МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Тольяттинский государственный университет»

Институт математики, физики и информационных технологий
(наименование института полностью)

Кафедра Прикладная математика и информатика
(наименование)

09.04.03 Прикладная информатика
(код и наименование направления подготовки)

Управление корпоративными информационными процессами
(направленность (профиль))

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ)

на тему «Технология People Data для прогнозирования оптимальных показателей о персонале»

Обучающийся А.И. Белоусов _____
(Инициалы Фамилия) (личная подпись)

Научный руководитель канд. пед. наук, доцент, О.М. Гущина
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2022

Содержание

Введение

Любое предприятие либо производит какую-либо продукцию, либо выполняет какие-либо работы, или же предоставляет какие-либо услуги. Чтобы предприятие, организация или фирма успешно производили свою работу, приносили прибыль, развивались, нужно эффективно управлять персоналом, прогнозировать какие-либо отклонения от нормы в работе людей, будь то возможные отпуска, больничные, непредвиденные увольнения и иные подобные ситуации. Все это позволит более грамотно составлять производственные планы, стараться предугадывать на перед возможные затраты, снижение эффективности производства, и все это нужно для того, чтобы можно было более эффективно выполнять непосредственные задачи, поставленные перед организацией.

Эффективное управление персоналом – одна из важнейших составляющих успешного ведения бизнеса. Ведь, когда персонал работает наиболее эффективно, прибыль предприятия начинает расти.

Сегодня во многих организациях для манипуляций с данными, такими как сбор, обработка и анализ, которые являются очень трудозатратными, используют информационные системы. Применение этих систем позволяет эффективно управлять ресурсами и людьми. Но проблема в том, что не всегда получается эффективно обрабатывать данные о людях в тех системах, которые применяются, так как такие системы в основном нацелены на обработку технической информации и не всегда дают аналитикам предприятия необходимую информацию о персонале. Что может влиять на эффективность прогнозирования показателей о персонале. Для эффективной обработки данных о людях необходима работа по совершенствованию способов обработки и анализа данных о людях, необходимых для оптимизации управления персоналом.

Таким образом, актуальность темы подтверждается необходимостью выявления методов и технологий, позволяющих более эффективно

обрабатывать информацию о людях и делать на основе этой информации выводы для оптимизации работы персонала.

Объектом исследования магистерской диссертации является People Data как часть понятия Big Data.

Предметом исследования магистерской диссертации являются методы и технологии работы с People Data.

Целью работы является разработка модели работы с People Data для оптимизации прогнозирования показателей о персонале.

Для достижения поставленной цели, нужно решить следующие задачи:

– ознакомиться с существующими научными работами на связанные и смежные тематики,

– определить значимость использования People Data на сегодняшний день,

– определить актуальные возможности применения подхода People Data и выявить современный вектор развития управления персоналом организации,

– ознакомиться с существующими методами работы с большими данными, изучить плюсы и минусы данных методов,

– выбрать метод и технологию, которые будут применены в настоящем диссертационном исследовании,

– рассмотреть возможные варианты реализации технологии MapReduce и метода статистического анализа,

– выбрать наиболее подходящий вариант реализации данных метода и технологии для нашего исследования,

– объединить полученные методы и технологии работы с People Data для наиболее эффективного решения задачи оптимизации прогнозирования показателей о персонале,

– применить предложенную модель и произвести анализ эффективности применения предложенной модели для решения задачи оптимизации прогнозирования показателей о персонале.

Гипотеза исследования: применение предложенной в рамках настоящего диссертационного исследования модели работы с данными о людях, предоставит возможность самым эффективным образом искать новые пути для решения управленческих задач в организации, а также формировать шаблон наиболее эффективного управления персоналом в будущем.

В ходе исследования были использованы следующие методы: теоретические методы для анализа существующих подходов и их сравнения, и предложения собственной модели работы с People Data; эмпирические методы для получения результатов апробации предложенной модели; и математические методы для анализа эффективности предложенной модели.

Новизна исследования заключается в предложении объединения технологии MapReduce и метода статистического анализа для оптимизации работы с данными о людях, а также для оптимизации показателей о персонале.

Практическая значимость исследования заключается в возможности применения в современных реалиях разрабатываемой модели по работе с People Data, которая сможет повысить эффективность прогнозирования показателей о персонале на реальных производствах.

Теоретической основой диссертационного исследования являются научные работы российских и зарубежных ученых, которые занимаются проблемами обработки больших данных и управлением персоналом.

Основные этапы исследования: исследование проводилось с 2020 по 2022 год в несколько этапов. На первом этапе (констатирующем этапе) формулировалась тема исследования, производилось накопление информации по теме работы, была сформулирована гипотеза, определены цель и задачи, а также предмет и объект исследования, и констатировалась проблема данного исследования. Второй этап (поисковый этап) – на этом этапе были проанализированы существующие методы работы с Big Data и возможность применения этих методов для работы с People Data, также был предложена собственная модель работы с People Data для оптимизации показателей о персонале. Третий этап (реализация), на этом этапе была апробирована

предложенная модель и произведен анализ эффективности применения предложенной модели, а также были сделаны выводы о полученных результатах по проведенному исследованию.

На защиту выносится:

– предложенная модель работы с People Data, которая заключается в совместном применении технологии MapReduce и метода статистического анализа для оптимизации прогнозирования показателей о персонале,

– результат апробации предложенной модели и анализ эффективности предложенной модели.

Диссертация состоит из введения, четырех глав, заключения и списка литературы. Во введении обоснована актуальность темы исследования, представлены объект, предмет, цели, задачи и положения, выносимые на защиту диссертации. Первая глава посвящена анализу понятия People Data и тому, как понятие People Data соотносится с понятием Big Data. Раскрыто понятие People Data, описаны особенности понятий Big Data и, в частности, People Data. Проанализированы возможные направления для применения People Data. Во второй главе исследованы методы и технологии работы с People Data, выявлены достоинства и недостатки существующих методов и технологий работы с People Data, предложена собственная модель работы с People Data, которая позволит более эффективно работать с People Data. В третьей главе было представлено описание собственной модели для более эффективной работы с People Data, которое применимо для оптимизации прогнозирования показателей о персонале. В четвертой главе была произведена апробация предложенной модели для решения задачи оптимизации показателей о персонале. Рассмотрена эффективность предложенной модели. В заключении описываются результаты выполнения магистерской диссертации.

Данная работа содержит 93 страниц, 3 таблицы, 8 формулы, 17 рисунков и 46 использованных источников.

1 People Data, как часть понятия Big Data

1.1 Анализ People Data и его связь с Big Data

Понятие «People Data» в своем становлении прошло несколько этапов.

В ходе первой стадии производилась агрегация данных и предпринимались первые шаги в анализе этих данных, чтобы сделать какие-либо выводы. А на основе этих выводов уже принимать решения по управлению персоналом в организации.

Второй этап подразумевает под собой использование уже собранной и накопленной информации на первом этапе для отчетности, но стоит отметить, что накопление информации все также продолжается.

Сравнительный анализ показателей за периоды является третьим этапом становления понятия «People Data».

Четвертым этапом выступает попытка понять, какими будут показатели на перспективу. Притом все большее применение находят специальные аналитические инструменты для анализа информации, с помощью которых можно оценить возможность изменения отдельных показателей.

С появлением и повсеместным использованием Интернета, стало возможным получать огромные объемы данных для анализа. Такие данные называются Big Data. На сегодняшний день продолжается пятый этап, который можно охарактеризовать, как способность предсказать показатели на перспективу.

Чтобы продолжить изучение данной тематики стоит понимать разницу между рядом терминов: People Data, Big Data и HR-аналитика. В 2015 году понятие Big Data выделилось в обособленное направление исследований, а позже и стало применяться на практике во многих IT-компаниях.

На основании анализа источников [45] и [39], можно увидеть и более ранее использование понятия «Big Data», так в 1918 году Ч. Тилли использует «Big Data», при написании работы, посвященной большим вопросам, которые

стоят перед человечеством [45]. Р. Вильямс в 2003 говорит, что «Big Data» очень перспективное направление в изучении и работе [39].

В Оксфордский словарь термин «Big Data» включили в 2013 году. В этом словаре говорится, что большие данные – это данные такого большого размера, что производить с ними DDL и DML операции проводить практически невозможно из-за возникающих логических ошибок. Помимо того, под большими данными понимается направление работы по вычислению с такими данными.

В более позднем трактовании Big Data связывают с анализом данных огромных объемов и их сбор. Результаты анализа таких данных представляют из себя утверждение о наличии или отсутствии связей между анализируемыми наборами данных. Эти утверждения выводятся с использованием специальных методов и инструментов.

Любому виду и объему данных присуще наличие восьми основополагающих отличительных черт:

- объем (volume),
- разнообразие (variety),
- скорость (velocity),
- жизнеспособность (viability),
- ценность (value),
- достоверность (veracity),
- наглядность (visualization),
- переменчивость (variability).

Выраженность этих черт прямо пропорциональна объему обрабатываемой информации.

People Data – это не просто Big Data, где обрабатываются данные о людях. People Data так же в себя взяло частичку HR-аналитики. По факту People Data стоит на стыке двух направлений в науке: Big Data и HR-аналитика.

С понятием Big Data разобрались. Теперь необходимо понять, что из себя представляет HR-аналитика.

HR-аналитика – это действия над данными, которые направлены на сбор, обработку, нормализацию и анализ данных о людях, поиск зависимостей между этими данными, чтобы принимать решения в отношении управления персоналом. Для анализа этих данных применяются информационные системы и математические методы.

За частую понятие HR-аналитики трактуют, как сбор данных о сотрудниках на регулярной основе и их анализ для принятия решений по управлению персоналом, которые могут поспособствовать достижению бизнес-целей. За последние десятилетия HR-аналитика вызывает не меньший интерес, чем Big Data.

HR-аналитика занимается расчетом HR-метрик, своевременным обнаружением и управлением выдающихся сотрудников в компании, представляет в наглядном виде данные о сотрудниках, производит поиск утечек информации. Все это позволяет более эффективно управлять сотрудниками в организации.

Если провести анализ компетенций HR-специалистов, инструментов и подходов, которые они применяют в своей работе, то можно увидеть, что сегодня HR-аналитика представляет из себя сбор HR-метрик и поиск связей между этими метриками и бизнес-показателями организации.

Поскольку, в соответствии с работой [16], в последние десять лет появился содержательный прирост в сфере HR-аналитики, это говорит о том, что есть смысл увеличить разнообразие качества данных, которые используются HR-аналитикой, для анализа данных о сотрудниках и на основе которых принимаются бизнес-решения по управлению сотрудниками. Для анализа уже сейчас можно начинать использовать данные о состоянии дел за пределами организации, помимо стандартных данных о сотрудниках и результатах их деятельности.

Все вышеперечисленное включает в себя новый подход, который носит название People Data. Впервые это понятие было применено в 2013 году. Давайте разберемся, что же такое People Data.

Если раньше People Data считалась исключительно информацией, которая собрана в информационных системах организации, о сотрудниках организации, которая включает в себя описание поведения сотрудников и их характеристики, то, начиная с 2018 года, People Data включило в себя всю полноту данных о сотрудниках организации (информация о сотрудниках и их родственниках) и еще данные извне (информация о клиентах компании, общественном секторе, власти, других жителей страны).

При таком подходе к People Data значительно возрастает количество используемых данных, направление их применения. А чем больше данных, тем тяжелее последствия их утечки, недобросовестного использования этих данных конкурентами, при их утечке.

People Data, как и HR-аналитика, и Big Data – это инструменты работы с данными. Big Data делает обработку больших данных на ЭВМ с применением различных методов и технологий, которые будут рассмотрены дальше в работе. HR-аналитика уже нацелена на анализ больших данных о людях с применением корреляционного анализа входных показателей. В свою очередь People Data обобщает все достоинства использования Big Data и HR-аналитики в едином месте для принятия управленческих решений.

Как говорилось ранее, при работе с данными о людях, имеются свои риски. Главным из этих рисков является возможная утечка персональных данных. Стоит также отметить, что работа с персональными данными регулируется Федеральным законом от 27 июля 2006 г. № 152-ФЗ «О персональных данных». Давайте дадим определение персональным данным.

Персональные данные – это всевозможная информация об определенном или определяемом физическом лице, субъекте персональных данных. Такая информация может прямо или косвенно относиться к данному субъекту.

Вопрос о персональных данных в РФ регулирует Федеральный закон от 27 июля 2006 г. № 152-ФЗ «О персональных данных». В рамках данного закона очень трудно использовать персональные данные людей, чтобы достичь организационных целей. Но данные о сотрудниках, полученные в установленном порядке, проще использовать в организационных целях предприятия.

В настоящей работе будут использоваться обезличенные данные о людях. Так как они будут необходимы для сбора статистики по организации в целом. Обезличенные данные – это информация о людях, по которой невозможно идентифицировать конкретного человека, не прибегая к использованию дополнительной информации или обработки. Основные риски использования данных о людях приведены в таблице 1.

Таблица 1 - Риски в использовании данных о людях в организационных целях

Сторона отношений	Риски	Варианты сокращения рисков
Работник	Риск несоблюдения интересов работника	Принятие государственных или внутренних НПА организации, которые запрещают использовать персональные данные в целях, отличных от целей сбора
Работодатель	Неумышленное нарушение государственных или внутренних НПА организации относительно обработки персональных данных	Четкое и неукоснительное соблюдение государственных или внутренних НПА организации относительно обработки персональных данных каждым сотрудником организации, шифрование данных
Органы власти	Неподтвержденность информации, которая собирается компаниями, парой слишком жесткие ограничения по обработке персональных данных, что ограничивает развитие технологий обработки персональных данных	Ужесточение законодательства за нарушение процедур обработки персональных данных

Продолжение таблицы 1

Сторона отношений	Риски	Варианты сокращения рисков
Муниципальные организации	Сложные процедуры для получения доступа к персональным для работодателя и субъекта персональных данных	Выстраивание современных сетевых коммуникаций, с высоким уровнем защиты данных, между организациями, владеющими необходимой информацией, субъектами, которым принадлежит эта информация, и работодателем
Сотрудники, работающие с персональными данными	Неумышленное нарушение государственных или внутренних НПА организации относительно обработки персональных данных, утечка персональных данных, нарушение целостности персональных данных	Соблюдение внутренних и государственных НПА относительно работы с персональными данными каждым сотрудником, наставничество над вновь прибывшими сотрудниками, которые работают с персональными данными

Стоит отметить, что помимо вышеперечисленных рисков, есть еще один. Люди могут негативно отнестись к сбору их персональных данных, так как не всегда понимают цели сбора этих данных. Это может в некоторой степени повлиять на рабочий настрой людей и снизить легитимность начальства в глазах работников. По этой причине, нужно максимально доходчивым образом доносить до людей цели сбора их данных, а также уверить людей в надежности хранения этих данных внутри организации, а при угрозе их компрометации – немедленно уничтожать, а также уничтожать по истечению сроков обработки и хранения данных. Все это однозначно влияет на репутацию компании на рынке, поэтому такими «рисками» тоже нужно уметь управлять.

Еще одной важной точкой можно выделить риск слепого следования рекомендациям искусственного интеллекта в ходе управления предприятием. Не всегда стоит слепо следовать рекомендациям, которые выдает нейронная сеть.

На сегодняшний день очень быстрыми темпами развивается цифровизация. Что в свою очередь влияет на работу с данными. В работах [43],

[44], [5] приводится обоснование того, что сегодня существует необходимость переходить бизнесу компаний на цифровую платформу ведения деятельности, так как количество клиентов, да и в принципе информации растет, а в таком случае ее удобнее обрабатывать в электронном виде. Для обработки больших данных следует использовать технологию Big Data.

Но в то же время не все компании готовы переходить на цифровизацию, а соответственно, и применение Big Data для обработки всех данных, хотя желание присутствует у многих компаний [9].

Но не стоит забывать и о безопасности данных, которые будут обрабатываться с применением технологий Big Data и People Data, о чем нам повествуют статьи [42], [30].

В работе [32] также говорится о безопасности обработки и хранения персональных данных, этических сторонах хранения персональных данных, а также возможных рисках, при хранении персональных данных.

В работе [40] производится анализ данных о поведении на дороге, а в работе [46] производится анализ причины скопления предпринимателей в одних регионах и их отсутствие в других. Эти анализы могут быть проведены с помощью технологии Big Data или технологии People Data, так как объем информации зачатую большой, что свидетельствует от том, что исследуемая тема актуальна.

Работа [13] демонстрирует реализацию анализа данных с помощью Python.

В работах [38], [21] рассмотрены предпосылки для возникновения и особенности управления эффективностью маркетинга и банкинга. Такая система рассматривается, как целостная многоуровневая система. В данных статьях предложен подход к управлению эффективностью маркетинга и банкинга с помощью интегрированных информационных систем, которые могут использовать технологию Big Data для анализа данных, которые могут повысить эффективность маркетинга и банкинга. То есть это нам говорит о

том, что сегодня направление Big Data продолжает набирать свою популярность.

В силу глобализации и цифровизации экономики сформировалась новая стратегия ведения бизнеса – Digital-стратегия, которая интегрирована в бизнес и HR-стратегию. Digital-стратегия подразумевает постепенный переход бизнеса в online, поэтому многие компании изменили применяемые подходы для управления персоналом, сформировав HR-тренды, которые используют новые модели для увеличения эффективности [28], что опять же свидетельствует о необходимости обработки больших данных, в частности People Data для управления человеческими ресурсами.

В статье [37] говорится о том, что такое HR-аналитика, а также, как она применяется в управлении персоналом.

В работе [33] исследуется проблема HR-аналитики в цифровой экономике. В работе представлены плюсы и минусы использования тех или иных подходов к HR-аналитике, как новых направлений автоматизации обработки данных о людях. Помимо того, в это работе предложена программа применения HR-аналитики, как направления автоматизации обработки данных о людях.

В статье [36] также говорится о результатах апробирования различных подходов HR-аналитики, как новых направлений автоматизации обработки данных о людях. Рассмотрены возможности каждого подхода HR-аналитики.

В статье [19] говорится об анализе больших данных из сети «Вконтакте», для анализа этих данных может также использоваться технология Big Data или People Data, что свидетельствует об актуальности рассматриваемой темы, так как существует множество сфер, где можно применить данную технологию.

Также в статье [41] описывается оценка на основе данных из социальных сетей для оценки кредитоспособности человека.

Работа [10] посвящена повышению эффективности управления персоналом за счет применения Data Science при работе с большими объемами данных. Приводятся примеры использования инструментов и платформ

обработки больших данных. Данная статья раскрывает целесообразные механизмы работы с Big Data в сфере управления персоналом.

В статье [34] говорится о применении Big Data в государственном и корпоративном управлении для анализа больших объемов данных. В результате работы были получены интересные данные для построения модели, как государственного, так и корпоративного управления при современном уровне цифровой трансформации.

Также Big Data может применяться не только в управлении, но и в маркетинговой деятельности для обработки большого потока информации о клиентах и не только [8]. Кстати, тут же можно использовать и другой подвид Big Data, такой как People Data.

Так же в [25] говорится о том, что в настоящее время вводятся ИС, которые способны произвести анализ эффективности управления социально-экономическим развитием области. Для данных целей целесообразно использовать как раз-таки технологии People Data, чтобы проанализировать все результаты эффективности управления на основе собранных данных, что еще раз подчеркивает тот факт, что на сегодняшний день очень актуально использовать технологии Big Data и, как ее частный случай, People Data.

В [11] предлагается использовать People Data на основе искусственного интеллекта для принятия управленческих решений в современных реалиях, в том числе о наборе новых сотрудников, основываясь на информации о них из общедоступных источников.

В статье [16] аккумулируется все вышесказанное и подводятся итоги по целесообразности применения People Data, ее плюсы и минусы. Делается вывод о том, что на сегодняшний день технология People Data очень актуальна и востребована практически во всех сферах жизни человека на сегодняшний день.

Вывод: на основании изученной литературы, мы получили представление о том, что такое People Data и как оно соотносится с понятием

Big Data. Также рассмотрение изучение литературы выявило актуальность рассматриваемой темы.

1.2 Направления использования данных о людях в системе управления персоналом

Данные о людях можно использовать для таких целей, как:

– воздействие на показатели эффективности (KPI),

– оптимизация производительности труда, повышение лояльности персонала,

– управление организационными процессами,

– анализ действий сотрудников и результатов, к которым они привели в соотношении с их компетенциями, знаниями, навыками и умениями,

– создание системы управления перспективными кадрами,

– улучшение обучения персонала,

– организация системы мониторинга за сотрудниками.

Воздействие на KPI посредством анализа зависимостей различных показателей для повышения эффективности управления сотрудниками.

Одними из главных показателей для производства служат производительность труда и лояльность подчиненных к начальству. За счет повышения эффективности управления этими показателями, можно повысить прибыль компании, что и является конечной целью любого анализа данных сотрудников организации.

Помимо того, основываясь на анализе данных о персонале, можно повысить эффективность управления информационными процессами в организации.

Анализ действий сотрудников и результатов, к которым они привели в соотношении с их компетенциями, знаниями, навыками и умениями может способствовать развитию эффективных процессов по обнаружению и

развитию талантов, а это может привести к повышению эффективности обучения персонала.

Все вышеперечисленное позволяет организовать систему мониторинга за сотрудниками, которая призвана помочь топ менеджменту по-новому взглянуть на управление персоналом, используя информационную систему для отслеживания соотношения качеств сотрудника, его действий и получаемых им результатов.

Такой комплексный подход (анализ, как внутренних для компании, так и внешних аспектов) позволит найти индивидуальный подход к каждому сотруднику.

Еще одной важной точкой для развития использования данных о персонале является создание отделов анализа информации о людях и их встраивание в существующую информационную систему организации. В большинстве случаев проще будет отказаться от используемых IT-решений в организации и сделать одну большую новую экосистему организации, которая позволит эффективно управлять персоналом организации.

В настоящее время HR-аналитика быстрыми темпами берет в оборот новые технологии для работы. Предполагается, что в скором будущем автоматизируют большую часть функций HR-аналитиков. Но это все будет происходить только в том случае, если в этом будет реальная необходимость при принятии взвешенных управленческих решений на основе автоматизированного анализа данных, в первую очередь, с применением технологии People Data.

Вывод: использование и исследование People Data на сегодня очень актуально, применение этой технологии может поспособствовать увеличению эффективности управления персоналом, что в свою очередь позволить поднять эффективность управления организацией в целом и увеличить производительность и прибыль.

2 Методы и технологии работы с People Data

2.1 Обзор существующих методов и технологий работы с Big Data применительно к работе с People Data

Прежде чем приступить к обзору и анализу существующих методов работы с People Data необходимо дать определение ему, а также понятию Big Data, так как это основополагающие понятия данной работы.

Начнем с определения Big Data, потом дадим определение понятию People Data, так как People Data является составной частью Big Data.

В Оксфордский словарь термин «Big Data» включили в 2013 году [29], [45]. В нем говорится, что «Большие данные – это компьютерные данные настолько большого размера, что манипулировать и управлять ими крайне трудно в силу логических проблем; так же это направление вычислений с подобными данными» [8]. Главным критерием, по которому определяется, относятся ли рассматриваемые данные к «Большим данным» является трудность вычислений.

В более позднем трактовании Big Data связывают с анализом данных огромных объемов и их сбор. Результаты анализа таких данных представляют из себя утверждение о наличии или отсутствии связей между анализируемыми наборами данных. Эти утверждения выводятся с использованием специальных методов и инструментов.

Теперь можно дать определение непосредственно People Data. Впервые это понятие было применено в 2013 году. Давайте разберемся, что же такое People Data.

Если раньше People Data считалась исключительно информацией, которая собрана в информационных системах организации, о сотрудниках организации, которая включает в себя описание поведения сотрудников и их характеристики, то, начиная с 2018 года, People Data включило в себя всю полноту данных о сотрудниках организации (информация о сотрудниках и их

родственниках) и еще данные извне (информация о клиентах компании, общественном секторе, власти, других жителей страны) [28], [33], [37].

После того, как мы разобрались с основными определениями, можно перейти непосредственно к обзору и анализу существующих методов работы с People Data.

Для начала рассмотрим основные характерные черты при работе с большими данными, которые применяются для работы с People Data, их всего три [43], [44], [5], [20]. Начнем с горизонтальной масштабируемости. Горизонтальная масштабируемость является основным принципом при обработке больших объемов данных. Исходя из того, что каждый день объемы данных растут, это касается и понятия People Data, необходимо пропорционально увеличивать количество узлов, на которых будут происходить вычисления и между которыми будут распределяться эти самые данные. В то же время нужно, чтобы производительность обработки данных не ухудшалась с возрастанием объема данных.

Следующий принцип – это отказоустойчивость. Он вытекает из горизонтальной масштабируемости. Поскольку количество вычислительных узлов возрастает, то и вероятность выхода из строя какого-либо узла возрастает. Поэтому методы работы с People Data должны учитывать возможность выхода из строя какого-либо вычислительного кластера, а также предоставлять меры для противодействия таким ситуациям.

И последний принцип, который мы рассмотрим – это локальность данных. Поскольку у нас есть множество машин для вычислений, то лучшим решением будет обрабатывать данные на той ЭВМ, где расположены эти данные, чтобы предупредить неоправданно большие затраты на передачу данных между машинами.

Вышеперечисленные принципы отличаются от принципов работы с централизованными, вертикальными моделями хранения структурированных данных. Исходя из этого можно сказать, что для работы с большими данными, в том числе с данными о людях, существуют свои подходы и технологии.

Начнем с рассмотрения методов работы с Big Data, которые также применимы к работе с People Data.

Международная консалтинговая компания McKinsey выделяет следующие методы и техники для анализа больших данных [26].

Методы класса Data Mining. Сюда входят методы, которые могут обнаружить во входных данных знания, которые были ранее неизвестны. А также знания, которые являются не простыми и практически полезными. Такие знания необходимы для принятия управленческих решений. Примерами таких методов могут послужить обучение ассоциативным правилам, кластерный анализ, регрессионный анализ, классификация и другие методы.

Краудсорсинг – обогащение и классификация данных усилиями многих людей, которые могут не знать друг друга.

Смешение и интеграция данных – совокупность приемов, которые позволяют соединить разнообразные и разнородные данные из множества источников для проведение глубинного анализа этих данных. Примером может послужить обработка речи человека, которая включает в себя тональный анализ, цифровая обработка различных сигналов и много другое.

Далее рассмотрим машинное обучение и нейронные сети. Машинное обучение может быть двух вариантов: с учителем или без учителя. В обоих случаях это будет применение моделей, которые были основаны на результатах статистического анализа. Такие модели позволяют получить прогноз развития ситуации посредством применения базовых моделей.

Искусственный интеллект использует в себе, как правило различные генетические алгоритмы или сетевой анализ, их оптимизации. Генетические алгоритмы – это познавательные алгоритмы для поиска, которые используются для решения ряда задач оптимизации или моделирования. Решение таких задач происходит посредством случайного перебора, вариации и комбинирования параметров, которые подлежат поиску, с использованием механизмов, которые аналогичны естественному отбору в природе.

Следующий метод для анализа больших данных – это прогнозная аналитика или как ее еще называют – предиктивная аналитика. Этот метод составляет прогнозы на ближайшее будущее на основе анализа и классификации входных данных.

Еще одним методом для анализа больших данных является имитационное моделирование. С помощью имитационного моделирования можно строить модели, которые способны описывать процессы таким образом, как они протекали бы в реальности. Этот метод можно считать разновидностью экспериментальных испытаний.

Еще один подход, который мы рассмотрим, будет пространственный анализ. Это совокупность методов, которые используют геометрическую, топологическую и географическую информацию, которая извлекается из входных данных.

Статистический анализ. Этот метод, который применяет A/B тестирование и анализ временных рядов. A/B тестирование – сравнение контрольной группы элементов с набором тестовых групп. При том в таких тестовых группах был изменен один или несколько показателей. Это изменение показателей используется для выяснения, какие из этих изменений улучшают целевой показатель [14].

И последний метод, который будет рассмотрен в данной работе – это визуализация аналитических данных. Под визуализацией аналитических данных понимается подача информации посредством диаграмм, рисунков, использование интерактивных возможностей и анимации для входных и выходных данных. Этот этап анализа больших данных представляет самые важные результаты анализа в удобном и понятном виде. Поэтому его следует использовать совместно с еще каким-нибудь методом работы с People Data, из перечисленных выше.

После рассмотрения методов работы с People Data, перейдем к рассмотрению конкретных технологий работы с People Data, как части понятия Big Data.

На заре развития Big Data в перечень технологий включались технологии массово-параллельной обработки слабо структурированных данных. К таким технологиям можно отнести NoSQL, средства проекта Hadoop, а также алгоритмы MapReduce. Со временем к этому списку начали относить и другие решения, с помощью которых можно добиться аналогичных по характеристикам возможностей для обработки больших данных. Помимо того, к этому списку технологий можно еще добавить некоторые аппаратные средства. Рассмотрим наиболее популярные технологии.

Прежде чем перейти непосредственно к рассмотрению конкретных технологий следует отметить, что большинство технологий связано с распределенной обработкой данных. Это связано с тем, что очень большие объемы данных практически не реально обрабатывать на одном физическом узле, так как мощности машины будет не хватать.

Начнем с технологии MapReduce. Эта технология была разработана корпорацией Google. Данная технология представляет из себя модель параллельных вычислений, которые распределены между некоторым числом компьютерных кластеров. Первый раз такой подход был использован для реализации распределенной файловой системы GFS (Google File System), а также для реализации не реляционной базы данных от Google, такой как Big Table.

При применении этой технологии наше приложение параллельно выполняется на большом количестве физических узлов (машинах) кластера. После полной обработки данных на каждом узле естественным образом сводится в единый конечный результат на одной машине (физическом узле).

Стоит отметить, что имплементация методов работы с входными данными для кластерных систем, включает в себя решение задачи разбиения и балансировки данных между узлами распределенной вычислительной системы, а также обработки отказов. Помимо того, нужно решить задачи сбора и агрегации промежуточных результатов в узлах распределенной системы [25].

Каждая партиция данных преобразуется на той машине, которая свободна и в приоритете используется та ЭВМ, на которой хранятся данные. Поскольку эти партиции данных не зависят друг от друга, то операции преобразования с ними могут выполняться параллельно. Точно также независимо – параллельно друг другу выполняются операции свертки. Но перед выполнением операции свертки должны успешно завершиться все операции преобразования с одним значением ключа, чтобы значения с одним значением ключа обрабатывать на одной машине. То есть перед операцией свертки необходимо синхронизировать все потоки с одинаковым ключом.

Технология MapReduce очень проста и удобна, поскольку от пользователя система скрывает детали реализации кластерной системы.

Следующей возможной технологией выступает NoSQL, что означает Not Only SQL. Это обобщающий термин для различных не реляционных баз данных и хранилищ.

Примерами систем управления не реляционных базами данных могут послужить Cassandra (рассчитана на создание высокомасштабируемых и надежных хранилищ огромных массивов данных, которые представлены в виде хэша, работает со структурами данных типа ключ-значение, данные хранятся в виде разряженной матрицы), MongoDB (документно-ориентированная система управления базами данных), Redis (СУБД, которая размещается в оперативной памяти – in memory, работает со структурами данных типа ключ-значение), HBase (данные хранятся в виде разряженной матрицы), Neo4j (графовая система управления базами данных), Amazon DynamoDB (документно-ориентированная СУБД, работает со структурами данных типа ключ-значение), Couchbase (документно-ориентированная СУБД, работает in memory, работает с данными в формате ключ-значение), Memcached (работает со структурами данных типа ключ-значение), CouchDB (документно-ориентированная СУБД).

Реляционные базы данных хорошо работают для простых запросов с хорошо структурированными данными, но, если данных много, и они слабо

структурированы, как бывает при работе с большими данными, нагрузка на СУБД становится слишком большой, поэтому использование реляционной СУБД становится неоправданно дорогостоящим по времени выполнения операций. В таком случае намного удобнее использовать NoSQL СУБД.

Hadoop – это фреймворк, который используется для создания распределенных программ, которые выполняются на физически отдельных кластерах. Данная технология является одной из основных для работы с большими данными. Данный инструмент для работы с большими данными был разработан Apache foundation. Hadoop изначально задумывался, как инструмент для хранения данных и как планировщик задач MapReduce. Сейчас же Hadoop обширный фреймворк для работы с большими данными, который включает в себя не только технологию MapReduce.

По состоянию на 2022 год технология Hadoop состоит из четырех модулей [15]. Рассмотрим их по порядку.

Первым модулем, который мы рассмотрим, будет Hadoop Common. Hadoop Common является связующим программным обеспечением. По факту это совокупность программных библиотек и вспомогательных компьютерных программ, которые применяются для других программных модулей и родственных проектов.

Следующий модуль, который мы рассмотрим, будет HDFS (Hadoop Distributed File System) – это не реляционная база данных, такие базы данных предназначены для хранения очень больших объемов информации.

Модуль YARN – это по сути своей фреймворк, который является планировщиком задач MapReduce, и управляет ресурсами кластера.

Модуль MapReduce от Hadoop является модулем программирования, где происходит выполнение задач MapReduce.

Hadoop Ozon применяется для хранения объектов в памяти для фреймворка Hadoop.

Также стоит отметить, что существуют проекты, которые тесно связаны с технологией Hadoop, но не входят в базовый набор фреймворка Hadoop. Рассмотрим их.

Hive является вспомогательным модулем для запросов на SQL подобных языках для работы с большими данными. Ее суть сводится к генерации MapReduce задач на основе SQL запросов.

Pig является языком программирования, который преобразует простые пользовательские команды в серию MapReduce задач.

HBase является базой данных, которая использует парадигму Big Table.

Cassandra это не реляционная база данных, работающая на основе HashMap. Данные в ней хранятся в виде разряженной матрицы.

ZooKeeper это сервис, который можно закинуть в контейнер и использовать для хранения конфигурационных файлов, он синхронизирует изменения вносимые в эти файлы на разных машинах.

Mahout – готовое решение для обучения искусственного интеллекта на больших объемах данных.

Также есть язык программирования R для с помощью которого можно легко вести статистический анализ данных. Обработанные данные представляются в виде графиков и диаграмм. Зачастую используется для проведения статистических исследований. Фактически он стал стандартом для написания статистических программ.

Стоит отметить, что R также – это свободно распространяемая среда вычислений, которая имеет открытый программный код.

В завершении нашего обзора технологий для работы с People Data, как части понятия Big Data, рассмотрим еще одну технологию работы с большими данными. Эта технология – аппаратные решения.

Некоторые корпорации, такие как EMC, Teradata и другие поставляют аппаратно-программные комплексы, с помощью которых можно производить обработку больших данных. Такие комплексы представляют из себя готовые к установке телекоммуникационные шкафы. В этих шкафах находятся кластер

серверов и программное обеспечение, которое управляет массово-параллельной обработкой входных данных.

Помимо того, к аппаратным решениям иногда относят аппаратные решения, которые аналитически обрабатывают большие данные в оперативной памяти. Примерами таких решений могут быть программные комплексы Exalytics от Oracle и Hana от компании SAP.

Стоит отметить, описанная выше обработка не массово-параллельная, по этой причине оперативная память на каждом вычислительном узле ограничивается парами терабайт.

Также стоит подчеркнуть, что компания McKinsey в перечень технологий, которые возможно применять для работы с Big Data включает системы управления реляционными базами данных и технологию Business Intelligence.

На этом можно закончить обзор существующих методов и технологий для работы с People Data. Теперь можно перейти к более подробному рассмотрению данных методов и технологий, определению их плюсов и минусов, а также к выбору той технологии, которая будет использоваться в данной работе.

Вывод: на сегодняшний день существует множество методов и технологий работы с Big Data. Необходимо выяснить какие из них наиболее подходят для работы с People Data, а также для прогнозирования оптимальных показателей о персонале.

2.2 Анализ методов и технологий работы с Big Data применительно к People Data

После того, как мы сделали обзор основных технологий и методов работы с People Data, можно перейти к анализу достоинств и недостатков данных методов и технологий и выбору метода и технологии, которые будут применены в данной работе.

Начнем анализ плюсов и минусов каждого метода в том же порядке, в котором они были рассмотрены выше. Начнем с методов класса Data Mining.

People Data, как и Big Data, в большинстве своем являются огромным массивом разнородных данных. Для того, чтобы можно было извлечь из них пользу, в этих огромных объемах данных нужно найти полезные закономерности. Примерами таких закономерностей могут послужить различия, сходства, общие категории и тому подобные вещи. А вот самим процессом поиска этих закономерностей и является Data Mining. В данном случае под Data Mining подразумевают добычу данных, или, как это еще называют, глубинный анализ данных.

Давайте посмотрим, как это работает. Берутся большие объемы данных, затем с помощью различных технологий из эти данных выбираем новые полезные для нас данные. Такими технологиями могут быть различные методы классификации, прогнозирования и моделирования, которые основаны на использовании деревьев принятия решений. Также это может быть использование нейросетей, генетических алгоритмов, применение статистических методов.

Далее приведем перечень основных задач, которые можно решить с помощью Data Mining: классификация, кластеризация, ассоциация, регрессионный анализ, а также анализ отклонений. Далее дадим определение каждому из этих понятий.

Классификация – это распределение данных по заранее определенным классам.

Кластеризация – определение схожести данных друг с другом и на основе этого происходит разделение данных на группы.

Ассоциация – поиск повторяющихся данных.

Регрессионный анализ – поиск факторов, влияющих на заданный параметр.

Анализ отклонений – поиск необычных данных, которые сильно отличаются от обычных.

Рассмотрим некоторые проблемы, связанные с Data Mining [34]. Эти проблемы можно разделить на три группы: методология Data Mining и взаимодействие с пользователем, проблемы с производительностью и проблема разнообразия типов данных. Рассмотрим каждую проблему более детально.

Методология Data Mining и взаимодействие с пользователем. Эта группа включает в себя такие трудности, как добыча разного рода знаний из базы данных, интерактивная добыча знаний на множественных уровнях абстракции, включение базовых знаний. Также к трудностям работы с Data Mining можно отнести язык запросов для Data Mining, сложность представления результатов работы, обработка неполных данных.

Если говорить о проблеме с производительностью, то она включает в себя эффективность и масштабируемость алгоритма Data Mining, а также параллельность работы алгоритма, распределения и инкрементальный алгоритм добычи данных.

И еще одна существующая проблема с Data Mining – это проблема разнообразия типов данных, которая в себя включает обработку реляционных и комплексных типов данных, а также добычу информации из неоднородных баз данных и глобальной информационной системы.

К плюсам данного метода можно относительную простоту нахождения множества готовых решений для классификации и кластеризации данных, присутствует возможность выбрать наиболее подходящий алгоритм классификации или кластеризации из множества существующих алгоритмов, возможность использования деревьев принятия решений.

После рассмотрения самого понятия Data Mining и того, как этот метод работает, поговорим о том, где он может применяться. Метод Data Mining применяется в основном там, где в большом объеме данных надо найти какие-то закономерности или тенденции. Большинство задач с большими данными в современных организациях можно привести к какому-либо классу задач Data Mining или комбинации таких классов задач.

Вкупе рассмотрев метод Data Mining, можно сказать, что для нашего исследования он не совсем подходит, так как есть необходимость работать с данными разного рода одновременно, есть вероятность того, что понадобится интерактивно представить результаты работы, так же есть проблемы с обработкой неполных данных, а такой случай тоже возможен, при работе с данными о людях. К тому же есть проблемы с параллельной работой этого метода, что тоже не приветствуется, при работе с большими данными.

Обратим наше внимание на краудсорсинг, выявляя плюсы и минусы данного метода. По большому счету краудсорсинг – это ручной анализ данных большим количеством людей. Давайте разберемся, как это работает.

Изначально у нас есть большой объем данных, которые неоднородны по способу записи, или имеют еще какие-либо недостатки, например одно и тоже название позиции может быть записано по-разному: на английском или русском, с сокращением или без. Набирается группа людей, которая вручную приводит эти данные к единообразному виду.

Рассмотрим плюсы и минусы данного метода. Начнем с минусов. Большим количеством людей сложно управлять, их сложно организовать, к тому же возможны утечки информации. Теперь перейдем к плюсам данного метода. Мотивация людей к работе, хотя это сложно отнести к плюсам, потому что не всегда и не все мотивированы к работе. Привлечение большого числа специалистов обеспечивает быстроту нахождения решения задачи, так как предлагается множество вариантов решения задачи. В процессе работы участвуют множество специалистов, что тоже непременно дает свои преимущества.

Метод работы с People Data применяется, при условии, что задача разовая, и для нее не рентабельно создавать сложную систему искусственного интеллекта.

Основываясь, на написанном выше, можно сказать, что данный метод не подходит для работы с People Data, потому что данные о людях – это очень конфиденциальная информация [42], [30], [32] и крайне нежелательно, чтобы

произошла ее утечка или информация попала не в те руки, так как есть риск нарушить закон о персональных данных.

Метод смешения и интеграции данных. Зачастую данные приходят из разных мест в разных форматах. Помимо краудсорсинга в такой ситуации может также помочь два шага: первый шаг – смешение и второй шаг – интеграция данных, то есть алгоритм преобразования разнотипной информации к единому знаменателю.

Для того, чтобы можно было использовать данные в разном формате из разных источников, применяются следующие шаги.

Сначала надо данные привести к единому формату. Этого можно достичь распознаванием текста и его конвертацией в документы, а затем из этих документов берется текст и переводится в цифры.

Далее данные могут быть дополнены в случае, если есть несколько источников данных об одном и том же объекте, например информация из первого источника может быть дополнена информацией из второго источника, для получения более полной картины.

И на завершающем этапе отсеиваются данные, которые являются избыточными. Например, в случае сбора не нужной информации каким-либо из источников, которая не доступна для анализа, то ее удаляют.

Рассмотрим достоинства и недостатки данного метода. Начнем с возможных минусов. При неоднородности или неполноте данных могут возникнуть проблемы, также могут отсеяться нужные данные. Помимо того, при использовании машинного распознавания текста могут возникнуть не точности его интерпретации и перевода. Перейдем к плюсам. Отсутствие ошибок людей, быстрота обработки данных, маловероятность утечки данных о людях.

Смешение и интеграцию данных можно применить, при наличии нескольких источников данных, которые нужно анализировать в комплексе.

Основываясь на плюсах и минусах данного подхода, можно сказать, что этот вариант лучше, чем краудсорсинг, но все же не совсем подходит для

обработки People Data, так как некоторые данные могут быть неправильно распознаны, а соответственно и неправильно обработаны.

Теперь обратим наше внимание на метод машинного обучения и нейронных сетей. Рассмотрим его достоинства и недостатки. Нейронные сети строятся из множества искусственных нейронов, образующих связи. Эти нейроны вкупе со связями после обучения на обучающих выборках, могут анализировать поступающие данные на вход нейронной сети.

Немного опишем, как работает нейронная сеть. В нейронных сетях для обучения может использоваться множество различных алгоритмов. Не будем на них останавливаться подробно, так как это не является темой данного исследования. Но работа самой нейронной сети сводится к следующему.

Нейронная сеть на вход получает данные, прогоняет их через свои слои нейронов, на выходе получается результат, например принадлежат ли данные к конкретной группе или нет.

Рассмотрим плюсы и минусы данного метода. Начнем с минусов. Точность работы данного метода сильно зависит от выбранного алгоритма машинного обучения, а также от обучающей выборки. Есть вероятность того, что не получится с первого раза достичь желаемого результата обучения сети. Из плюсов можно отметить, что конфиденциальность данных останется на высоком уровне, очень мал шанс утечки данных о людях, также к плюсам можно отнести отсутствие человеческого фактора [20], [21].

Данный метод работы можно применять, как альтернативу краудсорсингу, когда нужно сортировать, классифицировать данные, чтобы на основе этой классификации принимать какие-либо решения.

Основываясь на вышесказанном, можно сделать вывод, что нейронные сети и машинное обучение может не плохо справиться с работой с People Data, но все же не совсем подходит, так как возможно долгое обучение нейронной сети, возможна его низкая точность, что не приемлемо для работы с данными о людях. Так как ошибки в обработке этих данных могут стать критичными.

Прогнозная аналитика, или как его еще называют, предиктивная аналитика. Обычно этот метод применяется для составления прогнозов на будущее, основываясь на классификации и анализе старых данных [31].

Давайте разберемся, как работает предиктивная аналитика. Главной задачей прогнозной аналитики является определение набора параметров, которые воздействуют на данные. После выделения этих параметров нужно изучить, как их изменение влияло на результат в прошлом, на основе анализа таких результатов, можно будет делать выводы о том, к какому результату в будущем приведет изменение тех или иных параметров в настоящем. Для анализа вероятности наступления определенного события в результате изменения тех или иных параметров обычно используется нейронная сеть.

Рассмотрим плюсы и минусы данного метода. По факту он имеет те же плюсы и минусы, что и машинное обучение, поэтому для нас он тоже не очень подходит.

Этот метод применим для тех, случаев, когда нужно строить прогнозы. Сегодня предиктивную аналитику используют для того, чтобы предугадывать: количество продаж, а также поведение клиентов в маркетинге, время груза в пути в логистике, выявления мошенников в банковской и страховой сферах, экономический рост компании и финансовые показатели в любых сферах [36]. Как мы видим, для аналитики данных о людях, да и для обработки данных о людях, этот метод очень слабо применим.

Исходя из вышенаписанного, скажем, что этот метод мы тоже не будем использовать для работы с People Data.

Далее перейдем к рассмотрению возможности использования имитационного моделирования. Бывают ситуации, когда нужно увидеть, как будут вести себя зависимые величины, при изменении тех или иных показателей. Чтобы не ставить эксперименты в реальном мире, потому что это может привести к серьезным последствиям, есть имитационное моделирование.

Для имитационного моделирования мы строим математическую модель нашей ситуации со множеством факторов, потом меняем факторы, чтобы узнать, как это может повлиять на то, что мы исследуем [6].

Имитационное моделирование чем-то похоже на прогнозную аналитику, только в этом случае мы делаем выводы не на основе реальных данных, а на основе гипотетических данных.

Рассмотрим достоинства и недостатки данного метода. Начнем с минусов. Точность модели зависит от многих факторов, метод сам по себе довольно сложный. Из плюсов можно отметить можно выявить на начальном этапе первичного анализа большинство исследовательских ошибок. При построении модели мы не подвергаем риску реальные объекты исследования, так же этот метод хорош, когда мы не можем в силу объективных причин проверить теорию в реальном мире.

Давайте рассмотрим, где может применяться имитационное моделирование. Этот метод применим в тех случаях, когда необходимо проверить какие-либо предположения, но проверять это на реальном объекте будет слишком дорого.

Исходя из вышесказанного, можно сделать вывод о том, что имитационное моделирование совсем не подходит для работы с People Data, так как не позволяет обрабатывать большие данные в том ключе, который рассматривается в данной работе, хотя и нам нужно будет прогнозировать показатели о людях. Но прогнозы будут уже стояться на основе обработки больших данных о людях с применением конкретных технологий, которые мы выберем в данной работе. Имитационное моделирование, скорее может работать с большим количеством данных для построения математической модели объекта для исследования самого объекта, его поведения, но никак не для обработки больших данных и нахождения в них каких-либо закономерностей. Поэтому мы не будем использовать данный метод в настоящем исследовании.

Далее перейдем к рассмотрению применения статистического анализа для работы с People Data. Смысл данного метода сводится к тому, чтобы собрать входные данные, провести с ними расчеты по заранее определенным параметрам и в итоге получить некий результат в процентах.

Давайте посмотрим, как работает этот метод. Чтобы получить точные статистические данные есть множество методов. Перечислим некоторые из этих методов: расчет процентного соотношения; расчет средних значений по данным, эти данные иногда могут быть разбиты на группы; корреляционный анализ – дает возможность найти связи и помогает определить каким образом изменение тех или иных данных может оставить след на других данных; а также метод динамических рядов – оценивает частоту и интенсивность изменения данных на временном промежутке.

Рассмотрим плюсы и минусы данного метода. Начнем с недостатков данного метода. Главным минусом данного метода является зависимость его точности от объема выборки. Конечно, применительно к большим данным это не проблема, так как их объем гарантирует точность. Но стоит отметить, что не стоит всегда полагаться на статистику, так как всегда есть случаи, которые не вошли в большинство и рассматриваемый случай может оказаться именно таким. Из плюсов данного подхода можно выделить высокую точность на больших объемах данных.

Обычно данный метод применяется, когда нужно агрегировать данные для анализа, но в большинстве своем этот метод используется как часть других технологий.

Рассмотрев данный метод, можно сказать, что мы его будем использовать, как вспомогательный к основному в нашем исследовании для того, чтобы как раз сделать некие выводы, исходя из уже обработанных больших данных, которые готовы к анализу.

Теперь попробуем понять на сколько метод визуализации аналитических данных подходят для работы с People Data. Для удобства разбора и оценки результатов аналитики используется визуализация этих

самых данных. Визуализировать их можно с помощью графиков, диаграмм, гистограмм, трехмерных моделей, карт и пиктограмм.

Давайте разберемся, как с этим работать. В основном визуализация выступает заключительным этапом, демонстрацией результатов анализа, который проводился другими способами. Обычно для визуализации результатов используются такие как Qlik, Orange и Tableau.

Главным минусом данного метода можно отнести его несамостоятельность, то есть его можно применять только вкупе с другими методами. К плюсам данного метода можно отнести наглядность представляемых результатов – либо результатов обработки данных, либо уже результатов аналитики обработанных данных.

Данный метод можно применять в тех случаях, когда нужно визуализировать данные для людей.

Можно сказать, что этот метод мы тоже будем, скорее всего, применять в нашем исследовании для более наглядного представления результатов.

Последний метод, кандидатуру которого мы рассмотрим для работы с People Data, будет пространственный анализ. По сути, это набор методов, которые используют различную информацию, такую как топологическую, геометрическую и географическую, которая получена из входных данных. Для анализа таких данных используется статистический анализ.

Плюсы и минусы данного метода схожи со статистическим анализом. Данный метод мы не будем применять в нашем исследовании, так как он очень схож со статистическим анализом, который и так будет применен.

После анализа и выбора методов работы с People Data, можно перейти к анализу плюсов и минусов технологий работы с People Data, а также выбору наиболее подходящей, которая будет использоваться в настоящий работе.

Начнем с такой технологии, как NoSQL. Реляционные базы данных придерживаются принципов ACID: атомарность, согласованность, изолированность, надежность. А вот NoSQL базы данных не используют эти принципы, или использует их частично. У NoSQL баз данных есть свой набор

свойств: базовая доступность, гибкое состояние, согласованность в конечном счете. Рассмотрим каждое свойство отдельно.

Базовая доступность говорит нам о том, что любой запрос к базе данных завершится.

Гибкое состояние – означает, что с течением времени может меняться состояние системы, в том числе, если никаких действий с ней не производилось для достижения согласованности.

Согласованность в конечном счете – в какой-то момент времени данные могут быть не согласованы, но в конечном счете они становятся согласованными.

Минусами данного подхода является сложность работы в таких системах, малое количество бесплатных NoSQL баз данных.

Плюсами данного подхода можно считать возможность использовать множество типов хранилищ, нет необходимости использовать схему, при добавлении новых процессоров увеличивается производительность, можно использовать не только SQL.

Рассмотрев технологию NoSQL, можно сказать, что для работы с People Data, она слабо подходит, так как лишь позволяет хранить большие объемы неструктурированных данных, но не занимается их обработкой. Поэтому мы не будем применять данную технологию в нашем исследовании.

Далее рассмотрим технологию, даже не совсем технологию, как уже говорилось в данной работе, а язык программирования для работы с большими данными – язык программирования R. Этот язык программирования применяется для работы с большими данными, их обработки. Также этот язык программирования позволяет визуализировать результаты обработки больших данных. Эта технология применяется в разработке статистических программ.

К плюсам данной технологии стоит отнести направленность на обработку больших данных. А из минусов стоит отметить, что очень мало специалистов в этом языке на сегодняшний день. Сегодня будет более перспективно использовать Java с его фреймворками для работы с большими

данными. О них будет рассказано далее. Эту технологию мы не будем использовать в нашем исследовании по вышеприведенной причине.

Теперь обратим наше внимание на такую технологию работы с большими данными, как аппаратные решения. Такие решения представляют из себя готовые к работе машины, в которых установлены программные комплексы, которые обрабатывают данные больших объемов методом массово-параллельной обработки. Помимо того, к аппаратным решениям относят аппаратные решения, которые обрабатывают данные в оперативной памяти аналитическим методом. К таким аппаратным решениям можно отнести комплексы Hana от SAP и комплекс Exalytics от Oracle.

Из плюсов стоит отметит тот факт, что это уже полностью готовое решение для работы с People Data, что очень хорошо. Но есть существенный минус, из-за которого мы не будем использовать данную технологию в нашей работе. Этим минусом является тот факт, что это платная технология, которую самому не сделать, ее придется только приобретать [29].

Исходя из вышесказанного про аппаратные решения, делаем вывод, что в нашем исследовании мы не будем использовать эту технологию.

Теперь рассмотрим технологию Hadoop, как кандидата для применения в настоящей работе. Основными задачами платформы Hadoop являются хранение, обработка и управление данными. Эта платформа была разработана Google. Она включает в себя множество технологий в том числе MapReduce, которые осуществляют работу с большими данными. Этакое готовое решение.

Из минусов стоит отметить, что парой не все модули нужны для конкретной задачи, а они присутствуют в этой технологии, то есть она иногда становится избыточна. Как в нашем случае.

Перейдем к плюсам решения на основе фреймворка Hadoop. К ним относятся: быстрота обработки входных данных, общая стоимость оборудования становится меньше, повышается отказоустойчивость, присутствует линейная масштабируемость, возможность работы со слабоструктурированными или вовсе не структурированными данными.

Исходя из разбора технологии Hadoop, можно сказать, что мы ее не будем использовать в данной работе ввиду избыточности функционала.

В заключении рассмотрим еще одну технологию, которая является частью технологии Hadoop – это технология MapReduce. MapReduce – это модель распределенной обработки данных, ее предложила компания Google для обработки больших данных на распределенный компьютерных кластерах.

MapReduce работает с данными, которые представлены в виде записей. Данные проходят обработку в три стадии: Map, Shuffle и Reduce.

К недостаткам данного подхода можно отнести надобность иметь несколько физических узлов для работы с большими данными.

Эта технология имеет те же плюсы, что и технология Hadoop, но также отличается простотой и удобством использования, так как скрывает от пользователя детали вычислений в кластерной системе. Так же не содержит излишних инструментов обработки.

Рассмотрев данную технологию, можно прийти к выводу, что эта технология будет оптимальной для использования в данной работе.

Вывод: подводя итог, можно сказать, что после анализа всех методов и технологий работы с People Data, мы остановились на следующем списке, который будет применен в ходе данного исследования: метод статистического анализа, также будет применен метод визуализации данных, а также технология MapReduce.

Решение о том, какие именно реализации статистического анализа и алгоритма MapReduce будут использоваться, будет сделано в следующих частях работы.

2.3 Применение технологии MapReduce для работы с People Data и анализ результатов обработки информации

После разбора всех плюсов и минусов методов и технологий работы с большими данными, а также выбора применяемых методов и технологии,

рассмотрим более детально технологию MapReduce. Как уже говорилось выше технология MapReduce состоит из следующих этапов: этап Map, этап Shuffle и этап Reduce [13]. Рассмотрим кратко каждый этап.

Этап Map. На этом этапе к входным данным применяется функция map(), которая задается пользователем самостоятельно. Главный критерий к пользовательской функции – это чтобы функция выполняла преобразование входного потока данных, которые являются строками, в данные, которые представляют из себя карту ключ-значение, где ключ должен быть уникальный, а значения могут повторяться. Опционально можно добавить условие фильтрации к данным. Функция map() применяется к каждой записи во входном массиве. Эту функцию можно реализовать с помощью Stream API.

Стоит отметить, что для каждого значения может быть несколько ключей, так как значения могут повторяться, а ключи – уникальны. По этой причине мы можем получить для каждого значения во входном массиве от одной до нескольких записей. Какие значения будут находиться в ключе, а какие в значении завит от пользовательской реализации. По этому ключу в конце будут формироваться группы, которые попадут в один и тот же экземпляр функции reduce().

На этапе Shuffle результаты функции map() распределяются по корзинам, в одну корзину попадают значения, у которых значение хэш-функции попадают в одинаковый диапазон значений. Новая корзина формируется, когда результат хэш-функции выходит за диапазон значений этой корзины.

На этапе Reduce все корзины, полученные на предыдущем этапе, проходят финальные вычисления значений, формируя новую карту значений, которая и будет результатом работы алгоритма. Реализацию функции reduce() так же задает сам пользователь.

Также отметим несколько важных моментов в применении технологии MapReduce.

Первое. Каждая функция map() не зависит от работы других функций map(). Из этого можно сдать вывод, что мы можем запустить несколько функций map() параллельно на разных кластерах.

Второе. Так же, как и функция map(), функция reduce() не зависит от работы других функций reduce() и может запускаться параллельно на разных кластерах.

Третье. Функция shuffle() тоже работает не зависимо от других реплик этой функции, поэтому мы ее так же можем запускать параллельно на разных машинах.

Все эти вещи дают возможность использовать горизонтальную масштабируемость, при применении технологии MapReduce. То есть будет одновременно работать и обрабатывать одновременно данные сразу несколько реплик алгоритма MapReduce.

Хорошей практикой является использование функции map() на машине, где находятся данные, даже можно сказать, что это является стандартной практикой. Так как это снижает временные затраты, которые мы получаем при передачи данных по сети [11]. Также нельзя не упомянуть, что MapReduce не работает с индексами.

Как можно увидеть из разбора технологии MapReduce, самая трудоемкая стадия – это стадия Shuffle. Это так, потому что результаты работы метода map записываются на диск, читаются с диска, что довольно затратно повремени, особенно если используется медленный HDD, а не быстрый SSD, затем сортируются и передаются по сети. В нашем случае будет не плохо сначала агрегировать результаты выходов нескольких методов map на одном узле map-reduce задачи, а затем уже в метод reduce передавать уже просуммированные значения по нескольким машинам.

Для этого надо будет реализовать комбинирующую функцию – дополнительный шаг combine, которая будет обрабатывать выход части мапперов. Такую функцию можно подключить из набора технологий Hadoop, а можно сделать самим. Данная функция очень похожа на reduce. Чтобы из

функции reduce получить combiner нужно просто подавать ей на вход не все пары ключ-значение с одинаковым ключом, а набор таких пар разделить на несколько частей и передавать по отдельности и далее по алгоритму. Также стоит помнить, что функция combine просто включается в цепочку алгоритма, а не заменяет собой шаг reduce. Таким образом можно оптимизировать работу технологии MapReduce. В нашем случае такой вариант будет наиболее эффективен в применении.

Вывод: разобрав все технологии и методы работы с People Data, и изучив более детально технологию MapReduce, можно подтвердить тот факт, что она действительно лучше всего подходит для применения в данном исследовании, хотя, конечно, у нее есть свои минусы. Самым главным минусом стоит отметить невозможность работы с индексами, наличие которых приветствуется, ведь речь идет о работе с большими данными.

А для анализа данных и получения выводов, как уже говорилось, будем использовать метод статистического анализа. По результатам статистического анализа будет производиться построение графиков, за которое будет отвечать метод визуализации данных.

На основе полученных графиков прогнозирования показателей о персонале будет производиться оценка эффективности работы персонала на предприятии и приниматься меры по увеличению эффективности работы персонала в последующих расчетных периодах.

Такую оценку эффективности работы персонала, а также принятие мер по увеличению эффективности работы персонала будут принимать аналитики организации, в которой будет применено данное решение.

3 Использование технологии MapReduce и метода статистического анализа для решения задачи оптимизации показателей о персонале

3.1 Описание технологии MapReduce

Технология MapReduce основана на одноименной парадигме MapReduce. Парадигма MapReduce, которая была разработана в 2000-х годах американской транснациональной корпорацией Google, это один из самых эффективных способов работы с большими данными для распределенных систем [22], [29]. Изначально такой подход был придуман для обработки огромных страниц из Интернета. [39], [25].

Первое использование такого подхода было применено для GFS (Google File System) и в первой не реляционной базе данных для обработки больших данных от Google – Big Table [24], [31].

Этот подход для работы с большими данными скрывает от пользователя все детали обработки данных и их вычисления в кластерной системе [17].

Несомненным плюсом MapReduce можно выделить возможность выполнять обработку и свертку данных параллельно на нескольких кластерах, так как эти операции не зависят друг от друга.

Главный критерий для возможности проведения операции свертки параллельно – это все результаты Shuffle из одной корзины должны выполняться одной задачей Reduce.

Поскольку технология MapReduce использует параллельную обработку данных, у нас появляется возможность восстановления после сбоя одной или нескольких под сервера [13]. Например, если на какой-то реплике произошел сбой, то можно сделать перебалансировку нагрузки между подами, и работа вышедшей из строя поды будет передана наиболее свободной поде или распределена между несколькими подами, но при условии, необходимы данные для проведения шага алгоритма доступны.

Со стороны пользователя для работы парадигмы MapReduce необходимо будет реализовать всего несколько функций [15]. А балансировка нагрузки, обработки отказов под и контроль взаимодействия реплик приложения система берет на себя [43].

Для работы технологии MapReduce нужно, чтобы данные для обработки имели вид упорядоченного списка [11]. Входящий список будет обработан в три стадии: Map, Shuffle и Reduce, как показано на рисунке 1 [26].

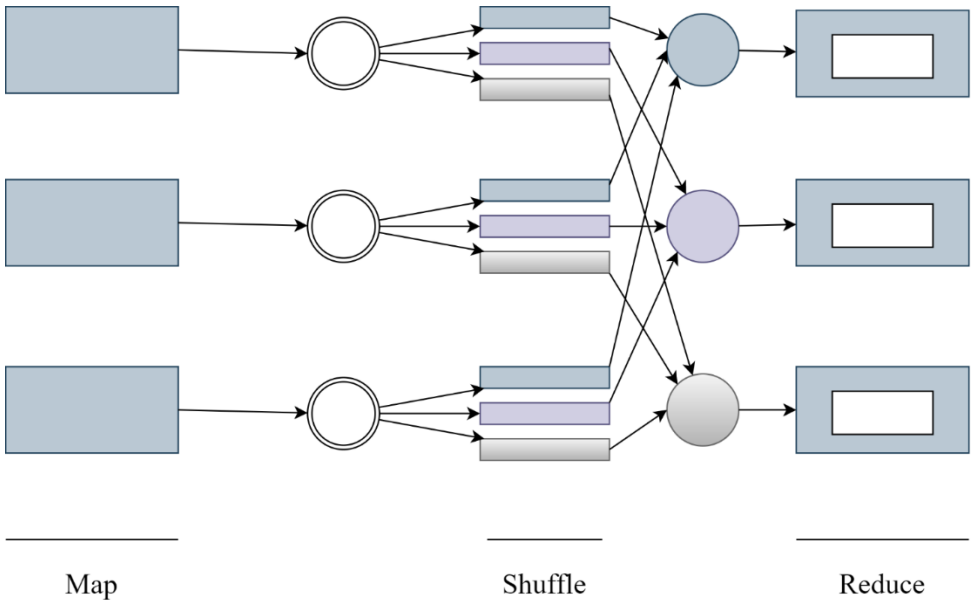

Рисунок 1 – Схема работы технологии MapReduce

Опишем все стадии работы алгоритма MapReduce.

Стадия Map производит преобразование входных данных в коллекцию значений, которая представляет из себя карту, где в качестве ключа лежит идентификатор, который является уникальным, а в качестве значения лежит любой пользовательский объект. Функцию map() задает сам пользователь, исходя из своих нужд. Эта функция очень похожа на функции map в языках программирования. Эта функция применяется к каждой записи из входного потока, который является массивом. В результате применения этой функции мы получаем множество пар ключ-значение. Как упоминалось ранее, операция

предварительной обработки данных может выполняться параллельно, а поскольку функция Map и есть шаг предварительной обработки данных, значит, мы можем распараллелить работу этой функции между несколькими подами на одной ЭВМ. Главным условием работы является доступность всех необходимых ресурсов для вычислений.

Идеальным случаем можно считать, когда метод map() обрабатывает данные с той же машины, на которой он запущен, так как это значительно снижает нагрузку на сеть, поскольку нет необходимости с одной физической машины передавать данные на другую физическую машину. Благодаря такому подходу сохраняется принцип локальности данных.

Далее рассмотрим шаг Shuffle. На этом шаге происходит разбиение выхода функции Map на отдельные бакеты. В каждом бакете лежат только записи с одним ключом вывода, который был получен на шаге Map. Метод shuffle() для работы использует входные корзины и распределяет эти корзины по репликам функции reduce(), в соответствии с ключом карты и числом корзин во входящей карте.

На выходе shuffle() выдает новую корзину значений, где в качестве ключа лежит номер выполняемой задачи. В том числе по этим номерам будут запускаться задачи reduce(). Ключ генерируется на основе функции хэширования.

Скорость работы shuffle() и ее эффективность сильно зависят от используемой функции хэширования, а также полей, которые участвуют в вычислении хэш-значения. Для увеличения скорости работы этой функции нужно подобрать максимально эффективную функцию хэширования и наиболее быстрый алгоритм параллельной сортировки. Коллизии разрешаются с помощью сортировки записей, и чем эффективнее функция хэширования и алгоритм сортировки, тем меньше вероятность возникновения коллизии.

На этапе Reduce по входящей карте проходятся итератором по всем корзинам и происходит вычисление конечного результата, после чего

формируется итоговая карта значений. Шаг Reduce, как и Map может быть распараллелен. А это значит, что эту функцию можно выполнять на разных кластерах (разных физических машинах).

Иногда для обработки данных не требуется стадия Reduce. После шага Shuffle возвращает набор отсортированных пар, которые были получены на стадиях Map и Shuffle.

Также стоит отметить, что технология MapReduce включает возможность расширения своего функционала. Дает возможность включать некие промежуточные шаги, которые могут повысить эффективность обработки информации.

Покажем возможные модификации алгоритма MapReduce посредством включения дополнительных вычислений.

Можно добавить использование функции combine(). Она применяется в том случае, если шаг Map имеет проблемы с решением коллизий функции хэширования. А функция reduce() была реализована таким образом, что ей присуще хотя бы одна из особенностей: как коммутативность или ассоциативность. При выполнении хотя бы одного из этих условий, функция combine() выполнит промежуточную агрегацию данных во временную карту с более эффективным расположением записей. После чего передаст выходной результат дальше по цепочке.

Чтобы еще больше увеличить эффективность алгоритма, стоит выполнять функцию combine() на той же машине, на который была произведена функция map(), это необходимо по той же причине, что и выполнение функции map() на машине, где хранятся данные. После выполнения промежуточной агрегации данных в карту значений с более эффективным расположением записей, все это помещается в файл для передачи по сети. Файл с результатами дальше отправляется на реплику или ЭВМ, где развернута функция shuffle() или reduce() в зависимости от модификации алгоритма, о которых мы поговорим далее.

Это то, что как раз нам нужно для решения нашей задачи, так как у нас может быть множество повторяющихся значений промежуточного ключа, так как данных, которые нужно обработать, может быть очень много даже по меркам BigData, так как эта технология нацелена на применение не только в маленьких фирмах, но и в больших корпорациях [8]. А пользовательская функция Reduce может запросто иметь такие свойства, как коммутативность и ассоциативность [4].

Давайте подумаем, а можем ли мы как-то упростить наше решение, чтобы давало меньшую нагрузку на сеть и железо. Рассмотрим несколько существующих вариантов.

Сегодня технология MapReduce используется часто в прикладных решениях. К таким решениям можно отнести фильтрацию данных, их упорядочивание по тем или иным критериям, обработку большого объема текста, вычисление индексов для документов, использование в искусственном интеллекте, подсчет индекса Хирша, использование при статистическом анализе [1].

И это далеко не все задачи, в которых может применяться технология MapReduce. Привести классификацию задач, где может использоваться данная технологий очень сложно, потому что многие задачи сильно связаны между собой. Но что бы хоть как-то упорядочить задачи, в которых может применяться MapReduce, давайте введем критерий. Он будет отображать стадии работы с данными, при применении MapReduce.

Существует совокупность задач, в которых необходимо применять MapReduce.

Самым первым типом задач можно назвать MapReduce. В этом типе задач применяются алгоритмы, которые, как минимум, используют стадию Map и стадию Reduce. Зачастую для выполнения реальных задач нет надобности использовать другие стадии работы алгоритма. В таком случае алгоритм работы будет содержать следующие пункты:

– В map() поступает список данных.

– Метод map() в потоке вызывается для каждого элемента входного массива и формирует промежуточную карту с результатами работы.

– Данные группируются по корзинам в соответствии со значением ключа.

– Промежуточная карта с результатами передается в функцию reduce().

– Метод reduce() производит окончательную обработку входных корзин данных. Часто результатом работы алгоритма бывает одно значение, например произведение.

Но в нашем случае шаги Combine и Shuffle как раз-таки нужны, так что решаемая задача не относится к этому классу.

Далее рассмотрим класс MapOnly. В класс MapOnly входят задачи, которые можно решить только шагом Map, как показано на рисунке 2.

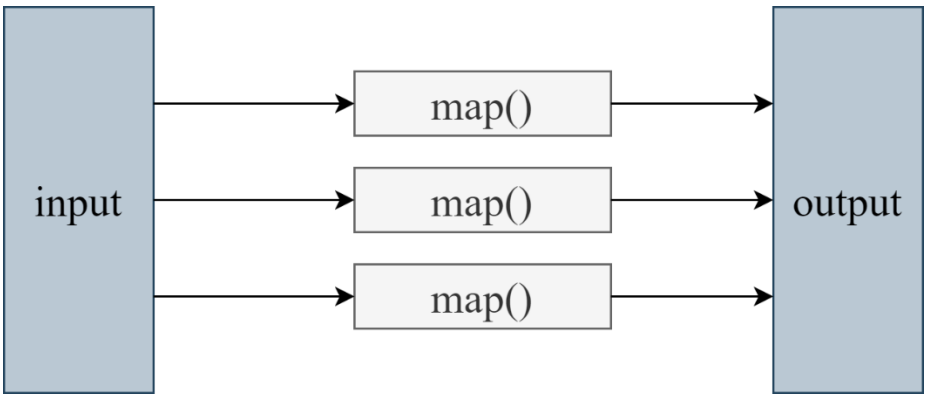

Рисунок 2 – Схема работы технологии MapOnly

К таким задачам можно отнести:

– преобразование данных из одного вида представления в другой, например изменение регистра или преобразование xml файла в json файл,

– выборка данных по условию из входной последовательности, например поиск информации некоторому фильтру,

– скачивание или загрузка данных из какого-либо хранилища, например сохранение или получение данных из базы данных.

В классе задач MapOnly происходит только получение данных, а значит шаги Combine, Shuffle и Reduce тут не используются, так как эти шаги предназначены для обработки данных. В нашем исследовании как раз требуется обработка данных, так что решаемая задача не относится к классу MapOnly. А это говорит о том, что мы не можем упросить алгоритм работы с большими данными описанным выше способом.

Перейдем к рассмотрению класса задач, который называется цепочки MapReduce. К этому классу задач относятся задачи, для решения которых одного прогона полного алгоритма MapReduce недостаточно. В таком случае нужно применить полный алгоритм MapReduce несколько раз. Такие цепочки задач MapReduce могут выполняться линейно или образовывать собой направленный нециклический граф.

Если цепочка предполагает линейное выполнение, то самым простым способом это реализовать будет – последовательный запуск задач MapReduce/ Каждая следующая задача MapReduce начнет выполняться только, после успешного завершения предыдущей задачи.

Если же цепочка задач MapReduce предполагает использование графа. Для построения такого графа нужно организовать поток выполнения задач MapReduce таким образом, чтобы сохранить целостность данных. Для достижения сохранности целостности данных необходимо не нарушать зависимость вызовов задач MapReduce между собой. Если хотя бы одна задача MapReduce не будет выполнена, то нужно прекратить выстраивание зависимостей задач и прервать выполнение цепочки задач MapReduce. Схематическое описание выполнения цепочки задач MapReduce показано на рисунке 3.

Решаемая проблема в данной работе, подпадает под этот класс задач, но использоваться в настоящем исследовании не будет, потому что этот подход

слишком нагромождает вычисления, а наша задача – оптимизировать процесс применения технологии MapReduce.

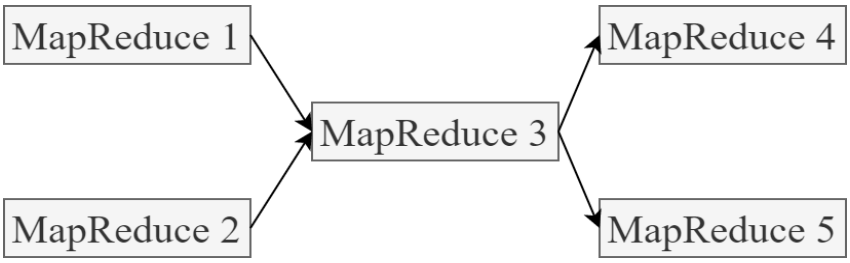

Рисунок 3 - Схема работы технологии цепочки MapReduce

Перейдем к рассмотрению такого типа задач, как ReduceJoin. Такой тип задач используется, когда информацию из множества источников нужно собрать в одном документе по заданному ключу. После группировки получаем ключ, который выдаем на выход вместе с самим объединенным документом. Объединяется информация, как правило, по какому-либо критерию. Результат применения ReduceJoin напоминает результат оператора JOIN в реляционных базах данных. JOIN объединяет данные нескольких таблиц по заданным условиям и внешним ключам. Примером задачи для такого типа может послужить задача объединения информации из нескольких объемных таблиц в одну временную таблицу для анализа данных. Также такую технологию можно использовать для ведения аудита сервиса, который работает на нескольких подах.

Опишем обязательные шаги алгоритма ReduceJoin:

– в функцию в качестве входных параметров поступает массив коллекций данных,

– отдельная задача MapOnly запускается для каждой коллекции, которая пришла во входных параметрах. Каждая задача MapOnly превращает полученную на вход коллекцию в карту значений. За ключ в данном случае принимается поле объекта из коллекции, по которому мы хотим объединить записи. За значение будем принимать карту, в которой в качестве ключа будет

лежать тип коллекции (Type), а за значение будем принимать данные, которые зависят от ключа и не могу быть отделены от него (Value),

– результат работы задачи MapOnly поступает в качестве входного параметра следующей задачи MapReduce. Снова получаем цепочку задач из задач MapReduce. Такая цепочка задач обязана иметь функцию map(), которая ничего не делает, кроме копирования данных в новую коллекцию. На этапе работы функции shuffle() происходит сортировка входной коллекции по значению ключа, образуя коллекцию с перераспределенными данными по корзинам. После чего, полученная коллекция передается на вход в функцию reduce().

Схематическое представление схемы работы модели ReduceJoin представлено на рисунке 4.

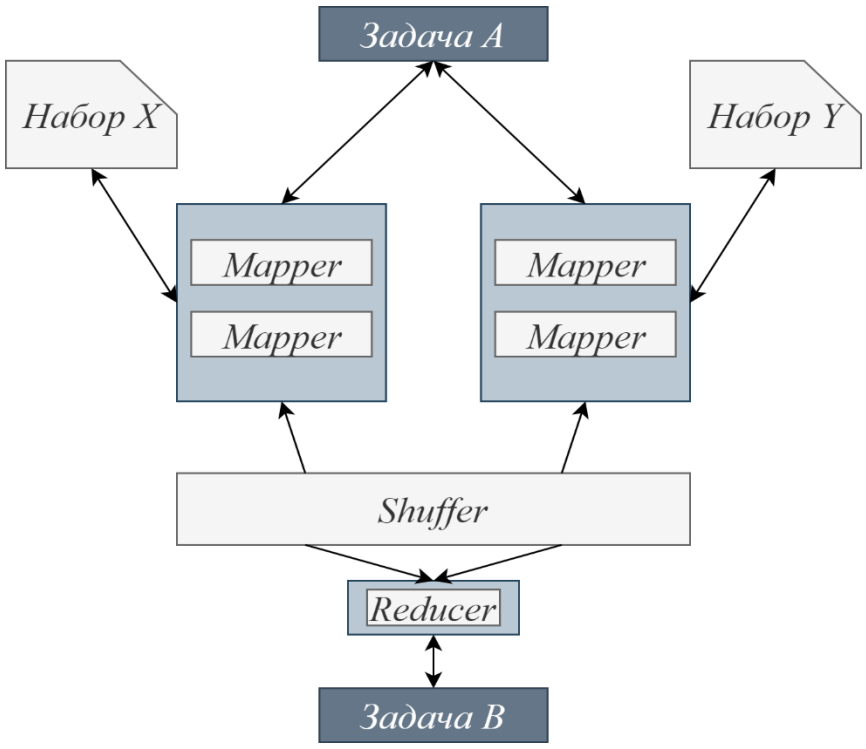

Рисунок 4 – Схема работы технологии ReduceJoin

В классе задач ReduceJoin в ходе решения используется класс задач MapOnly, поэтому задача, решаемая в данной работе, не относится к классу

задач ReduceJoin, а это говорит о том, что мы не можем упросить алгоритм работы с большими данными описанным выше способом.

Стоит отметить, что технология MapReduce предъявляет ряд требований к своей реализации и железу, на котором она будет выполняться. Например, для корректной работы этого подхода необходима реализация балансировщика задач, который будет распределять запросы от той или иной реплики распределенной файловой системы на ту или иную реплику реализации алгоритма MapReduce. За счет чего будет выполняться распараллеливание работы между несколькими репликами. Помимо того, этот балансировщик должен обрабатывать отказы той или иной реплики и перенаправлять их на наиболее подходящую, да так, чтобы клиент не заметил никаких особых задержек. Также напомним, что технология MapReduce не умеет работать с индексами, а это означает полную выгрузку данных их хранилища, чтобы на машине не было ошибки OutOfMemoryException, нужно распределять все данные между разными репликами распределенной файловой системы, при том в каждой реплике должно быть данных не больше, чем есть свободной оперативной памяти на данной машине. Это будет залогом высокой производительности, возможности к горизонтальному масштабированию, надежности и доступности данных [7]. Это позволит работать большими объемами данных, наиболее эффективно управлять нагрузкой на сеть и даст возможность поддерживать высокую пропускную способность системы, а также дать возможность работать с неполными данными. Такой артефакт, как нарушение целостности данных является следствием использования NoSQL баз данных.

Все MapReduce задачи запускает балансировщик. Помимо того, балансировщик постоянно проверят статус всех реплик приложения, в зависимости от этого распределяет задачи между репликами. Для поддержания параллельной работы нескольких под, все данные разбиваются также по подам распределенной файловой системы. Потом из каждой поды распределенной файловой системы данные попадают в функцию map(),

преимущественно запущенную на той же поде. Вызовы функций shuffle(), combine() и reduce() распределяются в зависимости от значения ключа. Записи, которые попали в одну корзину, обрабатываются одной функцией Shuffle и далее по алгоритму.

Вывод: подводя итог, скажем, что в нашей модели мы будем использовать технологию MapReduce с дополнительным шагом Combine. Упростить алгоритм у нас не получается, поэтому будем его использовать как есть с учетом шага Combine.

3.2 Описание метода статистического анализа

Для анализа данных существует множество подходов [34]. Самым популярным подходом к анализу данных является статистический анализ. Статистический анализ ищет скрытые связи между анализируемыми данными [14].

Существует множество методов в статистическом анализе. Которые делятся на категориальные и количественные.

Метод – подход к реализации на практике чего-либо или исследование чего-либо с теоретической точки зрения [23].

Количественные данные – это данные, которые имеют строгую структуру, которая не изменяется. Для измерения количественных данных используют либо интервальную шкалу, либо шкалу отношений [32].

Категориальные данные – это качественные данные, в которых за ранее определено множество категорий и множество значений для этих категорий. Категориальные данные бывают либо, номинальными, либо порядковыми [35].

Номинальные данные применяются для расчета предметов, а порядковые данные используются для упорядочивания предметов в рамках категорий.

Есть еще одна классификация статистических методов. Она делит статистические методы на многомерные и одномерные [18]. Если нам нужно оценить выборку одним общим измерителем [30].

Одномерные методы различают по типу данных, которые подвергаются анализу [27]. То есть в одномерных методах могут анализироваться данные двух категорий – метрические данные и не метрические данные. На рисунке 5 можно увидеть классификацию одномерных методов.

Рисунок 5 – Одномерные методы статистического анализа

В данном исследовании мы не сможем применять одномерные статистические методы по той причине, что данные о людях имеют сложную структуру, поэтому для анализа данных о людях будет применяться больше одного показателя одновременно.

Теперь поговорим о многомерных методах статистического анализа. Их используют, как раз в том случае, когда с применением одного показателя не

проанализировать данные [36]. И тогда нам нужно использовать несколько критериев с разными шкалами оценки [33]. Если мы используем несколько измерителей, то они должны показывать разные стороны изучаемого объекта. Классификация многомерного метода показана на рисунке 6.

Рисунок 6 - Многомерные методы статистического анализа

Статистическое наблюдение имеет следующие отличительные черты: плановость, научная организованность, регулярность сбора данных [3]. Наблюдение происходит посредством фиксации изменений у любого изучаемого объекта [2].

Статистическое наблюдение бывает [6], [37]:

– отчетность – в этом случае данные для статистики приходят в статистические органы в виде отчетов от юридических лиц или муниципальных учреждений,

– организованное статистическое наблюдение – выполняется для поиска данных, которые могут обеспечить единство других данных.

Чтобы можно было, исходя из наблюдений, вычленить какие-то важные факторы исследования, то используется сводка.

Сводка – множество действий, при помощи которых определяются конкретные факторы. По итогу сводки получается набор факторов [21].

Сводка включает в себя множество показателей, которые составляют систему итогов. А эта система итогов формируется в статистические таблицы [5]. Все это дает возможность проследить закономерности, которым подчиняется исследуемый объект [20].

Обратим наше внимание на регрессионный анализ, а именно метод наименьших квадратов, так как при анализе данных о людях будет необходимо использовать более одного показателя, которые не зависят друг от друга.

Метод наименьших квадратов – это метод в математике, который применяется для решения задач посредством поиска минимальной суммы квадратов отклонений функций от переменных, которые предстоит найти. Этот метод может использоваться в том числе для аппроксимации точечных значений некоторой функции.

Метод наименьших квадратов выполняет подбор таких значений набора неизвестных параметров (x), чтобы значения используемых функций были, как можно больше, приближены к некоторым значениям y_i. Таким образом, мы говорим о поиске решения для системы уравнений (1), в которой будет наибольшая близость обеих частей уравнений друг к другу.

$$f_i(x) = y_i, \tag{1}$$

где $i = 1, \dots, m$

Суть метода наименьших квадратов заключается в выборе в качестве меры близости суммы квадратов отклонений левых и правых частей уравнения (1). То есть суть метода наименьших квадратов может быть выражена формулой (2).

$$\sum_i e_i^2 = \sum_i (y_i - f_i(x))^2 \to \min_x \tag{2}$$

Когда у системы уравнений есть хотя бы одно решение, тогда сумма квадратов с наименьшим значением будет равна нулю. При условии, что такая сумма будет найдена, тогда будет существовать точное решение системы уравнений.

Когда количество неизвестных переменных меньше числа независимых уравнений, тогда система не будет иметь точного решения. В таком случае метод наименьших квадратов позволит найти наиболее приближенный вектор x, который максимально приближает вектор y и функцию $f(x)$, или наиболее близкий вектор отклонений e к нулю.

Если же мы применяем метод наименьших квадратов, как часть регрессионного анализа, то необходимо пользоваться следующими правилами и формулами.

У нас есть некоторое количество значений (n) для переменной (y), которые соотносятся с соответствующими значениями переменной x. Решение методом наименьших квадратов в контексте регрессионного анализа сводится к аппроксимации взаимосвязи между y и x некоторой функцией $f(x, b)$, точность которой соответствует неизвестным параметрам b. То есть нам нужно найти наиболее удовлетворяющее значение параметров b, с помощью которого можно максимально близко приблизить значение $f(x, b)$ к реальному значению y. Все это можно свести к решению системы уравнений (3).

$$f(x_t, b) = y_t, \tag{3}$$

где $t = 1, \dots, n$

Сущность решения метода наименьших квадратов в данном случае сведется к нахождению b, при котором сумма квадратов отклонений e_t будет минимальной.

Чтобы решить задачу нахождения минимального значения функции, необходимо вычислить стационарные точки функции посредством

нахождения дифференциала по неизвестным параметрам b. Для этого надо приравнять производные к нулю и найти решение системы уравнений (4).

$$\sum_{t=1}^{n}\left(y_t - f(x_t, b)\right) \times \frac{\partial f(x_t, b)}{\partial b} = 0 \qquad (4)$$

Выполнив приведенные выше вычисления, получим выходные коэффициенты, с помощью которых сможем построить графики для наглядного анализа показателей о персонале.

Вывод: подводя итог рассмотрению метода статистического анализа, можно сказать, что в предлагаемой модели предпочтительнее всего будет использовать такой подход к статистическому анализу, как регрессионный анализ, а именно метод наименьших квадратов. Поскольку в случае с данными о людях одним показателем для измерения не обойтись, так как данные о людях могут иметь сложную структуру и для их анализа будет необходимо использовать несколько показателей – это объясняет необходимость использовать многомерные методы. А вот необходимость использовать регрессионный анализ обуславливается тем, что все эти показатели будут независимы друг от друга, а конечный результат будет зависеть от этих независимых показателей.

3.3 Описание предлагаемой модели для решения задачи оптимизации показателей о персонале

Описав и выбрав методы и технологии, которые мы будем использовать для проектирования собственной модели в рамках данного исследования, перейдем к описанию собственной модели, которая позволит решить задачу оптимизации показателей о персонале. Начнем с описания компонентов, которые нам понадобятся для решения поставленной задачи в рамках выбранных технологий и методов работы с данными о людях.

Предлагаемая модель подразумевает интеграцию в существующую информационную систему предприятия, поэтому будут описаны только те, компоненты, которые нужны непосредственно для работы встраиваемой подсистемы оптимизации прогнозирования показателей о персонале. Внутреннее устройство информационной системы, которая уже есть на предприятии, рассматривать не будем, обойдемся лишь отображением этой системы на диаграмме компонентов.

В предлагаемой модели будет пять компонентов, не считая внутренней информационной системы организации:

- планировщик задач MapReduce (балансировщик),
- модуль MapReduce,
- модуль статистического анализа,
- модуль визуализации статистических данных,
- распределенная файловая система.

Кратко опишем каждый компонент. Начнем с планировщика задач. Данный компонент будет заниматься распределением входных данных на реплики компонента MapReduce в соответствии с тем с какой машины, на которой стоит распределенная файловая система, пришли эти данные и какие узлы сейчас менее нагружены, чтобы минимизировать затрату ресурсов на передачу по сети и минимизировать нагрузку на машины, которые загружены большим количеством задач на выполнение. Помимо того, для более эффективной передачи и обработки, балансировщик разбивает данные на партиции.

Модуль MapReduce будет производить операции обработки данных по алгоритму, который будет рассмотрен далее.

Модуль статистического анализа будет применять метод наименьших квадратов к входным значениям, полученным от модуля MapReduce.

В свою очередь модуль визуализации статистических данных будет производить отображение полученных коэффициентов в виде диаграмм и предавать его во внутреннюю систему организации для отображения в

браузере пользователя. Для удобства обработки запросов и обращения к модели и отображения результатов будет использоваться шаблон проектирования Model-View-Controller. Наглядное представление описанных компонентов и их взаимосвязей можно увидеть на рисунке 7.

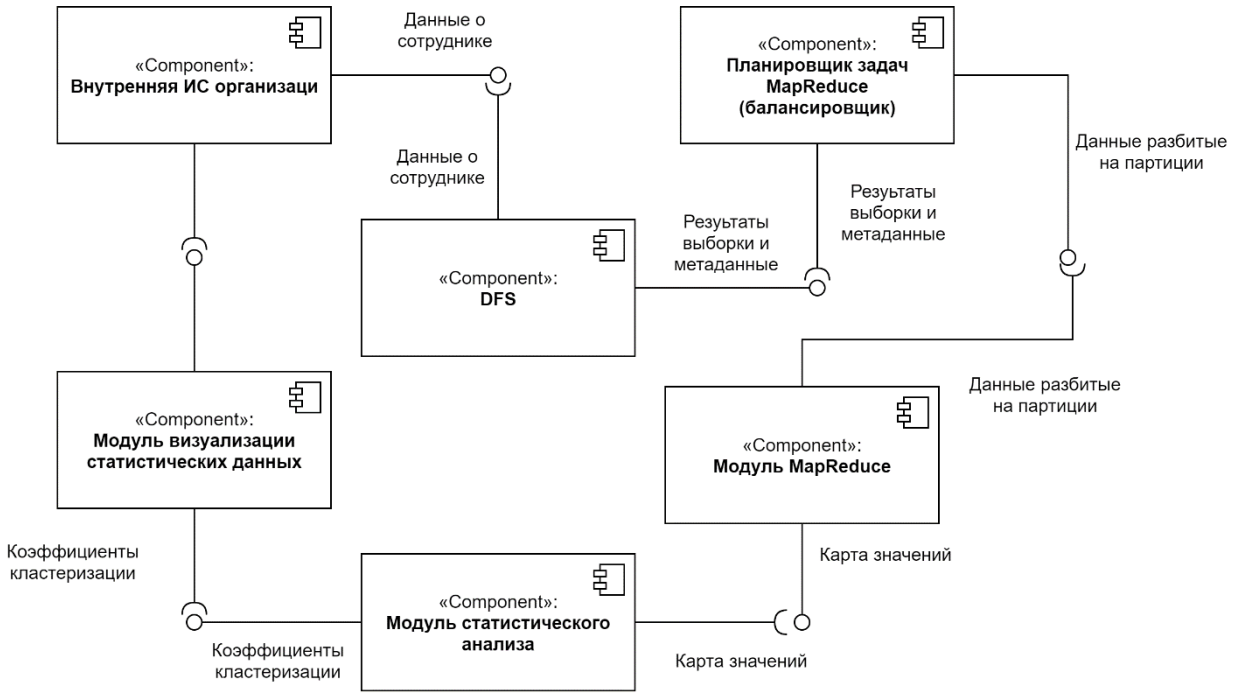

Рисунок 7 – Диаграмма компонентов

Для описываемой модели подсистемы оптимизации прогнозирования показателей о персонале было решено сделать диаграмму классов с несколькими пакетами, пакет API тоже должен быть, но отсутствует на схеме, так как относится к интеграции предложенной модели в существующую информационную систему и зависит от договоренностей об интерфейсах взаимодействия с внешней системой.

Описанные компоненты и их взаимосвязи программно можно представить в виде диаграммы классов, как показано на рисунке 8.

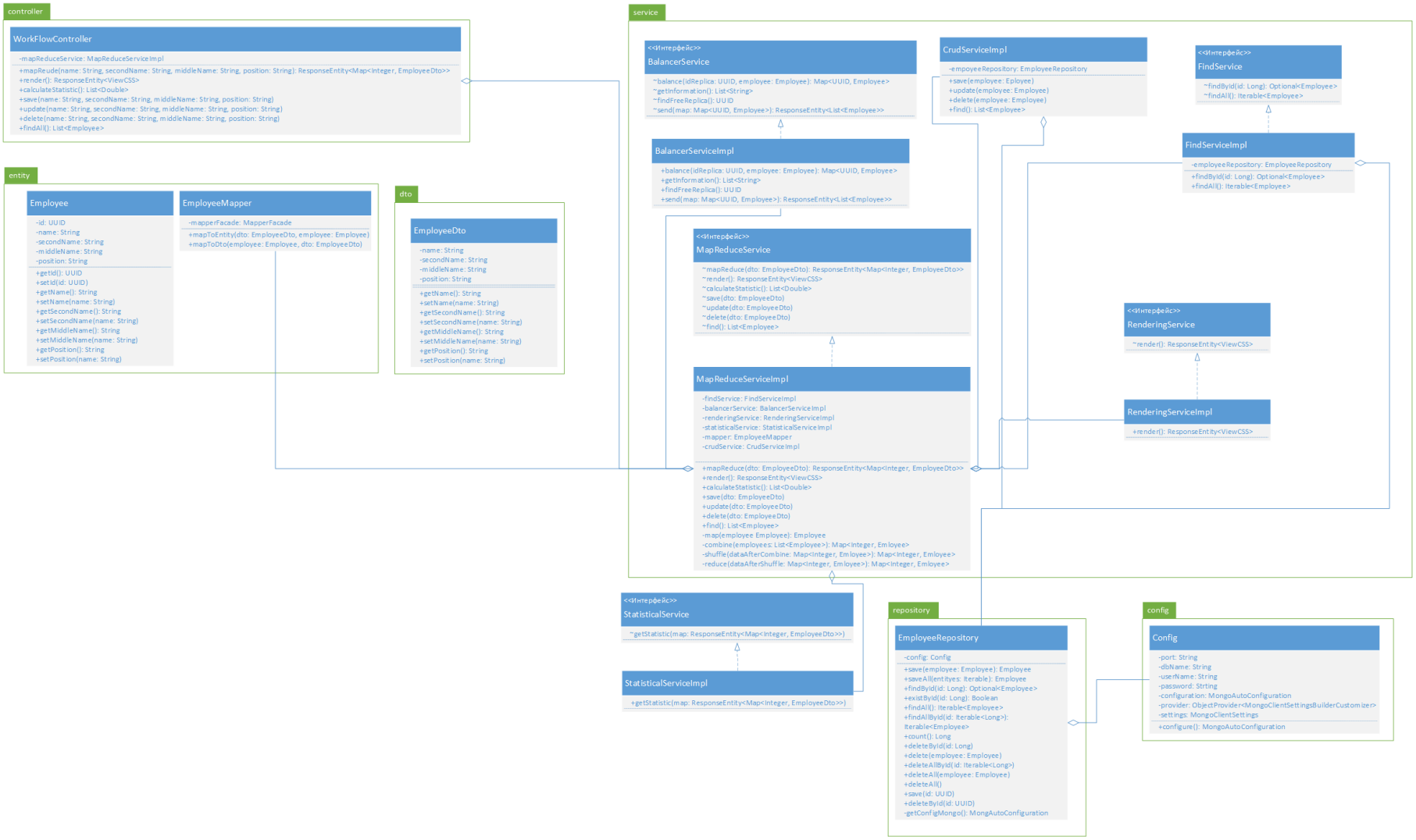

Рисунок 8 – Диаграмма классов

В пакете controller содержится основной контроллер WorkFlowController для обращения по REST к описываемой подсистеме и является презентативным слоем. В нем содержится единственный класс, который является rest controller’ом паттерна Model-View-Controller (MVC).

В свою очередь данный контроллер взаимодействует с сервисным слоем, который описан в пакете service. А сервисный слой обращается к репозиторию, который хранится в пакете repository и реализуется классом EmployeeRepository, для взаимодействия с базой данных. Сервисный слой и репозиторий образуют собой модель (model) паттерна MVC.

В сервисном слое у нас есть шесть сервисов, которые отвечают за работу предлагаемой модели. Сервис для алгоритма MapReduce, который обеспечивается классом MapReduceServiceImpl. Сервисы CrudServiceImple и FindServiceImpl, которые взаимодействуют с репозиторием – слоем персистентности. Сервис для распределения нагрузки между репликами MapReduce – BalancerServiceImpl.

Статистический сервис реализуется классом StatisticalServiceImpl, в котором применяется метод наименьших квадратов. И в заключении скажем про сервис отображения данных, полученных при регрессионном анализе – RenderingServiceImpl, в котором происходит отображение полученных коэффициентов методом наименьших квадратов на диаграммы.

Слой персистентности обеспечивается фреймоврком Spring Boot – его CRUD репозиторием, от который реализуется EmployeeRepository и обеспечивает связь с базой данных. Конфигурация базы данных находится в пакете config и осуществляется классом Config.

В свою очередь view генерируется при помощи ResponseEntity и jsp страниц, которые генерируются ViewResolver и передаются через контроллер на страницу браузера пользователя.

Для обмена данными между подсистемой и основной системой организации используются Data Transfer Object (DTO), который находится в

пакете dto. В самой же подсистеме используется сущность Employee, которая находится в пакете entity.

Для соответствия передаваемой внешней сущности и внутренней сущности используется маппер EmployeeMapper, который так же находится в пакете entity.

Разобравшись с внутренней структурой предлагаемой модели, давайте посмотрим, как это все будет работать по бизнес-процессам в связке с внутренней системой организации.

В бизнес-процессах у нас всего будет три независимых физически актора:

– пользователь, он же аналитик организации,
– внутренняя система организации,
– сама подсистема прогнозирования показателей о персонале.

Аналитик взаимодействует с внешней системой организации посредством графического интерфейса пользователя. Через который подает запрос в систему и получает ответ от системы.

Внутренняя система организации выполняет различный электронный документооборот и обеспечивает потребности организации в информационной системе. При запросе аналитиком функции прогнозирования оптимальных показателей о персонале, система организации обращается к подсистеме прогнозирования показателей о персонале для составления прогноза.

Подсистема прогнозирования показателей о персонале в свою очередь содержит несколько модулей, которые взаимодействуют между собой.

Как видно из модели, каждый модуль выполняет набор операций, позволяющий произвести прогнозирование показателей о персонале на основе входных данных, полученных на предыдущем шаге.

Все процессы описаны BPMN диаграммой на рисунке 9.

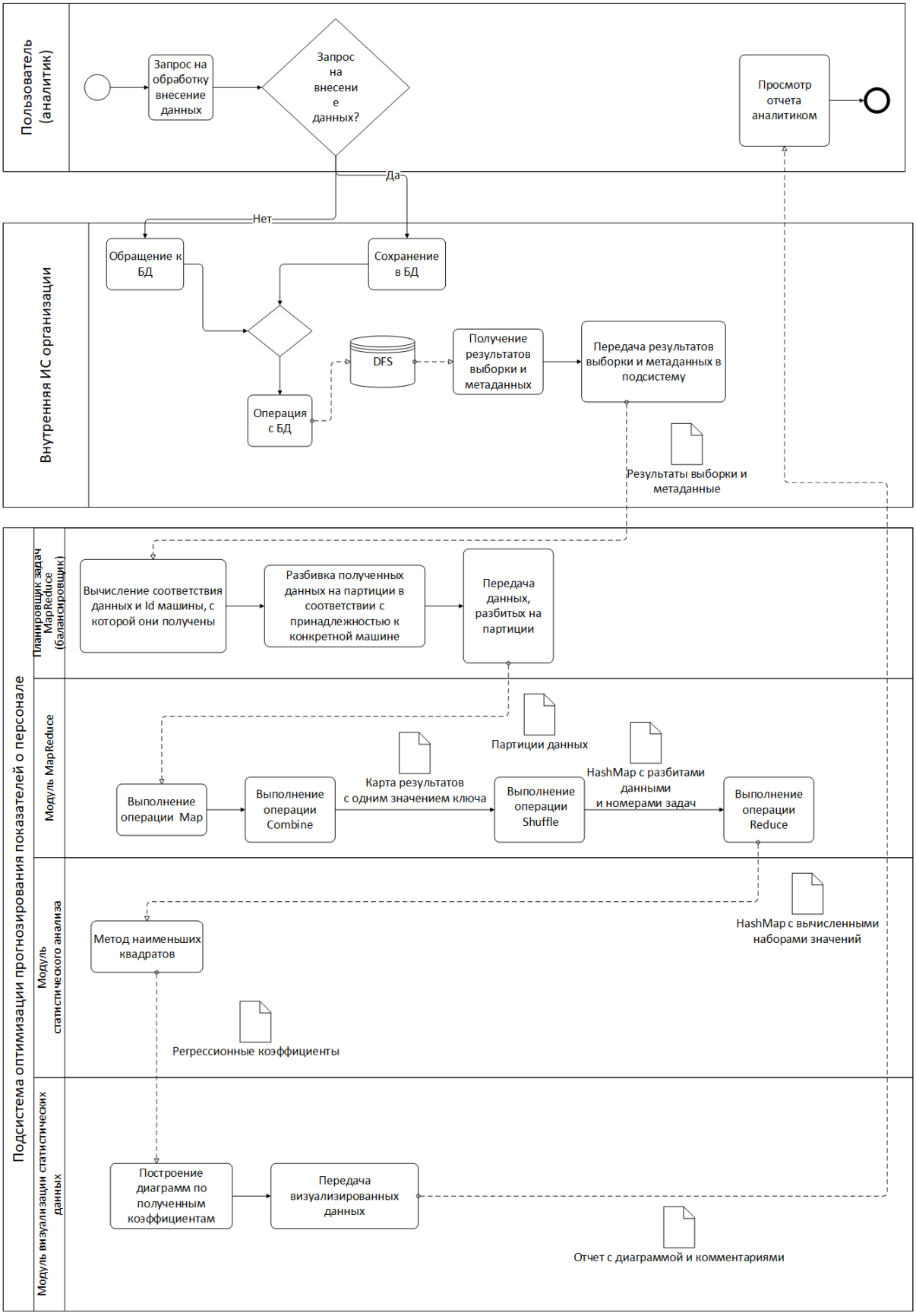

Рисунок 9 – Модель бизнес-процессов для прогнозирования показателей о персонале

Всю последовательность операций необходимых для проведения прогнозирования показателей о персонале можно увидеть на составляющих частях диаграммы последовательности, которая показана на рисунках 10-14.

На рисунке 10 показано взаимодействие пользователя системы, в нашем случае штатного аналитика, и внутренней системы организации. Пользователь делает запрос на выборку показателей о персонале через графический интерфейс пользователя.

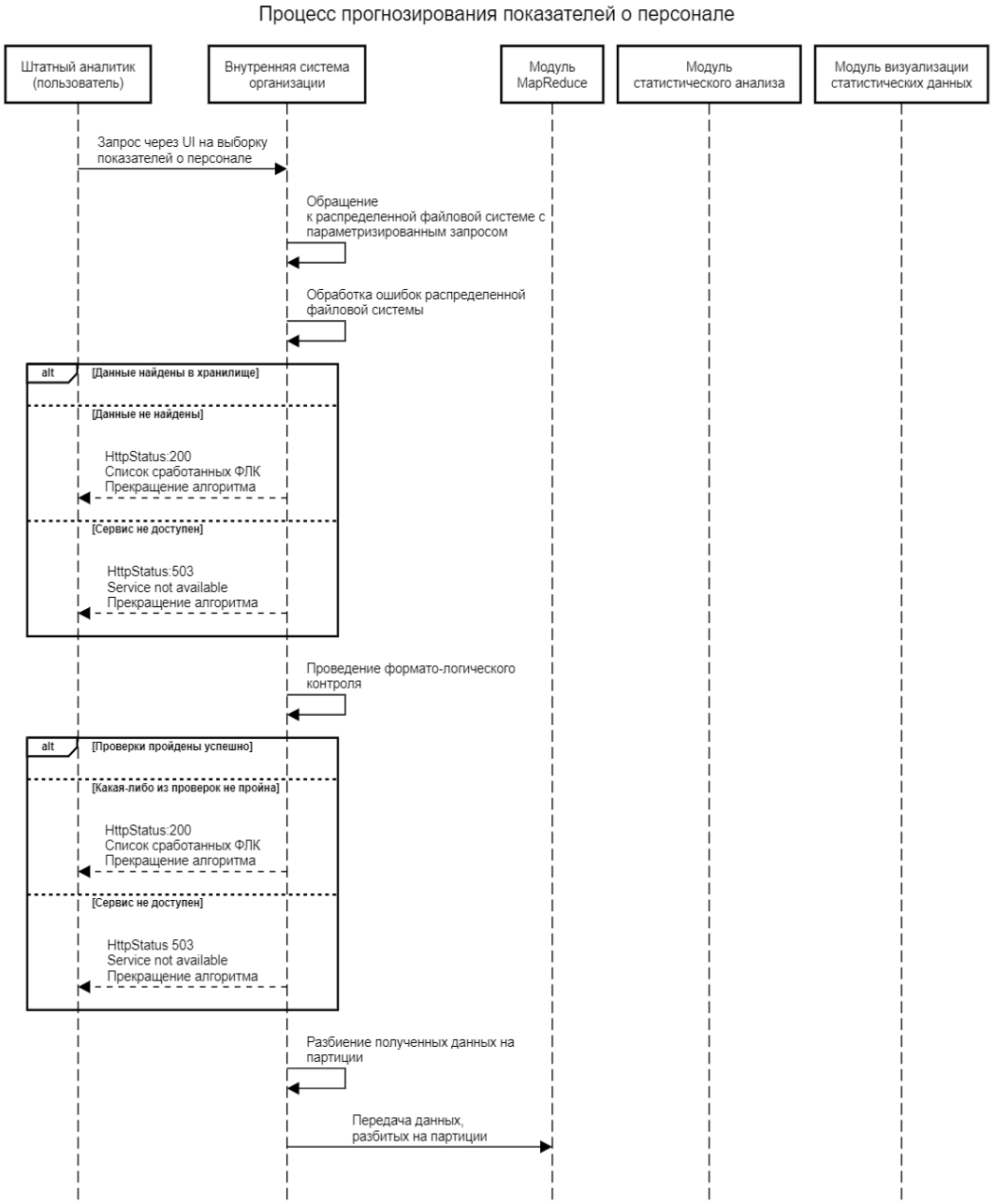

Рисунок 10 – Диаграмма последовательности часть 1

В свою очередь, внутренняя система организации обращается к распределенной файловой системе (DFS) с параметризированным запросом, который генерируется Spring Data на основе входных параметров. По получении ответа от DFS, внутренняя система организации осуществляет обработку ошибок DFS и формато-логический контроль (ФЛК) полученных данных.

При успешности всех проверок алгоритм продолжается и происходит разбивка данных на партиции балансировщиком для задач MapReduce и передача этих партиций в модуль MapReduce для дальнейшей обработки данных.

Если какие-то проверки не были пройдены или DFS не доступна, то возвращается соответствующий ответ пользователю. Так же, если не доступна внутренняя система организации, то пользователю тоже выдается соответствующее сообщение.

На рисунке 11 показано взаимодействие внутренней системы организации и модуля MapReduce. После того, как в модуль MapReduce поступили входные данные, разбитые на партиции, в модуле MapReduce происходит обработка входных данных по алгоритму MapReduce с применением шага Combine после метода map(). Далее в модуле MapRduce происходит выполнение ФЛК. При их успешном прохождении результаты обработки передаются в модуль статистического анализа. При срабатывании хотя бы одной ФЛК соответствующая ошибка передается во внутреннюю систему организации, так как все ФЛК в нашей подсистеме являются критичными, что значит, что при их появлении, процесс обработки данных прекращается. Так же при недоступности модуля MapReduce или какого-либо необходимого компонента для работы модуля MapReduce во внутреннюю систему организации посылается соответствующий ответ – HTTP статус 503, который говорит о том, что система не доступна, или какой-либо из ее модулей не доступен по причине того, что пода с репликой сервиса не активна, или проблемы с сетью.

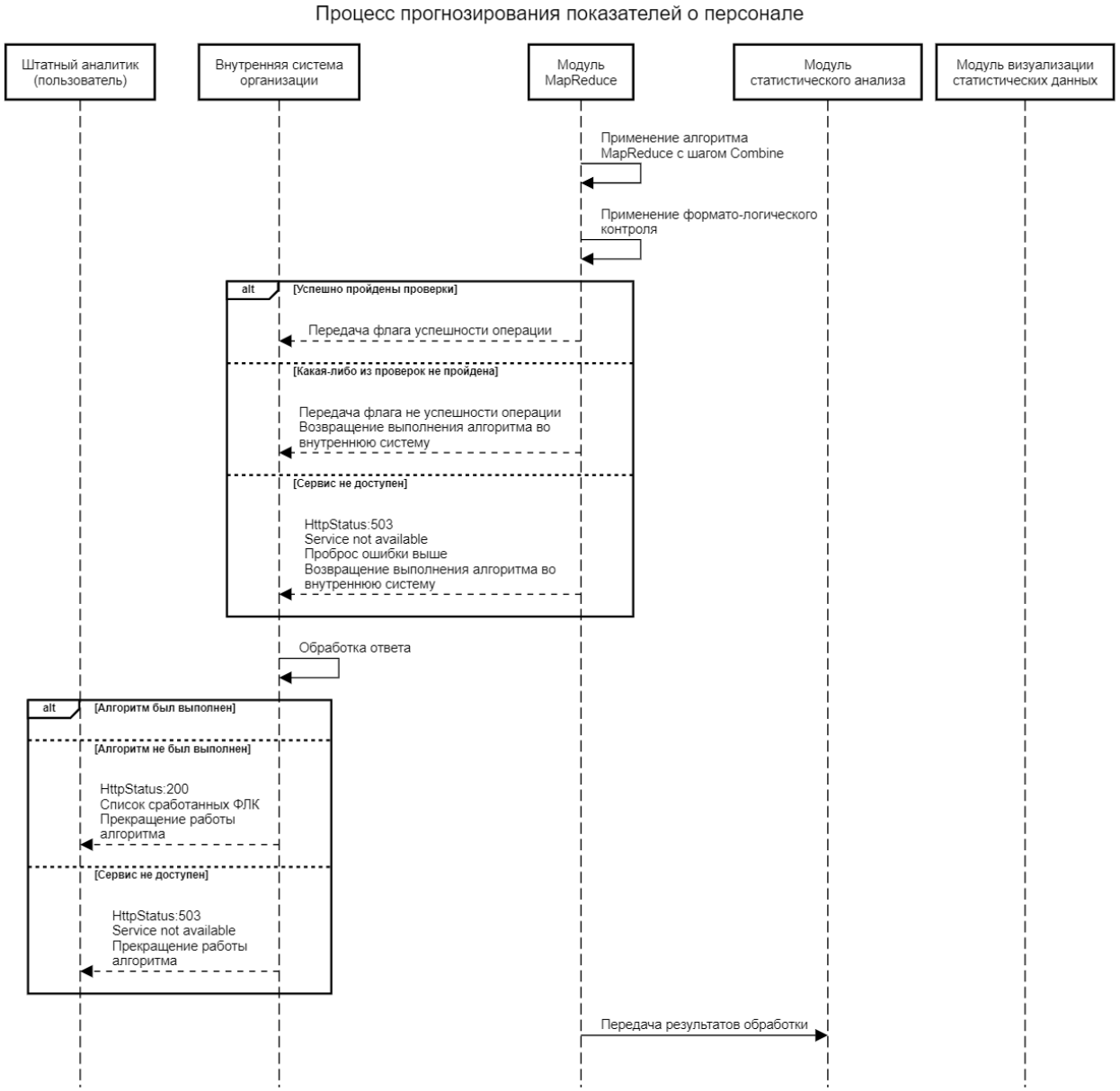

Рисунок 11 – Диаграмма последовательности часть 2

На рисунке 12 показано взаимодействие модуля MapReduce и модуля статистического анализа. После того, как в модуль статистического анализа поступили результаты работы MapReduce, модуль применяет метод наименьших квадратов. К результатам работы алгоритма производятся проверки ФЛК, при их успешности, полученные коэффициенты передаются в модуль визуализации статистических данных. Иначе, или при недоступности модуля статистического анализа, передается соответствующий ответ

вызывающей внутренней системе организации, а она в свою очередь выдает ответ пользователю.

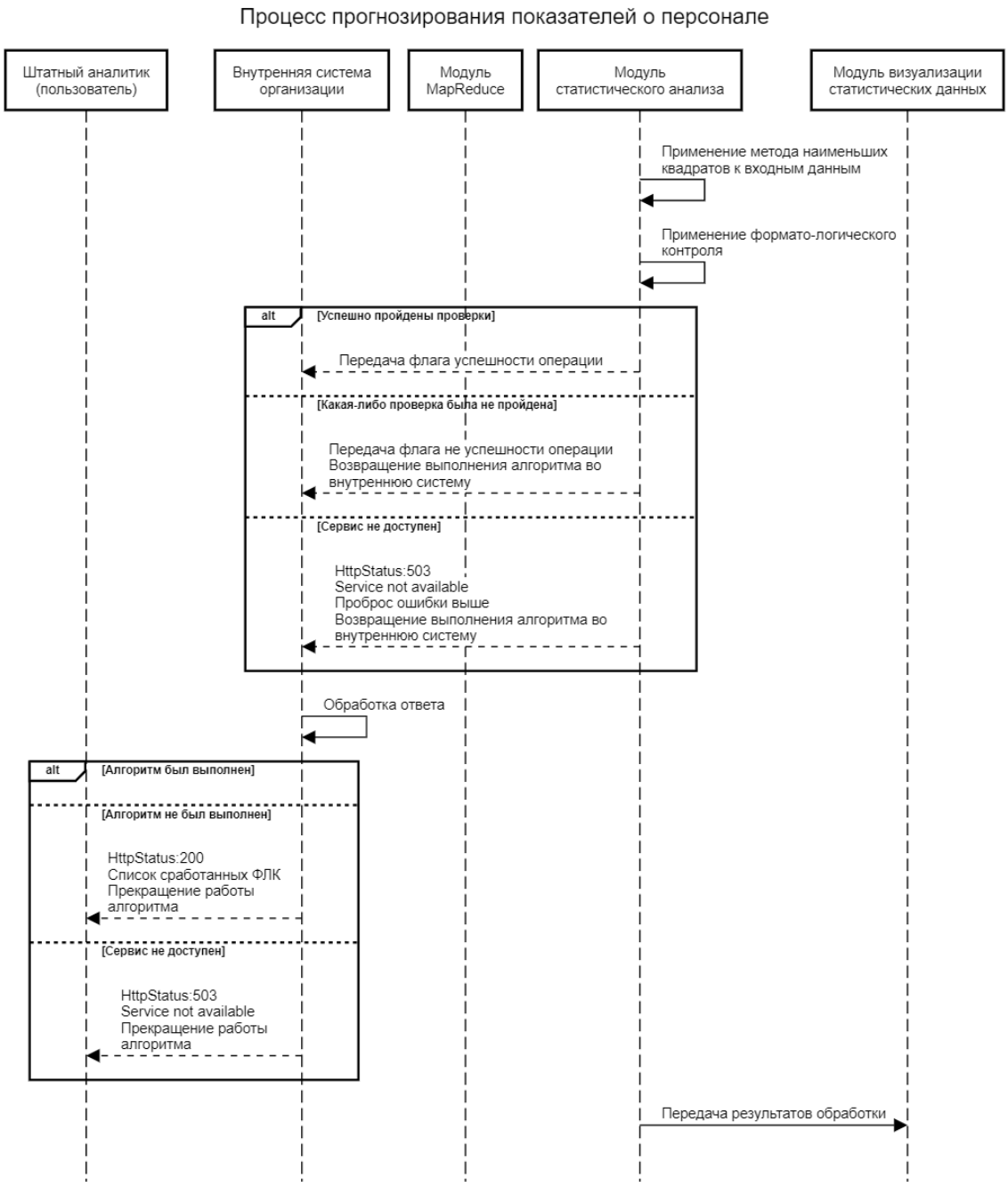

Рисунок 12 – Диаграмма последовательности часть 3

На рисунке 13 показано взаимодействие внутренней системы организации и модуля визуализации статистических данных. После получения коэффициентов от модуля статистического анализа, модуль визуализации

статистических данных визуализирует полученные коэффициенты с помощью фреймворка Swing и помещает полученные коэффициенты в текстовый документ. После чего производится проверка ФЛК и передается соответствующий ответ внутренней системе организации, которая в свою очередь его транслирует пользователю.

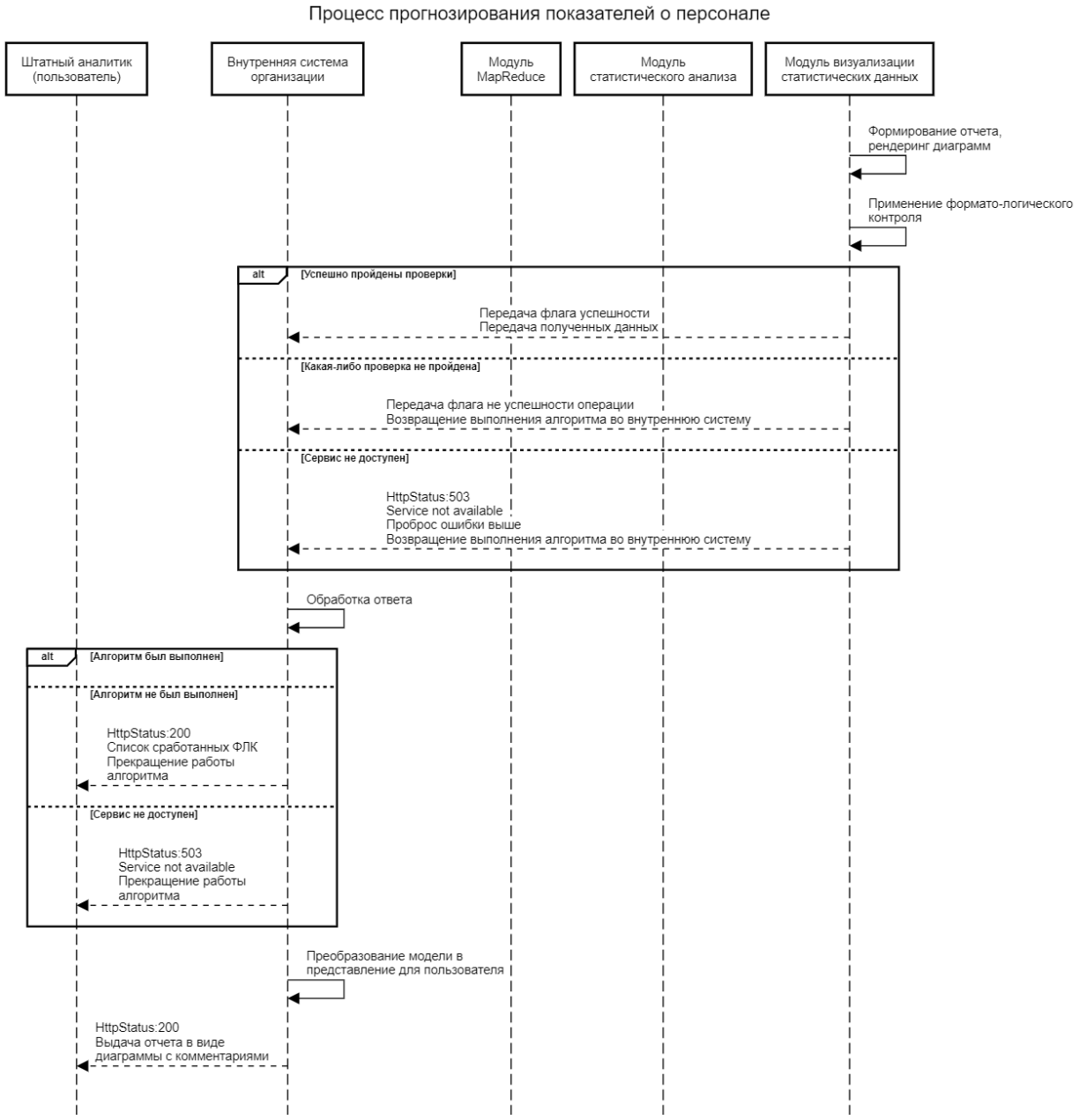

Рисунок 13 – Диаграмма последовательности часть 4

На рисунке 14 показано взаимодействие пользователя и внутренней системы организации, при сохранении новых данных о сотрудниках или обновлении уже имеющихся, а также при регистрации новых сотрудников во внутренней системе организации. Пользователь вводит необходимые данные через графический интерфейс пользователя, система сохраняет данные в DFS. Пользователю возвращается соответствующий ответ.

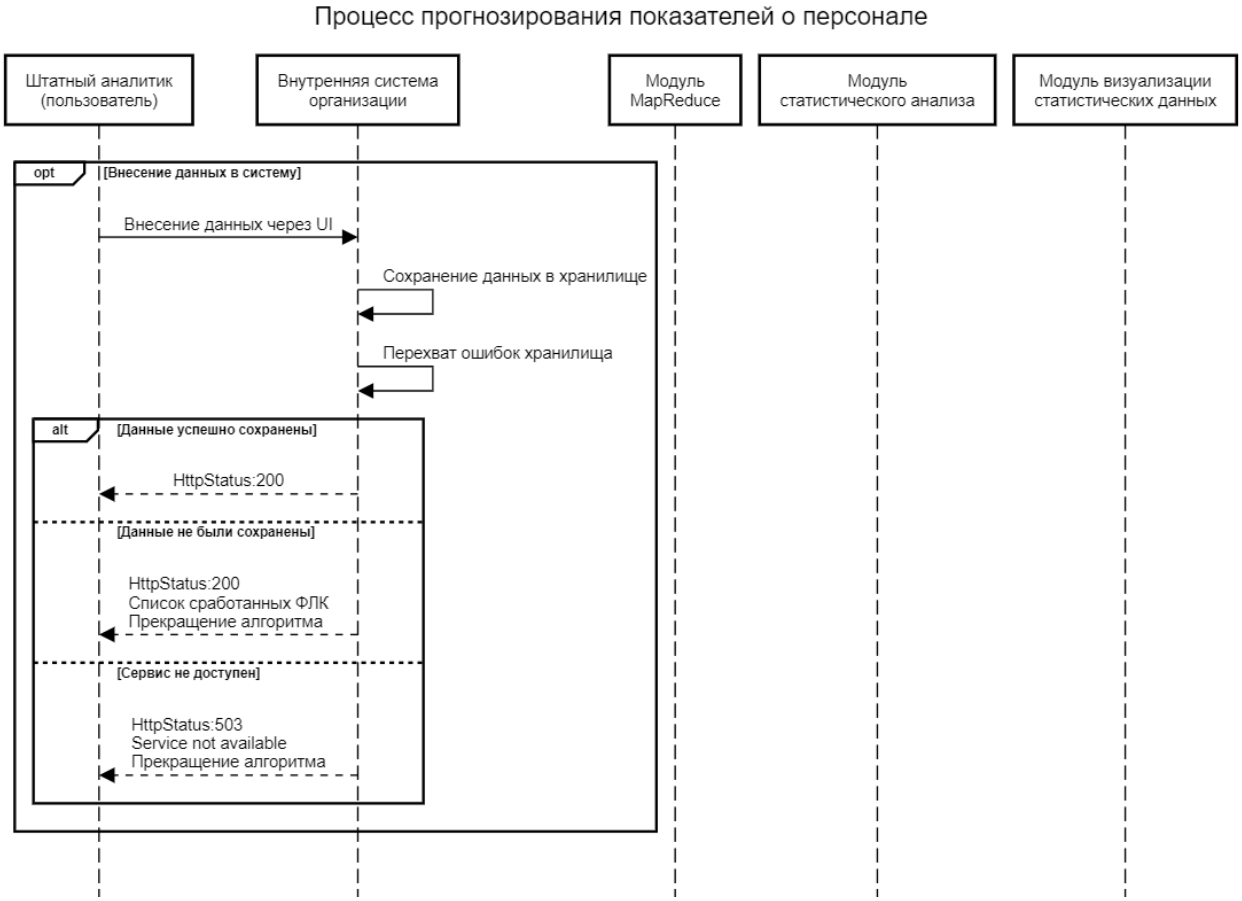

Рисунок 14 – Диаграмма последовательности часть 5

В итоге мы получим визуализированные данные для аналитиков компании, которые в свою очередь будут принимать те или иные решения по управлению персоналом для оптимизации работ. Эти визуализированные данные будут иметь вид временного графика, который отображает динамику изменения показателей о персонале с течением времени, делая прогноз на ближайший календарный год.

И, конечно, для того, чтобы это все заработало, нужно где-то разместить предложенное решение. Это будет происходить или на машинах организации или на арендованных машинах. Для развертывания систем будет необходимо несколько серверов: Web Server, где разместится несколько реплик подсистемы оптимизации прогнозирования показателей о персонале и одна реплика информационной системы организации.

Также будет необходим сервер базы данных, на котором будет крутиться несколько реплик распределенной базы данных с разными данными на каждой реплике.

И будут необходимы рабочие места для сотрудников, которые будут подключаться к серверу с приложениями, как показано на рисунке 15.

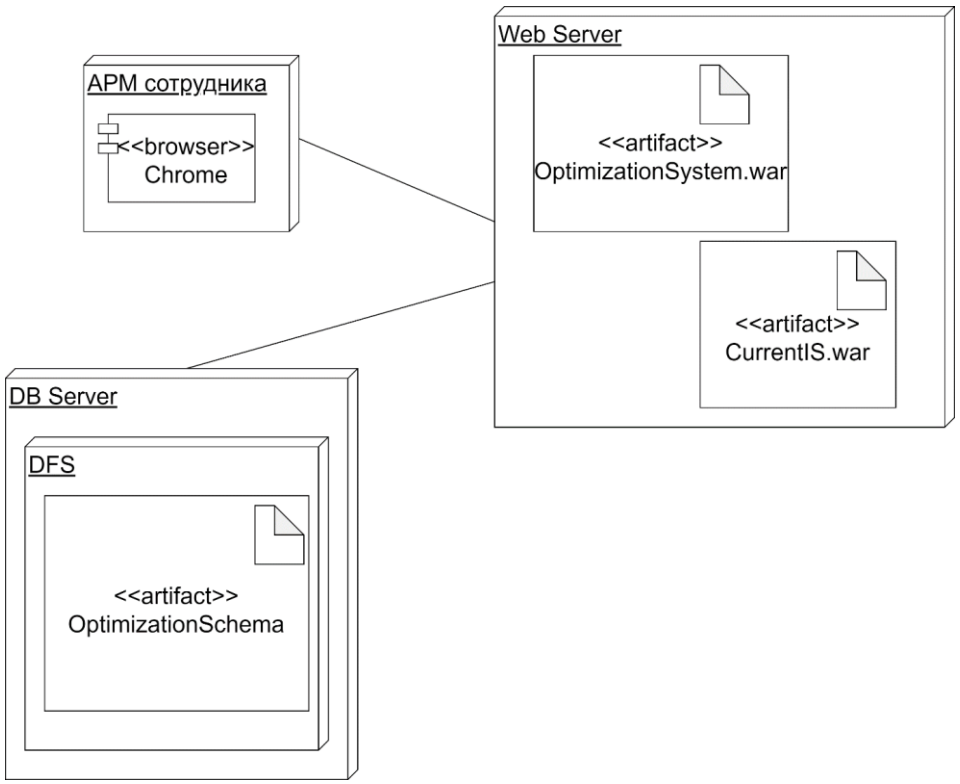

Рисунок 15 – Диаграмма развертывания

После деплоя на сервер всех необходимых инфраструктурных элементов и систем, можно будет пользоваться предложенной моделью.

Вывод: в данной части были детально изучены алгоритм MapReduce и его модификации, а также методы статистического анализа. После рассмотрения всех плюсов и минусов данных модификаций и всевозможных методов, был сделан окончательный выбор применяемой модификации алгоритма MapReduce. Данная модификация алгоритма MapReduce имеет дополнительный шаг Combine. Так же окончательно был выбран и конкретный метод статистического анализа, такой как регрессионный анализ и его конкретная реализация – метод наименьших квадратов.

Также была разработана модель для решения задачи оптимизации прогнозирования показателей о персонале: были разработаны следующие диаграммы:

- диаграмма компонентов,
- диаграмма классов,
- BPMN диаграмма,
- диаграмма последовательности,
- диаграмма развертывания.

На основе всех этих разработанных диаграмм будет программно реализована подсистема прогнозирования показателей о персонале на языке программирования Java с применением фреймворка Spring Framework, а в частности Spring Boot. Так как этот фреймфорк нацелен на разработку приложений на основе микросервисной архитектуры, что в нашем случае будет очень полезно, так как нам необходимо будет интегрировать данное решение в существующую информационную систему организации. Также несомненным преимуществом Spring Framework является возможность разрабатывать слабосвязанные приложения, что дает возможность проще писать код и не думать о внедрении всех необходимых зависимостей по ходу реализации приложения.

В следующем разделе будет проведена апробация данной модели и анализ эффективности ее работы.

4 Апробация предложенной модели для решения задачи оптимизации показателей о персонале

4.1 Применение технологии MapReduce и метода статистического анализа для работы с People Data

После детального рассмотрения технологий и методов, которые планируется использовать для достижения поставленных целей, а также демонстрации предлагаемой модели, давайте перейдем к применению модели для работы с People Data.

Как правило, в современных фирмах и корпорациях, особенно в корпорациях, работает множество сотрудников [31]. И чтобы поддерживать высокий уровень производительности, нужно уметь прогнозировать показатели сотрудников наперед и за ранее обходить узкие места или предотвращать, или, хотя бы, иметь в виду при планировании расчетного периода какие-либо издержки или же гипотетически возможные накладки в работе, такие как больничные сотрудников и тому подобное [6], [7].

В данном исследовании предлагается модель, в которой совместно применяется технология MapReduce и метод статистического анализа для анализа рисков и оптимизации прогнозирования показателей о персонале.

Начнем с технологии MapReduce. Для решения данной проблемы необходимо постоянно вести учет всех необходимых показателей о персонале, постоянно обновлять и добавлять данные в хранилище [8], [13], [17]. Это позволит делать более точные прогнозы [26]. Примерная структура показателей может иметь следующий вид [1], [29]:

- количество больничных за каждый месяц,
- количество отпусков в каждом месяце,
- количество часов переработок в каждом месяце,
- процент выполнения производственного плана в каждом месяце,

– общее количество затрат предприятия за месяц с учетом выплаты заработных плат сотрудникам,

– общий доход предприятия за месяц,

– общее количество сотрудников на предприятии.

В результате ведения такого учета на выходе можно получить фрагмент дампа данных, как показано в таблице 2, для дальнейшей обработки и анализа методом статистического анализа.

Таблица 2 - Фрагмент дампа базы данных предприятия

employee_id	month	amount_sick	amount_vac	amount_overtime	percentage_of_production_plan	spending_money	enterprise_income
dbb4c96e-8da6-4158-82c3-93038afb0207c	09	1	0	48	100	100 000	500 000
16318097-9f35-4063-a568-26519c17b784c	09	0	1	8	98	80 000	250 000

Приведенный фрагмент содержит данные по двум сотрудникам условного предприятия за сентябрь.

Применив технологию MapReduce, на основе этих показателей нужно будет провести статистический анализ для поиска сотрудников, которые недостаточно эффективно работают, после чего применить метод визуализации статистических данных, чтобы получить список таких сотрудников со всеми их показателями за текущий период [34], и выявить причины неудовлетворительных результатов производительности работы сотрудников. После чего можно будет искать пути решения данных проблем.

После получения всех необходимых данных в ход идет применение метода статистического анализа. Рассмотрим его применение.

Анализ первопричин несоответствий ожидаемых результатов и реальных результатов, а также выполнение корректировок решений – это главные задачи, которые необходимо выполнить для повышения качества. Всем этим занимается статистический анализ [20], [25].

Очень важно делать статистические выборки для анализа показателей о персонале на регулярной основе. Для более эффективного рабочего процесса и его оптимизации нужно использовать методы статистического анализа, которые имеют своей целью избегать причины некачественно сделанной работы, а не устранять последствия такой работы, такие как метод наименьших квадратов [21]. Для проведения статистического анализа существует множество методов и походов [22]. В данной работе предлагается использовать метод наименьших квадратов, так как все используемые показатели о персонале будут независимы друг от друга, а конечный результат будет зависеть от этих независимых показателей. Также метод наименьших квадратов можно автоматизировать, плюс его результаты можно вывести в виде графика или визуализировать любым другим способом.

В свою очередь использование графика для анализа полученных данных является одним из «семи инструментов контроля качества», предложенных союзом японских ученых и инженеров [27].

Применение графиков позволяет оценивать как текущее состояние процесса, так и прогнозировать возможные последствия результатов выполнения данного процесса. Все это делается, исключительно основываясь на предоставляемых графиках [32]. Но не стоит забывать, что такие прогнозы не могут быть на сто процентов верны, они лишь дают приближенную оценку тому, что может произойти в будущем, при сохранении текущих показателей [14]. Зачастую используют графики, которые отражают динамические изменения данных во времени, их называют временными рядами [33]. Наиболее популярными являются линейные, ленточные, столбчатые или круговые диаграммы [28].

Для успешного практического применения такого подхода к статистическому анализу, который поможет выявить причины неудовлетворительных результатов производительности работы сотрудников, необходимо, чтобы все сотрудники, которые принимают участие в планировании деятельности предприятия, знали и умели пользоваться этим подходом [23], [30], [36], [37].

Из таблицы 3, видно, что выбранный подход направлен на проработку архитектуры и реализацию продукции, а также на контроль качества продукции и тестирование [18]. Нас это вполне удовлетворяет.

Таблица 3 - Применение статистических методов на этапах жизненного цикла продукции

Стадии жизненного цикла продукта	Задачи, которые решаются системой управления качества	Применяемые методы статистики
Маркетинг и анализ рынка	Изучение рынка и расчет спроса, расчет вероятности изменения рынка или спроса	Экономические и математические методы, методы для анализа статистических совокупностей
	Исследование желаемой цены и качества продукции со стороны покупателей	Экономические и математические методы, др.
	Оценка и планирование количества выпускаемой продукции, ее стоимости, доли целевой аудитории на рынке	Экономические и математические методы
Архитектурная проработка и реализация продукции	Выявление требований к продукции, повышение качества продукции, оценка продукции	Графические методы (диаграммы)
	Проведение тестирования продукции, ее апробация	Графические методы, статистические методы, экономические и математические методы
	Проверка безопасности продукции	Экономические и математические методы

Продолжение таблицы 3

Стадии жизненного цикла продукта	Задачи, которые решаются системой управления качества	Применяемые методы статистики
Закупки	Составление планов закупок	Экономические и математические методы
	Анализ возможностей поставщиков	Экономические и математические методы
	Планирование доставок	Экономические и математические методы
	Оптимизация трат для обеспечения качества продукции	Экономические и математические методы
Выпуск продукции	Описание технологических процессов	Экономические и математические методы, методы для анализирования статистических совокупностей
	Контроль точности и стабильности технологических процессов	Методы для расчета статистической оценки точности и непрерывности технологических процессов
	Гарантия постоянности качества продукции на определенном уровне	Методы для статистического контроля технологических процессов
	Деятельность по выпуску продукции, ее доставки клиенту и сопровождение	Экономические и математические методы
Проведение тестирования и контроль качества	Четкое следование методологии при проработке, реализации испытаний и анализе результатов тестирования	Графические методы (диаграмма), методы для анализа статистических совокупностей
	Проверка соответствия качества продукции заявленному	Методы статистического контроля приемки
	Выведение выводов о качестве продукции	Графические методы (диаграмма), экономические и математические методы

Вывод: рассмотрев применение выбранных в нашей модели методов и технологий в теории, давайте апробируем нашу модель.

4.2 Апробация реализованной модели работы с People Data

После рассмотрения теоретического применение выбранных для нашей модели методов и технологий, давайте применим ее на практике.

В рамках настоящего исследования была разработана и реализована программа (подсистема) расчета показателей о персонале. Данная система берет данные для анализа из распределенной файловой системы, которые туда уже были загружены сотрудниками организации посредством взаимодействия с имеющейся ИС организации. После чего применяет к этим данным алгоритм MapReduce с шагом combine в модуле MapReduce и метод наименьших квадратов в статистическом модуле программы. Далее на основе коэффициентов примененного метода наименьших квадратов происходит рендеринг прогнозного графика, который показывает предполагаемую динамику показателей о персонале с течением времени.

Результат работы разработанного решения можно увидеть на рисунке 16.

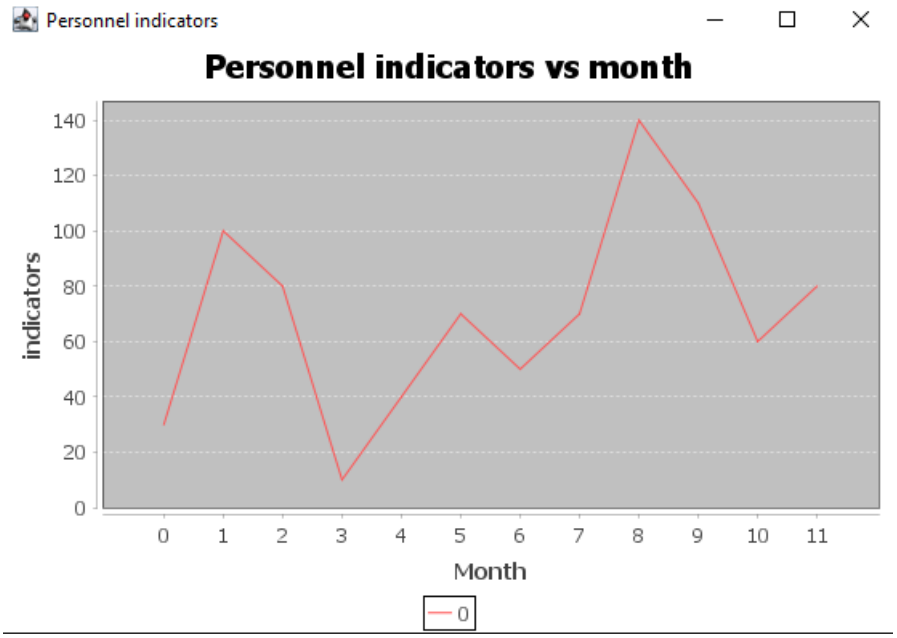

Рисунок 16 – Результат работы подсистемы прогнозирования показателей о персонале

На рисунке показана предполагаемая тенденция показателей о персонале на ближайшие 12 месяцев. Данный прогноз был сделан на основе метода наименьших квадратов. Данные о сотрудниках для анализа были получены случайным образом для тестирования решения.

Вывод: апробировав реализованную модель, давайте оценим эффективность применения этих методов и технологий.

4.3 Анализ эффективности работы предложенной модели

После применения предложенной модели, давайте перейдем к анализу эффективности применения этой модели для работы с People Data.

Оценку эффективности начнем с технологии MapReduce. Все большее значение играют стоимостные параметры и время реализации, когда речь идет о разработке проектов с применением технологии MapReduce.

Давайте определим, какие у нас могут быть издержки, когда мы используем подход MapReduce к обработке больших данных. Такие издержки можно разделить на четыре группы:

– издержки, связанные с приобретением аппаратных модулей для ЭВМ и их обслуживание. Обслуживание инфраструктуры организации,

– издержки, связанные с обслуживанием распределенной сети ЭВМ, издержки, связанные с поддержкой защищенного удаленного доступа к ЭВМ в каждом кластере,

– зарплаты сотрудникам, поддерживающим систему, затраты на покупку лицензии на программное обеспечение,

– затраты мощностей на управление и конфигурацию кластеров ЭВМ.

Так же стоит отметить, что при создании конфигурационных файлов с показателями, оптимальными для работы задач MapReduce, нужно рассчитать модель затрат для вычислений и стоимость таких затрат. Такая модель должна учитывать такие затраты, как:

– расходы ресурсов на запуск задач MapReduce,

– вычисление задач MapReduce,

– использование потоков ввода и вывода информации на различных стадиях работы алгоритма MapReduce.

Эта модель должна брать в расчет:

– объем входных данных, при решении задач,

– выставленные свойства в конфигурационных файлах для работы программы,

– свойства и конфигурации кластера, на котором происходит выполнение алгоритма.

Используя все вышеперечисленное, можно дать оценку масштабируемости и выбрать критерий, по которому будет производиться сравнение эффективности работы различных алгоритмов MapReduce.

Существует четыре разновидности настроек для конфигурации алгоритма MapReduce. От этих настроек зависит время, за которое будут выполняться последовательности задач MapReduce. Помимо того, от этих настроек будет зависеть объем ресурсов, которые были использованы кластером. Перечислим эти разновидности настроек:

– поток данных – это весь массив данных, который проходит через различные этапы цикла задач MapReduce,

– поля стоимости – это то время, за которое выполнятся задача MapReduce или ее определенная фаза,

– статистика потока данных – это сбор статистических данных о потоке,

– статистика полей стоимости – консолидация информации о времени выполнения задач MapReduce.

На выполнение того или иного действия алгоритма MapReduce затрачивается некоторое количество ресурсов. Это количество ресурсов зависит от типа операций, которые будут выполняться. К ресурсам можно причислить [24]:

– объем оперативной памяти,

– процессорное время,

– объем ПЗУ и так далее.

Посчитать общую стоимость вычислений на одном кластере можно, если сложить стоимость трат на подготовку ЭВМ, запуск задач MapReduce и цену ресурсов, которые понадобятся для решения задачи. Вычисление общей стоимости затрат показано формулой (5) [12].

$$C_O = C_{ST} + C_p, \tag{5}$$

где C_{ST} – затраты на запуск задач,

C_p – затраты на обработку данных.

Для расчета того, сколько будет стоить запустить одну задачу MapReduce, надо просуммировать стоимость конфигурации задачи, которая включает в себя длительность работы Splitter, а также траты на запуск всех этапов алгоритма MapReduce для каждой поды в скоупе одного задания, как показано в формуле (6).

$$C_{ST} = C_S + C_T, \tag{6}$$

где C_S – затраты для конфигурации кластера,

C_T – затраты на запуск задач.

Стоимость запуска всех функций Map и Reduce на кластере рассчитывается по формуле (7).

$$C_T = c_t \times \frac{N_m + N_r}{N_{cpu}}, \tag{7}$$

где c_t – стоимость запуска Map или Reduce задачи,

N_m – количество экземпляров Map функции,

N_r – количество экземпляров Reduce функций,

N_{cpu} – количество ядер процессора на кластере.

Чтобы рассчитать затраты на обработку данных нужно сложить цену за ввод и вывод данных, стоимость передачи информации между узлами, цену выполнения всех задач MapReduce. Рассчитать общие затраты на обработку того объема данных, который поступил на вход нашего алгоритма MapReduce можно по формуле (8).

$$C_p = C_{IO} + C_{MR}, \quad (8)$$

где C_{IO} – стоимость операций ввода и вывода,

C_{MR} – стоимость выполнения задач Map и Reduce.

Данных критериев будет достаточно для определения эффективности работы алгоритма MapReduce, так что будем использовать их. Хотя стоит отметить, что есть еще ряд критериев, по которым можно рассчитывать эффективность работы алгоритма, но мы их использовать не будем, чтобы не усложнять наше исследование большим количеством вычислений [4]. Так как вычисление эффективности работы алгоритма MapReduce не является темой данной работы.

Подводя итог, можно сказать, что расчет всего объема затрат на выполнение цепочки задач MapReduce осуществляется путем суммирования всех операций, которые применяются во всех этапах работы алгоритма MapReduce [11].

Полученные уравнения расчета стоимости затрат имеют линейную зависимость. Так что количество затрат (стоимость применения данного алгоритма) будет прямо пропорционально зависеть от количества операций, которые необходимо выполнить в цепочке MapReduce.

В рамках настоящего исследования была определена зависимость времени выполнения задач MapReduce от объема поступающей информации, как показано на рисунке 17.

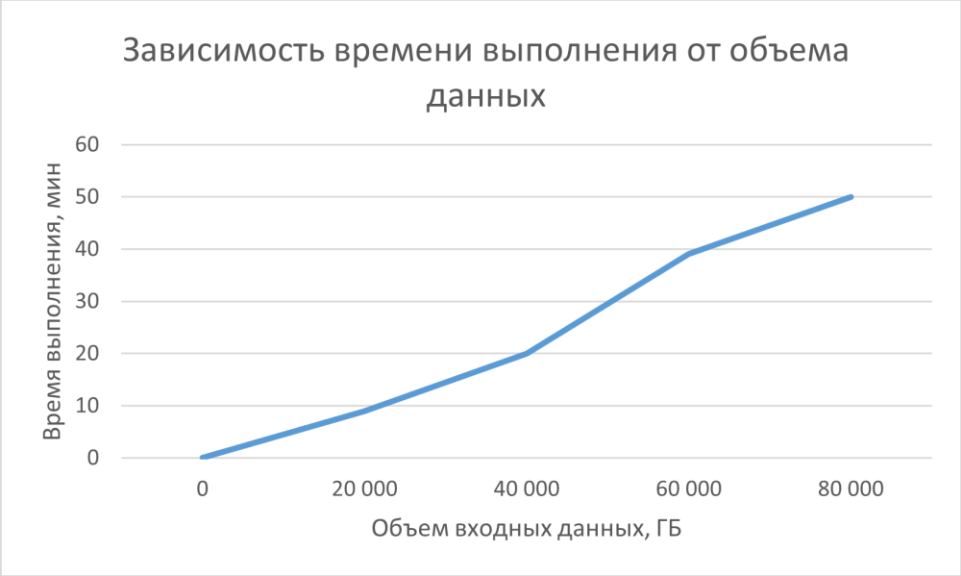

Рисунок 17 – Зависимость времени выполнения задачи MapReduce от объема поступающих данных

В свою очередь статистический анализ, а именно совместное применение регрессионного анализа (метод наименьших квадратов) и графиков, довольно эффективен, так как метод наименьших квадратов производится ЭВМ по нескольким простым формулам, а на выходе персонал оценивает результаты обработки показателей о персонале и делает выводы о том необходимо ли что-то менять в следующем расчетном периоде, чтобы увеличить производительность, исходя из построенных диаграмм программой, в которой применяется метод наименьших квадратов [3], [9]. Но стоит отметить, что во многом эффективность зависит от конкретной реализации метода наименьших квадратов, объема поступающих данных и компетенции сотрудников, которые занимаются анализом полученных результатов и делают на основе этих результатов выводы о продуктивности работы предприятия, а также принимают решения нужно ли что-то менять [2], [5].

Вывод: полученная оценка эффективности применения метода статистического анализа и технологии MapReduce может быть использована также для:

– оценки того, на сколько приемлема эффективность реализации алгоритмов и профилей, применяемых в программном комплексе Hadoop,

– проработки модели оптимальной реализации алгоритма MapReduce,

– вычисления уровня загрузки процессора в каждом узле кластера распределенной системы в определенное время,

– определения эффективности применения той или иной модификации решения задачи MapReduce,

– определения эффективности метода наименьших квадратов статистического анализа,

– определения эффективности любого подхода к статистическому анализу,

– определения эффективности использования того или иного метода визуализации полученных данных при работе алгоритма MapReduce.

Апробировав и эффективность разработанной модели, можно сказать, что модель получилась достаточно надежной и эффективной для прогнозирования оптимальных показателей о персонале.

Заключение

В ходе настоящего исследования были выявлены методы и технологии, которые легли в основу разработанного решения, а именно.

Была использована технология MapReduce с дополнительным шагом combine, которая занимается следующим:

– обрабатывает входные данные,

– разбивает их на части (партиции),

– распределяет нагрузку между подами приложения, в зависимости от загруженности той или иной поды,

– производит обработку по алгоритму.

Также был применен метода наименьших квадратов для расчета прогноза показателей о персонале, который основывается на множестве входных данных, которые имеют сложную структуру и требуют анализа зависимости одной переменной от нескольких независимых факторов.

Помимо того, в ходе реализации программного решения был применен метод визуализации данных, который был реализован с помощью фреймфорка Swing.

С применением всех этих методов и технологий была спроектирована и реализована программная составляющая работы, которая представляет из себя микросервисное приложение, написанное на Java с применением фреймфорка Spring, а именно его наиболее популярной части – Spring Boot.

Данное программное решение позволяет на основе входных данных, которые поступают из распределенной файловой системы организации, спрогнозировать показатели о персонале. Данный анализ производится с применением технологии MapReduce с шагом combine для выдачи небольших порций данных модулю статистического анализа.

В свою очередь модуль статистического анализа использует метод наименьших квадратов для прогнозирования показателей.

Далее полученные данные визуализируются в виде временного графика. С использованием этого графика аналитики организации смогут принимать какие-либо решения.

Также была проведена апробация данного решения и выявлена эффективность его работы.

Результаты настоящего исследования могут быть применены при дальнейшем изучении понятия People Data, а также на любых предприятиях, где необходимо обрабатывать данные о сотрудниках, где таких данных очень много, так много, что их можно причислить к большим данным.

Данные результаты исследования позволили посмотреть на понятие People Data, как часть понятия Big Data, с новой точки зрения, а также рассмотреть возможность применения подхода People Data для оптимизации прогнозирования показателей о персонале в современных реалиях.

Гипотеза, что применение предложенной в рамках настоящего диссертационного исследования модели для работы с данными о людях, предоставит возможность самым эффективным образом искать новые пути для решения управленческих задач в организации, а также формировать шаблон наиболее эффективного управления персоналом в будущем, была подтверждена.

Список используемой литературы

1. Авраменко Ю. В., Шумилов А. С. Спецификация распараллеливания обработки векторных данных в модели MapReduce // Информационные и математические технологии в науке и управлении. 2017. №4 (8). URL: https://cyberleninka.ru/article/n/spetsifikatsiya-rasparallelivaniya-obrabotki-vektornyh-dannyh-v-modeli-mapreduce (дата обращения: 15.12.2021).

2. Арькова Т. Ю. Управление человеческими ресурсами организации на основе системы HR-брендинга // Вестник АГТУ. Серия: Экономика. 2011. №1. URL: https://cyberleninka.ru/article/n/upravlenie-chelovecheskimi-resursami-organizatsii-na-osnove-sistemy-hr-brendinga (дата обращения: 25.06.2021).

3. Афлетунова Г.Э. Система управления персоналом // Инфраструктурные отрасли экономики: проблемы и перспективы развития. 2015. №8. URL: https://cyberleninka.ru/article/n/sistema-upravleniya-personalom (дата обращения: 25.06.2021).

4. Баданина О.В., Гиндин С.И., Хомоненко А.Д. Оценка оперативности передачи больших данных на примере базы данных PostgreSQL, платформы Hadoop и системы Sqoop // Интеллектуальные технологии на транспорте. 2020. №2 (22). URL: https://cyberleninka.ru/article/n/otsenka-operativnosti-peredachi-bolshih-dannyh-na-primere-bazy-dannyh-postgresql-platformy-hadoop-i-sistemy-sqoop (дата обращения: 15.12.2021).

5. Барановский В. Ю., Зайченко И. М. Формирование стратегической карты управления предприятием на основе концепции цифровой трансформации бизнеса // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Экономические науки. 2018. №3. URL: https://cyberleninka.ru/article/n/formirovanie-strategicheskoy-karty-upravleniya-predpriyatiem-na-osnove-kontseptsii-tsifrovoy-transformatsii-biznesa (дата обращения: 17.01.2021).

6. Баханов А. Г. Роль больших данных и имитационного моделирования в социально -экономических исследованиях // Социологический альманах. 2017. №8. URL: https://cyberleninka.ru/article/n/rol-bolshih-dannyh-i-imitatsionnogo-modelirovaniya-v-sotsialno-ekonomicheskih-issledovaniyah (дата обращения: 25.06.2021).

7. Борисенко О.Д., Пастухов Р.К., Кузнецов С.Д. Создание виртуальных кластеров Apache Spark в облачных средах с использованием систем оркестрации // Труды ИСП РАН. 2016. №6. URL: https://cyberleninka.ru/article/n/sozdanie-virtualnyh-klasterov-apache-spark-v-oblachnyh-sredah-s-ispolzovaniem-sistem-orkestratsii (дата обращения: 15.12.2021).

8. Бочкова Е. В., Авдеева Е. А., Щербаков Д. С. Особенности применения информационной технологии Big Data в маркетинговой деятельности российских компаний B2C-сектора // Концепт. 2016. №S17. URL: https://cyberleninka.ru/article/n/osobennosti-primeneniya-informatsionnoy-tehnologii-big-data-v-marketingovoy-deyatelnosti-rossiyskih-kompaniy-b2c-sektora (дата обращения: 17.01.2021).

9. Верещагина Л. С. Реализация цифровых технологий в HR системе корпорации // Вестник Саратовского государственного социально-экономического университета. 2019. №5 (79). URL: https://cyberleninka.ru/article/n/realizatsiya-tsifrovyh-tehnologiy-v-hr-sisteme-korporatsii (дата обращения: 17.01.2021).

10. Виниченко М. В., Шиховцова А.И. Применение datascience в HR // Материалы Афанасьевских чтений. 2016. №3 (16). URL: https://cyberleninka.ru/article/n/primenenie-datascience-v-hr (дата обращения: 17.01.2021).

11. Гладкий М. В. Модель распределенных вычислений MapReduce // Труды БГТУ. Серия 3: Физико-математические науки и информатика. 2016. №6 (188). URL: https://cyberleninka.ru/article/n/model-raspredelennyh-vychisleniy-mapreduce (дата обращения: 15.12.2021).

12. Гладкий М. В. Модель стоимости затрат MapReduce-вычислений // Труды БГТУ. Серия 3: Физико-математические науки и информатика. 2017. №3 (194). URL: https://cyberleninka.ru/article/n/model-stoimosti-zatrat-mapreduce-vychisleniy (дата обращения: 15.12.2021).

13. Григорьев Е. А., Климов Н. С. Разведочный анализ данных с помощью Python // E-Scio. 2020. №2 (41). URL: https://cyberleninka.ru/article/n/razvedochnyy-analiz-dannyh-s-pomoschyu-python (дата обращения: 17.01.2021).

14. Губа К. Большие данные в социологии: новые данные, новая социология? // Социологическое обозрение. 2018. №1. URL: https://cyberleninka.ru/article/n/bolshie-dannye-v-sotsiologii-novye-dannye-novaya-sotsiologiya (дата обращения: 25.06.2021).

15. Гусейнов А.А., Бочкова И.А. Исследование распределенной обработки данных на примере системы Hadoop // Актуальные проблемы авиации и космонавтики. 2016. №12. URL: https://cyberleninka.ru/article/n/issledovanie-raspredelennoy-obrabotki-dannyh-na-primere-sistemy-hadoop (дата обращения: 15.12.2021).

16. Долженко Р. А. People Data ("данные о людях") как новое направление работы с человеческими ресурсами // Вестник ОмГУ. Серия: Экономика. 2019. №1. URL: https://cyberleninka.ru/article/n/people-data-dannye-o-lyudyah-kak-novoe-napravlenie-raboty-s-chelovecheskimi-resursami (дата обращения: 08.01.2021).

17. Дорофеев Р. С., Дорофеев А. С., Рогачева С. А. Применение технологии big data на основе mapreduce для повышения уровня успеваемости обучающихся // Глобус: технические науки. 2019. №5 (29). URL: https://cyberleninka.ru/article/n/primenenie-tehnologii-big-data-na-osnove-mapreduce-dlya-povysheniya-urovnya-uspevaemosti-obuchayuschihsya (дата обращения: 15.12.2021).

18. Ершова Е. А. Методы статистического анализа // European research. 2016. №12 (23). URL: https://cyberleninka.ru/article/n/metody-statisticheskogo-analiza (дата обращения: 15.12.2021).

19. Забокрицкая Л. Д., Хлебников Н. А., Орешкина Т. А., Комоцкий Е. И. Возможности изучения ценностей молодежи через профиль социальной сети «ВКонтакте» // Мониторинг. 2020. №2 (156). URL: https://cyberleninka.ru/article/n/vozmozhnosti-izucheniya-tsennostey-molodezhi-cherez-profil-sotsialnoy-seti-vkontakte (дата обращения: 17.01.2021).

20. Калиновская И. Н. Социальные данные как инструмент специалиста по управлению человеческими ресурсами организации // Вестник ВГТУ. 2020. №1 (38). URL: https://cyberleninka.ru/article/n/sotsialnye-dannye-kak-instrument-spetsialista-po-upravleniyu-chelovecheskimi-resursami-organizatsii (дата обращения: 17.01.2021).

21. Китова О. В. Применение информационных систем для решения задач управления эффективностью маркетинга // Открытое образование. 2009. №2. URL: https://cyberleninka.ru/article/n/primenenie-informatsionnyh-sistem-dlya-resheniya-zadach-upravleniya-effektivnostyu-marketinga (дата обращения: 17.01.2021).

22. Колесниченко О. Ю. Векторы развития информационного аспекта глобализации // Вестник ГУУ. 2014. №16. URL: https://cyberleninka.ru/article/n/vektory-razvitiya-informatsionnogo-aspekta-globalizatsii (дата обращения: 25.06.2021).

23. Криницына З. В. Управление персоналом в организации на основе системы HR-брендинга // Векторы благополучия: экономика и социум. 2013. №4 (10). URL: https://cyberleninka.ru/article/n/upravlenie-personalom-v-organizatsii-na-osnove-sistemy-hr-brendinga (дата обращения: 25.06.2021).

24. Лыфарь Д. А. Обработка реляционных баз данных на графических процессорах // МСиМ. 2011. №1 (22). URL: https://cyberleninka.ru/article/n/obrabotka-relyatsionnyh-baz-dannyh-na-graficheskih-protsessorah (дата обращения: 15.12.2021).

25. Мавлютова Г.Ш., Савельев Д. Л. Информационная система Ямало-Ненецкого автономного округа как инструмент социально-экономического развития коренных малочисленных народов Севера на современном этапе // ARS ADMINISTRANDI. 2019. №1. URL: https://cyberleninka.ru/article/n/informatsionnaya-sistema-yamalo-nenetskogo-avtonomnogo-okruga-kak-instrument-sotsialno-ekonomicheskogo-razvitiya-korennyh (дата обращения: 17.01.2021).

26. Магеррамов З. Т., Абдуллаев В. Г., Магеррамова А. З. Big Data: проблемы, методы анализа, алгоритмы // Радиоэлектроника и информатика. 2017. №3. URL: https://cyberleninka.ru/article/n/big-data-problemy-metody-analiza-algoritmy (дата обращения: 25.06.2021).

27. Марченко М. А. Современные суперкомпьютерные статистические методы анализа больших данных // Марчуковские научные чтения. 2019. №2019. URL: https://cyberleninka.ru/article/n/sovremennye-superkompyuternye-statisticheskie-metody-analiza-bolshih-dannyh (дата обращения: 15.12.2021).

28. Нагибина Н. И., Щукина А. А. HR-Digital: цифровые технологии в управлении человеческими ресурсами // Вестник евразийской науки. 2017. №1 (38). URL: https://cyberleninka.ru/article/n/hr-digital-tsifrovye-tehnologii-v-upravlenii-chelovecheskimi-resursami (дата обращения: 17.01.2021).

29. Назаренко Ю. Л. Обзор технологии "большие данные" (Big Data) и программно-аппаратных средств, применяемых для их анализа и обработки // European science. 2017. №9 (31). URL: https://cyberleninka.ru/article/n/obzor-tehnologii-bolshie-dannye-big-data-i-programmno-apparatnyh-sredstv-primenyaemyh-dlya-ih-analiza-i-obrabotki (дата обращения: 15.12.2021).

30. Осавелюк Е. А. Роль средств массовой информации (сми) в обеспечении информационной безопасности России на евразийском пространстве // Международное сотрудничество евразийских государств: политика, экономика, право. 2016. №3 (8). URL: https://cyberleninka.ru/article/n/rol-sredstv-massovoy-informatsii-smi-v-

obespechenii-informatsionnoy-bezopasnosti-rossii-na-evraziyskom-prostranstve (дата обращения: 17.01.2021).

31. Осипов К. А. Использование технологий Big Data в формировании системы управления рисками предпринимательских структур // Теория и практика сервиса: экономика, социальная сфера, технологии. 2019. №2 (40). URL: https://cyberleninka.ru/article/n/ispolzovanie-tehnologii-big-data-v-formirovanii-sistemy-upravleniya-riskami-predprinimatelskih-struktur (дата обращения: 25.06.2021).

32. Попазова О.А., Шихова Н.Н. Управление персоналом на основе анализа больших данных: риски и возможности // Известия СПбГЭУ. 2019. №3 (117). URL: https://cyberleninka.ru/article/n/upravlenie-personalom-na-osnove-analiza-bolshih-dannyh-riski-i-vozmozhnosti (дата обращения: 16.01.2021).

33. Свиридова О. П., Чуланова О. Л. Программа реализации HR - аналитики как цифрового тренда // Материалы Афанасьевских чтений. 2020. №3 (32). URL: https://cyberleninka.ru/article/n/programma-realizatsii-hr-analitiki-kak-tsifrovogo-trenda (дата обращения: 17.01.2021).

34. Тимофеев А. Г., Лебединская О. Г. Data Mining и big data в бизнес-аналитике цифровой трансформации государственного и корпоративного управления // УЭкС. 2017. №9 (103). URL: https://cyberleninka.ru/article/n/data-mining-i-big-data-v-biznes-analitike-tsifrovoy-transformatsii-gosudarstvennogo-i-korporativnogo-upravleniya (дата обращения: 17.01.2021).

35. Чернышова М. А., Утешева А. Ю. Статистические методы анализа и управления качеством // Прикаспийский журнал: управление и высокие технологии. 2009. №3. URL: https://cyberleninka.ru/article/n/statisticheskie-metody-analiza-i-upravleniya-kachestvom (дата обращения: 15.12.2021).

36. Чуланова О. Л. Возможности применения дескриптивной, прогнозной, предиктивной и прескриптивной hr -аналитики как цифровых трендов // Материалы Афанасьевских чтений. 2020. №1 (30). URL: https://cyberleninka.ru/article/n/vozmozhnosti-primeneniya-deskriptivnoy-

prognoznoy-prediktivnoy-i-preskriptivnoy-hr-analitiki-kak-tsifrovyh-trendov (дата обращения: 17.01.2021).

37. Чуланова О. Л., Баймашева А. Б. Классификация данных и показателей для HR-аналитики // Материалы Афанасьевских чтений. 2020. №2 (31). URL: https://cyberleninka.ru/article/n/klassifikatsiya-dannyh-i-pokazateley-dlya-hr-analitiki (дата обращения: 17.01.2021).

38. Allamu M. Rogib Rova, Asmony Thatok, Sulaimiah Attitude analysis on the effect of career development on organizational commitment of frontline employees of banking sector in Mataram City // RJOAS. 2018. №7. URL: https://cyberleninka.ru/article/n/attitude-analysis-on-the-effect-of-career-development-on-organizational-commitment-of-frontline-employees-of-banking-sector-in-mataram-city (дата обращения: 17.01.2021).

39. Grid Computing: Making the Global Infrastructure a Reality / Eds. F. Berman, G. Fox, T. Hey. – New York : John Wiley and Sons, 2003. – 1060 p.

40. Khokhlova K.P. To the problem of aggression: road rage analysis // Актуальные проблемы авиации и космонавтики. 2016. №12. URL: https://cyberleninka.ru/article/n/to-the-problem-of-aggression-road-rage-analysis (дата обращения: 17.01.2021).

41. Masyutin A.A. Credit scoring based on social network data // Бизнес-информатика. 2015. №3 (33). URL: https://cyberleninka.ru/article/n/credit-scoring-based-on-social-network-data (дата обращения: 17.01.2021).

42. Oganesian Tigran D. The right to privacy and data protection in the information age // Журнал СФУ. Гуманитарные науки. 2020. №10. URL: https://cyberleninka.ru/article/n/the-right-to-privacy-and-data-protection-in-the-information-age (дата обращения: 17.01.2021).

43. Reeta Sony A.L., Sri Krishna Deva Rao, Devi Prasad Bhukya, Sai Deep B. Solving e-Governance Challenges in India through the Incremental Adoption of Cloud Services // Право. Журнал Высшей школы экономики. 2015. №1. URL: https://cyberleninka.ru/article/n/solving-e-governance-challenges-in-india-through-the-incremental-adoption-of-cloud-services (дата обращения: 17.01.2021).

44. Siniak Nikolai, Habib Avada, Sharif Nureddin Companies challenges and opportunities with Big Data // Труды БГТУ. Серия 5: Экономика и управление. 2018. №1 (208). URL: https://cyberleninka.ru/article/n/companies-challenges-and-opportunities-with-big-data (дата обращения: 17.01.2021).

45. Tilly Ch. The Old New Social History and the New Old Social History : CRSO Working Paper No. 218. – Ann Arbor, MI : Center for Research on Social Organization, University of Michigan, 1980. – 49 p. – URL : https://deepblue.lib.umich.edu/handle/2027.42/50992.

46. Zazdravnykh Yevgeny Aleksandrovich Why there are more entrepreneurs-manufacturers in one regions and less in others: an empirical evidence // Экономика региона. 2014. №3. URL: https://cyberleninka.ru/article/n/why-there-are-more-entrepreneurs-manufacturers-in-one-regions-and-less-in-others-an-empirical-evidence (дата обращения: 17.01.2021).