционной базы данных отчетных сведений государственных гражданских служащих. Однако, для это необходимо утверждение легальности сбора и цифровой обработки персональных данных, и структурированный формат подачи данных;

В качестве программы минимум — то есть осуществления детекции конфликта интересов — предполагается создание реляционной таблицы, записи которой будут содержать занимаемые должности лиц, а атрибутами выступать паспортные данные государственного или муниципального служащего (ФИО),

место жительства, место регистрации, сведения о доходах и расходах и составе семьи. Благодаря операции пересечения таблиц станет возможным автоматизированная проверка наличия конфликта интересов, а именно детекция привычного в муниципальной практике нарушения ― подчинение лиц, находящихся в близком родстве или свойстве. Конфликт интересов, согласно 10 статье Федерального закона №273-ФЗ трактуется как: «Ситуация, при которой личная заинтересованность (прямая или косвенная) лица, замещающего должность, замещение которой предусматривает обязанность принимать меры по предотвращению и урегулированию конфликта интересов, влияет или может повлиять на надлежащее, объективное и беспристрастное исполнение им должностных (служебных) обязанностей (осуществление полномочий», при этом в пункте указано: «В части 1 настоящей статьи под личной заинтересованностью понимается возможность получения доходов в виде денег, иного имущества, в том числе имущественных прав, услуг имущественного характера, результатов выполненных работ или каких-либо выгод (преимуществ) лицом, указанным в части 1 настоящей статьи, и (или) состоящими с ним в близком родстве или свойстве лицами (родителями, супругами, детьми, братьями, сестрами, а также братьями, сестрами, родителями, детьми супругов и супругами детей), гражданами или организациями, с которыми лицо, указанное в части 1 настоящей статьи, и (или) лица, состоящие с ним в близком родстве или свойстве, связаны имущественными, корпоративными или иными близкими отношениями.» [3]. На территории Самарской области расположено **341 муниципальных образований.** Каждое из них подлежит проверке, описанной по алгоритму, рассмотренному в главе 2 параграфе 1. В случае внедрения такой системы Отдел по обработке персональных данных претендует на развитие авангардной системы интеллектуального анализа данных. В рамках данной выпускной квалификационной работы предлагается также внедрение РБД, соответствующей «программе-максимум». Если в первом случае мы говорим о структурированных данных, то здесь речь пойдет о полу структурированных и вовсе неструктурированных данных, обладающих, во-первых, большим размером, во-

вторых, большим потенциалом, что полностью вписывается в концепцию Больших Данных.

На сегодняшний использование технологии Большие Данные во-многом затруднено отсутствием эффективных алгоритмов их обработки. Важно уточнить, что данные становятся Большими, если их больше 1 Пб ($2^{50}$ Б). В случае, если работа осуществляется с BLOB, изображениями, данными GPS, безусловно, это утяжеляет объем данных. Одним из ключевых компонентов является организация хранилища данных. Необходимо решить, будет ли это классический мейнфрейм ─ физический сервер или же будет организовано облачное хранение (например, использование облачных сервисов Яндекс, Amazon EC).

Рассмотрим проблему организации хранилища данных.

В случае, если при выборе облачного хранилища мы используем Hadoop, можно рассчитывать на широкую линейную масштабируемость (что актуально, учитывая ПМ и получение данных GPS-навигации, транзакций, получение и анализ ad hoc). При сравнительно невысокой стоимости, весомым аргументом против может служить то, что физические серверы Hadoop располагаются не на территории России. К тому же анализ возможен лишь при построении MapReduce процедур, что сложнее построения запросов с помощью обычного SQL.

Teradata намного лучше справляется с грубым методом работы с данными, таким как full scan. У Teradata идеология shared-nothing и она очень похожа на Hadoop/NoSQL. Данные лежат на множестве серверов, каждый сервер обрабатывает свою часть сам. Но у Teradata есть существенный недостаток — довольно бедный инструментарий. Цена, сервер Teradata и Exadata Full Rack стоят примерно одинаково, $5 млн.

Таким образом, если будет разрешена проблема с datawarehouse, следующим ключевым моментом является разработка алгоритма обработки данных. Здесь мы можем обратиться к опыту компании Wolfram. Но сначала обратимся к основам ─ язык SQL.

Язык SQL стал фактически стандартным языком доступа к базам данных. Все СУБД, претендующие на название «реляционные», реализуют тот или иной

диалект SQL. Многие нереляционные системы также имеют в настоящее время средства доступа к реляционным данным. Целью стандартизации является переносимость приложений между различными СУБД.

Нужно заметить, что в настоящее время, ни одна система не реализует стандарт SQL в полном объеме. Кроме того, во всех диалектах языка имеются возможности, не являющиеся стандартными. Таким образом, можно сказать, что каждый диалект - это надмножество некоторого подмножества стандарта SQL. Это затрудняет переносимость приложений, разработанных для одних СУБД в другие СУБД.

Язык SQL оперирует терминами, несколько отличающимися от терминов реляционной теории, например, вместо «отношений» используются «таблицы», вместо «кортежей» - «строки», вместо «атрибутов» - «колонки» или «столбцы».

Стандарт языка SQL, хотя и основан на реляционной теории, но во многих местах отходит он нее. Например, отношение в реляционной модели данных не допускает наличия одинаковых кортежей, а таблицы в терминологии SQL могут иметь одинаковые строки. Имеются и другие отличия.

Язык SQL является реляционно полным. Это означает, что любой оператор реляционной алгебры может быть выражен подходящим оператором SQL. Однако для работы с не/полуструктурированными данными и метаданными [32], необходимо задействовать NoSQL, так как возникает проблема масштабируемости и доступности данных. Для размещения массива данных с помощью NoSQL в Hadoop, платформе, позволяющей работать с распределенной базой данных больших размеров (ЭБ, ПБ) применяется алгоритм MapReduce.

Очевидно, что существует множество технологических и технических решений, однако все они являются западными, нероссийскими разработками. Одним из немногих признанных эффективных отечественных методов является PolyAnalyst.

Таким образом, для обеспечения безопасности необходимо способствовать развитию отечественной индустрии Данных, с тем, чтобы:

1. Вопросы их хранения и размещения могли решаться в пределах Российской Федерации;

2. Алгоритмы обработки данных находились также в интеллектуальной собственности гражданина/граждан Российской Федерации, с исключительным правом пользования;

3. Существовала специально созданная СУБД на основе решений, а также технологии облачных хранилищ и вычислений, с физическими серверами на территории Российской Федерации, что обеспечит максимальную безопасность работы с данными.

## 3.2. Программа-максимум — основные направления работы с персоналом в сфере технологии Большие Данные

Так как автор исследовав тему анализа применения БД-концепции, интеллектуального анализа данных находит применение технологий Большие Данные, методы Data Analytics и Data Mining наиболее перспективными и эффективными в области работы с данными в современных условиях информационного общества и нарастающего накопления данных; выяснив также, что применение технологий Большие Данные является абсолютной новацией в государственном секторе не только Российской Федерации, но и всего мира; предлагается разработать стратегию применения и работы с Большими Данными; учредить программу повышения квалификации для государственных гражданских служащих в сфере Big Data, Data-Mining, а также Анализа Больших Массивов Данных, искоренив безграмотность гордого Государственного Служащего Российской Федерации в сфере информационных технологий, цифровых и электронных коммуникаций.

Как говорилось ранее, планируется создание единого облачного хранилища всех баз данных госструктур к 2020 году, однако, на основе фактов проверки несоблюдения сроков, мы можем предположить некоторые отклонения от заданного временного вектора. В очередной раз мы подчеркиваем, что крупные масштабы

программ преобразования не только обладают повышенным риском неуспеха, но и являются коррупциогенными: величина выделяемых средств на крупный проект реформы соответствует ее масштабу, что создает дополнительные источники для махинаций. Мы же предлагаем применение конкретной технологии, в конкретной области и ведомстве, с определенными целями и результатами, прогнозным видением, методами и способами в купе с программой повышения квалификации и развития кадрового потенциала должностей государственной и муниципальной службы, а именно:

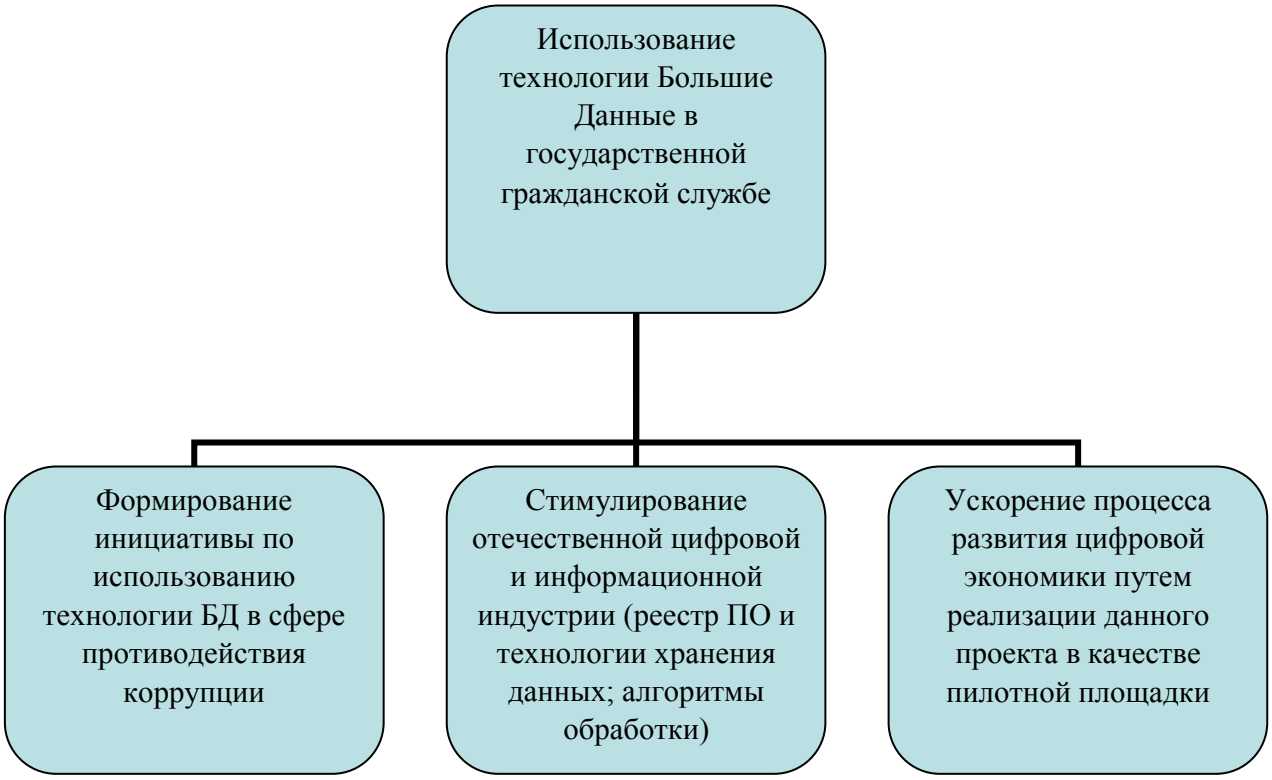


**Рисунок 4 - Внедрение технологий Большие Данные в ГГС**

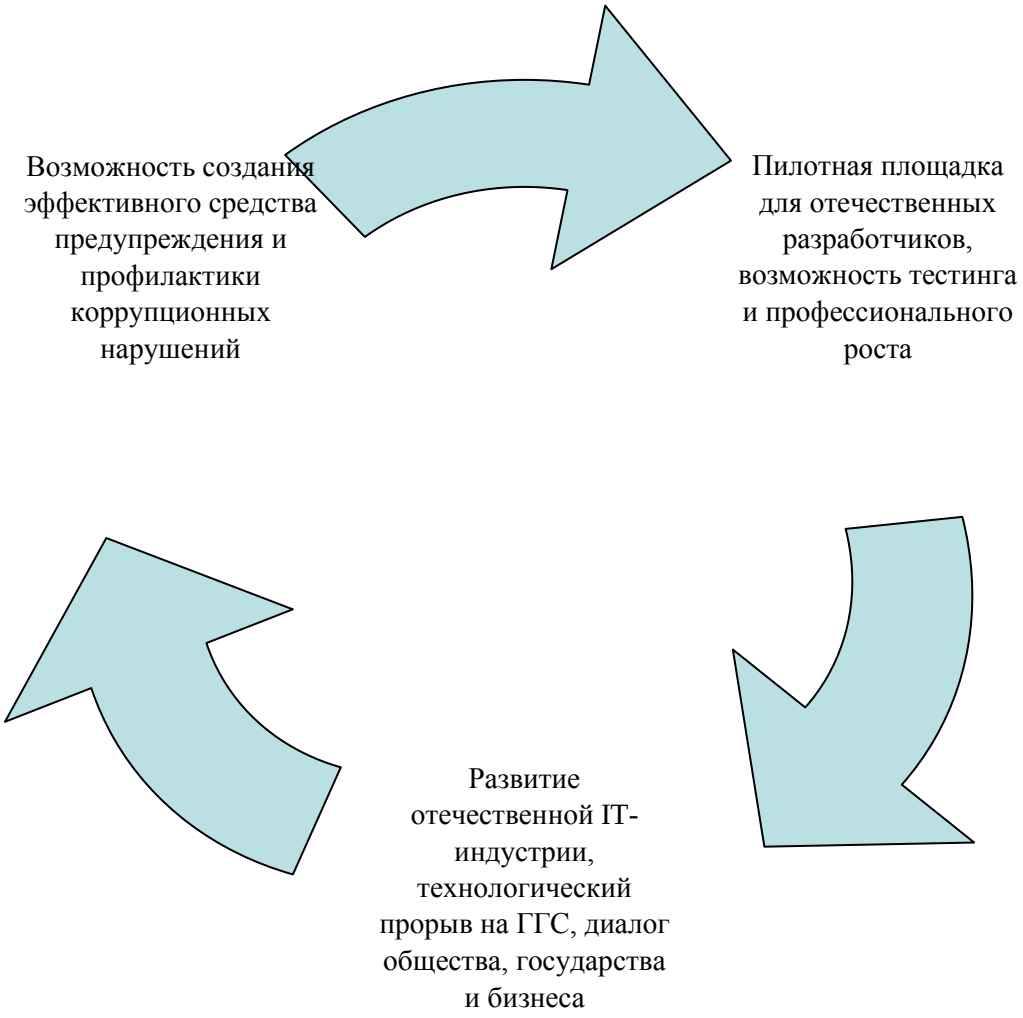


**Рисунок 5 - Перспективы запуска проекта Большие Данные в государственной гражданской службе**

В качестве программы-максимум предлагается формировать базу данных, содержащую индивидуальные (персональные данные) сведения лиц, замещающих должности государственной и муниципальной службы; сведения базы данных Россреестра об имеющемся имуществе и обязательствах имущественного характера; информация о счетах и вкладах в режиме реального времени; метки присутствия GPS соответствующие локации зарегистрированным на территории Российской Федерации учреждений, осуществляющих финансовую деятельность ─ банков.

Кроме того, для повышения грамотности государственных гражданских служащих и муниципальных служащих в области Data Analytics, Data Mining

предлагается разработать программу повышения квалификации и/или переподготовки. Сделать это можно на базе существующих образовательных учреждений Высшего образования, специализирующихся в данной области или обладающих образовательными ресурсами в данной области (кафедры прикладной математики и статистики) (НИУ ВШЭ-ГУ; МФТИ; МИЭМ; Самарский Университет). Возможно, перспективным является создания Центра Науки о Данных по опыту Великобритании. В таком случае подготовку смогут проходить не только государственные и гражданские служащие, но и специалисты других областей.

**Заключение**

Технология «Большие Данные» успешно применяется в маркетинге и в различных сферах бизнеса. Однако «Большие Данные» обладают потенциалом применения не только в коммерческом секторе, но и в государственной службе. В процессе накопления и увеличения все больших объемов данных, нарастающей информационной деструктуризации, представляется актуальным не только вопрос о создании высокоустойчивых баз данных, но также возможности интеллектуального анализа. Предлагается инновационный подход в работе с отчетными данными.

В рамках закона №273-ФЗ «О противодействии коррупции» государственные гражданские служащие обязаны предоставлять сведения о личных доходах, об имуществе и обязательствах имущественного характера и расходах. Численность государственных гражданских служащих в Российской Федерации составляет 715,9 тыс. человек. Очевидно, что поступающие сведения исчисляются в увесистом цифровом эквиваленте. Проблема заключается в том, что согласно пункту 3 статьи 8 Федерального Закона «О противодействии коррупции» №273-ФЗ сведения о доходах, об имуществе и обязательствах имущественного характера, относятся к информации ограниченного доступа. Несмотря на то, что данные сведения подлежат опубликованию и находятся в открытом доступе, создание единой базы данных не представляется возможным, так как существует закон о защите персональных данных (при одновременной обязанности компаний Мегафон, МТС, Билайн и В Контакте хранить персональные данные пользователей). Таким образом, открытым остается вопрос о законности интеллектуального анализа указанных данных, которые, по нашему мнению, могут быть эффективно проанализированы технологиями «Большие Данные».

Предложения:

- Разработать стратегию использования Больших Данных и ее применение в государственном секторе;
- Повысить грамотность государственных гражданских служащих в сфере интеллектуального анализа данных, извлечения полезных сведений из Данных и

современных методов обработки и анализа Данных путем учреждения программы переподготовки или повышения квалификации государственных гражданских служащих;

- Нормативно-правовым актом устранить препятствия (упростить процесс получения сведений одного ведомства другому) на пути обмена Данными между Правительственными организациями с целью объединения усилий и создания рабочих моделей по сбору и обработке данных;
- создать единую облачную базу данных отчетных сведений государственных гражданских служащих и муниципальных служащих Самарской области, законодательно устранив препятствия на пути свободного обмена данными между правительственными службами;
- расширить указанный перечень отчетных данных (GPS метки, информация банковских операций) с целью получения дополнительных сведений для интеллектуального анализа;
- разработать алгоритм обработки данных, структурированного, полуструктурированного и неструктурированного типа, позволяющий выявлять коррупциогенность государственных гражданских служащих ― выставлять индекс коррупциогенности; проводить проверку на соблюдение/несоблюдение требований к служебному поведению;
- способствовать развитию отечественной индустрии Данных, обеспечивать поддержку развития Науки о данных в образовательной и научно-прикладной среде (НИИ, НИОКР, образовательные учреждения).

**Список использованной литературы**

**Нормативно-правовые акты**

1. Конституция Российской Федерации (принята на всенародном голосовании 12.12. 1993) // Российская газета. 1993. 25 декабря.

2. Гражданский Кодекс Российской Федерации: Федеральный Закон от 30.11.1994 №51-ФЗ //СПС «КонсультантПлюс»

3. «О противодействии коррупции»: Федеральный закон от 22.12.2008, № 273 // СПС «КонсультантПлюс»

4. «О необходимости представления депутатами справок о доходах, об имуществе и обязательствах имущественного характера»: Федеральный закон от 03.11.2015, № 303 // СПС «КонсультантПлюс»

5. «О персональных данных»: Федеральный закон от 27.07.2006 №152 // СПС «КонсультантПлюс»

6. «О правовой охране программ для электронных вычислительных машин и баз данных»: Закон от 23.09.92, №3523-1 // СПС «КонсультантПлюс»

7. «О Национальном плане по борьбе с коррупцией»: Указ Президента Российской Федерации от 1.04.2016, №147 // СПС «Гарант»

8. «О проверке достоверности и полноты сведений, представляемых гражданами, претендующими на замещение должностей федеральной государственной службы, и федеральными государственными служащими, и соблюдения федеральными государственными служащими требований к служебному поведению»: Указ Президента Российской Федерации №1065 от 21.09.2009 // СПС «КонсультантПлюс»

9. «О координации мероприятий по использованию информационно-коммуникационных технологий в деятельности государственных органов»: Постановление Правительства Российской Федерации от 24.05.2010, №365 // СПС «КонсультантПлюс»

10. «Об утверждении Правил подготовки заключений об оценке мероприятий по информатизации и проектов планов информатизации федеральных органов

исполнительной власти и органов управления государственными внебюджетными фондами»: Приказ Министерства связи и массовых коммуникаций от 11.08.2016, №371 // СПС «КонсультантПлюс»

11. «Об утверждении Правил подготовки плана информатизации Министерства связи и массовых коммуникаций Российской Федерации»: Приказ Министерства связи и массовых коммуникаций от 23.06.2016, №282 // СПС «КонсультантПлюс»

12. «Об утверждении плана информатизации Министерства связи и массовых коммуникаций Российской Федерации на 2017 год и плановый период 2018 и 2019 годов»: Приказ Министерства связи и массовых коммуникаций от 25.04.2017, №204 // СПС «КонсультантПлюс»

13. «О департаменте по вопросам правопорядка и противодействия коррупции Самарской области»: Постановление Правительства Самарской области от 08.10.2012, №506 // СПС «КонсультантПлюс»

**Литература**

14. Айвазян С.А. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика. 2009. 607 с.

15. Барсегян А. Методы и модели анализа данных: OLAP и Data Mining. БХВ –Петербург. 2004. 336 с.

16. Виктор Гольцман. MySQL 5.0. Изд-во: Питер. 2009. 253 с.

17. Дак Джини Даниэль Монстр перемен. Причины успеха и провала организационных преобразований. Альпина Бизнес Букс. 2007. 389 с.

18. Дейт К.Дж. Введение в системы баз данных. СПб.: Издательский дом «Вильямс». 2000. 1328 с.

19. Диго С.М. Базы Данных. Проектирование и создание. М.:Изд. Центр ЕАОИ. 2008. 171 с.

20. Дюк В.А. Data Mining: Учебный курс/ В.А. Дюк, А.П. Самойленко. СПб: Питер. 2001. 368 с.

21. Инновационный менеджмент. Учебник для академического бакалавриата. М.: Юрайт. 2014. 640 с.

22. Когаловский М. Р. Энциклопедия технологий баз данных. М.: «Финансы и статистика». 2008. 800 с.

23. Конт Огюст Дух позитивной философии: Слово о положительном мышлении. УРСС. 2011. 80 с.

24. Костров А. В., Александров Д. В. Уроки информационного менеджмента – М.:Финансы и Статистика, 2005. – 304 с.

25. Мантенья Р., Стенли Х. Введение в эконофизику. Корреляции и сложность в финансах. Пер. с англ., М.: УРСС. 2007. 192 с.

26. Пугачев В.С. Теория вероятностей и математическая статистика. М.: Физматлит. 2002. 496 с.

27. Степанов Р. Г. Технология Data Mining: «Интеллектуальный Анализ Данных»/Р.Г. Степанов. -СПб.:БХВ –Петербург. 2010. 185 с.

28. Харрис Энди. PHP/MySQL для начинающих. Пер. с англ. Кудиц-образ. 2005. 384 с.

29. Чубукова И. А. Data Mining: учебное пособие. М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. 382 с.

30. Andreas Wichert Intelligent Big multimedia databases // World Scientific Publishing. 2015. 324 p.

31. Borak S., Hardle W., Weron R. Statistical Tools for Finance and Insurance. Stable Distributions. // Springer. 2005. 424 p.

32. Kimball Ralph The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data // Willey Publishing, Inc. 2004. 351-380 P. 528 p.

33. Mayer-Schonberger and Keneth Cukier Big Data – A revolution that will transform how we live, work and think. 2013. 198 p.

34. Schmarzo Bill Big Data MBA: Driving Business Strategies with Data Science Published // John Wiley & Sons, Inc. 2016. 372 p.

35. Thomas Erl, Waji Khattak, and Paul Buhler — Big Data Fundamentals Concepts, Drivers and Techniques // Arcitura Education Inc. 2016. 234 p.

36. Wichert Andreas Intelligent Big Multimedia Databases // World Scientific Publishing Co.Pte.Ltd. 2015. 322 p.

37. Working Knowledge Thomas H Davenport Laurence Prusak // Harvard Business School Press. 1998. 199 p.

**Монографии**

38. Богуслаев А.В. Прогрессивные технологии моделирования автоматизированных систем распознавания образов: монография / А.В. Богуслаев. 468 с.

39. Бодров А.А., Рамзаев В.М., Рамзаев М.В., Хаймович И.Н. Принципы и технологии Big Data в управлении динамичными экономическими и социальными средами территорий / монография. – Самара: Издательство СамНЦ РАН. 2017. 156 с.

**Интернет-ресурсы**

40. Big Data in Government: Challenges and Opportunuties // Сайт Правительства Соединенного Королевства Великобритании и Северной Ирландии. — URL: https://www.gov.uk/government/speeches/big-data-in-government-the-challenges-and-opportunities (дата обращения: 26.05.2017)

41. Building Data Science Team // Сайт компании O`Reilly Radar. ─URL: http://radar.oreilly.com/2011/09/building-data-science-teams.html (дата обращения: 20.04.2017)

42. Government Transformational Strategy // Сайт Правительства Объединенного Королевства Великобритании и Северной Ирландии. – URL: https://www.gov.uk/government/publications/government-transformation-strategy-2017-to-2020/government-transformation-strategy (дата обращения: 20.05.2017)

43. The Heritage Foundation: Рейтинг экономической свободы стран мира 2016 года // Центр гуманитарных технологий. — URL: http://gtmarket.ru/news/2016/02/01/7293 (дата обращения: 10.04.2017)

44. Минкомсвязи опоздало на 84 дня, принимая собственный план информатизации // Сайт Счетной палаты Российской Федерации. — URL: http://www.ach.gov.ru/press_center/news/25514 (дата обращения 13.05.2017)

45. Работа над законом о цифровой экономике начнется этим летом // Информационный портал «ТАСС». — URL: http://tass.ru/pmef-2017/articles/4310901 (дата обращения: 20.05.2017)

46. Сайт Федеральной службы государственной статистики. - URL: http://www.gks.ru/ (дата обращения: 20.03.2017)

**Приложение 1**

## Глоссарий
## Словарь использованных терминов

1. SQL – структурированный язык запросов –язык программирования, применяемый для создания, модификации и управления данными в реляционной базе данных

2. NoSQL – совокупность подходов и решений, направленных на реализацию хранилищ баз данных для решения проблем их масштабируемости и доступности за счет атомарности и согласованности данных.

3. O`Reilly Radar – мировая IT-компания, специализирующаяся на создании технологий хранения и обработки данных, а также консалтинге в сфере Data Analytics и Data Mining.

4. Data Analytics – серия подходов и технологий для сбора, анализа, управления и интерпретации больших массивов данных.

5. Data Mining – технологии добычи данных/интеллектуальный анализ данных – совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний.

6. XML – расширяемый язык разметки – формат файлов, написанных на языке простого формального синтаксиса

7. JSON – текстовый формат обмена данными, основанный на JavaScript

8. Конверсационный анализ – социологический метод получения информации о восприятии индивидом или социальными группами определённых ситуаций и явлений, заключающийся в детальном анализе разговора, который происходит в естественных ситуациях.

9. ER-модель – модель высокоуровневого проектирования баз данных, предназначенная для установления сущностей и установления связей между этими сущностями

10. Мейнфрейм — большой универсальный высокопроизводительный отказоустойчивый сервер

11. Физический сервер — самостоятельная и самодостаточная единица оборудования, позволяющая осуществлять полный спектр управления на аппаратном и программном уровне.

12. Виртуальный выделенный сервер (VDS) — Созданная непосредственно на физическом оборудовании единица, эмулирующая работу физического сервера, где каждая виртуальная машина состоит из определенного объема ресурсов

13. Виртуальный частный сервер (VPS) – созданная на базе одного серверного ядра единица с отдельным программным окружением, но без права изменения ядра и операционной системы

14. VPN – виртуальная частная сеть, предназначенная для обеспечения защищенного подключения внутри корпоративных соединений и доступа в интернет

15. Линейка объемов данных (1ТБ, 1ПБ, 1ЭБ): 1 ЭБ = 1024 ПБ; 1 ПБ = 1024 ТБ; 1 ТБ = 1024 ГБ

16. PHP – скриптовый язык общего назначения, применяемый для разработки веб-приложений

17. Wolfram Language — мультипарадигмальный язык программирования, позволяющий реализовывать произвольные структуры и данные

18. Булева Алгебра – раздел математики, изучающий высказывания, рассматриваемые со стороны их логических значений (истинности или ложности) и логических операций над ними.

19. Amazon Elastic Cloud — веб-сервис, предоставляющий вычислительные мощности в облаке

20. CRM системы — Прикладное программное обеспечение для организаций, предназначенное для автоматизации стратегий взаимодействия с заказчиками

21. ERP системы – корпоративная информационная система (кис), предназначенная для автоматизации учета и управления

22. Teradata – Американская корпорация, специализирующаяся на разработке и поставке аппаратно-программных комплексов для обработки и анализа данных.

23. Oracle –Американская транснациональная корпорация, второй по величине доходов производитель программного обеспечения, крупнейший производитель программного обеспечения для организаций, крупный поставщик серверного оборудования

24. AS/400 – Сервер, вычислительная система, представленная компанией IBM.

25. DB2 – Семейство систем управления реляционными базами данных, выпускаемых корпорацией IBM.

26. IBM ― Американская компания со штаб-квартирой в Армонке, один из крупнейших в мире производителей и поставщиков аппаратного и программного обеспечения, а также ИТ-сервисов и консалтинговых услуг.

27. Hadoop – платформа/фреймворк, предназначенный для построения распределённых приложений для работы с данными большого объёма.

28. Map-Reduce – вычислительная парадигма, используемая для параллельных вычислений над очень большими (ПБ), наборами данных в компьютерных кластерах.

29. PERL – Высокоуровневый интерпретируемый динамический язык программирования общего назначения.

30. BLOB – массив двоичных данных.

31. Масштабируемость ― способность системы работать с дополнительными пользователями или транзакциями путем наращивания ресурсов без фундаментальной перестройки архитектуры или модели реализации.

32. Дистрибутив – файлы и архивы, предназначенные для установки какой-либо программы.

33. ETL ― один из основных процессов в управлении хранилищами данных, который включает в себя извлечение данных из внешних источников; их трансформацию и очистку, чтобы они соответствовали потребностям бизнес-модели; загрузку их в хранилище данных.

34. Идеология shared-nothing ― это распределённая вычислительная архитектура, в которой каждый узел независим и самостоятелен, отсутствует единая для всей системы точка подключения.

35. Business Intelligence ― набор IT-технологий для сбора, хранения и анализа данных, позволяющих предоставлять пользователям достоверную аналитику в удобном формате, на основе которой можно принимать эффективные решения для управления бизнес-процессами.

36. Data Science ― раздел информатики, изучающий проблемы анализа, обработки и представления данных в цифровой форме.