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*

**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 «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, как дисциплина по выбору.

Дисциплина (модуль) является выборной и доступна для изучения на 2, 3 или 4 курсе в 4, 5, 6, 7 или 8 семестре 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 |
| 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-6 | Способен разрабатывать алгоритмы и компьютерные программы, пригодные | ОПК-6.1. | Знает алгоритмы разработки, компьютерные программы, а также алгоритмы вычислительной математики в области |

| | | | |
|---|---|---|---|
| | для практического применения | | искусственного интеллекта |
| | | ОПК-6.2. | Умеет разрабатывать математические программные продукты и комплексы с использованием современных технологий программирования в области искусственного интеллекта |
| | | ОПК-6.3. | Имеет практический опыт разработки интеллектуальных информационных систем для визуализации результатов исследований в области искусственного интеллекта |
| 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 |

## 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 | Кейс Коллоквиум |
| 2 | Инструменты анализа данных в экосистеме Hadoop | 7 | 7 | | 33 | Кейс Коллоквиум |
| 3 | Потоковая обработка данных и инструменты взаимодействия | 7 | 7 | | 32 | Кейс Коллоквиум |
| 4 | NoSQL хранилища и архитектуры данных | 7 | 7 | | 33 | Кейс Коллоквиум |
| | ***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 | 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. 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.

2. 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.

3. 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.

4. Ын, А. Теоретический минимум по Big Data. Всё что нужно знать о больших данных : практическое руководство / А. Ын, К. Су. - Санкт-Петербург : Питер, 2020. - 208 с. - (Серия «Библиотека программиста»). - ISBN 978-5-4461-1040-7. - Текст : электронный. - URL: https://znanium.com/catalog/product/1760820.

***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.

3. 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.

4. 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.

## 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.

*Коллоквиум* – устные ответы на вопросы, список которых известен студенту заранее.

В процессе подготовки к коллоквиуму необходимо проанализировать учебные материалы, ознакомившись с лекциями, учебниками и дополнительными источниками, акцентируя внимание на ключевых темах. Рекомендуется создать структурированные конспекты, выделяя основные идеи, термины и формулы.

*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 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 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, |
| 4 | Удовлетворительно | Зачтено | |

| Десятибалльная оценка | Пятибалльная оценка | Оценка за зачет | Общая характеристика результата обучения по дисциплине (модулю) |
|---|---|---|---|
| | | | касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования. |
| 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 | Описание |
|---|---|---|
| Кейс | 40% | Практическая работа студентов над реальными или смоделированными задачами, что позволяет студенту применять теоретические знания на практике |
| Коллоквиумы | 30% | Устные ответы на вопросы, список которых известен студенту заранее |
| Проект | 30% | Защита итогового проекта |

**Формула расчёта итоговой оценки по дисциплине (модулю) «Системы обработки больших данных»:** «0,4 × среднее за кейсы + 0,3 × среднее за коллоквиумы + 0,3 × проект»..

**Текущий контроль успеваемости обучающихся по дисциплине (модулю)**

**Примерные задания для кейсов**

**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.

## Примерные заданий для коллоквиума

## Основы системы Hadoop и MapReduce

1. Объясните основные компоненты архитектуры Hadoop (HDFS, YARN, MapReduce) и их роли в распределенной обработке данных. Приведите пример, как HDFS обеспечивает отказоустойчивость.
2. Компания анализирует логи веб-сервера с миллиардами записей. Разработайте MapReduce-задачу для подсчета уникальных посетителей по дням, оптимизировав ее для минимизации сетевых затрат в фазе shuffle.
3. Спроектируйте кластер Hadoop для обработки 1 ТБ данных в день. Укажите конфигурацию узлов, выбор файловой системы и стратегию репликации данных.
4. Опишите паттерн MapReduce для задачи WordCount. Как бы вы модифицировали его для обработки больших текстовых файлов с учетом стоп-слов и лемматизации?

5. Банк обрабатывает транзакции клиентов. Создайте MapReduce-программу для расчета среднего баланса по регионам, интегрируя ее с Hive для последующего SQL-анализа.

6. Сравните MapReduce с традиционными моделями параллельных вычислений (например, MPI). В каких сценариях MapReduce более эффективен, и почему?

7. Платформа социальных медиа собирает данные о постах. Реализуйте MapReduce для выявления трендовых хэштегов, оптимизируя для обработки потоковых данных в реальном времени.

8. Разработайте ETL-пайплайн с использованием MapReduce для очистки и агрегации данных из CSV-файлов в HDFS. Укажите шаги и потенциальные bottlenecks.

9. Как Hive упрощает аналитику на Hadoop? Приведите пример SQL-запроса в Hive для агрегации данных по категориям, эквивалентного MapReduce-задаче.

10. Компания IoT анализирует сенсорные данные. Создайте MapReduce-задачу для детекции аномалий (например, пиковых значений), интегрируя с YARN для управления ресурсами.

11. Объясните роль YARN в экосистеме Hadoop. Как он управляет ресурсами для MapReduce-задач, и какие оптимизации возможны для высоконагруженных кластеров?

12. Библиотека оцифровывает книги. Разработайте MapReduce для индексации слов и поиска синонимов, используя паттерны вроде Inverted Index.

13. Спроектируйте отказоустойчивую архитектуру Hadoop для обработки данных в облаке (AWS EMR). Укажите стратегии бэкапа и восстановления.

14. Как оптимизировать MapReduce-задачи для больших объемов данных? Приведите примеры техник, таких как combiner и partitioner.

15. Финансовая компания прогнозирует риски. Реализуйте MapReduce для расчета волатильности акций на основе исторических данных, интегрируя с Hive для визуализации результатов.

**Инструменты анализа данных в экосистеме Hadoop**

1. Объясните разницу между RDD, DataFrame и Dataset в Apache Spark. В каком случае DataFrame предпочтительнее для аналитики больших данных?

2. Розничная сеть анализирует продажи. Оптимизируйте Hive-запрос для расчета топ-10 продуктов по регионам, используя партиционирование и сравнив с Greenplum.

3. Разработайте Spark-приложение на RDD для фильтрации и агрегации логов сервера, переходя к DataFrame для оптимизации производительности.

4. Как Catalyst оптимизирует запросы в Spark? Приведите пример логического плана для SQL-запроса и его физической реализации.

5. Телеком-компания обрабатывает CDR-данные. Создайте Spark Streaming-задачу для подсчета трафика в реальном времени, интегрируя Structured Streaming.

6. Сравните ClickHouse и Hive по производительности для OLAP-запросов. Когда выбрать MPP-решение вместо Hadoop-экосистемы?

7. Интернет-магазин прогнозирует спрос. Используйте MLlib для кластеризации покупок, оптимизируя с Tungsten для ускорения вычислений.

8. Спроектируйте пайплайн для потоковой аналитики с Spark Streaming, включая обработку микробатчей и интеграцию с Kafka.

9. Как Tungsten улучшает производительность Spark? Объясните роль офф-хип памяти и векторизации в обработке данных.

10. Финансовая компания анализирует транзакции. Напишите запросы в Hive и Spark для расчета скользящих средних, оценив производительность на больших наборах данных.

11. Опишите процесс машинного обучения в MLlib. Как интегрировать модель кластеризации с DataFrame для предсказаний?

12. Сервис стриминга отслеживает просмотры. Реализуйте Structured Streaming для выявления популярных видео, оптимизируя для низкой латентности.

13. Разработайте сравнительный анализ MPP (Greenplum) и Spark для обработки 100 ТБ данных. Укажите метрики производительности и рекомендации.

14. Как оптимизировать Hive-запросы с использованием индексов и материализованных представлений? Приведите пример для запроса агрегации.

15. Аналитическая платформа обрабатывает сенсорные данные. Создайте Spark-приложение для anomaly detection с MLlib, интегрируя Catalyst для оптимизации.

**NoSQL хранилища и архитектуры данных**

1. Объясните архитектуру Lambda-архитектуры. Как она сочетает batch- и stream-обработку для аналитики больших данных?

2. Компания строит Data Lake. Спроектируйте архитектуру с Delta Lake для обработки исторических и потоковых данных, обеспечивая ACID-транзакции.

3. Разработайте ETL-пайплайн в Hadoop для загрузки данных из RDBMS в Iceberg, включая партиционирование и версионирование.

4. Чем отличается Delta Lake от Iceberg в управлении метаданными? Приведите пример использования для версионирования схем.

5. Стартап мигрирует на NoSQL. Реализуйте ELT-процесс с Iceberg для анонимизации пользовательских данных, интегрируя с Hadoop.

6. Как построить Data Lake с использованием Lambda-архитектуры? Опишите роли batch- и speed-слоев в обработке данных.

7. Платформа e-commerce анализирует поведение. Постройте Lambda-архитектуру с HBase для batch-аналитики и потоковой обработки.

8. Спроектируйте проектное решение для системы рекомендаций с ETL/ELT в Hadoop, включая защиту с демонстрацией масштабирования.

9. Как Iceberg обеспечивает совместимость с экосистемой Hadoop? Приведите пример интеграции с Spark для аналитики.

10. Здравоохранение анализирует медицинские данные. Разработайте Data Lake с Delta Lake для анонимизации и ETL, подготовив презентацию с разбором кейсов.

11. Опишите процесс подготовки проектного решения в Hadoop. Какие этапы включают анализ требований и реализацию?

12. Команда разрабатывает аналитику для IoT. Реализуйте архитектуру с Iceberg для обработки больших объемов данных, включая защиту проекта.

13. Разработайте разбор кейса для миграции на Data Lake: от RDBMS к Iceberg, с оценкой преимуществ и вызовов.

14. Как защищать проекты в Hadoop? Приведите пример презентации с демонстрацией ETL-пайплайна и метрик производительности.

15. Финансовая компания строит Data Lake. Создайте Lambda-архитектуру для аналитики транзакций, интегрируя Delta Lake и разбор реальных кейсов масштабирования.

**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. | Какие основные компоненты входят в экосистему Hadoop?<br>a) HDFS, MapReduce, Spark<br>b) HDFS, YARN, MapReduce<br>c) Kafka, Cassandra, Hive<br>d) YARN, Spark, ClickHouse | b | УК-1 |
| 2. | Что происходит на этапе Map в модели MapReduce?<br>a) Агрегация данных<br>b) Разбиение данных на блоки<br>c) Преобразование и фильтрация данных<br>d) Запись результатов в базу данных | c | ОПК-6 |
| 3. | Для чего используется Hive в экосистеме Hadoop?<br>a) Для потоковой обработки данных<br>b) Для выполнения SQL-подобных запросов к данным в Hadoop<br>c) Для управления ресурсами кластера<br>d) Для хранения данных в формате Key-Value | b | ПК-1 |
| 4. | Что из перечисленного является примером MPP-системы?<br>a) Apache Kafka<br>b) GreenPlum<br>c) Apache Spark<br>d) Apache Cassandra | b | УК-1 |
| 5. | Какова основная задача Spark Streaming?<br>a) Хранение больших данных<br>b) Пакетная обработка данных<br>c) Потоковая обработка данных с использованием микробатчей<br>d) Управление ресурсами кластера | c | ОПК-1 |
| 6. | Какую роль в архитектуре больших данных выполняет Apache Kafka?<br>a) Хранение данных<br>b) Распределённый брокер сообщений для передачи потоков данных<br>c) Выполнение SQL-запросов<br>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 | ОПК-6 |