

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное  
образовательное учреждение высшего образования  
«Астраханский государственный университет имени В. Н. Татищева»  
(Астраханский государственный университет им. В. Н. Татищева)

СОГЛАСОВАНО  
Руководитель ОПОП

Ю.А. Головки

«13» июня 2024 г.

УТВЕРЖДАЮ  
Заведующий кафедрой информационных  
технологий

А.Н. Марьенков

«13» июня 2024 г.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)**

**«Большие данные»**

Составитель(и)

**Кузнецова В. Ю., к.т.н., доцент кафедры ИБ;**

Направление подготовки /  
специальность

**09.03.02 ИНФОРМАЦИОННЫЕ СИСТЕМЫ И  
ТЕХНОЛОГИИ**

Направленность (профиль) ОПОП

**ТЕХНОЛОГИИ РАЗРАБОТКИ И  
АДМИНИСТРИРОВАНИЯ  
ИНФОРМАЦИОННЫХ СИСТЕМ**

Квалификация (степень)

**бакалавр**

Форма обучения

**очно-заочная**

Год приёма

**2022**

Курс

**4**

Семестр(ы)

**8**

## 1. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

**1.1. Целями освоения дисциплины (модуля) «Большие данные»** являются изучение математических методов и моделей, используемых в системах обработки и анализа больших данных для поддержки принятия решений, и развитие профессиональных навыков в этой области.

**1.2. Задачи освоения дисциплины (модуля): «Большие данные»**

- сформировать представление о проблемах анализа и обработки данных;
- сформировать навыки разработки алгоритмов анализа и обработки данных с применением моделей Data Mining.

## 2. МЕСТО ДИСЦИПЛИНЫ (МОДУЛЯ) В СТРУКТУРЕ ОПОП

**2.1. Учебная дисциплина (модуль) «Большие данные»** относится к обязательной части и осваивается в 8 семестре.

**2.2. Для изучения данной учебной дисциплины (модуля) необходимы следующие знания, умения, навыки, формируемые предшествующими учебными дисциплинами (модулями):**

– Информатика

Знания: базовые понятия информатики и вычислительной техники; понятие информационной системы и информационной технологии; технические и программные средства реализации информационных процессов; основные устройства, входящие в состав ЭВМ, их назначение и характеристики; формы представления и преобразования информации в компьютере.

Умения: применять вычислительную технику для решения практических задач; разработать алгоритм поставленной задачи.

Навыки: работы на персональном компьютере.

– Основы программирования

Знания: основные структуры данных, используемые в языках программирования; структуру программ; основные принципы алгоритмизации.

Умения: создавать схему алгоритма для задачи; проводить отладку и тестирование созданного программного продукта.

Навыки в области алгоритмизации, разработки, отладки и тестирования программных продуктов.

**2.3. Последующие учебные дисциплины (модули) и (или) практики, для которых необходимы знания, умения, навыки, формируемые данной учебной дисциплиной (модулем):**

– *Интеллектуальные системы и технологии;*

- *Производственная практика;*

- *Выпускная квалификационная работа.*

## 3. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Процесс освоения дисциплины (модуля) направлен на формирование элементов следующей компетенции в соответствии с ФГОС ВО и ОПОП ВО по данному направлению подготовки / специальности:

*а) общепрофессиональной(ых) (ОПК):*

ОПК-8 Способен применять математические модели, методы и средства проектирования информационных и автоматизированных систем

ПК-4 Способен проводить анализ требований к программному обеспечению, выполнять работы по проектированию программного обеспечения

**Таблица 1 – Декомпозиция результатов обучения**

Код и наименование компетенции	Планируемые результаты обучения по дисциплине (модулю)		
	Знать (1)	Уметь (2)	Владеть (3)
ОПК-8 Способен применять математические модели, методы и средства проектирования информационных и автоматизированных систем	ИОПК-8.1.1 Знать: методологию и основные методы математического моделирования, классификацию и условия применения моделей, основные методы и средства проектирования информационных и автоматизированных систем, инструментальные средства моделирования и проектирования информационных и автоматизированных систем.	ИОПК-8.2.1 Уметь: применять на практике математические модели, методы и средства проектирования и автоматизации систем на практике.	ИОПК-8.3.1 Владеть: навыками моделирования и проектирования информационных и автоматизированных систем.
ПК-4 Способен проводить анализ требований к программному обеспечению, выполнять работы по проектированию программного обеспечения	ПК-4.1. Знать методы проведения анализа и разработки требований к программному обеспечению.	ПК-4.2. Уметь выполнять работы по проектированию программного обеспечения	ПК4.3. Владеть методами проведения анализа требований к программному обеспечению, выполнять работы по проектированию программного обеспечения

#### 4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Объём дисциплины (модуля) составляет 5 зачётных единиц, в том числе 36 часов, выделенных на контактную работу обучающихся с преподавателем (из них 18 часов – лекции, 18 часов – лабораторные работы), и 144 часа – на самостоятельную работу обучающихся.

**Таблица 2 – Структура и содержание дисциплины (модуля)**

Раздел, тема дисциплины (модуля)	Семестр	Контактная работа (в часах)			Самост. работа		Форма текущего контроля успеваемости, форма промежуточной аттестации
		Л	ПЗ	ЛР	КР	СР	
Большие данные и экосистема больших данных	8	2	0	4	0	18	Лабораторная работа 1, Устный опрос Устный опрос на экзамене
Процесс исследования данных		2	0	2	0	18	Лабораторная работа 2, Устный опрос Устный опрос на экзамене
Машинное обучение и математические основы работы с данными		2	0	2	0	18	Лабораторная работа 3, Устный опрос Устный опрос на экзамене
Описательная статистика		2	0	2	0	18	Лабораторная работа, Устный опрос Устный опрос на экзамене
Теория вероятностей при работе с большими данными		2	0	2	0	18	Лабораторная работа, Устный опрос Устный опрос на экзамене
Проверка гипотез при анализе данных		2	0	2	0	18	Лабораторная работа, Устный опрос Устный опрос на экзамене
Визуализация больших данных		2	0	2	0	18	Лабораторная работа, Устный опрос Устный опрос на экзамене
Итоговый проект		4	0	2	0	18	Лабораторная работа, Устный опрос Устный опрос на экзамене
<b>Итого</b>		<b>18</b>	<b>0</b>	<b>18</b>	<b>0</b>	<b>144</b>	<b>Экзамен</b>

**Таблица 3 – Матрица соотнесения разделов, тем учебной дисциплины (модуля) и формируемых компетенций**

Раздел, тема дисциплины (модуля)	Кол-во часов	ОПК-8	ПК-4	Общее количество компетенций
Большие данные и экосистема больших данных	24	+	+	2
Процесс исследования данных	22	+	+	2
Машинное обучение и математические основы работы с данными	22	+	+	2
Описательная статистика	22	+	+	2
Теория вероятностей при работе с большими данными	22	+	+	2
Проверка гипотез при анализе данных	22	+	+	2

Раздел, тема дисциплины (модуля)	Кол-во часов	ОПК-8	ПК-4	Общее количество компетенций
Визуализация больших данных	24	+	+	2
Итоговый проект	26	+	+	2
<b>Итого</b>	<b>180</b>			<b>2</b>

### Краткое содержание каждой темы дисциплины (модуля)

#### Тема 1. Большие данные и экосистема больших данных.

Введение в анализ больших данных. Основные определения, термины, задачи анализа больших данных. Экосистема аналитики больших данных. Распределенные файловые системы.

#### Тема 2. Процесс исследования данных

Методы и методики процесса исследования данных, управление проектами в сфере аналитики данных.

#### Тема 3. Машинное обучение и математические основы работы с данными

Машинное обучение на больших данных. Обзор источников информации для Big Data. Методики сбора данных.

#### Тема 4. Описательная статистика

Модели данных. Подготовка исходных данных для анализа: первичная обработка и визуализация имеющихся данных. Описательная статистика имеющихся наборов данных и инструменты для этого.

#### Тема 5. Теория вероятностей при работе с большими данными

Событие. Случайная величина. Центральная предельная теорема. Распределение вероятностей. Закон больших чисел. Применение при работе с большими данными.

#### Тема 6. Проверка гипотез при анализе данных

Статистические тесты. Дисперсионный анализ. Критерий Хи-квадрат. Технические детали.

#### Тема 7. Визуализация больших данных

Создание дашбордов. Построение информационных панелей с использованием современных инструментов

#### Тема 8. Визуализация больших данных.

Анализ технического задания на проведение исследование. Интервьюирование заказчика. Выполнение самостоятельного рабочего проекта. Составление отчёта по проекту

## 5. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ПРЕПОДАВАНИЮ И ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

### 5.1. Указания для преподавателей по организации и проведению учебных занятий по дисциплине (модулю)

Учебная деятельность студента в процессе изучения строится из контактных форм работы с преподавателем (аудиторные занятия, экзамен) и самостоятельной работы.

Для успешного освоения дисциплины является обязательным посещение всех занятий, выполнение самостоятельной работы, которая назначается преподавателем.

Методическая поддержка дисциплины обеспечивается использованием дистанционных технологий. Студентам предлагается информационный ресурс, расположенный по адресу: <http://moodle.asu.edu.ru>, на сервере дистанционного обучения АГУ им. В.Н. Татищева. На сервере размещен методический материал по данной дисциплине, в содержание которого входит:

- теоретический материал;
- задания и указания по выполнению лабораторных работ.

Аудиторные занятия проводятся на основе теоретического материала, опубликованного на образовательном портале, это позволяет студентам изучить пропущенный материал или самостоятельно разобраться с темой, не освоенной на занятии.

### 5.2. Указания для обучающихся по освоению дисциплины (модулю)

В рамках дисциплины «Большие данные» предполагается организация следующих видов самостоятельной работы студентов:

- работа с учебно-методическим информационным обеспечением;
- подготовка к лабораторным работам, подготовка отчетов.

В качестве форм и методов контроля внеаудиторной самостоятельной работы используются: электронные отчеты по выполнению лабораторных работ; устный опрос.

**Таблица 4 – Содержание самостоятельной работы обучающихся**

Вопросы, выносимые на самостоятельное изучение	Кол-во часов	Форма работы
Тема 1. Большие данные и экосистема больших данных. Элементы экосистемы больших данных.	18	Отчет о выполнении ЛР 1 Устный опрос на экзамене
Тема 2. Процесс исследования данных. Apache Hadoop и его элементы. Межотраслевой стандартный процесс для исследования данных (CRISP-DM). Элементы процесса исследования данных. Как работает MapReduce. Пример кода (псевдокода).	18	Отчет о выполнении ЛР 2 Устный опрос на экзамене
Тема 3. Машинное обучение и математические основы работы с данными. Типы задач машинного обучения. Примеры алгоритмов.	18	Отчет о выполнении ЛР 3 Устный опрос на экзамене
Тема 4. Работа с большими наборами данных. MapReduce. Как работает MapReduce. Пример кода (псевдокода).	18	Отчет о выполнении ЛР 4 Устный опрос на экзамене
Тема 5. Hadoop и Spark. Подготовка данных в Spark. Работа со Spark в интерактивном режиме. Сохранение данных в Hive. Построение интерактивного отчета и дашборда с помощью платформы Qlik Sense.	18	Отчет о выполнении ЛР 5 Устный опрос на экзамене
Тема 6. Базы данных NoSQL. Использование Neo4j. Выполнение запросов на Cypher. Создание узла	18	Отчет о выполнении ЛР 6

Вопросы, выносимые на самостоятельное изучение	Кол-во часов	Форма работы
пользователя в базе данных Neo4j. Визуализация найденных закономерностей, построение графов.		Устный опрос на экзамене
Тема 7. Интеллектуальный анализ текста как источника больших данных. Алгоритм выделения основы. Классификация с использованием наивного баесовского классификатора и дерева принятия решений. Тренировка и оценка моделей. Визуализация выявленных закономерностей.	18	Отчет о выполнении ЛР 7 Устный опрос на экзамене
Тема 8. Визуализация больших данных. Построение графиков и диаграмм с использованием библиотеки dc.js. Создание информационной панели в браузере. Работа с библиотекой визуализации данных d3.js.	18	Отчет о выполнении ЛР 8 Устный опрос на экзамене

### **5.3. Виды и формы письменных работ, предусмотренных при освоении дисциплины (модуля), выполняемые обучающимися самостоятельно**

В процессе обучения студенты выполняют лабораторные работы. Результатом работы, выполняемой обучающимися, является электронный отчет по выполнению лабораторной работы.

Электронный отчет представляет собой файл формата doc, docx или pdf, содержащий программный код, результаты выполнения программы и текстовые пояснения. Файл передается на проверку преподавателю путем загрузки на ресурс <http://moodle.asu.edu.ru> в соответствующий заданию раздел.

Задания к лабораторным занятиям размещены на образовательном портале <http://moodle.asu.edu.ru>. Рекомендуется заранее ознакомиться с темой, основными вопросами и рекомендациями.

В процессе подготовки к лабораторным работам, необходимо обратить особое внимание на самостоятельное изучение рекомендованной литературы. Самостоятельная работа с учебниками, учебными пособиями, научной, справочной литературой, материалами периодических изданий и Интернета является наиболее эффективным методом получения дополнительных знаний, позволяет значительно активизировать процесс овладения информацией, способствует более глубокому усвоению изучаемого материала.

## **6. ОБРАЗОВАТЕЛЬНЫЕ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ**

### **6.1. Образовательные технологии**

В рамках реализации компетентностного подхода в соответствии с требованиями ФГОС ВО в учебном процессе предусмотрены активные и интерактивные формы проведения занятий. Основой для выстраивания аудиторных занятий служит технология развития критического мышления, которая, интегрируя элементы проблемного, проектного, дискуссионного обучения, позволяет достигать максимальной эффективности в достижении проектируемых компетенций.

Цели дисциплины достигаются путем сочетания контактной и самостоятельной работы студентов: лабораторных занятий на ПК и организации самостоятельной работы студентов.

Лабораторные работы выполняются студентами с применением ПК и ориентированы на формирование деятельностных компетентностей. Они заключаются в выполнении сквозного цикла лабораторных работ. В процессе выполнения лабораторных работ достигаются следующие цели:

- изучается экосистема работы с большими данными;

- формируются практические навыки работы с большими массивами информации при решении конкретных практических задач;
- формируется навык выявления ошибочных и нестандартных ситуаций и реагирования на них.

На лабораторных занятиях студент вначале знакомится с содержанием работы, пользуясь электронными методическими материалами, размещенными на <http://moodle.asu.edu.ru>, затем выполняет задание и показывает результаты преподавателю. Лабораторные работы, выполняются студентом самостоятельно, возникающие при их выполнении проблемы разрешаются в рамках учебного времени и индивидуальных и групповых консультаций. Для выставления баллов по итогам выполнения ЛР, студенты прикрепляют файлы с выполненными работами и отчеты на образовательный портал.

Для самостоятельного изучения теоретического материала дисциплины рекомендуется использовать интернет-ресурсы, информационные базы, методические разработки, специальную учебную и научную литературу.

В рамках организации самостоятельной работы студентам рекомендуется:

- работа с лекционным материалом;
- дополнительная подготовка к лабораторным работам или выполнение части лабораторной работы, которую они не успели сделать в аудитории;
- подготовка к текущей и промежуточной аттестации (экзамену).

**Таблица 5 – Образовательные технологии, используемые при реализации учебных занятий**

Раздел, тема дисциплины (модуля)	Форма учебного занятия		
	Лекция	Практическое занятие, семинар	Лабораторная работа
Большие данные и экосистема больших данных	Обзорная лекция с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Процесс исследования данных	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Машинное обучение и математические основы работы с данными	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Описательная статистика	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Теория вероятностей при работе с большими данными	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Проверка гипотез при анализе данных	Лекция-презентация с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle
Визуализация больших данных	Лекция-диалог с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы,

			отчет в Moodle
Итоговый проект	Лекция-диалог с применением ВКС	Не предусмотрено	Демонстрация выполнения лабораторной работы, отчет в Moodle

### 6.2. Информационные технологии

При реализации учебной и внеучебной работы используются следующие информационные технологии:

- образовательный портал <http://moodle.asu.edu.ru> (размещение учебно-методического материала, публикация заданий для предоставления студентами выполненных отчетов по всем видам работ, ознакомление учащихся с оценками и т.д., размещение объявлений, обсуждение вопросов в форуме и т.д.), как элемента интерактивного взаимодействия участников образовательного процесса (технологии дистанционного обучения);
- среда разработки моделей машинного обучения <https://colab.research.google.com>
- веб-сервис для хостинга IT-проектов и их совместной разработки <https://github.com>
- ресурсы ЭБС и сети Internet, как источников информации.

### 6.3. Программное обеспечение, современные профессиональные базы данных и информационные справочные системы

#### 6.3.1. Программное обеспечение

Наименование программного обеспечения	Назначение
Adobe Reader	Программа для просмотра электронных документов
LMS Moodle	Образовательный портал ФГБОУ ВО «АГУ»
Microsoft Office	Пакет офисных программ
OpenOffice	Пакет офисных программ
7-zip	Архиватор
Microsoft Windows	Операционная система
Kaspersky Endpoint Security	Средство антивирусной защиты
Google Chrome	Браузер
Opera	Браузер
Anaconda Navigator	Графический интерфейс для работы с библиотеками Python

#### 6.3.2. Современные профессиональные базы данных и информационные справочные системы

1. Электронная библиотека «Астраханский государственный университет» собственной генерации на платформе ЭБС «Электронный Читальный зал – БиблиоТех». <https://biblio.asu.edu.ru>.
2. Электронно-библиотечная система (ЭБС) ООО «Политехресурс» «Консультант студента». <https://www.studentlibrary.ru>.
3. Электронный каталог Научной библиотеки АГУ на базе MARK SQL НПО «Информ-систем». <https://library.asu.edu.ru>.

## 7. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ДЛЯ ПРОВЕДЕНИЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

При проведении текущего контроля и промежуточной аттестации по дисциплине (модулю) «Большие данные» проверяется сформированность у обучающихся компетенций, указанных в разделе 3 настоящей программы. Этапность формирования данных компетенций

в процессе освоения образовательной программы определяется последовательным освоением дисциплин (модулей) и прохождением практик, а в процессе освоения дисциплины (модуля) – последовательным достижением результатов освоения содержательно связанных между собой разделов, тем.

**Таблица 6 – Соответствие разделов, тем дисциплины (модуля), результатов обучения по дисциплине (модулю) и оценочных средств**

Контролируемый раздел, тема дисциплины (модуля)	Код контролируемой компетенции	Наименование оценочного средства
Большие данные и экосистема больших данных	ОПК–8, ПК-4	Отчет по лабораторной работе; Вопросы к экзамену
Процесс исследования данных	ОПК–8, ПК-4	Отчет по лабораторной работе; Вопросы к экзамену
Машинное обучение и математические основы работы с данными	ОПК–8, ПК-4	Отчет по лабораторной работе; Вопросы к экзамену
Описательная статистика	ОПК–8, ПК-4	Отчет по лабораторной работе; Вопросы к экзамену
Теория вероятностей при работе с большими данными	ОПК–8, ПК-4	Отчет по лабораторной работе; Вопросы к экзамену
Проверка гипотез при анализе данных	ОПК–8, ПК-4	Отчет по лабораторной работе; Вопросы к экзамену
Визуализация больших данных	ОПК–8, ПК-4	Отчет по лабораторной работе; Вопросы к экзамену
Итоговый проект	ОПК–8, ПК-4	Отчет по лабораторной работе; Вопросы к экзамену

## 7.2. Описание показателей и критериев оценивания компетенций, описание шкал оценивания

**Таблица 7 – Показатели оценивания результатов обучения в виде знаний**

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует глубокое знание теоретического материала, умение обоснованно излагать свои мысли по обсуждаемым вопросам, способность полно, правильно и аргументированно отвечать на вопросы, приводить примеры
4 «хорошо»	демонстрирует знание теоретического материала, его последовательное изложение, способность приводить примеры, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует неполное, фрагментарное знание теоретического материала, требующее наводящих вопросов преподавателя, допускает существенные ошибки в его изложении, затрудняется в приведении

Шкала оценивания	Критерии оценивания
	примеров и формулировке выводов
2 «неудовлетворительно»	демонстрирует существенные пробелы в знании теоретического материала, не способен его изложить и ответить на наводящие вопросы преподавателя, не может привести примеры

**Таблица 8 – Показатели оценивания результатов обучения в виде умений и владений**

Шкала оценивания	Критерии оценивания
5 «отлично»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы
4 «хорошо»	демонстрирует способность применять знание теоретического материала при выполнении заданий, последовательно и правильно выполняет задания, умеет обоснованно излагать свои мысли и делать необходимые выводы, допускает единичные ошибки, исправляемые после замечания преподавателя
3 «удовлетворительно»	демонстрирует отдельные, несистематизированные навыки, испытывает затруднения и допускает ошибки при выполнении заданий, выполняет задание по подсказке преподавателя, затрудняется в формулировке выводов
2 «неудовлетворительно»	не способен правильно выполнить задания

### **7.3. Контрольные задания и иные материалы, необходимые для оценки результатов обучения по дисциплине (модулю)**

#### **Вопросы к экзамену**

1. Что такое большие данные. Приведите примеры.
2. Элементы экосистемы больших данных.
3. Категории специалистов по аналитике.
4. Межотраслевой стандартный процесс для исследования данных (CRISP-DM). Элементы процесса исследования данных.
5. Что такое машинное обучение. Типы задач машинного обучения. Примеры алгоритмов.
6. Проблемы при работе с большими данными и способы их решения.
7. Как работает MapReduce. Пример кода (псевдокода).
8. Apache Hadoop и его элементы.
9. Apache Spark и его элементы.
10. Базы данных NoSQL. Типы баз данных NoSQL.
11. Принципы BASE для баз данных NoSQL. Графовые базы данных.
12. Методы интеллектуального анализа текста как источника больших данных.
13. Визуализация больших данных.

#### **Тематика и краткое содержание лабораторных работ**

Полная версия лабораторных работ представлена на платформе Moodle.

#### **Лабораторная работа 1**

Большие данные и Интернет Вещей.

Работа с платформой IBM Cloud. Построение аналитических потоков.

#### **Лабораторная работа 2**

Работа с большими наборами данных.

Задача: прогнозирование вредоносных веб-адресов. Использование разреженного представления данных. Использование сжатых данных вместо необработанных. Применение онлайн-алгоритма для прогнозирования. Построение рекомендательной системы внутри базы данных.

Метод k-ближайших соседей. Метод локально-чувствительного хеширования. Метрики расстояния.

### Лабораторная работа 3

Использование Hadoop для аналитики больших данных.

Настройка Vagrant. Работа с подходом Map Reduce.

### Лабораторная работа 4

Использование Hadoop и Spark.

Задача: оценка риска при кредитовании. Работа в среде vagrant на виртуальной машине. Взаимодействие с HDFS. Загрузка и сохранение данных в Hadoop. Подготовка данных в Spark. Работа со Spark в интерактивном режиме. Сохранение данных в Hive.

### Лабораторная работа 5

Машинное обучение на Spark

Настройка виртуальной машины. Распространение переменных по всем узлам кластера. Широковещательные переменные. Аккумуляторные переменные. Предобработка данных в среде Spark. Работа с пропущенными данными. RDD. Работа с объектами DataFrame.

### Лабораторная работа 6

Построение поисковой системы диагностики болезней на базе Elasticsearch.

Сбор данных и индексирование болезней. Диагностика болезней по симптомам. Обработка ошибок и опечаток с помощью метрики расстояния Дамерау-Левенштейна. Задача профилирования болезни через агрегирование значимых терминов. Отображение информации.

Построение рекомендательной системы для связанных данных с использованием графовой базы данных. Задача: построение рекомендаций рецептов блюд на основе предпочтений пользователей и набора ингредиентов. Использование индекса Elasticsearch для чистки данных и для заполнения графовой базы данных. Использование Neo4j. Выполнение запросов на Cypher. Создание узла пользователя в базе данных Neo4j. Визуализация найденных закономерностей, построение графов.

Классификация с использованием наивного байесовского классификатора и дерева принятия решений. Тренировка и оценка моделей. Визуализация выявленных закономерностей.

### Лабораторная работа 7

Создание информационной панели (дашборда).

Использование Crossfilter для фильтрации набора данных с применением MapReduce. Построение графиков и диаграмм с использованием библиотеки dc.js. Создание информационной панели в браузере. Работа с библиотекой визуализации данных d3.js.

### Лабораторная работа 8

Выполнение итогового проекта по реальным данным от заказчика. Анализ технического задания. Составление отчёта.

**Таблица 9 – Примеры оценочных средств с ключами правильных ответов**

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
		<b>ОПК-8</b> Способен применять математические модели, методы и средства проектирования информационных и автоматизированных систем.		

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
1.	Задание закрытого типа	Документ-ориентированной СУБД являются: 1. Oracle Database 2. MS SQL Server 3. Mongo DB 4. SQLite	3	3
2.		Графовой СУБД является 1. MS SQL Server 2. Mongo DB 3. Neo4j 4. SQLite	3	3
3.		Какое значение пропущено в выражении «Матрица с разреженностью больше ... является разреженной матрицей» 1. 0,4 2. 0,5 3. 0,6 4. 0,9	2	3
4.		Какой инструмент визуализации НЕ является облачным? 1. SAP Analytics Cloud, 2. Google Data Studio, 3. Yandex Datalens, 4. Tableau	4	3
5.		Какое свойство НЕ обязательно для больших данных? 1. Ценность данных 2. Структурированность данных 3. Объем 4. Скорость накопления	2	3
6.	Задание открытого типа	Какая матрица называется разреженной?	Разреженная матрица - это матрица, в которой большинство элементов равно нулю. Напротив, таблица, в которой большинство элементов отличны от нуля, называется плотной. Мы определяем разреженность матрицы как число нулевых элементов, деленное на общее количество элементов. Матрица с разреженностью	

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			больше 0,5 является разреженной матрицей.	
7.		Дайте определение усиковой диаграмме.	Усиковая диаграмма - график, использующийся в описательной статистике, компактно изображающий одномерное распределение вероятностей.	5
8.		Что может быть отображено на усиковой диаграмме?	Такой вид диаграммы в удобной форме показывает медиану (или, если нужно, среднее), нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы.	5
9.		Дайте определение пузырьковой диаграмме.	Пузырьковая диаграмма — это разновидность точечной диаграммы, в которой точки данных заменены пузырьками, причем их размер служит дополнительным измерением данных. На пузырьковой диаграмме, как и на точечной, нет оси категорий — и горизонтальная, и вертикальная оси являются осями значений. В дополнение к значениям X и значениям Y, наносимым на точечную диаграмму, на пузырьковой диаграмме показаны также значения Z (размер).	5
10.		В чем отличие пузырьковой диаграммы от точечной.	Вы можете использовать пузырьковую диаграмму вместо точечной, если данные состоят из трех рядов, каждый из которых содержит набор значений. Размеры пузырьков определяются значениями третьего ряда данных. Пузырьковые диаграммы часто используются для представления финансовых данных. Пузырьки разных размеров позволяют визуально выделить конкретные значения.	5

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
<b>ПК-4</b> Способен проводить анализ требований к программному обеспечению, выполнять работы по проектированию программного обеспечения.				
1.	Задание закрытого типа	<p>Какой модуль Hadoop представляет собой набор инфраструктурных программных библиотек и утилит, которые используются в других решениях и родственных проектах, в частности, для управления распределенными файлами и создания необходимой инфраструктуры</p> <ol style="list-style-type: none"> <li>1. Hadoop Common</li> <li>2. HDFS</li> <li>3. YARN</li> <li>4. Hadoop MapReduce</li> </ol>		3
2.		<p>В Python-библиотеке Scipy для создания разреженных матриц с сжатым хранением строкой используется функция</p> <ol style="list-style-type: none"> <li>1. csr_matrix</li> <li>2. coo_matrix</li> <li>3. tocsr</li> <li>4. tocsr</li> </ol>	1	3
5.		<p>В Python-библиотеке Scipy для конвертации разреженной матрицы в плотную сжатой строкой используется метод</p> <ol style="list-style-type: none"> <li>1. todense</li> <li>2. toarray</li> <li>3. tocsr</li> <li>4. tocsr</li> </ol>	3	3
5.		<p>Преимуществами сжатого хранения строкой являются:</p> <ol style="list-style-type: none"> <li>1. эффективные операции с матрицами такого же формата (CSR + CSR, CSR * CSR и т.д.);</li> <li>2. эффективное извлечение строк;</li> <li>3. быстрое извлечение</li> </ol>	1, 2	3

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
		столбцов; 4. быстрое преобразование в форматы список списков (LIL) и словарь ключей (DOK).		
5.		В Python-библиотеке Scipy для создания разреженных матриц с сжатым хранением столбцом используется функция 1. csr_matrix 2. csc_matrix 3. coo_matrix 4. todense	2	3
6.	Задание открытого типа	Перечислите основные свойства больших данных	Volume — объем данных. Velocity — скорость накопления и обработки массивов данных. Variety — разнообразие типов данных. Veracity — достоверность как самого набора данных, так и результатов его анализа; Variability — изменчивость. Value — ценность или значимость.	5
7.		Что такое Hadoop?	Hadoop – это свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Эта основополагающая технология хранения и обработки больших данных (Big Data) является проектом верхнего уровня фонда Apache Software Foundation.	5
8.		Опишите кратко концептуальную архитектуру Hadoop.	Изначально проект разработан на Java в рамках вычислительной парадигмы MapReduce, когда приложение разделяется на большое количество	5

№ п/п	Тип задания	Формулировка задания	Правильный ответ	Время выполнения (в минутах)
			одинаковых элементарных заданий, которые выполняются на распределенных компьютерах (узлах) кластера и сводятся в единый результат.	
9.		Дайте определение термину NoSQL	NoSQL — обозначение широкого класса разнородных систем управления базами данных (СУБД), появившихся в конце 2000-х — начале 2010-х годов и существенно отличающихся от традиционных реляционных СУБД с доступом к данным средствами языка SQL. Применяется к системам, в которых делается попытка решить проблемы масштабируемости и доступности за счёт полного или частичного отказа от требований атомарности и согласованности данных.	5
10.		Дайте определение термину «большие данные».	Большие данные — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.	5

Полный комплект оценочных материалов по дисциплине (модулю) (фонд оценочных средств) хранится в электронном виде на кафедре, утверждающей рабочую программу дисциплины (модуля), и в Центре мониторинга и аудита качества обучения.

#### **7.4. Методические материалы, определяющие процедуры оценивания результатов обучения по дисциплине (модулю)**

Итоговая оценка по промежуточной аттестации выставляется в соответствии с Положением АГУ о балльно-рейтинговой системе (БАРС). Итоговая оценка складывается из баллов, полученных студентами за текущую успеваемость в течение семестра и баллов, полученных студентом на экзамене. Для получения положительной оценки студенту необходимо набрать минимально 60 баллов.

**Таблица 10. Технологическая карта рейтинговых баллов по дисциплине (модулю)**

№ п/п	Контролируемые мероприятия	Количество мероприятий / баллы	Максимальное количество баллов	Срок представления
<b>Основной блок</b>				
1.	Ответ на занятии	2/1	2	
2.	Выполнение практического задания (лабораторная работа.)	8/6	48	
<b>Всего</b>			<b>50</b>	-
<b>Ответ на экзамене</b>			<b>40</b>	-
<b>Блок бонусов</b>				
3.	Отсутствие пропусков занятий	2,5	2,5	
4.	Своевременное выполнение всех заданий	3	3	
5.	Активное участие в семинарском занятии, обсуждении семинарских вопросов	9/0,5	4,5	
<b>Всего</b>			<b>10</b>	-
<b>ИТОГО</b>			<b>100</b>	-

**Таблица 11 – Система штрафов (для одного занятия)**

Показатель	Балл
<i>Опоздание на занятие</i>	-1
<i>Нарушение учебной дисциплины</i>	-1
<i>Неготовность к занятию</i>	-3
<i>Пропуск занятия без уважительной причины</i>	-2

**Таблица 12 – Шкала перевода рейтинговых баллов в итоговую оценку за семестр по дисциплине (модулю)**

Сумма баллов	Оценка по 4-балльной шкале
90–100	5 (отлично)
85–89	4 (хорошо)
75–84	
70–74	
65–69	3 (удовлетворительно)
60–64	
Ниже 60	2 (неудовлетворительно)

При реализации дисциплины (модуля) в зависимости от уровня подготовленности обучающихся могут быть использованы иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

Экзамен проходит в форме устного собеседования со студентом по билетам, составленным из вопросов (п. 7.3). Один билет включает в себя 2 вопроса. Выбор билета осуществляется в случайном порядке. На подготовку студенту отводится не менее 40 мин.

Во время проведения экзамена студенту запрещено пользоваться сотовым телефоном и иными средствами связи, персональным компьютером, сетью Интернет, заготовленными заранее ответами и т.п.

Для стимулирования развития творческого и научно-исследовательского потенциала студентов при промежуточном оценивании предусмотрена система дополнительных баллов, а именно начисление до 10 поощрительных баллов за участие в конференциях, семинарах, выставках и т.п. в области анализа данных, программировании с представлением индивидуальных проектов в области аналитики данных.

Начисление баллов зависит от статуса мероприятия и статуса участия в нем студента. Начисление баллов происходит при предоставлении диплома, сертификата, грамоты, материалов конференции, опубликованной статьи, тезисов и т.п.

Преподаватель, реализующий дисциплину, в зависимости от уровня подготовленности обучающихся может использовать иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

При реализации дисциплины (модуля) в зависимости от уровня подготовленности обучающихся могут быть использованы иные формы, методы контроля и оценочные средства, исходя из конкретной ситуации.

## **8. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)**

### **8.1. Основная литература**

1. Броневиц, А. Г. Нечеткие модели анализа данных и принятия решений : учебное пособие / А. Г. Броневиц, А. Е. Лепский. - Москва : Высшая школа экономики, 2022. - 266 с. Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". - ISBN 978-5-7598-2407-7. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785759824077.html>

2. Искусственный интеллект, аналитика и новые технологии / - Москва : Альпина Паблишер, 2022. - 200 с. (Серия "Harvard Business Review: 10 лучших статей") - ISBN 978-5-9614-4791-0. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785961447910.html>

3. Кошкарлов А.В. Аналитика больших данных. Астрахань: Издатель Сорокин Роман Васильевич, 2018. URL: <https://biblio.asu.edu.ru/Reader/Book/2019100910013323100002066826>. (Электронная библиотека "Астраханский государственный университет")

4. Маккинли, У. Python и анализ данных / Уэс Маккинли - Москва : ДМК Пресс, 2015. - 482 с. - ISBN 978-5-97060-315-4. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785970603154.html>

5. Ульман, Дж. Д. Анализ больших наборов данных / Дж. Д. Ульман, Ю. Лесковец, А. Раджараман; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 500 с. Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". - ISBN 978-5-89818-304-2. - Текст : электронный // ЭБС "Консультант студента" : [сайт]. - URL : <https://www.studentlibrary.ru/book/ISBN9785898183042.html>

### **8.2. Дополнительная литература**

1. Адлер Ю.П. Статистическое управление процессами. "Большие данные". М.: МИСиС, 2016. - 52 с. URL: <http://www.studentlibrary.ru/book/ISBN9785876239693.html>. (ЭБС «Консультант студента»).

2. Форман Дж. Много цифр: Анализ больших данных при помощи Excel. М.: Альпина Паблишер, 2016. - 461 с. URL : <http://www.studentlibrary.ru/book/ISBN9785961450323.html>. (ЭБС «Консультант студента»).

3. Будылдина Н.В. Сетевые технологии высокоскоростной передачи данных : Учебное пособие для вузов. М. : Горячая линия - Телеком, 2016. - 342 с. URL : <http://www.studentlibrary.ru/book/ISBN9785991205368.html>. (ЭБС «Консультант студента»).
4. Бэнкер К. MongoDB в действии. М.: ДМК Пресс, 2012. - 394 с. URL : <http://www.studentlibrary.ru/book/ISBN9785940748311.html> (ЭБС «Консультант студента»).
5. Кук Д. Машинное обучение с использованием библиотеки H2O. М. : ДМК Пресс, 2018. - 250 с. URL : <http://www.studentlibrary.ru/book/ISBN9785970605080.html>. (ЭБС «Консультант студента»).
6. Рашка С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения. М. : ДМК Пресс, 2017. - 418 с. URL : <http://www.studentlibrary.ru/book/ISBN9785970604090.html>. (ЭБС «Консультант студента»).
7. Сухов К.К. Node.js. Путеводитель по технологии. М.: ДМК Пресс, 2015. - 416 с. URL : <http://www.studentlibrary.ru/book/ISBN9785970601648.html>. (ЭБС «Консультант студента»).

### **8.3. Интернет-ресурсы, необходимые для освоения дисциплины (модуля)**

1. Электронная библиотека «Астраханский государственный университет» собственной генерации на платформе ЭБС «Электронный Читальный зал – БиблиоТех».<https://biblio.asu.edu.ru>.
2. Электронно-библиотечная система (ЭБС) ООО «Политехресурс» «Консультант студента». [www.studentlibrary.ru](http://www.studentlibrary.ru).

## **9. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)**

Для проведения лабораторных занятий необходима аудитория, оснащенная компьютерными рабочими местами студентов и доступом в Интернет.

Рабочая программа дисциплины (модуля) при необходимости может быть адаптирована для обучения (в том числе с применением дистанционных образовательных технологий) лиц с ограниченными возможностями здоровья, инвалидов. Для этого требуется заявление обучающихся, являющихся лицами с ограниченными возможностями здоровья, инвалидами, или их законных представителей и рекомендации психолого-медико-педагогической комиссии. Для инвалидов содержание рабочей программы дисциплины (модуля) может определяться также в соответствии с индивидуальной программой реабилитации инвалида (при наличии).