ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«07» марта 2024 г.
Протокол №1

Рабочая программа дисциплины (модуля) «Системы обработки больших данных»

Направление подготовки: *02.03.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Искусственный интеллект*

Квалификация (степень) выпускника: *бакалавр*

Форма обучения: *очная*

Срок освоения программы: *4 года*

Год набора: *2024*

Москва
2024

Электронный документ

Содержание

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Системы обработки больших данных» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по специальности 02.03.01 Математика и компьютерные науки, профиль Искусственный интеллект, утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) «Системы обработки больших данных» позволяет освоить современные методы и технологии эффективного хранения, обработки и анализа огромных объёмов информации, что критично для принятия обоснованных решений в бизнесе и науке. Эти знания обеспечивают конкурентоспособность специалистов в условиях стремительного роста данных и востребованности Big Data-решений в различных отраслях.

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 Искусственный интеллект и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений.

Дисциплина (модуль) является выборной и доступна для изучения на 3 или 4 курсе в 5, 6, 7, 8 семестрах на выбор.

Цель изучения дисциплины (модуля): формирование навыков проектирования, внедрения и эксплуатации масштабируемых систем для сбора, хранения и анализа больших объёмов данных с использованием современных технологий и инструментов.

Задачи изучения дисциплины (модуля) направлены на формирование у студентов следующий знаний, умений и навыков:

— знание основных понятий и архитектуры распределенных вычислительных систем;

— знание принципов работы MapReduce алгоритмов и их применения для обработки больших данных;

— знание особенностей и возможностей SQL-like интерфейса Hive для анализа данных в Hadoop;

— знание понятия и применения технологии потоковой обработки данных с использованием Spark Streaming;

— умение осуществлять развертывание и управление Hadoop кластером;

— умение работать с системой очередей Kafka и распределенной базой данных Cassandra;

— умение осуществлять написание и оптимизацию MapReduce задач на Hadoop для эффективной обработки данных;

— умение создавать и выполнять запросы в Hive, анализировать и устранять узкие места в запросах;

— умение понимать и применять Spark для быстрого анализа больших объемов данных;

— навык владения инструментами экосистемы Hadoop для создания и оптимизации систем обработки больших данных;

— навык применения алгоритмов на основе MapReduce для решения практических задач в области больших 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 консультировать и использовать фундаментальные знания в области математического анализа, комплексного и функционального анализа алгебры, аналитической	ОПК-1.1.	Знает основные концепции и теории в области математического анализа и смежных дисциплин; методы и подходы, используемые в различных областях математики

	геометрии, дифференциальной геометрии и топологии, дифференциальных уравнений, дискретной математики и математической логики, теории вероятностей, математической статистики и случайных процессов, численных методов, теоретической механики в профессиональной деятельности	ОПК-1.2.	Умеет применять математические методы для решения профессиональных задач
		ОПК-1.3.	Имеет практический опыт разработки и реализации математических моделей в профессиональной деятельности
ПК-1.	Способен формулировать задачи с математической точностью, обосновывать утверждения строго и анализировать полученные результаты в области математики и компьютерных наук	ПК-1.1.	Знает методы и подходы к формулированию задач, а также основные принципы математического доказательства и анализа результатов
		ПК-1.2.	Умеет корректно ставить и формулировать математические задачи, применять строгие методы доказательства и анализировать полученные результаты
		ПК-1.3.	Имеет опыт работы с задачами в области математики и компьютерных наук, включая применение математических методов для решения практических задач
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	Семинары (практические занятия)			
1	Основы системы Hadoop и MapReduce	7	7		32	Подготовка к семинару, Домашние задания
2	Инструменты анализа данных в экосистеме Hadoop	7	7		33	Подготовка к семинару, Домашние задания, Контрольная работа
3	Потоковая обработка данных и инструменты взаимодействия	7	7		32	Подготовка к семинару, Домашние задания
4	NoSQL хранилища и архитектуры данных	7	7		33	Подготовка к семинару, Домашние задания, Контрольная работа
	Зачет с оценкой			4		Проект
	Итого:	***28***	***28***	***4***	***130***	
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	Основы системы Hadoop и MapReduce	Развертывание и управление Hadoop кластером. Основы MapReduce: принципы работы и разработка алгоритмов.
2	Инструменты анализа данных в экосистеме Hadoop	Работа с Hive: запросы, управление данными и оптимизация. Использование MPP решений: ClickHouse и GreenPlum. Введение в Spark: основы и продвинутые функциональности.
3	Потоковая обработка данных и инструменты взаимодействия	Основы применения Spark Streaming для реализации микробатчевой обработки. Принципы работы и интеграция Kafka в сложной архитектуре систем больших данных.
4	NoSQL хранилища и архитектуры данных	Практическое использование Cassandra как Key-Value хранилища. Современные тренды и практики создания архитектур хранилищ 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. Парфенов Ю. П. Постреляционные хранилища данных : учебное пособие для вузов / Ю. П. Парфенов ; под научной редакцией Н. В. Папуловской. — Москва : Издательство Юрайт, 2025. — 97 с. — (Высшее образование). — ISBN 978-5-534-21173-3. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/55950.

2. Гордеев С. И. Организация баз данных : учебник для вузов / С. И. Гордеев, В. Н. Волошина. — 2-е изд., испр. и доп. — Москва : Издательство Юрайт, 2025. — 691 с. — (Высшее образование). — ISBN 978-5-534-21115-3. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/559377.

Дополнительная литература:

1. Советов Б. Я. Базы данных : учебник для вузов / Б. Я. Советов, В. В. Цехановский, В. Д. Чертовской. — 4-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 403 с. — (Высшее образование). — ISBN 978-5-534-18479-2. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/559898.

2. Нестеров С. А. Базы данных : учебник и практикум для вузов / С. А. Нестеров. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 258 с. — (Высшее образование). — ISBN 978-5-534-18107-4. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/560753.

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;
- 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 к 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:		

Электронный документ

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) «Системы обработки больших данных» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, контрольные работы, 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 (аудиторная работа) – 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.

Контрольная работа – письменная работа с набором задач, которые нужно решить за

ограниченное время.

Цель контрольной работы - получить специальные знания по одной или нескольким темам дисциплины (модуля) и продемонстрировать навыки их практического применения.

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.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Системы обработки больших данных»

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 ***зачета с оценкой***, 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**	**Оценка за зачет**	**Общая характеристика результата обучения по дисциплине (модулю)**
10	Отлично	Зачтено	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, а также решает сложные задачи. Студент хорошо знаком с методами исследования, необходимыми
9	Отлично	Зачтено	
8	Отлично	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
			для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Зачтено	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	Зачтено	
5	Удовлетворительно	Зачтено	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	Зачтено	
3	Не сдан	Не зачтено	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	Не зачтено	
1	Не сдан	Не зачтено	

Дисциплина (модуль) «Системы обработки больших данных» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	30%	13	Набор задач по темам недели
Аудиторная работа	10%	13	Активная работа студента на семинаре

Электронный документ

Активность	Вес	Количество	Описание
Контрольные работы	20%	2	Письменная работа с набором задач, которые нужно решить за ограниченное время
Зачет с оценкой	40%	1	Защита итогового проекта

Формула расчёта итоговой оценки по дисциплине (модулю) «Системы обработки больших данных»: «0,3 × среднее за домашние задания + 0,1 × аудиторная работа + 0,2 × среднее за контрольные работы + 0,4 × зачет с оценкой».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание по теме «Основы MapReduce: принципы работы и разработка алгоритмов»

1. Опишите этапы выполнения MapReduce задачи, объясните роль функций Map и Reduce на примере подсчёта слов в тексте.
2. Реализуйте алгоритм MapReduce для подсчёта количества уникальных слов в большом текстовом файле (на выбранном языке или с использованием Hadoop).
3. Напишите MapReduce программу для вычисления среднего значения числовых данных из набора входных файлов.
4. Проанализируйте, как происходит распределение задач Map и Reduce в кластере и опишите, как это влияет на производительность.
5. Разработайте MapReduce алгоритм для нахождения топ-10 самых часто встречающихся слов в наборе данных.

Домашнее задание по теме «Работа с Hive: запросы, управление данными и оптимизация»

1. Создайте таблицу Hive для хранения данных о пользователях (id, имя, возраст, город) и загрузите в неё примерный набор данных.
2. Напишите запрос Hive, который выбирает пользователей из определённого города и сортирует их по возрасту.
3. Реализуйте партиционирование таблицы по городу и сравните время выполнения запроса из предыдущего задания с использованием партиций и без них.
4. Создайте индекс на колонке «возраст» и объясните, как индекс влияет на производительность запросов.
5. Напишите сложный запрос с использованием JOIN двух таблиц (например, пользователи и заказы) и оптимизируйте его с помощью подходящих техник Hive.

Домашнее задание по теме «Основы применения Spark Streaming для реализации микробатчевой обработки»

1. Настройте Spark Streaming приложение, которое читает поток текстовых данных из сокета и выводит количество строк в каждом микробатче.
2. Реализуйте подсчёт количества уникальных слов в каждом 10-секундном микробатче входящего текстового потока.
3. Объясните принцип работы микробатчевой обработки в Spark Streaming и опишите, как можно изменять размер микробатча.
4. Напишите Spark Streaming приложение, которое фильтрует поток сообщений по определённому условию (например, содержит ключевое слово) и сохраняет результат в файл.
5. Проанализируйте, как параметры batchInterval и windowDuration влияют на задержку и пропускную способность приложения, приведите примеры настройки.

Примерные вопросы для подготовки к семинарам

Вопросы к семинару по теме «Использование MPP решений: ClickHouse и GreenPlum»

1. В чем основные архитектурные отличия ClickHouse и GreenPlum как MPP-систем?
2. Какие типы задач и сценарии лучше всего подходят для ClickHouse, а какие — для GreenPlum?
3. Как реализуется распределение данных и параллельная обработка запросов в ClickHouse и GreenPlum?
4. Какие механизмы обеспечения отказоустойчивости и масштабируемости присутствуют в этих системах?
5. Как оптимизировать запросы и структуру данных в ClickHouse и GreenPlum для повышения производительности?

Вопросы к семинару по теме «Принципы работы и интеграция Kafka в сложной архитектуре систем больших данных»

1. Каковы ключевые компоненты архитектуры Apache Kafka и как они взаимодействуют?
2. Какие подходы существуют для обеспечения гарантированной доставки сообщений в Kafka?
3. Как интегрировать Kafka с другими системами обработки данных (например, Spark, Flink, Hive)?
4. Какие стратегии масштабирования и балансировки нагрузки применимы в Kafka кластерах?
5. Как организовать мониторинг и управление производительностью Kafka в продакшн-среде?

Вопросы к семинару по теме «Практическое использование Cassandra как Key-Value хранилища»

1. Как устроена архитектура Apache Cassandra и чем она отличается от традиционных реляционных СУБД?
2. Какие модели данных поддерживает Cassandra и как правильно проектировать таблицы под конкретные запросы?
3. Как реализуется репликация и обеспечение согласованности данных в Cassandra?
4. Какие инструменты и методы используются для мониторинга и оптимизации производительности Cassandra?
5. В каких сценариях Cassandra является предпочтительным решением по сравнению с другими NoSQL базами?

Примерные задания по контрольным работам

Контрольная работа № 1

Задание 1. Опишите архитектуру Hadoop кластера и его основные компоненты.

Задание 2. Перечислите основные этапы запуска MapReduce задачи и опишите роль каждого этапа.

Задание 3. Напишите простой MapReduce алгоритм для подсчёта количества слов в текстовом файле.

Задание 4. Объясните, как происходит распределение задач Map и Reduce в кластере Hadoop.

Задание 5. Опишите процесс развертывания Hadoop кластера на нескольких узлах.

Задание 6. Создайте таблицу в Hive и загрузите в неё данные из CSV-файла.

Задание 7. Напишите Hive-запрос для выборки пользователей старше 30 лет из таблицы.

Задание 8. Объясните принципы партиционирования и индексирования в Hive и их влияние на производительность.

Задание 9. Сравните основные архитектурные особенности ClickHouse и GreenPlum.

Задание 10. Приведите примеры задач, которые лучше решать с помощью ClickHouse, а какие — с помощью GreenPlum.

Задание 11. Опишите основные принципы работы Spark и его отличия от MapReduce.

Задание 12. Напишите пример кода на Spark для подсчёта среднего значения числового столбца в наборе данных.

Задание 13. Объясните, как Spark реализует ленивые вычисления и почему это важно для оптимизации.

Задание 14. Опишите процесс оптимизации запросов в Hive и Spark.

Задание 15. Сравните подходы к масштабированию в Hadoop, ClickHouse и Spark.

Контрольная работа № 2

Задание 1. Опишите архитектуру Spark Streaming и принцип работы микробатчей.

Задание 2. Напишите Spark Streaming приложение для подсчёта количества строк в каждом 5-секундном микробатче.

Задание 3. Объясните, как настроить размер микробатча и как это влияет на задержку и пропускную способность.

Задание 4. Опишите основные компоненты Kafka и их функции.

Задание 5. Объясните, как Kafka обеспечивает гарантированную доставку сообщений.

Задание 6. Приведите пример интеграции Kafka с Spark Streaming.

Задание 7. Опишите стратегии масштабирования Kafka кластера.

Задание 8. Объясните, как организовать мониторинг Kafka в продакшн-среде.

Задание 9. Опишите архитектуру Apache Cassandra и её особенности как key-value хранилища.

Задание 10. Напишите пример создания таблицы в Cassandra и вставки данных.

Задание 11. Объясните, как работает репликация и согласованность данных в Cassandra.

Задание 12. Опишите методы мониторинга и оптимизации производительности Cassandra.

Задание 13. Приведите сценарии, в которых Cassandra предпочтительнее реляционных баз данных.

Задание 14. Объясните современные тренды в построении архитектур хранилищ данных (например, Data Lake, Lakehouse).

Задание 15. Опишите, как комбинировать разные типы хранилищ (MPP, NoSQL, потоковые) в единой архитектуре больших данных.

Примерное описание и критерии оценивания к итоговому проекту

Описание проекта:

В рамках итогового проекта необходимо разработать комплексное решение для обработки и анализа больших данных, демонстрирующее глубокое понимание и практические навыки работы с основными технологиями и инструментами, изученными в ходе курса. Проект должен включать развертывание и управление Hadoop кластером, реализацию MapReduce алгоритмов, использование Hive для анализа данных, применение MPP-решений (ClickHouse или GreenPlum) для эффективной аналитики, а также внедрение Spark для пакетной и потоковой обработки данных. Кроме того, проект должен демонстрировать интеграцию с Kafka для организации потоковой передачи данных и использование Cassandra в качестве NoSQL хранилища. В итоговом решении необходимо продемонстрировать современные архитектурные подходы к построению систем обработки больших данных.

Требования к проекту:

- Развертывание и настройка Hadoop кластера с демонстрацией его работоспособности.
- Разработка и запуск MapReduce задач для решения прикладных задач обработки данных.
- Создание и оптимизация Hive-запросов для анализа данных, включая работу с партиционированием и индексами.
- Использование MPP-решений (ClickHouse или GreenPlum) для выполнения сложных аналитических запросов.
- Реализация пакетной обработки данных с помощью Apache Spark, включая оптимизацию производительности.
- Разработка потокового приложения на Spark Streaming с интеграцией Kafka для обработки данных в реальном времени.
- Использование Cassandra для хранения и управления ключ-значение данными с учётом требований к репликации и согласованности.
- Описание архитектуры решения с обоснованием выбранных технологий и подходов.
- Документирование проекта с подробным описанием этапов разработки, конфигураций и результатов.

Критерии оценивания:

1. **Техническая корректность и полнота решения**
   - Полнота реализации всех основных компонентов согласно требованиям.
   - Корректность и работоспособность развернутого Hadoop кластера.
   - Правильность и эффективность MapReduce алгоритмов.
   - Корректность и оптимизация Hive-запросов.
   - Эффективное применение MPP-решений для аналитики.
   - Корректная реализация пакетной обработки в Spark.
   - Работоспособность потоковой обработки с использованием Spark Streaming и Kafka.
   - Правильное использование Cassandra для хранения данных.
2. **Архитектурное решение и интеграция компонентов**
   - Обоснованность выбора технологий и архитектурных подходов.
   - Грамотная интеграция различных компонентов системы.
   - Соответствие современным практикам построения систем больших данных.
3. **Оптимизация и масштабируемость**
   - Наличие оптимизаций производительности на уровне запросов и обработки данных.
   - Рассмотрение вопросов масштабируемости и отказоустойчивости.
4. **Документирование и презентация проекта**
   - Полнота и качество технической документации.
   - Чёткость описания архитектуры, процессов и результатов.
   - Качество и информативность презентации итогового решения.
5. **Практическая ценность и инновационность**
   - Практическая применимость решения к реальным задачам.
   - Использование современных тенденций и инновационных подходов в обработке больших данных.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Что происходит на этапе Map в модели MapReduce? a) Агрегация данных b) Разбиение данных на блоки c) Преобразование и фильтрация данных d) Запись результатов в базу данных	c	УК-2
2.	Для чего используется Hive в экосистеме Hadoop? a) Для потоковой обработки данных b) Для выполнения SQL-подобных запросов к данным в Hadoop c) Для управления ресурсами кластера d) Для хранения данных в формате Key-Value	b	ПК-1
3.	Что из перечисленного является примером MPP-системы? a) Apache Kafka b) GreenPlum c) Apache Spark d) Apache Cassandra	b	УК-1
4.	Какова основная задача Spark Streaming? a) Хранение больших данных b) Пакетная обработка данных c) Потоковая обработка данных с использованием микробатчей d) Управление ресурсами кластера	c	ОПК-1
5.	Какие основные компоненты входят в экосистему Hadoop?	HDFS, YARN, MapReduce	УК-1
6.	Какую роль в архитектуре больших данных выполняет Apache Kafka? a) Хранение данных b) Распределённый брокер сообщений для передачи потоков данных c) Выполнение SQL-запросов d) Аналитическая обработка данных	b	УК-1
7.	Какая технология Hadoop отвечает за управление ресурсами в кластере?	YARN	УК-1
8.	Как называется функция MapReduce, которая агрегирует промежуточные результаты?	Reduce	ПК-1
9.	Назовите язык запросов, используемый в Hive.	HiveQL	ОПК-1
10.	Какой тип базы данных представляет Cassandra?	NoSQL	ПК-2