Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный индустриальный университет»

Кафедра автоматизации и информационных систем

УТВЕРЖДАЮ
Директор института
информационных технологий и
автоматизированных систем
_____ А.В. Зимин
подпись
«____» _____ 20___ г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Инструменты для работы с большими данными

09.03.02 «Информационные системы и технологии»
(направленность (профиль): «Разработка и эксплуатация информационных систем и технологий»)

Квалификация выпускника
Бакалавр

Форма обучения
Очная форма

Срок обучения: 4 года

Год начала подготовки 2025

Новокузнецк
2025

СибГИУ
Подлинник электронного документа хранится в системе электронного согласования документов СибГИУ
ДОКУМЕНТ ПОДПИСАН ЭП
Сертификат: daba56d7b892efca3bf5c04ed0251c98b19b099c
Владелец: Директор ИИТиАС Зимин Алексей Валерьевич
Действителен: с 15 января 2025 г. по 10 апреля 2026 г.
Дата и время: 22-04-2025 14:43:19 ПРОВЕРИТЬ ПОДПИСЬ

1 Цели и задачи освоения учебной дисциплины

Целями учебной дисциплины являются:

– формирование у обучающихся профессиональной компетенции в области разработки и использования систем обработки и анализа больших массивов данных.

Задачами учебной дисциплины являются:

– освоение обучающимся методов анализа и хранения больших объемов данных, этапов жизненного цикла обработки больших данных, языков, наиболее приспособленные для обработки и аналитики больших данных, способов организации хранения и доступа к большим данным;

– освоение обучающимися технологий разработки специализированных программных систем, отвечающих за обработку больших данных.

2 Место учебной дисциплины в структуре ООП по направлению подготовки (специальности)

Учебная дисциплина относится к учебным дисциплинам обязательной части **Блока 1 «Дисциплины (модули)»** ООП по направлению подготовки (специальности) 09.03.02 «Информационные системы и технологии».

Учебная дисциплина базируется на предварительном усвоении обучающимися учебных дисциплин:

– Техническое обеспечение информационно-управляющих систем;

– Основы облачных платформ и распределенных систем;

– Технологии сбора и обмена информацией в системах больших данных;

– Современные СУБД и хранилища данных.

Учебная дисциплина дополняет знания и умения, получаемые по одновременно изучаемым и последующим учебным дисциплинам:

– Проектирование информационно-управляющих систем;

– Программное обеспечение информационно-управляющих систем.

3 Планируемые результаты обучения по учебной дисциплине

Процесс изучения учебной дисциплины направлен на формирование следующих компетенций:

– Общепрофессиональные компетенции

Наименование категории (группы) ОПК	Код и наименование ОПК	Код и наименование индикатора	Планируемые результаты обучения

		достижения ОПК	
	ОПК-1: Способен применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности	ОПК-1.3 Применяет методы математического анализа и моделирования в теоретических и экспериментальных исследованиях профессиональной деятельности	– знать: методы математического анализа больших данных. – уметь: использовать регрессионные модели при работе с большими данными.
	ОПК-2: Способен понимать принципы работы современных информационных технологий и программных средств, в том числе отечественного производства, и использовать их при решении задач профессиональной деятельности	ОПК-2.3 Применяет современные информационные технологии и программные средства при решении прикладных задач	– знать: прикладной инструментарий работы с большими данными. – уметь: использовать информационные технологии решения прикладных задач больших данных.
	ОПК-3: Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	ОПК-3.3 Решает поставленные задачи на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий	– знать: методы и способы поиска информации. – уметь: применять информационно-коммуникационные технологии для поиска информации.

4 Объем и содержание учебной дисциплины

Учебные занятия по учебной дисциплине проводятся в форме контактной работы и в форме самостоятельной работы обучающихся.

Контактная работа включает в себя занятия лекционного типа (лекции), занятия семинарского типа (семинары, практические занятия, практикумы), промежуточную аттестацию обучающихся и иные формы взаимодействия обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации ООП на иных условиях, в том числе при проведении промежуточной аттестации обучающихся. Контактная работа может проводиться с применением электронного обучения, дистанционных образовательных технологий.

Объем учебной дисциплины

Семестр / курс		***ИТОГО***	***7 семестр***
Форма промежуточной аттестации			*экзамен*
Трудоёмкость	*академ. час.*	***144***	*144*
	зачетных единиц	***4***	*4*
Лекции, *академ. час.*		***32***	*32*
в форме практической подготовки		***0***	*0*
Лабораторные работы, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Практические занятия, *академ. час.*		***16***	*16*
в форме практической подготовки		***0***	*0*
Курсовая работа / проект, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Консультации, *академ. час.*		***0***	*0*
в форме практической подготовки		***0***	*0*
Самостоятельная работа, *академ. час.*		***51***	*51*
в форме практической подготовки		***0***	*0*
Контроль, *академ. час.*		***45***	*45*
в форме практической подготовки		***0***	*0*

Содержание учебной дисциплины

Раздел 1 Введение в большие данные (Понятие Data Minig. Прикладные инструменты для работы с Big Data. Технология MapRaduce. Hadoop.);

Тема 1.1 Понятие Data Mining (Основные понятия Data Mining. Понятия аналитической обработки данных, витрины данных. Data Warehouse. Data Lakes. Современные методы построения хранилищ данных.);

Тема 1.2 Прикладные инструменты для работы с Big Data (Инструменты для работы с большими данными. Технологии параллельной (parallel) обработки информации. Технология mpp. Общие сведения о пакетах Apache Hadoop, Apache Spark, Kafka, RabbitMQ.);

Тема 1.3 Технология MapReduce (Технология работы с большими массивами данных. Основные этапы технологии MapReduce. Использование MapReduce в современных пакетах работы с большими данными.);

Тема 1.4 Apache Hadoop (Концепции Apache Hadoop. Типы узлов Hadoop. Концепция арбитра в распределенных многопоточных системах. Файловая система HDFS. Принцип работы Hadoop.);

Тема 1.5 Apache Spark (Концепции Apache Spark. Типы узлов Spark. Механизмы передачи и обработки данных Apache Spark. Файловая система RDD. Разъединенные универсальные коллекции Apache Spark.);

Раздел 2 Технологии анализа данных (Жизненный цикл анализа больших данных, стандарты. Когнитивный анализ данных. Визуализация больших данных.);

Тема 2.1 Жизненный цикл анализа больших данных (Основные этапы подготовки и хранения больших данных для анализа. Сбор больших данных. Подготовка к анализу больших данных. Анализ данных.);

Тема 2.2 Статистический анализ больших данных (Методы анализа данных. Цель анализа данных. Синтаксический анализ данных. Когнитивный анализ данных. Способы и технологии когнитивного анализа данных);

Тема 2.3 Визуализация больших данных (Задача визуализации данных. Технологии визуализации данных. Инструментарий визуализации больших данных).

5 Перечень тем лекций

№ раздела / темы дисциплины	**Темы лекций**	**Трудоемкость, *академ.час***	
		всего	**в форме практической подготовки**
Раздел 1.	Введение в большие данные		
Тема 1.1.	Понятие Data Mining	2	
Тема 1.2.	Прикладные инструменты для работы с Big Data	4	
Тема 1.3.	Технология MapReduce	4	
Тема 1.4.	Apache Hadoop	4	
Тема 1.5.	Apache Spark	6	
Раздел 2.	Технологии анализа данных		
Тема 2.1.	Жизненный цикл анализа больших данных	4	
Тема 2.2.	Статистический анализ больших данных	6	
Тема 2.3.	Визуализация больших данных	2	
Итого:		**32**	**0**

6 Перечень тем практических занятий (семинаров)

№ раздела / темы дисциплины	Темы практических занятий (семинаров)	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	Введение в большие данные		
Тема 1.1.	Понятие Data Mining	2	
Тема 1.2.	Прикладные инструменты для работы с Big Data	2	
Тема 1.3.	Технология MapReduce	2	
Тема 1.4.	Apache Hadoop	2	
Тема 1.5.	Apache Spark	2	
Раздел 2.	Технологии анализа данных		
Тема 2.1.	Жизненный цикл анализа больших данных	2	
Тема 2.2.	Статистический анализ больших данных	2	
Тема 2.3.	Визуализация больших данных	2	
Итого:		**16**	**0**

7 Перечень тем лабораторных работ

№ раздела / темы дисциплины	Темы лабораторных работ	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
	Отсутствуют		
Итого:		**0**	**0**

8 Перечень тем курсовых работ (проектов)

№ раздела / темы дисциплины	Темы курсовых работ (проектов)	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
	Отсутствуют		
Итого:		**0**	**0**

9 Виды самостоятельной работы

№ раздела / темы дисциплины	Виды самостоятельной работы	Трудоемкость, *академ.час*	
		всего	**в форме практической подготовки**
Раздел 1.	1. Выполнение домашнего задания; 2. Оформление отчета о практической работе; 3. Подготовка к	30	

	практическому занятию; 4. Прохождение тестирования.		
Раздел 2.	1. Выполнение домашнего задания; 2. Оформление отчета о практической работе; 3. Подготовка к практическому занятию; 4. Прохождение тестирования.	21	
Контроль	*Подготовка к экзамену*	45	
Итого:		**96**	**0**

10 Матрица компетентностных задач по учебной дисциплине

Контролируемые разделы (темы) дисциплины	Контролируемые компетенции	Оценочные средства
Раздел 1.	ОПК-1; ОПК-2; ОПК-3	1. Домашнее задание; 2. Наблюдение за работой на практическом занятии; 3. Отчет о практической работе; 4. Тестовые задания
Раздел 2.	ОПК-1; ОПК-2; ОПК-3	1. Домашнее задание; 2. Наблюдение за работой на практическом занятии; 3. Отчет о практической работе; 4. Тестовые задания
Экзамен.	ОПК-1; ОПК-2; ОПК-3	1. Вопросы к экзамену

11 Входной контроль

11.1 Цель входного контроля − определить начальный уровень подготовленности обучающихся и выстроить индивидуальную траекторию обучения. В условиях личностно-ориентированной образовательной среды результаты входного оценивания обучающегося используются как начальные значения в индивидуальном профиле академической успешности обучающегося.

11.2 Описание оценочных средств

Форма проведения входного контроля – тестирование.

Длительность тестирования – 30 минут.

Тестирование проводится в течение первых двух недель изучения учебной дисциплины в СУО «Moodle».

Для проведения тестирования используются тестовые задания для входного контроля (для ответа на 1 вопрос, как правило, выделяется 1 минута).

Шкала оценивания результатов тестирования

За каждый правильный ответ выставляется 1 балл. Оценка выставляется в зависимости от процента правильных ответов:

Оценка формируется в соответствии с критериями таблицы:

Процент результативности (правильных ответов)	Оценка
86 до 100	«отлично»
от 67до 85	«хорошо»
от 50 до 66	«удовлетворительно»
менее 50	«неудовлетворительно»

Основываясь на результатах тестирования, формируется индивидуальная траектория обучающегося (группы обучающихся), исходя из следующих соображений:

– 86 – 100 % – обучающийся готов к изучению учебной дисциплины без дополнительного повторения изученного ранее материала;

– 50 – 67 % – обучающийся в целом готов к изучению учебной дисциплины, требуется повторение отдельных вопросов(тем) изученного ранее материала, необходимого для успешного освоения учебной дисциплины;

– 49% и менее – обучающийся имеет слабую подготовку к изучению учебной дисциплины, необходимы корректирующие занятия и консультации.

Тестовые задания для входного контроля знаний обучающихся приведены в приложении.

12 Текущий контроль

12.1 Текущий контроль знаний используется для оперативного и регулярного управления учебной деятельностью (в том числе самостоятельной) обучающихся. Данный вид контроля стимулирует у обучающихся стремление к систематической самостоятельной работе по изучению учебной дисциплины и позволяет оценить степень освоения изученного материала.

Текущий контроль успеваемости осуществляется в течение семестра, в ходе повседневной учебной работы по индивидуальной инициативе педагогического работника, в том числе посредством использования СУО «Moodle».

Формы проведения текущего контроля – домашнее задание; наблюдение за работой на практическом занятии; отчет о практической работе; тестовые задания.

12.2 Описание оценочных средств

Шкалы оценивания (методики оценки)

12.2.1 Рекомендации по оцениванию домашнего задания

Домашнее задание – это самостоятельная форма учебной деятельности обучающегося, которая осуществляется без непосредственного руководства и контроля со стороны педагогического работника, но по его поручению.

Критерии оценки:

– правильность ответа по содержанию задания (учитывается количество и характер ошибок при ответе);

– полнота и глубина ответа (учитывается количество усвоенных фактов, понятий и т.п.);

– сознательность ответа (учитывается понимание излагаемого материала);

– логика изложения материала (учитывается умение строить целостный, последовательный ответ, грамотно пользоваться специальной терминологией);

– рациональность использованных приемов и способов решения поставленной учебной задачи (учитывается умение использовать наиболее прогрессивные и эффективные способы достижения цели);

– использование при ответе требований нормативных документов, учебной литературы (обязательное условие);

– использование графических элементов с целью выделения особо значимой информации: таблиц, схем, рисунков и т.п.

Критерии оценки домашнего задания

Наименование параметра оценки	Оценка
Обучающийся выполнил задание в полном объёме; полно и аргументировано отвечает по содержанию задания; обнаруживает понимание материала; может обосновать свои суждения, применить знания на практике, привести необходимые примеры, в том числе самостоятельно составленные; излагает материал последовательно и правильно.	отлично
Обучающийся выполнил задание, удовлетворяющее тем же требованиям, что и для оценки «отлично», но допускает несколько незначительных ошибок.	хорошо
Обучающийся обнаруживает знание и понимание основных положений задания, но излагает материал неполно и допускает неточности в определении понятий или формулировке ответа; не умеет достаточно глубоко и доказательно обосновать свои суждения и привести примеры; излагает материал непоследовательно и допускает ошибки.	удовлетворительно
Обучающийся обнаруживает незнание ответа на соответствующее задание, допускает ошибки в	неудовлетворительно

Наименование параметра оценки	Оценка
формулировке определений или ответа в целом, искажающие их смысл, беспорядочно и неуверенно излагает материал. Оценка «неудовлетворительно» отмечает такие недостатки в подготовке обучающегося, которые являются серьезным препятствием к успешному освоению последующего материала.	

12.2.2 Рекомендации по оцениванию результатов тестирования обучающихся

Тестирование проводится по изученным разделам (темам) учебной дисциплины для выявления уровня знаний и умений обучающегося, мотивирования обучающегося к активизации самостоятельной работы по усвоению учебного материала, выявления и устранения пробелов в знаниях, повышения дисциплины и организации учебной деятельности обучающегося.

Тестирование проводится в СУО «Moodle» в рамках часов, отведенных на самостоятельную работу обучающихся. Для тестирования используется 20 вопросов из базы типовых тестовых заданий по изученным разделам (темам) учебной дисциплины. Обучающийся выбирает один ответ из предложенных вариантов ответа. В случае выбора неправильного ответа обучающийся получает за данный вопрос 0 баллов; в случае правильного ответа – 1 балл.

Оценка формируется в соответствии с критериями таблицы:

Максимальный балл	Проходной балл	Оценка
20	не менее 18	отлично
17	не менее 14	хорошо
13	не менее 10	удовлетворительно
9	–	неудовлетворительно

12.2.3 Рекомендации по оцениванию в процессе наблюдения за работой на практическом занятии

Наблюдение за работой на практическом занятии – важнейший компонент контроля деятельности обучающихся и качества их обучения.

Критерии оценки в процессе наблюдения за работой на практическом занятии:

– самостоятельность выполнения задания и работы с методическими указаниями (учитывается работа в течение занятия, быстрота и способность нахождения в методических указаниях нужной информации);

– правильность выполнения работы (учитывается последовательность выполняемых действий, умение работать с инструментами и/или средствами вычислительной техники);

– рациональность использования времени, отведенного на задание (не одобряется затянутость выполнения работы с учетом индивидуальных особенностей обучающихся);

– активность обучающегося при выполнении практической работы (учитывается индивидуальная работа и работа в группе).

Критерии оценки по результатам наблюдения за работой на практическом занятии

Наименование параметра оценки	Максимальное количество баллов
Самостоятельно выполняет задания; умеет работать с методическими указаниями.	2,5
Указанные в методических указаниях действия выполняет в правильной последовательности; умеет работать с инструментами и/или средствами вычислительной техники.	1,25
Рационально использует время, отведенное на задание; проявляет активность при выполнении практической работы.	1,25
Итого	**5**

Максимальное количество – 5 баллов соответствует оценке «отлично», 4 балла – «хорошо», 3 балла – «удовлетворительно», 2 балла – «неудовлетворительно».

12.2.4 Рекомендации по оцениванию отчета о практической работе

Оформление отчета о проделанной практической работе обучающиеся осуществляют в рамках самостоятельной работы.

Критерии оценки отчета о практической работе:

– правильность выполнения работы (учитывается полнота отчета о практической работе, наличие всех необходимых данных, наличие выводов, наличие ошибок и неточностей);

– аккуратность оформления отчета о практической работе.

Критерии оценки отчета о практической работе

Наименование параметра оценки	Максимальное количество баллов
Отчет о практической работе содержит все необходимые данные и выводы.	2,5
Отчет о практической работе не содержит ошибок и неточностей.	1,25
Отчет о практической работе оформлен аккуратно.	1,25
Итого	**5**

Максимальное количество – 5 баллов соответствует оценке «отлично», 4 балла – «хорошо», 3 балла – «удовлетворительно», 2 балла – «неудовлетворительно».

Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины приведены в приложении.

13 Промежуточная аттестация

13.1 Оценочные материалы для промежуточной аттестации обучающихся по учебной дисциплине предназначены для оценки степени достижения запланированных результатов обучения по завершению изучения учебной дисциплины в установленной учебным планом форме и позволяют определить качество усвоения изученного материала.

Итоговой формой контроля сформированности компетенций у обучающихся по учебной дисциплине является: экзамен.

Оценочные материалы для промежуточной аттестации обучающихся включают: вопросы к экзамену.

13.2 Оценивание обучающегося на экзамене

Оценка	Требования к знаниям
отлично	Оценка «отлично» выставляется обучающемуся, если он глубоко и прочно усвоил программный материал, исчерпывающе, последовательно, четко и логически его излагает, умеет тесно увязывать теорию с практикой, свободно справляется с задачами, вопросами и другими видами применения знаний, причем не затрудняется с ответом при видоизменении заданий, использует в ответе материал различной литературы, правильно обосновывает принятое нестандартное решение, обладает приемами выполнения практических задач по формированию компетенций.
хорошо	Оценка «хорошо» выставляется обучающемуся, если он твердо знает материал, грамотно и по существу излагает его, не допуская существенных неточностей в ответе на вопрос, правильно применяет теоретические положения при решении практических вопросов и задач, обладает необходимыми умениями и приемами их выполнения.
удовлетворительно	Оценка «удовлетворительно» выставляется обучающемуся, если он имеет знания только основного материала, но не усвоил его деталей, допускает неточности, недостаточно правильные формулировки, испытывает сложности при выполнении практических работ и затрудняется связать теорию вопроса с практикой.
неудовлетворительно	Оценка «неудовлетворительно» выставляется обучающемуся, который не знает значительной части программного материала, неуверенно отвечает, допускает серьезные ошибки, не имеет представлений по методике выполнения практической работы. Как правило, оценка «неудовлетворительно» ставится обучающимся, которые не могут продолжить обучение без дополнительных занятий по данной учебной дисциплине.

Задания и вопросы для промежуточной аттестации обучающихся приведены в приложении.

14 Учебно-методическое и информационное обеспечение учебной дисциплины

а) литература:

1 Базы данных в высокопроизводительных информационных системах : учебное пособие / авт.-сост. Е. И. Николаев ; Северо-Кавказский федеральный университет. – Ставрополь : СКФУ, 2016. – 163 с. – URL: https://biblioclub.ru/index.php?page=book&id=466799 (дата обращения: 20.04.2025);

2 Жуковский, О. И. Информационные технологии и анализ данных : учебное пособие / О. И. Жуковский ; Томский Государственный университет систем управления и радиоэлектроники. – Томск : Эль Контент, 2014. – 130 с. – ISBN 978-5-4332-0158-3. – URL: https://biblioclub.ru/index.php?page=book&id=480500 (дата обращения: 20.04.2025);

3 Бутаков, Н. А. Обработка больших данных с Apache Spark : учебно-методическое пособие / Н. А. Бутаков, М. В. Петров, Д. Насонов. – Санкт-Петербург : Университет ИТМО, 2019. – 52 с. – URL: https://biblioclub.ru/index.php?page=book&id=566771 (дата обращения: 20.04.2025);

4 Советов, Б. Я. Базы данных : учебник для вузов / Б. Я. Советов, В. В. Цехановский, В. Д. Чертовской. – 3-е изд., перераб. и доп. – Москва : Издательство Юрайт, 2023. – 420 с. – ISBN 978-5-534-07217-4. – URL: https://urait.ru/bcode/510752 (дата обращения: 20.04.2025);

5 Нестеров, С. А. Базы данных : учебник и практикум для вузов / С. А. Нестеров. — Москва : Издательство Юрайт, 2023. – 230 с. – ISBN 978-5-534-00874-6. – URL: https://urait.ru/bcode/511650 (дата обращения: 20.04.2025);

6 Лэм, Ч. Hadoop в действии / Чак Лэм. – Москва : ДМК-пресс, 2012. – 424 с. – ISBN 978-5-94074-785-7. – URL: https://www.studentlibrary.ru/book/ISBN9785940747857.html (дата обращения: 20.04.2025).

б) ресурсы информационно-телекоммуникационной сети «Интернет»:

1 Электронный каталог : сайт / Научно-техническая библиотека СибГИУ. – Новокузнецк, [199 –]. – URL: http://libr.sibsiu.ru. – URL: https://libr.sibsiu.ru;

2 Электронная библиотека // Научно-техническая библиотека СибГИУ : сайт. – Новокузнецк, [200 –]. – URL: http://library.sibsiu.ru/LibrELibraryFullText.asp. – Режим доступа: для авторизир. пользователей. – URL: https://library.sibsiu.ru/LibrELibraryFullText.asp;

3 Консультант студента : электронно-библиотечная система / ООО «КОНСУЛЬТАНТ СТУДЕНТА». – Москва, [200 –]. – URL: http://www.studentlibrary.ru. – Режим доступа: для авторизир. пользователей;

4 ЛАНЬ : электронно-библиотечная система : [коллекция «Единая профессиональная база знаний для технических вузов - Издательство Лань" ЭБС»] / ООО «Издательство ЛАНЬ». – Санкт-Петербург, [200 –]. – URL: http://e.lanbook.com. – Режим доступа: для авторизир. пользователей;

5 Образовательная платформа ЮРАЙТ / ООО «Электронное издательство ЮРАЙТ». – Москва, [200 –]. – URL: https://urait.ru. – Режим доступа: для авторизир. пользователей;

6 Университетская библиотека онлайн : электронно-библиотечная система / ООО «Директ-Медиа». – Москва, [200 –]. – URL: https://biblioclub.ru. – Режим доступа: для авторизир. пользователей. – URL: http://www.biblioclub.ru;

7 Электронная библиотека УМЦ ЖДТ : [коллекция «Эксплуатация железных дорог»] / ФГБУ ДПО «Учебно-методический центр по образованию на железнодорожном транспорте». – Москва, [2013 –]. – URL: https://umczdt.ru/books/. – Режим доступа: для авторизир. пользователей;

8 НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА eLIBRARY.RU : база данных / ООО «НЭБ». – Москва, [200 –]. – URL: http://elibrary.ru. – Режим доступа: по подписке;

9 Электронные периодические издания ИВИС : универсальная база данных / ООО «ИВИС». – Москва, [200 –]. – URL: http://eivis.ru. – Режим доступа: по подписке.

в) лицензионное и свободно распространяемое программное обеспечение:

– 7-Zip;
– Adobe Acrobat Reader;
– Astra Linux Special Edition;
– Kaspersky Endpoint Security;
– Microsoft Office;
– Microsoft Windows;
– OnlyOffice;
– Р7-Офис.

г) базы данных и информационно-справочные системы:

1 ГАРАНТ : справочно-правовая система / ООО «Правовой центр «Гарант». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.;

2 КонсультантПлюс : справочно-правовая система / ООО «Информационный центр АНВИК». – Новокузнецк, [199 –]. – Режим доступа: компьютерная сеть библиотеки Сиб. гос. индустр. ун-та.;

3 Техэксперт : информационно-справочная система / ООО «Группа компаний «Кодекс». – Кемерово, [200 –]. – Режим доступа: компьютерная сеть Сиб. гос. индустр. ун-та.

15 Материально-техническое обеспечение учебной дисциплины

Материально-техническое обеспечение учебной дисциплины включает учебные аудитории, оснащенные оборудованием, компьютерной техникой, и техническими средствами обучения, в том числе:
- учебную аудиторию для проведения занятий лекционного типа, оборудованную учебной доской, экраном и мультимедийным проектором;
- учебную аудиторию для проведения занятий семинарского типа (практических занятий), оснащенную компьютерным и мультимедийным оборудованием;
- учебную аудиторию (помещения) для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации;
- помещения для самостоятельной работы, оснащенные компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду, научно-техническую библиотеку СибГИУ.

Рабочая программа дисциплины составлена в соответствии с требованиями ФГОС ВО по направлению подготовки (специальности) 09.03.02 «Информационные системы и технологии».

Составитель(и):

доцент Добрынин Алексей Сергеевич (кафедра автоматизации и информационных систем).

Рабочая программа дисциплины рассмотрена и утверждена на заседании кафедры.

Согласована:
Начальник РЦС С/
начальник Регионального центра сопровождения Сибирь ООО Евразтехника _____ Кириенко Е.В.
должность, степень, звание представителя работодателя — подпись — инициалы, фамилия

Приложение

Тестовые задания для входного контроля знаний обучающихся

Выберите один правильный ответ.

1. Тройками из нулей и единиц можно закодировать … различных символов.

a) 6;

b) 8;

c) 5;

d) 9.

2. Архитектура компьютера — это:

a) техническое описание деталей устройств компьютера;

b) описание устройств для ввода-вывода информации;

c) описание программного обеспечения для работы компьютера;

d) описание устройства и принципов работы компьютера, достаточное для понимания пользователя.

3. Что такое микропроцессор?

a) Интегральная микросхема, которая выполняет поступающие на ее вход команды (например, вычисление) и управляет работой машины;

b) устройство для хранения той информации, которая часто используется в работе;

c) устройство для вывода текстовой или графической информации;

d) устройство для ввода алфавитно-цифровых данных.

4. Подключение отдельных периферийных устройств компьютера к магистрали на физическом уровне возможно:

a) с помощью драйвера;

b) с помощью контроллера;

c) без дополнительного устройства;

d) с помощью утилиты.

5. Внешняя память необходима для:

a) для хранения часто изменяющейся информации в процессе решения задачи;

b) для долговременного хранения информации после выключения компьютера;

c) для обработки текущей информации;

d) для постоянного хранения информации о работе компьютера.

6. Для построения с помощью компьютера сложных чертежей в системах автоматизированного проектирования используют:

a) плоттер;

b) графический планшет (дигитайзер);

c) сканер;

d) джойстик.

7. К устройствам накопления информации относится:

a) принтер;

b) В) процессор;
c) ПЗУ;
d) ВЗУ.
8. Что из перечисленного не относится к программным средствам?
a) Системное программирование;
b) драйвер;
c) процессор;
d) текстовые и графические редакторы.
9. Файлом называется:
a) набор данных для решения задачи;
b) поименованная область на диске или другом машинном носителе;
c) программа на языке программирования для решения задачи;
d) нет верного ответа.
10. В каком файле может храниться рисунок?
a) TEST.EXE;
b) ZADAN.TXT;
c) COMMAND.COM;
d) CREML.BMP.
11. Могут ли два каталога 2-го уровня иметь одинаковые имена?
a) Нет;
b) да;
c) да, если они принадлежат разным каталогам 1-го уровня;
d) затрудняюсь ответить.
12. Необходимым компонентом операционной системы является:
a) оперативная память;
b) командный процессор;
c) центральный процессор;
d) файл конфигурации системы.
13. Что такое система счисления?
a) Цифры 1, 2, 3, 4, 5, 6, 7, 8, 9;
b) правила арифметических действий;
c) компьютерная программа для арифметических вычислений;
d) это знаковая система, в которой числа записываются по определенным правилам, с помощью знаков некоторого алфавита, называемых цифрами.
14. Какие системы счисления не используются специалистами для взаимодействия с ЭВМ?
a) Десятичная;
b) троичная;
c) двоичная;
d) шестнадцатеричная.
15. Что называется основанием системы счисления?
a) Количество цифр, используемых для записи чисел;
b) отношение значений единиц соседних разрядов;

c) арифметическая основа ЭВМ;

d) сумма всех цифр системы счисления.

16. Переведите число 27 из десятичной системы счисления в двоичную.

a) 11011;

b) 1011;

c) 1101;

d)

e) 11111.

17. Алгоритм — это:

a) некоторые истинные высказывания, которые должны быть направлены на достижение поставленной цели;

b) отражение предметного мира с помощью знаков и сигналов, предназначенное для конкретного исполнителя;

c) понятное и точное предписание исполнителю совершить последовательность действий, направленных на решение поставленной задачи или цели;

d) инструкция по технике безопасности.

18. Какой тип алгоритма должен быть выбран при решении квадратного уравнения?

a) Линейный;

b) циклический;

c) разветвляющийся;

d) циклически-разветвляющийся.

19. Разветвляющийся алгоритм — это:

a) присутствие в алгоритме хотя бы одного условия;

b) набор команд, которые выполняются последо-; вательно друг за другом;

c) многократное исполнение одних и тех же действий;

d) другое.

20. Что такое сетевой протокол?

a) Соглашение о способе обмена информацией;

b) файл на сервере;

c) устройство связи в сети;

d) сетевая программа.

Задания и вопросы для текущего контроля знаний обучающихся по разделам (темам) учебной дисциплины

Раздел 1 Введение в большие данные

1 Устный опрос на лекции

Перечень вопросов для устного опроса:

1 Что такое Data Mining?

2 Какие прикладные инструменты для работы с Big Data вы знаете?

3 Опишите технологию Map Reduce.

4 Опишите технологию Apache Hadoop.

5 Опишите технологию Apache Spark.

6 Что такое база данных (БД)?

7 Опишите принцип построения облачных платформ и распределенных систем.

8 Какие технологии сбора и обмена информацией в системах больших данных вы знаете?

9 Какие современные СУБД и хранилища данных вам известны?

10 Что такое Python (общее описание и область применения)?

2 Выполнение практических работ

Задания для практических работ, теоретический и методический материал для самостоятельного изучения и вопросы для собеседования представлены в методических указаниях к практическим занятиям.

Варианты заданий по разделу «Введение в большие данные».

Тема 1.1. Понятие Data Mining.

1 Решение задачи сбора данных из реляционной СУБД (MS SQL Server, mySQL).

2 Решение задачи сбора данных из нереляционной СУБД (MongoDB).

3 Решение задачи получения сводных данных из реляционного хранилища данных.

4 Решение задачи получения сводных данных из нереляционного хранилища.

5 Формирование агрегирующих запросов.

Тема 1.2. Прикладные инструменты для работы с Big Data.

1 Решите задачу сбора больших данных с использованием распределенного сервиса обмена сообщениями Apache Kafka.

2 Решите задачу сбора больших данных с использованием распределенного сервиса обмена сообщениями RabbitMQ.

3 Выполните сбор данных с использованием стримингового сервиса.

4 Выполните сбор данных из нескольких источников.

5 Настройте инструмент RabbitMQ для работы с несколькими источниками данных.

Тема 1.3. Технология MapReduce.

1 Решите задачу параллельной обработки данных, на основе потоковых вычислений.

2 Решите задачу параллельной обработки данных, на основе массово-векторных вычислений.

3 Решите задачу параллельной обработке данных в парадигме «map – shuffle - reduce».

4 Реализуйте концепцию «map – shuffle - reduce» на языке программирования высокого уровня.

5 Выполните обработку информации с использованием технологии «map – shuffle - reduce».

6 Решение типовой задачи обработки больших данных на основе парадигмы «Map Reduce». Задача word count problem.

7 Решение типовой задачи обработки больших данных на основе парадигмы «Map Reduce». Задача iris classification problem.

8 Решение типовой задачи обработки больших на основе парадигмы «Map Reduce». Задача выживания пассажиров Титаника.

Тема 1.4. Apache Hadoop.

1 Выполните настройку Apache Hadoop.

2 Выполните настройку Apache Hadoop Yarn.

3 Выполните конфигурирование файловой системы в экосистеме Hadoop.

4 Решите задачу загрузки больших данных в систему Apache Hadoop.

5 Решите задачу обработки больших данных в системе Apache Spark.

6 Изучение базовых элементов и основных статистических функций пакета Apache Hadoop.

7 Подготовка, формирование и предварительная обработка больших данных с использованием концепции «трубопровод» (pipeline).

8 Установка и развертывание фреймворка для работы с большими данными (Apache Hadoop) в облачной среде Google Colab. Монтирование HDFS.

9 Формирование универсальных коллекций Apache Hadoop.

10 Обработка универсальных коллекций Apache Hadoop.

Тема 1.5. Apache Spark.

1 Выполните установку фреймворка для обработки больших данных Apache Spark.

2 Выполните настройку фреймворка для обработки больших данных Apache Spark.

3 Решите задачу загрузки больших данных в систему Apache Spark.

4 Решите задачу обработки больших данных в Apache Spark.

5 Решите задачу статистического анализа больших данных в Apache Spark.

6 Установка и развертывание фреймворка для работы с большими данными (Apache Spark) в облачной среде Google Colab. Монтирование RDD.

7 Компоненты и подсистемы машинного обучения на больших данных (Big Data) в технологиях Hadoop, Spark.

8 Решение задачи многомерной математической регрессии с использованием фреймворка Apache Hadoop.

9 Решение задачи многомерной математической регрессии с использованием фреймворка Apache Spark.

10 Изучение базовых элементов и основных статистических функций пакета Apache Spark.

11 Изучение универсальных коллекций Apache Spark.

3 Выполнение домашнего задания

Примерный перечень тем домашних заданий:

1 Решите задачу анализа данных методами элементарной статистики.

2 Опишите основные библиотеки Spark для работы с данными и их назначение.

3 Опишите основные инструменты, которые используются для анализа данных.

4 Решите задачу анализа данных методами машинного обучения.

5 Решите задачу анализа данных с использованием искусственных нейронных сетей.

6 Опишите основные роли и задачи ученого по данным.

7 Способы анализа данных в Spark.

8 Опишите основные этапы развертывания локального кластера Apache Hadoop.

9 Опишите принцип работы параллельной технологии (MPP) на примере редактирования текста с использованием регулярных выражений (на базе подсчета текста).

10 Выполните по статистик в тексте по эвристическим правилам.

4 Тестирование по результатам освоения лекционного материала и самостоятельной работы

Тестирование проводится для определения степени усвоения материалов лекций, а также выполнения самостоятельной работы в рамках часов, отведенных на самостоятельную работу обучающихся, в СУО «Moodle». Примерный перечень вопросов показан ниже. Тест закрытого типа.

Вопрос 1. Для машинного обучения подходят данные:

а) Предварительно подготовленные, очищенные от ошибок, пропусков и выбросов, а также нормализованные и представленные в виде числовых векторов;

б) Бинарные;

в) Любых форматов в цифровом виде;

г) Числовые типа int.

Вопрос 2. Выберите технологию потоковой обработки событий в режиме реального времени:

а) Apache Kafka;

б) Spark Streaming;

в) Apache Hadoop;

Вопрос 3. Apache NiFi используется для:

а) эффективного хранения больших данных;

б) оптимизации SQL-запросов к DWH;

в) маршрутизации потоков Big Data и построения ETL-конвейеров;

г) визуализации результатов аналитики.

Вопрос 4. Для распределенного глубокого машинного обучения (Deep Learning) больше подходит фреймворк:

а) PyTorch;

б) TensorFlow;

в) Flask;

г) Scikit-learn.

Вопрос 5. Для полнотекстового интеллектуального поиска и аналитики по полуструктурированным данным в формате JSON отлично подходит СУБД:

а) Hive;

б) Elasticsearch;

в) Cassandra;

г) HBase.

Вопрос 6. Анализировать данные, хранящиеся в Apache Hadoop, с помощью стандартного инструментария SQL-запросов.

а) Можно;

б) Нельзя.

Вопрос 7. Повысить производительность Apache Kafka можно с помощью:

а) Повышения коэффициента репликации;

б) Увеличения размера сообщений;

в) Замены HDD-дисков на SSD;

г) Увеличения плотности разделов на каждом брокере.

Вопрос 8. Формат Parquet считается:

а) строковым;

б) неструктурированным;

в) колоночным (столбцовым);

г) полуструктурированным.

Вопрос 9. Для реализации микросервисной архитектуры и интеграции разрозненных систем подходит:

а) Apache Spark;

б) Apache Kafka;

в) Apache AirFlow;

г) Apache Hadoop.

Вопрос 10. Автоматизировать запуск пакетных задач в рамках конвейера обработки больших данных по расписанию можно с помощью:

а) Apache Kafka;

б) Apache AirFlow;

в) Apache Hive;

г) Apache Hadoop.

Вопрос 11. Что такое SQL?

а) Реляционная база данных;

б) Язык неструктурированных запросов;

в) Язык структурированных запросов;

г) Средство для создания пайплайнов.

Вопрос 12. Одна из главных целей Big Data – это…

а) привлечение все больше пользователей;

б) рост числа обработок данных;

в) снижение издержек операций;

г) таргетирование пользователей.

Вопрос 13. Закончите следующее предложение: С точки зрения машины, информация становится структурированной, если...

а) машина проинструктирована, каким образом её обрабатывать;

б) информация разделена на части и озаглавлена;

в) информация имеет логическую взаимосвязь внутри себя;

г) машина знает из каких частей состоит информация.

Вопрос 14. Отметьте причину создания NoSQL баз данных…

а) высокая стоимость горизонтальной масштабируемости RDBMS при сохранении требования высокой доступности;

б) недостаточная гибкость языка запросов SQL;

в) невозможность хранить большие объёмы;

г) данных дороговизна лицензий RDBMS;

Вопрос 15. Искусственные нейронные сети (ИНС) — модели машинного обучения, использующие комбинации распределенных простых операций, зависящих от обучаемых параметров, для обработки входных данных. Какого вида ИНС не существует?

а) Наивные;

б) Рекуррентные;

в) Импульсные;

г) Противоборствующие.

Раздел 2 Технологии анализа данных

1 Устный опрос на лекции

Перечень вопросов для устного опроса:

1 Что такое жизненный цикл анализа больших данных?

2 Какие существую стандарты при анализе больших данных?

3 Что такое когнитивный анализ данных?

4 Виды визуализация больших данных?

5 Что такое индустриальный жизненный цикл?

6 Основные технологии анализа данных.

7 Возможности анализа больших данных.

8 Обработка больших данных.

9 Алгоритмы обработки больших данных.

10 Комплексная оценка больших данных.

2 Выполнение практических работ

Задания для практических работ, теоретический и методический материал для самостоятельного изучения и вопросы для собеседования представлены в методических указаниях к практическим занятиям.

Варианты заданий по разделу «Технологии анализа данных».

Тема 1.1. Жизненный цикл анализа больших данных.

1 Опишите этапы подготовки и анализа больших данных.

2 Решите задачу обработки больших данных, с использованием фреймворка Apache Hadoop или Apache Spark.

3 Выполните сбор больших данных с использованием стриминговой технологии Kafka или RabbitMQ и обработки их.

4 Выполните предварительную подготовку данных, собранных с использованием стриминговых технологий (Kafka, RabbitMQ).

5 Сформируйте DWH, на основе данных сервиса Kafka.

6 Сформируйте DWH, на основе данных сервиса RabbitMQ.

7 Решите задачу обработки больших данных, по выбору преподавателя.

9 Выполните обработку больших данных, в соответствии с заданным регулярным выражением.

10 Опишите все этапы жизненного цикла больших данных.

Тема 1.2. Статистический анализ больших данных.

1 Решите задачу определения основных статистических параметров для больших данных.

2 Решите задачу определения статистических моделей на основе больших данных.

3 Решите задачу машинного обучения на основе больших данных.

4 Решите задачу статистической обработки больших данных.

5 Ознакомление с возможностями Apache Spark для анализа данных.

6 Реализуйте механизм подготовки и отображения данных в Apache Spark. Возможности ввода/вывода.

7 Решите задачу на больших графах.

8 Выполните анализ данных в Apache Spark. Получите первичные элементарные характеристики о наборах данных (элементарные статистики). Способы импорта/экспорта данных.

9 Реализуйте корреляционный анализ и регрессионный анализ данных с использованием встроенных средств Apache Spark.

10 Решите задачу комплексной оценки данных, с использованием Apache Spark.

Тема 1.3. Визуализация больших данных.

1 Настройте инструмент Power BI для визуализации данных.

2 Настройте инструмент Tableau для визуализации данных.

3 Постройте диаграмму на основе коллекций с данными HDFS.

4 Постройте диаграмму на основе коллекций с данными RDD

5 Постройте диаграммы и графики Spark с использованием временных отсоединенных коллекции RDD (RDD – Memory Allocations)

3 Выполнение домашнего задания

Примерный перечень тем домашних заданий:

1 Опишите форматы хранения структурированных и неструктурированных данных.

2 Механизмы обработки данных в формате xml.

3 Механизмы обработки данных в формате json.

4 Опишите системы хранения и обработки больших данных.

5 Опишите основные задачи больших данных (Big Data).

6 Способы анализа данных в Python. Получение первичных элементарных характеристик о наборах данных.

7 Развертывание локального кластера Apache Hadoop.

8 Решение задач Data Mining.

9 Задачи классификации, кластеризации: деревья решений, RandomForest, k-means.

10 Развертывание локального кластера Apache Hadoop. Подсчёт слов в тексте, с помощью MapReduce.

4 Тестирование по результатам освоения лекционного материала и самостоятельной работы

Тестирование проводится для определения степени усвоения материалов лекций, а также выполнения самостоятельной работы в рамках часов, отведенных на самостоятельную работу обучающихся, в СУО «Moodle». Примерный перечень вопросов показан ниже. Тест открытого типа.

1 Наиболее распространенная технология хранения и обработки больших данных называется _____

2 Язык запросов, который используется в реляционных СУБД называется _____

3 Технология запросов, используемая в нереляционных БД (на примере MongoDB) называется _____

4 Основная (базовая) библиотека анализа данных языка Python носит название _____

5 Библиотека языка программирования Python, используемая для искусственных нейронных сетей называется _____

6 Основной инструмент анализа данных от компании Google носит название _____

7 Основной инструмент анализа данных в составе компании Anaconda носит название _____

8 Модель прогнозирования для задач классификации носит название_____.

9 Модель прогнозирования для линейных задач называется_____.

10. Наиболее распространённый язык программирования в задачах анализа данных носит название _____.

5 Вопросы к экзамену для промежуточной аттестации

1 Что такое Data Mining?

2 Какие прикладные инструменты для работы с Big Data вы знаете?

3 Опишите технологию MapRaduce.

4 Опишите технологию Apache Hadoop.

5 Опишите технологию Apache Spark.

6 Что такое жизненный цикл анализа больших данных?

7 Какие существую стандарты при анализе больших данных?

8 Что такое когнитивный анализ данных?

9 Виды визуализация больших данных?

10 Что такое индустриальный жизненный цикл?

11 Опишите общие принципы и инструменты аналитики.

12 Какие бывают шкалы измерений?

13 Что такое генеральная совокупность и выборка?

14 Что такое нормальное распределение?

15 Что такое уровень статистической достоверности?

16 Что такое дескриптивный анализ?

17 Что такое Data Mining?

18 Какие прикладные инструменты для работы с Big Data вы знаете?

19 Опишите технологию MapRaduce.

20 Опишите технологию Apache Hadoop.

21 Опишите технологию Apache Spark.

22 Что такое база данных (БД)?

23 Опишите принцип построения облачных платформ и распределенных систем.

24 Какие технологии сбора и обмена информацией в системах больших данных вы знаете?

25 Какие современные СУБД и хранилища данных вам известны?

26 Что такое Python (общее описание и область применения)?

27 Что такое жизненный цикл анализа больших данных?

28 Какие существую стандарты при анализе больших данных?

29 Что такое когнитивный анализ данных?

30 Виды визуализация больших данных?

31 Что такое индустриальный жизненный цикл?

32 Основные технологии анализа данных.

33 Возможности анализа больших данных.

34 Обработка больших данных.