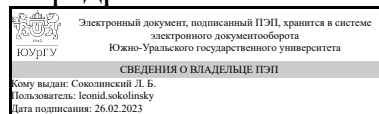


УТВЕРЖДАЮ:
Заведующий выпускающей
кафедрой



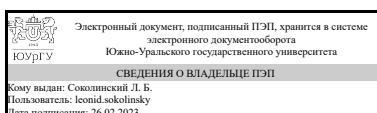
Л. Б. Соколинский

РАБОЧАЯ ПРОГРАММА

дисциплины 1.Ф.П0.08 Основы распределенной обработки данных
для направления 09.03.04 Программная инженерия
уровень Бакалавриат
профиль подготовки Инженерия информационных и интеллектуальных систем
форма обучения очная
кафедра-разработчик Системное программирование

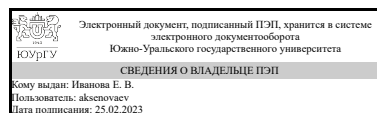
Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 09.03.04 Программная инженерия, утверждённым приказом Минобрнауки от 19.09.2017 № 920

Зав.кафедрой разработчика,
д.физ.-мат.н., проф.



Л. Б. Соколинский

Разработчик программы,
к.физ.-мат.н., доцент



Е. В. Иванова

1. Цели и задачи дисциплины

Целью курса является изучение студентами задач, связанных с распределенным хранением и обработкой больших данных. При изучении этого курса должны быть решены следующие задачи: изучение понятия и проблематики больших данных, способы распределенного хранения и обработки больших данных, хранение и обработка больших данных в экосистеме Hadoop.

Краткое содержание дисциплины

Понятие больших данных. Распределенная обработка больших данных. Основы Hadoop, HDFS, MapReduce. Экосистема Hadoop: Pig, Apache Hive, HBase, Apache Spark, MLlib, Hadoop YARN, Zookeeper, Apache Kafka.

2. Компетенции обучающегося, формируемые в результате освоения дисциплины

Планируемые результаты освоения ОП ВО (компетенции)	Планируемые результаты обучения по дисциплине
ПК-2 Способен разрабатывать компоненты системных программных продуктов на основе соответствующей технической документации	Знает: основы работы компонентов экосистемы Hadoop Умеет: строить программную систему на основе компонентов экосистемы Hadoop для решения поставленной задачи Имеет практический опыт: создания программной системы на основе компонентов экосистемы Hadoop
ПК-7 (ПК-8 модели) Способен разрабатывать системы анализа больших данных	Знает: ПК-8.1. 3-2. Знает принципы работы экосистемы Hadoop, фреймворка SPARK; ПК-8.2. 3-1. Знает принципы и методы анализа больших данных, включая спецификации и стандартизацию метаданных; ПК-8.2. 3-2. Знает устройство и принципы работы систем обработки и анализа больших массивов данных (SQL, NoSQL, Hadoop, ETL); ПК-8.2. 3-3. Знает архитектуру и принципы работы промышленных решений, созданных на основе искусственного интеллекта; Умеет: ПК-8.1. У-4. Умеет использовать шины данных (Apache Kafka); ПК-8.2. У-3. Умеет использовать системы обработки и анализа больших массивов данных (SQL, NoSQL, Hadoop, ETL процессы и инструменты); ПК-8.2. У-4. Умеет использовать технологии науки о данных и больших данных в разработке для решения практических задач промышленности; Имеет практический опыт: обработки и анализа данных в экосистеме Hadoop

3. Место дисциплины в структуре ОП ВО

Перечень предшествующих дисциплин, видов работ учебного плана	Перечень последующих дисциплин, видов работ
---------------------------------------------------------------	---------------------------------------------

<p>Основы разработки систем управления большими данными, Технологии аналитической обработки информации, Администрирование и развертывание программных компонент систем искусственного интеллекта в ОС Linux, Основы интеллектуального анализа данных, Архитектура ЭВМ, Базы данных</p>	<p>Не предусмотрены</p>
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------

Требования к «входным» знаниям, умениям, навыкам студента, необходимым при освоении данной дисциплины и приобретенным в результате освоения предшествующих дисциплин:

Дисциплина	Требования
Базы данных	<p>Знает: основы устройства систем баз данных, основы работы современных систем управления базами данных, устройство интерфейсов между реляционными SQL-хранилищами данных и нереляционными NoSQL-хранилищами данных Умеет: устанавливать и настраивать реляционные и нереляционные системы баз данных, создавать реляционные и нереляционные базы данных и запросы к ним, использовать языки запросов, в том числе нереляционных, для поддержки различных типов данных (например, XML, RDF, JSON, мультимедиа) и операций с большими данными (например , матричные операции) Имеет практический опыт: инсталляции систем баз данных, разработки реляционных и нереляционных баз данных, написания запросов к реляционным и нереляционным большим базам данных</p>
Технологии аналитической обработки информации	<p>Знает: постановку базовых задач интеллектуального анализа данных (поиск шаблонов, классификация, кластеризация) и базовые методы их решения, методы и критерии оценки качества моделей машинного обучения, общедоступные репозитории и специализированные библиотеки, содержащие наборы больших данных Умеет: планировать и выполнять машинные эксперименты, оценивать точность и качество построенных моделей, сопоставить задачам предметной области классы задач машинного обучения, настраивать и оптимизировать конфигурацию программного и аппаратного обеспечения с целью интеграции больших данных Имеет практический опыт: разработки приложений для аналитической обработки информации с помощью современных инструментальных средств, анализа требований и идентификации классов задач для реализации приложений машинного обучения, разработки</p>

	<p>программных компонент для извлечения и подготовки больших данных для аналитической обработки информации</p>
<p>Архитектура ЭВМ</p>	<p>Знает: типы архитектур ЭВМ, требования к системному и прикладному ПО, понятие архитектуры ЭВМ, способы представления данных в ЭВМ, принципы организации вычислений, основные положения и концепции в области архитектуры ЭВМ, базовые принципы проектирования системного ПО Умеет: проектировать ПО с учетом принципов организации ЭВМ, разрабатывать алгоритмические и программные решения с использованием низкоуровневых языков программирования, решать стандартные задачи в профессиональной деятельности с учетом способов представления и обработки данных в ЭВМ Имеет практический опыт: проектирования системного ПО с учетом принципов организации ЭВМ, системного программирования с использованием низкоуровневых языков программирования, разработки программ на низкоуровневых языках программирования с учетом способов представления и обработки данных в ЭВМ</p>
<p>Администрирование и развертывание программных компонент систем искусственного интеллекта в ОС Linux</p>	<p>Знает: основные программные платформы и компоненты систем искусственного интеллекта: механизмы логического вывода (рассуждений), объяснений, приобретения знаний, интеллектуальных интерфейсов, принципы Data Ops и Dev Ops , принципы разработки системных утилит в Linux, основные принципы устройства файловой системы в Linux, межпроцессное и многопоточное взаимодействие, основные принципы устройства и администрирования ОС семейства Linux Умеет: применять на практике принципы и инструменты Data Ops и Dev Ops при развертывании компонентов систем искусственного интеллекта в ОС семейства Linux, реализовывать системные скрипты для решения задач профессиональной деятельности, разрабатывать системные решения обработки файлов в Linux, реализацию многопоточных приложений, клиент-серверных приложений в Linux, выполнять задачи администрирования ОС семейства Linux Имеет практический опыт: работы с основными утилитами командной строки в Linux</p>
<p>Основы интеллектуального анализа данных</p>	<p>Знает: постановку и методы решения основных задач интеллектуального анализа данных (поиск шаблонов, классификация, кластеризация), общедоступные репозитории и специализированные библиотеки, содержащие наборы больших данных, методы и критерии оценки качества моделей машинного обучения Умеет: планировать и выполнять машинные</p>

	эксперименты, оценивать точность и качество построенных моделей, настраивать и оптимизировать конфигурацию программного и аппаратного обеспечения с целью интеграции больших данных, сопоставить задачам предметной области классы задач машинного обучения Имеет практический опыт: разработки моделей машинного обучения для решения основных задач интеллектуального анализа данных (поиск шаблонов, классификация, кластеризация) и проведения вычислительных экспериментов по оценке точности и качества построенных моделей, разработки программных компонент для извлечения и подготовки больших данных для интеллектуального анализа, анализа требований и определения необходимых классов задач для реализации приложений машинного обучения; определения метрик и критериев качества оценки моделей машинного обучения
Основы разработки систем управления большими данными	Знает: методы и технологии массово параллельной обработки и анализа данных, методы оперативной обработки потоков данных Умеет: использовать методы и технологии массово параллельной обработки и анализа данных, выполнять потоковую обработку данных (data streaming, event processing) Имеет практический опыт: применения методов и технологий массово параллельной обработки и анализа данных, использования методов оперативной обработки потоков данных

4. Объём и виды учебной работы

Общая трудоемкость дисциплины составляет 4 з.е., 144 ч., 58,5 ч. контактной работы

Вид учебной работы	Всего часов	Распределение по семестрам в часах
		Номер семестра
		8
Общая трудоёмкость дисциплины	144	144
<i>Аудиторные занятия:</i>	48	48
Лекции (Л)	24	24
Практические занятия, семинары и (или) другие виды аудиторных занятий (ПЗ)	24	24
Лабораторные работы (ЛР)	0	0
<i>Самостоятельная работа (СРС)</i>	85,5	85,5
Подготовка к экзамену	25,5	25,5
Изучение тем и проблем, не выносимых на лекции и практические занятия	60	60
Консультации и промежуточная аттестация	10,5	10,5
Вид контроля (зачет, диф.зачет, экзамен)	-	экзамен

5. Содержание дисциплины

№ раздела	Наименование разделов дисциплины	Объем аудиторных занятий по видам в часах			
		Всего	Л	ПЗ	ЛР
1	Введение в большие данные и Hadoop	14	8	6	0
2	Экосистема Hadoop	14	8	6	0
3	Анализ данных в Hadoop	20	8	12	0

5.1. Лекции

№ лекции	№ раздела	Наименование или краткое содержание лекционного занятия	Кол-во часов
1	1	Введение в большие данные и распределенные вычисления	2
2	1	Введение в платформу Hadoop	2
3	1	Распределенная файловая система Hadoop (HDFS)	2
4	1	Технология MapReduce	2
5	2	Pig и СУБД Apache Hive	2
6	2	Введение в HBase	2
7	2	Архитектура MapReduce 2.0. Hadoop YARN	2
8	2	Брокер сообщений Apache Kafka. Координация распределенных приложений с Zookeeper	2
9	3	Apache Spark	4
10	3	Анализ и визуализация данных в Hadoop	4

5.2. Практические занятия, семинары

№ занятия	№ раздела	Наименование или краткое содержание практического занятия, семинара	Кол-во часов
1	1	Установка Hadoop. Работа с HDFS	2
2	1	Разработка MapReduce-приложения	4
3	2	Разработка базы данных в СУБД Hive	6
4	3	Разработка приложения в Apache Spark	6
5	3	Разработка приложения для анализ и визуализация данных в Hadoop	6

5.3. Лабораторные работы

Не предусмотрены

5.4. Самостоятельная работа студента

Выполнение СРС			
Подвид СРС	Список литературы (с указанием разделов, глав, страниц) / ссылка на ресурс	Семестр	Кол-во часов
Подготовка к экзамену	[Осн. лит., 1], Часть 3, с. 331–400; [Осн. лит., 2], Гл. 24-25, с. 446-470; [Доп. лит., 5] с. 5-47	8	25,5
Изучение тем и проблем, не выносимых на лекции и практические занятия	[Доп. лит., 3] Гл. 1-11, с. 22-335; [Осн. лит., 4], с. 35-495	8	60

6. Фонд оценочных средств для проведения текущего контроля успеваемости, промежуточной аттестации

Контроль качества освоения образовательной программы осуществляется в соответствии с Положением о балльно-рейтинговой системе оценивания результатов учебной деятельности обучающихся.

6.1. Контрольные мероприятия (КМ)

№ КМ	Се-местр	Вид контроля	Название контрольного мероприятия	Вес	Макс. балл	Порядок начисления баллов	Учитывается в ПА
1	8	Текущий контроль	Минитест 1: большие данные	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	экзамен
2	8	Текущий контроль	Минитест 2: Hadoop	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	экзамен
3	8	Текущий контроль	Минитест 3: HDFS	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	экзамен
4	8	Текущий контроль	Минитест 4: MapReduce	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за	экзамен

						тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	
5	8	Текущий контроль	Минитест 5: Pig и Hive	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	экзамен
6	8	Текущий контроль	Минитест 6: HBase	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	экзамен
7	8	Текущий контроль	Минитест 7: YARN	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	экзамен
8	8	Текущий контроль	Минитест 8: Apache Kafka, Zookeeper	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	экзамен
9	8	Текущий контроль	Минитест 9: Apache Spark	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за	экзамен

						тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	
10	8	Текущий контроль	Минитест 10: анализ и визуализация данных	2	5	Минитест проводится в виде электронного теста в конце лекционного занятия. Тест содержит 5 вопросов, за каждый из которых можно получить максимум 1 балл. Студент получает 1 балл за вопрос, если ответ полностью верный, 0 баллов - иначе. Оценка студента за тест - это сумма баллов за каждый вопрос. Время, отведенное на опрос, 10 минут.	экзамен
11	8	Текущий контроль	ПЗ 1. Установка Hadoop. Работа с HDFS	16	1	1 балл: задание полностью выполнено 0 баллов: задание не выполнено	экзамен
12	8	Текущий контроль	ПЗ 2. Разработка MapReduce-приложения	16	1	1 балл: задание полностью выполнено 0 баллов: задание не выполнено	экзамен
13	8	Текущий контроль	ПЗ 3. Разработка статистических отчетов с использованием Apache Hive	16	3	3 балла: реализованы все отчеты. 2 балла: реализованы не все, но более половины отчетов. 0 баллов: задание не выполнено	экзамен
14	8	Текущий контроль	ПЗ 4. Разработка приложения в Apache Spark	16	1	1 балл: задание полностью выполнено 0 баллов: задание не выполнено	экзамен
15	8	Текущий контроль	ПЗ 5. Разработка приложения для анализа и визуализации данных в Hadoop	16	1	1 балл: задание полностью выполнено 0 баллов: задание не выполнено	экзамен
16	8	Промежуточная аттестация	Итоговое тестирование	-	100	Промежуточная аттестация включает компьютерное тестирование. Контрольное мероприятие промежуточной аттестации проводится во время экзамена. Тест состоит из 20 случайных равноценных вопросов, позволяющих оценить сформированность компетенций. На ответы отводится 1 час. На экзамене происходит оценивание учебной деятельности обучающихся по дисциплине на основе полученных оценок за контрольно-рейтинговые мероприятия текущего контроля и промежуточной аттестации. При оценивании результатов учебной деятельности обучающегося по дисциплине используется балльно-рейтинговая система оценивания результатов учебной деятельности обучающихся (утверждена приказом ректора от 24.05.2019 г. № 179)	экзамен

					Отлично: Величина рейтинга обучающегося по дисциплине 85...100 % Хорошо: Величина рейтинга обучающегося по дисциплине 75...84 % Удовлетворительно: Величина рейтинга обучающегося по дисциплине 60...74 % Неудовлетворительно: Величина рейтинга обучающегося по дисциплине 0...59 %	
--	--	--	--	--	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--

6.2. Процедура проведения, критерии оценивания

Вид промежуточной аттестации	Процедура проведения	Критерии оценивания
экзамен	<p>При оценивании результатов учебной деятельности обучающегося по дисциплине используется балльно-рейтинговая система оценивания результатов учебной деятельности обучающихся (Положение о БРС утверждено приказом ректора от 24.05.2019 г. № 179, в редакции приказа ректора от 10.03.2022 г. № 25-13/09). Оценка за дисциплину формируется на основе полученных оценок за контрольно-рейтинговые мероприятия текущего контроля. Отлично: Величина рейтинга обучающегося по дисциплине 85...100 %. Хорошо: Величина рейтинга обучающегося по дисциплине 75...84 %. Удовлетворительно: Величина рейтинга обучающегося по дисциплине 60...74 %. Неудовлетворительно: Величина рейтинга обучающегося по дисциплине 0...59 %.</p> <p>Если студент не согласен с оценкой, полученной по результатам текущего контроля, студент проходит мероприятие промежуточной аттестации в виде тестирования. Тестирование проводится в системе edu.susu.ru. Тест содержит 20 вопросов. На выполнение теста дается 60 минут. В этом случае оценка за дисциплину рассчитывается на основе полученных оценок за контрольно-рейтинговые мероприятия текущего контроля и промежуточной аттестации. Фиксация результатов учебной деятельности по дисциплине проводится в день экзамена при личном присутствии студента.</p>	В соответствии с пп. 2.5, 2.6 Положения

6.3. Паспорт фонда оценочных средств

Компетенции	Результаты обучения	№ KM															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ПК-2	Знает: основы работы компонентов экосистемы Nadoop	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ПК-2	Умеет: строить программную систему на основе компонентов экосистемы Nadoop для решения поставленной задачи	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ПК-2	Имеет практический опыт: создания программной системы на основе компонентов экосистемы Nadoop	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ПК-7	Знает: ПК-8.1. 3-2. Знает принципы работы	+								+	+				+	+	+

		форме	
1	Основная литература	Электронно-библиотечная система издательства Лань	Григорьев, Ю. А. Реляционные базы данных и системы NoSQL : учебное пособие / Ю. А. Григорьев, А. Д. Плутенко, О. Ю. Плужникова. — Благовещенск : АмГУ, 2018. — 424 с. — ISBN 978-5-93493-308-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/156492
2	Основная литература	Электронно-библиотечная система издательства Лань	Осипов, Д. Л. Технологии проектирования баз данных / Д. Л. Осипов. — Москва : ДМК Пресс, 2019. — 498 с. — ISBN 978-5-97060-737-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131692
3	Дополнительная литература	Электронно-библиотечная система издательства Лань	Маккинни, У. Python и анализ данных / У. Маккинни ; перевод с английского А. А. Слинкина. — 2-ое изд., испр. и доп. — Москва : ДМК Пресс, 2020. — 540 с. — ISBN 978-5-97060-590-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131721
4	Основная литература	Электронно-библиотечная система издательства Лань	Перрен, Ж. -. Spark в действии / Ж. -. Перрен ; перевод с английского А. В. Снастина. — Москва : ДМК Пресс, 2021. — 636 с. — ISBN 978-5-97060-879-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/241001
5	Дополнительная литература	Электронно-библиотечная система издательства Лань	Бутаков, Н. А. Обработка больших данных с Apache Spark : учебно-методическое пособие / Н. А. Бутаков, М. В. Петров, Д. Насонов. — Санкт-Петербург : НИУ ИТМО, 2019. — 50 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/136573

Перечень используемого программного обеспечения:

1. РСК Технологии-Система "Персональный виртуальный компьютер" (ПВК) (MS Windows, MS Office, открытое ПО)(бессрочно)

Перечень используемых профессиональных баз данных и информационных справочных систем:

Нет

8. Материально-техническое обеспечение дисциплины

Вид занятий	№ ауд.	Основное оборудование, стенды, макеты, компьютерная техника, предустановленное программное обеспечение, используемое для различных видов занятий
Лекции		Мультимедийный проектор
Экзамен		Компьютерный класс или ПВК-класс
Практические занятия и семинары		ПВК-класс