TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Исследование методов машинного обучения без учителя для анализа задач в больших вычислительных сетях.

УДК 519.237.8:004.93'14

Студент

Группа	ФИО	Подпись	Дата
8ПМ7И	Шкабара Анастасия Игоревна		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		

Консультант

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ведущий программист ОИТ ИШИТР	Губин Максим Юрьевич	-		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ОСГН ШБИП	Потехина Нина Васильевна	-		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Горбенко Михаил Владимирович	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		

Томск – 2019 г.

Планируемые результаты обучения

Код результата	Результат обучения (выпускник должен быть готов)
Общие по направлению подготовки 09.04.04 «Программная инженерия»	
Р1	Проводить научные исследования, связанные с объектами профессиональной деятельности.
Р2	Разрабатывать новые и улучшать существующие методы и алгоритмы обработки данных в информационно-вычислительных системах.
Р3	Составлять отчеты о проведенной научно-исследовательской работе и публиковать научные результаты.
Р4	Проектировать системы с параллельной обработкой данных и высокопроизводительные системы.
Р5	Осуществлять программную реализацию информационно-вычислительных систем, в том числе распределенных.
Р6	Осуществлять программную реализацию систем с параллельной обработкой данных и высокопроизводительных систем.
Р7	Организовывать промышленное тестирование создаваемого программного обеспечения.
Профиль «Технологии больших данных»/ «Big data solutions»	
Р8	Исследовать и анализировать большие данные, создавать их модели и интерпретировать структуры данных в таких моделях.
Р9	Понимать принципы создания, хранения, управления, передачи и анализа больших данных с использованием новейших технологий, инструментов и систем обработки данных в высокопроизводительных сетях.
Р10	Применять теорию распределенной системы управления базами данных к традиционным распределенным системам реляционных баз данных, облачным базам данных, крупномасштабным системам машинного обучения и хранилищам данных.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП

_____ _____ _____
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	**ФИО**
8ПМ7И	Шкабара Анастасия Игоревна

Тема работы:

Исследование методов машинного обучения без учителя для анализа задач в больших вычислительных сетях.	
Утверждена приказом директора (дата, номер)	№3794/с от 15.05.2019

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Исследование и разработрка алгоритма предварительной кластеризации данных журналов обработки задач в системе WLCG, снятых с детектора ATLAS Большого Адронного Коллайдера.

Перечень подлежащих исследованию, проектированию и разработке вопросов	1. Изучение предметной области; 2. Обзор существующих решений 3. Подготовка данных 4. Реализация методов снижения размерности 5. Реализация методов кластеризации 6. Анализ результатов исследования 7. Расчет показателей ресурсоэффективности 8. Оценка показателей социальной ответственности
Перечень графического материала	Графики визуализации результатов кластерного анализа

Консультанты по разделам выпускной квалификационной работы

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Старший преподаватель ОСГН ШБИП Потехина Нина Васильевна
Социальная ответственность	Доцент ООД ШБИП Горбенко Михаил Владимирович
Раздел на иностранном языке	Доцент ОИЯ ШБИП Диденко Анастасия Владимировна

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Обзор существующих методов кластерного анализа (Cluster analysis methods)

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель / консультант (при наличии):

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		
Ведущий программист ОИТ ИШИТР	Губин Максим Юрьевич	-		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Шкабара Анастасия Игоревна		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Уровень образования Магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения весенний семестр 2018 /2019 учебного года

Форма представления работы:

Магистерская диссертация

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
31.03.2019	*Раздел 1. Аналитический обзор предметной области*	*25*
28.04.2019	*Раздел 2. Применение подходов кластерного анализа к предметной области*	*20*
04.05.2019	*Раздел 3. Результаты исследования*	*35*
17.05.2019	*Раздел 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение*	*10*
23.05.2019	*Раздел 5. Социальная ответственность*	*10*

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		

Консультант (при наличии)

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ведущий программист ОИТ ИШИТР	Губин Максим Юрьевич	-		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Губин Евгений Иванович	к.ф.-м.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ПМ7И	Шкабара Анастасия Игоревна

Школа	ИШИТР	**Отделение школы (НОЦ)**	ОИТ
Уровень образования	Магистратура	**Направление/специальность**	09.04.04. Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*- Оклад научного руководителя – 33664;* *- Оклад консультанта – 26624;* *- Оклад инженера – 21760;*
2. *Нормы и нормативы расходования ресурсов*	*- Годовая норма амортизации составляет 33.3 %.*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*- Размер страховых взносов равный 30%;* *- Районный коэффициент г. Томск. 30%;*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*Анализ перспективности исследования с помощью технологии QuaD-анализа;* *Оценка готовности проекта к коммерциализации*
2. *Разработка устава научно-технического проекта*	*Постановка целей проекта, определение ожидаемого результата и критериев приемки проекта, рабочей группы*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*Планирование работ* *Построение диаграммы Ганта;* *Формирование бюджета затрат;* *Анализ рисков проекта.*
4. *Определение ресурсной, финансовой, экономической эффективности*	*Расчет интегрального показателя ресурсоэффективности. Общие выводы.*
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. *Технология QuaD* 2. *Оценка степени готовности проекта к коммерциализации* 3. *Заинтересованные стороны проекта* 4. *Диаграмма Ганта* 5. *Структура затрат* 6. *Реестр рисков*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель ОСГН ШБИП	Потехина Нина Васильевна	-		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Шкабара Анастасия Игоревна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ПМ7И	Шкабара Анастасия Игоревна

Школа	ИШИТР	**Отделение школы (НОЦ)**	ОИТ
Уровень образования	Магистратура	**Направление/специальность**	09.04.04 Программная инженерия

Исходные данные к разделу «Социальная ответственность»:	
1. Описание рабочего места (рабочей зоны, технологического процесса, механического оборудования)	*В соответствии с ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования».*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности.	*- ГОСТ 12.2.032-78 ССБТ - СанПиН 2.2.4.548–96 - СанПиН 2.2.4/2.1.8.562–96 - СанПиН 2.2.2/2.4.1340–03 - ГОСТ 12.1.009-2009 - ГОСТ 12.1.038-82 ССБТ - ГОСТ Р 22.3.03-94*
2. Выявление и анализ вредных факторов проектируемой производственной среды	*- Освещение - Микроклимат - Шум - Психофизиологические факторы: нервно-психические перегрузки*
3. Выявление и анализ опасных факторов проектируемой произведённой среды.	*- Электрический ток (источник – ПК) - Короткое замыкание - Статическое заземление (источник – ПК)*
4. Охрана окружающей среды.	*Воздействие объекта на атмосферу, гидросферу отсутствует. Воздействие на литосферу происходит при утилизации ПК, используемого для разработки, а также утилизации люминесцентных ламп освещения.*
5. Защита в чрезвычайных ситуациях.	*Возможной чрезвычайной ситуацией при разработке алгоритма является возникновение пожара на рабочем месте.*

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ООД ШБИП	Горбенко М. В.	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Шкабара Анастасия Игоревна		

РЕФЕРАТ

Выпускная квалификационная работа **111** с, **20** рис., **22** табл., **33** источников, **2** прил.

Ключевые слова: кластеризация, кластерный анализ, снижение размерности, анализ данных, большая вычислительная сеть, обработка лог-файлов, машинное обучение, Большой Адронный Коллайдер, WLCG.

Объектом исследования являются анализ и обработка данных журналов обработки задач в больших вычислительных системах, снятые с Большого Адронного Коллайдера.

Цель работы - поиск закономерностей, влияющих на предсказание длительности выполнения задач в цепочках с помощью предварительной кластеризации.

В процессе исследования проводились изучение существующих методов кластеризации, предварительная обработка данных журналов обработки задач, реализация методов снижения размерности – PCA и T-SNE, и методов кластеризации – k-means, иерархический и DBSCAN.

В результате исследования было произведено обогащение данных несколькими способами, выбран наиболее оптимальный метод кластеризации для данного набора данных.

Область применения: Планировщик задач WLCG.

В будущем планируется встроить данный метод кластеризации в существующую систему предсказания длительности обработки задач.

УСЛОВНЫЕ ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

БАК – Большой адронный коллайдер;

LHC – Large Hadron Collider;

ЦЕРН (CERN) – Европейский Центр Ядерных Исследований;

WLCG (Worldwide LHC Computing Grid) – Всемирная Распределенная Сеть Большого Адронного Коллайдера или Всемирный Грид для Большого Адронного Коллайдера;

НТИ – Научно-техническое исследование;

PCA – Principal component analysis;

DBSCAN – Density-Based Spatial Clustering and Application with Noise

T-SNE - t-Distributed Stochastic Neighbor Embedding

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В современном мире одним из наиболее актуальных видов физических исследований являются эксперименты по физике высоких энергий, вносящие неоспоримый вклад в фундаментальную науку. В результате таких исследований образуются огромное число данных, которые фиксируются детекторами ускорителя заряженных частиц. Обработка такого объема данных требует больших вычислительных мощностей, поэтому и создаются распределенные системы обработки данных, в которые входят большое количество суперкомпьютеров. Однако использование такого оборудования обходится дорого, поэтому необходима организация рационального планирования обработки данных, во избежание простоя оборудования и неравномерного распределения задач по обработки, среди суперкомпьютеров системы обработки данных. Очевидно, что определение времени обработки данных являются ключевой задачей для организации системы планирования. Решением такой задачи является система, способная предсказывать время обработки данных.

Данное исследование является частью работы по предсказыванию длительности выполнения заданий в большой вычислительной сети Большого Адронного Коллайдера. Было высказано предположение, что есть закономерности, которые влияют на выполнение задач и время их окончания. Предполагается, что предварительный кластерный анализ этих задач поможет предсказывать длительность обработки точнее.

Цель данной работы - поиск закономерностей, влияющих на предсказание длительности выполнения задач в цепочках с помощью предварительной кластеризации.

Из этой цели вытекают следующие задачи:

1. Исследование методов кластеризации многомерных данных без учителя.
2. Реализация алгоритмов снижения размерности и кластеризации
3. Сравнение и выбор наиболее оптимального метода для данного набора данных

Данная работа является актуальной, потому что ученым важно знать, будет ли задача выполнена завтра или через год, чтобы планировать эксперименты. Часто задача состоит из более чем 1000 заданий и если какие-то события вызывают ошибки, то одно необработанное задание приводит к тому, что вся задача считается необработанным.

Важно уметь предсказывать данные аномалии и устранять ошибки. Одним из этапов данной работы является кластеризация заданий.

Научная Новизна исследования заключается в том, что применение алгоритмов кластеризации к журналам WLCG ранее не делалось. Были попытки применить алгоритмы машинного обучения, но с учителем, а в данном случае необходимо предсказывать тип, к которому может относиться задание заранее.

ГЛАВА 1 АНАЛИТИЧЕСКИЙ ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1 Особенности распределенных вычислений в WLCG

В современном мире широкое распространение получают глобальные исследования, которые находятся на стыке различных наук, таких как физика высоких энергий, астрофизика, биология, науки о Земле и другие. Для осуществления таких исследований необходима обработка большого объема данных в достаточно ограниченное время. Для этого стали использоваться географически распределенные вычислительные системы. В их возможности входит прием и передача сотни терабайт данных, обработка сотни тысяч задач и долговременное хранение большого количества информации.

Современные грид-инфраструктуры обеспечивают интеграцию аппаратурных и программных ресурсов, находящихся в разных организациях в масштабах стран, регионов, континентов в единую вычислительную среду, позволяющую решать задачи по обработке сверхбольших объемов данных, чего в настоящее время невозможно достичь в локальных вычислительных центрах.

На данный момент, самые эффективные результаты в области распределенных вычислений принадлежат проекту WLCG (Worldwide LHC Computing Grid или Всемирный грид для Большого адронного коллайдера), который базируется в ЦЕРН и создан для обработки данных с экспериментов, проводимых на Большом Адронном Коллайдере (англ. LHC- Large Hadron Collider). В рамках данного проекта была выстроена иерархическая система региональных центров, которая включает в себя несколько уровней [1].

Суть данной вычислительной модели состоит в том, что все данные с детекторов коллайдера проходят первоначальную обработку в реальном времени а так же первичную реконструкцию (восстановление треков частиц, их импульсов и других характеристик из хаотического набора сигналов от различных регистрирующих систем). Затем эти данные отправляются для обработки и анализа в региональные центры разных уровней (Tier-ы):

Tier0 (CERN) => Tier1 => Tier2 => Tier3 => компьютеры пользователей

Уровни различаются по масштабу ресурсов (сетевые, вычислительные, дисковые, архивные) и по выполняемым функциям:

- Tier0 (ЦЕРН) - первичная реконструкция событий, калибровка, хранение копий полных баз данных

- Tier1 - полная реконструкция событий, хранение актуальных баз данных по событиям, создание и хранение наборов анализируемых событий, моделирование, анализ

- Tier2 - репликация и хранение наборов анализируемых событий, моделирование, анализ.

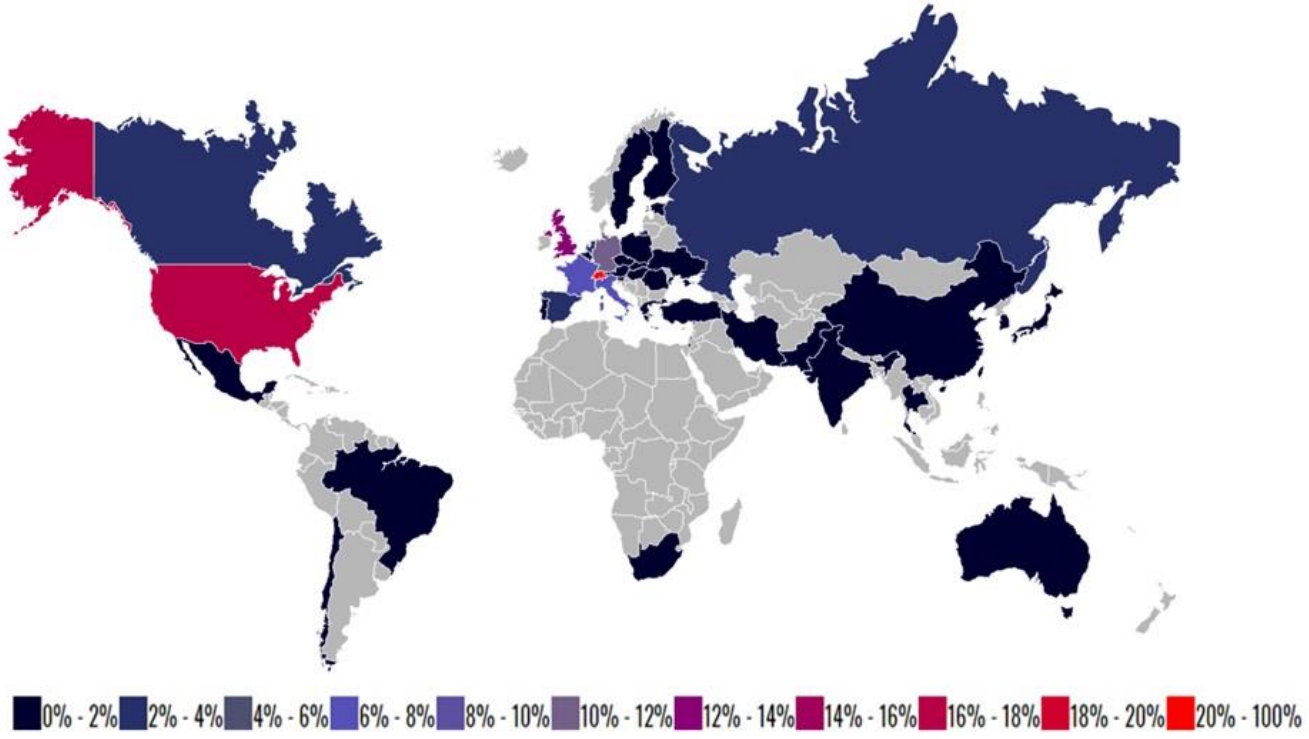

Рисунок 1 – Страны, в которых происходит обработка заданий на узлах.

В настоящее время проект WLCG объединяет более 150 грид-сайтов, более 300000 ЦПУ, более 250 Пбайт систем хранения данных на дисках и ленточных роботах. [13].

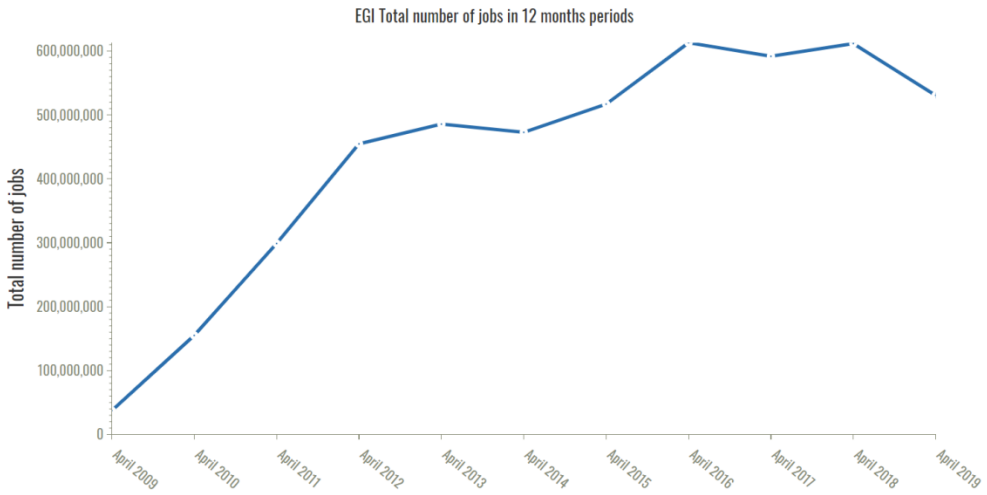

Рисунок 2 – Число выполненных заданий на узлах WLCG по годам.

1.2 Предсказание длительности выполнения задач по данным с БАК

Столкновения частиц описываются определенным набором данных (dataset), содержащим информацию о каждом столкновении. В один набор входит информация о нескольких часах работы коллайдера – один цикл. В эти данные входят дата, время работы коллайдера, настройки, с которыми проходили данные эксперименты, а также описания независимых событий – столкновений, собранных с различных датчиков детектора.

Физики присылают запрос на работу с определенным набором данных, они описывают действия, необходимые для получения результатов. Часто эти действия содержат в себе несколько задач (Task), которые могут выполняться друг за другом или параллельно, а также содержать в себе работу над несколькими наборами данных одновременно.

Задача разбивается на задания (Job), каждое задание состоит из нескольких событий (Event). Каждое задание содержит столько событий, сколько один узел может обработать одновременно. Событие описывает одно столкновение, так как все данные из набора независимы друг от друга.

Сначала, задача делит набор данных на задания. Эти задания распределяются между узлами Распределенной Сети и начинают выполнение. Как только все события в задании обработаны, эти данные готовы перейти к следующей задаче, не дожидаясь остальных. Следующая задача может

перераспределить количество необходимых событий внутри задания. И как только достаточное количество событий готово, чтобы укомплектовать новое задание, задача начинает свое выполнение.

Отсюда следует, что задачи могут выполняться не последовательно, а с наложением, которое зависит от заданий и времени их выполнения.

Но не все задачи могут обрабатываться параллельно. Такие задачи как сравнение или слияние двух наборов данных должны иметь два готовых обработанных набора, соответственно могут выполняться только последовательно.

То, насколько задачи перекрывают друг друга, зависит от параметров цепочки задач, которые мы можем дать модели машинного обучения.

Как правило, задачи, которые физики ставят для обработки, являются стандартными задачами. По текущей цепочке и ее параметрам, а также по имеющимся уже примерам цепочек задач с определенным временем выполнения мы можем предсказать время работы текущей цепочки. В результате предыдущих исследований был разработан алгоритм, предсказывающий длительность заданий по параметрам, описывающим столкновения, а также длительность выполнения цепочек задач.

Только когда все 100% заданий обработаны, задачу можно считать успешно выполненной. Часто задача состоит из более чем 1000 заданий и если какие-то события вызывают ошибки, то одно необработанное задание приводит к тому, что все задание считается необработанным. Также, важно уметь предсказывать данные аномалии и устранять баги.

Решением данной проблемы может послужить предварительная кластеризация данных журналов заданий.

1.3 Обзор подходов к кластерному анализу в научной литературе

В настоящее время вопросам кластеризации посвящено большое множество исследований, так как данный метод анализа данных подходит для любых сфер и типов задач. Многие исследования сравнивают имеющиеся методы

для конкретного набора данных. Например, в работе [12] исследовались финансовые данные и в этом случае метод иерархических деревьев DBHT и k-medoids показывают лучшие характеристики, но последний больше зависит от шума чем DBHT.

Показатели точности прогнозирования в [11] показывают, что кластерная регрессия дает чрезвычайно точные прогнозы, но несколько нестабильные кластеры, в то время как k-means дает более стабильные кластеры, но очень плохие прогнозы в отдельных кластерах на наборе данных энергопотребления.

В работе [10] выполнено сравнение 18 методов кластеризации для анализа многомерных данных массы и потока цитометрии. В данном случае наиболее оптимальным оказался метод FlowSOM с выбором числа кластеров вручную.

Значительные улучшения в кластеризации можно наблюдать, если использовать подходящий метод снижения размерности. В работе [8] рассмотрен принцип предварительной кластеризации базы данных отпечатков пальцев для более скорого поиска совпадений в базе в следующем порядке: Сначала для дальнейшей обработки подготавливается база изображений отпечатков пальцев. Затем каждый отпечаток представляется в виде числового вектора признаков для дальнейшего анализа. После этого данные векторов кластеризуются, а затем с помощью результатов кластеризации и машинного обучения уже новые отпечатки сопоставляются с полученными кластерами. В процессе кластеризации использовались методы снижения размерности PCA и поле направлений, и методы кластеризации K-means и агломеративный иерархический. Для данных изображений наиболее точные результаты дал метод главных компонент вместе с алгоритмом K-means.

Большое влияние на точность оказывает тип признаков. Категориальные признаки могут значительно улучшить алгоритм или, по крайней мере снизить негативное влияние на результат за счёт удаления категориального столбца. Так, например, в работе [9] предложен пространственный структурно основанный метод для кластеризации категориальных признаков, что дает выигрыш в точности кластеризации на разных наборах данных. Так же в работе [13] был

представлен алгоритм k-prototypes для кластеризации больших наборов реальных данных. Сохраняя эффективность алгоритма k-средних, и убирая его ограничения на только числовые данные было показано, что он эффективен для кластеризации больших наборов данных со смешанными числовыми и категориальными значениями.

Практически для всех данных процесс кластеризации сводится к следующим универсальным шагам: предварительная обработка, снижение размерности, кластеризация, визуализация, оценка результатов. Но, как показывает обзор, для каждого набора данных и цели анализа необходим свой уникальный подход с подобранными параметрами алгоритмов кластеризации, а так же грамотная предварительная обработка и очистка данных.

1.4 Классификация методов кластеризации.

Кластерный анализ представляет собой класс методов, используемых для классификации объектов в группы, которые называют кластерами. Кластерный анализ также называется классификационным анализом или численной таксономией. Главная его особенность - отсутствие предварительной информации о принадлежности к группе или кластеру любого из объектов [16].

Задача кластеризации – используя все имеющиеся данные, предсказать соответствие объектов выборки их классам, сформировав, таким образом, кластеры.

Кластеризация включает формулирование задачи, выбор меры расстояния, выбор метода кластеризации, определение количества кластеров, интерпретацию профильных кластеров и, наконец, оценку результатов.

Признаки, по которым проводится кластерный анализ, должны выбираться с учетом предыдущих исследований, они должны быть приведены в пригодный для анализа вид, очищены от ненужной информации. Для кластеризации следует выбрать соответствующую меру расстояния или подобия; наиболее часто используемая мера это евклидово расстояние или его квадрат.

Все атрибуты, или признаки объектов делятся на числовые (numerical) и категорийные (categorical). Числовые атрибуты – это такие, которые могут быть упорядочены в пространстве, соответственно категорийные – которые не могут быть упорядочены. Например, атрибут "возраст" – числовой, а "цвет" – категорийный. Приписывание атрибутам значений происходит во время измерений выбранным типом шкалы, а это, вообще говоря, представляет собой отдельную задачу.

В общем виде методология кластерного анализа заключается в следующем:

1. Выбор объектов для кластеризации.
2. Определение признаков, для оценки объектов в выборке. При необходимости – нормализация значений переменных.
3. Вычисление значений меры сходства между объектами.

4. Применение метода кластерного анализа для создания групп сходных объектов (кластеров).

5. Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Основные типы методов кластерного анализа:

- Разделяющая кластеризация
- Иерархическая кластеризация
- Нечеткая кластеризация
- Кластеризация на основе плотности
- Кластеризация на основе моделей

Разделяющая кластеризация - это методы кластеризации, относящая объекты в наборе данных к разным группам на основе их сходства. Алгоритмы требуют от аналитика указания количества кластеров, которые будут сгенерированы. Самые часто популярные методы разделяющей кластеризации:

- K-means или кластеризация k-средних, в которой каждый кластер представлен центроидом или средним точек, принадлежащих кластеру. Метод k - средних чувствителен к аномальным точкам и выбросам.
- K-medoids или PAM (Partitioning Around Medoids - разделение вокруг медоидов), в которой каждый кластер представлен одним из объектов в кластере. PAM менее чувствительна к выбросам по сравнению с k-means.
- Алгоритм CLARA (Clustering Large Applications - кластеризация больших приложений), является версией PAM, адаптированный для больших наборов данных.

Метод иерархической кластеризации является альтернативным подходом к разделяющей кластеризации для группировки объектов на основе их сходства. В отличие от нее, иерархическая кластеризация не требует

предварительного указания количества создаваемых кластеров. Результатом иерархической кластеризации является древовидное представление объектов, которое также известно как дендрограмма.

Нечеткая кластеризация рассматривается как гибкая кластеризация, в которой каждый элемент может принадлежать к каждому кластеру. Другими словами, каждый элемент имеет набор коэффициентов членства степени нахождения в каждом кластере. Этот метод отличается от кластеризации k-means и k-medoid, где каждый объект точно принадлежит к одному кластеру.

В нечеткой кластеризации точки, близкие к центру кластера, могут находиться в кластере в большей степени, чем точки на краю кластера. Степень, в которой элемент принадлежит данному кластеру, является числовым значением от 0 до 1.

Алгоритм fuzzy c-means (FCM) является одним из наиболее широко используемых алгоритмов нечеткой кластеризации. Центроид кластера вычисляется как среднее всех точек, взвешенных по степени их принадлежности к кластеру.

Кластеризация на основе моделей. Традиционные методы кластеризации, такие как иерархическая кластеризация и k-means, являются эвристическими и не основаны на формальных моделях. Кроме того, алгоритм k-means обычно инициализируется случайным образом, поэтому разные прогоны часто дают разные результаты. Кроме того, k-means требует заранее указывать оптимальное количество кластеров.

Альтернативой является кластеризация на основе моделей, которая рассматривает данные как исходящие из распределения, представляющего собой смесь двух или более кластеров. В отличие от k-средних, кластеризация на основе модели использует мягкое назначение, где каждая точка данных имеет вероятность принадлежности к каждому кластеру.

Основная идея подхода **кластеризации на основе плотности** основана на интуитивном методе кластеризации человека. Основанный на плотности алгоритм

кластеризации, может использоваться для идентификации кластеров любой формы в данных, содержащих шум и выбросы.

Кластеры в данном методе – это плотные области в пространстве данных, разделенные областями с меньшей плотностью точек. Алгоритм DBSCAN основан на этом интуитивном понятии кластеров и шума. Ключевая идея заключается в том, что для каждой точки кластера окрестности заданного радиуса должны содержать, по крайней мере, минимальное количество точек.

1.5 Методы кластеризации

1.5.1 Метод k-средних

Наиболее популярным алгоритмом кластеризации данных является метод k-средних. Это итеративный алгоритм кластеризации, основанный на минимизации суммарных квадратичных отклонений точек кластеров от центроидов (средних координат) этих кластеров.

Пусть $X=\{x_i\}$, $i=1,\ldots,n$ – множество n d-мерных точек для кластеризации в набор из K кластеров, $C = \{c_k, k = 1,\ldots, K\}$. Алгоритм находит такой разделение, чтобы квадрат ошибки между эмпирическим средним значением кластера и точек в кластере был сведен к минимуму.

Пусть μ_k - среднее значение кластера c_k. Квадрат ошибки между μ_k и точки в кластере c_k определены как

$$J(c_k) = \sum_{x_i \in c_k} \|x_i - \mu_k\|^2.$$

Задача k-means – минимизировать сумму квадратичной ошибки во всех K кластерах:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} \|x_i - \mu_k\|^2.$$

Алгоритм k-means начинается с первоначального отделения K кластеров и назначает шаблоны кластерам, чтобы уменьшить квадрат ошибки. Поскольку квадратичная ошибка всегда уменьшается с увеличением числа кластеров K (J(C)

= 0 при K=n), она может быть сведена к минимуму только для фиксированного числа кластеров. Основные шаги алгоритма k-means следующие:

1. Выбор начального разделения с K кластерами; Повтор 2 и 3 этапов, пока членство в кластере не стабилизируется.

2. Создание нового разделения, назначая каждый объект к ближайшему центру кластера.

3. Вычисление новых центров кластеров.

На рис. 5 изображены этапы кластеризации методом k-means.

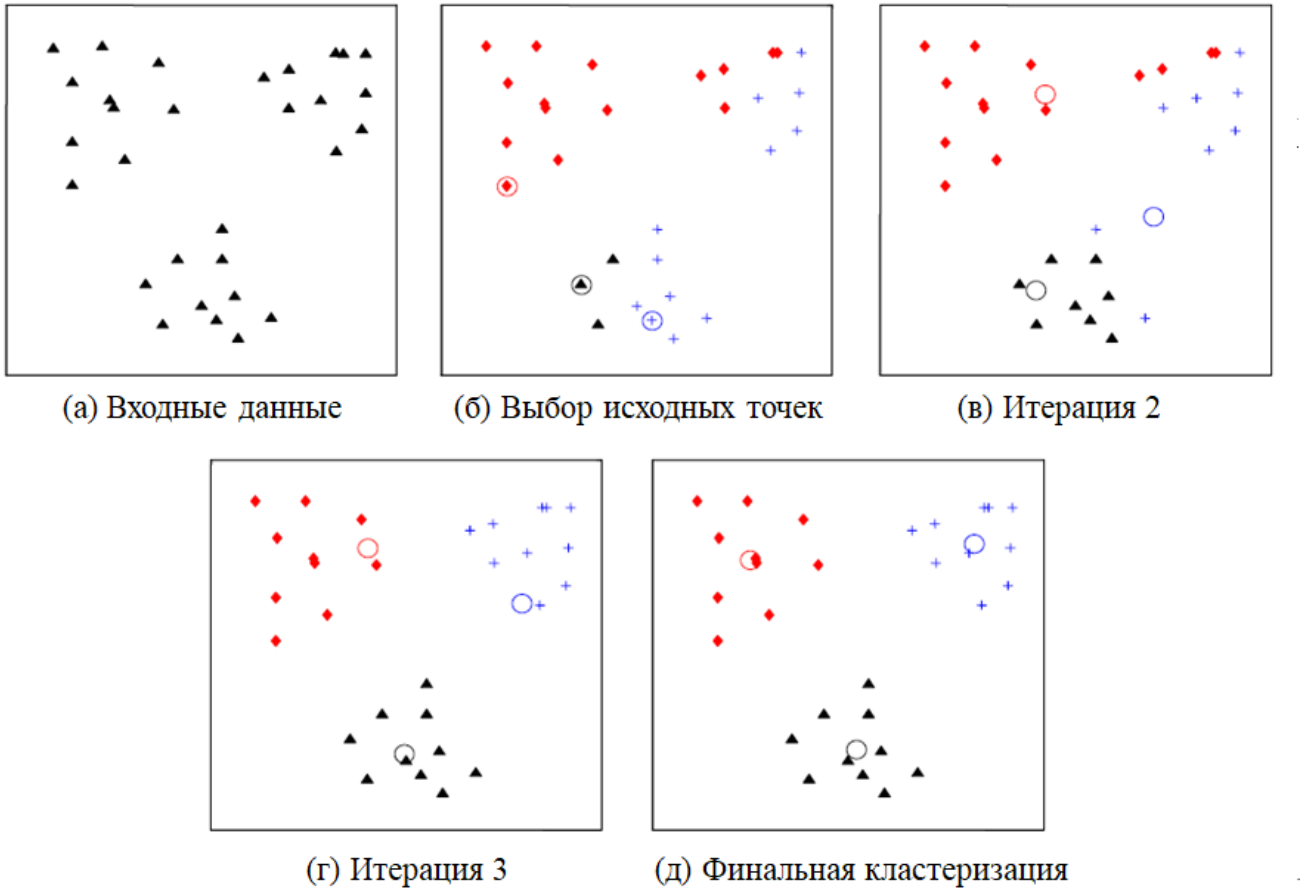

Рисунок 3 – Иллюстрация алгоритма k-means. (а) двумерные входные данные с тремя кластерами; (б) три начальные точки, выбранные в качестве центров кластера, и первоначальное распределение точек по кластерам; (в) и (г) промежуточные итерации, обновляющие метки кластеров и их центры; (д) окончательная кластеризация, полученная с помощью алгоритма k-means [17].

1.5.2 Алгоритмы иерархической кластеризации

Среди алгоритмов иерархической кластеризации выделяются два основных типа: восходящие и нисходящие алгоритмы. Нисходящие алгоритмы работают по принципу «сверху-вниз»: в начале, все объекты помещаются в один кластер, который затем разбивается на все более мелкие кластеры. Более распространены восходящие алгоритмы, которые в начале работы помещают каждый объект в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержаться в одном кластере. Таким образом строится система вложенных разбиений. Результаты таких алгоритмов обычно представляют в виде дерева – дендрограммы. Классический пример такого дерева на рис. 6.

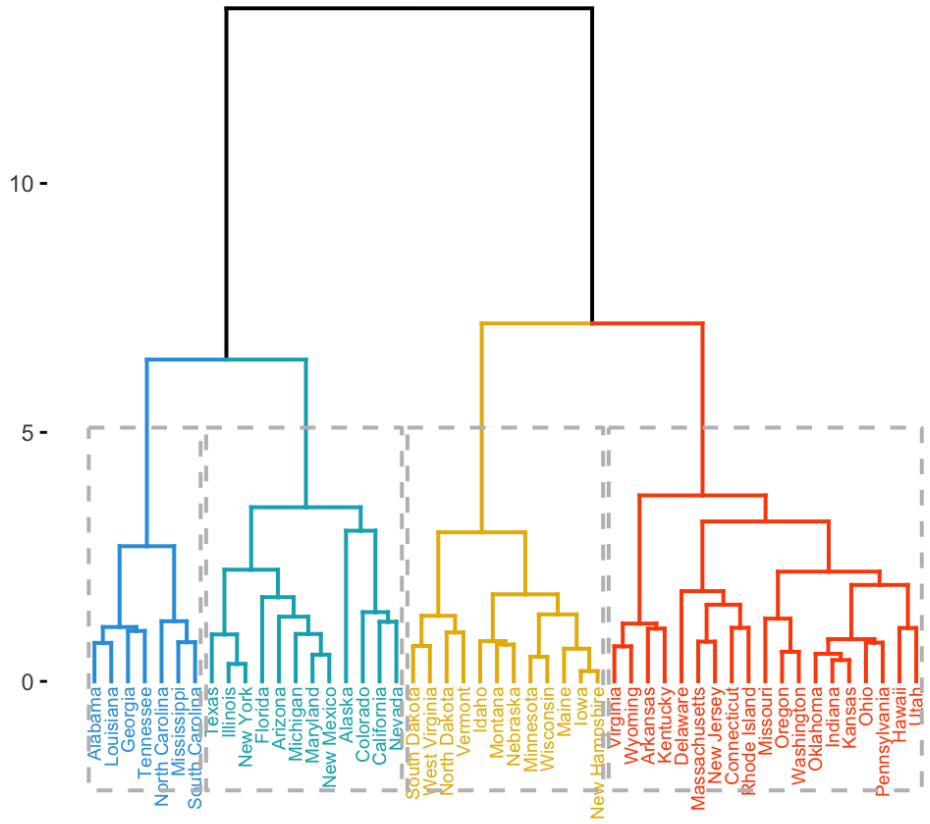

Рисунок 4 – Иллюстрация восходящего алгоритма иерархической кластеризации.

Для вычисления расстояний между кластерами чаще все пользуются двумя расстояниями: одиночной связью или полной связью.

1.5.3 Метод кластеризации на основе плотности DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise, плотностной алгоритм пространственной кластеризации с присутствием шума) – популярный алгоритм кластеризации, используемый в анализе данных в качестве одной из замен метода k-средних.

DBSCAN, использует простую оценку уровня минимальной плотности, основанную на пороге для числа соседей, *minPts*, в пределах радиуса ε (с произвольной мерой расстояния). Объекты с более чем *minPts* соседями в пределах этого радиуса (включая точку запроса) считаются базовой точкой. Суть DBSCAN заключается в том, чтобы найти те области, которые удовлетворяют этой минимальной плотности, и которые разделены областями более низкой плотности. По соображениям эффективности DBSCAN не выполняет оценку плотности между точками. Вместо этого все соседи в радиусе ε основной точки считаются частью того же кластера, что и основная точка (называемая достижимой прямой плотностью). Если любой из этих соседей снова является основной точкой, их окрестности транзитивно включены (плотность достижима). Неосновные точки в этом наборе называются пограничными точками, и все точки в пределах одного набора связаны плотностью. Точки, которые не являются плотностью, достижимой из любой основной точки, считаются шумом и не принадлежат ни к одному кластеру [18].

Рисунок 5 иллюстрирует основные понятия из DBSCAN. Параметр *minPts* равен 4, а радиус ε обозначается окружностями. N является точкой шума, A является основной точкой, а точки B и C пограничными.

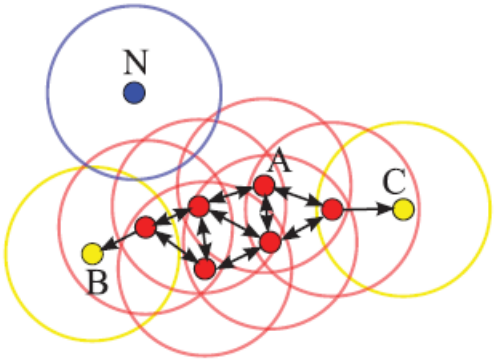

Рисунок 5 – Иллюстрация алгоритма DBSCAN

Стрелки указывают на прямую достижимость плотности. Точки B и C связаны плотностью, потому что обе плотности достижимы из A. N не достижима плотностью, и поэтому считается точкой шума.

1.6 Очистка данных

Выделяют следующее этапы очистки данных:

1. Анализ данных (Data analysis). Для того чтобы определить, какие виды ошибок и несоответствий должны быть удалены, требуется детальный анализ данных. В дополнение к инспекции данных или отдельных выборок данных «вручную», следует использовать и метаданные.

2. Определение способов трансформации потоков данных и правил отображения (Definition of transformation workflow and mapping rules). На данном этапе выполняется оценка количества источников данных, степени их неоднородности и «загрязненности». На основе этой информации создаются схемы потоков данных, позволяющих преобразовать множество источников данных в один, избегая создания ошибок Multi-Source слияния (например, появление дублирующих записей).

3. Верификация (Verification). Оценка корректности и результативности выполнения предыдущего этапа (например, на небольшой выборке данных). При необходимости производиться возврат к этапу 2 для его повторного выполнения.

4. Трансформация (Transformation). Загрузка данных в единое хранилище с использованием правил трансформации, определенных и отлаженных на этапах 2–3. Очистка данных уровня Single-Source.

5. Обратная загрузка очищенных данных (Backflow of cleaned data). Имея на этапе 4 очищенный набор данных в едином хранилище, целесообразно этими «чистыми» данными заменить аналогичные «грязные» данные в исходных источниках. Это позволит в будущем во многом не выполнять повторно все этапы преобразований по очистке данных.

Реализовать эти этапы можно самыми различными путями с использованием существующих и созданных специально способов и технологий. Рассмотрим наиболее интересные из них.

Этап анализа данных предполагает анализ использования метаданных, которых, как правило, недостаточно для оценки качества данных из имеющихся источников. Поэтому важно анализировать реальные примеры данных, оценивая их характеристики и сигнатуры значений. Это позволяет находить взаимосвязи между атрибутами в схемах данных различных источников. Выделяют два подхода решения этой задачи – профилирование данных (англ. Data profiling) и извлечение данных (data mining).

Профилирование данных сориентировано на анализ индивидуальных атрибутов, характеризующихся их конкретными свойствами: тип данных, длина, диапазон значений, частота встречаемости дискретных значений, дисперсия, уникальность, встречаемость «null» значений, типичная сигнатура записи (например, у телефонного номера). Именно набор подобных свойств (профиль) позволяет оценить различные аспекты качества данных.

Извлечение данных предполагает поиск взаимосвязей между несколькими атрибутами достаточно большого набора данных. Учитывая то, что этот способ получил название data mining, здесь используют упоминавшиеся выше (см. табл. 1) методы кластеризации, подведения итогов, поиска ассоциаций и последовательностей.

Кроме того, для дополнения пропущенных значений, корректировки недопустимых значений или идентификации дубликатов могут быть использованы существующие ограничения целостности (англ. integrity constraints), принятые в реляционных базах данных, наложенные дополнительно

на бизнес-связи между атрибутами. Например, известно, что «Total = Quantity×Unit_Price». Все записи, не удовлетворяющие этому условию, должны быть изучены более внимательно, исправлены или исключены из рассмотрения.

Для разрешения проблем очистки данных в одном источнике (single-source problems), в том числе перед его интеграцией с другими источниками данных, реализуют следующие этапы:

- Извлечение значений из атрибутов свободной формы (разбиение атрибутов, англ. Extracting values from free-form attributes (attribute split)). В данном случае речь может идти о строковых значениях, сохраняющих несколько слов подряд (например, адрес или полное имя человека). В данном случае требуется четкое понимание того, на какой позиции этого значения находится интересующая нас часть атрибута. Возможно, потребуется даже сортировка составных частей такого атрибута.

- Валидация и коррекция (англ. Validation and correction). Данный этап предполагает поиск ошибок ввода данных и их исправление наиболее автоматическим способом. Например, используя автоматическую проверку правописания во избежание орфографических ошибок и опечаток. Словарь географических названий и почтовых кодов также следует использовать для корректировки значений вводимых адресов. Зависимость атрибутов (дата рождения – возраст, Total = Quantity×Unit_Price и т.п.) также способствует избеганию множества ошибок в данных.

- Стандартизация (англ. Standardization). Этот этап предполагает приведение всех данных к единому универсальному формату. Примерами таких форматов являются формат написания даты и времени, размер регистра в написании строковых значений. Текстовые поля должны исключать префиксы и суффиксы, аббревиатуры в них должны быть унифицированы, исключены проблемы с различной кодировкой.

Одной из основных проблем, вызванных интеграцией различных источников (multi-source problems) данных, является устранение дублирования записей. Этот этап выполняется после подавляющего большинства

преобразований и чисток. Он предполагает сначала идентификацию сходных в некотором смысле записей, а затем их слияние с объединением атрибутов. Очевидно, решение этой задачи при наличии у дублирующих записей первичного ключа достаточно просто. Если такого однозначно идентифицирующего признака нет, то задача устранения дубликатов значительно осложняется, требуя применения нечетких (англ. fuzzy) подходов сравнения (близости в некотором смысле) записей между собой.

1.7 Снижение размерности.

Основное отличие методов обучения без учителя от привычных классификаций и регрессий машинного обучения в том, что разметки для данных в этом случае нет. От этого образуются сразу несколько особенностей - во-первых это возможность использования несопоставимо больших объёмов данных, поскольку их не нужно будет размечать руками для обучения, а во-вторых это неясность измерения качества методов, из-за отсутствия таких же прямолинейных и интуитивно понятных метрик, как в задачах обучения с учителем.

Одной из самых очевидных задач, которые возникают в голове в отсутствие явной разметки, является задача снижения размерности данных. С одной стороны её можно рассматривать как помощь в визуализации данных. С другой стороны, подобное снижение размерности может убрать лишние сильно скоррелированные признаки у наблюдений и подготовить данные для дальнейшей обработки в режиме обучения с учителем, например сделать входные данные более "перевариваемыми" для деревьев решений. На рис. 6 представлена классификация основных методов снижения размерности.

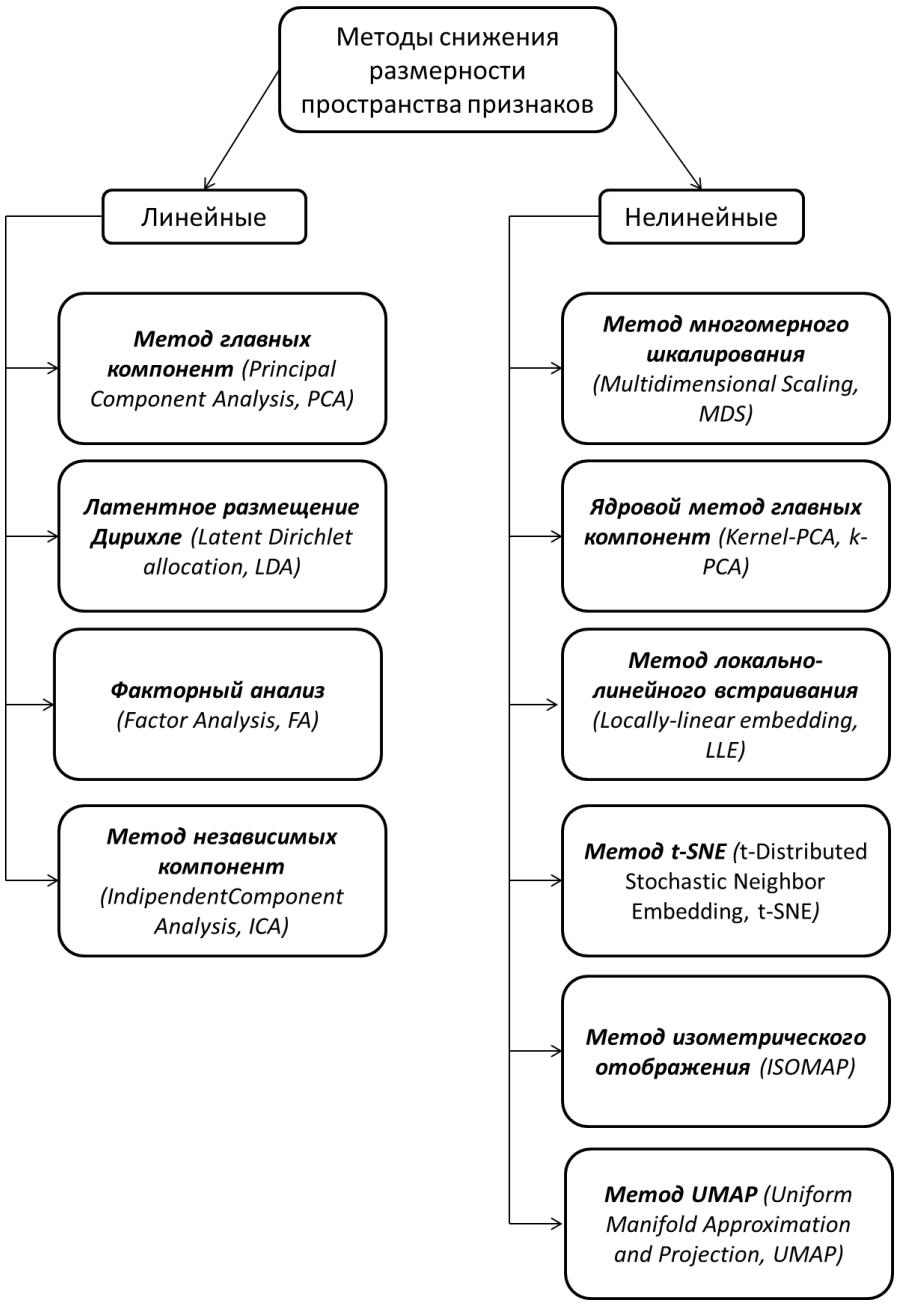

Рисунок 6 – Классификация методов снижения пространства.

Для исследования были выбраны следующие методы снижения размерности:

- Principal Component Analysis (PCA),
- t-distributed Stochastic Neighbor Embedding (t-SNE)

1.7.1 Метод главных компонент (PCA)

Метод главных компонент (разложение Карунена-Лоева, principal component analysis, PCA) является простейшим методом уменьшения размерности в данных. Идея метода заключается в поиске в исходном пространстве

гиперплоскости заданной размерности с последующим проектированием выборки на данную гиперплоскость. При этом выбирается та гиперплоскость, ошибка проектирования данных на которую является минимальной в смысле суммы квадратов отклонений [2].

Пусть набор данных представлен матрицей X размера M×N. Каждый столбец, $X^{(n)}$, *n = 0, …, N -1*, содержит все признаки одного атрибута. Кроме того, предполагается, что каждый столбец имеет средний центр, т. е. если, $\tilde{X}^{(n)}$ является исходным вектором, то

$$X^{(n)} = \tilde{X}^{(n)} - N^{-1} \sum_{n=0}^{N-1} \tilde{X}^{(n)} ,$$

PCA преобразует множество входных столбцов векторов $\left[X^{(0)}|\ldots|X^{(N-1)}\right]$ в другой набор столбцов векторов $\left[T^{(0)}|\ldots|T^{(N-1)}\right]$, называемый баллами главных компонент. Это преобразование имеет такое свойство, что большая часть информационного содержимого исходных данных (или большая часть его дисперсии) хранится в первых нескольких баллах компонентов.

Это позволяет уменьшить данные до меньшего количества измерений, с низкой потерей информации, просто отбрасывая последние оценки компонент. Каждая компонента является линейной комбинацией исходных данных и каждая компонента ортогональна. Это линейное преобразование матрицы X задается матрицей P N×N, так что матрица X факторизуется как:

$$X = TP^T,$$

где P известна как матрица нагрузки [19].

Несмотря на относительную простоту, PCA обладает двумя существенными недостатками. Во-первых, при использовании PCA предполагается, что распределение исходных многомерных данных подчинено нормальному закону и трансформация происходит относительно многомерного гиперэллипсоида рассеивания (хотя исходные измерения могут быть распределены не в рамках такого гиперэллипсоида). Во-вторых, трансформация исходного признакового пространства может повлечь за собой значительные искажения признакового пространства, что может привести к снижению

разделимости в таком новом признаковом пространстве для объектов и снизить итоговое качество классификации.

1.7.2 Стохастическое вложение соседей с t-распределением (T-SNE)

T-SNE - это алгоритм для понижения размерности, разработанный Лоренсом ван дер Маатеном и Джеффри Хинтоном. Метод проецирует каждый объект высокой размерности в заданную размерность таким образом, что похожие объекты проецируются близко расположенными точками, а непохожие точки проецируются расположенными далеко друг от друга [4].

На начальном этапе t-SNE состоит из набора попарных (евклидовых) расстояний δ_{ij} между n входными объектами, которые преобразуются в условные вероятности

$$p_{i|j} = \frac{\exp(^{-\delta_{ij}^2}/_{2\sigma_i})}{\sum_{k \neq l} \exp(^{-\delta_{ik}^2}/_{2\sigma_i})},$$ (где $p_{i|j} = 0$ и σ_i задается таким образом, чтобы

получить условные распределения вероятностей с фиксированной сложностью), которые, в свою очередь, симметричны для получения совместной матрицы вероятностей P над парами точек с записями $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$. Попарные расстояния между точками y_iна карте входных объектов аналогично преобразуются в совместную матрицу вероятностей Q, но вместо гауссовых плотностей используются плотности при распределении Стьюдента-t

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|y_k - y_l\|^2)^{-1}}.$$

Точки y_i создаются таким образом, чтобы минимизировать расхождение Кульбака-Лейблера между совместными вероятностными распределениями P и Q, т. е. целевая функция $C(Y) = const - \sum_i \sum_{j \neq i} p_{ij} \log q_{ij}$ [20.]

ГЛАВА 2. ПРИМЕНЕНИЕ ПОДХОДОВ КЛАСТЕРНОГО АНАЛИЗА К ПРЕДМЕТНОЙ ОБЛАСТИ

Для исследований и анализа данных, а так же для машинного обучения существует много инструментов, но самыми распространенными являются Python и R. Оба языка имеют большое количество библиотек для работы с данными, кластерного анализа, визуализации и машинного обучения. Для данного исследования был выбран Python 3.7, так как он является наиболее актуальным и более гибким для современных методов машинного обучения. В качестве оболочки использован Jupyter Notebook.

Алгоритм работы следующий

1. Отбор выборки объектов для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.
3. Вычисление значений меры сходства между объектами.
4. Применение метода кластерного анализа для создания групп сходных объектов (кластеров).
5. Представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата

3.1 Описание входных данных

Данные для данного исследования были взяты из планировщика задач WLCG. Данные содержат информацию о заданиях, поставленных на выполнение в WLCG с 04.06.2017 по 17.06.2017. В таблице описываются основные характеристики заданий, такие как время выполнения, даты начала и конца задачи, количество ядер, вычислительный узел и т.д. Размер исходной таблицы 10517 × 46 Описание таблицы с числом и типом значений представлены в таблице 1.

Таблица 1 – Описание входных данных.

№	Имя столбца	Число заполненных ячеек	Тип переменных
1	IObytesRead	6304	float64
2	IObytesReadRate	6304	float64
3	IObytesWriteRate	6304	float64
4	IObytesWritten	6304	float64
5	IOcharRead	6304	float64
6	IOcharReadRate	6304	float64
7	IOcharWriteRate	6304	float64
8	IOcharWritten	6304	float64
9	actualcorecount	9626	float64
10	assignedpriority	10517	int64
11	avgpss	7718	float64
12	avgrss	7718	float64
13	avgswap	7718	float64
14	avgvmem	7718	float64
15	computingsite	10517	object
16	cpu_eff	10517	float64
17	cpuconsumptiontime	10517	int64
18	currentpriority	10517	int64
19	dbTime	6320	float64
20	endtime	10517	object

21	failedattempt	10517	int64
22	hs06	10517	int64
23	hs06sec	9313	float64
24	inputfilebytes	10517	int64
25	jobname	10517	object
26	jobstatus	10517	object
27	maxcpucount	10517	int64
28	maxdiskcount	10517	int64
29	maxpss	7718	float64
30	maxrss	7718	float64
31	maxvmem	7718	float64
32	minramcount	10517	int64
33	nevents	10517	int64
34	ninputdatafiles	10517	int64
35	noutputdatafiles	10517	int64
36	nucleus	10517	object
37	outputfilebytes	10517	int64
38	pandaid	10517	int64
39	queue_time	10517	int64
40	starttime	10517	object
41	timeExe	10517	int64

42	timeSetup	10517	int64
43	timeStageIn	10517	int64
44	timeStageOut	10517	int64
45	wall_time	10517	int64
46	workDirSize	8789	float64
	Всего по типу переменных: float64(20), int64(20), object(6)		

3.2 Подготовка данных

Первый этап перед любым анализом – это подготовка и очистка данных.

Во-первых, во многих строках содержатся пустые ячейки, в рядах столбцов частично отсутствуют значения. Было сделано предположение, что пустые значения зависят от определенных типов вычислительных узлов. Проверка данного предположения показала, что такой зависимости нет. Так как эти столбцы содержали больше 40% пропущенной информации, было принято решение их удалить

Так же, были сброшены столбцы, не несущие никакой значимой информации для анализа. Это имя задания и его статус.

```
X = X.drop(columns=['IOcharWritten','IObytesRead', 'IObytesReadRate', 'IObytesWriteRate','IObytesWritten','IOcharRead','IOcharReadRate','IOcharWriteRate', 'jobstatus', 'jobname' ])
```

Каждое задание имеет приоритет. Причем значение этого приоритета меняется с ходом выполнения задания. Начальный приоритет (assignedpriority) содержит значения [320, 901], низкий и высокий соответственно. Но чем дольше выполняется задача, тем выше становится ее приоритет. Конечный приоритет (currentpriority) уже содержит следующие значения [320, 322, 321, 5000, 323, 324, 325, 338, 336, 340, 331, 329, 332, 342, 337, 333, 326, 344, 328, 343, 339, 4990, 330, 334, 335, 341, 327, 901, 902, 903, 904, 900].

Значит, приоритет может повлиять на модель обучения. В качестве нового столбца была выбрана разница приоритетов.

```
X['PriorityDiff'] = X['currentpriority']-X['assignedpriority']
```

В таблице все еще остается много нулевых значений. Необходимо было проанализировать оставшиеся нулевые строки и то, насколько они важны для дальнейшего анализа

```
X.isnull().sum()
```

```
actualcorecount        891
avgpss                2799
avgrss                2799
avgswap               2799
avgvmem               2799
dbTime                4197
hs06sec               1204
maxpss                2799
maxrss                2799
maxvmem               2799
workDirSize           1728
```

После анализа таблицы, было принято решение, что некоторые столбцы не несут большой ценности.

```
X.drop(columns = ['workDirSize', 'maxvmem', 'maxrss', 'maxpss'
, 'avgpss','avgrss','dbTime','avgswap' ,'avgvmem', 'nucleus'])
```

Также в таблице содержатся данные о времени. Время начала (starttime) и время конца выполнения задания (endtime) заданы в виде 2018-06-05T10:07:10.0 Данные в таком формате не принесут много результата для анализа. Гораздо больше смысла будет нести информация о длительности выполнения всего задания (эта информация уже содержится в столбце ‘wall_time’), а так же о времени, прошедшего от начала дня, когда выполняется это задание, ведь тогда можно будет проследить какие-либо закономерности в течение дня во время обработки заданий. Так же на скорость выполнения может повлиять день недели. Было высказано предположение, что если задача поставлена в выходные, то она срочная и у нее присутствует большая вероятность сбоев. Также существует предположение, что члены физической группы, которые ставят задачи рано в

начале дня и поздно в конце имеют большую вероятность ошибки. В результате были созданы столбцы ‘startsec’, ‘endsec’, ‘startweekday’, ‘endweekday’

```
start = pd.to_datetime(X2.starttime)
startday = start.dt.normalize()
delta = start-startday
startsec = delta.dt.seconds
X2['startsec'] = startsec
startweekday = start.dt.dayofweek
X2['startweekday'] = startweekday
```

Теперь пустые значения имеют только следующие столбцы:

```
actualcorecount        891
hs06sec               1204
```

В данных столбцах пропуски были заполнены нулями и 1. Также, были переведены float в int.

```
X2['actualcorecount'] = X2['actualcorecount'].fillna(1)
X2['actualcorecount'] = X2.actualcorecount.astype('int64')
X2['hs06sec'] = X2['hs06sec'].fillna(0)
X2['hs06sec'] = X2.hs06sec.astype('int64')
X2['cpu_eff'] = (X2.cpu_eff * 1000000).astype('int64')
```

Столбец ‘computingsite’ содержит текстовую информацию о вычислительном узле. Его необходимо перевести в категориальный вид:

```
le = preprocessing.LabelEncoder()
X2.computingsite = le.fit_transform(X2.computingsite)
```

Категориальный тип переменных имеет достаточно бесполезный эффект для алгоритмов кластеризации, так как они основаны на расстоянии между точками, а в случае категориальных переменных оно равное между всеми категориями. Для решения данной проблемы категориальные столбцы переводятся в отдельные признаки, по одному для каждой категории. Это касается таких признаков как ‘computingsite’, ‘startweekday’ и ‘endweekday’.

```
onehot_encoder = OneHotEncoder(sparse=False, dtype = np.int64)
new_features = onehot_encoder.fit_transform(X2.computingsite.values.reshape(-1, 1))
tmp = pd.DataFrame(new_features, columns=['computingsite' + st
```

```
r(i) for i in range(new_features.shape[1])])
      X3 = pd.concat([X2, tmp], axis=1)
```

При построении модели можно столкнуться с наличием линейной или близкой к ней связи между всеми или некоторыми признаками. Это явление называется мультиколлинеарностью. В математической статистике этот термин используется для обозначения тесной корреляционной взаимосвязи между отбираемыми для анализа факторами, совместно воздействующими на общий результат [15].

Для исследуемой таблицы была выполнена проверка на мультиколлинеарность. После построения таблицы корреляции следующие признаки показали линейную зависимость друг от друга:

Таблица 2 – Признаки, коррелирующие более чем на 90%.

	Коррелирующие признаки	**Коэффициент корреляции, %**
1	currentpriority -> PriorityDiff	99,94
2	outputfilebytes -> PriorityDiff	99,18
3	currentpriority -> outputfilebytes	99,14
4	outputfilebytes -> computingsite3	98,70
5	PriorityDiff -> computingsite3	98,35
6	currentpriority -> computingsite3	98,31
7	timeExe -> wall_time	96,42
8	cpuconsumptiontime -> hs06sec	93,28
9	SWD4 -> EWD4	90,77
10	minramcount -> computingsite12	90,35

Можно увидеть, что такие признаки как разница приоритетов, текущий приоритет, размер выходных файлов и третий вычислительный узел имеют сильную связь. Также, время выполнения и общее время сильно коррелируют между собой. Немного меньший коэффициент корреляции имеют пятница как начало обработки задания и как конец.

В связи с этим было принято решение удалить такие столбцы как 'currentpriority', 'hs06sec', 'minramcount', 'wall_time', 'outputfilebytes', 'EWD4', 'PriorityDiff'. Теперь, когда таблица была избавлена от сильно коррелирующих признаков, она имеет размер 10517×51 и следующую структуру:

Таблица 3 – Структура данных после преобразований.

	Имя столбца	**Число заполненных ячеек**	**Тип переменных**
1	actualcorecount	10517	int64
2	assignedpriority	10517	int64
3	cpu_eff	10517	int64
4	cpuconsumptiontime	10517	int64
5	failedattempt	10517	int64
6	hs06	10517	int64
7	inputfilebytes	10517	int64
8	maxcpucount	10517	int64
9	maxdiskcount	10517	int64
10	nevents	10517	int64
11	ninputdatafiles	10517	int64
12	noutputdatafiles	10517	int64
13	pandaid	10517	int64
14	queue_time	10517	int64
15	timeExe	10517	int64
16	timeSetup	10517	int64
17	timeStageIn	10517	int64
18	timeStageOut	10517	int64
19	startsec	10517	int64
20	endsec	10517	int64
21	computingsite0	10517	int64
22	computingsite1	10517	int64
23	computingsite2	10517	int64
24	computingsite3	10517	int64
25	computingsite4	10517	int64
26	computingsite5	10517	int64
27	computingsite6	10517	int64
28	computingsite7	10517	int64
29	computingsite8	10517	int64
30	computingsite9	10517	int64
31	computingsite10	10517	int64

32	computingsite11	10517	int64
33	computingsite12	10517	int64
34	computingsite13	10517	int64
35	computingsite14	10517	int64
36	computingsite15	10517	int64
37	computingsite16	10517	int64
38	computingsite17	10517	int64
39	SWD0	10517	int64
40	SWD1	10517	int64
41	SWD2	10517	int64
42	SWD3	10517	int64
43	SWD4	10517	int64
44	SWD5	10517	int64
45	SWD6	10517	int64
46	EWD0	10517	int64
47	EWD1	10517	int64
48	EWD2	10517	int64
49	EWD3	10517	int64
50	EWD5	10517	int64
51	EWD6	10517	int64

После этого была проведена нормализация и шкалирование всей таблицы.

```
StandardScaler().fit_transform(X3)
```

3.3 Снижение размерности признакового пространства

Для точной визуализации необходимо уменьшить размерность пространства данных с 51 столбца до двухмерного пространства, доступного для визуализации точек и кластеров.

Первым способом для снижения признакового пространства был выбран **метод главных компонент**.

В результате анализы было выделено от 2 до 5 главных компонент. На итоговой визуализации это значительно не отразилось. Ниже приведен график распределения точек для 2 главных компонент.

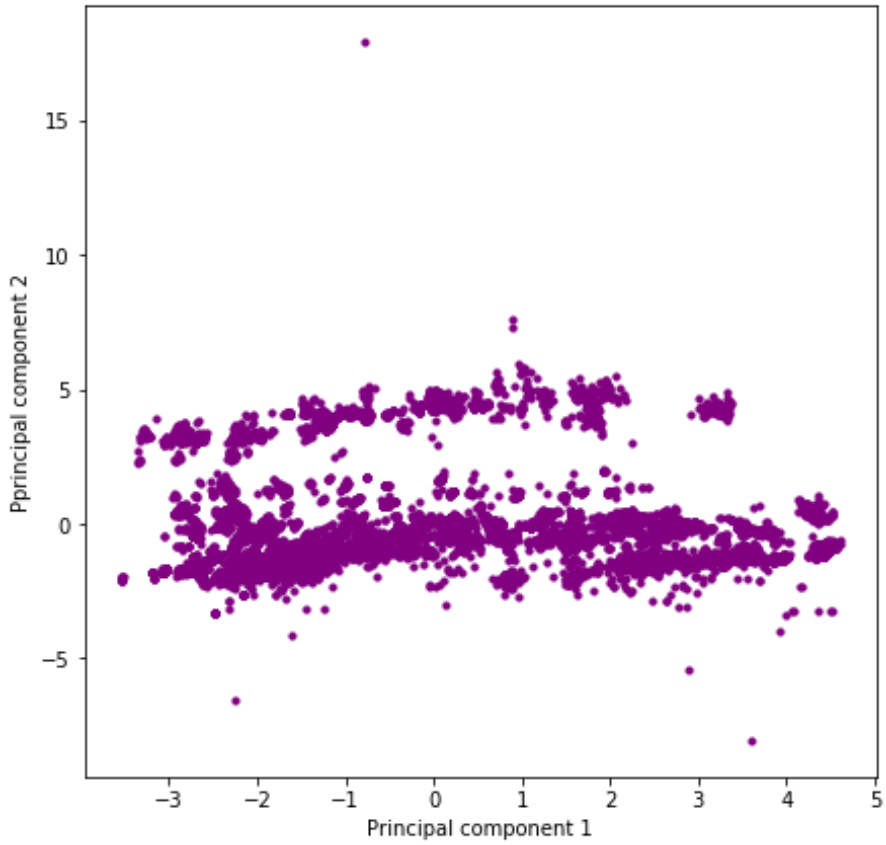

Рисунок 7 – Визуализация точек после метода PCA

На графике видно явное разделение на несколько кластеров. Особенно это видно на вертикальной оси.

Переменные, которые влияют на компоненты (вес > 0,3 по модулю):

```
Component  1
    maxcpucount 0.30798042478224846
    maxdiskcount 0.3621007531826703
    SWD4 -0.3559489863080462
    EWD6 0.30719739289749076
Component  2
    actualcorecount 0.47050088169920634
    cpuconsumptiontime 0.344149281857442
    timeSetup 0.4628014555386573
    computingsite12 0.45472783823315654
```

Особенно в данном случае вызывает интерес вторая компонента. Наибольший вес для нее имеют число ядер, задействованных в вычислении, время процессора, время настройки и двенадцатый вычислительный узел.

Второй способ снижения размерности – это T-SNE. Если сравнивать рис. 8 и 9, можно явно увидеть как настройка параметра perplexity (перплексивность - степень неопределенности вероятностной модели) влияет на форму графика. Для дальнейшего исследования будет взят график на рис. 9. На нем хорошо видно, что данный метод выделяет достаточно много мелких кластеров.

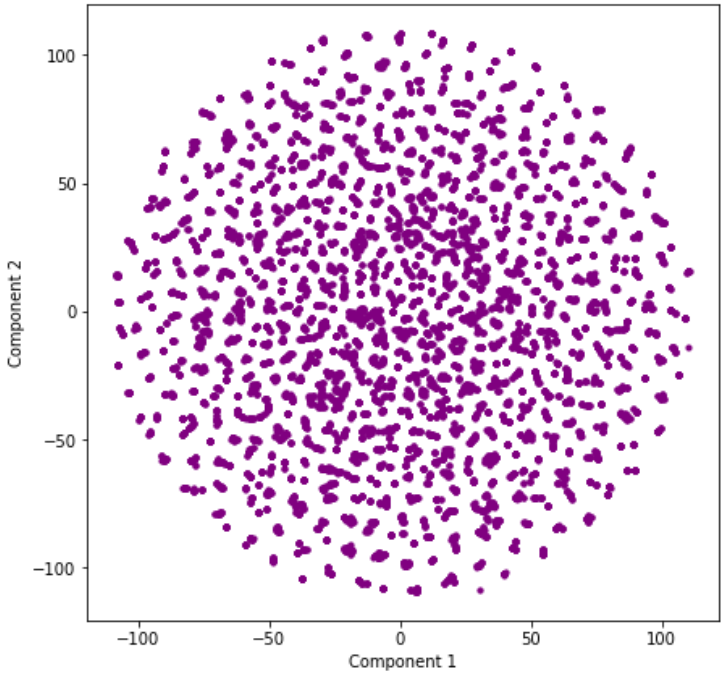

Рисунок 8 – Визуализация данных методом T-SNE (perplexity = 10)

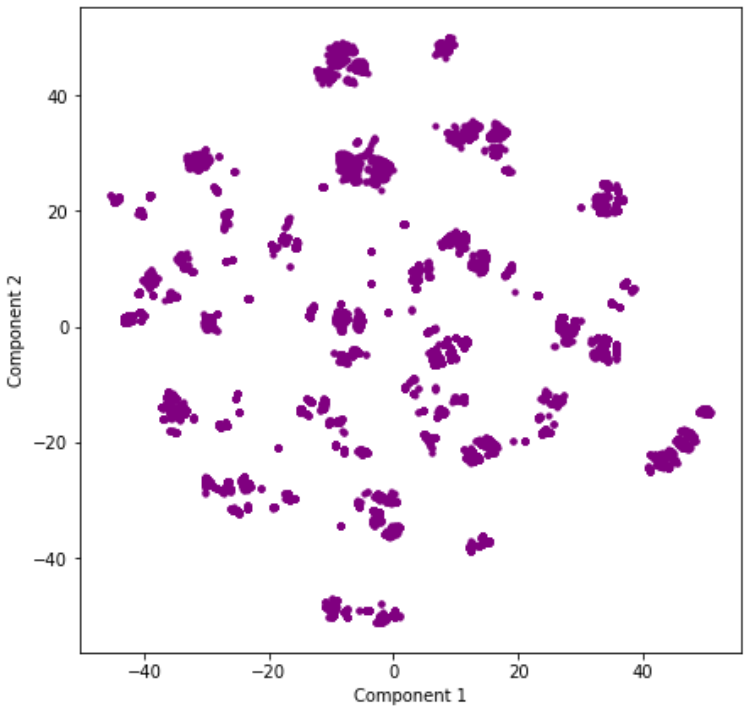

Рисунок 8 – Визуализация данных методом T-SNE (perplexity = 200)

ГЛАВА 3. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

3.1 K-means

К-средних – самый быстродействующий и часто самый эффективный за счёт своей простоты алгоритм. Это итеративный алгоритм, который ищет центры кластеров и точки, которые находятся ближе всего к ним. Алгоритм k-means обычно находит набор стабильных кластеров за несколько десятков итераций.

На графиках можно заметить, что визуализация данных очень полезна. Мы видим, что деление на кластеры в случае PCA очень условно. На графиках (9) видно, что данный алгоритм не лучшим образом подходит для геометрии именно этих данных.

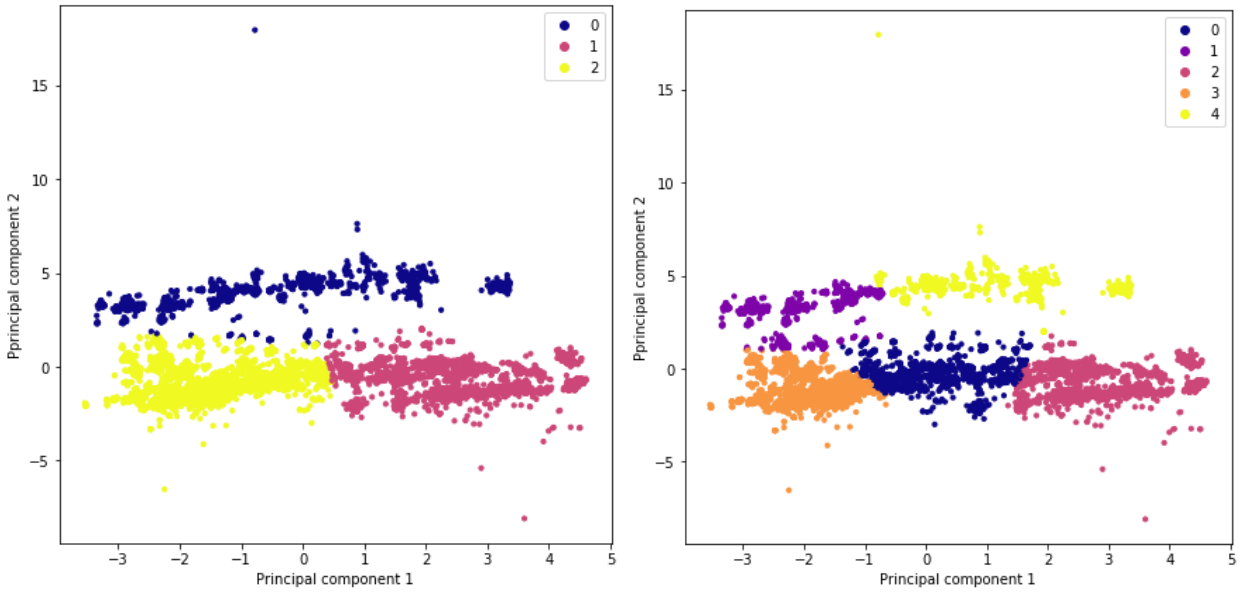

Рисунок 9 –Кластеризация методом K-means для PCA на 3 (а) и 5 кластеров (б)

Для T-SNE в данном методе необходимо указывать большое количество кластеров, иначе разделение будет формальным. На рис. 10 видно различие между разделением на 3 и на 20 кластеров.

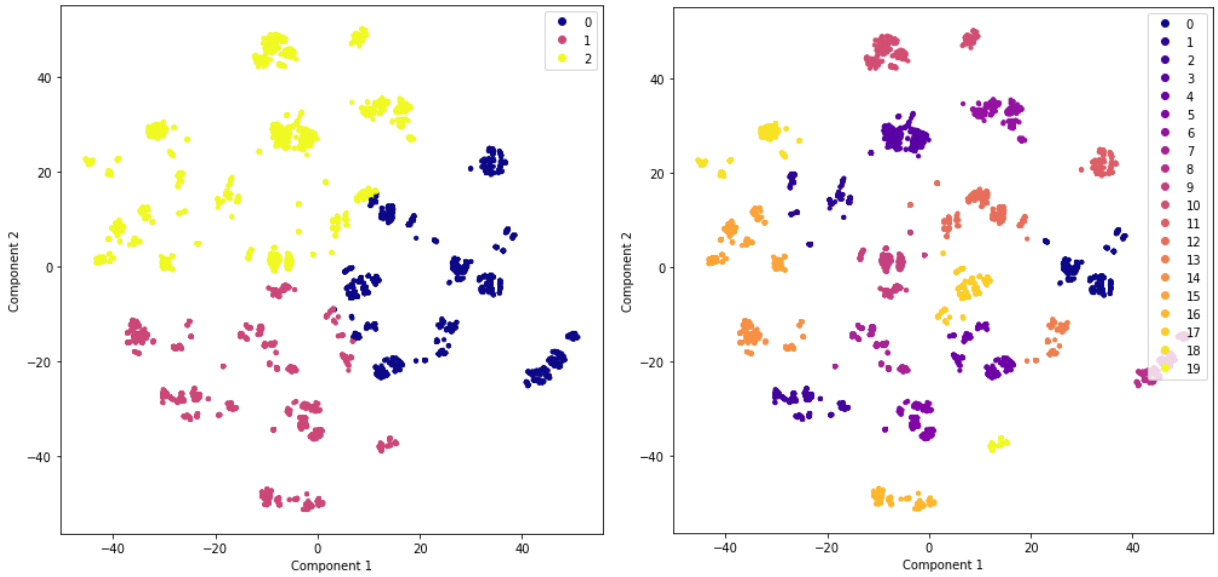

Рисунок 10 – Кластеризация методом K-means для T-SNE на 3 (а) и на 20 (б) кластеров.

3.2 Иерархическая кластеризация

В данной кластеризации строится иерархическое дерево, которое объединяет точки по схожим признакам, в итоге образуя кластеры, которое представлено на рисунке 11:

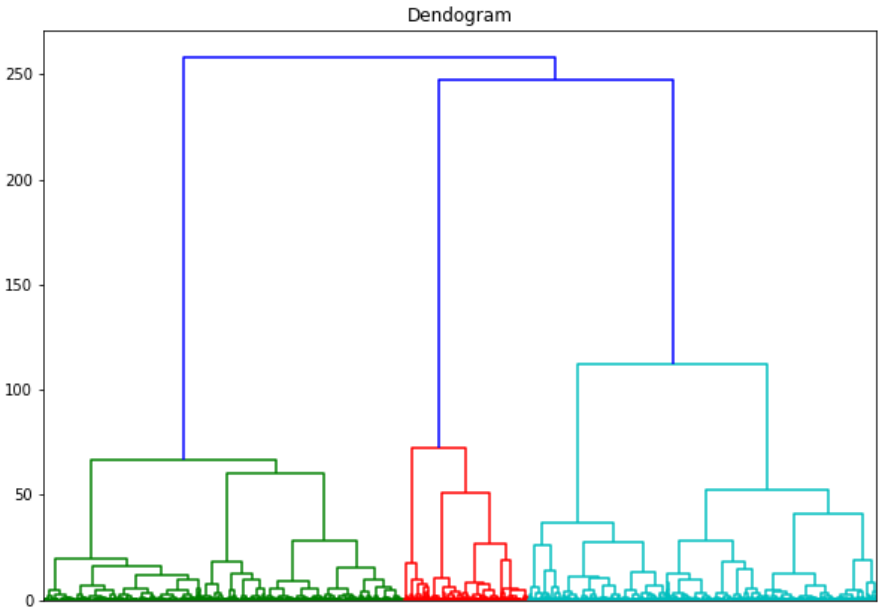

Рисунок 11 – Дендрограмма для метода PCA

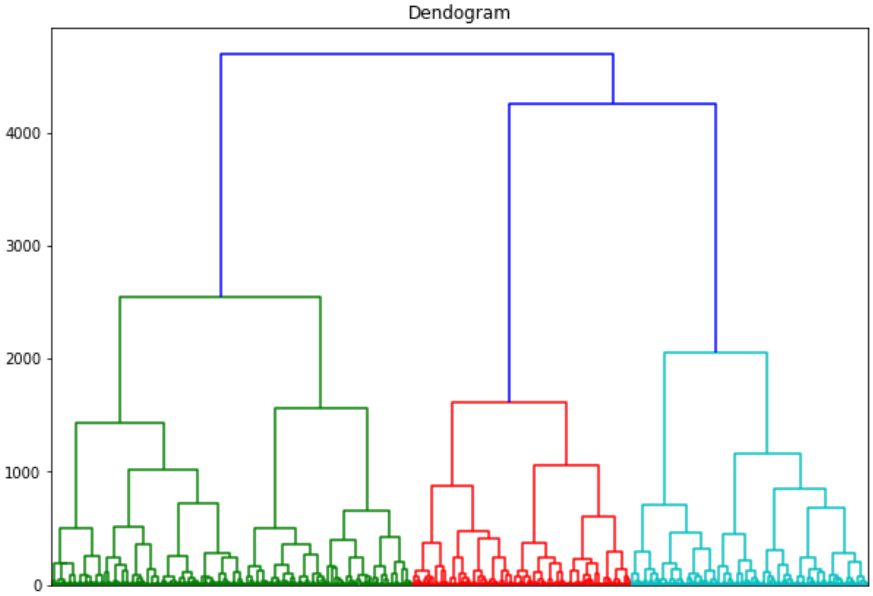

Рисунок 12 - Дендрограмма для метода T-SNE

Чтобы визуализировать данное разделение необходимо решить какое количество кластеров оптимально для исследования и сделать «срез» данной дендограммы относительно определенного количества кластеров. На рисунке 13 представлены графики для 3-7 кластеров.

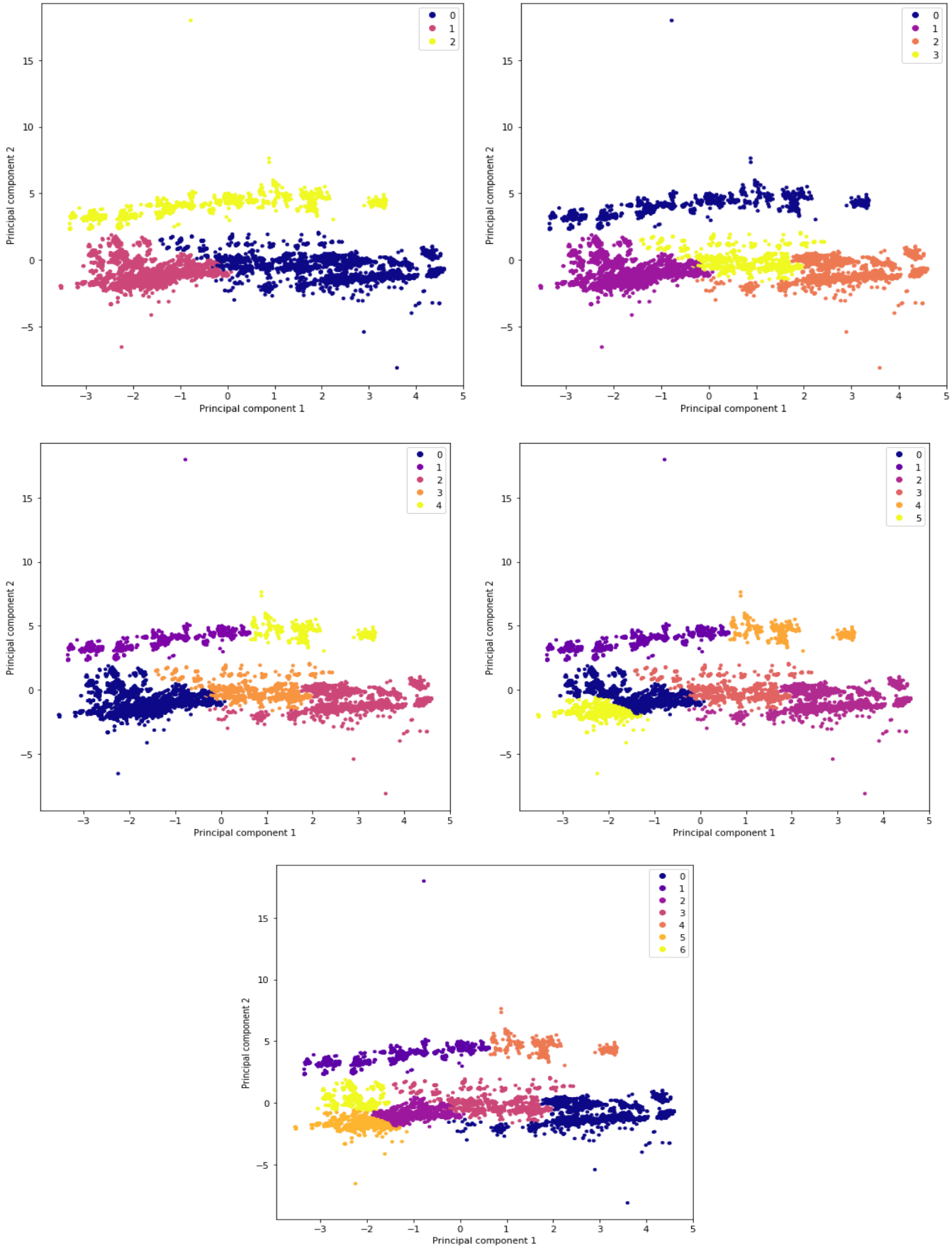

Рисунок 13 – Кластеризация иерархическим методом для PCA на 3(а), 4(б), 5(в), 6(г), 7(д) кластеров.

Для T-SNE результаты представлены на рис. 14.

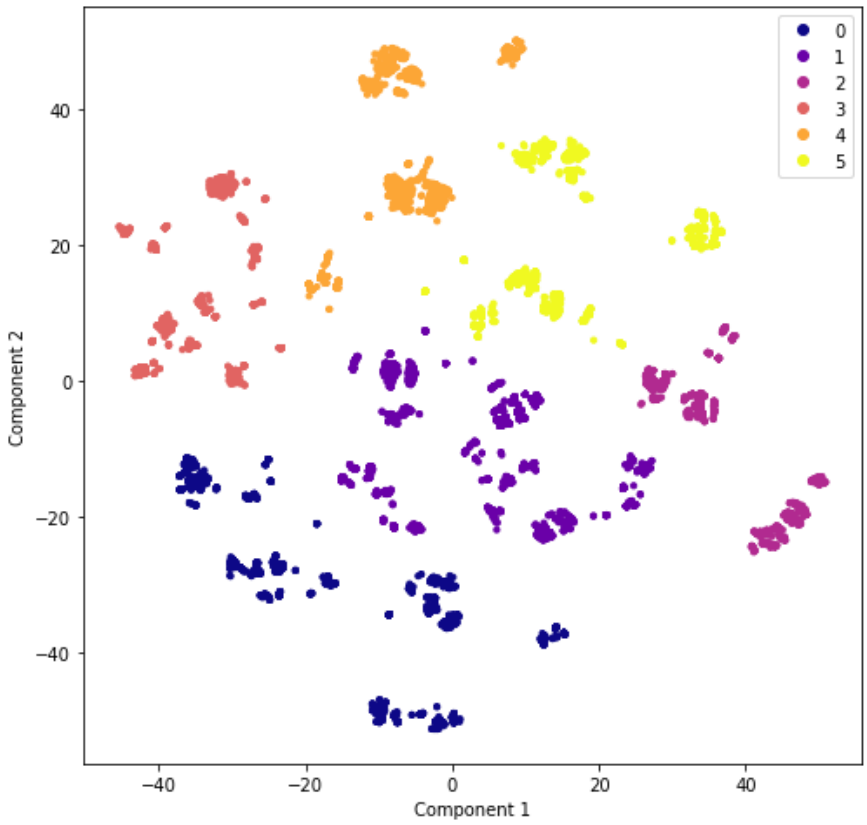

Рисунок 14 – Кластеризация иерархическим методом для T-SNE на 6 кластеров

Для PCA данный алгоритм показывает более хорошие результаты по сравнению с k-means. Для T-SNE были выделены кластеры достаточно логично, но все равно видно, что реальная численность кластеров гораздо больше, что бесполезно для исследования, так как нас интересуют около 10 кластеров.

3.3 DBSCAN

DBSCAN – плотностной алгоритм пространственной кластеризации с присутствием шума – популярный алгоритм кластеризации, используемый в анализе данных в качестве одной из замен метода k-средних.

Если текущий объект – окруженная точка, то все объекты, достижимые по плотности от текущего объекта, соединяем в новый кластер. В противном случае, если объект не является окруженной точкой и не достижим по плотности ни от какого объекта, то текущий объект – выброс.

Для PCA метода оптимальные параметры:

– eps=0.5 – плотность;

– min_samples = 10 – минимальное количество точек для принадлежности к кластеру.

Фиолетовым на графике изображены выбросы.

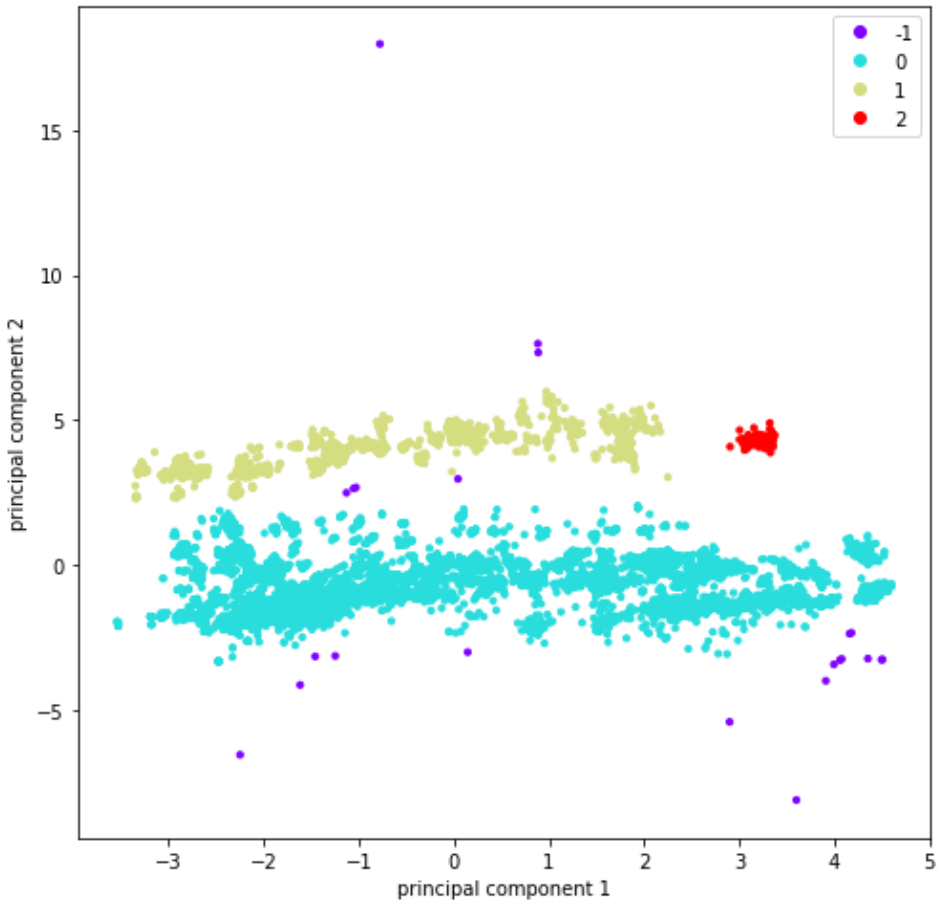

Рисунок 15 – Визуализация DBSCAN для PCA

В случае с T-SNE параметры были следующие:

– eps=0.25 – плотность;

– min_samples = 20 – минимальное количество точек для принадлежности к кластеру.

В результате алгоритм нашел 55 кластеров и выбросы.

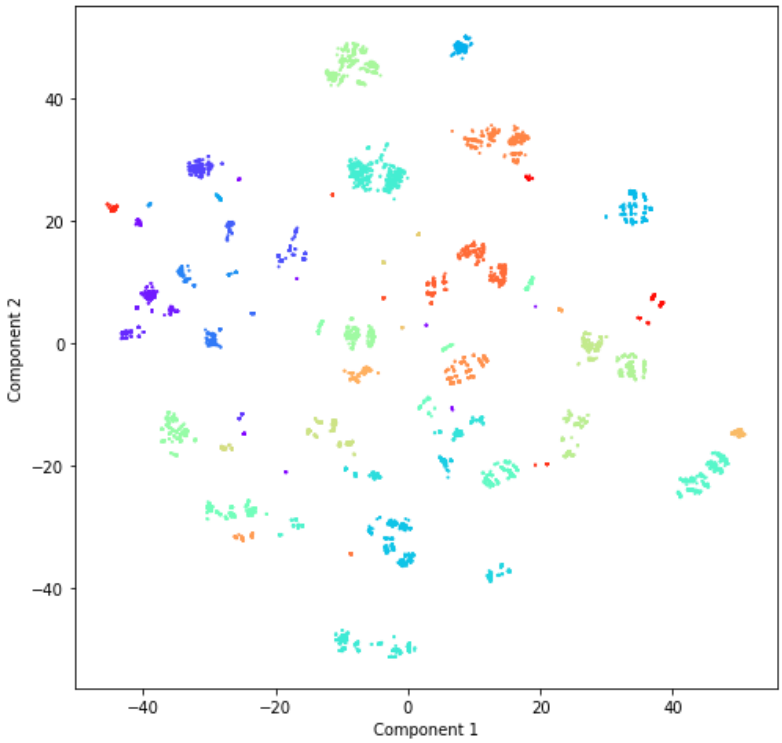

Рисунок 16 – Визуализация DBSCAN для T-SNE (55 кластеров)

Данный метод показывает наиболее оптимальный результат для PCA, он более всего совпадает с визуальным делением и хорошо определяет выбросы.

Но и для T-SNE он работает достаточно точно, так как выделяет очень большое число кластеров, но это число, 55, слишком велико для анализа.

ЗАКЛЮЧЕНИЕ

В ходе проведенного исследования были получены следующие результаты:

1. Проведен анализ подходов к кластеризации без учителя существующих методов снижения размерности, очистки данных;
2. Проведена подготовка данных, в том числе их очистка, обогащение, заполнение пустых значений, снижение признакового пространства.
3. Было выяснено, что сильное влияние на результаты кластеризации оказывают такие признаки как число ядер, задействованных в вычислении, время процессора, время настройки и двенадцатый вычислительный узел. Так же, время выполнения заданий напрямую зависят от третьего вычислительного узла. Помимо этого большой вес на результат оказывают пятница, как день недели начала обработки, и воскресенье как день недели конца.
4. Метод снижения размерности PCA хорошо визуализирует данные, разделяя их на небольшие кластеры, при этом не требуя точной настройки.
5. T-SNE показывает совершенно разную геометрию при разных параметрах настройки. В данном случае удалось достичь оптимального результата с перплексивностью 200.
6. Среди методов кластерного анализа k-means наиболее «прямолинейный». Из-за того, что он основывается на расстоянии между точками, его результаты часто оказываются неподходящими под геометрию определенных, например, вытянутых, данных, как это было представлено на графиках для PCA. Для T-SNE форма графиков больше подходила для k-means, но при малом количестве кластеров разделение слишком условно, а при большом количестве их число слишком велико для анализа, хотя разделение достаточно четкое.
7. Иерархический метод в случае PCA выделяет кластеры более точно, чем k-means, однако для данной формы он так же не является оптимальным. Для T-SNE, результаты примерно аналогичны с k-means, однако в данном случае они выглядят более убедительно на малом количестве кластеров.

8. Для PCA наиболее точное разделение на кластеры показал метод DBSCAN после настройки параметров, при этом он нашел и шумы, которые явно выделяются на графиках. Для T-SNE данный метод находил от 55 до 1145 кластеров. Это обусловлено тем, что изначально на данном наборе присутствует слишком много тесных данных.

9. Наиболее точные результаты кластеризации на данном наборе данных показывают метод кластеризации DBSCAN совместно с методом снижения размерности PCA.

В планах использовать результат данной связки, чтобы на основании исходных данные ввести дополнительный столбец с результатами кластеризации и принадлежностью задачи к определенному кластеру и обучать следующие данные уже на модели с учителем, планируя, что это может помочь сделать предсказания длительности времени выполнения заданий более точными.

Если результаты анализа будут коррелировать с длительностью выполнения задач, и данный подход повысит качество предсказаний, то данный метод будет внедрен в используемую систему предсказания длительности. Если нет, то будут рассмотрены и исследованы другие подходы. В любом из вариантов существует возможность дальше пробовать другие связки методов, повышая качество и точность предсказаний.

ГЛАВА 4. ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

Работа посвящена исследованию и реализации алгоритма предварительной кластеризации задач в больших вычислительных сетях. В ней используются конкретные данные, полученные с системы обработки задач экспериментов Большого Адронного Коллайдера в ЦЕРНе. Потенциальными потребителями являются физики ЦЕРНа, которые ставят задачи на обработку в распределенную вычислительную систему.

Актуальность данной работы состоит в том, что в данный момент существует системы предсказания длительности выполнения задач в этих вычислительных системах, но данная работа упростит предсказание длительности за счет предварительной кластеризации и отнесению задачи к определенному типу.

При разработке научно-технического проекта одним из важных этапов является его технико-экономическое обоснование. При помощи него можно выделить преимущества и недостатки разработки данного программного продукта в разрезе экономической эффективности, социальной значимости и других аспектов.

Целью выполнения данного раздела является анализ и выявление потенциальной выгоды научного исследования и реализации алгоритма предварительной кластеризации задач в больших вычислительных сетях.

Задачи, которые следуют из данной цели:

1. Провести анализ перспективности исследования;
2. Провести оценку готовности проекта к коммерциализации;
3. Составить цели проекта, определить ожидаемые результаты, рабочую группу;
4. Спланировать работы, распределить задачи между участниками проекта, построить график Ганта;
5. Сформировать бюджет затрат;
6. Провести анализ рисков проекта.

4.1.1Технология QuaD

Для оценки экономической эффективности необходимо произвести анализ конкурентных технических решений. Но так как данная разработка является уникальной и создается для конкретного пользователя, то в таблице 1 приведен анализ критериев конкурентоспособности. Для этого необходимо произвести процедуру QuaD, используя оценочную карту.

Таблица 4 – Quad анализ исследования

Критерии оценки	**Вес критерия**	**Баллы**	**Максимальный балл**	**Относительное значение (3/4)**	**Средневзвешенное значение (5x2)**
1	2	3	4	5	
Показатели оценки качества разработки					
1.Читаемость кода для команды	0,05	60	100	0,6	0,03
2.Соответствие требованиям	0,1	80	100	0,8	0,08
3.Надежность	0,1	60	100	0,6	0,06
4.Точность кластеризации	0,2	85	100	0,85	0,17
5.Простота встраивания в большую систему	0,05	50	100	0,5	0,025
6.Уменьшение затраченного времени сотрудниками	0,05	40	100	0,4	0,02
7. Кол-во ресурсов памяти	0,05	30	100	0,3	0,015
8. Защита данных в процессе	0,05	40	100	0,4	0,02
Показатели оценки коммерческого потенциала разработки					
9. Конкурентоспособность продукта	0,05	70	100	0,7	0,035
10. Срок выполнения научной разработки	0,05	60	100	0,6	0,03
11. Цена	0,1	50	100	0,5	0,05
12.Финансирование научной разработки	0,1	85	100	0,85	0,085
13. Срок внедрения метода в систему.	0,05	50	100	0,5	0,025
Итого	1				0,645

Анализируя результаты, полученные по технологии QuaD, можно сделать вывод, что качество и перспективность разработки информационной системы

выше среднего. Об этом говорит средневзвешенное значение показателя качества и перспективности, равное 0,645. Для того, чтобы произвести повышение качества и перспективности необходимо выполнять ряд действий. Например, ускорить процесс разработки и внедрения информационной системы.

5.1.2 Оценка готовности проекта к коммерциализации

На любой стадии жизненного цикла проекта необходимо оценить степень ее готовности к коммерциализации и выяснить уровень знаний для ее проведения. В таблице 5 заполнена форма с показателями о степени проработанности проекта с позиции коммерциализации и компетенциям разработчика научного проекта.

Таблица 5 - Оценка степени готовности проекта к коммерциализации

№ п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся знаний у разработчика
1	Определен имеющийся научно-технический задел	5	5
2	Определены перспективы направления коммерциализации научно-технического задела	5	4
3	Определены отрасли и технологии (товары, услуги) для предложения на рынке	5	4
4	Определена товарная форма задела для представления на рынок	4	4
5	Определены авторы и осуществлена охрана их прав	4	3
6	Проведена оценка стоимости интеллектуальной собственности	4	4
7	Проведены маркетинговые исследования рынков сбыта	3	4
8	Разработан бизнес-план коммерциализации научной разработки	3	2
9	Определены пути продвижения научной разработки на рынок	4	3
10	Разработана стратегия (форма) реализации научной разработки	4	3
11	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	3	3
12	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	2	3

13	Проработаны вопросы финансирования коммерциализации научной разработки	2	3
14	Имеется команда для коммерциализации научной разработки	2	2
15	Проработан механизм реализации научного проекта	3	2
	ИТОГО БАЛЛОВ	**53**	**49**

По итогам анализа можно сделать вывод, что перспективность проекта для коммерциализации выше среднего. Для улучшения данных показателей необходимо улучшать компетенции разработчика научного проекта. Так же, для улучшения качества вопросов коммерциализации необходимо привлечь специалистов из сектора бизнес-планирования и внедрения продуктов на рынок.

Проанализировав имеющиеся методы коммерциализации, наиболее подходящим выбран **инжиниринг**, так как данная научная разработка является частью более полной системы, реализуемой специально для данного предприятия под конкретные требования конкретных систем. И наилучшим решением будет полная поддержка системы в рамках совместного проекта.

1.2 Инициация проекта

Группа процессов инициации состоит из процессов, которые выполняются для определения нового проекта или новой фазы существующего. В рамках процессов инициации определяются изначальные цели и содержание и фиксируются изначальные финансовые ресурсы. Определяются внутренние и внешние заинтересованные стороны проекта, которые будут взаимодействовать и влиять на общий результат научного проекта.

5.2.1 Цели и результаты проекта

Под заинтересованными сторонами проекта понимаются лица или организации, которые активно участвуют в проекте или интересы которых могут быть затронуты как положительно, так и отрицательно в ходе исполнения или в

результате завершения проекта. Информация по заинтересованным сторонам проекта представлена в таблице 6

Таблица 6 – Заинтересованные стороны проекта

Заинтересованные стороны проекта	**Ожидания заинтересованных сторон**
Ученые-физики, которые обрабатывают задачи во Всемирной Распределенной Сети Коллайдера (WLCG)	Реализованный метод кластеризации, относящий задачу к определенному типу.
Разработчики программного обеспечения для WLCG	Предварительная кластеризация задач на группы для более точного дальнейшего предсказания длительности выполнения задач
Высшее Учебное Заведение	Совместная работа над проектом с ЦЕРНом повышает престиж университета и способствует дальнейшему сотрудничеству и партнерству

Ниже представлена информация о целях исследования и задачах, вытекающих из этих целей.

Таблица 7 – Цели и результат проекта

Цели проекта:	Поиск закономерностей, влияющих на предсказание длительности выполнения задач в цепочках с помощью предварительной кластеризации.
Задачи:	1. Исследование методов кластеризации многоразмерных данных без учителя. 2. Реализация алгоритмов снижения размерности и кластеризации 3. Сравнение и выбор наиболее оптимального метода.
Ожидаемые результаты:	1. Проведен анализ и сравнение методов кластеризации. 2. Проведена предварительная подготовка данных. 3. Реализованы методы снижения размерности и алгоритмы кластеризации. 4. Выбран наиболее точный метод в связке снижение размерности + кластеризация. 5. Найдены признаки, влияющие на результат кластеризации в большей степени. 6. Результаты кластеризации внедрены в таблицу для дальнейшего обучения с учителем алгоритма предсказывания длительностью.

Таблица 8 – Рабочая группа проекта

№	**ФИО, место работы, должность**	**Роль в проекте**	**Функции**
1	Губин Евгений Иванович, ТПУ,	Научный	1. Составление научных

	кандидат физико-математических наук	руководитель	целей и задач 2. Проверка документации
2	Губин Максим Юрьевич, ТПУ, ведущий программист.	Технический консультант	1. Предоставление данных 2. Проверка алгоритма
3	Шкабара Анастасия Игоревна, ТПУ, студент	Инженер	1. Проектирование 2. Реализация 3. Создание документации

Из вышеприведенной таблицы следует, что обязанности среди трех участников проекта распределены следующим образом: Инженер выполняет само исследование, технический консультант проверяет работу на промежуточных этапах, вносит коррективы в процессе. Научный руководитель ответственен за глобальные задачи и итоговые проверки результатов и документации.

5.2.2 Организация и планирование работы

От количества исполнителей напрямую зависит продолжительность проекта и его затраты. Так как число исполнителей равно трем, линейный график работ является наиболее удобным и компактным способом представления данных планирования. Данные по перечню работ и степени участия представлены в таблице 9. Исполнителей в данном проекте – Инженер (И), консультант (К) и научный руководитель (НР).

Таблица 9 – Перечень работ и исполнители

	Этапы работы	**Исполнители**
1	Постановка задачи, определение целей, получение исходных данных	НР, К
2	Выявлений требований к программе	К, И
3	Календарное планирование	НР, К, И
4	Обзор литературы и существующих решений	И
5	Подготовка данных	И
6	Реализация методов снижения размерности	И
7	Реализация методов кластеризации	И
8	Анализ результатов исследования	К, И
9	Расчет показателей ресурсоэффективности	И
10	Оценка показателей социальной ответственности	И
11	Оформление пояснительной записки	И
12	Проверка работы	НР, К, И

Для инициации любого проекта, и данного в частности, важно на начальном этапе проработать цели исследования, задачи и результаты, которые будут служить показателем эффективной работы. Так же, следует определять роли участников в проекте, описать распределение обязанностей.

5.3 Планирование управления научно-техническим исследованием

5.3.1 План исследования

Для составления плана был сделан перечень этапов работ в исследовании, затем были определены участники каждой работы. Все эти данные сведены в таблицу 10.

Таблица 10 – Перечень этапов работ и распределение исполнителей

Код работы (из ИСР)	Название	Длительность, дни	Дата начала работ	Дата окончания работ	Состав участников (ФИО ответственных исполнителей)
1	Постановка задачи, определение целей, получение исходных данных	8	01.03.2019	09.03.2019	Губин Евгений Иванович, Губин Максим Юрьевич
2	Выявлений требований к программе	3	11.03.2019	13.03.2019	Шкабара Анастасия Игоревна, Губин Максим Юрьевич
3	Календарное планирование	2	14.03.2019	15.03.2019	Губин Евгений Иванович, Шкабара Анастасия Игоревна, Губин Максим Юрьевич
4	Обзор литературы и существующих решений	9	16.03.2019	26.03.2019	Шкабара Анастасия Игоревна
5	Подготовка данных	9	27.03.2019	05.04.2019	Шкабара Анастасия Игоревна
6	Реализация методов снижения размерности	6	08.04.2019	13.04.2019	Шкабара Анастасия Игоревна
7	Реализация методов кластеризации	6	15.04.2019	20.04.2019	Шкабара Анастасия Игоревна
8	Анализ результатов исследования	12	22.04.2019	04.05.2019	Шкабара Анастасия Игоревна, Губин Максим Юрьевич
9	Расчет показателей ресурсоэффективности	3	06.05.2019	08.05.2019	Шкабара Анастасия Игоревна

10	Оценка показателей социальной ответственности	4	08.05.2019	13.05.2019	Шкабара Анастасия Игоревна
11	Оформление пояснительной записки	11	14.05.2019	25.05.2019	Шкабара Анастасия Игоревна
12	Подведение итогов	6	27.05.2019	31.05.2019	Губин Евгений Иванович, Шкабара Анастасия Игоревна, Губин Максим Юрьевич

Планирование работ – один из важнейших эиапов для любого проекта. Оно позволяет распределить время и оенить объем выполненной и оставшейся работы. Согласно данному плану, исследование планируется выполнить за 3 месяца, основная рабочая нагрузка приходится на инженера – основного исполнителя.

5.3.2 Определение трудоемкости выполнения работ

Вычислим трудоёмкость выполнения научного исследования. Для определения ожидаемого (среднего) значения трудоемкости $t_{ожi}$ используем следующую формулу:

$$t_{\text{ожі}} = \frac{3t_{\min i} + 2t_{\max i}}{5},$$

где $t_{ожi}$ – ожидаемая трудоемкость выполнения *i*-ой работы чел.-дн.;

$t_{\min i}$ – минимально возможная трудоемкость выполнения заданной *i*-ой работы (оптимистическая оценка: в предположении наиболее благоприятного стечения обстоятельств), чел.-дн.;

$t_{\max i}$ – максимально возможная трудоемкость выполнения заданной *i*-ой работы (пессимистическая оценка: в предположении наиболее неблагоприятного стечения обстоятельств), чел.-дн.

Исходя из ожидаемой трудоемкости работ, определим продолжительность каждой работы в рабочих днях $T_{\text{р}}$, учитывая параллельность выполнения работ несколькими исполнителями.

$$T_{\text{р}_i} = \frac{t_{\text{ожі}}}{\text{Ч}_i},$$

где T_{pi} – продолжительность одной работы, раб. дн.;

$t_{ожi}$ – ожидаемая трудоемкость выполнения одной работы, чел.-дн.

$Ч_i$ – численность исполнителей, выполняющих одновременно одну и ту же работу на данном этапе, чел.

5.3.3 Разработка графика проведения НТИ

Для построения графика проведения научного исследования рассчитаем предварительно некоторые необходимые параметры.

Переведём длительность каждого из этапов работ из рабочих дней в календарные дни. Для этого необходимо воспользоваться следующей формулой:

$$T_{кi} = T_{рi} \cdot k_{кал},$$

где $T_{кi}$– продолжительность выполнения *i*-й работы в календарных днях;

$T_{рi}$ – продолжительность выполнения *i*-й работы в рабочих днях;

$k_{кал}$– коэффициент календарности.

Коэффициент календарности определяется по следующей формуле:

$$k_{кал} = \frac{T_{кал}}{T_{кал} - T_{вых} - T_{пр}},$$

где $T_{кал}$ – количество календарных дней в году;

$T_{вых}$ – количество выходных дней в году;

$T_{пр}$ – количество праздничных дней в году.

Согласно производственному календарю (для 6-дневной рабочей недели) в 2019 году 365 календарных дней, 299 рабочих дней, 66 выходных/праздничных дней.

$$kкал = \frac{Ткал}{Ткал - Твых - Тпр} = \frac{365}{365 - 66} = 1{,}22$$

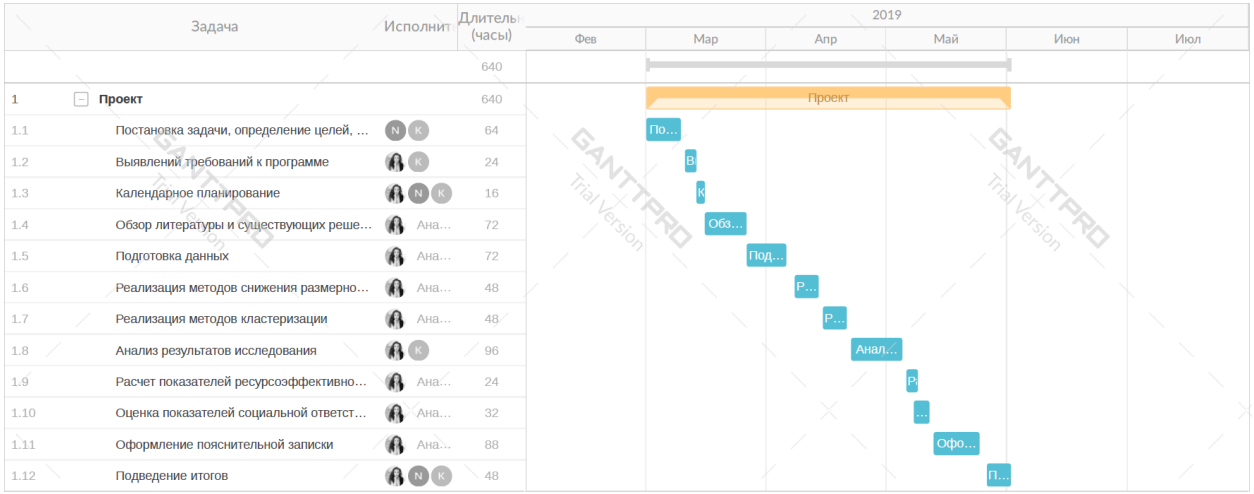

Рисунок 17– Диаграмма Ганта

Для визуализации сроков проведения НИОКР воспользуемся диаграммой Ганта (рис. 17). Диаграмма Ганта – горизонтальный ленточный график, на котором работы по теме представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

Далее вычислим количество календарных дней согласно общему графику:

$$T_к = 80 * 1{,}22 = 97{,}6 \text{ кал. дней}$$

Из данных вычислений следует, что, согласно плану, исследование займет 80 рабочих дней или 97,6 календарных дней по 6-ти дневной рабочей неделе.

5.4 Бюджет НТИ

При планировании бюджета НТИ должно быть обеспечено полное и достоверное отражение всех видов расходов, связанных с его выполнением. В процессе формирования бюджета НТИ используется следующая группировка затрат по статьям:

1. материальные затраты;
2. амортизации;
3. основная заработная плата исполнителей;
4. дополнительная заработная плата исполнителей;
5. отчисления во внебюджетные фонды (страховые отчисления);
6. накладные расходы.

Для разработки данного продукта отдельные материалы не закупались, кроме канцелярских расходов. Материальные затраты были только на канцелярию в размере 1000 руб.

5.4.1 Амортизационные отчисления

Для создания НТИ использовался ПК. В таком случае необходимо рассчитать амортизацию основных средств линейным способом.

Она рассчитывалась следующим образом: первоначальная стоимость ПК 70000 рублей; срок полезного использования для машин офисных код 330.28.23.23 составляет 2-3 года, в данном случае 3 года; Время, планируемое использовать ПК для написания ВКР - 3 месяцев. Тогда:

Норма амортизации:

$$\text{Ан} = \frac{1}{n} * 100\% = \frac{1}{3} \times 100\% = 33{,}33\%$$

Годовые амортизационные отчисления:

$$\text{Аг} = 70000 \times 0{,}33 = 23100 \text{ рублей}$$

Ежемесячные амортизационные отчисления:

$$\text{Ам} = \frac{13200}{12} = 1925 \text{ рублей}$$

Итоговая сумма амортизации основных средств:

$$\text{А} = 1925 \times 3 = 5775 \text{ рублей}$$

5.4.2 Основная заработная плата исполнителей

Затраты на заработную плату:

$$\text{Зп} = \text{Зосн} + \text{Здоп},$$

Где *Зосн* – основная заработная плата, руб.

Здоп – дополнительная заработная плата, руб.

Заработная плата основная:

$$\text{Зосн} = \text{Здн} \times \text{Тр}$$

Где *Здн* – среднедневная заработная плата, руб.

Кпр – премиальный коэффициент (0,3);

Кд – коэффициент доплат и надбавок (0,2-0,5);

Кр – районный коэффициент (для Томска 1,3);

Тр – продолжительность работ, выполняемых работником, раб. Дни

Среднедневная заработная плата:

$$Здн = \frac{Зм \times М}{F_д},$$

Где *Зм* –оклад работника за месяц, руб.

М – количество месяцев работы без отпуска в течение года:

при отпуске в 48 раб. дней М=10,4 месяца, 6-дневная неделя;

Fд – действительный годовой фонд рабочего времени персонала, раб. дн.

Таблица 11 – Баланс рабочего времени (для 6-дневной недели)

Показатели рабочего времени	**Дни**
Календарные дни	365
Нерабочие дни (праздники/выходные)	66
Потери рабочего времени (отпуск/невыходы по болезни)	56
Действительный годовой фонд рабочего времени	243

Для инженера:

$$Здн = \frac{Зо \times (1 + Кпр + Кд) \times Кр \times М}{Fд} == \frac{21760 \times (1 + 0{,}3 + 0{,}2) \times 1{,}3 \times 10{,}4}{243}$$

$$= 1816{,}49 \text{ руб},$$

Для консультанта:

$$Здн = \frac{Зо \times (1 + Кпр + Кд) \times Кр \times М}{Fд} == \frac{26624 \times (1 + 0{,}3 + 0{,}2) \times 1{,}3 \times 10{,}4}{243}$$

$$= 2221{,}95 \text{ руб.}$$

Для руководителя:

$$Здн = \frac{Зо \times (1 + Кпр + Кд) \times Кр \times М}{Fд} == \frac{33664 \times (1 + 0{,}3 + 0{,}2) \times 1{,}3 \times 10{,}4}{243}$$

$$= 2809{,}49 \text{ руб.}$$

Таблица 12 – Расчет основной заработной платы

Исполнители	**Здн, руб.**	**Кпр**	**Кд**	**Кр**	**Тр**	**Зосн**
Научный руководитель	2809,49	0,3	0,2	1,3	5	14047,45
Консультант	2221,95	0,3	0,2	1,3	8	17775,63

Инженер	1816,02	0,3	0,2	1,3	71	128937,40
Итого:						160760,48

5.4.3 Дополнительная заработная плата исполнителей

Дополнительная заработная плата включает заработную плату за не отработанное рабочее время, но гарантированную действующим законодательством.

Расчет дополнительной заработной платы ведется по формуле 9:

$$З_{доп} = k_{доп} \cdot З_{осн} \quad (9)$$

где $k_{доп}$ – коэффициент дополнительной заработной платы (на стадии проектирования принимается равным 0,12 – 0,15).

$k_{доп}$ равен 0,15. Результаты по расчетам дополнительной заработной платы сведены в таблицу 13.

Таблица 13 – Расчет дополнительной заработной платы

Заработная плата	Научный руководитель	Консультант	Инженер
Основная зарплата	14047,45	35551,15	105328,89
Дополнительная зарплата	2107,117037	5332,67	15799,33
Зарплата исполнителя	16154,56	40883,82	121128,23
Итого по статье	184874,54		

5.4.4 Отчисления во внебюджетные фонды (страховые отчисления)

Величина отчислений во внебюджетные фонды определяется исходя из формулы:

$$З_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп}),$$

где $k_{внеб}$ – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.). Данный коэффициент равен 30%. Отчисления во внебюджетные фонды представлены в таблице 14.

Таблица 14 – Отчисления во внебюджетные фонды

Заработная плата	Научный руководитель	Консультант	Инженер
Зарплата исполнителя	16154,56	20441,97	148278,01
Отчисления во внебюджетные фонды	4846,36	6132,59	44483,40

5.4.5 Накладные расходы

Накладные расходы ― расходы на организацию, управление и обслуживание процесса производства товара, оказания услуги; носят комплексный характер, т. е. включают различные экономические элементы затрат; при выпуске предприятием нескольких видов продукции (услуг) накладные расходы должны быть распределены между ними и включены в себестоимость каждого товара (услуги).

Накладные расходы в ТПУ составляют 16% от суммы прямых затрат на разработку, которые, в свою очередь, включают затраты на сырье и материалы, основную и дополнительную заработную плату, отчисления на социальные нужды, стоимость затраченного машинного времени, услуги сторонних организаций:

Накладные расходы - 55462,36 руб.

5.4.6 Формирование бюджета затрат НТИ

Рассчитанная величина затрат научно-исследовательской работы (темы) является основой для формирования бюджета затрат проекта, который при формировании договора с заказчиком защищается научной организацией в качестве нижнего предела затрат на разработку научно-технической продукции.

Определение бюджета затрат на научно-исследовательский проект по каждому варианту исполнения приведен в таблице 15 и на рис. 18.

Таблица 15 – Расчет бюджета затрат НТИ

Наименование	**Сумма, руб.**	**Удельный вес, %**
Материальные затраты	1000	0,35%
Затраты на специальное оборудование	5775	2,01%

Затраты на основную заработную плату	160760,48	56,08%
Затраты на дополнительную заработную плату	24114,07	8,41%
Страховые взносы	55462,36	19,35%
Накладные расходы	39537,91	13,79%
Общий бюджет	286649,82	100,00%

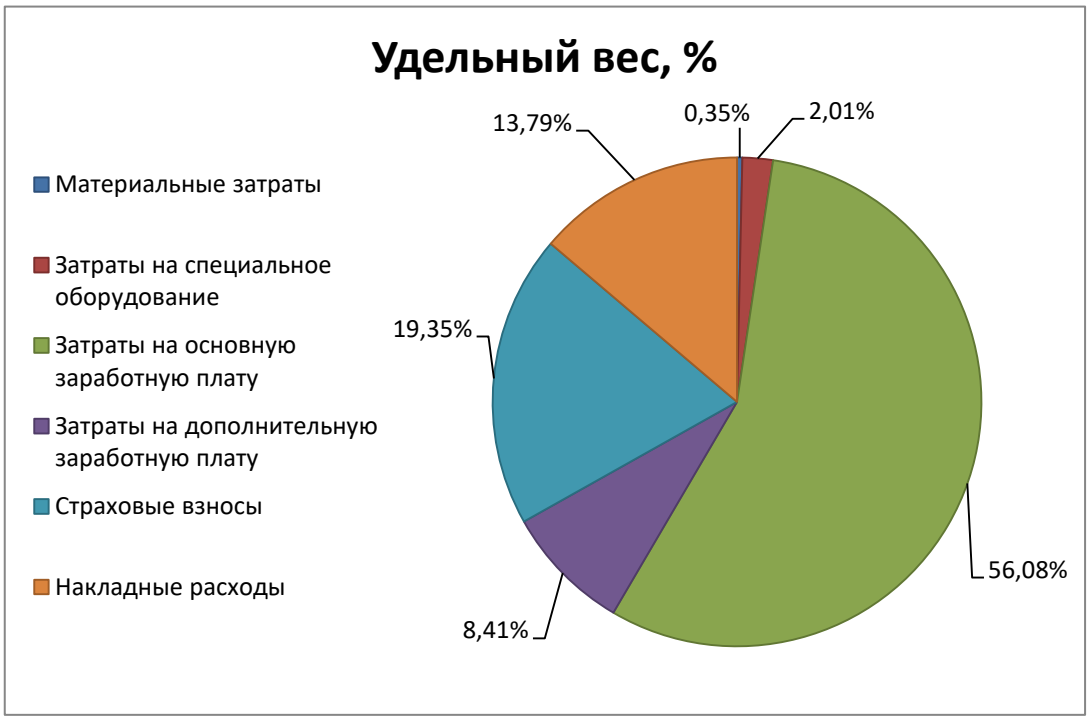

Рисунок 18 – Структура затрат

Основываясь на данных, полученных в предыдущих пунктах, был рассчитан бюджет затрат научно-исследовательской работы для трех исполнителей. Затраты на полную реализацию составляют **286649,82** рублей. Исходя из расчета бюджета затрат следует, что наибольшая его часть приходится на основную и дополнительную заработную плату исполнителей (64,49 %). Также необходимо отметить, что расходы на страховые взносы (19,35 %) составляют немаловажную часть расходов. Затраты на амортизацию, материалы и накладные расходы составляют небольшую долю (суммарно 16,16 %). Это связано с отсутствие необходимости использования значительно дорогостоящего оборудования и материалов.

5.5 Риски проекта

Риск – это возможность наступления некоторого неблагоприятного события, влекущего за собой возникновение различного рода потерь. Для предотвращения их наступления необходимо составить реестр возможных рисков.

Таблица 16 – Реестр рисков

№	Риск	Вероятность наступления (1-5)	Влияние риска (1-5)	Уровень риска*	Способы смягчения риска
1	Отсутствие поддержки бесплатных библиотек	1	4	Умеренный	Использовать проверенные библиотеки
2	Алгоритм ухудшит предсказания	2	2	Существенный	Проверять алгоритм на модели машинного обучения
3	Изменения формата входных данных	3	5	Существенный	Коммуникация с заказчиком
4	Система сломается от большого количества информации	2	4	Умеренный	Повышать устойчивость, осуществлять параллельные вычисления
5	Заказчик не захочет усложнять имеющийся алгоритм	1	3	Незначительный	Проводить анализ и презентации возможных решений.
6	Заказчик изменит требования к системе	2	3	Существенный	Коммуникация внутри проекта с заказчиком.

В данном разделе были проанализированы риски НТИ, были предложены методы по их смягчению и предотвращению. Одним из самых главных условий для избежания рисков является хорошая коммуникация между исполнителем и заказчиком и грамотная архитектура проекта на начальном этапе.

5.6 Определение интегрального показателя ресурсоэффективности

Расчет интегрального показателя ресурсоэффективности работы над объектом исследования определяется по следующей формуле:

$$I_p = \sum a * b,$$

где I_p– интегральный показатель ресурсоэффективности;

a – весовой коэффициент;

b – бальная оценка, устанавливается экспертным путем по выбранной шкале оценивания;

n – число параметров сравнения.

В таблице 17 отображены расчеты интегрального показателя ресурсоэффективности

Таблица 17 Оценка характеристик исполнения проекта

Объект исследования / Критерии	**Весовой коэффициент параметра**	**Оценка выполнения**
1. Рост производительности труда пользователя	0,30	5
2. Удобство в эксплуатации и соответствие требованиям потребителей	0,35	4
3. Надежность и безопасность	0,15	4
4. Минимизация времени	0,15	5
5. Единство исполнения	0,05	5
ИТОГО	1,00	

$$I_p = 0{,}30 * 5 + 0{,}35 * 4 + 0{,}15 * 4 + 0{,}15 * 5 + 0{,}05 * 5 = 4{,}5$$

В ходе разработки части диссертации, затрагивающей финансовую и ресурсную эффективность, был проведен анализ потребителей алгоритма предварительной кластеризации задач в больших вычислительных сетях.

После рассмотрения конкурентных решений с помощью оценочной карты QuaD можно сделать вывод, что данное исследование может быть перспективным.

Далее была определена структура работ проекта и назначены ответственные исполнители. На основе этого была рассчитана общая длительность проекта, которая составила 97,6 календарных дней. Для рассматриваемого исследования был рассчитан общий бюджет НТИ. Для текущей разработки бюджет НТИ составил **286649,82** рублей, большая часть приходится на оплату труда исполнителей (64,49%).

В результате проведения данного исследования реализован наиболее оптимальный алгоритм кластеризации для данного набора данных. Планируется, что он повысит точность предсказания длительности обработки задач для ученых

ЦЕРНа, тем самым повысив эффективность их работы, улучшит планирование их временного ресурса и понизит материальные расходы на простои и ожидание результатов.

ГЛАВА 5. СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

Цели, поставленные в процессе работы над магистерской диссертацией, предполагают ускорение процесса предсказания длительности выполнения задач во Всемирной Распределенной Сети Большого Адронного Коллайдера. В процессе работы были реализованы несколько методов кластеризации данных и был проведен сравнительный анализ их эффективности для поиска наилучшего метода.

Выполняемая работа заключалась в исследовании и реализации методов кластеризации. Таким образом, работу можно классифицировать как работу разработчика программного обеспечения.

С появлением компьютеров произошли серьезные изменения в условиях производственной деятельности работников умственного труда. Их труд стал более интенсивным, напряженным, требующим значительных затрат умственной, эмоциональной и физической энергии.

Обеспечение безопасной жизнедеятельности человека в значительной степени зависит от правильной оценки опасных, вредных производственных факторов. Одинаковые по тяжести изменения в организме человека могут быть вызваны различными причинами. Это могут быть какие-либо факторы производственной среды, чрезмерная физическая и умственная нагрузка, нервно-эмоциональное напряжение, а также разное сочетание этих причин.

В данном разделе рассмотрены вопросы безопасной жизнедеятельности на этапе исследования методов машинного обучения без учителя для анализа задач в больших вычислительных сетях. Поскольку на данной стадии разработки проекта практически все работы велись в аудитории за компьютером, правомерно будет рассмотреть вредные факторы, связанные с этим видом работ, воздействие на окружающую среду и возможные чрезвычайные ситуации.

5.1 Правовые и организационные вопросы обеспечения безопасности

5.1.1 Организационные мероприятия при компоновке рабочей зоны

При организации рабочего места необходимо учитывать требования безопасности, промышленной санитарии, эргономики, технической эстетики. Невыполнение этих требований может привести к получению работником производственной травмы или развитию у него профессионального заболевания [30].

При организации работы на ПЭВМ должны выполняться следующие условия:

- рабочее место с персональным компьютером (ПК) должно располагаться по отношению к оконным проемам так, чтобы свет падал сбоку, предпочтительнее слева;
- нужно избегать расположения рабочего места в углах комнаты или лицом к стене (расстояние от ПК до стены должно быть не менее 1 м), экраном и лицом к окну;
- ПК желательно устанавливать так, чтобы, подняв глаза от экрана, можно было увидеть самый удаленный предмет в комнате, так как перевод взгляда на дальнее расстояние – один из самых эффективных способов разгрузки зрительной системы при работе на ПК;
- при наличии нескольких компьютеров расстояние между экраном одного монитора и задней стенкой другого должно быть не менее 2 м, а расстояние между боковыми стенками соседних мониторов – не менее 1,2 м [32];
- окна в помещениях с ПЭВМ должны быть оборудованы регулируемыми устройствами (жалюзи, занавески, внешние козырьки и т.д.);
- монитор, клавиатура и корпус компьютера должны находиться прямо перед оператором; высота рабочего стола с клавиатурой должна составлять 680 – 800 мм над уровнем стола; а высота экрана (над полом) –900–1280см;
- монитор должен находиться от оператора на расстоянии 60 – 70 см на 20 градусов ниже уровня глаз [31];

- пространство для ног должно быть: высотой не менее 600 мм, шириной не менее 500 мм, глубиной не менее 450 мм. Должна быть предусмотрена подставка для ног работающего шириной не менее 300 мм с регулировкой угла наклона 0-20 градусов;
- рабочее кресло должно иметь мягкое сиденье и спинку, с регулировкой сиденья по высоте, с удобной опорой для поясницы;
- Следовать руководству.
- Положение тела пользователя относительно монитора должно соответствовать направлению просмотра под прямым углом или под углом 75 градусов [33].
- Правильная поза и положение рук оператора являются весьма важными для исключения нарушений в опорно-двигательном аппарате и возникновения синдрома постоянных нагрузок.

Согласно СанПиНу 2.2.2.542-96 при 8-ми часовой рабочей смене на ВДТ и ПЭВМ перерывы в работе должны составлять от 10 до 20 минут каждые два часа работы.

5.1.2 Особенности законодательного регулирования проектных решений

При работе с персональным компьютером очень важную роль играет соблюдение правильного режима труда и отдыха.

Вид трудовой деятельности при проведении исследований в лаборатории за компьютером входит в группу В – творческая работа в режиме с диалогом ПЭВМ. Категория тяжести и напряженности работы с ПЭВМВ определяется в зависимости от суммарного времени непосредственной работы с ПЭВМ за рабочую смену, но не более 6 ч за смену. В табл. 1 представлены сведения о регламентированных перерывах, которые необходимо делать при работе на компьютере, в зависимости от продолжительности рабочей смены, видов и категорий трудовой деятельности с ВДТ (видеодисплейный терминал) и ПЭВМ [30].

Таблица 18 - Время регламентированных перерывов при работе на компьютере

Категория работы с ВДТ или ПЭВМ	Уровень нагрузки за рабочую смену при видах работы с ВДТ	Суммарное время регламентированных перерывов, мин
	Группа В, часов	При 8-часовой смене
I	до 2,0	30
II	до 4,0	50
III	до 6,0	70

Время перерывов дано при соблюдении указанных Санитарных правил и норм. При несоответствии фактических условий труда требованиям Санитарных правил и норм время регламентированных перерывов следует увеличить на 30%.

Эффективность перерывов повышается при сочетании с производственной гимнастикой или организации специального помещения для отдыха персонала с удобной мягкой мебелью, аквариумом, зеленой зоной и т.п.

5.2 Профессиональная социальная ответственность

В данном разделе дипломной работы приведены: оценка условий труда на рабочем месте, анализ вредных и опасных факторов труда, разработка мер защиты от них. Темой выпускной работы является «Исследование методов машинного обучения без учителя для анализа задач в больших вычислительных сетях». Объектом исследования выступает рабочее место, оборудование, помещение, в котором находится это рабочее место. Все исследования производились во время работы над дипломным проектом в помещении, где выполнялась эта работа. Лаборатория, в которой проводилось исследование и разработка алгоритмов находится в научно-технической библиотеке Томского Политехнического университета (НТБ ТПУ).

Полностью безопасных и безвредных производств не бывает, поэтому с целью уменьшения воздействия различных неблагоприятных факторов прибегают к такой дисциплине как охрана труда.

Основным оборудованием для выполнения исследований является компьютер. При этом, опасным для разработчика фактором является высокое напряжение в электрической сети и как следствие, опасность поражения электрическим током. Напряжение в сети составляет 220В при частоте 50Гц, что является смертельно опасным в случае поражения работающего электрическим током.

К вредным производственным факторам, при работе с компьютером следует отнести:

1. повышенный уровень электромагнитных излучений, основными источниками которых является монитор компьютера;
2. отклонение показателей микроклимата
3. повышенный уровень шума, источниками которого являются вентиляторы внутри системного блока и блока питания компьютера, накопители на жестких и магнитных дисках, светильники люминесцентных ламп и др.
4. повышенный уровень ионизирующих излучений, источником которых является дисплей монитора компьютера
5. недостаточная освещённость рабочей зоны [21]

5.2.1 Повышенный уровень электромагнитных излучений

Как любые электрические приборы, видеотерминалы (ВДТ) и системные блоки производят электромагнитное излучение, воздействие которого на человека зависит от напряжённостей электрического и магнитного полей, потока энергии, частоты колебаний, размера облучаемого тела.

Нарушения в организме человека при воздействии электромагнитных полей незначительных напряженностей носят обратимый характер. При воздействии полей, имеющих напряженность выше предельно допустимого уровня, развиваются нарушения со стороны нервной, сердечно-сосудистой систем, органов пищеварения и некоторых биологических показателей крови.

Большая часть электромагнитных излучений происходит не от экрана монитора, а от видеокабеля и системного блока. В портативных компьютерах

практически всё электромагнитное излучение идет от системного блока, располагающегося под клавиатурой. Современные машины выпускаются заводом-изготовителем со специальной металлической защитой внутри системного блока для уменьшения фона электромагнитного излучения.

Напряженность электромагнитного поля на расстоянии 50см вокруг ВДТ по электрической составляющей должна быть не более:

- В диапазоне частот 5 Гц ÷ 2 кГц – 25 В/м;
- В диапазоне частот 2 кГц ÷ 400кГц – 2,5 В/м.

Плотность магнитного потока должна быть не более:

- В диапазоне частот 5 Гц ÷ 2 кГц – 250 нТл;
- В диапазоне частот 2 кГц ÷ 400кГц – 25 нТл.

Возможные способы защиты от ЭМП:

Основной способ – увеличение расстояния от источника, экран видеомонитора должен находится на расстоянии не менее 50 см от пользователя;

Применение приэкранных фильтров, специальных экранов и других средств индивидуальной защиты, прошедших испытание в аккредитованных лабораториях и имеющих соответствующий гигиенический сертификат.

5.2.2 Отклонение показателей микроклимата

Проанализируем микроклимат в помещении, где находится рабочее место. Воздух рабочей зоны (микроклимат) производственных помещений определяют следующие параметры: температура, относительная влажность, скорость движения воздуха. Параметры микроклимата оказывают непосредственное влияние на тепловое самочувствие человека и его работоспособность.

Например, понижение температуры и повышение скорости воздуха может привести к переохлаждению организма.

При повышении температуры воздуха возникают обратные явления. Исследователями установлено, что при температуре воздуха более 30 ^{0}C работоспособность человека начинает падать.

Для человека определены максимальные температуры в зависимости от длительности их воздействия и используемых средств защиты. Предельная температура вдыхаемого воздуха, при которой человек в состоянии дышать в течение нескольких минут без специальных средств защиты, около 116 ^{0}C.

Существенное значение имеет равномерность температуры. Вертикальный градиент ее не должен выходить за пределы 5 ^{0}C/метр.

Переносимость человеком температуры, как и его теплоощущение, в значительной мере зависит от влажности окружающего воздуха. Чем больше относительная влажность, тем меньше испаряется пота в единицу времени и тем быстрее наступает перегрев тела. Особенно неблагоприятное воздействие на тепловое самочувствие человека оказывает высокая влажность при $t_{oc} > 30$ ^{0}C, так как при этом почти все выделяемая теплота отдается в окружающую среду при испарении пота.

Недостаточная влажность воздуха также может оказаться неблагоприятной для человека вследствие интенсивного испарения влаги со слизистых оболочек, их пересыхания и растрескивания, а затем и загрязнения болезнетворными микроорганизмами. Поэтому при длительном пребывании людей в закрытых помещениях рекомендуется ограничиваться относительной влажностью в пределах 30...70 %.

Вместе с потом организм теряет значительное количество минеральных солей (до 1%, в том числе 0,4...0,6 NaCl). При неблагоприятных условиях потеря жидкости может достигать 8…10 л за смену и в ней до 60 г поваренной соли (всего в организме около 140 г NaCl). Потеря соли лишает кровь способности удерживать воду и приводит к нарушению деятельности сердечно-сосудистой системы.

Атмосферное давление оказывает существенное влияние на процесс дыхания и самочувствие человека. При работе в условиях избыточного давления снижаются показатели вентиляции легких за счет некоторого урежения частоты дыхания и пульса. Длительное пребывание при избыточном давлении приводит к токсическому действию некоторых газов, входящих в состав вдыхаемого воздуха.

Оно проявляется в нарушении координации движений, возбуждении или угнетении, галлюцинациях, ослаблении памяти, расстройстве зрения и слуха [22].

Оптимальные значения характеристик микроклимата приведены в таблицах 2 и 3

По степени физической тяжести работа инженера-программиста относится к лёгкой физической работе категории I а, с энергозатратами организма до 120 Дж/с, т.к. работа проводилась сидя, не требуя систематического физического напряжения.

Таблица 19 – Оптимальные значения характеристик микроклимата

Период года	**Категория работ по уровню энергозатрат, Вт**	**Температура воздуха, °С**	**Температура поверхностей, °С**	**Относительная влажность воздуха, %**	**Скорость движения воздуха, м/с**
Холодный	Ia (до 139)	22 - 24	21 - 25	60 - 40	0,1
Теплый	Ia (до 139)	23 - 25	22 - 26	60 - 40	0,1

Таблица 20 –Допустимые значения характеристик микроклимата

Период года	**Категория работ по уровню энерго-затрат, Вт**	**Температура воздуха, °С**	**Температура поверхностей°С**	**Относит. влажность воздуха, %**	**Скорость движения воздуха, м/с**
Холодный	Ia(до 139)	20,0-25,0	19,0-26,0	15-75	0,1
Теплый	Ia (до 139)	21,0-28,0	20,0-29,0	15-75	0,1-0,2

Параметры микроклимата в помещении, где находится рабочее место, регулируются системой центрального отопления и приточно-вытяжной вентиляцией, и имеют следующие значения: влажность 40%, скорость движения воздуха 0,1 м/с, температура летом 20-25°С, зимой 20-22°С, что соответствует требованиям, представленных в таблице 1.1.

К мероприятиям по оздоровлению воздушной среды в производственном помещении относятся: правильная организация вентиляции и кондиционирования

воздуха, отопление помещений. Вентиляция может осуществляться естественным и механическим путём. В рабочем помещении должны подаваться следующие объёмы наружного воздуха: при объёме помещения до 20м^3 на человека – не менее 30м^3 в час на человека; при объёме помещения более 40м^3 на человека и отсутствии выделения вредных веществ допускается естественная вентиляция.

В аудитории отсутствует принудительная вентиляция. Имеется лишь естественная, т.е. воздух поступает и удаляется через щели, окна, двери. Основной недостаток такой вентиляции в том, что приточный воздух поступает в помещение без предварительной очистки и нагревания. Естественная вентиляция допускается при условии, что на одного работающего приходится более 40м^3 объема воздуха в помещении. Поскольку в помещении не выполняется требование к объему воздуха на одного работающего (объём на одного человека — 28,88м^3), то наличие принудительной вентиляции просто необходимо.

В зимнее время в помещении необходимо предусмотреть систему отопления. Она должна обеспечивать достаточное, постоянное и равномерное нагревание воздуха. В помещениях с повышенными требованиями к чистоте воздуха должно использоваться водяное отопление. В рассматриваемой аудитории используется водяное отопление со встроенными нагревательными элементами и стояками.

5.2.3 Недостаточная освещённость рабочей зоны

Недостаточное освещение влияет на функционирование зрительного аппарата, то есть определяет зрительную работоспособность, на психику человека, его эмоциональное состояние, вызывает усталость центральной нервной системы, возникающей в результате прилагаемых усилий для опознания четких или сомнительных сигналов. [24]

Для оптимизации условий труда имеет большое значение освещение рабочих мест. Задачи организации освещённости рабочих мест следующие: обеспечение различаемости рассматриваемых предметов, уменьшение напряжения и утомляемости органов зрения. Производственное освещение

должно быть равномерным и устойчивым, иметь правильное направление светового потока, исключать слепящее действие света и образование резких теней.

Среди качественных показателей световой среды очень важным является коэффициент пульсации освещенности (Кп). Требования к коэффициенту пульсации освещенности наиболее жесткие для рабочих мест с ПЭВМ — не более 5%. [24] Оптимальная яркость экрана дисплея составляет 75–100 кд/м2. При такой яркости экрана и яркости поверхности стола в пределах 100–150 кд/м2 обеспечивается продуктивность работы зрительного аппарата на уровне 80–90 %, сохраняется постоянство размера зрачка на допустимом уровне 3–4 мм.

Таблица 21 – Нормируемые показатели естественного, искусственного и совмещенного освещения в соответствии с СанПиН 2.2.1/2.1.1.1278-03

Помещения	Рабочая поверхность и плоскость нормирования КЕО и освещенности и высота плоскости над полом, м	Естественное освещение		Совмещенное освещение		Искусственное освещение				
		КЕО е.н, %		КЕО е.н, %		Освещенность, лк				
		При верхнем или комбинированном освещении	При боковом освещении	При верхнем или комбинированном освещении	При боковом освещении	При комбинированном освещении		При общем освещении	Показатель дискомфорта, М, не более	Коэффициент пульсации освещенности, К_п, %, не более
						всего	от общего			
1	2	3	4	5	6	7	8	9	10	11
Помещения для работы с дисплеями и видеотерминалами, залы ЭВМ	Г – 0,8 Экран монитора: В – 1,2	3,5 -	1,2 -	2,1 -	0,7 -	500 -	300 -	400 200	15 -	10 -
Кабинеты, рабочие комнаты	Г – 0,8	3,0	1,0	1,8	0,6	400	200	300	40	15

Местное освещение не должно создавать бликов на поверхности экрана и увеличивать освещенность экрана ПЭВМ более 300 лк. Следует ограничивать прямую и отраженную блесткость от любых источников освещения.

В лаборатории, где проводятся исследования, используется смешанное освещение, т.е. сочетание естественного и искусственного освещения.

Естественным освещением является освещение через окна. Искусственное освещение используется при недостаточном естественном освещении. В данном помещении используется общее искусственное освещение.

Лаборатория освещается 3 светильниками, в каждом из которых установлено 4 люминесцентных лампы типа ЛБ-40. Светильники расположены равномерно по всей площади потолка в ряд, создавая при этом равномерное освещение рабочих мест. Световой поток каждой из ламп в помещении свидетельствует о соблюдении норм освещенности.

5.2.4 Повышенный уровень шума на рабочем месте

Одним из важнейших параметров, наносящим большой ущерб для здоровья и резко снижающим производительность труда, является шум.

Шум может создаваться работающим оборудованием, установками кондиционирования воздуха, преобразователями напряжения, работающими осветительными приборами дневного света, а также проникать извне.

В результате исследований установлено, что шум и вибрация ухудшают условия труда, оказывают вредное воздействие на организм человека. Действие шума различно: он затрудняет разборчивость речи, вызывает снижение работоспособности, повышает утомляемость, вызывает необратимые изменения в органах слуха человека. Шум воздействует не только на органы слуха, но и на весь организм человека через центральную нервную систему. Ослабляется внимание, ухудшается память, снижается реакция, увеличивается число ошибок при работе.

Производственные помещения, в которых для работы используются ПЭВМ, не должны граничить с помещениями, в которых уровень шума и вибрации превышают нормируемые значения. При выполнении основной работы на ПЭВМ уровень шума на рабочем месте не должен превышать 50 дБ. Допустимые уровни звукового давления в помещениях для персонала,

осуществляющего эксплуатацию ПЭВМ при разных значениях частот, приведены в таблице 22. [25]

Таблица 22 – Допустимые уровни звукового давления на рабочих местах расчетчиков, программистов вычислительных машин

Уровни звукового давления, дБ, в октавных полосах со среднегеометрическими частотами, Гц								Уровни звука и эквивалентные уровни звука, дБ А
63	125	250	500	1000	2000	4000	8000	
71	61	54	49	45	42	40	38	50

По субъективным ощущениям шумовая обстановка на рабочем месте соответствует норме.

5.2.5 Электробезопасность

Статическое электричество возникает в результате сложных процессов, связанных с перераспределением электронов и ионов при соприкосновении двух поверхностей неоднородных жидких или твердых веществ, на которых образуется двойной электрический слой. При механическом разделении поверхностей происходит разделение зарядов этого двойного электрического слоя. При этом между разделенными поверхностями, несущими электрический заряд, образуется разность потенциалов и возникает электрическое поле.

В помещении разрядные токи статического электричества чаще всего возникают при прикосновении пользователей к любому из элементов ЭВМ. Такие разряды опасности для человека не представляют, однако, кроме неприятных ощущений, они могут привести к выходу из строя ЭВМ.

Для снижения величин возникающих зарядов статического электричества в помещении покрытие полов выполнено из однослойного линолеума.

При работе с электроприборами очень важно соблюдать технику безопасности.

Под техникой безопасности понимается система организационных мероприятий и технических средств, направленная на предотвращения воздействия на работника вредных и опасных производственных факторов.

Электрические установки представляют для человека большую потенциальную опасность, которая усугубляется тем, что органы чувств человека не могут на расстоянии обнаружить наличие электрического напряжения на оборудовании.

В зависимости от условий в помещении опасность поражения человека электрическим током увеличивается или уменьшается. Не следует работать с компьютером в условиях повышенной влажности (относительная влажность воздуха длительно превышает 75%), высокой температуры (более 35°С), наличии токопроводящей пыли, токопроводящих полов и возможности одновременного соприкосновения к имеющим соединение с землей металлическим элементам и металлическим корпусом электрооборудования. Таким образом, работа может проводиться только в помещениях без повышенной опасности, при этом существует опасность электропоражения:

1) при непосредственном прикосновении к токоведущим частям во время ремонта ПЭВМ;

2) при прикосновении к нетоковедущим частям, оказавшимся под напряжением (в случае нарушения изоляции токоведущих частей ПЭВМ);

3) при соприкосновении с полом, стенами, оказавшимися под напряжением;

4) имеется опасность короткого замыкания в высоковольтных блоках: блоке питания и блоке дисплейной развёртки [26].

Лаборатория, в которой проводились работы, по опасности электропоражения относится к помещениям без повышенной опасности, то есть отсутствуют условия, создающие повышенную опасность.

В помещении используются приборы, потребляющие напряжение 220В переменного тока с частотой 50Гц. Это напряжение опасно для жизни, поэтому обязательны следующие меры предосторожности:

1) перед началом работы нужно убедиться, что выключатели и розетка закреплены и не имеют оголённых токоведущих частей;

2) при обнаружении неисправности оборудования и приборов необходимо не делая никаких самостоятельных исправлений сообщить ответственному за оборудование;

3) запрещается загромождать рабочее место лишними предметами. При возникновении несчастного случая следует немедленно освободить пострадавшего от действия электрического тока и, вызвав врача, оказать ему необходимую помощь.

5.3 Экологическая безопасность

Вследствие развития научно-технического прогресса, постоянно увеличивается возможность воздействия на окружающую среду, создаются предпосылки для возникновения экологических кризисов. В то же время прогресс расширяет возможности устранения создаваемых человеком ухудшений природной среды.

Под окружающей нас средой понимается совокупность «чистой» природы и среды созданной человеком.

Защита окружающей среды - это комплексная проблема, требующая усилий всего человечества. Наиболее активной формой защиты окружающей среды от вредного воздействия выбросов промышленных предприятий является полный переход к безотходным и малоотходным технологиям и производствам. Это потребует решения целого комплекса сложных технологических, конструкторских и организационных задач, основанных на использовании новейших научно-технических достижений [27].

5.3.1 Загрязнение атмосферного воздуха

Во время проведения исследований выбросы вредных веществ в атмосферу не осуществляются. Загрязнение атмосферного воздуха может возникнуть в случае возникновения пожара в учебном корпусе, в этом случае дым и газы от пожара будут являться антропогенным загрязнением атмосферного воздуха.

5.3.2 Отходы

Основные виды загрязнения литосферы – твердые бытовые и промышленные отходы.

При проведении исследований, образовывались различные твердые отходы. К ним можно отнести: бумагу, батарейки, лампочки, использованные картриджи, отходы от продуктов питания и личной гигиены, отходы от канцелярских принадлежностей и т.д. [27]

Защита почвенного покрова и недр от твердых отходов реализуется за счет сбора, сортирования и утилизации отходов и их организованного захоронения.

5.4 Безопасность в чрезвычайных ситуациях

5.4.1 Пожарная профилактика

Мероприятия по пожарной профилактике разделяются на организационные, технические, эксплуатационные и режимные.

Организационные мероприятия предусматривают правильную эксплуатацию оборудования, правильное содержание зданий и территорий, противопожарный инструктаж рабочих и служащих, обучение производственного персонала правилам противопожарной безопасности, издание инструкций, плакатов, наличие плана эвакуации.

К техническим мероприятиям относятся: соблюдение противопожарных правил, норм при проектировании зданий, при устройстве электропроводов и оборудования, отопления, вентиляции, освещения, правильное размещение оборудования.

К режимным относятся установление правил организации работ и соблюдение противопожарных мер [26].

5.4.2 Оценка пожарной безопасности помещения

Согласно нормам технологического проектирования, в зависимости от характеристики используемых в производстве веществ и их количества, по

пожарной и взрывной опасности помещения подразделяются на категории А, Б, В, Г, Д.

Наличие в лаборатории деревянных изделий (столы, шкафы), электропроводов напряжением 220В, а также применение электронагревательных приборов с открытыми нагревательными элементами – паяльниками дает право отнести помещение по степени пожаро и взрывобезопасности к категории В. Категория помещения «В»: помещения, в которых горючие и трудногорючие жидкости, твердые горючие и трудногорючие вещества и материалы (в том числе пыли и волокна), вещества и материалы, находящиеся в помещении, способны при взаимодействии с водой, кислородом воздуха или друг с другом гореть, при условии, что помещения, в которых они имеются в наличии или обращаются, не относятся к категориям А или Б.

Необходимо предусмотреть ряд профилактических мероприятий технического, эксплуатационного, организационного плана.

В качестве возможных причин пожара можно указать следующие:

1) наличие горючей пыли (некоторые осевшие частицы пыли способны к самовозгоранию);

2) короткие замыкания;

3) опасная перегрузка сетей, которая ведет за собой сильный нагрев токоведущих частей и загорание изоляции;

4) нередко пожары происходят при пуске оборудования после ремонта [29].

Для предупреждения пожаров от коротких замыканий и перегрузок необходимы правильный выбор, монтаж и соблюдение установленного режима эксплуатации электрических сетей, дисплеев и других электрических средств автоматизации.

Следовательно, необходимо предусмотреть ряд профилактических мероприятий технического, эксплуатационного, организационного плана.

5.4.3 Анализ возможных причин загорания

Причиной возгорания может быть:

1) неисправность токоведущих частей установок;

2) работа с открытой электроаппаратурой;

3) короткие замыкания в блоке питания или высоковольтном блоке дисплейной развертки;

4) несоблюдение правил пожарной безопасности;

5) наличие горючих компонентов: документы, двери, столы, изоляция кабелей и т.п.

5.4.4 Мероприятия по устранению и предупреждению пожаров

Для предупреждения возникновения пожара необходимо соблюдать следующие правила пожарной безопасности:

1) исключение образования горючей среды (герметизация оборудования, контроль воздушной среды, рабочая и аварийная вентиляция);

2) применение при строительстве и отделке зданий несгораемых или трудно сгораемых материалов.

Необходимо в аудитории проводить следующие пожарно-профилактические мероприятия:

1) организационные мероприятия, касающиеся технического процесса с учетом пожарной безопасности объекта;

2) эксплуатационные мероприятия, рассматривающие эксплуатацию имеющегося оборудования;

3) технические и конструктивные, связанные с правильным размещением и монтажом электрооборудования и отопительных приборов.

Организационные мероприятия:

1) противопожарный инструктаж обслуживающего персонала;

2) обучение персонала правилам техники безопасности;

3) издание инструкций, плакатов, планов эвакуации.

Эксплуатационные мероприятия:

1) соблюдение эксплуатационных норм оборудования;

2) обеспечение свободного подхода к оборудованию;

3) содержание в исправности изоляции токоведущих проводников.

Технические мероприятия:

1) соблюдение противопожарных мероприятий при устройстве электропроводок, оборудования, систем отопления, вентиляции и освещения. В лаборатории имеется углекислотный огнетушитель типа ОУ–2, установлен рубильник, обесточивающий всю аудиторию, на двери аудитории приведен план эвакуации в случае пожара, и на досягаемом расстоянии находится пожарный щит (2 этаж НТБ). Если возгорание произошло в электроустановке, для его устранения должны использоваться углекислотные огнетушители типа ОУ–2.

2) профилактический осмотр, ремонт и испытание оборудования.

Кроме устранения самого очага пожара, нужно своевременно организовать эвакуацию людей [26].

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Кореньков В. В. Распределенная система для обработки, хранения и анализа экспериментальных данных Большого адронного коллайдера //Современные информационные технологии и ИТ-образование. – 2012. – №. 8.

2. Конспект лекции «Уменьшение размерности описания данных: метод главных компонент» по курсу «Математические основы теории прогнозирования» 2011

3. Замятин А. В. и др. Введение в интеллектуальный анализ данных: учебное пособие. – 2016.

4. Береснев Д. В., Шараев Е. В. Анализ методов понижения размерности пространства //КОМП'ЮТЕРНІ ТА ІНФОРМАЦІЙНІ СИСТЕМИ І ТЕХНОЛОГІЇ. – 2019.

5. Игорь М., Мотренко А. Модификация метода t-SNE для задачи классификации. – 2016.

6. Воронцов К.В. Алгоритмы кластеризации и многомерного шкалирования. Курс лекций. МГУ, 2007

Дутов И. Ю. Применение методов машинного обучения для предсказания длительности цепочек вычислительных задач. – 2017.

8. Алиев М. Решение задачи кластеризации отпечатков пальцев методами глубокого обучения. – Санкт-Петербург 2016

9. Qian Y. et al. Space structure and clustering of categorical data //IEEE transactions on neural networks and learning systems. – 2015. – Т. 27. – №. 10. – С. 2047-2059.

10. Weber L. M., Robinson M. D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data //Cytometry Part A. – 2016. – Т. 89. – №. 12. – С. 1084-1096.

11. Hsu D. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data //Applied energy. – 2015. – Т. 160. – С. 153-163.

12. Musmeci N., Aste T., Di Matteo T. Relation between financial market structure and the real economy: comparison between clustering methods //PloS one. – 2015. – Т. 10. – №. 3. – С. e0116201.

13. Huang Z. Clustering large data sets with mixed numeric and categorical values //Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD). – 1997. – С. 21-34.

14. Портал функционирования Европейской грид-инфраструктуры [Электронный ресурс]. URL: http://accounting.egi.eu, свободный. – Дата обращения: 27.04.2019 г.

15. Орлова И. В., Филонова Е. С. Выбор экзогенных факторов в модель регрессии при мультиколлинеарности данных //Международный журнал прикладных и фундаментальных исследований. – 2015. – №. 5-1. – С. 108-116.

16. Abonyi J., Feil B. Cluster analysis for data mining and system identification. – Springer Science & Business Media, 2007.

17. Jain A. K. Data clustering: 50 years beyond K-means //Pattern recognition letters. – 2010. – Т. 31. – №. 8. – С. 651-666.

18. Schubert E. et al. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN //ACM Transactions on Database Systems (TODS). – 2017. – Т. 42. – №. 3. – С. 19.

19. Andrecut M. Parallel GPU implementation of iterative PCA algorithms //Journal of Computational Biology. – 2009. – Т. 16. – №. 11. – С. 1593-1599.

20. Van Der Maaten L. Fast optimization for t-SNE //In Neural Information Processing Systems (NIPS) 2010 Workshop on Challenges in Data Visualization. – 2010. – Т. 100.

21. Охрана труда. Вредные и опасные факторы производства [Электронный ресурс]. URL: http://www.grandars.ru/shkola/bezopasnost-zhiznedeyatelnosti/ohrana-truda.html, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

22. СанПиН 2.2.4.548-96 «Гигиенические требования к микроклимату производственных помещений. Санитарные правила и нормы» [Электронный ресурс]. URL: https://ohranatruda.ru/ot_biblio/normativ/data_normativ/5/5225/, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

23. СанПиН 2.2.2/2.4.1340-03. Санитарно-эпидемиологические правила и нормы. Гигиенические требования к персональным электронно- вычислительным машинам и организации работы // Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс]. URL: http://docs.cntd.ru/document/901865498, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

24. СанПиН 2.2.1/2.1.1.1278-03. Гигиенические требования к естественному, искусственному и совмещенному освещению жилых и общественных зданий. // Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс]. URL: http://docs.cntd.ru/document/901865479, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

25. СНиП 23 – 03 – 2003. Защита от шума. // Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс]. URL: http://docs.cntd.ru/document/901836458, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

26. ГОСТ Р 12.1.019-2009 ССБТ. Электробезопасность. Общие требования и номенклатура видов защиты // Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс]. URL: http://docs.cntd.ru/document/1200080203, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

27. Постановление Правительства РФ от 03.09.2010 N 681 (ред. от 01.10.2013) "Об утверждении Правил обращения с отходами производства и потребления в части осветительных устройств, электрических ламп,

ненадлежащие сбор, накопление, использование, обезвреживание, транспортирование и размещение которых может повлечь причинение вреда жизни, здоровью граждан, вреда животным, растениям и окружающей среде // Государственная система правовой информации [Электронный ресурс]. URL: http://pravo.gov.ru/proxy/ips/?docbody=&nd=102141053, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

28. Как утилизировать люминесцентную лампу? | Экологические проблемы и их решения [Электронный ресурс]. URL: http://eco63.ru/lampalum.html, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

29. Долин П.А. Справочник по технике безопасности. М.: Энергоатомиздат, 1984 г. – 824 с.

30. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 3.07.2016) // Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс]. URL: http://docs.cntd.ru/document/901807664, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

31. ГОСТ Р 50923-96 Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения // Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс]. URL: http://docs.cntd.ru/document/1200025975, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

32. ГОСТ 22269-76 Система "Человек-машина". Рабочее место оператора. Взаимное расположение элементов рабочего места. Общие эргономические требования // Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс]. URL:

http://docs.cntd.ru/document/1200012834, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

33. ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования // Электронный фонд правовой и нормативно-технической документации. [Электронный ресурс]. URL: http://docs.cntd.ru/document/1200003913, свободный. – Загл. с экрана. – Дата обращения: 06.04.2019 г.

СПИСОК ПУБЛИКАЦИЙ И ОСНОВНЫХ НАУЧНЫХ ДОСТИЖЕНИЙ

1. Шкабара А.И. Исследование методов кластеризации для анализа задач в больших вычислительных сетях // Труды Томского государственного университета - 2019 - принято в печать.
2. Победитель конкурса на получение именной стипендии Стипендиальной программы Владимира Потанина 2017/18 .
3. Стипендиат международной грантовой программы Erasmus+, студент по обмену в Мариборском Университете, г. Марибор, Словения 2018.
4. Стипендиат Правительственной стипендии по ПНР, 2018г.
5. Диплом победителя VII Международной молодежной научной конференции «Математическое и программное обеспечение информационных, технических и экономических систем», Томск 2019.
6. Стендовый доклад на международной конференции «The 2nd International School on Heterogeneous Computing Infrastructure» Черногория, 2017
7. Международная школа «Big data analysis for smart cities», Участник, куратор команды, Томск, 2018.
8. Летняя международная школа в Пекинском Технологическом Институте, Диплом за лучший дизайн проекта, Пекин, 16.08.2018

Приложение А

Раздел 1

ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА

CLUSTER ANALYSIS METHODS

Студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Шкабара Анастасия Игоревна		

Консультант ОИТ ИШИТР:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Соколова Вероника Валерьевна	к.т.н.		

Консультант – лингвист ОИЯ ШБИП:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИЯ ШБИП	Диденко Анастасия Владимировна	к.ф.н		

1 SUBJECT AREA OVERVIEW

The development of research in high-energy physics, astrophysics, biology, Earth science and other scientific fields requires a lot of work among different organizations to process huge amount of data in quite a short time. It demands geographically distributed computing systems that are capable of transmitting and receiving hundreds of terabytes of data per day, handling hundreds of thousands of tasks at one time and storing hundreds of petabytes of data for a long time.

Modern grid infrastructure provides integration of hardware and software resources located in different organizations across countries, regions, continents in a single computing environment. It allows solving the problem of processing large amounts of data, which is currently impossible to achieve in local computing centers.

The most impressive results on the global distributed computing infrastructure organization are obtained in the WLCG project (Worldwide LHC Computing Grid) at CERN, where data from LHC experiments (Large Hadron Collider) is processed. It developed a basic computing model for LHC experiments as a hierarchical centralized structure of regional centers, which includes centers of several levels [1].

The essence of distributed computing model is that the entire volume of information from LHC detectors after real-time processing and primary reconstruction (recovered particle tracks, their pulses and other characteristics from a chaotic set of signals from different recording systems) should be sent for further processing and analysis to regional centers of different levels (Tiers):

Tier0 (CERN) => Tier1 => Tier2 => Tier3 => user computers

The levels differ by resources (network, computing, disk, archive) and functions they perform:

- Tier0 (CERN) - primary event reconstruction, calibration, storage of complete databases copies;
- Tier1 - complete reconstruction of events, storage of relevant events databases, creation and storage of analyzed events sets, modeling, analysis;
- Tier2 - replication and storage of analyzed events sets, modeling, analysis.

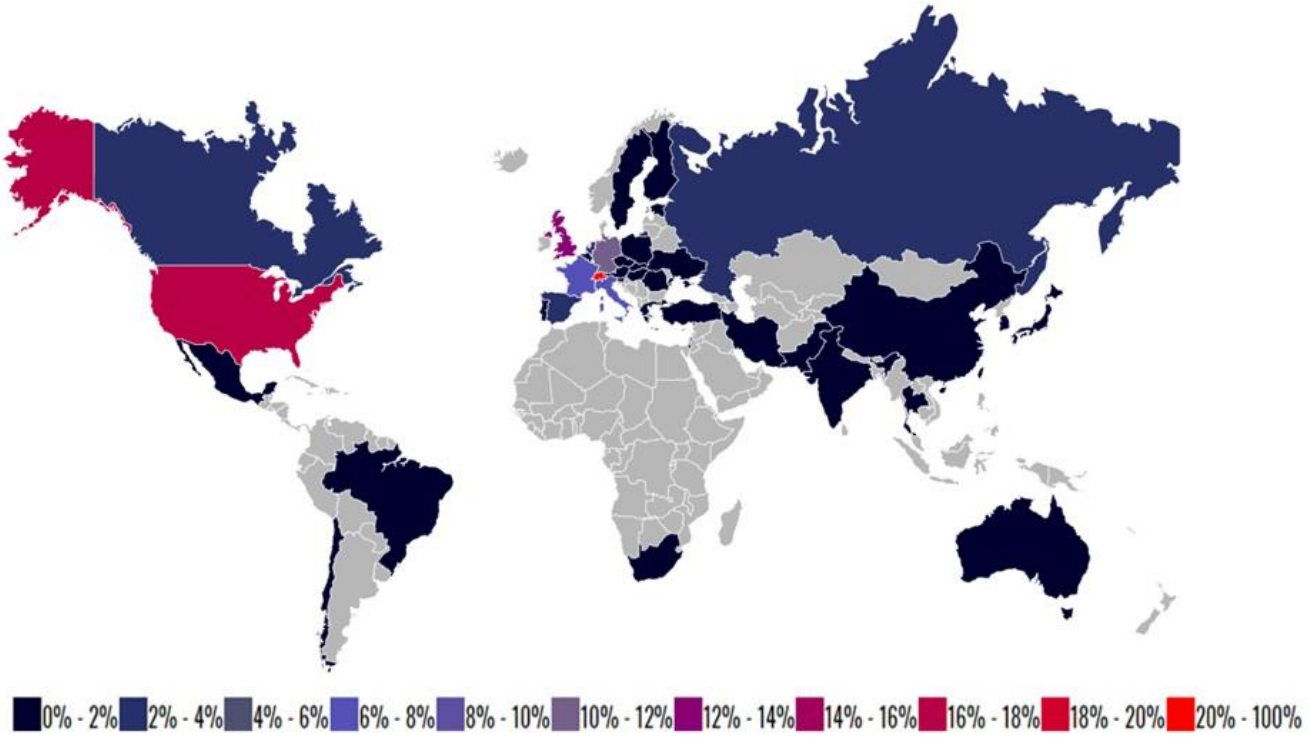

Figure 19 –Countries with nodes for jobs to be processed

Currently, the WLCG project brings together more than 150 grid sites, more than 300,000 CPUs, more than 250 Pb of storage systems on disks and tape robots. Particle collisions are described by a specific dataset that contains information about each collision. One set includes information about several hours of operation of the Collider – one cycle. These data include the date, time of operation of the Collider, the settings with which these experiments took place, as well as descriptions of independent events – collisions collected from various sensors of the detector.

Physicists send a request to work with a certain set of data, they describe the actions that are necessary to obtain results. Often those actions contain multiple tasks that can be executed one after another or in parallel, and also contain work on several datasets at the same time.

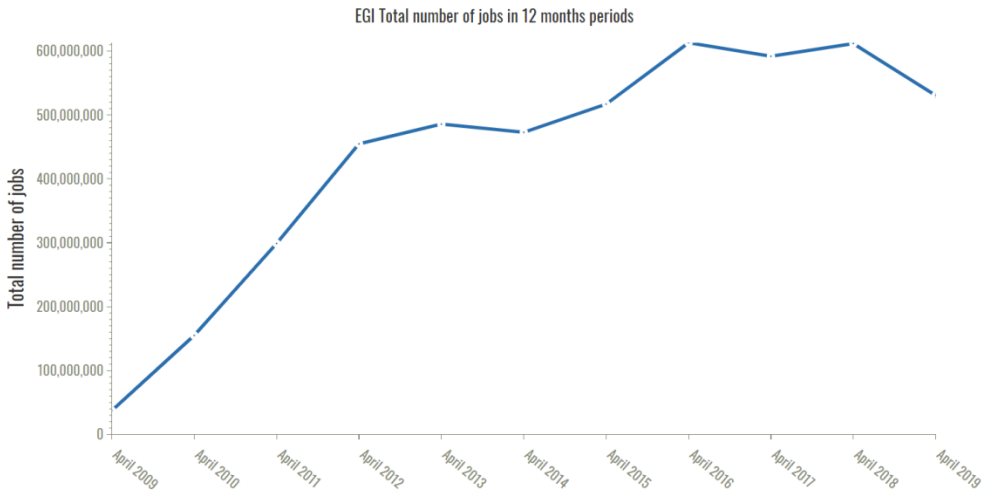

Figure 20 – Number of completed jobs on the nodes of the WLCG over the years.

A task is divided into several jobs, each job consists of several events. Each job contains as many events as one node can handle at a time. The event describes only one encounter, because all data from a dataset is independent from each other.

Firstly, the task divides the dataset into jobs. These jobs are distributed among the nodes of the WLCG and its execution begins. Once all of the events in the job are processed, the data is ready to move on to the next task without waiting for the others. The following task can redistribute the number of required events within the task. And as soon as enough events are ready to start and complete a new task, the task begins its execution.

Therefore, tasks can be performed not only sequentially, but with an overlay that depends on the tasks and their execution time. However, not all of the tasks can be processed in one time. Tasks like comparison or merge of two datasets require two preprocessed datasets so they can be performed only sequentially. The task overlap depends on the parameters of the task chain that we give to the machine learning model.

Usually, tasks that physicists set for processing are standard. We can predict the running time of the current chain by the current chain itself and its parameters and by existing examples of task chains with a certain execution time. In previous studies, there are developed an algorithm that predicts the duration of tasks as well as the duration of the tasks chains by parameters that describe the collision.

Only when all 100% of the jobs are processed the task can be considered successfully completed. Often a task consists of more than 1000 jobs and if something causes errors, then one unprocessed job leads to the fact that the entire task is considered unprocessed. Also, it is important to be able to predict these anomalies and fix bugs. The solution to this problem can be preliminary clustering of job log data.

2 SYSTEM DESIGN

2.1 Cluster analysis

Clustering (or cluster analysis) is the process of dividing a set of objects into groups called clusters. Within each group there should be "similar" objects, and objects of different groups should be as different as possible. The main difference between clustering and classification is that the list of groups is not defined clearly and is decided in the process [7]. The task is to use all available data to predict the correspondence of the sample objects to their classes, forming clusters in the process.

Clustering in data analysis becomes valuable when it is one of the stages of data analysis, built in a complete analytical solution. It is often easier for an analyst to identify groups of similar objects, explore their features, and build a separate model for each group rather than create one common model for all data. This technique is constantly used in marketing to highlight groups of customers, buyers, products and develop a separate strategy for each of them.

All attributes are divided into numeric and categorical types. Numerical attributes can be ordered, while categorical ones cannot. For example, the age attribute is numeric and the color attribute is categorical. Setting values to attributes occurs during measurements by the selected scale type, and this is usually a separate task.

In general, application of cluster analysis is reduced to the following stages:

1. Selecting objects for clustering.
2. Defining the set of variables and features in the sample for evaluation. Normalize values of the variables, if necessary.
3. Calculating the similarity measure between objects.

4. Using the cluster analysis method to create groups of similar objects (clusters).

5. Presenting the results of analysis.

After receiving and analyzing the results, it is recommended to adjust selected metric and clustering method to obtain the optimal result.

2.2 Classification of clustering methods

The classification of clustering algorithms divides into scalable and non-scalable types. Scalability is the most important property of the algorithm, that depends on its computational complexity and software implementation. There is also a more comprehensive definition. The algorithm is called scalable if its memory capacity is constant and the number of records in the database increases linearly.

There are two types of clustering algorithms methods: hierarchical and non-hierarchical. Classical hierarchical algorithms work only with categorical attributes when a complete tree of nested clusters is built. Usually agglomerative methods of constructing cluster hierarchies are used in these cases – they produce a consistent union of the original objects and a corresponding reduction in the number of clusters. Hierarchical algorithms provide a relatively high quality of clustering and do not require specifying the number of clusters in advance.

Non-hierarchical algorithms are based on the optimization of some objective function. It determines the optimal division of the set of objects into clusters. This group includes popular algorithms of the K-means family (k-means, fuzzy c-means, Gustafson-Kessel). In these algorithms we search for spherical or ellipsoidal shaped clusters.

Clustering methods are also divided into clear and fuzzy types. Clear (or disjoint) algorithms assign a cluster number to each sample object, as if each object belongs to only one cluster. Fuzzy (or intersecting) algorithms put a set of real values for each object accordingly, showing the degree of relationship of the object to the clusters. This way, each object refers to each cluster with some probability.

2.3 Clustering Methods

2.3.1 K-means Method

The most popular data clustering algorithm is the k-means method. This is an iterative clustering algorithm based on minimizing the total quadratic deviations of cluster points from the centroids (mean coordinates) of these clusters.

The algorithm includes the following steps:

1. The number of clusters k is set to be formed from the objects of the initial selection.

2. K records are randomly selected to serve as the initial cluster centers. The starting points where the cluster then grows are often called "seeds." Each record is a kind of "embryo" cluster and consists of one element only.

3. For each record in the source sample the closest cluster center is assigned.

4. Centroids are calculated – they are the centers of gravity of clusters. This is done by determining the average for each feature of all records in the cluster. Then old cluster centers shift to its centroid. Thus, centroids become new cluster centers for the next iteration of the algorithm.

5. The algorithm stops when the boundaries of the clusters and the location of the centroids cease to change, when at each iteration in each cluster a set of records remains the same. The k-means algorithm usually finds a set of stable clusters in a few dozens of iterations.

One of the disadvantages of k-means is the lack of a clear criterion for choosing the optimal number of clusters. To solve this problem, a large number of algorithms have been developed already.

2.3.2 Hierarchical clustering algorithms

Among the algorithms of hierarchical clustering there are two main types: ascending and descending algorithms. Descending algorithms operate on a “top-down” basis: at the beginning, all objects are placed in one cluster, which is then divided into smaller clusters. More common are “bottom-up” algorithms, which initially put each object in a separate cluster, and then combine the clusters into larger and larger clusters

until all objects are contained in a single cluster. Thus the system of nested partitions is constructed. The results of those algorithms are usually represented as a dendrogram tree. A classic example of such a tree is the classification of animals and plants. To calculate distances between clusters it is typical to use two distances: single linkage or complete linkage.

2.3.3 DBSCAN density-based clustering method

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular clustering algorithm used in data analysis as a replacement for the k-means method. In this method, clustering means division of a given set of data points (objects) into subgroups, and each of them is homogeneous as far as possible. The base of DBSCAN method is the clustering of certain objects in accordance with their intra-group "connection". To implement the correct clustering procedure, we must specify the criteria for the clustering.

The implementation of the DBSCAN algorithm can be divided into two stages. First of all, from the whole data set D it is necessary to select those points that are surrounded. Then, we performe the following procedure: for each x object in the D dataset we define whether:

1. the current object belongs to any of clusters;

2. the current object is a surrounded point.

If the current object is a surrounded point, then all objects that are reachable in density from the current object are connected to a new cluster. Otherwise, if the object is not a surrounded point and is not reachable in density from any object, then the current object is an outlier.

2.4 Data cleaning

There are the following stages of data cleaning:

1. *Data analysis*. A detailed analysis of the data is required to determine which types of errors and inconsistencies should be removed. In this case data inspection, individual manual data samples or metadata should be used.

2. *Definition of transformation workflow and mapping rules*. At this stage, the number of data sources, their heterogeneity and "contamination" are estimated. Based on this information, data flow diagrams are created to convert many data sources into one, avoiding the creation of Multi-Source merge errors (for example, duplicated records).

3. *Verification*. This includes an assessment of the accuracy and effectiveness of the previous stage (for example, on a small sample of data). A return to step 2 for its re-execution is possible if it is necessary.

4. *Transformation*. In this step we load data into a single repository using transformation rules defined and debugged in steps 2-3 and clear Single-Source level data.

5. *Backflow of cleaned data.* If we now have a cleaned data set in a single storage at step 4, it is recommended to replace similar "dirty" data in the original sources with this "clean" data. This will help not to repeat all the stages of data cleansing transformations in the future.

These stages can be implemented in a variety of ways using existing and specially created methods and technologies. The data analysis phase involves an analysis of the use of metadata, which is generally not enough to assess the quality of data from available sources. Therefore, it is important to analyze real-world data examples by evaluating their characteristics and value signatures. This enables to find relationships between attributes in data schemas from different sources. There are two approaches to solving this problem – Data profiling and Data mining.

Data profiling is focused on the analysis of individual attributes characterized by their specific properties: data type, length, range of values, frequency of occurrence of discrete values, variance, uniqueness, occurrence of "null" values, typical signature of the record (for example, a phone number). It is a set of properties (profile) that allows evaluating various aspects of data quality. Data mining involves finding relationships between multiple attributes in a large dataset.

In addition, existing integrity constraints can be used to complement missing values, correct invalid values or identify duplicates. They are adopted in relational

databases and imposed in addition to the business relationships between attributes. For example, it is known that "Total = Quantity×Unit_Price". All records that do not meet this condition should be examined more closely, corrected or excluded from consideration. To solve single-source problems in data cleaning, including its integration with other data sources, the following steps are implemented.

In the case of extraction of values from free-form attributes (attribute split) we can talk about string values that store several words in a row (for example, address or full name of a person). Here is needed a clear understanding of the position of the attribute part value. We may need even to sort component parts of an attribute.

Validation and correction involves finding data entry errors and correcting them in the most automatic way. For example, using automatic spell checking to avoid spelling errors and typos. The dictionary of geographical names and postal codes should also be used to correct the values of the addresses entered. The dependence of attributes (date of birth – age, Total = Quantity×Unit_Price, etc.) also helps to avoid many errors in the data.

Standardization involves bringing all data into a single, universal format. Examples of such formats are date and time format, case size in writing string values. Text fields should exclude prefixes and suffixes, abbreviations in them should be unified, problems with different encoding should be excluded.

One of the main problems caused by the integration of different data sources (multi-source problems) is the elimination of duplicated records. This stage is performed after the majority of transformations and cleanings. It involves first identifying records that are similar in some sense, and then merging them with combining attributes. Obviously, the solution to this problem in the presence of duplicated records of the primary key is quite simple. If such a uniquely identifying feature is not present, the task of eliminating duplicates is much more complicated and requires the use of fuzzy approaches comparison (proximity) of the records among themselves.

ПРИЛОЖЕНИЕ Б

Таблица 23 – Образец входных данных

	IObytesRead	IObytesReadRate	IObytesWriteRate	IObytesWritten	IOcharRead	IOcharReadRate	IOcharWriteRate	IOcharWritten	actualcorecount	assignedpriority	avgpss	avgrss	avgswap
0	1746132.0	77047.0	2140.0	48508.0	1240508.0	54736.0	1689.0	38295.0	1.0	320	1980901.0	5247065.0	219665.0
1	2179944.0	60713.0	3891.0	139708.0	3770857.0	105022.0	3376.0	121244.0	1.0	320	2091186.0	5463918.0	0.0
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.0	320	3322352.0	18273821.0	0.0
3	1712080.0	64383.0	2479.0	65940.0	1340762.0	50420.0	1914.0	50910.0	1.0	320	1605039.0	4061987.0	1407185.0
4	1599668.0	300396.0	2544.0	13552.0	747106.0	140296.0	2305.0	12277.0	1.0	320	1901975.0	4883290.0	257537.0
5	4160660.0	26425.0	857.0	134956.0	12966104.0	82352.0	636.0	100232.0	1.0	320	2101237.0	5423748.0	41042.0
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	320	NaN	NaN	NaN
7	4967400.0	20924.0	2387.0	566756.0	21421123.0	90233.0	2046.0	485794.0	1.0	320	2143473.0	5532207.0	19360.0
8	1626964.0	476547.0	5862.0	20016.0	853407.0	249968.0	6351.0	21684.0	1.0	320	1946034.0	4932609.0	0.0
9	3036204.0	33785.0	3954.0	355336.0	8510363.0	94700.0	3549.0	318993.0	1.0	320	2121763.0	5513443.0	1727.0
10	1638748.0	5466051.0	1464740.0	439136.0	825750.0	2754296.0	1416233.0	424593.0	1.0	320	481420.0	500330.0	28204.0
11	1712520.0	5712118.0	1542457.0	462436.0	867847.0	2894710.0	1541739.0	462220.0	1.0	320	501231.0	519190.0	10524.0
12	3438328.0	31537.0	4076.0	444464.0	10198631.0	93546.0	3592.0	391623.0	1.0	320	2064492.0	5447511.0	86264.0
13	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.0	320	2343526.0	13068532.0	0.0
14	3358140.0	23063.0	2425.0	353128.0	13372809.0	91842.0	2089.0	304304.0	1.0	320	2115707.0	5510625.0	12454.0
15	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	320	NaN	NaN	NaN
16	1556748.0	2600505.0	3588.0	2148.0	638229.0	1066144.0	9242.0	5532.0	1.0	320	1168113.0	2452275.0	0.0
17	2813548.0	276044.0	56987.0	580840.0	1367472.0	134166.0	55372.0	564377.0	1.0	320	376074.0	396537.0	0.0
18	3023760.0	27738.0	2528.0	275620.0	10108777.0	92734.0	2201.0	239999.0	1.0	320	2111303.0	5503137.0	3462.0
19	3686936.0	47287.0	3695.0	288144.0	7446795.0	95510.0	3276.0	255441.0	1.0	320	2119854.0	5496707.0	3990.0
20	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	8.0	320	3133632.0	17311489.0	0.0
21	2888512.0	19847.0	2545.0	370416.0	13203607.0	90723.0	2192.0	319043.0	1.0	320	2085521.0	5475352.0	43588.0
22	2997636.0	25478.0	2491.0	293160.0	10943926.0	93016.0	2151.0	253138.0	1.0	320	2117409.0	5510823.0	0.0
23	1854012.0	121707.0	2566.0	39096.0	1041623.0	68377.0	2219.0	33816.0	1.0	320	1640781.0	4241162.0	1173097.0

Продолжение таблицы 23

	avgvmem	computingsite	cpu_eff	cpuconsumptiontime	currentpriority	dbTime	endtime	failedattempt	hs06	hs06sec	inputfilebytes
0	8505939.0	MWT2_SL6	0.867353	20486	320	0.98	2018-06-05T16:40:49.0	0	10	245215.0	305588060
1	8616607.0	RAL-LCG2_ES	0.948564	35131	320	13547,00	2018-06-05T16:41:51.0	0	10	370360.0	306456422
2	26879076.0	LRZ-LMU_C2PAP_ES_MCORE	7.426346	51019	322	NaN	2018-06-05T11:39:42.0	0	10	NaN	305588060
3	8556741.0	MWT2_SL6	0.872314	24075	321	0.75	2018-06-07T09:32:16.0	0	10	286536.0	306456422
4	8025891.0	MWT2_SL6	0.831018	4785	320	21551,00	2018-06-07T03:28:21.0	0	10	59780.0	307539367
5	8616881.0	RAL-LCG2_ES	0.873891	141587	320	43560,00	2018-06-07T03:47:03.0	0	10	1620190.0	307539367
6	NaN	BNL_PROD	0.000000	0	320	0.00	2018-06-07T04:21:51.0	0	10	NaN	305588060
7	8724738.0	RAL-LCG2_ES	0.986263	240080	320	20972,00	2018-06-07T12:46:44.0	0	10	2434240.0	305588060
8	7695042.0	BNL_PROD	0.813694	3066	322	57.64	2018-06-07T02:53:11.0	0	10	37610.0	306690486
9	8682943.0	RAL-LCG2_ES	0.964670	89067	320	18295,00	2018-06-06T08:02:44.0	0	10	923290.0	306456422
10	1701840.0	CERN-PROD_UCORE	0.303754	178	5000	27576,00	2018-06-07T04:31:42.0	0	10	9892.0	305588060
11	1719685.0	CERN-PROD_UCORE	0.250000	167	5000	16923,00	2018-06-07T06:40:58.0	0	10	11277.0	306690486
12	8697421.0	RAL-LCG2_ES	0.972027	108765	320	11324,00	2018-06-06T13:30:24.0	0	10	1118950.0	306690486
13	19800958.0	LRZ-LMU_C2PAP_ES_MCORE	3.643593	2474	321	NaN	2018-06-05T03:36:15.0	0	10	NaN	305588060
14	8689772.0	RAL-LCG2_ES	0.987282	147500	320	13516,00	2018-06-06T23:14:27.0	0	10	1494000.0	310471665
15	NaN	BNL_PROD	0.000000	0	322	0.00	2018-06-07T03:06:21.0	0	10	NaN	310471665
16	4146906.0	BNL_PROD	0.465455	384	320	65.52	2018-06-07T12:42:20.0	0	10	8234.0	305588060
17	1472578.0	CERN-PROD_UCORE	0.025188	288	5000	31929,00	2018-06-07T12:52:31.0	0	10	193027.0	306456422
18	8673650.0	RAL-LCG2_ES	0.964358	107982	320	20821,00	2018-06-06T13:30:53.0	0	10	1119730.0	306690486
19	8659941.0	RAL-LCG2_ES	0.962275	77058	321	43499,00	2018-06-06T16:00:25.0	0	10	800790.0	306456422
20	25544090.0	LRZ-LMU_C2PAP_ES_MCORE	7.081167	22334	322	NaN	2018-06-05T11:39:12.0	0	10	NaN	305588060
21	8686245.0	RAL-LCG2_ES	0.960821	143564	320	23377,00	2018-06-06T23:14:51.0	0	10	1494180.0	310471665
22	8685467.0	RAL-LCG2_ES	0.985080	118978	320	12055,00	2018-06-06T15:57:33.0	0	10	1207800.0	306456422

23	8500693.0	MWT2_SL6	0.898744	14317	321	1.00	2018-06-07T06:17:48.0	0	10	165387.0	306456422	

Продолжение таблицы 23

	jobname	jobstatus	maxcpucount	maxdiskcount	maxpss	maxrss	maxvmem	minramcount	nevents	ninputdatafiles	noutputdatafiles
0	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	795341	1472	2145511.0	5610708.0	8656128.0	1961	51	1	1
1	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	2131844.0	5602172.0	8708904.0	1961	189	1	1
2	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	795341	1472	3510403.0	18859192.0	27588388.0	9388	106	1	0
3	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	2152953.0	5609456.0	8656100.0	1961	73	1	1
4	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1476	2057916.0	5439148.0	8632868.0	1961	11	1	1
5	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1476	2143575.0	5539788.0	8716320.0	1961	150	1	1
6	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	closed	795341	1472	NaN	NaN	NaN	1961	0	1	0
7	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	795341	1472	2226791.0	5617304.0	8822032.0	1961	800	1	1
8	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	2163347.0	5622344.0	8647264.0	1961	24	1	1
9	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	2166792.0	5579724.0	8747252.0	1961	500	1	1
10	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	795341	1472	773638.0	782540.0	3072240.0	2000	1000	1	1
11	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	813186.0	821536.0	3071688.0	2000	1000	1	1
12	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	2132785.0	5607432.0	8769956.0	1961	623	1	1
13	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	795341	1472	3198132.0	18464884.0	27584092.0	9388	2	1	0
14	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1482	2156321.0	5602504.0	8752156.0	1961	486	1	1
15	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	closed	825711	1482	NaN	NaN	NaN	1961	0	1	0
16	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	795341	1472	2062296.0	5569348.0	8644456.0	1961	1	1	1
17	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	1035083.0	1101104.0	3545040.0	2000	1000	1	1
18	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	2157387.0	5596092.0	8742100.0	1961	377	1	1
19	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	2158671.0	5602328.0	8731892.0	1961	407	1	1
20	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	795341	1472	3414071.0	18803912.0	27588412.0	9388	2	1	0
21	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1482	2122481.0	5600016.0	8747848.0	1961	514	1	1
22	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	2153280.0	5603248.0	8748020.0	1961	404	1	1
23	mc16_13TeV.423210.Pythia8B_A14_CTEQ6L1_bb_Jpsi...	finished	825711	1474	1897121.0	4956744.0	8655960.0	1961	45	1	1

Продолжение таблицы 23

	nucleus	outputfilebytes	pandaid	queue_time	starttime	timeExe	timeSetup	timeStageIn	timeStageOut	wall_time	workDirSize
0	CERN-PROD	0	3953497359	84310	2018-06-05T10:07:10.0	23269	29	36	0	23619	42018980.0
1	CERN-PROD	0	3954179248	19390	2018-06-05T06:24:35.0	36784	15	46	0	37036	216877210.0
2	CERN-PROD	0	3954385811	12462	2018-06-05T09:45:12.0	0	0	0	0	6870	NaN
3	CERN-PROD	0	3955485399	63807	2018-06-07T01:52:17.0	27296	13	21	0	27599	90077338.0
4	CERN-PROD	0	3954178900	175866	2018-06-07T01:52:23.0	5481	15	31	0	5758	20087385.0
5	CERN-PROD	0	3954179315	20717	2018-06-05T06:46:44.0	161287	83	174	0	162019	179356249.0
6	CERN-PROD	0	3953496446	227578	2018-06-07T01:54:29.0	0	0	0	0	8842	0.0
7	CERN-PROD	0	3953497530	23256	2018-06-04T17:09:40.0	243113	10	47	0	243424	926005412.0
8	CERN-PROD	0	3956075753	19808	2018-06-07T01:50:23.0	3545	23	15	0	3768	36077568.0
9	CERN-PROD	0	3954179243	19350	2018-06-05T06:23:55.0	92078	8	33	0	92329	599353498.0
10	CERN-PROD	555286163	3956484876	5	2018-06-07T04:21:56.0	374	20	111	21	586	7745536.0
11	CERN-PROD	566605134	3956587541	29	2018-06-07T06:29:50.0	374	31	118	27	668	610394112.0
12	CERN-PROD	0	3954179250	19444	2018-06-05T06:25:29.0	111672	5	24	0	111895	738924106.0
13	CERN-PROD	0	3954190708	7489	2018-06-05T03:24:56.0	0	0	0	0	679	NaN
14	CERN-PROD	0	3954179139	16984	2018-06-05T05:44:27.0	149119	11	50	0	149400	585643023.0
15	CERN-PROD	0	3956037724	22760	2018-06-07T01:43:42.0	0	0	0	0	4959	0.0
16	CERN-PROD	0	3953496454	265624	2018-06-07T12:28:35.0	654	7	51	0	825	9980068.0
17	CERN-PROD	562534255	3956705763	63	2018-06-07T09:41:57.0	10685	64	361	67	11434	613259824.0
18	CERN-PROD	0	3954179249	19395	2018-06-05T06:24:40.0	111683	17	56	0	111973	438273610.0
19	CERN-PROD	0	3954806690	3508	2018-06-05T17:45:46.0	79884	9	29	0	80079	480086170.0
20	CERN-PROD	0	3954385818	16145	2018-06-05T10:46:38.0	0	0	0	0	3154	NaN
21	CERN-PROD	0	3954179142	16990	2018-06-05T05:44:33.0	149063	19	66	0	149418	617489423.0
22	CERN-PROD	0	3954179247	19388	2018-06-05T06:24:33.0	120519	11	47	0	120780	485271706.0
23	CERN-PROD	0	3955887262	35436	2018-06-07T01:52:18.0	15661	10	24	0	15930	63862938.0