

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ
ФЕДЕРАЦИИ
БУРЯТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ ДОРЖИ БАНАЗАРОВА
ИНСТИТУТ МАТЕМАТИКИ, ФИЗИКИ И КОМПЬЮТЕРНЫХ НАУК
КАФЕДРА ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ И ИНФОРМАТИКИ

Утверждена на заседании
Ученого совета ИМФКН
«__»_____ 202__ г.
Протокол № __

Рабочая программа дисциплины
Технологии сбора и обработки больших данных

Направление подготовки / специальность
09.04.02 Информационные системы и технологии
Профиль

Проектирование, разработка и эксплуатация информационных систем

Квалификация (степень) выпускника
Магистр

Форма обучения
Очная

Улан-Удэ
2025

Пояснительная записка

Цели освоения дисциплины

сформировать у студентов системное представление о методах и инструментах обработки больших объемов данных, обеспечить практические навыки работы с современными платформами и технологиями Big Data для эффективного извлечения, хранения, обработки и анализа данных с учетом требований масштабируемости, производительности и безопасности.

12 практических , 168 СРС

Место дисциплины в структуре образовательной программы

Дисциплина Б1.В.ДВ.01.01 Технологии сбора и обработки больших данных входит в часть, формируемую участниками образовательных отношений, учебного плана 09.04.02 Информационные системы и технологии и является дисциплиной по выбору Б1.В.ДВ.01 .

Планируемые результаты обучения по дисциплине и индикаторы достижения компетенций.

УК-1. Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий

УК.М-1.2 определяет пробелы в информации, необходимой для решения проблемной ситуации, и проектирует процессы по их устранению

УК-2. Способен управлять проектом на всех этапах его жизненного цикла

УК.М-2.3 разрабатывает план реализации проекта с учетом возможных рисков реализации и возможностей их устранения, планирует необходимые ресурсы

ПК-1. Способен планировать работы в проектах в области ИТ малого и среднего уровня сложности

ПК.М-1.2. Анализирует входные данные проектов в области ИТ малого и среднего уровня сложности

В результате освоения дисциплины студент должен:

Знать:

- Основные понятия и принципы работы с большими данными
- Архитектуру и компоненты экосистемы Hadoop и Spark
- Основные типы систем хранения данных (HDFS, NoSQL базы)
- Принципы MapReduce и основы работы с Apache Spark
- Основы потоковой обработки данных и инструменты для этого (Kafka, Spark Streaming)
- Методы оптимизации и масштабирования обработки данных
- Основы безопасности и управления доступом в системах Big Data

Уметь:

- Устанавливать и настраивать среды для обработки больших данных
- Работать с Hadoop и Spark: запускать задачи MapReduce и Spark-приложения
- Загружать, очищать и предварительно обрабатывать большие объемы данных
- Создавать потоковые приложения для обработки данных в реальном времени
- Писать ETL-процессы с использованием технологий больших данных
- Анализировать и визуализировать результаты обработки данных
- Реализовывать базовые задачи машинного обучения с помощью Spark MLlib

Владеть:

- Навыками работы с кластерами Big Data (установка, настройка, мониторинг)
- Практическими навыками программирования на Scala, Python или Java для обработки больших данных
- Умением оптимизировать производительность Hadoop и Spark приложений
- Использованием инструментов для построения и автоматизации конвейеров обработки данных

Планируемые результаты освоения образовательной программы:

Объем дисциплины в зачетных единицах с указанием количества часов, выделенных на контактную работу обучающихся с преподавателем и на самостоятельную работу обучающихся

Общая трудоемкость дисциплины составляет 5 зачетные единицы, 180 часа.

№	Название разделов дисциплины	Лабораторная работа	Самостоятельная работа
Семестр 3		12	168
1	Введение большие данные	4	90
2	Обработка данных	8	78

Тематическое планирование курса

Темы

Введение большие данные

Семестр 3

Введение. Обзор экосистемы Hadoop и Spark

Лабораторная работа. 1(0) ч. Обзор экосистемы Hadoop и Spark

Лабораторная работа. 1(0) ч. Установка и настройка среды для обработки больших данных

Самостоятельная работа. 40(0) ч. Экосистемы Hadoop и Spark.

Системы хранения данных

Лабораторная работа. 1(0) ч. Работа с HDFS

Лабораторная работа. 1(0) ч. NoSQL базы данных: HBase, Cassandra

Самостоятельная работа. 50(0) ч. Распределённые файловые системы. Объектные хранилища. Колонковые базы данных NoSQL базы данных. Системы хранения на основе блоков. Технологии кеширования и in-memory хранилища.

Обработка данных

Семестр 3

Обработка данных с помощью Apache Hadoop

Лабораторная работа. 2(0) ч. Обработка данных с помощью Hadoop MapReduce

Самостоятельная работа. 20(0) ч. Сравнение подходов обработки данных MapReduce vs Apache Spark на кластере Hadoop.

Обработка данных с Apache Spark

Лабораторная работа. 2(0) ч. Работа с RDD. Использование DataFrame

Самостоятельная работа. 20(0) ч. Изучение архитектуры Apache Spark и ее компонентов (Driver, Executors, RDD, DataFrame< DataSet)

Загрузка и предварительная обработка данных

Лабораторная работа. 2(0) ч. Загрузка данных в больших объемах. Предварительная обработка данных

Самостоятельная работа. 18(0) ч. Исследование способов загрузки данных и методов их предварительной обработки с использованием Apache Spark для подготовки к последующему анализу

Анализ потоковых данных

Лабораторная работа. 2(0) ч. Apache Kafka. Spark Streaming

Самостоятельная работа. 20(0) ч. Реализация потоковой обработки данных с использованием Apache Kafka и Spark Streaming: настройка, интеграция и обработка потоков в реальном времени

Семестр	Контрольные точки	Баллы
3	Текущий контроль в разделе «Введение большие данные»	
	Контрольные вопросы	10
	проверка выполнения практической работы	20
3	Текущий контроль в разделе «Обработка данных »	
	Контрольные вопросы	10
	проверка выполнения практической работы	20
3	Зачет	
	Защита проекта	40
Итого за семестр 3:		100

Учебно-методическое и информационное обеспечение учебного процесса

Образовательные технологии (в том числе на занятиях, проводимых в интерактивных формах).

Учебно-методические материалы, в том числе методические указания для обучающихся по освоению дисциплины

Оценочные средства

По данной дисциплине разработаны оценочные средства, критерии их оценивания, а также методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций.

- [1057583_fos_ТехнСбораОбрБольшДанных_magistr — копия — копия.doc](#)

Список литературы

Перечень основной и дополнительной литературы, необходимой для освоения дисциплины.

Основная

1. [Большие данные. Big Data](#): учебник для вузов/Макшанов А. В., Журавлев А. Е., Тындыкарь Л. Н.; Журавлев А. Е., Тындыкарь Л. Н.. —Санкт-Петербург: Лань, 2023. —188 с.
Режим доступа: <https://e.lanbook.com/book/322664>

Дополнительная

1. [Цифровизация гражданского оборота: big data в механизме гражданско-правового регулирования \(цивилистическое исследование\)](#)/Василевская Л. Ю., Подузова Е. Б., Тасалов Ф. А., Василевская Л. Ю.. —, Т. 5: Цифровизация гражданского оборота: big data в механизме гражданско-правового регулирования (цивилистическое исследование). Том 5, Т. 5. —2023. —344 с.
Режим доступа: <https://e.lanbook.com/book/298427>
2. [Интеллектуальное право в условиях развития технологии Big Data. База данных как объект интеллектуальных и иных прав](#): монография/Войниканис Е. А., Кольздорф М. А., Корнеев В. А., Ульянова Е. В., Шебанова Н. А.. —Москва: Проспект, 2022. —177 с.
Режим доступа: <https://e.lanbook.com/book/298049>

Перечень ресурсов информационно-коммуникационной сети «Интернет», необходимых для освоения дисциплины

Официальная документация и сайты проектов:

- Apache Hadoop: <https://hadoop.apache.org/>
- Apache Spark: <https://spark.apache.org/>
- Apache Kafka: <https://kafka.apache.org/>
- Apache Flink: <https://flink.apache.org/>
- Apache Cassandra: <https://cassandra.apache.org/>

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем (при необходимости)

- Apache Hadoop (HDFS, MapReduce)
- Apache Spark
- Apache Flink
- Apache Kafka (система обработки потоковых данных)
- Apache Cassandra (распределённая база данных)

Языки программирования и среды разработки:

- Python (с библиотеками для анализа данных: Pandas, NumPy, PySpark)
- Java и Scala (основные языки для Apache Spark и Hadoop)
- SQL (работа с базами данных и обработка данных)

Инструменты для визуализации данных:

- Tableau
- Power BI
- Apache Superset
- Jupyter Notebook / JupyterLab (интерактивные ноутбуки)

Среды разработки и управления проектами:

- IntelliJ IDEA, Eclipse (IDE для разработки на Java/Scala)
- Visual Studio Code

- Git (системы контроля версий)
- Docker (контейнеризация приложений)

Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Аудитория 0417

Корпус:главный

Назначение аудитории:учебная аудитория для проведения занятий лекционного типа, занятий семинарского типа, курсового проектирования (выполнения курсовых работ), групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации

Число посадочных мест:19

Площадь (кв. м):55.4

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГБОУ ВО «БУРЯТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ ДОРЖИ БАНЗАРОВА»
Институт математики, физики и компьютерных наук
Кафедра вычислительной техники и информатики

**Фонд оценочных средств по дисциплине
Технологии сбора и обработки больших данных**

Направление подготовки/ специальность

09.04.02– Информационные системы и технологии

Профиль подготовки /специализация

Проектирование, разработка и эксплуатация информационных систем

Квалификация (степень) выпускника

Магистр

Форма обучения

очная

Улан-Удэ
2025

Паспорт фонда оценочных средств

УК-1.2 - определяет пробелы в информации, необходимой для решения проблемной ситуации, и проектирует процессы по их устранению

УК-2.3 – разрабатывает план реализации проекта с учетом возможных рисков реализации и возможностей их устранения, планирует необходимые ресурсы

ПК-1.2 - Анализирует входные данные проектов в области ИТ малого и среднего уровня сложности

№	Контролируемые разделы, темы, модули	Формируемые компетенции	Этапы формирования	Оценочные средства	
				Вид	Количество
1.	Введение большие данные	УК-1.2, УК-2.3	3 семестр	Опрос, проверка выполнения практических работ	1
2.	Обработка данных	УК-1.2, УК-2.3, ПК-1.2	3 семестр	Опрос, проверка выполнения практических работ	1

Описание показателей и критериев оценивания уровня приобретенных компетенций на различных этапах их формирования

Результаты обучения	Уровень сформированности компетенций	Показатели оценивания компетенций	Шкала оценивания
Знает: - основные понятия и принципы работы с большими данными; - архитектуру и компоненты экосистемы Hadoop и Spark;- Основные типы систем хранения данных (HDFS, NoSQL базы); - принципы MapReduce и основы работы с Apache Spark; - основы потоковой обработки данных и инструменты для этого (Kafka, Spark Streaming); - методы	Пороговый уровень (как обязательный для всех студентов)	Знает: - основные понятия и принципы работы с большими данными; - архитектуру и компоненты экосистемы Hadoop и Spark; - основные типы систем хранения данных (HDFS, NoSQL базы); Умеет: - устанавливать и настраивать среды для обработки больших данных; - работать с Hadoop и Spark: запускать задачи MapReduce и Spark-приложения - загружать, очищать и предварительно обрабатывать большие объемы данных; - создавать потоковые приложения для обработки данных в реальном времени, Владеет: - навыками работы с кластерами Big Data (установка, настройка, мониторинг); - практическими навыками программирования на Scala, Python или Java для обработки больших данных.	60-69 баллов

<p>оптимизации и масштабирования обработки данных;</p> <p>-основы безопасности и управления доступом в системах Big Data</p> <p>Умеет:</p> <ul style="list-style-type: none"> - устанавливать и настраивать среды для обработки больших данных; - работать с Hadoop и Spark: запускать задачи MapReduce и Spark-приложения - загружать, очищать и предварительно обрабатывать большие объемы данных; - создавать потоковые приложения для обработки данных в реальном времени, - писать ETL-процессы с использованием технологий больших данных; 	<p>Базовый уровень</p>	<p>Знает:</p> <ul style="list-style-type: none"> - основные понятия и принципы работы с большими данными; - архитектуру и компоненты экосистемы Hadoop и Spark; - основные типы систем хранения данных (HDFS, NoSQL базы); - принципы MapReduce и основы работы с Apache Spark; - основы потоковой обработки данных и инструменты для этого (Kafka, Spark Streaming); <p>Умеет:</p> <ul style="list-style-type: none"> - устанавливать и настраивать среды для обработки больших данных; - работать с Hadoop и Spark: запускать задачи MapReduce и Spark-приложения - загружать, очищать и предварительно обрабатывать большие объемы данных; - создавать потоковые приложения для обработки данных в реальном времени, - писать ETL-процессы с использованием технологий больших данных; <p>Владеет:</p> <ul style="list-style-type: none"> - навыками работы с кластерами Big Data (установка, настройка, мониторинг); - практическими навыками программирования на Scala, Python или Java для обработки больших данных; - умением оптимизировать производительность Hadoop и Spark приложений; 	<p>70-84 баллов</p>
<p>- анализировать и визуализировать результаты обработки данных</p> <p>- реализовывать базовые задачи машинного обучения с помощью Spark MLlib</p> <p>Владеет:</p> <ul style="list-style-type: none"> - навыками работы с кластерами Big Data (установка, настройка, мониторинг); - практическими навыками программирования на Scala, Python или 	<p>Высокий уровень</p>	<p>Знает:</p> <ul style="list-style-type: none"> - основные понятия и принципы работы с большими данными; - архитектуру и компоненты экосистемы Hadoop и Spark; - основные типы систем хранения данных (HDFS, NoSQL базы); - принципы MapReduce и основы работы с Apache Spark; - основы потоковой обработки данных и инструменты для этого (Kafka, Spark Streaming); - методы оптимизации и масштабирования обработки данных; - основы безопасности и управления доступом в системах Big Data <p>Умеет:</p> <ul style="list-style-type: none"> - устанавливать и настраивать среды для обработки больших данных; 	<p>85-100 баллов</p>

Java для обработки больших данных; - умением оптимизировать производительность Hadoop и Spark приложений; - использованием инструментов для построения и автоматизации конвейеров обработки данных		- работать с Hadoop и Spark: запускать задачи MapReduce и Spark-приложения - загружать, очищать и предварительно обрабатывать большие объемы данных; - создавать потоковые приложения для обработки данных в реальном времени, - писать ETL-процессы с использованием технологий больших данных; - анализировать и визуализировать результаты обработки данных - реализовывать базовые задачи машинного обучения с помощью Spark MLlib Владеет: - навыками работы с кластерами Big Data (установка, настройка, мониторинг); - практическими навыками программирования на Scala, Python или Java для обработки больших данных; - умением оптимизировать производительность Hadoop и Spark приложений; - использованием инструментов для построения и автоматизации конвейеров обработки данных	
--	--	---	--

Балльно-рейтинговая система

Общая максимальная сумма баллов, которую студент может набрать по дисциплине в течение семестра – 100 баллов: 60 баллов текущий контроль и рубежный контроль + 40 баллов зачет/экзамен (итоговый контроль);

– общая максимальная сумма баллов, которую студент может набрать в течение семестра за выполнение всех видов работ во время аудиторных и внеаудиторных занятий, активность и посещаемость, должна быть равна 60 баллам;

– минимальная сумма баллов, при которой студент допускается к зачету/экзамену (итоговому контролю), равна 36 баллам (60% от 60 баллов);

– минимальная сумма баллов, при которой студент получает положительную итоговую оценку по дисциплине равна 60 баллам (60% от 100 баллов).

Связь между четырехбалльной и столбальной системами оценки качества обучения студентов

Оценка	Буквенный эквивалент оценки	Рейтинговые баллы
Отлично	A+	95-100
	A	90-94
	A-	85-89
Хорошо	B+	80-84
	B	75-79
	B-	70-74
Удовлетворительно	C+	67-69
	C	64-66
	C-	60-63
Неудовлетворительно	D	40-59
–	F	<40

Зачтено	S	60-100
Не зачтено	U	<60

Оценочные средства и критерии их оценки

Примерные вопросы к зачету (5 семестр):

1. BigData, назначение
2. Технологии BigData
3. BigData, история появления и основные принципы BigData.
4. Достоинства и недостатки BigData.
5. Технологии управления знаниями, визуализации знаний и интеллектуальные карты.
6. Данные, информация, знания, модели. Наука о данных.
7. Эволюционное развитие архитектур и данных.
8. Критерии больших данных. Источники больших данных. Интернет вещей. Робототехника.
9. Навыки, специфичные для науки о данных
10. Хранение больших данных. Масштабируемость СУБД.
11. Определение и классификация СУБД: MongoDB, Google BigTable, HBase, Redis, DynamoDB, Apache Cassandra. Графовая СУБД Neo4j. NoSQL.
12. Сравнение СУБД.
13. Параллельные архитектуры.
14. Метрики производительности
15. Аппаратные платформы
16. Обзор экосистемы Hadoop и Spark
17. Установка и настройка среды для обработки больших данных
18. Работа с HDFS
19. NoSQL базы данных: HBase
20. NoSQL базы данных: Cassandra
21. Распределённые файловые системы.
22. Объектные хранилища.
23. Колонковые базы данных
24. NoSQL базы данных.
25. Системы хранения на основе блоков.
26. Технологии кеширования и in-memory хранилища.
27. Обработка данных с помощью Hadoop MapReduce
28. Работа с RDD. Использование DataFrame
29. Архитектура Apache Spark и ее компонентов (Driver, Executors. RDD, DataFrame<DataSet)
30. Apache Kafka. Spark Streaming
31. Загрузка данных в больших объемах. Предварительная обработка данных
32. Поточковая обработка данных с использованием Apache Kafka и Spark Streaming: настройка, интеграция и обработка потоков в реальном времени
33. Модели распределения и развертывания облачных услуг.
34. Облачные услуги для больших данных
35. Возможные этапы работы с большими даны
36. Облачные сервисы для Больших данных. Сравнение ведущих компаний.
37. Интеллектуальный анализ данных: краткий обзор подходов.
38. Генетические алгоритмы.
39. Деревья принятия решений.
40. Визуализация больших данных.
41. Специфика хранения и обработки больших данных.
42. Парадигма MapReduce

43. Файловая система HDFS.
44. Особенности хранилищ данных NoSQL.
45. Архитектура высоконагруженных систем.

Критерии оценки теоретической части:

- оценка «отлично» (38-40 баллов) *выставляется студенту, если он*
 - Четко знает принципы и базовые концепции BigData, технологии BigData;
 - Дает четкий и правильный ответ, выявляющий понимание учебного материала и характеризующий прочные знания, излагает материал в логической последовательности с использованием принятой терминологии;
 - Ошибок не делает, но допускает оговорки по невнимательности при работе с программными продуктами, которые легко исправляет по требованию преподавателя;
 - Ответ логичен, последователен, технически грамотен.
- оценка «хорошо» (35-38 баллов) *выставляется студенту, если он*
 - Овладел программным материалом, ориентируется в базовых концепциях BigData, умеет применять технологии BigData небольшим затруднением, но знает основные теги и их атрибуты;
 - Дает правильный ответ в определенной логической последовательности;
- оценка «удовлетворительно» (29-34 баллов) *выставляется студенту, если он*
 - Основной программный материал знает нетвердо, но большинство изученных понятий и обозначений усвоил;
 - Ответ дает неполный, построенный несвязно, но выявивший общее понимание вопросов;
- оценка «неудовлетворительно» (0-28 баллов) *выставляется студенту, если он*
 - Обнаруживает незнание или непонимание большей или наиболее важной части учебного материала;
 - Ответы строит несвязно, допускает существенные ошибки, которые не может исправить даже с помощью преподавателя.

Примерные контрольные вопросы

Введение. Обзор экосистемы Hadoop и Spark:

1. Что такое экосистема Hadoop и какие основные компоненты она включает?
2. Какие задачи решает Hadoop и в чем ключевые особенности его архитектуры?
3. Что такое HDFS и какова его роль в экосистеме Hadoop?
4. Чем отличается обработка данных в Hadoop MapReduce и Apache Spark?
5. Какие преимущества предоставляет Apache Spark по сравнению с традиционной моделью MapReduce?
6. Какие основные модули входят в состав Apache Spark?
7. Как устроена распределённая обработка данных в Apache Spark?
8. В каких сценариях лучше использовать Hadoop, а в каких — Spark?
9. Что такое YARN и какова его роль в экосистеме Hadoop?
10. Какие существуют основные инструменты и фреймворки для работы с данными в экосистемах Hadoop и Spark?

Технологии Big Data

1. Как работает MapReduce и какая его роль в обработке больших данных?
2. Какие преимущества предоставляет Apache Spark по сравнению с Hadoop MapReduce?
3. Что такое потоковая обработка данных и какие инструменты её реализуют?
4. Какие существуют популярные инструменты для визуализации и анализа Big Data?
5. Как обеспечивается безопасность и конфиденциальность данных в современных Big Data технологиях?

Примерные задания для практических работ:

1. Исследование источников данных

- Опишите основные источники больших данных. Приведите примеры структурированных и неструктурированных данных. Объясните, какие технологии подходят для сбора каждого типа данных.

2. Разработка ETL-процесса

Создайте схему процесса ETL (Extract, Transform, Load) для примера большого набора данных, например, данных из социальных сетей или интернет-магазина. Опишите этапы извлечения, трансформации и загрузки.

3. Анализ данных с использованием Apache Hadoop

- Настройте Hadoop-кластер (локально или в облаке). Загрузите пример большого набора данных и выполните базовый анализ (например, подсчёт частоты слов в текстах).

Опишите, как работает MapReduce.

- Анализ данных с помощью MapReduce

Напишите и запустите простую задачу MapReduce на примере подсчёта частоты слов в большом тексте, используя Apache Hadoop.

Настройте локальный HDFS-кластер и загрузите туда большой набор данных. Оцените время чтения и записи данных.

- Используя pandas, загрузите большой CSV-файл, выполните очистку (удалите пропуски, исправьте типы данных, уберите дубликаты) и сохраните итоговый файл.

4. Использование инструментов потоковой обработки данных (Stream Processing)

Напишите простой скрипт на Apache Kafka или Apache Flink, который собирает данные с сенсоров или веб-логов в реальном времени, обрабатывает их и сохраняет результаты.

Разверните потоковую обработку с Apache Kafka, создайте продюсера и консьюмера, которые принимают и обрабатывают данные в режиме реального времени.

5. Визуализация больших данных

Выберите набор больших данных (например, данные о движении транспорта в городе). Подготовьте отчет с графиками, диаграммами, которые помогут объяснить ключевые выводы. Поясните, какие инструменты вы использовали для визуализации.

- С помощью библиотек `matplotlib` или `seaborn` визуализируйте ключевые метрики (тренды, распределения) из обработанного набора данных. Подготовьте краткий отчет с графиками.

6. Сравнение технологий хранения данных

Сравните плюсы и минусы баз данных SQL (например, MySQL, PostgreSQL) и NoSQL (MongoDB, Cassandra) в контексте больших данных. Приведите примеры задач, для которых лучше подходит каждая из технологий.

7. Обработка данных с помощью Python

Напишите скрипт на Python, который считывает большой CSV-файл (не менее 1 млн строк), выполняет очистку данных (удаление дубликатов, заполнение пропусков) и сохраняет результаты в новый файл.