МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Национальный исследовательский ядерный университет «МИФИ»

Обнинский институт атомной энергетики –

филиал федерального государственного автономного образовательного учреждения высшего образования
«Национальный исследовательский ядерный университет «МИФИ»

(ИАТЭ НИЯУ МИФИ)

Одобрено на заседании УМС
ИАТЭ НИЯУ МИФИ Протокол
от 30.08.2022 № 2-8/2022

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

«Большие данные»

название дисциплины

для студентов направления подготовки

09.04.01 Информатика и вычислительная техника

профиль:

Большие данные и машинное обучение для атомной энергетики

Форма обучения: очная

г. Обнинск 2022 г.

Программа составлена в соответствии с образовательным стандартом высшего образования НИЯУ МИФИ по направлению подготовки 09.04.01 «Информатика и вычислительная техника».

Программу составил:

_____ С.В. Грицюк, доцент, к.т.н.

Рецензент:

_____

Программа рассмотрена на заседании отделения интеллектуальных кибернетических систем (О)
(протокол № ____ от «____»_____2022 г.)

Руководитель образовательной программы
090401 «Информатика и вычислительная техника»
_____ Старков С.О.
«____»_____2022 г.

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения ОПОП магистратуры обучающийся должен овладеть следующими результатами обучения по дисциплине:

Коды компетенций	Результаты освоения ООП *Содержание компетенций*	Перечень планируемых результатов обучения по дисциплине
ПК-1	способен применять научно обоснованные перспективные методы исследования и решать задачи на основе знания мировых тенденций развития вычислительной техники и информационных технологий с внедрением результатов исследований в реальный сектор экономики	**Знать**: существующие подходы и технологии работы с большими объемами данных, их особенности, предпосылки возникновения, сильные и слабые стороны. **Уметь**: составлять список требований к разрабатываемой информационной системе и технологиям работы с большими данными. **Владеть**: навыками создания новых алгоритмов и систем обработки больших объемов данных.
СПК-1	способен использовать и развивать методы научных исследований и инструментарий в области интеллектуального анализа данных	**Знать**: подходы к обработке данных, их особенности. **Уметь**: использовать современные информационные технологии поиска, анализа и обработки данных для последующего использования в рамках информационной системы. **Владеть**: навыками практического применения методов и современных компьютерных технологий поиска и анализа информации.
		Знать: теоретические основы больших данных; возможности современных технологий и их использование для решения прикладных задач в различных областях профессиональной деятельности. **Уметь**: использовать технологии и системы анализа больших данных **Владеть**: технологиями анализа больших данных
		Знать: актуальные подходы и технологии в области обработки больших объемов данных **Уметь**: применять современные подходы и технологии больших данных

		в рамках реальных создаваемых систем **Владеть**: навыками проектирования и создания систем на базе современных технологий обработки больших данных
		Знать: современные разработки в области интеллектуального анализа данных **Уметь**: создавать новые подходы/инструменты в области интеллектуального анализа данных **Владеть**: навыками поиска и интеллектуального анализа данных

2. Место дисциплины в структуре ОПОП магистратуры

Дисциплина реализуется в рамках обязательной части.

Для освоения дисциплины необходимы компетенции, сформированные в рамках изучения следующих дисциплин: «Программирование», «Технологии программирования», «Нереляционные базы данных».

Дисциплины и/или практики, для которых освоение данной дисциплины необходимо как предшествующее: «Технологии программирования для Больших данных», «Высокопроизводительные вычисления».

Дисциплина изучается на 1 курсе в 1 семестре.

3. Объем дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающихся с преподавателем (по видам занятий) и на самостоятельную работу обучающихся

Вид работы	**Форма обучения** (вносятся данные по реализуемым формам)	
	Очная	
	Семестр	**Курс**
	№ 1	**№ 1**
	Количество часов на вид работы:	
Контактная работа обучающихся с преподавателем		
Аудиторные занятия *(всего)*	48	
В том числе:		
лекции *(лекции в интерактивной форме)*	16	
практические занятия	16	

(практические занятия в интерактивной форме)	
лабораторные занятия	16
Промежуточная аттестация	
В том числе:	
зачет	-
экзамен	36
Самостоятельная работа обучающихся	
Самостоятельная работа обучающихся *(всего)*	96
В том числе:	
проработка учебного (теоретического) материала	16
выполнение индивидуальных практических заданий	48
подготовка ко всем видам контрольных испытаний текущего контроля успеваемости	16
подготовка ко всем видам контрольных испытаний промежуточной аттестации	16
Всего (часы):	**180**
Всего (зачетные единицы):	**5**

4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины и трудоемкость по видам учебных занятий (в академических часах)

№ п/п	Наименование раздела /темы дисциплины	**Виды учебной работы** в часах				
		Очная форма обучения				
		Лек	**Пр**	**Лаб**	**Внеауд**	**СРО**
1.	Введение в область больших данных	2				6
2.	Введение в программирование на языке Scala		4	2		16
3.	MapReduce, Hadoop и Apache Spark	6				6
4.	Системы хранения	6				6
5.	Экосистема Hadoop		4			6
6.	Анализ естественного языка		4			6
7.	Машинное обучение		4			6
8.	Практические примеры из области больших данных	2				6
9.	Анализ естественного языка на примере литературного произведения			4		10
10.	Spark SQL для анализа данных			5		14
11.	Основы машинного обучения на Spark ML			5		14
	Всего:	**16**	**16**	**16**		**96**

Прим.: Лек – лекции, Пр – практические занятия / семинары, Лаб – лабораторные занятия, Внеауд – внеаудиторная работа, СРО – самостоятельная работа обучающихся

4.2. Содержание дисциплины, структурированное по разделам (темам)

Лекционный курс

№	Наименование раздела /темы дисциплины	Содержание
1.	Введение в область больших данных	Информация о лекторе. Информация о структуре курса. История возникновения понятия «большие данные». Определение. Свойства 3V. Распределенные системы. Целостность данных в распределенных системах. Некоторые примеры работы в реальном мире.
2.	MapReduce, Hadoop и Apache Spark	Проблемы обработки больших объемов данных. Идея Google MapReduce. Apache Hadoop, как открытая реализация идей Google. WordCount. Технологии в составе Apache Hadoop. Apache Hadoop MapReduce. Ограничения и особенности. Apache Spark. Сравнение с Apache Hadoop. Основные концепции и подходы.
3.	Системы хранения	Нереляционные базы данных. Классификация. Основные свойства каждого из классов. Плюсы и минусы баз данных ключ-значение. Примеры. Области применения.
4.	Практические примеры из области больших данных	Современные реалии рынка «больших данных». Примеры и результаты внедрения. Перспективы и новые направления, возможные последствия.

Практические/семинарские занятия

№	Наименование раздела /темы дисциплины	Содержание
1.	Введение в программирование на языке Scala	Определение языка Scala, причины и цель его создания. История возникновения. Основные свойства и конструкции языка. Сравнение с языком Java. Объектно-ориентированные возможности. Функциональные возможности. Изменяемые и неизменяемые данные. Работа с коллекциями.
2.	Экосистема Hadoop	Распределенная инфраструктура. Проблемы построения. Дистрибутивы. Cloudera. Установка. Основные компоненты и их краткая характеристика. Рекомендации по оборудованию.
3.	Анализ естественного языка	Определение. Предпосылки зарождения области. Направления развития, классические задачи. Уровни анализа. Языковые модели. Инструменты.
4.	Машинное обучение	Зарождение и развитие области машинного

		обучения. Краткая история. Терминология. Категории и техники. Классические задачи. Качество данных. Стадии построения модели. Deep Learning. Примеры использования.

Лабораторные занятия

№	Наименование раздела /темы дисциплины	Название лабораторной работы
1.	Введение в программирование на языке Scala	*Лабораторная работа №1:* Изучение основ языка Scala. Установка и запуск среды разработки Scala IDE. Создание рабочего Maven/SBT-проекта. Понимание императивного и функционального подхода написания кода на примере своего варианта задания.
2.	Анализ естественного языка на примере литературного произведения	*Лабораторная работа №2:* Реализация программы на языке Scala. Подключение и использование Apache Spark в standalone режиме. Анализ по своему варианту литературного произведения. Очистка текста. Реализация WordCount. Стемминг. Ранжирование встречаемости слов.
3.	Spark SQL для анализа данных	*Лабораторная работа №3:* Реализация программы на языке Scala. Анализ и описание набора данных в соответствии с заданным вариантом. Построение запросов к данным с помощью Spark SQL. Визуализация данных.
4.	Основы машинного обучения на Spark ML	*Лабораторная работа №4:* Реализация программы на языке Scala. Постановка простейшей задачи машинного обучения (регрессия, классификация). Изучение основ Spark ML. Решение поставленной задачи и описание полученных результатов.

5. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

В качестве учебно-методических материалов используется рекомендованная литература и рекомендованные ресурсы сети Интернет (разделы 7 и 8).

6. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

6.1. Паспорт фонда оценочных средств по дисциплине

№ п/п	Контролируемые разделы (темы) дисциплины (результаты по разделам)	Код контролируемой компетенции (или её части) / и ее формулировка	Наименование оценочного средства
1-8.	1. Введение в область больших данных 2. Введение в программирование на языке Scala 3. MapReduce, Hadoop и Apache Spark 4. Системы хранения 5. Экосистема Hadoop 6. Анализ естественного языка 7. Машинное обучение 8. Практические примеры из области больших данных	ПК-1 (знать, уметь, владеть) СПК-1 (знать, уметь, владеть)	**Лабораторная работа №1** (демонстрация на компьютере выполненного проекта и защита работы в форме собеседования с преподавателем); **Контрольная работа №1** (в форме письменных ответов и устного собеседования на теоретические вопросы); **Экзамен** (в форме письменных ответов и устного собеседования на теоретические вопросы)
9.	Анализ естественного языка на примере литературного произведения	СПК-1 (знать, уметь, владеть)	**Лабораторная работа №2** (демонстрация на компьютере выполненного проекта и защита работы в форме собеседования с преподавателем)
10.	Spark SQL для анализа данных	ПК-1 (знать, уметь, владеть) СПК-1 (знать, уметь,	**Лабораторная работа №3** (демонстрация на компьютере выполненного проекта и защита работы в форме собеседования с преподавателем)
11.	Основы машинного обучения на Spark ML	ПК-1 (знать, уметь, владеть) СПК-1 (знать, уметь,	**Лабораторная работа №4** (демонстрация на компьютере выполненного проекта и защита работы в форме собеседования с преподавателем)

6.2. ***Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующие этапы формирования компетенций в процессе освоения образовательной программы***

6.2.1. Экзамен

Экзамен проводится в виде письменных ответов на 2 вопроса, с последующим устным собеседованием. Критерий оценки – правильность и полнота ответа на вопросы.
Оценка выставляется в баллах от 0 до 40 в равных долях за каждый вопрос.
Экзамен считается сданным при оценке не ниже 60% от максимального балла.

Список билетов на экзамен:

Вариант №1
1. Apache Spark. Работа в кластере.
2. Масштабируемость. Виды.

Вариант №2
1. NoSQL. Колоночные БД.
2. CAP теорема.

Вариант №3
1. Apache Hadoop. Особенности. Плюсы и минусы.
2. Примеры применения машинного обучения (2-3).

Вариант №4
1. Apache Hadoop. HDFS. MapReduce.
2. Классификация и различия NoSQL хранилищ.

Вариант №5
1. NoSQL. Графовые БД.
2. Основы языка Scala. Особенности. Плюсы и минусы.

Вариант №6
1. Анализ естественного языка. Мотивация. Направления.
2. Классификация и различия NoSQL хранилищ.

Вариант №7
1. Машинное обучение. Deep Learning.
2. NoSQL и транзакции.

Вариант №8
1. Машинное обучение. Техники/задачи.
2. Основы языка Scala. Методы на коллекциях.

Вариант №9
1. Анализ естественного языка. Мотивация. Направления.
2. Шардинг.

Вариант №10
1. Apache Spark. Стэк технологий.
2. Масштабируемость. Виды.

Вариант №11
1. Машинное обучение. Категории.
2. Репликация.

Вариант №12
1. NoSQL. Документные БД.
2. Популярные Hadoop-дистрибутивы и их различия.

Вариант №13
1. Парадигма MapReduce.
2. Популярные Hadoop-дистрибутивы и их различия.

Вариант №14
1. Apache Spark. Работа в кластере.
2. Примеры работы с Большими данными (2-3).

Вариант №15
1. Apache Hadoop. HDFS. MapReduce.

2. Репликация.
Вариант №16
1. NoSQL. Ключ-значение.
2. Дистрибутив Cloudera. Сервисы в составе дистрибутива.
Вариант №17
1. Apache Spark. Особенности. Плюсы и минусы.
2. Примеры применения машинного обучения (2-3).
Вариант №18
1. Парадигма MapReduce.
2. Области применения технологий Больших данных.
Вариант №19
1. NoSQL. Мотивация. Классификация.
2. Что такое Большие данные? Характеристика.
Вариант №20
1. Машинное обучение. Основы.
2. Дистрибутив Cloudera. Сервисы в составе дистрибутива.

6.2.2. Контрольная работа №1

Контрольная работа предназначена для выявления качества усвоения теоретических знаний по основным темам в курсе:
- Парадигма MapReduce;
- Apache Hadoop;
- Apache Spark;
- Нереляционные базы данных.

Контрольная работа включает в себя 2 вопроса, на которые студент должен дать исчерпывающий устный ответ. Контрольная работа оценивается в баллах от 0 до 10 и считается сданной при оценке не ниже 60% от максимального балла.

Варианты заданий составляются из двух вопросов: первый вопрос из 1-5, второй вопрос из 6-15.

Вопросы контрольной работы №1:
1. Что такое Большие данные? Характеристика.
2. Масштабируемость. Виды.
3. Репликация и Шардинг.
4. NoSQL и транзакции.
5. CAP теорема
6. Парадигма MapReduce.
7. Apache Hadoop. Особенности. Плюсы и минусы.
8. Apache Hadoop. HDFS. MapReduce.
9. Apache Spark. Особенности. Плюсы и минусы.
10. Apache Spark. Работа в кластере.
11. Apache Spark. Стэк технологий.
12. NoSQL. Ключ-значение.

13. NoSQL. Документные БД.
14. NoSQL. Колоночные БД.
15. NoSQL. Графовые БД.

6.2.3. Лабораторные работы №1, №2, №3, №4

Лабораторные работы предназначены для выработки практических навыков по материалу, полученному в рамках предмета (курс лекций), а также выявления качества усвоения знаний по дисциплине.
По завершению каждой из лабораторных работ студент должен продемонстрировать ее результат на компьютере и защитить в форме собеседования с преподавателем. На собеседование выносятся вопросы, касающиеся теоретических аспектов выполняемой работы, последовательности используемых для решения задачи шагов/процедур, а также анализа полученных результатов.
Критерий оценки – полнота, качество, своевременность выполненной работы и успешная ее защита. Лабораторные работы №1 и №2 оцениваются в баллах от 0 до 10, а лабораторные работы №3 и №4 от 0 до 15. Каждая лабораторная работа считается сданной при получении оценки не ниже 60% от максимального балла.

Лабораторная работа №1 включает установку, запуск, изучение интерфейса и встроенных средств среды разработки Scala IDE, а также основы программирования на языке Scala. Студент получает практические навыки создания и конфигурирования Maven-проекта в среде Scala IDE. По завершению лабораторной работы №1 в ходе устного опроса у компьютера студент показывает реализацию Scala программы в соответствии со своим вариантом.

Лабораторная работа №2 включает подключение и запуск библиотек Apache Spark в standalone режиме, а также реализацию программы для анализа текста своего варианта литературного произведения (очистка текста, реализация WordCount, стемминг, ранжирование встречаемости слов). По завершению лабораторной работы №2 в ходе устного опроса у компьютера студент демонстрирует код программы на языке Scala, объясняет основные проблемы, с которыми пришлось столкнуться и методы их решения, показывает результаты.

Лабораторная работа №3 включает освоение основ библиотеки Spark SQL, а также анализ и описание набора данных в соответствии с вариантом. Студент выполняет запросы к данным через Spark SQL, проводит визуализацию данных. По завершению лабораторной работы №3 в ходе устного опроса у компьютера студент демонстрирует код программы на языке Scala, объясняет полученные результаты.

Лабораторная работа №4 включает освоение основ библиотеки Spark ML. Студент формулирует простейшую задачу машинного обучения (регрессия, классификация) по набору данных для своего варианта. Далее, с помощью библиотеки Spark ML студент решает поставленную задачу. По завершению

лабораторной работы №4 в ходе устного опроса у компьютера студент демонстрирует код программы на языке Scala, объясняет полученные результаты.

6.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Рейтинговая оценка знаний является интегральным показателем качества теоретических и практических знаний и навыков студентов по дисциплине и складывается из оценок, полученных в ходе текущего контроля и промежуточной аттестации.

Текущий контроль в семестре проводится с целью обеспечения своевременной обратной связи, для коррекции обучения, активизации самостоятельной работы студентов.

Промежуточная аттестация предназначена для объективного подтверждения и оценивания достигнутых результатов обучения после завершения изучения дисциплины.

Текущий контроль осуществляется два раза в семестр: контрольная точка № 1 (КТ № 1) и контрольная точка № 2 (КТ № 2).

Результаты текущего контроля и промежуточной аттестации подводятся по шкале балльно-рейтинговой системы.

Вид контроля	**Этап рейтинговой системы Оценочное средство**	**Балл**	
		Минимум	Максимум
Текущий	**Контрольная точка № 1**	**18**	**30**
	Лабораторная работа №1	6	10
	Лабораторная работа №2	6	10
	Контрольная работа №1 (2 вопроса – 5 и 5 баллов)	6	10
	Контрольная точка № 2	**18**	**30**
	Лабораторная работа №3	9	15
	Лабораторная работа №4	9	15
Промежуточный	**Экзамен**	**24**	**40**
ИТОГО по дисциплине		60	100

За несвоевременную сдачу любого из указанных в таблице оценочных средств оценка может быть снижена от 1 до 2 баллов.

Процедура оценивания знаний, умений, владений по дисциплине включает учет успешности по всем видам заявленных оценочных средств.

Устный опрос проводится на каждом практическом занятии и затрагивает как тематику прошедшего занятия, так и лекционный материал. Ответ оценивается преподавателем.

По окончании освоения дисциплины проводится промежуточная аттестация в виде экзамена, что позволяет оценить совокупность приобретенных в процессе обучения компетенций. При выставлении итоговой оценки применяется балльно-

рейтинговая система оценки результатов обучения.

Экзамен предназначен для оценки работы обучающегося в течение всего срока изучения дисциплины и призван выявить уровень и систематичность полученных обучающимся теоретических знаний, приобретенных навыков самостоятельной работы.

Оценка сформированных компетенций на экзамене для тех обучающихся, которые пропускали занятия и не участвовали в проверке компетенций во время изучения дисциплины, проводится после индивидуального собеседования с преподавателем по пропущенным или не усвоенным обучающимся темам с последующей оценкой самостоятельно усвоенных знаний на экзамене.

7. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

1. Martin Odersky, Lex Spoon, Bill Benners Programming in Scala.
2. Alex Payne, Dean Wampler Programming Scala. 2014. 583 p.
3. Paul Chiusano, Rúnar Bjarnason Functional Programming in Scala. 2014. 320 p.
4. Joshua D. Suereth Scala in Depth. 2012. 304 p.
5. Кей С. Хорстманн Функциональное программирование. SCALA для нетерпеливых. 2013. 408 с.
6. Garry Turkington, Gabriele Modena. Learning Hadoop 2. 2015. 382 p.
7. Thilina Gunarathne. Hadoop MapReduce v2 Cookbook - Second Edition. 2015. 322 p.
8. Alex Holmes. Hadoop in Practice, Second Edition. 2014. 512 p.
9. Holden Karau, Rachel Warren. High Performance Spark. 2017. 175 p.
10. Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell. Learning Spark, Lightning-Fast Big Data Analysis. 2015. 276 p.
11. Petar Zecevic. Spark in Action. 2016. 468 p.
12. Mike Frampton. Mastering Apache Spark. 2015. 318 p.

8. Перечень ресурсов информационно-телекоммуникационной сети «Интернет» (далее - сеть «Интернет»), необходимых для освоения дисциплины

1. Язык программирования Scala [Официальный сайт]. — https://www.scala-lang.org/
2. Среда разработки Scala IDE [Официальный сайт]. — http://scala-ide.org/
3. Scala Школа. — https://twitter.github.io/scala_school/ru/index.html
4. Упражнения по Scala. — https://www.scala-exercises.org/
5. Курс по принципам функционального программирования. — https://www.coursera.org/learn/progfun1
6. Специализация по Scala. — https://www.coursera.org/specializations/scala
7. Apache Hadoop [Официальный сайт]. — http://hadoop.apache.org/

8. Apache Spark [Официальный сайт]. — https://spark.apache.org/

9. Методические указания для обучающихся по освоению дисциплины

Вид учебного занятия	Организация деятельности студента
Лекция	Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначить вопросы, термины, материал, который вызывает трудности, пометить и попытаться найти ответ в рекомендуемой литературе. Если самостоятельно не удается разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на лекции или лабораторной работе. Уделить внимание следующим базовым понятиям: большие данные, масштабирование, распределенная система, целостность данных, репликация данных, шардинг данных, CAP теорема.
Контрольная работа	Работа с конспектами лекций, знакомство с основной и дополнительной литературой, включая справочные издания, зарубежные источники.
Лабораторная работа	При выполнении лабораторных работ необходимо ориентироваться на конспекты лекций, рекомендуемую литературу. Лабораторная работа считается выполненной после ее успешной защиты, включающей: – демонстрацию на компьютере решаемой задачи с разъяснением разработанного программного кода и демонстрацией выполнения; – собеседование с преподавателем для выявления уровня освоения теоретических основ в области больших данных.
Подготовка к экзамену	При подготовке к экзамену необходимо ориентироваться на конспекты лекций и лабораторные работы, а также рекомендуемую литературу.

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

• Операционные системы Windows 7/10, Linux (CentOS / RedHat, OpenSUSE, Ubuntu);
• Среда для программирования на языке Scala – Scala IDE (http://scala-ide.org/);
• Java Runtime Environment v.1.8 (http://www.java.com/);
• Электронные презентации лекций в формате PDF, демонстрируемые с использованием мультимедийного проектора или дистанционно.

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

• Компьютерный класс сетевых технологий. Класс оснащен 10 компьютерами (Intel Core i5/8GB/1 TB) и 1 компьютером (Intel Celeron 1.6 GHz, 2 GB RAM, 250 GB) с операционной системой Windows 7, а также мультимедийным проектором. Есть доступ к Wi-Fi.
• Аудиторный класс, оборудованный проекционным экраном, мультимедийным проектором и персональным компьютером (AMD, ATHLON64, 2.7 GHz, 4 GB RAM, 250 GB). Есть доступ к Wi-Fi.

12. Иные сведения и (или) материалы

12.1. Перечень образовательных технологий, используемых при осуществлении образовательного процесса по дисциплине

Лекционные и практические занятия проходят с обсуждением учебного материала, демонстрируемого в форме презентаций на экране с использованием мультимедиа-проектора. Все лабораторные занятия проводятся в интерактивной форме при тесном контакте студентов с преподавателем.

В рамках лабораторных работ студенты выполняют 4 лабораторные работы, призванные дать представление о возможностях применения больших данных, как инструментария для решения самых разнообразных практических задач. Лабораторные работы проводятся при активном взаимодействии студентов и преподавателя, в ходе которого обсуждаются детали создания проекта задачи, проблемы и ошибки, возникающие на всех этапах их разработки, проводится проверка корректности полученных результатов.

12.2. Формы организации самостоятельной работы обучающихся (темы, выносимые для самостоятельного изучения; вопросы для самоконтроля; типовые задания для самопроверки

На самостоятельное изучение студентам предлагается более глубоко рассмотреть темы, кратко затрагиваемые в лекционных курсах. Контроль освоения материала осуществляется в ходе приема лабораторных работ и в рамках экзамена по дисциплине.

№	Тема	Часть, осваиваемая самостоятельно
1.	Введение в область больших данных	Распределенные системы. Особенности организации и работы. Согласованность данных в распределенных системах.
2.	Введение в программирование на языке Scala	Объектно-ориентированное программирование на Scala. Коллекции в Scala. Mutability и Immutability.
3.	MapReduce, Hadoop и Apache Spark	Архитектура и принципы работы Apache Hadoop. Принцип работы HDFS. Spark кластер. Spark SQL. Spark Streaming. MLlib. GraphX.
4.	Системы хранения	Key-value хранилища. Документные базы данных. Колоночные базы данных. Графовые базы данных.
5.	Экосистема Hadoop	Hortonworks. MapR. Компоненты платформы Cloudera. YARN. Apache Kafka. Apache Solr. Apache Hive.
6.	Анализ естественного языка	Задачи анализа естественного языка. Information extraction. Онтологии. Языковые модели.
7.	Машинное обучение	Задачи и применение машинного обучения. Техники. Feature engineering. Уменьшение размерности. Underfitting и Overfitting. Reinforcement Learning. Deep Learning.
8.	Практические примеры из области больших данных	Примеры использования Big Data в реальной жизни.

Контроль освоения самостоятельно изученного теоретического материала осуществляется в виде собеседования во время защиты лабораторных, в виде устного опроса на экзамене.

Кроме этого, студенты также самостоятельно выполняют большую часть предусмотренных практических работ, промежуточный результат которых

представляется на лабораторных занятиях, а конечный результат - на защите лабораторных работ.

Вопросы для самоконтроля:

- Организация Spark кластера.
- Принцип работы HDFS.
- CAP теорема.
- Кластер. Распределенные системы. Масштабирование.
- BASE свойства.
- Репликация.
- Шардинг.
- Целостность данных. Целостность в конечном итоге.
- Язык Scala. Коллекции.
- Язык Scala. Mutability и Immutability.

12.3. Краткий терминологический словарь

Приводятся русские, а также общепринятые сокращения/акронимы на английском языке

BASE – Basically Available, Soft state, Eventual consistency

CAP – Consistency, Availability, Partition tolerance

IDE – Integrated Development Environment

HDFS – Hadoop Distributed File System

NoSQL – Not Only SQL

SQL – Structured Query Language

YARN – Yet Another Resource Negotiator

БД – База Данных