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: *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: *2024*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B0C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Москва
2024

Содержание

1. **Краткая характеристика дисциплины (модуля)** 3
2. **Перечень планируемых результатов обучения** 5
3. **Тематический план** 7
4. **Содержание дисциплины (модуля)** 7
5. **Учебно-методическое обеспечение** 8
6. **Материально-техническое обеспечение** 8
7. **Методические и оценочные материалы** 10

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 направлению 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) «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 «Программа двух дипломов НИУ ВШЭ и ЦУ «Прикладная математика и информатика» и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений.

Дисциплина (модуль) доступна для изучения на 3 курсе в 5 или 6 семестре на выбор после успешного освоения дисциплин «Разработка на Rust» или «Разработка на C++» или «Разработка на Java», «Распределенные системы», «Базы данных».

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):

— изучить основные технологии и архитектуры систем обработки больших данных, включая Hadoop, Spark и NoSQL базы данных;

— освоить методы анализа и визуализации больших данных с применением алгоритмов машинного обучения и статистических моделей;

— освоить навыки проектирования и реализации масштабируемых решений для обработки больших данных в облачных и распределенных средах.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

— основные понятия и архитектуры распределенных вычислительных систем;

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

— особенности и возможности SQL-подобного интерфейса Hive для анализа данных в Hadoop;

уметь:

— осуществлять развертывание и управление Hadoop кластером;

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

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

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

владеть:

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

— навыком 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 разработки, 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
ОПК-1.	Способен консультировать и использовать фундаментальные знания в области математического анализа, комплексного и функционального анализа алгебры, аналитической 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-1.1.	Знает основные концепции и теории в области математического анализа и смежных дисциплин; методы и подходы, используемые в различных областях математики
		ОПК-1.2.	Умеет применять математические методы для решения профессиональных задач
		ОПК-1.3.	Имеет практический опыт разработки и реализация математических моделей в профессиональной деятельности
ОПК-4.	Способен находить, анализировать, реализовывать программно и использовать на практике математические алгоритмы, в том числе с применением современных вычислительных систем	ОПК-4.1.	Знает базовые основы современного математического аппарата, связанного с проектированием, разработкой, реализацией и оценкой качества программных продуктов и программных комплексов в различных областях человеческой деятельности

		ОПК-4.2.	Умеет использовать этот математический аппарат в профессиональной деятельности
		ОПК-4.3.	Имеет практический опыт применения современного математического аппарата, связанного с проектированием, разработкой, реализацией и оценкой качества программных продуктов и программных комплексов в различных областях человеческой деятельности
ОПК-6.	Способен разрабатывать алгоритмы и компьютерные программы, пригодные для практического применения	ОПК-6.1.	Знает алгоритмы разработки, компьютерные программы, а также алгоритмы вычислительной математики
		ОПК-6.2.	Умеет разрабатывать математические программные продукты и комплексы с использованием современных технологий программирования
		ОПК-6.3.	Имеет практический опыт разработки интеллектуальных информационных систем для визуализации результатов исследований
ПК-3.	Способен использовать методы математического и алгоритмического моделирования при решении теоретических и прикладных задач	ПК-3.1.	Знает основные методы математического и алгоритмического моделирования, а также их применение для решения теоретических и прикладных задач
		ПК-3.2.	Умеет разрабатывать и применять математические модели и алгоритмы для решения различных задач, анализируя полученные результаты
		ПК-3.3.	Имеет практический опыт использования методов математического и алгоритмического моделирования в реальных проектах или исследованиях

3. Тематический план

№п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары (практические занятия)			
1	Основы системы Hadoop и MapReduce	8	8		30	Кейс
2	Инструменты анализа данных в экосистеме Hadoop	8	8	2	30	Кейс
3	Потоковая обработка данных и инструменты взаимодействия	6	6	2	30	Кейс
4	NoSQL хранилища и архитектуры данных	8	8	2	32	Кейс Проект
	Зачет с оценкой			2		
	Итого:	***30***	***30***	***8***	***122***	
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	Основы Hadoop и MapReduce. Архитектура распределённых вычислительных систем. Паттерны и оптимизация MapReduce задач. Hive и SQL-on-Hadoop для аналитики
2	Инструменты анализа данных в экосистеме Hadoop	Оптимизация запросов Hive и MPP-решения (ClickHouse, Greenplum). Введение в Apache Spark: RDD, DataFrame, Dataset. Продвинутый Spark: Catalyst, Tungsten, MLlib. Потоковая обработка: Spark Streaming и Structured Streaming
3	Потоковая обработка данных и инструменты взаимодействия	Основы Apache Kafka и интеграция потоков данных. NoSQL хранилища: Cassandra модель данных и запросы. Оптимизация Cassandra и масштабирование кластера
4	NoSQL хранилища и архитектуры данных	Архитектуры Data Lake: Lambda, Delta Lake, Iceberg. Построение ETL/ELT пайплайнов в экосистеме Hadoop. Подготовка и реализация проектных решений. Защита проектов и разбор кейсов

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. Волкова, П. А. Статистическая обработка данных в учебно-исследовательских работах : учебное пособие / П.А. Волкова, А.Б. Шипунов. — Москва : ФОРУМ : ИНФРА-М, 2022. — 96 с. — (Высшее образование: Бакалавриат). - ISBN 978-5-00091-710-7. - Текст : электронный. - URL: https://znanium.com/catalog/product/1862854.

2. Мартишин, С. А. Базы данных. Практическое применение СУБД SQL и NoSQL-типа для проектирования информационных систем : учебное пособие / С.А. Мартишин, В.Л. Симонов, М.В. Храпченко. — Москва : ФОРУМ : ИНФРА-М, 2024. — 368 с. — (Высшее образование). - ISBN 978-5-8199-0946-1. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2096940.

Дополнительная литература:

1. 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, 2023. — 235 с. + Доп. материалы [Электронный ресурс]. — (Среднее профессиональное образование). - ISBN 978-5-16-015643-9. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2111334.

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) «Системы обработки больших данных» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, коллоквиумы, кейсы, 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.

Семинар — это форма учебного занятия, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Кейс – практическая работа студентов над реальными или смоделированными задачами, что позволяет студенту применять теоретические знания на практике.

Студент самостоятельно разрабатывает стратегию решения поставленной задачи, что способствует развитию навыков критического мышления и самостоятельного принятия решений. Такой подход помогает подготовить будущих специалистов к реальным вызовам в их профессиональной деятельности.

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, если иное не указано в силлабусе дисциплины:

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 для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
9	Отлично	Зачтено	
8	Отлично	Зачтено	
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
6	Хорошо	Зачтено	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
			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.
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
Кейс	40%	4	Практическая работа студентов над реальными или смоделированными задачами, что позволяет студенту применять теоретические знания на практике
Проект	30%	1	Исследовательская работа по дисциплине (модулю) и презентация результатов
Зачет с оценкой	30%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Формула расчёта итоговой оценки по дисциплине (модулю) «Системы обработки больших данных»: «0,4 × среднее за кейсы + 0,3 × проекты + 0,3 × зачёт с оценкой».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные задания для кейсов

Электронный документ

Основы системы Hadoop и MapReduce

1. **Кейс: Анализ логов веб-сервера.** Компания обрабатывает миллионы записей логов с веб-сервера. Используя MapReduce, разработайте задачу для подсчета количества запросов по IP-адресам и выявления наиболее активных пользователей, оптимизировав ее для снижения сетевых затрат.
2. **Кейс: Обработка текстовых данных.** Библиотека оцифровывает книги и нуждается в индексации слов. Создайте MapReduce-задачу для подсчета частоты слов в корпусе текстов, используя паттерны вроде WordCount, и интегрируйте ее с архитектурой Hadoop для распределенной обработки.
3. **Кейс: Агрегация финансовых транзакций.** Банк анализирует ежедневные транзакции клиентов. Реализуйте MapReduce-программу для агрегации сумм по категориям расходов, оптимизировав shuffle-фаза для минимизации времени выполнения на кластере.
4. **Кейс: Аналитика социальных сетей.** Платформа социальных медиа собирает данные о взаимодействиях пользователей. Используя Hive, напишите SQL-запрос для анализа паттернов связей между пользователями на основе MapReduce-вычислений, с оптимизацией для больших объемов данных.
5. **Кейс: Мониторинг IoT-устройств.** Компания с IoT-датчиками собирает сенсорные данные. Разработайте MapReduce-задачу для выявления аномалий в потоках данных (например, пиковые значения), интегрируя ее с архитектурой распределенных систем для реального времени.

Инструменты анализа данных в экосистеме Hadoop

1. **Кейс: Оптимизация запросов к данным продаж.** Розничная сеть анализирует продажи в Hive. Оптимизируйте запрос для расчета средних продаж по регионам, используя партиционирование и сравнив производительность с MPP-решением вроде Greenplum.
2. **Кейс: Анализ данных с использованием Spark RDD.** Телеком-компания обрабатывает CDR-данные звонков. Создайте Spark-приложение на RDD для фильтрации и агрегации данных по длительности звонков, переходя к DataFrame для оптимизации производительности.
3. **Кейс: Машинное обучение на Spark MLlib.** Интернет-магазин прогнозирует спрос на товары. Используйте MLlib для обучения модели кластеризации на данных о покупках, интегрируя Catalyst для оптимизации запросов и Tungsten для ускорения.
4. **Кейс: Потоковая обработка событий.** Сервис стриминга анализирует просмотры в реальном времени. Реализуйте Spark Streaming для подсчета популярных видео, переходя к Structured Streaming для обработки микробатчей и визуализации трендов.
5. **Кейс: Сравнение MPP и Spark.** Финансовая компания сравнивает ClickHouse и Spark для аналитики транзакций. Напишите запросы в обоих инструментах для расчета волатильности цен, оптимизируя Hive-запросы и оценивая производительность на больших данных.

NoSQL хранилища и архитектуры данных

1. **Кейс: Построение Data Lake с Delta Lake.** Компания строит Data Lake для хранения исторических и потоковых данных. Спроектируйте архитектуру с использованием Delta Lake для ACID-транзакций, интегрируя ETL-пайплайн для загрузки данных из различных источников.
2. **Кейс: Миграция на Iceberg.** Стартап мигрирует от традиционных хранилищ к Iceberg. Реализуйте ETL-процесс для партиционирования и версионирования данных о пользователях, обеспечивая совместимость с Hadoop-экосистемой.
3. **Кейс: Lambda-архитектура для аналитики.** Платформа e-commerce анализирует поведение покупателей. Постройте Lambda-архитектуру с NoSQL (HBase) для batch-обработки и потоковой аналитики, включая ETL для очистки данных.

4. **Кейс: Реализация проектного решения.** Команда разрабатывает систему рекомендаций. Подготовьте проектное решение с использованием ELT-пайплайнов в Hadoop для обработки данных отзывов, включая защиту проекта с демонстрацией кейсов масштабирования.

5. **Кейс: Разбор кейсов и защита.** Студенческая группа анализирует реальный кейс обработки больших данных в здравоохранении. Разработайте архитектуру Data Lake с Iceberg, реализуйте ETL/ELT для анонимизации данных и подготовьте презентацию с разбором преимуществ и вызовов.

Примерное описание задания и критерии оценивания к 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	Укажите основную архитектуру распределённых вычислительных систем в Hadoop.	HDFS / HDFS-архитектура / HDFS (варианты: HDFS, Hadoop Distributed File System)	УК-1
2	Укажите паттерн оптимизации MapReduce задач для снижения сетевого трафика.	Combiner / Комбайнер / Combiner (варианты: Combiner, комбайнер)	УК-1
3	Укажите инструмент SQL-on-Hadoop для аналитики в Hive.	Hive / Apache Hive / Hive (варианты: Hive, Apache Hive)	УК-1
4	Укажите метод поиска релевантных источников в экосистеме Hadoop.	GitHub / Репозитории / GitHub (варианты: GitHub, репозитории, открытые источники)	УК-1
5	Укажите основную концепцию математического анализа в MapReduce.	Функционал / Функциональный анализ / Функционал (варианты: функционал, функциональный анализ)	ОПК-1
6	Укажите математический метод для оценки вероятностей в больших данных.	Теория вероятностей / Вероятностные модели / Теория вероятностей (варианты: теория вероятностей, вероятностные модели)	ОПК-1
7	Укажите подход к реализации математических моделей в Spark.	MLlib / Машинное обучение / MLlib (варианты: MLlib, машинное обучение в Spark)	ОПК-1
8	Укажите математический алгоритм для оптимизации запросов в ClickHouse.	Индексирование / Индексы / Индексирование (варианты: индексирование, индексы, оптимизация индексов)	ОПК-4
9	Укажите базовый математический аппарат для RDD в Spark.	Линейная алгебра / Матричные операции / Линейная алгебра (варианты: линейная алгебра, матричные операции)	ОПК-4
10	Укажите алгоритм для оценки качества программных продуктов в DataFrame.	Катализатор / Catalyst / Catalyst (варианты: Catalyst, катализатор оптимизации)	ОПК-4
11	Укажите алгоритм для потоковой обработки в Spark Streaming.	Окна / Временные окна / Окна (варианты: окна, временные	ОПК-4

		окна)	
12	Укажите алгоритм разработки программ для интеграции Kafka.	Потоковый процессор / Stream processor / Потоковый процессор (варианты: потоковый процессор, stream processor)	ОПК-6
13	Укажите метод программирования для запросов в Cassandra.	CQL / Cassandra Query Language / CQL (варианты: CQL, Cassandra Query Language)	ОПК-6
14	Укажите технологию для визуализации результатов в экосистеме Hadoop.	Tableau / BI-инструменты / Tableau (варианты: Tableau, BI-инструменты, визуализация)	ОПК-6
15	Укажите алгоритм для масштабирования кластера Cassandra.	Репликация / Репликация данных / Репликация (варианты: репликация, репликация данных)	ОПК-6
16	Укажите архитектуру для построения ETL-пайплайнов.	Lambda / Lambda-архитектура / Lambda (варианты: Lambda, Lambda-архитектура)	ОПК-6
17	Укажите метод математического моделирования в Delta Lake.	ACID / Транзакции / ACID (варианты: ACID, транзакции, атомарность)	ПК-3
18	Укажите алгоритм для решения задач в Iceberg.	Партиционирование / Партиции / Партиционирование (варианты: партиционирование, партиции)	ПК-3
19	Укажите подход к анализу результатов в проектных решениях Hadoop.	A/B-тестирование / Эксперименты / A/B-тестирование (варианты: A/B-тестирование, эксперименты)	ПК-3
20	Укажите метод алгоритмического моделирования для защиты проектов.	Шифрование / Криптография / Шифрование (варианты: шифрование, криптография)	ПК-3