

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
BigData**

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

- Формирование знаний о технологиях работы с большими данными и получение начальных навыков развертывания инфраструктуры Big data;
- Освоение современных технологий и инструментов обработки больших данных (Hadoop, Spark, NoSQL, потоковая обработка и др.).
- Формирование навыков проектирования и реализации систем обработки больших данных.
- Развитие умений работы с распределёнными вычислительными средами и облачными платформами.

Место дисциплины в структуре образовательной программы

Дисциплина В1.О.02.03 BigData входит в Блок 1. Дисциплины (модули). Обязательная часть Б1.О.02 Математическое моделирование учебного плана 09.04.02 Информационные системы и технологии.

Планируемые результаты обучения по дисциплине и индикаторы достижения компетенций.

ОПК-3. Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями

ОПК-3.1. Понимает принципы, методы и средства анализа и структурирования профессиональной информации

ОПК-3.2. Анализирует профессиональную информацию, выделяет в ней главное, структурирует, оформляет и представляет в виде аналитических обзоров

ОПК-6. Способен использовать методы и средства системной инженерии в области получения, передачи, хранения, переработки и представления информации посредством информационных технологий

ОПК-6.1. Формулирует основные положения системной инженерии и определяет методы их приложения в области получения, переработки и представления информации посредством информационных технологий

ОПК-7. Способен разрабатывать и применять математические модели процессов и объектов при решении задач анализа и синтеза распределённых информационных систем и систем поддержки принятия решений

ОПК-7.1. Понимает математические алгоритмы функционирования, принципы построения, модели хранения и обработки данных распределённых информационных систем и систем поддержки принятия решений

В результате освоения дисциплины студент должен:

Знать:

- Основные понятия и принципы работы с большими данными.
- Архитектуру и компоненты экосистемы Big Data (Hadoop, Spark, NoSQL базы данных и др.).
- Методы сбора, хранения и обработки больших объёмов данных.
- Основы распределённых вычислений и потоковой обработки данных.
- Вопросы безопасности, качества и конфиденциальности данных.
- Современные инструменты и технологии анализа и визуализации больших данных.

Уметь:

- Работать с инструментами для обработки больших данных (например, Hadoop MapReduce, Apache Spark).
- Настраивать и использовать распределённые файловые системы и базы данных.
- Писать программы и запросы для анализа больших данных.

- Выполнять обработку как пакетных (batch), так и потоковых данных.
- Проводить визуализацию результатов анализа для принятия решений.
- Оценивать качество и полноту данных, обеспечивать их защиту и анонимизацию.

Владеть:

- Навыками проектирования и построения систем обработки и анализа Big Data.
- Способностью интегрировать различные инструменты и технологии для построения комплексных решений.
- Умением работать с распределёнными вычислительными средами и облачными платформами.
- Критическим мышлением и аналитическими способностями при интерпретации больших 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 4 зачетные единицы, 144 часа.

№ Название разделов дисциплины	Лекция	Лабораторная работа	Самостоятельная работа
Семестр 1	14	14	116
1 Концепции больших данных	6	4	46
2 Инфраструктура больших данных	4	4	40
3 Технологии Big Data	4	6	30

Тематическое планирование курса

Темы

Концепции больших данных

Семестр 1

Принципы и характеристики больших данных

Лекция. 2(0) ч. Процессы науки о данных

Лабораторная работа. 2(0) ч. Источники больших данных: поисковые машины, социальные сети, банковские транзакции, телеком, биоинформатика, Интернет вещей.

Самостоятельная работа. 16(0) ч. Навыки, специфичные для науки о данных

Хранение больших данных.

Лекция. 4(0) ч. Хранение больших данных. Масштабируемость СУБД.

Лабораторная работа. 2(0) ч. Определение и классификация СУБД: MongoDB, Google BigTable, HBase, Redis, DynamoDB, Apache Cassandra. Графовая СУБД Neo4j. NoSQL.

Самостоятельная работа. 30(0) ч. Сравнение СУБД.

Инфраструктура больших данных

Семестр 1

Распределенные архитектуры больших данных. Параллельные и распределенные вычисления.

Лекция. 2(0) ч. Базовые понятия и определения. Параллельные архитектуры. Аппаратные платформы

Лабораторная работа. 2(0) ч. Метрики производительности

Самостоятельная работа. 20(0) ч. Аппаратные платформы. Требования к аппаратному и сетевому обеспечению.

Облачные вычисления

Лекция. 2(0) ч. Модели распределения и развертывания облачных услуг.

Лабораторная работа. 2(0) ч. Облачные услуги для больших данных

Самостоятельная работа. 20(0) ч. Облачные сервисы для Больших данных. Сравнение ведущих компаний.

Технологии Big Data

Семестр 1

Технологии распределенных вычислений и хранения данных

Лекция. 4(0) ч. Инструменты программирования на основе модели MapReduce.

Лабораторная работа. 1(0) ч. Hadoop – технология распределенных вычисления на основе модели Map Reduce. HDFS – технология распределенной файловой системы Hadoop.

Apache Hadoop

Лабораторная работа. 1(0) ч. Инструменты программирования на основе рабочих потоков.

Apache Spark. Apache Storm. Apache Airflow.

Лабораторная работа. 1(0) ч. Инструменты программирования на основе передачи сообщений.

Лабораторная работа. 1(0) ч. Инструменты программирования на основе массового синхронного параллелизма. Spark GraphX.

Лабораторная работа. 2(0) ч. Инструменты SQL-подобного программирования. Apache Hive. Apache Pig Инструменты программирования на основе разделенного глобального адресного пространства.

Самостоятельная работа. 30(0) ч. Сравнение инструментов программирования

БРС

Семестр	Контрольные точки	Баллы
1	Текущий контроль в разделе «Концепции больших данных»	
	Контрольные вопросы	20
1	Текущий контроль в разделе «Инфраструктура больших данных»	
	Контрольные вопросы	20
1	Текущий контроль в разделе «Технологии Big Data»	
	Контрольные вопросы	20
	Зачёт	40

Итого за семестр 1: 100

Учебно-методическое и информационное обеспечение учебного процесса

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

Теоретическая часть курса, общие вопросы излагаются в лекционном курсе. Отдельные вопросы могут выноситься на самостоятельное изучение. Для приобретения навыков работы на ПК предназначены лабораторные занятия. При изучении дисциплины используются интерактивные формы занятий (лекция-дискуссия, защита рефератов) в объеме 10 часов.

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

Теоретическая часть курса, общие вопросы методики и технологий применения компьютерных средств излагаются преподавателем в лекционном курсе. Отдельные вопросы могут выноситься на самостоятельное изучение. Студент должен иметь в виду, что на лекциях преподаватель определяет такие вопросы и рекомендует необходимую для их изучения литературу, источники и др. ресурсы. Для успешного освоения курса необходимо внимательно фиксировать основные положения лекции, своевременно их усваивать, при необходимости самостоятельно прорабатывать, используя основную и дополнительную литературу.

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_bigdata_magistr.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. [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>
2. [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>

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

1. Технологии хранения и обработки больших данных:

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

2. 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)

3. Инструменты для визуализации данных:

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

4. Среда разработки и управления проектами:

- IntelliJ IDEA, Eclipse (IDE для разработки на Java/Scala)
- Visual Studio Code
- Git (системы контроля версий)
- Docker (контейнеризация приложений)

5. Облачные платформы и сервисы:

- AWS (Amazon EMR, S3, Redshift)
- Microsoft Azure (HDInsight, Data Lake)
- Google Cloud Platform (BigQuery, Dataflow)
- Databricks (облачная платформа для Apache Spark)

6. Системы управления базами данных и NoSQL решения:

- MySQL, PostgreSQL (реляционные СУБД)
- MongoDB (NoSQL)
- HBase (распределённая колоночная СУБД)

7. Информационные справочные системы и ресурсы:

- Официальная документация Apache проектов (Hadoop, Spark, Kafka и др.)
- Stack Overflow (технические вопросы и решения)
- GitHub (репозитории с обучающими проектами)
- Онлайн-курсы и платформы (Coursera, edX, Udemy)
- Блоги и сайты с примером кода (Medium, Towards Data Science)

8. Средства обучения и взаимодействия:

- Платформы для видеоконференций (Zoom, Microsoft Teams)
- Электронные образовательные платформы (LMS: Moodle, Canvas)
- Интерактивные онлайн-платформы (Google Colab)

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

Личный кабинет преподавателя или студента БГУ <https://my.bsu.ru/>

Федеральное интернет-тестирование: проекты «Интернет-тренажеры в сфере профессионального образования» и «Федеральный интернет-экзамен в сфере

профессионального образования»

База данных «Университет»

Электронные библиотечные системы: Руконт, издательство «Лань», Консультант студента
Технологии хранения и обработки больших данных:

- 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)

Облачные платформы и сервисы:

- AWS (Amazon EMR, S3, Redshift)
- Microsoft Azure (HDInsight, Data Lake)
- Google Cloud Platform (BigQuery, Dataflow)
- Databricks (облачная платформа для Apache Spark)

Системы управления базами данных и NoSQL решения:

- MySQL, PostgreSQL (реляционные СУБД)
- MongoDB (NoSQL)
- HBase (распределённая колоночная СУБД)

Информационные справочные системы и ресурсы:

- Официальная документация Apache проектов (Hadoop, Spark, Kafka и др.)
- Stack Overflow (технические вопросы и решения)
- GitHub (репозитории с обучающими проектами)
- Онлайн-курсы и платформы (Coursera, edX, Udemy)
- Блоги и сайты с примером кода (Medium, Towards Data Science)

Средства обучения и взаимодействия:

- Платформы для видеоконференций (Zoom, Microsoft Teams)
- Электронные образовательные платформы (LMS: Moodle, Canvas)
- Интерактивные онлайн-платформы (Google Colab)

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
Big Data**

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

ОПК-3.1 - Понимает принципы, методы и средства анализа и структурирования профессиональной информации

ОПК-3.2 – Анализирует профессиональную информацию, выделяет в ней главное, структурирует, оформляет и представляет в виде аналитических обзоров

ОПК-6.1. Формулирует основные положения системной инженерии и определяет методы их приложения в области получения, переработки и представления информации посредством информационных технологий

ОПК-7.1 - Понимает математические алгоритмы функционирования, принципы построения, модели хранения и обработки данных распределенных информационных систем и систем поддержки принятия решений

№	Контролируемые разделы, темы, модули	Формируемые компетенции	Этапы формирования	Оценочные средства	
				Вид	Количество
1.	Концепции больших данных	ОПК-3.1, ОПК-3.2	1 семестр	опрос	1
2.	Инфраструктура больших данных	ОПК-3.1, ОПК-3.2, ОПК-7.1.	1 семестр	опрос	1
3.	Технологии Big Data	ОПК-3.2, ОПК-7.1, ОПК-6.1	1 семестр	опрос	1

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

Результаты обучения	Уровень сформированности компетенций	Показатели оценивания компетенций	Шкала оценивания
Знать: - основные понятия и принципы работы с большими данными. - архитектуру и компоненты экосистемы Big Data (Hadoop, Spark, NoSQL базы данных и др.). - методы сбора, хранения и обработки больших объемов данных. - основы распределённых вычислений и потоковой обработки данных. - вопросы безопасности, качества и конфиденциальности данных. - Современные инструменты и	Пороговый уровень (как обязательный для всех студентов)	Знает: основные понятия и принципы работы с большими данными, методы сбора, хранения и обработки больших объёмов данных Умеет: работать с современными инструментами и технологиями анализа и визуализации больших данных для решения простых задач. Владеет: умением работать с распределёнными вычислительными средами и облачными платформами.	60-69 баллов
	Базовый уровень	Знает: основные понятия и принципы работы с большими данными; архитектуру и компоненты экосистемы Big Data (Hadoop, Spark, NoSQL базы данных и др.); методы сбора, хранения и обработки больших объёмов данных. Умеет: работать с инструментами для обработки больших данных (например, Hadoop MapReduce, Apache Spark); настраивать и использовать распределённые файловые системы и	70-84 баллов

<p>технологии анализа и визуализации больших данных.</p> <p>Уметь:</p> <ul style="list-style-type: none"> - работать с инструментами для обработки больших данных (например, Hadoop MapReduce, Apache Spark). - настраивать и использовать распределённые файловые системы и базы данных. - писать программы и запросы для анализа больших данных. - выполнять обработку как пакетных (batch), так и потоковых данных. - проводить визуализацию результатов анализа для принятия решений. - оценивать качество и полноту данных, обеспечивать их защиту и анонимизацию. 		<p>базы данных.</p> <p>Владеет: навыками проектирования и построения систем обработки и анализа Big Data; способностью интегрировать различные инструменты и технологии для построения комплексных решений.</p>	
<p>Владеть:</p> <ul style="list-style-type: none"> - навыками проектирования и построения систем обработки и анализа Big Data. - способностью интегрировать различные инструменты и технологии для построения комплексных решений. - умением работать с распределёнными вычислительными средами и облачными платформами. - критическим мышлением и аналитическими способностями при интерпретации больших данных. 	<p>Высокий уровень</p>	<p>Знает: на высоком уровне основные понятия и принципы работы с большими данными; архитектуру и компоненты экосистемы Big Data (Hadoop, Spark, NoSQL базы данных и др.); методы сбора, хранения и обработки больших объёмов данных; основы распределённых вычислений и потоковой обработки данных; вопросы безопасности, качества и конфиденциальности данных.</p> <p>Умеет: работать с инструментами для обработки больших данных на высоком уровне (например, Hadoop MapReduce, Apache Spark); настраивать и использовать распределённые файловые системы и базы данных; писать программы и запросы для анализа больших данных; выполнять обработку как пакетных (batch), так и потоковых данных; проводить визуализацию результатов анализа для принятия решений.</p> <p>Владеет: навыками проектирования и построения систем обработки и анализа Big Data на высоком уровне; способностью интегрировать различные инструменты и технологии для построения комплексных решений; умением работать с распределёнными вычислительными средами и облачными платформами; критическим мышлением и аналитическими способностями при интерпретации больших данных.</p>	<p>85-100 баллов</p>

Балльно-рейтинговая система по дисциплине «Big Data»

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 столбалльной 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):

Концепции больших данных

1. Концепция больших данных.
2. Основные подходы к обработке и анализу больших данных.
3. Основные этапы жизненного цикла данных в Big Data.
4. Источники больших данных: поисковые машины, социальные сети, банковские транзакции, телеком, биоинформатика, Интернет вещей.
5. Навыки, специфичные для науки о данных.
6. Хранение больших данных. Масштабируемость СУБД.
7. Определение и классификация СУБД: MongoDB, Google BigTable, HBase, Redis, DynamoDB, Apache Cassandra.
8. Графовая СУБД Neo4j.
9. БД NoSQL.
10. Распределённая файловая система и её роль в Big Data.
11. Основные компоненты архитектуры Hadoop.
12. Аппаратные платформы. Требования к аппаратному и сетевому обеспечению систем Big Data
13. Параллельные архитектуры
14. Метрики производительности
15. Модели распределения и развертывания облачных услуг.

16. Инструменты программирования на основе модели MapReduce.
17. Hadoop – технология распределенных вычисления на основе модели Map Reduce.
18. HDFS – технология распределенной файловой системы Hadoop.
19. Apache Hadoop
20. Инструменты программирования на основе рабочих потоков. Apache Spark.
21. Инструменты программирования на основе рабочих потоков. Apache Airflow.
22. Инструменты программирования на основе рабочих потоков. Apache Storm.
23. Инструменты программирования на основе передачи сообщений.
24. Инструменты программирования на основе массового синхронного параллелизма. Spark GraphX.
25. Инструменты SQL-подобного программирования. Apache Hive.
26. Инструменты SQL-подобного программирования. Apache Pig.
27. Инструменты программирования на основе разделенного глобального адресного пространства.
28. Сравнение инструментов программирования
29. Обеспечение безопасности и конфиденциальности данных в современных Big Data технологиях
30. Инструменты для визуализации и анализа Big Data

Критерии оценки теоретической части:

- оценка «отлично» (19-20 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» (17-18 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» (15-16 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» (0-14 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

Концепции больших данных

1. Что такое большие данные и в чем их основные характеристики (5V)?
2. Какие основные этапы жизненного цикла данных в Big Data?
3. Какие выгоды и вызовы связаны с использованием больших данных?
4. В чем отличие структурированных, полуструктурированных и неструктурированных данных?
5. Какие существуют основные подходы к обработке и анализу больших данных?

Инфраструктура больших данных

1. Что такое распределённая файловая система и какова её роль в Big Data?
2. Назовите и опишите основные компоненты архитектуры Hadoop.
3. Что представляет собой кластеры и как они используются в инфраструктуре Big Data?
4. Чем отличаются NoSQL базы данных от реляционных в контексте Big Data?
5. Какие требования предъявляются к аппаратному и сетевому обеспечению систем Big Data?

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?

Критерии оценки устного опроса теоретической части:

- *оценка «отлично» (19-20 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» (17-18 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» (15-16 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» (0-14 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 задания для лабораторных работ

Источники больших данных (поисковые машины, социальные сети, банковские транзакции, телеком, биоинформатика и Интернет вещей).

1. Поисковые машины

- Соберите или найдите датасет поисковых запросов (например, Google Trends) и проанализируйте, как меняется популярность определенных тем в зависимости от времени.

- Опишите, какие виды данных (клики, запросы, геолокация) собирает поисковая система и как они могут быть использованы для улучшения качества поиска.

2. Социальные сети

- Проведите анализ настроений (sentiment analysis) твитов или постов из социальной сети по заданной теме (например, к выборам или спортивному событию).

- Определите ключевых инфлюенсеров в небольшом сообществе на основе количества подписчиков, лайков и репостов.

3. Банковские транзакции

- Исследуйте данные банковских транзакций и попробуйте выявить подозрительные операции, которые могут указывать на мошенничество.

- Смоделируйте поток транзакций и оцените категории расходов: какие категории доминируют (питание, транспорт, развлечения и т.д.).

4. Телеком

- Используя данные о звонках и обмене сообщениями, исследуйте поведение пользователей: часы пик, частые контакты, длительность звонков.

- Постройте модель предсказания оттока абонентов (churn prediction) на основе исторических данных.

5. Биоинформатика

- Проанализируйте набор геномных данных и выявите мутации, связанные с определённым заболеванием.

- Сравните последовательности ДНК нескольких видов и найдите общие гены и отличия.

6. Интернет вещей (IoT)

- Соберите и исследуйте данные с умного дома: температурные датчики, датчики движения, энергопотребления.

- Создайте простую систему мониторинга состояния устройства или предсказания неисправностей на основе IoT-данных (например, датчиков вибрации).

Определение и классификация СУБД: MongoDB, Google BigTable, HBase, Redis, DynamoDB, Apache Cassandra. Графовая СУБД Neo4j. NoSQL.

1. Общие задачи по NoSQL СУБД

- Изучить архитектуру и особенности различных типов NoSQL СУБД (документно-ориентированные, графовые, колоночные, key-value). Составить сравнительную таблицу по характеристикам (масштабируемость, тип данных, возможности запросов, схема данных).

2. MongoDB (Документно-ориентированная СУБД)

- Установить MongoDB и создать базу данных с коллекцией «Пользователи» со следующими полями: имя, возраст, город, интересы (список).

- Выполнить запросы на добавление, обновление и удаление документов.

- Выполнить поиск пользователей, живущих в определённом городе и имеющих интересы из заданного списка.

3. Google BigTable / HBase (Колонковые СУБД)

- Ознакомиться с архитектурой HBase (или BigTable).

- Создать таблицу с несколькими колонками (например, данные о клиентах: ID, имя, покупки).

- Выполнить операции вставки и чтения данных по ключу, а также сканирование по диапазону ключей.

- Рассмотреть применение семейств колонок и версионность данных.

4. Redis (Key-Value)

- Установить Redis, создать несколько key-value пар с разными типами значений (строка, список, множество).

- Выполнить операции добавления, удаления, поиска по ключу.

- Реализовать простой кэш с установкой времени жизни (TTL) для объектов.

5. DynamoDB (Колонковая NoSQL СУБД AWS)

- Ознакомиться с моделированием данных в DynamoDB.

- Создать таблицу для хранения информации о продуктах: ID, название, категория, цена.

- Написать запросы на выборку по ключу и вторичным индексам.

- Определить стратегию масштабирования таблицы.

6. Apache Cassandra (Колонковая СУБД)

- Установить Apache Cassandra.

- Создать ключевое пространство и таблицу, например, для хранения данных о событиях (ID события, время, тип события, пользователь).

- Вставить несколько записей и выполнить выборки с использованием WHERE по первичному ключу и партициям.

- Пояснить, как работает репликация и консистентность в Cassandra.

7. Neo4j (Графовая СУБД)

- Установить Neo4j, создать небольшой граф социальных связей (Пользователи и их связи: друзья, коллеги).

- Выполнить CRUD-операции с узлами и ребрами.

- Использовать язык запросов Cypher для поиска всех друзей определённого пользователя, поиска друзей друзей, обнаружения циклов и определения кратчайшего пути.

Облачные услуги для больших данных

- Зарегистрироваться и настроить аккаунт в одной из популярных облачных платформ (AWS, Google Cloud, Azure).
- Изучить основные сервисы для обработки больших данных: Amazon EMR, Google Dataproc, Azure HDInsight.
- Запустить тестовый кластер Hadoop или Spark в облаке.
- Изучить возможности облачных хранилищ (Amazon S3, Google Cloud Storage, Azure Blob Storage).
- Загрузить набор данных в облачное хранилище.
- Организовать доступ и управление правами к данным.
- Подключить облачное хранилище к кластеру для обработки данных.
- Написать и запустить простую задачу MapReduce или Spark для анализа данных, хранящихся в облачном хранилище.
- Отследить выполнение задачи и изучить детали логов работы.
- Подсчитать время выполнения и ресурсы, затраченные на обработку.
- Ознакомиться с потоковой обработкой данных в облаке (например, Amazon Kinesis, Google Cloud Dataflow, Azure Stream Analytics).
- Настроить поток данных и выполнить простую трансформацию.
- Собрать результаты и проанализировать производительность.
- Использовать инструменты мониторинга и отчетности облачной платформы.
- Провести анализ затрат на хранение и обработку данных.
- Настроить автоматическое масштабирование кластера под нагрузкой.
- Подготовить рекомендации по оптимизации использования облачных сервисов для больших данных.

Hadoop, MapReduce и HDFS

- Установить и настроить Apache Hadoop в локальной или кластерной среде.
- Ознакомиться с основными компонентами Hadoop: HDFS, MapReduce, YARN.
- Запустить демо-приложение WordCount и проанализировать результат выполнения.
- Изучить архитектуру HDFS и принцип хранения данных.
- Создать пользовательскую директорию в HDFS.
- Загрузить файлы в HDFS с локальной машины и прочитать их содержимое.

- Выполнить команды HDFS: ls, cat, du, rm.
- Провести репликацию и проверить отказоустойчивость данных.
- Ознакомиться с моделью программирования MapReduce.
- Написать и отладить программу WordCount на Java или Python.
- Запустить программу на подготовленных данных в Hadoop.
- Проанализировать логи и понять распределение задач по узлам.
- Изучить параметры настройки MapReduce (размер блоков, количество reduce задач).
- Переписать приложение для решения задачи подсчёта частоты слов с учетом регистра и удаления стоп-слов.
- Провести сравнение времени выполнения задач при различных настройках.
- Ознакомиться с YARN – менеджером ресурсов Hadoop.
- Работа с несколькими MapReduce задачами одновременно.
- Изучить методы мониторинга и диагностики кластера Hadoop.
- Запуск и анализ работы нескольких приложений на общем кластере.