ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«24» июня 2025 г.
Протокол №2

Рабочая программа дисциплины (модуля)
«Большие данные»

Направление подготовки: *02.03.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика»*

Квалификация (степень) выпускника: *бакалавр*

Форма обучения: *очная*

Срок освоения программы: *4 года*

Год набора: *2025*

Москва
2025

Электронный документ

Содержание

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Большие данные» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по специальности 02.03.01 Математика и компьютерные науки, профиль «Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) «Большие данные» формирует у студентов навыки работы с современными технологиями обработки больших объемов информации, что обеспечивает их конкурентоспособность на рынке труда и способствует развитию инновационных решений на стыке математики, компьютерных наук и прикладных дисциплин.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки бакалавриата по направлению 02.03.01 Математика и компьютерные науки, профиль «Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика» и входит в часть Блока 1, формируемую участниками образовательных отношений.

Дисциплина (модуль) является выборной, изучается на 3 или 4 курсе в 5, 6 или 7 семестре на выбор. Доступна после успешного освоения дисциплин (модулей): «Базы данных», «Алгоритмы и структуры данных II».

Цель изучения дисциплины (модуля): освоение методов сбора, хранения, обработки и анализа масштабных и разнообразных данных для решения прикладных задач в различных областях науки и техники.

Задачи изучения дисциплины (модуля):

— формирование знаний основных понятий и архитектуры распределенных вычислительных систем, концепции работы корпоративных хранилищ данных;

— умение работать с системой очередей Kafka и распределенной базой данных Cassandra, решать задачи с использованием различных хранилищ;

— владение построением автоматизированных потоков данных с использованием различных инструментов.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

— основные концепции работы корпоративных хранилищ данных;

— основы распределенных систем хранения, основные компоненты и методы работы с S3;

— основные компоненты и методы работы с Apache Spark;

— основные компоненты и методы работы с ClickHouse;

— основные компоненты и методы работы с Apache Airflow;

— основные компоненты и методы работы с Kafka;

— базовые принципы оптимизации и проверки качества данных;

— основные концепции в проектировании баз данных и корпоративных хранилищ данных;

— основные концепции и принципы проектирования автоматизированных потоков данных;

уметь:

— решать задачи с использованием S3;

— решать задачи с использованием Apache Spark;

— решать задачи с использованием ClickHouse;

— решать задачи с использованием Apache Airflow;

— решать задачи с использованием Kafka;

— проектировать схемы хранения данных в базах данных и корпоративных хранилищах данных;

— применять базовые принципы оптимизации используемых ресурсов;

— применять базовые принципы проверки качества данных;

владеть:

— построением автоматизированных потоков данных с использованием: S3, Kafka, Apache Spark, ClickHouse, Apache Airflow;

— разработкой и внедрением автоматизированного контроля качества данных;

— оптимизацией используемых ресурсов в автоматизированных потоках данных.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
УК-1.	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1.	Знает методы поиска и анализа информации в области искусственного интеллекта, основные принципы критической оценки источников информации и их релевантности.
		УК-1.2.	Умеет критически оценивать источники информации и синтезировать данные из различных источников для решения задач, применять системный подход к анализу и решению комплексных проблем
		УК-1.3.	Имеет практический опыт работы с современными инструментами и технологиями для обработки информации, формулировании и структурировании задач на основе полученной информации
УК-2.	Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1.	Знает действующие правовые нормы, регулирующие деятельность в области решения задач, основные методы и подходы к определению круга задач
		УК-2.2.	Умеет определять круг задач в рамках поставленной цели, выбирать оптимальные способы решения задач, учитывая имеющиеся ресурсы и ограничения
		УК-2.3.	Имеет практический опыт применения знаний о правовых нормах и ресурсах в реальных ситуациях, разработки и реализации решений в соответствии с установленными ограничениями
ОПК-1.	Способен находить, формулировать и решать актуальные и значимые проблемы прикладной и компьютерной математики	ОПК-1.1.	Знает основные методы и подходы к решению задач прикладной и компьютерной математики, включая алгоритмы, математическое моделирование и теорию оптимизации, а также современные инструменты и технологии, используемые в этой области
		ОПК-1.2.	Умеет анализировать и формулировать математические задачи, применять соответствующие методы и алгоритмы для их решения, а также интерпретировать и представлять результаты в понятной и доступной форме
		ОПК-1.3.	Имеет практический опыт работы над проектами или исследованиями в

			области прикладной и компьютерной математики, включая участие в конкурсах, олимпиадах или научных публикациях, где были решены актуальные и значимые задачи
ПК-1.	Способен определять общие формы и закономерности области машинного обучения	ПК-1.1.	Знает основные теоретические концепции и принципы, относящиеся к области машинного обучения, а также ключевые закономерности и модели, которые помогают в анализе и интерпретации данных
		ПК-1.2.	Умеет проводить систематический анализ области разработки, выявлять и формулировать общие закономерности и тенденции, а также применять методы исследования для получения новых знаний и понимания
		ПК-1.3.	Имеет практический опыт работы в области машинного обучения, включая участие в научных проектах, исследованиях или практических заданиях, где были выявлены и описаны общие формы и закономерности
ПК-2.	Способен решать типовые задачи профессиональной деятельности в области искусственного интеллекта, опираясь на информационную и библиографическую культуру, используя информационно-коммуникационные технологии и учитывая основные требования информационной безопасности	ПК-2.1.	Знает основы информационной и библиографической культуры, а также принципы информационной безопасности и применения информационно-коммуникационных технологий в профессиональной деятельности
		ПК-2.2.	Умеет эффективно использовать информационно-коммуникационные технологии для решения стандартных задач профессиональной деятельности, учитывая требования информационной безопасности
		ПК-2.3.	Имеет опыт работы с информационными ресурсами и технологиями в области искусственного интеллекта, включая соблюдение норм информационной безопасности

3. Тематический план

№п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Контактная работа		Контроль	Самостоятельная работа	
		Лекции	Практические занятия			
1	Основные концепции	3	3		15	Домашние задания
2	Файловые хранилища	3	3		15	Домашние задания
3	Распределённые системы вычисления	3	3		15	Домашние задания
4	Автоматизация ETL процессов	3	3		15	Домашние задания
5	Построение корпоративных хранилищ данных	4	4		15	Домашние задания
6	Clickhouse	3	3		15	Домашние задания
7	Оптимизация	3	3		14	Домашние задания
8	Качество данных	3	3		14	Домашние задания
9	Kafka	3	3		14	Домашние задания
	Экзамен			2		Проект
	Итого:	***28***	***28***	***2***	***132***	
Объем дисциплины (модуля) (в ак. ч.)		***190***				
Объем дисциплины (модуля) (в зач. ед.)		***5***				

4. Содержание дисциплины (модуля)

№ п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Основные концепции	Развитие подходов к хранению и обработке данных. Инструменты и инфраструктура платформ данных. Корпоративные хранилища данных
2	Файловые хранилища	Данные и большие данные. Форматы файлов. Использование объектного хранилища в контексте платформы данных
3	Распределённые системы вычисления	Отделение вычислительного слоя от слоя хранения данных. Пакетная обработка данных.
4	Автоматизация ETL процессов	Основные принципы разработки процессов загрузки данных. Построение потоков из озера данных в структурированные витрины данных.
5	Построение корпоративных хранилищ данных	Многослойная масштабируемая архитектура хранилища данных. Озера данных и Lakehouse. Проектирование витрин данных
6	Clickhouse	Распределенные базы данных shared nothing. Механизмы хранения данных в Clickhouse. Применение Clickhouse для создания процессов обработки данных на SQL
7	Оптимизация	Основы оптимизации в задачах извлечения данных. Ситуации, в которых оптимизация является необходимой. Примеры оптимизации распределенной обработки данных.
8	Качество данных	Измерения качества данных. Метрики качества данных. Интеграция правил проверки данных в процессы ETL
9	Kafka	Архитектура брокера сообщений. Потоковая обработка данных.

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Лесковец, Ю. Анализ больших наборов данных : практическое руководство / Д. Дж. Ульман, Ю. Лесковец, А. Раджараман ; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 500 с. - ISBN 978-5-89818-304-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2102592.

2. Карау, Х. Изучаем Spark. Молниеносный анализ данных : практическое руководство / Х. Карау, Э. Конвински, П. Венделл, З. Матей ; пер. с англ. - 2-е изд. - Москва : ДМК Пресс, 2023. - 305 с. - ISBN 978-5-89818-320-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2102607.

3. Харенслак, Б. Apache Airflow и конвейеры обработки данных : практическое руководство / Б. Харенслак, Д. дэ Руйтер ; пер. с англ. Д. А. Беликова. - Москва : ДМК Пресс, 2022. - 502 с. - ISBN 978-5-97060-970-5. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2155905.

4. Apache Kafka. Потоковая обработка и анализ данных : практическое руководство / Г. Шапира, Т. Палино, Р. Сиварам, К. Петти. - 2-е изд. - Санкт-Петербург : Питер, 2023. - 512 с. - (Серия «Бестселлеры O'Reilly»). - ISBN 978-5-4461-2288-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2123357.

Дополнительная литература:

1. Дьяконов, А.Г. Машинное обучение и анализ данных / А.Г. Дьяконов. — URL: https://github.com/Dyakonov/MLDM_BOOK/blob/main/README.md.

2. Мартишин, С. А. Базы данных: работа с распределенными базами данных и файловыми системами на примере MongoDB и HDFS с использованием Node.js, Express.js, Apache Spark и Scala : учебное пособие / С. А. Мартишин, В. Л. Симонов, М. В. Храпченко. — Москва : ИНФРА-М, 2024. — 235 с. - ISBN 978-5-16-019845-3. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2139860.

3. Стружкин, Н. П. Базы данных: проектирование. Практикум : учебник для вузов / Н. П. Стружкин, В. В. Годин. — Москва : Издательство Юрайт, 2025. — 291 с. — (Высшее образование). — ISBN 978-5-534-00739-8. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/561215.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также

помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

— столами и стульями;
— компьютерной техникой;
— механическими калькуляторами;
— специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование"	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое

Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		
КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное
Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое
Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Большие данные» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, практические занятия, домашние задания, проект, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Участие в семинаре (практическом занятии) – активная работа студента на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре студентам рекомендуется заранее ознакомиться с темой обсуждения, прочитать необходимые материалы и подготовить вопросы. Важно активно слушать и вовлекаться в дискуссию, высказывая свои мнения и аргументируя их. При ответах на вопросы преподавателя стоит быть уверенным, четким и логичным,

опираясь на изученный материал. Также полезно поддерживать диалог с однокурсниками, чтобы обогатить обсуждение и расширить свои знания.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал. Использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Проект – исследовательская работа по курсу и презентация результатов.

Для успешной подготовки к проекту: четко определите цели и задачи проекта, распределите роли и обязанности между участниками, а также установите сроки выполнения каждой части работы. Регулярно проводите встречи для обсуждения прогресса и решения возникающих вопросов.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Бонусные баллы — это оценки, которые студенты могут получить за выполнение дополнительных заданий.

Формат бонусных баллов позволяет студентам улучшить общую оценку по курсу и стимулирует углубленное изучение материала.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Большие данные»

Оценивание уровня учебных достижений, обучающихся по дисциплине (модулю), осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме ***экзамена,*** при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом:

Десятибалльная оценка	**Пятибалльная оценка**	**Общая характеристика результата обучения по дисциплине (модулю)**
10	Отлично	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований,
9	Отлично	
8	Отлично	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		а также решает сложные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	
5	Удовлетворительно	Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	
3	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	
1	Не сдан	

Дисциплина (модуль) «Большие данные» оценивается следующим образом:

Активность	Вес	Описание
Домашние задания	80%	Набор задач по темам недели
Проекты	20%	Исследовательская работа по курсу и презентация результатов

Формула расчёта итоговой оценки по дисциплине (модулю) «Большие данные»: «0,8 × среднее за домашние задания + 0,2 × среднее за проекты».

При изучении дисциплины (модуля) так же возможно получение бонусных баллов.

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1

1. Объясните эволюцию подходов к хранению данных от традиционных баз данных к современным платформам. Приведите 3 примера изменений за последние 20 лет.

2. Опишите роль инструментов инфраструктуры (например, Hadoop, Spark) в платформах данных. Как они решают проблемы масштабируемости?

3. Проанализируйте сценарий: компания хочет перейти от локального хранения к облачному. Какие преимущества и риски вы видите?

4. Сравните корпоративное хранилище данных (EDW) с обычной базой данных. В каких случаях EDW предпочтительнее?

5. Придумайте и опишите гипотетическую платформу данных для малого бизнеса (например, кафе). Какие компоненты вы включите?

6. Рассчитайте потенциальную экономию: если миграция на облачное хранилище снижает затраты на 40% для компании с бюджетом 1 млн долларов, сколько это составит?

Домашнее задание 2

1. Перечислите 5 ключевых инструментов инфраструктуры платформ данных и их основные функции.

2. Как корпоративные хранилища данных интегрируются с инструментами вроде AWS S3 или Google Cloud Storage? Приведите пример архитектуры.

3. Решите задачу: компания обрабатывает 1 ТБ данных ежедневно. Подберите подходящую инфраструктуру и объясните выбор.

4. Обсудите этические аспекты хранения данных в корпоративных хранилищах (например, приватность).

5. Создайте диаграмму простой платформы данных для анализа продаж, используя бесплатные инструменты (например, Draw.io).

6. Проведите анализ: сравните стоимость развертывания EDW на-premises vs. в облаке для 100 пользователей.

Домашнее задание 3

1. Перечислите 5 ключевых инструментов инфраструктуры платформ данных и их основные функции.

2. Как корпоративные хранилища данных интегрируются с инструментами вроде AWS S3 или Google Cloud Storage? Приведите пример архитектуры.

3. Решите задачу: компания обрабатывает 1 ТБ данных ежедневно. Подберите подходящую инфраструктуру и объясните выбор.

4. Обсудите этические аспекты хранения данных в корпоративных хранилищах (например, приватность).

5. Создайте диаграмму простой платформы данных для анализа продаж, используя бесплатные инструменты (например, Draw.io).

6. Проведите анализ: сравните стоимость развертывания EDW на-premises vs. в облаке для 100 пользователей.

Домашнее задание 4

1. Объясните, почему инструменты инфраструктуры необходимы для больших данных. Приведите доказательства из реальных кейсов.

2. Как корпоративные хранилища данных обеспечивают консистентность данных? Опишите механизмы.

3. Решите: если задержка в обработке данных превышает 5 секунд, как оптимизировать инфраструктуру?

4. Анализируйте тренды: как развитие облачных платформ изменило роль IT-специалистов в компаниях.

5. Создайте отчет: оцените преимущества платформ данных для стартапа в сфере AI (например, обработка изображений).

6. Математическая задача: если рост данных экспоненциальный (удваивается ежегодно), рассчитайте объем через 5 лет при старте 1 ТБ

Примерное описание к проекту

Проект по темем: Развертывание Hadoop кластера и разработка MapReduce алгоритма для анализа текстовых данных

Цель проекта:

Научиться развертывать и управлять Hadoop кластером, понять принципы работы MapReduce и разработать собственный алгоритм для обработки и анализа больших объемов текстовых данных.

Задачи проекта

1. Изучить архитектуру Hadoop и основные компоненты кластера.
2. Развернуть локальный или распределенный Hadoop кластер (минимум 3 узла).
3. Изучить основы MapReduce, написать и запустить простую программу на Java или Python.
4. Разработать MapReduce алгоритм для подсчёта частоты слов в большом текстовом файле.
5. Проанализировать результаты работы алгоритма и подготовить отчет по производительности.

Этапы выполнения проекта

1. **Подготовительный этап**
    - Изучение теоретических материалов по Hadoop и MapReduce.
    - Планирование архитектуры кластера и выбор среды для развертывания (виртуальные машины, Docker и т.п.).
2. **Развертывание Hadoop кластера**
    - Установка и настройка Hadoop на выбранных узлах.
    - Проверка состояния кластера, запуск тестовых заданий.
3. **Разработка MapReduce алгоритма**
    - Создание программы для подсчета количества слов.
    - Запуск и отладка MapReduce задачи на Hadoop кластере.
4. **Тестирование и анализ**
    - Запуск алгоритма на большом наборе текстовых данных.
    - Сбор и анализ метрик производительности (время выполнения, использование ресурсов).
5. **Подготовка итогового отчета и презентации**
    - Описание архитектуры кластера, алгоритма и результатов.
    - Выводы и рекомендации по улучшению.

Критерии оценивания проекта

- Полнота и корректность развертывания Hadoop кластера (20%)
- Правильность и качество реализации MapReduce алгоритма (25%)

- Эффективность и корректность обработки данных (20%)
- Качество анализа результатов и выводов (15%)
- Оформление отчета и презентация проекта (20%)

Критерии защиты проекта

- Умение объяснить архитектуру Hadoop кластера и роль каждого компонента.
- Демонстрация работы MapReduce алгоритма на практике, объяснение логики работы.
- Ответы на вопросы по оптимизации и возможным улучшениям.
- Обоснование выбора инструментов и подходов.
- Качество и полнота итогового отчета и презентации.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Какой компонент Hadoop отвечает за распределение задач MapReduce и управление ресурсами кластера? a) HDFS b) Hive c) YARN d) Spark	c	ОПК-1
2.	Как называется язык запросов, используемый в Hive для анализа больших данных?	HiveQL	УК-1
3.	Какой инструмент в экосистеме Hadoop применяется для потоковой передачи и обработки данных в реальном времени?	Kafka	ПК-2
4.	Какой тип базы данных представляет Cassandra?	Key-Value хранилище	УК-2
5.	Как называется основной принцип работы MapReduce, разделяющий задачу на две фазы?	Map и Reduce	ПК-1
6.	Какой из перечисленных инструментов экосистемы Hadoop предназначен для SQL-подобных запросов к большим данным? а) Spark Streaming б) Hive в) Kafka г) Cassandra	б	УК-1
7.	Какой инструмент экосистемы Hadoop используется для микробатчевой потоковой обработки данных?	Spark Streaming	УК-1
8.	Какой высокопроизводительный MPP-движок для обработки данных, не входящий в экосистему Hadoop, часто используется для аналитических запросов?	ClickHouse / GreenPlum	УК-2
9.	Какой распределенный мессенджер-брокер используется для построения конвейеров реального времени и интеграции различных систем?	Kafka	ПК-1
10.	Как называется ключевой компонент Hadoop, отвечающий за распределенное хранение данных больших объемов?	HDFS/Hadoop Distributed File System	ПК-2
11.	Какой общий термин описывает архитектуру, где несколько процессоров работают над разными частями одной задачи одновременно?	Массово-параллельная обработка/MPP	ОПК-1
12.	Какие меры информационной безопасности важны при работе с большими данными? В ответе укажите 3 меры	Шифрование, контроль доступа, аудит логов	ПК-1
13.	Как называется процесс подготовки данных для построения модели машинного обучения?	Предобработка/ очистка/нормализаци	ОПК-1

		я/feature engineering	
14.	Перечислите три слоя корпоративного хранилища данных.	Staging/промежуточный, Core /ядро, Presentation/презентационный	УК-1
15.	Подтверждение личности пользователя для доступа к системам (пароли, биометрия) – это?	Аутентификация	ПК-2
16.	Модель программирования для параллельной обработки больших данных (карта-свертка) – это?	MapReduce	ОПК-1